Yucky Genome Annotation File

Feature 1- Stop 547

Glimmer/GeneMark

What feature number is this? 1

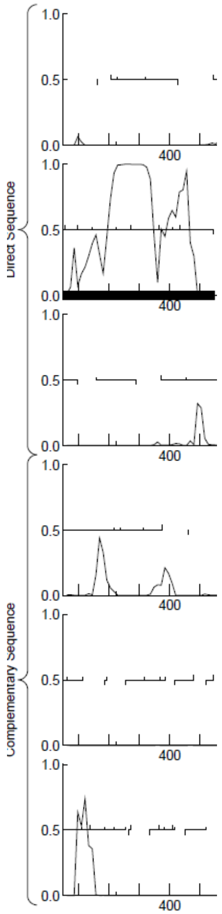
What is the stop site? 547

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? both

What is the autoannotated start? 98

Gap: ____N/A____ or overlap: ___N/A____ (with gene in front of it) for the autoannotated start

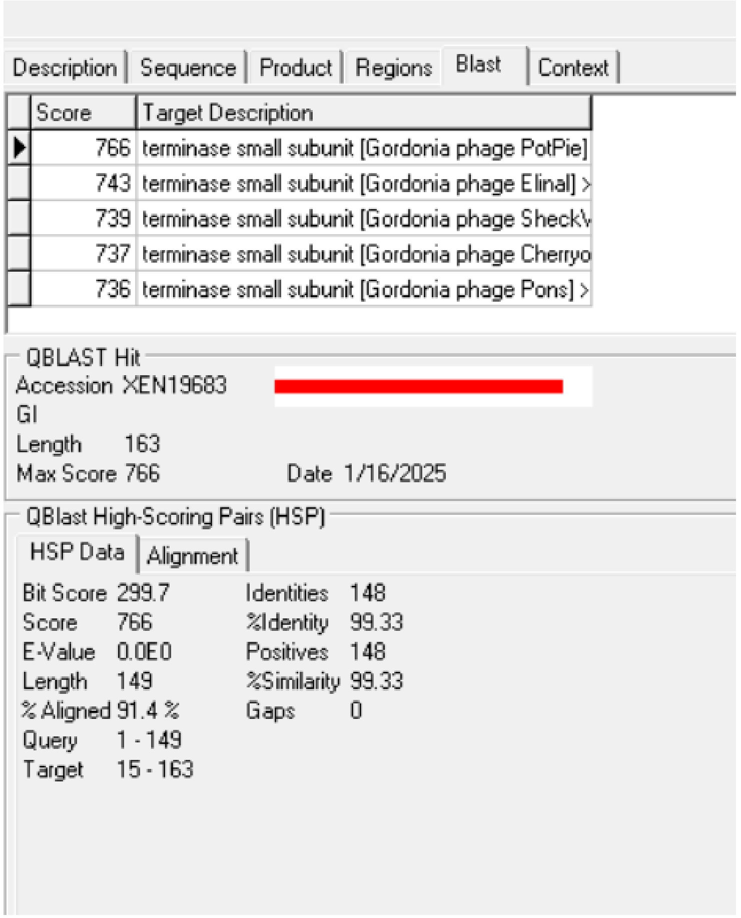
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Strong coding potential through about half of the feature in reading frame two with some dips, particular at the beginning.

- Some coding potential, particular in frames -1 and -3, but not enough to overtake coding potential in reading frame 2

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
766	terminase small subunit [Gordonia phage PotPie]
743	terminase small subunit [Gordonia phage Elinal] >
739	terminase small subunit [Gordonia phage SheckV
737	terminase small subunit [Gordonia phage Cherryo
736	terminase small subunit [Gordonia phage Pons] >

QBLAST Hit

Accession XEN19683

GI

Length 163

Max Score 766

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

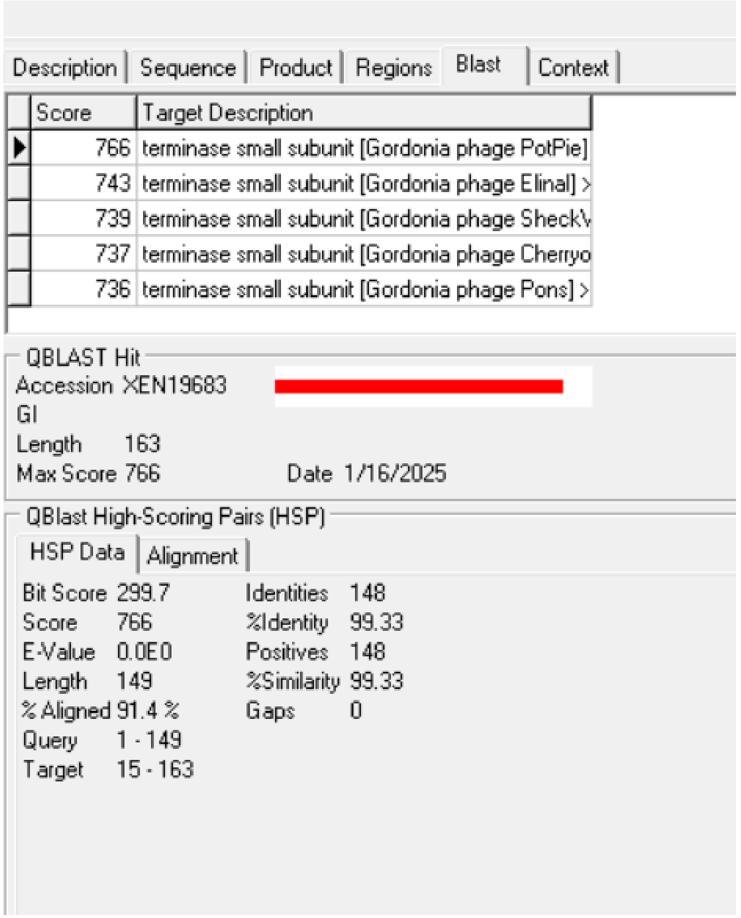
Bit Score	299.7	Identities	148
Score	766	%Identity	99.33
E-Value	0.0E0	Positives	148
Length	149	%Similarity	99.33
% Aligned	91.4 %	Gaps	0
Query	1 - 149		
Target	15 - 163		

- 25 other highly similar genes with E-values close to zero

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, it is a gene. Called by both glimmer and genemark, strong coding potential and many similar matches in BLAST

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
766	terminase small subunit [Gordonia phage PotPie]
743	terminase small subunit [Gordonia phage Elinal] >
739	terminase small subunit [Gordonia phage SheckV
737	terminase small subunit [Gordonia phage Cherryo
736	terminase small subunit [Gordonia phage Pons] >

QBLAST Hit

Accession XEN19683

GI

Length 163

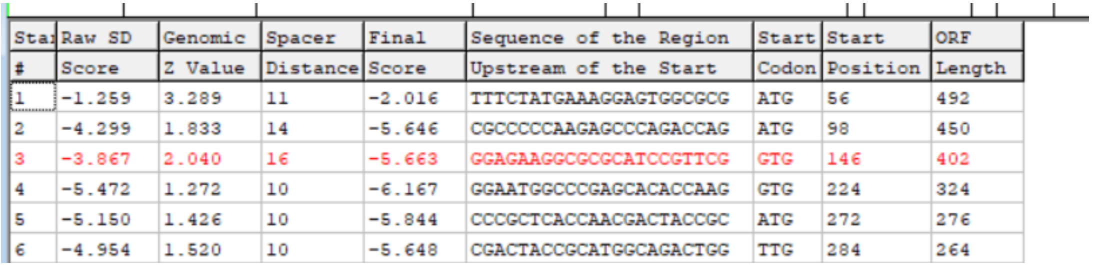
Max Score 766 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	299.7	Identities	148
Score	766	%Identity	99.33
E-Value	0.0E0	Positives	148
Length	149	%Similarity	99.33
% Aligned	91.4 %	Gaps	0
Query	1 - 149		
Target	15 - 163		

- 3 1:1 hits for start at start 98
- >12 1:1 hits at start 56
- No info available for hits starting at 2 – not a location of a start according to RBS chart

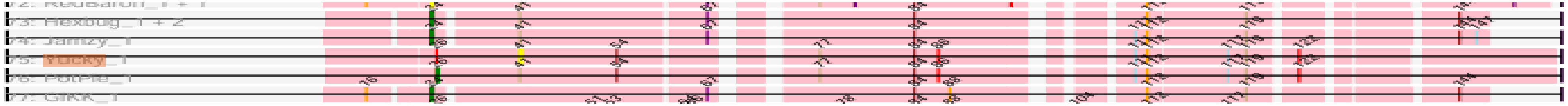
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

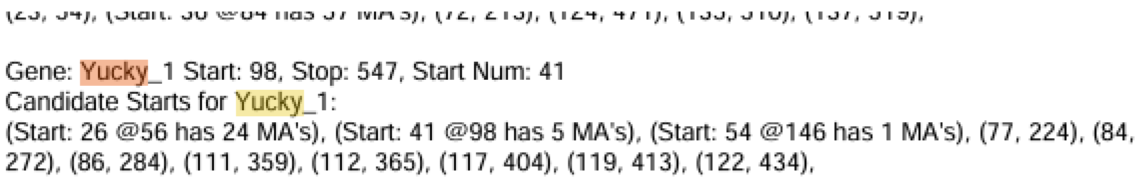


Sta #	Raw SD Score	Genomic Z Value	Spacer Distance	Final Score	Sequence of the Region Upstream of the Start	Start Codon	Start Position	ORF Length
1	-1.259	3.289	11	-2.016	TTTCTATGAAAGGAGTGGCGCG	ATG	56	492
2	-4.299	1.833	14	-5.646	CGCCCCCAAGAGCCCAGACCAG	ATG	98	450
3	-3.867	2.040	16	-5.663	GGAGAAGGCGCGCATCCGTTCG	GTG	146	402
4	-5.472	1.272	10	-6.167	GGAATGGCCCGAGCACACCAAG	GTG	224	324
5	-5.150	1.426	10	-5.844	CCCGCTCACCAACGACTACCGC	ATG	272	276
6	-4.954	1.520	10	-5.648	CGACTACCGCATGGCAGACTGG	TTG	284	264

- Z value for start at 56 is 3.289 with a FS of -2.016
- Z value for start at 98 is 1.833 with a FS of -5.646.
- Z value and final score for start 56 preferred.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

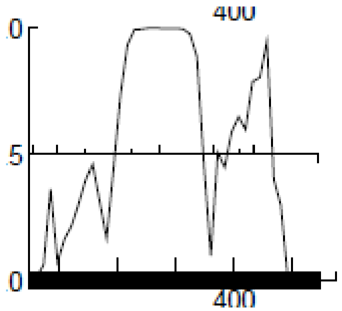



(23, 54), (Start: 38 @84 has 37 MA's), (72, 213), (124, 471), (135, 510), (137, 519),

Gene: Yucky_1 Start: 98, Stop: 547, Start Num: 41
Candidate Starts for Yucky_1:
(Start: 26 @56 has 24 MA's), (Start: 41 @98 has 5 MA's), (Start: 54 @146 has 1 MA's), (77, 224), (84, 272), (86, 284), (111, 359), (112, 365), (117, 404), (119, 413), (122, 434),

- Start at 56 has 24 MA’s while start at 98 has 5, indicating that start at 56 is preferred. In addition, the start at 56 is the first start noted in starterator, maximizing coding potential

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- More coding potential will be cut off at 98 than at 56

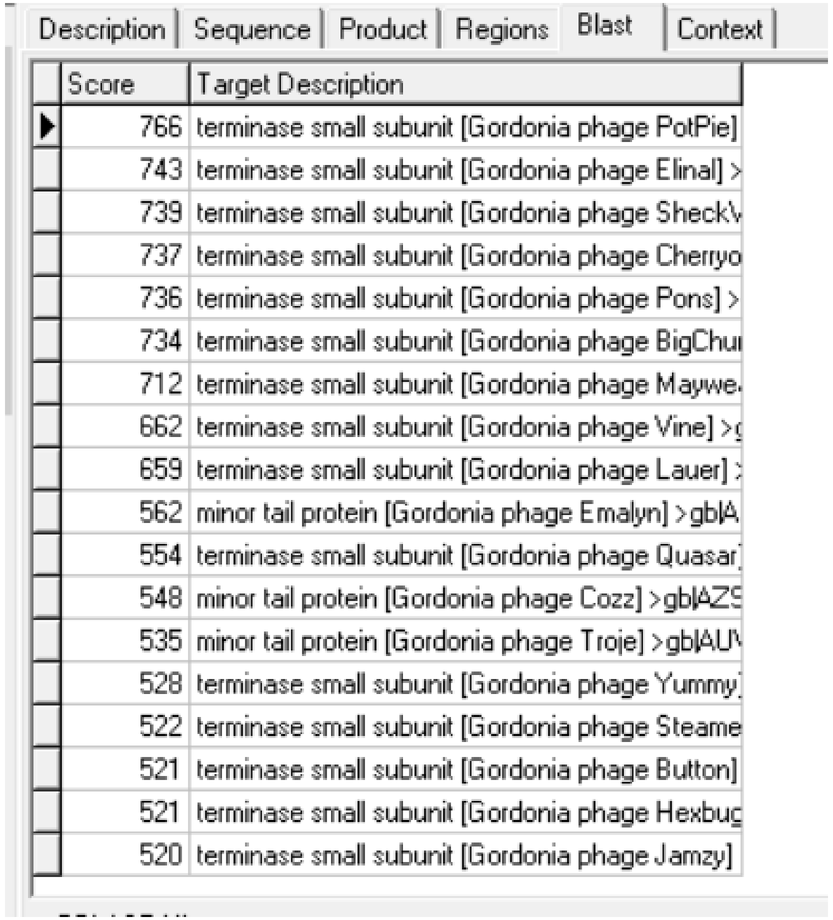
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- This is the first feature, so there is no Gap/Overlap evidence

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- BLAST, coding potential, starterator, and RBS Scores all favor a start at 56. I am calling the start at 56, because it maximizes coding potential, has better RBS scores and also BLAST data favors the 56 start site.

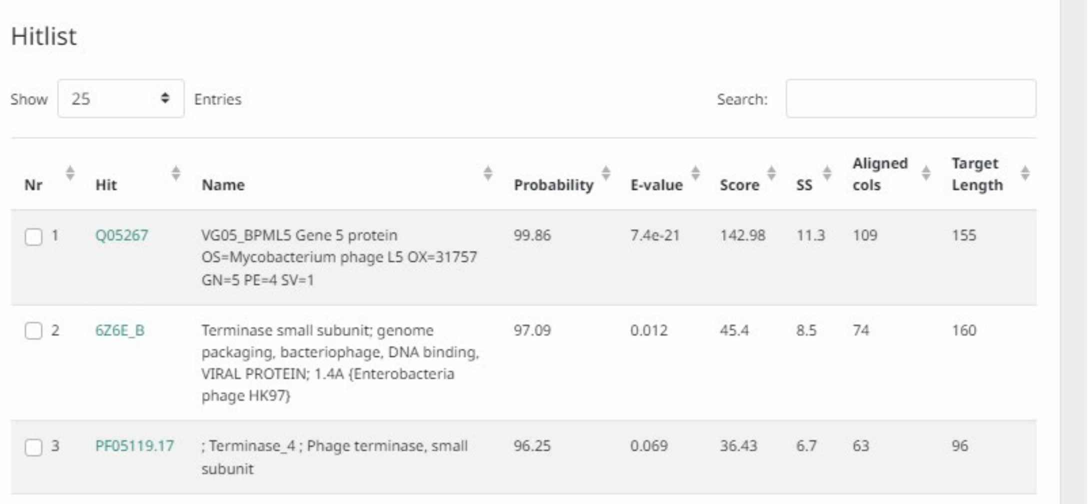
BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
766	terminase small subunit [Gordonia phage PotPie]
743	terminase small subunit [Gordonia phage Elinal] >
739	terminase small subunit [Gordonia phage SheckV
737	terminase small subunit [Gordonia phage Cherryo
736	terminase small subunit [Gordonia phage Pons] >
734	terminase small subunit [Gordonia phage BigChu
712	terminase small subunit [Gordonia phage Maywe
662	terminase small subunit [Gordonia phage Vine] >
659	terminase small subunit [Gordonia phage Lauer]
562	minor tail protein [Gordonia phage Emalyn] >gb|A
554	terminase small subunit [Gordonia phage Quasar]
548	minor tail protein [Gordonia phage Cozz] >gb|AZS
535	minor tail protein [Gordonia phage Troje] >gb|AU
528	terminase small subunit [Gordonia phage Yummy]
522	terminase small subunit [Gordonia phage Steame
521	terminase small subunit [Gordonia phage Button]
521	terminase small subunit [Gordonia phage Hexbug
520	terminase small subunit [Gordonia phage Jamzy]

- Other highly similar genes have assigned functions of terminase small subunit and a few have minor tail protein. Those most closely related (PotPie, Elinal, and SheckWes) have the function of terminase small subunit.

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

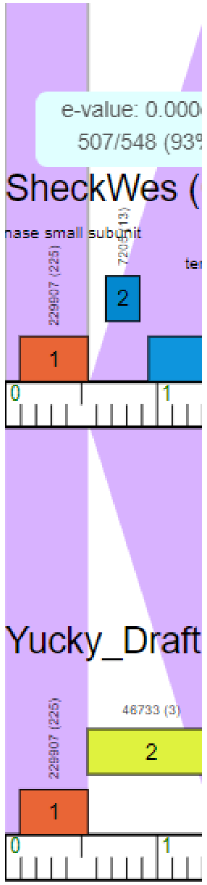


Hitlist

Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	Q05267	VG05_BPML5 Gene 5 protein OS=Mycobacterium phage L5 OX=31757 GN=5 PE=4 SV=1	99.86	7.4e-21	142.98	11.3	109	155
□ 2	6Z6E_B	Terminase small subunit; genome packaging, bacteriophage, DNA binding, VIRAL PROTEIN; 1.4A {Enterobacteria phage HK97}	97.09	0.012	45.4	8.5	74	160
□ 3	PF05119.17	; Terminase_4 ; Phage terminase, small subunit	96.25	0.069	36.43	6.7	63	96

- Several hits greater than 90% probability indicate the function is a terminase small subunit.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- No conserved domain noted in Phamerator, but Feature 1 in Phamerator is in the same pham as those in other phages, including SheckWes, which lists the function as a terminase small subunit.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- I am calling this a terminase small subunit, therefore Deep TMHMM evidence is not applicable.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I am submitting the function as a terminase, small subunit as feature five is identified as a terminase, large subunit. Both HHPRED and BLAST indicate that this is a terminase, small subunit, even though there is no conserved domain indicated in Phamerator.

Feature 2 – Stop 1389

Instructions

Fill this out for each gene you annotate. This should be thought of as the minimum amount of information that needs to be provided for each gene. You can always add more slides or information as necessary

- Is it a gene?
  - Yes!
- Where does it start?
  - 544
- What is the function?
  - PAPS reductase-like domain

- This PowerPoint is for feature 2.

Glimmer/GeneMark

What feature number is this? **DNAM_2**

What is the stop site? **1389**

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? **Glimmer**

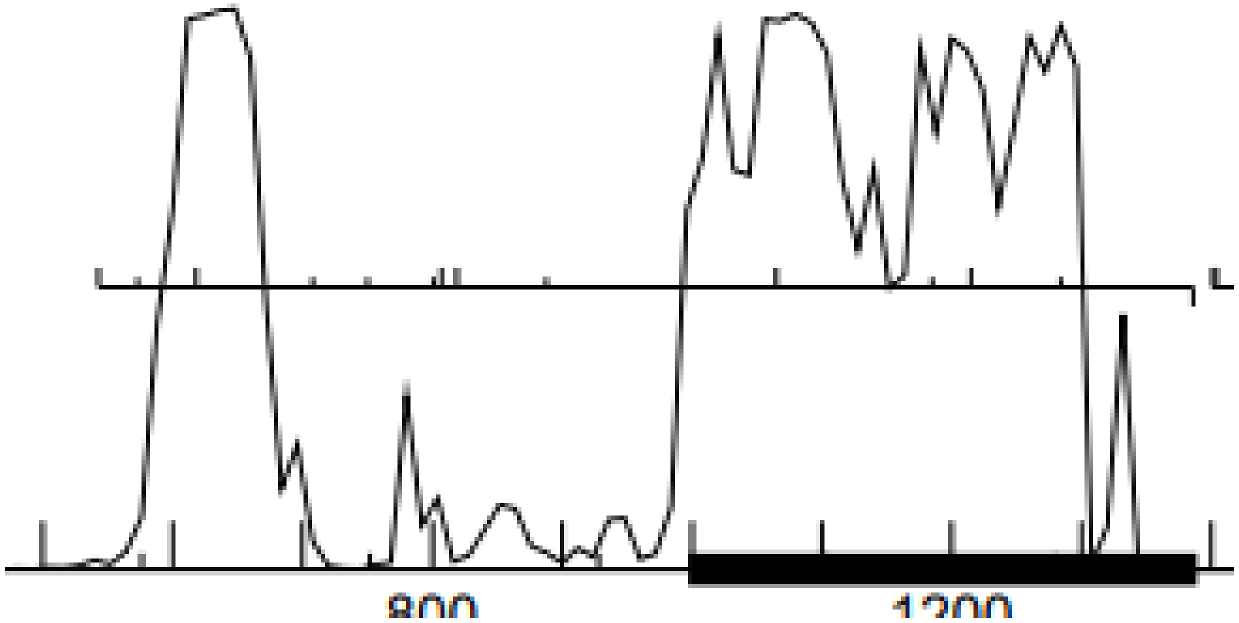
What is the autoannotated start? **544**

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

Overlap from 544-547, there is an overlap of 4 nucleotides

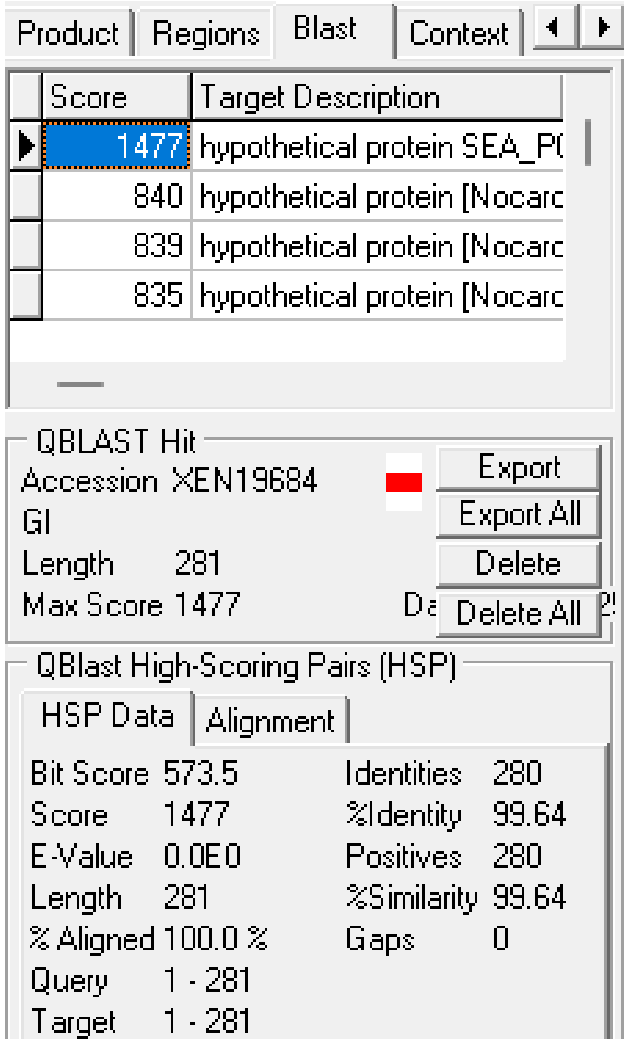
- GeneMark called the gene starting at 997
  - Gap from 547-997, gap of 449 nucleotides

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- GeneMark called the feature running from 997 to 1389
- The GeneMark file shows strong coding potential from around 544 to around 700 where it drops to weak coding potential until around 980 where it increases back to strong until it drops off around 1380.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
1477	hypothetical protein SEA_PC
840	hypothetical protein [Nocarc
839	hypothetical protein [Nocarc
835	hypothetical protein [Nocarc

Accession XEN19684
GI
Length 281
Max Score 1477

Bit Score	573.5	Identities	280
Score	1477	%Identity	99.64
E-Value	0.0E0	Positives	280
Length	281	%Similarity	99.64
% Aligned	100.0 %	Gaps	0
Query	1 - 281		
Target	1 - 281		

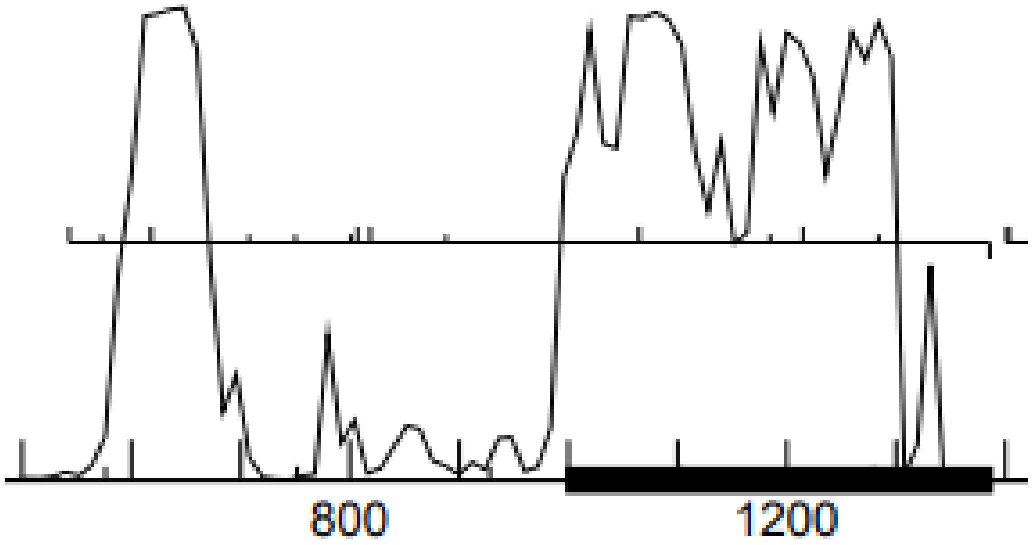
- There were 25 BLAST hits for this feature that all have an e-value of almost zero.
- There was 1 1:1 alignment with SEA_POTPIE_2

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- It is a gene!
- GeneMark shows coding potential running throughout where glimmer and GeneMark shows the feature running. There were also several BLAST hits showing similar features all having e-values close to zero.

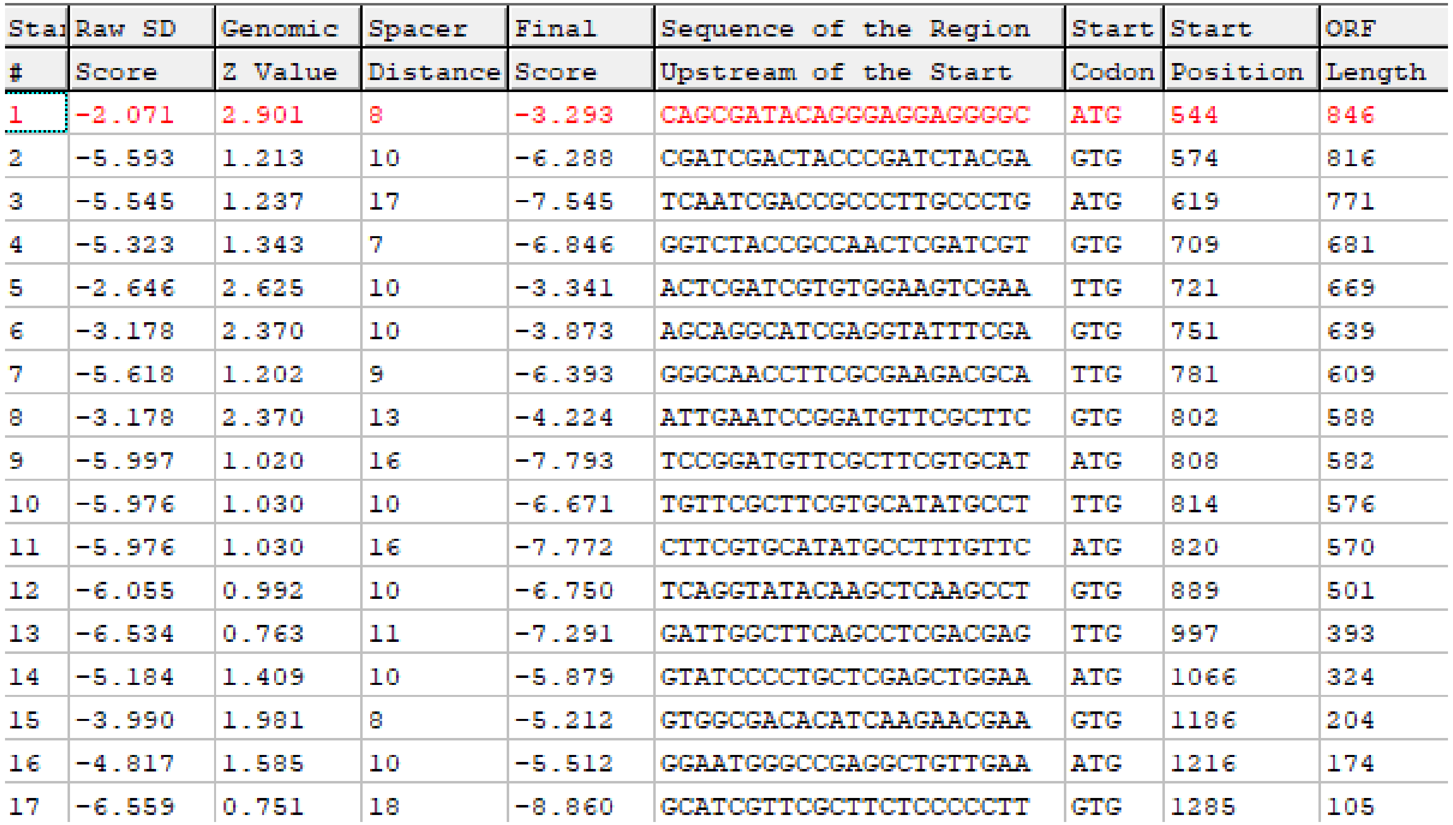
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- GeneMark called the start at 997
- Coding potential starts off strong at 544 then tapers off to weak coding potential around 750 until around 997 where it peaks again to strong coding potential until 1389
- If the start was at 997 then a lot of the coding potential would be cut out, but if it stated at 544 then all of it would be included.



RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

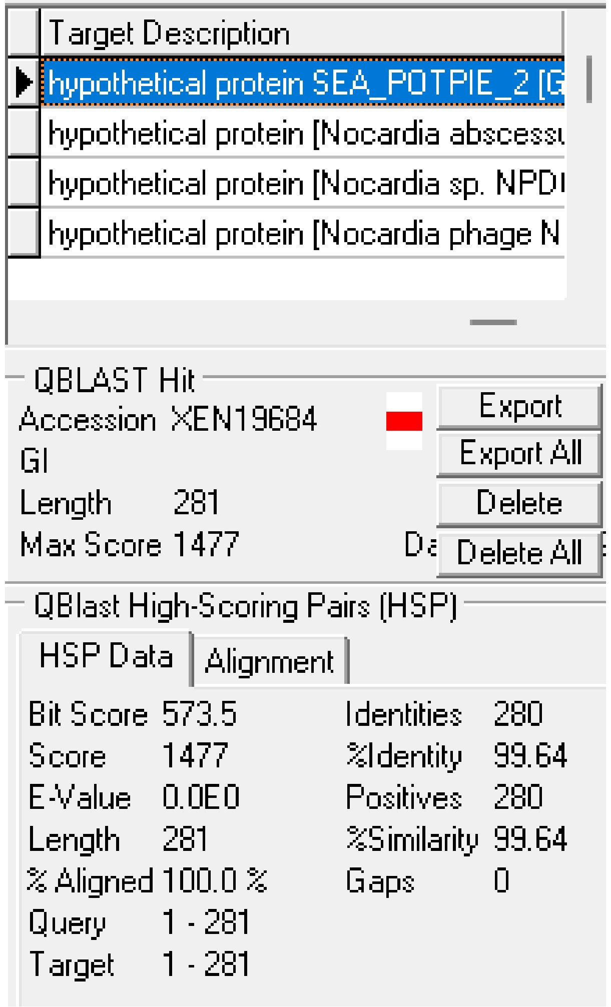
- At 544, the z-value is 2.901 and the final score is -3.293.
- At 997, then z-value if 0.763 and the final score is -7.291
- Based on the RBS values 544 is the favored start



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	-2.071	2.901	8	-3.293	CAGCGATACAGGGAGGAGGGGC	ATG	544	846
2	-5.593	1.213	10	-6.288	CGATCGACTACCCGATCTACGA	GTG	574	816
3	-5.545	1.237	17	-7.545	TCAATCGACCGCCCTTGCCCTG	ATG	619	771
4	-5.323	1.343	7	-6.846	GGTCTACCGCCAACTCGATCGT	GTG	709	681
5	-2.646	2.625	10	-3.341	ACTCGATCGTGTGGAAGTCGAA	TTG	721	669
6	-3.178	2.370	10	-3.873	AGCAGGCATCGAGGTATTTCGA	GTG	751	639
7	-5.618	1.202	9	-6.393	GGGCAACCTTCGCGAAGACGCA	TTG	781	609
8	-3.178	2.370	13	-4.224	ATTGAATCCGGATGTTCGCTTC	GTG	802	588
9	-5.997	1.020	16	-7.793	TCCGGATGTTCGCTTCGTGCAT	ATG	808	582
10	-5.976	1.030	10	-6.671	TGTTCGCTTCGTGCATATGCCT	TTG	814	576
11	-5.976	1.030	16	-7.772	CTTCGTGCATATGCCTTTGTTC	ATG	820	570
12	-6.055	0.992	10	-6.750	TCAGGTATACAAGCTCAAGCCT	GTG	889	501
13	-6.534	0.763	11	-7.291	GATTGGCTTCAGCCTCGACGAG	TTG	997	393
14	-5.184	1.409	10	-5.879	GTATCCCCTGCTCGAGCTGGAA	ATG	1066	324
15	-3.990	1.981	8	-5.212	GTGGCGACACATCAAGAACGAA	GTG	1186	204
16	-4.817	1.585	10	-5.512	GGAATGGGCCGAGGCTGTTGAA	ATG	1216	174
17	-6.559	0.751	18	-8.860	GCATCGTTCGCTTCTCCCCCTT	GTG	1285	105

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There is only 1 1:1 alignment for starting at 544 (PotPie)
- All the BLAST hits have e-values that are close to zero
- At 997 there is a 1:152 alignment



Target Description
hypothetical protein SEA_POTPIE_2 [G
hypothetical protein [Nocardia abscessu
hypothetical protein [Nocardia sp. NPDI
hypothetical protein [Nocardia phage N

QBLAST Hit

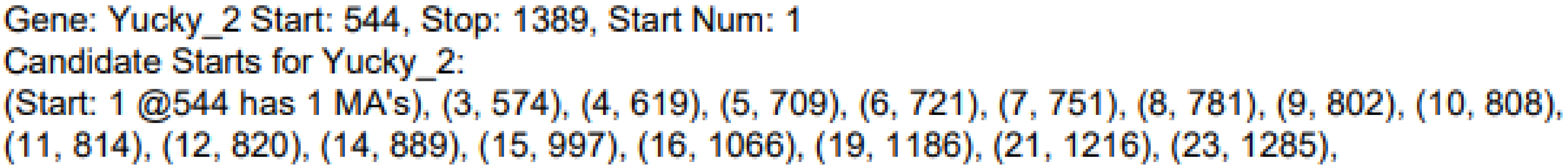
Accession XEN19684 🟥
GI
Length 281
Max Score 1477
Da

QBlast High-Scoring Pairs (HSP)

Bit Score	573.5	Identities	280
Score	1477	%Identity	99.64
E-Value	0.0E0	Positives	280
Length	281	%Similarity	99.64
% Aligned	100.0 %	Gaps	0
Query	1 - 281		
Target	1 - 281		

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Start at 544 has 1 MA (PotPie)
- There are no manual annotations for starting at 997



Gene: Yucky_2 Start: 544, Stop: 1389, Start Num: 1
Candidate Starts for Yucky_2:
(Start: 1 @544 has 1 MA's), (3, 574), (4, 619), (5, 709), (6, 721), (7, 751), (8, 781), (9, 802), (10, 808), (11, 814), (12, 820), (14, 889), (15, 997), (16, 1066), (19, 1186), (21, 1216), (23, 1285),

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Starting at 544:
  - Previous gene ended at 547 and this gene starts theoretically starts at 544
    - Overlaps by 4 nucleotides

- Starting at 997:
  - Previous gene ended at 547 and this gene theoretically starts at 997
    - Gap of 449 nucleotides

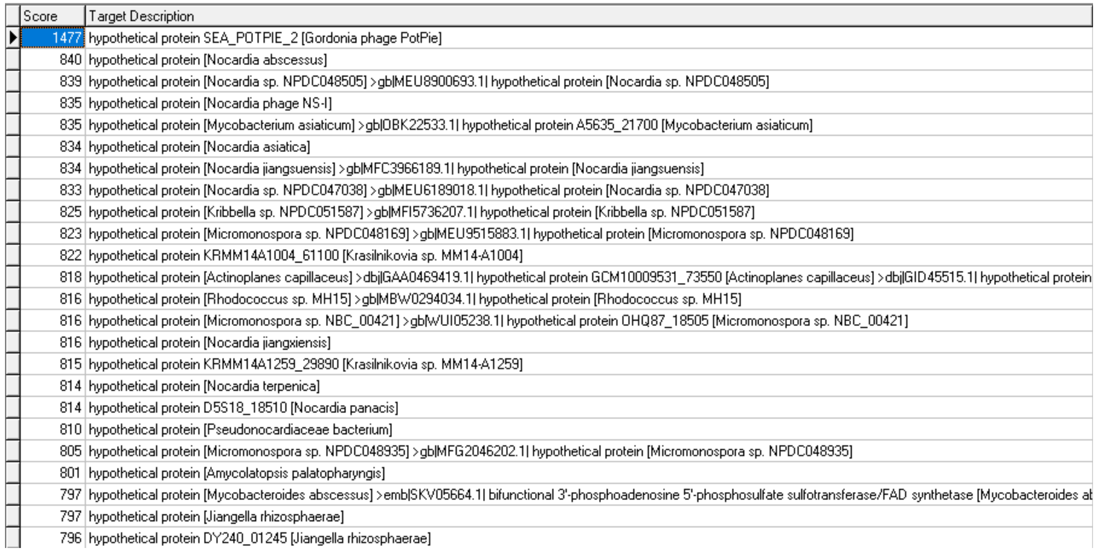
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	Start @ 544	Start @ 997
Glimmer/GeneMark	Glimmer	GeneMark
Coding Potential	Includes all coding potential beginning at a strong peak at 544	Cuts of a large amount of coding potential and instead starts at a strong peak occurring at 997
RBS	Z-value = 2.901 Final score = -3.293	Z-value = 0.763 Final score = -7.291
BLAST	1 1:1 hit with PotPie	1:152 alignment
Starterator	1 MA – PotPie	0 MA
Gap/Overlap	Overlap of 4 nucleotides	Gap of 449 nucleotides

The start site is 544! This start site was called by Glimmer and includes all of the coding potential of the gene. This starting point also has the largest z-value sitting at 2.901 and a final score of -3.293. There was only 1 1:1 alignment on BLAST with PotPie, but the other possible start had no 1:1 alignment. There was one manual annotation according to Starterator for starting at 544 (PotPie). There is an overlap of 4 nucleotides for starting at 544, but this is favorable in comparison to starting at 997 with a gap of 449 nucleotides.

BLAST function evidence. What assigned functions do other highly similar genes have?

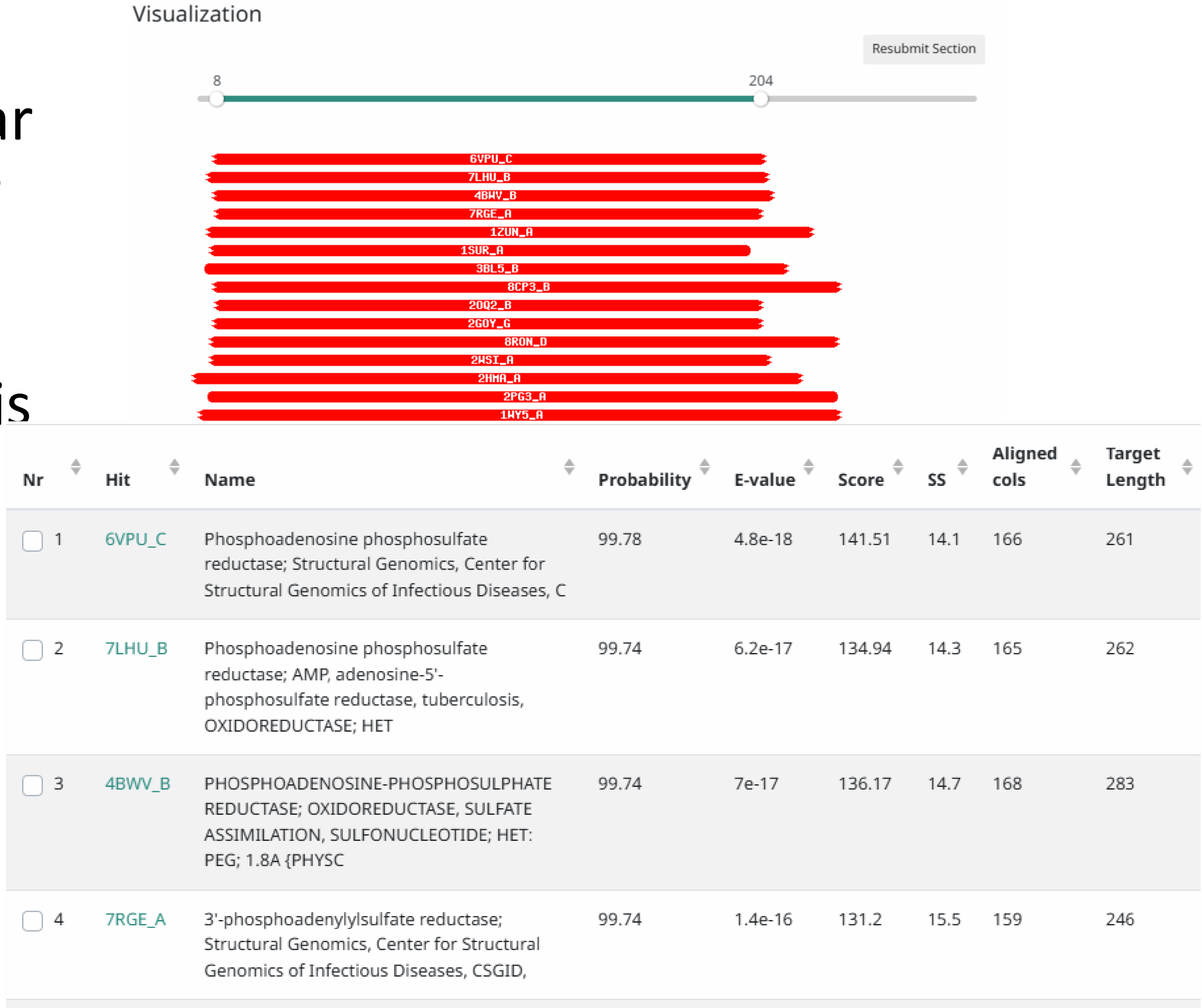
- >17 similar genes have the assigned function of "hypothetical protein"
- The highest match was the phage PotPie which was the only Gordonia phage in the group



Score	Target Description
1477	hypothetical protein SEA_POTPIE_2 [Gordonia phage PotPie]
840	hypothetical protein [Nocardia abscessus]
839	hypothetical protein [Nocardia sp. NPDC048505] >gb|MEU8900693.1| hypothetical protein [Nocardia sp. NPDC048505]
835	hypothetical protein [Nocardia phage NS-I]
835	hypothetical protein [Mycobacterium asiaticum] >gb|OBK22533.1| hypothetical protein A5635_21700 [Mycobacterium asiaticum]
834	hypothetical protein [Nocardia asiatica]
834	hypothetical protein [Nocardia jiangsuensis] >gb|MFC3966189.1| hypothetical protein [Nocardia jiangsuensis]
833	hypothetical protein [Nocardia sp. NPDC047038] >gb|MEU6189018.1| hypothetical protein [Nocardia sp. NPDC047038]
825	hypothetical protein [Kribbella sp. NPDC051587] >gb|MFI5736207.1| hypothetical protein [Kribbella sp. NPDC051587]
823	hypothetical protein [Micromonospora sp. NPDC048169] >gb|MEU9515883.1| hypothetical protein [Micromonospora sp. NPDC048169]
822	hypothetical protein KRMM14A1004_61100 [Krasilnikovia sp. MM14-A1004]
818	hypothetical protein [Actinoplanes capillaceus] >dbj|GAA0469419.1| hypothetical protein GCM10009531_73550 [Actinoplanes capillaceus] >dbj|GID45515.1| hypothetical protein
816	hypothetical protein [Rhodococcus sp. MH15] >gb|MBW0294034.1| hypothetical protein [Rhodococcus sp. MH15]
816	hypothetical protein [Micromonospora sp. NBC_00421] >gb|WUI05238.1| hypothetical protein OHQ87_18505 [Micromonospora sp. NBC_00421]
816	hypothetical protein [Nocardia jiangxiensis]
815	hypothetical protein KRMM14A1259_29890 [Krasilnikovia sp. MM14-A1259]
814	hypothetical protein [Nocardia terpenica]
814	hypothetical protein D5S18_18510 [Nocardia panacis]
810	hypothetical protein [Pseudonocardiaceae bacterium]
805	hypothetical protein [Micromonospora sp. NPDC048935] >gb|MFG2046202.1| hypothetical protein [Micromonospora sp. NPDC048935]
801	hypothetical protein [Amycolatopsis palatopharyngis]
797	hypothetical protein [Mycobacteroides abscessus] >emb|SKV05664.1| bifunctional 3'-phosphoadenosine 5'-phosphosulfate sulfotransferase/FAD synthetase [Mycobacteroides ab
797	hypothetical protein [Jiangella rhizosphaerae]
796	hypothetical protein DY240_01245 [Jiangella rhizosphaerae]

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

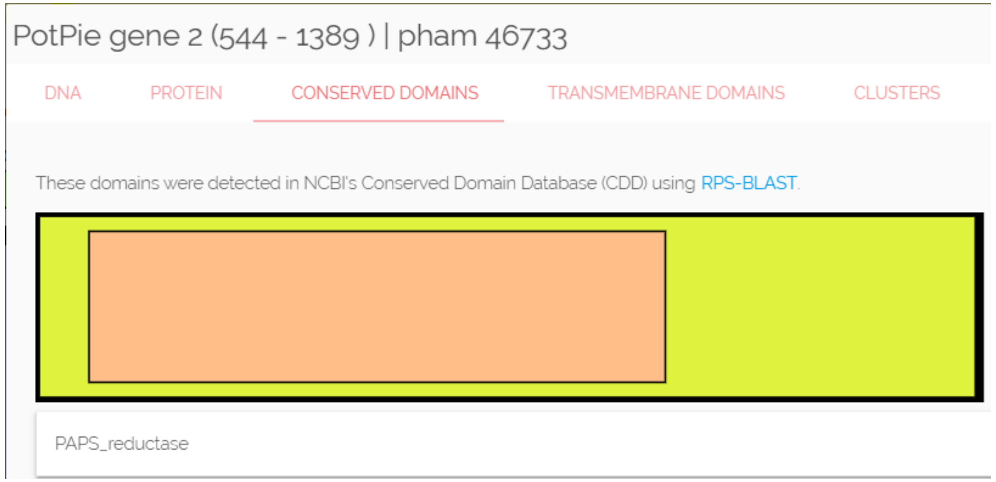
- There were several highly similar matches with probabilities over 99 with a function labeled as phosphoadenosine phosphosulfate reductase that is homologous with about 2/3 of the gene.



	Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
☐	1	6VPU_C	Phosphoadenosine phosphosulfate reductase; Structural Genomics, Center for Structural Genomics of Infectious Diseases, C	99.78	4.8e-18	141.51	14.1	166	261
☐	2	7LHU_B	Phosphoadenosine phosphosulfate reductase; AMP, adenosine-5'-phosphosulfate reductase, tuberculosis, OXIDOREDUCTASE; HET	99.74	6.2e-17	134.94	14.3	165	262
☐	3	4BWV_B	PHOSPHOADENOSINE-PHOSPHOSULPHATE REDUCTASE; OXIDOREDUCTASE, SULFATE ASSIMILATION, SULFONUCLEOTIDE; HET: PEG; 1.8A {PHYSC	99.74	7e-17	136.17	14.7	168	283
☐	4	7RGE_A	3'-phosphoadenylylsulfate reductase; Structural Genomics, Center for Structural Genomics of Infectious Diseases, CSGID,	99.74	1.4e-16	131.2	15.5	159	246

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Phamerator does show phages with genes in the same pham having conserved domains labeled as PAPS-reductase, but there is no labeled function.



PotPie gene 2 (544 - 1389) | pham 46733

These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- Not applicable since it likely has the function of PAPS reductase-like domain

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Official Function List assignment → PAPS reductase-like domain
- The BLAST hits for this gene all show their functions being labeled as hypothetical protein, but upon putting the protein sequence into HHpred several results show up with probabilities over 99 with functions labeled as phosphoadenosine phosphosulfate reductase. Phamerator also shows that phages with genes in the pham having a conserved domain labeled as PAPS-reductase which provides evidence supporting the function of this gene being labeled as a PAPS reductase-like domain.

Removed Reverse Feature with Stop 568

Glimmer/GeneMark

What feature number is this? Removed

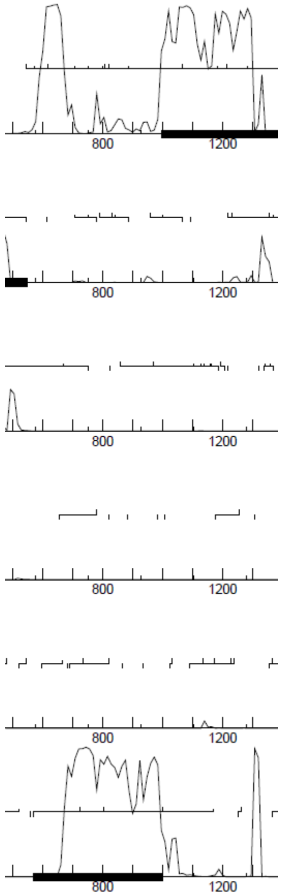
What is the stop site?568 (reverse gene)

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? 1002 Called by Genemark, not called by glimmer

What is the autoannotated start? 1002

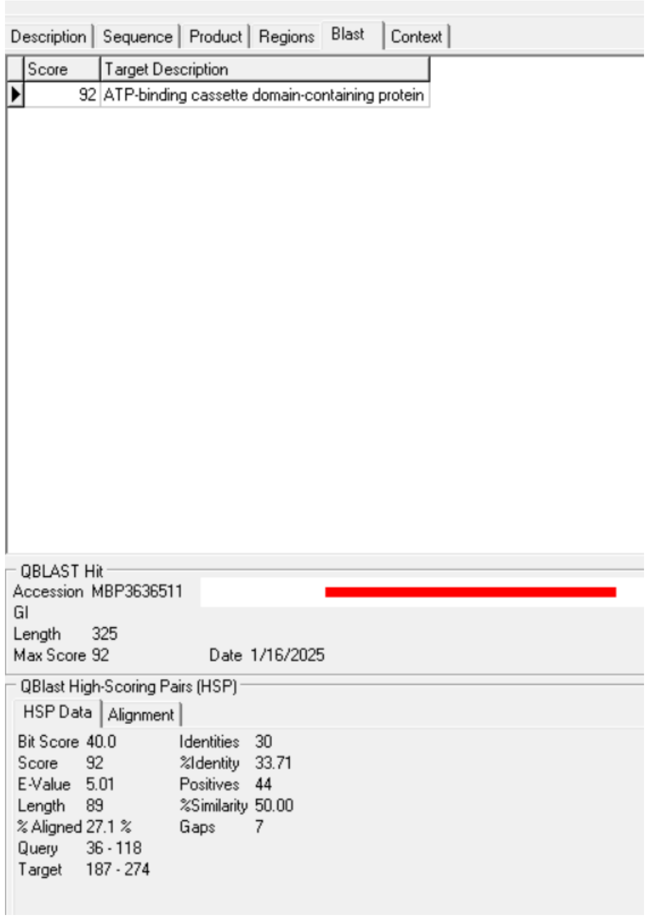
Gap: __496_ with feature 4_ or overlap: _______ (with gene in front of it) for the autoannotated start – However, this feature completely overlaps with feature number 2

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Strong coding potential in reading frame -3, however fully overlaps with coding potential of feature two in reading frame 1

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
92	ATP-binding cassette domain-containing protein

QBLAST Hit

Accession MBP3636511

GI

Length 325

Max Score 92 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	40.0	Identities	30
Score	92	%Identity	33.71
E-Value	5.01	Positives	44
Length	89	%Similarity	50.00
% Aligned	27.1 %	Gaps	7
Query	36 - 118		
Target	187 - 274		

- Only one BLAST hit with an e-value of 5. No close matches with e-values close to zero.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- No, this isn’t a gene. The feature stands alone as a reverse gene, which does not agree with guiding principles. There are no close matches in BLAST. Even though it is called by Genemark, it is not called by Glimmer. In addition, it completely overlaps with feature 2, which is in reading frame 1.

Feature 3 – stop 1675

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

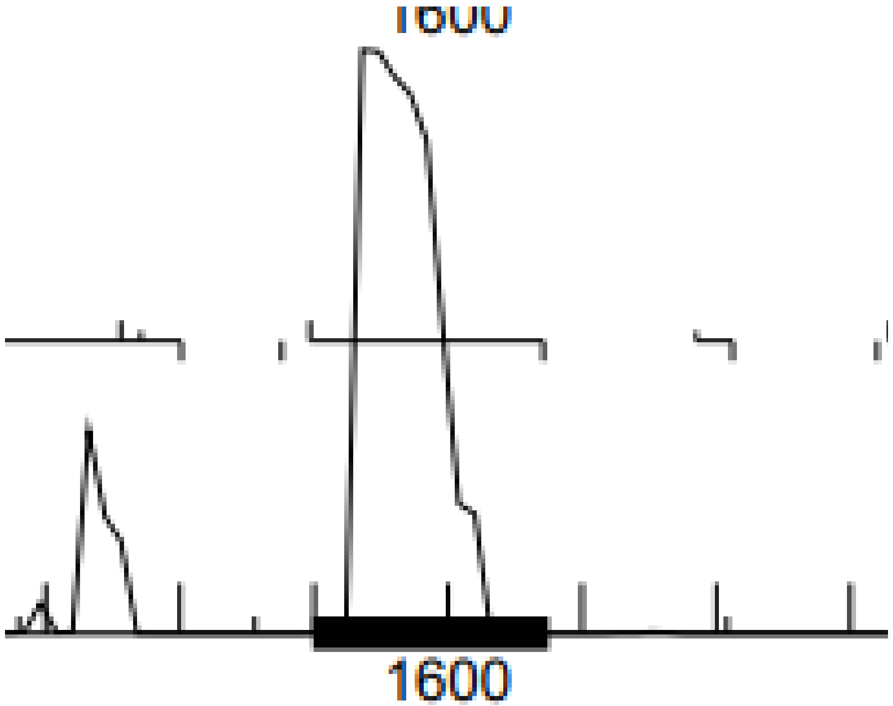
What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

- Feature: 3
- Stop site: 1675
- Called by both Glimmer and GeneMark
- Autonannotated start: 1499
- Gap: 109

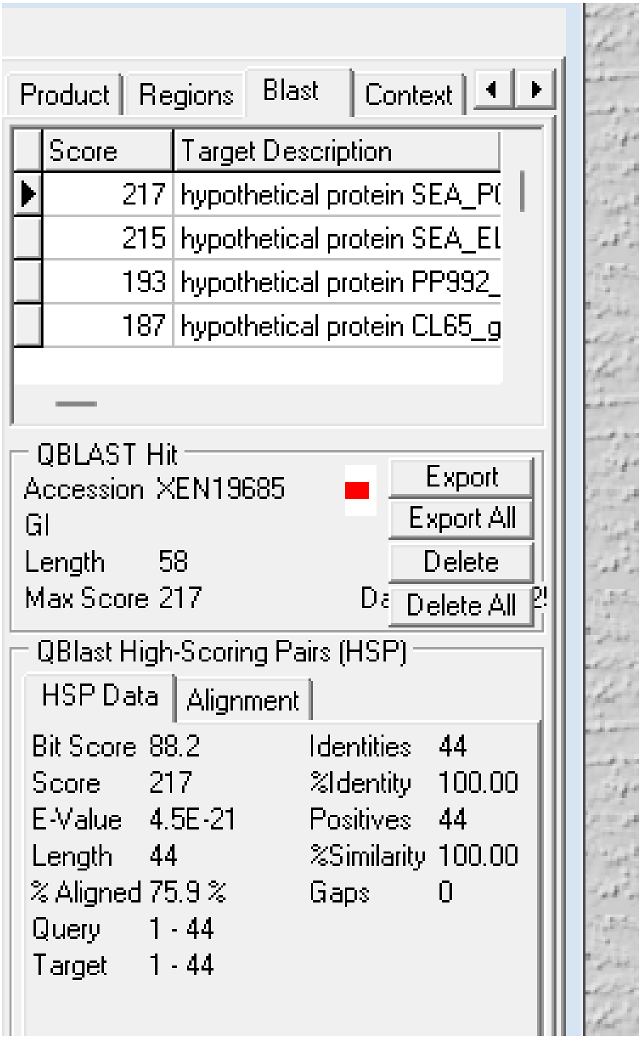
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Coding potential found in frame 2
- Not the only frame with coding potential
- Includes all coding potential at start site 1499



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- Has 25 highly similar genes
- Anything smaller than E-7 is what we want to include as a similar gene



Score	Target Description
217	hypothetical protein SEA_P(
215	hypothetical protein SEA_EL
193	hypothetical protein PP992_
187	hypothetical protein CL65_g

QBLAST Hit

Accession XEN19685 🟥

GI

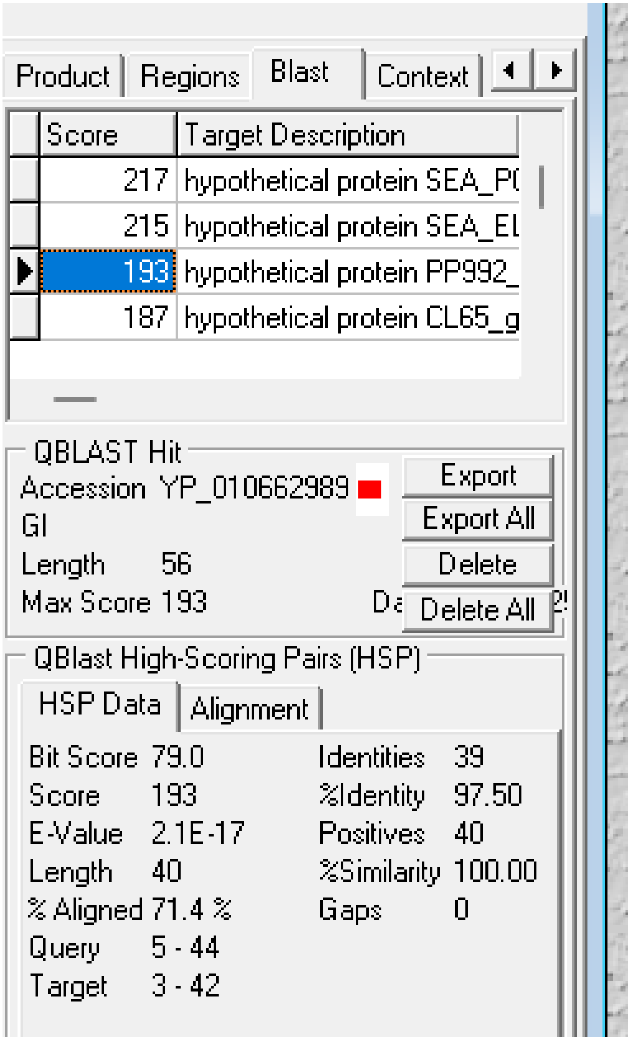
Length 58

Max Score 217

QBlast High-Scoring Pairs (HSP)

HSP Data | Alignment

Bit Score	88.2	Identities	44
Score	217	%Identity	100.00
E-Value	4.5E-21	Positives	44
Length	44	%Similarity	100.00
% Aligned	75.9 %	Gaps	0
Query	1 - 44		
Target	1 - 44		



Score	Target Description
217	hypothetical protein SEA_P(
215	hypothetical protein SEA_EL
193	hypothetical protein PP992_
187	hypothetical protein CL65_g

QBLAST Hit

Accession YP_010662989 🟥

GI

Length 56

Max Score 193

QBlast High-Scoring Pairs (HSP)

HSP Data | Alignment

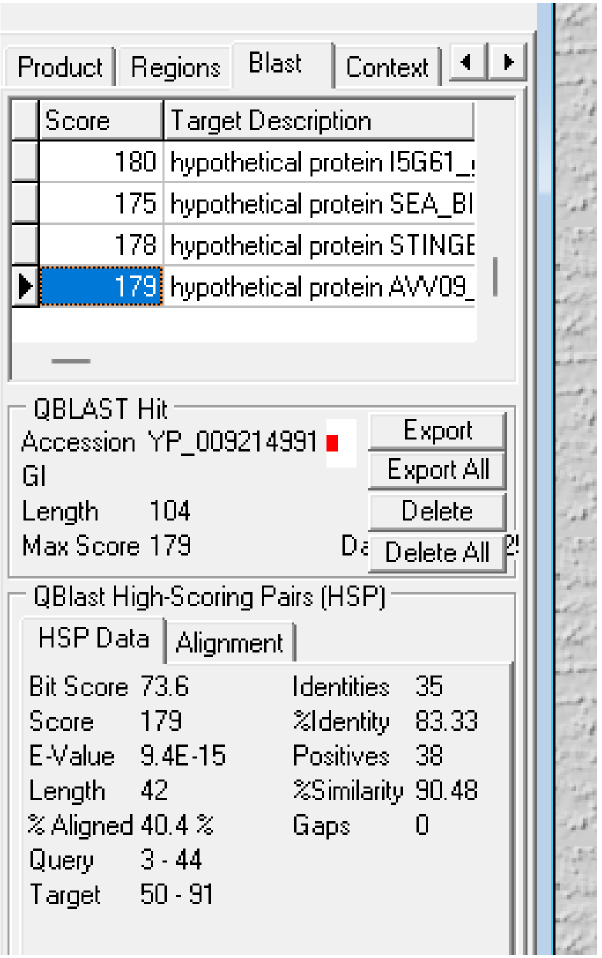
Bit Score	79.0	Identities	39
Score	193	%Identity	97.50
E-Value	2.1E-17	Positives	40
Length	40	%Similarity	100.00
% Aligned	71.4 %	Gaps	0
Query	5 - 44		
Target	3 - 42		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, it is a gene because both Glimmer and GeneMark call it at 1499. The start site at 1499, includes all coding potential, and the BLAST evidence displays 25 highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- Has 0 1:1 alignments



Score	Target Description
180	hypothetical protein I5G61_
175	hypothetical protein SEA_BI
178	hypothetical protein STINGE
179	hypothetical protein AVV09_

QBLAST Hit

Accession YP_009214991

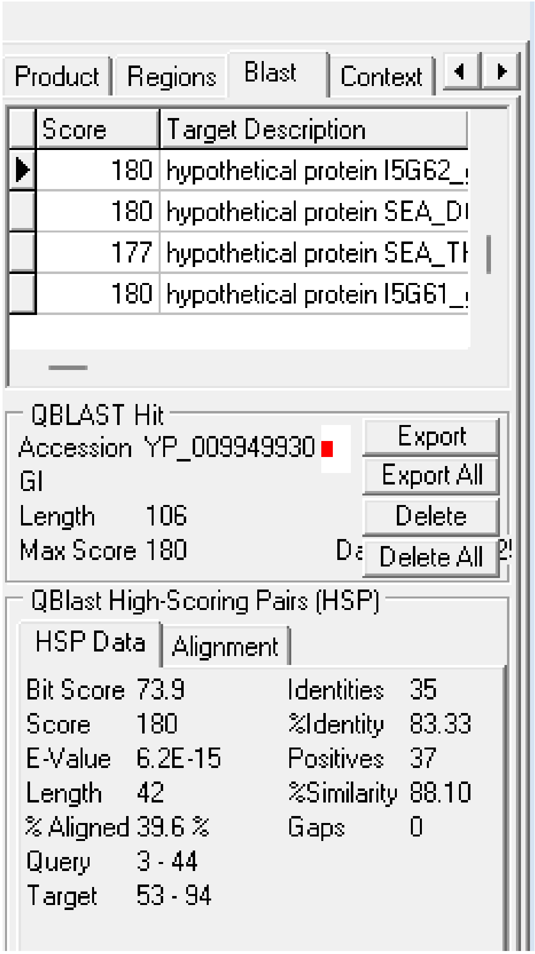
GI

Length 104

Max Score 179

QBlast High-Scoring Pairs (HSP)

Bit Score	73.6	Identities	35
Score	179	%Identity	83.33
E-Value	9.4E-15	Positives	38
Length	42	%Similarity	90.48
% Aligned	40.4 %	Gaps	0
Query	3 - 44		
Target	50 - 91		



Score	Target Description
180	hypothetical protein I5G62_
180	hypothetical protein SEA_DI
177	hypothetical protein SEA_TI
180	hypothetical protein I5G61_

QBLAST Hit

Accession YP_009949930

GI

Length 106

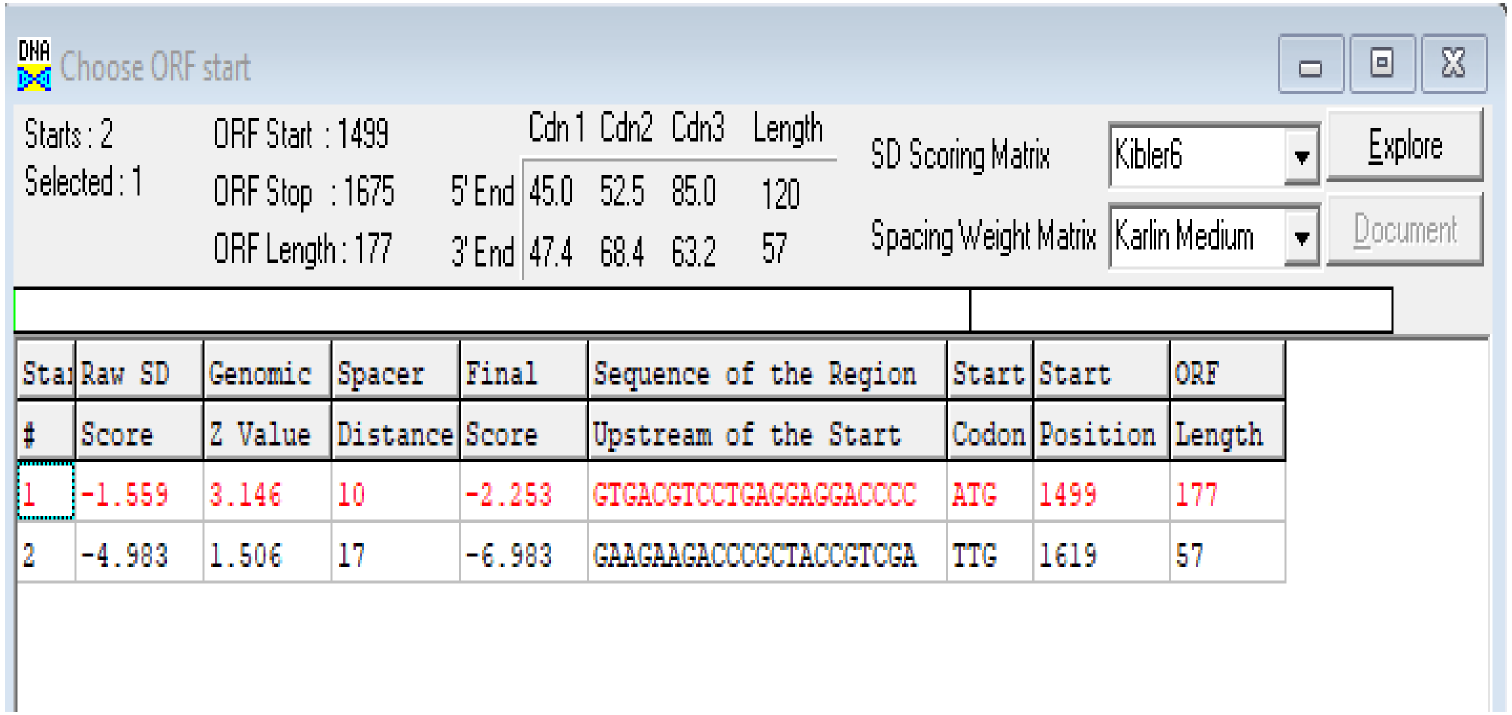
Max Score 180

QBlast High-Scoring Pairs (HSP)

Bit Score	73.9	Identities	35
Score	180	%Identity	83.33
E-Value	6.2E-15	Positives	37
Length	42	%Similarity	88.10
% Aligned	39.6 %	Gaps	0
Query	3 - 44		
Target	53 - 94		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- 1499
- Z value: 3.146
- Final Score: -2.253



Starts : 2
Selected : 1

ORF Start : 1499
ORF Stop : 1675
ORF Length : 177

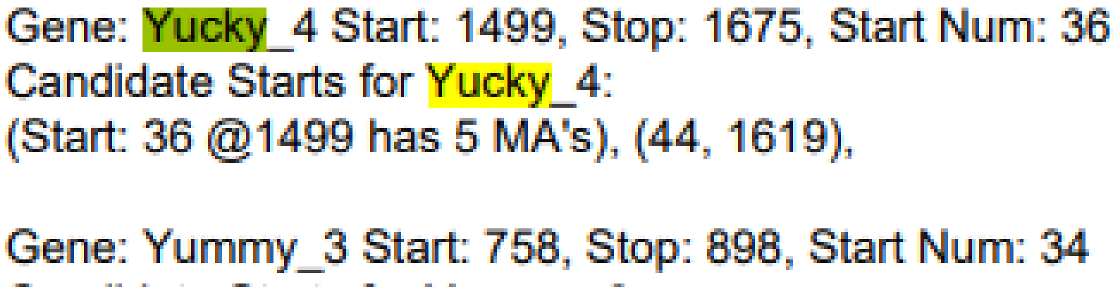
	Cdn 1	Cdn2	Cdn3	Length
5' End	45.0	52.5	85.0	120
3' End	47.4	68.4	63.2	57

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-1.559	3.146	10	-2.253	GTGACGTCCTGAGGAGGACCCC	ATG	1499	177
2	-4.983	1.506	17	-6.983	GAAGAAGACCCGCTACCGTCGA	TTG	1619	57

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

Yucky start 36 @1499 has 5 MA’s

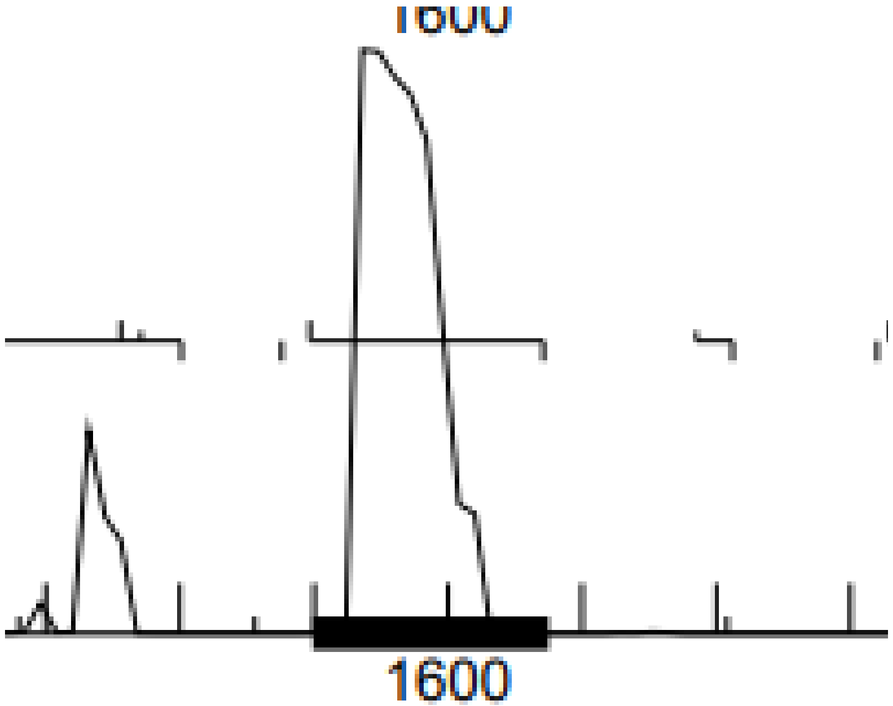


Gene: Yucky_4 Start: 1499, Stop: 1675, Start Num: 36
Candidate Starts for Yucky_4:
(Start: 36 @1499 has 5 MA's), (44, 1619),

Gene: Yummy_3 Start: 758, Stop: 898, Start Num: 34

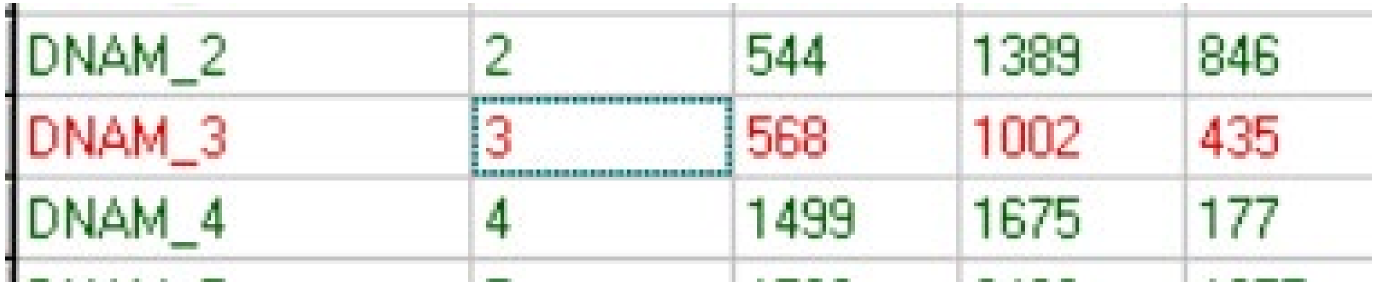
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- At start 1499, none of the coding potential is cut off
- There is no listed alternative start site



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Gap: 1499-1389 (feature 2) = 110-1 =109



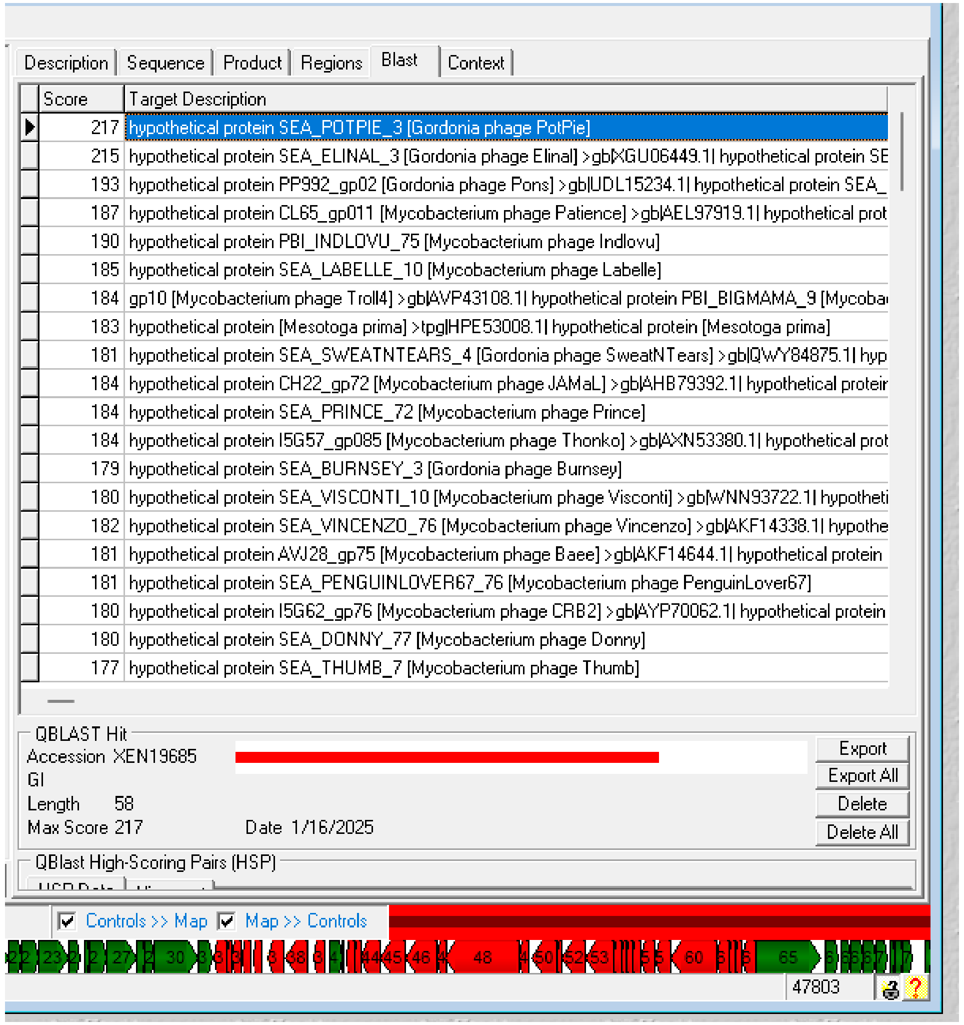
DNAM_2	2	544	1389	846
DNAM_3	3	568	1002	435
DNAM_4	4	1499	1675	177

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	1499
GeneMark	Glimmer & GeneMark
Coding potential	Includes all cp
RBS	Z value: 3.146 Final score: -2.253
BLAST	0 1:1 alignments
Starterator	5 MA’s
Gap	109

Yes, 1499 is the start site because both Glimmer and GeneMark call it. Frame 2 includes all coding potential, and it has a high z value. The start site also includes 5 manual annotations. Starterator evidence did not reveal an alternative start, so the auto annotated start found in the DNAM file is the

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
217	hypothetical protein SEA_POTPIE_3 [Gordonia phage PotPie]
215	hypothetical protein SEA_ELINAL_3 [Gordonia phage Elinal] >gb|XGU06449.1| hypothetical protein SE
193	hypothetical protein PP992_gp02 [Gordonia phage Pons] >gb|UDL15234.1| hypothetical protein SEA_
187	hypothetical protein CL65_gp011 [Mycobacterium phage Patience] >gb|AEL97919.1| hypothetical prot
190	hypothetical protein PBI_INDLOVU_75 [Mycobacterium phage Indlovu]
185	hypothetical protein SEA_LABELLE_10 [Mycobacterium phage Labelle]
184	gp10 [Mycobacterium phage Troll4] >gb|AVP43108.1| hypothetical protein PBI_BIGMAMA_9 [Mycoba
183	hypothetical protein [Mesotoga prima] >tpg|HPE53008.1| hypothetical protein [Mesotoga prima]
181	hypothetical protein SEA_SWEATNTEARS_4 [Gordonia phage SweatNTears] >gb|QWY84875.1| hyp
184	hypothetical protein CH22_gp72 [Mycobacterium phage JAMaL] >gb|AHB79392.1| hypothetical protei
184	hypothetical protein SEA_PRINCE_72 [Mycobacterium phage Prince]
184	hypothetical protein I5G57_gp085 [Mycobacterium phage Thonko] >gb|AXN53380.1| hypothetical prot
179	hypothetical protein SEA_BURNSEY_3 [Gordonia phage Burnsey]
180	hypothetical protein SEA_VISCONTI_10 [Mycobacterium phage Visconti] >gb|WNN93722.1| hypotheti
182	hypothetical protein SEA_VINCENZO_76 [Mycobacterium phage Vincenzo] >gb|AKF14338.1| hypothe
181	hypothetical protein AVJ28_gp75 [Mycobacterium phage Baee] >gb|AKF14644.1| hypothetical protein
181	hypothetical protein SEA_PENGUINLOVER67_76 [Mycobacterium phage PenguinLover67]
180	hypothetical protein I5G62_gp76 [Mycobacterium phage CRB2] >gb|AYP70062.1| hypothetical protein
180	hypothetical protein SEA_DONNY_77 [Mycobacterium phage Donny]
177	hypothetical protein SEA_THUMB_7 [Mycobacterium phage Thumb]

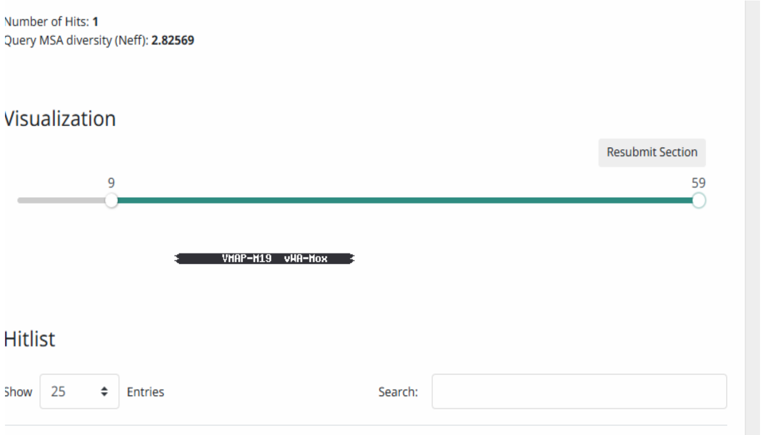
QBLAST Hit
Accession XEN19685
GI
Length 58
Max Score 217
Date 1/16/2025

Has 25 similar genes with assigned function "hypothetical protein"

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

Has 1 alignment
However, that alignment has an 11.55% probability and an E value of 540. The probability should be higher than 90% and have an E-value less than 1 to assign a function

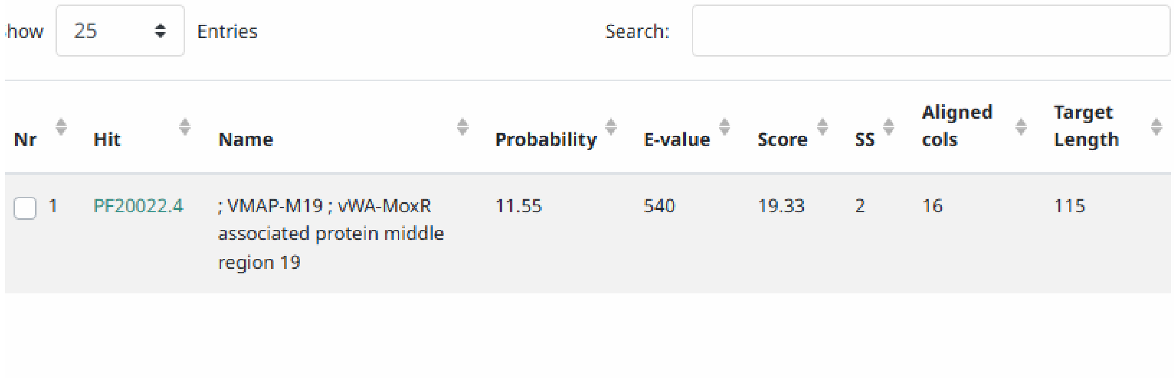
So, the HHpred evidence does not assign a function to Yucky



Number of Hits: 1
Query MSA diversity (Neff): 2.82569

Visualization

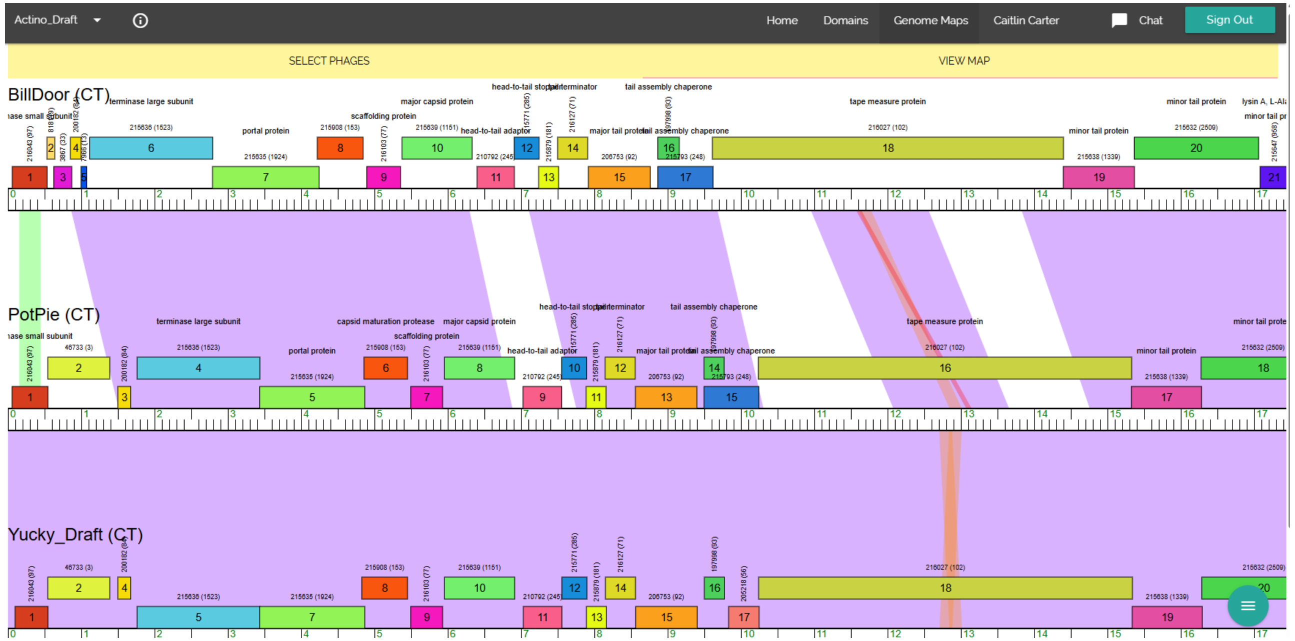
Hitlist

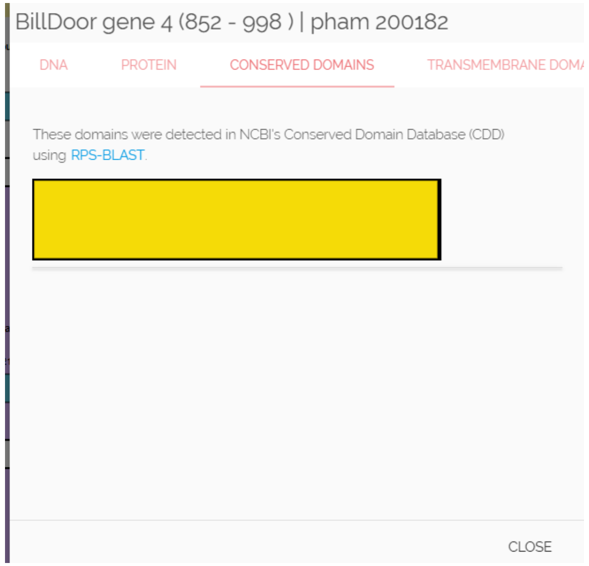


Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	PF20022.4	; VMAP-M19 ; vWA-MoxR associated protein middle region 19	11.55	540	19.33	2	16	115

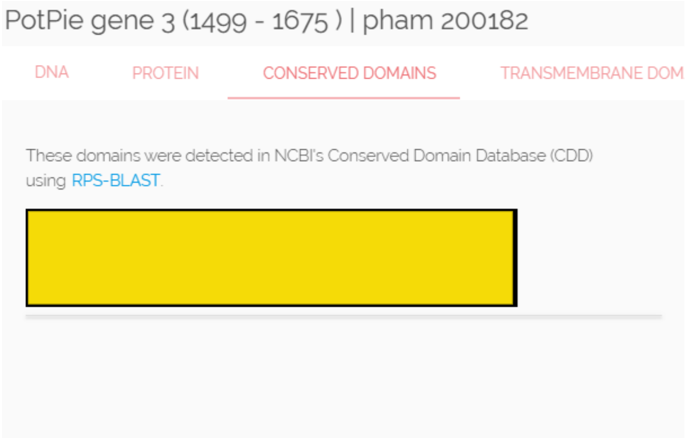
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

Yucky feature 4: No conserved domain and no function
BillDoor Feature 4: Has no function and no conserved domain
PotPie feature 3: Has no function and no conserved domain

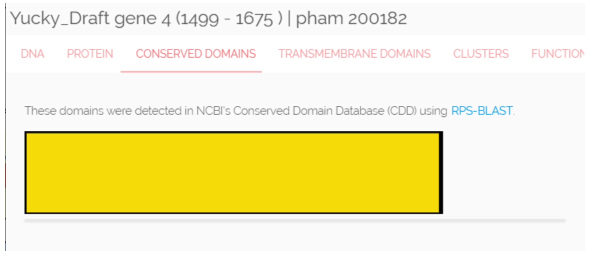



BillDoor gene 4 (852 - 998) | pham 200182

These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.


PotPie gene 3 (1499 - 1675) | pham 200182

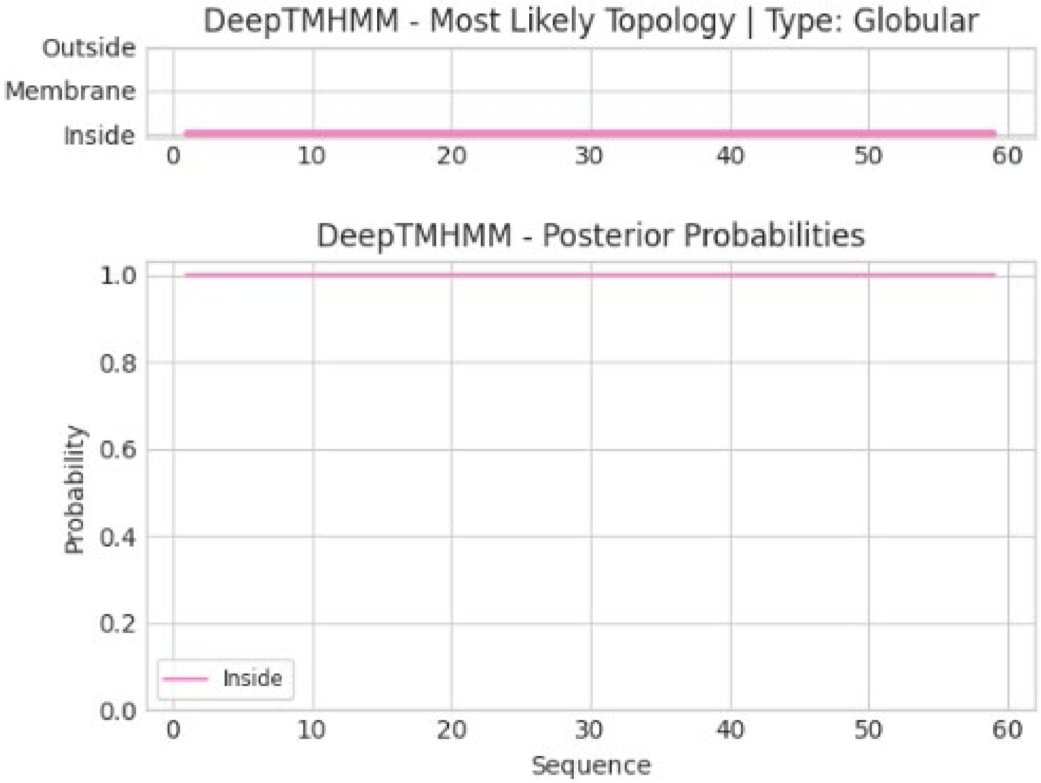
These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.


Yucky_Draft gene 4 (1499 - 1675) | pham 200182

These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

# of Unnamed Number of predicted TMRs: 0



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

There is no function, so it is a hypothetical protein because Hhpred evidence shows 1 alignment. However, that alignment is not considered because it has a low probability and an E value that is not less than 1.

The Phamerator evidence for highly similar genes (PotPie and BillDoor), also have no conserved domain or function assigned to Yucky. The Deep TMHMM evidence has zero Unnamed Number of predicted TMRs.

Feature 4 – Stop 3438

Glimmer/GeneMark

What feature number is this? 4

What is the stop site? 3438

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

It is called by both, but Glimmer and GeneMark disagree.

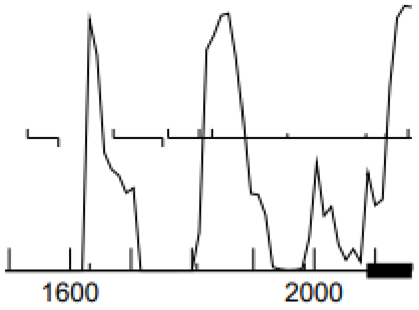
What is the autoannotated start?

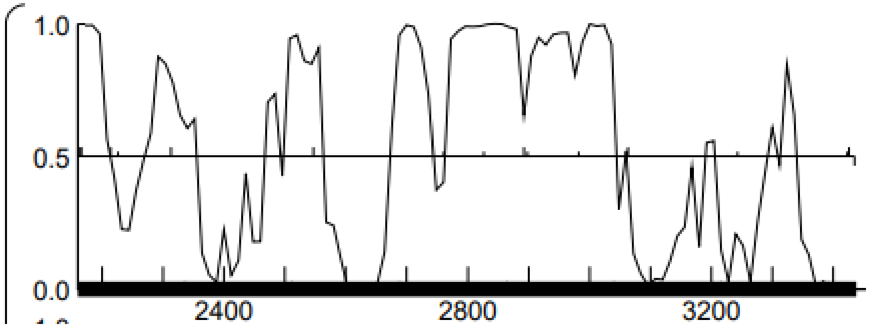
Glimmer called the start site at 1762. GeneMark called it at 2086,

Gap: ___86/410_____ or overlap: _______ (with gene in front of it) for the autoannotated start

- Glimmer and GeneMark disagree on the start site.
- There is no overlap
- Glimmer gap: 86
- GeneMark gap: 410

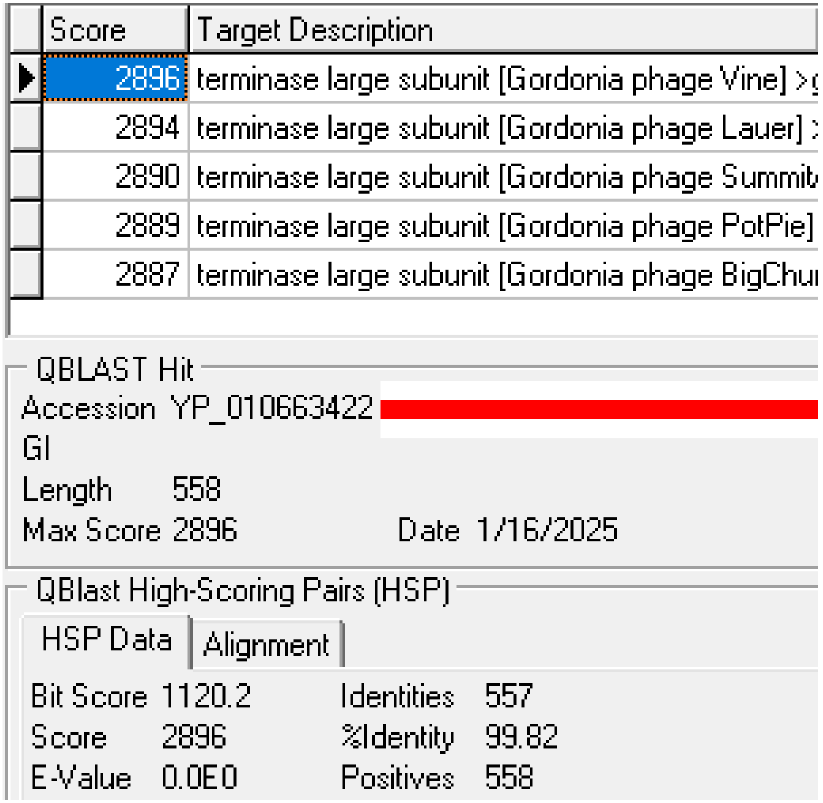
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?





- There is a strong peak at approximately nucleotide 1800, which quickly drops and weakly peaks again around nucleotide 2000. There are many wavering strong and weak peaks throughout the rest of the feature, getting particularly strong and consistent from about nucleotide 2650-3000. There is a peak of coding potential in reading frame 6 as well, which is a reverse reading frame.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
2896	terminase large subunit [Gordonia phage Vine] >(
2894	terminase large subunit [Gordonia phage Lauer] :
2890	terminase large subunit [Gordonia phage Summit
2889	terminase large subunit [Gordonia phage PotPie]
2887	terminase large subunit [Gordonia phage BigChu

QBLAST Hit

Accession YP_010663422

GI

Length 558

Max Score 2896 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

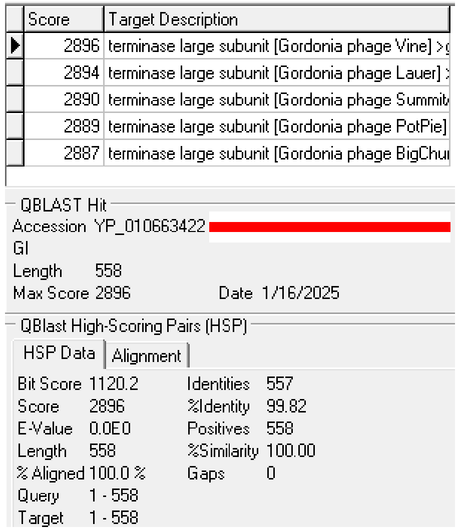
Bit Score	1120.2	Identities	557
Score	2896	%Identity	99.82
E-Value	0.0E0	Positives	558

- There are 24 1:1 alignment hits.
- All 25 close matches have an E-value close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This feature is a gene. Both Glimmer and GeneMark autoannotated it as a gene, despite disagreeing on start site. BLAST found at least 25 close matches containing an E-value close to 0. Lastly, there is a lot of strong peaks in coding potential near the start site, and throughout the sequence of the autoannoted gene.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



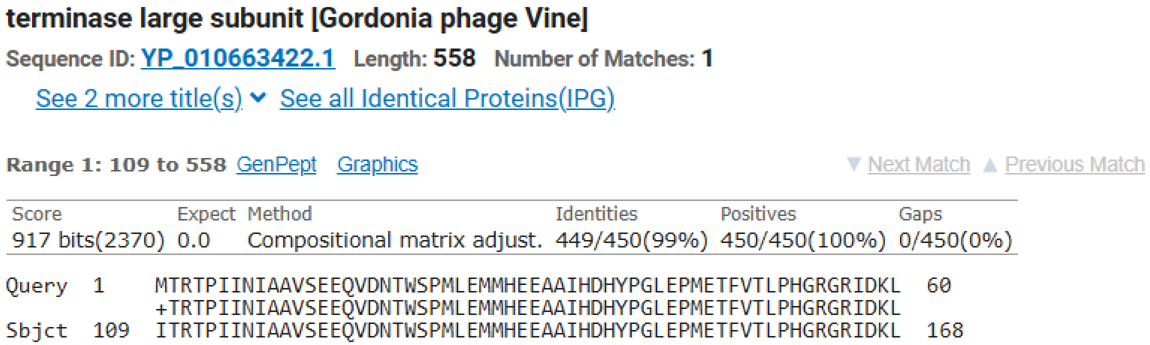
Score	Target Description
2896	terminase large subunit [Gordonia phage Vine] >
2894	terminase large subunit [Gordonia phage Lauer] >
2890	terminase large subunit [Gordonia phage Summit
2889	terminase large subunit [Gordonia phage PotPie]
2887	terminase large subunit [Gordonia phage BigChu

QBLAST Hit
Accession YP_010663422
GI
Length 558
Max Score 2896 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	1120.2	Identities	557
Score	2896	%Identity	99.82
E-Value	0.0E0	Positives	558
Length	558	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 558		
Target	1 - 558		

- Glimmer call (1762): There are 24 1:1 alignments on BLAST
- GeneMark call (2086): There are several 1:109 alignments, several 2:11 alignments, and one 1:92 alignment



terminase large subunit [Gordonia phage Vine]

Sequence ID: YP_010663422.1 Length: 558 Number of Matches: 1

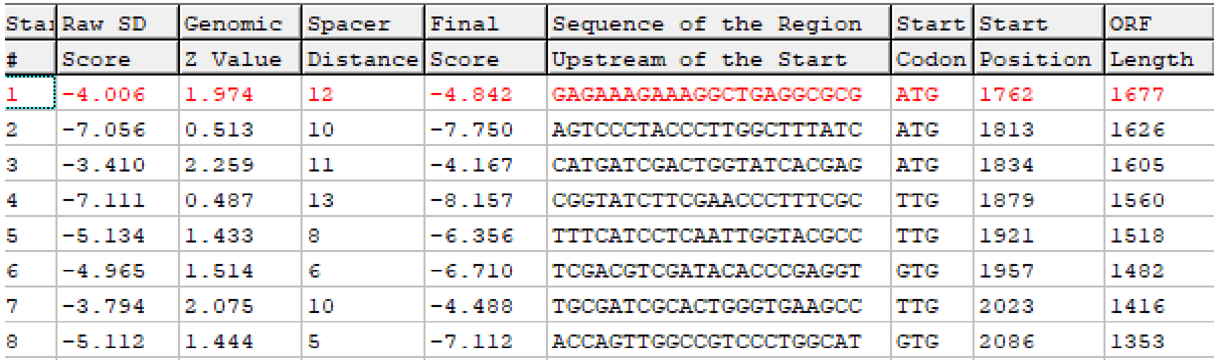
See 2 more title(s) ⌄ See all Identical Proteins(IPG)

Range 1: 109 to 558 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
917 bits(2370)	0.0	Compositional matrix adjust.	449/450(99%)	450/450(100%)	0/450(0%)

```
Query  1    MTRTPIINIAAVSEEQVDNTWSPMLEMMHEEAAIHDHYPGLEPMETFVTLPHGRGRIDKL  60
             +TRTPIINIAAVSEEQVDNTWSPMLEMMHEEAAIHDHYPGLEPMETFVTLPHGRGRIDKL
Sbjct  109  ITRTPIINIAAVSEEQVDNTWSPMLEMMHEEAAIHDHYPGLEPMETFVTLPHGRGRIDKL  168
```

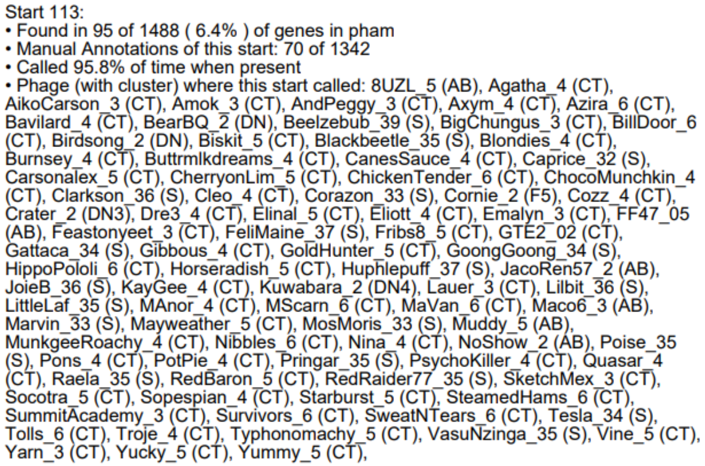
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



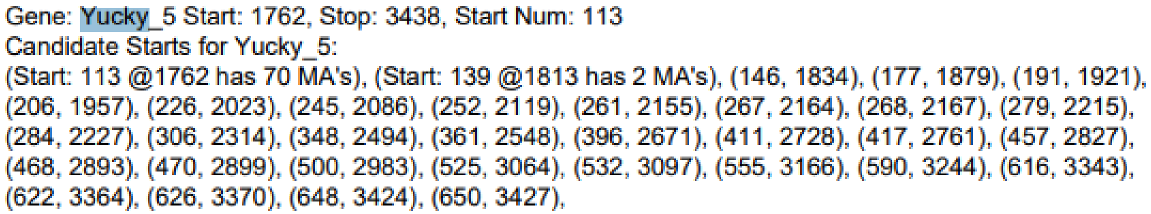
Sta	Raw SD	Genomic	Spacer	Final	Sequence of the Region	Start	Start	ORF
#	Score	Z Value	Distance	Score	Upstream of the Start	Codon	Position	Length
1	-4.006	1.974	12	-4.842	GAGAAAGAAAGGCTGAGGCGCG	ATG	1762	1677
2	-7.056	0.513	10	-7.750	AGTCCCTACCCTTGGCTTTATC	ATG	1813	1626
3	-3.410	2.259	11	-4.167	CATGATCGACTGGTATCACGAG	ATG	1834	1605
4	-7.111	0.487	13	-8.157	CGGTATCTTCGAACCCTTTCGC	TTG	1879	1560
5	-5.134	1.433	8	-6.356	TTTCATCCTCAATTGGTACGCC	TTG	1921	1518
6	-4.965	1.514	6	-6.710	TCGACGTCGATACACCCGAGGT	GTG	1957	1482
7	-3.794	2.075	10	-4.488	TGCGATCGCACTGGGTGAAGCC	TTG	2023	1416
8	-5.112	1.444	5	-7.112	ACCAGTTGGCCGTCCCTGGCAT	GTG	2086	1353

- Glimmer call (1762): Z-value= 1.974. Final score=-4.842
- GeneMark call (2086): Z-value= 1.444. Final score= -7.112
- There was another start site (1834) with good RBS numbers. Will be looked further into in Starterator: Z-value=2.259. Final score= -4.167

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



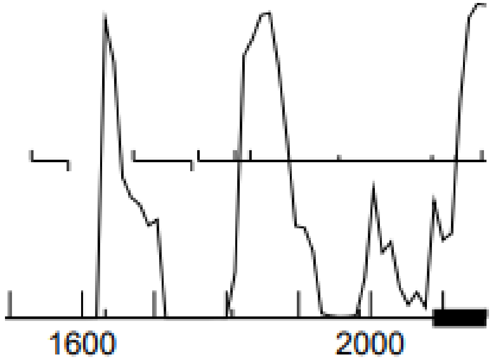
Start 113:
- Found in 95 of 1488 (6.4%) of genes in pham
- Manual Annotations of this start: 70 of 1342
- Called 95.8% of time when present
- Phage (with cluster) where this start called: 8UZL_5 (AB), Agatha_4 (CT), AikoCarson_3 (CT), Amok_3 (CT), AndPeggy_3 (CT), Axym_4 (CT), Azira_6 (CT), Bavilard_4 (CT), BearBQ_2 (DN), Beelzebub_39 (S), BigChungus_3 (CT), BillDoor_6 (CT), Birdsong_2 (DN), Biskit_5 (CT), Blackbeetle_35 (S), Blondies_4 (CT), Burnsey_4 (CT), Buttrmlkdreams_4 (CT), CanesSauce_4 (CT), Caprice_32 (S), Carsonalex_5 (CT), CherryonLim_5 (CT), ChickenTender_6 (CT), ChocoMunchkin_4 (CT), Clarkson_36 (S), Cleo_4 (CT), Corazon_33 (S), Cornie_2 (F5), Cozz_4 (CT), Crater_2 (DN3), Dre3_4 (CT), Elinal_5 (CT), Eliott_4 (CT), Emalyn_3 (CT), FF47_05 (AB), Feastonyeet_3 (CT), FeliMaine_37 (S), Fribs8_5 (CT), GTE2_02 (CT), Gattaca_34 (S), Gibbous_4 (CT), GoldHunter_5 (CT), GoongGoong_34 (S), HippoPololi_6 (CT), Horseradish_5 (CT), Huphlepuff_37 (S), JacoRen57_2 (AB), JoieB_36 (S), KayGee_4 (CT), Kuwabara_2 (DN4), Lauer_3 (CT), Lilbit_36 (S), LittleLaf_35 (S), MAnor_4 (CT), MScarn_6 (CT), MaVan_6 (CT), Maco6_3 (AB), Marvin_33 (S), Mayweather_5 (CT), MosMoris_33 (S), Muddy_5 (AB), MunkgeeRoachy_4 (CT), Nibbles_6 (CT), Nina_4 (CT), NoShow_2 (AB), Poise_35 (S), Pons_4 (CT), PotPie_4 (CT), Pringar_35 (S), PsychoKiller_4 (CT), Quasar_4 (CT), Raela_35 (S), RedBaron_5 (CT), RedRaider77_35 (S), SketchMex_3 (CT), Socotra_5 (CT), Sopespian_4 (CT), Starburst_5 (CT), SteamedHams_6 (CT), SummitAcademy_3 (CT), Survivors_6 (CT), SweatNTears_6 (CT), Tesla_34 (S), Tolls_6 (CT), Troje_4 (CT), Typhonomachy_5 (CT), VasuNzinga_35 (S), Vine_5 (CT), Yarn_3 (CT), Yucky_5 (CT), Yummy_5 (CT),

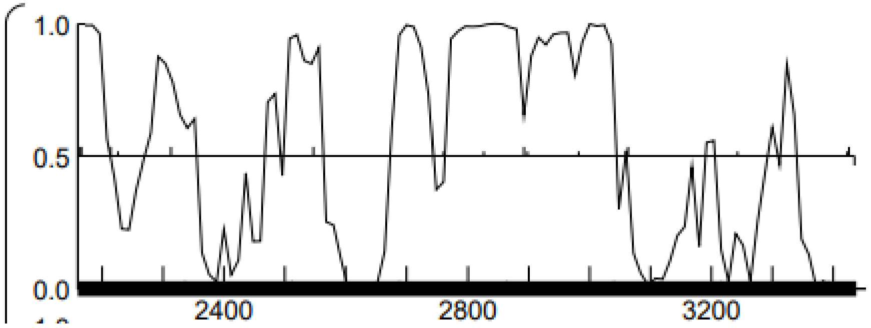


Gene: Yucky_5 Start: 1762, Stop: 3438, Start Num: 113
Candidate Starts for Yucky_5:
(Start: 113 @1762 has 70 MA's), (Start: 139 @1813 has 2 MA's), (146, 1834), (177, 1879), (191, 1921), (206, 1957), (226, 2023), (245, 2086), (252, 2119), (261, 2155), (267, 2164), (268, 2167), (279, 2215), (284, 2227), (306, 2314), (348, 2494), (361, 2548), (396, 2671), (411, 2728), (417, 2761), (457, 2827), (468, 2893), (470, 2899), (500, 2983), (525, 3064), (532, 3097), (555, 3166), (590, 3244), (616, 3343), (622, 3364), (626, 3370), (648, 3424), (650, 3427),

- Glimmer call (1762): Has 70 MAs, called 95.8% of the time when present.
- GeneMark call (2068): 0 MAs, never called before
- Potential alternative start site (1834): 0 MAs, never called before, also not autoannotated.

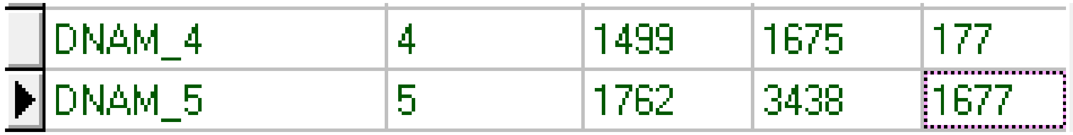
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.





- Glimmer call (1762): Small strong coding potential peak cut off at roughly nucleotide 1600. Many strong peaks throughout.
- GeneMark call (2086): More coding potential cut off: 3 strong peaks. Many strong peaks throughout, similar to the Glimmer call.

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.



DNAM_4	4	1499	1675	177
DNAM_5	5	1762	3438	1677

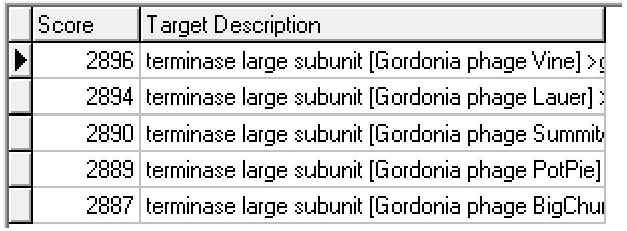
- Glimmer gap: 1762- 1675= 87-1 for gap= 86
- GeneMark gap: 2068-1675= 411-1 for gap= 410

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

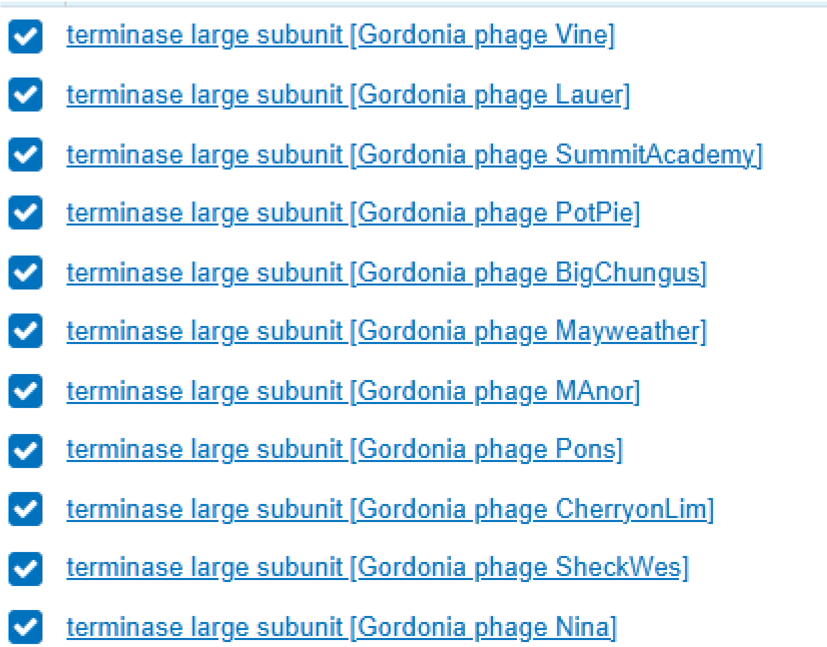
	1762	2068
GeneMark	Glimmer	GeneMark
Coding Potential	Cuts off slight coding potential at 1600, strong coding potential throughout	Cuts off 3 strong peaks. Contains strong coding potential throughout.
RBS	Z-value: 1.974 Final score: -4.842	Z-value: 1.444 Final score: -7.112
BLAST	24 1:1 alignments	Several 1:109 alignments, several 2:111 alignments, one 1:92 alignment
Starterator	70 MAs	0 MAs
Gap/Overlap	86	410

Based on this evidence I believe 1762 to be the true start site. It cuts off less coding potential, has better RBS members, has more manual annotations, and has less gap between the previous gene.

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
2896	terminase large subunit [Gordonia phage Vine] >g
2894	terminase large subunit [Gordonia phage Lauer] >
2890	terminase large subunit [Gordonia phage Summit
2889	terminase large subunit [Gordonia phage PotPie]
2887	terminase large subunit [Gordonia phage BigChu



☑ terminase large subunit [Gordonia phage Vine]

☑ terminase large subunit [Gordonia phage Lauer]

☑ terminase large subunit [Gordonia phage SummitAcademy]

☑ terminase large subunit [Gordonia phage PotPie]

☑ terminase large subunit [Gordonia phage BigChungus]

☑ terminase large subunit [Gordonia phage Mayweather]

☑ terminase large subunit [Gordonia phage MAnor]

☑ terminase large subunit [Gordonia phage Pons]

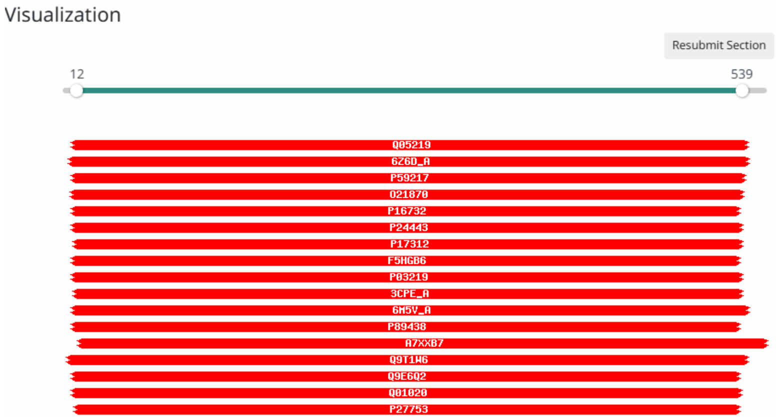
☑ terminase large subunit [Gordonia phage CherryonLim]

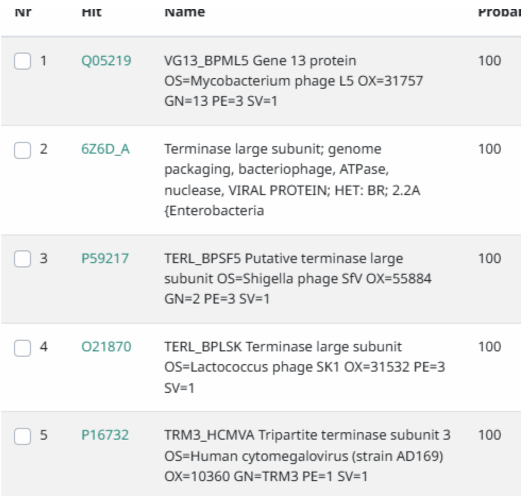
☑ terminase large subunit [Gordonia phage SheckWes]

☑ terminase large subunit [Gordonia phage Nina]

- All 25 highly similar genes shown by BLAST have been assigned a terminase large subunit function.
- BLASTing on NCBI yielded the same result.

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

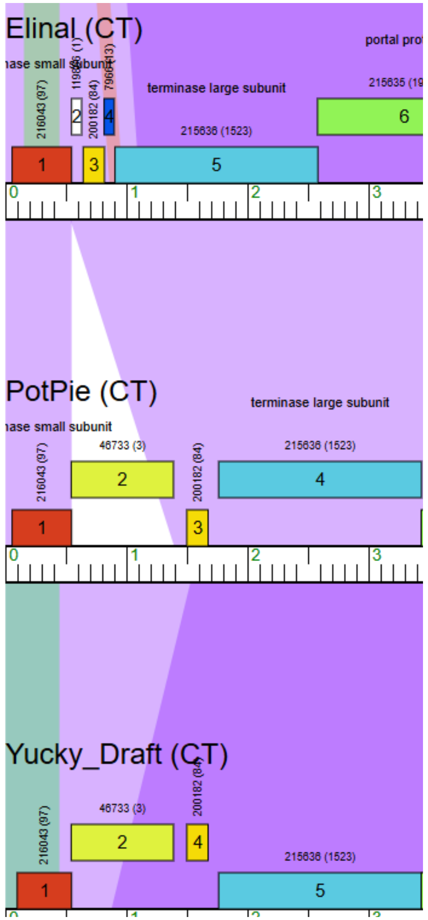




Nr		Hit	Name	Probal
☐	1	Q05219	VG13_BPML5 Gene 13 protein OS=Mycobacterium phage L5 OX=31757 GN=13 PE=3 SV=1	100
☐	2	6Z6D_A	Terminase large subunit; genome packaging, bacteriophage, ATPase, nuclease, VIRAL PROTEIN; HET: BR; 2.2A {Enterobacteria	100
☐	3	P59217	TERL_BPSF5 Putative terminase large subunit OS=Shigella phage SfV OX=55884 GN=2 PE=3 SV=1	100
☐	4	O21870	TERL_BPLSK Terminase large subunit OS=Lactococcus phage SK1 OX=31532 PE=3 SV=1	100
☐	5	P16732	TRM3_HCMVA Tripartite terminase subunit 3 OS=Human cytomegalovirus (strain AD169) OX=10360 GN=TRM3 PE=1 SV=1	100

- Hhpred shows at least 25 excellent hits as terminase large subunits. For the 25 shown, most of the gene is homologous.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- PotPie, Elinal, and BigChungus shows this gene as being a terminase large subunit.
- PotPie and BigChungus have a conserved domain as a terminase.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- This gene has a function of a large terminase subunit so deep TMHMM is not applicable.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The official function I am assigning to this gene is terminase, large subunit. On both DNA master and NCBI there were at least 25 BLAST hits saying this gene is a large terminase subunit. Hhpred backs this information, showing many excellent hits as a large terminase subunit. Lastly, Phamerator showed 3 phages with a similar gene in the same cluster and pham that had that function.

Feature Removed - Stop 1777

Glimmer/GeneMark

What feature number is this? Removed

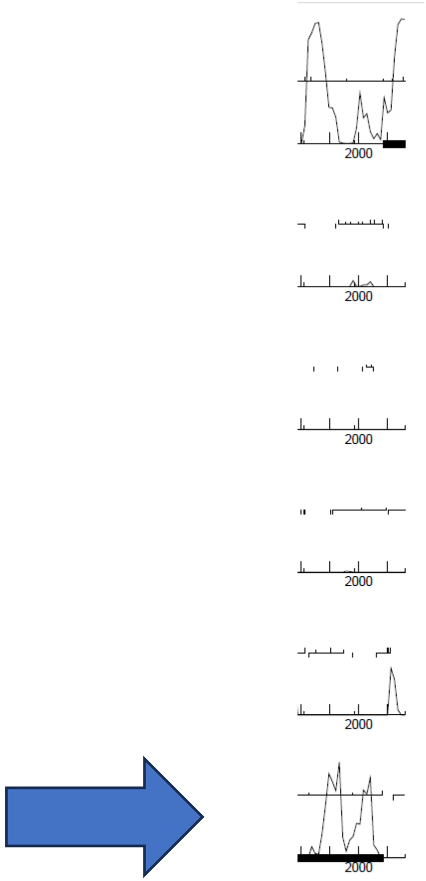
What is the stop site? 1777 (reverse)

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? Called by GeneMark at 2085, not called by Glimmer

What is the autoannotated start? 2085

Gap: ____1349 with feature 7____ or overlap: ________ (with gene in front of it) for the autoannotated start   - Overlaps completely with feature 5

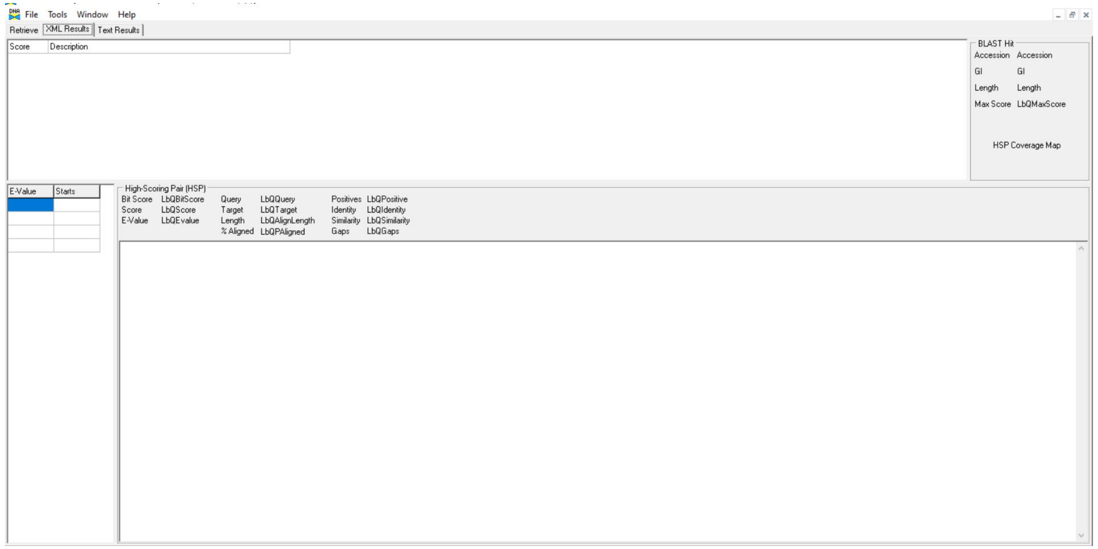
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Overlaps with feature 5 in reading frame 1. Fair amount of coding potential, but appears as single reverse gene in many forward genes, going against guiding principles.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are no BLAST hits.



Answer: Is it a gene? Give evidence why you think this is a gene or not.

- No, this isn’t a gene. The feature stands alone as a reverse gene, which does not agree with guiding principles. There are no close matches in BLAST. Even though it is called by Genemark, it is not called by Glimmer. In addition, it completely overlaps with feature 5, which is in reading frame 1.

Feature 5 – Stop 4868

Glimmer/GeneMark

What feature number is this? 5

What is the stop site? **4868**

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Glimmer and GeneMark

What is the autoannotated start?

3435

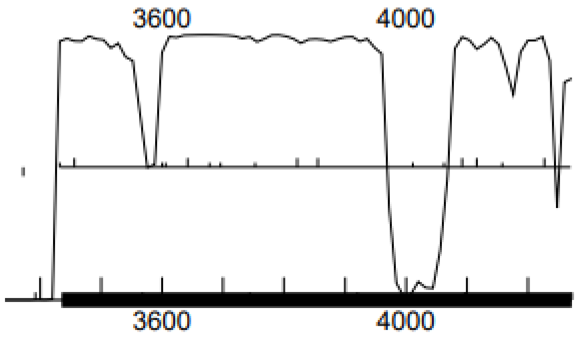
Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

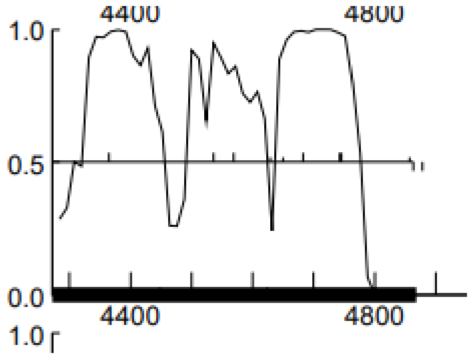
Overlap of 4

- Previous ends at 3438

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

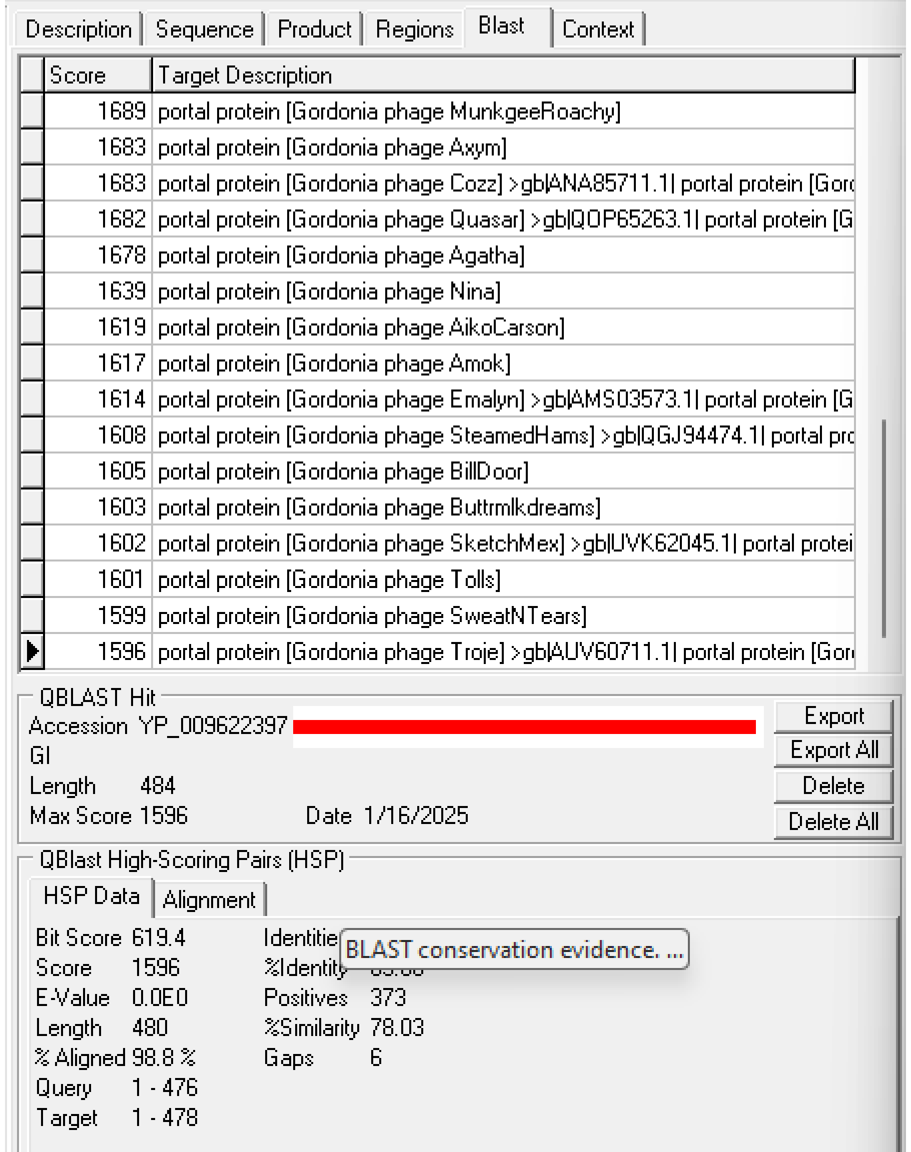
- There is strong coding potential throughout where the feature is called to be with a few small dips into weak coding potential throughout the feature.
- The initial peak of potential starts before the feature is called to being, but a majority of the potential is included.





BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are at least 25 BLAST hits of highly similar genes from other phages that all have e-values extremely close to zero.
- 6 1:1 alignments



Score	Target Description
1689	portal protein [Gordonia phage MunkgeeRoachy]
1683	portal protein [Gordonia phage Axym]
1683	portal protein [Gordonia phage Cozz] >gb|ANA85711.1| portal protein [Gord
1682	portal protein [Gordonia phage Quasar] >gb|QOP65263.1| portal protein [G
1678	portal protein [Gordonia phage Agatha]
1639	portal protein [Gordonia phage Nina]
1619	portal protein [Gordonia phage AikoCarson]
1617	portal protein [Gordonia phage Amok]
1614	portal protein [Gordonia phage Emalyn] >gb|AMS03573.1| portal protein [G
1608	portal protein [Gordonia phage SteamedHams] >gb|QGJ94474.1| portal pro
1605	portal protein [Gordonia phage BillDoor]
1603	portal protein [Gordonia phage Buttrmlkdreams]
1602	portal protein [Gordonia phage SketchMex] >gb|UVK62045.1| portal protei
1601	portal protein [Gordonia phage Tolls]
1599	portal protein [Gordonia phage SweatNTears]
1596	portal protein [Gordonia phage Troje] >gb|AUV60711.1| portal protein [Gor

QBLAST Hit

Accession YP_009622397

GI

Length 484

Max Score 1596

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 619.4

Score 1596

E-Value 0.0E0

Length 480

% Aligned 98.8 %

Query 1 - 476

Target 1 - 478

Identitie

%Identity

Positives 373

%Similarity 78.03

Gaps 6

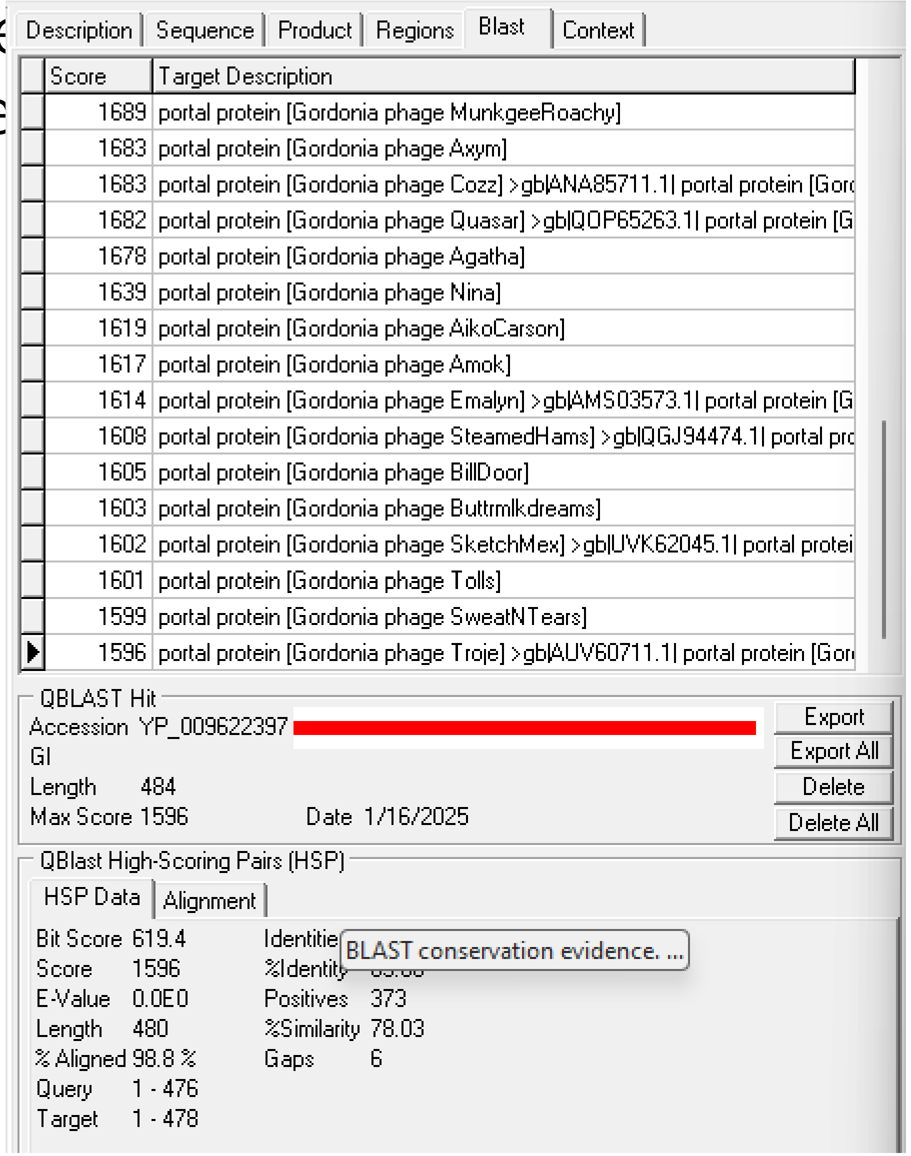
BLAST conservation evidence. ...

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This feature is a gene! There is strong coding potential throughout where the feature is called to be and there are at least 25 BLAST hits of highly similar genes from other phages that all have e-values close to zero.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answe
is favored based on BLAST alignment evide

- There are 6 1:1 alignments for starting at 3435



Score	Target Description
1689	portal protein [Gordonia phage MunkgeeRoachy]
1683	portal protein [Gordonia phage Axym]
1683	portal protein [Gordonia phage Cozz] >gb|ANA85711.1| portal protein [Gord
1682	portal protein [Gordonia phage Quasar] >gb|QOP65263.1| portal protein [G
1678	portal protein [Gordonia phage Agatha]
1639	portal protein [Gordonia phage Nina]
1619	portal protein [Gordonia phage AikoCarson]
1617	portal protein [Gordonia phage Amok]
1614	portal protein [Gordonia phage Emalyn] >gb|AMS03573.1| portal protein [G
1608	portal protein [Gordonia phage SteamedHams] >gb|QGJ94474.1| portal pro
1605	portal protein [Gordonia phage BillDoor]
1603	portal protein [Gordonia phage Buttrmlkdreams]
1602	portal protein [Gordonia phage SketchMex] >gb|UVK62045.1| portal protei
1601	portal protein [Gordonia phage Tolls]
1599	portal protein [Gordonia phage SweatNTears]
1596	portal protein [Gordonia phage Troje] >gb|AUV60711.1| portal protein [Gor

QBLAST Hit

Accession YP_009622397
GI
Length 484
Max Score 1596
Date 1/16/2025

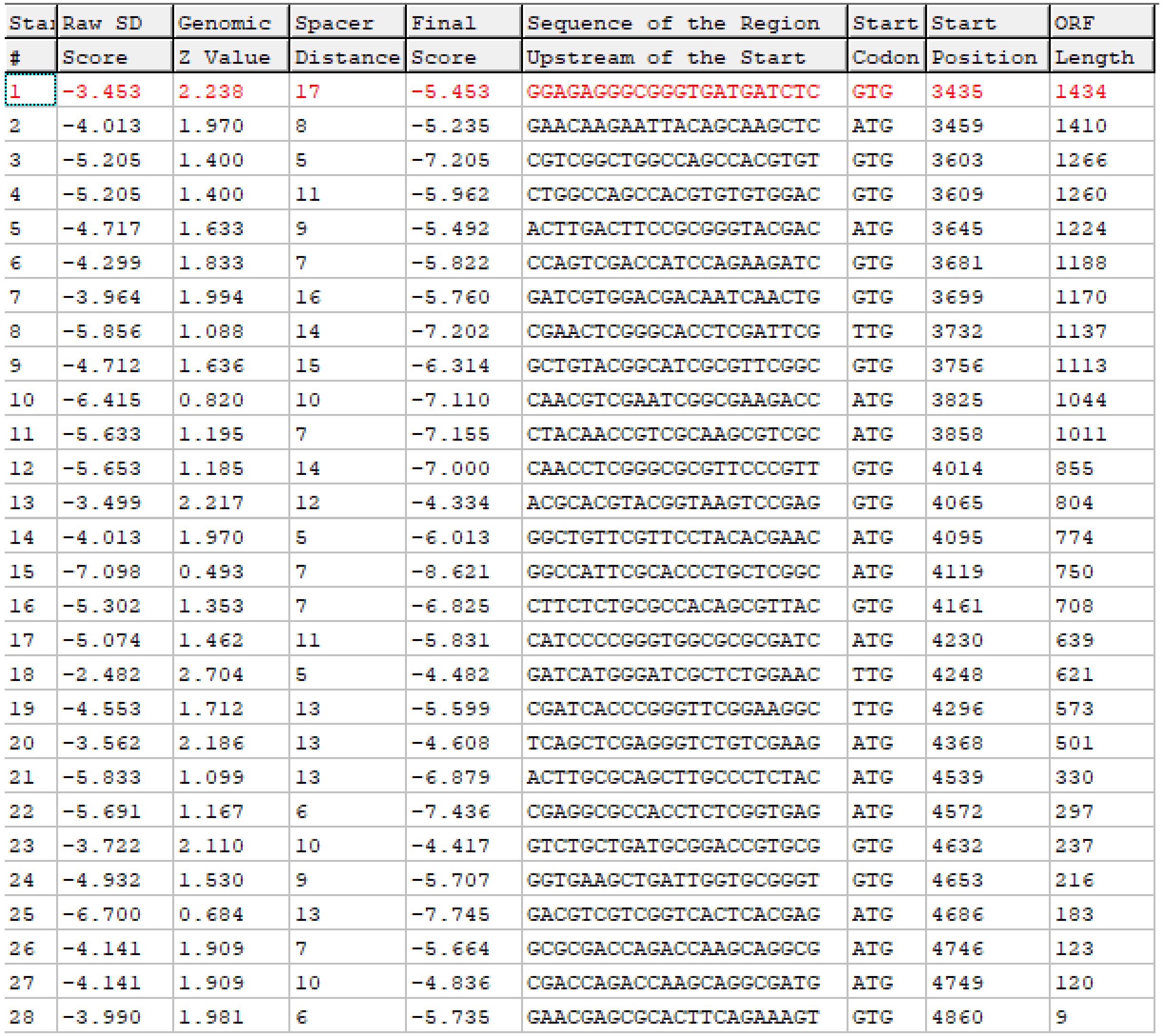
QBlast High-Scoring Pairs (HSP)

Bit Score	619.4	Identitie	
Score	1596	%Identity	
E-Value	0.0E0	Positives	373
Length	480	%Similarity	78.03
% Aligned	98.8 %	Gaps	6
Query	1 - 476		
Target	1 - 478		

BLAST conservation evidence. ...

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts?   Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

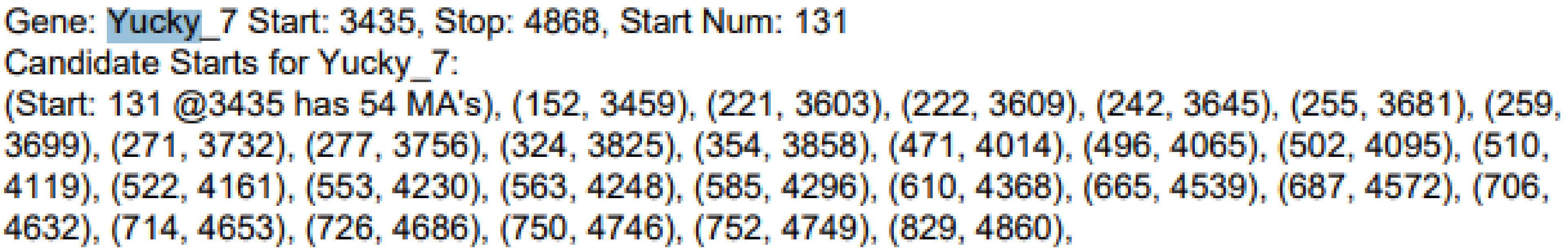
- Starting at 3435:
  - Z-value = 2.238
  - Final score = -5.453
- There was only one start site that had slightly better RBS scores than 3435, but it cut off a significantly larger amount of coding potential and was not mentioned in the starterator report.



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	-3.453	2.238	17	-5.453	GGAGAGGGCGGGTGATGATCTC	GTG	3435	1434
2	-4.013	1.970	8	-5.235	GAACAAGAATTACAGCAAGCTC	ATG	3459	1410
3	-5.205	1.400	5	-7.205	CGTCGGCTGGCCAGCCACGTGT	GTG	3603	1266
4	-5.205	1.400	11	-5.962	CTGGCCAGCCACGTGTGTGGAC	GTG	3609	1260
5	-4.717	1.633	9	-5.492	ACTTGACTTCCGCGGGTACGAC	ATG	3645	1224
6	-4.299	1.833	7	-5.822	CCAGTCGACCATCCAGAAGATC	GTG	3681	1188
7	-3.964	1.994	16	-5.760	GATCGTGGACGACAATCAACTG	GTG	3699	1170
8	-5.856	1.088	14	-7.202	CGAACTCGGGCACCTCGATTCG	TTG	3732	1137
9	-4.712	1.636	15	-6.314	GCTGTACGGCATCGCGTTCGGC	GTG	3756	1113
10	-6.415	0.820	10	-7.110	CAACGTCGAATCGGCGAAGACC	ATG	3825	1044
11	-5.633	1.195	7	-7.155	CTACAACCGTCGCAAGCGTCGC	ATG	3858	1011
12	-5.653	1.185	14	-7.000	CAACCTCGGGCGCGTTCCCGTT	GTG	4014	855
13	-3.499	2.217	12	-4.334	ACGCACGTACGGTAAGTCCGAG	GTG	4065	804
14	-4.013	1.970	5	-6.013	GGCTGTTCGTTCCTACACGAAC	ATG	4095	774
15	-7.098	0.493	7	-8.621	GGCCATTCGCACCCTGCTCGGC	ATG	4119	750
16	-5.302	1.353	7	-6.825	CTTCTCTGCGCCACAGCGTTAC	GTG	4161	708
17	-5.074	1.462	11	-5.831	CATCCCCGGGTGGCGCGCGATC	ATG	4230	639
18	-2.482	2.704	5	-4.482	GATCATGGATCGCTCTGGAAC	TTG	4248	621
19	-4.553	1.712	13	-5.599	CGATCACCCGGGTTCGGAAGGC	TTG	4296	573
20	-3.562	2.186	13	-4.608	TCAGCTCGAGGGTCTGTCGAAG	ATG	4368	501
21	-5.833	1.099	13	-6.879	ACTTGCGCAGCTTGCCCTCTAC	ATG	4539	330
22	-5.691	1.167	6	-7.436	CGAGGCGCCACCTCTCGGTGAG	ATG	4572	297
23	-3.722	2.110	10	-4.417	GTCTGCTGATGCGGACCGTGCG	GTG	4632	237
24	-4.932	1.530	9	-5.707	GGTGAAGCTGATTGGTGCGGGT	GTG	4653	216
25	-6.700	0.684	13	-7.745	GACGTCGTCGGTCACTCACGAG	ATG	4686	183
26	-4.141	1.909	7	-5.664	GCGCGACCAGACCAAGCAGGCG	ATG	4746	123
27	-4.141	1.909	10	-4.836	CGACCAGACCAAGCAGGCGATG	ATG	4749	120
28	-3.990	1.981	6	-5.735	GAACGAGCGCACTTCAGAAAGT	GTG	4860	9

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

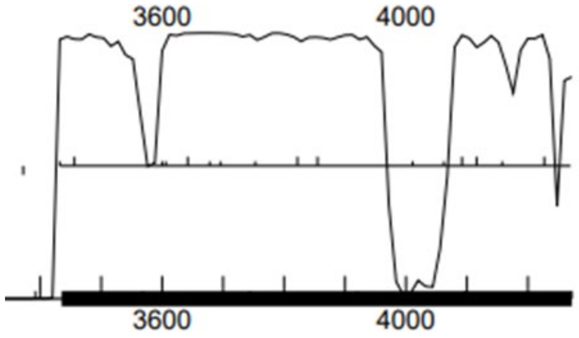
- Start 3435 was the only start site that had manual annotations, and it had 54 total.

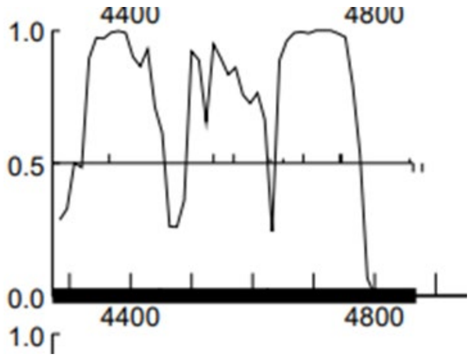


Gene: Yucky_7 Start: 3435, Stop: 4868, Start Num: 131
Candidate Starts for Yucky_7:
(Start: 131 @3435 has 54 MA's), (152, 3459), (221, 3603), (222, 3609), (242, 3645), (255, 3681), (259, 3699), (271, 3732), (277, 3756), (324, 3825), (354, 3858), (471, 4014), (496, 4065), (502, 4095), (510, 4119), (522, 4161), (553, 4230), (563, 4248), (585, 4296), (610, 4368), (665, 4539), (687, 4572), (706, 4632), (714, 4653), (726, 4686), (750, 4746), (752, 4749), (829, 4860),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Starting at 3454 would cut off part of the initial peak of coding potential, but most of the possible coding potential for the feature would be included.
- This is the earliest start possible, so any start after this one would cut off a larger amount of coding potential.





Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

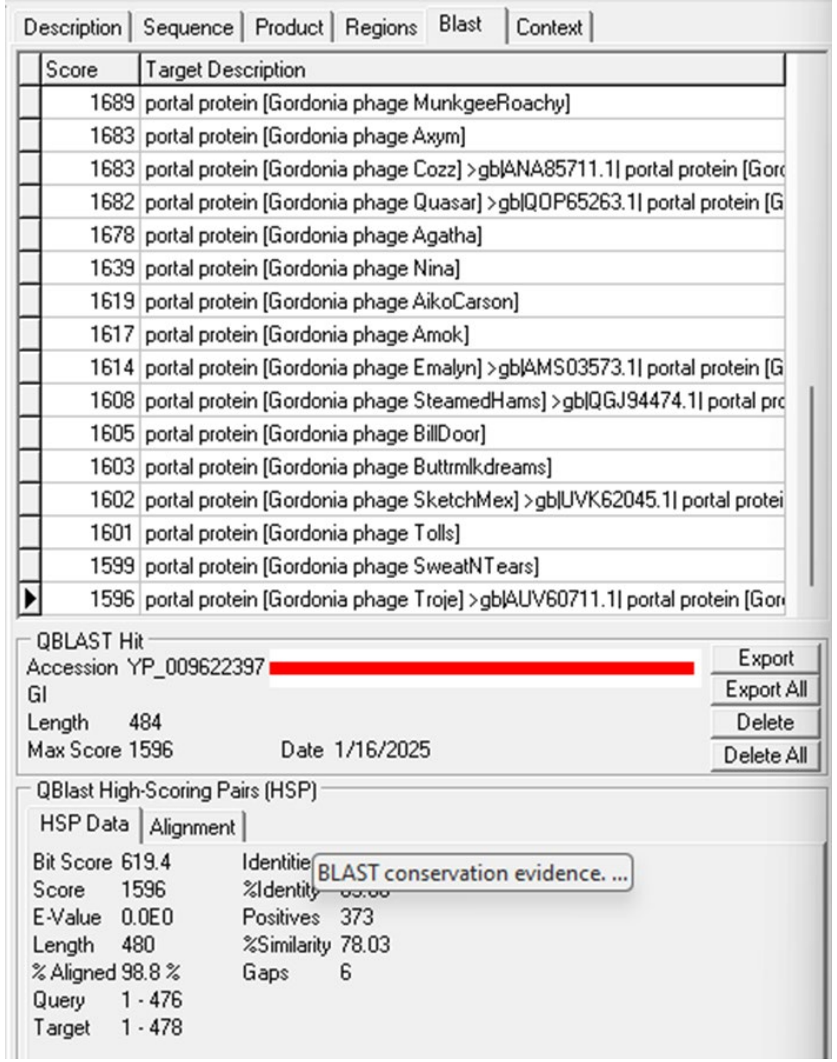
- 3435 → overlap of 4

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is 3435, and this was the only proposed start site possible based off all the evidence collected. There were 6 1:1 alignments with highly similar genes for starting at this position, and it includes the most coding potential possible. There were 54 manual annotations for starting at 3435, and it had the best RBS scores possible other than a really late start site that would cut off a large amount of coding potential. There would only be an overlap of 4 nucleotides with the previous gene starting at here which is a favorable condition as well.

BLAST function evidence. What assigned functions do other highly similar genes have?

- At least 25 BLAST hits had their functions listed as portal protein.



Score	Target Description
1689	portal protein [Gordonia phage MunkgeeRoachy]
1683	portal protein [Gordonia phage Axym]
1683	portal protein [Gordonia phage Cozz] >gb|ANA85711.1| portal protein [Gor
1682	portal protein [Gordonia phage Quasar] >gb|QOP65263.1| portal protein [G
1678	portal protein [Gordonia phage Agatha]
1639	portal protein [Gordonia phage Nina]
1619	portal protein [Gordonia phage AikoCarson]
1617	portal protein [Gordonia phage Amok]
1614	portal protein [Gordonia phage Emalyn] >gb|AMS03573.1| portal protein [G
1608	portal protein [Gordonia phage SteamedHams] >gb|QGJ94474.1| portal pro
1605	portal protein [Gordonia phage BillDoor]
1603	portal protein [Gordonia phage Buttrmlkdreams]
1602	portal protein [Gordonia phage SketchMex] >gb|UVK62045.1| portal protei
1601	portal protein [Gordonia phage Tolls]
1599	portal protein [Gordonia phage SweatNTears]
1596	portal protein [Gordonia phage Troje] >gb|AUV60711.1| portal protein [Gor

QBLAST Hit

Accession YP_009622397

GI

Length 484

Max Score 1596 Date 1/16/2025

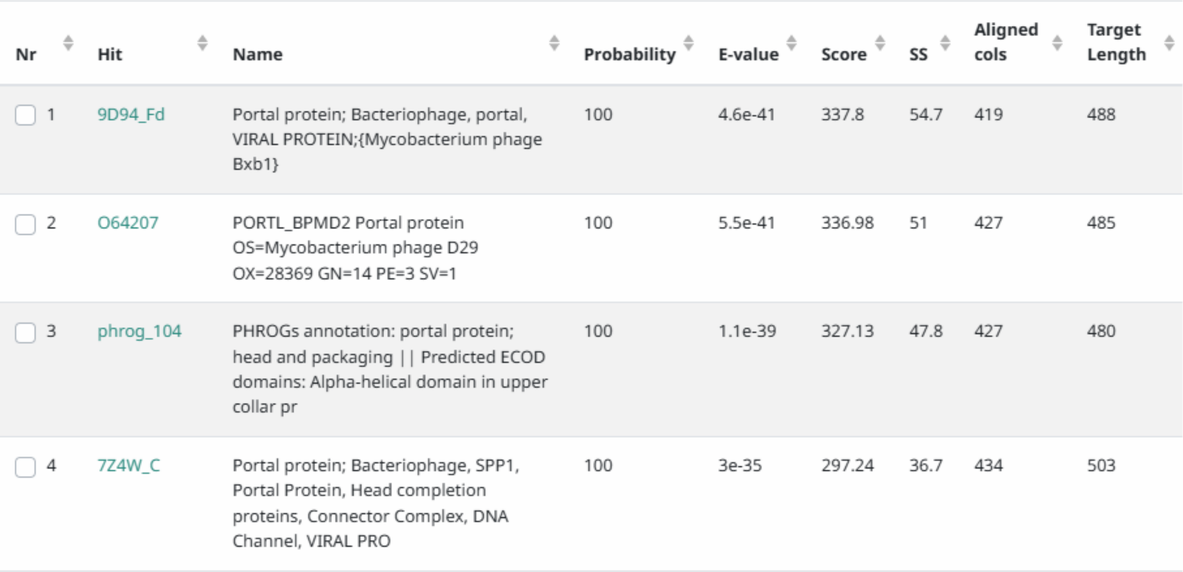
QBlast High-Scoring Pairs (HSP)

Bit Score	619.4	Identitie	
Score	1596	%Identity	
E-Value	0.0E0	Positives	373
Length	480	%Similarity	78.03
% Aligned	98.8 %	Gaps	6
Query	1 - 476		
Target	1 - 478		

BLAST conservation evidence. ...

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

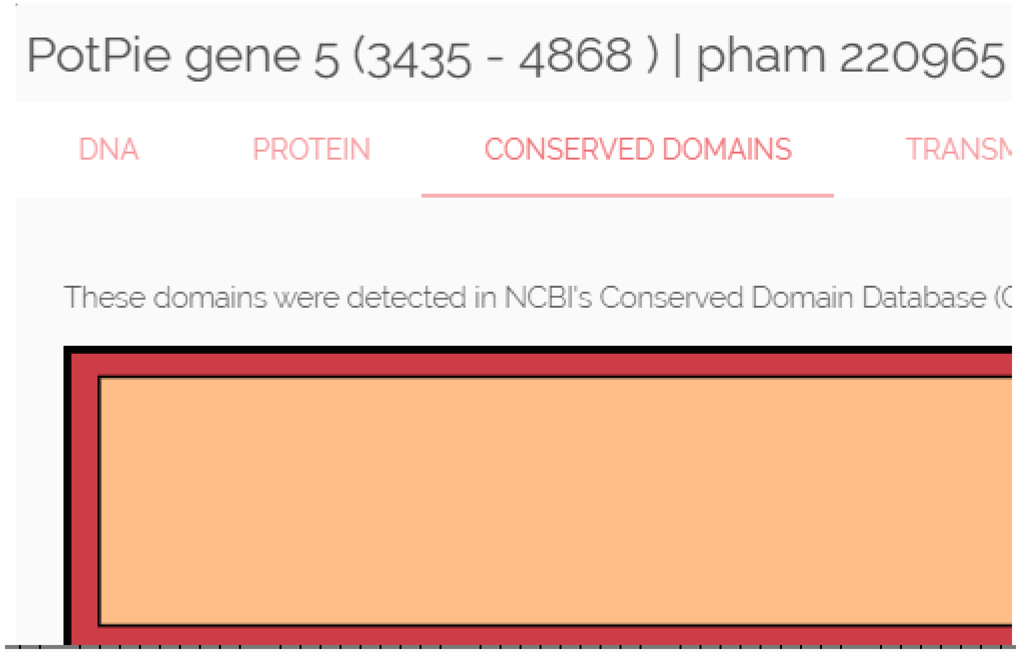
- There were several hits with probabilities over 90 (and several with 100) that suggested the function of portal protein as well.
- These hits were also homologous for a majority of the gene.



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	9D94_Fd	Portal protein; Bacteriophage, portal, VIRAL PROTEIN;{Mycobacterium phage Bxb1}	100	4.6e-41	337.8	54.7	419	488
□ 2	O64207	PORTL_BPMD2 Portal protein OS=Mycobacterium phage D29 OX=28369 GN=14 PE=3 SV=1	100	5.5e-41	336.98	51	427	485
□ 3	phrog_104	PHROGs annotation: portal protein; head and packaging || Predicted ECOD domains: Alpha-helical domain in upper collar pr	100	1.1e-39	327.13	47.8	427	480
□ 4	7Z4W_C	Portal protein; Bacteriophage, SPP1, Portal Protein, Head completion proteins, Connector Complex, DNA Channel, VIRAL PRO	100	3e-35	297.24	36.7	434	503

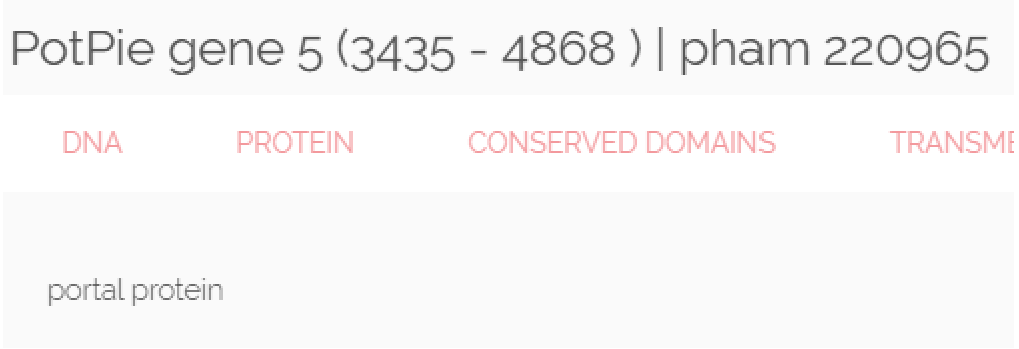
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Phamerator showed that phages with genes in the same pham as this one had functions listed as portal protein and conserved domains of Phage_prot_Gp6.
- This supports the function of this gene being labeled as a portal protein.



PotPie gene 5 (3435 - 4868) | pham 220965

These domains were detected in NCBI's Conserved Domain Database (C



PotPie gene 5 (3435 - 4868) | pham 220965

portal protein

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- **Not applicable since there is a probable function**

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Official function → portal protein
- The function of this gene should be labeled as a portal protein. At least 25 BLAST hits show that highly similar genes to this one have been listed as portal proteins. Hhpred also show several hit with high probabilities suggesting that the function of this gene should be labeled as a portal protein. Phamerator showed that phages with genes in the same pham as this one were also listed as portal proteins, and they showed conserved domains listed as Phage_prot_Gp6. Since this gene had a probable function a Deep TMHMM graph was not necessary.

Feature 6 Stop 5451

Instructions

Fill this out for each gene you annotate. This should be thought of as the minimum amount of information that needs to be provided for each gene. You can always add more slides or information as necessary

- Is it a gene?
  - Yes!
- Where does it start?
  - Gene starts at 4858!
- What is the function?
  - Hypothetical Protein

Glimmer/GeneMark

What feature number is this? **6**

What is the stop site? **5451**

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Glimmer only

What is the autoannotated start? **4828**

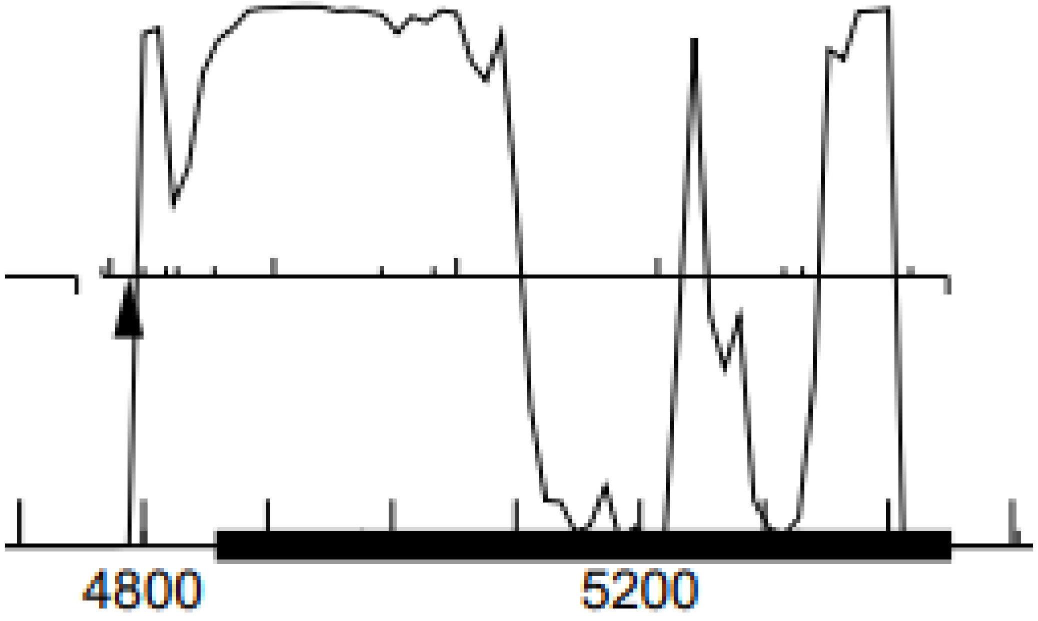
Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

For the autoannotated start there would be an overlap of 41 nucleotides

- *GeneMark called the feature starting at 4858*
- *The previous gene stopped at 4868*
- *If the start was 4858 then there would be an overlap of 11 nucleotides.*

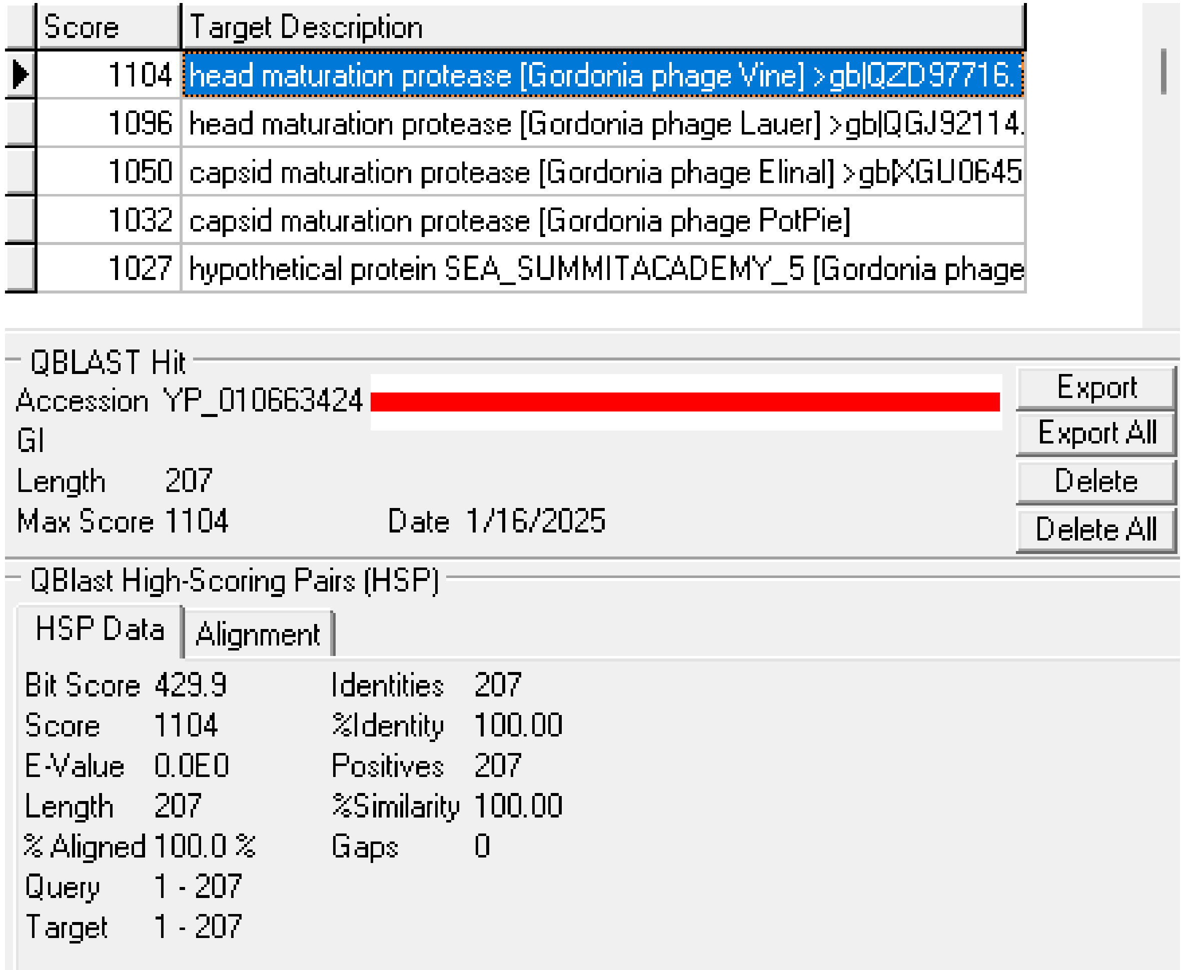
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- The coding potential of this feature starts off strong at around 4828 and remains that way until around 5100 where it dips until right after 5200 where it peaks back to strong. It dips again around 5300, but it peaks again right after and ends at 5451.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 2 1:1 alignments (Vine & Lauer)
- There are at least 25 hits on BLAST
- All BLAST hits have e-values close to zero



Score	Target Description
1104	head maturation protease [Gordonia phage Vine] >gb|QZD97716.
1096	head maturation protease [Gordonia phage Lauer] >gb|QGJ92114.
1050	capsid maturation protease [Gordonia phage Elinal] >gb|XGU0645
1032	capsid maturation protease [Gordonia phage PotPie]
1027	hypothetical protein SEA_SUMMITACADEMY_5 [Gordonia phage

QBLAST Hit

Accession YP_010663424

GI

Length 207

Max Score 1104 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

HSP Data

Bit Score 429.9

Score 1104

E-Value 0.0E0

Length 207

% Aligned 100.0 %

Query 1 - 207

Target 1 - 207

Identities 207

%Identity 100.00

Positives 207

%Similarity 100.00

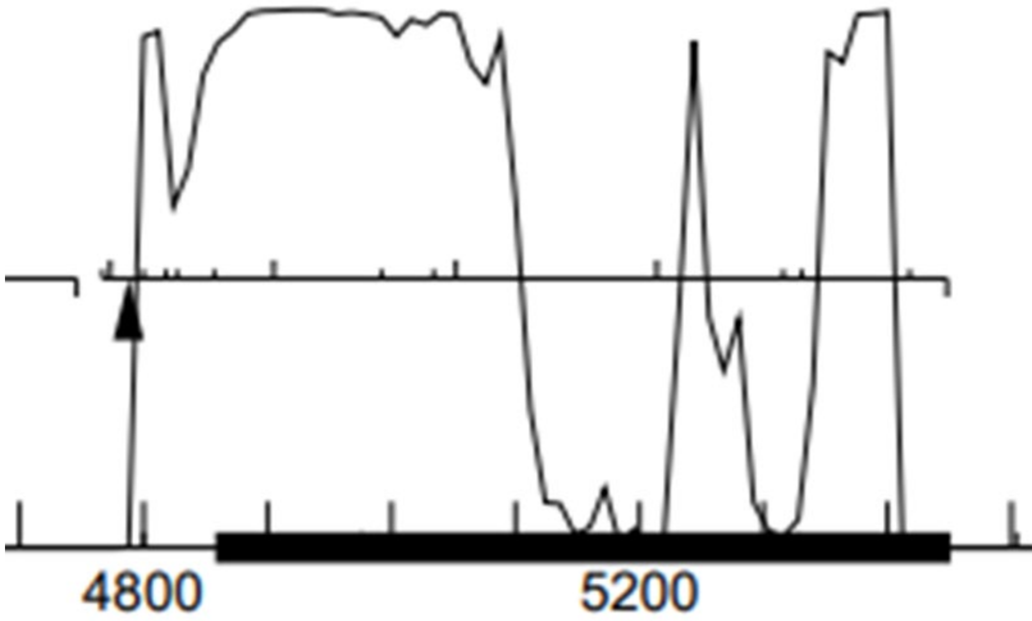
Gaps 0

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes! This feature is a gene as there is strong coding potential throughout it with only a couple dips. There are also several BLAST hits that all have e-values close to zero as well as two 1:1 alignments.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

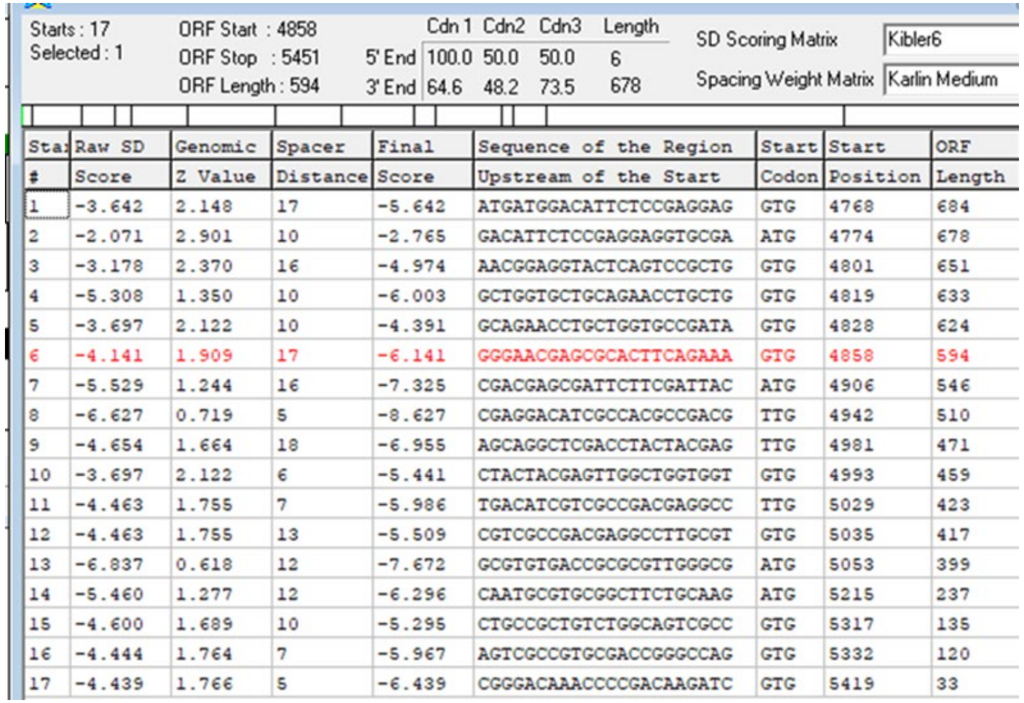
- Starting at 4828 (Glimmer call)
  - If the gene starts at nucleotide 4828, then all the coding potential would be included.
- Starting at 4858 (GeneMark call)
  - If the gene starts at nucleotide 4858, then nearly all the coding potential would be included except for the first small peak.



RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Starting at 4828:
  - Z-value = 2.122
  - Final score = -4.391
- Starting at 4858:
  - Z-value = 1.909
  - Final score = -6.141

- Based on the RBS values the favored start would be 4828.



Starts : 17
Selected : 1

ORF Start : 4858
ORF Stop : 5451
ORF Length : 594

	Cdn 1	Cdn2	Cdn3	Length
5' End	100.0	50.0	50.0	6
3' End	64.6	48.2	73.5	678

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-3.642	2.148	17	-5.642	ATGATGGACATTCTCCGAGGAG	GTG	4768	684
2	-2.071	2.901	10	-2.765	GACATTCTCCGAGGAGGTGCGA	ATG	4774	678
3	-3.178	2.370	16	-4.974	AACGGAGGTACTCAGTCCGCTG	GTG	4801	651
4	-5.308	1.350	10	-6.003	GCTGGTGCTGCAGAACCTGCTG	GTG	4819	633
5	-3.697	2.122	10	-4.391	GCAGAACCTGCTGGTGCCGATA	GTG	4828	624
6	-4.141	1.909	17	-6.141	GGGAACGAGCGCACTTCAGAAA	GTG	4858	594
7	-5.529	1.244	16	-7.325	CGACGAGCGATTCTTCGATTAC	ATG	4906	546
8	-6.627	0.719	5	-8.627	CGAGGACATCGCCACGCCGACG	TTG	4942	510
9	-4.654	1.664	18	-6.955	AGCAGGCTCGACCTACTACGAG	TTG	4981	471
10	-3.697	2.122	6	-5.441	CTACTACGAGTTGGCTGGTGGT	GTG	4993	459
11	-4.463	1.755	7	-5.986	TGACATCGTCGCCGACGAGGCC	TTG	5029	423
12	-4.463	1.755	13	-5.509	CGTCGCCGACGAGGCCTTGCGT	GTG	5035	417
13	-6.837	0.618	12	-7.672	GCGTGTGACCGCGCGTTGGGCG	ATG	5053	399
14	-5.460	1.277	12	-6.296	CAATGCGTGCGGCTTCTGCAAG	ATG	5215	237
15	-4.600	1.689	10	-5.295	CTGCCGCTGTCTGGCAGTCGCC	GTG	5317	135
16	-4.444	1.764	7	-5.967	AGTCGCCGTGCGACCGGGCCAG	GTG	5332	120
17	-4.439	1.766	5	-6.439	CGGGACAAACCCCGACAAGATC	GTG	5419	33

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- Starting at 4828:
  - 2 1:1 Alignments

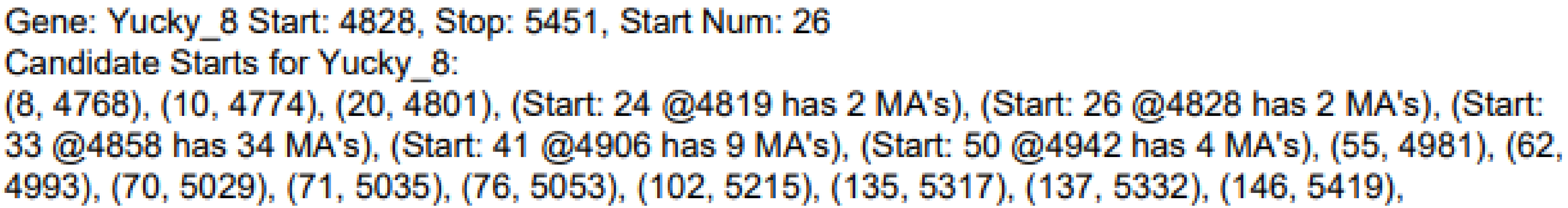
- Starting at 4858:
  - Over 10 1:1 alignments

Based off this evidence the favored start would be 4858!

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- There are 2 MA's for starting at 4828.

- There are 34 MA's for starting at 4858.



Gene: Yucky_8 Start: 4828, Stop: 5451, Start Num: 26
Candidate Starts for Yucky_8:
(8, 4768), (10, 4774), (20, 4801), (Start: 24 @4819 has 2 MA's), (Start: 26 @4828 has 2 MA's), (Start: 33 @4858 has 34 MA's), (Start: 41 @4906 has 9 MA's), (Start: 50 @4942 has 4 MA's), (55, 4981), (62, 4993), (70, 5029), (71, 5035), (76, 5053), (102, 5215), (135, 5317), (137, 5332), (146, 5419),

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Starting at 4828
  - There would be an overlap of 41 nucleotides.

- Starting at 4858
  - There would be an overlap of 11 nucleotides.

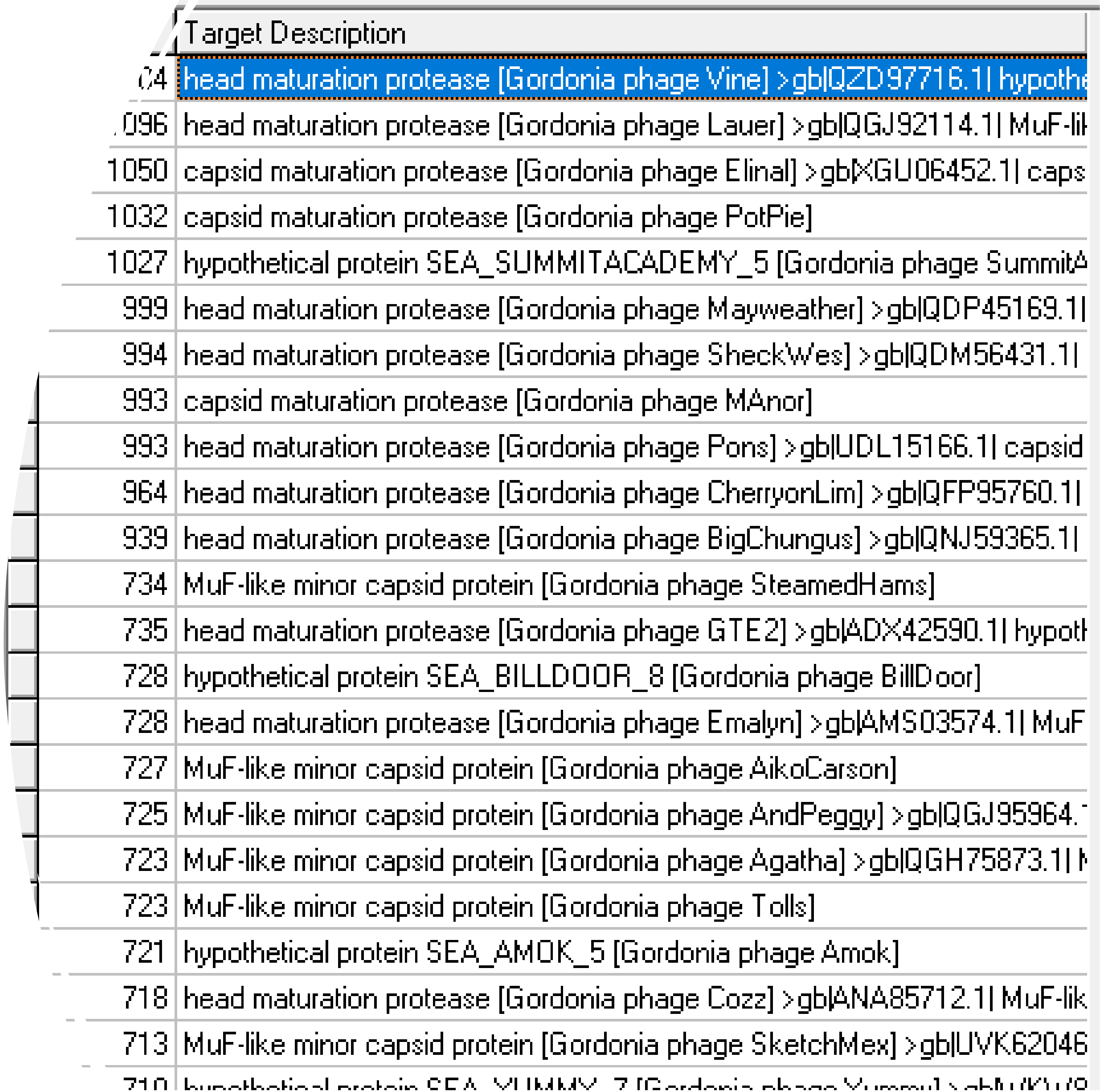
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	Start at 4828	Start at 4858
Glimmer/GeneMark	Glimmer	GeneMark
Coding Potential	Includes all coding potential of the gene	Cuts out a little of the first peak
RBS	Z-value =2.122 Final score = -4.391	Z-value = 1.909 Final score = -6.141
BLAST	2 1:1 Alignments	Over 10 1:1 alignments
Starterator	2 MA's	34 MA's
Gap/Overlap	Overlap of 41 nucleotides	Overlap of 11 nucleotides

The gene starts at 4858! Starting at 4858 cuts of a small portion of the first peak, but it includes nearly all the coding potential. The z-value of 1.909 and final score of -6.141 were not as preferable as the z-value and final score given by starting at 4828 (2.122 and -4.391), but due to the overlap of nucleotides these numbers do not hold as much value. Starting at 4828 would end up with an overlap of 41 nucleotides, and starting at 4858 would only leave an overlap of 11 nucleotides. Over these possible overlaps having an overlap of 11 nucleotides would be preferable. Starting at 4858 there were over 10 1:1 alignments which was better than the 2 1:1 alignments that there were starting at 4828. The final piece of evidence that supports 4858 being the start site is that there were 34 MA's of that being the start site according to Starterator.

BLAST function evidence. What assigned functions do other highly similar genes have?

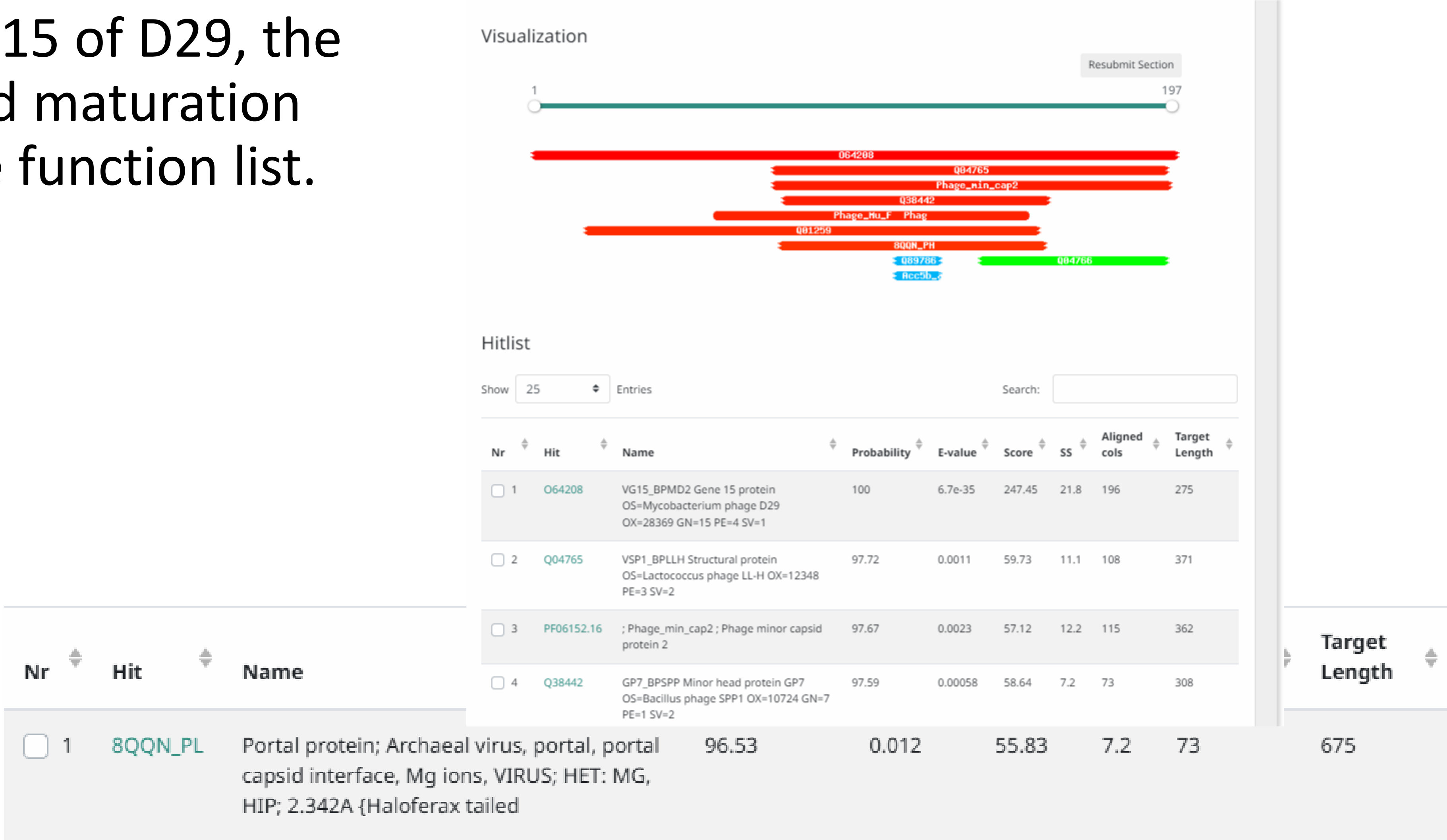
- BLAST hit functions:
    - Head maturation protease – at least 10 hits (not on Official Function List)
    - Capsid maturation protease – at least 3 hits
    - Hypothetical protein – at least 4 hits
    - MuF-like minor capsid protein – at least 7 hits (not usable according to Official Function List)
- All BLAST hits were Gordonia phages



	Target Description
04	head maturation protease [Gordonia phage Vine] >gb|QZD97716.1| hypothe
096	head maturation protease [Gordonia phage Lauer] >gb|QGJ92114.1| MuF-lik
1050	capsid maturation protease [Gordonia phage Elinal] >gb|XGU06452.1| caps
1032	capsid maturation protease [Gordonia phage PotPie]
1027	hypothetical protein SEA_SUMMITACADEMY_5 [Gordonia phage SummitA
999	head maturation protease [Gordonia phage Mayweather] >gb|QDP45169.1|
994	head maturation protease [Gordonia phage SheckWes] >gb|QDM56431.1|
993	capsid maturation protease [Gordonia phage MAnor]
993	head maturation protease [Gordonia phage Pons] >gb|UDL15166.1| capsid
964	head maturation protease [Gordonia phage CherryonLim] >gb|QFP95760.1|
939	head maturation protease [Gordonia phage BigChungus] >gb|QNJ59365.1|
734	MuF-like minor capsid protein [Gordonia phage SteamedHams]
735	head maturation protease [Gordonia phage GTE2] >gb|ADX42590.1| hypoth
728	hypothetical protein SEA_BILLDOOR_8 [Gordonia phage BillDoor]
728	head maturation protease [Gordonia phage Emalyn] >gb|AMS03574.1| MuF
727	MuF-like minor capsid protein [Gordonia phage AikoCarson]
725	MuF-like minor capsid protein [Gordonia phage AndPeggy] >gb|QGJ95964.
723	MuF-like minor capsid protein [Gordonia phage Agatha] >gb|QGH75873.1|
723	MuF-like minor capsid protein [Gordonia phage Tolls]
721	hypothetical protein SEA_AMOK_5 [Gordonia phage Amok]
718	head maturation protease [Gordonia phage Cozz] >gb|ANA85712.1| MuF-lik
713	MuF-like minor capsid protein [Gordonia phage SketchMex] >gb|UVK62046
710	hypothetical protein SEA_YUMMY_7 [Gordonia phage Yummy] >gb|UKV

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- Matches with gp15 of D29, the example for capsid maturation protease from the function list.



Visualization

Hitlist

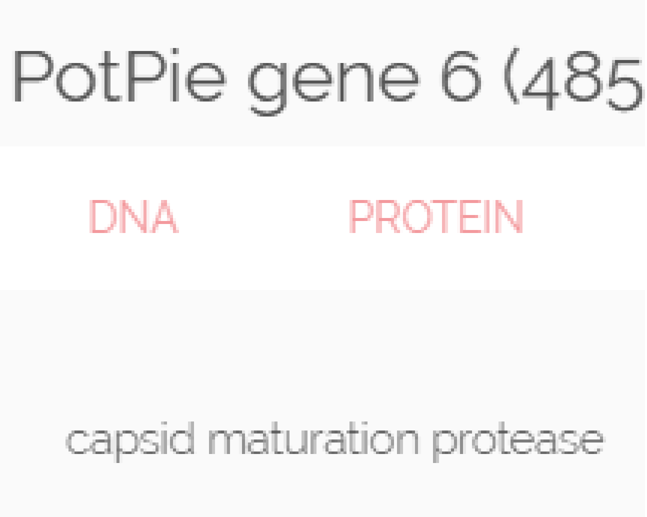
Show 25 Entries Search:

	Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
☐	1	O64208	VG15_BPMD2 Gene 15 protein OS=Mycobacterium phage D29 OX=28369 GN=15 PE=4 SV=1	100	6.7e-35	247.45	21.8	196	275
☐	2	Q04765	VSP1_BPLLH Structural protein OS=Lactococcus phage LL-H OX=12348 PE=3 SV=2	97.72	0.0011	59.73	11.1	108	371
☐	3	PF06152.16	; Phage_min_cap2 ; Phage minor capsid protein 2	97.67	0.0023	57.12	12.2	115	362
☐	4	Q38442	GP7_BPSPP Minor head protein GP7 OS=Bacillus phage SPP1 OX=10724 GN=7 PE=1 SV=2	97.59	0.00058	58.64	7.2	73	308

	Nr	Hit	Name						Target Length
☐	1	8QQN_PL	Portal protein; Archaeal virus, portal, portal capsid interface, Mg ions, VIRUS; HET: MG, HIP; 2.342A {Haloferax tailed	96.53	0.012	55.83	7.2	73	675

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

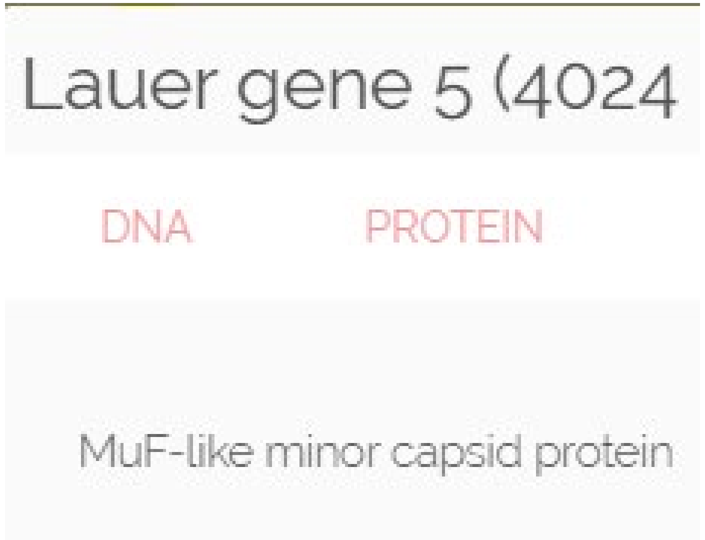
- There were no conserved domains.
- Other genes in the same pham called for functions of capsid maturation protease and MuF-like minor capsid protein.
- PotPie has been a gene that is highly similar in the past and it was listed as capsid maturation protein as the function.



PotPie gene 6 (485

DNA PROTEIN

capsid maturation protease



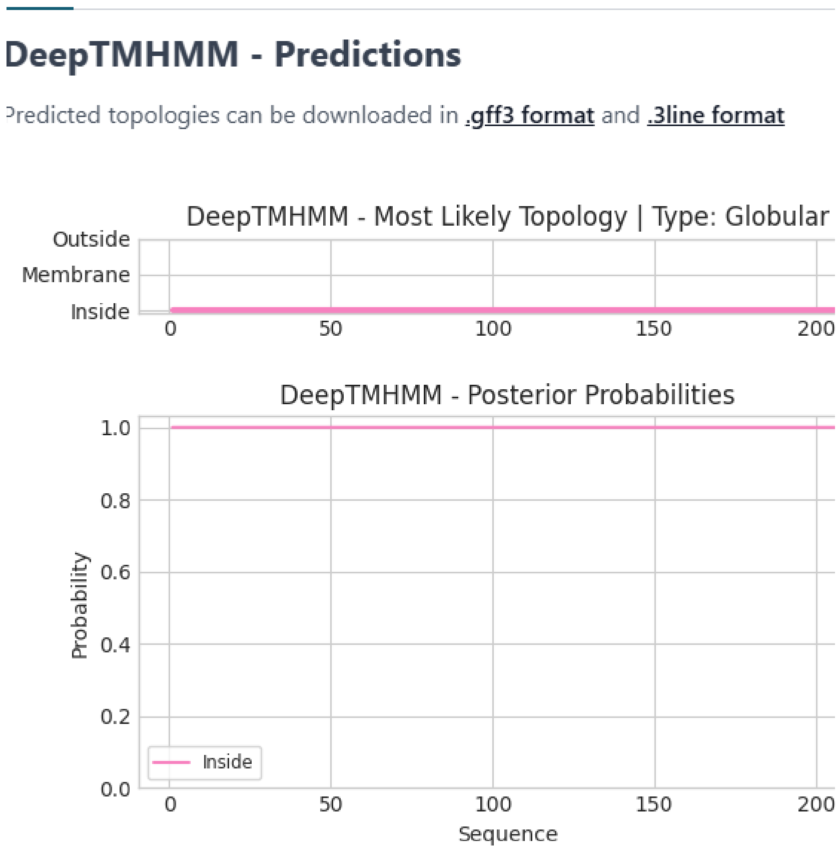
Lauer gene 5 (4024

DNA PROTEIN

MuF-like minor capsid protein

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- According to Deep TMHMM, there were no transmembrane domains.



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- capsid maturation protease
- Matches D29_gp15 in HHPRED, the example for capsid maturation protease.

Feature 7 Stop 5930

Glimmer/GeneMark

What feature number is this? 7

What is the stop site? 5930

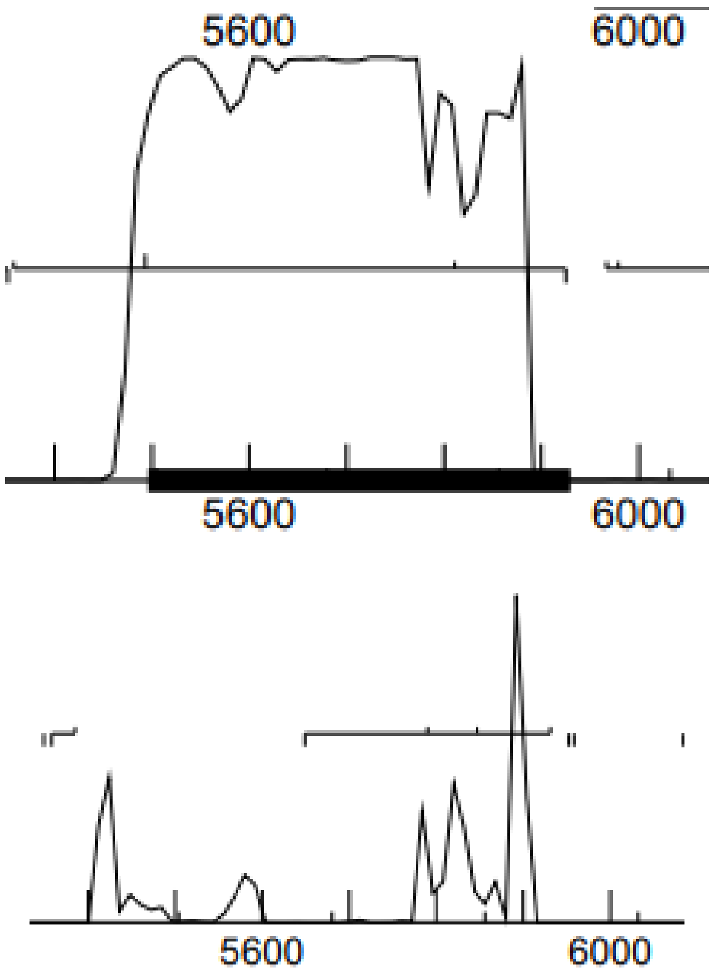
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? Glimmer and GeneMark both call the start at 5496.

What is the autoannotated start? 5496

Gap: _______or overlap: _44_none_____ (with gene in front of it) for the autoannotated start

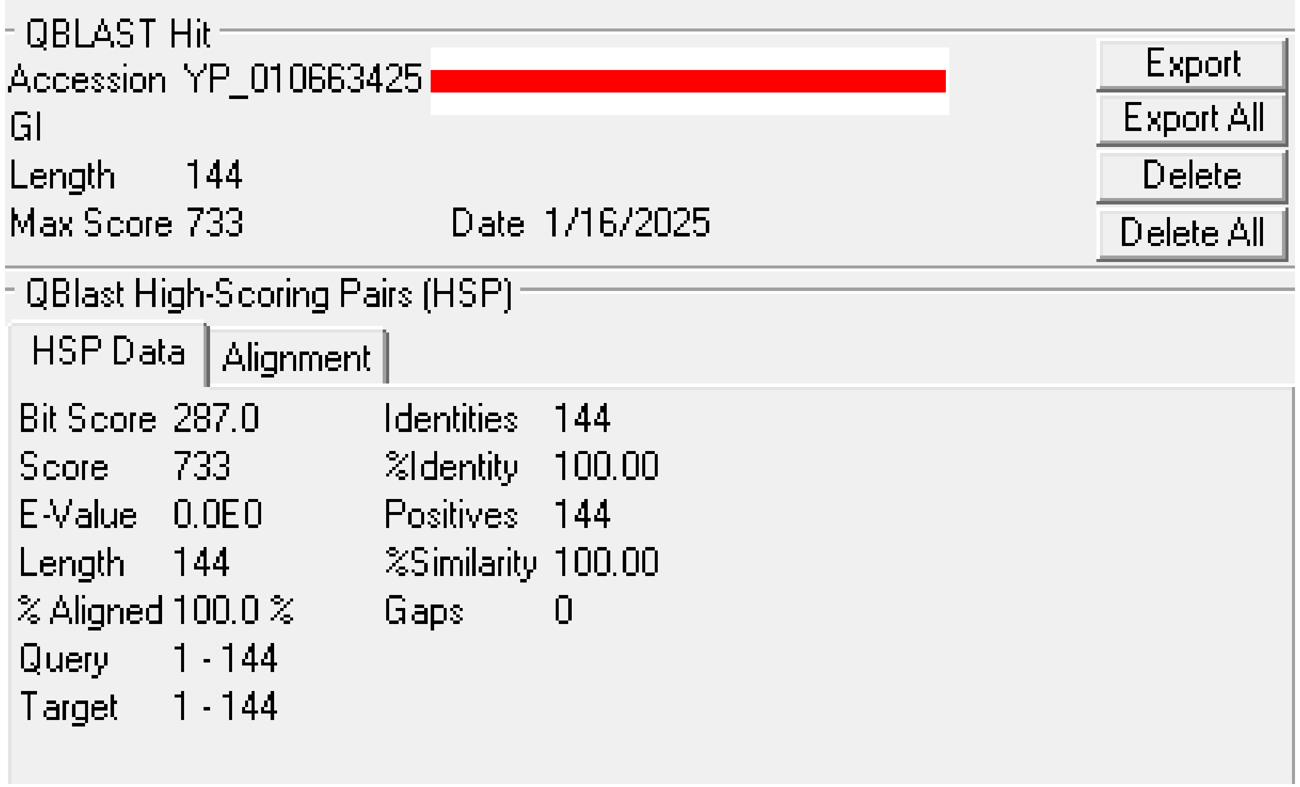
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- There is strong cp until the stopping point from about 5450-5890. Start 5496 excludes about 50 nucleotides. The bottom shows one other reading frame with very weak cp except for at the very end (5930) there is a very strong peak. The reading frame I chose to base cp off is the top picture.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There is an E Value of 0.0E0 for > 25 similar genes.



QBLAST Hit

Accession YP_010663425

GI

Length 144

Max Score 733

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

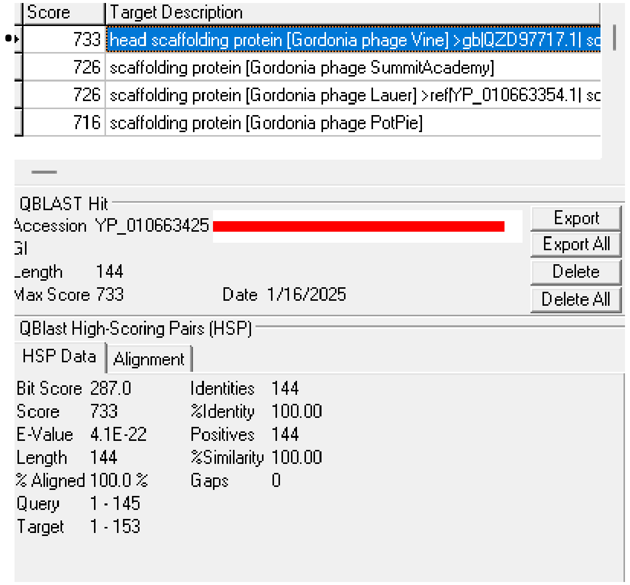
Bit Score	287.0	Identities	144
Score	733	%Identity	100.00
E-Value	0.0E0	Positives	144
Length	144	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 144		
Target	1 - 144		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, feature 9 is a gene because there is strong CP, there are over 25 genes with similar BLAST hits with an E Value of 0.0E0, and Glimmer and GeneMark both call it a gene.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- Starting site 5496 has at least 7 1:1 alignments. (Vine, SummitAcademy,PotPie)
- Starting site 5427 has 2 1:1 alignments (Fabs8)
- The top screenshot is from BLAST in DNA Master while the bottom screenshot is from Blast from NCBI



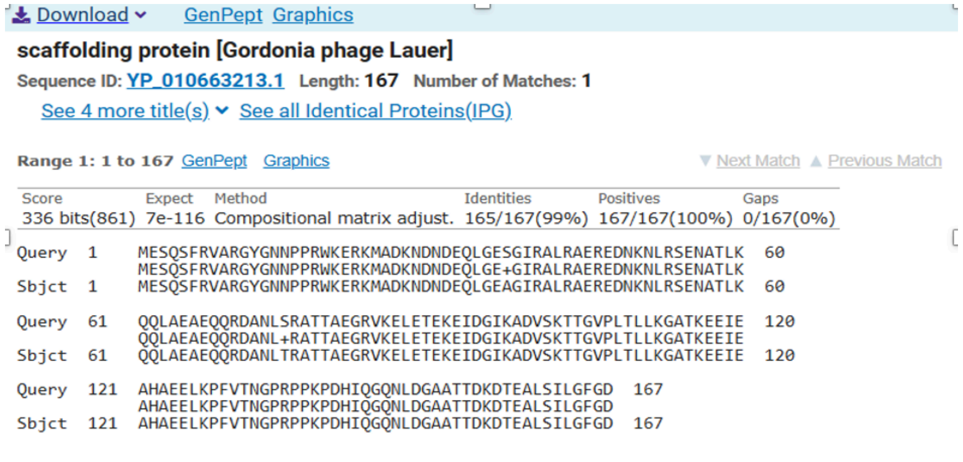
Score	Target Description
733	head scaffolding protein [Gordonia phage Vine] >gb|QZD97717.1| sc
726	scaffolding protein [Gordonia phage SummitAcademy]
726	scaffolding protein [Gordonia phage Lauer] >ref|YP_010663354.1| sc
716	scaffolding protein [Gordonia phage PotPie]

QBLAST Hit

Accession YP_010663425
GI
Length 144
Max Score 733 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

```
Bit Score 287.0      Identities  144
Score     733        %Identity   100.00
E-Value   4.1E-22    Positives   144
Length    144        %Similarity 100.00
% Aligned 100.0 %    Gaps        0
Query     1 - 145
Target    1 - 153
```



scaffolding protein [Gordonia phage Lauer]

Sequence ID: YP_010663213.1 Length: 167 Number of Matches: 1

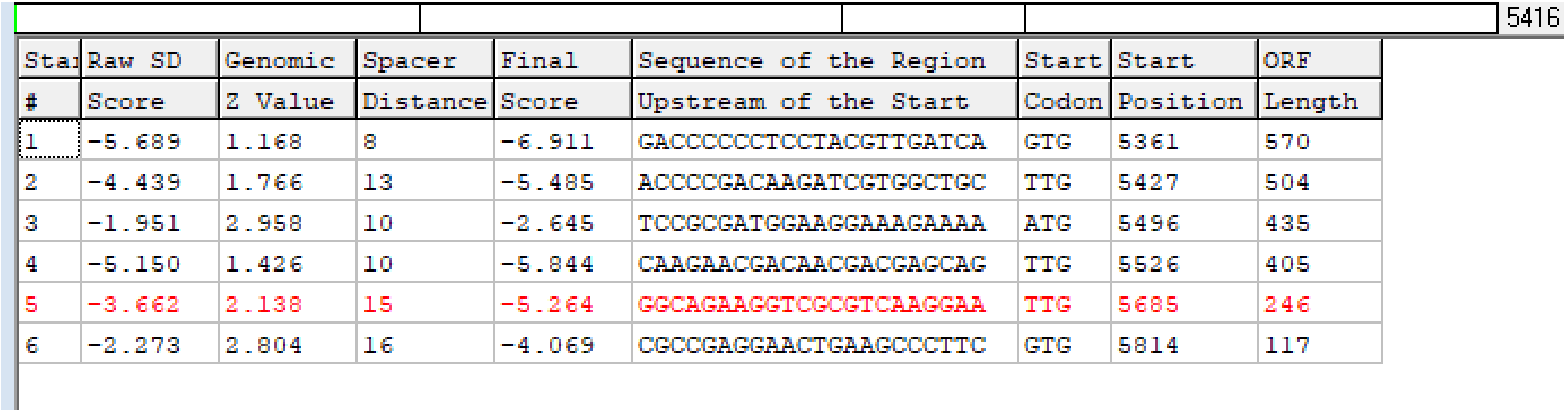
See 4 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 167 GenPept Graphics

```
Score          Expect  Method                          Identities      Positives        Gaps
336 bits(861)  7e-116  Compositional matrix adjust.  165/167(99%)  167/167(100%)  0/167(0%)

Query  1    MESQSFRVARGYGNNPPRWKERKMADKNDNDEQLGESGIRALRAEREDNKNLRSENATLK  60
            MESQSFRVARGYGNNPPRWKERKMADKNDNDEQLGE+GIRALRAEREDNKNLRSENATLK
Sbjct  1    MESQSFRVARGYGNNPPRWKERKMADKNDNDEQLGEAGIRALRAEREDNKNLRSENATLK  60

Query  61   QQLAEAEQQRDANLSRATTAEGRVKELETEKEIDGIKADVSKTTGVPLTLLKGATKEEIE  120
            QQLAEAEQQRDANL+RATTAEGRVKELETEKEIDGIKADVSKTTGVPLTLLKGATKEEIE
Sbjct  61   QQLAEAEQQRDANLTRATTAEGRVKELETEKEIDGIKADVSKTTGVPLTLLKGATKEEIE  120

Query  121  AHAEELKPFVTNGPRPPKPDHIQGQNLDGAATTDKDTEALSILGFGD  167
            AHAEELKPFVTNGPRPPKPDHIQGQNLDGAATTDKDTEALSILGFGD
Sbjct  121  AHAEELKPFVTNGPRPPKPDHIQGQNLDGAATTDKDTEALSILGFGD  167
```

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

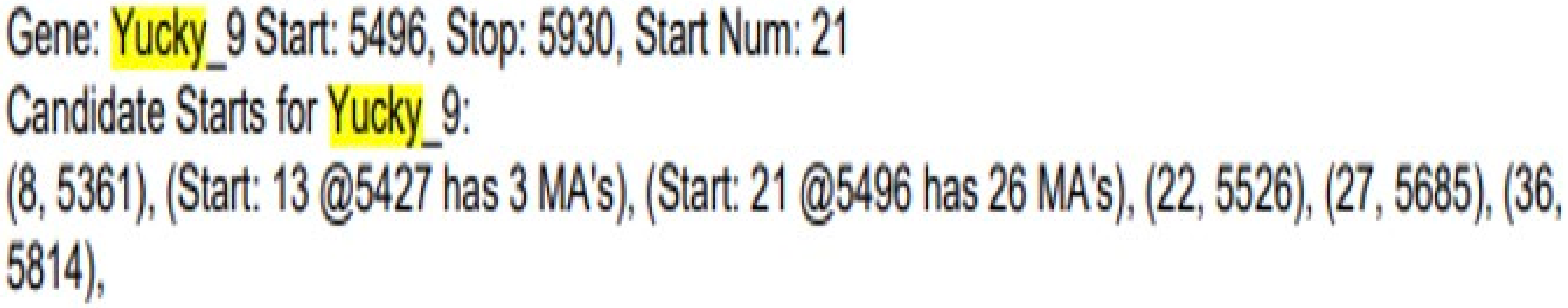
- 5496: The z value is 2.958 and The fs is -2.645
- 5427: The z value is 1.766 and the fs is –5.485
- Based only RBS values start site 5496 is favored because the z value is closer to 3 compared to the z value of start site 5427 and the fs is closer to 0 than the fs of start site 5427.



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	-5.689	1.168	8	-6.911	GACCCCCCTCCTACGTTGATCA	GTG	5361	570
2	-4.439	1.766	13	-5.485	ACCCCGACAAGATCGTGGCTGC	TTG	5427	504
3	-1.951	2.958	10	-2.645	TCCGCGATGGAAGGAAAGAAAA	ATG	5496	435
4	-5.150	1.426	10	-5.844	CAAGAACGACAACGACGAGCAG	TTG	5526	405
5	-3.662	2.138	15	-5.264	GGCAGAAGGTCGCGTCAAGGAA	TTG	5685	246
6	-2.273	2.804	16	-4.069	CGCCGAGGAACTGAAGCCCTTC	GTG	5814	117

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- At starting point 5496 there are 26 MAs
- At starting point 5427 there are 3MAs



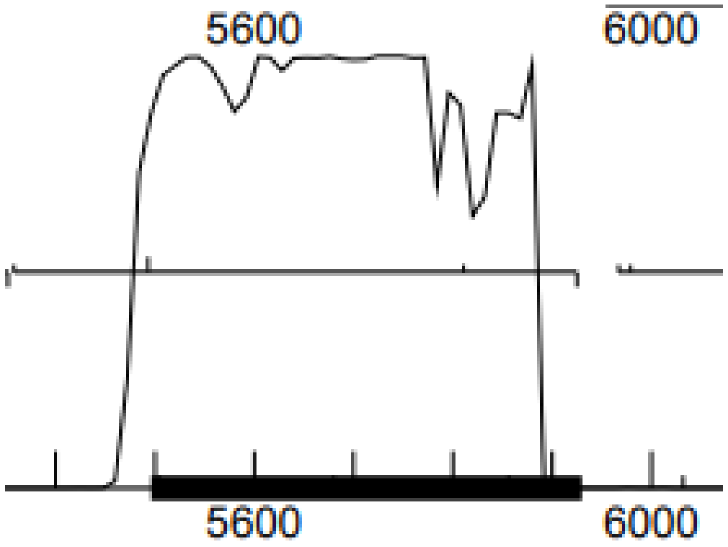
Gene: Yucky_9 Start: 5496, Stop: 5930, Start Num: 21

Candidate Starts for Yucky_9:

(8, 5361), (Start: 13 @5427 has 3 MA's), (Start: 21 @5496 has 26 MA's), (22, 5526), (27, 5685), (36, 5814),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Starting Point 5496 cuts off about 50 nucleotides.
- Starting Point 5427 includes all nucleotides.



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- There is a gap off 44 with no overlap for starting site 5496
- There is an overlap of 25 for starting site 5427

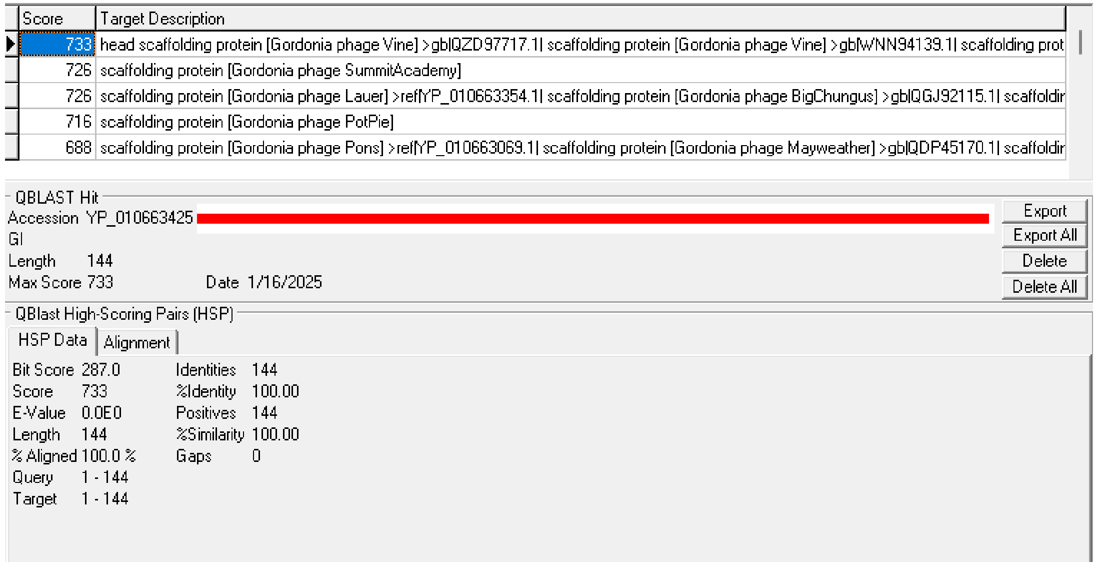
	5496	5427
GeneMark/Glimmer	Both GeneMark and Glimmer call this the start.	Glimmer nor GeneMark called this as astart
Coding Potential	There is strong cp , but it cuts off about 50 nucleotides	Includes all nucleotides
RBS	The z value is 2.958 The fs is -2.645	The z value is 1.766 The fs is –5.435
Blast	There are >10 1:1 alignments	There are 2 1:1 alignments
Starterator	26 MAs	Starteratror called this starting point. 3MAs
Gap/Overlap	There is a gap of 44	There is an overlap of 25

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site of feature 9 is 5496 as both Glimmer and GeneMark agreed upon this start site. There is a strong cp but starting at 5496 cuts off about 50 nucleotides. The RBS values are also in the range they are supposed to be. The z value is 2.958 and the fs is -2.645. Blast also called for >10 1:1 alignments. Starterator also stated that there was 26 MAs. There was a gap of 44. Start site 5427 does include all cp and has a gap of 25 but Glimmer and GeneMark did not call it a gene, there are only 2 1:1 alignments compared to 5496 which has more than 10 1:1 alignments, and only 3 MAs compared to 5496 which has 26s MAs.

BLAST function evidence. What assigned functions do other highly similar genes have?

- There are 19 “Scaffolding proteins” and 6 “Head scaffolding proteins”



Score	Target Description
733	head scaffolding protein [Gordonia phage Vine] >gb|QZD97717.1| scaffolding protein [Gordonia phage Vine] >gb|WNN94139.1| scaffolding prot
726	scaffolding protein [Gordonia phage SummitAcademy]
726	scaffolding protein [Gordonia phage Lauer] >ref|YP_010663354.1| scaffolding protein [Gordonia phage BigChungus] >gb|QGJ92115.1| scaffoldir
716	scaffolding protein [Gordonia phage PotPie]
688	scaffolding protein [Gordonia phage Pons] >ref|YP_010663069.1| scaffolding protein [Gordonia phage Mayweather] >gb|QDP45170.1| scaffoldir

QBLAST Hit

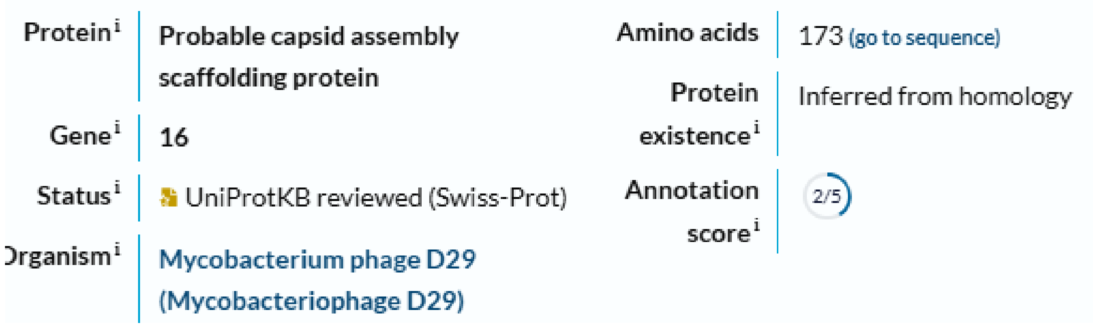
Accession YP_010663425
GI
Length 144
Max Score 733 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

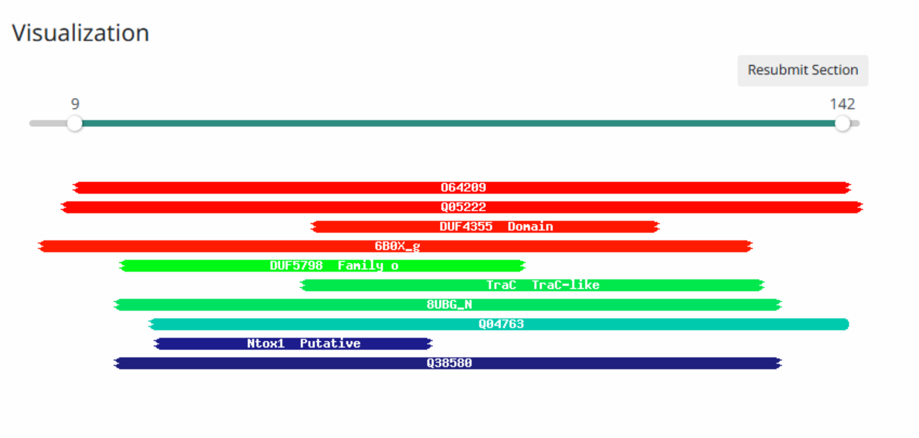
```
Bit Score  287.0       Identities   144
Score      733         %Identity    100.00
E-Value    0.0E0       Positives    144
Length     144         %Similarity  100.00
% Aligned  100.0 %     Gaps         0
Query      1 - 144
Target     1 - 144
```

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

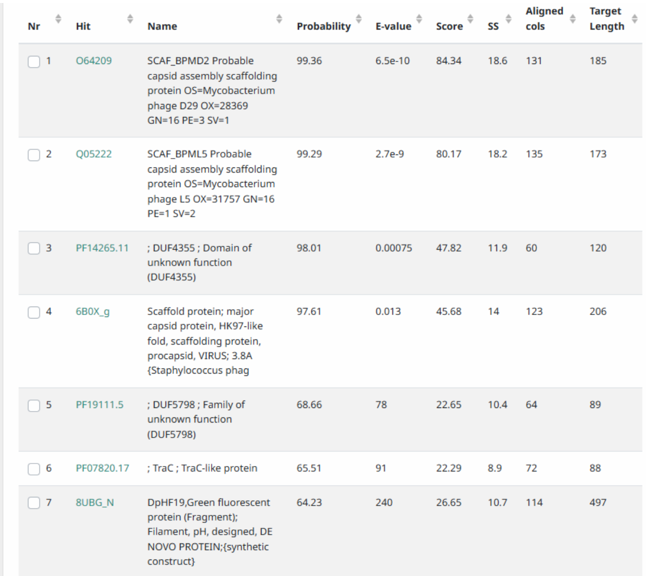
- There are 24 hits
- The first hit is an assembly Scaffold protein
- Probability is 99.36
- E value is 6.5e^-10
- Score is 84.34
- SS is 18.6
- Aligned Cols 132
- Target Length 185



Protein	Probable capsid assembly scaffolding protein	Amino acids	173 (go to sequence)
Gene	16	Protein existence	Inferred from homology
Status	UniProtKB reviewed (Swiss-Prot)	Annotation score	2/5
Organism	Mycobacterium phage D29 (Mycobacteriophage D29)		

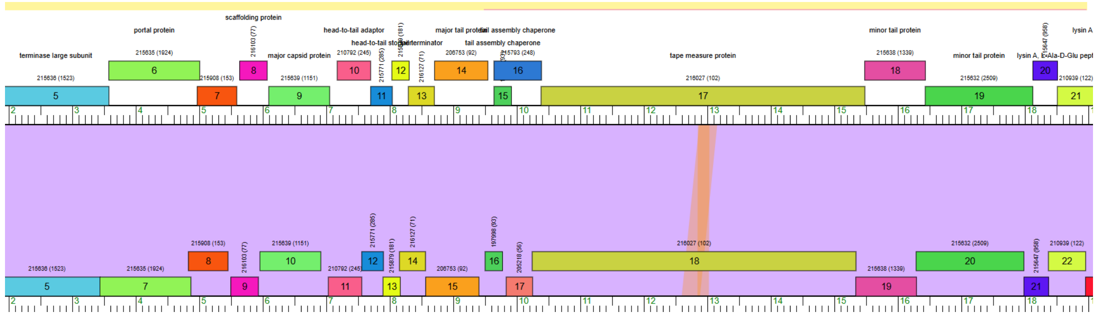


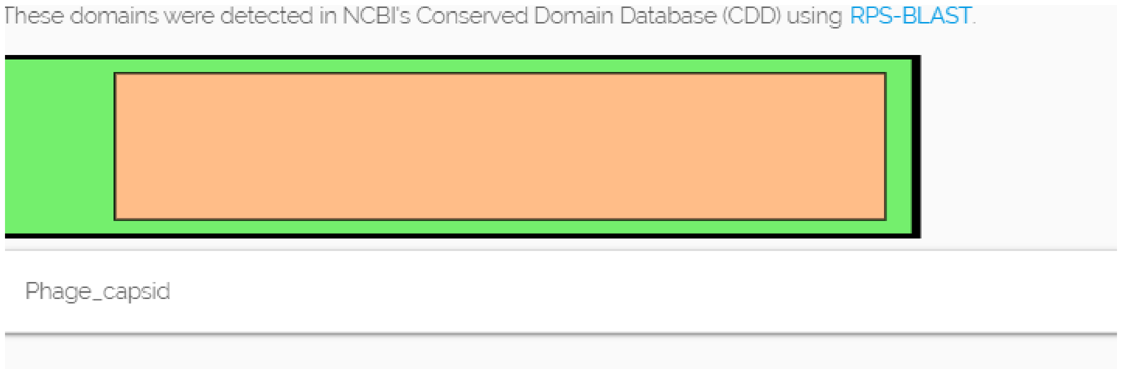
Visualization



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	O64209	SCAF_BPMD2 Probable capsid assembly scaffolding protein OS=Mycobacterium phage D29 OX=28369 GN=16 PE=3 SV=1	99.36	6.5e-10	84.34	18.6	131	185
□ 2	Q05222	SCAF_BPML5 Probable capsid assembly scaffolding protein OS=Mycobacterium phage L5 OX=31757 GN=16 PE=1 SV=2	99.29	2.7e-9	80.17	18.2	135	173
□ 3	PF14265.11	; DUF4355 ; Domain of unknown function (DUF4355)	98.01	0.00075	47.82	11.9	60	120
□ 4	6B0X_g	Scaffold protein; major capsid protein, HK97-like fold, scaffolding protein, procapsid, VIRUS; 3.8A {Staphylococcus phag	97.61	0.013	45.68	14	123	206
□ 5	PF19111.5	; DUF5798 ; Family of unknown function (DUF5798)	68.66	78	22.65	10.4	64	89
□ 6	PF07820.17	; TraC ; TraC-like protein	65.51	91	22.29	8.9	72	88
□ 7	8UBG_N	DpHF19,Green fluorescent protein (Fragment); Filament, pH, designed, DE NOVO PROTEIN;{synthetic construct}	64.23	240	26.65	10.7	114	497

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?




These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

Phage_capsid

- Gene 9 is the same with SummitAcademy and Vine.
- When looking at conserved domains there is one called Phage_capsid.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- Not a hypothetical protein so this evidence is not applicable.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Gene #9 is a scaffolding protein. Blast and Hhpred indicated this as a scaffolding protein while Phamerator gave one conserved domain which was Phage_capsid.

Feature 8 Stop 6913

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

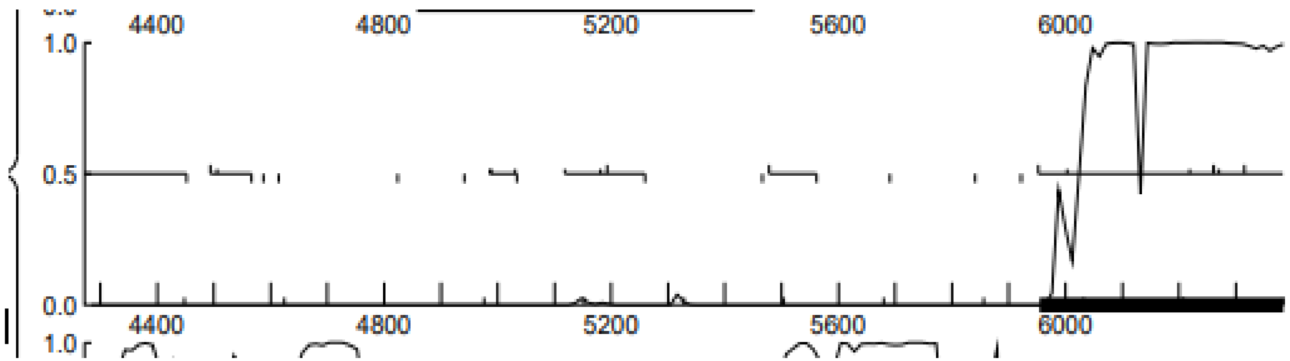
What is the autoannotated start?

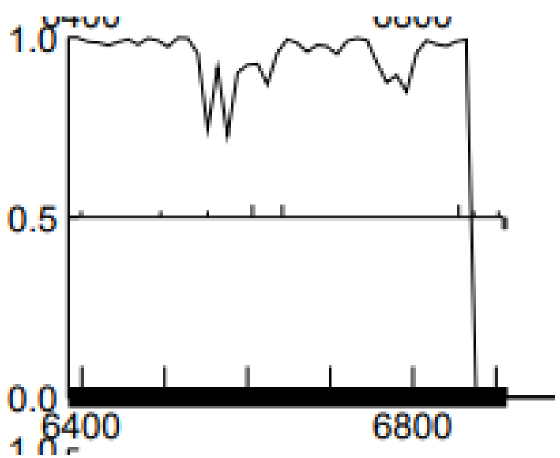
Gap: ___23____ or overlap: _______
(with gene in front of it) for the autoannotated start

- Feature 8
- Stop site: 6913
- Both Glimmer and Genemark call it @bp 5954
- Gap of 23

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Coding potential occurs at 5954
- The coding potential starts at 5954 and cp potential ranges from 5954-6895
- The coding potential is found i reading frame 2 and extends t frame 2 on the next page





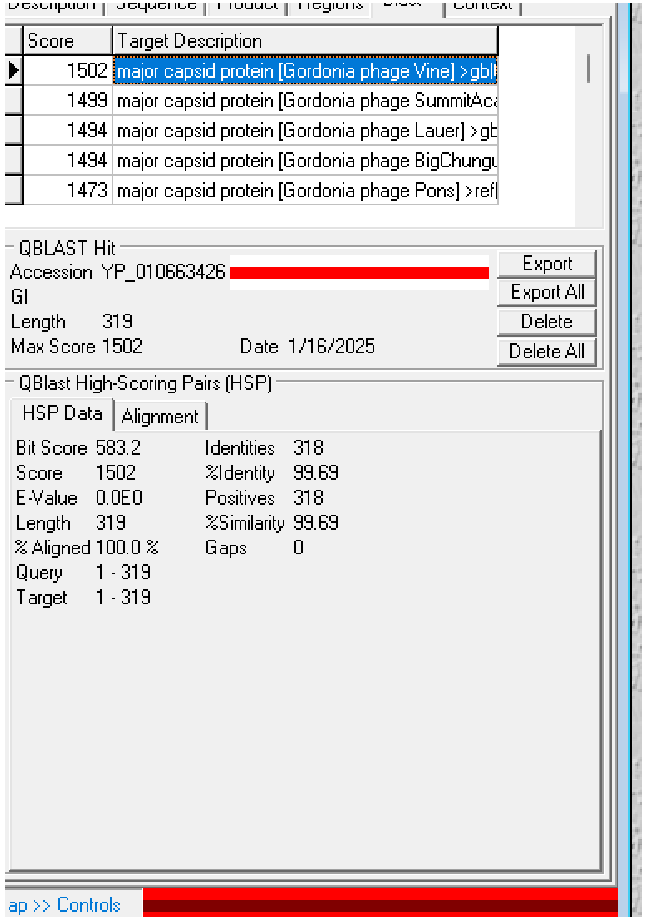
BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 1:1 Alignment with Vine
- 1:1 Alignment with SummitAcademy
- 1:1 Alignment with Lauer
- 1:1 Alignment with BigChungus
- 1:1 Alignment with Pons
- 1:1 Alignment with SheckWes
- 1:1 Alignment with MAnor
- 1:1 Alignment with CherryonLim
- 1:1 Alignment with GTE2

9 1:1 alignments

25 highly similar genes with 0E0:

- Vine
- SummitAcademy
- Lauer
- BigChungus
- Pons
- SheckWes
- MAnor
- CherryonLim
- GTE2
- Amok
- Emalyn
- SteamedHams
- AndPeggy
- AikoCarson
- BillDoor
- Nodigi
- Yakult
- Orla
- Yummy
- Cozz
- Troje
- Button
- GiKK
- Margaret
- MunkgeeRoachy



Score	Target Description
1502	major capsid protein [Gordonia phage Vine] >gbl
1499	major capsid protein [Gordonia phage SummitAca
1494	major capsid protein [Gordonia phage Lauer] >gb
1494	major capsid protein [Gordonia phage BigChungu
1473	major capsid protein [Gordonia phage Pons] >ref|

QBLAST Hit
Accession YP_010663426
GI
Length 319
Max Score 1502 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 583.2
Score 1502
E-Value 0.0E0
Length 319
% Aligned 100.0 %
Query 1 - 319
Target 1 - 319

Identities 318
%Identity 99.69
Positives 318
%Similarity 99.69
Gaps 0

Screenshot of Vine that has a 1:1 alignment with Yucky and has an E-value of 0.0E0 making it a highly similar gene to Yucky as well

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Feature 10 is definitely a gene, because both Glimmer and Genemark agree on the start site to be 5954. Feature 10 has a gap of 23 with feature 9, and there is strong coding potential from 5954-6895.

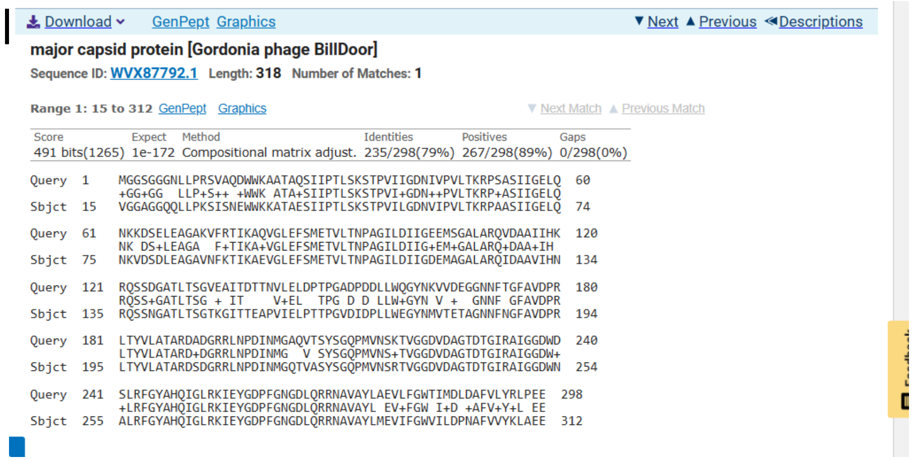
- Feature 10 also has 100% alignment with Vine, SummitAcademy, Lauer, BigChungus, Pons, SheckWes, Manor, CherryonLim, and GTE2 and matches 0E0 value with 25 other highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evide

In DNAM file at start site 5954, there was a 1:1 alignment with 9 other genes. *Start site 5954 is favored*

However, on the NCBI website for start site 6005, I counted that 15 genes had a 1:18 alignment with feature 10 of Yucky

6 genes had a 1:15 alignment with feature 10 of Yucky on NCBI website for start site 6005

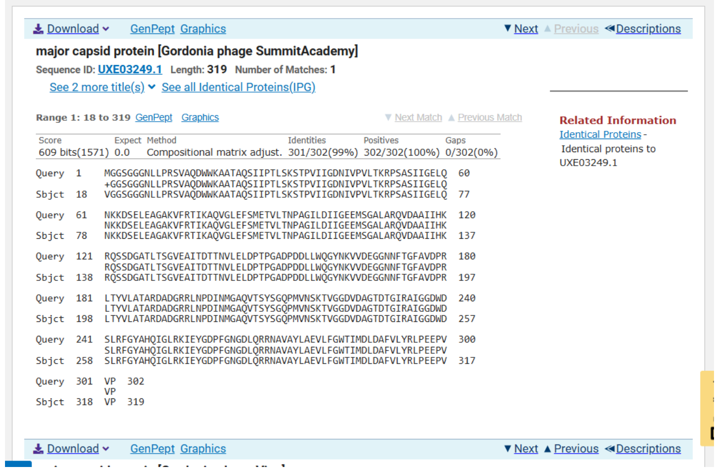


major capsid protein [Gordonia phage BillDoor]

Sequence ID: WVX87792.1 Length: 318 Number of Matches: 1

Range 1: 15 to 312 GenPept Graphics

```
Score           Expect   Method                          Identities      Positives       Gaps
491 bits(1265)  1e-172   Compositional matrix adjust.    235/298(79%)    267/298(89%)    0/298(0%)

Query  1    MGGSGGGNLLPRSVAQDWWKAATAQSIIPTLSKSTPVIIGDNIVPVLTKRPSASIIGELQ  60
            +GG+GG  LLP+S++ +WWK ATA+SIIPTLSKSTPVI+GDN++PVLTKRP+ASIIGELQ
Sbjct  15   VGGAGGQQLLPKSISNEWWKKATAESIIPTLSKSTPVILGDNVIPVLTKRPAASIIGELQ  74

Query  61   NKKDSELEAGAKVFRTIKAQVGLEFSMETVLTNPAGILDIIGEEMSGALARQVDAAIIHK  120
            NK DS+LEAGA  F+TIKA+VGLEFSMETVLTNPAGILDIIG+EM+GALARQ+DAA+IH
Sbjct  75   NKVDSDLEAGAVNFKTIKAEVGLEFSMETVLTNPAGILDIIGDEMAGALARQIDAAVIHN  134

Query  121  RQSSDGATLTSGVEAITDTTNVLELDPTPGADPDDLLWQGYNKVVDEGGNNFTGFAVDPR  180
            RQSS+GATLTSG + IT    V+EL  TPG D D LLW+GYN V +  GNNF GFAVDPR
Sbjct  135  RQSSNGATLTSGTKGITTEAPVIELPTTPGVDIDPLLWEGYNMVTETAGNNFNGFAVDPR  194

Query  181  LTYVLATARDADGRRLNPDINMGAQVTSYSGQPMVNSKTVGGDVDAGTDTGIRAIGGDWD  240
            LTYVLATARD+DGRRLNPDINMG  V SYSGQPMVNS+TVGGDVDAGTDTGIRAIGGDW+
Sbjct  195  LTYVLATARDSDGRRLNPDINMGQTVASYSGQPMVNSRTVGGDVDAGTDTGIRAIGGDWN  254

Query  241  SLRFGYAHQIGLRKIEYGDPFGNGDLQRRNAVAYLAEVLFGWTIMDLDAFVLYRLPEE  298
            +LRFGYAHQIGLRKIEYGDPFGNGDLQRRNAVAYL EV+FGW I+D +AFV+Y+L EE
Sbjct  255  ALRFGYAHQIGLRKIEYGDPFGNGDLQRRNAVAYLMEVIFGWVILDPNAFVVYKLAEE  312
```



major capsid protein [Gordonia phage SummitAcademy]

Sequence ID: UXE03249.1 Length: 319 Number of Matches: 1

See 2 more title(s) See all Identical Proteins(IPG)

Range 1: 18 to 319 GenPept Graphics

```
Score           Expect  Method                          Identities      Positives        Gaps
609 bits(1571)  0.0     Compositional matrix adjust.    301/302(99%)    302/302(100%)    0/302(0%)

Query  1    MGGSGGGNLLPRSVAQDWWKAATAQSIIPTLSKSTPVIIGDNIVPVLTKRPSASIIGELQ  60
            +GGSGGGNLLPRSVAQDWWKAATAQSIIPTLSKSTPVIIGDNIVPVLTKRPSASIIGELQ
Sbjct  18   VGGSGGGNLLPRSVAQDWWKAATAQSIIPTLSKSTPVIIGDNIVPVLTKRPSASIIGELQ  77

Query  61   NKKDSELEAGAKVFRTIKAQVGLEFSMETVLTNPAGILDIIGEEMSGALARQVDAAIIHK  120
            NKKDSELEAGAKVFRTIKAQVGLEFSMETVLTNPAGILDIIGEEMSGALARQVDAAIIHK
Sbjct  78   NKKDSELEAGAKVFRTIKAQVGLEFSMETVLTNPAGILDIIGEEMSGALARQVDAAIIHK  137

Query  121  RQSSDGATLTSGVEAITDTTNVLELDPTPGADPDDLLWQGYNKVVDEGGNNFTGFAVDPR  180
            RQSSDGATLTSGVEAITDTTNVLELDPTPGADPDDLLWQGYNKVVDEGGNNFTGFAVDPR
Sbjct  138  RQSSDGATLTSGVEAITDTTNVLELDPTPGADPDDLLWQGYNKVVDEGGNNFTGFAVDPR  197

Query  181  LTYVLATARDADGRRLNPDINMGAQVTSYSGQPMVNSKTVGGDVDAGTDTGIRAIGGDWD  240
            LTYVLATARDADGRRLNPDINMGAQVTSYSGQPMVNSKTVGGDVDAGTDTGIRAIGGDWD
Sbjct  198  LTYVLATARDADGRRLNPDINMGAQVTSYSGQPMVNSKTVGGDVDAGTDTGIRAIGGDWD  257

Query  241  SLRFGYAHQIGLRKIEYGDPFGNGDLQRRNAVAYLAEVLFGWTIMDLDAFVLYRLPEEPV  300
            SLRFGYAHQIGLRKIEYGDPFGNGDLQRRNAVAYLAEVLFGWTIMDLDAFVLYRLPEEPV
Sbjct  258  SLRFGYAHQIGLRKIEYGDPFGNGDLQRRNAVAYLAEVLFGWTIMDLDAFVLYRLPEEPV  317

Query  301  VP  302
            VP
Sbjct  318  VP  319
```

Related Information

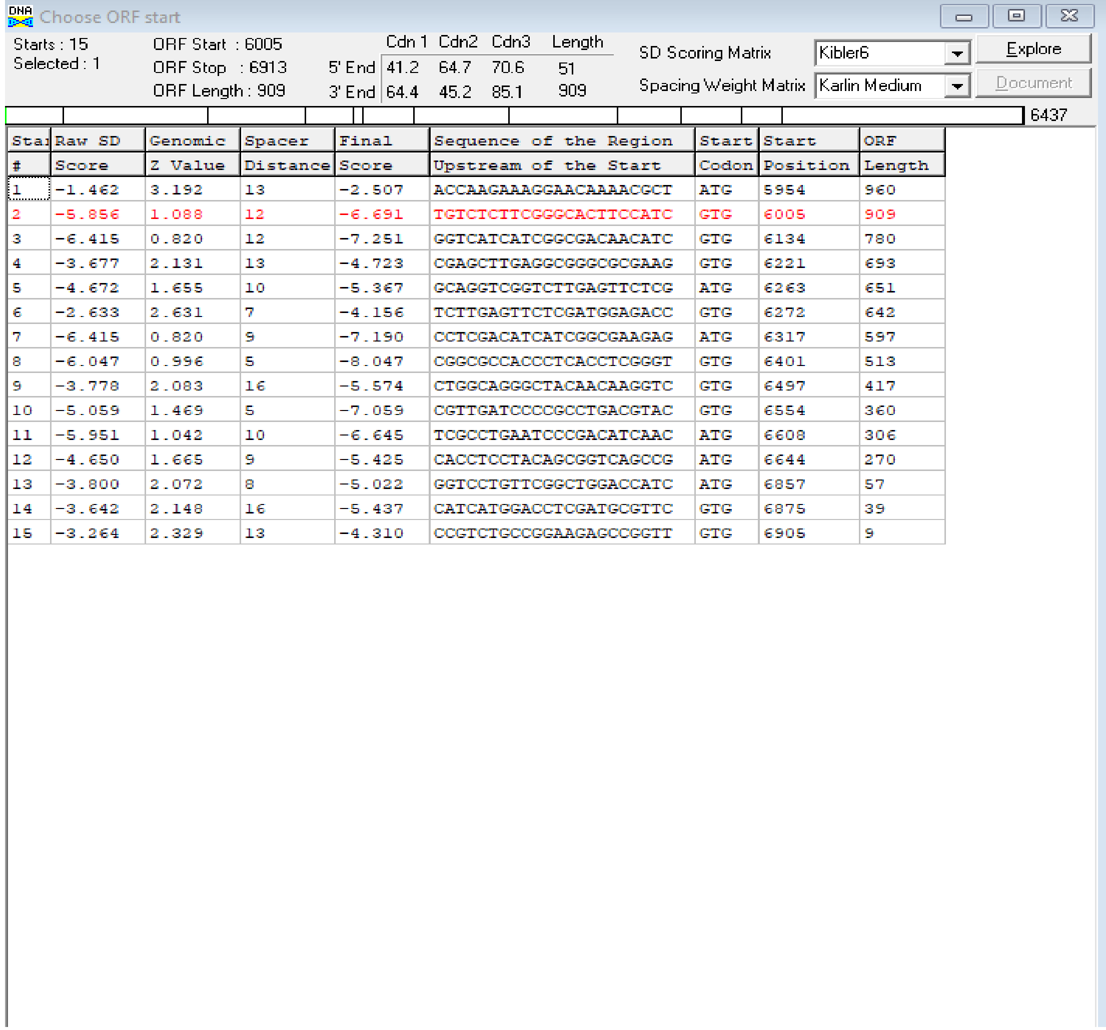
Identical Proteins - Identical proteins to UXE03249.1

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Start site 5954
- Z value: 3.192
- Final score: -2.507

Start site 5954 is favored

- Start site 6005
- Z value: 1.088
- Final score: -6.691



Starts : 15
Selected : 1

ORF Start : 6005
ORF Stop : 6913
ORF Length : 909

	Cdn 1	Cdn2	Cdn3	Length
5' End	41.2	64.7	70.6	51
3' End	64.4	45.2	85.1	909

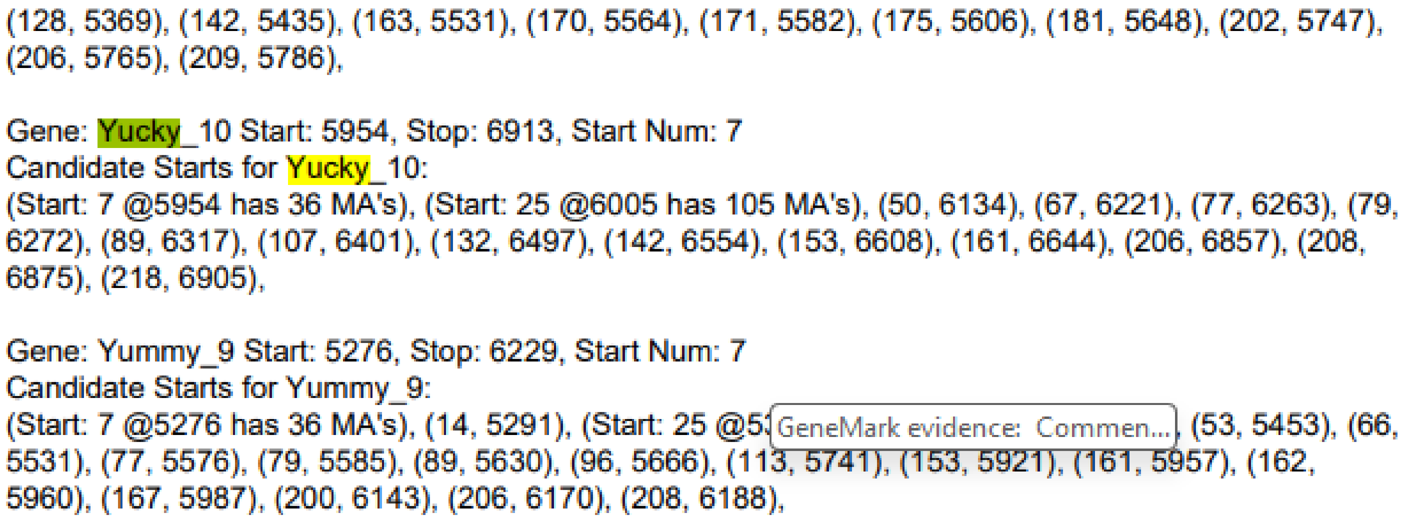
SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

6437

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	-1.462	3.192	13	-2.507	ACCAAGAAAGGAACAAAACGCT	ATG	5954	960
2	-5.856	1.088	12	-6.691	TGTCTCTTCGGGCACTTCCATC	GTG	6005	909
3	-6.415	0.820	12	-7.251	GGTCATCATCGGCGACAACATC	GTG	6134	780
4	-3.677	2.131	13	-4.723	CGAGCTTGAGGCGGGCGCGAAG	GTG	6221	693
5	-4.672	1.655	10	-5.367	GCAGGTCGGTCTTGAGTTCTCG	ATG	6263	651
6	-2.633	2.631	7	-4.156	TCTTGAGTTCTCGATGGAGACC	GTG	6272	642
7	-6.415	0.820	9	-7.190	CCTCGACATCATCGGCGAAGAG	ATG	6317	597
8	-6.047	0.996	5	-8.047	CGGCGCCACCCTCACCTCGGGT	GTG	6401	513
9	-3.778	2.083	16	-5.574	CTGGCAGGGCTACAACAAGGTC	GTG	6497	417
10	-5.059	1.469	5	-7.059	CGTTGATCCCCGCCTGACGTAC	GTG	6554	360
11	-5.951	1.042	10	-6.645	TCGCCTGAATCCCGACATCAAC	ATG	6608	306
12	-4.650	1.665	9	-5.425	CACCTCCTACAGCGGTCAGCCG	ATG	6644	270
13	-3.800	2.072	8	-5.022	GGTCCTGTTCGGCTGGACCATC	ATG	6857	57
14	-3.642	2.148	16	-5.437	CATCATGGACCTCGATGCGTTC	GTG	6875	39
15	-3.264	2.329	13	-4.310	CCGTCTGCCGGAAGAGCCGGTT	GTG	6905	9

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- At start 7 @ 5954 Yucky has 36 MA's
- At start 25 @ 6005 Yucky has 105 MA's



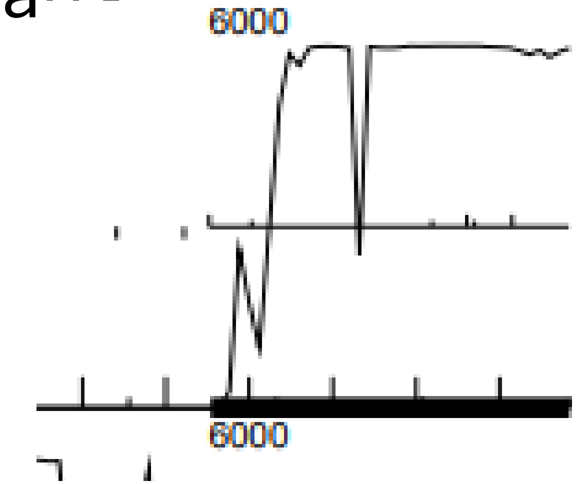
(128, 5369), (142, 5435), (163, 5531), (170, 5564), (171, 5582), (175, 5606), (181, 5648), (202, 5747), (206, 5765), (209, 5786),

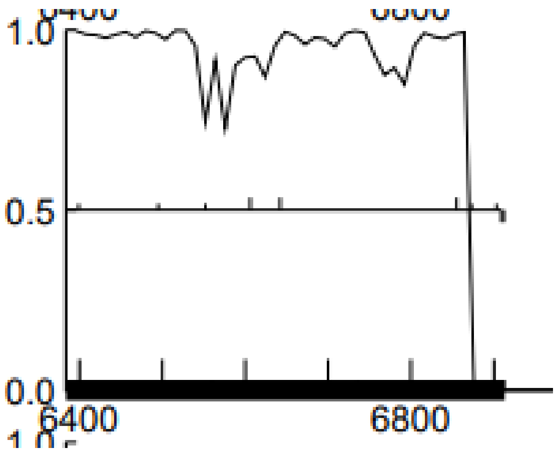
Gene: Yucky_10 Start: 5954, Stop: 6913, Start Num: 7
Candidate Starts for Yucky_10:
(Start: 7 @5954 has 36 MA's), (Start: 25 @6005 has 105 MA's), (50, 6134), (67, 6221), (77, 6263), (79, 6272), (89, 6317), (107, 6401), (132, 6497), (142, 6554), (153, 6608), (161, 6644), (206, 6857), (208, 6875), (218, 6905),

Gene: Yummy_9 Start: 5276, Stop: 6229, Start Num: 7
Candidate Starts for Yummy_9:
(Start: 7 @5276 has 36 MA's), (14, 5291), (Start: 25 @5 GeneMark evidence: Commen... , (53, 5453), (66, 5531), (77, 5576), (79, 5585), (89, 5630), (96, 5666), (113, 5741), (153, 5921), (161, 5957), (162, 5960), (167, 5987), (200, 6143), (206, 6170), (208, 6188),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- At start 7 @5954, all the coding potential is included while at start 25 @6005, the coding potential is cut off but cp still exists. The frame is extended to other page. As discussed in class, there is not that much of a significant difference for cp between the 2 potential starts





Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Start at 5954:
- Has gap of 23

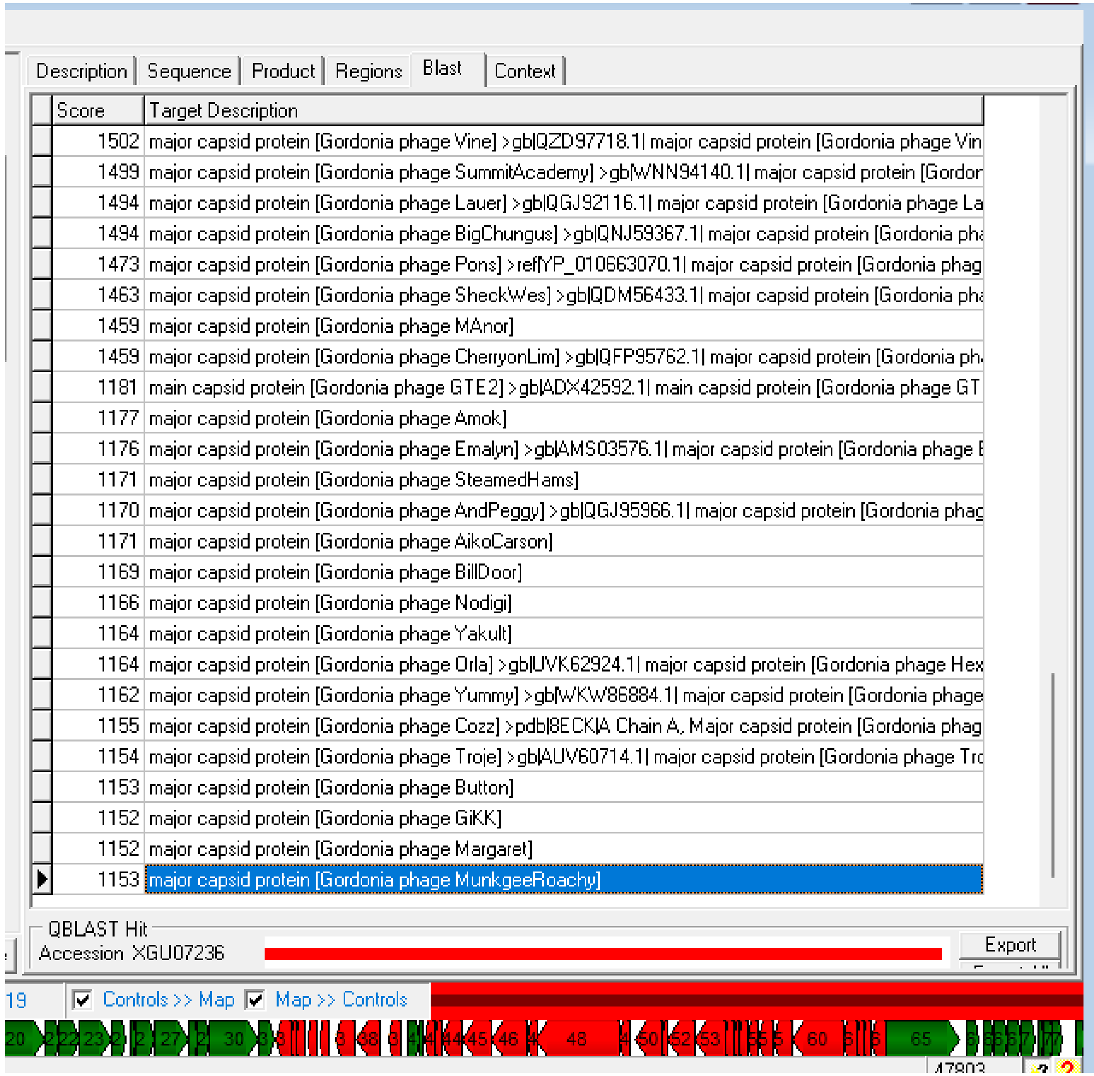
- Start at 6005:
- Has gap of 74

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	5954	6005
Genemark	Glimmer & Genemark	Nothing
Coding potential	Includes all cp	Cuts off peak of cp
RBS	Z value: 3.192 Final score: -2.607	Z value: 1.088 Final score: -6.691
BLAST	9 1:1 alignments	6 1:15 alignments 15 1:18 alignments
Starterator	36 MA	105 MA
Gap	23	74

While it was a close call for 5954 and 6005, 5954 is considered the best start as it was called by both Glimmer and Genemark. The start site 5954 also includes all coding potential. The Z score was also greater than 1 and had a final score closer to 0. The 5954 start also had the highest number of 1:1 alignments and has the smaller gap. The only evidence to support the 6005 start was that it had the highest number of manual annotations at 105.

BLAST function evidence. What assigned functions do other highly similar genes have?



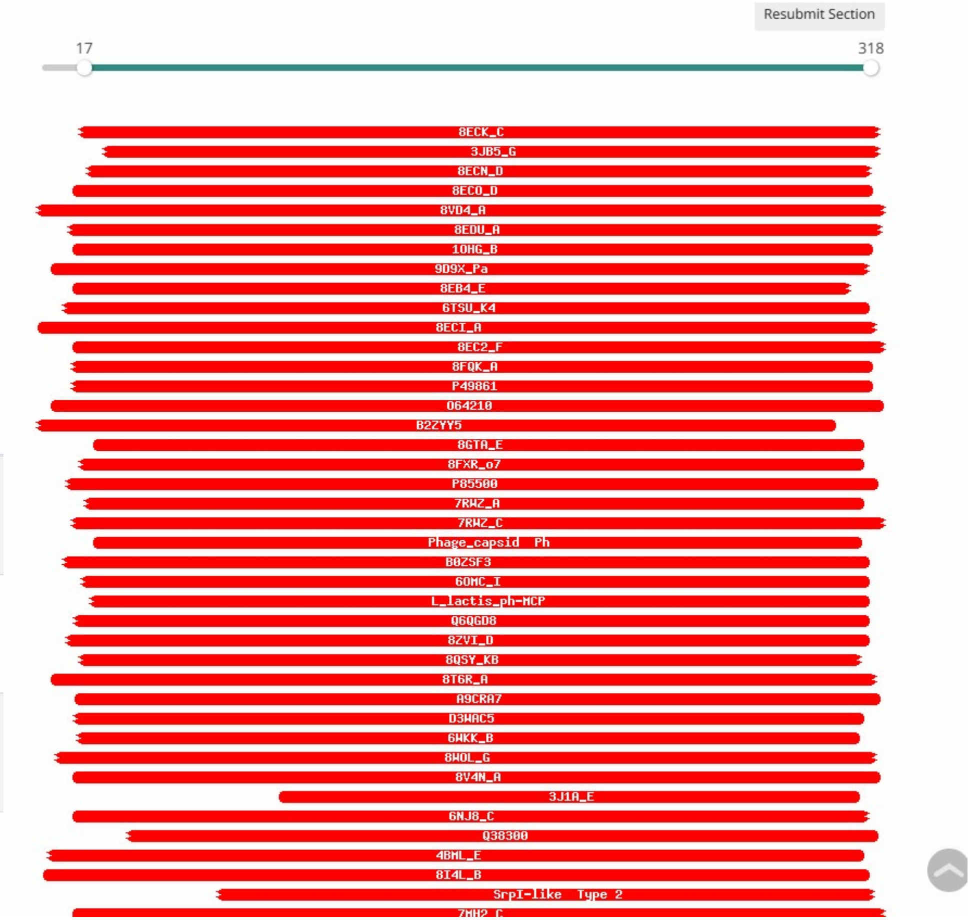
Score	Target Description
1502	major capsid protein [Gordonia phage Vine] >gb|QZD97718.1| major capsid protein [Gordonia phage Vin
1499	major capsid protein [Gordonia phage SummitAcademy] >gb|WNN94140.1| major capsid protein [Gordor
1494	major capsid protein [Gordonia phage Lauer] >gb|QGJ92116.1| major capsid protein [Gordonia phage La
1494	major capsid protein [Gordonia phage BigChungus] >gb|QNJ59367.1| major capsid protein [Gordonia pha
1473	major capsid protein [Gordonia phage Pons] >ref|YP_010663070.1| major capsid protein [Gordonia phag
1463	major capsid protein [Gordonia phage SheckWes] >gb|QDM56433.1| major capsid protein [Gordonia pha
1459	major capsid protein [Gordonia phage MAnor]
1459	major capsid protein [Gordonia phage CherryonLim] >gb|QFP95762.1| major capsid protein [Gordonia ph
1181	main capsid protein [Gordonia phage GTE2] >gb|ADX42592.1| main capsid protein [Gordonia phage GT
1177	major capsid protein [Gordonia phage Amok]
1176	major capsid protein [Gordonia phage Emalyn] >gb|AMS03576.1| major capsid protein [Gordonia phage E
1171	major capsid protein [Gordonia phage SteamedHams]
1170	major capsid protein [Gordonia phage AndPeggy] >gb|QGJ95966.1| major capsid protein [Gordonia phag
1171	major capsid protein [Gordonia phage AikoCarson]
1169	major capsid protein [Gordonia phage BillDoor]
1166	major capsid protein [Gordonia phage Nodigi]
1164	major capsid protein [Gordonia phage Yakult]
1164	major capsid protein [Gordonia phage Orla] >gb|UVK62924.1| major capsid protein [Gordonia phage Hex
1162	major capsid protein [Gordonia phage Yummy] >gb|WKW86884.1| major capsid protein [Gordonia phage
1155	major capsid protein [Gordonia phage Cozz] >pdb|8ECK|A Chain A, Major capsid protein [Gordonia phag
1154	major capsid protein [Gordonia phage Troje] >gb|AUV60714.1| major capsid protein [Gordonia phage Tro
1153	major capsid protein [Gordonia phage Button]
1152	major capsid protein [Gordonia phage GiKK]
1152	major capsid protein [Gordonia phage Margaret]
1153	major capsid protein [Gordonia phage MunkgeeRoachy]

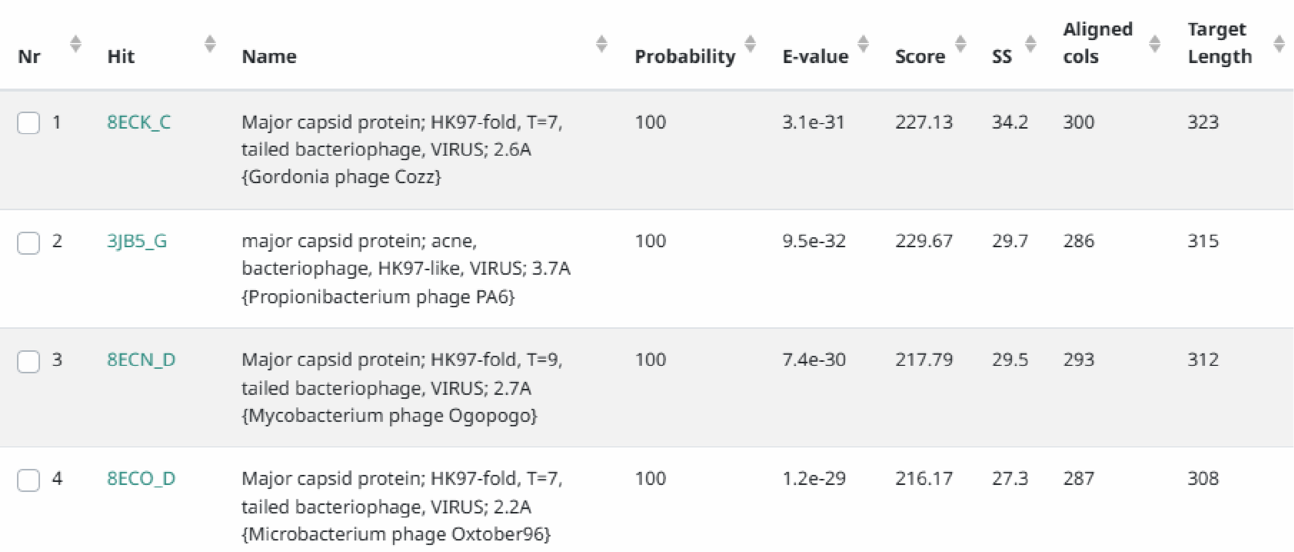
QBLAST Hit
Accession XGU07236

- Has 25 highly similar genes with "major capsid protein"

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- Highly similar matches: all major capsid protein

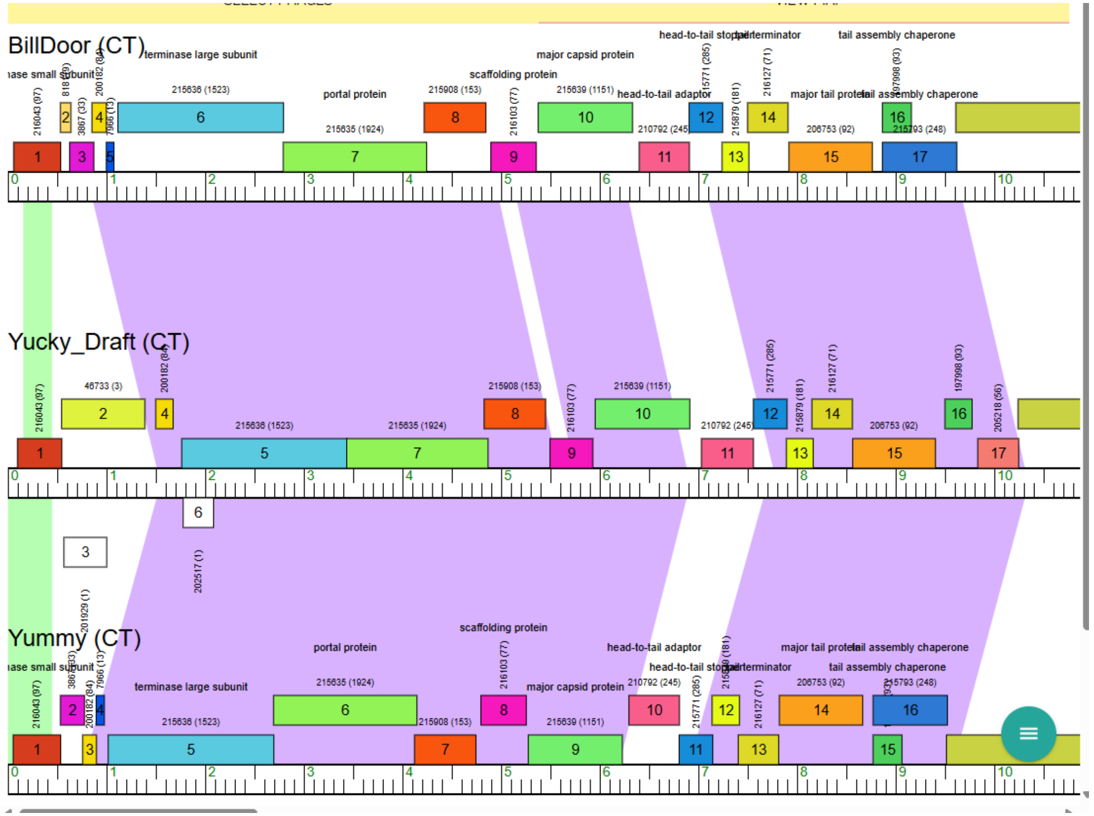




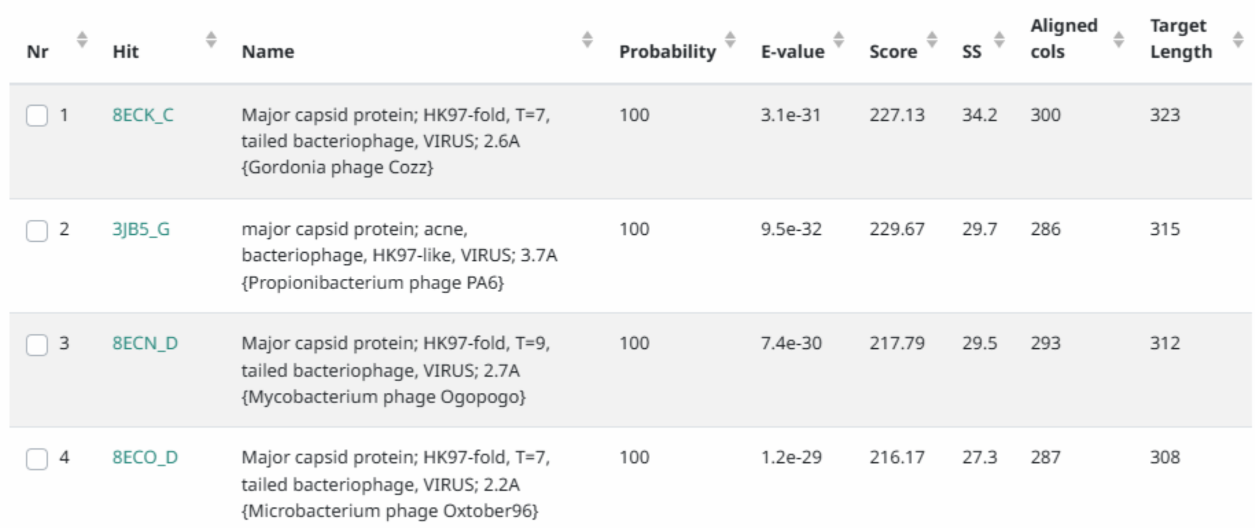
Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
☐ 1	8ECK_C	Major capsid protein; HK97-fold, T=7, tailed bacteriophage, VIRUS; 2.6A {Gordonia phage Cozz}	100	3.1e-31	227.13	34.2	300	323
☐ 2	3JB5_G	major capsid protein; acne, bacteriophage, HK97-like, VIRUS; 3.7A {Propionibacterium phage PA6}	100	9.5e-32	229.67	29.7	286	315
☐ 3	8ECN_D	Major capsid protein; HK97-fold, T=9, tailed bacteriophage, VIRUS; 2.7A {Mycobacterium phage Ogopogo}	100	7.4e-30	217.79	29.5	293	312
☐ 4	8ECO_D	Major capsid protein; HK97-fold, T=7, tailed bacteriophage, VIRUS; 2.2A {Microbacterium phage Oxtober96}	100	1.2e-29	216.17	27.3	287	308

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

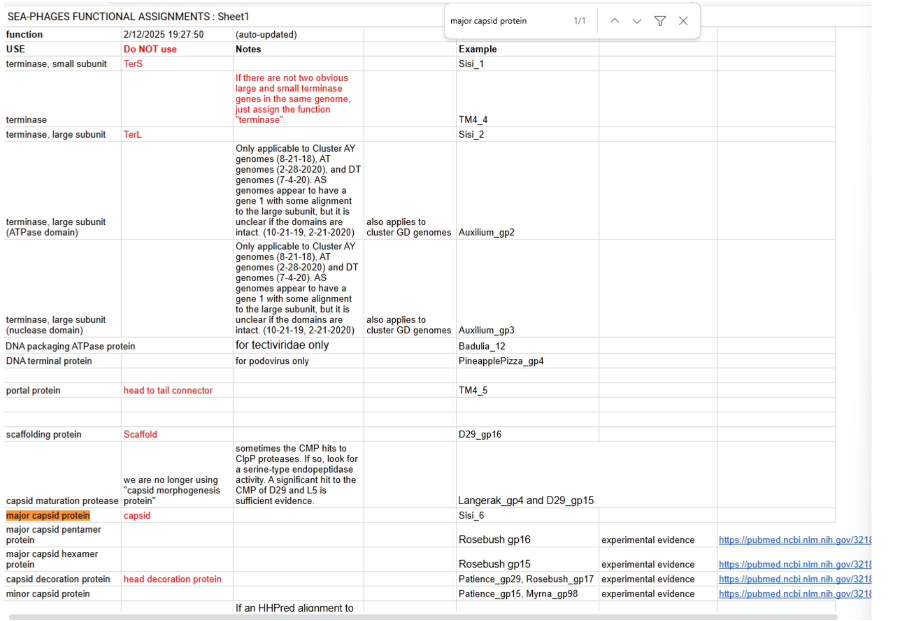
- Yucky feature 10 conserved domain: Phage_capsid function: none
- Yummy feature 9 conserved domain: prophage_Lp3_protein_18 and Phage_capsid function: major capsid protein
- BillDoor feature 10 conserved domain: prophage_Lp3_protein_18 and Phage_capsid function: major capsid protein



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	8ECK_C	Major capsid protein; HK97-fold, T=7, tailed bacteriophage, VIRUS; 2.6A {Gordonia phage Cozz}	100	3.1e-31	227.13	34.2	300	323
□ 2	3JB5_G	major capsid protein; acne, bacteriophage, HK97-like, VIRUS; 3.7A {Propionibacterium phage PA6}	100	9.5e-32	229.67	29.7	286	315
□ 3	8ECN_D	Major capsid protein; HK97-fold, T=9, tailed bacteriophage, VIRUS; 2.7A {Mycobacterium phage Ogopogo}	100	7.4e-30	217.79	29.5	293	312
□ 4	8ECO_D	Major capsid protein; HK97-fold, T=7, tailed bacteriophage, VIRUS; 2.2A {Microbacterium phage Oxtober96}	100	1.2e-29	216.17	27.3	287	308



SEA-PHAGES FUNCTIONAL ASSIGNMENTS : Sheet1

function	2/12/2025 19:27:50	(auto-updated)				
USE	**Do NOT use**	**Notes**		**Example**		
terminase, small subunit	TerS			Sisi_1		
terminase		If there are not two obvious large and small terminase genes in the same genome, just assign the function "terminase".		TM4_4		
terminase, large subunit	TerL			Sisi_2		
terminase, large subunit (ATPase domain)		Only applicable to Cluster AY genomes (8-21-18), AT genomes (2-28-2020), and DT genomes (7-4-20). AS genomes appear to have a gene 1 with some alignment to the large subunit, but it is unclear if the domains are intact. (10-21-19, 2-21-2020)	also applies to cluster GD genomes	Auxilium_gp2		
terminase, large subunit (nuclease domain)		Only applicable to Cluster AY genomes (8-21-18), AT genomes (2-28-2020) and DT genomes (7-4-20). AS genomes appear to have a gene 1 with some alignment to the large subunit, but it is unclear if the domains are intact. (10-21-19, 2-21-2020)	also applies to cluster GD genomes	Auxilium_gp3		
DNA packaging ATPase protein		for tectiviridae only		Badulia_12		
DNA terminal protein		for podovirus only		PineapplePizza_gp4		
portal protein	head to tail connector			TM4_5		
scaffolding protein	Scaffold			D29_gp16		
capsid maturation protease	we are no longer using "capsid morphogenesis protein"	sometimes the CMP hits to ClpP proteases. If so, look for a serine-type endopeptidase activity. A significant hit to the CMP of D29 and L5 is sufficient evidence.		Langerak_gp4 and D29_gp15		
major capsid protein	capsid			Sisi_6		
major capsid pentamer protein				Rosebush gp16	experimental evidence	https://pubmed.ncbi.nlm.nih.gov/3218
major capsid hexamer protein				Rosebush gp15	experimental evidence	https://pubmed.ncbi.nlm.nih.gov/3218
capsid decoration protein	head decoration protein			Patience_gp29, Rosebush_gp17	experimental evidence	https://pubmed.ncbi.nlm.nih.gov/3218
minor capsid protein				Patience_gp15, Myrna_gp98	experimental evidence	https://pubmed.ncbi.nlm.nih.gov/3218
		If an HHPred alignment to				

The function of this Yucky gene is major capsid protein, because the BLAST evidence had 25 similar matches with “major capsid protein”. The Hhpred evidence also had highly similar matches with function as major capsid protein. The Phamerator evidence also gave Yucky its function as two of its highly similar genes, Yummy and BillDoor had the function major capsid protein.

Feature 9 – Stop 7558

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

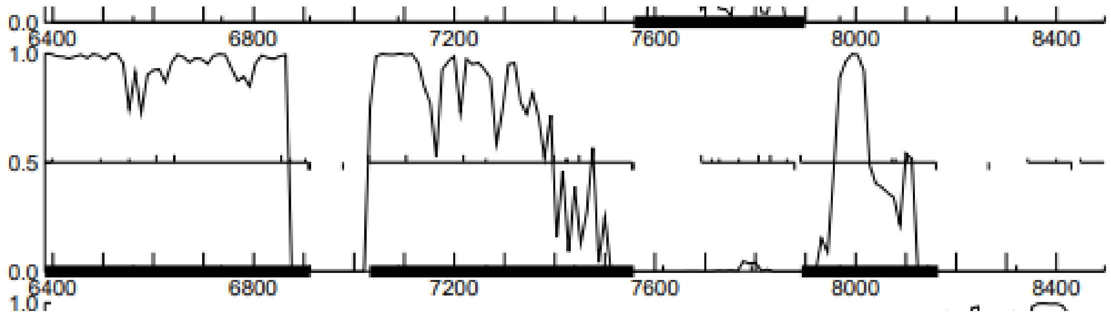
What is the autoannotated start?

Gap: ___117_____ or overlap: _______ (with gene in front of it) for the autoannotated start

- Feature 9
- Stop site: 7558
- Both Glimmer and GeneMark call @bp 7031
- Gap: 117

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

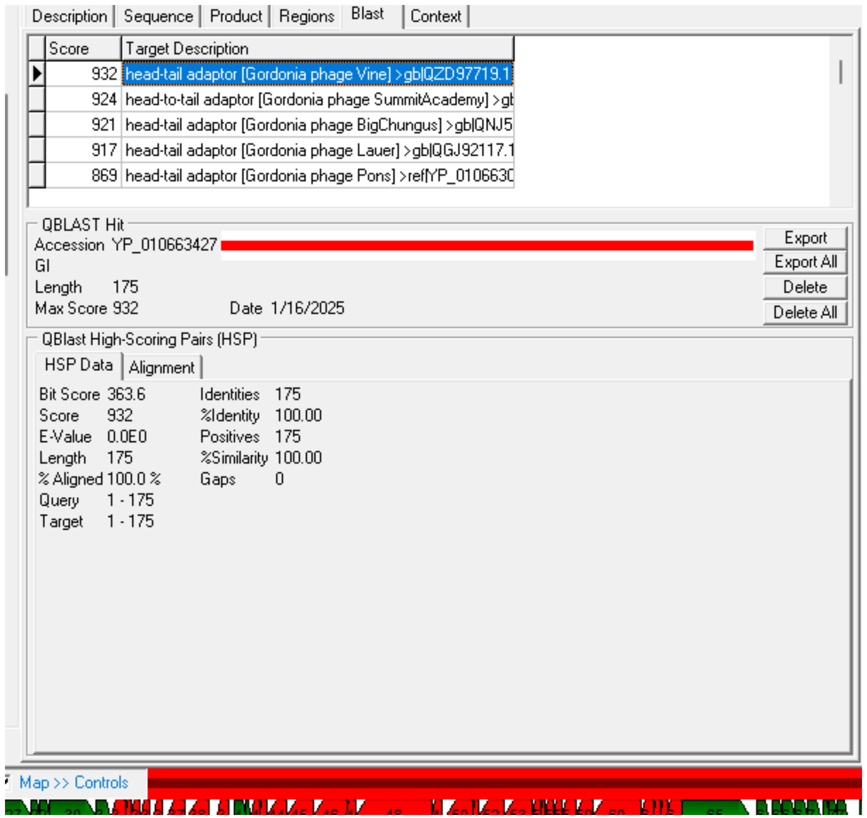
At start site 7031, some of the coding potential is cut off. Coding potential is found in frame 2. No other forward frames include cp from 7031-7558



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 1:1 alignment with Vine
- 1:1 alignment with SummitAcademy
- 1:1 alignment with BigChungus
- 1:1 alignment with Lauer
- 1:1 alignment with Pons

TOTAL: 5 1:1 alignments



Score	Target Description
932	head-tail adaptor [Gordonia phage Vine] >gb|QZD97719.1
924	head-to-tail adaptor [Gordonia phage SummitAcademy] >gb
921	head-tail adaptor [Gordonia phage BigChungus] >gb|QNJ5
917	head-tail adaptor [Gordonia phage Lauer] >gb|QGJ92117.1
869	head-tail adaptor [Gordonia phage Pons] >ref|YP_0106630

QBLAST Hit

Accession YP_010663427
GI
Length 175
Max Score 932 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

```
Bit Score  363.6      Identities   175
Score      932        %Identity    100.00
E-Value    0.0E0      Positives    175
Length     175        %Similarity  100.00
% Aligned  100.0 %    Gaps         0
Query      1 - 175
Target     1 - 175
```

24 highly similar genes:

SteamedHams
Vine
SummitAcademy
BigChungus
Lauer
Pons
AikoCarson
Emalyn
Quasar
Cozz
Nina

BillDoor
AndPeggy
Troje
SweatNTears
SketchMex
GTE2
Yummy
Fribs8
Gibbous
Cleo
Azira
Survivors
HippoPololi

Answer: Is it a gene? Give evidence why you think this is a gene or not.

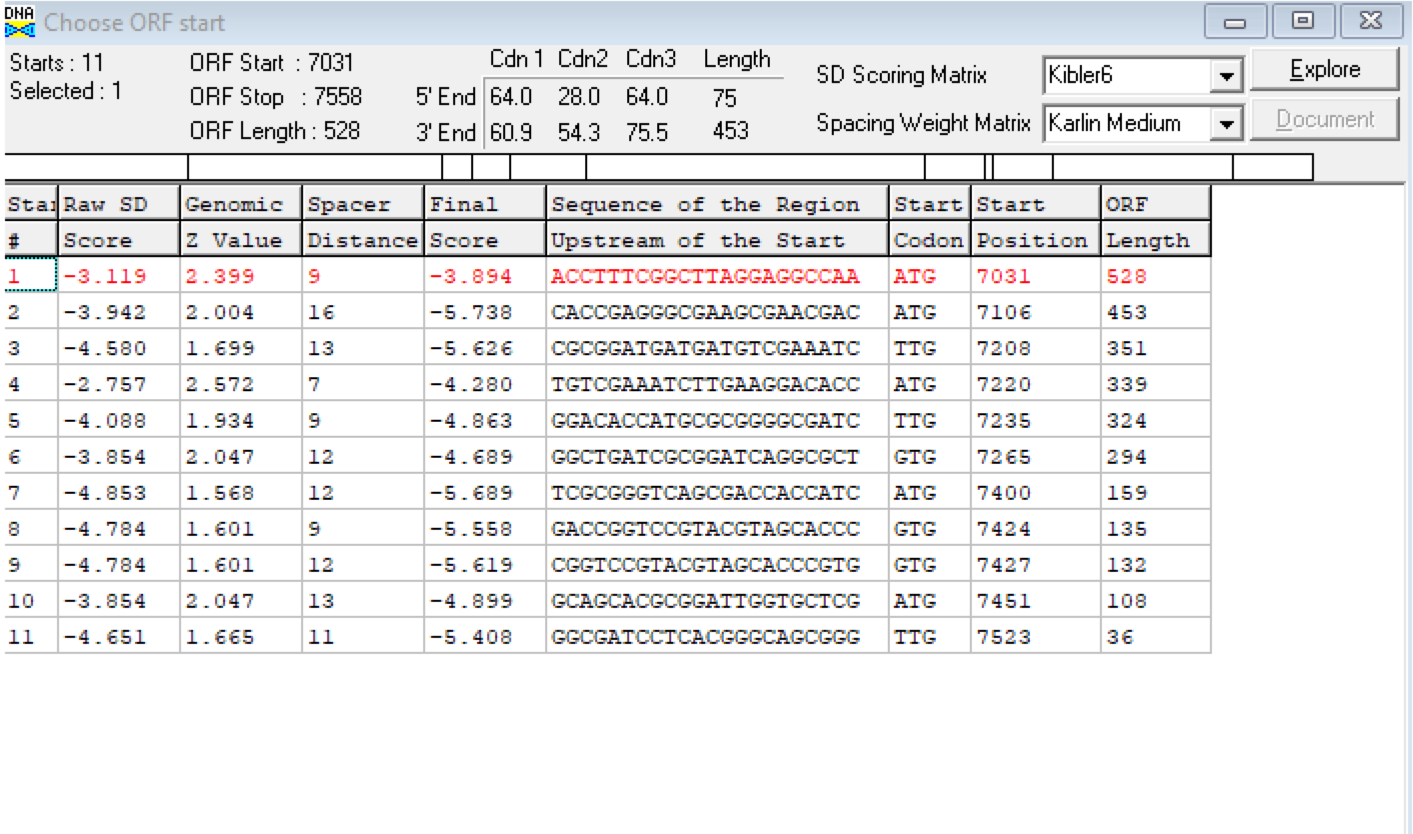
- Yes, it is a gene, because at start site 7031, all coding potential is included.
- It is also a gene because both Glimmer and GeneMark call @bp 7031. And the gene has 5 1:1 alignments based on BLAST conservation evidence and 24 highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- In DNAM file, for start site 7031, there was a 1:1 alignment with 5 other genes
- It is the favored start site, because it was the only start site found in Starterator evidence and had 1:1 alignments which is ideal. And the NCBI website is only used for alternative starts, which does not apply to feature 11.

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Start 7031
- Z value = 2.399
- Final score = -3.894



Starts : 11
Selected : 1

ORF Start : 7031
ORF Stop : 7558
ORF Length : 528

	Cdn 1	Cdn2	Cdn3	Length
5' End	64.0	28.0	64.0	75
3' End	60.9	54.3	75.5	453

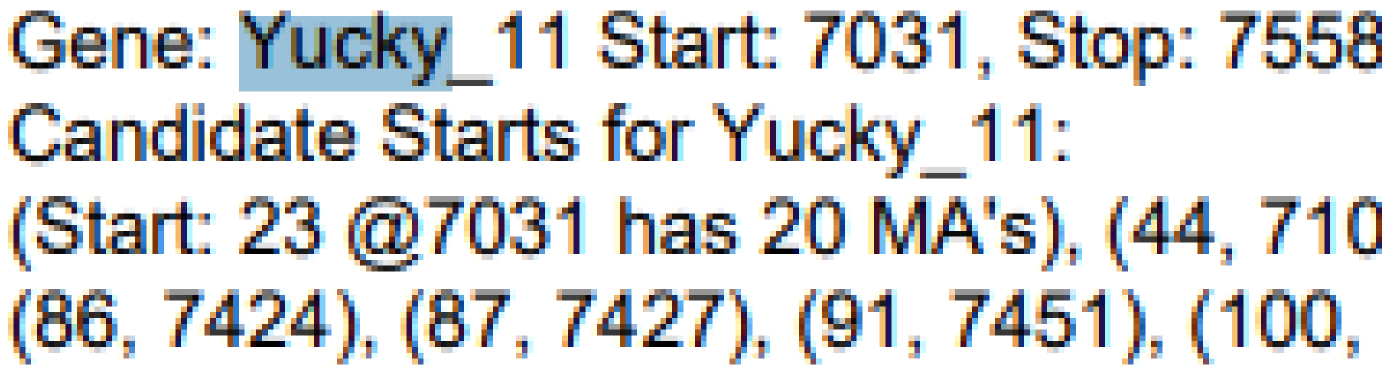
SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-3.119	2.399	9	-3.894	ACCTTTCGGCTTAGGAGGCCAA	ATG	7031	528
2	-3.942	2.004	16	-5.738	CACCGAGGGCGAAGCGAACGAC	ATG	7106	453
3	-4.580	1.699	13	-5.626	CGCGGATGATGATGTCGAAATC	TTG	7208	351
4	-2.757	2.572	7	-4.280	TGTCGAAATCTTGAAGGACACC	ATG	7220	339
5	-4.088	1.934	9	-4.863	GGACACCATGCGCGGGGCGATC	TTG	7235	324
6	-3.854	2.047	12	-4.689	GGCTGATCGCGGATCAGGCGCT	GTG	7265	294
7	-4.853	1.568	12	-5.689	TCGCGGGTCAGCGACCACCATC	ATG	7400	159
8	-4.784	1.601	9	-5.558	GACCGGTCCGTACGTAGCACCC	GTG	7424	135
9	-4.784	1.601	12	-5.619	CGGTCCGTACGTAGCACCCGTG	GTG	7427	132
10	-3.854	2.047	13	-4.899	GCAGCACGCGGATTGGTGCTCG	ATG	7451	108
11	-4.651	1.665	11	-5.408	GGCGATCCTCACGGGCAGCGGG	TTG	7523	36

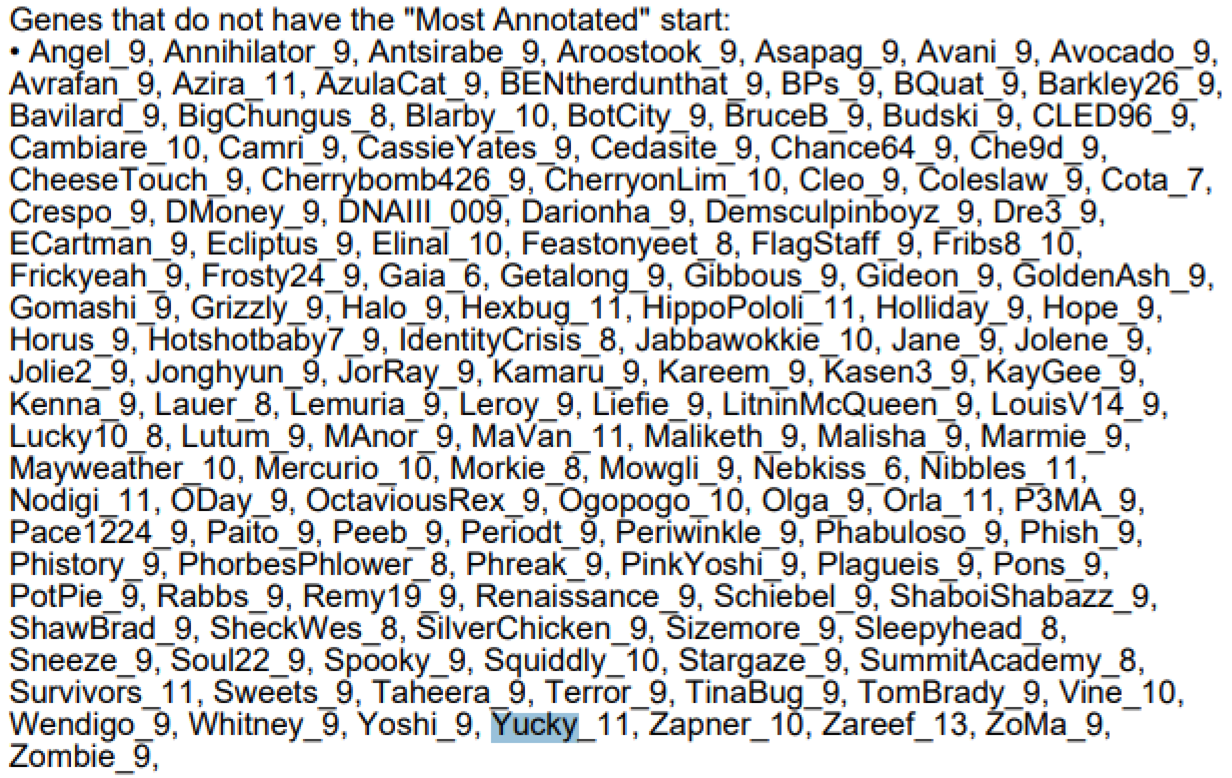
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Yucky only has one listed start site of 23 @7031 which has 20 MA's

- Yucky _11 does not have the "Most Annotated" start like Pons_9 and PotPie_9



Gene: Yucky_11 Start: 7031, Stop: 7558
Candidate Starts for Yucky_11:
(Start: 23 @7031 has 20 MA's), (44, 710
(86, 7424), (87, 7427), (91, 7451), (100,

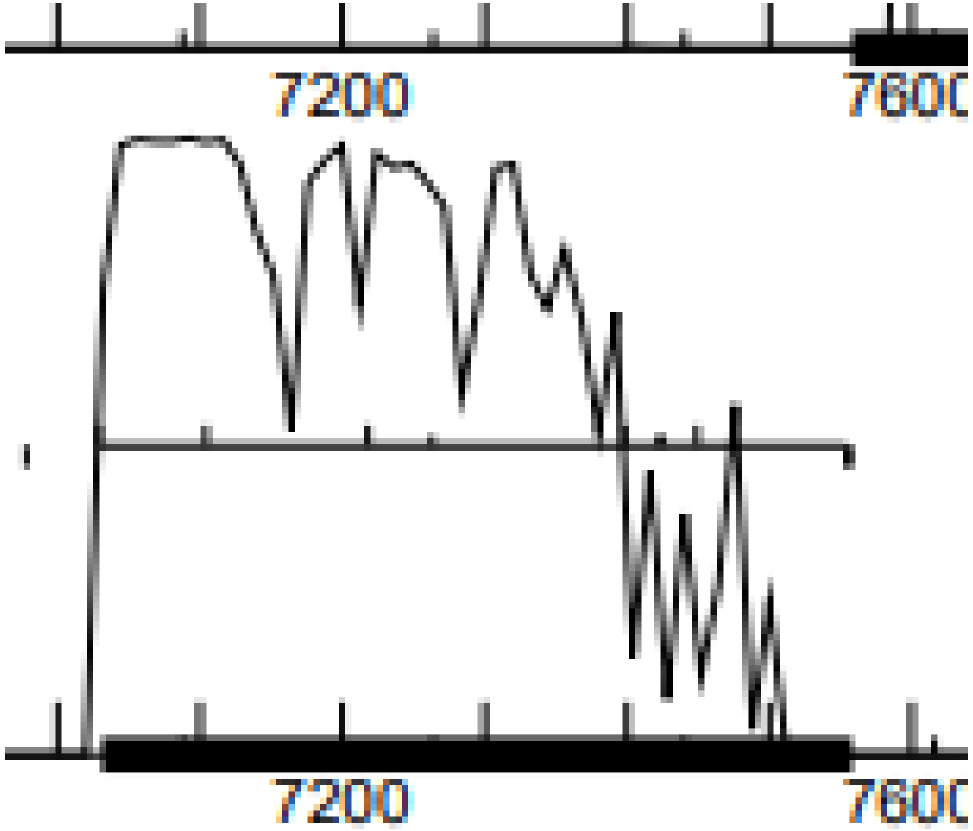


Genes that do not have the "Most Annotated" start:

• Angel_9, Annihilator_9, Antsirabe_9, Aroostook_9, Asapag_9, Avani_9, Avocado_9, Avrafan_9, Azira_11, AzulaCat_9, BENtherdunthat_9, BPs_9, BQuat_9, Barkley26_9, Bavilard_9, BigChungus_8, Blarby_10, BotCity_9, BruceB_9, Budski_9, CLED96_9, Cambiare_10, Camri_9, CassieYates_9, Cedasite_9, Chance64_9, Che9d_9, CheeseTouch_9, Cherrybomb426_9, CherryonLim_10, Cleo_9, Coleslaw_9, Cota_7, Crespo_9, DMoney_9, DNAIII_009, Darionha_9, Demsculpinboyz_9, Dre3_9, ECartman_9, Ecliptus_9, Elinal_10, Feastonyeet_8, FlagStaff_9, Fribs8_10, Frickyeah_9, Frosty24_9, Gaia_6, Getalong_9, Gibbous_9, Gideon_9, GoldenAsh_9, Gomashi_9, Grizzly_9, Halo_9, Hexbug_11, HippoPololi_11, Holliday_9, Hope_9, Horus_9, Hotshotbaby7_9, IdentityCrisis_8, Jabbawokkie_10, Jane_9, Jolene_9, Jolie2_9, Jonghyun_9, JorRay_9, Kamaru_9, Kareem_9, Kasen3_9, KayGee_9, Kenna_9, Lauer_8, Lemuria_9, Leroy_9, Liefie_9, LitninMcQueen_9, LouisV14_9, Lucky10_8, Lutum_9, MAnor_9, MaVan_11, Maliketh_9, Malisha_9, Marmie_9, Mayweather_10, Mercurio_10, Morkie_8, Mowgli_9, Nebkiss_6, Nibbles_11, Nodigi_11, ODay_9, OctaviousRex_9, Ogopogo_10, Olga_9, Orla_11, P3MA_9, Pace1224_9, Paito_9, Peeb_9, Periodt_9, Periwinkle_9, Phabuloso_9, Phish_9, Phistory_9, PhorbesPhlower_8, Phreak_9, PinkYoshi_9, Plagueis_9, Pons_9, PotPie_9, Rabbs_9, Remy19_9, Renaissance_9, Schiebel_9, ShaboiShabazz_9, ShawBrad_9, SheckWes_8, SilverChicken_9, Sizemore_9, Sleepyhead_8, Sneeze_9, Soul22_9, Spooky_9, Squiddly_10, Stargaze_9, SummitAcademy_8, Survivors_11, Sweets_9, Taheera_9, Terror_9, TinaBug_9, TomBrady_9, Vine_10, Wendigo_9, Whitney_9, Yoshi_9, Yucky_11, Zapner_10, Zareef_13, ZoMa_9, Zombie_9,

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- At and past start 23 @7031 all coding potential is included. Some coding potential is cut off before the start site.



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start?   Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

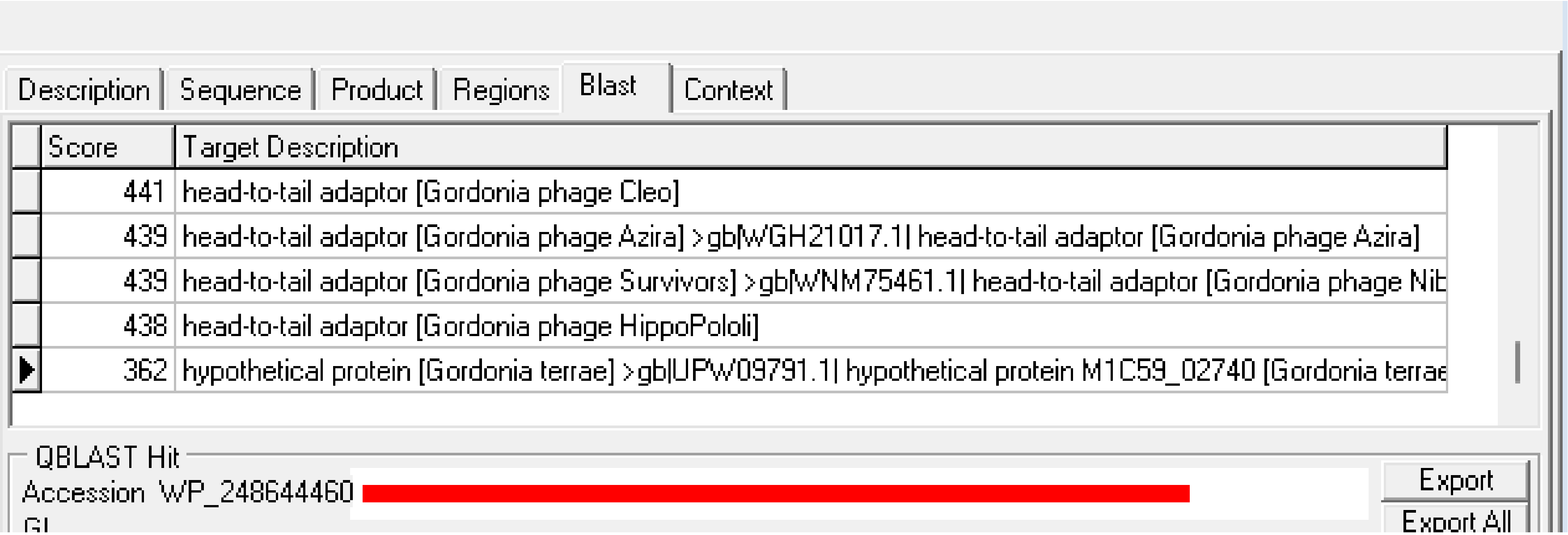
- Gap: 117 at start site 7031

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	7031
GeneMark	Glimmer & GeneMark
Coding potential	Includes some coding potential
RBS	Z value: 2.399 final score: -3.894
BLAST	5 1:1 alignments
Starterator	20 MA’s
Gap	117

While gap is greater than 100, the start site of 7031 is the best and only choice as the start site, because both Glimmer and GeneMark call it, it includes coding potential in frame 2, has a z value greater than 2, and has 5 1:1 alignments.

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
441	head-to-tail adaptor [Gordonia phage Cleo]
439	head-to-tail adaptor [Gordonia phage Azira] >gb|WGH21017.1| head-to-tail adaptor [Gordonia phage Azira]
439	head-to-tail adaptor [Gordonia phage Survivors] >gb|WNM75461.1| head-to-tail adaptor [Gordonia phage Nib
438	head-to-tail adaptor [Gordonia phage HippoPololi]
362	hypothetical protein [Gordonia terrae] >gb|UPW09791.1| hypothetical protein M1C59_02740 [Gordonia terrae

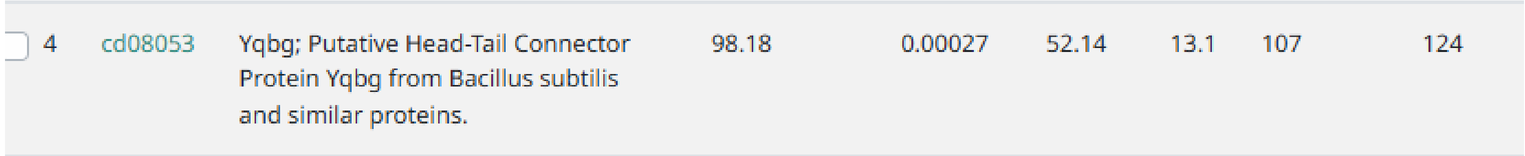
QBLAST Hit
Accession WP_248644460
GI

24 head-to-tail adaptor
1 hypothetical protein
(Gordonia terrae)

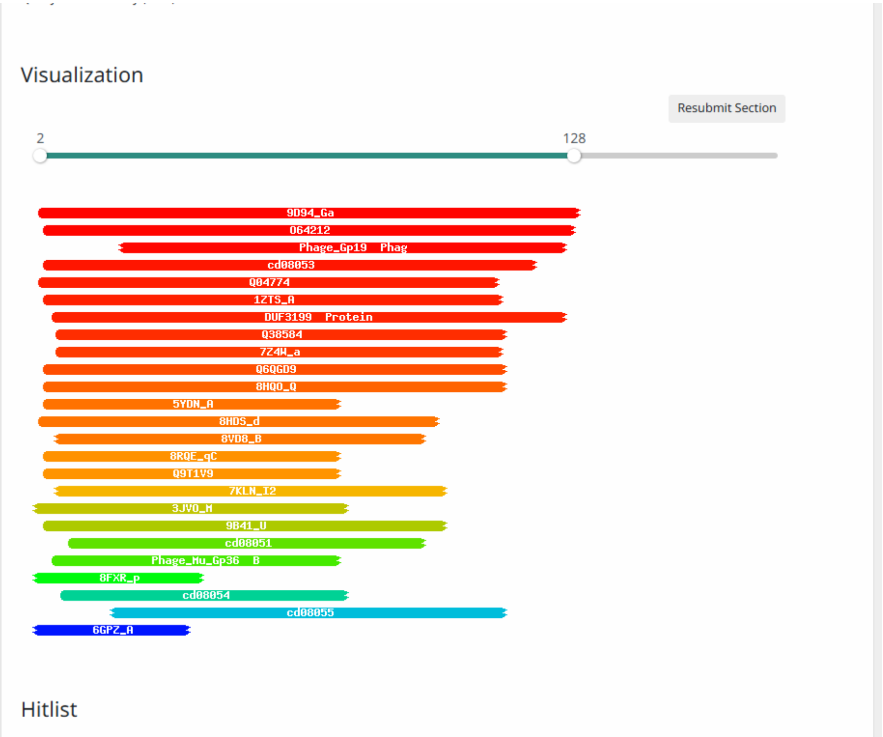
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

Yes, the HHpred evidence supports the function head-to-tail adaptor, but also the function hypothetical protein.

For it to have the function head-to-tail adaptor, HHPRED alignment had to be with crystal structures: SPP1 15 or HK97 gp6 or Bacillus protein yqbG. I found one gene that had the structure yqBG and function head-to-tail adaptor.



☐	4	cd08053	Yqbg; Putative Head-Tail Connector Protein Yqbg from Bacillus subtilis and similar proteins.	98.18	0.00027	52.14	13.1	107	124

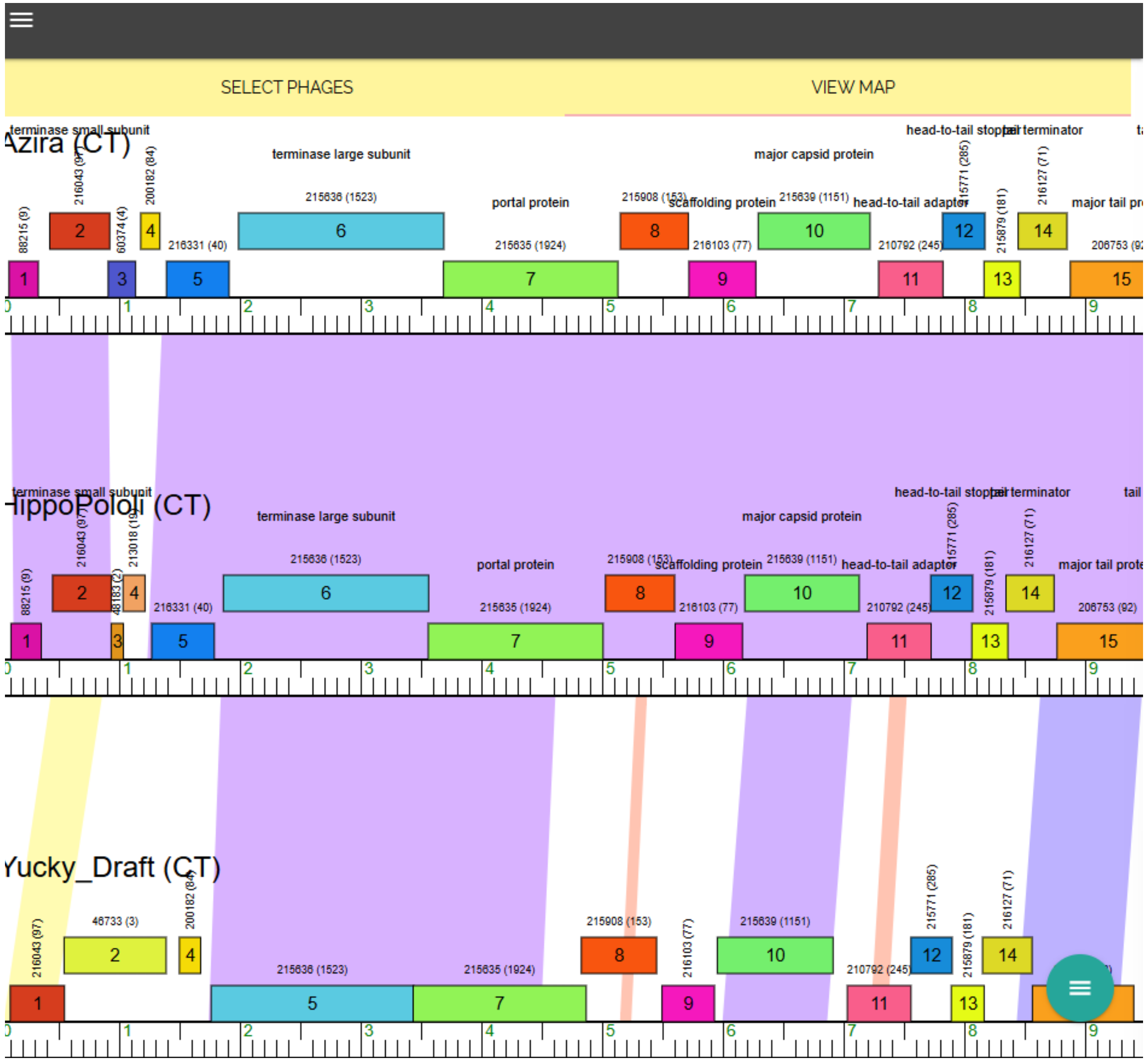


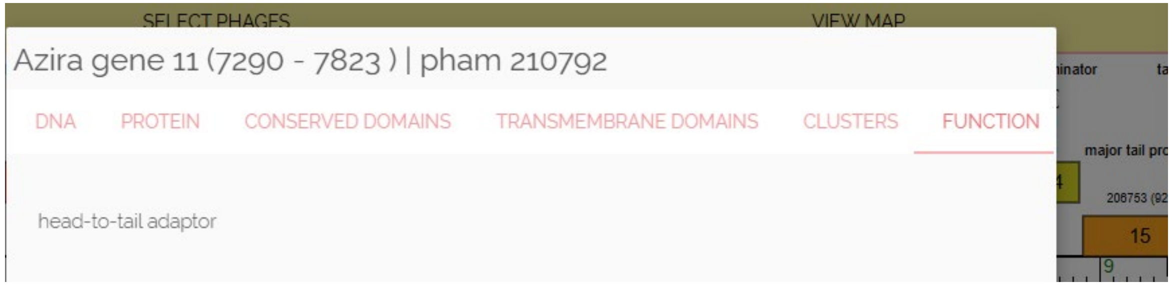
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

Yucky feature 11 conserved domain: none function: none

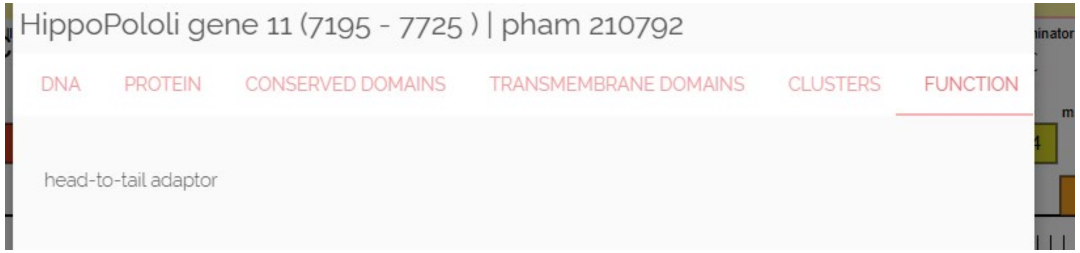
HippoPololi feature 11 conserved domain: none function: head-to-tail adaptor

Azira feature 11 conserved domain: none function: head-to-tail adaptor




Azira gene 11 (7290 - 7823) | pham 210792

head-to-tail adaptor


HippoPololi gene 11 (7195 - 7725) | pham 210792

head-to-tail adaptor

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function of the gene is head-to-tail adapter because there were 24 highly similar genes from BLAST that had the head-to-tail adapter. Hhpred evidence also showed highly similar genes with above a 90% probability and an E value less than 1 that had the function head-to-tail adapter. The Phamerator evidence displayed similar genes, HippoPololi and Azira alongside Yucky. While the similar genes did not have a conserved domain their functions were both the same being the head-to-tail adapter.

Feature 10 – Stop 7899

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

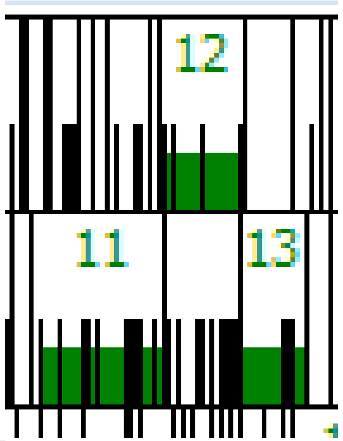
What is the autoannotated start?

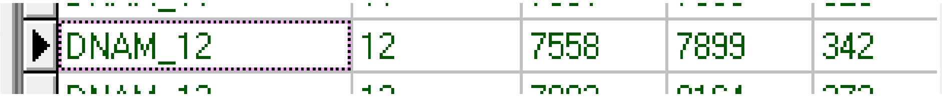
Gap: ____0____ or overlap: _______ (with gene in front of it) for the autoannotated start

- Feature 10
- Stop site: 7899

- Auto-annotated start is called by both Glimmer and GeneMark

- Both call @bp 7558

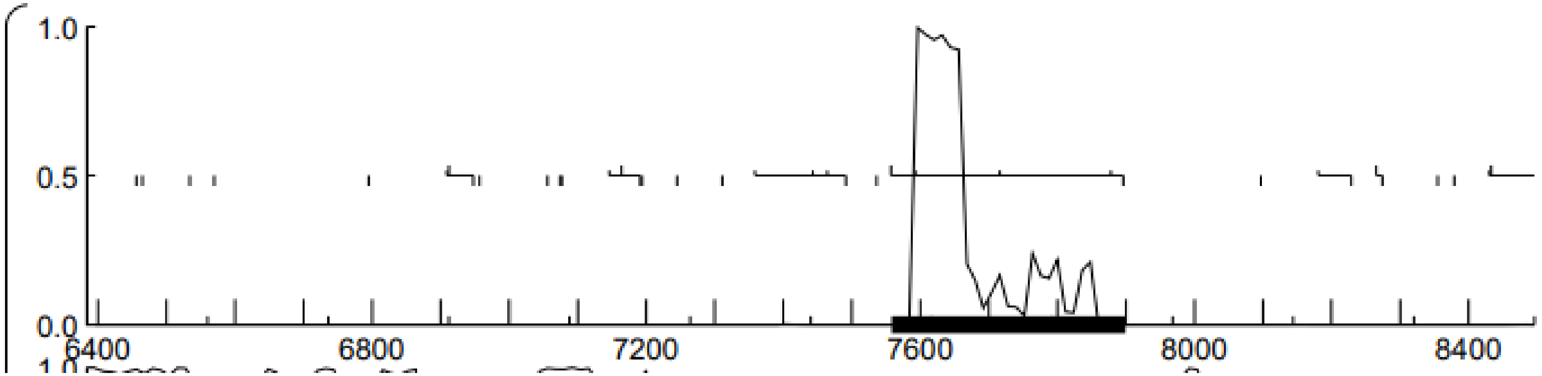




DNAM_12	12	7558	7899	342

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

The start site 7558 includes all the coding potential. None of the coding potential is cut off. The coding potential ranges from 7558-7900. It is the only forward reading frame with cp from 7558-7899.

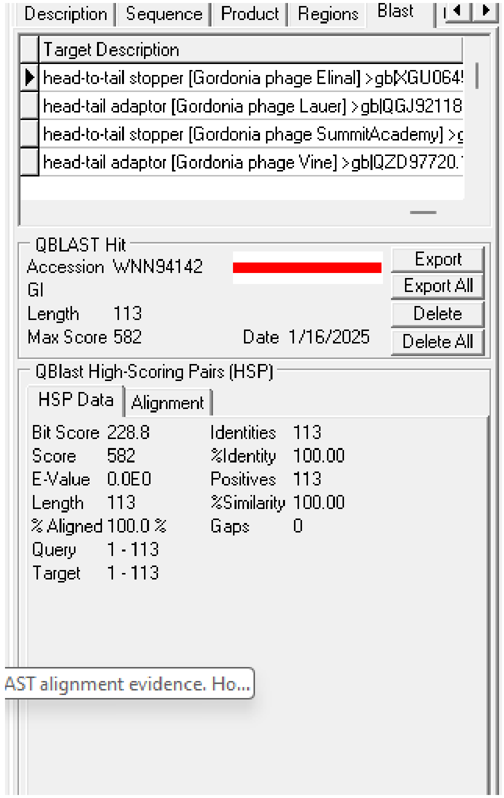


BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 1:1 alignment with Elinal
- 1:1 alignment with Lauer
- 1:1 alignment with SummitAcademy
- 1:1 alignment with Vine
- 1:1 alignment with BigChungus
- 1:1 alignment with SheckWes
- 1:1 alignment with Pons
- 1:1 alignment with Manor
- 1:1 alignment with CherryonLim
- 1:1 alignment with Mayweather

- 1:1 alignment with Cozz
- 1:1 alignment with AikoCarson
- 1:1 alignment with Quasar
- 1:1 alignment with Emalyn
- 1:1 alignment with Nina
- 1:1 alignment with SteamedHams
- 1:1 alignment with Yummy
- 1:1 alignment with GTE2
- 1:1 alignment with SketchMex
- 1:1 Alignment with Troje
- 1:1 alignment with Margaret

21 1:1 alignments with Feature 12! ALSO 21 highly similar genes to feature 12



Target Description
head-to-tail stopper [Gordonia phage Elinal] >gb|XGU064!
head-tail adaptor [Gordonia phage Lauer] >gb|QGJ92118
head-to-tail stopper [Gordonia phage SummitAcademy] >g
head-tail adaptor [Gordonia phage Vine] >gb|QZD97720.

QBLAST Hit

Accession WNN94142
GI
Length 113
Max Score 582
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 228.8
Score 582
E-Value 0.0E0
Length 113
% Aligned 100.0 %
Query 1 - 113
Target 1 - 113

Identities 113
%Identity 100.00
Positives 113
%Similarity 100.00
Gaps 0

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, it is a gene because both Glimmer and GeneMark call it at start site 7558. The start site at 7558 also includes all coding potential (nothing is cut off).

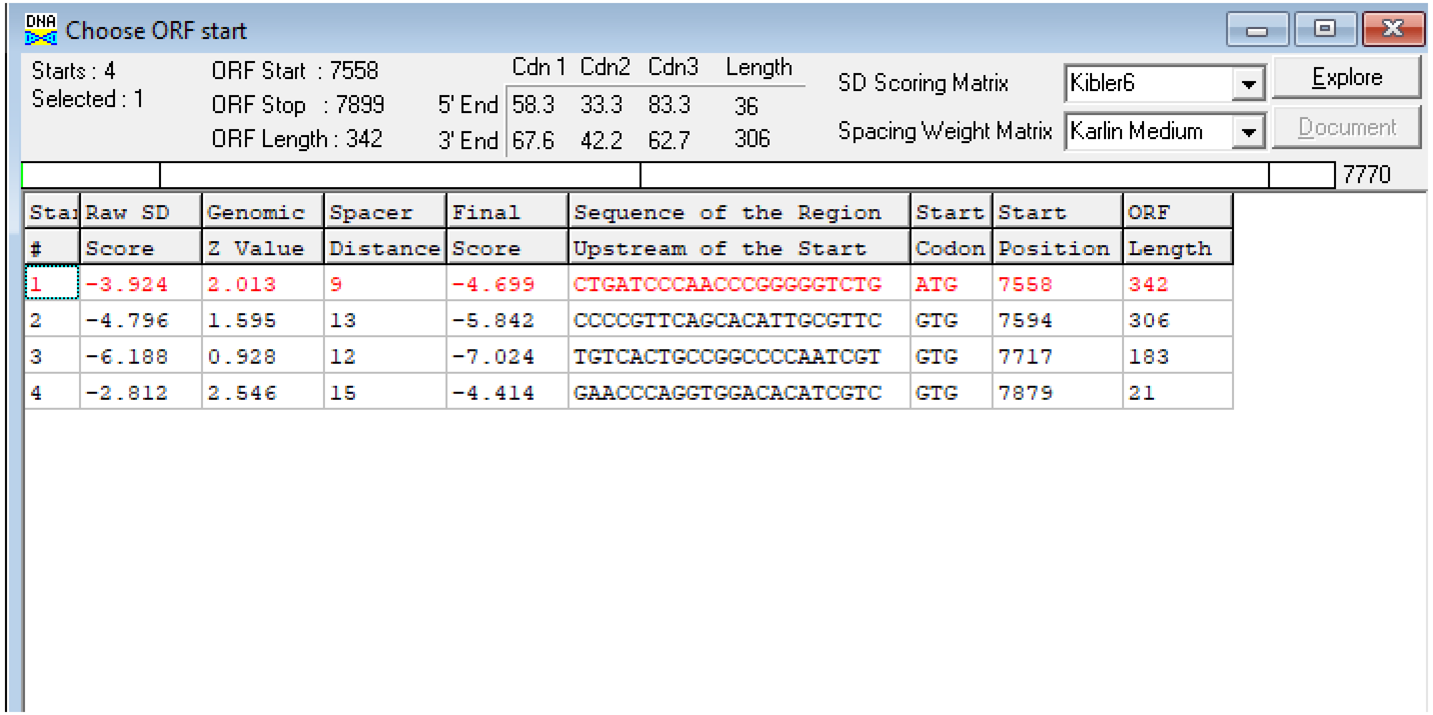
- According to BLAST conservation evidence, feature 12 has 21 1:1 alignments with other genes such as Nina, Cozz, Yummy, and GTE2.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- Start 7558 had 21 1:1 alignments
- There were no alternative starts

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Start site 7558
- Z value: 2.013
- Final score: -4.699



Choose ORF start

Starts : 4
Selected : 1

ORF Start : 7558
ORF Stop : 7899
ORF Length : 342

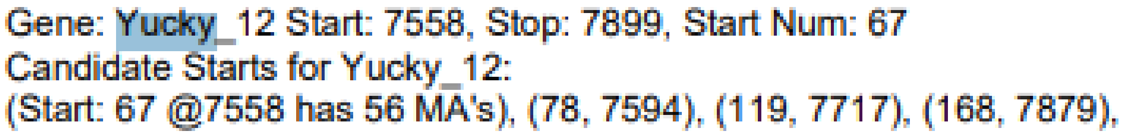
	Cdn 1	Cdn2	Cdn3	Length
5' End	58.3	33.3	83.3	36
3' End	67.6	42.2	62.7	306

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-3.924	2.013	9	-4.699	CTGATCCCAACCCGGGGGTCTG	ATG	7558	342
2	-4.796	1.595	13	-5.842	CCCCGTTCAGCACATTGCGTTC	GTG	7594	306
3	-6.188	0.928	12	-7.024	TGTCACTGCCGGCCCCAATCGT	GTG	7717	183
4	-2.812	2.546	15	-4.414	GAACCCAGGTGGACACATCGTC	GTG	7879	21

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Yucky Feature 12 was one of the genes that did not have the "Most Annotated" start. Other genes a part of this section were Ziko_44, Zombie_10, an PotPie_10.

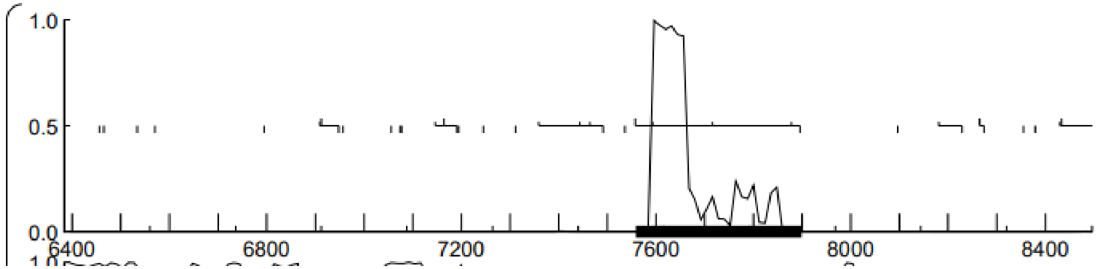


Gene: Yucky_12 Start: 7558, Stop: 7899, Start Num: 67
Candidate Starts for Yucky_12:
(Start: 67 @7558 has 56 MA's), (78, 7594), (119, 7717), (168, 7879),

- Only one start was listed which was 67 @7558 which had 56 MA's

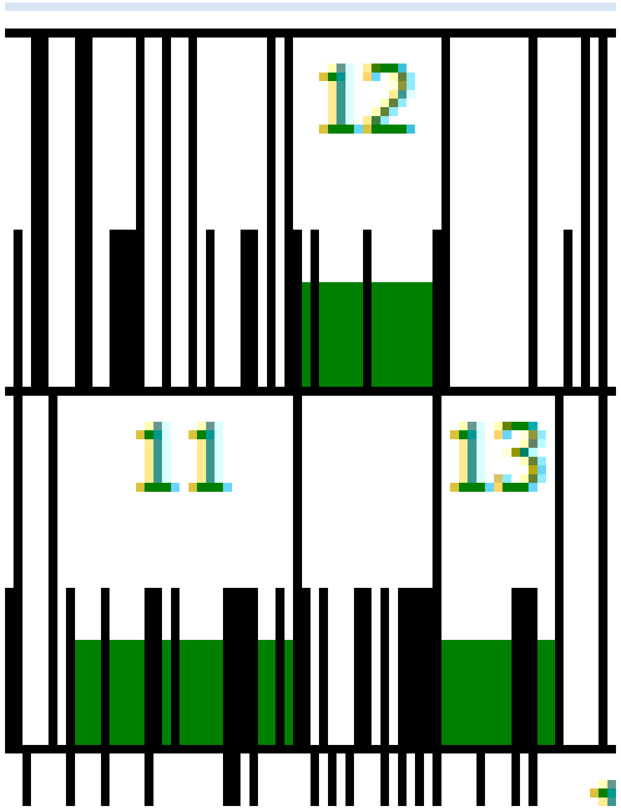
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- For start site 7558, all coding potential was included. Nothing was cut off.



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Overlap of 1
- Previous feature ends at 7558, this feature starts at 7558



What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

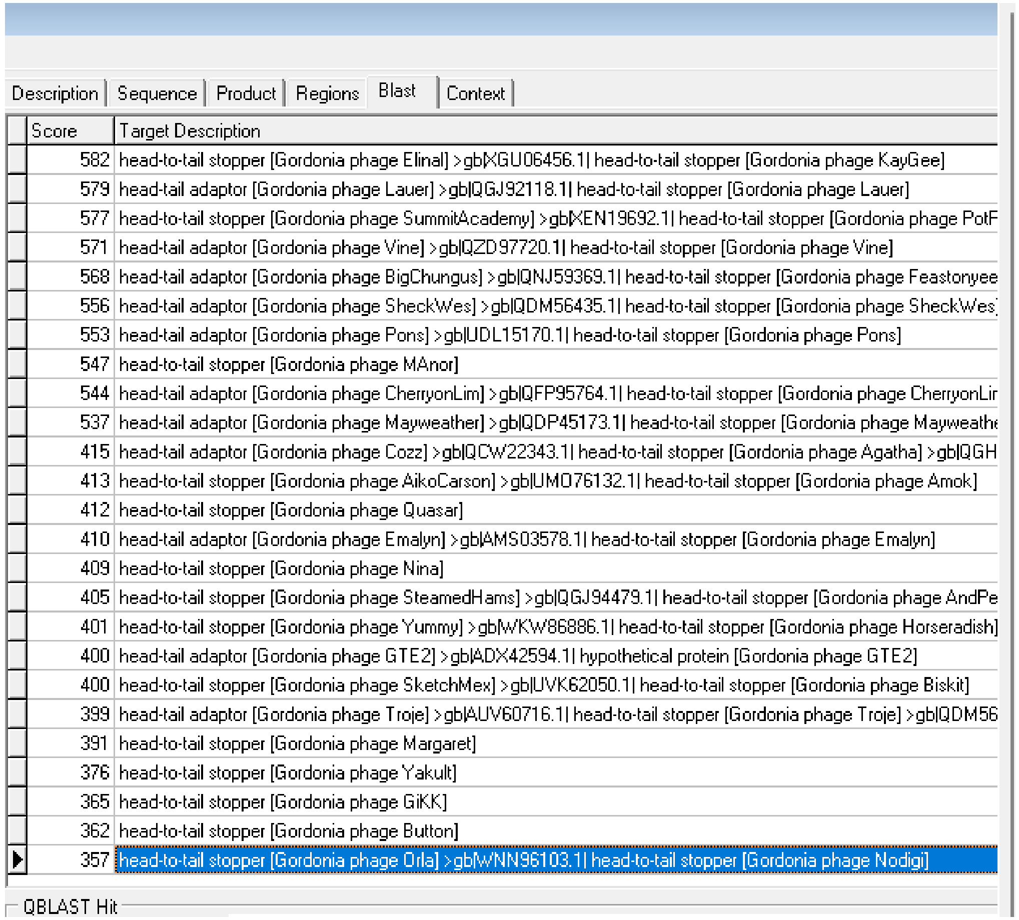
	7558
Genemark	Glimmer & GeneMark
Coding potential	All coding potential is included
RBS	Z value: 2.013 Final score: -4.699
BLAST	21 1:1 alignments
Starterator	56 MA’s
Overlap	1

The start site is 7558 because both Glimmer and GeneMark call it at 7558 and all coding potential is included within the frame. The Z value is greater than 1 and it has 21 1:1 alignments with other genes. It agrees with the auto-annotated start site.

BLAST function evidence. What assigned functions do other highly similar genes have?

- Head-to-tail stopper: 14 highly similar genes

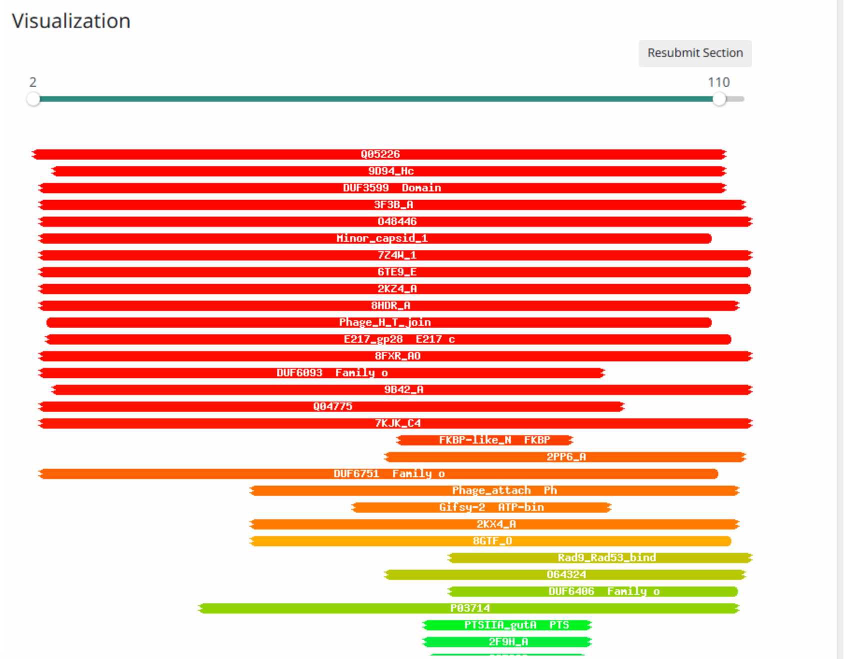
- Head-to-tail adapter: 11 highly similar genes

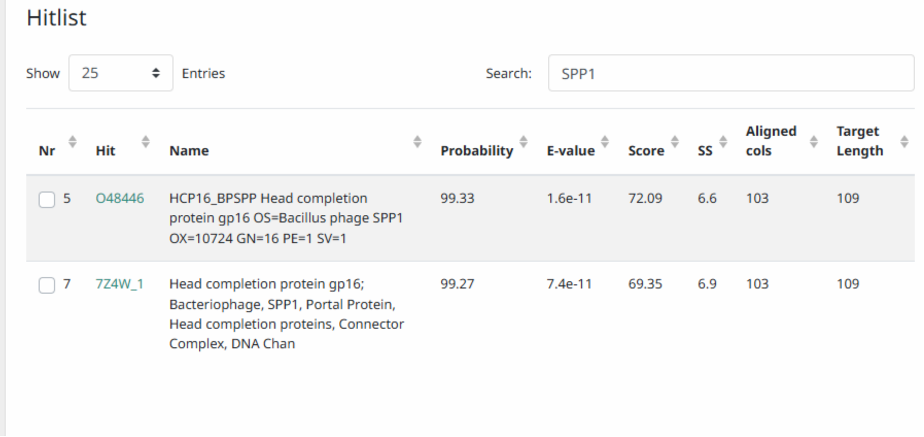


Score	Target Description
582	head-to-tail stopper [Gordonia phage Elinal] >gb|XGU06456.1| head-to-tail stopper [Gordonia phage KayGee]
579	head-tail adaptor [Gordonia phage Lauer] >gb|QGJ92118.1| head-to-tail stopper [Gordonia phage Lauer]
577	head-to-tail stopper [Gordonia phage SummitAcademy] >gb|XEN19692.1| head-to-tail stopper [Gordonia phage PotF
571	head-tail adaptor [Gordonia phage Vine] >gb|QZD97720.1| head-to-tail stopper [Gordonia phage Vine]
568	head-tail adaptor [Gordonia phage BigChungus] >gb|QNJ59369.1| head-to-tail stopper [Gordonia phage Feastonyee
556	head-tail adaptor [Gordonia phage SheckWes] >gb|QDM56435.1| head-to-tail stopper [Gordonia phage SheckWes]
553	head-tail adaptor [Gordonia phage Pons] >gb|UDL15170.1| head-to-tail stopper [Gordonia phage Pons]
547	head-to-tail stopper [Gordonia phage MAnor]
544	head-tail adaptor [Gordonia phage CherryonLim] >gb|QFP95764.1| head-to-tail stopper [Gordonia phage CherryonLir
537	head-tail adaptor [Gordonia phage Mayweather] >gb|QDP45173.1| head-to-tail stopper [Gordonia phage Mayweathe
415	head-tail adaptor [Gordonia phage Cozz] >gb|QCW22343.1| head-to-tail stopper [Gordonia phage Agatha] >gb|QGH
413	head-to-tail stopper [Gordonia phage AikoCarson] >gb|UMO76132.1| head-to-tail stopper [Gordonia phage Amok]
412	head-to-tail stopper [Gordonia phage Quasar]
410	head-tail adaptor [Gordonia phage Emalyn] >gb|AMS03578.1| head-to-tail stopper [Gordonia phage Emalyn]
409	head-to-tail stopper [Gordonia phage Nina]
405	head-to-tail stopper [Gordonia phage SteamedHams] >gb|QGJ94479.1| head-to-tail stopper [Gordonia phage AndPe
401	head-to-tail stopper [Gordonia phage Yummy] >gb|WKW86886.1| head-to-tail stopper [Gordonia phage Horseradish]
400	head-tail adaptor [Gordonia phage GTE2] >gb|ADX42594.1| hypothetical protein [Gordonia phage GTE2]
400	head-to-tail stopper [Gordonia phage SketchMex] >gb|UVK62050.1| head-to-tail stopper [Gordonia phage Biskit]
399	head-tail adaptor [Gordonia phage Troje] >gb|AUV60716.1| head-to-tail stopper [Gordonia phage Troje] >gb|QDM56
391	head-to-tail stopper [Gordonia phage Margaret]
376	head-to-tail stopper [Gordonia phage Yakult]
365	head-to-tail stopper [Gordonia phage GiKK]
362	head-to-tail stopper [Gordonia phage Button]
357	head-to-tail stopper [Gordonia phage Orla] >gb|WNN96103.1| head-to-tail stopper [Gordonia phage Nodigi]

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- For head-to-tail stopper, must have HHPRED alignment to following structures: SPP1 16 or Bacillus protein yqbH
- Did have 2 similar alignments with crystal structure SPP1 16.
- 0 alignment for crystal structure Bacillus protein yqbH
- Both of the 2 alignments for SPP1 16 had function of head-to-tail stopper.




Hitlist

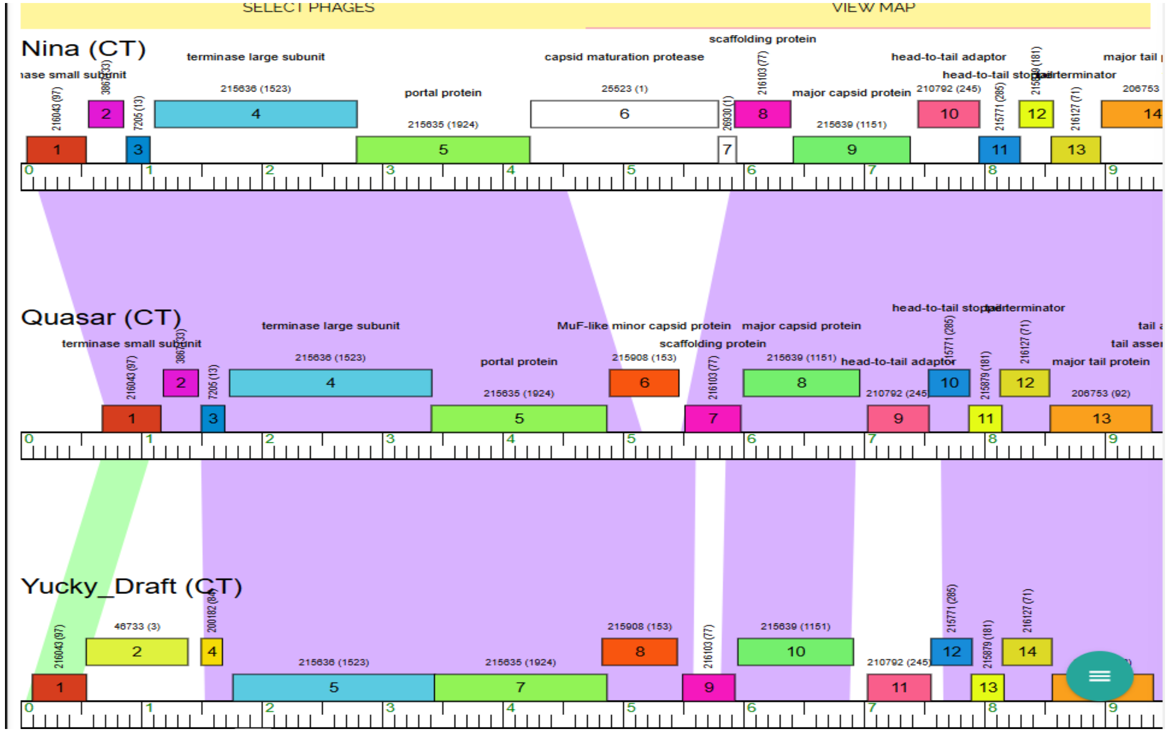
Show 25 Entries

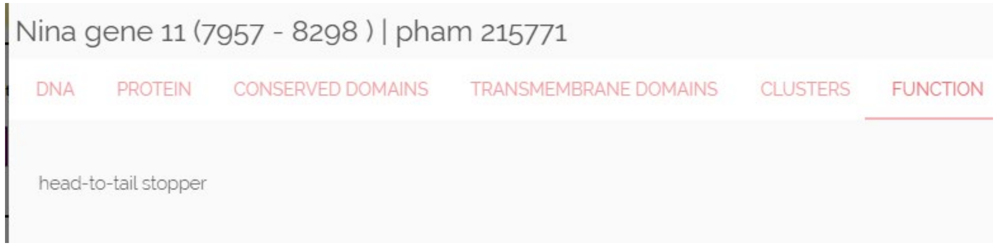
Search: SPP1

	Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
☐	5	O48446	HCP16_BPSPP Head completion protein gp16 OS=Bacillus phage SPP1 OX=10724 GN=16 PE=1 SV=1	99.33	1.6e-11	72.09	6.6	103	109
☐	7	7Z4W_1	Head completion protein gp16; Bacteriophage, SPP1, Portal Protein, Head completion proteins, Connector Complex, DNA Chan	99.27	7.4e-11	69.35	6.9	103	109

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

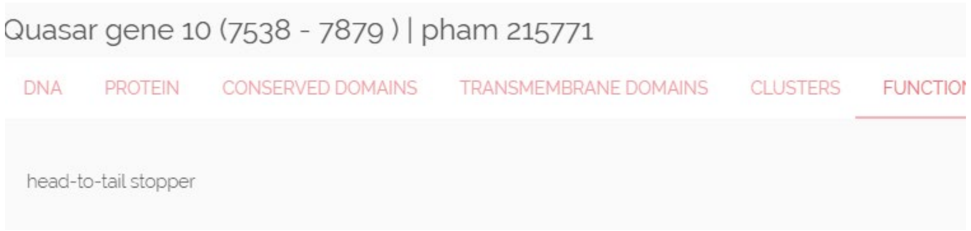
- Yucky feature 12 conserved domain: none function: none
- Quasar feature 10 conserved domain: none function: head-to-tail stopper
- Nina feature 11 conserved domain: none function: head-to-tail stopper





Nina gene 11 (7957 - 8298) | pham 215771

head-to-tail stopper



Quasar gene 10 (7538 - 7879) | pham 215771

head-to-tail stopper

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function is head-to-tail stopper because the BLAST evidence had 14 highly similar genes with function head-to-tail stopper. Also, the Hhpred evidence had certain crystalline structures that had the function head-to-tail stopper. The Phamerator evidence also supported the function as two highly similar genes to Yucky (Nina and Quasar) had the function of head-to-tail stopper.

Feature 11 – Stop 8164

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

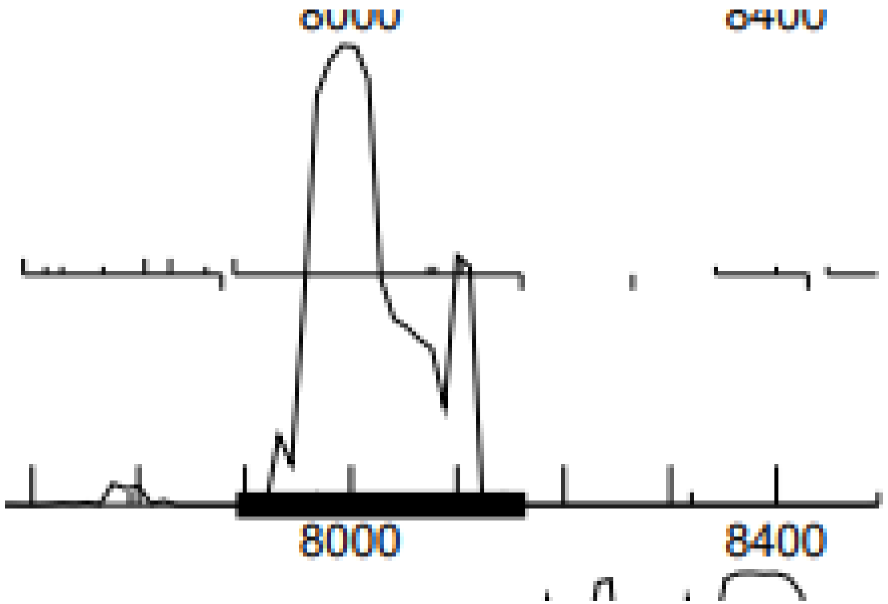
What is the autoannotated start?

Gap: ________ or overlap: ___14____ (with gene in front of it) for the autoannotated start

- Feature 11
- Stop site: 8164
- Both Glimmer and GeneMark call the autoannotated start
- The autoannotated start is called @bp 7892
- Overlap: 14

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

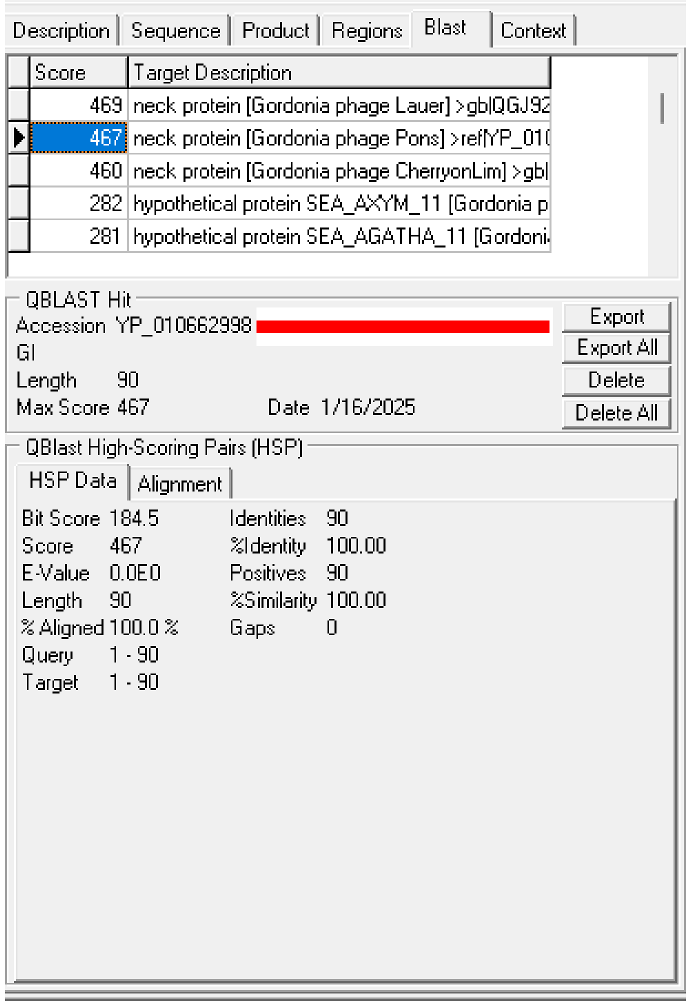
- Start site 7892, starts before cp and includes all coding potential. It is the only frame with cp from 7892-8164



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 1:1 alignment with Pons
- 1:1 alignment with CherryonLim

3 highly similar genes (0E0):
Lauer
Pons
CherryonLim

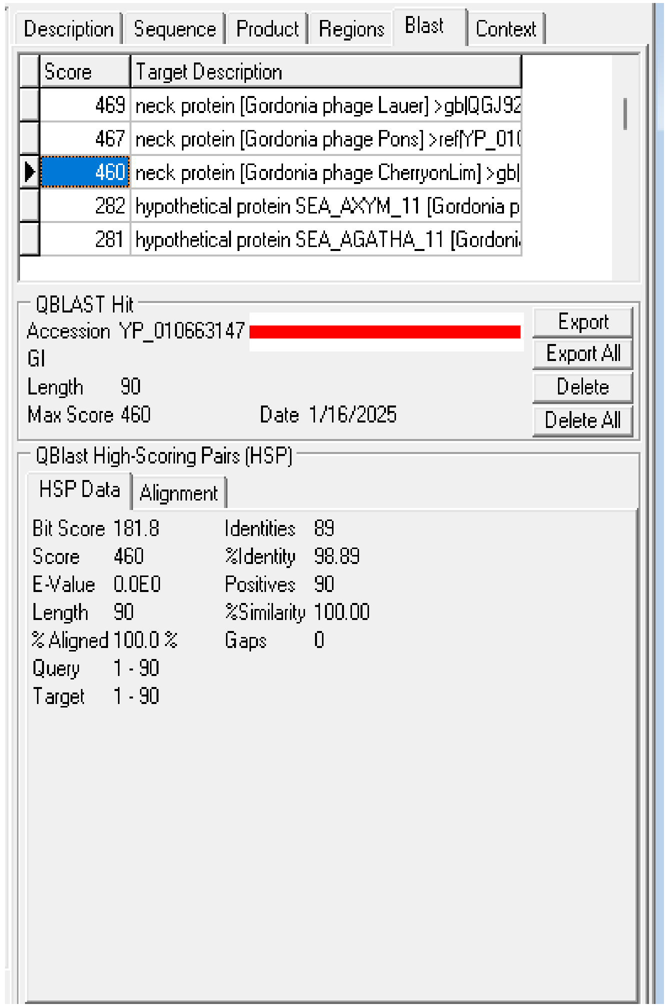


Score	Target Description
469	neck protein [Gordonia phage Lauer] >gb|QGJ92
467	neck protein [Gordonia phage Pons] >ref|YP_010
460	neck protein [Gordonia phage CherryonLim] >gb|
282	hypothetical protein SEA_AXYM_11 [Gordonia p
281	hypothetical protein SEA_AGATHA_11 [Gordoni

QBLAST Hit
Accession YP_010662998
GI
Length 90
Max Score 467
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	184.5	Identities	90
Score	467	%Identity	100.00
E-Value	0.0E0	Positives	90
Length	90	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 90		
Target	1 - 90		



Score	Target Description
469	neck protein [Gordonia phage Lauer] >gb|QGJ92
467	neck protein [Gordonia phage Pons] >ref|YP_010
460	neck protein [Gordonia phage CherryonLim] >gb|
282	hypothetical protein SEA_AXYM_11 [Gordonia p
281	hypothetical protein SEA_AGATHA_11 [Gordoni

QBLAST Hit
Accession YP_010663147
GI
Length 90
Max Score 460
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	181.8	Identities	89
Score	460	%Identity	98.89
E-Value	0.0E0	Positives	90
Length	90	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 90		
Target	1 - 90		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

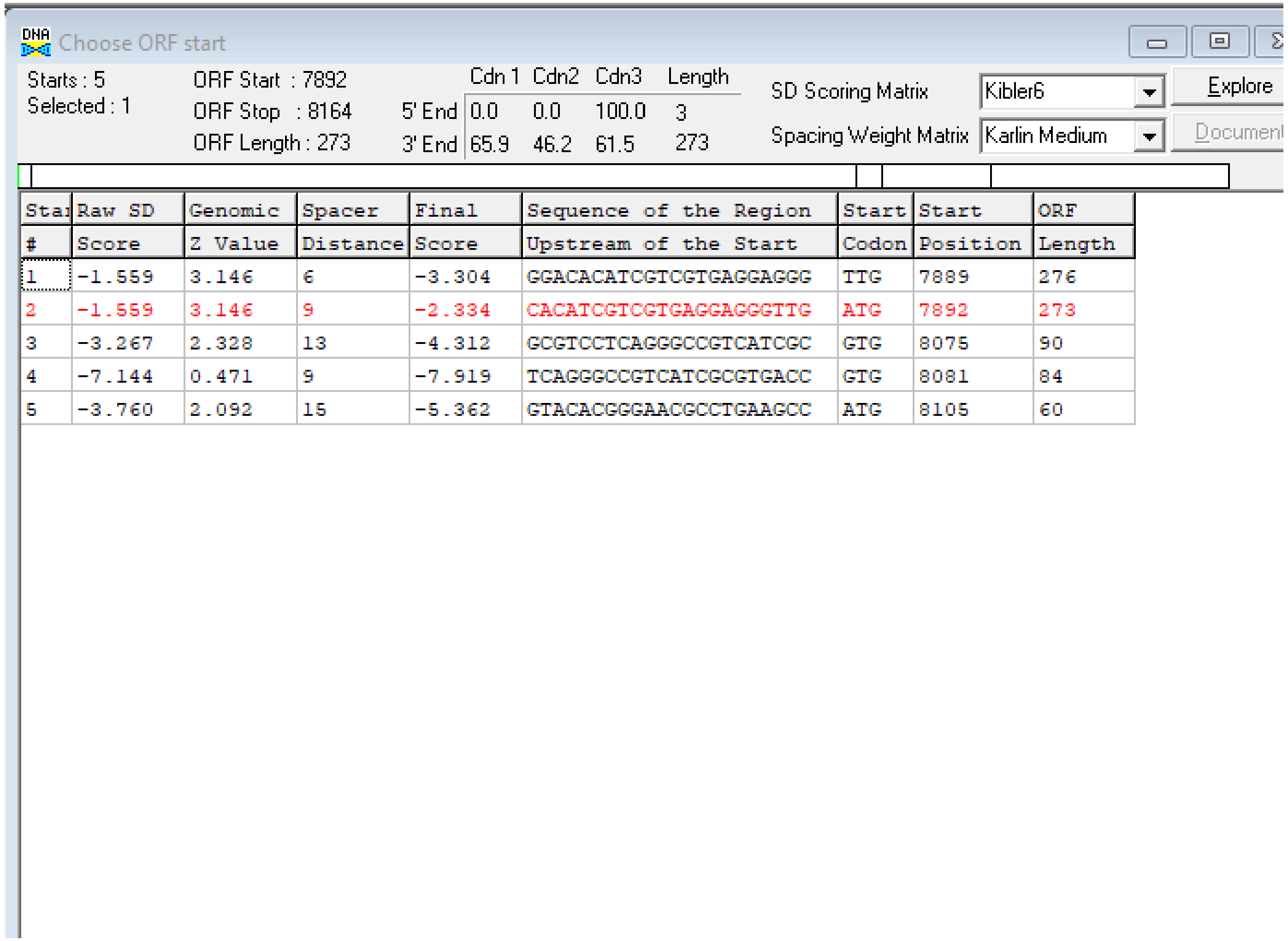
- Yes, it is a gene, because both Glimmer and GeneMark call it at 7892. The start site 7892 also starts before the coding potential and includes all the coding potential. The feature 13 also has 2 1:1 alignments according to BLAST conservation evidence and 3 highly similar genes as well.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- The BLAST evidence for start site 7892 had 2 1:1 alignments
- There were no alternative starts

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Start site 7892
- Z value: 3.146
- Final score: -2.334



Starts : 5
Selected : 1

ORF Start : 7892
ORF Stop : 8164
ORF Length : 273

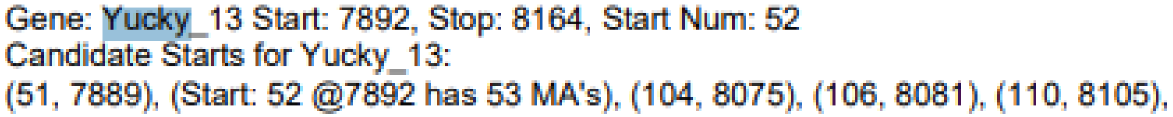
	Cdn 1	Cdn2	Cdn3	Length
5' End	0.0	0.0	100.0	3
3' End	65.9	46.2	61.5	273

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-1.559	3.146	6	-3.304	GGACACATCGTCGTGAGGAGGG	TTG	7889	276
2	-1.559	3.146	9	-2.334	CACATCGTCGTGAGGAGGGTTG	ATG	7892	273
3	-3.267	2.328	13	-4.312	GCGTCCTCAGGGCCGTCATCGC	GTG	8075	90
4	-7.144	0.471	9	-7.919	TCAGGGCCGTCATCGCGTGACC	GTG	8081	84
5	-3.760	2.092	15	-5.362	GTACACGGGAACGCCTGAAGCC	ATG	8105	60

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Yucky_13 had start 52 @7892 with 53 MA’s

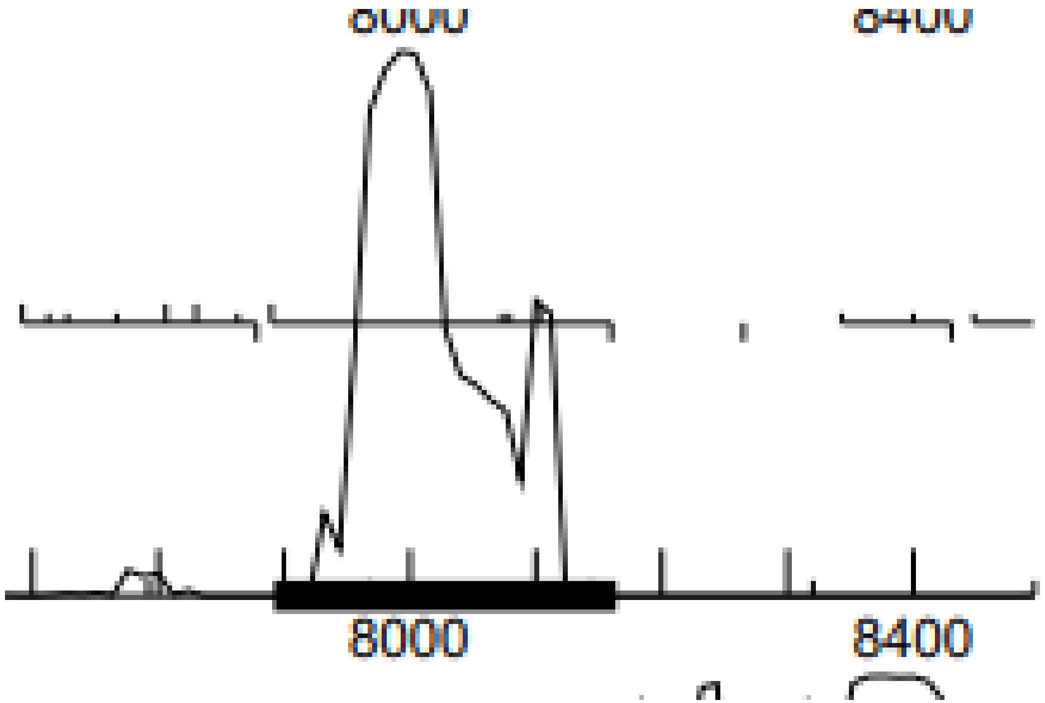


```
Gene: Yucky_13 Start: 7892, Stop: 8164, Start Num: 52
Candidate Starts for Yucky_13:
(51, 7889), (Start: 52 @7892 has 53 MA's), (104, 8075), (106, 8081), (110, 8105),
```

- Yucky was a part of the genes that did not have the “Most Annotated” start along with ChilliPepper_10, Floral_12, and Emalyn_10.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

All coding potential is included at start site 7892.



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Overlap for start site 7892 is 14.

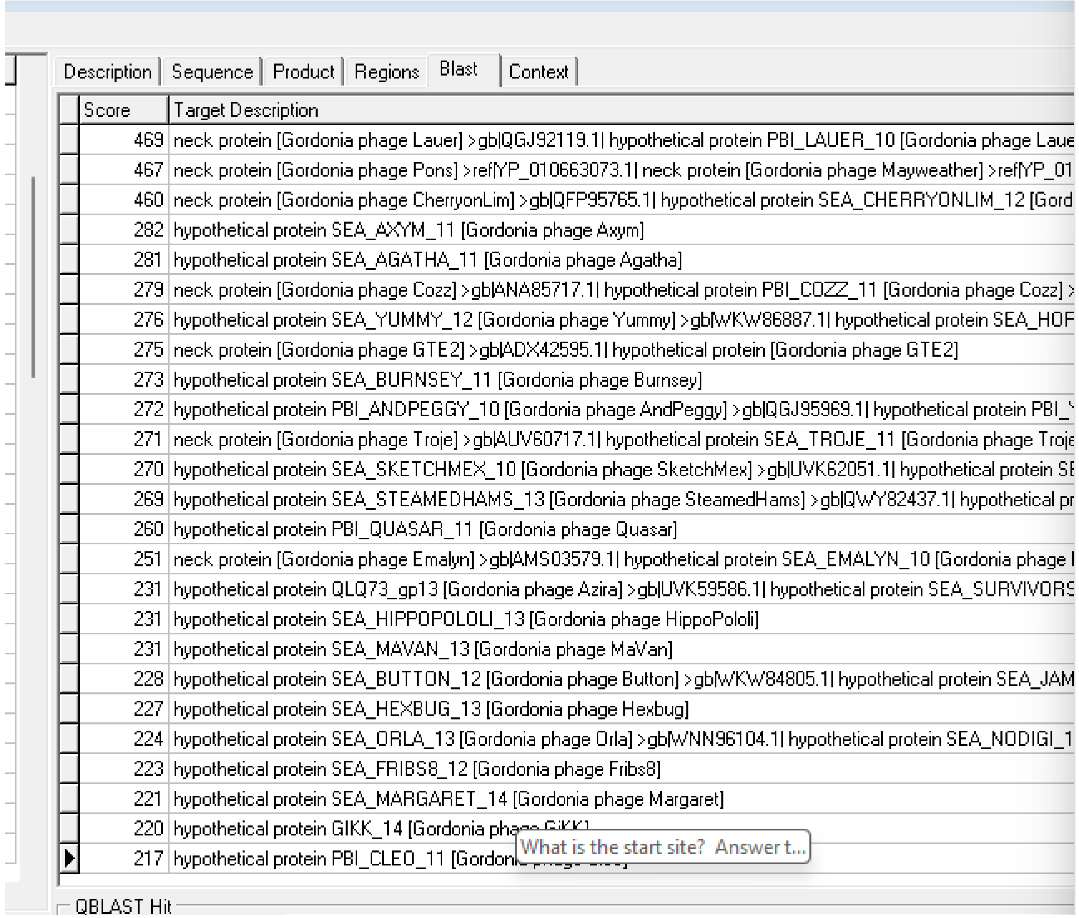
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	7892
GeneMark	Glimmer and GeneMark
Coding potential	All coding potential
RBS	Z value: 3.146 Final score: -2.334
BLAST	2 1:1 alignment
Starterator	53 MAs
Overlap	14

The start site is 7892 because both Glimmer and GeneMark call it at 7892 and the coding potential was within the start site 7892. The z value for start site 7892 is greater than 1 and the start site aligned 1:1 with two other genes Pons, and CherryonLim. The starterator evidence also showed that at start site 7892 there were 53 manual annotations. While there was no gap, feature 13 did overlap with feature 14 and the overlap was 14.

BLAST function evidence. What assigned functions do other highly similar genes have?

- 7 genes with function neck protein
- 18 genes with function hypothetical protein



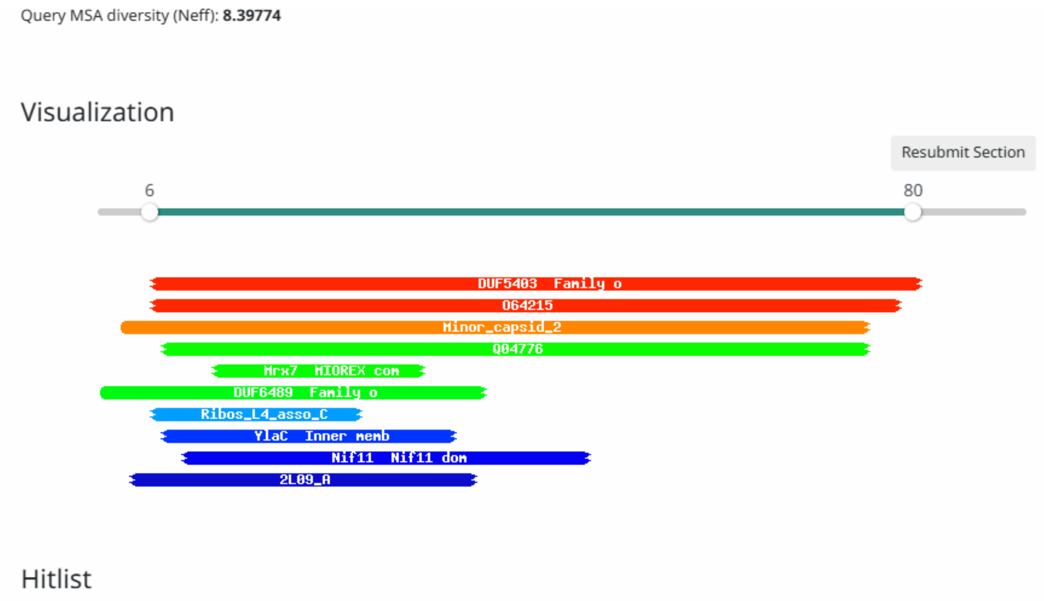
Score	Target Description
469	neck protein [Gordonia phage Lauer] >gb|QGJ92119.1| hypothetical protein PBI_LAUER_10 [Gordonia phage Laue
467	neck protein [Gordonia phage Pons] >ref|YP_010663073.1| neck protein [Gordonia phage Mayweather] >ref|YP_01
460	neck protein [Gordonia phage CherryonLim] >gb|QFP95765.1| hypothetical protein SEA_CHERRYONLIM_12 [Gord
282	hypothetical protein SEA_AXYM_11 [Gordonia phage Axym]
281	hypothetical protein SEA_AGATHA_11 [Gordonia phage Agatha]
279	neck protein [Gordonia phage Cozz] >gb|ANA85717.1| hypothetical protein PBI_COZZ_11 [Gordonia phage Cozz] >
276	hypothetical protein SEA_YUMMY_12 [Gordonia phage Yummy] >gb|WKW86887.1| hypothetical protein SEA_HOF
275	neck protein [Gordonia phage GTE2] >gb|ADX42595.1| hypothetical protein [Gordonia phage GTE2]
273	hypothetical protein SEA_BURNSEY_11 [Gordonia phage Burnsey]
272	hypothetical protein PBI_ANDPEGGY_10 [Gordonia phage AndPeggy] >gb|QGJ95969.1| hypothetical protein PBI_
271	neck protein [Gordonia phage Troje] >gb|AUV60717.1| hypothetical protein SEA_TROJE_11 [Gordonia phage Troje
270	hypothetical protein SEA_SKETCHMEX_10 [Gordonia phage SketchMex] >gb|UVK62051.1| hypothetical protein SE
269	hypothetical protein SEA_STEAMEDHAMS_13 [Gordonia phage SteamedHams] >gb|QWY82437.1| hypothetical pr
260	hypothetical protein PBI_QUASAR_11 [Gordonia phage Quasar]
251	neck protein [Gordonia phage Emalyn] >gb|AMS03579.1| hypothetical protein SEA_EMALYN_10 [Gordonia phage I
231	hypothetical protein QLQ73_gp13 [Gordonia phage Azira] >gb|UVK59586.1| hypothetical protein SEA_SURVIVORS
231	hypothetical protein SEA_HIPPOPOLOLI_13 [Gordonia phage HippoPololi]
231	hypothetical protein SEA_MAVAN_13 [Gordonia phage MaVan]
228	hypothetical protein SEA_BUTTON_12 [Gordonia phage Button] >gb|WKW84805.1| hypothetical protein SEA_JAM
227	hypothetical protein SEA_HEXBUG_13 [Gordonia phage Hexbug]
224	hypothetical protein SEA_ORLA_13 [Gordonia phage Orla] >gb|WNN96104.1| hypothetical protein SEA_NODIGI_1
223	hypothetical protein SEA_FRIBS8_12 [Gordonia phage Fribs8]
221	hypothetical protein SEA_MARGARET_14 [Gordonia phage Margaret]
220	hypothetical protein GIKK_14 [Gordonia phage GiKK]
217	hypothetical protein PBI_CLEO_11 [Gordonia phage Cleo]

What is the start site? Answer t...

QBLAST Hit

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

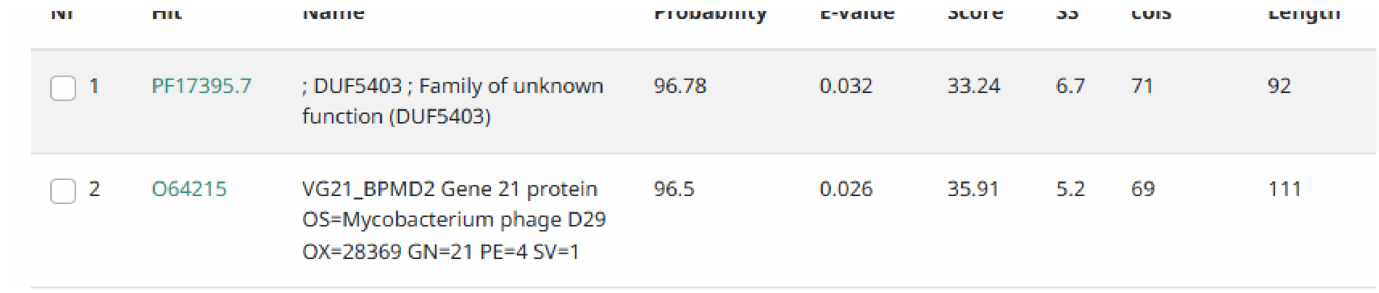
- HHpred data does not support the function as while it has two hits with a probability higher than 90 and an E value less than one, their functions are unknown.



Query MSA diversity (Neff): **8.39774**

Visualization

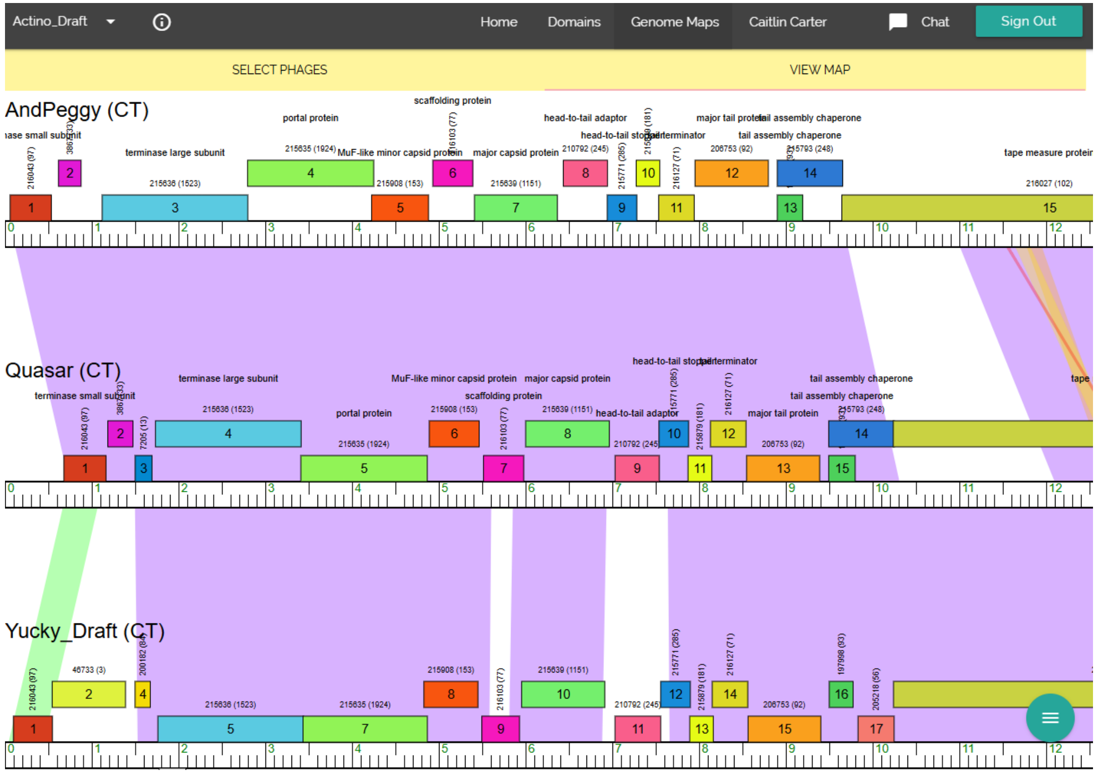
Hitlist



Nr	Hit	Name	Probability	E-value	Score	SS	Cols	Length
□ 1	PF17395.7	; DUF5403 ; Family of unknown function (DUF5403)	96.78	0.032	33.24	6.7	71	92
□ 2	O64215	VG21_BPMD2 Gene 21 protein OS=Mycobacterium phage D29 OX=28369 GN=21 PE=4 SV=1	96.5	0.026	35.91	5.2	69	111

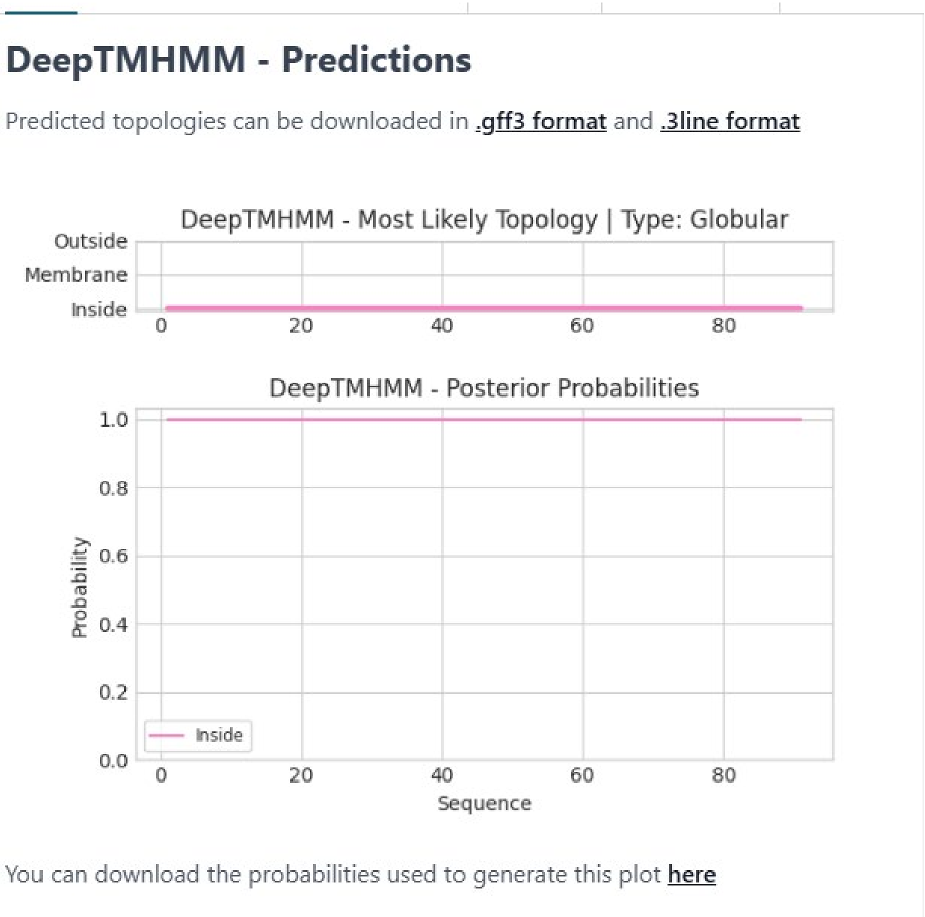
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Yucky feature 13 conserved domain: none function: none
- Quasar feature 11 conserved domain: none function: none
- AndPeggy feature 10 conserved domain: none function: none



Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- # of unnamed predicted TMRs: 0



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

DeepTMHMM - Most Likely Topology | Type: Globular
Outside
Membrane
Inside
0
20
40
60
80
DeepTMHMM - Posterior Probabilities
1.0
0.8
0.6
0.4
0.2
0.0
Probability
Inside
0
20
40
60
80
Sequence

You can download the probabilities used to generate this plot here

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- There is no function, so it is a hypothetical protein, because the Hhpred evidence does not show any matches with a known function and an E value less than 1, and the Phamerator evidence does not show any function for the two highly similar genes; Quasar, AndPeggy. Since no function was defined, I turned to DeepTMHMM evidence, which did not determine the function as there were zero unnamed number of predicted TMRS.

Feature 12 – Stop 8561

Glimmer/GeneMark

What feature number is this? 12

What is the stop site? 8561

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Called by Glimmer and GeneMark

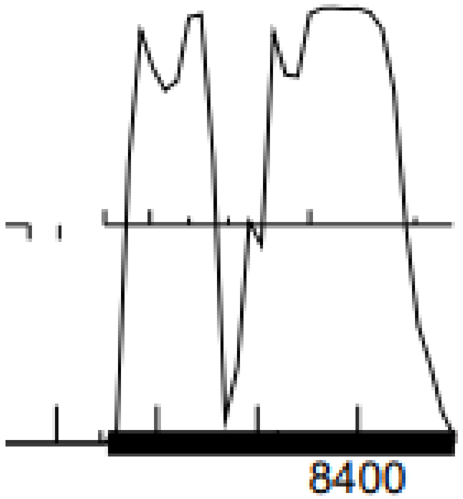
What is the autoannotated start?

8151

Gap: or overlap: 14 (with gene in front of it) for the autoannotated start

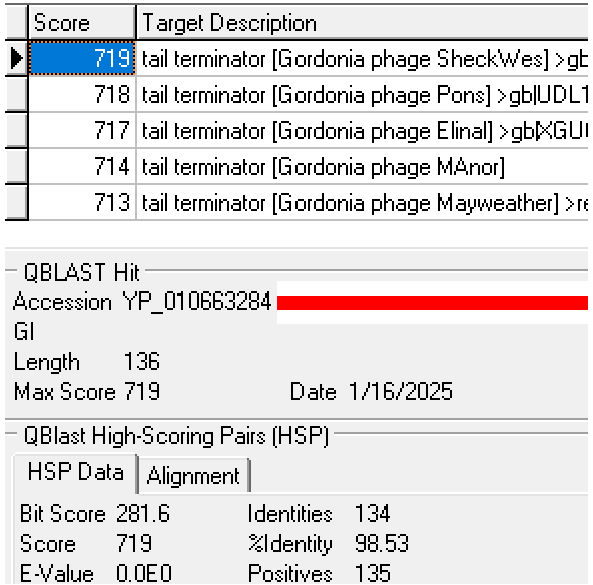
- Overlap of 14
- Called by both

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- There is a strong peak of coding potential briefly, before it falls and respikes into another strong peak. Reading frame 3 is the only frame with coding potential for this subsequence of nucleotides.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
719	tail terminator [Gordonia phage SheckWes] >gb
718	tail terminator [Gordonia phage Pons] >gb|UDL1
717	tail terminator [Gordonia phage Elinal] >gb|XGU
714	tail terminator [Gordonia phage MAnor]
713	tail terminator [Gordonia phage Mayweather] >r

QBLAST Hit

Accession YP_010663284

GI

Length 136

Max Score 719 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

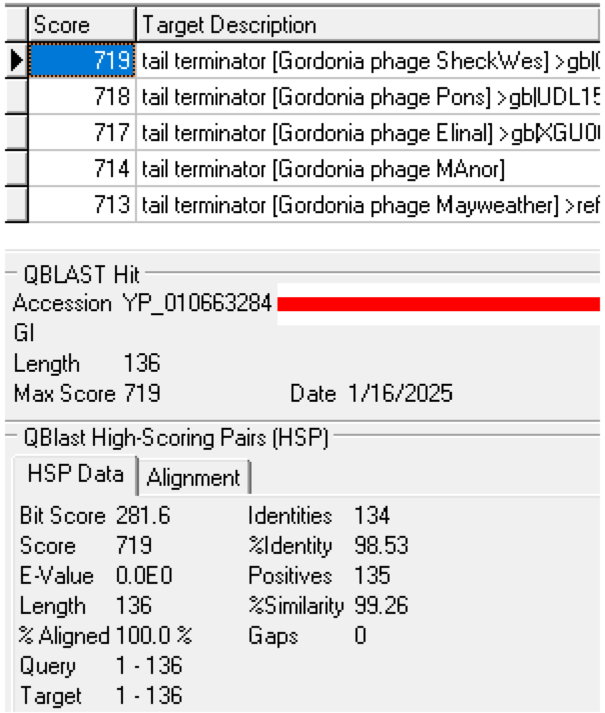
Bit Score	281.6	Identities	134
Score	719	%Identity	98.53
E-Value	0.0E0	Positives	135

- There are 16 BLAST hits with an E-value close to 0.
- There are 8 1:1 alignments.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- I believe this is a gene. Both Glimmer and GeneMark called it a gene. There is some coding potential in the sequence of nucleotides of this gene. There are also several highly similar BLAST results. This evidence leads me to believe this feature is a gene.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
719	tail terminator [Gordonia phage SheckWes] >gb|
718	tail terminator [Gordonia phage Pons] >gb|UDL15
717	tail terminator [Gordonia phage Elinal] >gb|XGU0
714	tail terminator [Gordonia phage MAnor]
713	tail terminator [Gordonia phage Mayweather] >ref

QBLAST Hit

Accession YP_010663284

GI

Length 136

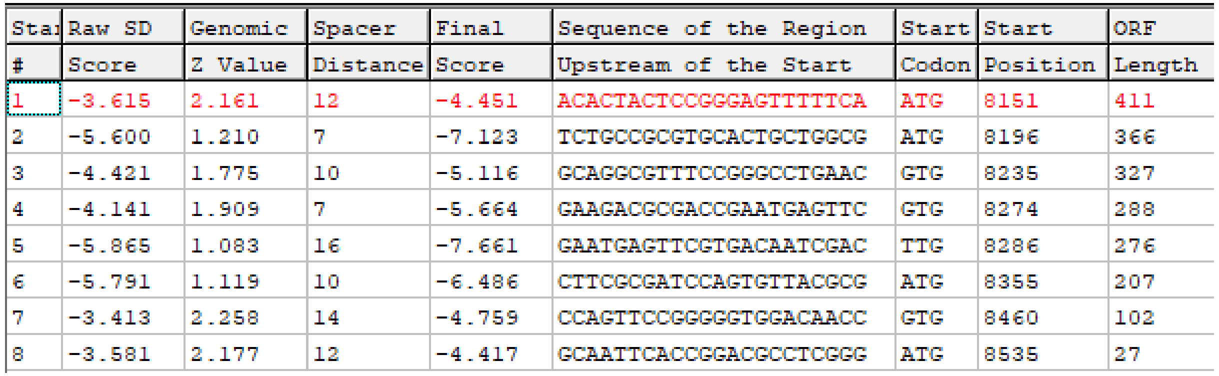
Max Score 719 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	281.6	Identities	134
Score	719	%Identity	98.53
E-Value	0.0E0	Positives	135
Length	136	%Similarity	99.26
% Aligned	100.0 %	Gaps	0
Query	1 - 136		
Target	1 - 136		

- There are 8 1:1 alignments in the BLAST data. There are many other close alignments, such as 8:9 or 8:10.

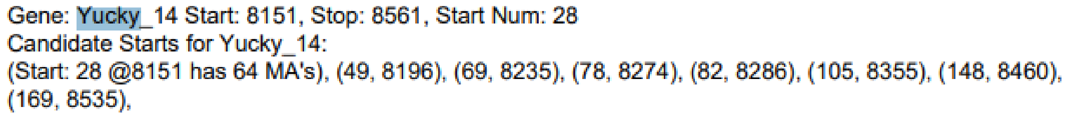
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-3.615	2.161	12	-4.451	ACACTACTCCGGGAGTTTTTCA	ATG	8151	411
2	-5.600	1.210	7	-7.123	TCTGCCGCGTGCACTGCTGGCG	ATG	8196	366
3	-4.421	1.775	10	-5.116	GCAGGCGTTTCCGGGCCTGAAC	GTG	8235	327
4	-4.141	1.909	7	-5.664	GAAGACGCGACCGAATGAGTTC	GTG	8274	288
5	-5.865	1.083	16	-7.661	GAATGAGTTCGTGACAATCGAC	TTG	8286	276
6	-5.791	1.119	10	-6.486	CTTCGCGATCCAGTGTTACGCG	ATG	8355	207
7	-3.413	2.258	14	-4.759	CCAGTTCCGGGGGTGGACAACC	GTG	8460	102
8	-3.581	2.177	12	-4.417	GCAATTCACCGGACGCCTCGGG	ATG	8535	27

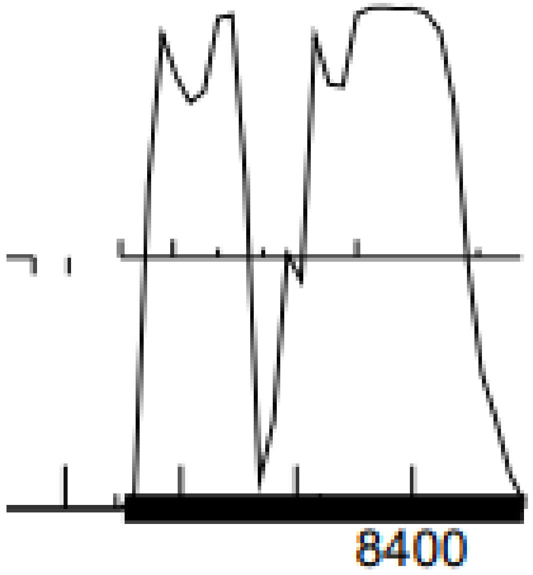
- When looking at RBS values, the autoannotated start site was the only site that looked possible as the true start site. It had a Z-value of 2.161 and a final score of -4.451. These numbers are better than most of the other available starts. The ones that are better are too far along the sequence to be the start.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.


Gene: Yucky_14 Start: 8151, Stop: 8561, Start Num: 28
Candidate Starts for Yucky_14:
(Start: 28 @8151 has 64 MA's), (49, 8196), (69, 8235), (78, 8274), (82, 8286), (105, 8355), (148, 8460), (169, 8535),

- There are 64 MAs for the autoannotated start of 8151. It is the only start site to have any manual annotations, and it is called 98.8% of the time when present.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- The only start site that makes sense, 8151, cuts off no coding potential. The coding potential looks like it bgins at around 8160.

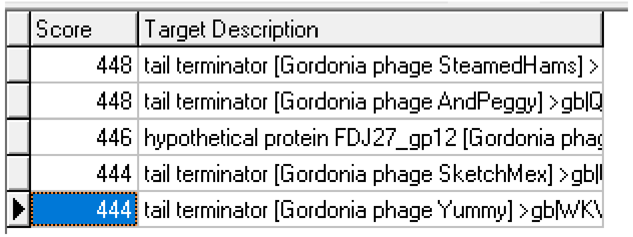
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- There is an overlap of 14 for the autoannotated start site. This is within the 30 or less range we consider acceptable.
- 8164-8151=13+1 for overlap= 14

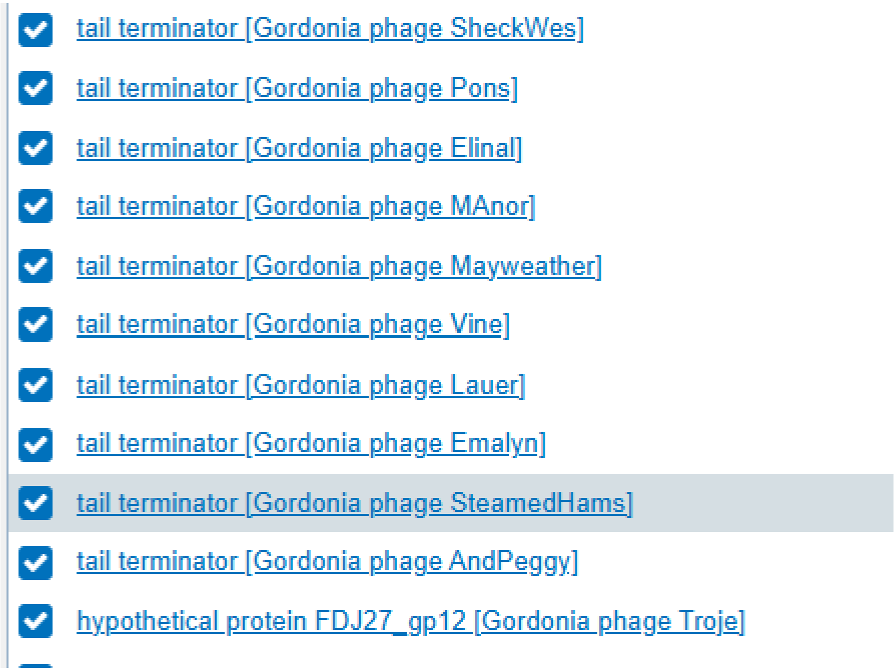
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is 8151. My call agrees with the automated start site. This site has 8 1:1 alignments. It also has good RBS numbers: Z-value of 2.161 and a final score of -4.451. It is the only site to ever be manually annotated, and it has 64 Mas. It cuts off no coding potential and has an acceptable overlap. It is the only start site that makes sense based on this evidence, and its placement in the sequence.

BLAST function evidence. What assigned functions do other highly similar genes have?



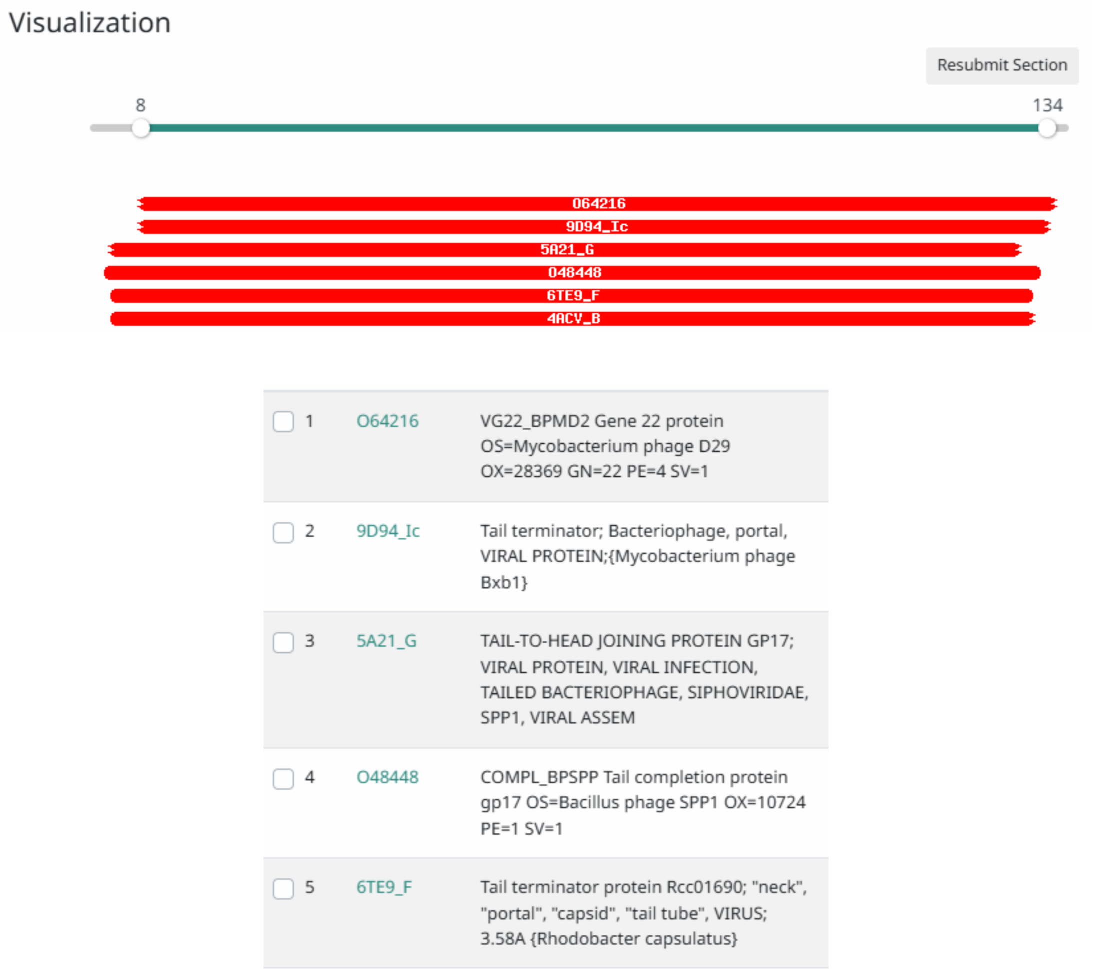
Score	Target Description
448	tail terminator [Gordonia phage SteamedHams] >
448	tail terminator [Gordonia phage AndPeggy] >gb|Q
446	hypothetical protein FDJ27_gp12 [Gordonia phag
444	tail terminator [Gordonia phage SketchMex] >gb|
444	tail terminator [Gordonia phage Yummy] >gb|WK\



- ☑ tail terminator [Gordonia phage SheckWes]
- ☑ tail terminator [Gordonia phage Pons]
- ☑ tail terminator [Gordonia phage Elinal]
- ☑ tail terminator [Gordonia phage MAnor]
- ☑ tail terminator [Gordonia phage Mayweather]
- ☑ tail terminator [Gordonia phage Vine]
- ☑ tail terminator [Gordonia phage Lauer]
- ☑ tail terminator [Gordonia phage Emalyn]
- ☑ tail terminator [Gordonia phage SteamedHams]
- ☑ tail terminator [Gordonia phage AndPeggy]
- ☑ hypothetical protein FDJ27_gp12 [Gordonia phage Troje]

- DNA master BLAST showed 22 similar genes with the function tail terminator, and 3 with hypothetical proteins.
- BLASTing on NCBI revealed that the best matches were labeled as tail terminators, however there were still some hypothetical proteins and a couple tail completion proteins, however these were not as good of matches.

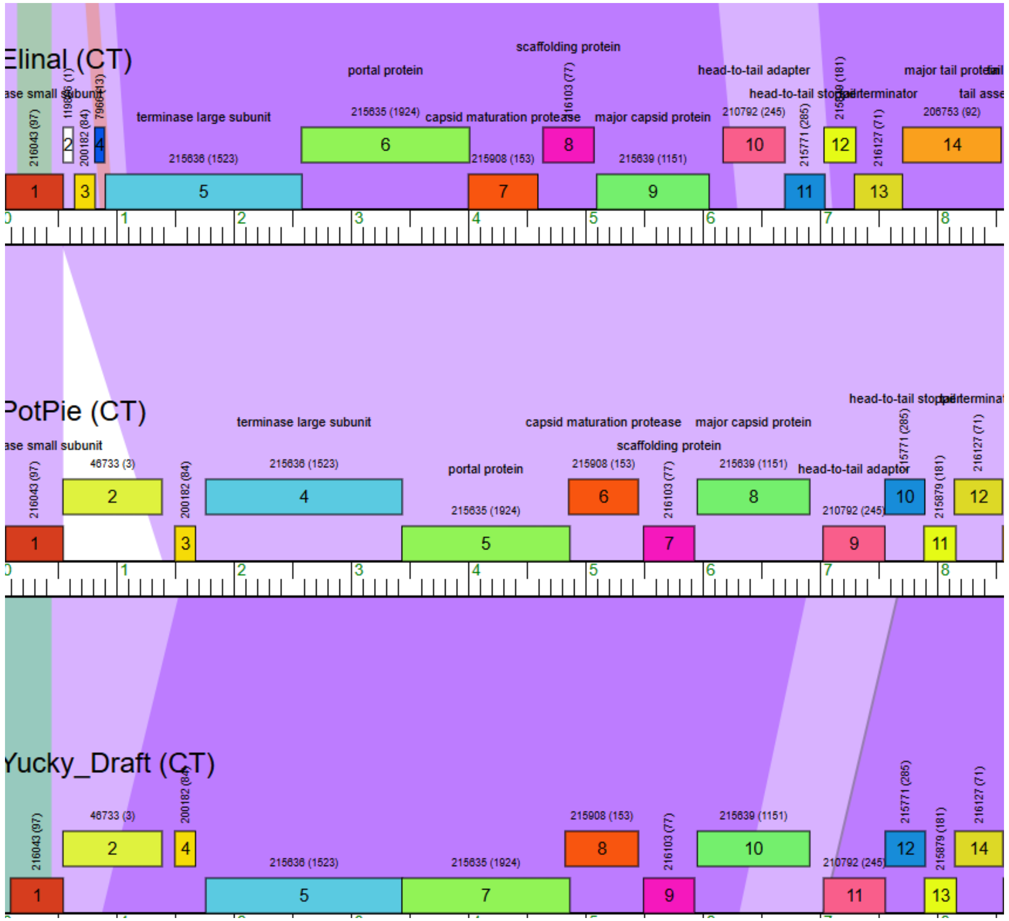
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



☐	1	O64216	VG22_BPMD2 Gene 22 protein OS=Mycobacterium phage D29 OX=28369 GN=22 PE=4 SV=1
☐	2	9D94_Ic	Tail terminator; Bacteriophage, portal, VIRAL PROTEIN;{Mycobacterium phage Bxb1}
☐	3	5A21_G	TAIL-TO-HEAD JOINING PROTEIN GP17; VIRAL PROTEIN, VIRAL INFECTION, TAILED BACTERIOPHAGE, SIPHOVIRIDAE, SPP1, VIRAL ASSEM
☐	4	O48448	COMPL_BPSPP Tail completion protein gp17 OS=Bacillus phage SPP1 OX=10724 PE=1 SV=1
☐	5	6TE9_F	Tail terminator protein Rcc01690; "neck", "portal", "capsid", "tail tube", VIRUS; 3.58A {Rhodobacter capsulatus}

- There are 24 good hits with a couple of functions, mostly tail terminator. Out of the 24 red colored hits they were all mostly homologous, some were less homologous at the beginning.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- PotPie, BigChungus, and Elinal all have this gene and in all 3 it is a tail terminator. There are no conserved domains in any of the phages.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- I would like to call this gene a tail terminator.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I believe this gene to be a tail terminator. The majority of DNA master and NCBI BLAST hits show this. HHpred also has many hits showing a tail terminator. Lastly, Phamerator shows that other phages in the same cluster have this gene and that tail terminator is the function on these genes.

Feature 13 – Stop 9403

Glimmer/GeneMark

What feature number is this? 13

What is the stop site? 9403

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Called by both Glimmer and GeneMark

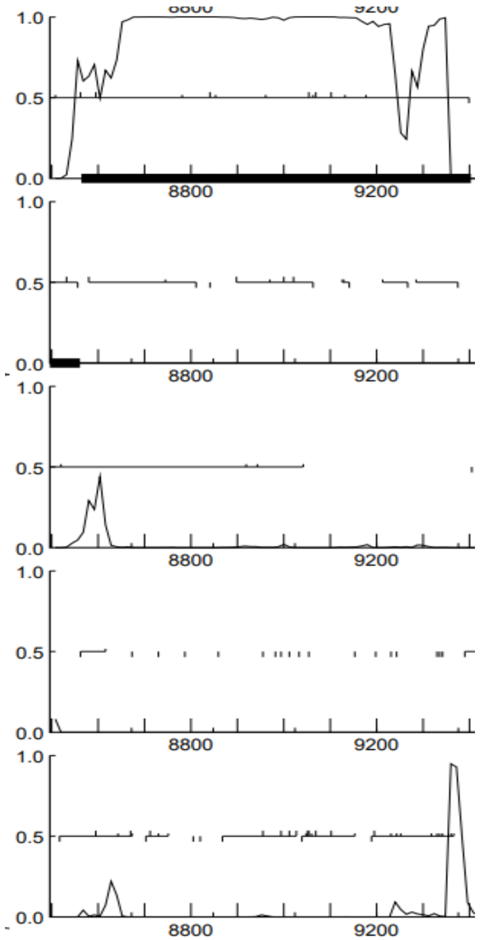
What is the autoannotated start?

8564

Gap: 2 or overlap: (with gene in front of it) for the autoannotated start

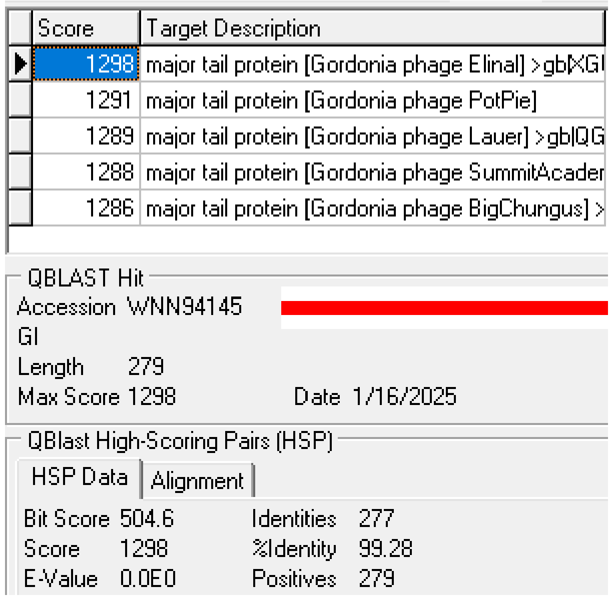
- Called by both
- Gap of 2

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Reading frame 2 contains a massive, strong spike of coding potential that lasts for a very long time. Reverse reading frame 4 contains 1 weak peak of coding potential. Reverse reading frame 6 contains 1 weak peak and one strong peak, neither sustained.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
1298	major tail protein [Gordonia phage Elinal] >gb|XGI
1291	major tail protein [Gordonia phage PotPie]
1289	major tail protein [Gordonia phage Lauer] >gb|QG
1288	major tail protein [Gordonia phage SummitAcader
1286	major tail protein [Gordonia phage BigChungus] >

QBLAST Hit

Accession WNN94145

GI

Length 279

Max Score 1298

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

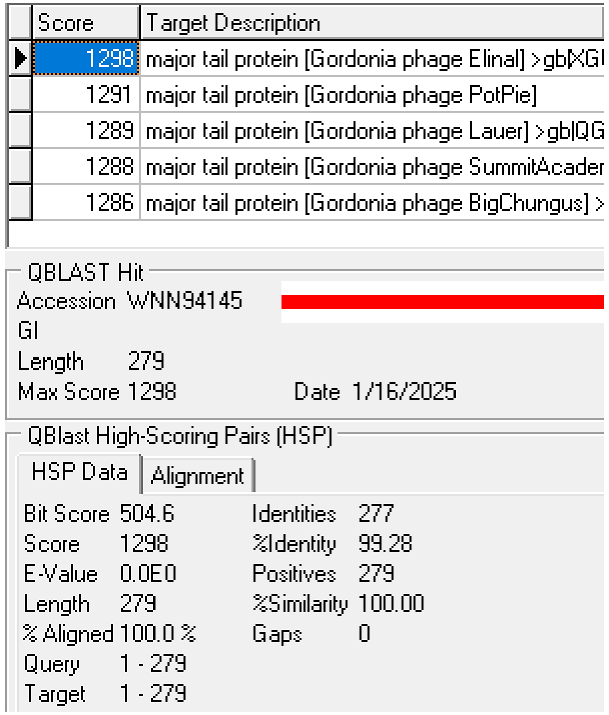
Bit Score	504.6	Identities	277
Score	1298	%Identity	99.28
E-Value	0.0E0	Positives	279

- There are at least 25 similar genes with an E-value close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This is a gene. Both Glimmer and GeneMark call it a gene, it has a massive strong peak of coding potential that is sustained through the entire feature, and it has at least 25 BLAST hits with an E-value close to 0. This evidence makes it clear that this is a gene.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
1298	major tail protein [Gordonia phage Elinal] >gb|XGI
1291	major tail protein [Gordonia phage PotPie]
1289	major tail protein [Gordonia phage Lauer] >gb|QG
1288	major tail protein [Gordonia phage SummitAcader
1286	major tail protein [Gordonia phage BigChungus] >

QBLAST Hit

Accession WNN94145

GI

Length 279

Max Score 1298

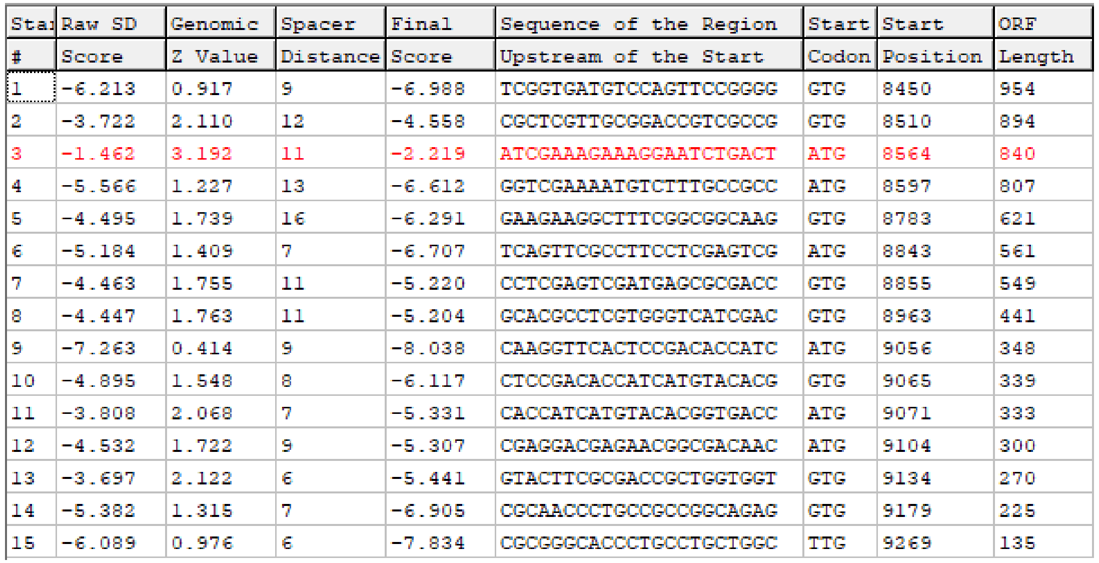
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	504.6	Identities	277
Score	1298	%Identity	99.28
E-Value	0.0E0	Positives	279
Length	279	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 279		
Target	1 - 279		

- There are at least 25 1:1 alignments revealed by BLAST. There are no known alternate starts yet, as Glimmer and GeneMark agreed.

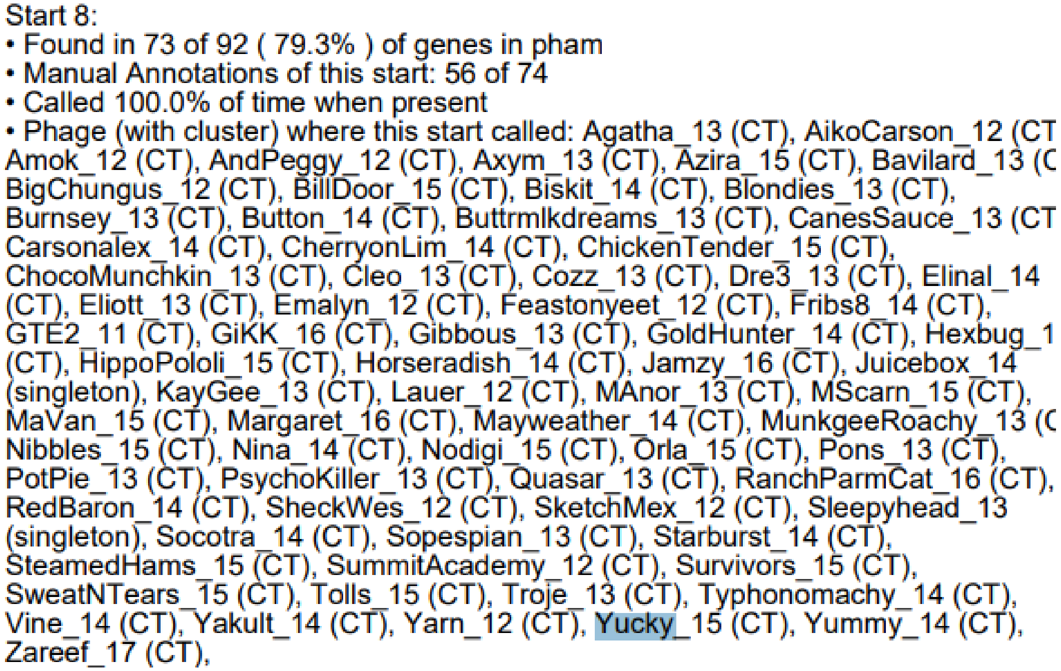
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-6.213	0.917	9	-6.988	TCGGTGATGTCCAGTTCCGGGG	GTG	8450	954
2	-3.722	2.110	12	-4.558	CGCTCGTTGCGGACCGTCGCCG	GTG	8510	894
3	-1.462	3.192	11	-2.219	ATCGAAAGAAAGGAATCTGACT	ATG	8564	840
4	-5.566	1.227	13	-6.612	GGTCGAAAATGTCTTTGCCGCC	ATG	8597	807
5	-4.495	1.739	16	-6.291	GAAGAAGGCTTTCGGCGGCAAG	GTG	8783	621
6	-5.184	1.409	7	-6.707	TCAGTTCGCCTTCCTCGAGTCG	ATG	8843	561
7	-4.463	1.755	11	-5.220	CCTCGAGTCGATGAGCGCGACC	GTG	8855	549
8	-4.447	1.763	11	-5.204	GCACGCCTCGTGGGTCATCGAC	GTG	8963	441
9	-7.263	0.414	9	-8.038	CAAGGTTCACTCCGACACCATC	ATG	9056	348
10	-4.895	1.548	8	-6.117	CTCCGACACCATCATGTACACG	GTG	9065	339
11	-3.808	2.068	7	-5.331	CACCATCATGTACACGGTGACC	ATG	9071	333
12	-4.532	1.722	9	-5.307	CGAGGACGAGAACGGCGACAAC	ATG	9104	300
13	-3.697	2.122	6	-5.441	GTACTTCGCGACCGCTGGTGGT	GTG	9134	270
14	-5.382	1.315	7	-6.905	CGCAACCCTGCCGCCGGCAGAG	GTG	9179	225
15	-6.089	0.976	6	-7.834	CGCGGGCACCCTGCCTGCTGGC	TTG	9269	135

- The Z-value of the automated start is 3.192. The final score of the automated start is -2.219. No other RBS numbers are even close to good.

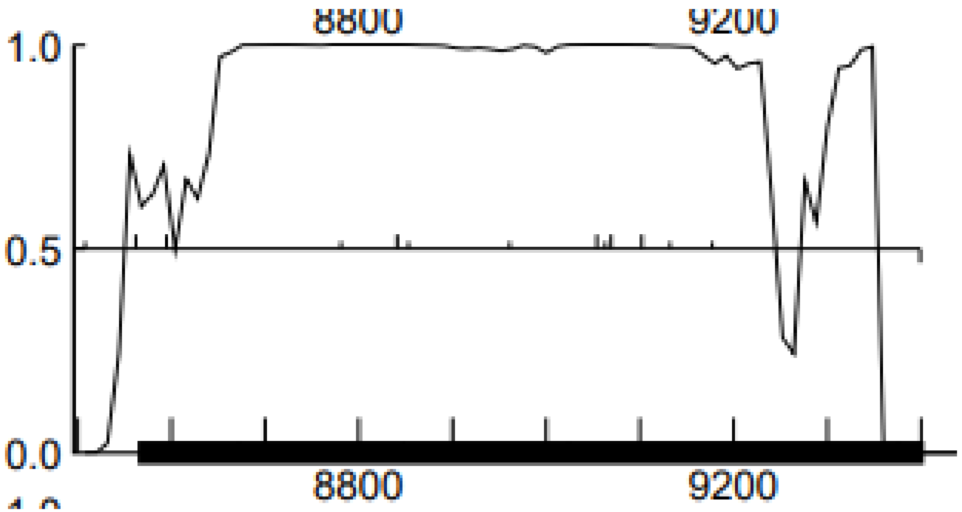
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



Start 8:
- Found in 73 of 92 (79.3%) of genes in pham
- Manual Annotations of this start: 56 of 74
- Called 100.0% of time when present
- Phage (with cluster) where this start called: Agatha_13 (CT), AikoCarson_12 (CT
Amok_12 (CT), AndPeggy_12 (CT), Axym_13 (CT), Azira_15 (CT), Bavilard_13 (C
BigChungus_12 (CT), BillDoor_15 (CT), Biskit_14 (CT), Blondies_13 (CT),
Burnsey_13 (CT), Button_14 (CT), Buttrmlkdreams_13 (CT), CanesSauce_13 (CT
Carsonalex_14 (CT), CherryonLim_14 (CT), ChickenTender_15 (CT),
ChocoMunchkin_13 (CT), Cleo_13 (CT), Cozz_13 (CT), Dre3_13 (CT), Elinal_14
(CT), Eliott_13 (CT), Emalyn_12 (CT), Feastonyeet_12 (CT), Fribs8_14 (CT),
GTE2_11 (CT), GiKK_16 (CT), Gibbous_13 (CT), GoldHunter_14 (CT), Hexbug_1
(CT), HippoPololi_15 (CT), Horseradish_14 (CT), Jamzy_16 (CT), Juicebox_14
(singleton), KayGee_13 (CT), Lauer_12 (CT), MAnor_13 (CT), MScarn_15 (CT),
MaVan_15 (CT), Margaret_16 (CT), Mayweather_14 (CT), MunkgeeRoachy_13 (C
Nibbles_15 (CT), Nina_14 (CT), Nodigi_15 (CT), Orla_15 (CT), Pons_13 (CT),
PotPie_13 (CT), PsychoKiller_13 (CT), Quasar_13 (CT), RanchParmCat_16 (CT),
RedBaron_14 (CT), SheckWes_12 (CT), SketchMex_12 (CT), Sleepyhead_13
(singleton), Socotra_14 (CT), Sopespian_13 (CT), Starburst_14 (CT),
SteamedHams_15 (CT), SummitAcademy_12 (CT), Survivors_15 (CT),
SweatNTears_15 (CT), Tolls_15 (CT), Troje_13 (CT), Typhonomachy_14 (CT),
Vine_14 (CT), Yakult_14 (CT), Yarn_12 (CT), Yucky_15 (CT), Yummy_14 (CT),
Zareef_17 (CT),

- The automated start site has 56 MAs. No other start site has ever been manually annotated. 8564 is called 100% of the time when present.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- 8564 does cut off a slight bit of coding potential. It seemingly cuts off the beginning of a peak, however it cuts very little.

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

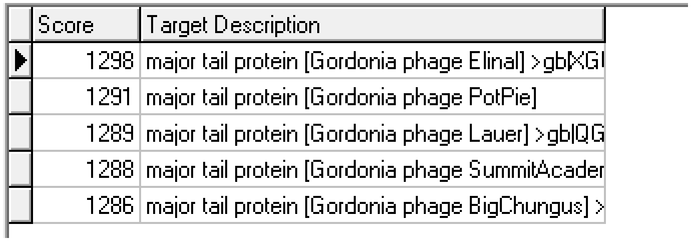
- 8564-8561=3-1 for gap =2

- There is a gap of 2 with the previous gene.

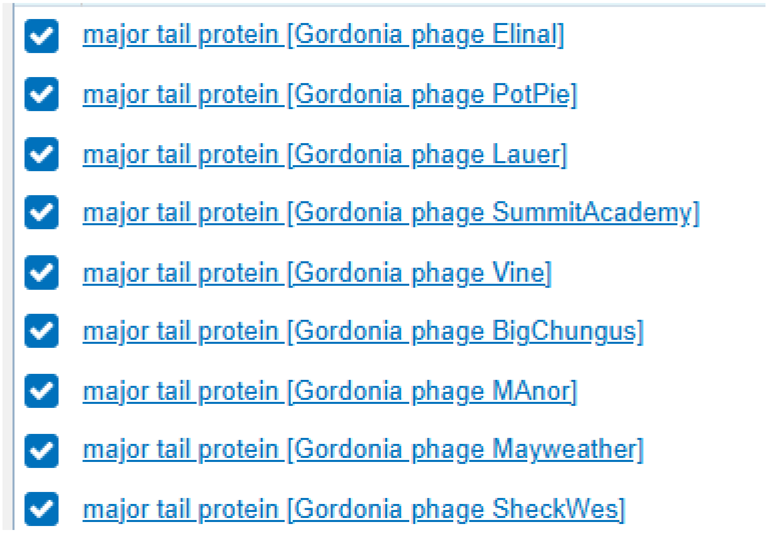
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is 8564. It has at least 25 BLAST hits with 1:1 alignments, great RBS numbers, especially when compared to other start sites, it is the only site to ever be manually annotated for this gene, it cuts off very little coding potential, and it has an acceptable gap with the previous gene.

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
1298	major tail protein [Gordonia phage Elinal] >gb|XGI
1291	major tail protein [Gordonia phage PotPie]
1289	major tail protein [Gordonia phage Lauer] >gb|QG
1288	major tail protein [Gordonia phage SummitAcader
1286	major tail protein [Gordonia phage BigChungus] >



☑ major tail protein [Gordonia phage Elinal]

☑ major tail protein [Gordonia phage PotPie]

☑ major tail protein [Gordonia phage Lauer]

☑ major tail protein [Gordonia phage SummitAcademy]

☑ major tail protein [Gordonia phage Vine]

☑ major tail protein [Gordonia phage BigChungus]

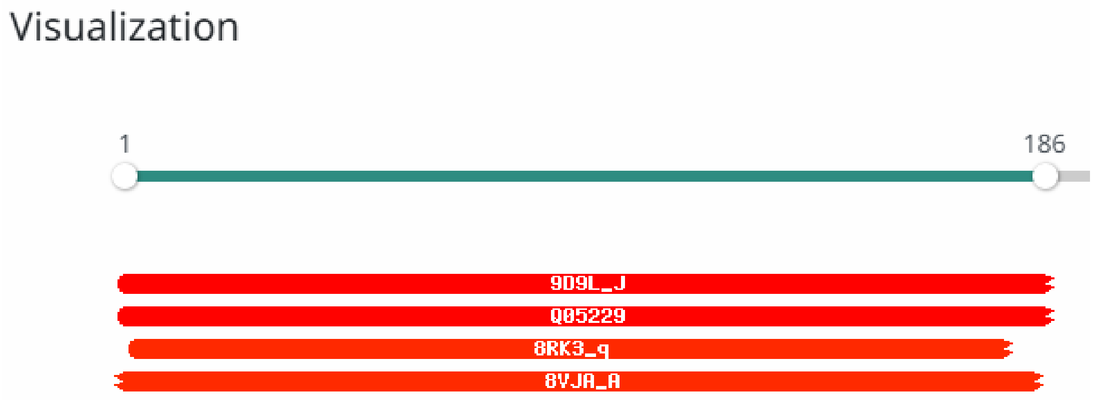
☑ major tail protein [Gordonia phage MAnor]

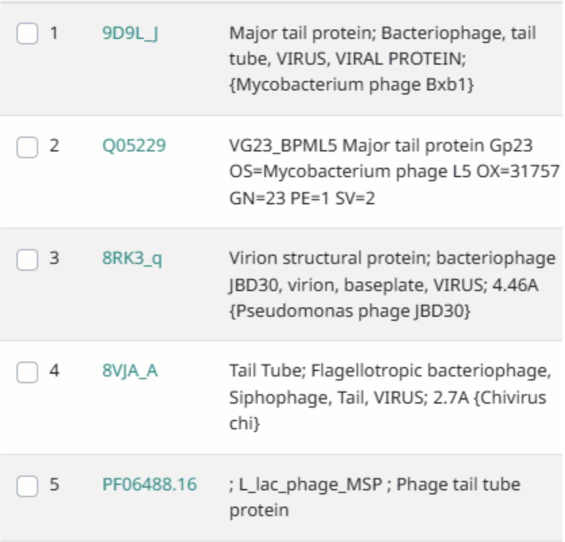
☑ major tail protein [Gordonia phage Mayweather]

☑ major tail protein [Gordonia phage SheckWes]

- DNA master BLAST shows at least 25 hits as a major tail protein.
- BLASTing on NCBI yielded the same results.

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

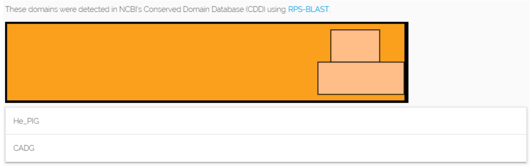




□	1	9D9L_J	Major tail protein; Bacteriophage, tail tube, VIRUS, VIRAL PROTEIN; {Mycobacterium phage Bxb1}
□	2	Q05229	VG23_BPML5 Major tail protein Gp23 OS=Mycobacterium phage L5 OX=31757 GN=23 PE=1 SV=2
□	3	8RK3_q	Virion structural protein; bacteriophage JBD30, virion, baseplate, VIRUS; 4.46A {Pseudomonas phage JBD30}
□	4	8VJA_A	Tail Tube; Flagellotropic bacteriophage, Siphophage, Tail, VIRUS; 2.7A {Chivirus chi}
□	5	PF06488.16	; L_lac_phage_MSP ; Phage tail tube protein

- There are 4 strong Hhpred hits, only 2 of them showed as being a major tail protein. The strong hits are largely homologous throughout.

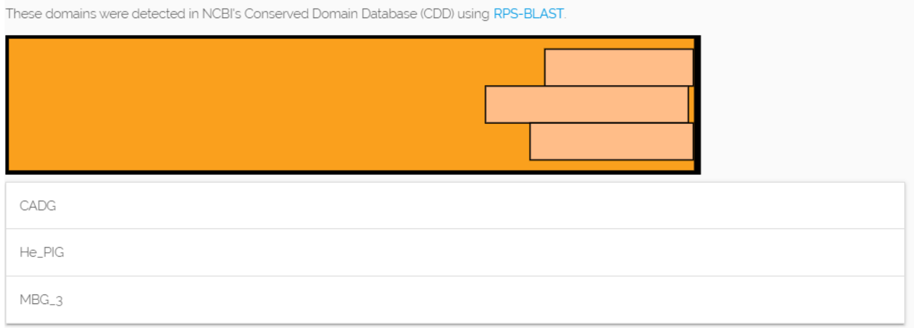
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST

He_PIG

CADG

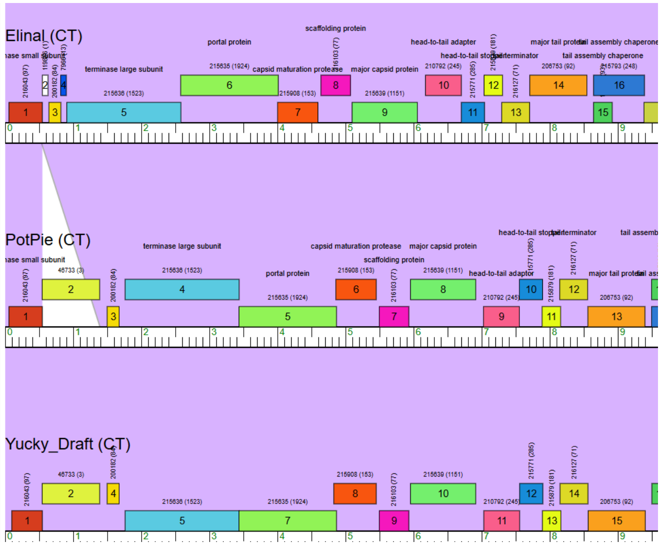


These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

CADG

He_PIG

MBG_3



- BigChungus, Elinal, and PotPie all show have this gene and have it called as a major tail protein.
- PotPie has 3 conserved domains.
- Elinal has the same 3 conserved domains.
- BigChungus has 2 conserved domains.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- I would like to call this gene as a major tail protein.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I officially call this gene as a major tail protein. Both DNA master and NCBI BLAST showed many highly similar phages with this feature being a major tail protein. Phamerator also showed 3 phages very similar to ours as having this gene being a major tail protein. The Hhpred evidence is the best, but the 2 strongest hits are still showing the gene as a major tail protein. Thus, I believe this gene to be a major tail protein.

Feature 14 – Stop 9775

Glimmer/GeneMark

What feature number is this? 14

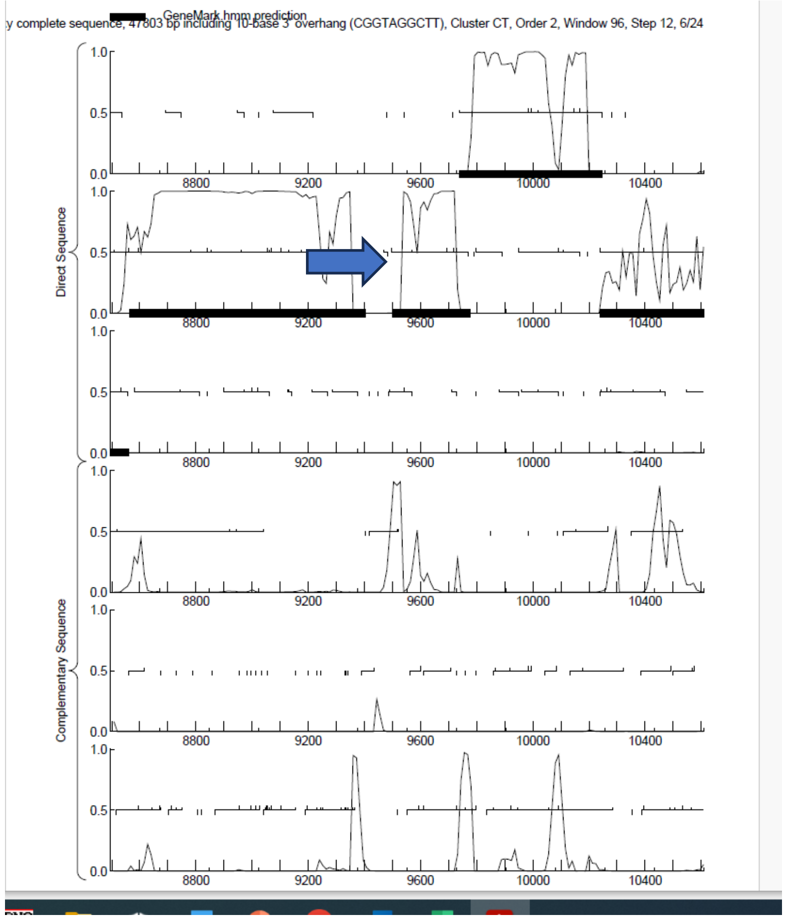
What is the stop site? 9775

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? Called by glimmer and GeneMark

What is the autoannotated start? 9500

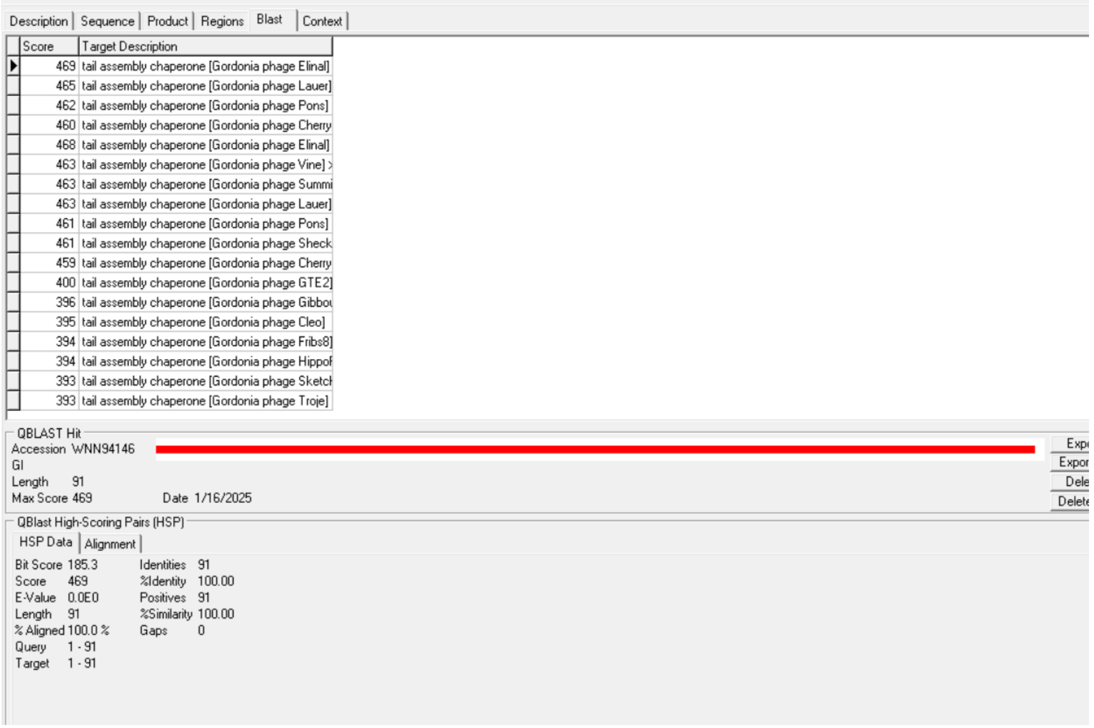
Gap: _____96 with feature in front of it___ or overlap: ________ (with gene in front of it) for the autoannotated start

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Good coding potential in Forward frame 2. Some coding potential in reading frame -1, but reverse reading frame not called.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
469	tail assembly chaperone [Gordonia phage Elinal]
465	tail assembly chaperone [Gordonia phage Lauer]
462	tail assembly chaperone [Gordonia phage Pons]
460	tail assembly chaperone [Gordonia phage Cherry
468	tail assembly chaperone [Gordonia phage Elinal]
463	tail assembly chaperone [Gordonia phage Vine] >
463	tail assembly chaperone [Gordonia phage Summi
463	tail assembly chaperone [Gordonia phage Lauer]
461	tail assembly chaperone [Gordonia phage Pons]
461	tail assembly chaperone [Gordonia phage Sheck
459	tail assembly chaperone [Gordonia phage Cherry
400	tail assembly chaperone [Gordonia phage GTE2]
396	tail assembly chaperone [Gordonia phage Gibbou
395	tail assembly chaperone [Gordonia phage Cleo]
394	tail assembly chaperone [Gordonia phage Fribs8]
394	tail assembly chaperone [Gordonia phage HippoF
393	tail assembly chaperone [Gordonia phage Sketch
393	tail assembly chaperone [Gordonia phage Troje]

QBLAST Hit
Accession WNN94146
GI
Length 91
Max Score 469 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

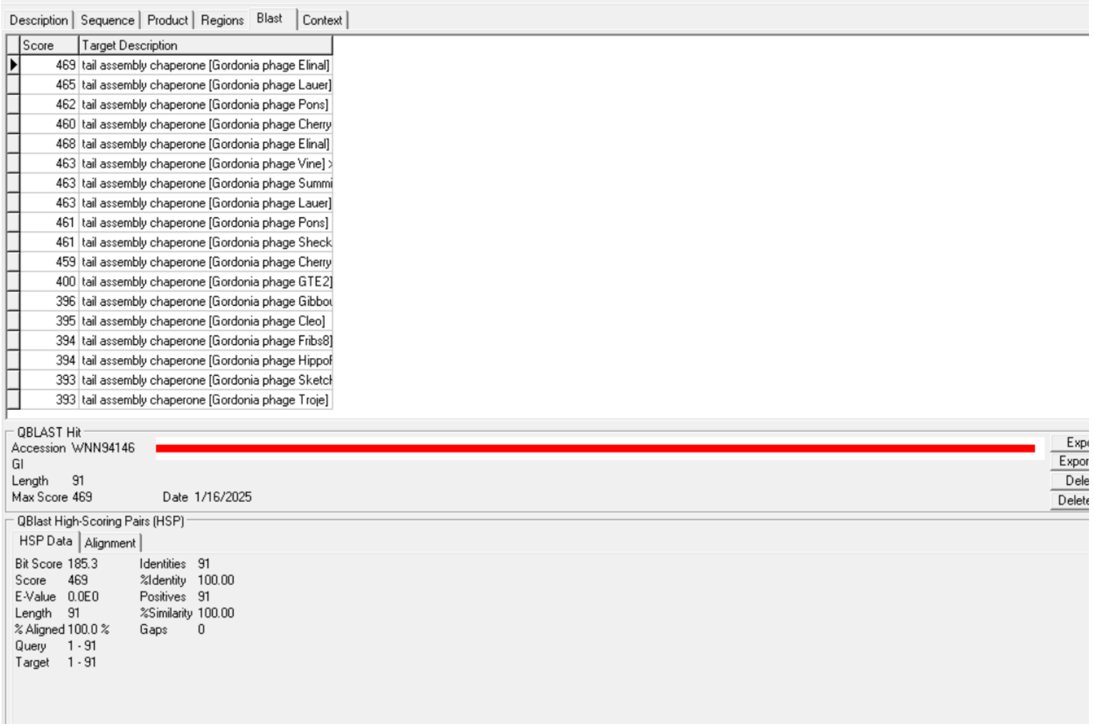
Bit Score	185.3	Identities	91
Score	469	%Identity	100.00
E-Value	0.0E0	Positives	91
Length	91	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 91		
Target	1 - 91		

- At least 25 genes with E values at 0 indidicating close matches with similar genes.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, this is a gene. Good coding potential. 25 blast matches with e values close to zero. Called by both glimmer and genemark.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
469	tail assembly chaperone [Gordonia phage Elinal]
465	tail assembly chaperone [Gordonia phage Lauer]
462	tail assembly chaperone [Gordonia phage Pons]
460	tail assembly chaperone [Gordonia phage Cherry
468	tail assembly chaperone [Gordonia phage Elinal]
463	tail assembly chaperone [Gordonia phage Vine] >
463	tail assembly chaperone [Gordonia phage Summi
463	tail assembly chaperone [Gordonia phage Lauer]
461	tail assembly chaperone [Gordonia phage Pons]
461	tail assembly chaperone [Gordonia phage Sheck
459	tail assembly chaperone [Gordonia phage Cherry
400	tail assembly chaperone [Gordonia phage GTE2]
396	tail assembly chaperone [Gordonia phage Gibbou
395	tail assembly chaperone [Gordonia phage Cleo]
394	tail assembly chaperone [Gordonia phage Fribs8]
394	tail assembly chaperone [Gordonia phage Hippof
393	tail assembly chaperone [Gordonia phage Sketcl
393	tail assembly chaperone [Gordonia phage Troje]

QBLAST Hit

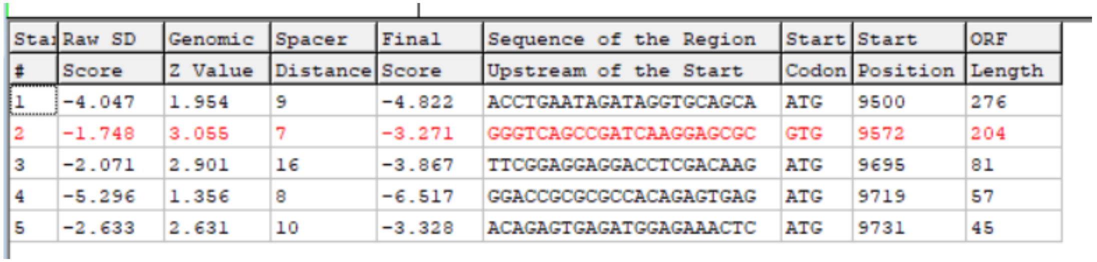
Accession WNN94146
GI
Length 91
Max Score 469
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

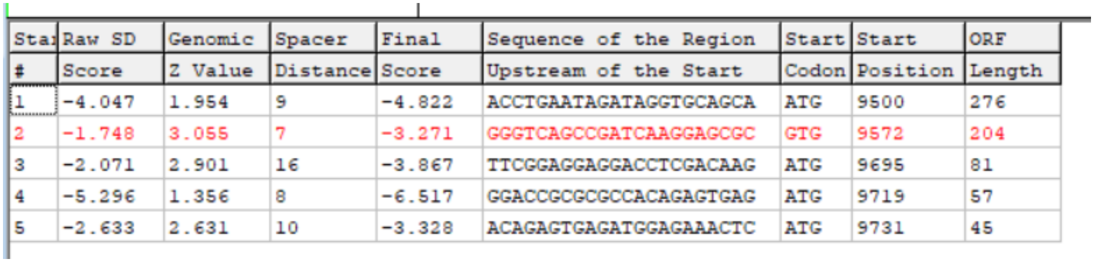
Bit Score	185.3	Identities	91
Score	469	%Identity	100.00
E-Value	0.0E0	Positives	91
Length	91	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 91		
Target	1 - 91		

- 24 1:1 alignments for predicted start of 9500. This start is favored based on BLAST alignment evidence.

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

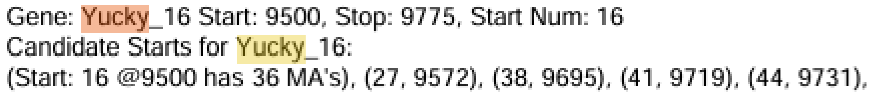


Sta	Raw SD	Genomic	Spacer	Final	Sequence of the Region	Start	Start	ORF
#	Score	Z Value	Distance	Score	Upstream of the Start	Codon	Position	Length
1	-4.047	1.954	9	-4.822	ACCTGAATAGATAGGTGCAGCA	ATG	9500	276
2	-1.748	3.055	7	-3.271	GGGTCAGCCGATCAAGGAGCGC	GTG	9572	204
3	-2.071	2.901	16	-3.867	TTCGGAGGAGGACCTCGACAAG	ATG	9695	81
4	-5.296	1.356	8	-6.517	GGACCGCGCGCCACAGAGTGAG	ATG	9719	57
5	-2.633	2.631	10	-3.328	ACAGAGTGAGATGGAGAAACTC	ATG	9731	45



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	-4.047	1.954	9	-4.822	ACCTGAATAGATAGGTGCAGCA	ATG	9500	276
2	-1.748	3.055	7	-3.271	GGGTCAGCCGATCAAGGAGCGC	GTG	9572	204
3	-2.071	2.901	16	-3.867	TTCGGAGGAGGACCTCGACAAG	ATG	9695	81
4	-5.296	1.356	8	-6.517	GGACCGCGCGCCACAGAGTGAG	ATG	9719	57
5	-2.633	2.631	10	-3.328	ACAGAGTGAGATGGAGAAACTC	ATG	9731	45

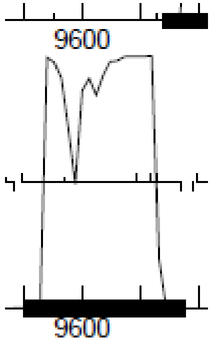
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



Gene: Yucky_16 Start: 9500, Stop: 9775, Start Num: 16
Candidate Starts for Yucky_16:
(Start: 16 @9500 has 36 MA's), (27, 9572), (38, 9695), (41, 9719), (44, 9731),

- 9500 has 36 Manual annotation. The proposed start aligns well with other pham members, as it is the most annotated start and called 98% of the time when it is present.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- 9500 is the earliest start available, maximizing coding potential. Later starts would cut of coding potential.

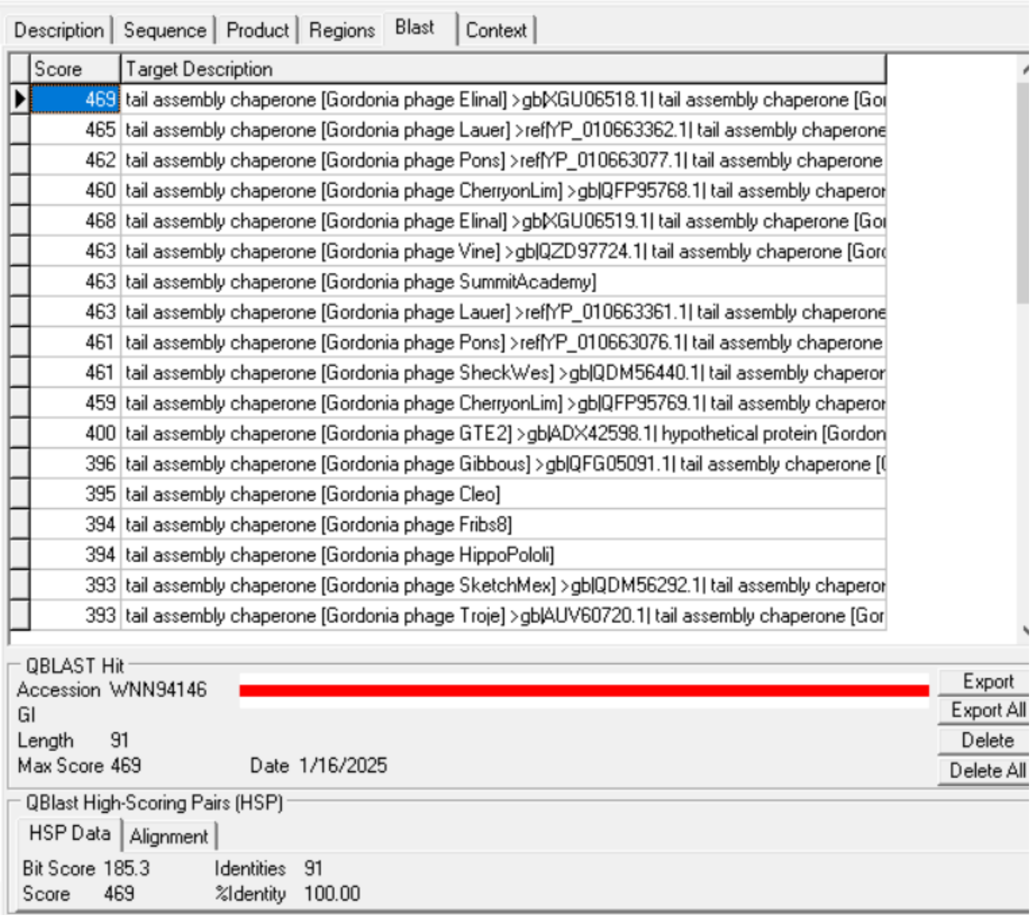
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- This feature has a 96 bp gap with the previous feature, which ends at 9403. However, no earlier start exists, leaving us with a gap.

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is 9500. This is the first start available. It agrees with the automated start site. Even though it does not have the best RBS values, it maximizes coding potential inclusion, as well as has many 1:1 BLAST hits with highly similar features. This start is called 98% of the time when it is present.

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
469	tail assembly chaperone [Gordonia phage Elinal] >gb|XGU06518.1| tail assembly chaperone [Goi
465	tail assembly chaperone [Gordonia phage Lauer] >ref|YP_010663362.1| tail assembly chaperone
462	tail assembly chaperone [Gordonia phage Pons] >ref|YP_010663077.1| tail assembly chaperone
460	tail assembly chaperone [Gordonia phage CherryonLim] >gb|QFP95768.1| tail assembly chaperor
468	tail assembly chaperone [Gordonia phage Elinal] >gb|XGU06519.1| tail assembly chaperone [Goi
463	tail assembly chaperone [Gordonia phage Vine] >gb|QZD97724.1| tail assembly chaperone [Gorc
463	tail assembly chaperone [Gordonia phage SummitAcademy]
463	tail assembly chaperone [Gordonia phage Lauer] >ref|YP_010663361.1| tail assembly chaperone
461	tail assembly chaperone [Gordonia phage Pons] >ref|YP_010663076.1| tail assembly chaperone
461	tail assembly chaperone [Gordonia phage SheckWes] >gb|QDM56440.1| tail assembly chaperor
459	tail assembly chaperone [Gordonia phage CherryonLim] >gb|QFP95769.1| tail assembly chaperor
400	tail assembly chaperone [Gordonia phage GTE2] >gb|ADX42598.1| hypothetical protein [Gordon
396	tail assembly chaperone [Gordonia phage Gibbous] >gb|QFG05091.1| tail assembly chaperone [(
395	tail assembly chaperone [Gordonia phage Cleo]
394	tail assembly chaperone [Gordonia phage Fribs8]
394	tail assembly chaperone [Gordonia phage HippoPololi]
393	tail assembly chaperone [Gordonia phage SketchMex] >gb|QDM56292.1| tail assembly chaperor
393	tail assembly chaperone [Gordonia phage Troje] >gb|AUV60720.1| tail assembly chaperone [Gor

QBLAST Hit

Accession WNN94146

GI

Length 91

Max Score 469

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 185.3

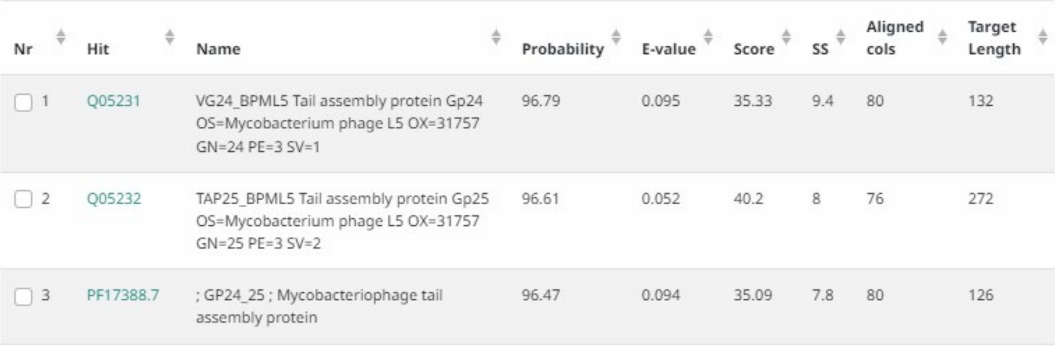
Score 469

Identities 91

%Identity 100.00

- Highly similar genes all call the function of a tail assembly chaperone.

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	Q05231	VG24_BPML5 Tail assembly protein Gp24 OS=Mycobacterium phage L5 OX=31757 GN=24 PE=3 SV=1	96.79	0.095	35.33	9.4	80	132
□ 2	Q05232	TAP25_BPML5 Tail assembly protein Gp25 OS=Mycobacterium phage L5 OX=31757 GN=25 PE=3 SV=2	96.61	0.052	40.2	8	76	272
□ 3	PF17388.7	; GP24_25 ; Mycobacteriophage tail assembly protein	96.47	0.094	35.09	7.8	80	126

- 3 hits over 90% probability indicate similarity to tail assembly protein with the top two indicating similarity to GP24 and GP25 of Mycobacterium phage L5.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Other features in the same pham in closely related phages such as elinal, potpie and SheckWes are annotated as tail assembly chaperones

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- This has a putative function of tail assembly chaperone, so the Deep TMHMM evidence is not applicable.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Tail Assembly Chaperone. Both BLAST evidence as well as HHPRED and Phamerator support Tail Assembly Chaperone as official function.
- Recoding site from Baranov, et al. 2006 GGGGGAA found in L5 phage found beginning at 9766. The shared nucleotide is G found at 9769.

Feature 15 – Stop 10248

Glimmer/GeneMark

What feature number is this? 15

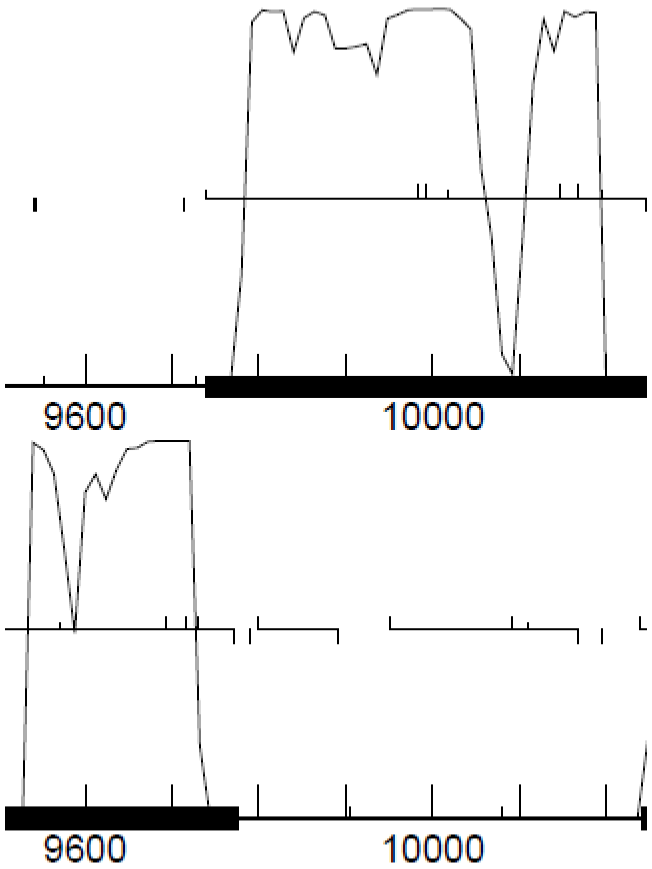
What is the stop site? 10248

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? N/A

What is the autoannotated start? N/A

Gap: __96______ or overlap: _______ (with gene in front of it) for the autoannotated start

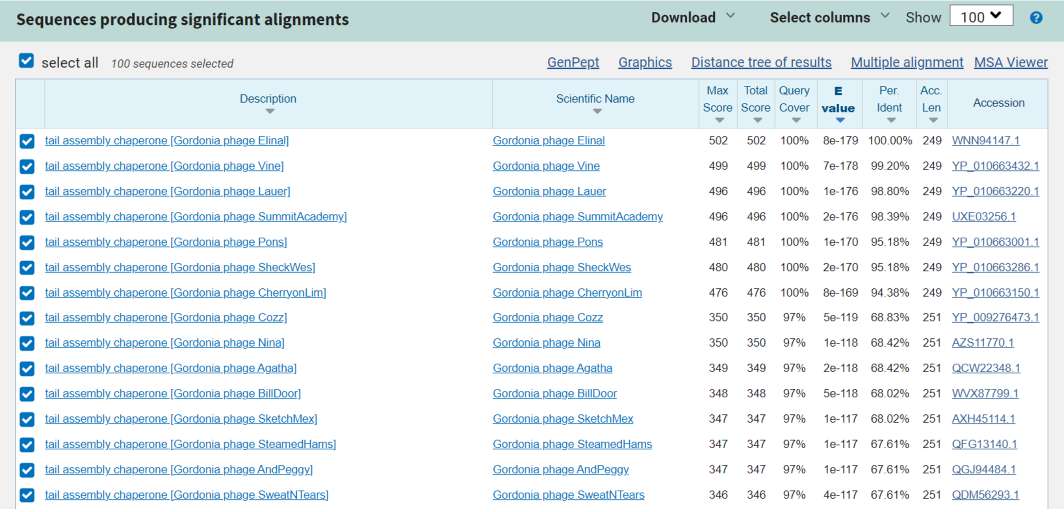
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Only reading frame with coding potential. Overlaps with coding potential in frame 2. Mostly strong with one dip near 10,100

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- Many hits with e value close to zero



Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	tail assembly chaperone [Gordonia phage Elinal]	Gordonia phage Elinal	502	502	100%	8e-179	100.00%	249	WNN94147.1
☑	tail assembly chaperone [Gordonia phage Vine]	Gordonia phage Vine	499	499	100%	7e-178	99.20%	249	YP_010663432.1
☑	tail assembly chaperone [Gordonia phage Lauer]	Gordonia phage Lauer	496	496	100%	1e-176	98.80%	249	YP_010663220.1
☑	tail assembly chaperone [Gordonia phage SummitAcademy]	Gordonia phage SummitAcademy	496	496	100%	2e-176	98.39%	249	UXE03256.1
☑	tail assembly chaperone [Gordonia phage Pons]	Gordonia phage Pons	481	481	100%	1e-170	95.18%	249	YP_010663001.1
☑	tail assembly chaperone [Gordonia phage SheckWes]	Gordonia phage SheckWes	480	480	100%	2e-170	95.18%	249	YP_010663286.1
☑	tail assembly chaperone [Gordonia phage CherryonLim]	Gordonia phage CherryonLim	476	476	100%	8e-169	94.38%	249	YP_010663150.1
☑	tail assembly chaperone [Gordonia phage Cozz]	Gordonia phage Cozz	350	350	97%	5e-119	68.83%	251	YP_009276473.1
☑	tail assembly chaperone [Gordonia phage Nina]	Gordonia phage Nina	350	350	97%	1e-118	68.42%	251	AZS11770.1
☑	tail assembly chaperone [Gordonia phage Agatha]	Gordonia phage Agatha	349	349	97%	2e-118	68.42%	251	QCW22348.1
☑	tail assembly chaperone [Gordonia phage BillDoor]	Gordonia phage BillDoor	348	348	97%	5e-118	68.02%	251	WVX87799.1
☑	tail assembly chaperone [Gordonia phage SketchMex]	Gordonia phage SketchMex	347	347	97%	1e-117	68.02%	251	AXH45114.1
☑	tail assembly chaperone [Gordonia phage SteamedHams]	Gordonia phage SteamedHams	347	347	97%	1e-117	67.61%	251	QFG13140.1
☑	tail assembly chaperone [Gordonia phage AndPeggy]	Gordonia phage AndPeggy	347	347	97%	1e-117	67.61%	251	QGJ94484.1
☑	tail assembly chaperone [Gordonia phage SweatNTears]	Gordonia phage SweatNTears	346	346	97%	4e-117	67.61%	251	QDM56293.1

Answer: Is it a gene? Give evidence why you think this is a gene or not.

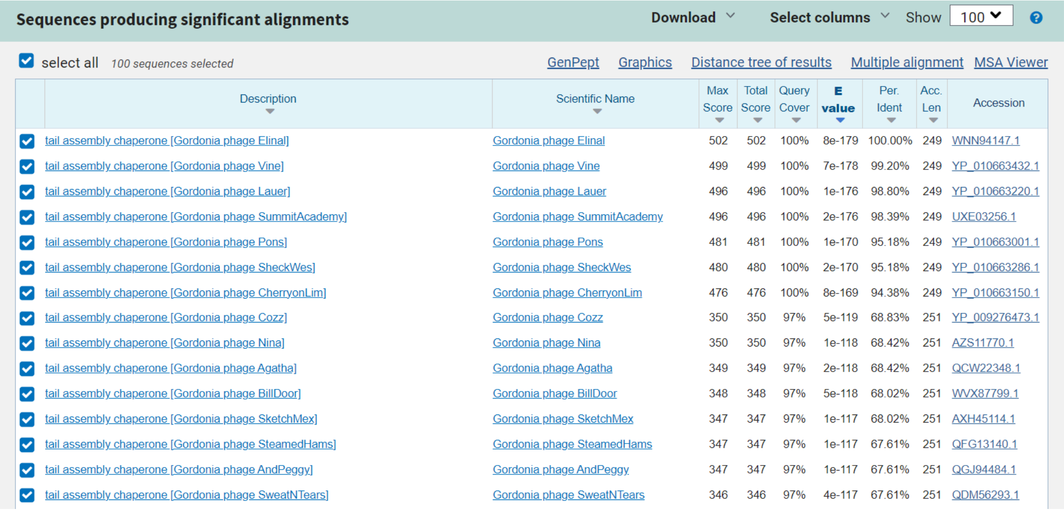
- Yes, it has coding potential, Is called to be a gene by both Glimmer and Genemark, and has many blast hits

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- This is the tail assembly chaperone. See evidence in feature 14. Start at 9500, as directed in the genomics guide

BLAST function evidence. What assigned functions do other highly similar genes have?

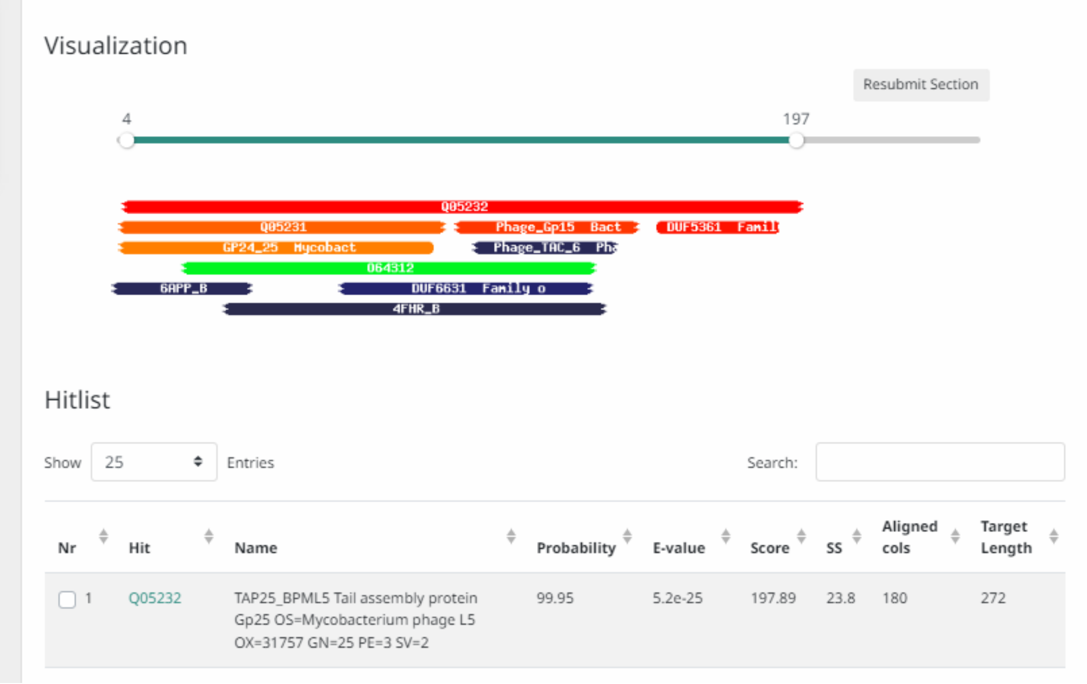
- Other similar genes call it a tail assembly chaperone



100 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	tail assembly chaperone [Gordonia phage Elinal]	Gordonia phage Elinal	502	502	100%	8e-179	100.00%	249	WNN94147.1
☑	tail assembly chaperone [Gordonia phage Vine]	Gordonia phage Vine	499	499	100%	7e-178	99.20%	249	YP_010663432.1
☑	tail assembly chaperone [Gordonia phage Lauer]	Gordonia phage Lauer	496	496	100%	1e-176	98.80%	249	YP_010663220.1
☑	tail assembly chaperone [Gordonia phage SummitAcademy]	Gordonia phage SummitAcademy	496	496	100%	2e-176	98.39%	249	UXE03256.1
☑	tail assembly chaperone [Gordonia phage Pons]	Gordonia phage Pons	481	481	100%	1e-170	95.18%	249	YP_010663001.1
☑	tail assembly chaperone [Gordonia phage SheckWes]	Gordonia phage SheckWes	480	480	100%	2e-170	95.18%	249	YP_010663286.1
☑	tail assembly chaperone [Gordonia phage CherryonLim]	Gordonia phage CherryonLim	476	476	100%	8e-169	94.38%	249	YP_010663150.1
☑	tail assembly chaperone [Gordonia phage Cozz]	Gordonia phage Cozz	350	350	97%	5e-119	68.83%	251	YP_009276473.1
☑	tail assembly chaperone [Gordonia phage Nina]	Gordonia phage Nina	350	350	97%	1e-118	68.42%	251	AZS11770.1
☑	tail assembly chaperone [Gordonia phage Agatha]	Gordonia phage Agatha	349	349	97%	2e-118	68.42%	251	QCW22348.1
☑	tail assembly chaperone [Gordonia phage BillDoor]	Gordonia phage BillDoor	348	348	97%	5e-118	68.02%	251	WVX87799.1
☑	tail assembly chaperone [Gordonia phage SketchMex]	Gordonia phage SketchMex	347	347	97%	1e-117	68.02%	251	AXH45114.1
☑	tail assembly chaperone [Gordonia phage SteamedHams]	Gordonia phage SteamedHams	347	347	97%	1e-117	67.61%	251	QFG13140.1
☑	tail assembly chaperone [Gordonia phage AndPeggy]	Gordonia phage AndPeggy	347	347	97%	1e-117	67.61%	251	QGJ94484.1
☑	tail assembly chaperone [Gordonia phage SweatNTears]	Gordonia phage SweatNTears	346	346	97%	4e-117	67.61%	251	QDM56293.1

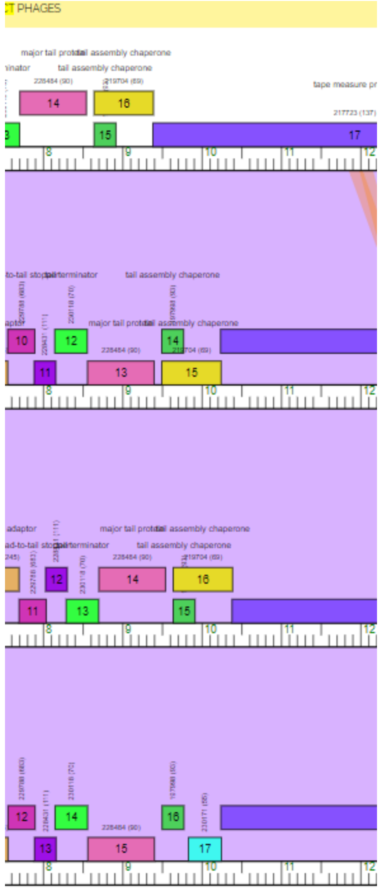
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	Q05232	TAP25_BPML5 Tail assembly protein Gp25 OS=Mycobacterium phage L5 OX=31757 GN=25 PE=3 SV=2	99.95	5.2e-25	197.89	23.8	180	272

- Topmost hit corresponds with Tail Assembly protein in phage L5.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- Closely related genes have a gene in a different pham called a tail assembly chaperone (bottom row for Yucky, features 16/17 on phamerator, but I fully expect the phams to change to be congruent with elinal, vine and potpie in the top three genomes.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- N/A since function will be called
  a Tail Assembly Chaperone

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Tail Assembly Chaperone. Many BLAST hits called a tail assembly chaperone and slippery sequence found in feature 14.

Feature 16 – Stop 15340

Glimmer/GeneMark

What feature number is this? 16

What is the stop site? 15340

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Called by both Glimmer and GeneMark and they agree on start site.

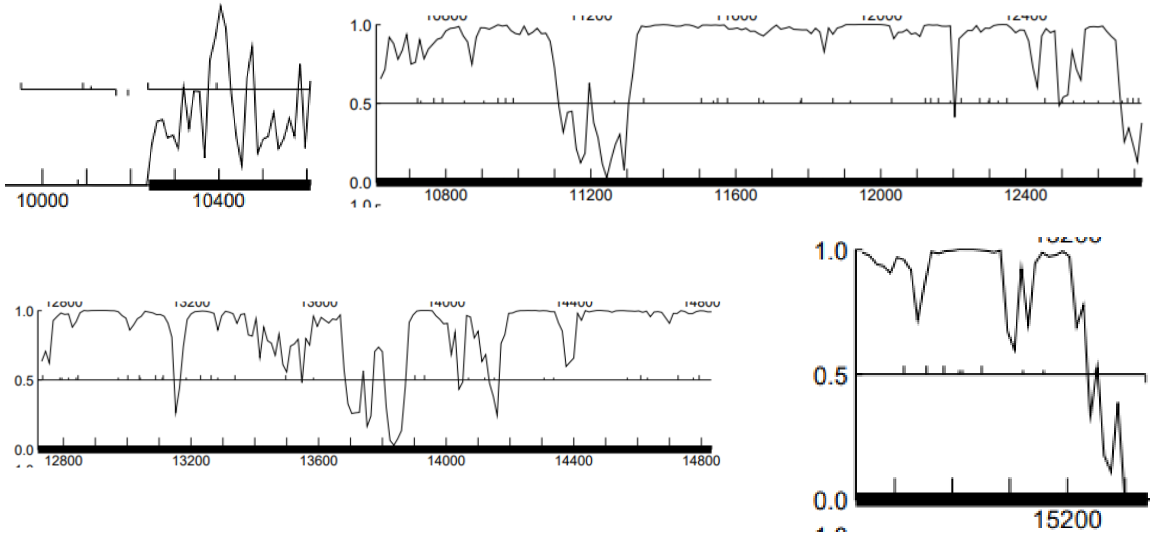
What is the autoannotated start?

10241

Gap: _________ or overlap: 8 (with gene in front of it) for the autoannotated start

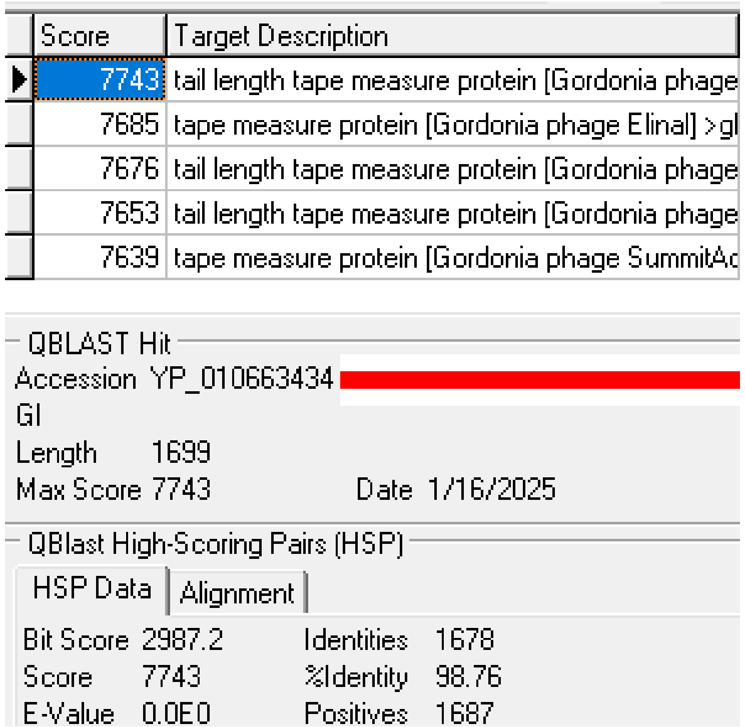
- Glimmer and GeneMark agree
- Overlap of 8

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak?   How do you know?



- This gene is particularly long, 5100 nucleotides in length. Despite this, there are consistent strong peaks of coding potential throughout the entire nucleotide sequence. Reading frame 2 contains the most coding potential for this feature. There is also overlapping coding potential on frames 3,4,6, and one strong peak on frame 5.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
7743	tail length tape measure protein [Gordonia phage
7685	tape measure protein [Gordonia phage Elinal] >gl
7676	tail length tape measure protein [Gordonia phage
7653	tail length tape measure protein [Gordonia phage
7639	tape measure protein [Gordonia phage SummitAc

QBLAST Hit

Accession YP_010663434

GI

Length 1699

Max Score 7743 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

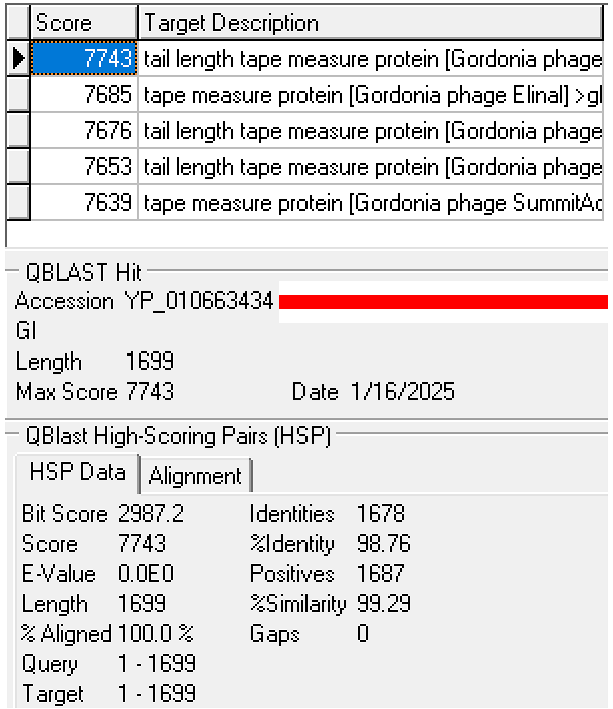
Bit Score	2987.2	Identities	1678
Score	7743	%Identity	98.76
E-Value	0.0E0	Positives	1687

- There are at least 25 highly similar genes as revealed by BLAST, all containing an E-value of close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, it is a gene. I believe this because both Glimmer and GeneMark called it a gene, there is coding potential throughout the entire sequence of nucleotides, and there are least 25 BLAST hits for similar genes with an E-value close to 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
7743	tail length tape measure protein [Gordonia phage
7685	tape measure protein [Gordonia phage Elinal] >gl
7676	tail length tape measure protein [Gordonia phage
7653	tail length tape measure protein [Gordonia phage
7639	tape measure protein [Gordonia phage SummitAc

QBLAST Hit

Accession YP_010663434

GI

Length 1699

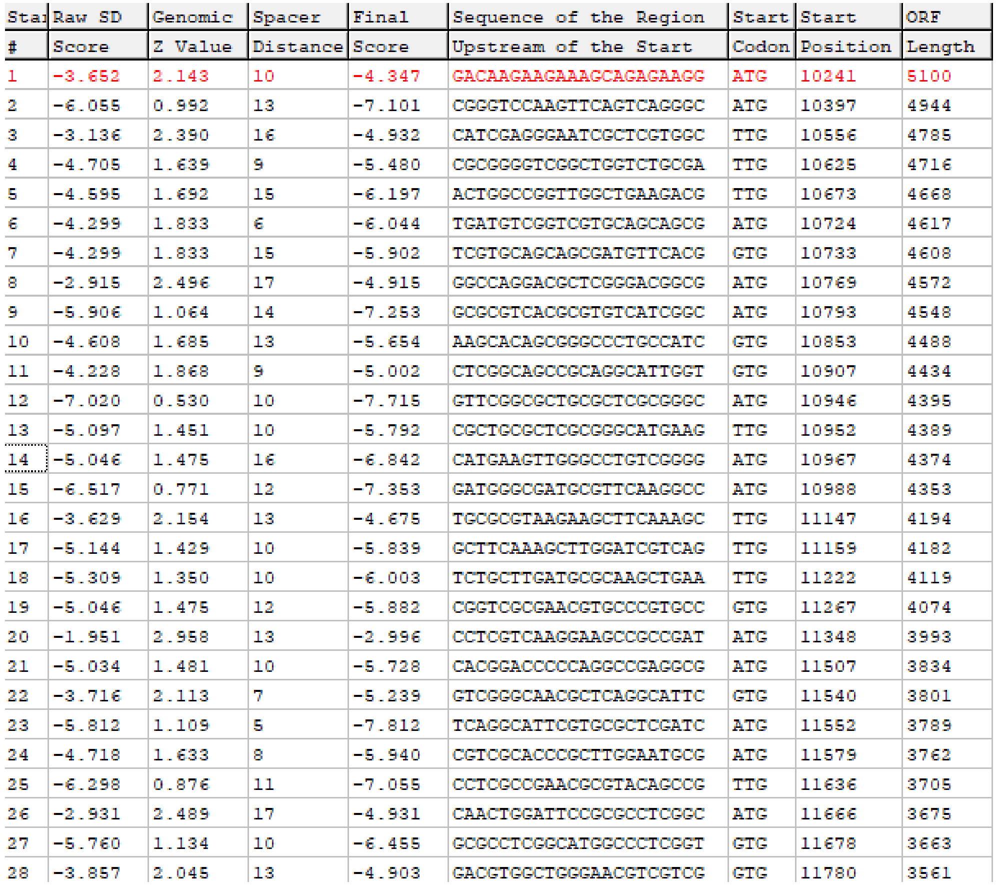
Max Score 7743 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	2987.2	Identities	1678
Score	7743	%Identity	98.76
E-Value	0.0E0	Positives	1687
Length	1699	%Similarity	99.29
% Aligned	100.0 %	Gaps	0
Query	1 - 1699		
Target	1 - 1699		

- There are at least 25 1:1 alignments shown by BLAST. No alternative starts are known at this time since Glimmer and GeneMark agree on the start site.

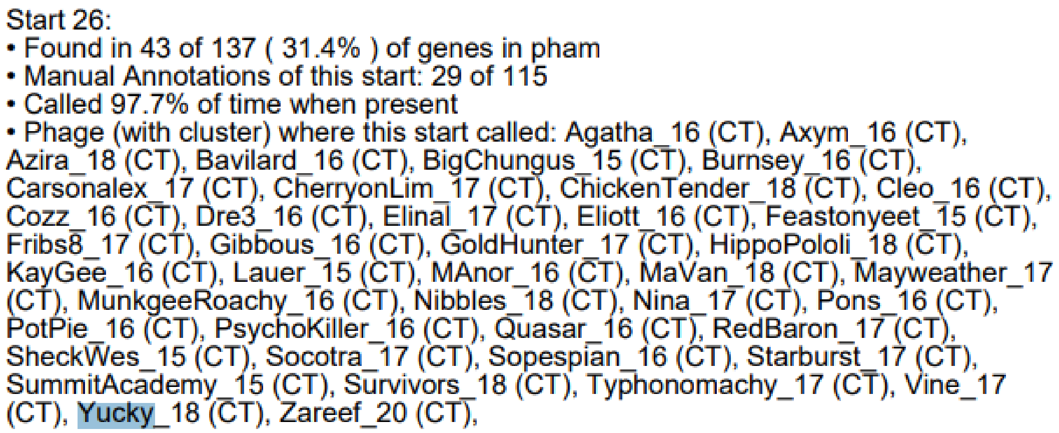
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-3.652	2.143	10	-4.347	GACAAGAAGAAAGCAGAGAAGG	ATG	10241	5100
2	-6.055	0.992	13	-7.101	CGGGTCCAAGTTCAGTCAGGGC	ATG	10397	4944
3	-3.136	2.390	16	-4.932	CATCGAGGGAATCGCTCGTGGC	TTG	10556	4785
4	-4.705	1.639	9	-5.480	CGCGGGGTCGGCTGGTCTGCGA	TTG	10625	4716
5	-4.595	1.692	15	-6.197	ACTGGCCGGTTGGCTGAAGACG	TTG	10673	4668
6	-4.299	1.833	6	-6.044	TGATGTCGGTCGTGCAGCAGCG	ATG	10724	4617
7	-4.299	1.833	15	-5.902	TCGTGCAGCAGCGATGTTCACG	GTG	10733	4608
8	-2.915	2.496	17	-4.915	GGCCAGGACGCTCGGGACGGCG	ATG	10769	4572
9	-5.906	1.064	14	-7.253	GCGCGTCACGCGTGTCATCGGC	ATG	10793	4548
10	-4.608	1.685	13	-5.654	AAGCACAGCGGGCCCTGCCATC	GTG	10853	4488
11	-4.228	1.868	9	-5.002	CTCGGCAGCCGCAGGCATTGGT	GTG	10907	4434
12	-7.020	0.530	10	-7.715	GTTCGGCGCTGCGCTCGCGGGC	ATG	10946	4395
13	-5.097	1.451	10	-5.792	CGCTGCGCTCGCGGGCATGAAG	TTG	10952	4389
14	-5.046	1.475	16	-6.842	CATGAAGTTGGGCCTGTCGGGG	ATG	10967	4374
15	-6.517	0.771	12	-7.353	GATGGGCGATGCGTTCAAGGCC	ATG	10988	4353
16	-3.629	2.154	13	-4.675	TGCGCGTAAGAAGCTTCAAAGC	TTG	11147	4194
17	-5.144	1.429	10	-5.839	GCTTCAAAGCTTGGATCGTCAG	TTG	11159	4182
18	-5.309	1.350	10	-6.003	TCTGCTTGATGCGCAAGCTGAA	TTG	11222	4119
19	-5.046	1.475	12	-5.882	CGGTCGCGAACGTGCCCGTGCC	GTG	11267	4074
20	-1.951	2.958	13	-2.996	CCTCGTCAAGGAAGCCGCCGAT	ATG	11348	3993
21	-5.034	1.481	10	-5.728	CACGGACCCCCAGGCCGAGGCG	ATG	11507	3834
22	-3.716	2.113	7	-5.239	GTCGGGCAACGCTCAGGCATTC	GTG	11540	3801
23	-5.812	1.109	5	-7.812	TCAGGCATTCGTGCGCTCGATC	ATG	11552	3789
24	-4.718	1.633	8	-5.940	CGTCGCACCCGCTTGGAATGCG	ATG	11579	3762
25	-6.298	0.876	11	-7.055	CCTCGCCGAACGCGTACAGCCG	TTG	11636	3705
26	-2.931	2.489	17	-4.931	CAACTGGATTCCGCGCCTCGGC	ATG	11666	3675
27	-5.760	1.134	10	-6.455	GCGCCTCGGCATGGCCCTCGGT	GTG	11678	3663
28	-3.857	2.045	13	-4.903	GACGTGGCTGGGAACGTCGTCG	GTG	11780	3561

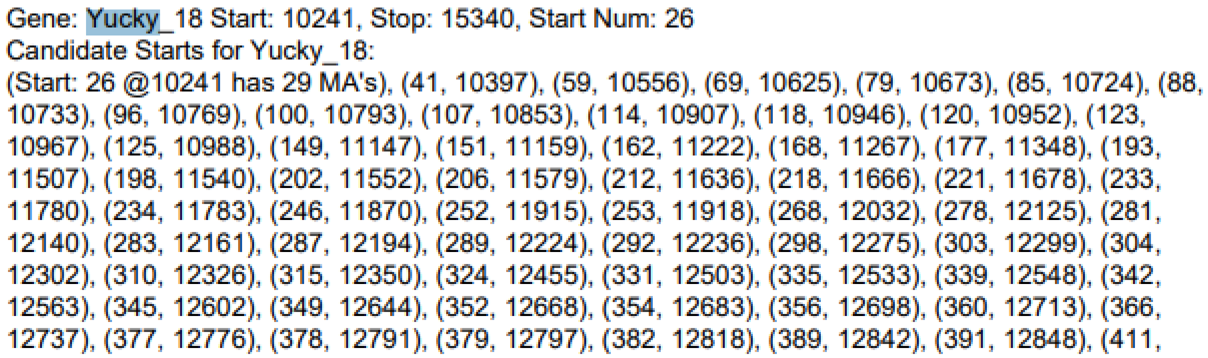
- The Z-value is 2.143.
- The final score is -4.347
- I see no other RBS values indicating a start site better than the autoannotated one.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



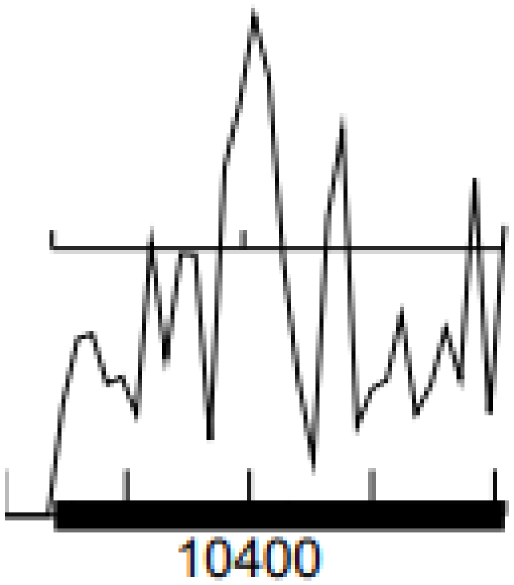
Start 26:
• Found in 43 of 137 (31.4%) of genes in pham
• Manual Annotations of this start: 29 of 115
• Called 97.7% of time when present
• Phage (with cluster) where this start called: Agatha_16 (CT), Axym_16 (CT), Azira_18 (CT), Bavilard_16 (CT), BigChungus_15 (CT), Burnsey_16 (CT), Carsonalex_17 (CT), CherryonLim_17 (CT), ChickenTender_18 (CT), Cleo_16 (CT), Cozz_16 (CT), Dre3_16 (CT), Elinal_17 (CT), Eliott_16 (CT), Feastonyeet_15 (CT), Fribs8_17 (CT), Gibbous_16 (CT), GoldHunter_17 (CT), HippoPololi_18 (CT), KayGee_16 (CT), Lauer_15 (CT), MAnor_16 (CT), MaVan_18 (CT), Mayweather_17 (CT), MunkgeeRoachy_16 (CT), Nibbles_18 (CT), Nina_17 (CT), Pons_16 (CT), PotPie_16 (CT), PsychoKiller_16 (CT), Quasar_16 (CT), RedBaron_17 (CT), SheckWes_15 (CT), Socotra_17 (CT), Sopespian_16 (CT), Starburst_17 (CT), SummitAcademy_15 (CT), Survivors_18 (CT), Typhonomachy_17 (CT), Vine_17 (CT), Yucky_18 (CT), Zareef_20 (CT),

- The proposed start has 29 MAs. It is the only proposed start site with any MAs. It is called 97.7% of the time when present.



Gene: Yucky_18 Start: 10241, Stop: 15340, Start Num: 26
Candidate Starts for Yucky_18:
(Start: 26 @10241 has 29 MA's), (41, 10397), (59, 10556), (69, 10625), (79, 10673), (85, 10724), (88, 10733), (96, 10769), (100, 10793), (107, 10853), (114, 10907), (118, 10946), (120, 10952), (123, 10967), (125, 10988), (149, 11147), (151, 11159), (162, 11222), (168, 11267), (177, 11348), (193, 11507), (198, 11540), (202, 11552), (206, 11579), (212, 11636), (218, 11666), (221, 11678), (233, 11780), (234, 11783), (246, 11870), (252, 11915), (253, 11918), (268, 12032), (278, 12125), (281, 12140), (283, 12161), (287, 12194), (289, 12224), (292, 12236), (298, 12275), (303, 12299), (304, 12302), (310, 12326), (315, 12350), (324, 12455), (331, 12503), (335, 12533), (339, 12548), (342, 12563), (345, 12602), (349, 12644), (352, 12668), (354, 12683), (356, 12698), (360, 12713), (366, 12737), (377, 12776), (378, 12791), (379, 12797), (382, 12818), (389, 12842), (391, 12848), (411,

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- The beginning of a peak of coding potential is cut off, though I would estimate it cuts off less than 10 nucleotides of coding potential.

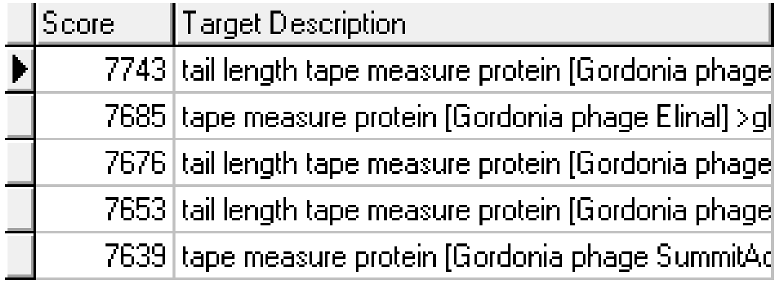
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- There is an overlap of 8 with the previous gene. As this is not a large overlap it is still an acceptable start site.
- 10248-10241=7+1=8

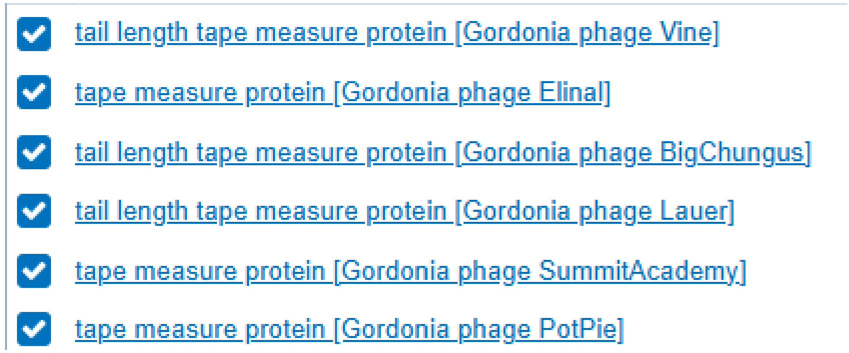
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is the same as the automates start site of 10241. There are at least 25 BLAST 1:1 alignments, it has a good Z-value (2.143) and Final score (-4.347), it is the only start site to ever be manually annotated and it is called very frequently when present, it cuts off minimal coding potential, and it has an acceptable overlap value.

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
7743	tail length tape measure protein [Gordonia phage
7685	tape measure protein [Gordonia phage Elinal] >gl
7676	tail length tape measure protein [Gordonia phage
7653	tail length tape measure protein [Gordonia phage
7639	tape measure protein [Gordonia phage SummitAc



☑ tail length tape measure protein [Gordonia phage Vine]

☑ tape measure protein [Gordonia phage Elinal]

☑ tail length tape measure protein [Gordonia phage BigChungus]

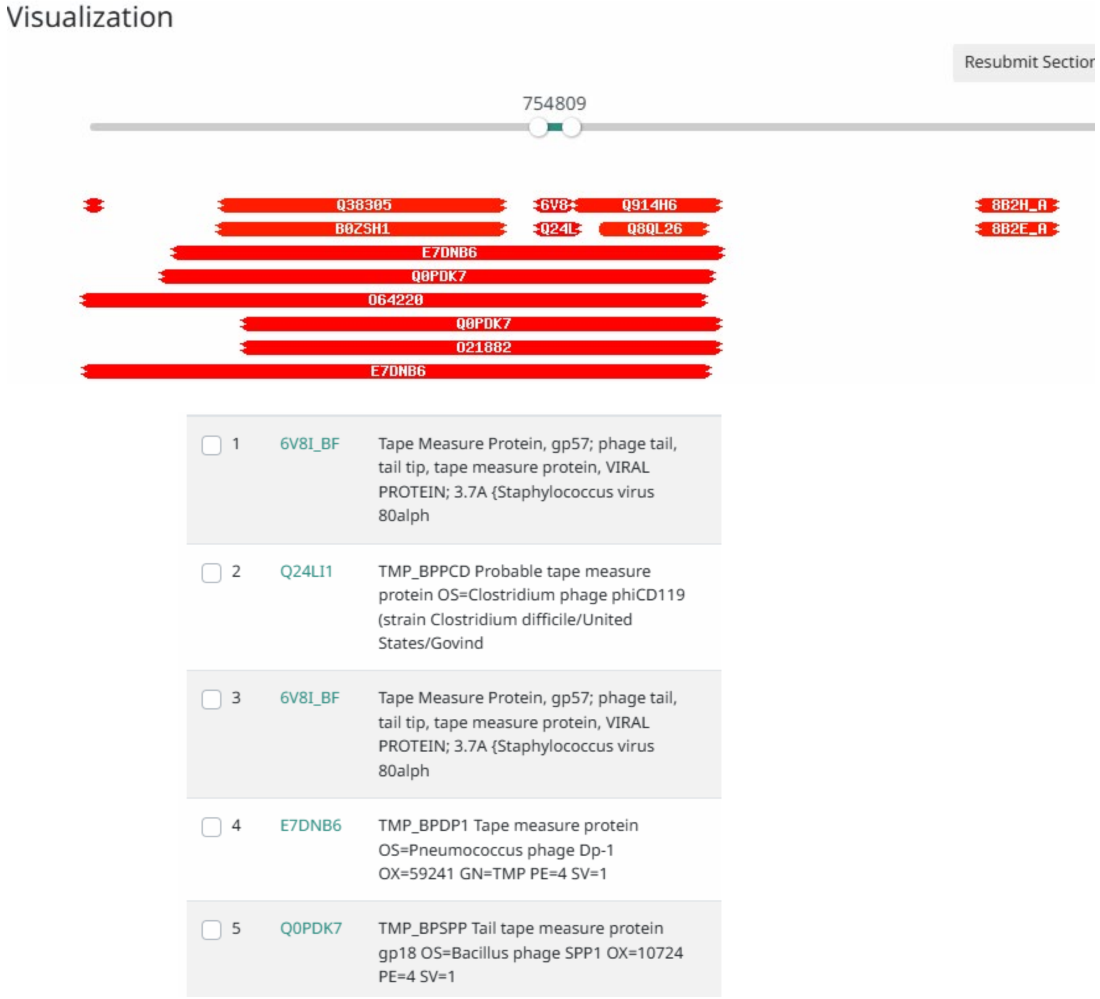
☑ tail length tape measure protein [Gordonia phage Lauer]

☑ tape measure protein [Gordonia phage SummitAcademy]

☑ tape measure protein [Gordonia phage PotPie]

- DNA master BLAST showed 17 hits as a tape measure protein and 8 hits as a tail length tape measure protein.
- BLASTing on NCBI yielded similar results, showing results for both tape measure protein and tail length tape measure protein.

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



☐ 1 6V8I_BF Tape Measure Protein, gp57; phage tail, tail tip, tape measure protein, VIRAL PROTEIN; 3.7A {Staphylococcus virus 80alph

☐ 2 Q24LI1 TMP_BPPCD Probable tape measure protein OS=Clostridium phage phiCD119 (strain Clostridium difficile/United States/Govind

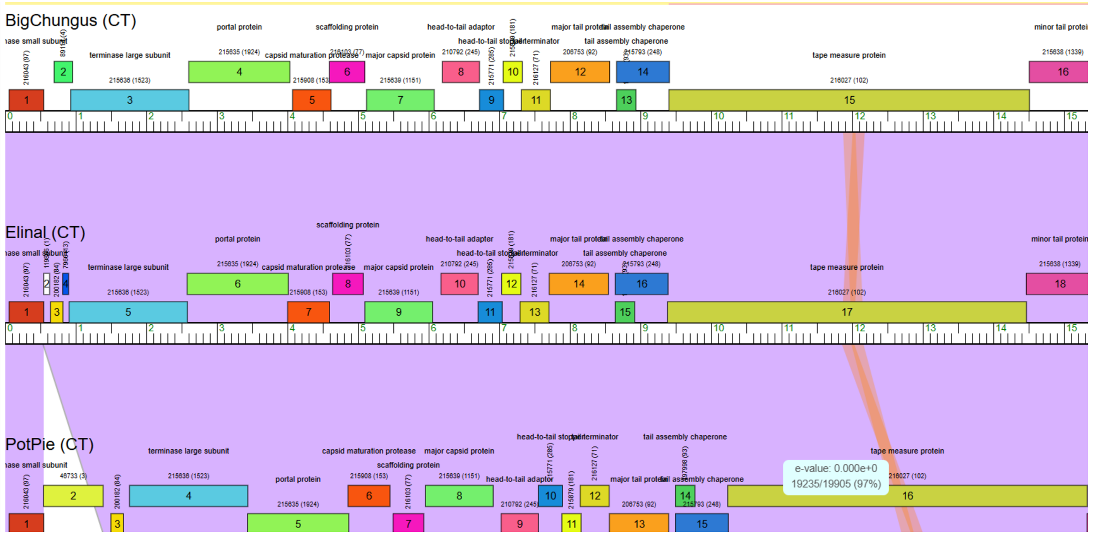
☐ 3 6V8I_BF Tape Measure Protein, gp57; phage tail, tail tip, tape measure protein, VIRAL PROTEIN; 3.7A {Staphylococcus virus 80alph

☐ 4 E7DNB6 TMP_BPDP1 Tape measure protein OS=Pneumococcus phage Dp-1 OX=59241 GN=TMP PE=4 SV=1

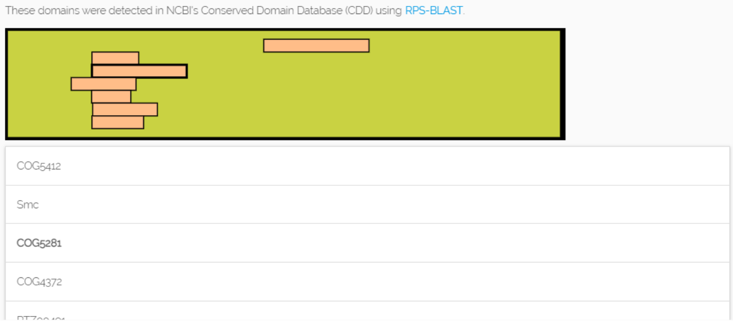
☐ 5 Q0PDK7 TMP_BPSPP Tail tape measure protein gp18 OS=Bacillus phage SPP1 OX=10724 PE=4 SV=1

- Hhpred evidence points towards this being a tape measure protein. There are many strong, some homologous hits showing this gene as a tape measure protein.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- PotPie, BigChungus, and Elinal all have this gene and in all 3 it is a tape measure protein.
- PotPie has 6 conserved domains.
- Elinal has 7 conserved domains.
- BigChungus has no conserved domains.


These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

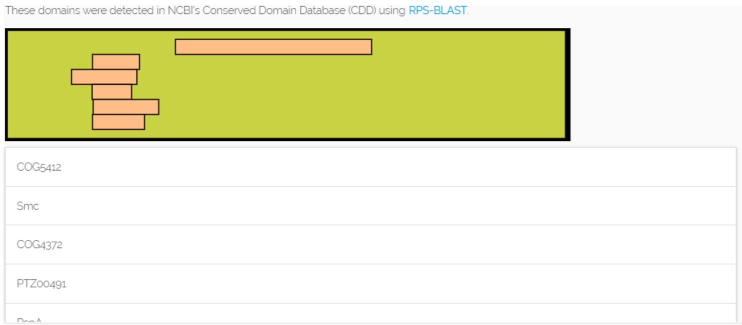
COG5412

Smc

COG5281

COG4372

PTZ00491


These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

COG5412

Smc

COG4372

PTZ00491

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- I would like to call this a tape measure protein.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I officially call this a tape measure protein. Both DNAmaster and NCBI BLAST showed many strong hits of similar genes with this gene as a tape measure protein. HHpred also showed many strong hits of this gene being a tape measure protein. Lastly, Phamerator showed 3 very similar phages with this gene, and it was called a tape measure protein in all of them.

Feature 17 Stop 16293

Glimmer/GeneMark

What feature number is this? 17

What is the stop site? 16293

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Glimmer and GeneMark both called it.

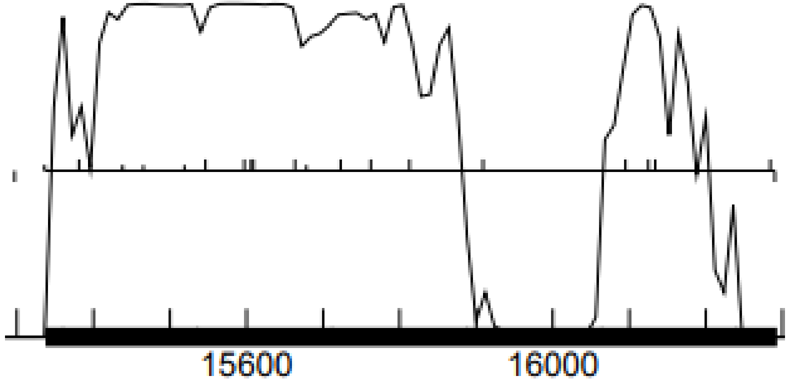
What is the autoannotated start?

15337

Gap: ________ or overlap 4 (with gene in front of it) for the autoannotated start

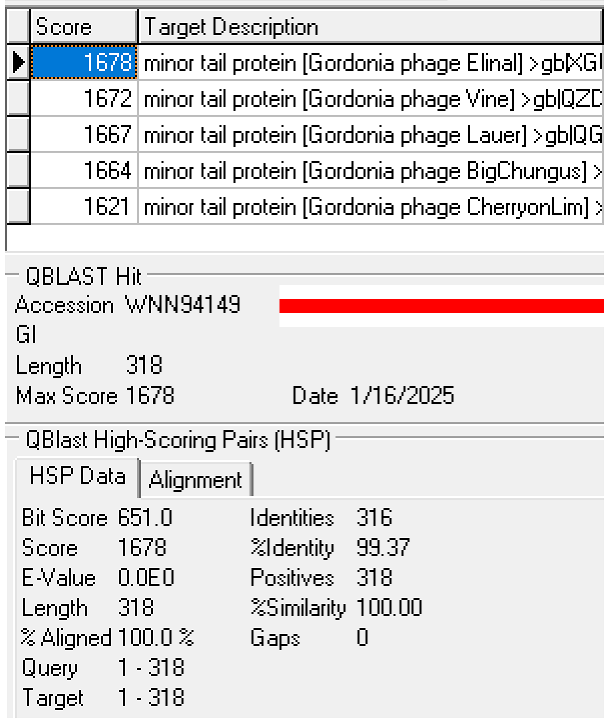
- Called by both
- Overlap of 4

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- There is a consistent, strong peak of coding potential on reading frame 2 that tapers off before returning to a strong peak. It then completely drops off before returning for one more strong peak. There is a singular weak peak of coding potential on the 6th reading frame.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
1678	minor tail protein [Gordonia phage Elinal] >gb|XGI
1672	minor tail protein [Gordonia phage Vine] >gb|QZD
1667	minor tail protein [Gordonia phage Lauer] >gb|QG
1664	minor tail protein [Gordonia phage BigChungus] >
1621	minor tail protein [Gordonia phage CherryonLim] >

QBLAST Hit

Accession WNN94149

GI

Length 318

Max Score 1678 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 651.0
Score 1678
E-Value 0.0E0
Length 318
% Aligned 100.0 %
Query 1 - 318
Target 1 - 318

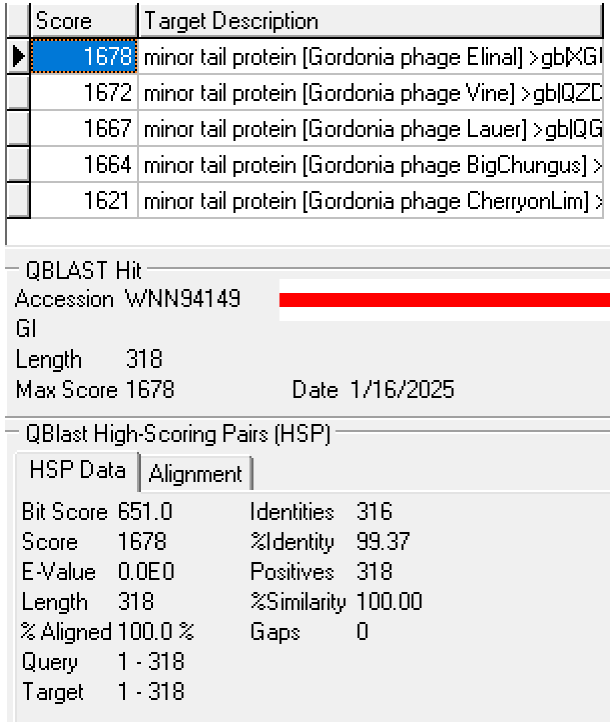
Identities 316
%Identity 99.37
Positives 318
%Similarity 100.00
Gaps 0

- There are at least 25 highly similar genes with an E-value close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, it is a gene. I believe this because both Glimmer and GeneMark called it a gene, there are many strong peaks of coding potential throughout the sequence of nuceltodies, and there are least 25 BLAST hits for similar genes with an E-value close to 0

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
1678	minor tail protein [Gordonia phage Elinal] >gb|XGI
1672	minor tail protein [Gordonia phage Vine] >gb|QZD
1667	minor tail protein [Gordonia phage Lauer] >gb|QG
1664	minor tail protein [Gordonia phage BigChungus] >
1621	minor tail protein [Gordonia phage CherryonLim] >

QBLAST Hit

Accession WNN94149

GI

Length 318

Max Score 1678

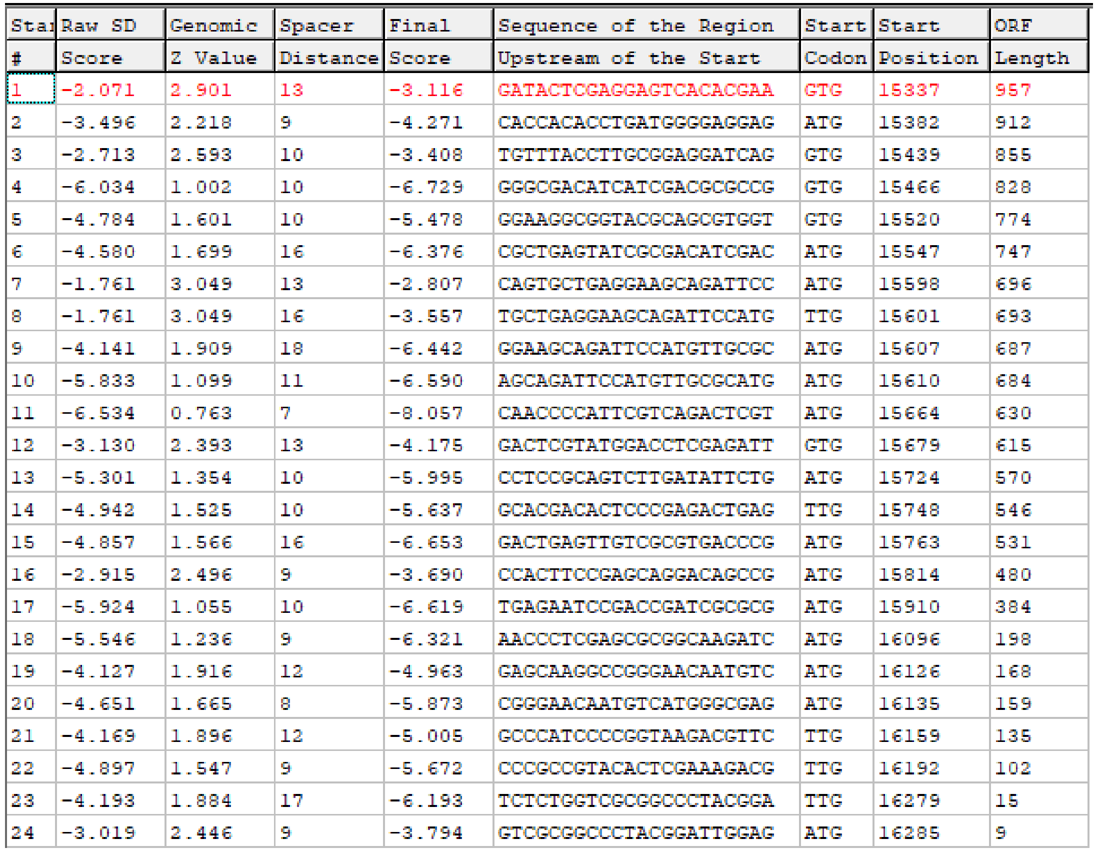
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	651.0	Identities	316
Score	1678	%Identity	99.37
E-Value	0.0E0	Positives	318
Length	318	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 318		
Target	1 - 318		

- There are 24 1:1 alignments and 1 16:1 alignment. No alternative starts are known at this time since Glimmer and GeneMark agree on the start site.

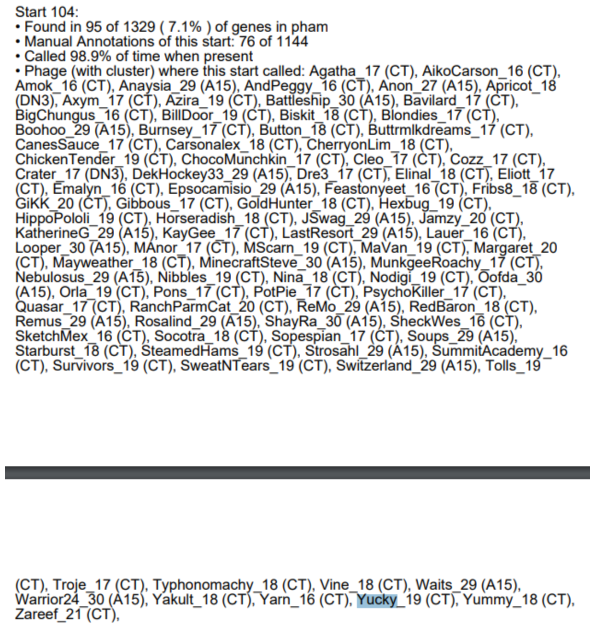
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-2.071	2.901	13	-3.116	GATACTCGAGGAGTCACACGAA	GTG	15337	957
2	-3.496	2.218	9	-4.271	CACCACACCTGATGGGGAGGAG	ATG	15382	912
3	-2.713	2.593	10	-3.408	TGTTTACCTTGCGGAGGATCAG	GTG	15439	855
4	-6.034	1.002	10	-6.729	GGGCGACATCATCGACGCGCCG	GTG	15466	828
5	-4.784	1.601	10	-5.478	GGAAGGCGGTACGCAGCGTGGT	GTG	15520	774
6	-4.580	1.699	16	-6.376	CGCTGAGTATCGCGACATCGAC	ATG	15547	747
7	-1.761	3.049	13	-2.807	CAGTGCTGAGGAAGCAGATTCC	ATG	15598	696
8	-1.761	3.049	16	-3.557	TGCTGAGGAAGCAGATTCCATG	TTG	15601	693
9	-4.141	1.909	18	-6.442	GGAAGCAGATTCCATGTTGCGC	ATG	15607	687
10	-5.833	1.099	11	-6.590	AGCAGATTCCATGTTGCGCATG	ATG	15610	684
11	-6.534	0.763	7	-8.057	CAACCCCATTCGTCAGACTCGT	ATG	15664	630
12	-3.130	2.393	13	-4.175	GACTCGTATGGACCTCGAGATT	GTG	15679	615
13	-5.301	1.354	10	-5.995	CCTCCGCAGTCTTGATATTCTG	ATG	15724	570
14	-4.942	1.525	10	-5.637	GCACGACACTCCCGAGACTGAG	TTG	15748	546
15	-4.857	1.566	16	-6.653	GACTGAGTTGTCGCGTGACCCG	ATG	15763	531
16	-2.915	2.496	9	-3.690	CCACTTCCGAGCAGGACAGCCG	ATG	15814	480
17	-5.924	1.055	10	-6.619	TGAGAATCCGACCGATCGCGCG	ATG	15910	384
18	-5.546	1.236	9	-6.321	AACCCTCGAGCGCGGCAAGATC	ATG	16096	198
19	-4.127	1.916	12	-4.963	GAGCAAGGCCGGGAACAATGTC	ATG	16126	168
20	-4.651	1.665	8	-5.873	CGGGAACAATGTCATGGGCGAG	ATG	16135	159
21	-4.169	1.896	12	-5.005	GCCCATCCCCGGTAAGACGTTC	TTG	16159	135
22	-4.897	1.547	9	-5.672	CCCGCCGTACACTCGAAAGACG	TTG	16192	102
23	-4.193	1.884	17	-6.193	TCTCTGGTCGCGGCCCTACGGA	TTG	16279	15
24	-3.019	2.446	9	-3.794	GTCGCGGCCCTACGGATTGGAG	ATG	16285	9

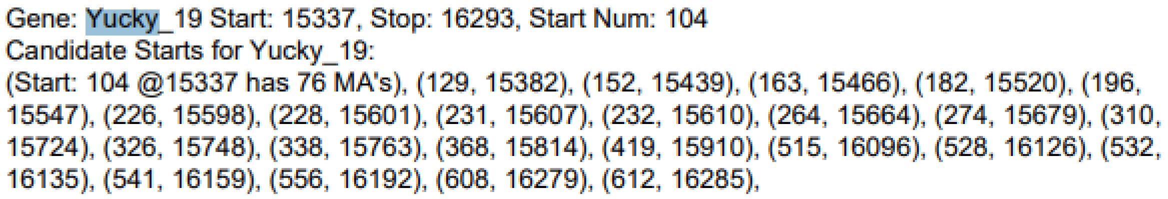
- The Z-value is 2.901.
- The final score is -3.116
- No other RBS numbers indicate and alternative start

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



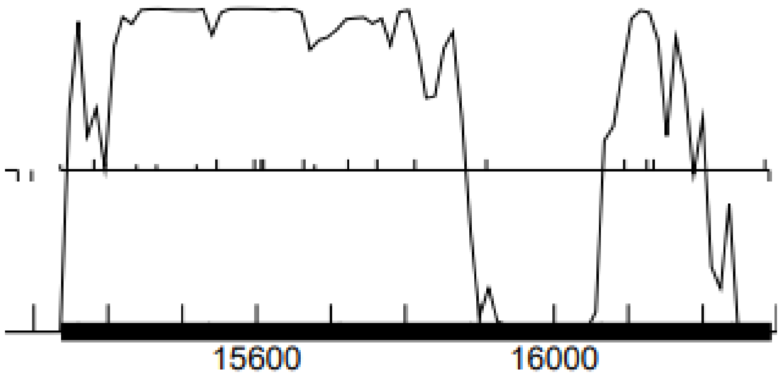
Start 104:
• Found in 95 of 1329 (7.1%) of genes in pham
• Manual Annotations of this start: 76 of 1144
• Called 98.9% of time when present
• Phage (with cluster) where this start called: Agatha_17 (CT), AikoCarson_16 (CT), Amok_16 (CT), Anaysia_29 (A15), AndPeggy_16 (CT), Anon_27 (A15), Apricot_18 (DN3), Axym_17 (CT), Azira_19 (CT), Battleship_30 (A15), Bavilard_17 (CT), BigChungus_16 (CT), BillDoor_19 (CT), Biskit_18 (CT), Blondies_17 (CT), Boohoo_29 (A15), Burnsey_17 (CT), Button_18 (CT), Buttrmlkdreams_17 (CT), CanesSauce_17 (CT), Carsonalex_18 (CT), CherryonLim_18 (CT), ChickenTender_19 (CT), ChocoMunchkin_17 (CT), Cleo_17 (CT), Cozz_17 (CT), Crater_17 (DN3), DekHockey33_29 (A15), Dre3_17 (CT), Elinal_18 (CT), Eliott_17 (CT), Emalyn_16 (CT), Epsocamisio_29 (A15), Feastonyeet_16 (CT), Fribs8_18 (CT), GiKK_20 (CT), Gibbous_17 (CT), GoldHunter_18 (CT), Hexbug_19 (CT), HippoPololi_19 (CT), Horseradish_18 (CT), JSwag_29 (A15), Jamzy_20 (CT), KatherineG_29 (A15), KayGee_17 (CT), LastResort_29 (A15), Lauer_16 (CT), Looper_30 (A15), MAnor_17 (CT), MScarn_19 (CT), MaVan_19 (CT), Margaret_20 (CT), Mayweather_18 (CT), MinecraftSteve_30 (A15), MunkgeeRoachy_17 (CT), Nebulosus_29 (A15), Nibbles_19 (CT), Nina_18 (CT), Nodigi_19 (CT), Oofda_30 (A15), Orla_19 (CT), Pons_17 (CT), PotPie_17 (CT), PsychoKiller_17 (CT), Quasar_17 (CT), RanchParmCat_20 (CT), ReMo_29 (A15), RedBaron_18 (CT), Remus_29 (A15), Rosalind_29 (A15), ShayRa_30 (A15), SheckWes_16 (CT), SketchMex_16 (CT), Socotra_18 (CT), Sopespian_17 (CT), Soups_29 (A15), Starburst_18 (CT), SteamedHams_19 (CT), Strosahl_29 (A15), SummitAcademy_16 (CT), Survivors_19 (CT), SweatNTears_19 (CT), Switzerland_29 (A15), Tolls_19 (CT), Troje_17 (CT), Typhonomachy_18 (CT), Vine_18 (CT), Waits_29 (A15), Warrior24_30 (A15), Yakult_18 (CT), Yarn_16 (CT), Yucky_19 (CT), Yummy_18 (CT), Zareef_21 (CT),

- The proposed start has 76 MAs. No other start site has ever been manually annotated. It is called 98.9% of the time when present.



Gene: Yucky_19 Start: 15337, Stop: 16293, Start Num: 104
Candidate Starts for Yucky_19:
(Start: 104 @15337 has 76 MA's), (129, 15382), (152, 15439), (163, 15466), (182, 15520), (196, 15547), (226, 15598), (228, 15601), (231, 15607), (232, 15610), (264, 15664), (274, 15679), (310, 15724), (326, 15748), (338, 15763), (368, 15814), (419, 15910), (515, 16096), (528, 16126), (532, 16135), (541, 16159), (556, 16192), (608, 16279), (612, 16285),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- I do not believe the start site cuts off any coding potential, if it does it is very minimal.

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

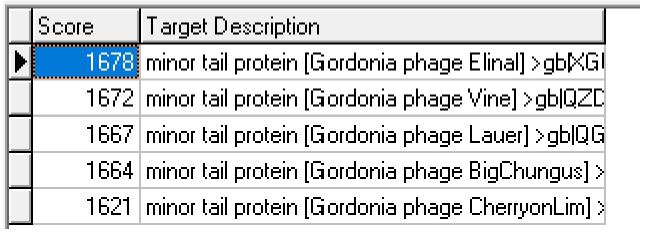
- 15340-15337=3+1=4

- There is an overlap of 4 with the previous gene. This is an acceptable overlap.

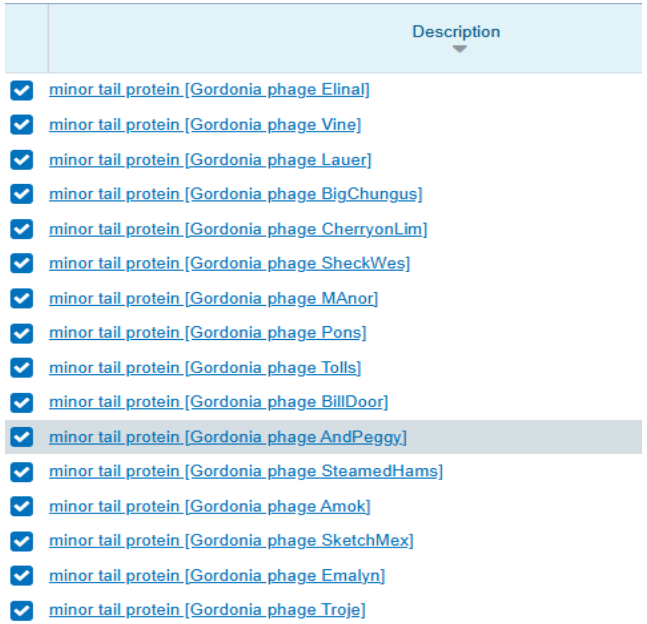
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is the same as the automated start site of 15337. I believe this because it has a lot of 1:1 alignments, 24 to be exact. It also has very good RBS numbers with a Z-value of 2.901 and a Final score of -3.116. It is the only start site to ever be manually annotated, and it is called very frequently when present. It also has an acceptable overlap value. Lastly, it cuts off very little, if any, coding potential.

BLAST function evidence. What assigned functions do other highly similar genes have?



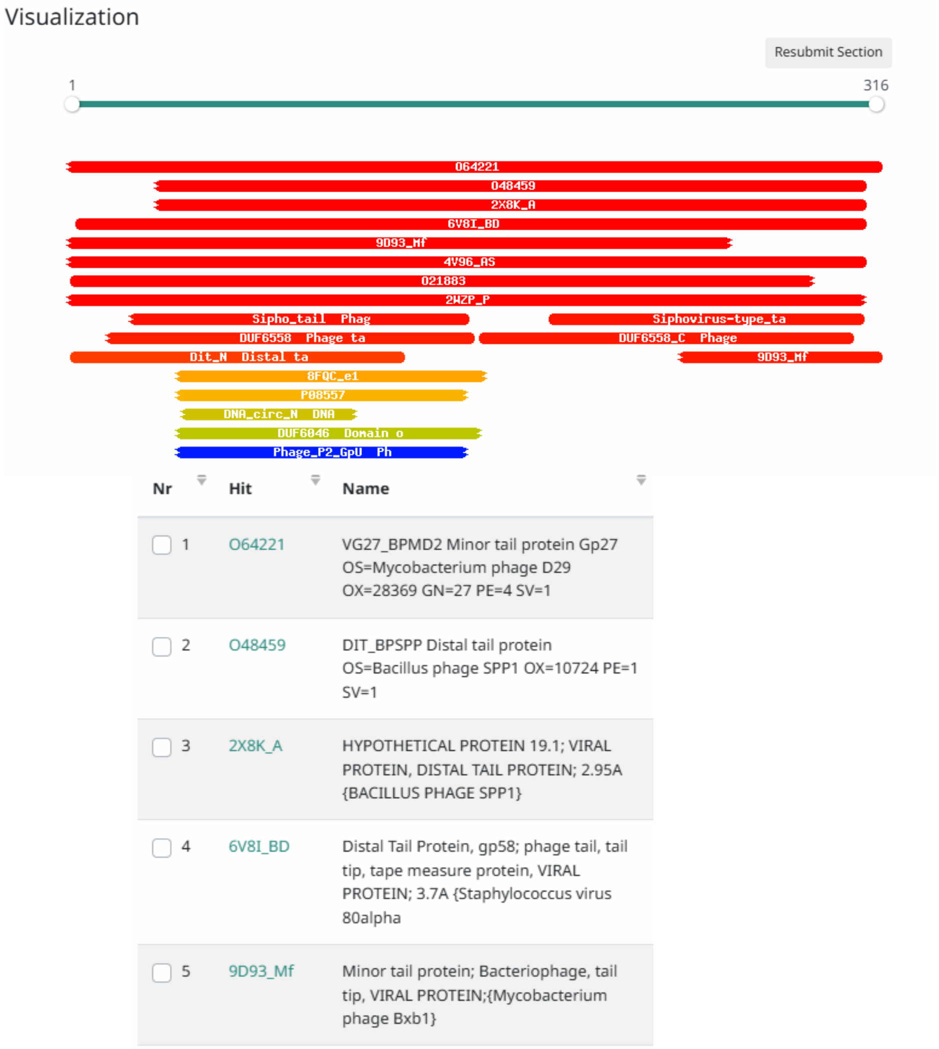
Score	Target Description
1678	minor tail protein [Gordonia phage Elinal] >gb|XGI
1672	minor tail protein [Gordonia phage Vine] >gb|QZD
1667	minor tail protein [Gordonia phage Lauer] >gb|QG
1664	minor tail protein [Gordonia phage BigChungus] >
1621	minor tail protein [Gordonia phage CherryonLim] >



	Description
☑	minor tail protein [Gordonia phage Elinal]
☑	minor tail protein [Gordonia phage Vine]
☑	minor tail protein [Gordonia phage Lauer]
☑	minor tail protein [Gordonia phage BigChungus]
☑	minor tail protein [Gordonia phage CherryonLim]
☑	minor tail protein [Gordonia phage SheckWes]
☑	minor tail protein [Gordonia phage MAnor]
☑	minor tail protein [Gordonia phage Pons]
☑	minor tail protein [Gordonia phage Tolls]
☑	minor tail protein [Gordonia phage BillDoor]
☑	minor tail protein [Gordonia phage AndPeggy]
☑	minor tail protein [Gordonia phage SteamedHams]
☑	minor tail protein [Gordonia phage Amok]
☑	minor tail protein [Gordonia phage SketchMex]
☑	minor tail protein [Gordonia phage Emalyn]
☑	minor tail protein [Gordonia phage Troje]

- DNA master BLAST shows at least 25 highly similar genes with the function minor tail protein.
- NCBI BLASTing showed the same results.

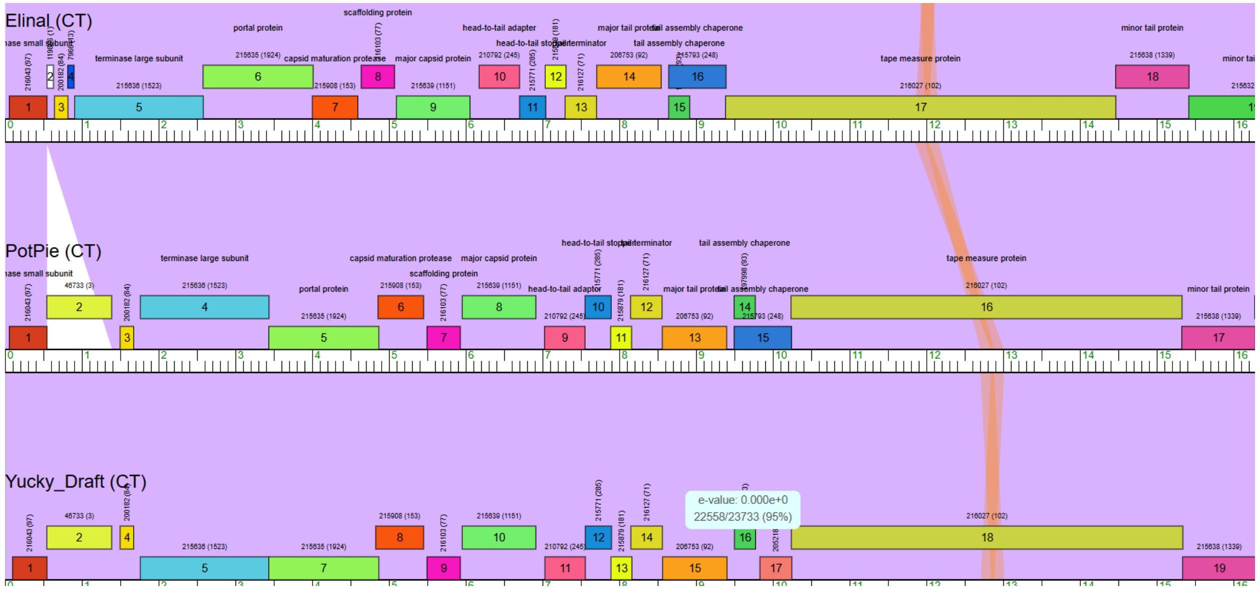
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



Nr	Hit	Name
□ 1	O64221	VG27_BPMD2 Minor tail protein Gp27 OS=Mycobacterium phage D29 OX=28369 GN=27 PE=4 SV=1
□ 2	O48459	DIT_BPSPP Distal tail protein OS=Bacillus phage SPP1 OX=10724 PE=1 SV=1
□ 3	2X8K_A	HYPOTHETICAL PROTEIN 19.1; VIRAL PROTEIN, DISTAL TAIL PROTEIN; 2.95A {BACILLUS PHAGE SPP1}
□ 4	6V8I_BD	Distal Tail Protein, gp58; phage tail, tail tip, tape measure protein, VIRAL PROTEIN; 3.7A {Staphylococcus virus 80alpha
□ 5	9D93_Mf	Minor tail protein; Bacteriophage, tail tip, VIRAL PROTEIN;{Mycobacterium phage Bxb1}

- HHpred shows results for many different functions including many for minor tail proteins. The hits are largely homologous throughout.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- PotPie, BigChungus, and Elinal all have this gene and in all 3 it is a minor tail protein in all 3 phages.
- None of the phages have a conserved domain.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- I would like to call this gene a minor tail protein.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function is a minor tail protein. BLAST evidence on both DNA master and NCBI shows very strong evidence for this being a minor tail protein. HHpred's evidence isn't as strong as I would like it to be, but it is strong enough for me to be confident in calling it still. HHpred shows a couple hits for a minor tail protein. Lastly, BigChungus, Elinal, and PotPie contain this gene and has it called as a minor tail protein. Also, synteny indicates this as a minor tail protein.

Feature 18 – Stop 17984

Instructions

Fill this out for each gene you annotate. This should be thought of as the minimum amount of information that needs to be provided for each gene. You can always add more slides or information as necessary

- Is it a gene?
  - Yes
- Where does it start?
  - 16290
- What is the function?
  - Minor tail protein

Glimmer/GeneMark

What feature number is this? **18**

What is the stop site? **17984**

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Glimmer and GeneMark

What is the autoannotated start?

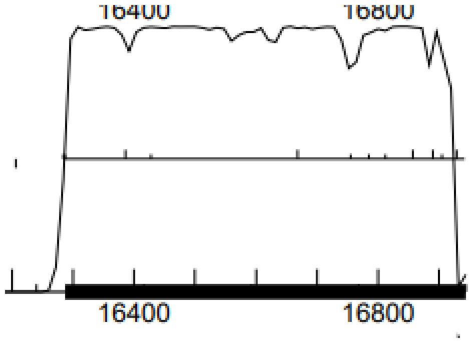
16287

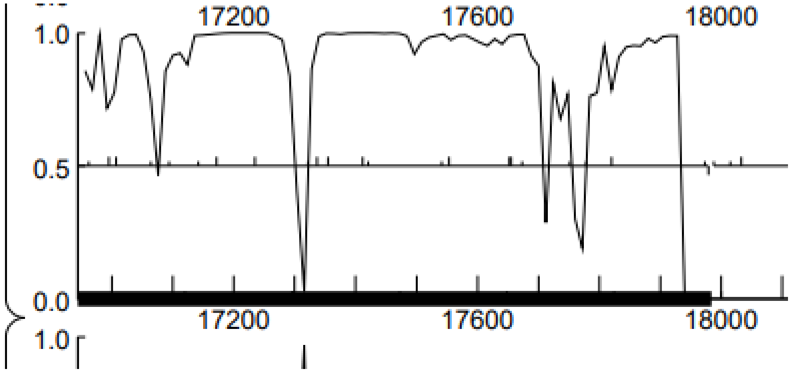
Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

Overlap of 7

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

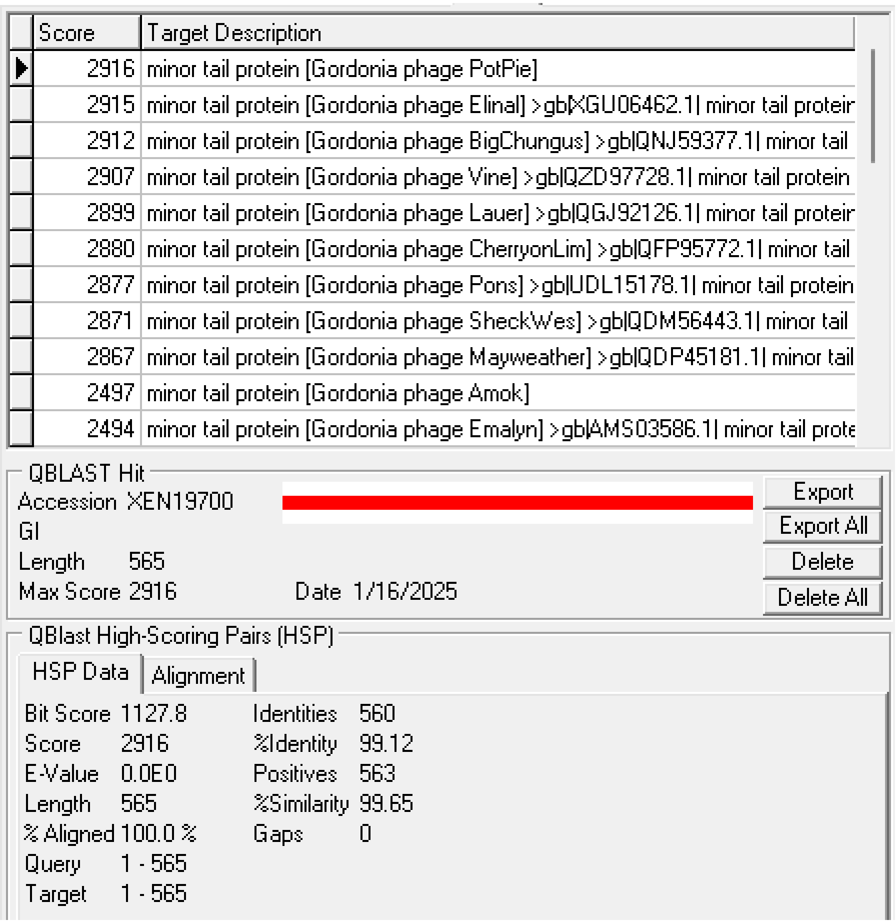
- There is strong coding potential throughout where the feature is called to be. The potential does start slightly before where the feature is called to start, but 16287 was the earliest possible start site.





BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- At least 25 BLAST hits of highly similar genes from other phages
- All e-values are extremely close to zero
- 14 1:1 alignments for auto annotated starts



Score	Target Description
2916	minor tail protein [Gordonia phage PotPie]
2915	minor tail protein [Gordonia phage Elinal] >gb|XGU06462.1| minor tail protei
2912	minor tail protein [Gordonia phage BigChungus] >gb|QNJ59377.1| minor tail
2907	minor tail protein [Gordonia phage Vine] >gb|QZD97728.1| minor tail protein
2899	minor tail protein [Gordonia phage Lauer] >gb|QGJ92126.1| minor tail protei
2880	minor tail protein [Gordonia phage CherryonLim] >gb|QFP95772.1| minor tail
2877	minor tail protein [Gordonia phage Pons] >gb|UDL15178.1| minor tail protein
2871	minor tail protein [Gordonia phage SheckWes] >gb|QDM56443.1| minor tail
2867	minor tail protein [Gordonia phage Mayweather] >gb|QDP45181.1| minor tail
2497	minor tail protein [Gordonia phage Amok]
2494	minor tail protein [Gordonia phage Emalyn] >gb|AMS03586.1| minor tail prote

QBLAST Hit

Accession XEN19700

GI

Length 565

Max Score 2916

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

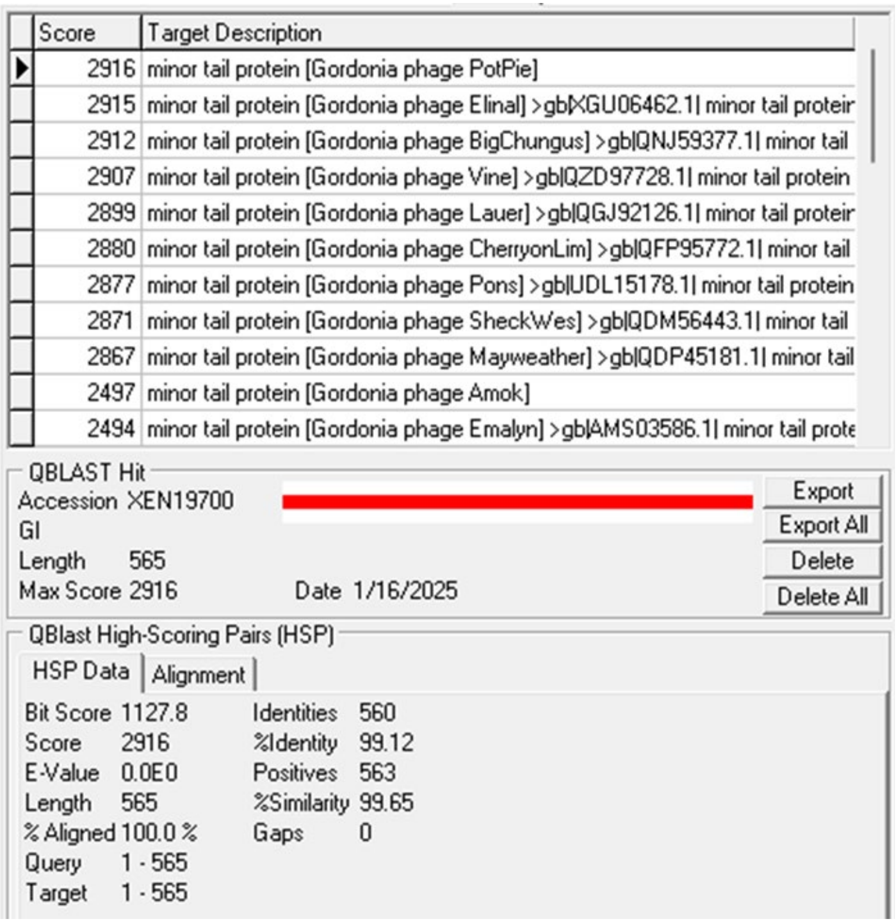
Bit Score	1127.8	Identities	560
Score	2916	%Identity	99.12
E-Value	0.0E0	Positives	563
Length	565	%Similarity	99.65
% Aligned	100.0 %	Gaps	0
Query	1 - 565		
Target	1 - 565		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This feature is a gene! There is strong coding potential throughout where the feature is called to be, and there are at least 25 BLAST hits of highly similar genes from other phages that all have e-values extremely close to zero.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- 16287 - 14 1:1 alignments
- 16290 - 9 1:1 alignments
- 16287 is favored based off this evidence alone



Score	Target Description
2916	minor tail protein [Gordonia phage PotPie]
2915	minor tail protein [Gordonia phage Elinal] >gb|XGU06462.1| minor tail proteir
2912	minor tail protein [Gordonia phage BigChungus] >gb|QNJ59377.1| minor tail
2907	minor tail protein [Gordonia phage Vine] >gb|QZD97728.1| minor tail protein
2899	minor tail protein [Gordonia phage Lauer] >gb|QGJ92126.1| minor tail proteir
2880	minor tail protein [Gordonia phage CherryonLim] >gb|QFP95772.1| minor tail
2877	minor tail protein [Gordonia phage Pons] >gb|UDL15178.1| minor tail protein
2871	minor tail protein [Gordonia phage SheckWes] >gb|QDM56443.1| minor tail
2867	minor tail protein [Gordonia phage Mayweather] >gb|QDP45181.1| minor tail
2497	minor tail protein [Gordonia phage Amok]
2494	minor tail protein [Gordonia phage Emalyn] >gb|AMS03586.1| minor tail prote

QBLAST Hit

Accession XEN19700

GI

Length 565

Max Score 2916

Date 1/16/2025

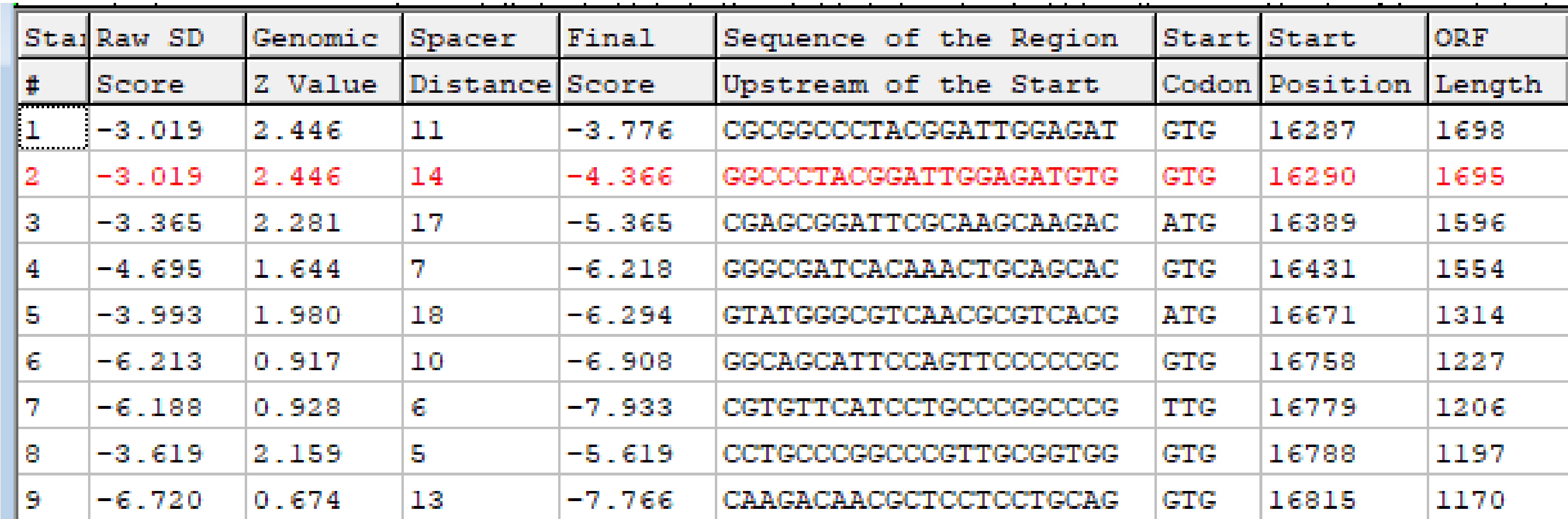
QBlast High-Scoring Pairs (HSP)

Bit Score	1127.8	Identities	560
Score	2916	%Identity	99.12
E-Value	0.0E0	Positives	563
Length	565	%Similarity	99.65
% Aligned	100.0 %	Gaps	0
Query	1 - 565		
Target	1 - 565		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- 16287
  - Z-value = 2.446
  - Final score = -3.776
- 16290
  - Z-value = 2.446
  - Final score = -4.366

- 16287 is the favored start based off this evidence alone

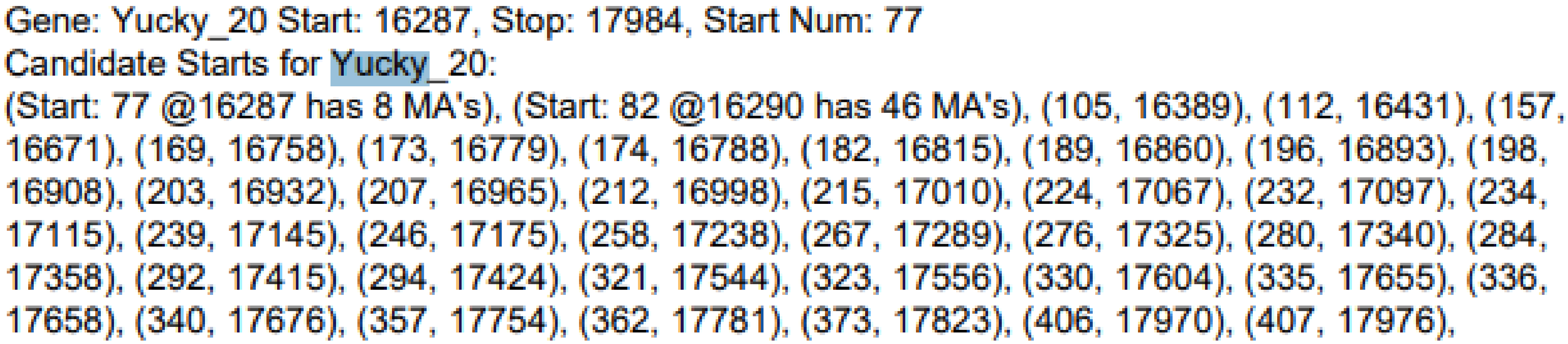


Sta	Raw SD	Genomic	Spacer	Final	Sequence of the Region	Start	Start	ORF
#	Score	Z Value	Distance	Score	Upstream of the Start	Codon	Position	Length
1	-3.019	2.446	11	-3.776	CGCGGCCCTACGGATTGGAGAT	GTG	16287	1698
2	-3.019	2.446	14	-4.366	GGCCCTACGGATTGGAGATGTG	GTG	16290	1695
3	-3.365	2.281	17	-5.365	CGAGCGGATTCGCAAGCAAGAC	ATG	16389	1596
4	-4.695	1.644	7	-6.218	GGGCGATCACAAACTGCAGCAC	GTG	16431	1554
5	-3.993	1.980	18	-6.294	GTATGGGCGTCAACGCGTCACG	ATG	16671	1314
6	-6.213	0.917	10	-6.908	GGCAGCATTCCAGTTCCCCCGC	GTG	16758	1227
7	-6.188	0.928	6	-7.933	CGTGTTCATCCTGCCCGGCCCG	TTG	16779	1206
8	-3.619	2.159	5	-5.619	CCTGCCCGGCCCGTTGCGGTGG	GTG	16788	1197
9	-6.720	0.674	13	-7.766	CAAGACAACGCTCCTCCTGCAG	GTG	16815	1170

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- 16287 – 8 MA’s
- 16290 – 46 MA’s

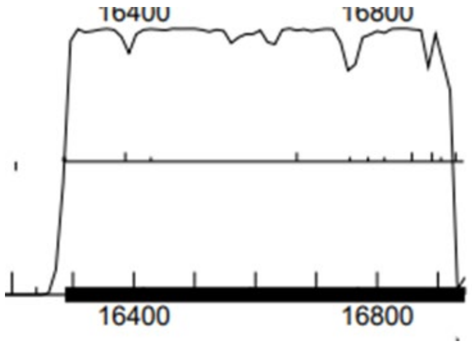
- 16290 is favored off this evidence alone

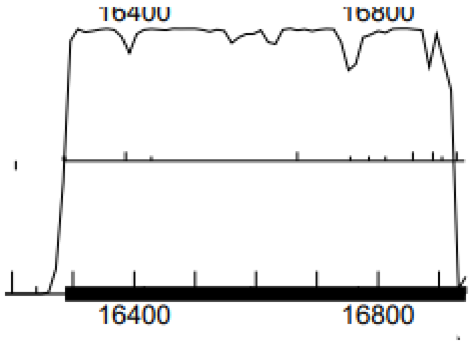


Gene: Yucky_20 Start: 16287, Stop: 17984, Start Num: 77
Candidate Starts for Yucky_20:
(Start: 77 @16287 has 8 MA's), (Start: 82 @16290 has 46 MA's), (105, 16389), (112, 16431), (157, 16671), (169, 16758), (173, 16779), (174, 16788), (182, 16815), (189, 16860), (196, 16893), (198, 16908), (203, 16932), (207, 16965), (212, 16998), (215, 17010), (224, 17067), (232, 17097), (234, 17115), (239, 17145), (246, 17175), (258, 17238), (267, 17289), (276, 17325), (280, 17340), (284, 17358), (292, 17415), (294, 17424), (321, 17544), (323, 17556), (330, 17604), (335, 17655), (336, 17658), (340, 17676), (357, 17754), (362, 17781), (373, 17823), (406, 17970), (407, 17976),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- 16287 – cuts off the initial peak of coding potential, but a majority of the coding potential is included
- 16290 – includes about the same amount of coding potential as 16287, but it does cut of a bit more
- 16287 would be the favored start based off this evidence alone





Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 16287 – overlap of 7
- 16290 – overlap of 4

- 16290 would be favored based off this evidence alone

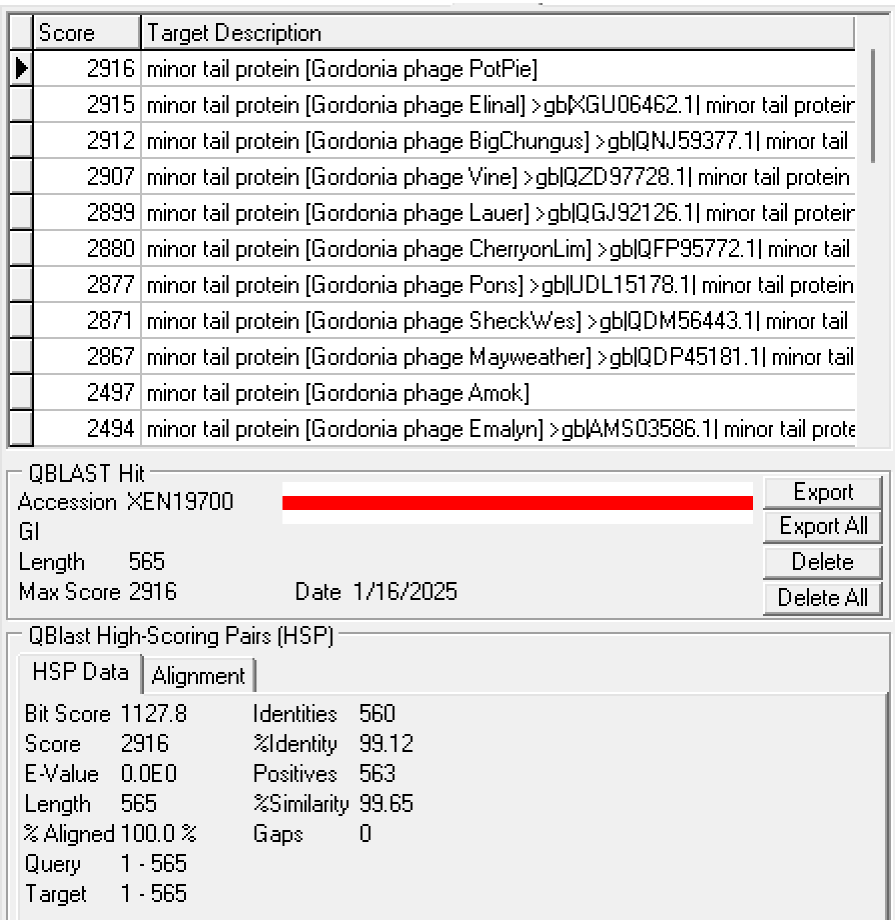
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	Starting 16287	Starting 16290
Glimmer/GeneMark	Glimmer & GeneMark	Starterator
Coding potential	cuts off the initial peak of coding potential, but a majority of the coding potential is included	includes about the same amount of coding potential as 16287, but it does cut of a bit more
BLAST	14 1:1 alignments	9 1:1 alignments
RBS Score	Z-value = 2.446 Final score = -3.776	Z-value = 2.446 Final score = -4.366
Starterator	8 MA's	46 MA's
Gap/Overlap	7 overlap	4 overlap

The start for this gene is likely 16290. 16287 and 16290 are tandem starts, so both potential start include about the same amount of coding potential, but based of the guiding principles the second start should be used. The RBS scores for both start sites were also similar. They had the same z-value and 16287 had a slightly better final score. 16290 had 46 manual annotation whereas 16287 only had 8. 16290 also a the more favorable overlap of 4 over 7.

BLAST function evidence. What assigned functions do other highly similar genes have?

- There were at least 25 BLAST hits that called the function of minor tail protein for highly similar genes to this one.



Score	Target Description
2916	minor tail protein [Gordonia phage PotPie]
2915	minor tail protein [Gordonia phage Elinal] >gb|XGU06462.1| minor tail protein
2912	minor tail protein [Gordonia phage BigChungus] >gb|QNJ59377.1| minor tail
2907	minor tail protein [Gordonia phage Vine] >gb|QZD97728.1| minor tail protein
2899	minor tail protein [Gordonia phage Lauer] >gb|QGJ92126.1| minor tail protein
2880	minor tail protein [Gordonia phage CherryonLim] >gb|QFP95772.1| minor tail
2877	minor tail protein [Gordonia phage Pons] >gb|UDL15178.1| minor tail protein
2871	minor tail protein [Gordonia phage SheckWes] >gb|QDM56443.1| minor tail
2867	minor tail protein [Gordonia phage Mayweather] >gb|QDP45181.1| minor tail
2497	minor tail protein [Gordonia phage Amok]
2494	minor tail protein [Gordonia phage Emalyn] >gb|AMS03586.1| minor tail prote

QBLAST Hit

Accession XEN19700

GI

Length 565

Max Score 2916

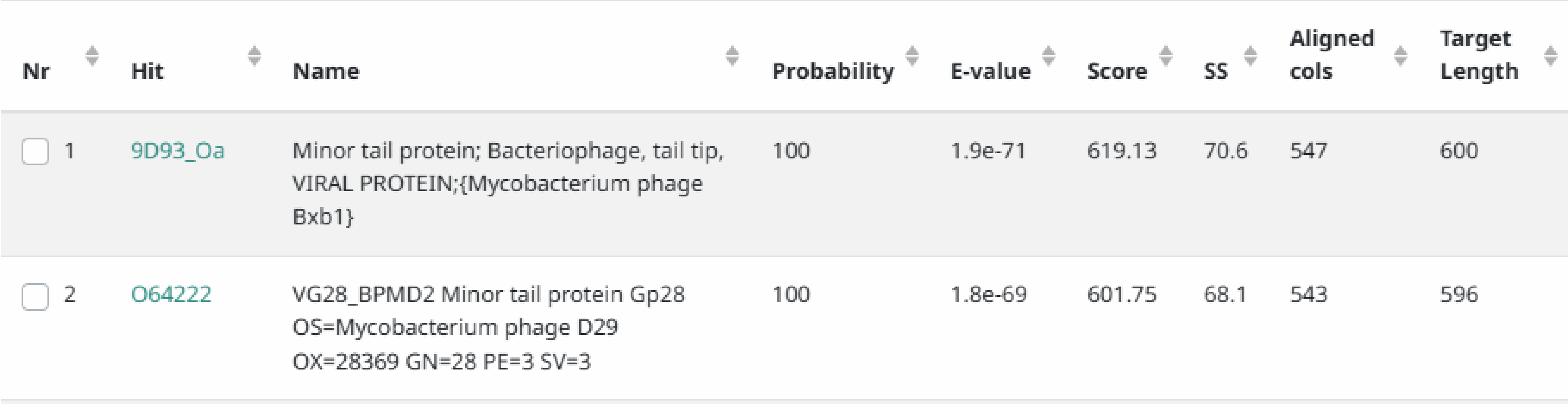
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	1127.8	Identities	560
Score	2916	%Identity	99.12
E-Value	0.0E0	Positives	563
Length	565	%Similarity	99.65
% Aligned	100.0 %	Gaps	0
Query	1 - 565		
Target	1 - 565		

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

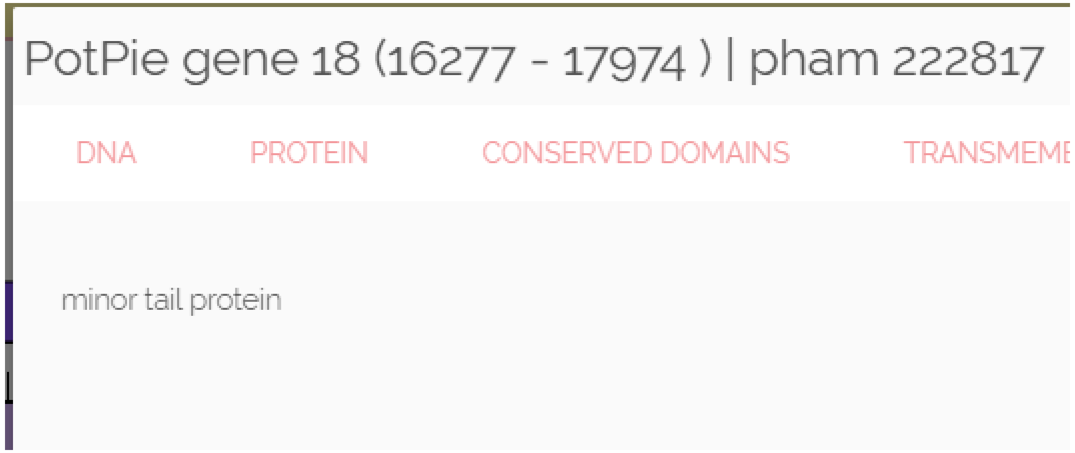
- Hhpred showed several hits with over 90 and e-values close to zero. These hits labeled the function as a minor tail protein as well and were homologous for a majority of the gene. There were no conserved domains shown.



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	9D93_Oa	Minor tail protein; Bacteriophage, tail tip, VIRAL PROTEIN;{Mycobacterium phage Bxb1}	100	1.9e-71	619.13	70.6	547	600
□ 2	O64222	VG28_BPMD2 Minor tail protein Gp28 OS=Mycobacterium phage D29 OX=28369 GN=28 PE=3 SV=3	100	1.8e-69	601.75	68.1	543	596

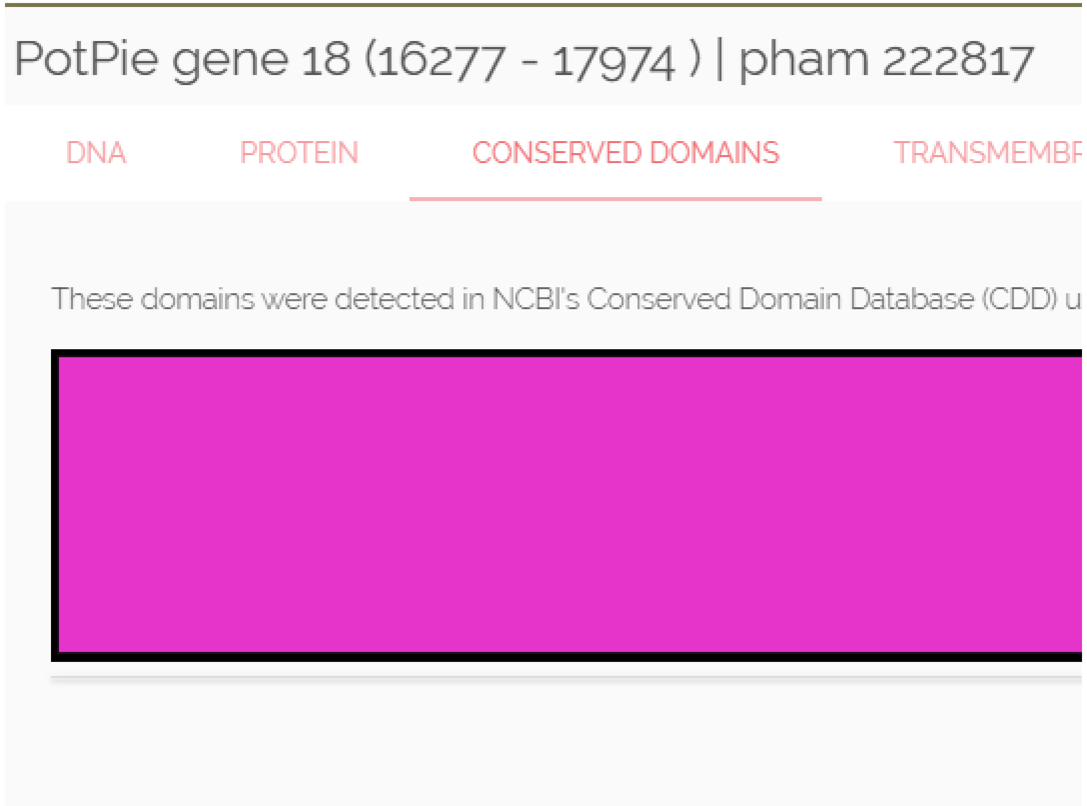
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this ge domains?

- Phamerator showed that phages with genes in the same pham as this one called the function as a minor tail protein and they did not have any conserved domains.



PotPie gene 18 (16277 - 17974) | pham 222817

minor tail protein



PotPie gene 18 (16277 - 17974) | pham 222817

These domains were detected in NCBI's Conserved Domain Database (CDD) u

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- **Not applicable since there is a probable function**

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Official function → minor tail protein
- The function for this gene should be labeled as a minor tail protein. There were at least 25 BLAST hits that showed highly similar genes from other phages having the designated function of minor tail protein, and all the e-values for those hits were extremely close to zero. Hhpred also showed several hits with probabilities above 90 that suggested the function of this gene should be labeled as a minor tail protein. Phamerator showed that phages with genes in the same pham as this one called their function as a minor tail protein without the presence of conserved domains. Since there was a probable function for this gene a graph from Deep TMHMM was not necessary.

Feature 19 – Stop 18373

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

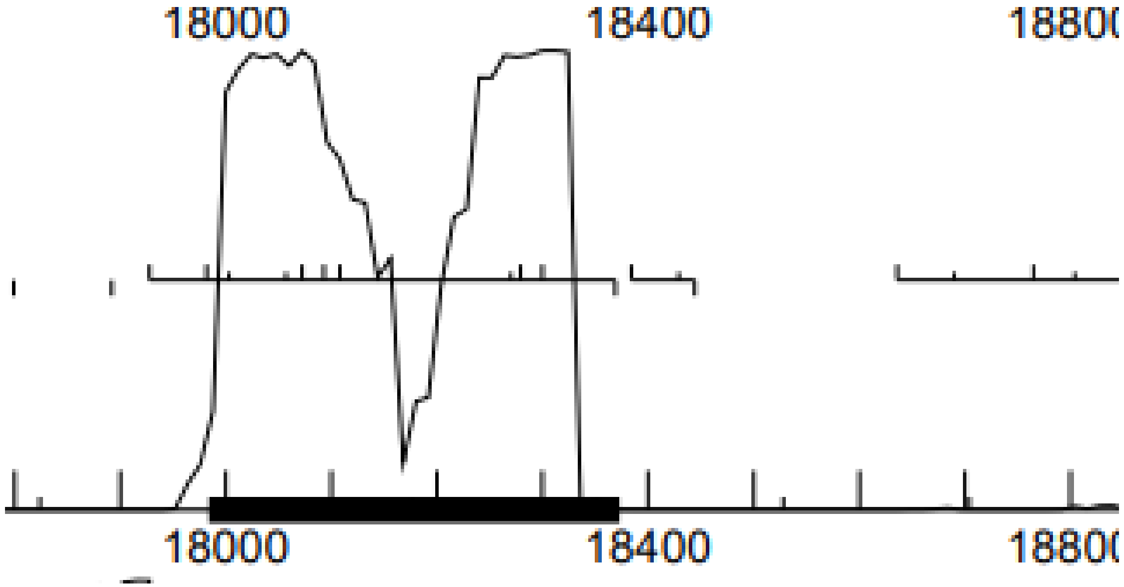
What is the autoannotated start?

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

- Feature: 19
- Stop site: 18373
- Called by both Glimmer & GeneMark
- Autoannotated start: 17984
- Overlap: 4

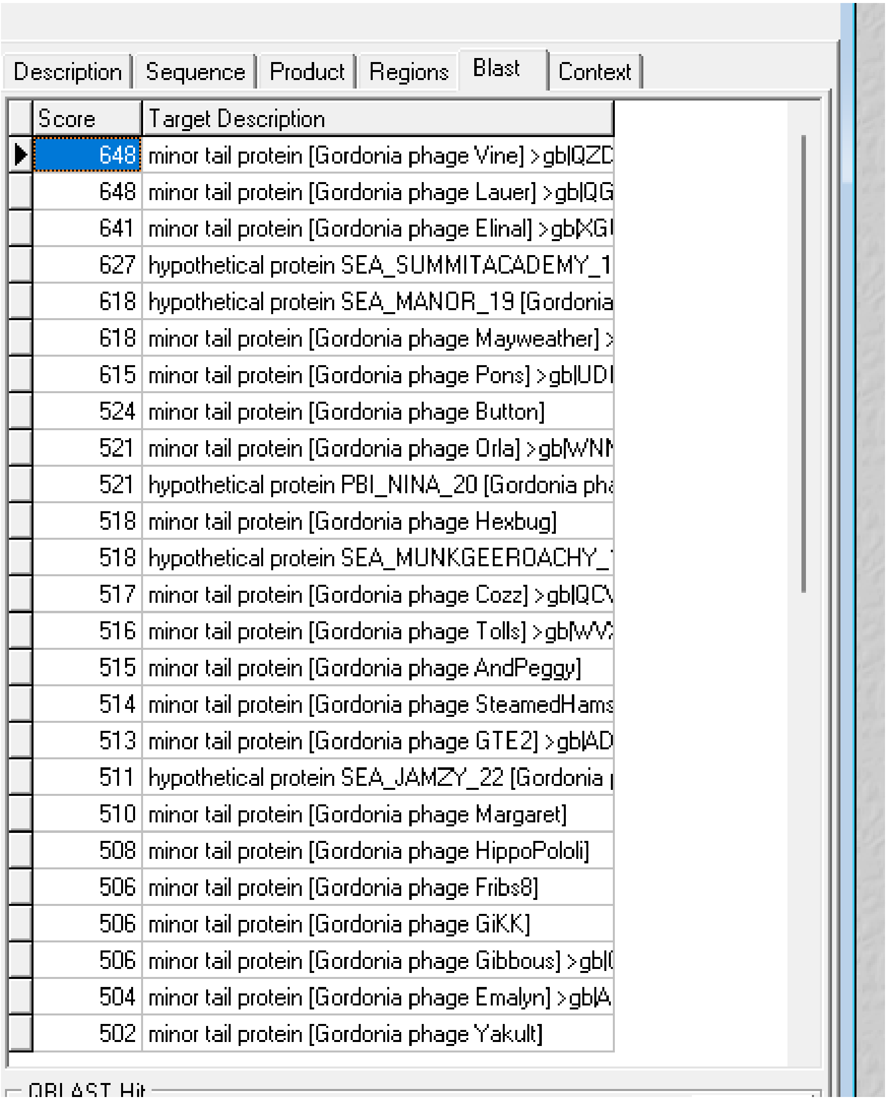
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Start site: 17984
- CP in reading frame 2
- Cuts off some coding potential



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 25 highly similar genes
- All with a 0.0E0 value



Score	Target Description
648	minor tail protein [Gordonia phage Vine] >gb|QZD
648	minor tail protein [Gordonia phage Lauer] >gb|QG
641	minor tail protein [Gordonia phage Elinal] >gb|XGI
627	hypothetical protein SEA_SUMMITACADEMY_1
618	hypothetical protein SEA_MANOR_19 [Gordonia
618	minor tail protein [Gordonia phage Mayweather] >
615	minor tail protein [Gordonia phage Pons] >gb|UDI
524	minor tail protein [Gordonia phage Button]
521	minor tail protein [Gordonia phage Orla] >gb|WNN
521	hypothetical protein PBI_NINA_20 [Gordonia pha
518	minor tail protein [Gordonia phage Hexbug]
518	hypothetical protein SEA_MUNKGEEROACHY_1
517	minor tail protein [Gordonia phage Cozz] >gb|QCV
516	minor tail protein [Gordonia phage Tolls] >gb|WV
515	minor tail protein [Gordonia phage AndPeggy]
514	minor tail protein [Gordonia phage SteamedHams
513	minor tail protein [Gordonia phage GTE2] >gb|AD
511	hypothetical protein SEA_JAMZY_22 [Gordonia p
510	minor tail protein [Gordonia phage Margaret]
508	minor tail protein [Gordonia phage HippoPololi]
506	minor tail protein [Gordonia phage Fribs8]
506	minor tail protein [Gordonia phage GiKK]
506	minor tail protein [Gordonia phage Gibbous] >gb|(
504	minor tail protein [Gordonia phage Emalyn] >gb|A
502	minor tail protein [Gordonia phage Yakult]

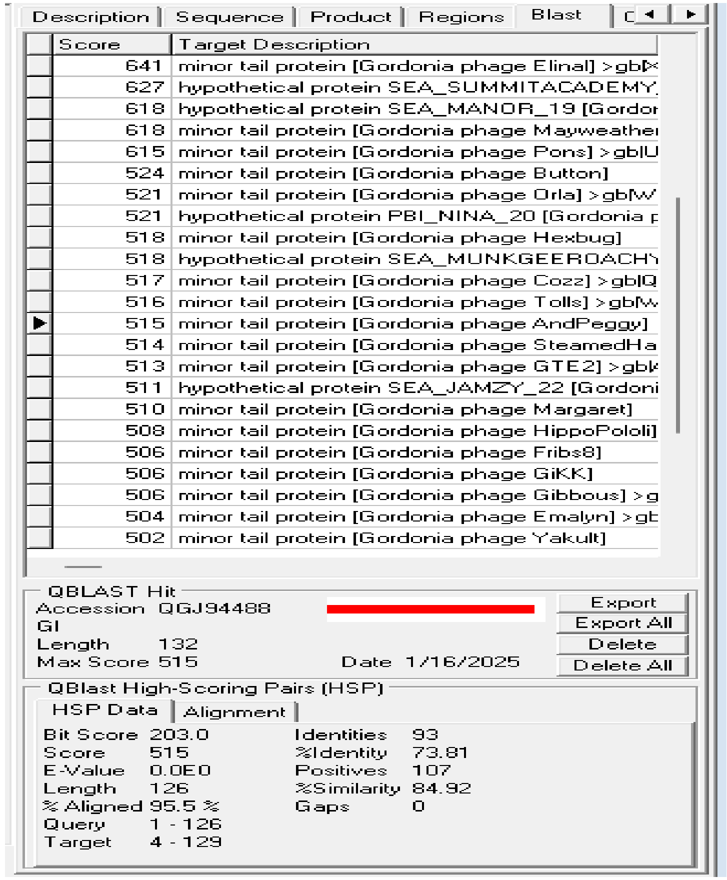
Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes it is a gene, because both Glimmer and GeneMark call the same start, includes strong coding potential within the reading frame, and has 25 highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

Start: 17984

- 24 1:1 alignments

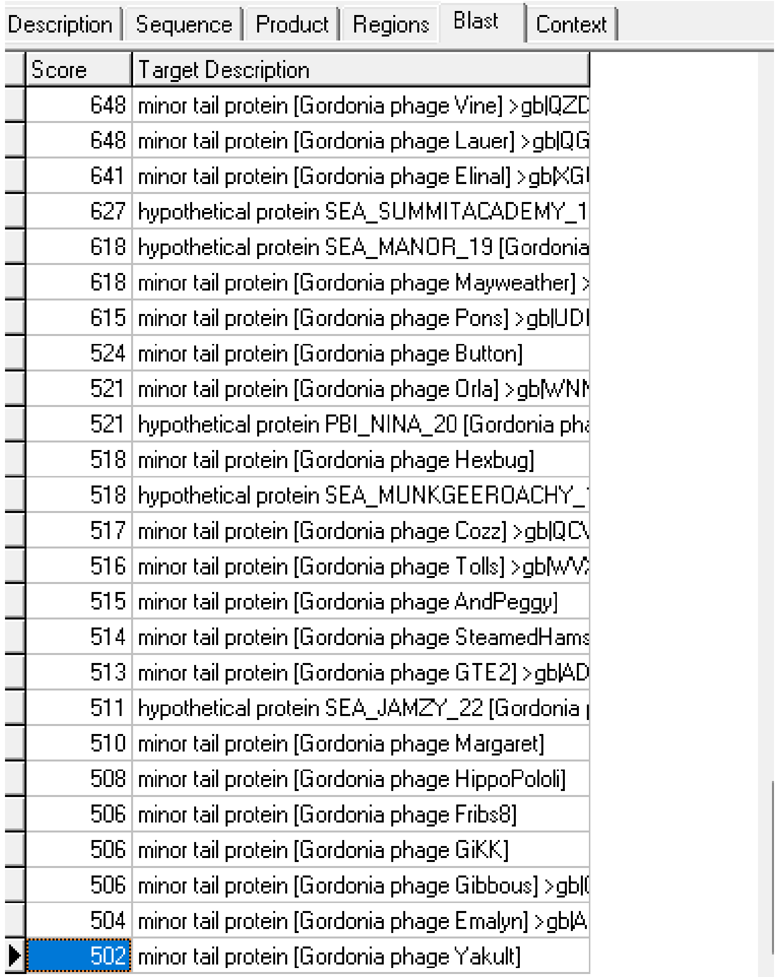


Score	Target Description
641	minor tail protein [Gordonia phage Elinal] >gb|X
627	hypothetical protein SEA_SUMMITACADEMY_
618	hypothetical protein SEA_MANOR_19 [Gordor
618	minor tail protein [Gordonia phage Mayweather
615	minor tail protein [Gordonia phage Pons] >gb|U
524	minor tail protein [Gordonia phage Button]
521	minor tail protein [Gordonia phage Orla] >gb|W
521	hypothetical protein PBI_NINA_20 [Gordonia p
518	minor tail protein [Gordonia phage Hexbug]
518	hypothetical protein SEA_MUNKGEEROACHY
517	minor tail protein [Gordonia phage Cozz] >gb|Q
516	minor tail protein [Gordonia phage Tolls] >gb|W
▶ 515	minor tail protein [Gordonia phage AndPeggy]
514	minor tail protein [Gordonia phage SteamedHa
513	minor tail protein [Gordonia phage GTE2] >gb|A
511	hypothetical protein SEA_JAMZY_22 [Gordoni
510	minor tail protein [Gordonia phage Margaret]
508	minor tail protein [Gordonia phage HippoPololi]
506	minor tail protein [Gordonia phage Fribs8]
506	minor tail protein [Gordonia phage GiKK]
506	minor tail protein [Gordonia phage Gibbous] >g
504	minor tail protein [Gordonia phage Emalyn] >gb
502	minor tail protein [Gordonia phage Yakult]

QBLAST Hit
Accession QGJ94488
GI
Length 132
Max Score 515
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

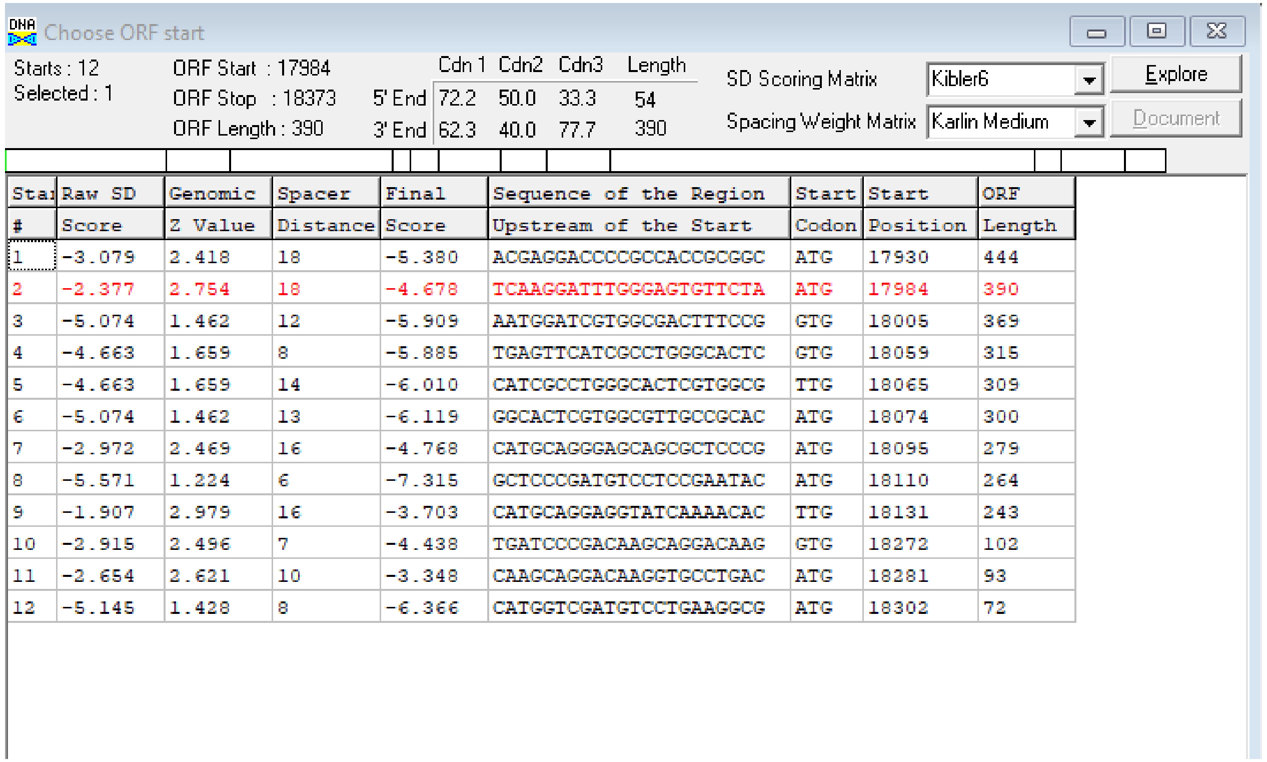
Bit Score	203.0	Identities	93
Score	515	%Identity	73.81
E-Value	0.0E0	Positives	107
Length	126	%Similarity	84.92
% Aligned	95.5 %	Gaps	0
Query	1 - 126		
Target	4 - 129		



Score	Target Description
648	minor tail protein [Gordonia phage Vine] >gb|QZD
648	minor tail protein [Gordonia phage Lauer] >gb|QG
641	minor tail protein [Gordonia phage Elinal] >gb|XG
627	hypothetical protein SEA_SUMMITACADEMY_1
618	hypothetical protein SEA_MANOR_19 [Gordonia
618	minor tail protein [Gordonia phage Mayweather] >
615	minor tail protein [Gordonia phage Pons] >gb|UD
524	minor tail protein [Gordonia phage Button]
521	minor tail protein [Gordonia phage Orla] >gb|WN
521	hypothetical protein PBI_NINA_20 [Gordonia pha
518	minor tail protein [Gordonia phage Hexbug]
518	hypothetical protein SEA_MUNKGEEROACHY_
517	minor tail protein [Gordonia phage Cozz] >gb|QC
516	minor tail protein [Gordonia phage Tolls] >gb|WV
515	minor tail protein [Gordonia phage AndPeggy]
514	minor tail protein [Gordonia phage SteamedHams
513	minor tail protein [Gordonia phage GTE2] >gb|AD
511	hypothetical protein SEA_JAMZY_22 [Gordonia
510	minor tail protein [Gordonia phage Margaret]
508	minor tail protein [Gordonia phage HippoPololi]
506	minor tail protein [Gordonia phage Fribs8]
506	minor tail protein [Gordonia phage GiKK]
506	minor tail protein [Gordonia phage Gibbous] >gb|
504	minor tail protein [Gordonia phage Emalyn] >gb|A
▶ 502	minor tail protein [Gordonia phage Yakult]

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Start 17984:
- Z Value: 2.754
- Final score: -4.678



Starts : 12
Selected : 1

ORF Start : 17984
ORF Stop : 18373
ORF Length : 390

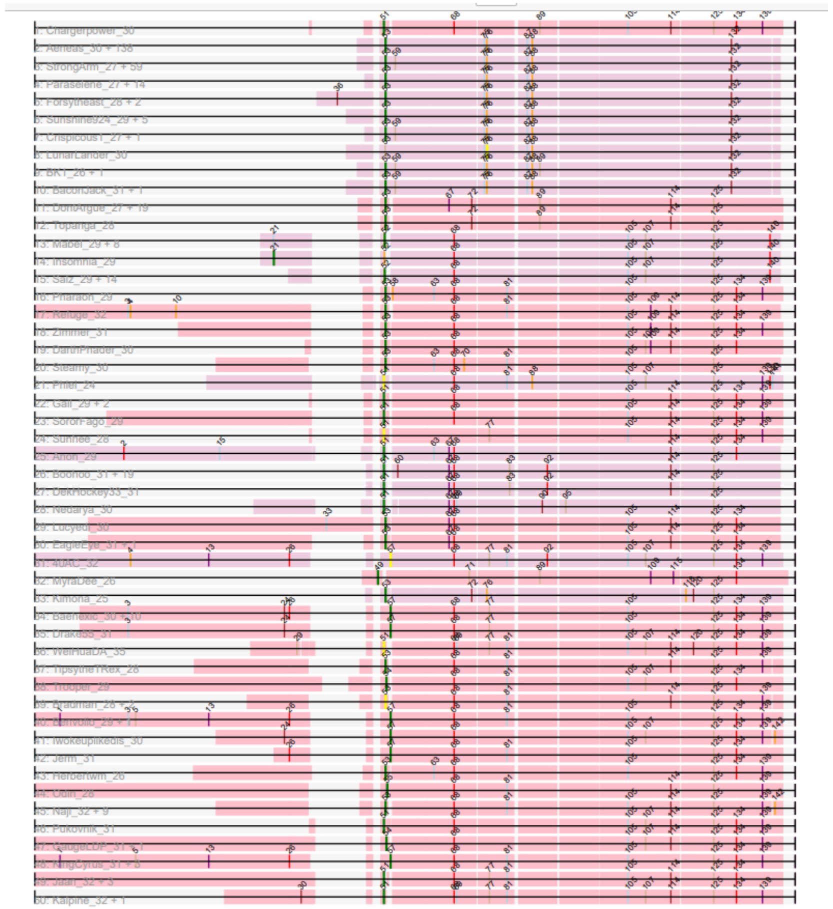
	Cdn 1	Cdn2	Cdn3	Length
5' End	72.2	50.0	33.3	54
3' End	62.3	40.0	77.7	390

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-3.079	2.418	18	-5.380	ACGAGGACCCCGCCACCGCGGC	ATG	17930	444
2	-2.377	2.754	18	-4.678	TCAAGGATTTGGGAGTGTTCTA	ATG	17984	390
3	-5.074	1.462	12	-5.909	AATGGATCGTGGCGACTTTCCG	GTG	18005	369
4	-4.663	1.659	8	-5.885	TGAGTTCATCGCCTGGGCACTC	GTG	18059	315
5	-4.663	1.659	14	-6.010	CATCGCCTGGGCACTCGTGGCG	TTG	18065	309
6	-5.074	1.462	13	-6.119	GGCACTCGTGGCGTTGCCGCAC	ATG	18074	300
7	-2.972	2.469	16	-4.768	CATGCAGGGAGCAGCGCTCCCG	ATG	18095	279
8	-5.571	1.224	6	-7.315	GCTCCCGATGTCCTCCGAATAC	ATG	18110	264
9	-1.907	2.979	16	-3.703	CATGCAGGAGGTATCAAAACAC	TTG	18131	243
10	-2.915	2.496	7	-4.438	TGATCCCGACAAGCAGGACAAG	GTG	18272	102
11	-2.654	2.621	10	-3.348	CAAGCAGGACAAGGTGCCTGAC	ATG	18281	93
12	-5.145	1.428	8	-6.366	CATGGTCGATGTCCTGAAGGCG	ATG	18302	72

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

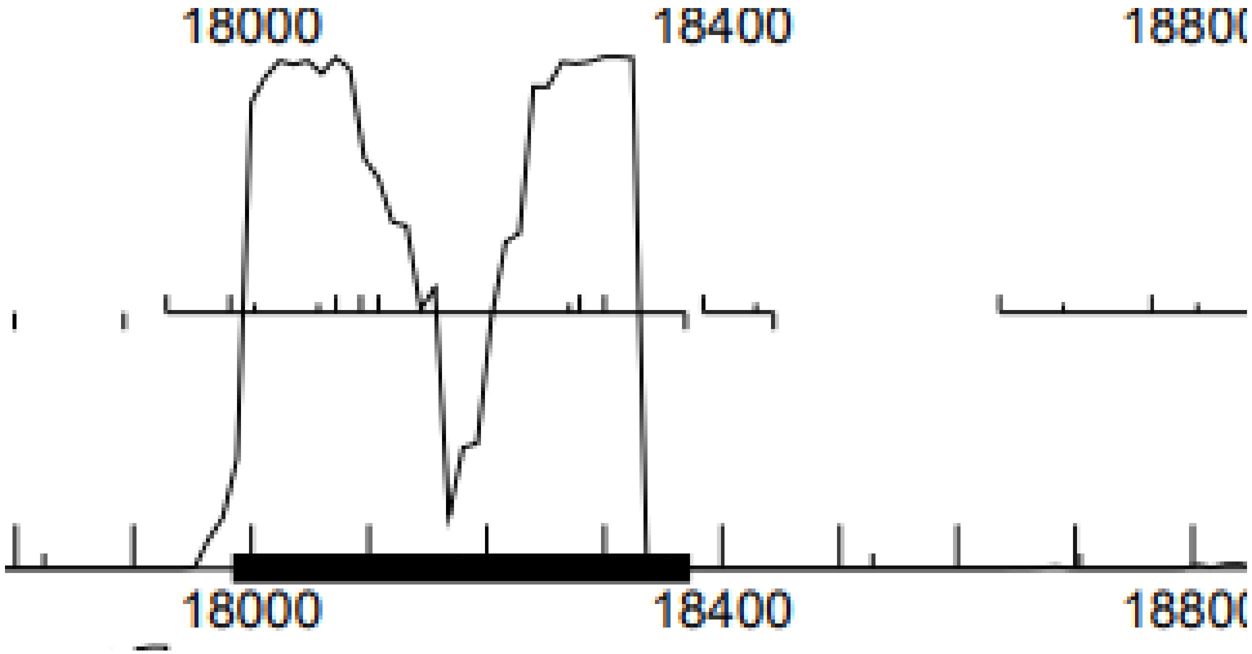
- Start: 49 @ 17984 has 26 MA’s



GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Start site: 17984

Coding potential is cut off



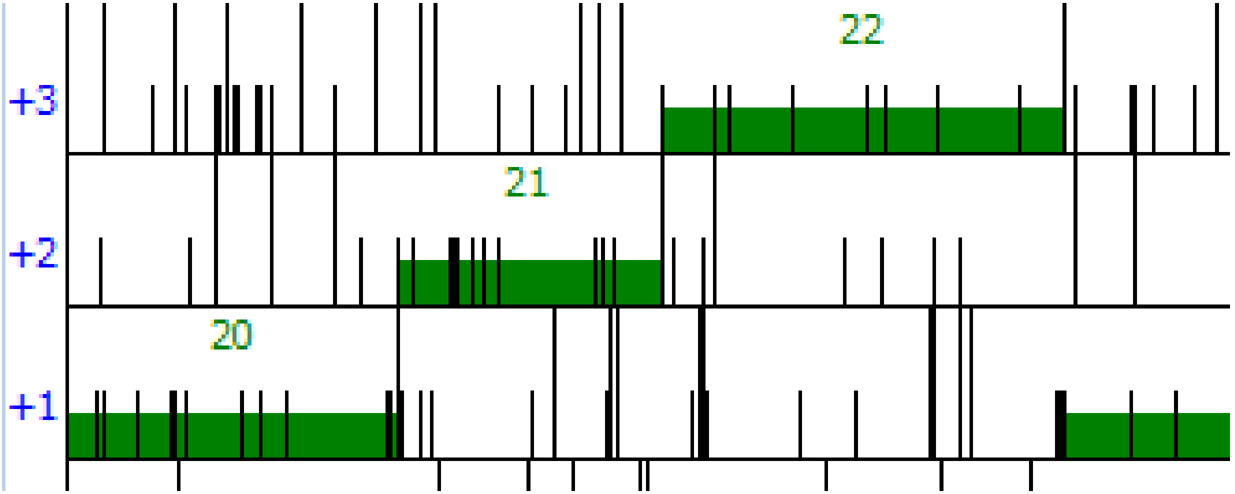
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Start site: 17984

Overlap: 1

Previous feature ends at

17984



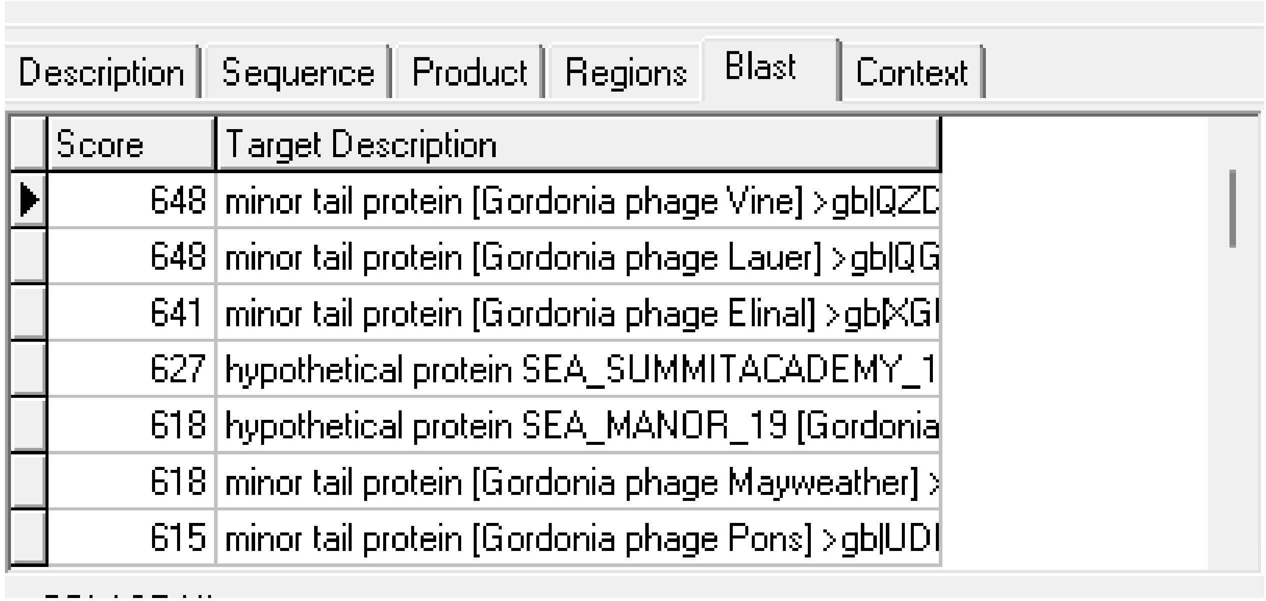
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	17984
GeneMark	Both Glimmer and GeneMark call it
Coding potential	Includes some cp
RBS	Z value: 2.754 Final score: -4.678
BLAST	24 1:1 alignments
Starterator	26 MA's
Overlap	1

Start site is 17984, because both Glimmer and GeneMark call the same start site, the frame includes some coding potential, the z value is greater than 1, and the overlap is 1 which is ideal.

BLAST function evidence. What assigned functions do other highly similar genes have?

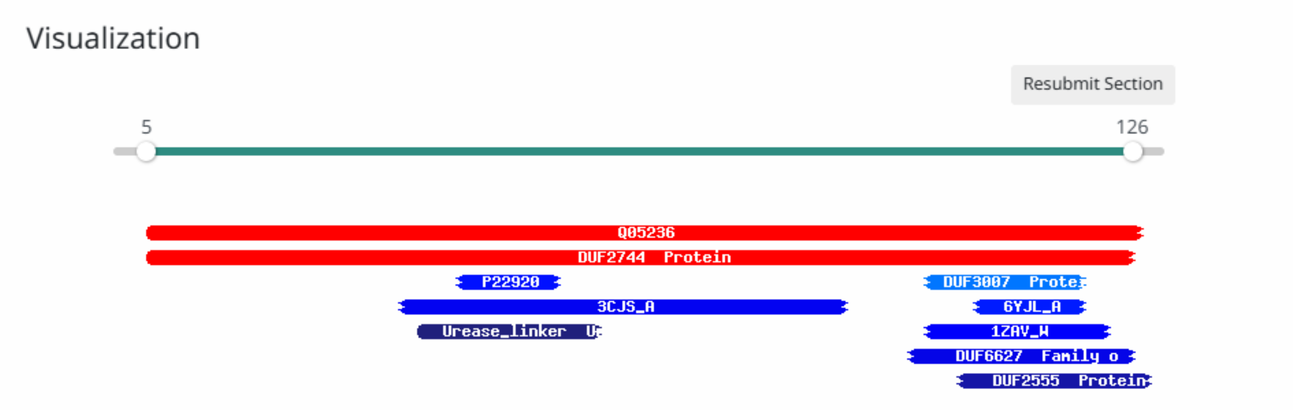
- 20 minor tail protein
- 5 hypothetical protein

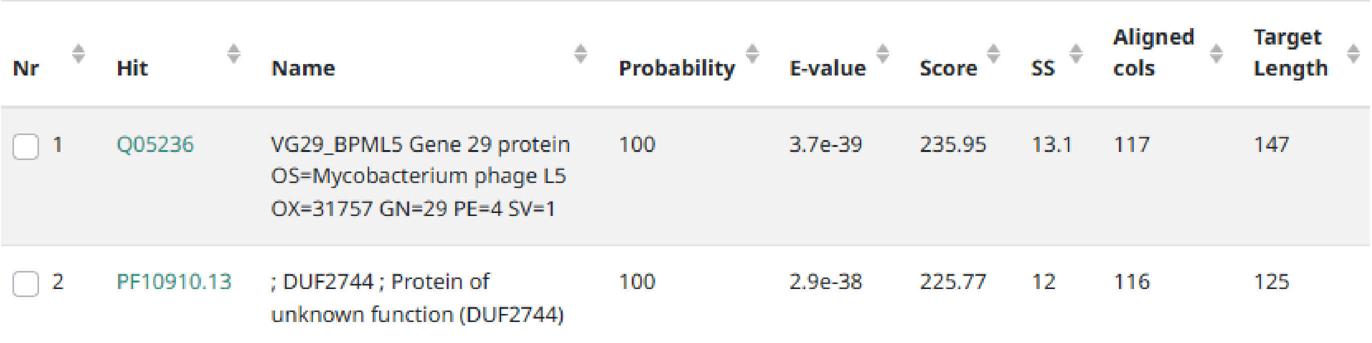


Score	Target Description
648	minor tail protein [Gordonia phage Vine] >gb|QZD
648	minor tail protein [Gordonia phage Lauer] >gb|QG
641	minor tail protein [Gordonia phage Elinal] >gb|XGI
627	hypothetical protein SEA_SUMMITACADEMY_1
618	hypothetical protein SEA_MANOR_19 [Gordonia
618	minor tail protein [Gordonia phage Mayweather] >
615	minor tail protein [Gordonia phage Pons] >gb|UDI

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- Top two hits based on protein of unknown function
- For it to have function minor tail protein, requires collagen-like or glycine-rich proteins which these hits do not have

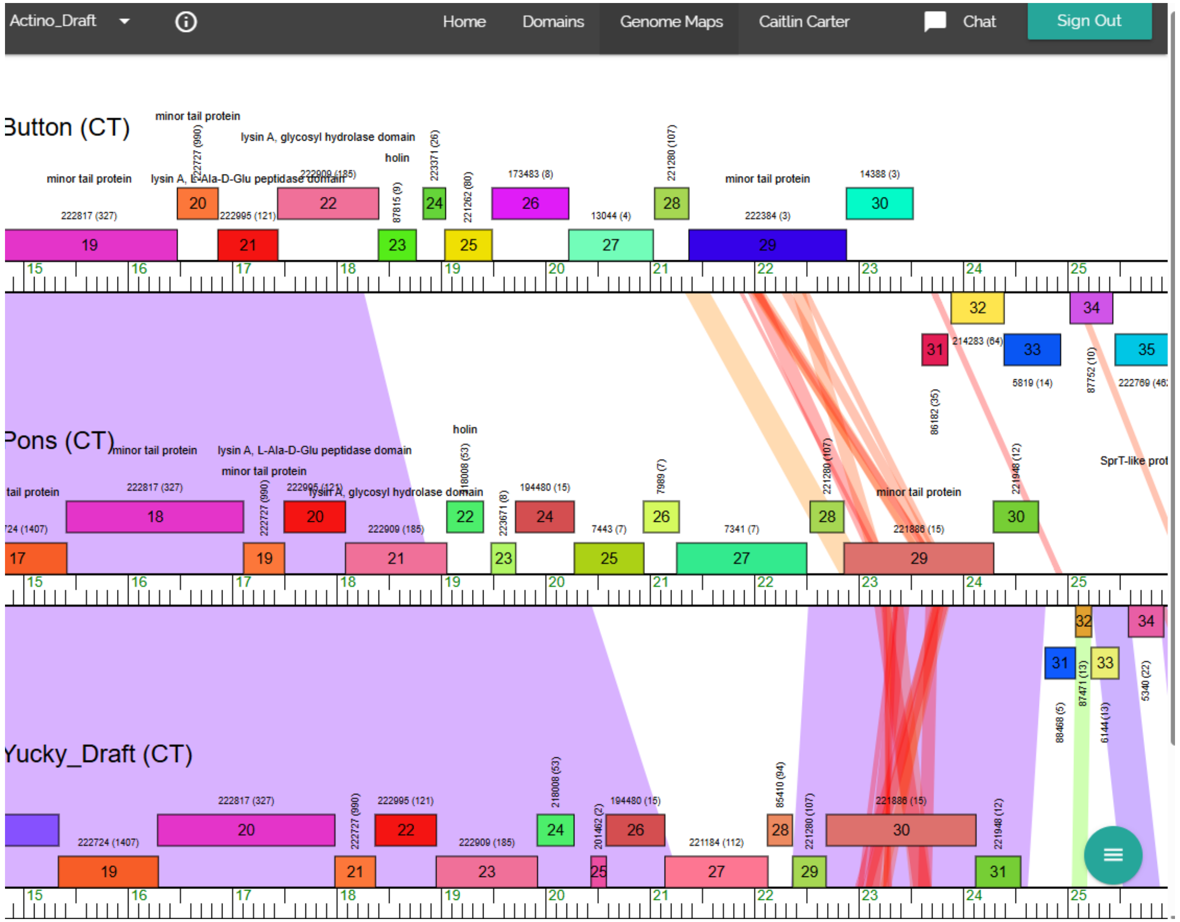




Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	Q05236	VG29_BPML5 Gene 29 protein OS=Mycobacterium phage L5 OX=31757 GN=29 PE=4 SV=1	100	3.7e-39	235.95	13.1	117	147
□ 2	PF10910.13	; DUF2744 ; Protein of unknown function (DUF2744)	100	2.9e-38	225.77	12	116	125

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Yucky feature 21 conserved domain: DUF2744 function: none
- Button feature 20 conserved domain: DUF2744 function: minor tail protein
- Pons feature 19 conserved domain: DUF2744 function: minor tail protein



Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function is minor tail protein because the genes around feature 21 all have the function minor tail protein. Call minor tail protein based on synteny.

Feature 20 – Stop 18957

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

What is the autoannotated start?

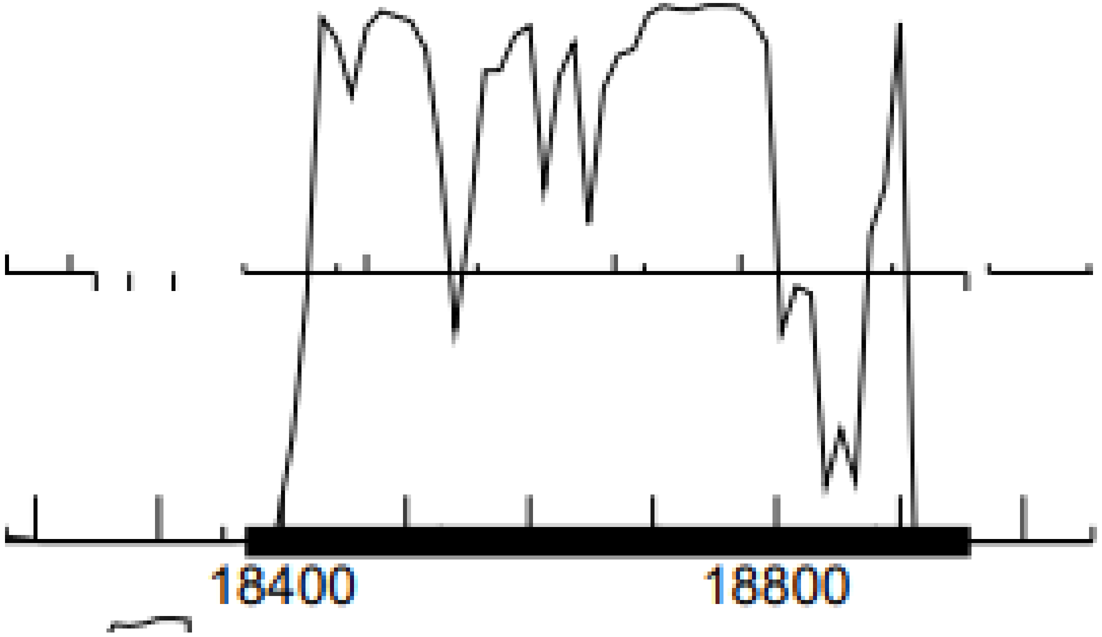
Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

- Feature 20
- Stop site: 18957
- Both Glimmer and GeneMark call the same start site
- Autoannotated start: 18370
- Overlap: 4

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

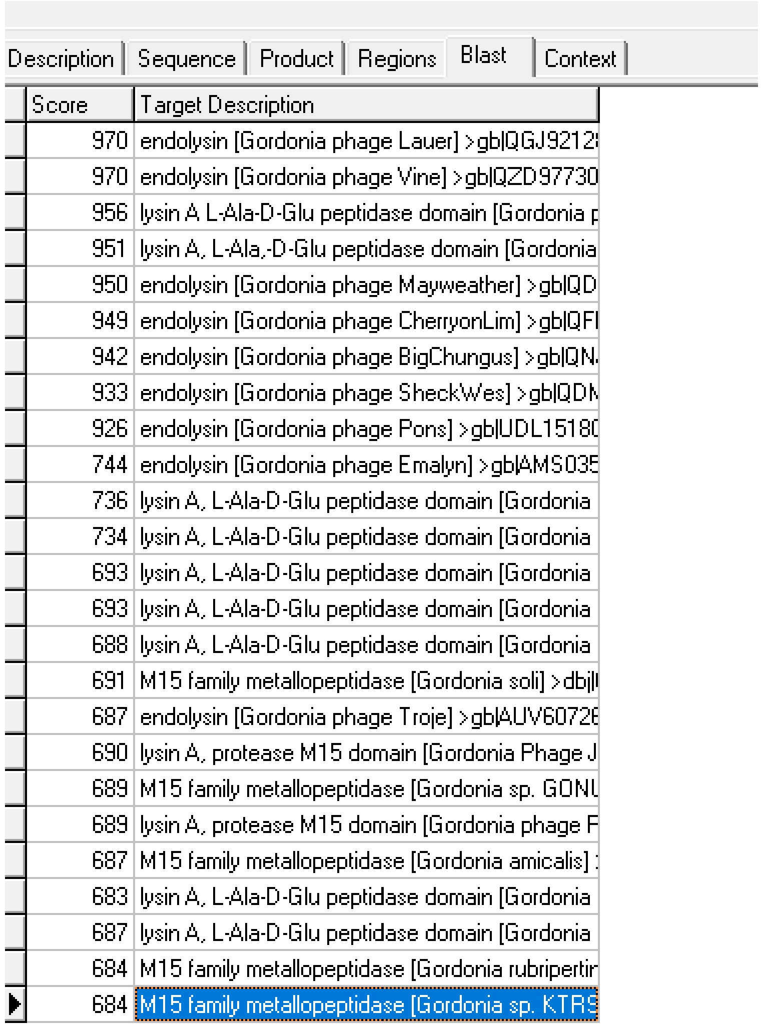
Start 18370

- Reading frame 1
- Includes all coding potential



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 25 highly similar genes



Score	Target Description
970	endolysin [Gordonia phage Lauer] >gb|QGJ9212
970	endolysin [Gordonia phage Vine] >gb|QZD97730
956	lysin A L-Ala-D-Glu peptidase domain [Gordonia p
951	lysin A, L-Ala,-D-Glu peptidase domain [Gordonia
950	endolysin [Gordonia phage Mayweather] >gb|QD
949	endolysin [Gordonia phage CherryonLim] >gb|QF
942	endolysin [Gordonia phage BigChungus] >gb|QN
933	endolysin [Gordonia phage SheckWes] >gb|QD
926	endolysin [Gordonia phage Pons] >gb|UDL1518
744	endolysin [Gordonia phage Emalyn] >gb|AMS035
736	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
734	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
693	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
693	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
688	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
691	M15 family metallopeptidase [Gordonia soli] >dbj|
687	endolysin [Gordonia phage Troje] >gb|AUV6072
690	lysin A, protease M15 domain [Gordonia Phage J
689	M15 family metallopeptidase [Gordonia sp. GON
689	lysin A, protease M15 domain [Gordonia phage F
687	M15 family metallopeptidase [Gordonia amicalis]
683	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
687	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
684	M15 family metallopeptidase [Gordonia rubripertir
684	M15 family metallopeptidase [Gordonia sp. KTRS

Answer: Is it a gene? Give evidence why you think this is a gene or not.

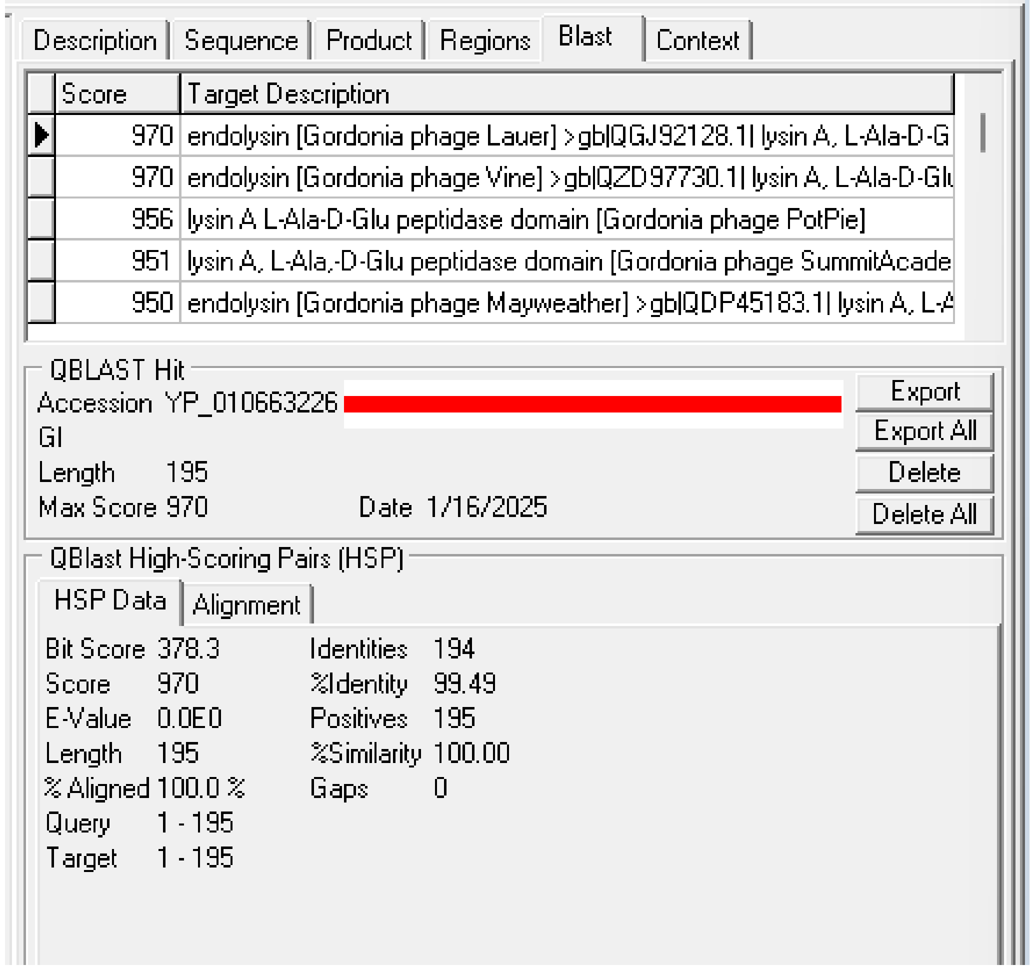
- Yes, it is a gene because both Glimmer and GeneMark call it at the same start site, the frame includes all coding potential, and there are 25 other highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

Start 18370

- **11 1:1 alignments**

- Lauer
- Vine
- PotPie
- SummitAcademy
- Mayweather
- CherryonLim
- BigChungus
- SheckWes
- Pons
- Emalyn
- AikoCarson



Score	Target Description
970	endolysin [Gordonia phage Lauer] >gb|QGJ92128.1| lysin A, L-Ala-D-G
970	endolysin [Gordonia phage Vine] >gb|QZD97730.1| lysin A, L-Ala-D-Glu
956	lysin A L-Ala-D-Glu peptidase domain [Gordonia phage PotPie]
951	lysin A, L-Ala,-D-Glu peptidase domain [Gordonia phage SummitAcade
950	endolysin [Gordonia phage Mayweather] >gb|QDP45183.1| lysin A, L-A

QBLAST Hit
Accession YP_010663226
GI
Length 195
Max Score 970
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

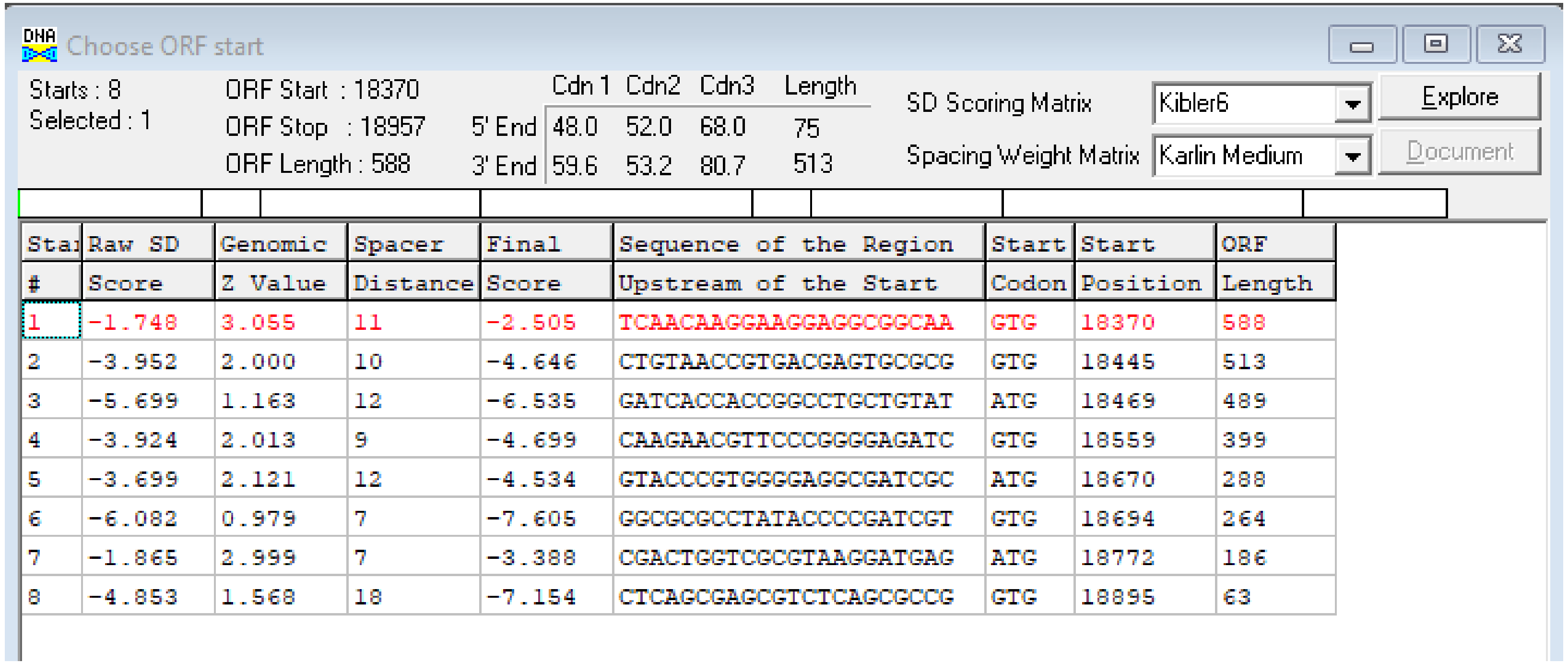
Bit Score	378.3	Identities	194
Score	970	%Identity	99.49
E-Value	0.0E0	Positives	195
Length	195	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 195		
Target	1 - 195		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Start site: 18370

Z value: 3.055

Final score: -2.505



Starts : 8
Selected : 1

ORF Start : 18370
ORF Stop : 18957
ORF Length : 588

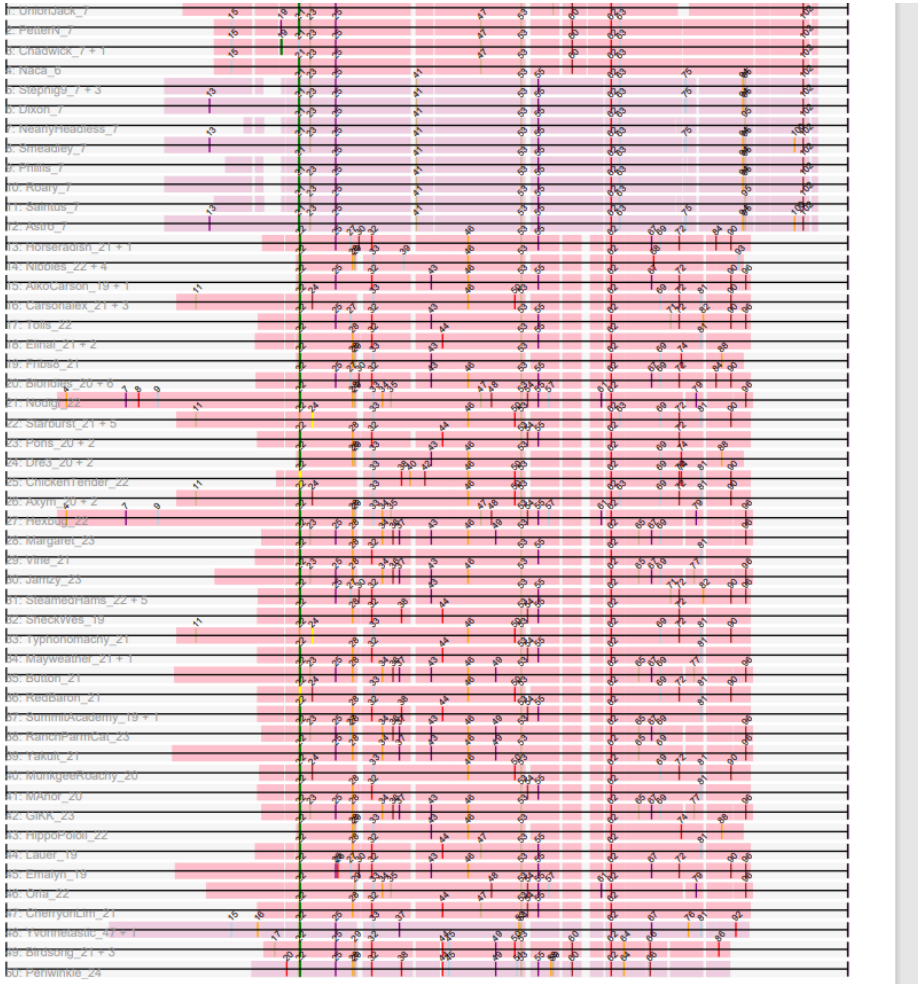
	Cdn 1	Cdn2	Cdn3	Length
5' End	48.0	52.0	68.0	75
3' End	59.6	53.2	80.7	513

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-1.748	3.055	11	-2.505	TCAACAAGGAAGGAGGCGGCAA	GTG	18370	588
2	-3.952	2.000	10	-4.646	CTGTAACCGTGACGAGTGCGCG	GTG	18445	513
3	-5.699	1.163	12	-6.535	GATCACCACCGGCCTGCTGTAT	ATG	18469	489
4	-3.924	2.013	9	-4.699	CAAGAACGTTCCCGGGGAGATC	GTG	18559	399
5	-3.699	2.121	12	-4.534	GTACCCGTGGGGAGGCGATCGC	ATG	18670	288
6	-6.082	0.979	7	-7.605	GGCGCGCCTATACCCCGATCGT	GTG	18694	264
7	-1.865	2.999	7	-3.388	CGACTGGTCGCGTAAGGATGAG	ATG	18772	186
8	-4.853	1.568	18	-7.154	CTCAGCGAGCGTCTCAGCGCCG	GTG	18895	63

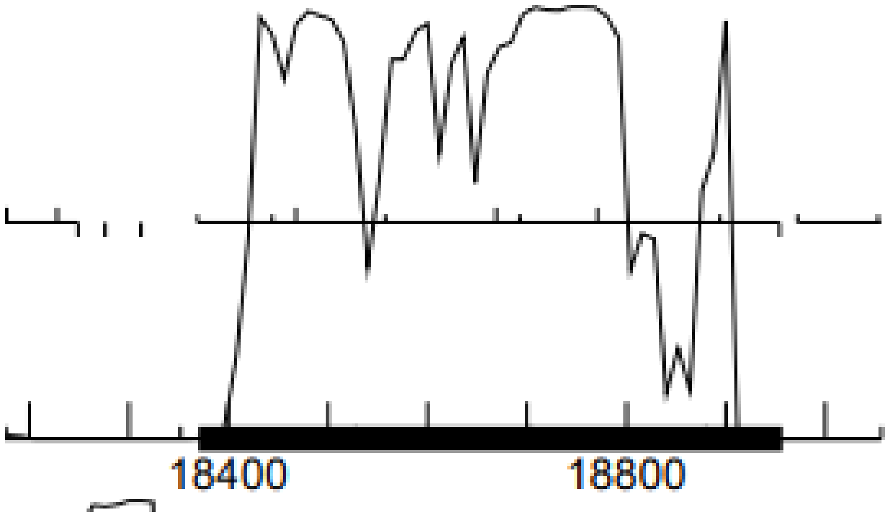
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Start: 22 @18370 has 76 MA’s



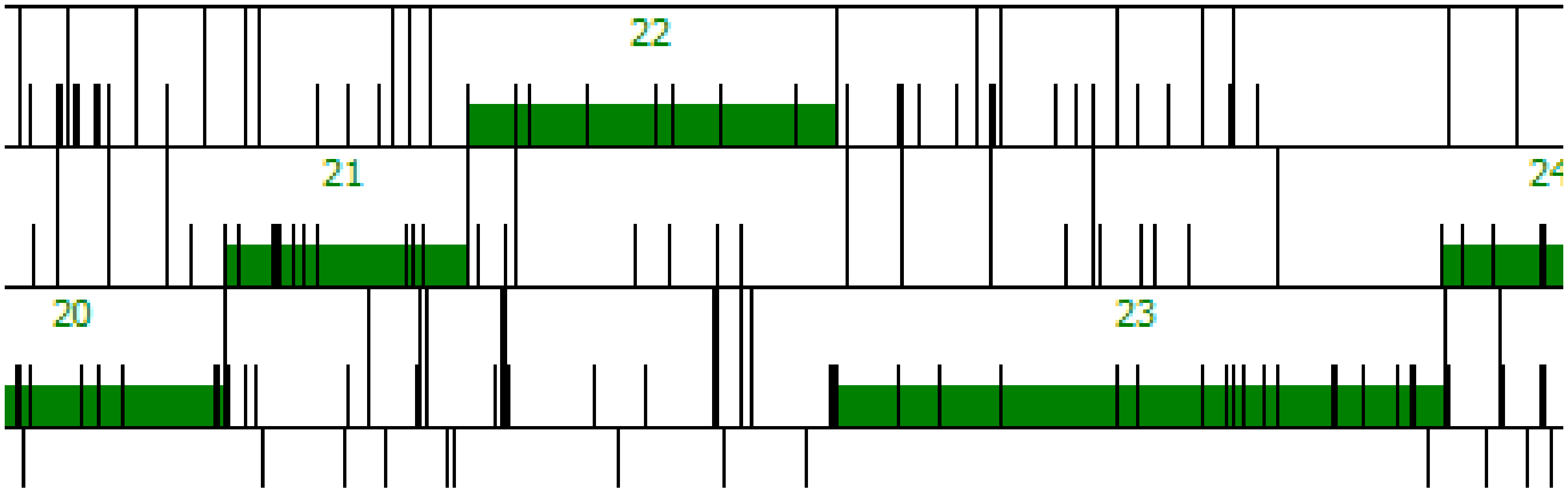
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Start site 18370
- Includes all coding potential
- None of the coding potential is cut off



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Start site 18370 - previous end sight 18373
- Overlap: 4



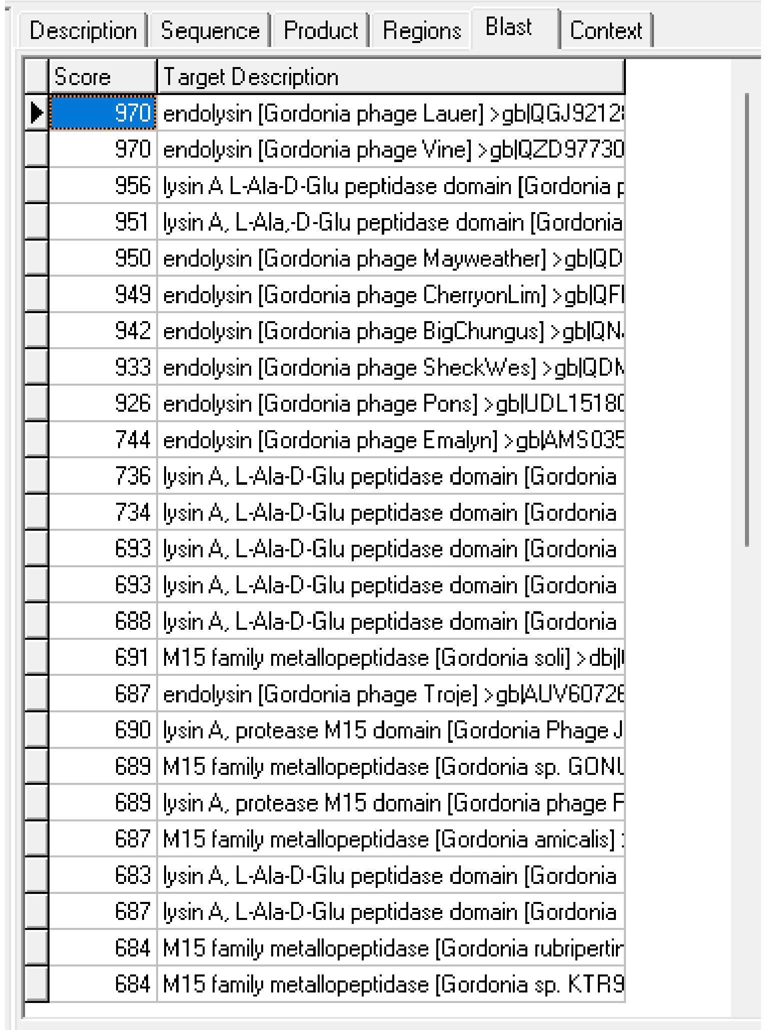
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	18370
GeneMark	Called by both Glimmer & GeneMark
Coding potential	Includes all cp
RBS	Z value: 3.055 Final score: -2.505
BLAST	11 1:1 alignments
Starterator	76 MA’s
Overlap	4

The start site is 18370 because it is called by both Glimmer and GeneMark, the frame includes all coding potential, the Z value is greater than 1, and it has an overlap of 4 which is ideal.

BLAST function evidence. What assigned functions do other highly similar genes have?

- 9 endolysin
- 11 lysin A
- 5 M15 family metallopeptidase

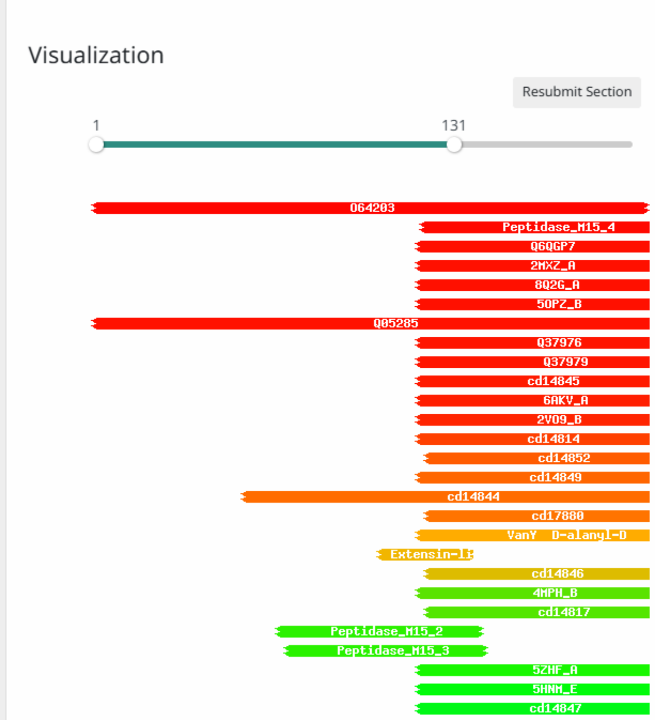


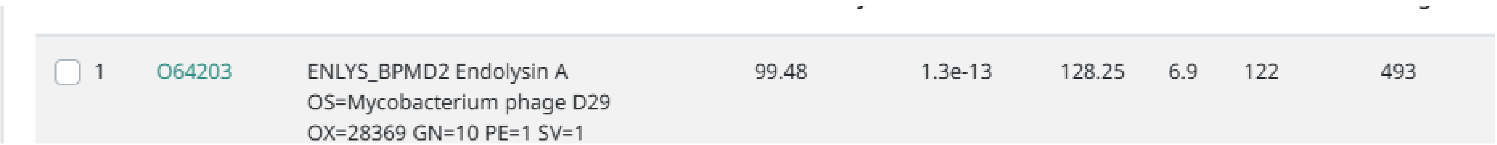
Score	Target Description
970	endolysin [Gordonia phage Lauer] >gb|QGJ9212
970	endolysin [Gordonia phage Vine] >gb|QZD97730
956	lysin A L-Ala-D-Glu peptidase domain [Gordonia p
951	lysin A, L-Ala,-D-Glu peptidase domain [Gordonia
950	endolysin [Gordonia phage Mayweather] >gb|QD
949	endolysin [Gordonia phage CherryonLim] >gb|QFI
942	endolysin [Gordonia phage BigChungus] >gb|QN
933	endolysin [Gordonia phage SheckWes] >gb|QD
926	endolysin [Gordonia phage Pons] >gb|UDL1518
744	endolysin [Gordonia phage Emalyn] >gb|AMS035
736	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
734	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
693	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
693	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
688	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
691	M15 family metallopeptidase [Gordonia soli] >dbj|
687	endolysin [Gordonia phage Troje] >gb|AUV6072
690	lysin A, protease M15 domain [Gordonia Phage J
689	M15 family metallopeptidase [Gordonia sp. GON
689	lysin A, protease M15 domain [Gordonia phage F
687	M15 family metallopeptidase [Gordonia amicalis]
683	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
687	lysin A, L-Ala-D-Glu peptidase domain [Gordonia
684	M15 family metallopeptidase [Gordonia rubripertir
684	M15 family metallopeptidase [Gordonia sp. KTR9

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

Function list does not include function M15 family metallopeptidase.

It is also not endolysin A as the phage does infect Mycobacterium, so it is lysin A with conserved domain L-Ala-D-Glu_peptidase_

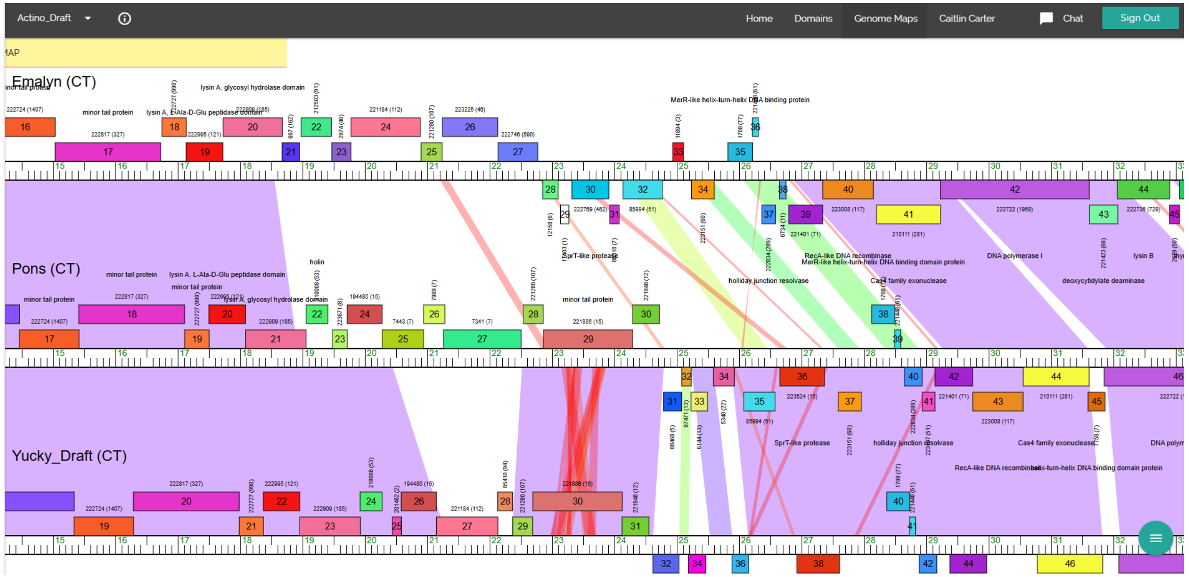




☐	1	O64203	ENLYS_BPMD2 Endolysin A OS=Mycobacterium phage D29 OX=28369 GN=10 PE=1 SV=1	99.48	1.3e-13	128.25	6.9	122	493

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Yucky feature 22 conserved domain: L-Ala-D-Glu_peptidase_, Peptidase_M15_4 function: none
- Pons feature 20 conserved domain: L-Ala-D-Glu_peptidase_, Peptidase_M15_4 function: lysin A, L-Ala-D-Glu peptidase domain
- Emalyn feature 19 conserved domain: L-Ala-D-Glu_peptidase_, Peptidase_M15_4 function: lysin A, L-Ala-D-Glu peptidase domain


Actino_Draft
Home
Domains
Genome Maps
Caitlin Carter
Chat
Sign Out
MAP
Emalyn (CT)
minor tail protein
lysin A, glycosyl hydrolase domain
lysin A, L-Ala-D-Glu peptidase domain
MerR-like helix-turn-helix DNA binding protein
Pons (CT)
holin
SprT-like protease
holliday junction resolvase
RecA-like DNA recombinase
MerR-like helix-turn-helix DNA binding protein
Cas4 family exonuclease
DNA polymerase I
lysin B
deoxycytidylate deaminase
Yucky_Draft (CT)
SprT-like protease
holliday junction resolvase
Cas4 family exonuclease
RecA-like DNA recombinase
helix-turn-helix DNA binding protein
DNA polym

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- None

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function is lysin A with conserved domain L-Ala-D-Glu_peptidase_ because it has the highest amount of hits in BLAST evidence, it is the function for highly similar genes Pons and Emalyn on Phamerator

Feature 21 – Stop 19925

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

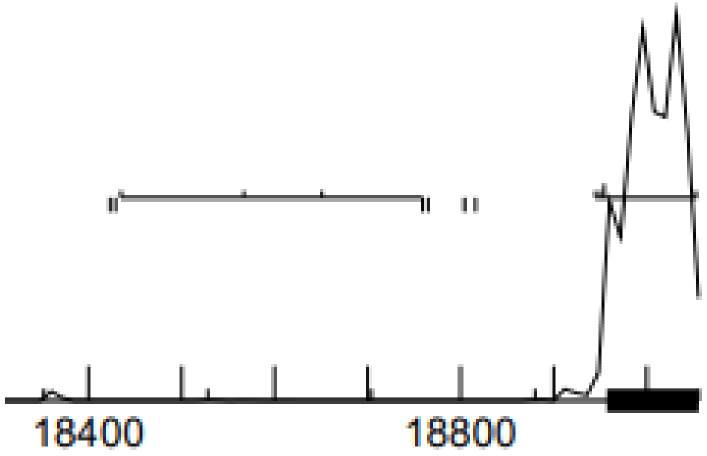
What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

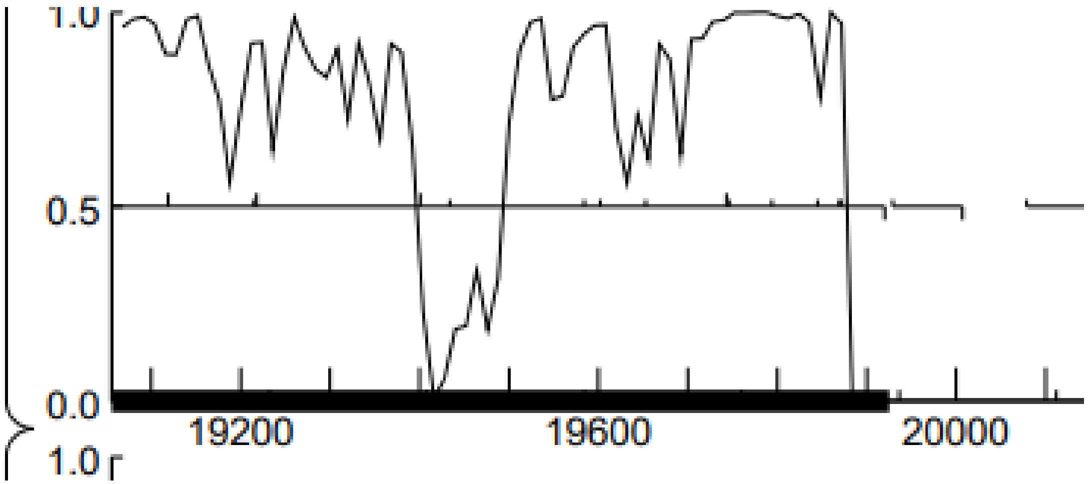
- Feature: 21
- Stop site: 19925
- Both Glimmer and GeneMark call it but at different start sites
- Glimmer call @bp 18954
- GeneMark calls start at 18957
- Overlap: 4

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

Start 18954
Some of the coding potential is cut off before the start site. Located in frame 3.



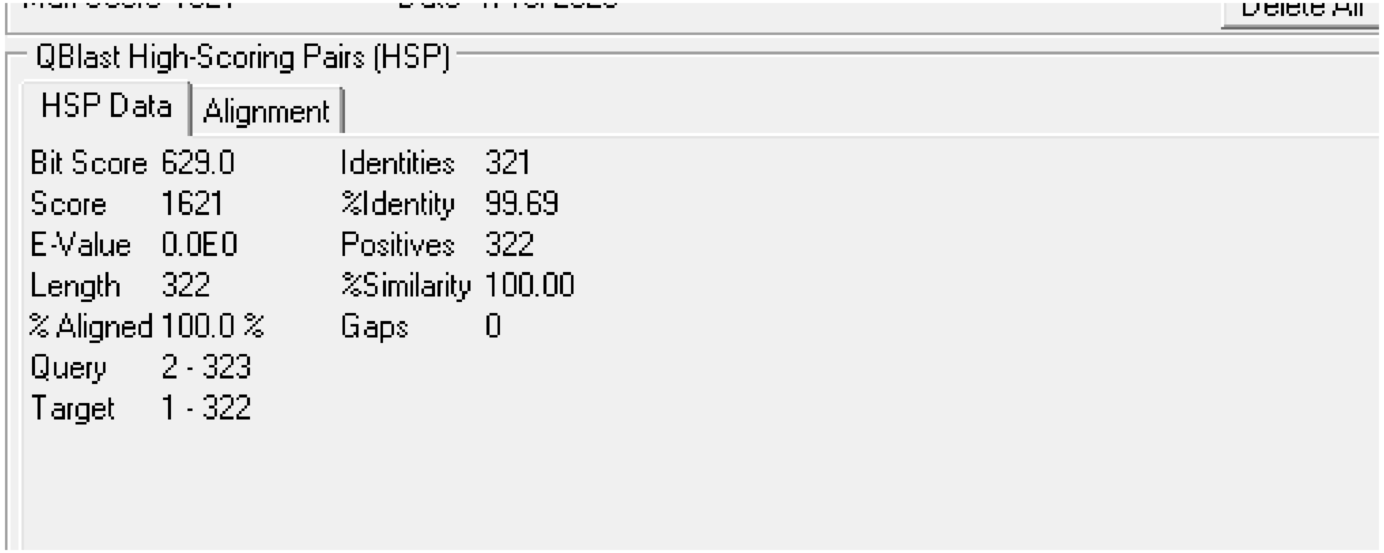
Start 18957
Some of the coding potential is cut off before the start site. Located in frame 3.



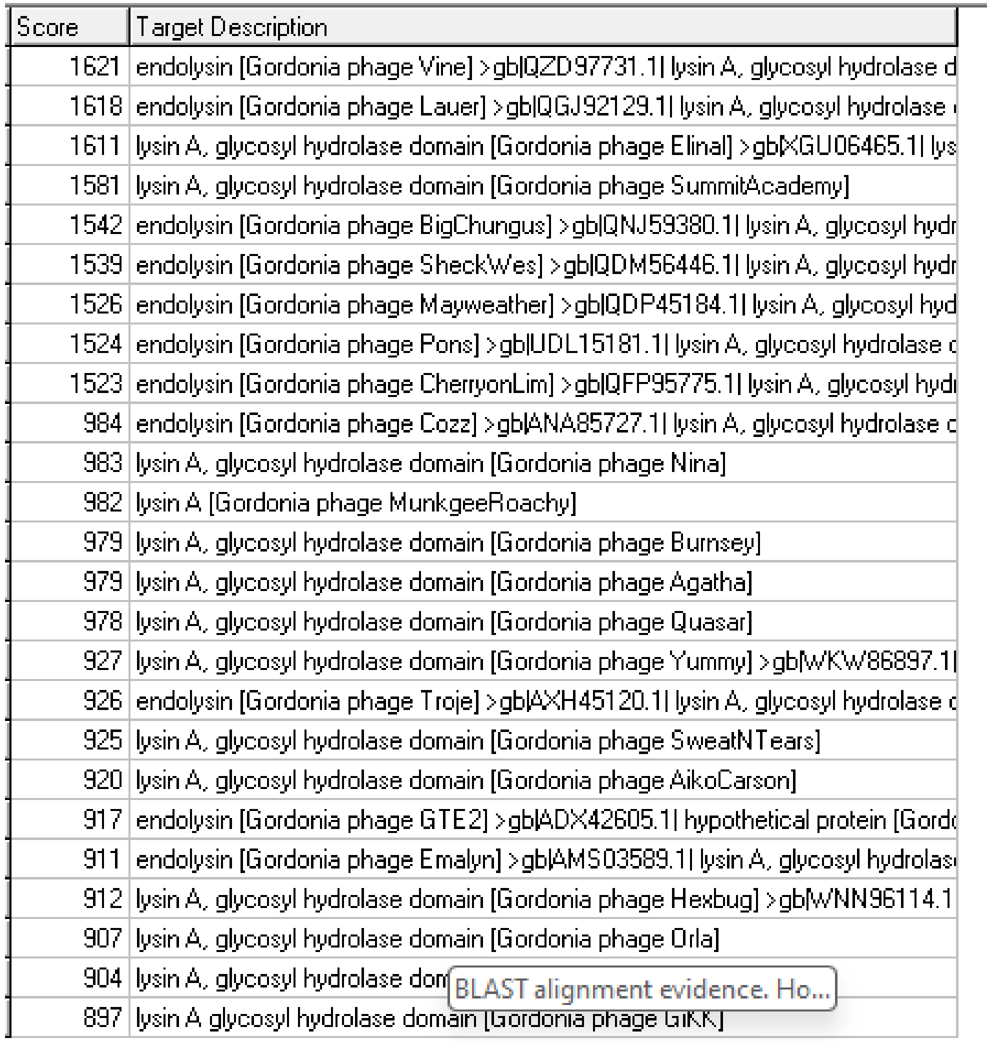
BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 25 highly similar genes

25 E-value 0.0E0



Bit Score	629.0	Identities	321
Score	1621	%Identity	99.69
E-Value	0.0E0	Positives	322
Length	322	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	2 - 323		
Target	1 - 322		



Score	Target Description
1621	endolysin [Gordonia phage Vine] >gb|QZD97731.1| lysin A, glycosyl hydrolase d
1618	endolysin [Gordonia phage Lauer] >gb|QGJ92129.1| lysin A, glycosyl hydrolase
1611	lysin A, glycosyl hydrolase domain [Gordonia phage Elinal] >gb|XGU06465.1| lys
1581	lysin A, glycosyl hydrolase domain [Gordonia phage SummitAcademy]
1542	endolysin [Gordonia phage BigChungus] >gb|QNJ59380.1| lysin A, glycosyl hydr
1539	endolysin [Gordonia phage SheckWes] >gb|QDM56446.1| lysin A, glycosyl hydr
1526	endolysin [Gordonia phage Mayweather] >gb|QDP45184.1| lysin A, glycosyl hyd
1524	endolysin [Gordonia phage Pons] >gb|UDL15181.1| lysin A, glycosyl hydrolase c
1523	endolysin [Gordonia phage CherryonLim] >gb|QFP95775.1| lysin A, glycosyl hydr
984	endolysin [Gordonia phage Cozz] >gb|ANA85727.1| lysin A, glycosyl hydrolase c
983	lysin A, glycosyl hydrolase domain [Gordonia phage Nina]
982	lysin A [Gordonia phage MunkgeeRoachy]
979	lysin A, glycosyl hydrolase domain [Gordonia phage Burnsey]
979	lysin A, glycosyl hydrolase domain [Gordonia phage Agatha]
978	lysin A, glycosyl hydrolase domain [Gordonia phage Quasar]
927	lysin A, glycosyl hydrolase domain [Gordonia phage Yummy] >gb|WKW86897.1|
926	endolysin [Gordonia phage Troje] >gb|AXH45120.1| lysin A, glycosyl hydrolase c
925	lysin A, glycosyl hydrolase domain [Gordonia phage SweatNTears]
920	lysin A, glycosyl hydrolase domain [Gordonia phage AikoCarson]
917	endolysin [Gordonia phage GTE2] >gb|ADX42605.1| hypothetical protein [Gordo
911	endolysin [Gordonia phage Emalyn] >gb|AMS03589.1| lysin A, glycosyl hydrolas
912	lysin A, glycosyl hydrolase domain [Gordonia phage Hexbug] >gb|WNN96114.1
907	lysin A, glycosyl hydrolase domain [Gordonia phage Orla]
904	lysin A, glycosyl hydrolase dom
897	lysin A glycosyl hydrolase domain [Gordonia phage GiKK]

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, it is a gene because both Glimmer and GeneMark call it, includes a large majority of coding potential, and there are 25 other highly similar genes with an E value of 0.0E0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

Start 18948: (NCBI)

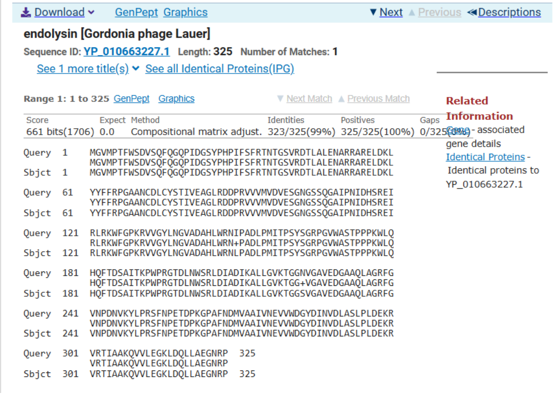
2 1:1 alignments

Start 18954: (DNAM)

2 1:1 alignments

Start 18957: (NCBI)

6 1:1 alignments



endolysin [Gordonia phage Lauer]

Sequence ID: YP_010663227.1 Length: 325 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

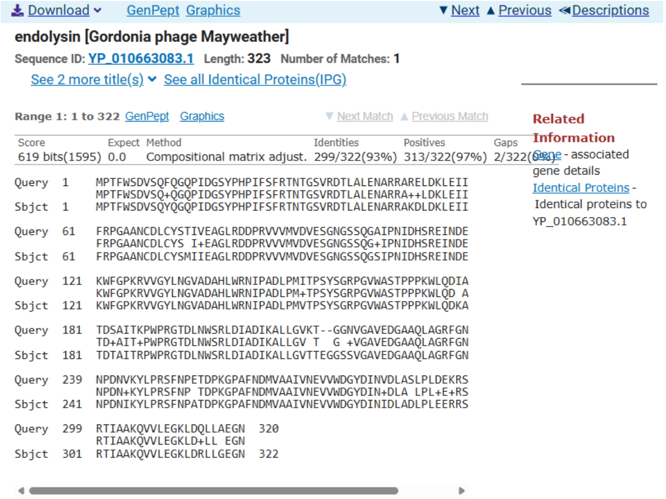
Range 1: 1 to 325 GenPept Graphics

```
Score           Expect Method                        Identities     Positives       Gaps
661 bits(1706)  0.0    Compositional matrix adjust.  323/325(99%)   325/325(100%)   0/325(0%)

Query  1    MGVMPTFWSDVSQFQGQPIDGSYPHPIFSFRTNTGSVRDTLALENARRARELDKL
            MGVMPTFWSDVSQFQGQPIDGSYPHPIFSFRTNTGSVRDTLALENARRARELDKL
Sbjct  1    MGVMPTFWSDVSQFQGQPIDGSYPHPIFSFRTNTGSVRDTLALENARRARELDKL

Query  61   YYFFRPGAANCDLCYSTIVEAGLRDDPRVVVMVDVESGNGSSQGAIPNIDHSREI
            YYFFRPGAANCDLCYSTIVEAGLRDDPRVVVMVDVESGNGSSQGAIPNIDHSREI
Sbjct  61   YYFFRPGAANCDLCYSTIVEAGLRDDPRVVVMVDVESGNGSSQGAIPNIDHSREI

Query  121  RLRKWFGPKRVVGYLNGVADAHLWRNIPADLPMITPSYSGRPGVWASTPPPKWLQ
            RLRKWFGPKRVVGYLNGVADAHLWRN+PADLPMITPSYSGRPGVWASTPPPKWLQ
Sbjct  121  RLRKWFGPKRVVGYLNGVADAHLWRNLPADLPMITPSYSGRPGVWASTPPPKWLQ

Query  181  HQFTDSAITKPWPRGTDLNWSRLDIADIKALLGVKTGGNVGAVEDGAAQLAGRFG
            HQFTDSAITKPWPRGTDLNWSRLDIADIKALLGVKTGG+VGAVEDGAAQLAGRFG
Sbjct  181  HQFTDSAITKPWPRGTDLNWSRLDIADIKALLGVKTGGSVGAVEDGAAQLAGRFG

Query  241  VNPDNVKYLPRSFNPETDPKGPAFNDMVAAIVNEVVWDGYDINVDLASLPLDEKR
            VNPDNVKYLPRSFNPETDPKGPAFNDMVAAIVNEVVWDGYDINVDLASLPLDEKR
Sbjct  241  VNPDNVKYLPRSFNPETDPKGPAFNDMVAAIVNEVVWDGYDINVDLASLPLDEKR

Query  301  VRTIAAKQVVLEGKLDQLLAEGNRP  325
            VRTIAAKQVVLEGKLDQLLAEGNRP
Sbjct  301  VRTIAAKQVVLEGKLDQLLAEGNRP  325
```

Related Information

Gene - associated gene details

Identical Proteins - Identical proteins to YP_010663227.1



endolysin [Gordonia phage Mayweather]

Sequence ID: YP_010663083.1 Length: 323 Number of Matches: 1

See 2 more title(s) See all Identical Proteins(IPG)

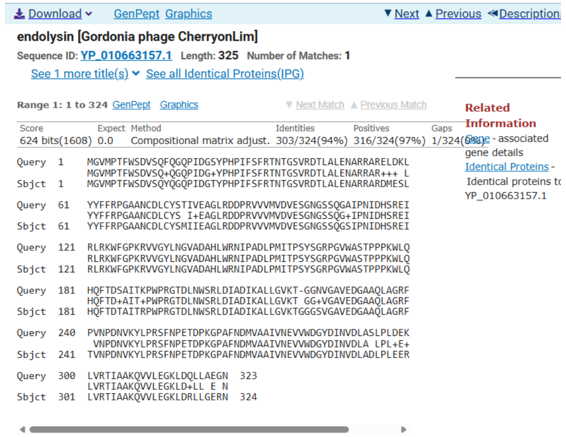
Range 1: 1 to 322 GenPept Graphics

```
Score           Expect Method                        Identities     Positives      Gaps
619 bits(1595)  0.0    Compositional matrix adjust.  299/322(93%)   313/322(97%)   2/322(0%)

Query  1    MPTFWSDVSQFQGQPIDGSYPHPIFSFRTNTGSVRDTLALENARRARELDKLEII
            MPTFWSDVSQ+QGQPIDGSYPHPIFSFRTNTGSVRDTLALENARRA++LDKLEII
Sbjct  1    MPTFWSDVSQYQGQPIDGSYPHPIFSFRTNTGSVRDTLALENARRAKDLDKLEII

Query  61   FRPGAANCDLCYSTIVEAGLRDDPRVVVMVDVESGNGSSQGAIPNIDHSREINDE
            FRPGAANCDLCYS I+EAGLRDDPRVVVMVDVESGNGSSQG+IPNIDHSREINDE
Sbjct  61   FRPGAANCDLCYSMIIEAGLRDDPRVVVMVDVESGNGSSQGSIPNIDHSREINDE

Query  121  KWFGPKRVVGYLNGVADAHLWRNIPADLPMITPSYSGRPGVWASTPPPKWLQDIA
            KWFGPKRVVGYLNGVADAHLWRNIPADLPM+TPSYSGRPGVWASTPPPKWLQD A
Sbjct  121  KWFGPKRVVGYLNGVADAHLWRNIPADLPMVTPSYSGRPGVWASTPPPKWLQDKA

Query  181  TDSAITKPWPRGTDLNWSRLDIADIKALLGVKT--GGNVGAVEDGAAQLAGRFGN
            TD+AIT+PWPRGTDLNWSRLDIADIKALLGV T  G +VGAVEDGAAQLAGRFGN
Sbjct  181  TDTAITRPWPRGTDLNWSRLDIADIKALLGVTTEGGSSVGAVEDGAAQLAGRFGN

Query  239  NPDNVKYLPRSFNPETDPKGPAFNDMVAAIVNEVVWDGYDINVDLASLPLDEKRS
            NPDN+KYLPRSFNP TDPKGPAFNDMVAAIVNEVVWDGYDIN+DLA LPL+E+RS
Sbjct  241  NPDNIKYLPRSFNPATDPKGPAFNDMVAAIVNEVVWDGYDINIDLADLPLEERRS

Query  299  RTIAAKQVVLEGKLDQLLAEGN  320
            RTIAAKQVVLEGKLD+LL EGN
Sbjct  301  RTIAAKQVVLEGKLDRLLGEGN  322
```

Related Information

Gene - associated gene details

Identical Proteins - Identical proteins to YP_010663083.1



endolysin [Gordonia phage CherryonLim]

Sequence ID: YP_010663157.1 Length: 325 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

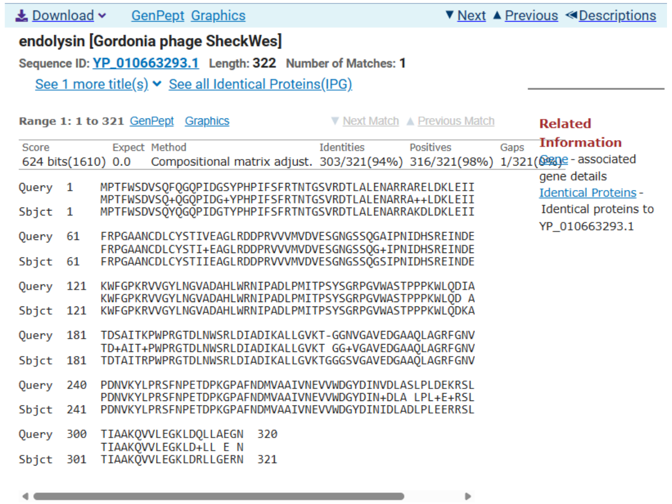
Range 1: 1 to 324 GenPept Graphics

```
Score           Expect Method                        Identities     Positives      Gaps
624 bits(1608)  0.0    Compositional matrix adjust.  303/324(94%)   316/324(97%)   1/324(0%)

Query  1    MGVMPTFWSDVSQFQGQPIDGSYPHPIFSFRTNTGSVRDTLALENARRARELDKL
            MGVMPTFWSDVSQ+QGQPIDG+YPHPIFSFRTNTGSVRDTLALENARRAR+++ L
Sbjct  1    MGVMPTFWSDVSQYQGQPIDGTYPHPIFSFRTNTGSVRDTLALENARRARDMESL

Query  61   YYFFRPGAANCDLCYSTIVEAGLRDDPRVVVMVDVESGNGSSQGAIPNIDHSREI
            YYFFRPGAANCDLCYS I+EAGLRDDPRVVVMVDVESGNGSSQG+IPNIDHSREI
Sbjct  61   YYFFRPGAANCDLCYSMIIEAGLRDDPRVVVMVDVESGNGSSQGSIPNIDHSREI

Query  121  RLRKWFGPKRVVGYLNGVADAHLWRNIPADLPMITPSYSGRPGVWASTPPPKWLQ
            RLRKWFGPKRVVGYLNGVADAHLWRNIPADLPMITPSYSGRPGVWASTPPPKWLQ
Sbjct  121  RLRKWFGPKRVVGYLNGVADAHLWRNIPADLPMITPSYSGRPGVWASTPPPKWLQ

Query  181  HQFTDSAITKPWPRGTDLNWSRLDIADIKALLGVKT-GGNVGAVEDGAAQLAGRF
            HQFTD+AIT+PWPRGTDLNWSRLDIADIKALLGVKT GG+VGAVEDGAAQLAGRF
Sbjct  181  HQFTDTAITRPWPRGTDLNWSRLDIADIKALLGVKTGGGSVGAVEDGAAQLAGRF

Query  240  PVNPDNVKYLPRSFNPETDPKGPAFNDMVAAIVNEVVWDGYDINVDLASLPLDEK
             VNPDNVKYLPRSFNPETDPKGPAFNDMVAAIVNEVVWDGYDINVDLA LPL+E+
Sbjct  241  TVNPDNVKYLPRSFNPETDPKGPAFNDMVAAIVNEVVWDGYDINVDLADLPLEER

Query  300  LVRTIAAKQVVLEGKLDQLLAEGN  323
            LVRTIAAKQVVLEGKLD+LL E N
Sbjct  301  LVRTIAAKQVVLEGKLDRLLGERN  324
```

Related Information

Gene - associated gene details

Identical Proteins - Identical proteins to YP_010663157.1



endolysin [Gordonia phage SheckWes]

Sequence ID: YP_010663293.1 Length: 322 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 321 GenPept Graphics

```
Score           Expect Method                        Identities     Positives      Gaps
624 bits(1610)  0.0    Compositional matrix adjust.  303/321(94%)   316/321(98%)   1/321(0%)

Query  1    MPTFWSDVSQFQGQPIDGSYPHPIFSFRTNTGSVRDTLALENARRARELDKLEII
            MPTFWSDVSQ+QGQPIDG+YPHPIFSFRTNTGSVRDTLALENARRA++LDKLEII
Sbjct  1    MPTFWSDVSQYQGQPIDGTYPHPIFSFRTNTGSVRDTLALENARRAKDLDKLEII

Query  61   FRPGAANCDLCYSTIVEAGLRDDPRVVVMVDVESGNGSSQGAIPNIDHSREINDE
            FRPGAANCDLCYSTI+EAGLRDDPRVVVMVDVESGNGSSQG+IPNIDHSREINDE
Sbjct  61   FRPGAANCDLCYSTIIEAGLRDDPRVVVMVDVESGNGSSQGSIPNIDHSREINDE

Query  121  KWFGPKRVVGYLNGVADAHLWRNIPADLPMITPSYSGRPGVWASTPPPKWLQDIA
            KWFGPKRVVGYLNGVADAHLWRNIPADLPMITPSYSGRPGVWASTPPPKWLQD A
Sbjct  121  KWFGPKRVVGYLNGVADAHLWRNIPADLPMITPSYSGRPGVWASTPPPKWLQDKA

Query  181  TDSAITKPWPRGTDLNWSRLDIADIKALLGVKT-GGNVGAVEDGAAQLAGRFGNV
            TD+AIT+PWPRGTDLNWSRLDIADIKALLGVKT GG+VGAVEDGAAQLAGRFGNV
Sbjct  181  TDTAITRPWPRGTDLNWSRLDIADIKALLGVKTGGGSVGAVEDGAAQLAGRFGNV

Query  240  PDNVKYLPRSFNPETDPKGPAFNDMVAAIVNEVVWDGYDINVDLASLPLDEKRSL
            PDNVKYLPRSFNPETDPKGPAFNDMVAAIVNEVVWDGYDIN+DLA LPL+E+RSL
Sbjct  241  PDNVKYLPRSFNPETDPKGPAFNDMVAAIVNEVVWDGYDINIDLADLPLEERRSL

Query  300  TIAAKQVVLEGKLDQLLAEGN  320
            TIAAKQVVLEGKLD+LL E N
Sbjct  301  TIAAKQVVLEGKLDRLLGERN  321
```

Related Information

Gene - associated gene details

Identical Proteins - Identical proteins to YP_010663293.1

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

Start 18948

Z value: 2.555

Final score: -4.316

Start 18954

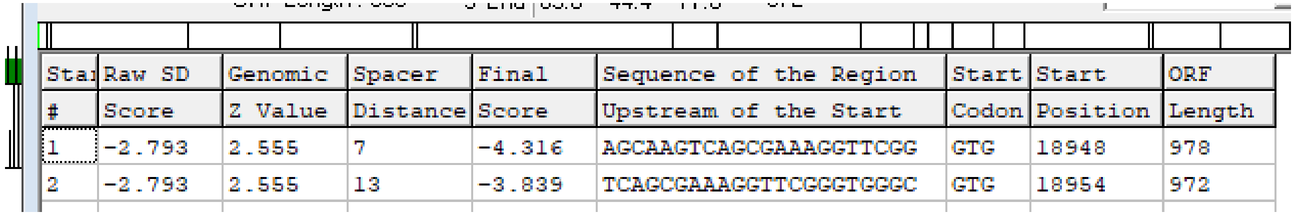
Z value: 2.555

Final score: -3.839

Start 18957

Z value: 2.555

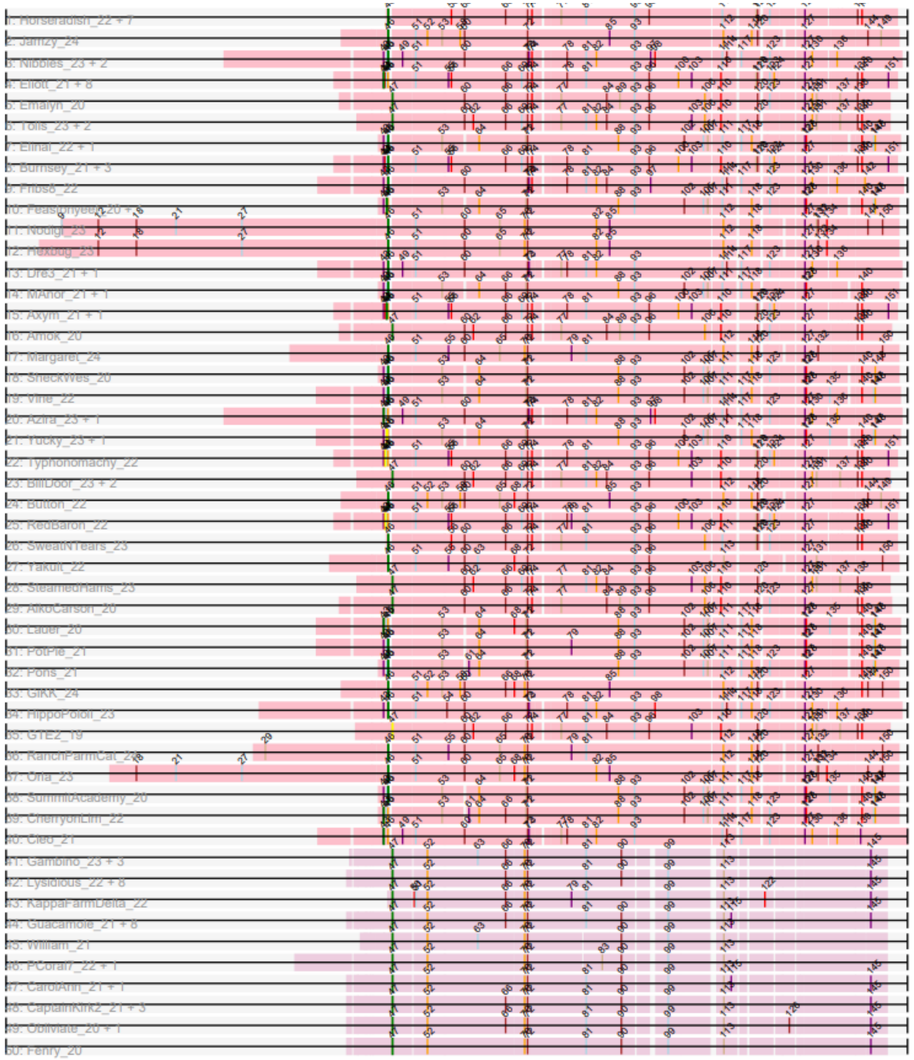
Final score: -4.589 *Preferred start*



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	-2.793	2.555	7	-4.316	AGCAAGTCAGCGAAAGGTTCGG	GTG	18948	978
2	-2.793	2.555	13	-3.839	TCAGCGAAAGGTTCGGGTGGGC	GTG	18954	972

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- 43 @18948 has 5 MA's
- 45 @18954 has 4 MA's
- 46 @18957 has 37 MA's



GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Start 18948

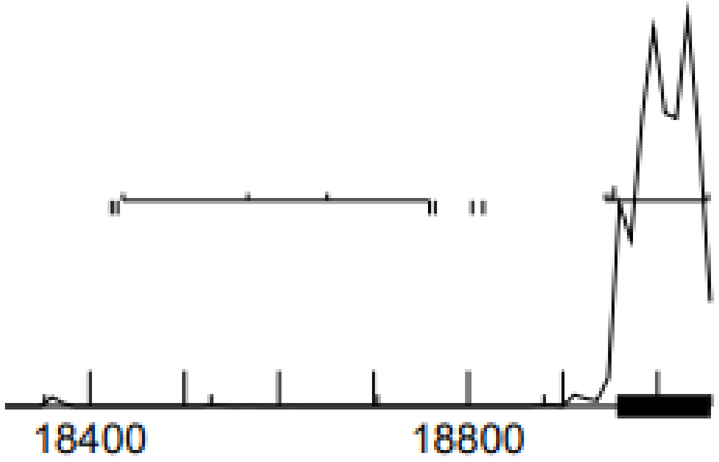
Some cp is cut off *preferred start*

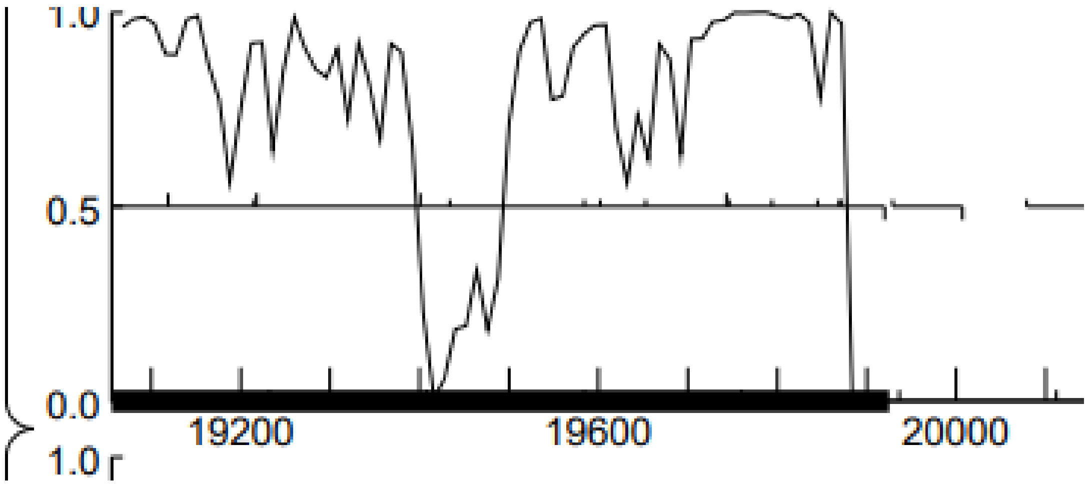
- Start 18954

Some cp is cut off

- Start 18957

Some cp is cut off





Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Start 18948

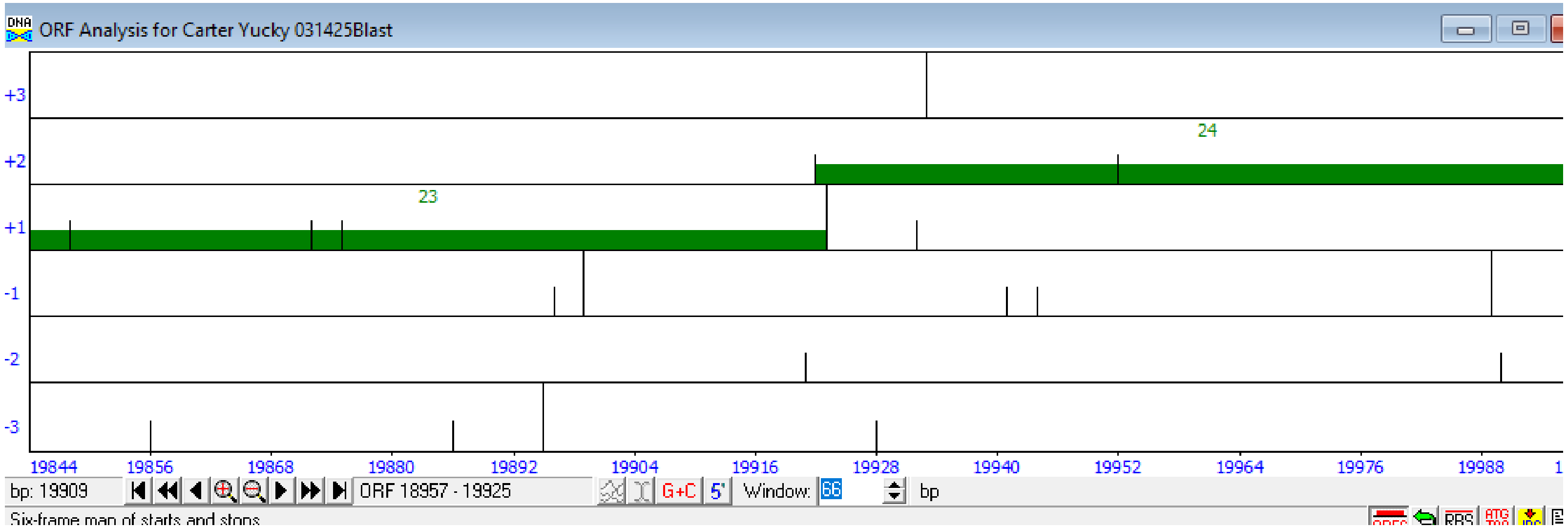
Overlap of 10

- Start 18954

Overlap of 4

- Start 18957

Overlap of 1



ORF Analysis for Carter Yucky 031425Blast

bp: 19909

ORF 18957 - 19925

Window: 66 bp

Six-frame map of starts and stops

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

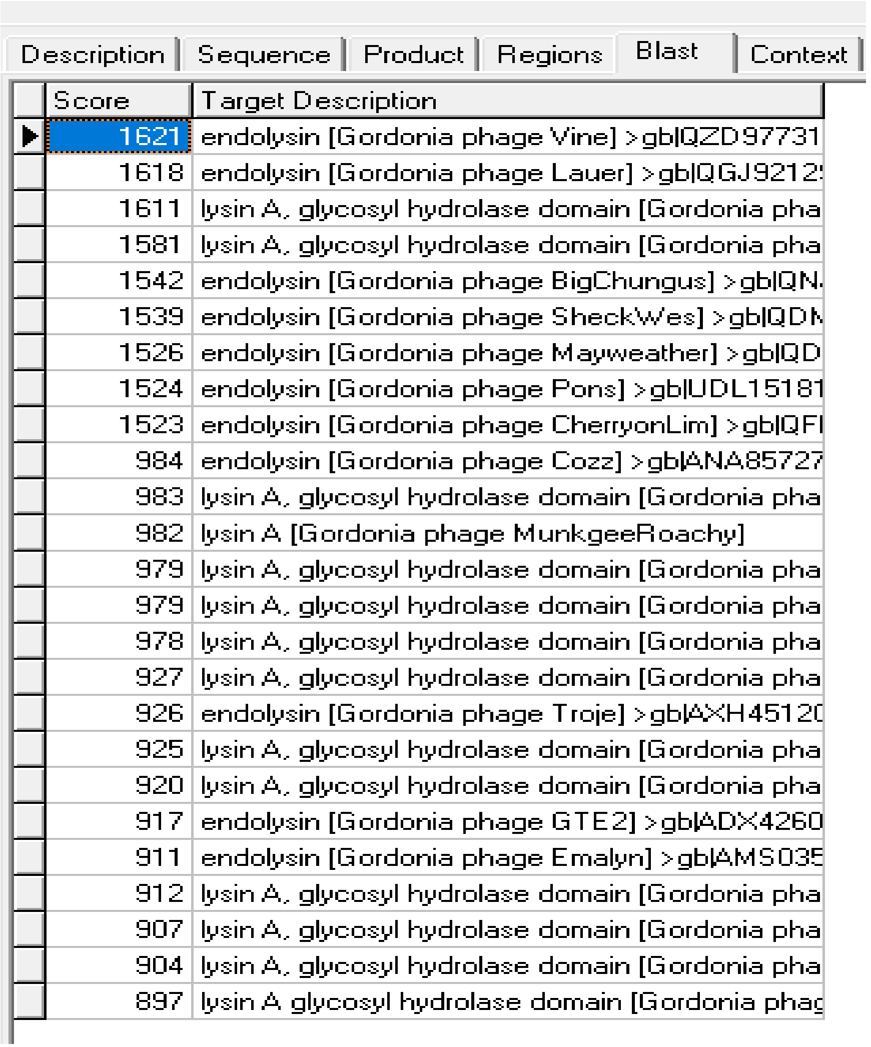
	18948	18954	18957
GeneMark	None	Glimmer	GeneMark
Coding potential	Includes some cp	Includes some cp	Includes some cp
RBS	Z value: 2.555 Final score: -4.316	Z value: 2.555 Final score: -3.839	Z value: 2.555 Final score: -4.589
BLAST	2 1:1 alignments	2 1:1 alignments	6 1:1 alignments
Starterator	5	4	37
Overlap	10	4	1

Start site is 18957 because it was called by GeneMark, it had the best z value and final score out of all the possible start sites and had the most manual annotations. It also had the most 1:1 alignments out of all the possible start sites.

BLAST function evidence. What assigned functions do other highly similar genes have?

11 endolysin

14 lysin A

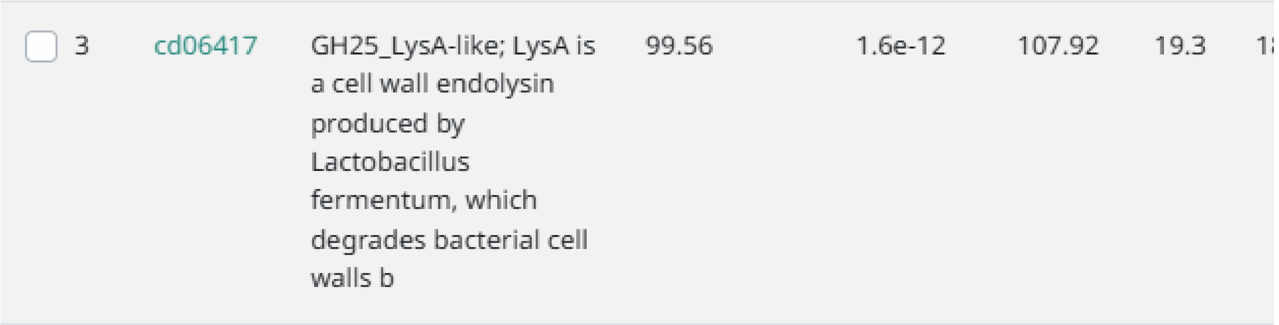


Score	Target Description
1621	endolysin [Gordonia phage Vine] >gb|QZD97731
1618	endolysin [Gordonia phage Lauer] >gb|QGJ9212
1611	lysin A, glycosyl hydrolase domain [Gordonia pha
1581	lysin A, glycosyl hydrolase domain [Gordonia pha
1542	endolysin [Gordonia phage BigChungus] >gb|QN
1539	endolysin [Gordonia phage SheckWes] >gb|QD
1526	endolysin [Gordonia phage Mayweather] >gb|QD
1524	endolysin [Gordonia phage Pons] >gb|UDL15181
1523	endolysin [Gordonia phage CherryonLim] >gb|QF
984	endolysin [Gordonia phage Cozz] >gb|ANA85727
983	lysin A, glycosyl hydrolase domain [Gordonia pha
982	lysin A [Gordonia phage MunkgeeRoachy]
979	lysin A, glycosyl hydrolase domain [Gordonia pha
979	lysin A, glycosyl hydrolase domain [Gordonia pha
978	lysin A, glycosyl hydrolase domain [Gordonia pha
927	lysin A, glycosyl hydrolase domain [Gordonia pha
926	endolysin [Gordonia phage Troje] >gb|AXH4512
925	lysin A, glycosyl hydrolase domain [Gordonia pha
920	lysin A, glycosyl hydrolase domain [Gordonia pha
917	endolysin [Gordonia phage GTE2] >gb|ADX4260
911	endolysin [Gordonia phage Emalyn] >gb|AMS035
912	lysin A, glycosyl hydrolase domain [Gordonia pha
907	lysin A, glycosyl hydrolase domain [Gordonia pha
904	lysin A, glycosyl hydrolase domain [Gordonia pha
897	lysin A glycosyl hydrolase domain [Gordonia phag

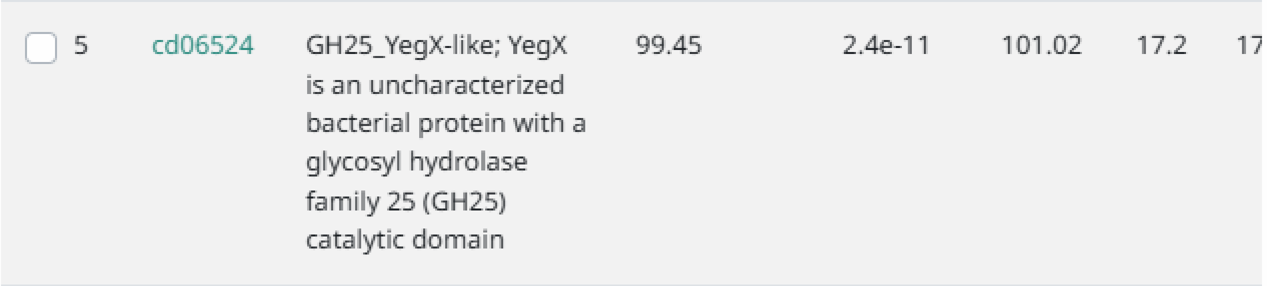
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- Numerous hits for lysin A. To be lysin A, must have a lysin B if mycobacteriophage is not present. Otherwise, it is endolysin.

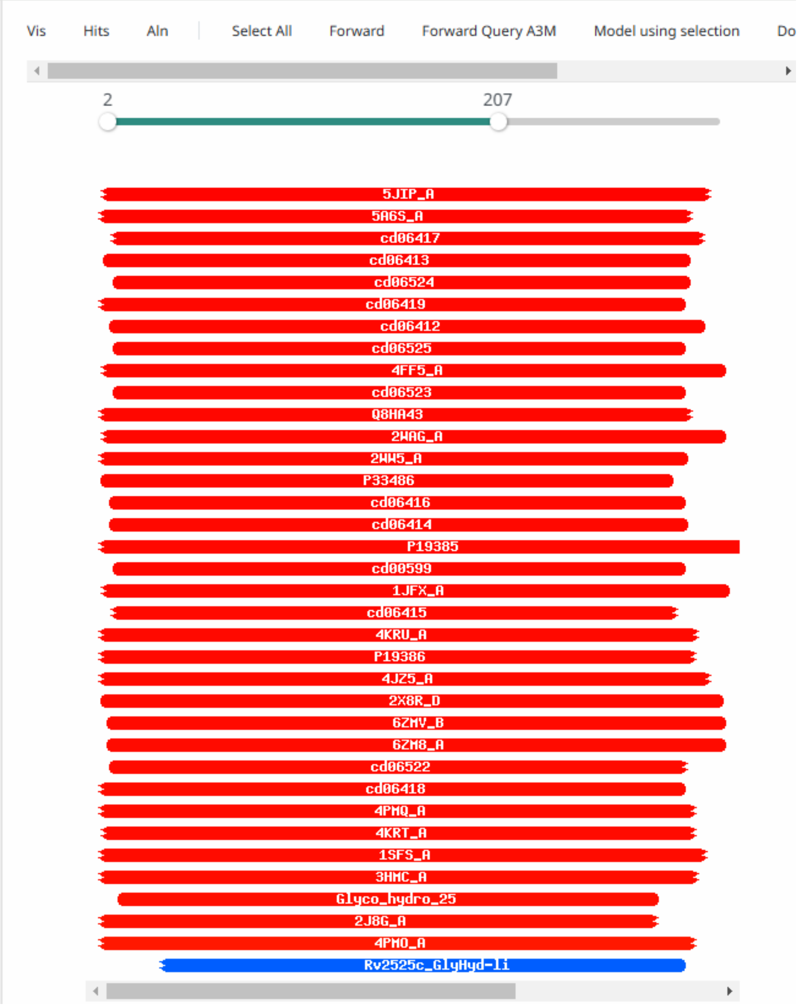
- Multiple hits for the domain: glycosyl hydrolase



□	3	cd06417	GH25_LysA-like; LysA is a cell wall endolysin produced by Lactobacillus fermentum, which degrades bacterial cell walls b	99.56	1.6e-12	107.92	19.3	1



□	5	cd06524	GH25_YegX-like; YegX is an uncharacterized bacterial protein with a glycosyl hydrolase family 25 (GH25) catalytic domain	99.45	2.4e-11	101.02	17.2	17

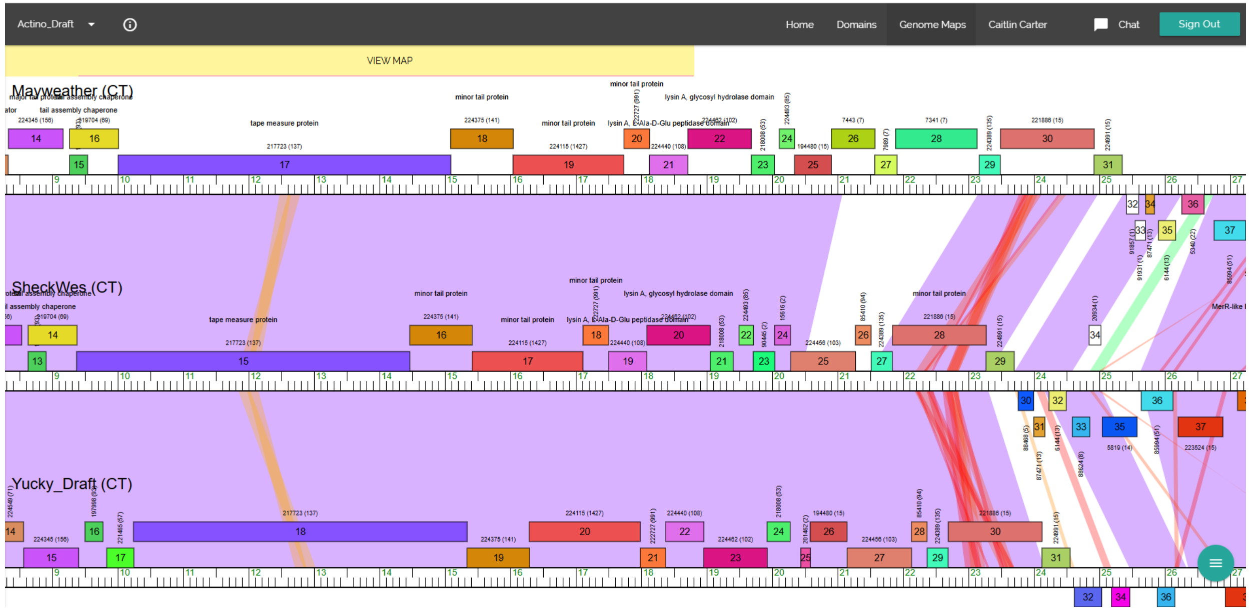


Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Yucky feature 23 function: none
  conserved domain: none

- SheckWes feature 20 function: lysin A, glycosyl hydrolase domain
  conserved domain: none

- Mayweather feature 22 function: lysin A, glycosyl hydrolase domain
  conserved domain: none



Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- None

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

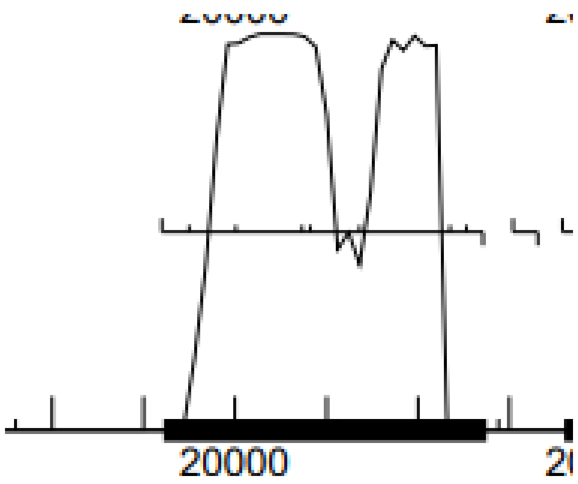
- The function is lysin A with glycosyl hydrolase domain because it had the highest amount of hits in BLAST, was the given function and conserved domain for two other highly similar genes and had the highest probability with lowest E values on Hhpred.

Feature 22 – Stop 20275

Glimmer/GeneMark

What feature number is this?	• 22
What is the stop site?	• 20275
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?	• Both and they are the same
What is the autoannotated start?	• 19922
Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start	• There is an overlap of 4

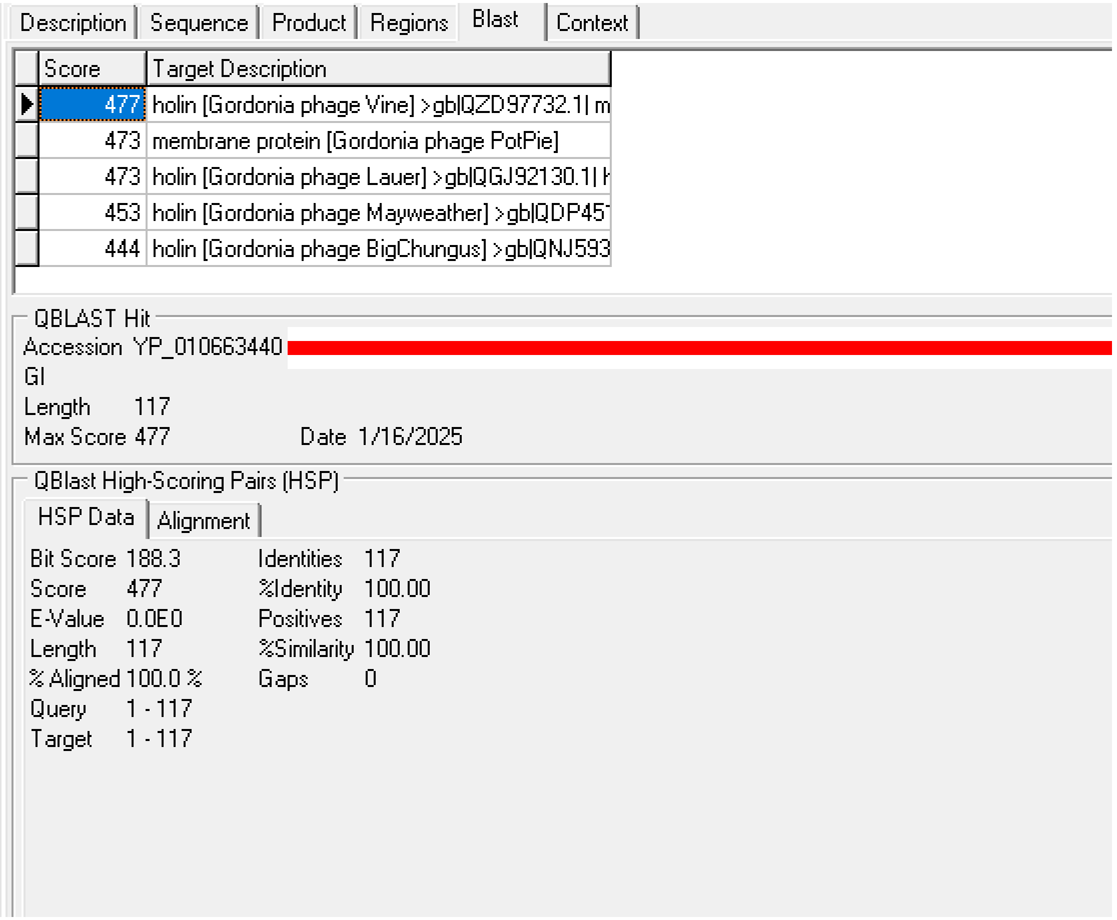
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- There is strong coding potential for this feature. There is a gap in between 20100 and about 20125. It is the only direct frame with coding potential but some of the complementary frames do have coding potential

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There is 9 1:1 hits
- There is also about 9 E-values that are close to zero
- Vine, PotPie, Lauer



Score	Target Description
477	holin [Gordonia phage Vine] >gb|QZD97732.1| m
473	membrane protein [Gordonia phage PotPie]
473	holin [Gordonia phage Lauer] >gb|QGJ92130.1| I
453	holin [Gordonia phage Mayweather] >gb|QDP45
444	holin [Gordonia phage BigChungus] >gb|QNJ593

Accession YP_010663440
GI
Length 117
Max Score 477 Date 1/16/2025

Bit Score	188.3	Identities	117
Score	477	%Identity	100.00
E-Value	0.0E0	Positives	117
Length	117	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 117		
Target	1 - 117		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

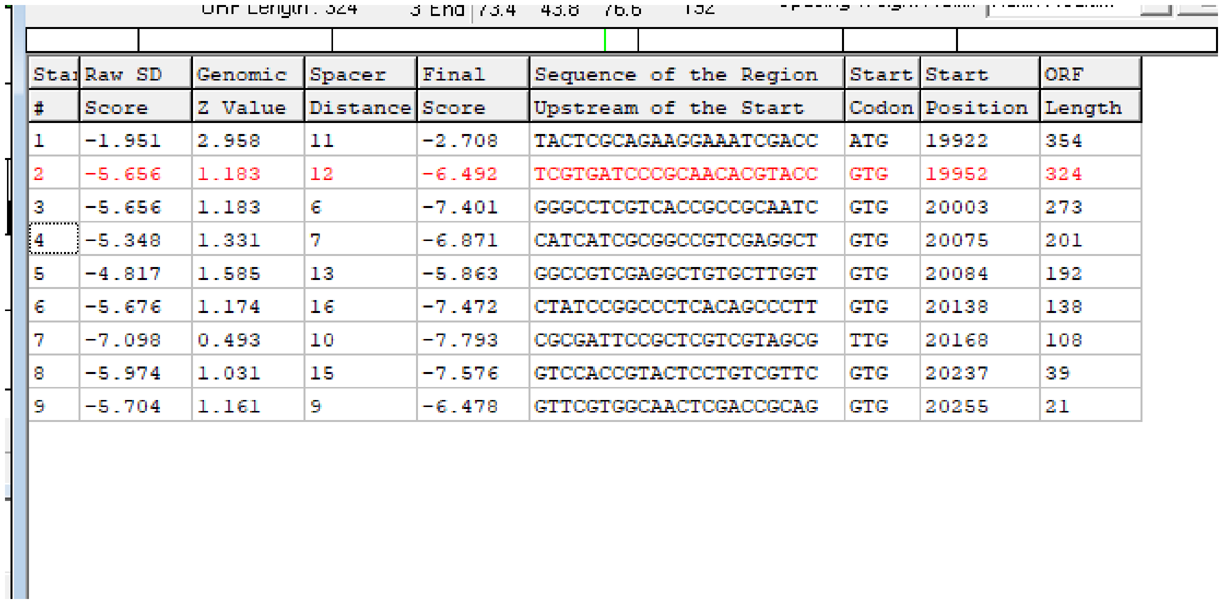
- Yes this feature is a gene dur to the multiple 1:1 blast hits and having strong coding potential with multiple peaks through the length of the feature. It was also called by glimmer and genemark.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- Start 19922 had 9 1:1 blast hits with others like PotPie, Lauer, and Vine

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

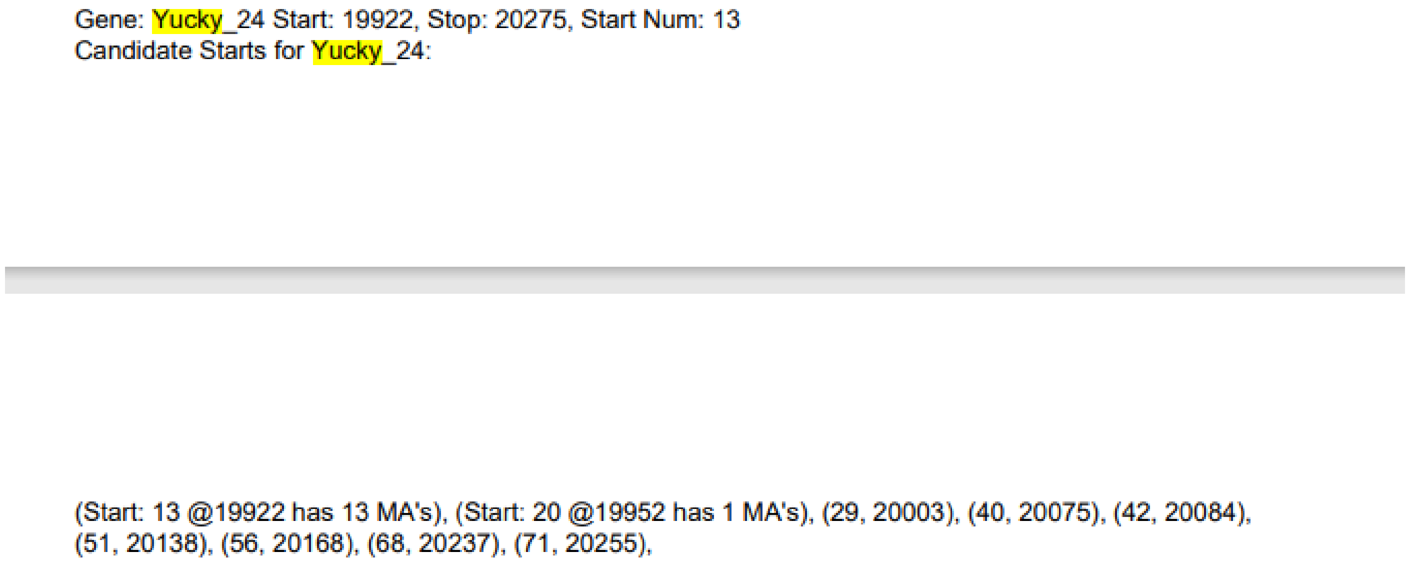
- Z-value:2.958
- Final: -2.708
- These are great values to have since the z-value is close to 3 and the final score is the closest to zero out of all of these



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	-1.951	2.958	11	-2.708	TACTCGCAGAAGGAAATCGACC	ATG	19922	354
2	-5.656	1.183	12	-6.492	TCGTGATCCCGCAACACGTACC	GTG	19952	324
3	-5.656	1.183	6	-7.401	GGGCCTCGTCACCGCCGCAATC	GTG	20003	273
4	-5.348	1.331	7	-6.871	CATCATCGCGGCCGTCGAGGCT	GTG	20075	201
5	-4.817	1.585	13	-5.863	GGCCGTCGAGGCTGTGCTTGGT	GTG	20084	192
6	-5.676	1.174	16	-7.472	CTATCCGGCCCTCACAGCCCTT	GTG	20138	138
7	-7.098	0.493	10	-7.793	CGCGATTCCGCTCGTCGTAGCG	TTG	20168	108
8	-5.974	1.031	15	-7.576	GTCCACCGTACTCCTGTCGTTC	GTG	20237	39
9	-5.704	1.161	9	-6.478	GTTCGTGGCAACTCGACCGCAG	GTG	20255	21

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

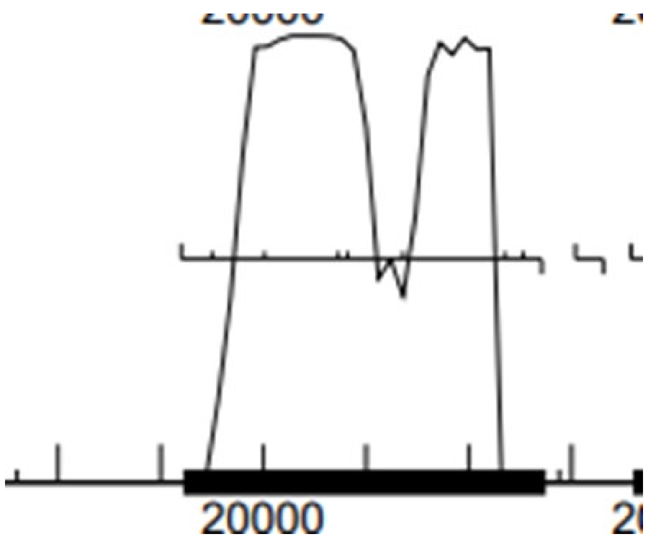
- Start 19922 has 13 Manual Annotated starts which has the best numbers out of all the others since the only other proposed (19952) only has 1 MA



Gene: Yucky_24 Start: 19922, Stop: 20275, Start Num: 13
Candidate Starts for Yucky_24:

(Start: 13 @19922 has 13 MA's), (Start: 20 @19952 has 1 MA's), (29, 20003), (40, 20075), (42, 20084), (51, 20138), (56, 20168), (68, 20237), (71, 20255),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- The start at 19922 includes all of the coding potential of the entire length of the feature which makes it the best candidate here.

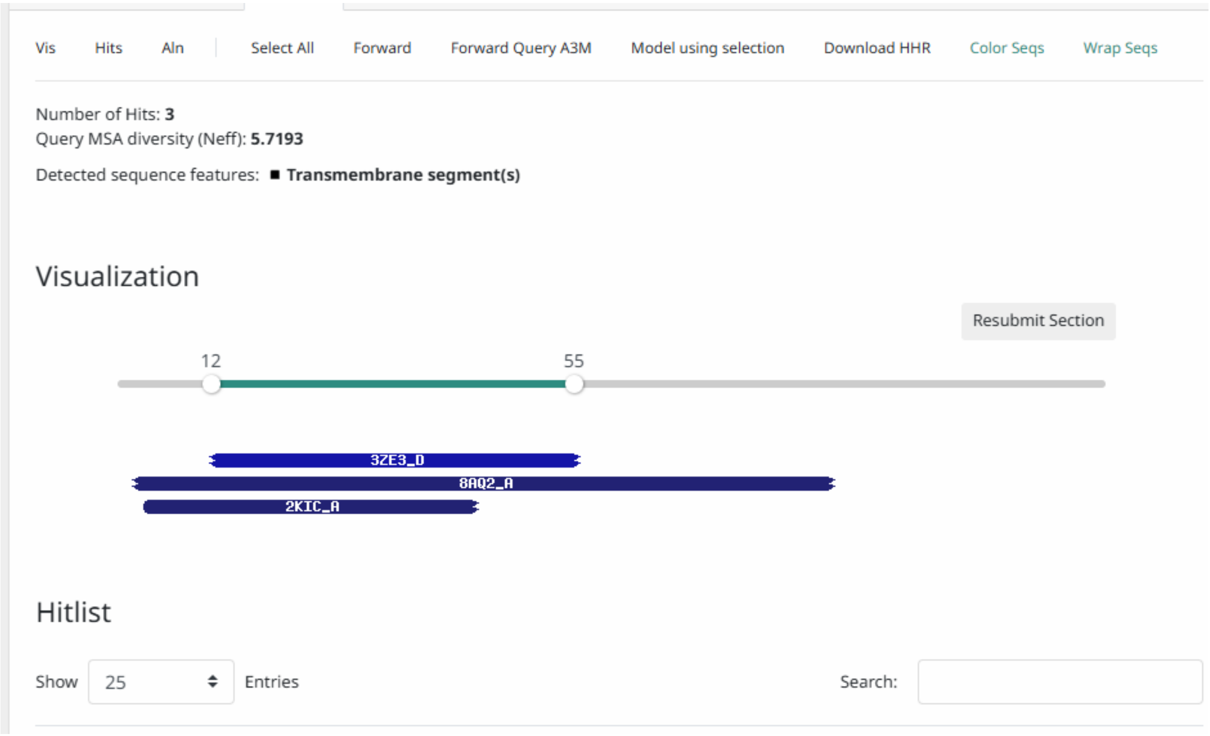
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Start 19922 has an overlap of 4

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- My only proposed start was 19922 which has great RBS scores. Has 13 MA in starterator. Includes all of the coding potential for the entire feature length. And also, has 9 1:1 blast hits. So, with that evidence I’m going to say that 19922 is the best start for feature 24.

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



Number of Hits: **3**
Query MSA diversity (Neff): **5.7193**
Detected sequence features: ■ **Transmembrane segment(s)**

Visualization

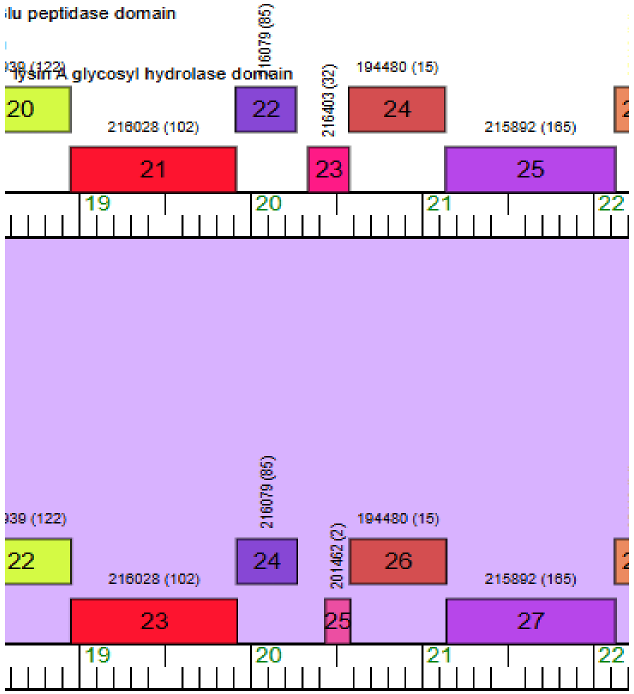
Hitlist

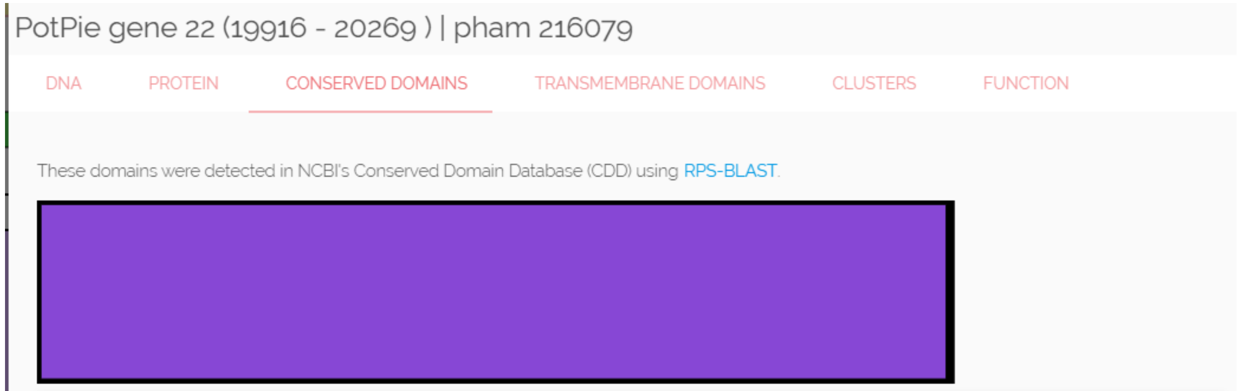
Show 25 Entries

Search:

- There is no hits that support there being a function of this gene.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

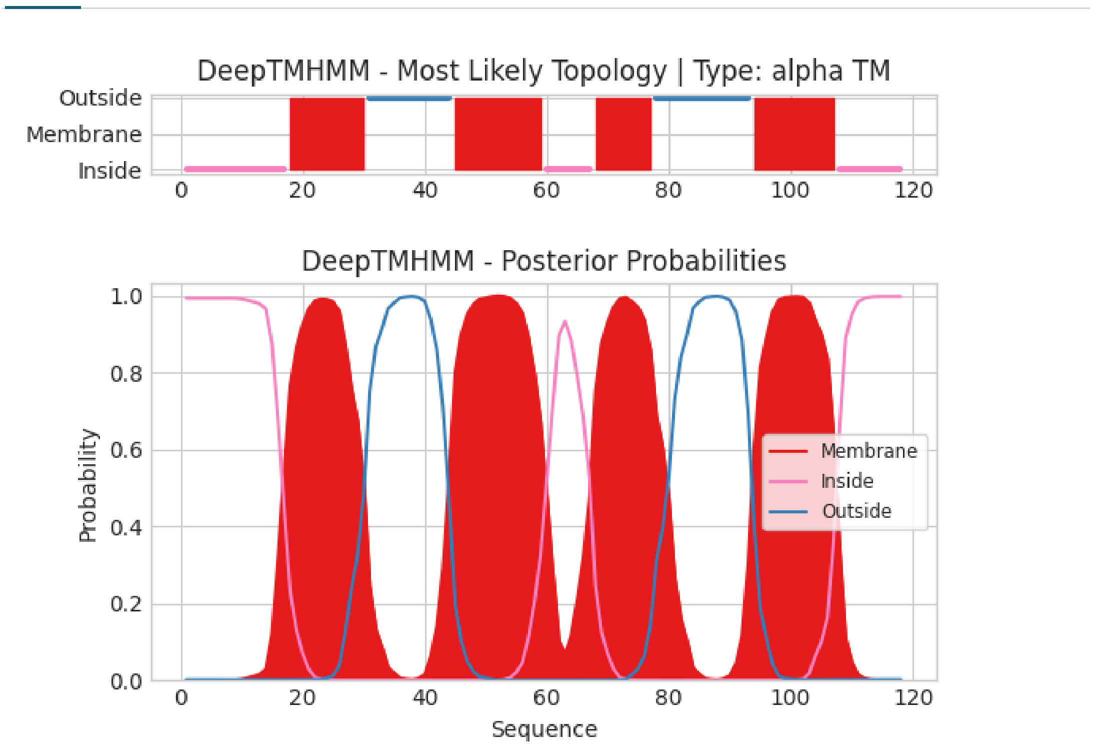



PotPie gene 22 (19916 - 20269) | pham 216079

These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

- There was no function predicted for either of the genes and there was no conserved domains provided
- In phamerator it is next to an endolysin which gives proof it may be a holin

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- This shows that there are 4 distinct transmembrane domains. This helps lead me to believe this could be a holin due to the number of TMDs being 4 which is what you need for it to be considered a holin.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function of this feature is a holin. This is due to it being adjacent to an endolysin in phamerator. It has numerous blast hits with other features that are holins in DNA master and in ncbi blast. It has 4 transmembrane domains which is more than the minimum holin requirement of 2. The only issue is there is no evidence for a function in HHPRED but the evidence from the other resources make up for this. Alternatively, this would be a membrane protein.

Feature 23 – Stop 20584

Glimmer/GeneMark

What feature number is this?

- 23

What is the stop site?

- 20584

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

- Glimmer:20438 Genemark: 20360

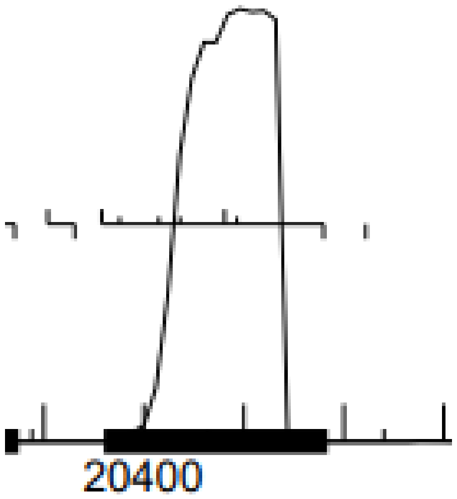
What is the autoannotated start?

- 20438

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

- 20438 has a gap of 162
- 29360 has a gap of 84

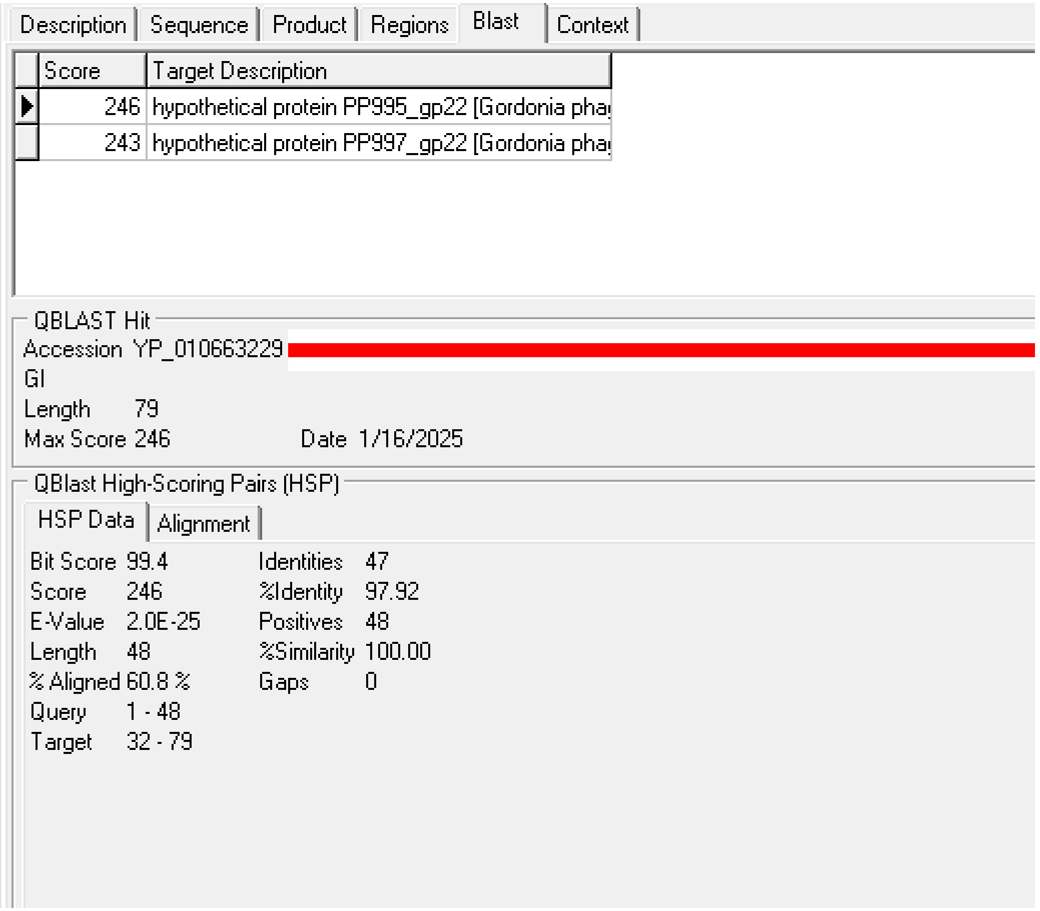
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- There is strong coding potential for this feature dur to the strong peak it has that goes for the majority of its length. There is proposed pieces of the feature that have little to no coding potential at all. Complementary frames have coding potential in this place in their frames.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There is only 2 blast hits for this feature both 1:32 and they have e-values that go to 10^-25



Score	Target Description
246	hypothetical protein PP995_gp22 [Gordonia pha
243	hypothetical protein PP997_gp22 [Gordonia pha

QBLAST Hit

Accession YP_010663229

GI

Length 79

Max Score 246 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 99.4

Score 246

E-Value 2.0E-25

Length 48

% Aligned 60.8 %

Query 1 - 48

Target 32 - 79

Identities 47

%Identity 97.92

Positives 48

%Similarity 100.00

Gaps 0

Answer: Is it a gene? Give evidence why you think this is a gene or not.

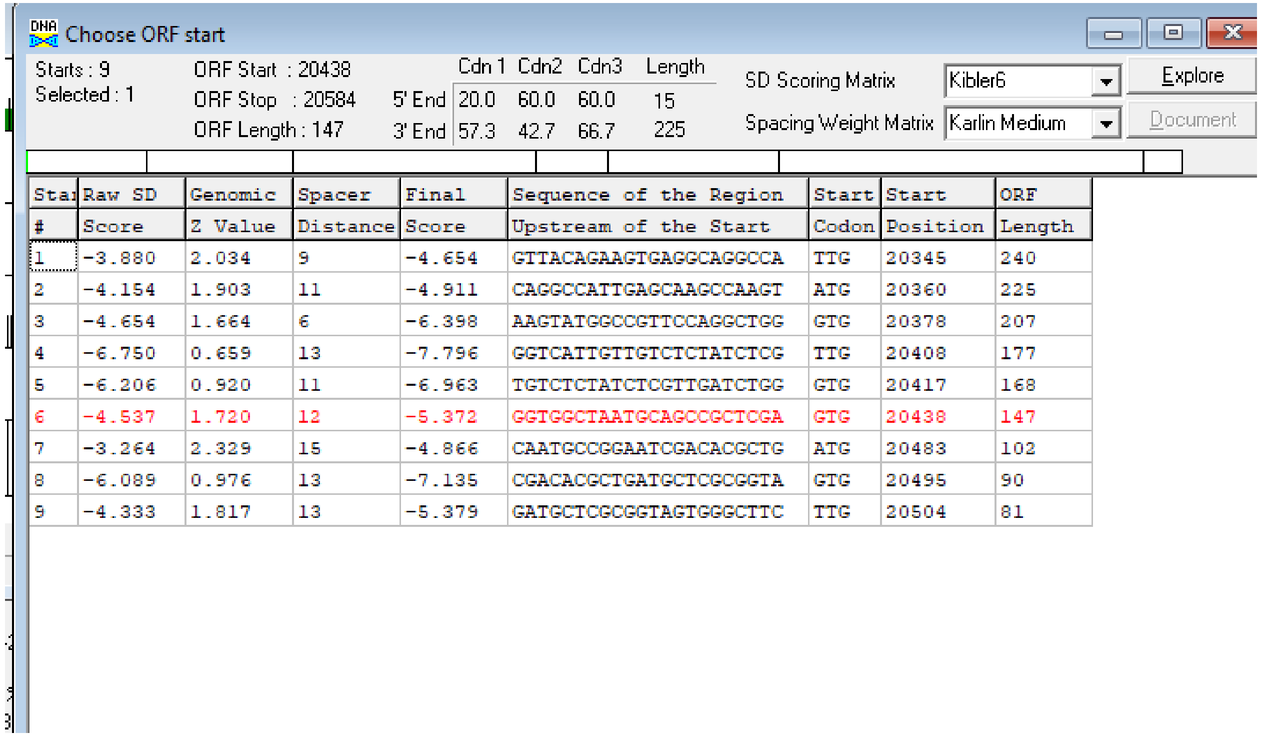
- Yes, it is a gene because it was called by both glimmer and gene mark, it is shown to have strong coding potential throughout, and it has 2 blast hits that have e-values of 10^-25 which is way below the required 10^-7.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- Start 20345 has 6 blast hits with 1:1 alignments with
- Start 20438 has 2 blast alignments both at 1:32
- Start 20360 has 3 blast hits of 1:6 and 1 blast hit of 1:1 with CherryonLim
- Start 20360 is favored here since it has the 1:1 blast hit

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- 20345 Z value 2.034 and FS -4.654
- 20438 has
- Z-value: 1.720
- FS: -5.372
- 20360 has
- Z-value: 1.903
- FS: -4.911



Starts : 9
Selected : 1

ORF Start : 20438
ORF Stop : 20584
ORF Length : 147

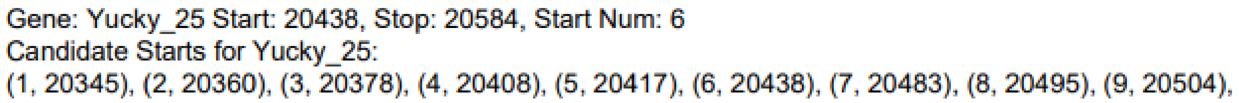
	Cdn 1	Cdn2	Cdn3	Length
5' End	20.0	60.0	60.0	15
3' End	57.3	42.7	66.7	225

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-3.880	2.034	9	-4.654	GTTACAGAAGTGAGGCAGGCCA	TTG	20345	240
2	-4.154	1.903	11	-4.911	CAGGCCATTGAGCAAGCCAAGT	ATG	20360	225
3	-4.654	1.664	6	-6.398	AAGTATGGCCGTTCCAGGCTGG	GTG	20378	207
4	-6.750	0.659	13	-7.796	GGTCATTGTTGTCTCTATCTCG	TTG	20408	177
5	-6.206	0.920	11	-6.963	TGTCTCTATCTCGTTGATCTGG	GTG	20417	168
6	-4.537	1.720	12	-5.372	GGTGGCTAATGCAGCCGCTCGA	GTG	20438	147
7	-3.264	2.329	15	-4.866	CAATGCCGGAATCGACACGCTG	ATG	20483	102
8	-6.089	0.976	13	-7.135	CGACACGCTGATGCTCGCGGTA	GTG	20495	90
9	-4.333	1.817	13	-5.379	GATGCTCGCGGTAGTGGGCTTC	TTG	20504	81

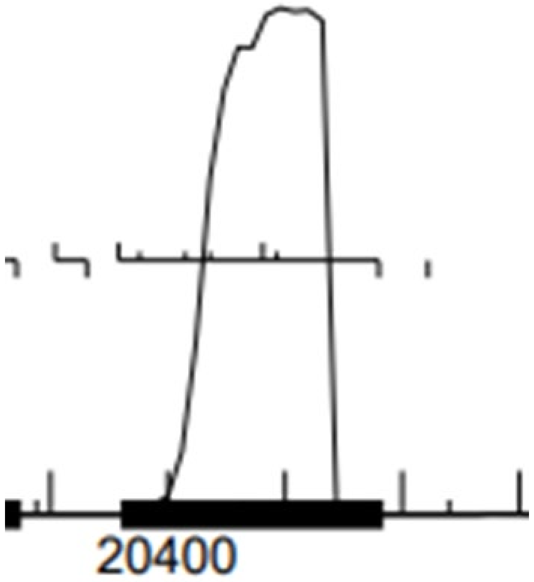
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- There are no manual annotations proposed for any start



Gene: Yucky_25 Start: 20438, Stop: 20584, Start Num: 6
Candidate Starts for Yucky_25:
(1, 20345), (2, 20360), (3, 20378), (4, 20408), (5, 20417), (6, 20438), (7, 20483), (8, 20495), (9, 20504),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- 20345 includes all coding potential
- Start 20438 cuts off a tiny piece of starting coding potential but it is not very strong
- Start 20360 includes all of the coding potential in the feature

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

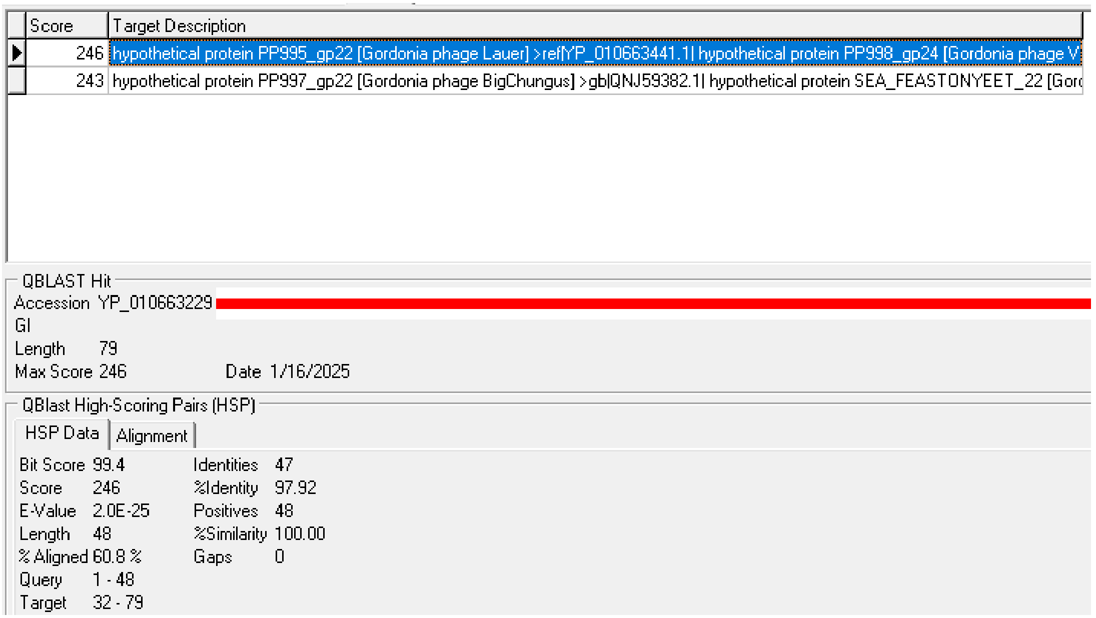
- 20438 has a gap of 162
- 29360 has a gap of 84
- 20345 has a gap of 69
- Start 20345 would have the better stats here since it has a smaller overall gap than the other two considered starts

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	20345	20360
Glimmer/Genemark		Genemark
Blast	6 1:1 hits	1 1:1 hit 3 1:6 hits
RBS	Z value 2.034 and FS -4.654	Z-value: 1.903 FS: -4.911
Genemark	Includes all coding potential	Includes all coding potential
Starterator	No MA	No MA
Gap/Overlap	Gap of 69	Gap of 84

- 20345 would be the better starting site here because it has it has 6 1:1 Blast hits, has a better RBS Scocres, includes all coding potential and has a smaller gap of 69.

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
246	hypothetical protein PP995_gp22 [Gordonia phage Lauer] >ref|YP_010663441.1| hypothetical protein PP998_gp24 [Gordonia phage V
243	hypothetical protein PP997_gp22 [Gordonia phage BigChungus] >gb|QNJ59382.1| hypothetical protein SEA_FEASTONYEET_22 [Gord

QBLAST Hit

Accession YP_010663229

GI

Length 79

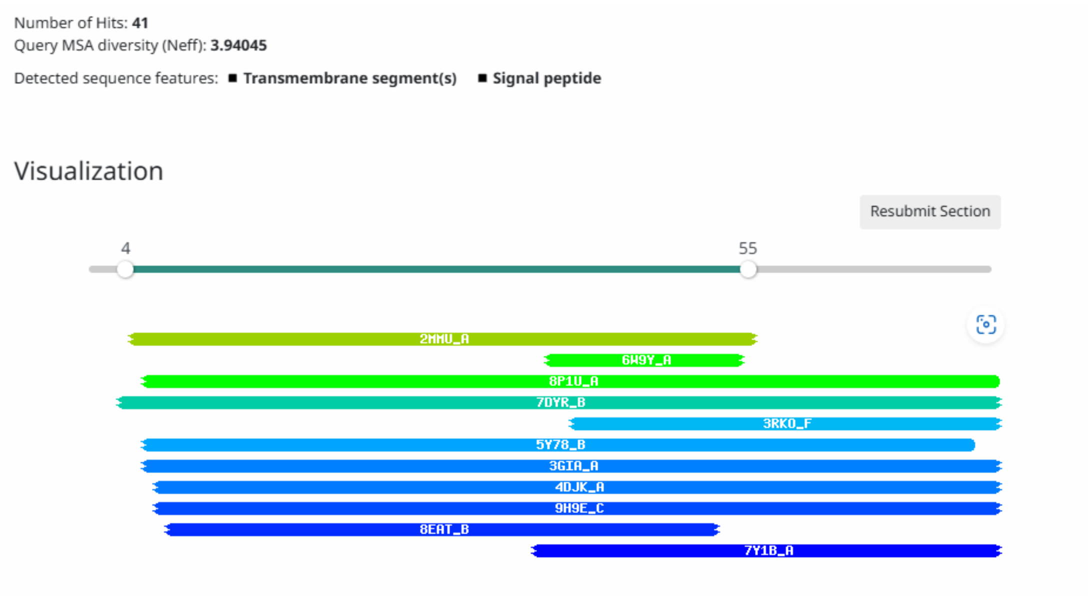
Max Score 246 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	99.4	Identities	47
Score	246	%Identity	97.92
E-Value	2.0E-25	Positives	48
Length	48	%Similarity	100.00
% Aligned	60.8 %	Gaps	0
Query	1 - 48		
Target	32 - 79		

- The only evidence this has due to blast is that it’s just a hypothetical protein which matches with other phages like Lauer and BigChungus.

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



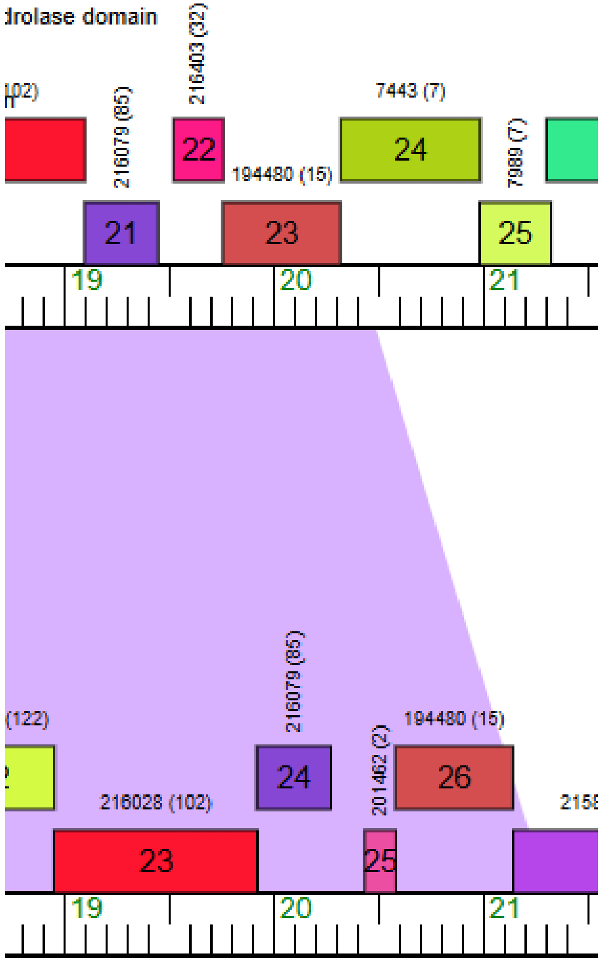
Number of Hits: **41**
Query MSA diversity (Neff): **3.94045**

Detected sequence features: ■ **Transmembrane segment(s)** ■ **Signal peptide**

Visualization

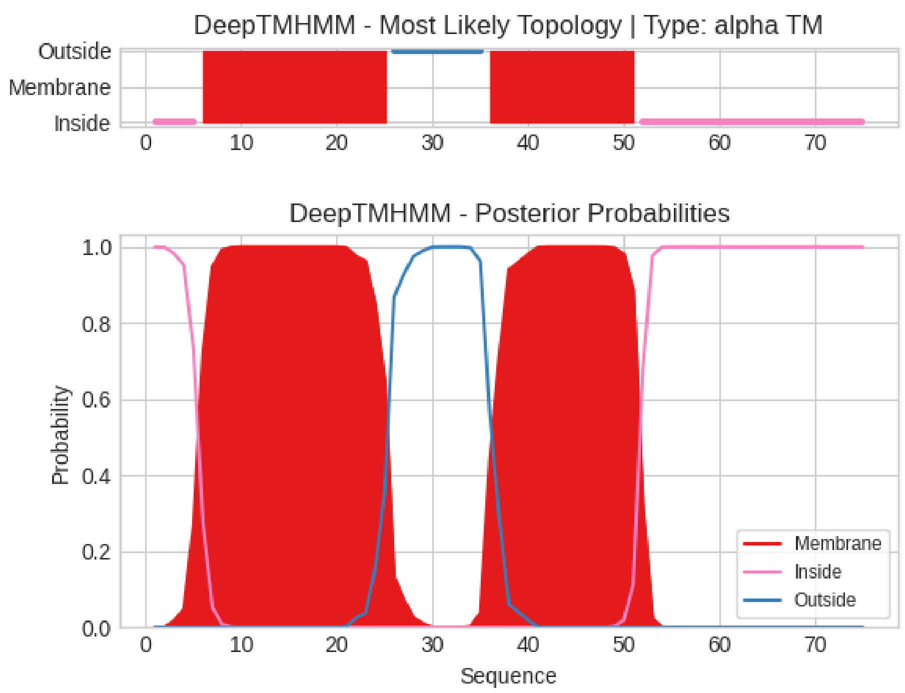
- None of these hits provide any evidence that there is a function for this gene

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- For Lauer and BigChungus the features that relate both do not have any function that is announced on the conserved domain list of phamerator

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- There are 2 transmembrane domains for this gene, so this gives evidence that it is most definitely a membrane protein is not anything else

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I believe that this feature is just a membrane protein due to there being almost no evidence that this could have a function other than the 2 transmembrane domains that only provide evidence for the function of membrane protein.

Feature 24 – Stop 21144

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

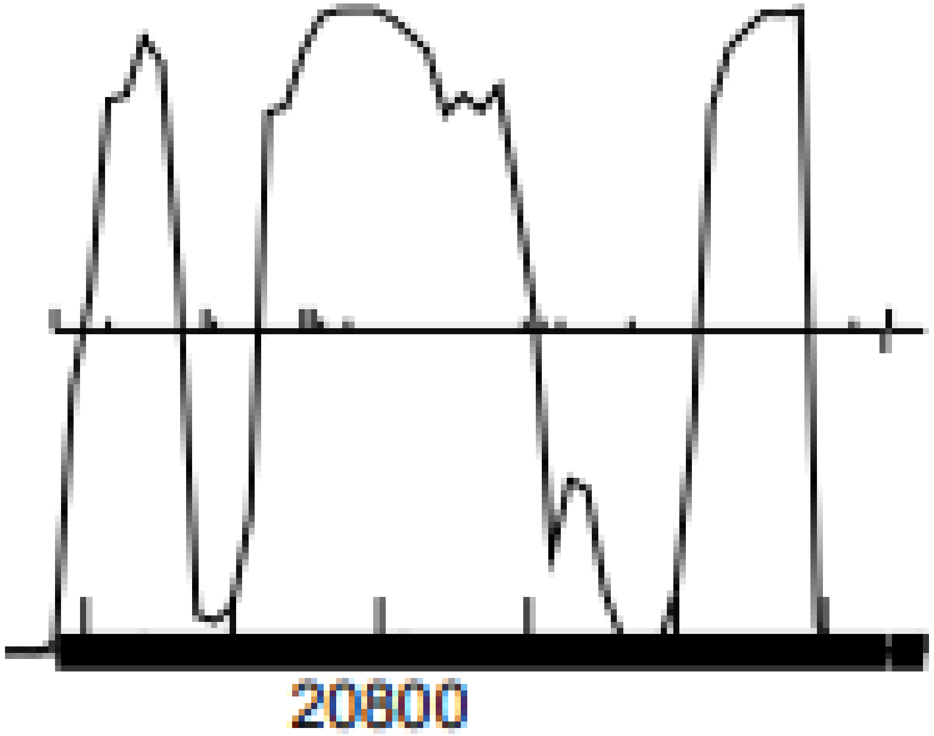
- Feature 24
- Stop site: 21144
- Called by both Glimmer and GeneMark
- Autoannotated start: 20581
- Overlap: 4

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak?   How do you know?

- Start 20581

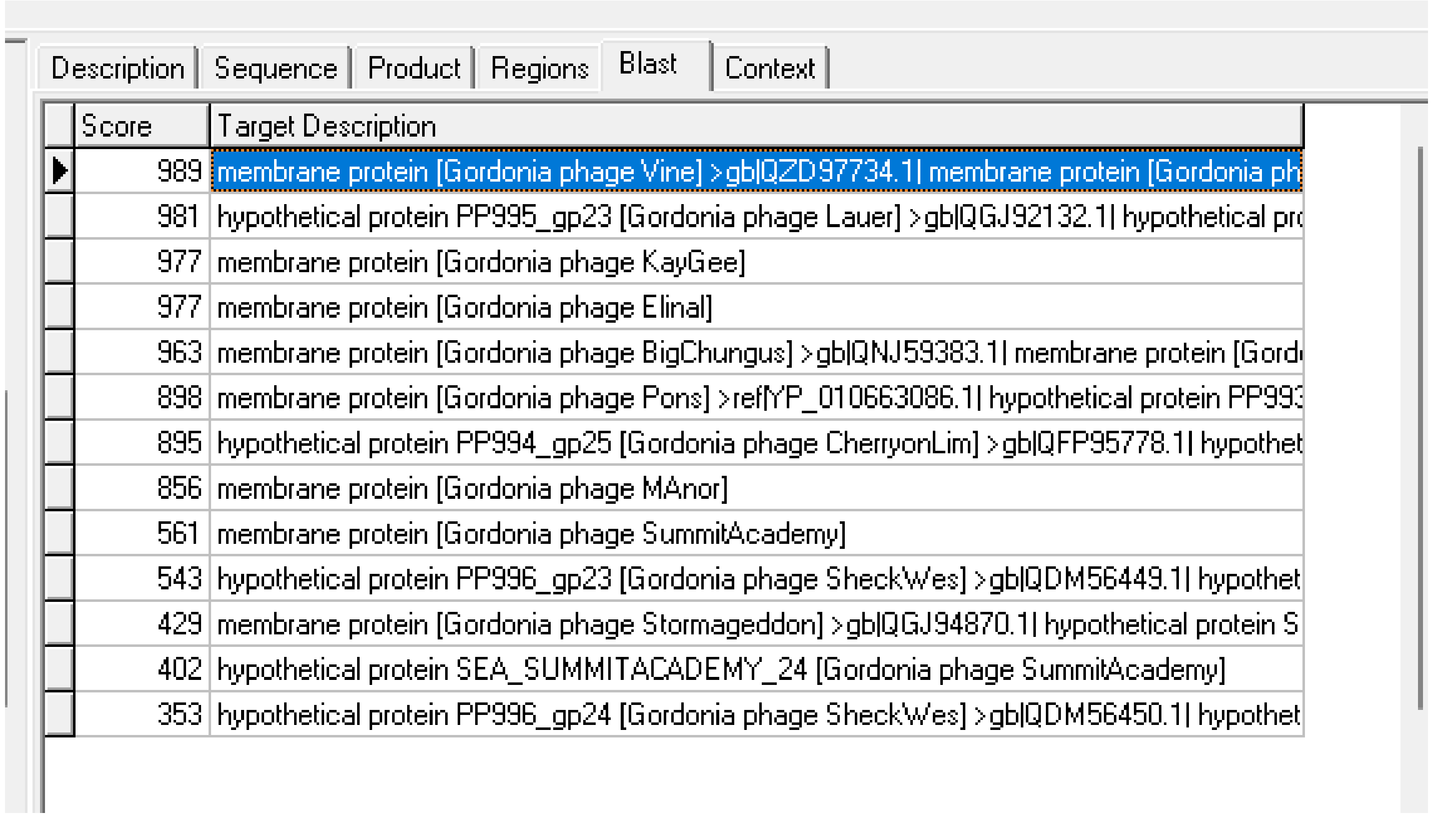
Found in forward frame 1

Includes all coding potential



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 13 highly similar genes
- Vine
- Lauer
- KayGee
- Elinal
- BigChungus
- Pons
- CherryonLim
- Manor
- SummitAcademy
- SheckWes
- Stormageddon
- SEA_SUMMITACADEMY_24
- SheckWes



Score	Target Description
989	membrane protein [Gordonia phage Vine] >gb|QZD97734.1| membrane protein [Gordonia ph
981	hypothetical protein PP995_gp23 [Gordonia phage Lauer] >gb|QGJ92132.1| hypothetical pro
977	membrane protein [Gordonia phage KayGee]
977	membrane protein [Gordonia phage Elinal]
963	membrane protein [Gordonia phage BigChungus] >gb|QNJ59383.1| membrane protein [Gordo
898	membrane protein [Gordonia phage Pons] >ref|YP_010663086.1| hypothetical protein PP993
895	hypothetical protein PP994_gp25 [Gordonia phage CherryonLim] >gb|QFP95778.1| hypothet
856	membrane protein [Gordonia phage MAnor]
561	membrane protein [Gordonia phage SummitAcademy]
543	hypothetical protein PP996_gp23 [Gordonia phage SheckWes] >gb|QDM56449.1| hypothet
429	membrane protein [Gordonia phage Stormageddon] >gb|QGJ94870.1| hypothetical protein S
402	hypothetical protein SEA_SUMMITACADEMY_24 [Gordonia phage SummitAcademy]
353	hypothetical protein PP996_gp24 [Gordonia phage SheckWes] >gb|QDM56450.1| hypothet

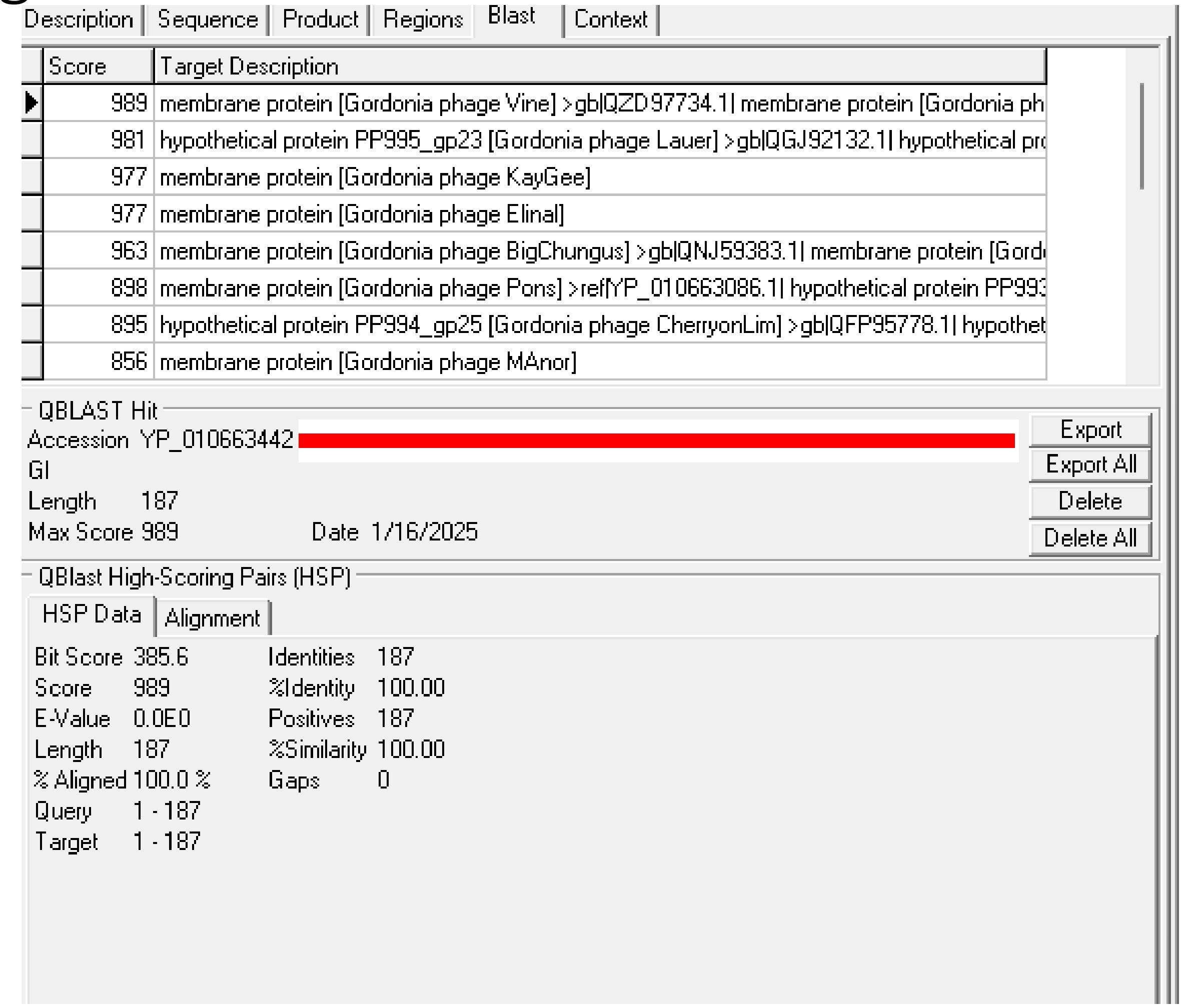
Answer: Is it a gene? Give evidence why you think this is a gene or not.

- It is a gene because it is called by both Glimmer and GeneMark, the reading frame includes all coding potential, and the feature has 13 highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

Start 20581

- 10 1:1 alignments



Score	Target Description
989	membrane protein [Gordonia phage Vine] >gb|QZD97734.1| membrane protein [Gordonia ph
981	hypothetical protein PP995_gp23 [Gordonia phage Lauer] >gb|QGJ92132.1| hypothetical pro
977	membrane protein [Gordonia phage KayGee]
977	membrane protein [Gordonia phage Elinal]
963	membrane protein [Gordonia phage BigChungus] >gb|QNJ59383.1| membrane protein [Gord
898	membrane protein [Gordonia phage Pons] >ref|YP_010663086.1| hypothetical protein PP99
895	hypothetical protein PP994_gp25 [Gordonia phage CherryonLim] >gb|QFP95778.1| hypothet
856	membrane protein [Gordonia phage MAnor]

QBLAST Hit

Accession YP_010663442
GI
Length 187
Max Score 989
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 385.6
Score 989
E-Value 0.0E0
Length 187
% Aligned 100.0 %
Query 1 - 187
Target 1 - 187

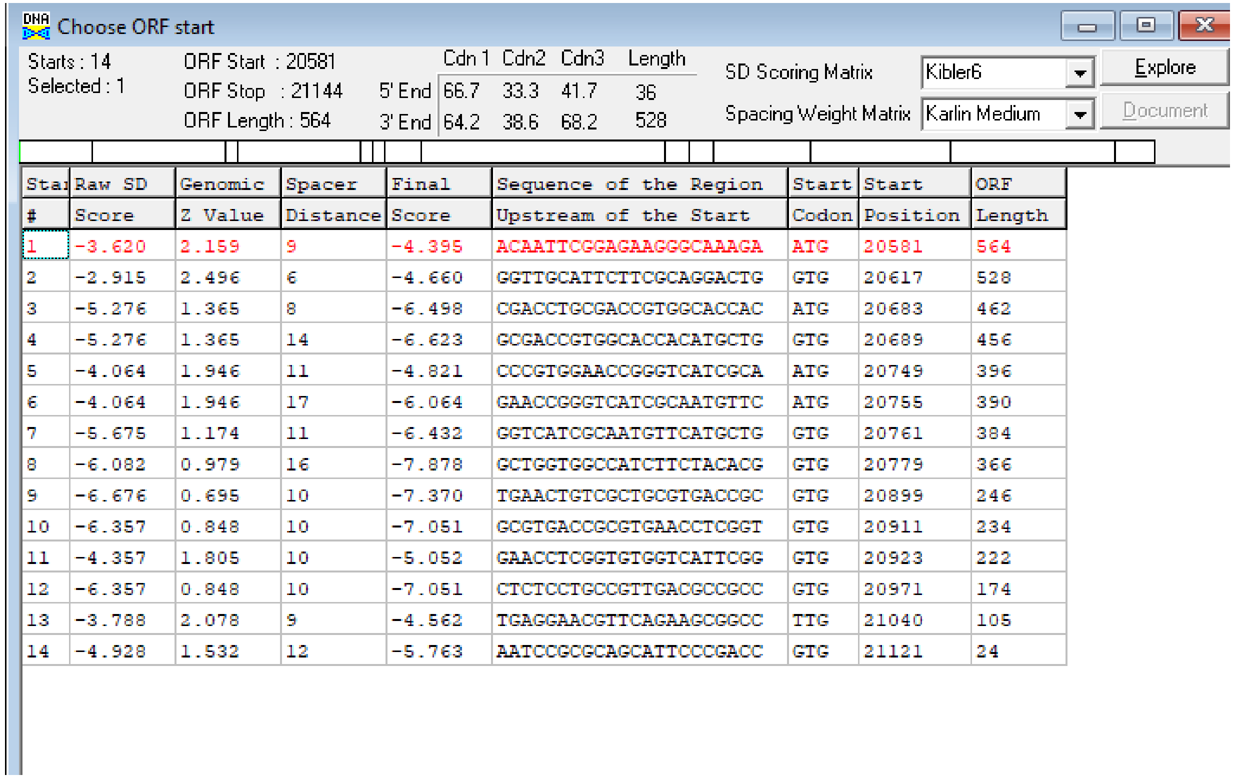
Identities 187
%Identity 100.00
Positives 187
%Similarity 100.00
Gaps 0

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Start 20581

Z Value: 2.159

Final Score: -4.395



Starts : 14
Selected : 1

ORF Start : 20581
ORF Stop : 21144
ORF Length : 564

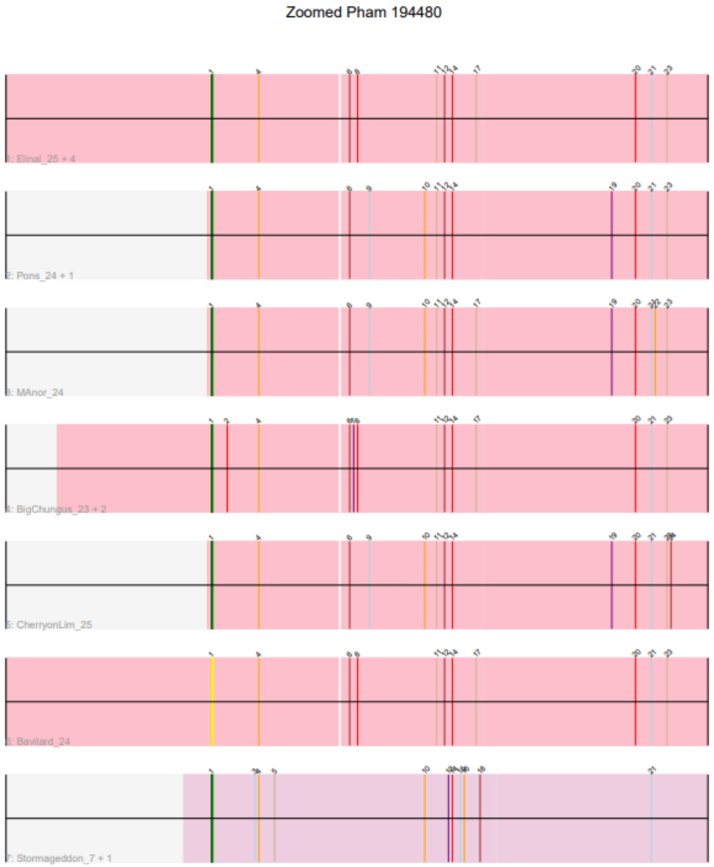
	Cdn 1	Cdn2	Cdn3	Length
5' End	66.7	33.3	41.7	36
3' End	64.2	38.6	68.2	528

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-3.620	2.159	9	-4.395	ACAATTCGGAGAAGGGCAAAGA	ATG	20581	564
2	-2.915	2.496	6	-4.660	GGTTGCATTCTTCGCAGGACTG	GTG	20617	528
3	-5.276	1.365	8	-6.498	CGACCTGCGACCGTGGCACCAC	ATG	20683	462
4	-5.276	1.365	14	-6.623	GCGACCGTGGCACCACATGCTG	GTG	20689	456
5	-4.064	1.946	11	-4.821	CCCGTGGAACCGGGTCATCGCA	ATG	20749	396
6	-4.064	1.946	17	-6.064	GAACCGGGTCATCGCAATGTTC	ATG	20755	390
7	-5.675	1.174	11	-6.432	GGTCATCGCAATGTTCATGCTG	GTG	20761	384
8	-6.082	0.979	16	-7.878	GCTGGTGGCCATCTTCTACACG	GTG	20779	366
9	-6.676	0.695	10	-7.370	TGAACTGTCGCTGCGTGACCGC	GTG	20899	246
10	-6.357	0.848	10	-7.051	GCGTGACCGCGTGAACCTCGGT	GTG	20911	234
11	-4.357	1.805	10	-5.052	GAACCTCGGTGTGGTCATTCGG	GTG	20923	222
12	-6.357	0.848	10	-7.051	CTCTCCTGCCGTTGACGCCGCC	GTG	20971	174
13	-3.788	2.078	9	-4.562	TGAGGAACGTTCAGAAGCGGCC	TTG	21040	105
14	-4.928	1.532	12	-5.763	AATCCGCGCAGCATTCCCGACC	GTG	21121	24

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Start: 1 @20581 has 13 MAs

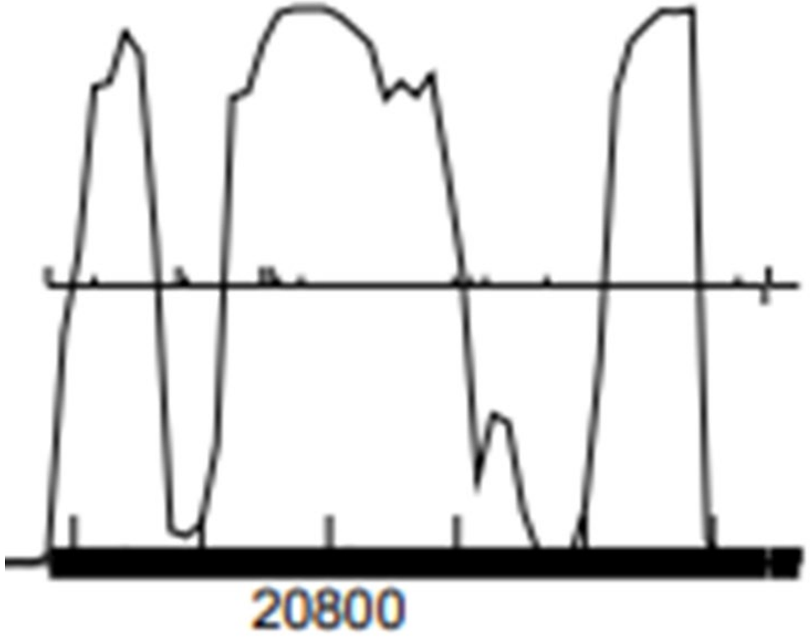


GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Start 20581

Found in forward frame 1

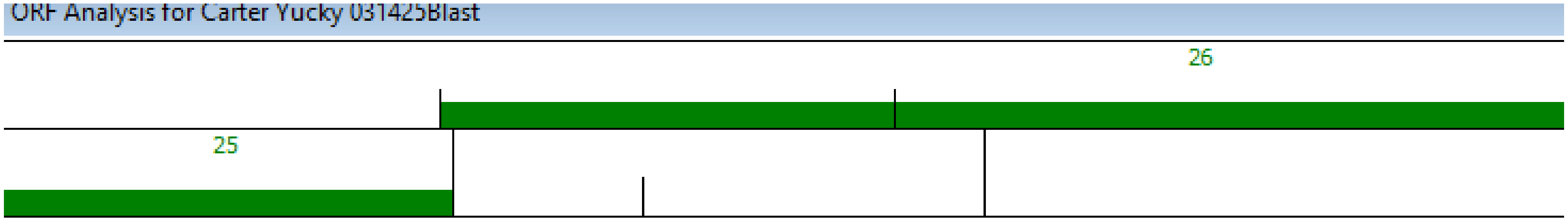
Includes all coding potential



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Start site: 20581

- Overlap: 4 (Previous feature ends at 20684)


ORF Analysis for Carter Yucky 031425Blast

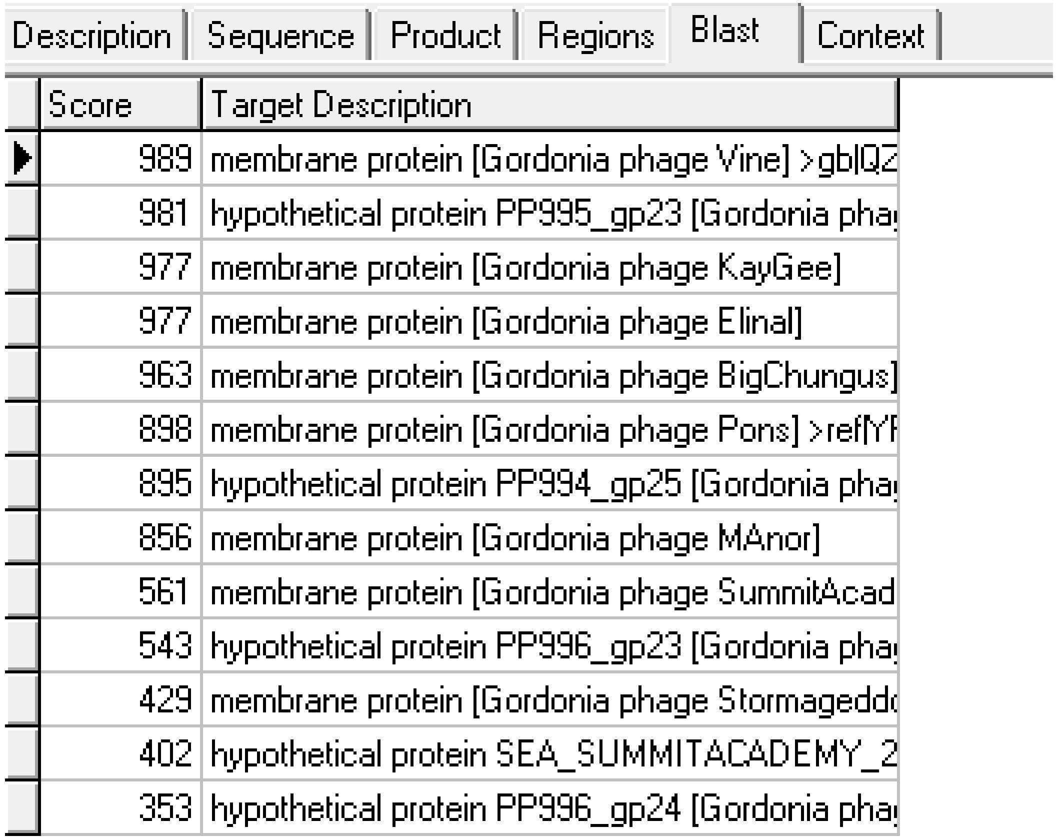
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	20581
Genemark	Glimmer & GeneMark
Coding potential	Includes all cp
RBS	Z Value: 2.159 Final Score: -4.395
BLAST	10 1:1 alignments
Starterator	13 MAs
Overlap	4

Start site is 20581 because it includes both Glimmer and GeneMark, the frame includes all coding potential, the z value is greater than 2, and there are 10 1:1 alignments.

BLAST function evidence. What assigned functions do other highly similar genes have?

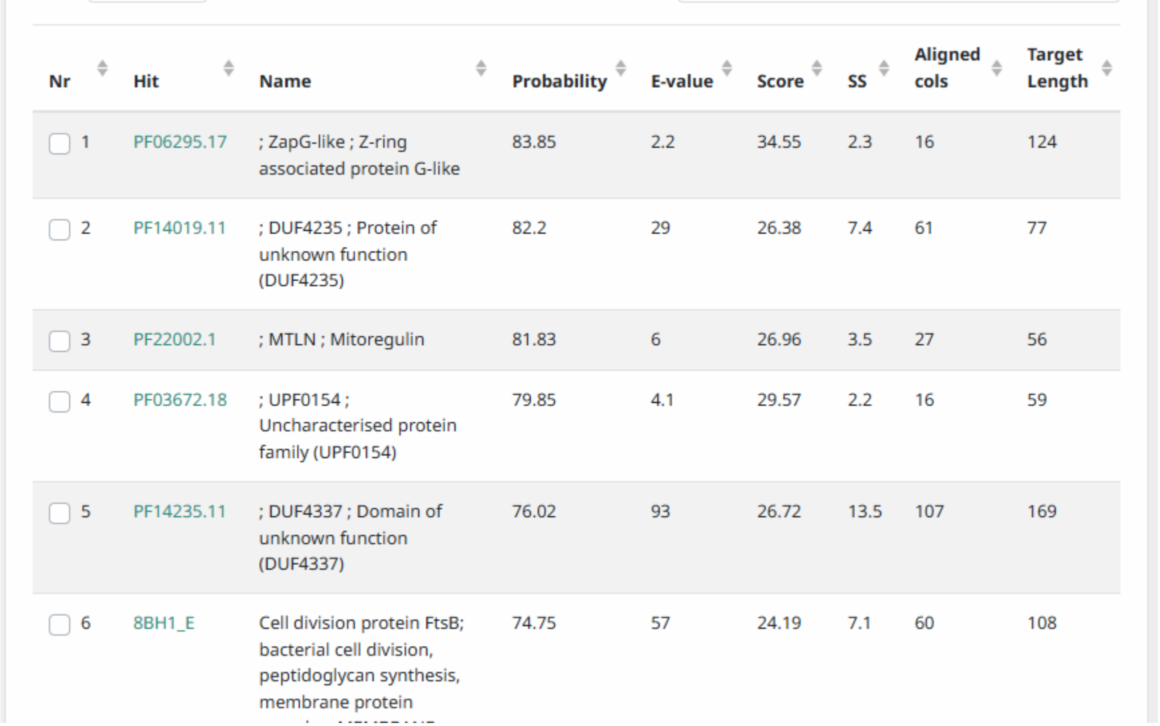
- 8 membrane protein
- 5 hypothetical protein



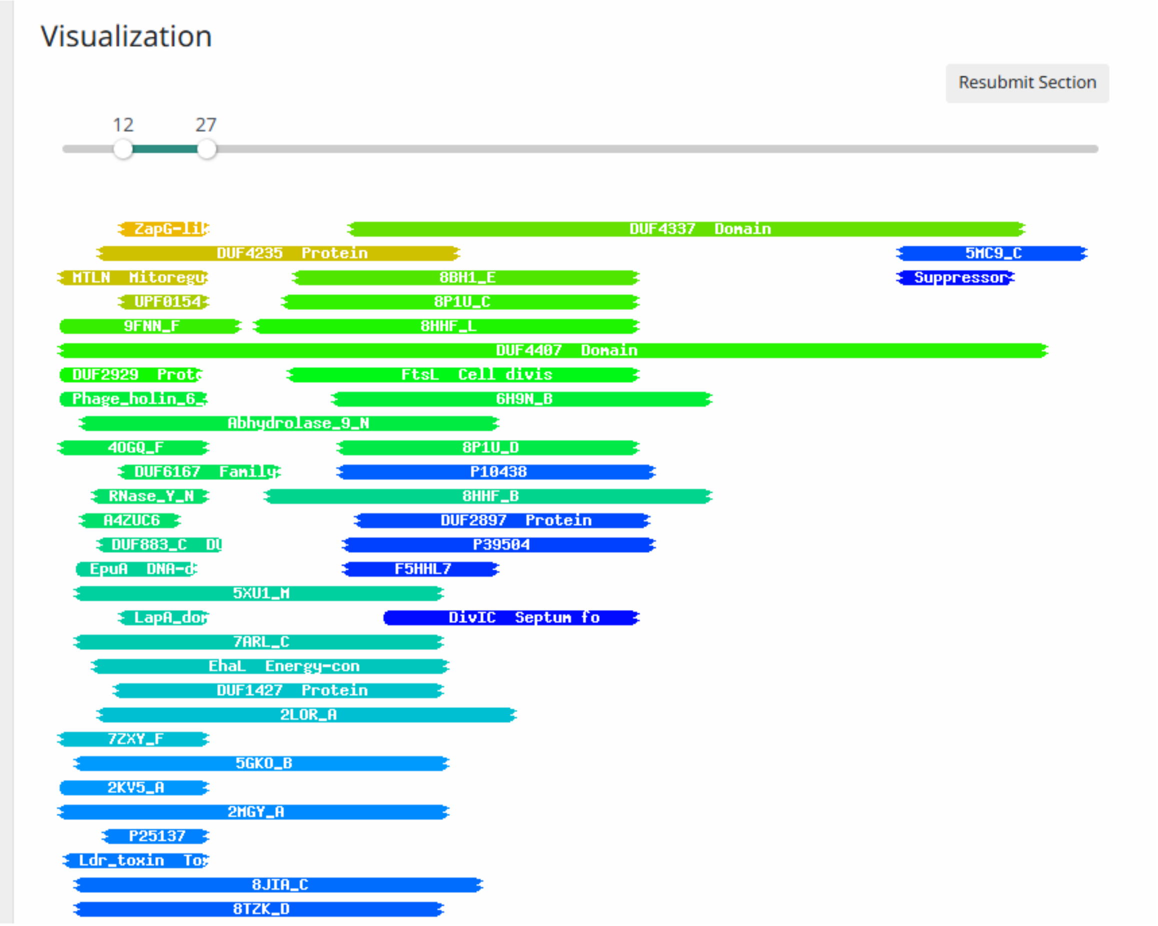
Score	Target Description
989	membrane protein [Gordonia phage Vine] >gb|QZ
981	hypothetical protein PP995_gp23 [Gordonia phag
977	membrane protein [Gordonia phage KayGee]
977	membrane protein [Gordonia phage Elinal]
963	membrane protein [Gordonia phage BigChungus]
898	membrane protein [Gordonia phage Pons] >ref|YP
895	hypothetical protein PP994_gp25 [Gordonia phag
856	membrane protein [Gordonia phage MAnor]
561	membrane protein [Gordonia phage SummitAcad
543	hypothetical protein PP996_gp23 [Gordonia phag
429	membrane protein [Gordonia phage Stormageddo
402	hypothetical protein SEA_SUMMITACADEMY_2
353	hypothetical protein PP996_gp24 [Gordonia phag

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- No hits as all probabilities are less than 90%.

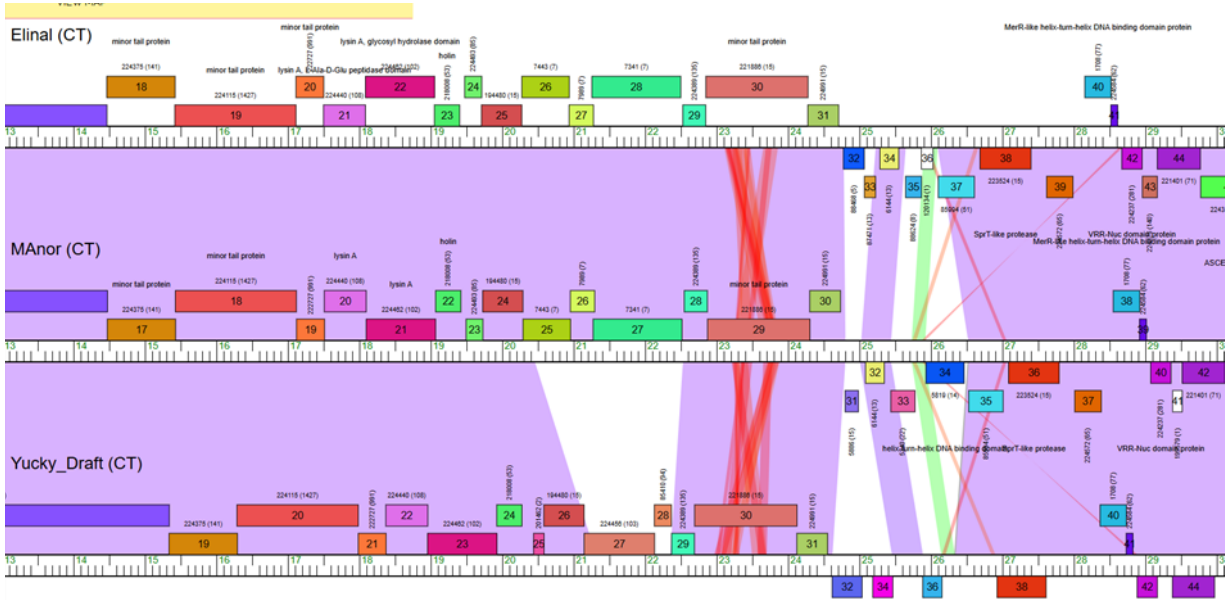


Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	PF06295.17	; ZapG-like ; Z-ring associated protein G-like	83.85	2.2	34.55	2.3	16	124
□ 2	PF14019.11	; DUF4235 ; Protein of unknown function (DUF4235)	82.2	29	26.38	7.4	61	77
□ 3	PF22002.1	; MTLN ; Mitoregulin	81.83	6	26.96	3.5	27	56
□ 4	PF03672.18	; UPF0154 ; Uncharacterised protein family (UPF0154)	79.85	4.1	29.57	2.2	16	59
□ 5	PF14235.11	; DUF4337 ; Domain of unknown function (DUF4337)	76.02	93	26.72	13.5	107	169
□ 6	8BH1_E	Cell division protein FtsB; bacterial cell division, peptidoglycan synthesis, membrane protein	74.75	57	24.19	7.1	60	108



Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

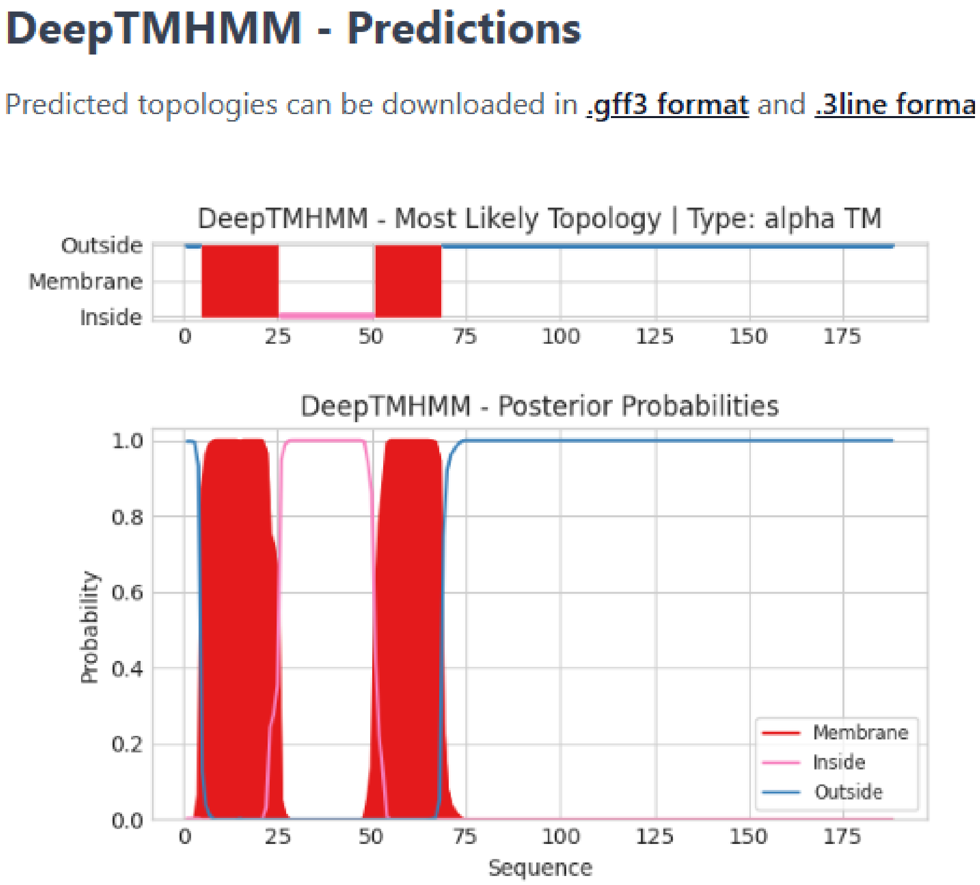
- Yucky feature 26 conserved domain: none function: none
- Elinal feature 25 conserved domain: none function: none
- MAnor feature 24 conserved domain: none function: none



Elinal (CT)
minor tail protein
lysin A, glycosyl hydrolase domain
lysin A, L-Ala-D-Glu peptidase domain
holin
MerR-like helix-turn-helix DNA binding domain protein
MAnor (CT)
lysin A
SprT-like protease
VRR-Nuc domain protein
Yucky_Draft (CT)
helix-turn-helix DNA binding domain

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- # Unnamed Number of predicted TMRs: 2



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line forma

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

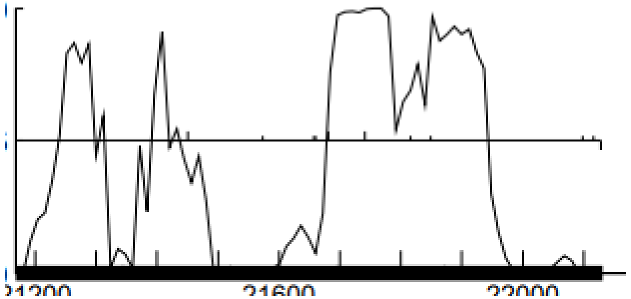
- The function is membrane protein because while no function was determined by Hhpred or Phamerator, BLAST did include 8 hits for membrane protein, and Deep TMHMM had 2 unnamed number of predicted TMRs.

Feature 25 – Stop 22131

Glimmer/GeneMark

What feature number is this?	• 25
What is the stop site?	• 22131
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?	• Both
What is the autoannotated start?	• 21145
Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start	• There is no gap or overlap they are adjacent. (Previous feture ends at 21144)

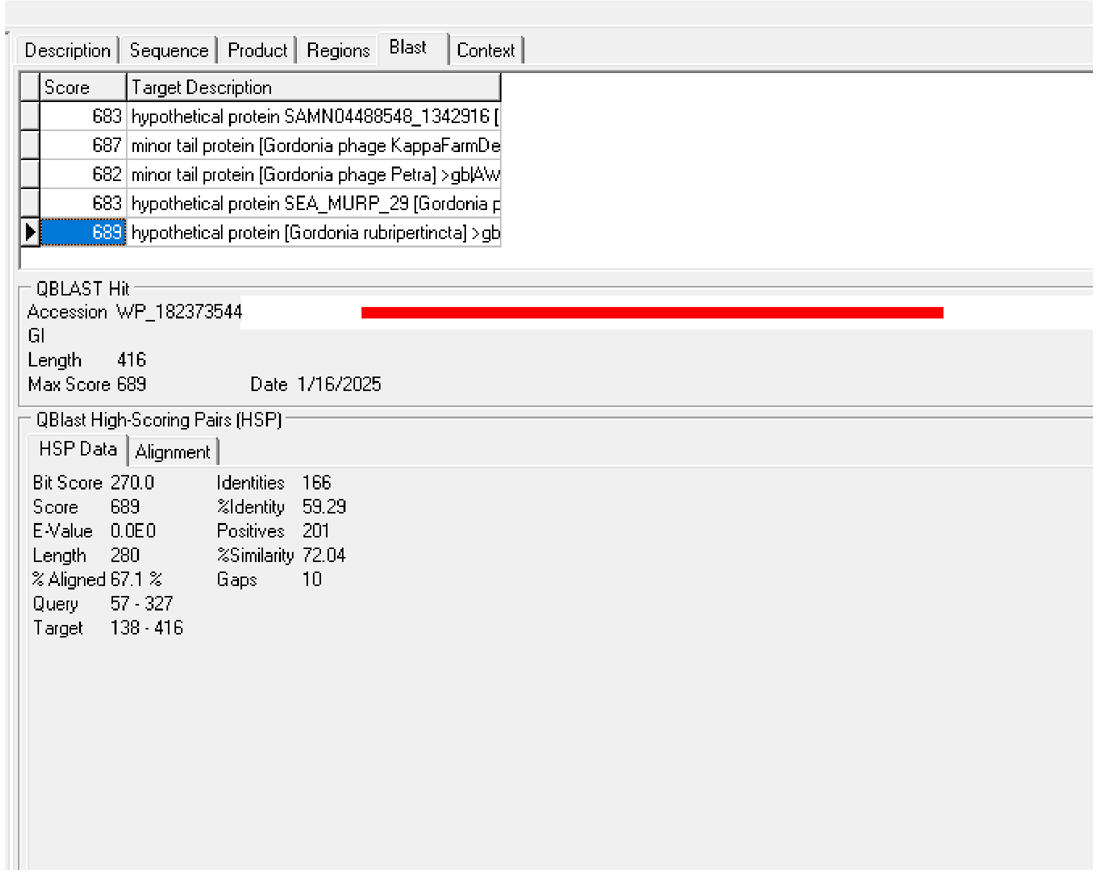
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- The coding potential for this graph is spread out but it has several peaks throughout it with areas that have none.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are 25 blast hits that have an e-value that is zero.



Score	Target Description
683	hypothetical protein SAMN04488548_1342916 [
687	minor tail protein [Gordonia phage KappaFarmDe
682	minor tail protein [Gordonia phage Petra] >gb|AW
683	hypothetical protein SEA_MURP_29 [Gordonia p
689	hypothetical protein [Gordonia rubripertincta] >gb

QBLAST Hit

Accession WP_182373544
GI
Length 416
Max Score 689 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	270.0	Identities	166
Score	689	%Identity	59.29
E-Value	0.0E0	Positives	201
Length	280	%Similarity	72.04
% Aligned	67.1 %	Gaps	10
Query	57 - 327		
Target	138 - 416		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

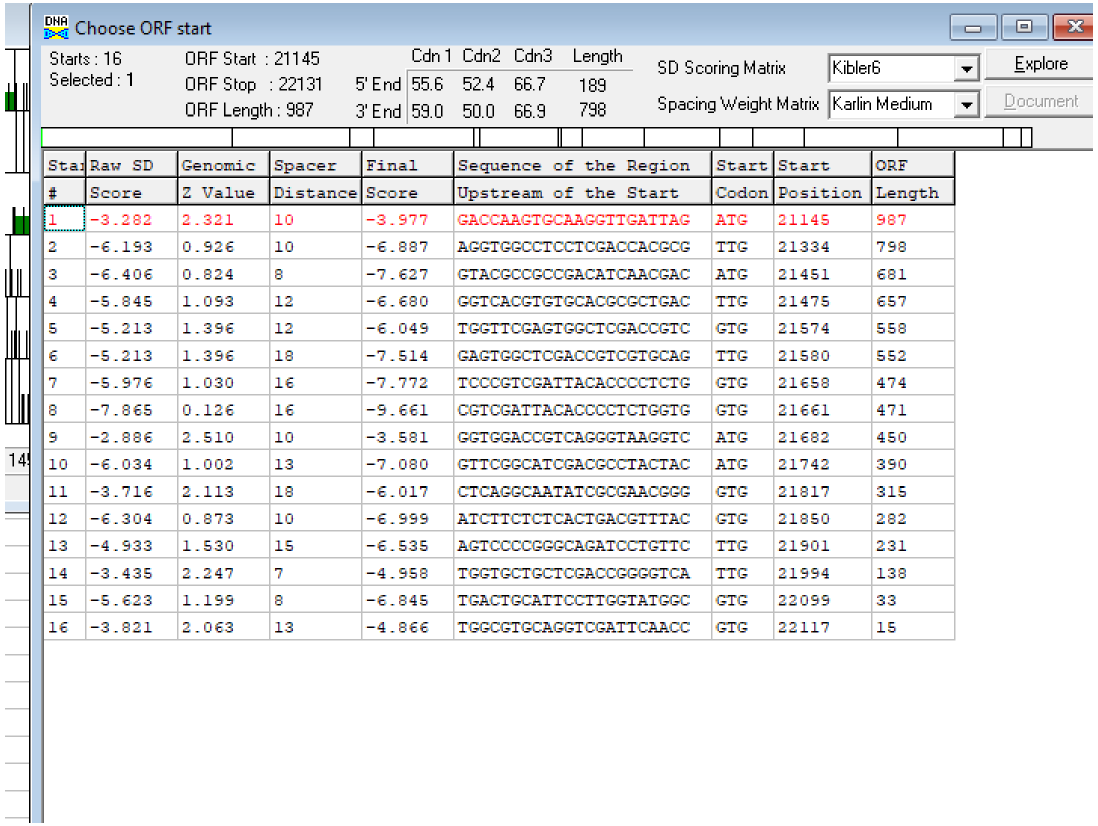
- Yes, this feature is a gene because it has a lot of coding potential, has over 25 blast hits that have an e-value of zero, and it was called by both genemark and glimmer.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- For the start of 21145 there are 5 1:1 blast hits which make this a great start site. There is no other compelling evidence for any of the other start sites so far

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts?   Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- The start of 21145 has
- Z-value:2.321
- FS:-3.977
- These are by far the best scores of all the other RBS values



Starts : 16
Selected : 1

ORF Start : 21145
ORF Stop : 22131
ORF Length : 987

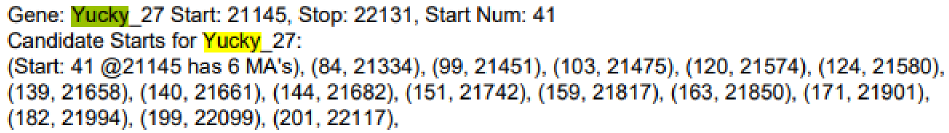
	Cdn 1	Cdn2	Cdn3	Length
5' End	55.6	52.4	66.7	189
3' End	59.0	50.0	66.9	798

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-3.282	2.321	10	-3.977	GACCAAGTGCAAGGTTGATTAG	ATG	21145	987
2	-6.193	0.926	10	-6.887	AGGTGGCCTCCTCGACCACGCG	TTG	21334	798
3	-6.406	0.824	8	-7.627	GTACGCCGCCGACATCAACGAC	ATG	21451	681
4	-5.845	1.093	12	-6.680	GGTCACGTGTGCACGCGCTGAC	TTG	21475	657
5	-5.213	1.396	12	-6.049	TGGTTCGAGTGGCTCGACCGTC	GTG	21574	558
6	-5.213	1.396	18	-7.514	GAGTGGCTCGACCGTCGTGCAG	TTG	21580	552
7	-5.976	1.030	16	-7.772	TCCCGTCGATTACACCCCTCTG	GTG	21658	474
8	-7.865	0.126	16	-9.661	CGTCGATTACACCCCTCTGGTG	GTG	21661	471
9	-2.886	2.510	10	-3.581	GGTGGACCGTCAGGGTAAGGTC	ATG	21682	450
10	-6.034	1.002	13	-7.080	GTTCGGCATCGACGCCTACTAC	ATG	21742	390
11	-3.716	2.113	18	-6.017	CTCAGGCAATATCGCGAACGGG	GTG	21817	315
12	-6.304	0.873	10	-6.999	ATCTTCTCTCACTGACGTTTAC	GTG	21850	282
13	-4.933	1.530	15	-6.535	AGTCCCCGGGCAGATCCTGTTC	TTG	21901	231
14	-3.435	2.247	7	-4.958	TGGTGCTGCTCGACCGGGGTCA	TTG	21994	138
15	-5.623	1.199	8	-6.845	TGACTGCATTCCTTGGTATGGC	GTG	22099	33
16	-3.821	2.063	13	-4.866	TGGCGTGCAGGTCGATTCAACC	GTG	22117	15

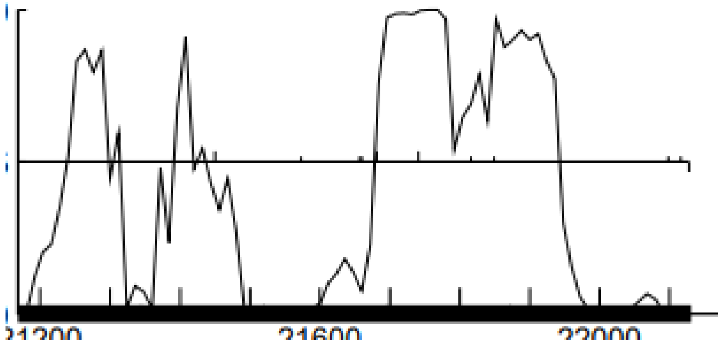
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- The start of 21145 has 6 MA's and is the only start site with MA's so this is the best option



Gene: Yucky_27 Start: 21145, Stop: 22131, Start Num: 41
Candidate Starts for Yucky_27:
(Start: 41 @21145 has 6 MA's), (84, 21334), (99, 21451), (103, 21475), (120, 21574), (124, 21580),
(139, 21658), (140, 21661), (144, 21682), (151, 21742), (159, 21817), (163, 21850), (171, 21901),
(182, 21994), (199, 22099), (201, 22117),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- The start site of 21145 includes all of the coding potential of the feature

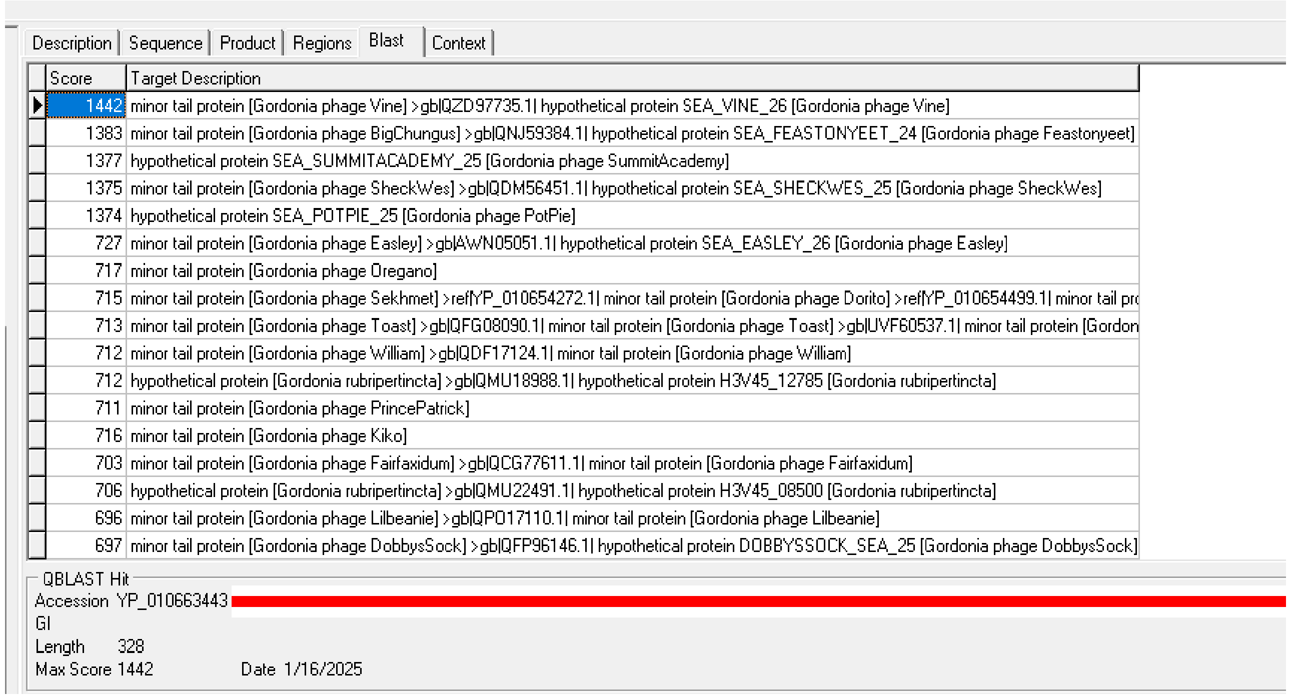
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- There was no gap overlap for this feature as it and the feature before it are adjacent to each other.

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start for feature 27 is 21145 because of it having 5 1:1 blast alignments, the start including all of the coding potential, having no gap/overlap being adjacent to the feature before, and having 6 MA’s in starterator.

BLAST function evidence. What assigned functions do other highly similar genes have?



<table>
<tr><th>Score</th><th>Target Description</th></tr>
<tr><td>1442</td><td>minor tail protein [Gordonia phage Vine] >gb|QZD97735.1| hypothetical protein SEA_VINE_26 [Gordonia phage Vine]</td></tr>
<tr><td>1383</td><td>minor tail protein [Gordonia phage BigChungus] >gb|QNJ59384.1| hypothetical protein SEA_FEASTONYEET_24 [Gordonia phage Feastonyeet]</td></tr>
<tr><td>1377</td><td>hypothetical protein SEA_SUMMITACADEMY_25 [Gordonia phage SummitAcademy]</td></tr>
<tr><td>1375</td><td>minor tail protein [Gordonia phage SheckWes] >gb|QDM56451.1| hypothetical protein SEA_SHECKWES_25 [Gordonia phage SheckWes]</td></tr>
<tr><td>1374</td><td>hypothetical protein SEA_POTPIE_25 [Gordonia phage PotPie]</td></tr>
<tr><td>727</td><td>minor tail protein [Gordonia phage Easley] >gb|AWN05051.1| hypothetical protein SEA_EASLEY_26 [Gordonia phage Easley]</td></tr>
<tr><td>717</td><td>minor tail protein [Gordonia phage Oregano]</td></tr>
<tr><td>715</td><td>minor tail protein [Gordonia phage Sekhmet] >ref|YP_010654272.1| minor tail protein [Gordonia phage Dorito] >ref|YP_010654499.1| minor tail pr</td></tr>
<tr><td>713</td><td>minor tail protein [Gordonia phage Toast] >gb|QFG08090.1| minor tail protein [Gordonia phage Toast] >gb|UVF60537.1| minor tail protein [Gordon</td></tr>
<tr><td>712</td><td>minor tail protein [Gordonia phage William] >gb|QDF17124.1| minor tail protein [Gordonia phage William]</td></tr>
<tr><td>712</td><td>hypothetical protein [Gordonia rubripertincta] >gb|QMU18988.1| hypothetical protein H3V45_12785 [Gordonia rubripertincta]</td></tr>
<tr><td>711</td><td>minor tail protein [Gordonia phage PrincePatrick]</td></tr>
<tr><td>716</td><td>minor tail protein [Gordonia phage Kiko]</td></tr>
<tr><td>703</td><td>minor tail protein [Gordonia phage Fairfaxidum] >gb|QCG77611.1| minor tail protein [Gordonia phage Fairfaxidum]</td></tr>
<tr><td>706</td><td>hypothetical protein [Gordonia rubripertincta] >gb|QMU22491.1| hypothetical protein H3V45_08500 [Gordonia rubripertincta]</td></tr>
<tr><td>696</td><td>minor tail protein [Gordonia phage Lilbeanie] >gb|QPO17110.1| minor tail protein [Gordonia phage Lilbeanie]</td></tr>
<tr><td>697</td><td>minor tail protein [Gordonia phage DobbysSock] >gb|QFP96146.1| hypothetical protein DOBBYSSOCK_SEA_25 [Gordonia phage DobbysSock]</td></tr>
</table>

QBLAST Hit

Accession YP_010663443

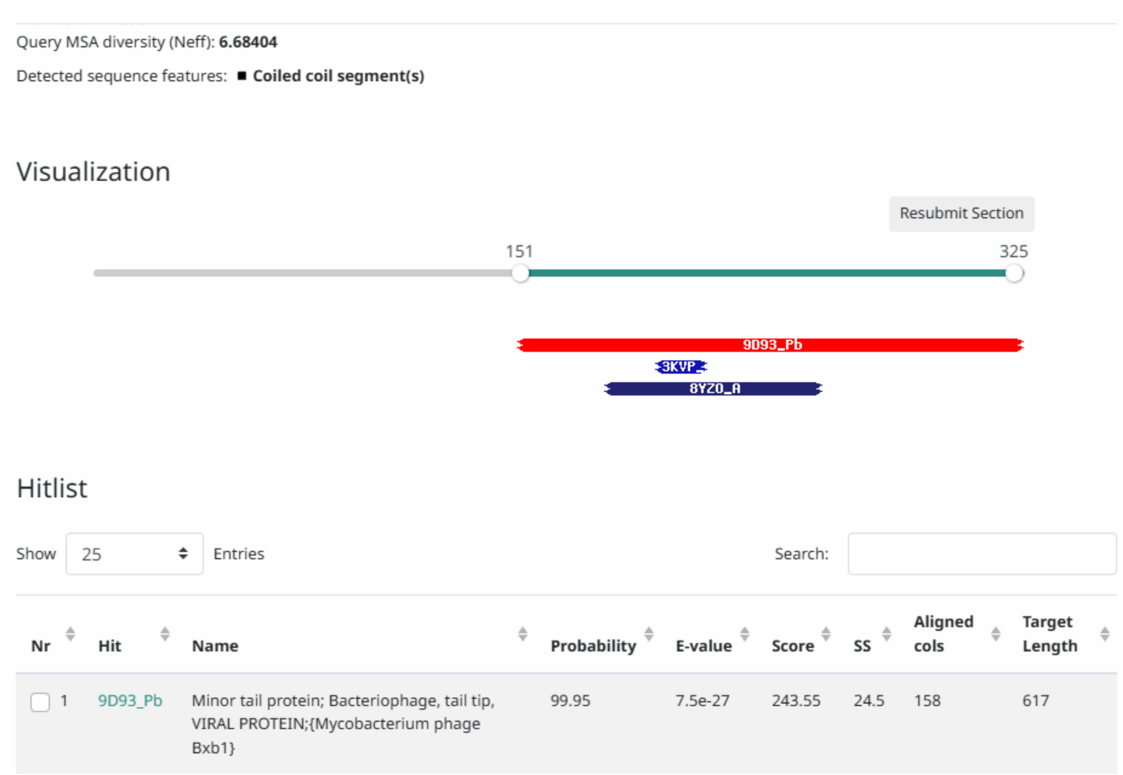
GI

Length 328

Max Score 1442 Date 1/16/2025

- Blast shows evidence that this may be a minor tail protein because a lot of other similar genes like Vine have this as a function for this feature

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



Query MSA diversity (Neff): **6.68404**

Detected sequence features: ■ **Coiled coil segment(s)**

Visualization

Hitlist

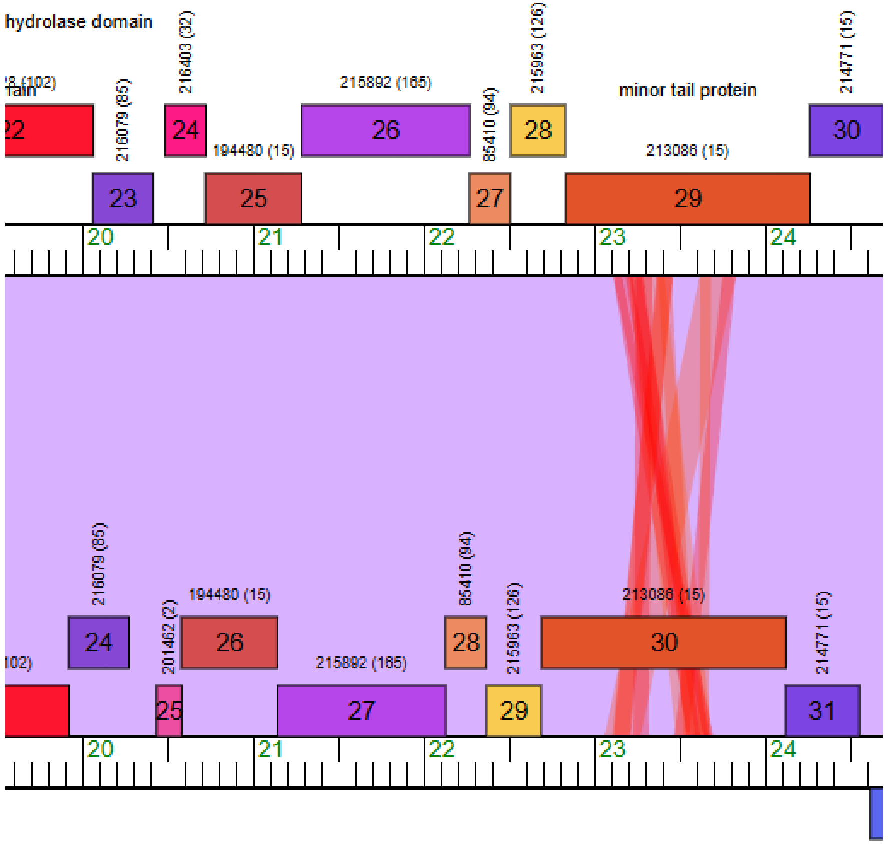
Show 25 Entries

Search:

Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	9D93_Pb	Minor tail protein; Bacteriophage, tail tip, VIRAL PROTEIN;{Mycobacterium phage Bxb1}	99.95	7.5e-27	243.55	24.5	158	617

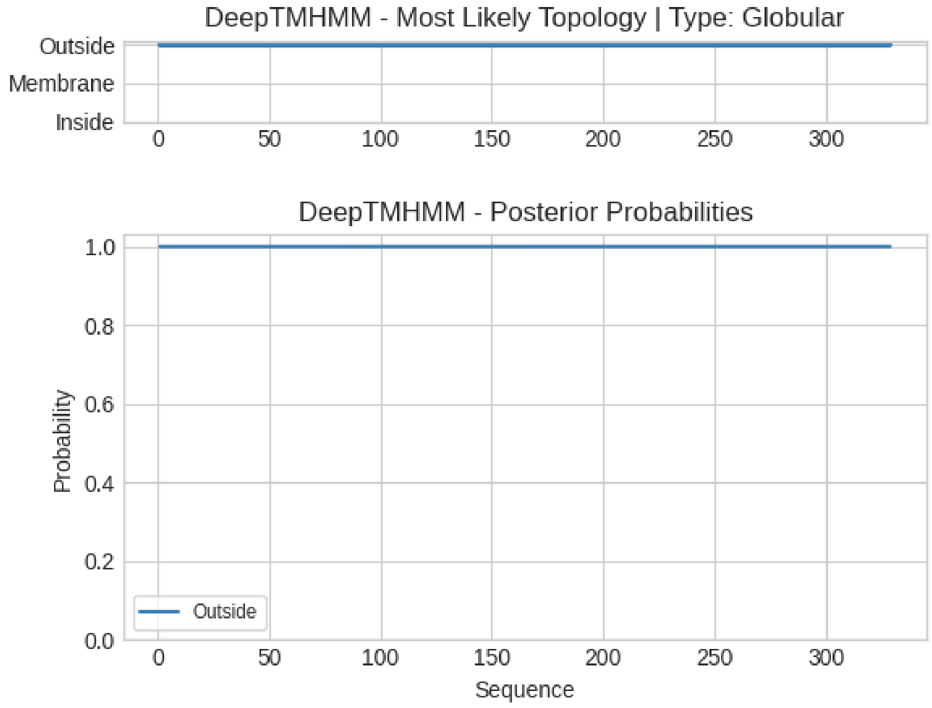
- HHPRED gave evidence that this may be a minor tail protein. The coding with also rich with glycine which gives further evidence that it could be a minor tail protein

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- Phamerator gives no evidence of a function for this gene due to no conserved domains popping up and it doesn’t have a name corresponding to the colored block that it relates to. This gene is close to another gene that has the function of minor tail protein which is 29 on the top (Vine) which gives evidence it may be a minor tail protein.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- There is no evidence that supports a function here since there are no transmembrane domain hits

This feature is a hypothetical protein

- There is no compelling evidence that this has a function, and it is not a transmembrane domain

Feature 26 – Stop 22367

Instructions

Fill this out for each gene you annotate. This should be thought of as the minimum amount of information that needs to be provided for each gene. You can always add more slides or information as necessary

- Is it a gene?
    - Yes!
- Where does it start?
    - 22128!
- What is the function?
    - Hypothetical Protein

Glimmer/GeneMark

What feature number is this? **26**

What is the stop site? **22367**

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? **Called by both Glimmer and GeneMark**

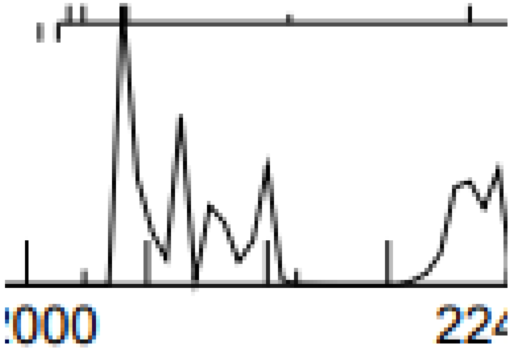
What is the autoannotated start? **22128**

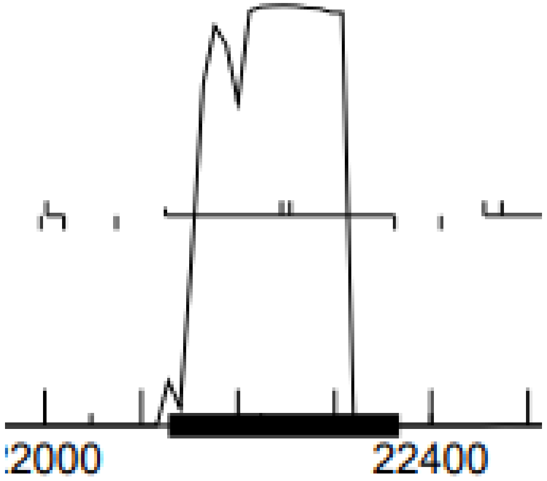
Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

Overlap of 4 nucleotides

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

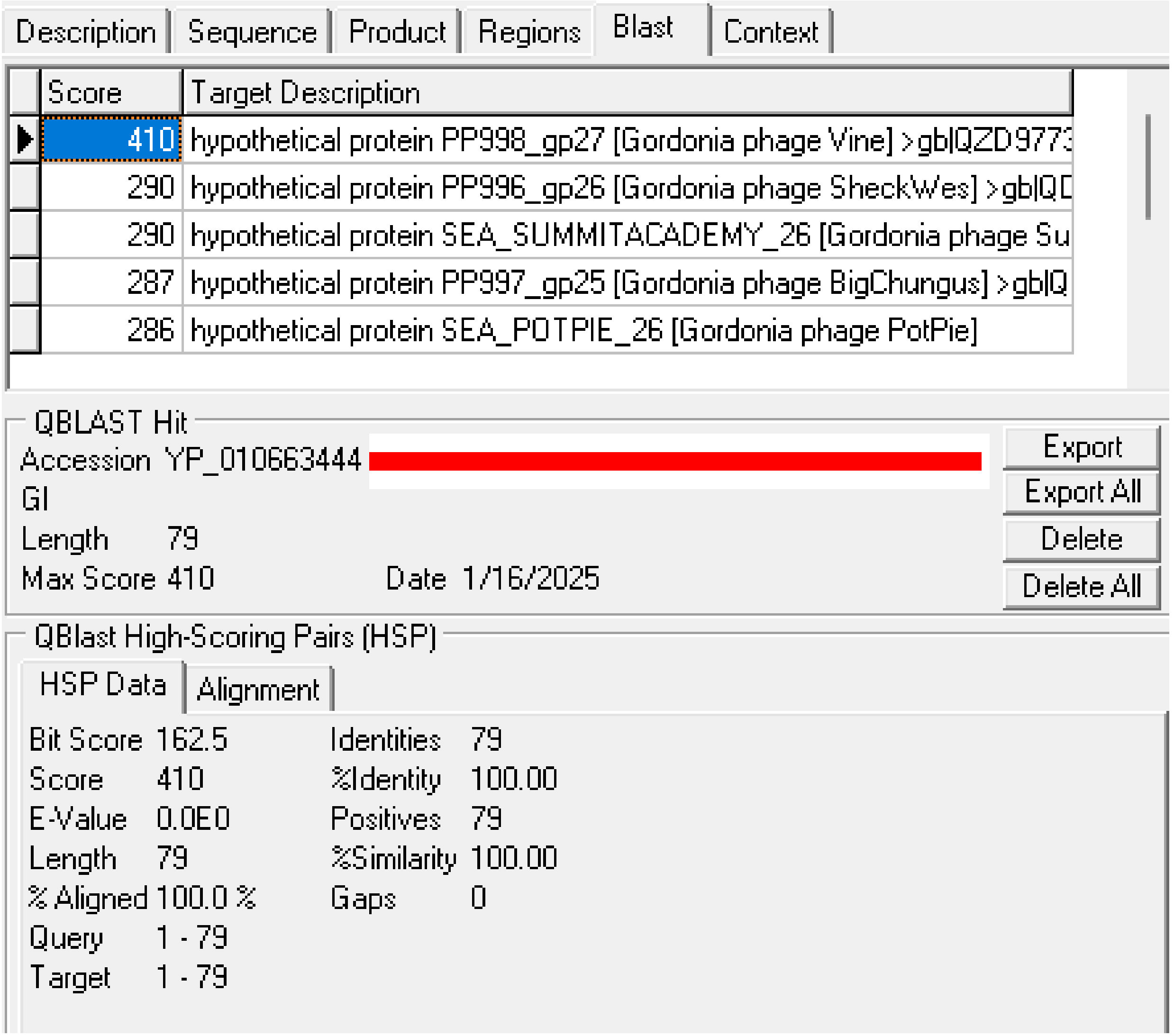
- There is coding potential throughout where the gene is supposed to be starting off weak at 22128 and then peaking to strong potential around 22150 before dropping of a small amount. The coding potential then remains strong until it drops off at the stop of 22367.
- Another reading frame has some coding potential, but it is not consistent throughout where the gene is supposed to be.





BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There were 5 1:1 Alignments
- There were seven BLAST hits of phages with genes highly similar to this feature.
- All BLAST hits had e-values that were relatively close to zero or zero.



Score	Target Description
410	hypothetical protein PP998_gp27 [Gordonia phage Vine] >gb|QZD977
290	hypothetical protein PP996_gp26 [Gordonia phage SheckWes] >gb|QD
290	hypothetical protein SEA_SUMMITACADEMY_26 [Gordonia phage Su
287	hypothetical protein PP997_gp25 [Gordonia phage BigChungus] >gb|Q
286	hypothetical protein SEA_POTPIE_26 [Gordonia phage PotPie]

Accession YP_010663444
GI
Length 79
Max Score 410
Date 1/16/2025

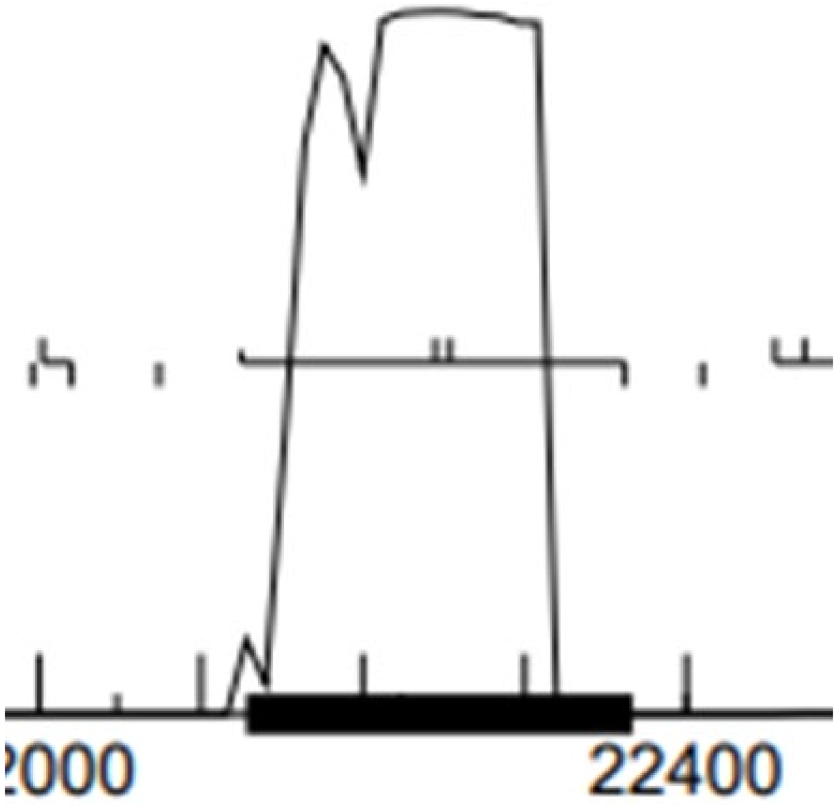
Bit Score	162.5	Identities	79
Score	410	%Identity	100.00
E-Value	0.0E0	Positives	79
Length	79	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 79		
Target	1 - 79		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This feature is a gene! There is strong coding potential throughout where the gene is called to be, and there are several BLAST hits of phages with genes that are highly similar to this feature with e-values close to zero.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Starting at 22128:
  - If the gene starts at 22128, then a small part of the initial peak would be lost. A majority of the coding potential would be included based on this starting point.



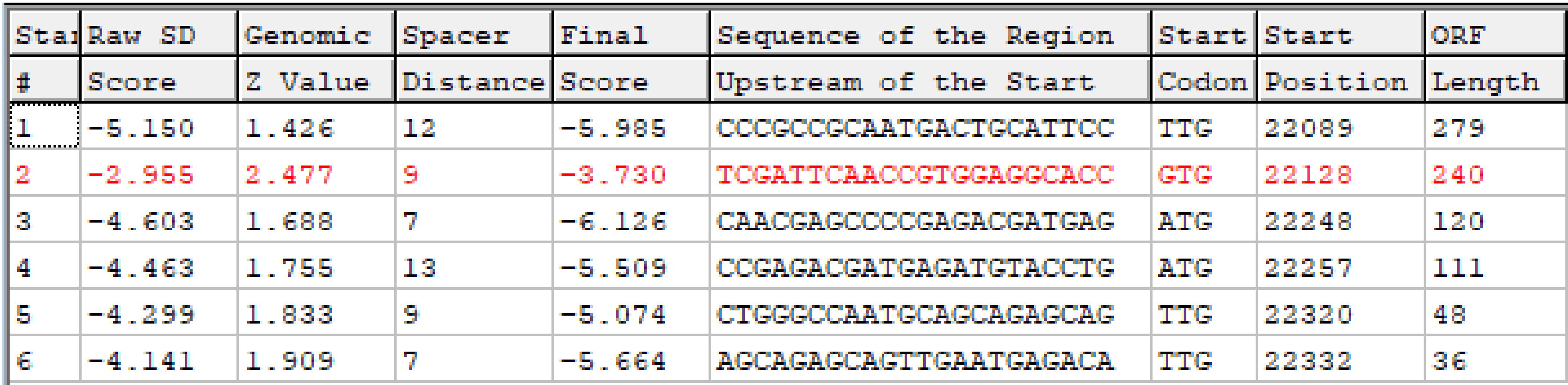
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts?   Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

Starting at 22128:

z-value = 2.477

final score = -3.730

- This is the only proposed start based of the evidence, so it is favored.



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	-5.150	1.426	12	-5.985	CCCGCCGCAATGACTGCATTCC	TTG	22089	279
2	-2.955	2.477	9	-3.730	TCGATTCAACCGTGGAGGCACC	GTG	22128	240
3	-4.603	1.688	7	-6.126	CAACGAGCCCCGAGACGATGAG	ATG	22248	120
4	-4.463	1.755	13	-5.509	CCGAGACGATGAGATGTACCTG	ATG	22257	111
5	-4.299	1.833	9	-5.074	CTGGGCCAATGCAGCAGAGCAG	TTG	22320	48
6	-4.141	1.909	7	-5.664	AGCAGAGCAGTTGAATGAGACA	TTG	22332	36

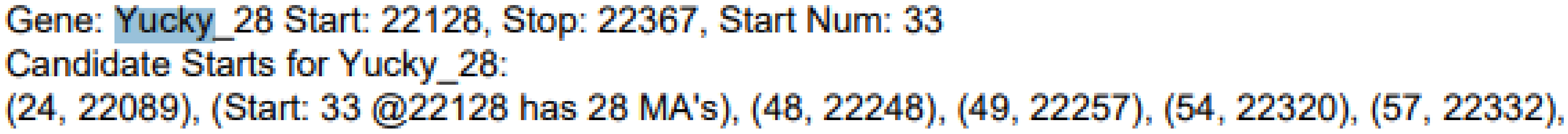
BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- Starting at 22128:
  - There are 5 1:1 alignments of other highly similar genes with the start of this predicted start based of the 7 BLAST hits.

This is the only proposed start based of the evidence, so it is favored.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Yucky has the Most Annotated start for this pham and it is called 49.2% of the time when present
- 33 MA's for this start (only start for this gene that has manual annotations)
- 22128 is the only proposed start suggested by the Starterator report.



Gene: Yucky_28 Start: 22128, Stop: 22367, Start Num: 33
Candidate Starts for Yucky_28:
(24, 22089), (Start: 33 @22128 has 28 MA's), (48, 22248), (49, 22257), (54, 22320), (57, 22332),

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Starting at 22128 would leave an overlap of 4 nucleotides with the previous feature.

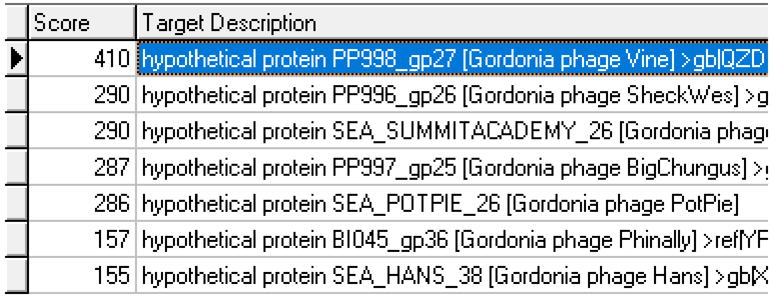
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	Start 22128
Glimmer/GeneMark	Glimmer & GeneMark
Coding Potential	Starting at 22128 would result in the loss of a small portion of the initial small peak of coding potential
RBS	z-value = 2.477 Final score = -3.370
BLAST	5 1:1 Alignments
Starterator	33 MA's
Gap/Overlap	Overlap of 4 nucleotides

The start is 22128! This was; however, the only proposed start based off all the evidence. 22128 was called as the start of this gene by Glimmer and GeneMark, and by starting at this nucleotide only a small portion of the initial peak of coding potential is lost. At this starting point a z-value of 2.477 and a final score of -3.370 were given. There were 5 1:1 alignments according to BLAST of phages with highly similar genes, and the Starterator report showed 33 manual annotations for starting at 22128. There would be an overlap of 4 nucleotides with the previous gene

BLAST function evidence. What assigned functions do other highly similar genes have?

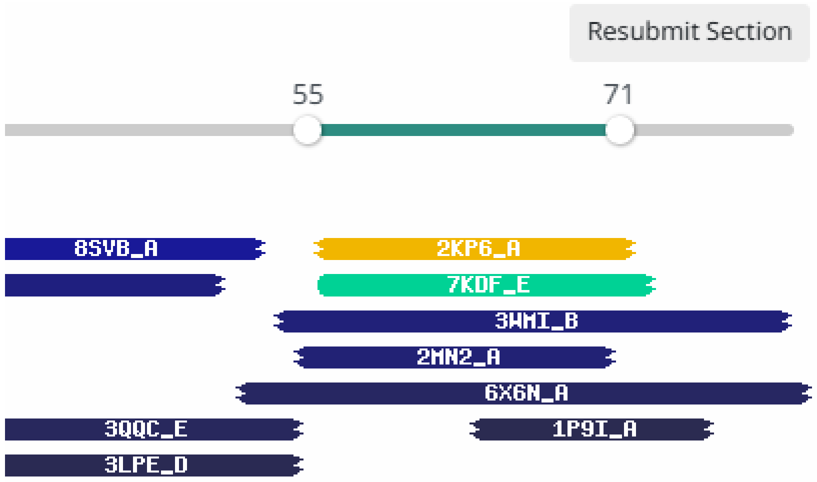
- There were 7 BLAST hits that all had functions labeled as hypothetical protein.

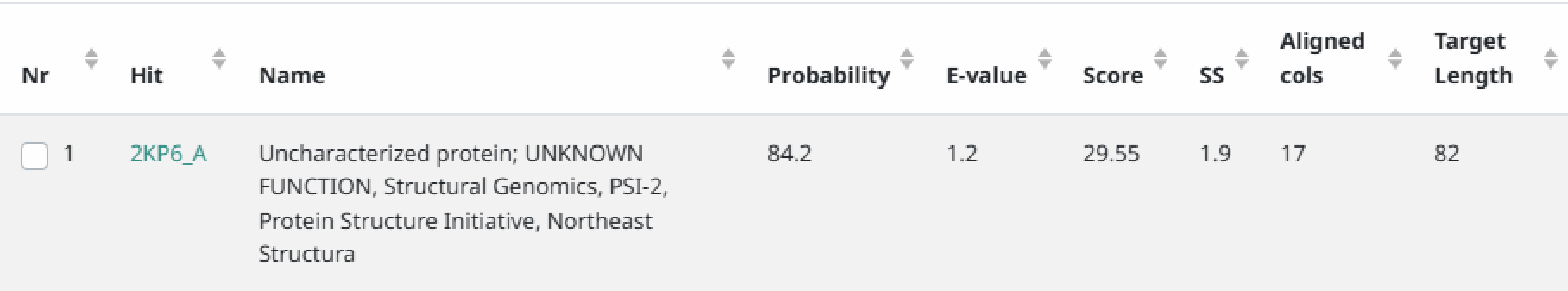


Score	Target Description
410	hypothetical protein PP998_gp27 [Gordonia phage Vine] >gb|QZD
290	hypothetical protein PP996_gp26 [Gordonia phage SheckWes] >g
290	hypothetical protein SEA_SUMMITACADEMY_26 [Gordonia phag
287	hypothetical protein PP997_gp25 [Gordonia phage BigChungus] >
286	hypothetical protein SEA_POTPIE_26 [Gordonia phage PotPie]
157	hypothetical protein BI045_gp36 [Gordonia phage Phinally] >ref|YF
155	hypothetical protein SEA_HANS_38 [Gordonia phage Hans] >gb|X

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- The highest probability hit according to HHpred was labeled as 84.2 with function labeled as “uncharacterized protein”, and none of the hits regardless of their probability value matched up with more than a small portion of the gene.

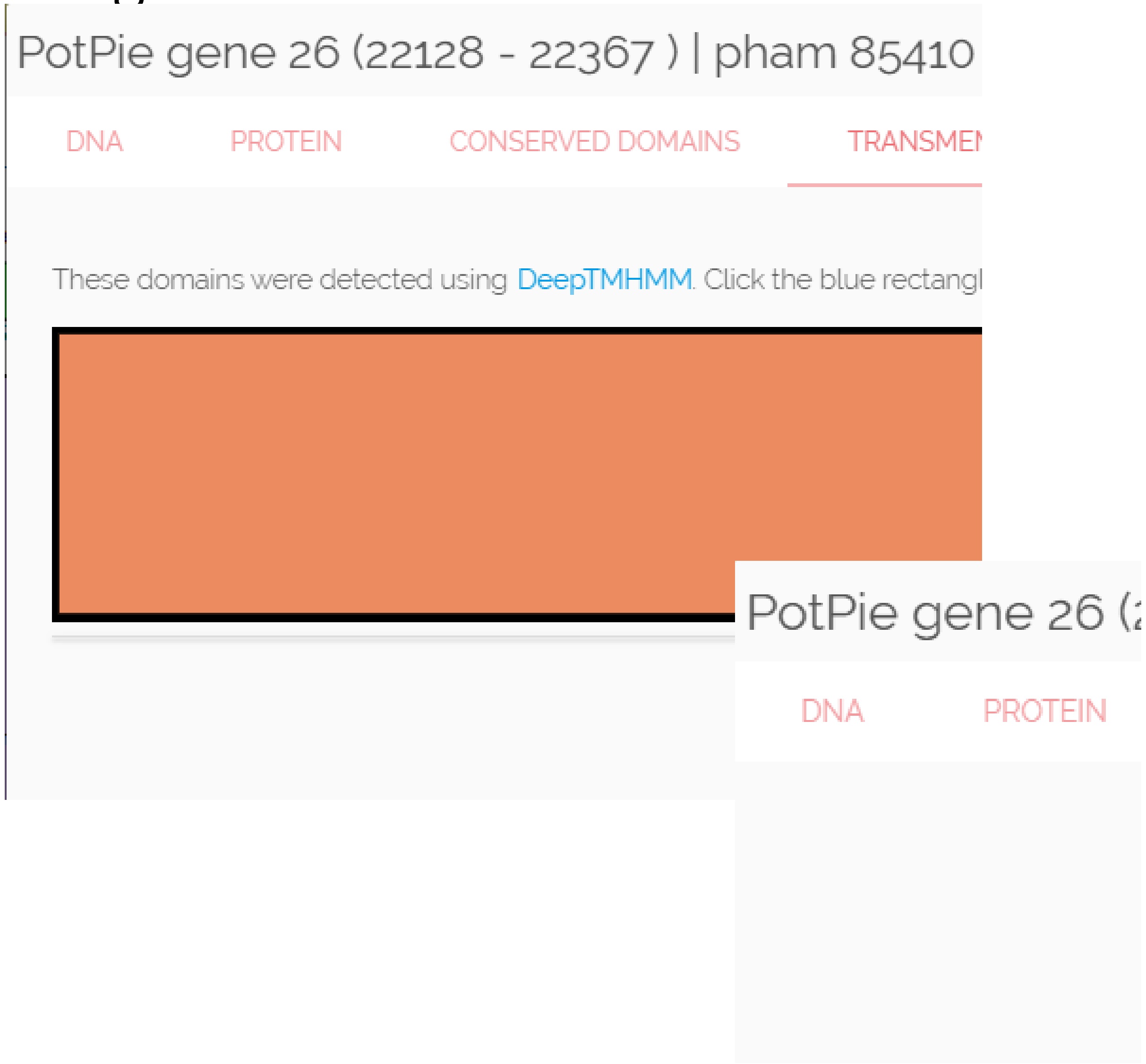




Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
☐ 1	2KP6_A	Uncharacterized protein; UNKNOWN FUNCTION, Structural Genomics, PSI-2, Protein Structure Initiative, Northeast Structura	84.2	1.2	29.55	1.9	17	82

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- None of the closely related phages with genes in the same pham predict a function for this gene and there are no conserved domains.
- This evidence supports the function of this gene being labeled as hypothetical protein.



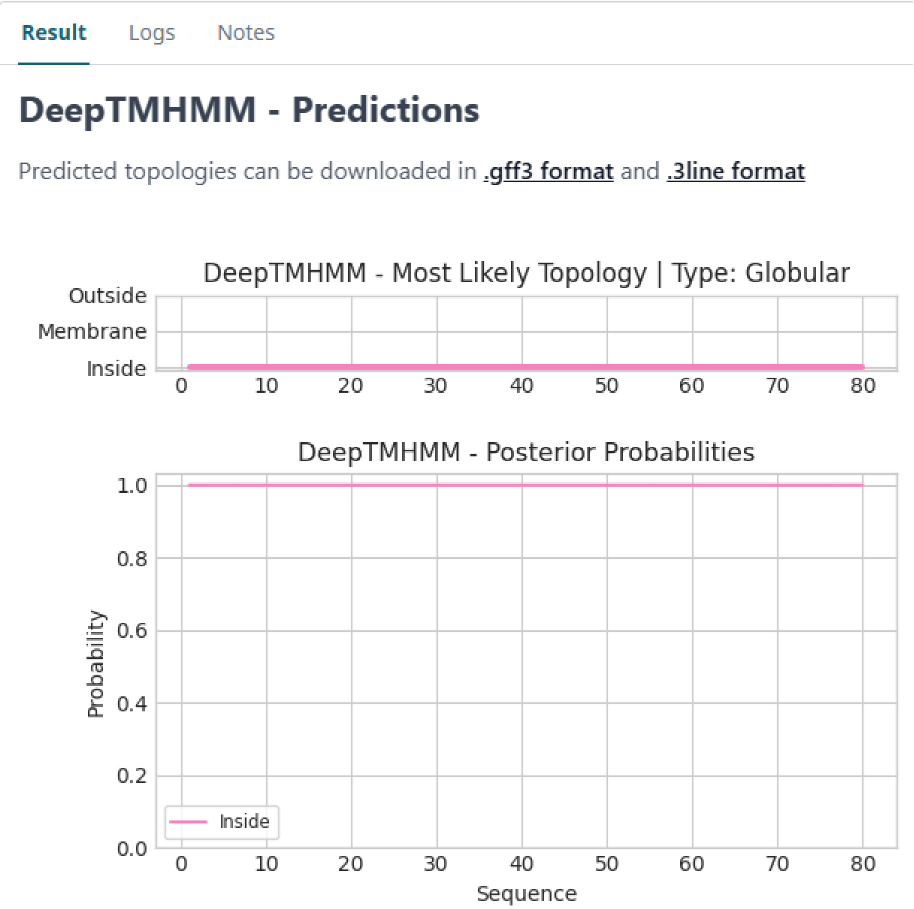
PotPie gene 26 (22128 - 22367) | pham 85410

These domains were detected using DeepTMHMM. Click the blue rectangl

PotPie gene 26 (

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- According the results from Deep TMHMM there are no transmembrane domains.



Result Logs Notes

DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

DeepTMHMM - Most Likely Topology | Type: Globular
Outside
Membrane
Inside
0
10
20
30
40
50
60
70
80
DeepTMHMM - Posterior Probabilities
Probability
1.0
0.8
0.6
0.4
0.2
0.0
Inside
Sequence

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function for this gene should be labeled as hypothetical protein which is also the official SEA-PHAGES function that should be assigned to this gene. All of the 7 BLAST hits for this gene had functions labeled as hypothetical protein, and the HHpred results do not support this gene having any alternative function to being labeled as a hypothetical protein as the highest probability hit was 84.2 and was also labeled as having an unknown function. The Phamerator map of phages with genes in the same pham as this one have no conserved domains or official function assigned.

Feature 27 – Stop 22689

Instructions

Fill this out for each gene you annotate. This should be thought of as the minimum amount of information that needs to be provided for each gene. You can always add more slides or information as necessary

- Is it a gene?
  - Yes!
- Where does it start?
  - 22128!
- What is the function?
  - Hypothetical Protein

Glimmer/GeneMark

What feature number is this? **27**

What is the stop site? **22367**

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? **Called by both Glimmer and GeneMark**

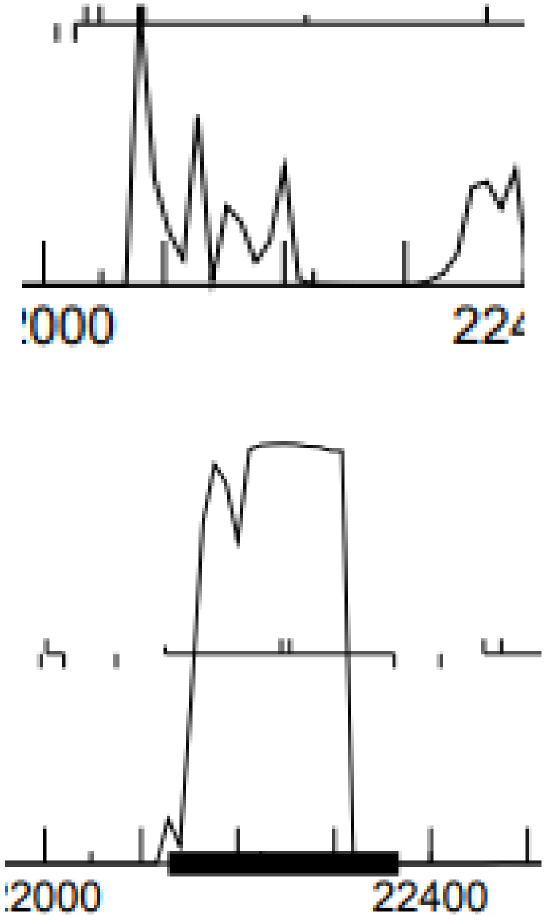
What is the autoannotated start? **22128**

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

Overlap of 4 nucleotides

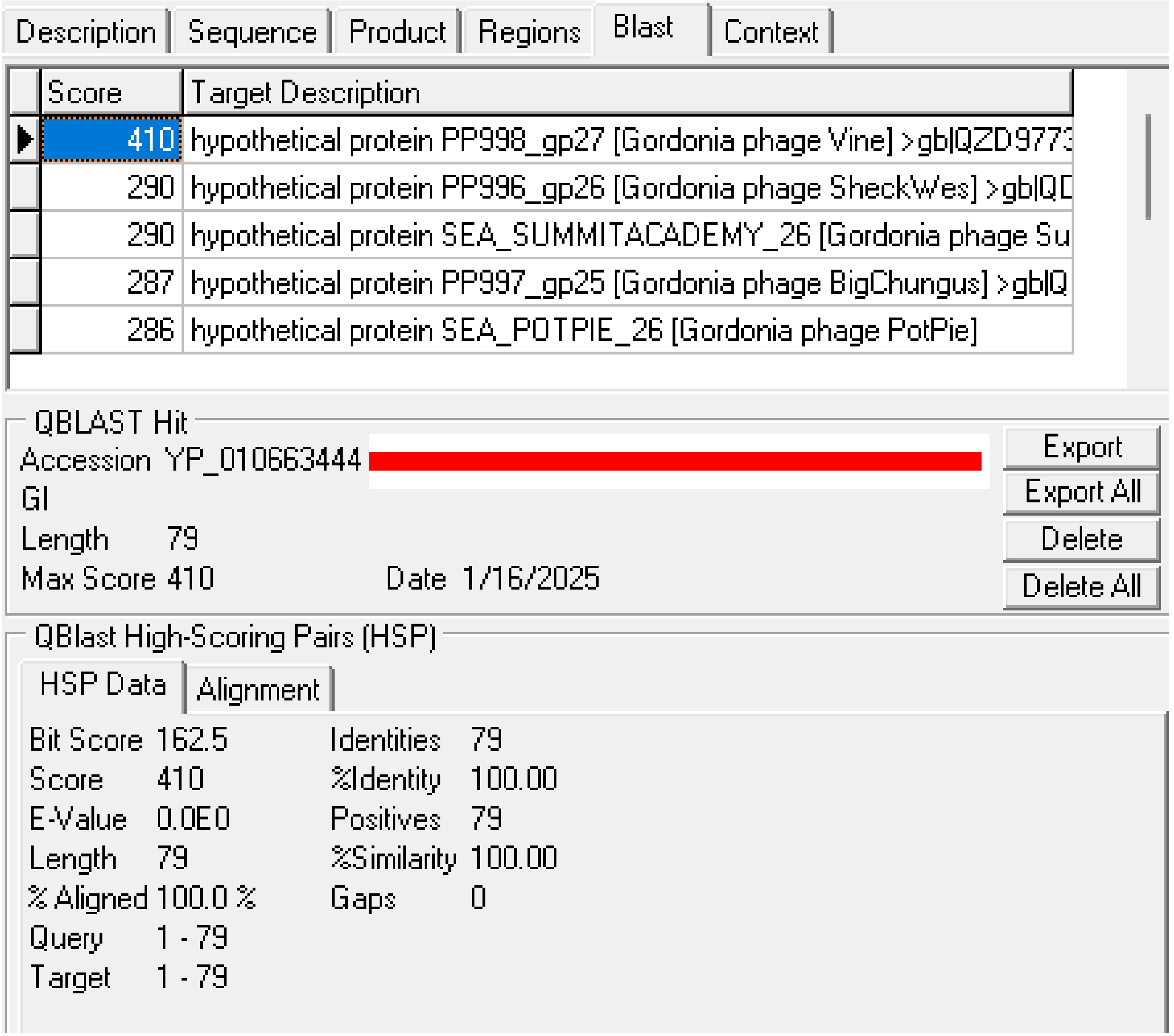
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- There is coding potential throughout where the gene is supposed to be starting off weak at 22128 and then peaking to strong potential around 22150 before dropping of a small amount. The coding potential then remains strong until it drops off at the stop of 22367.
- Another reading frame has some coding potential, but it is not consistent throughout where the gene is supposed to be.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There were 5 1:1 Alignments
- There were seven BLAST hits of phages with genes highly similar to this feature.
- All BLAST hits had e-values that were relatively close to zero or zero.



Score	Target Description
410	hypothetical protein PP998_gp27 [Gordonia phage Vine] >gb|QZD977
290	hypothetical protein PP996_gp26 [Gordonia phage SheckWes] >gb|QD
290	hypothetical protein SEA_SUMMITACADEMY_26 [Gordonia phage Su
287	hypothetical protein PP997_gp25 [Gordonia phage BigChungus] >gb|Q
286	hypothetical protein SEA_POTPIE_26 [Gordonia phage PotPie]

QBLAST Hit

Accession YP_010663444

GI

Length 79

Max Score 410 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

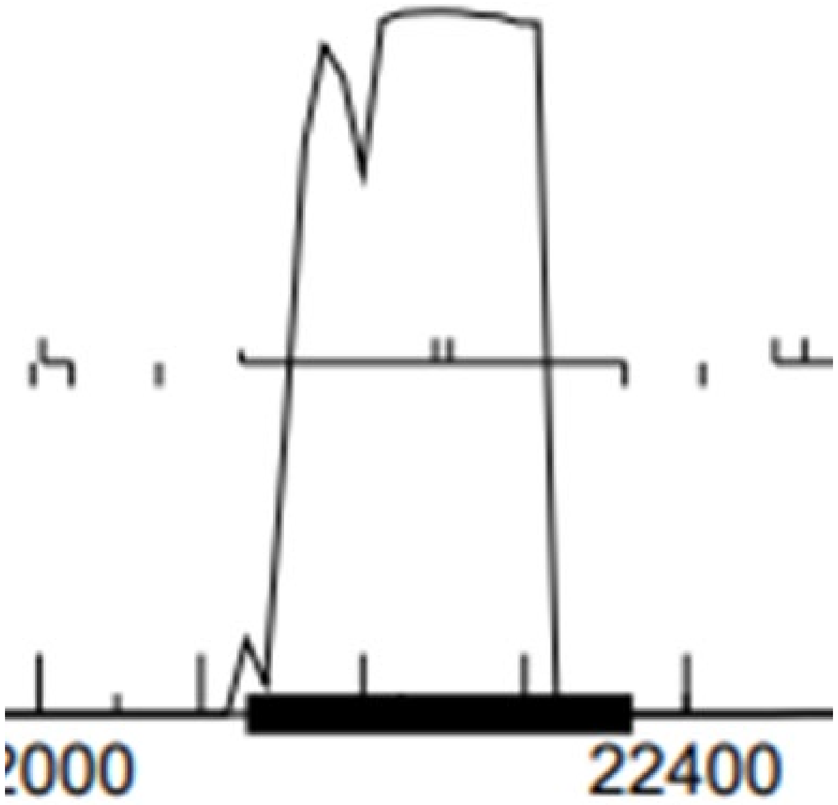
Bit Score	162.5	Identities	79
Score	410	%Identity	100.00
E-Value	0.0E0	Positives	79
Length	79	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 79		
Target	1 - 79		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This feature is a gene! There is strong coding potential throughout where the gene is called to be, and there are several BLAST hits of phages with genes that are highly similar to this feature with e-values close to zero.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Starting at 22128:
  - If the gene starts at 22128, then a small part of the initial peak would be lost. A majority of the coding potential would be included based on this starting point.



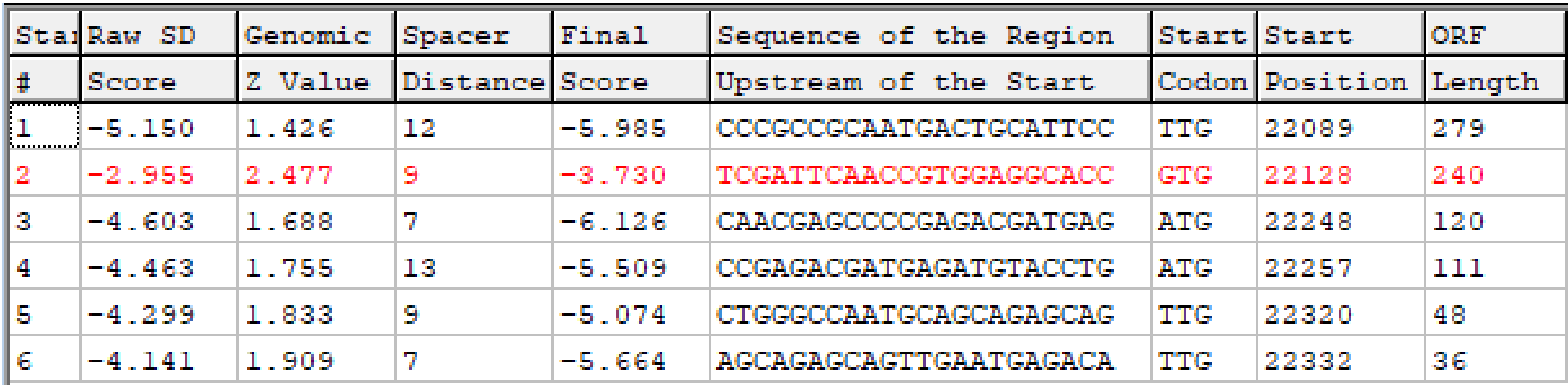
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts?   Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

Starting at 22128:

z-value = 2.477

final score = -3.730

- This is the only proposed start based of the evidence, so it is favored.



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	-5.150	1.426	12	-5.985	CCCGCCGCAATGACTGCATTCC	TTG	22089	279
2	-2.955	2.477	9	-3.730	TCGATTCAACCGTGGAGGCACC	GTG	22128	240
3	-4.603	1.688	7	-6.126	CAACGAGCCCCGAGACGATGAG	ATG	22248	120
4	-4.463	1.755	13	-5.509	CCGAGACGATGAGATGTACCTG	ATG	22257	111
5	-4.299	1.833	9	-5.074	CTGGGCCAATGCAGCAGAGCAG	TTG	22320	48
6	-4.141	1.909	7	-5.664	AGCAGAGCAGTTGAATGAGACA	TTG	22332	36

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

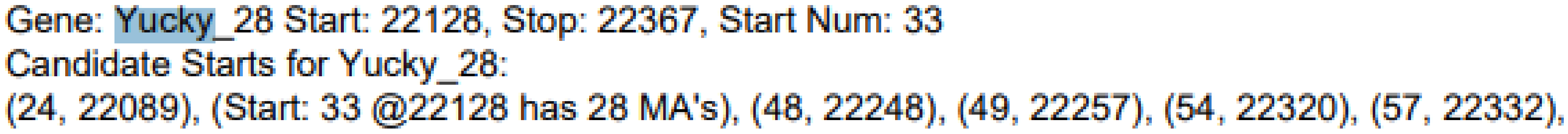
- Starting at 22128:
  - There are 5 1:1 alignments of other highly similar genes with the start of this predicted start based of the 7 BLAST hits.

This is the only proposed start based of the evidence, so it is favored.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Yucky has the Most Annotated start for this pham and it is called 49.2% of the time when present
- 33 MA's for this start (only start for this gene that has manual annotations)

- 22128 is the only proposed start suggested by the Starterator report.



Gene: Yucky_28 Start: 22128, Stop: 22367, Start Num: 33
Candidate Starts for Yucky_28:
(24, 22089), (Start: 33 @22128 has 28 MA's), (48, 22248), (49, 22257), (54, 22320), (57, 22332),

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start?   Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Starting at 22128 would leave an overlap of 4 nucleotides with the previous feature.

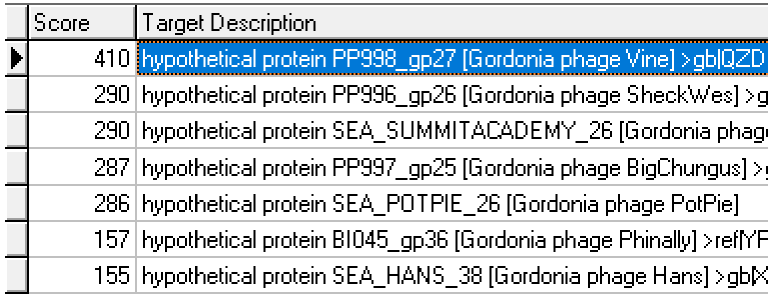
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	Start 22128
Glimmer/GeneMark	Glimmer & GeneMark
Coding Potential	Starting at 22128 would result in the loss of a small portion of the initial small peak of coding potential
RBS	z-value = 2.477 Final score = -3.370
BLAST	5 1:1 Alignments
Starterator	33 MA’s
Gap/Overlap	Overlap of 4 nucleotides

The start is 22128! This was; however, the only proposed start based off all the evidence. 22128 was called as the start of this gene by Glimmer and GeneMark, and by starting at this nucleotide only a small portion of the initial peak of coding potential is lost. At this starting point a z-value of 2.477 and a final score of -3.370 were given. There were 5 1:1 alignments according to BLAST of phages with highly similar genes, and the Starterator report showed 33 manual annotations for starting at 22128. There would be an overlap of 4 nucleotides with the previous gene

BLAST function evidence. What assigned functions do other highly similar genes have?

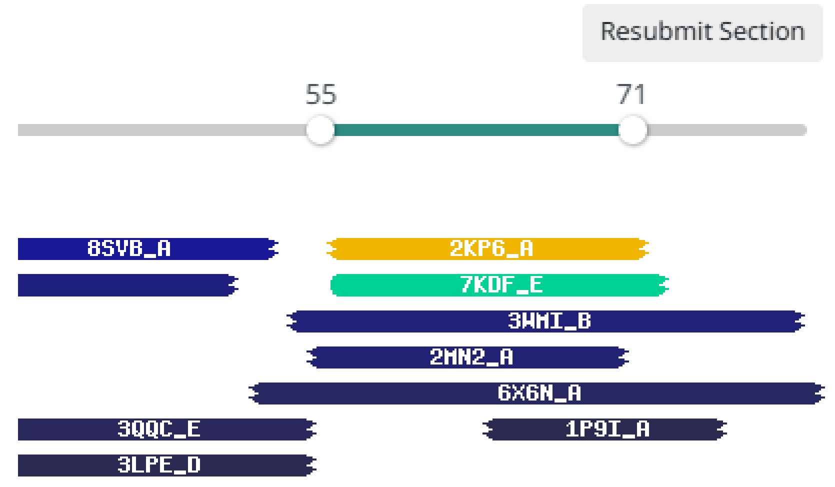
- There were 7 BLAST hits that all had functions labeled as hypothetical protein.

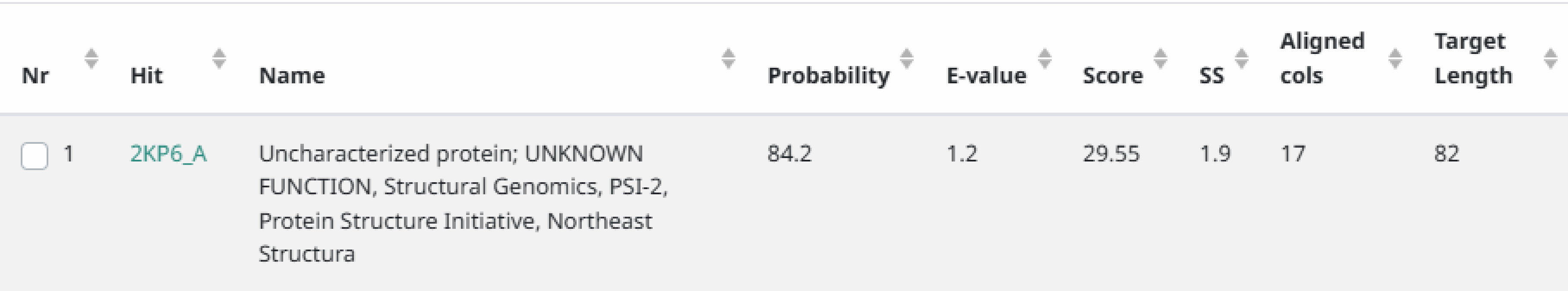


Score	Target Description
410	hypothetical protein PP998_gp27 [Gordonia phage Vine] >gb|QZD
290	hypothetical protein PP996_gp26 [Gordonia phage SheckWes] >g
290	hypothetical protein SEA_SUMMITACADEMY_26 [Gordonia phag
287	hypothetical protein PP997_gp25 [Gordonia phage BigChungus] >
286	hypothetical protein SEA_POTPIE_26 [Gordonia phage PotPie]
157	hypothetical protein BI045_gp36 [Gordonia phage Phinally] >ref|YF
155	hypothetical protein SEA_HANS_38 [Gordonia phage Hans] >gb|X

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- The highest probability hit according to HHpred was labeled as 84.2 with function labeled as “uncharacterized protein”, and none of the hits regardless of their probability value matched up with more than a small portion of the gene.

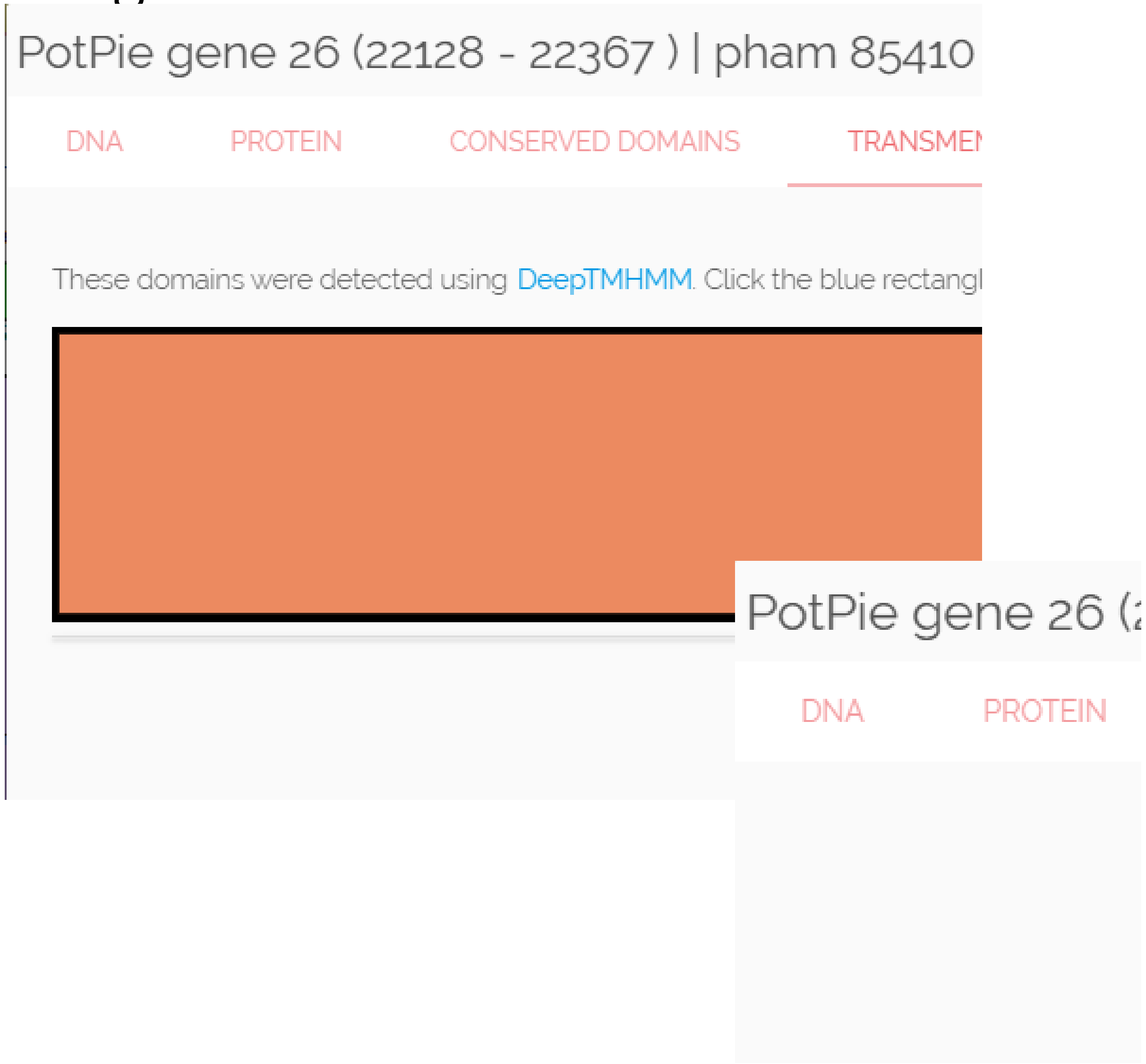




Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	2KP6_A	Uncharacterized protein; UNKNOWN FUNCTION, Structural Genomics, PSI-2, Protein Structure Initiative, Northeast Structura	84.2	1.2	29.55	1.9	17	82

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- None of the closely related phages with genes in the same pham predict a function for this gene and there are no conserved domains.
- This evidence supports the function of this gene being labeled as hypothetical protein.



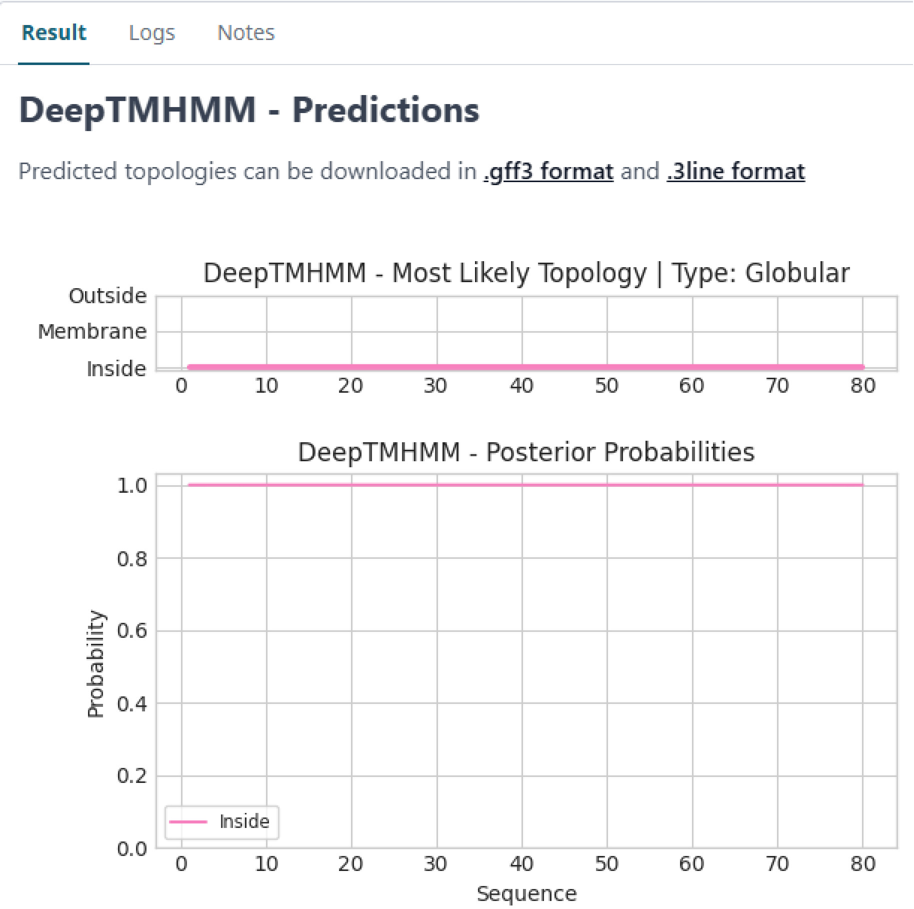
PotPie gene 26 (22128 - 22367) | pham 85410

These domains were detected using DeepTMHMM. Click the blue rectangl

PotPie gene 26 (

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- According the results from Deep TMHMM there are no transmembrane domains.



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

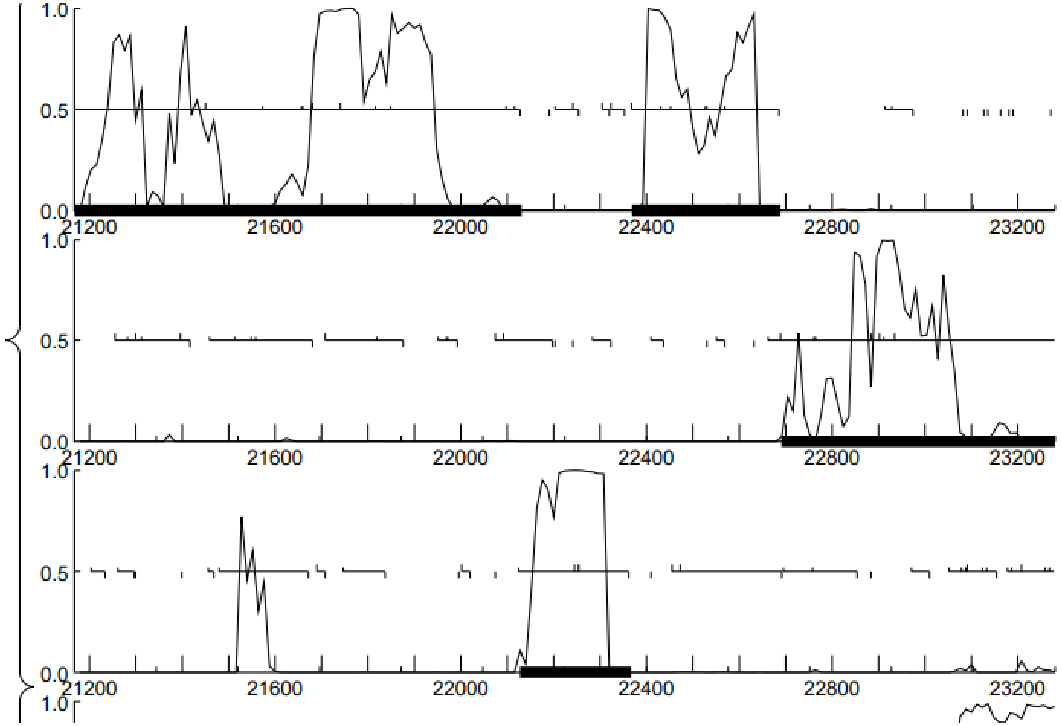
- The function for this gene should be labeled as hypothetical protein which is also the official SEA-PHAGES function that should be assigned to this gene. All of the 7 BLAST hits for this gene had functions labeled as hypothetical protein, and the HHpred results do not support this gene having any alternative function to being labeled as a hypothetical protein as the highest probability hit was 84.2 and was also labeled as having an unknown function. The Phamerator map of phages with genes in the same pham as this one have no conserved domains or official function assigned.

Feature 28 – Stop 24124

Glimmer/GeneMark

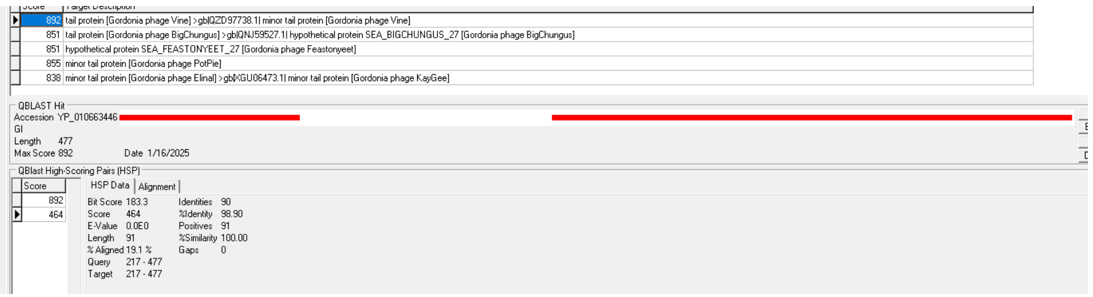
What feature number is this?	• 28
What is the stop site?	• 24124
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?	• Both Glimmer and GeneMark call it.
What is the autoannotated start?	• Nucleotide number 22691.
Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start	• There is 1 nucleotide gap.

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- The reading frame 2 has a strong coding potential.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
892	tail protein [Gordonia phage Vine] >gb|QZD97738.1| minor tail protein [Gordonia phage Vine]
851	tail protein [Gordonia phage BigChungus] >gb|QNJ59527.1| hypothetical protein SEA_BIGCHUNGUS_27 [Gordonia phage BigChungus]
851	hypothetical protein SEA_FEASTONYEET_27 [Gordonia phage Feastonyeet]
855	minor tail protein [Gordonia phage PotPie]
838	minor tail protein [Gordonia phage Elinal] >gb|XGU06473.1| minor tail protein [Gordonia phage KayGee]

QBLAST Hit

Accession YP_010663446

GI

Length 477

Max Score 892 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Score
892
464

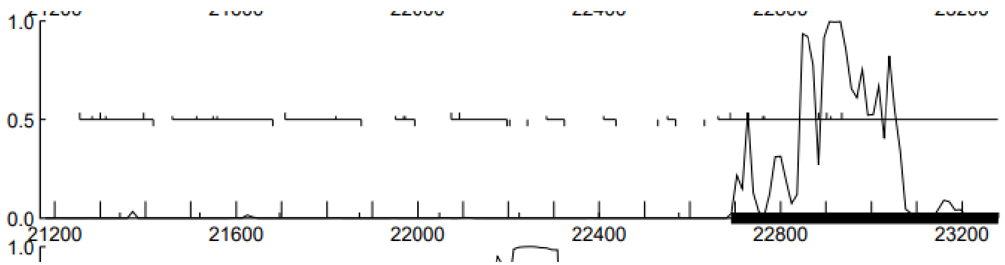
Bit Score	183.3	Identities	90
Score	464	%Identity	98.90
E-Value	0.0E0	Positives	91
Length	91	%Similarity	100.00
% Aligned	19.1 %	Gaps	0
Query	217 - 477		
Target	217 - 477		

- There are 25 highly similar genes with E value of 0 or less than 1x10-7 (Vine, BigChungus).

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- It is a gene because:
- Both Glimmer and GeneMark called it a gene.
- Coding potential is strong.
- There are many highly similar genes with E value of 0 or less than 1x10-7.

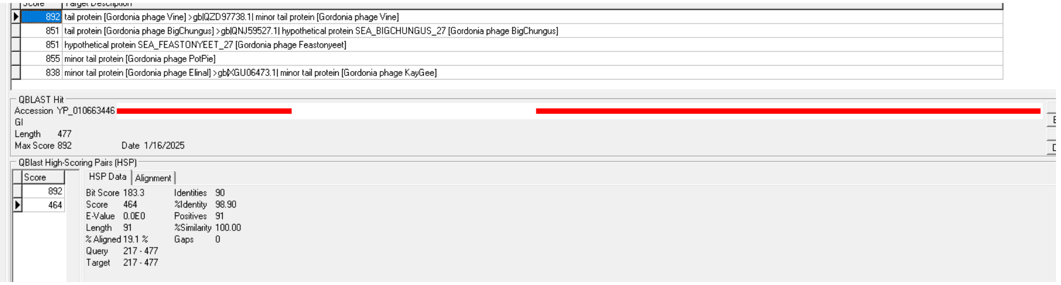
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- Reading frame 2 has a coding potential where feature 30 starts. So, included.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- **There are 10 1:1 alignments.**



Score	Target Description
892	tail protein [Gordonia phage Vine] >gb|QZD97738.1| minor tail protein [Gordonia phage Vine]
851	tail protein [Gordonia phage BigChungus] >gb|QNJ59527.1| hypothetical protein SEA_BIGCHUNGUS_27 [Gordonia phage BigChungus]
851	hypothetical protein SEA_FEASTONYEET_27 [Gordonia phage Feastonyeet]
855	minor tail protein [Gordonia phage PotPie]
838	minor tail protein [Gordonia phage Elinal] >gb|XGU06473.1| minor tail protein [Gordonia phage KayGee]

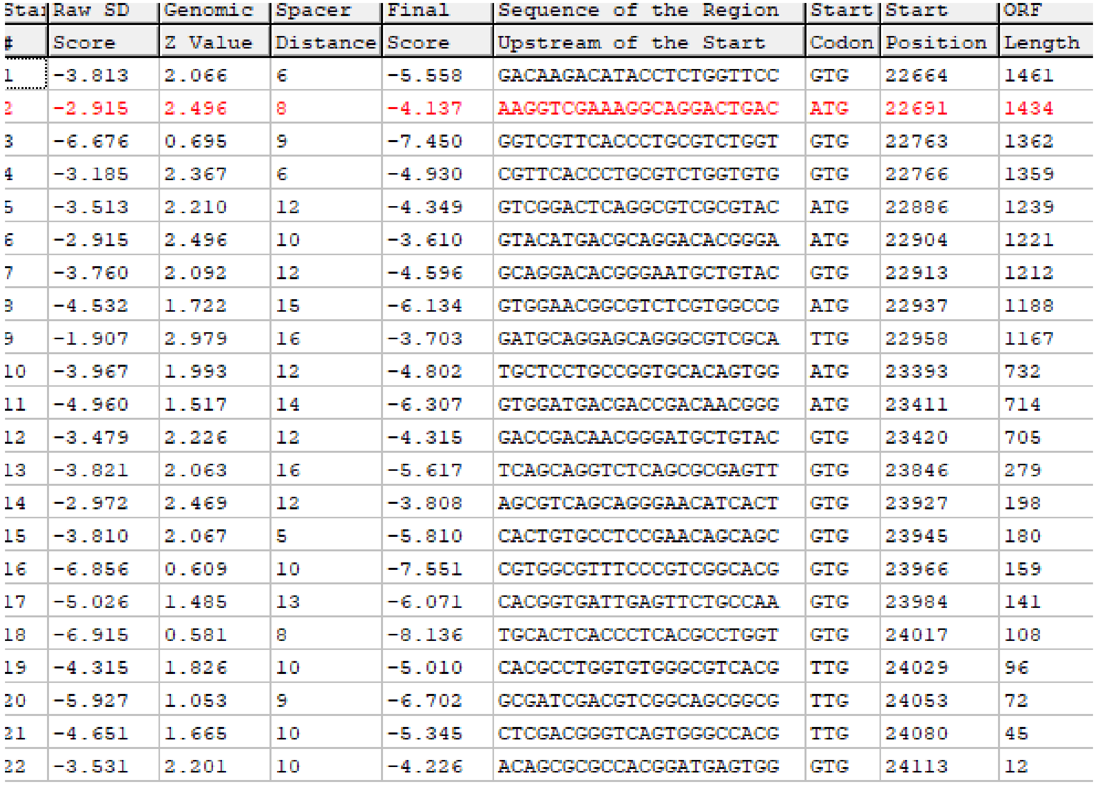
QBLAST Hit
Accession YP_010663446
GI
Length 477
Max Score 892 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Score
892
464

Bit Score 183.3 Identities 90
Score 464 %Identity 98.90
E-Value 0.0E0 Positives 91
Length 91 %Similarity 100.00
% Aligned 19.1 % Gaps 0
Query 217 - 477
Target 217 - 477

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts?   Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

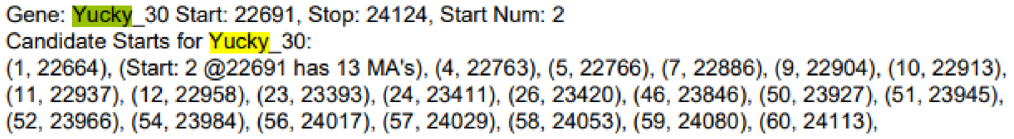


Sta #	Raw SD Score	Genomic Z Value	Spacer Distance	Final Score	Sequence of the Region Upstream of the Start	Start Codon	Start Position	ORF Length
1	-3.813	2.066	6	-5.558	GACAAGACATACCTCTGGTTCC	GTG	22664	1461
2	-2.915	2.496	8	-4.137	AAGGTCGAAAGGCAGGACTGAC	ATG	22691	1434
3	-6.676	0.695	9	-7.450	GGTCGTTCACCCTGCGTCTGGT	GTG	22763	1362
4	-3.185	2.367	6	-4.930	CGTTCACCCTGCGTCTGGTGTG	GTG	22766	1359
5	-3.513	2.210	12	-4.349	GTCGGACTCAGGCGTCGCGTAC	ATG	22886	1239
6	-2.915	2.496	10	-3.610	GTACATGACGCAGGACACGGGA	ATG	22904	1221
7	-3.760	2.092	12	-4.596	GCAGGACACGGGAATGCTGTAC	GTG	22913	1212
8	-4.532	1.722	15	-6.134	GTGGAACGGCGTCTCGTGGCCG	ATG	22937	1188
9	-1.907	2.979	16	-3.703	GATGCAGGAGCAGGGCGTCGCA	TTG	22958	1167
10	-3.967	1.993	12	-4.802	TGCTCCTGCCGGTGCACAGTGG	ATG	23393	732
11	-4.960	1.517	14	-6.307	GTGGATGACGACCGACAACGGG	ATG	23411	714
12	-3.479	2.226	12	-4.315	GACCGACAACGGGATGCTGTAC	GTG	23420	705
13	-3.821	2.063	16	-5.617	TCAGCAGGTCTCAGCGCGAGTT	GTG	23846	279
14	-2.972	2.469	12	-3.808	AGCGTCAGCAGGGAACATCACT	GTG	23927	198
15	-3.810	2.067	5	-5.810	CACTGTGCCTCCGAACAGCAGC	GTG	23945	180
16	-6.856	0.609	10	-7.551	CGTGGCGTTTCCCGTCGGCACG	GTG	23966	159
17	-5.026	1.485	13	-6.071	CACGGTGATTGAGTTCTGCCAA	GTG	23984	141
18	-6.915	0.581	8	-8.136	TGCACTCACCCTCACGCCTGGT	GTG	24017	108
19	-4.315	1.826	10	-5.010	CACGCCTGGTGTGGGCGTCACG	TTG	24029	96
20	-5.927	1.053	9	-6.702	GCGATCGACGTCGGCAGCGGCG	TTG	24053	72
21	-4.651	1.665	10	-5.345	CTCGACGGGTCAGTGGGCCACG	TTG	24080	45
22	-3.531	2.201	10	-4.226	ACAGCGCGCCACGGATGAGTGG	GTG	24113	12

- The z value is the greatest with 2.496.
- The final score is less negative than most of them. But it is not the least negative.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

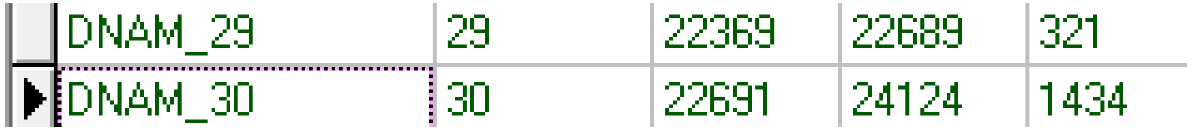
- 13 MA’s for starting site 22691.



```
Gene: Yucky_30 Start: 22691, Stop: 24124, Start Num: 2
Candidate Starts for Yucky_30:
(1, 22664), (Start: 2 @22691 has 13 MA's), (4, 22763), (5, 22766), (7, 22886), (9, 22904), (10, 22913),
(11, 22937), (12, 22958), (23, 23393), (24, 23411), (26, 23420), (46, 23846), (50, 23927), (51, 23945),
(52, 23966), (54, 23984), (56, 24017), (57, 24029), (58, 24053), (59, 24080), (60, 24113),
```

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 22691-22689 = 2
- 2-1 = 1 gap



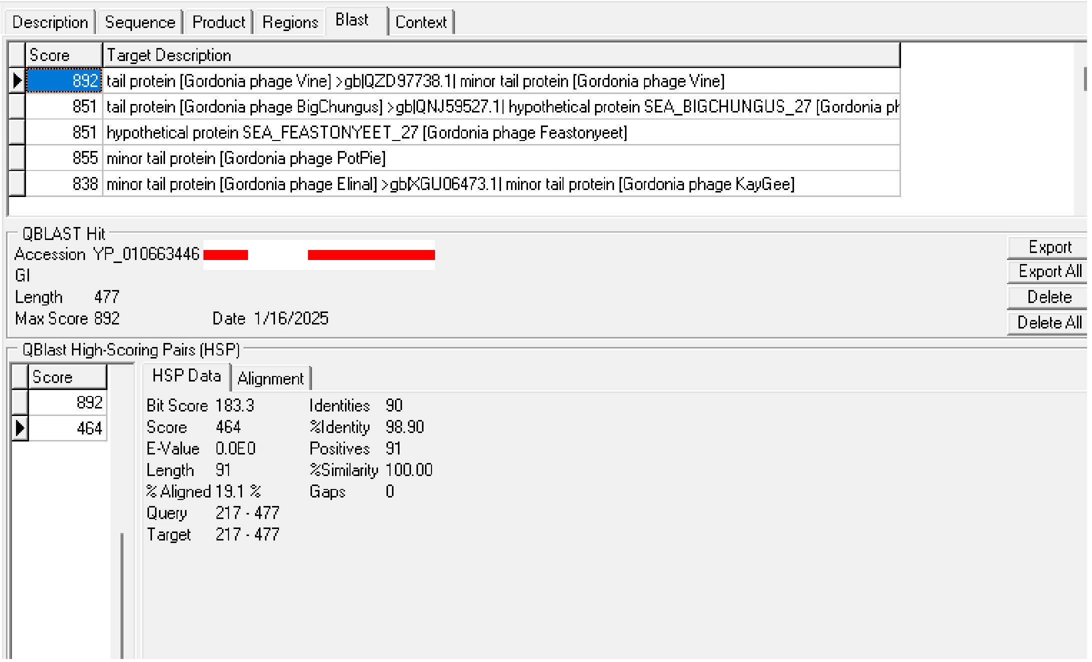
	DNAM_29	29	22369	22689	321
▶	DNAM_30	30	22691	24124	1434

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	22691
GeneMark	Both Glimmer and GeneMark.
Coding potential	Included
RBS	Z value: 2.496 Final Score: -4.137 (Not least negative)
Blast	10 1:1 alignments
Starterator	13 MA’s
Gap/overlap	1 gap

Both Glimmer and GeneMark call it a start site. Coding potential is included. RBS score does not completely, just a little bit, favor it with the final score that is not the least negative. But Blast information with 1:1 alignemtns and the MA’s favor this starting site. So, 22691 is a start site of feature 30. Gap of 1 is negligible.

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
892	tail protein [Gordonia phage Vine] >gb|QZD97738.1| minor tail protein [Gordonia phage Vine]
851	tail protein [Gordonia phage BigChungus] >gb|QNJ59527.1| hypothetical protein SEA_BIGCHUNGUS_27 [Gordonia ph
851	hypothetical protein SEA_FEASTONYEET_27 [Gordonia phage Feastonyeet]
855	minor tail protein [Gordonia phage PotPie]
838	minor tail protein [Gordonia phage Elinal] >gb|XGU06473.1| minor tail protein [Gordonia phage KayGee]

QBLAST Hit

Accession YP_010663446 [redacted] [redacted]

GI

Length 477

Max Score 892 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

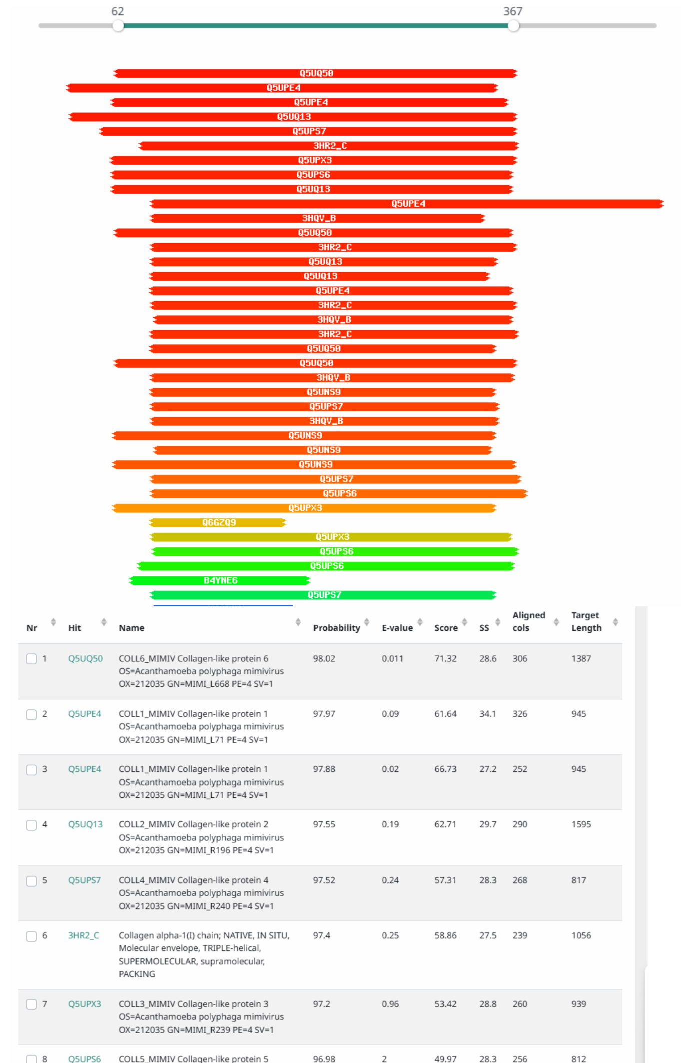
Score
892
464

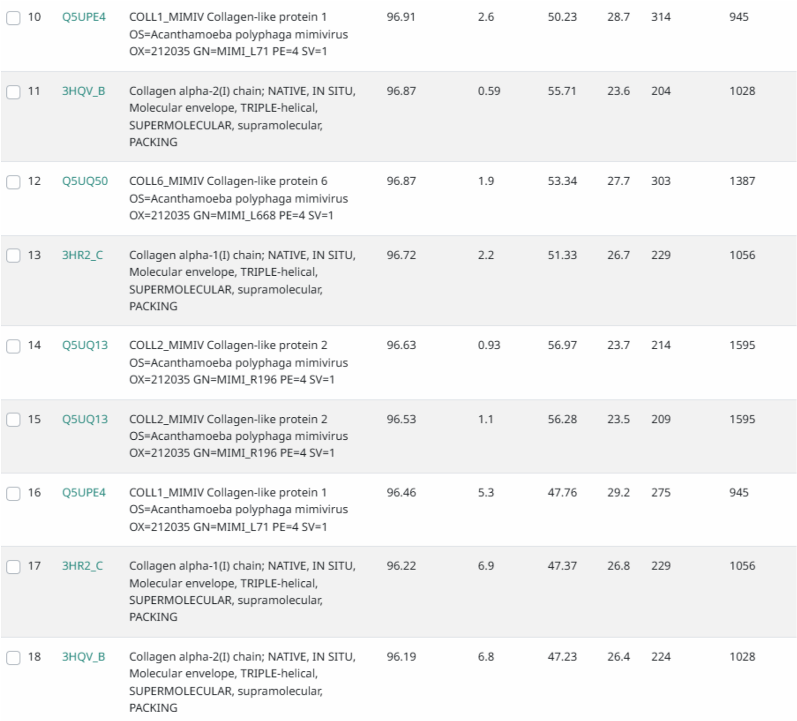
Bit Score	183.3	Identities	90
Score	464	%Identity	98.90
E-Value	0.0E0	Positives	91
Length	91	%Similarity	100.00
% Aligned	19.1 %	Gaps	0
Query	217 - 477		
Target	217 - 477		

- This gene is suggested with 3 functions.

- Tail protein (Vine, BigChungus)
- Minor tail protein (Vine, Potpie)
- Hypothetical protein (BigChungus, Feastonyeet)

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



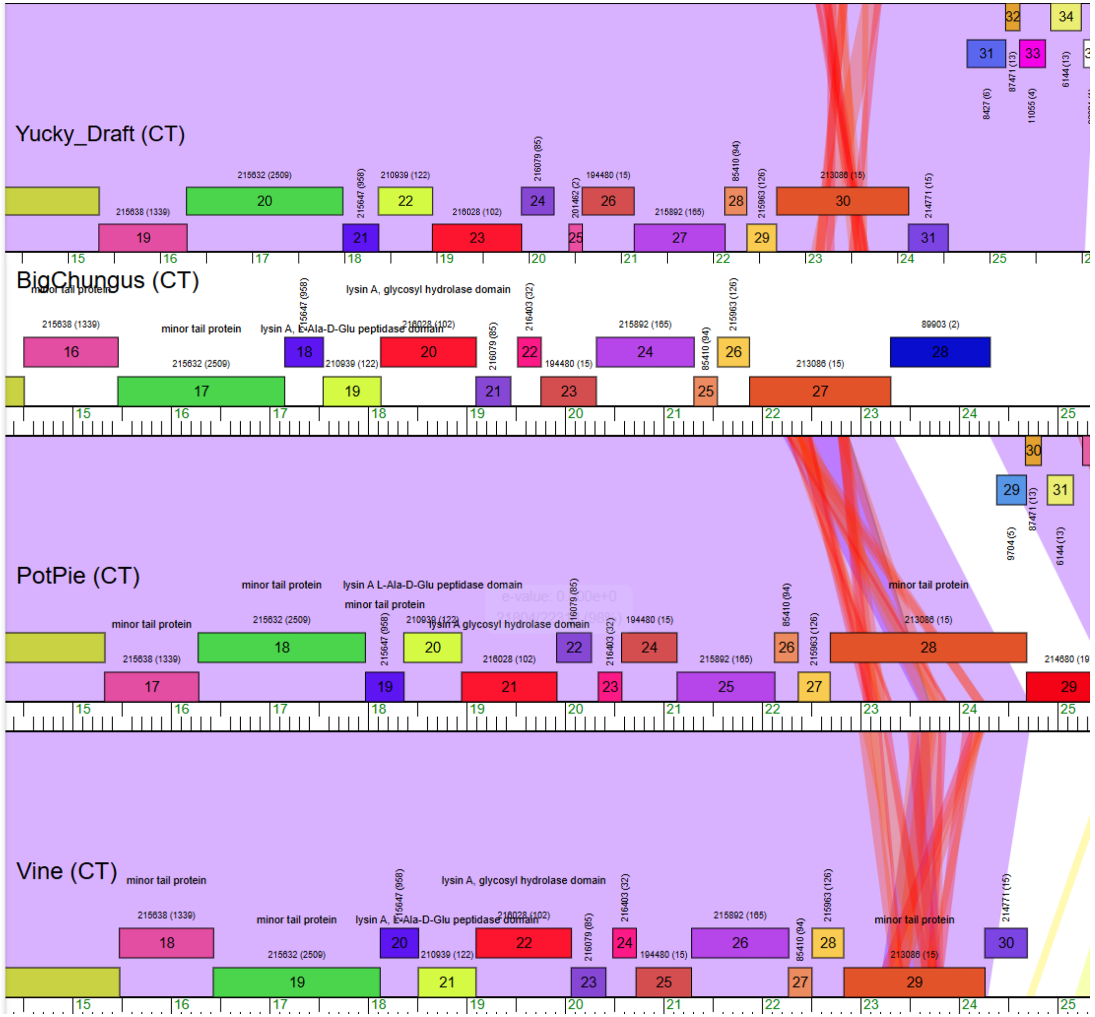


Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	Q5UQ50	COLL6_MIMIV Collagen-like protein 6 OS=Acanthamoeba polyphaga mimivirus OX=212035 GN=MIMI_L668 PE=4 SV=1	98.02	0.011	71.32	28.6	306	1387
□ 2	Q5UPE4	COLL1_MIMIV Collagen-like protein 1 OS=Acanthamoeba polyphaga mimivirus OX=212035 GN=MIMI_L71 PE=4 SV=1	97.97	0.09	61.64	34.1	326	945
□ 3	Q5UPE4	COLL1_MIMIV Collagen-like protein 1 OS=Acanthamoeba polyphaga mimivirus OX=212035 GN=MIMI_L71 PE=4 SV=1	97.88	0.02	66.73	27.2	252	945
□ 4	Q5UQ13	COLL2_MIMIV Collagen-like protein 2 OS=Acanthamoeba polyphaga mimivirus OX=212035 GN=MIMI_R196 PE=4 SV=1	97.55	0.19	62.71	29.7	290	1595
□ 5	Q5UPS7	COLL4_MIMIV Collagen-like protein 4 OS=Acanthamoeba polyphaga mimivirus OX=212035 GN=MIMI_R240 PE=4 SV=1	97.52	0.24	57.31	28.3	268	817
□ 6	3HR2_C	Collagen alpha-1(I) chain; NATIVE, IN SITU, Molecular envelope, TRIPLE-helical, SUPERMOLECULAR, supramolecular, PACKING	97.4	0.25	58.86	27.5	239	1056
□ 7	Q5UPX3	COLL3_MIMIV Collagen-like protein 3 OS=Acanthamoeba polyphaga mimivirus OX=212035 GN=MIMI_R239 PE=4 SV=1	97.2	0.96	53.42	28.8	260	939
□ 8	Q5UPS6	COLL5_MIMIV Collagen-like protein 5	96.98	2	49.97	28.3	256	812
□ 10	Q5UPE4	COLL1_MIMIV Collagen-like protein 1 OS=Acanthamoeba polyphaga mimivirus OX=212035 GN=MIMI_L71 PE=4 SV=1	96.91	2.6	50.23	28.7	314	945
□ 11	3HQV_B	Collagen alpha-2(I) chain; NATIVE, IN SITU, Molecular envelope, TRIPLE-helical, SUPERMOLECULAR, supramolecular, PACKING	96.87	0.59	55.71	23.6	204	1028
□ 12	Q5UQ50	COLL6_MIMIV Collagen-like protein 6 OS=Acanthamoeba polyphaga mimivirus OX=212035 GN=MIMI_L668 PE=4 SV=1	96.87	1.9	53.34	27.7	303	1387
□ 13	3HR2_C	Collagen alpha-1(I) chain; NATIVE, IN SITU, Molecular envelope, TRIPLE-helical, SUPERMOLECULAR, supramolecular, PACKING	96.72	2.2	51.33	26.7	229	1056
□ 14	Q5UQ13	COLL2_MIMIV Collagen-like protein 2 OS=Acanthamoeba polyphaga mimivirus OX=212035 GN=MIMI_R196 PE=4 SV=1	96.63	0.93	56.97	23.7	214	1595
□ 15	Q5UQ13	COLL2_MIMIV Collagen-like protein 2 OS=Acanthamoeba polyphaga mimivirus OX=212035 GN=MIMI_R196 PE=4 SV=1	96.53	1.1	56.28	23.5	209	1595
□ 16	Q5UPE4	COLL1_MIMIV Collagen-like protein 1 OS=Acanthamoeba polyphaga mimivirus OX=212035 GN=MIMI_L71 PE=4 SV=1	96.46	5.3	47.76	29.2	275	945
□ 17	3HR2_C	Collagen alpha-1(I) chain; NATIVE, IN SITU, Molecular envelope, TRIPLE-helical, SUPERMOLECULAR, supramolecular, PACKING	96.22	6.9	47.37	26.8	229	1056
□ 18	3HQV_B	Collagen alpha-2(I) chain; NATIVE, IN SITU, Molecular envelope, TRIPLE-helical, SUPERMOLECULAR, supramolecular, PACKING	96.19	6.8	47.23	26.4	224	1028

There are many hits that have higher probability than 90. Many of them call it collagen-like protein.

The correct name for collagen-like protein is minor tail protein.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



Other genes from same pham are minor tail protein.

Gene 30 of Yucky share one conserved domain: collagen.

Other genes from same pham have collagen but also PHA03169.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- It looks like it is a minor tail protein, so don’t need to do this part.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Gene 30 is a minor tail protein because

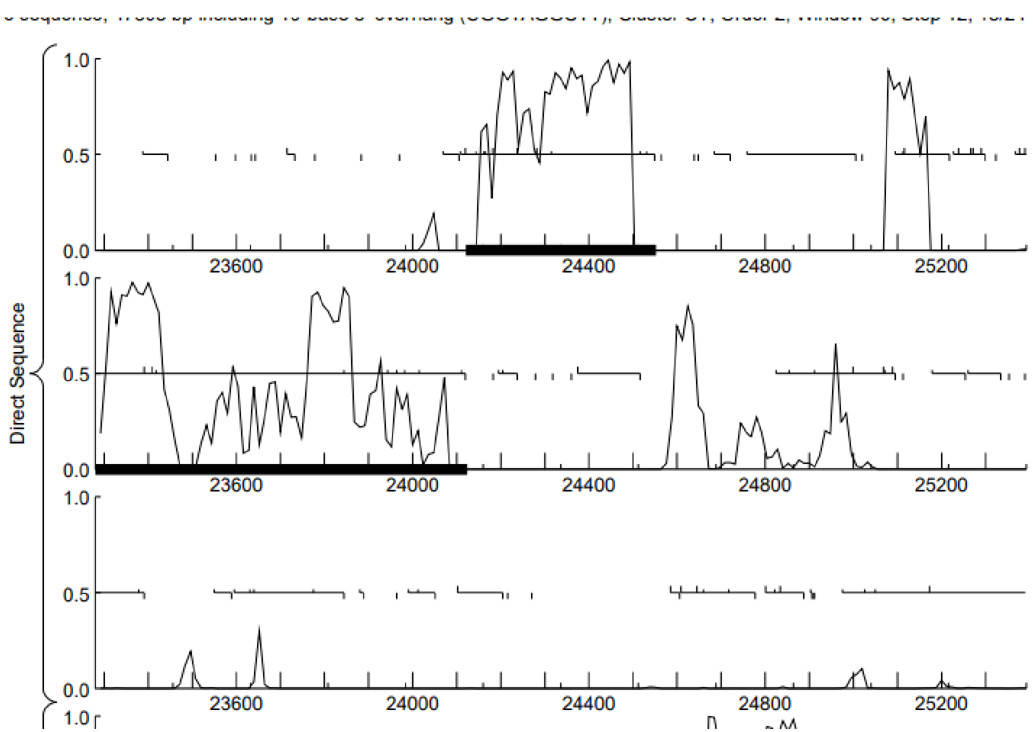
- One of the suggestion that BLAST provided was minor tail.
- Hhpred gave a strong evidence with many hits that call it minor tail protein with higher probability than 90.
- Phamerator shows that other genes in same pham are minor tail protein.

Feature 29 – Stop 24552

Glimmer/GeneMark

What feature number is this?	• 29
What is the stop site?	• 24552
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?	• Both Glimmer and GeneMark
What is the autoannotated start?	• 24121 • RBS score has one more suggestion for start site: 24109.
Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start	• 4 overlap

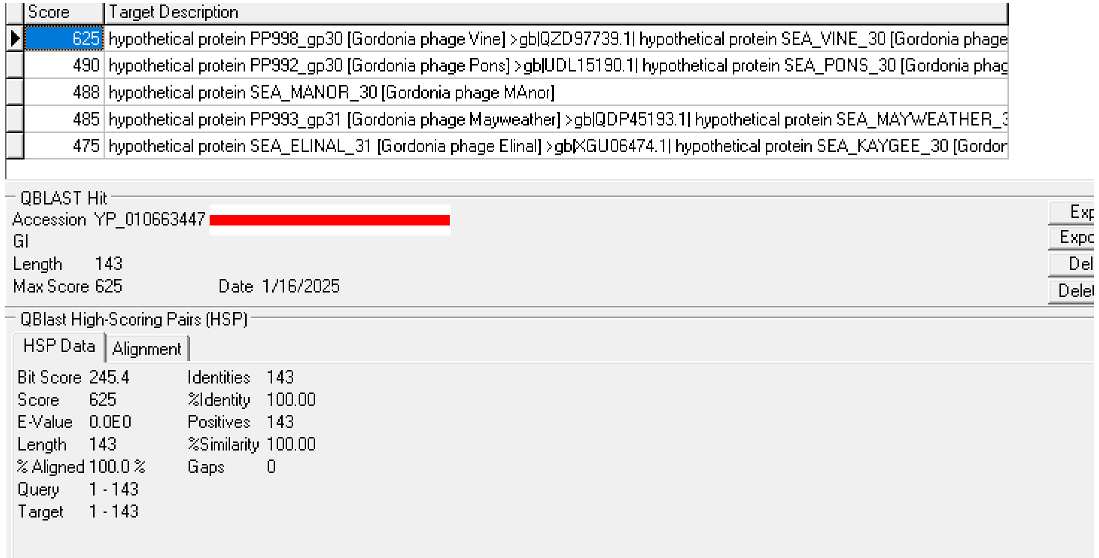
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Coding potential in reading frame 1 is strong in the area of feature 31.

- 24109:
- Coding potential is strong as well.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



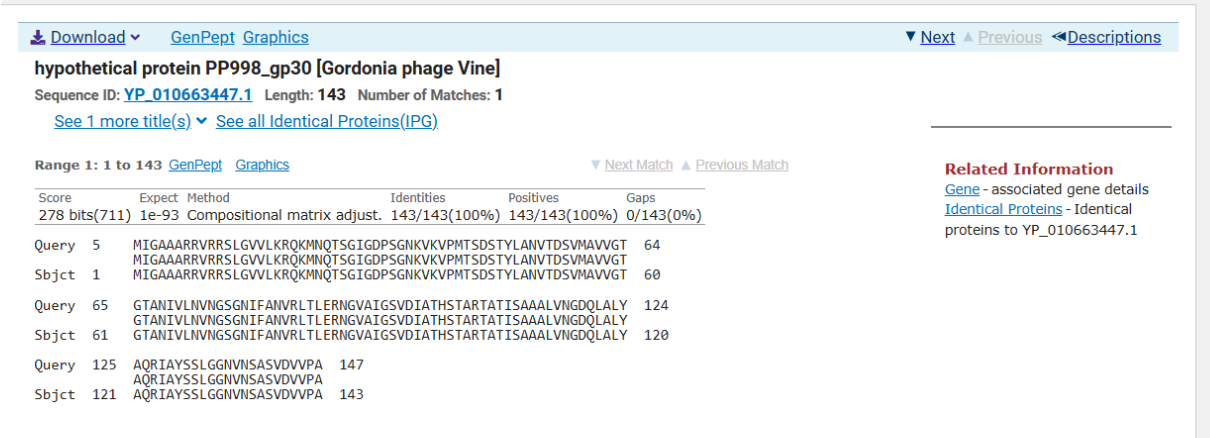
Score	Target Description
625	hypothetical protein PP998_gp30 [Gordonia phage Vine] >gb|QZD97739.1| hypothetical protein SEA_VINE_30 [Gordonia phage
490	hypothetical protein PP992_gp30 [Gordonia phage Pons] >gb|UDL15190.1| hypothetical protein SEA_PONS_30 [Gordonia phag
488	hypothetical protein SEA_MANOR_30 [Gordonia phage MAnor]
485	hypothetical protein PP993_gp31 [Gordonia phage Mayweather] >gb|QDP45193.1| hypothetical protein SEA_MAYWEATHER_3
475	hypothetical protein SEA_ELINAL_31 [Gordonia phage Elinal] >gb|XGU06474.1| hypothetical protein SEA_KAYGEE_30 [Gordor

QBLAST Hit
Accession YP_010663447
GI
Length 143
Max Score 625 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 245.4
Score 625
E-Value 0.0E0
Length 143
% Aligned 100.0 %
Query 1 - 143
Target 1 - 143

Identities 143
%Identity 100.00
Positives 143
%Similarity 100.00
Gaps 0



Download
GenPept Graphics
▼ Next ▲ Previous ◀Descriptions

hypothetical protein PP998_gp30 [Gordonia phage Vine]

Sequence ID: YP_010663447.1 Length: 143 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 143 GenPept Graphics

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
278 bits(711)	1e-93	Compositional matrix adjust.	143/143(100%)	143/143(100%)	0/143(0%)

```
Query  5    MIGAAARRVRRSLGVVLKRQKMNQTSGIGDPSGNKVKVPMTSDSTYLANVTDSVMAVVGT  64
            MIGAAARRVRRSLGVVLKRQKMNQTSGIGDPSGNKVKVPMTSDSTYLANVTDSVMAVVGT
Sbjct  1    MIGAAARRVRRSLGVVLKRQKMNQTSGIGDPSGNKVKVPMTSDSTYLANVTDSVMAVVGT  60

Query  65   GTANIVLNVNGSGNIFANVRLTLERNGVAIGSVDIATHSTARTATISAAALVNGDQLALY  124
            GTANIVLNVNGSGNIFANVRLTLERNGVAIGSVDIATHSTARTATISAAALVNGDQLALY
Sbjct  61   GTANIVLNVNGSGNIFANVRLTLERNGVAIGSVDIATHSTARTATISAAALVNGDQLALY  120

Query  125  AQRIAYSSLGGNVNSASVDVVPA  147
            AQRIAYSSLGGNVNSASVDVVPA
Sbjct  121  AQRIAYSSLGGNVNSASVDVVPA  143
```

Related Information

Gene - associated gene details

Identical Proteins - Identical proteins to YP_010663447.1

- There are 13 highly similar genes with E value of 0 or less than 1x10-7.

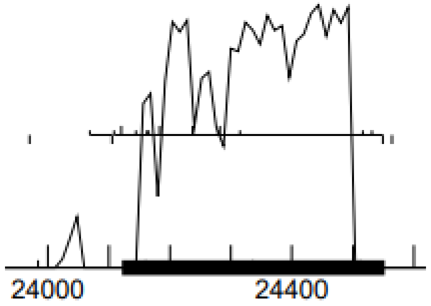
- 24109:
- There are highly similar genes with E value that’s less than 1x10-7.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- It is a gene because:

- Both Glimmer and GeneMark called it a gene.
- Coding potential is strong.
- There are 13 highly similar genes with favorable E value.

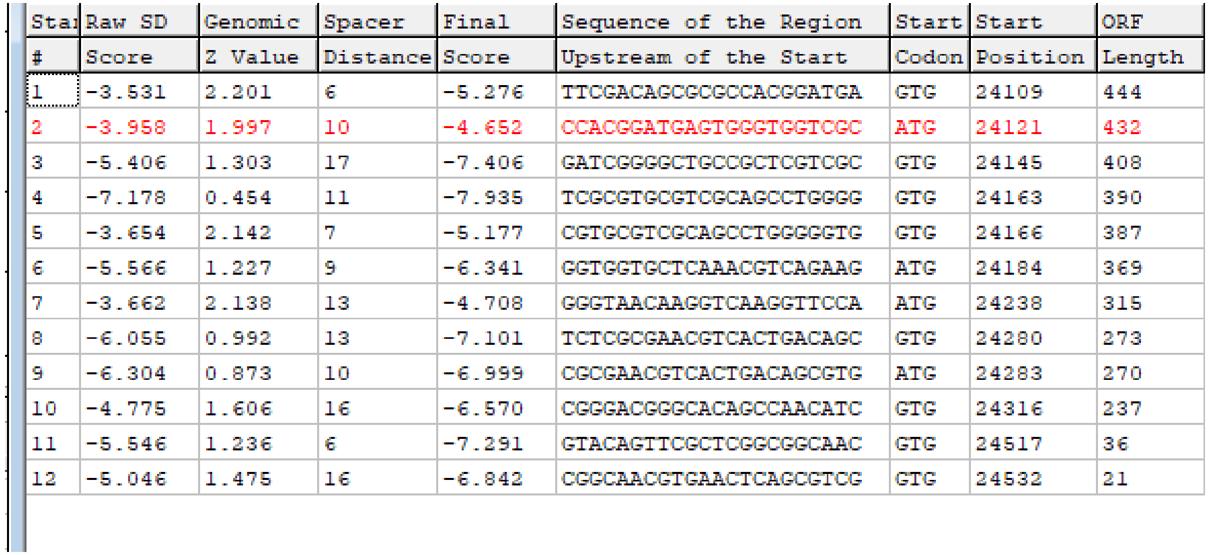
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- Coding potential in reading frame 1 starts around 24120.
- Feature 31 starts at 24121.
- Therefore, all coding potential is included.

- 24109:
- coding potential is included in reading frame 1.

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-3.531	2.201	6	-5.276	TTCGACAGCGCGCCACGGATGA	GTG	24109	444
2	-3.958	1.997	10	-4.652	CCACGGATGAGTGGGTGGTCGC	ATG	24121	432
3	-5.406	1.303	17	-7.406	GATCGGGGCTGCCGCTCGTCGC	GTG	24145	408
4	-7.178	0.454	11	-7.935	TCGCGTGCGTCGCAGCCTGGGG	GTG	24163	390
5	-3.654	2.142	7	-5.177	CGTGCGTCGCAGCCTGGGGGTG	GTG	24166	387
6	-5.566	1.227	9	-6.341	GGTGGTGCTCAAACGTCAGAAG	ATG	24184	369
7	-3.662	2.138	13	-4.708	GGGTAACAAGGTCAAGGTTCCA	ATG	24238	315
8	-6.055	0.992	13	-7.101	TCTCGCGAACGTCACTGACAGC	GTG	24280	273
9	-6.304	0.873	10	-6.999	CGCGAACGTCACTGACAGCGTG	ATG	24283	270
10	-4.775	1.606	16	-6.570	CGGGACGGGCACAGCCAACATC	GTG	24316	237
11	-5.546	1.236	6	-7.291	GTACAGTTCGCTCGGCGGCAAC	GTG	24517	36
12	-5.046	1.475	16	-6.842	CGGCAACGTGAACTCAGCGTCG	GTG	24532	21

The z value is not the greatest with 1.997 but close to 2.000.

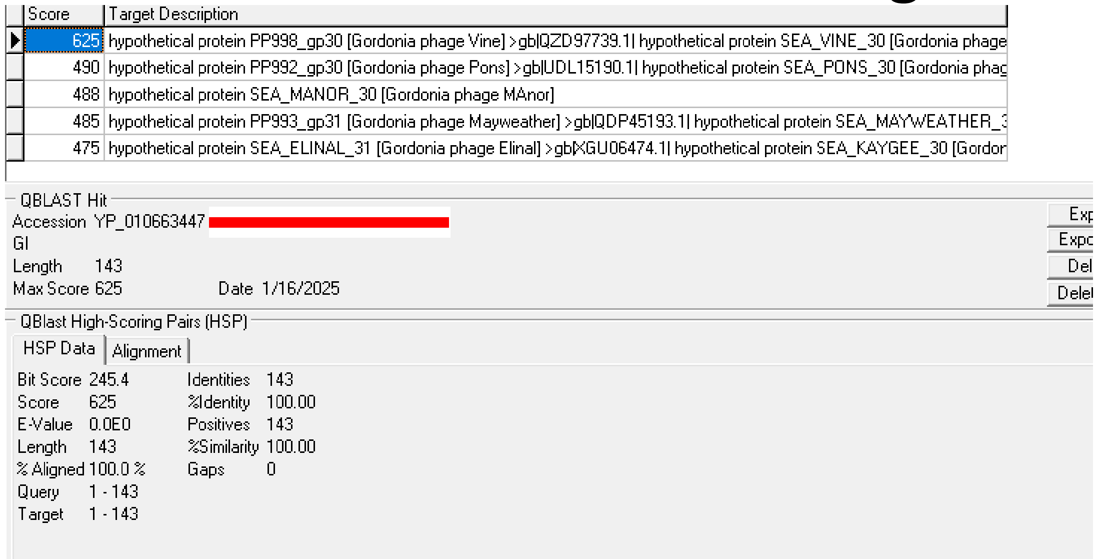
Final score is the least negative number with -4.652.

24109:

Z value: greatest with 2.201.

Final score: -5.276 not least negative.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
625	hypothetical protein PP998_gp30 [Gordonia phage Vine] >gb|QZD97739.1| hypothetical protein SEA_VINE_30 [Gordonia phage
490	hypothetical protein PP992_gp30 [Gordonia phage Pons] >gb|UDL15190.1| hypothetical protein SEA_PONS_30 [Gordonia phag
488	hypothetical protein SEA_MANOR_30 [Gordonia phage MAnor]
485	hypothetical protein PP993_gp31 [Gordonia phage Mayweather] >gb|QDP45193.1| hypothetical protein SEA_MAYWEATHER_3
475	hypothetical protein SEA_ELINAL_31 [Gordonia phage Elinal] >gb|XGU06474.1| hypothetical protein SEA_KAYGEE_30 [Gordon

QBLAST Hit

Accession YP_010663447

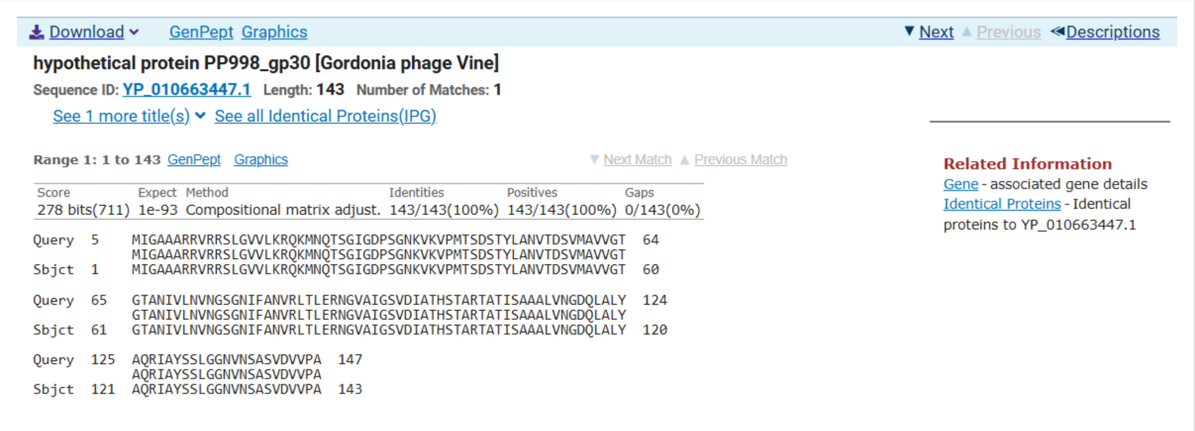
GI

Length 143

Max Score 625 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	245.4	Identities	143
Score	625	%Identity	100.00
E-Value	0.0E0	Positives	143
Length	143	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 143		
Target	1 - 143		



hypothetical protein PP998_gp30 [Gordonia phage Vine]

Sequence ID: YP_010663447.1 Length: 143 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 143 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
278 bits(711)	1e-93	Compositional matrix adjust.	143/143(100%)	143/143(100%)	0/143(0%)

```
Query  5    MIGAAARRVRRSLGVVLKRQKMNQTSGIGDPSGNKVKVPMTSDSTYLANVTDSVMAVVGT  64
            MIGAAARRVRRSLGVVLKRQKMNQTSGIGDPSGNKVKVPMTSDSTYLANVTDSVMAVVGT
Sbjct  1    MIGAAARRVRRSLGVVLKRQKMNQTSGIGDPSGNKVKVPMTSDSTYLANVTDSVMAVVGT  60

Query  65   GTANIVLNVNGSGNIFANVRLTLERNGVAIGSVDIATHSTARTATISAAALVNGDQLALY  124
            GTANIVLNVNGSGNIFANVRLTLERNGVAIGSVDIATHSTARTATISAAALVNGDQLALY
Sbjct  61   GTANIVLNVNGSGNIFANVRLTLERNGVAIGSVDIATHSTARTATISAAALVNGDQLALY  120

Query  125  AQRIAYSSLGGNVNSASVDVVPA  147
            AQRIAYSSLGGNVNSASVDVVPA
Sbjct  121  AQRIAYSSLGGNVNSASVDVVPA  143
```

Related Information

Gene - associated gene details

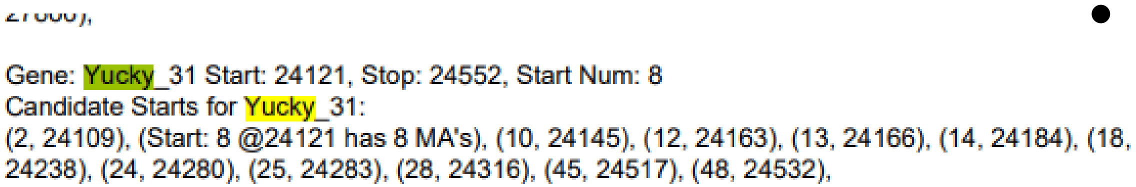
Identical Proteins - Identical proteins to YP_010663447.1

- Six 1:1 alignments.
- 24109:
- No 1:1 alignments.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- 24121 has 8 MA's

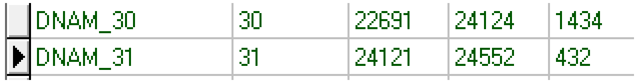
- 24109
- No MA's



27000),

Gene: Yucky_31 Start: 24121, Stop: 24552, Start Num: 8
Candidate Starts for Yucky_31:
(2, 24109), (Start: 8 @24121 has 8 MA's), (10, 24145), (12, 24163), (13, 24166), (14, 24184), (18, 24238), (24, 24280), (25, 24283), (28, 24316), (45, 24517), (48, 24532),

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.



DNAM_30	30	22691	24124	1434
DNAM_31	31	24121	24552	432

- 24124-24121 = 3
- 3+1 = 4 overlap

- 24109:
- 24124-24109: 15
- 15+1=16 overlap

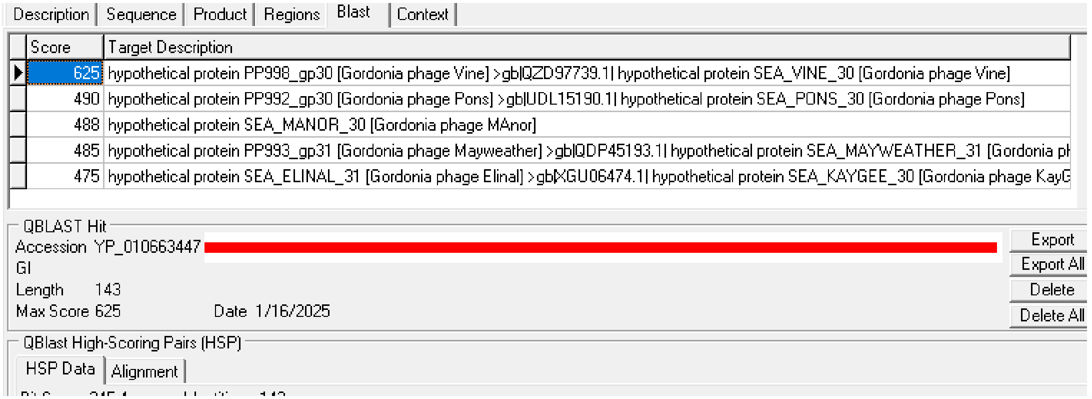
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	24121	24109
GeneMark	Both Glimmer and GeneMark	NA
Coding potential	Included	Included
RBS	Z value: 1.997 Final score: -4.652	Z value: 2.201 Final Score: -5.276
Blast	6 1:1 alignments	0
Starterator	8 MA's	0
Gap/overlap	4 overlap	16 overlap

24121 is a start because both Glimmer and GeneMark called it. Coding potential is included. There are some number of 1:1 alignments and manual annotation. Since there are 4 nucleotides overlap, RBS score is considered important. Z value is not the greatest but close enough to 2.000, and the final score is least negative. So, RBS score favors the start site 24121.

BLAST function evidence. What assigned functions do other highly similar genes have?

- BLAST call it a hypothetical protein (Vine).



Score	Target Description
625	hypothetical protein PP998_gp30 [Gordonia phage Vine] >gb|QZD97739.1| hypothetical protein SEA_VINE_30 [Gordonia phage Vine]
490	hypothetical protein PP992_gp30 [Gordonia phage Pons] >gb|UDL15190.1| hypothetical protein SEA_PONS_30 [Gordonia phage Pons]
488	hypothetical protein SEA_MANOR_30 [Gordonia phage MAnor]
485	hypothetical protein PP993_gp31 [Gordonia phage Mayweather] >gb|QDP45193.1| hypothetical protein SEA_MAYWEATHER_31 [Gordonia pl
475	hypothetical protein SEA_ELINAL_31 [Gordonia phage Elinal] >gb|XGU06474.1| hypothetical protein SEA_KAYGEE_30 [Gordonia phage KayG

QBLAST Hit

Accession YP_010663447

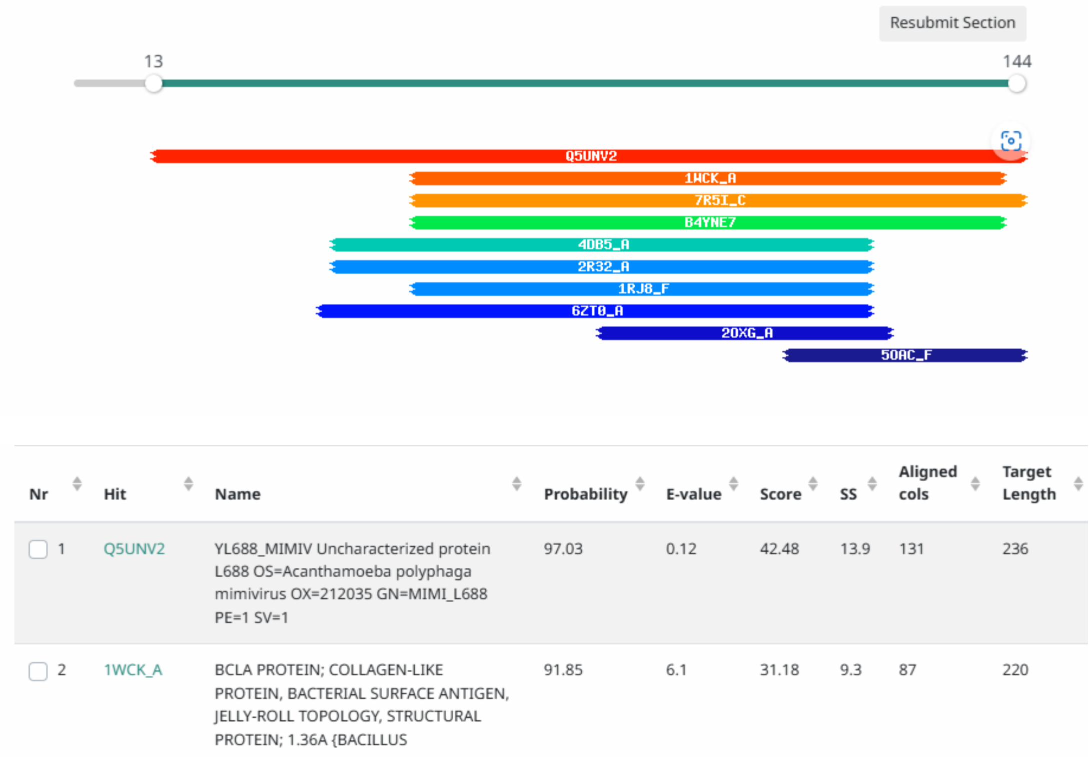
GI

Length 143

Max Score 625 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
☐ 1	Q5UNV2	YL688_MIMIV Uncharacterized protein L688 OS=Acanthamoeba polyphaga mimivirus OX=212035 GN=MIMI_L688 PE=1 SV=1	97.03	0.12	42.48	13.9	131	236
☐ 2	1WCK_A	BCLA PROTEIN; COLLAGEN-LIKE PROTEIN, BACTERIAL SURFACE ANTIGEN, JELLY-ROLL TOPOLOGY, STRUCTURAL PROTEIN; 1.36A {BACILLUS	91.85	6.1	31.18	9.3	87	220

There are 2 hits with probability greater than 90.

One is uncharacterized protein, and one is collagen-like protein (Minor tail protein).

There should be many hits with minor tail protein in order to call it a minor tail protein.

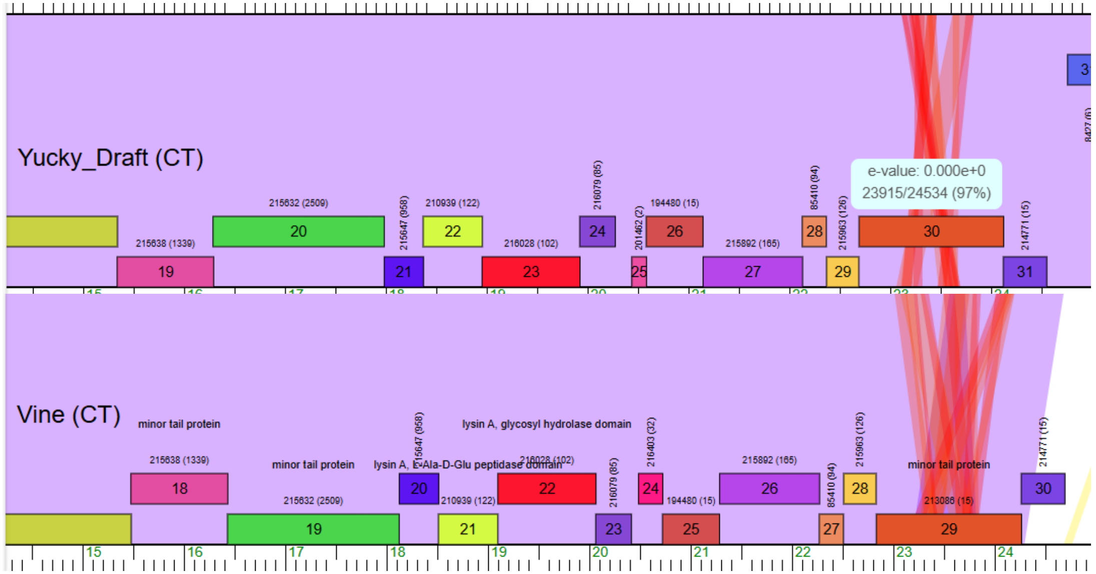
Since there is only one, it is not likely to be a minor tail protein.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

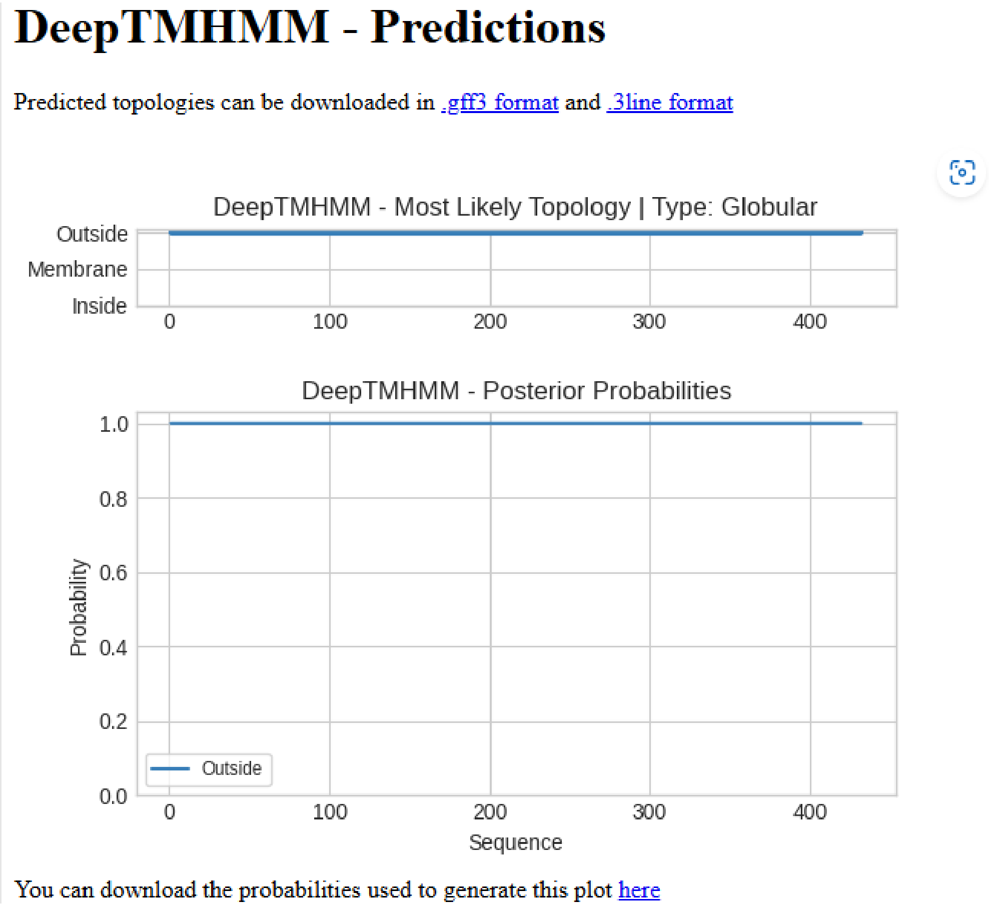
Only vine has the same gene as gene 31 of Yucky.

No function is provided.

There is no conserved domain.



Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

You can download the probabilities used to generate this plot here

- The graph does not seem to cross the membrane axis.

- So it is a hypothetical protein.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- It is a hypothetical protein because

- BLAST called it a hypothetical protein.
- Hhpred called it hypothetical protein and minor tail protein. But there is no strong evidence for minor tail protein.
- Phamerator show that there is no function assigned in the same gene in the same pham.
- Deep THMHH gave a graph that do not cross membrane axis.

Feature 30 – reverse – stop 24617

Glimmer/GeneMark

What feature number is this? 30

What is the stop site? 24617

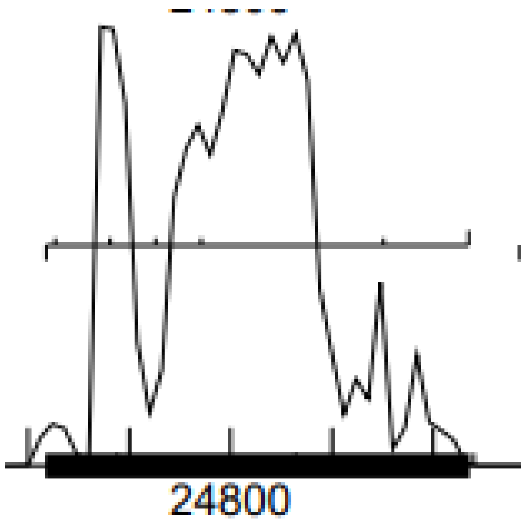
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Called by both.

What is the autoannotated start? 25036

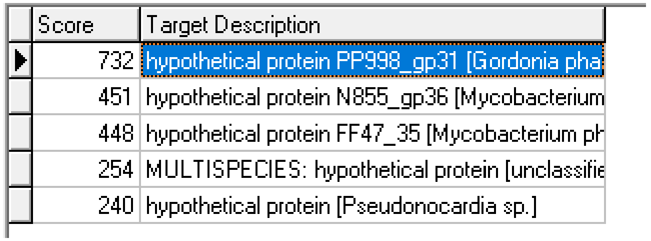
Gap: _________ or overlap: __4______ (with gene in front of it) for the autoannotated start

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Reading frame 4 has a lot of strong coding potential. It is the only frame with coding potential.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
732	hypothetical protein PP998_gp31 [Gordonia pha
451	hypothetical protein N855_gp36 [Mycobacterium
448	hypothetical protein FF47_35 [Mycobacterium ph
254	MULTISPECIES: hypothetical protein [unclassifie
240	hypothetical protein [Pseudonocardia sp.]

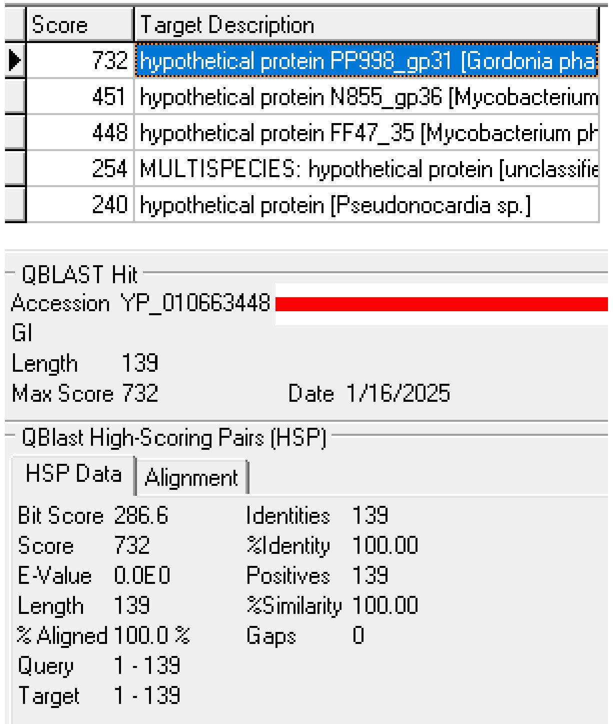
- All 5 BLAST hits have an E-value close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This is a gene, it is called by Glimmer and GeneMark, has 5 BLAST hits with E-values close to 0 and has strong coding potential.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- **There is 1 1:1 alignment. No other start is known yet.**



Score	Target Description
732	hypothetical protein PP998_gp31 [Gordonia pha
451	hypothetical protein N855_gp36 [Mycobacterium
448	hypothetical protein FF47_35 [Mycobacterium ph
254	MULTISPECIES: hypothetical protein [unclassifie
240	hypothetical protein [Pseudonocardia sp.]

QBLAST Hit

Accession YP_010663448

GI

Length 139

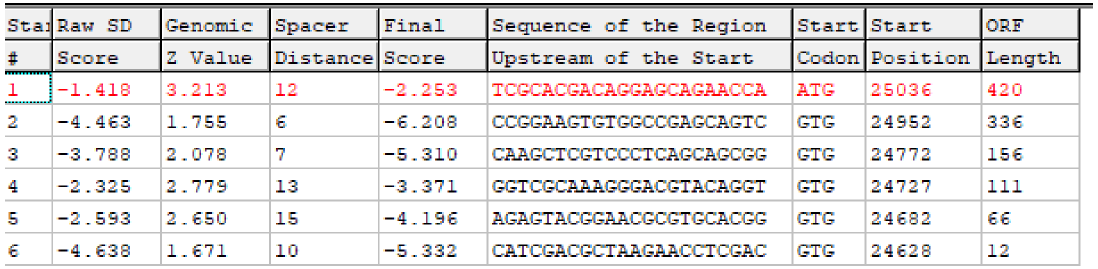
Max Score 732 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	286.6	Identities	139
Score	732	%Identity	100.00
E-Value	0.0E0	Positives	139
Length	139	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 139		
Target	1 - 139		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

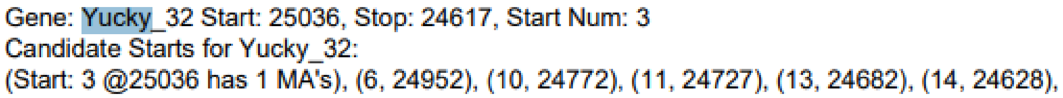
- The Z-value is 3.213 and the final score is -2.253. No other site has decent RBS numbers.



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	-1.418	3.213	12	-2.253	TCGCACGACAGGAGCAGAACCA	ATG	25036	420
2	-4.463	1.755	6	-6.208	CCGGAAGTGTGGCCGAGCAGTC	GTG	24952	336
3	-3.788	2.078	7	-5.310	CAAGCTCGTCCCTCAGCAGCGG	GTG	24772	156
4	-2.325	2.779	13	-3.371	GGTCGCAAAGGGACGTACAGGT	GTG	24727	111
5	-2.593	2.650	15	-4.196	AGAGTACGGAACGCGTGCACGG	GTG	24682	66
6	-4.638	1.671	10	-5.332	CATCGACGCTAAGAACCTCGAC	GTG	24628	12

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

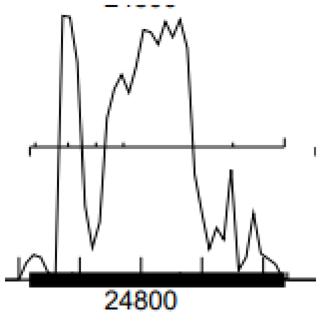
- The automated start has 1 MA and it is the only site to ever receive MA's.



Gene: Yucky_32 Start: 25036, Stop: 24617, Start Num: 3
Candidate Starts for Yucky_32:
(Start: 3 @25036 has 1 MA's), (6, 24952), (10, 24772), (11, 24727), (13, 24682), (14, 24628),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- This start does not cut off any coding potential.



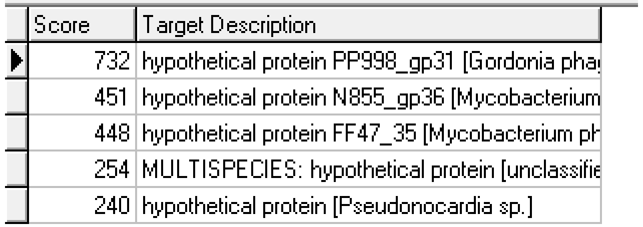
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 25036-25033=3+1 for overlap of 4

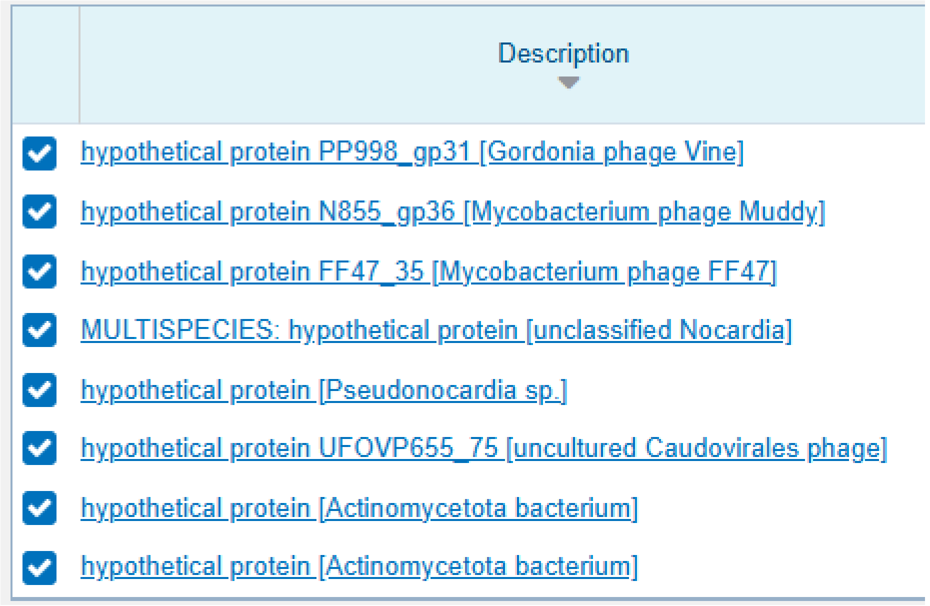
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The automated start is the true start. The BLAST evidence shows a 1:1 alignment, it is the only site with good RBS numbers, it is the only site to ever receive MA’s, it cuts off no coding potential, and it has a overlap of 4

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
732	hypothetical protein PP998_gp31 [Gordonia pha
451	hypothetical protein N855_gp36 [Mycobacterium
448	hypothetical protein FF47_35 [Mycobacterium ph
254	MULTISPECIES: hypothetical protein [unclassifie
240	hypothetical protein [Pseudonocardia sp.]

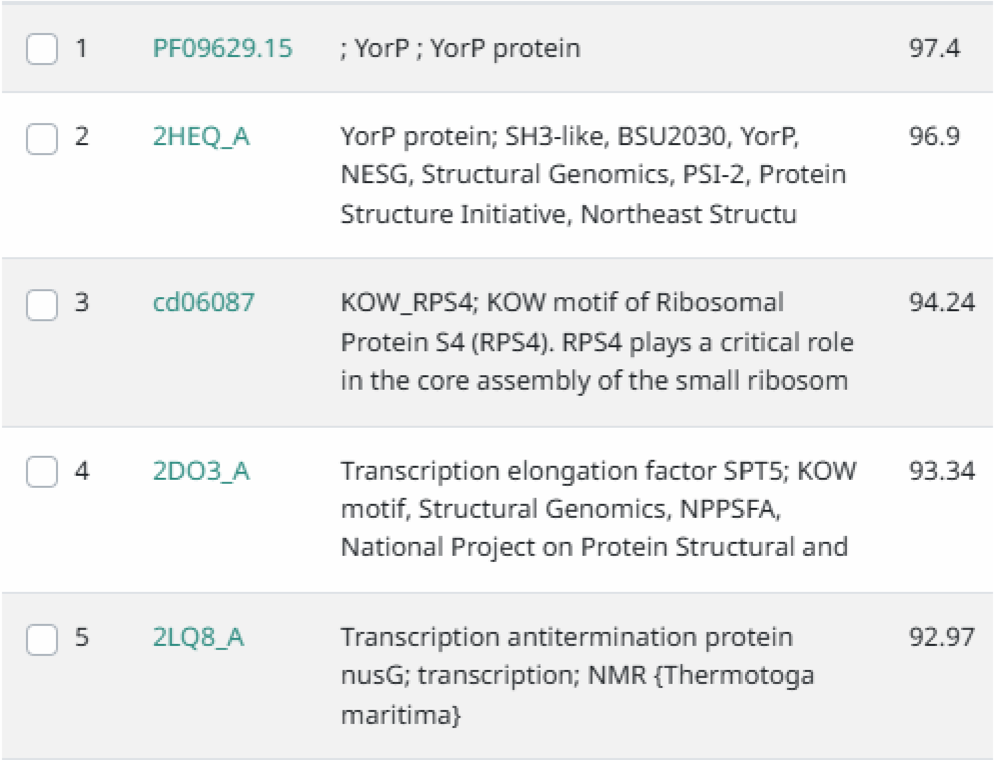


Description

- ☑ hypothetical protein PP998_gp31 [Gordonia phage Vine]
- ☑ hypothetical protein N855_gp36 [Mycobacterium phage Muddy]
- ☑ hypothetical protein FF47_35 [Mycobacterium phage FF47]
- ☑ MULTISPECIES: hypothetical protein [unclassified Nocardia]
- ☑ hypothetical protein [Pseudonocardia sp.]
- ☑ hypothetical protein UFOVP655_75 [uncultured Caudovirales phage]
- ☑ hypothetical protein [Actinomycetota bacterium]
- ☑ hypothetical protein [Actinomycetota bacterium]

- All 5 BLAST hits are hypothetical proteins.
- NCBI BLAST yielded the same results.

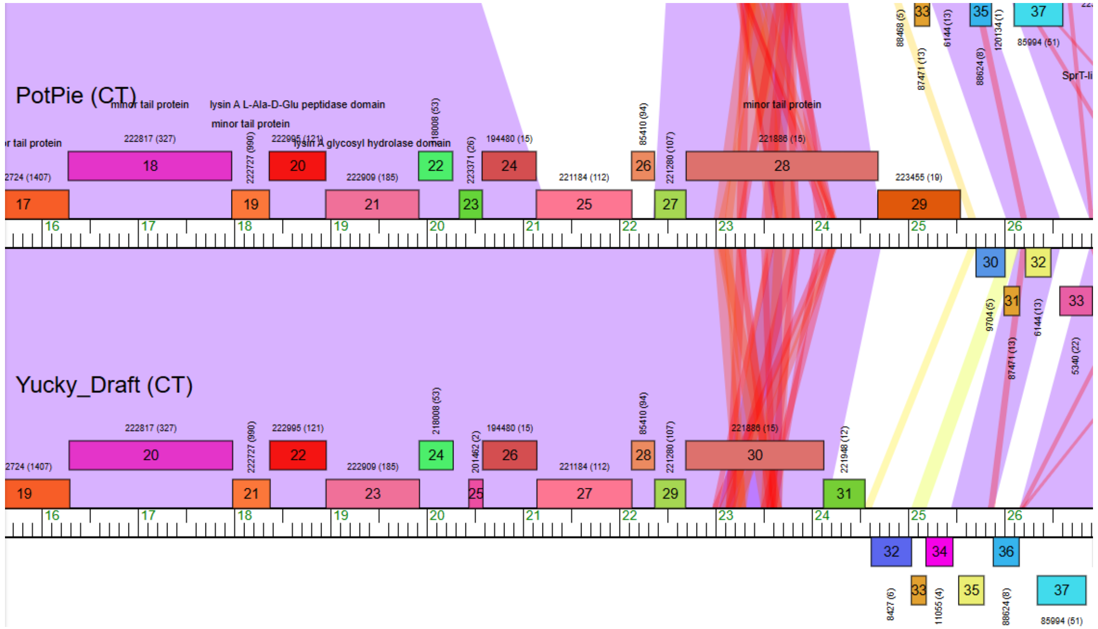
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



□	1	PF09629.15	; YorP ; YorP protein	97.4
□	2	2HEQ_A	YorP protein; SH3-like, BSU2030, YorP, NESG, Structural Genomics, PSI-2, Protein Structure Initiative, Northeast Structu	96.9
□	3	cd06087	KOW_RPS4; KOW motif of Ribosomal Protein S4 (RPS4). RPS4 plays a critical role in the core assembly of the small ribosom	94.24
□	4	2DO3_A	Transcription elongation factor SPT5; KOW motif, Structural Genomics, NPPSFA, National Project on Protein Structural and	93.34
□	5	2LQ8_A	Transcription antitermination protein nusG; transcription; NMR {Thermotoga maritima}	92.97

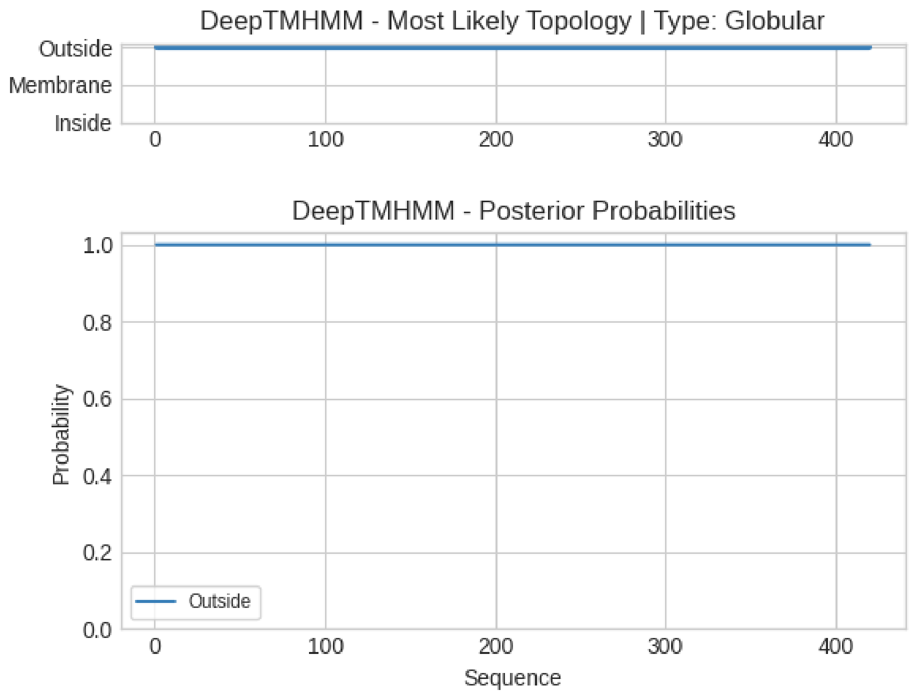
- Hhpred shows many hits for a ribosomal protein. I don’t believe this to be strong enough evidence to overwrite the other evidence.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- None of the phages I’ve been looking at have this gene.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- It is not an intermembrane protein and it functions outside of the membrane.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I am assigning this as a hypothetical protein. BLAST via both DNA Master and NCBI yield only results for hypothetical proteins. Hhpred shows results for a ribosomal protein, but I don’t believe this to be enough evidence. None of the similar phages I have been looking at have this gene. Lastly, it is not a transmembrane protein.

Feature 31 – reverse – stop 25033

Glimmer/GeneMark

What feature number is this?

What is the stop site?

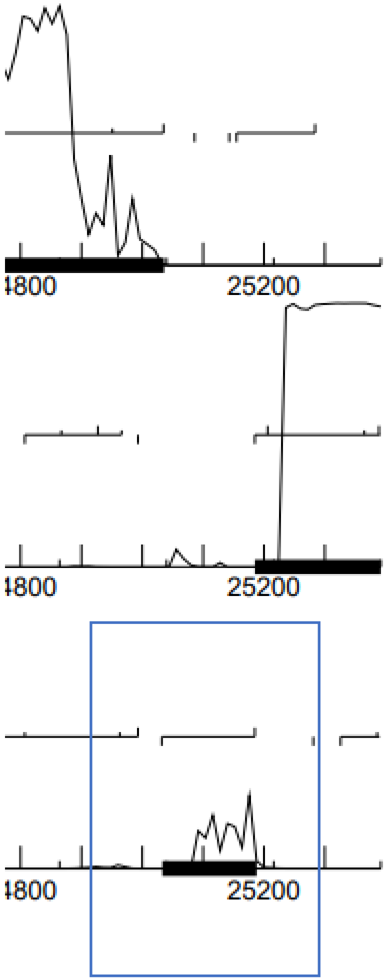
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

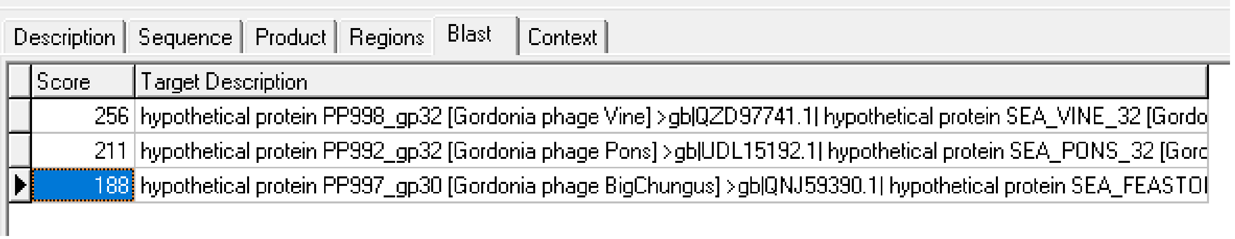
- Feature 31
- Stop Site: 25033
- Start is called by both Glimmer and GeneMark
- Auto-annotated start site: 25188
- Start 25188 has a 4 bp overlap with feature 34

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Weak coding potential
- There is another reading frame with very weak coding potential

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



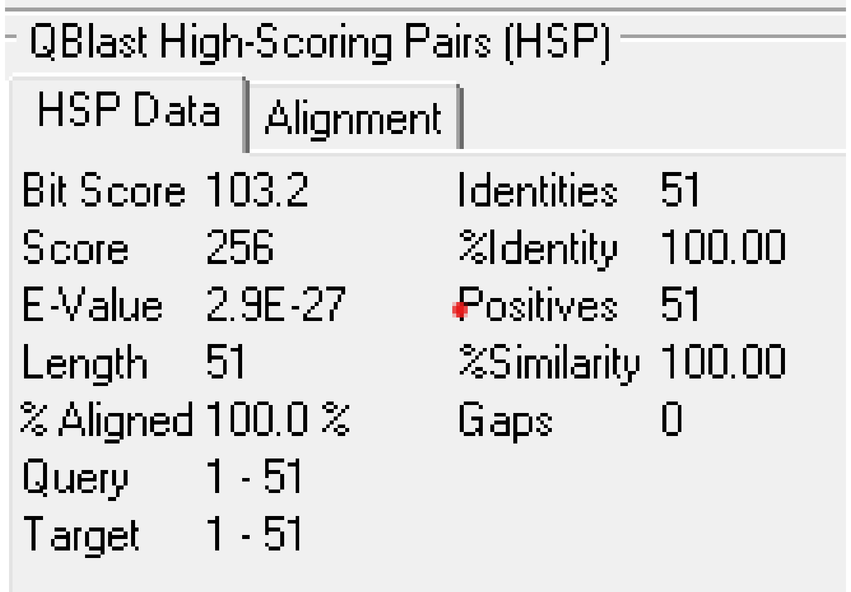
<table>
<tr><th>Score</th><th>Target Description</th></tr>
<tr><td>256</td><td>hypothetical protein PP998_gp32 [Gordonia phage Vine] >gb|QZD97741.1| hypothetical protein SEA_VINE_32 [Gordo</td></tr>
<tr><td>211</td><td>hypothetical protein PP992_gp32 [Gordonia phage Pons] >gb|UDL15192.1| hypothetical protein SEA_PONS_32 [Gorc</td></tr>
<tr><td>188</td><td>hypothetical protein PP997_gp30 [Gordonia phage BigChungus] >gb|QNJ59390.1| hypothetical protein SEA_FEASTOI</td></tr>
</table>

- There are hits to 3 other highly similar genes in Gordonia CT cluster phages

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, it is a gene.
- Glimmer and GeneMark called the gene
- There is coding potential
- Has BLAST hits to 3 other Gordonia CT cluster phage

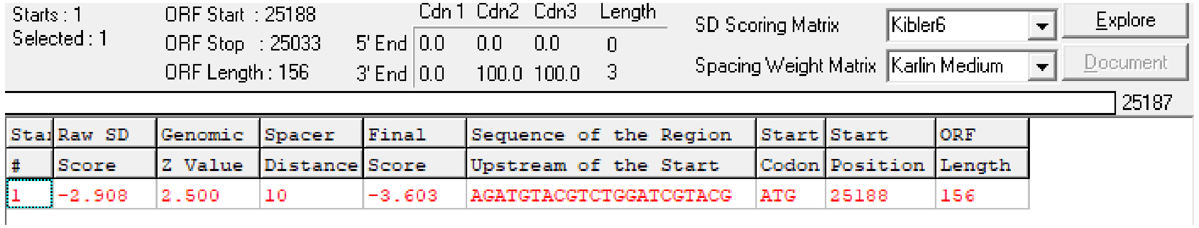
BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Bit Score	103.2	Identities	51
Score	256	%Identity	100.00
E-Value	2.9E-27	Positives	51
Length	51	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 51		
Target	1 - 51		

- There are 3 Q1:S1 alignments with other Gordonia CT cluster phage
- 94-100% alignment, good E-values
- There are no alternative starts
- Start 25188 is favored

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



Starts : 1
Selected : 1

ORF Start : 25188
ORF Stop : 25033
ORF Length : 156

	Cdn 1	Cdn2	Cdn3	Length
5' End	0.0	0.0	0.0	0
3' End	0.0	100.0	100.0	3

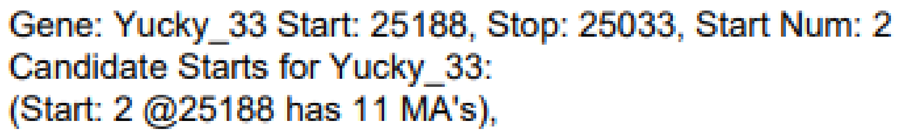
SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-2.908	2.500	10	-3.603	AGATGTACGTCTGGATCGTACG	ATG	25188	156

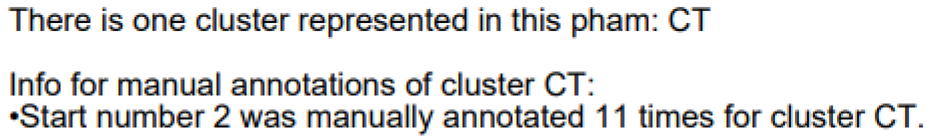
- Z-Value: 2.500
- Final Score: -3.603

- 25188 is the favored and only start

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



Gene: Yucky_33 Start: 25188, Stop: 25033, Start Num: 2
Candidate Starts for Yucky_33:
(Start: 2 @25188 has 11 MA's),

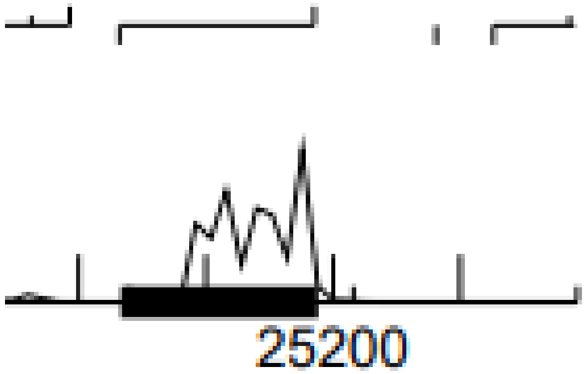


There is one cluster represented in this pham: CT

Info for manual annotations of cluster CT:
•Start number 2 was manually annotated 11 times for cluster CT.

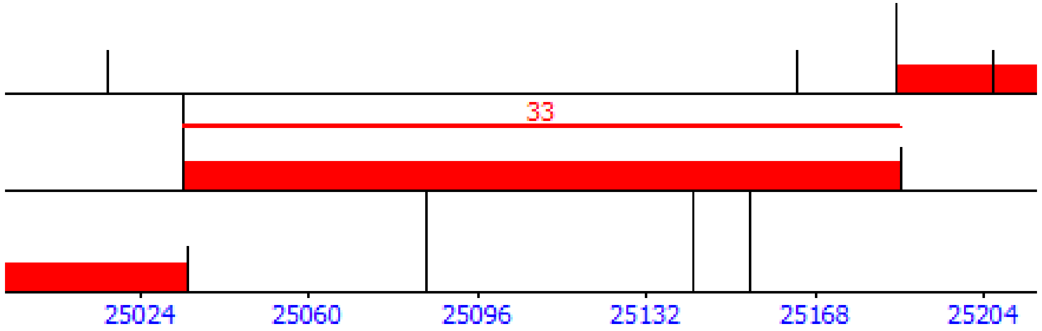
- There are 11 manual annotations for the proposed start
- The proposed start aligns with all other pham members
- There are no other possible starts

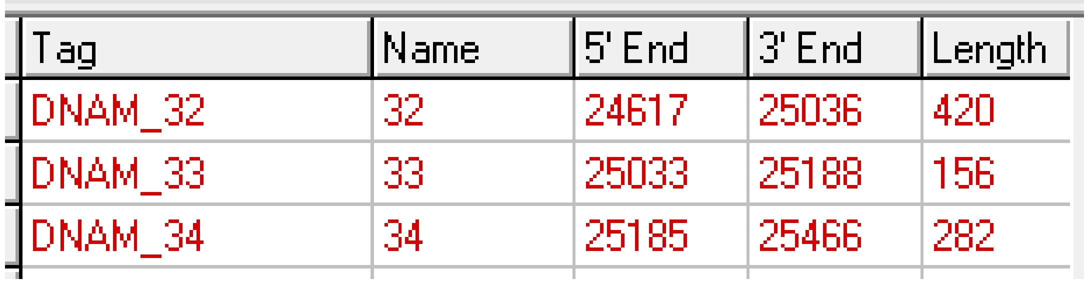
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- A tiny amount of CP is cut off, but there are no other possible starts to include the cut off CP

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.





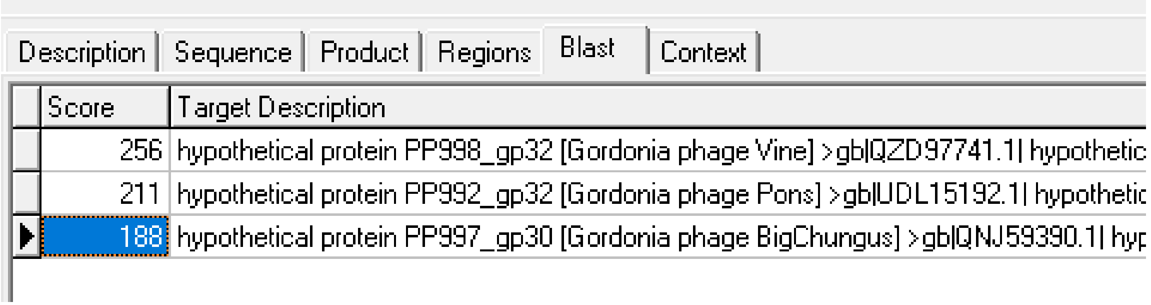
Tag	Name	5' End	3' End	Length
DNAM_32	32	24617	25036	420
DNAM_33	33	25033	25188	156
DNAM_34	34	25185	25466	282

- There is a 4 bp overlap with the with the stop of the downstream gene

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- Start Site: 25188
- This agrees with the auto-annotated start. It is the only possible start for this feature
- There are 11 manual annotations for this start from other Gordonia CT cluster phage

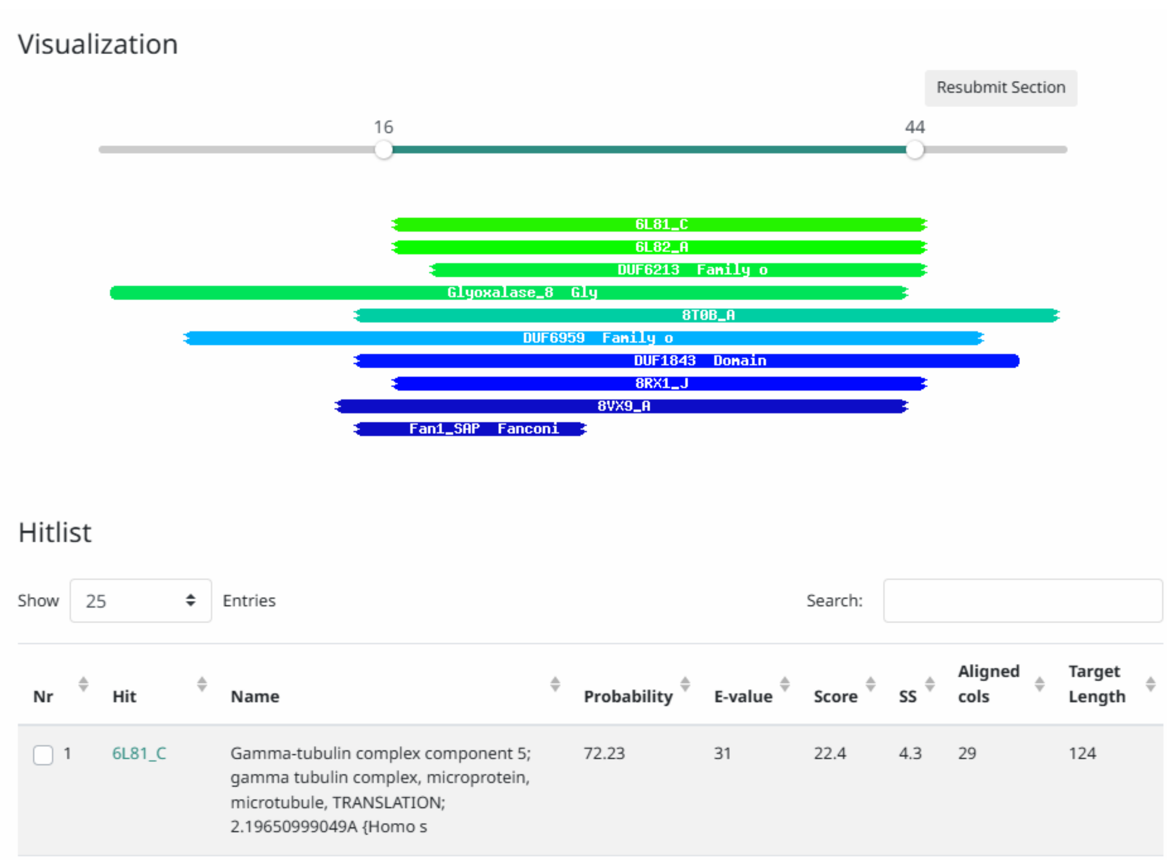
BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
256	hypothetical protein PP998_gp32 [Gordonia phage Vine] >gb|QZD97741.1| hypothetic
211	hypothetical protein PP992_gp32 [Gordonia phage Pons] >gb|UDL15192.1| hypothetic
188	hypothetical protein PP997_gp30 [Gordonia phage BigChungus] >gb|QNJ59390.1| hyp

- Other Gordonia phage assigned the function Hypothetical Protein

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



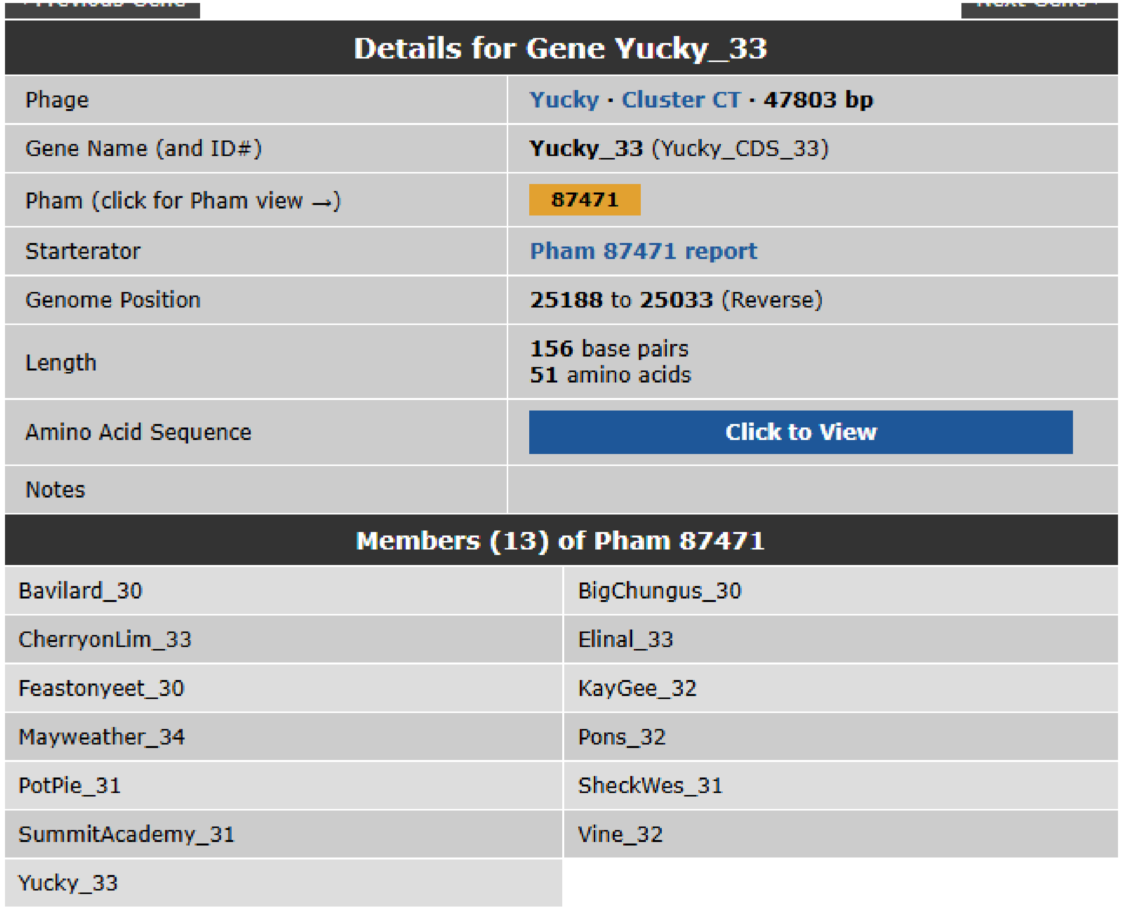
Hitlist

Show 25 Entries Search:

Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	6L81_C	Gamma-tubulin complex component 5; gamma tubulin complex, microprotein, microtubule, TRANSLATION; 2.19650999049A {Homo s	72.23	31	22.4	4.3	29	124

- NKF, there are no hits with a probability >90%

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

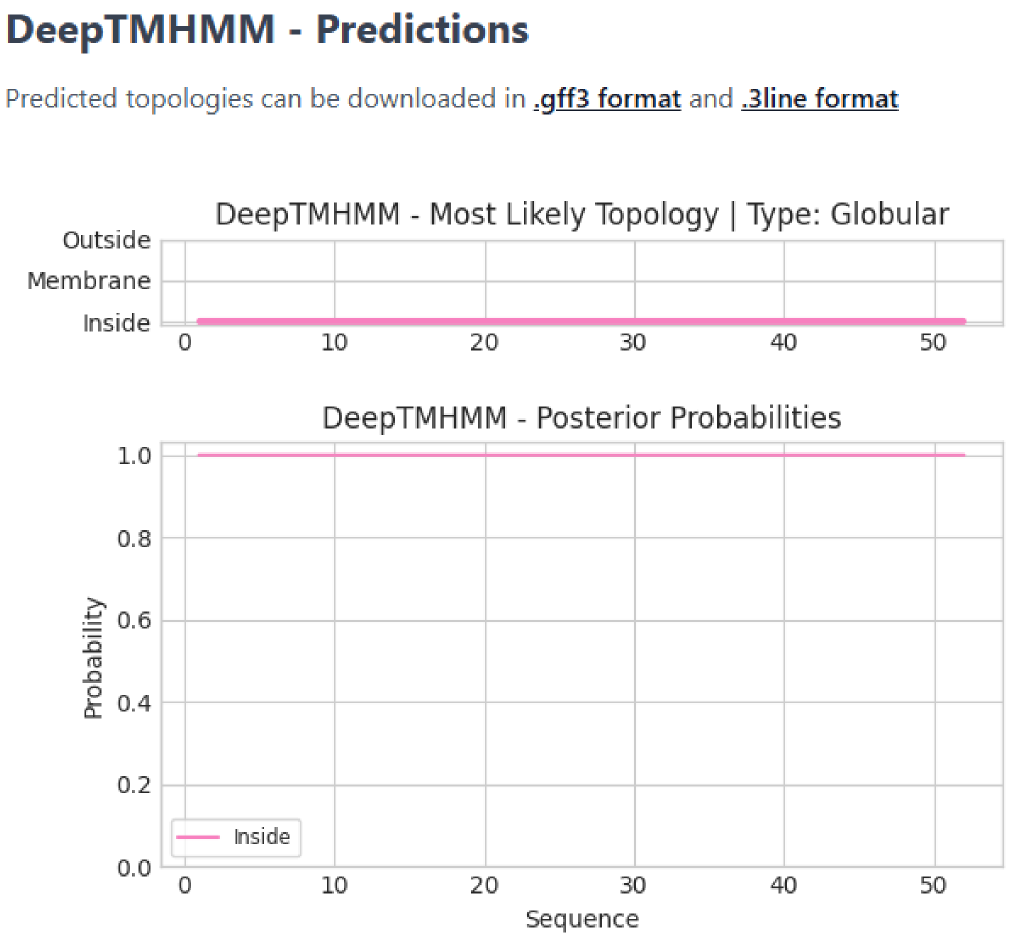


Details for Gene Yucky_33	
Phage	Yucky · Cluster CT · 47803 bp
Gene Name (and ID#)	**Yucky_33** (Yucky_CDS_33)
Pham (click for Pham view →)	87471
Starterator	Pham 87471 report
Genome Position	**25188** to **25033** (Reverse)
Length	**156** base pairs **51** amino acids
Amino Acid Sequence	Click to View
Notes	

Members (13) of Pham 87471	
Bavilard_30	BigChungus_30
CherryonLim_33	Elinal_33
Feastonyeet_30	KayGee_32
Mayweather_34	Pons_32
PotPie_31	SheckWes_31
SummitAcademy_31	Vine_32
Yucky_33	

- 12 other Gordonia CT phages have this gene; all are hypothetical proteins
- There are no conserved domains

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

- There are no predicted TMRs

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Hypothetical Protein

- All BLAST hits are to Hypothetical Proteins
- HHPred had no hits with probability >90%
- The 12 other Gordonia CT phages in the pham have assigned the gene funcation as hypothetical protein. There are no conserved domains.

Feature 32 – reverse – stop 25185

Glimmer/GeneMark

What feature number is this? 32

What is the stop site? 25185

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Called by both Glimmer and GeneMark

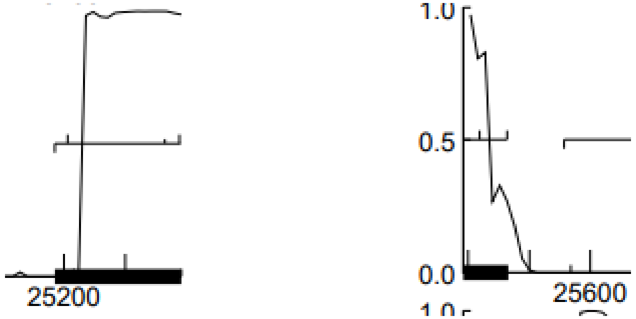
What is the autoannotated start?

25446

Gap: 58 or overlap: _______ (with gene in front of it) for the autoannotated start

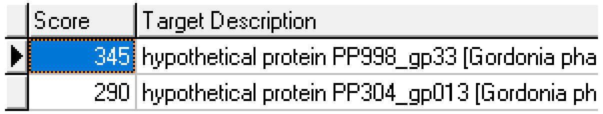
- Called by both
- Gap of 58

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

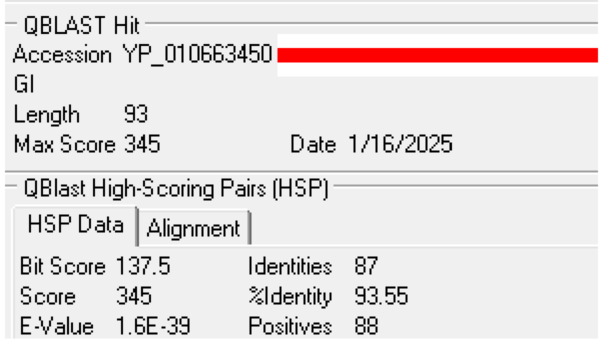


- There is a consistent peak of strong coding potential on reading frame 5. No other frame has coding potential

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



	Score	Target Description
▶	345	hypothetical protein PP998_gp33 [Gordonia pha
	290	hypothetical protein PP304_gp013 [Gordonia ph



QBLAST Hit

Accession YP_010663450

GI

Length 93

Max Score 345 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 137.5 Identities 87

Score 345 %Identity 93.55

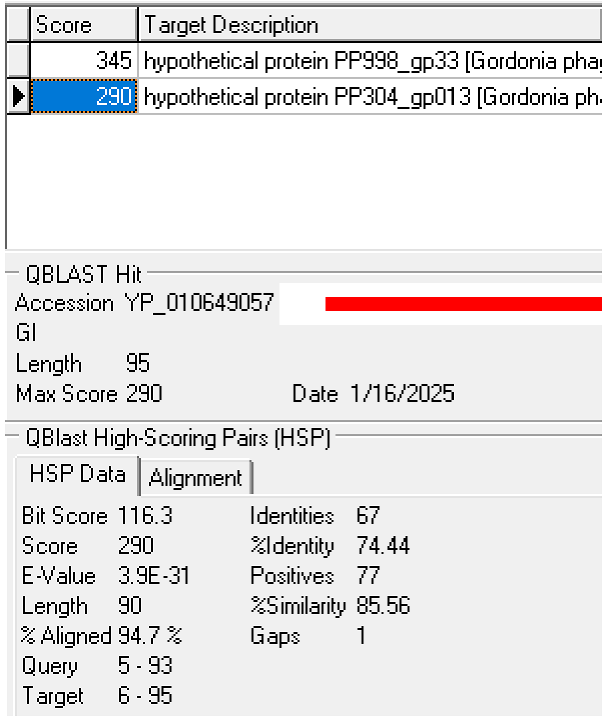
E-Value 1.6E-39 Positives 88

- There are 2 BLAST hits E-values close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- I believe this to be a gene. It is called by both Glimmer and GeneMark. There is a strong peak of coding potential throughout the nucleotide sequence. There are 2 BLAST hits wit E-values close to 0. These pieces of evidence lead me to believe this is a gene.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
345	hypothetical protein PP998_gp33 [Gordonia pha
290	hypothetical protein PP304_gp013 [Gordonia ph

QBLAST Hit

Accession YP_010649057

GI

Length 95

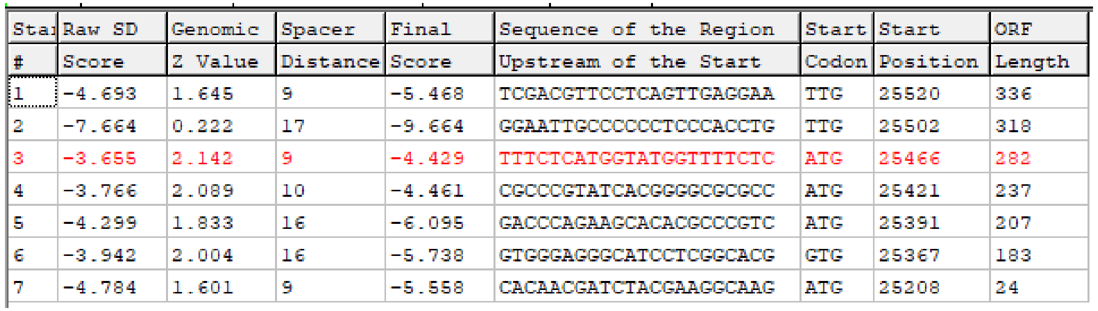
Max Score 290 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	116.3	Identities	67
Score	290	%Identity	74.44
E-Value	3.9E-31	Positives	77
Length	90	%Similarity	85.56
% Aligned	94.7 %	Gaps	1
Query	5 - 93		
Target	6 - 95		

- There is one 1:1 alignment and a 5:6 alignment. No alternative starts are known at this time since Glimmer and GeneMark agree on the start site.

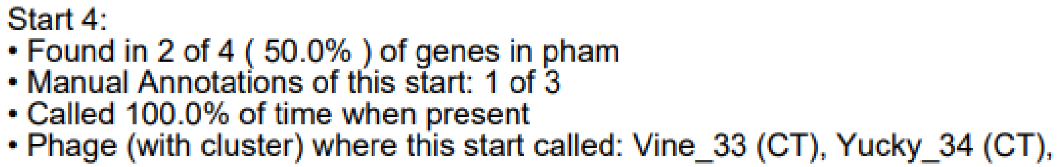
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



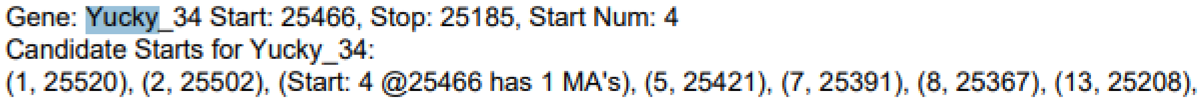
Sta	Raw SD	Genomic	Spacer	Final	Sequence of the Region	Start	Start	ORF
#	Score	Z Value	Distance	Score	Upstream of the Start	Codon	Position	Length
1	-4.693	1.645	9	-5.468	TCGACGTTCCTCAGTTGAGGAA	TTG	25520	336
2	-7.664	0.222	17	-9.664	GGAATTGCCCCCCTCCCACCTG	TTG	25502	318
3	-3.655	2.142	9	-4.429	TTTCTCATGGTATGGTTTTCTC	ATG	25466	282
4	-3.766	2.089	10	-4.461	CGCCCGTATCACGGGGCGCGCC	ATG	25421	237
5	-4.299	1.833	16	-6.095	GACCCAGAAGCACACGCCCGTC	ATG	25391	207
6	-3.942	2.004	16	-5.738	GTGGGAGGGCATCCTCGGCACG	GTG	25367	183
7	-4.784	1.601	9	-5.558	CACAACGATCTACGAAGGCAAG	ATG	25208	24

- Z-value: 2.142
- Final score:-4.429
- No other RBS values indicate a start site

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



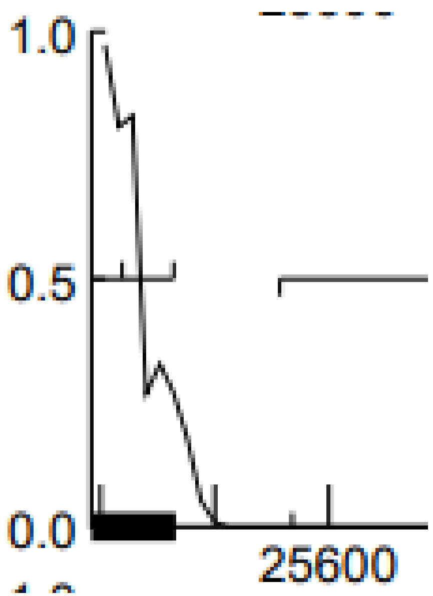
Start 4:
- Found in 2 of 4 (50.0%) of genes in pham
- Manual Annotations of this start: 1 of 3
- Called 100.0% of time when present
- Phage (with cluster) where this start called: Vine_33 (CT), Yucky_34 (CT),



Gene: Yucky_34 Start: 25466, Stop: 25185, Start Num: 4
Candidate Starts for Yucky_34:
(1, 25520), (2, 25502), (Start: 4 @25466 has 1 MA's), (5, 25421), (7, 25391), (8, 25367), (13, 25208),

- The autoannotated start has 1 MA. It is the only proposed site to receive a manual annotation. It is called 100% of the time when present.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- There is a slight bit of coding potential cut off at the start site, seems to be the beginning of a peak.

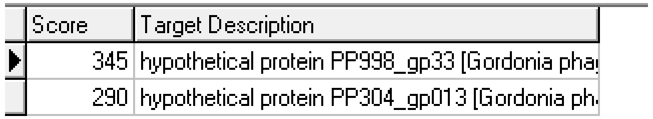
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- There is a gap of 78, this is not ideal but it is acceptable.
- 25525-25466= 59-1= 58

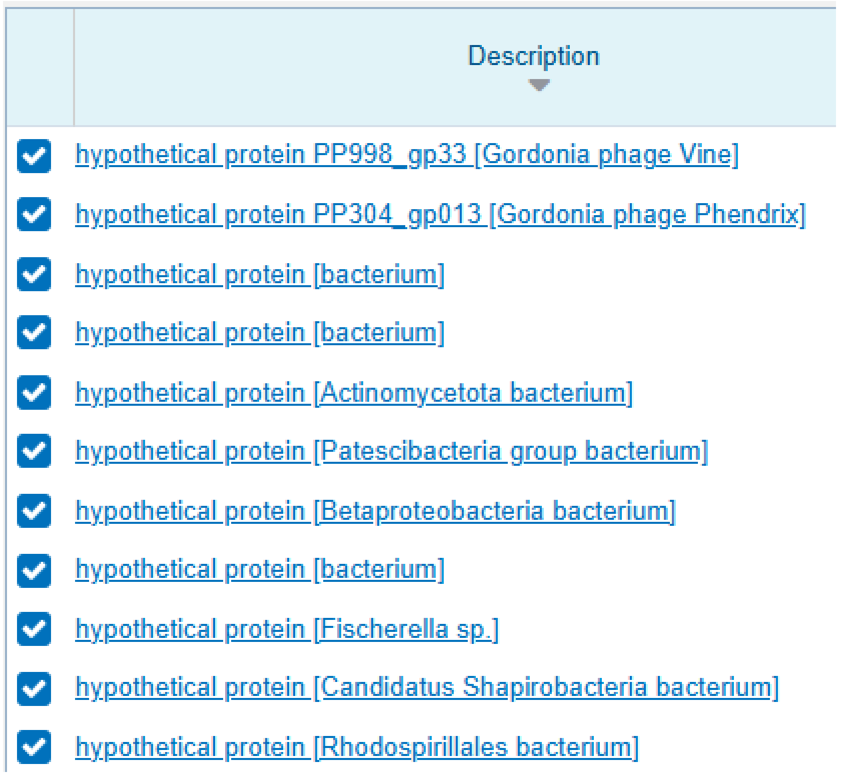
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The autoannotated start site is the start site (25466). There is a 1:1 alignment on BLAST. The RBS numbers are good with a Z-value of 2.142 and a final score of -4.429. It has a MA and is called 100% of the time when present. It only cuts off a slight piece of coding potential, and it has a big, but acceptable gap.

BLAST function evidence. What assigned functions do other highly similar genes have?



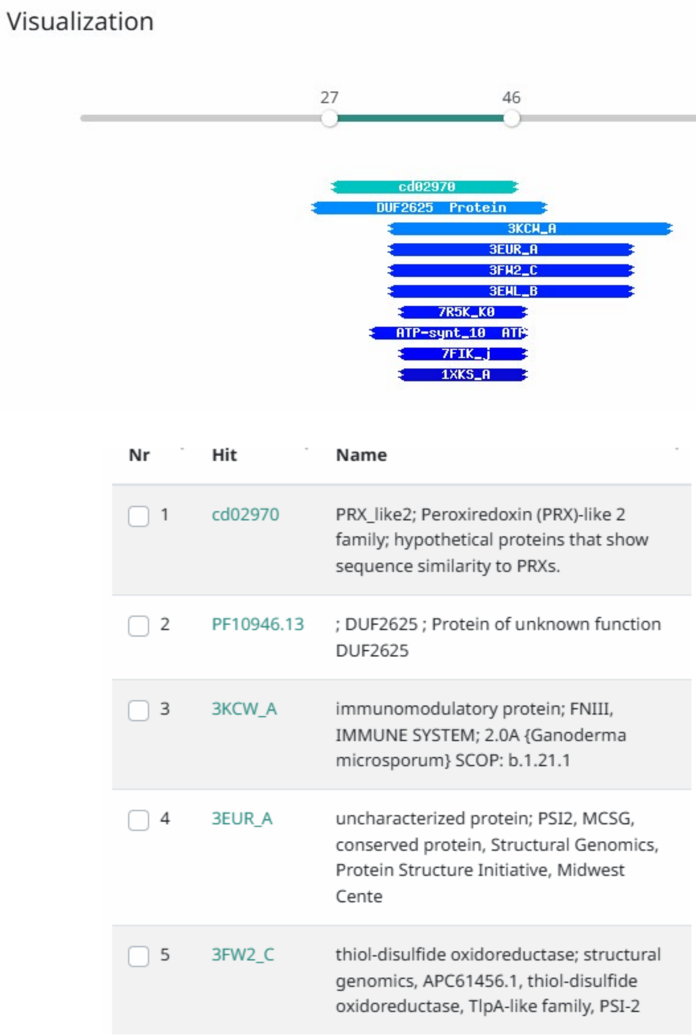
Score	Target Description
345	hypothetical protein PP998_gp33 [Gordonia pha
290	hypothetical protein PP304_gp013 [Gordonia ph



	Description
☑	hypothetical protein PP998_gp33 [Gordonia phage Vine]
☑	hypothetical protein PP304_gp013 [Gordonia phage Phendrix]
☑	hypothetical protein [bacterium]
☑	hypothetical protein [bacterium]
☑	hypothetical protein [Actinomycetota bacterium]
☑	hypothetical protein [Patescibacteria group bacterium]
☑	hypothetical protein [Betaproteobacteria bacterium]
☑	hypothetical protein [bacterium]
☑	hypothetical protein [Fischerella sp.]
☑	hypothetical protein [Candidatus Shapirobacteria bacterium]
☑	hypothetical protein [Rhodospirillales bacterium]

- DNA master BLAST results show only 2 hits, and both are as hypothetical proteins.
- NCBI only shows hits as a hypothetical protein.

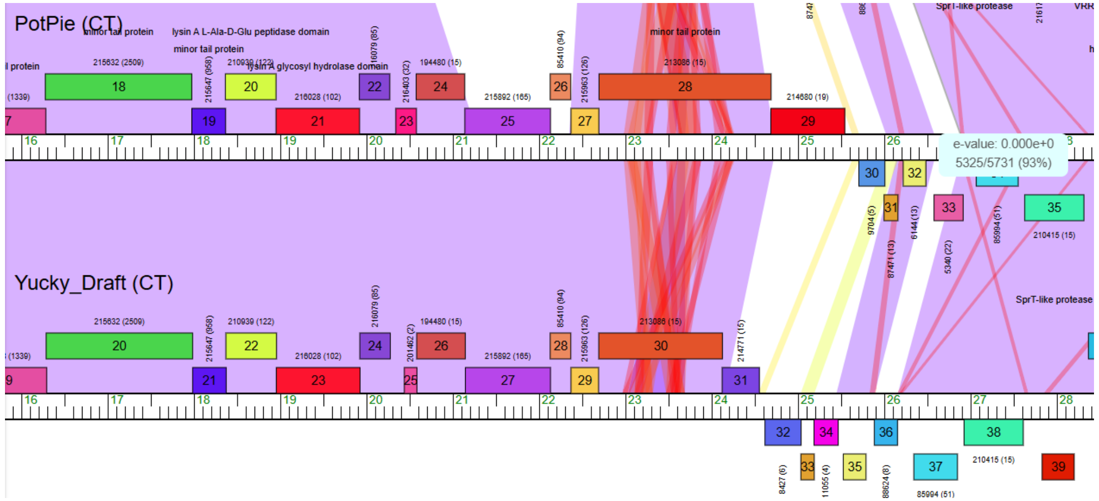
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



Nr	Hit	Name
□ 1	cd02970	PRX_like2; Peroxiredoxin (PRX)-like 2 family; hypothetical proteins that show sequence similarity to PRXs.
□ 2	PF10946.13	; DUF2625 ; Protein of unknown function DUF2625
□ 3	3KCW_A	immunomodulatory protein; FNIII, IMMUNE SYSTEM; 2.0A {Ganoderma microsporum} SCOP: b.1.21.1
□ 4	3EUR_A	uncharacterized protein; PSI2, MCSG, conserved protein, Structural Genomics, Protein Structure Initiative, Midwest Cente
□ 5	3FW2_C	thiol-disulfide oxidoreductase; structural genomics, APC61456.1, thiol-disulfide oxidoreductase, TlpA-like family, PSI-2

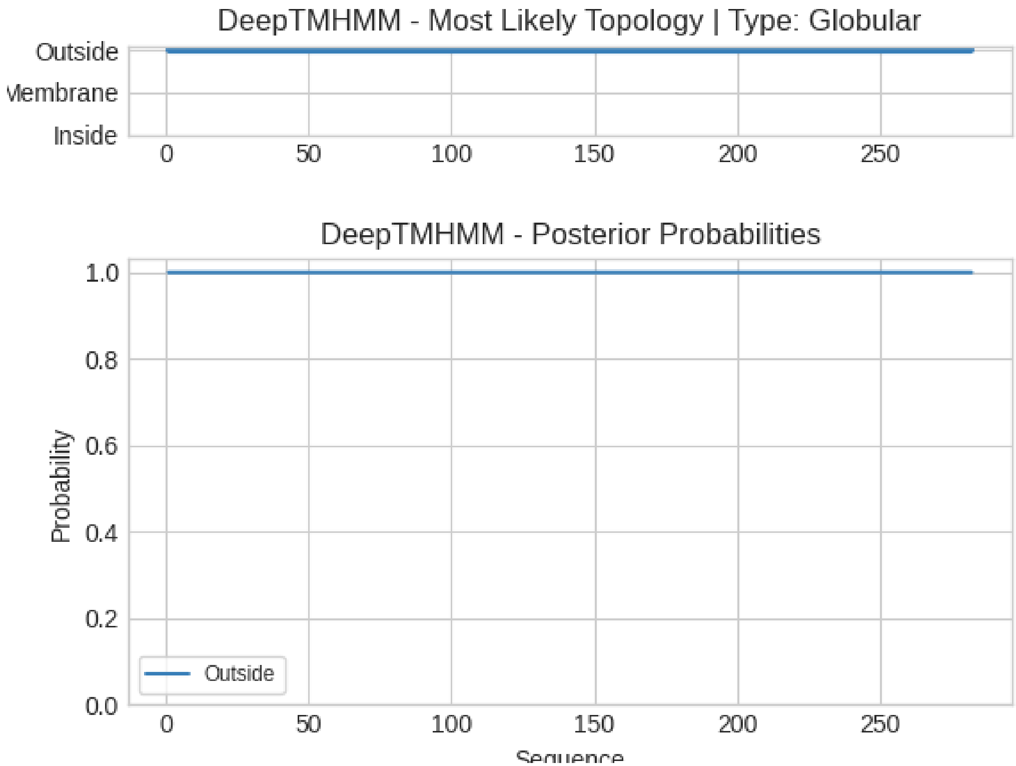
- HHpred shows hits as mostly hypothetical proteins, definitely not enough evidence to call a function.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- BigChungus, PotPie, and Elinal all do not contain this gene.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- This is not a transmembrane protein as it never crosses the membrane.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- This is a hypothetical protein. All shown BLAST hits on DNA master and NCBI are only hypothetical proteins. Hhpred shows a couple proteins with other functions but enough hypothetical proteins and nothing definitive enough to assign a different function.

Feature 33 – reverse – end 25525

Glimmer/GeneMark

What feature number is this? 33

What is the stop site? 25525

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Called by both Glimmer and GeneMark

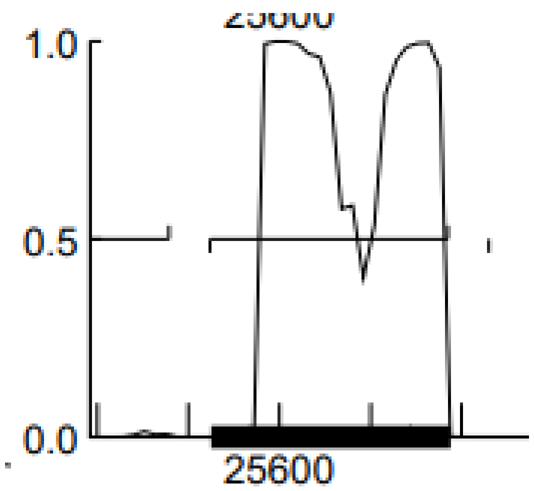
What is the autoannotated start?

25788

Gap: 95 or overlap: _______ (with gene in front of it) for the autoannotated start

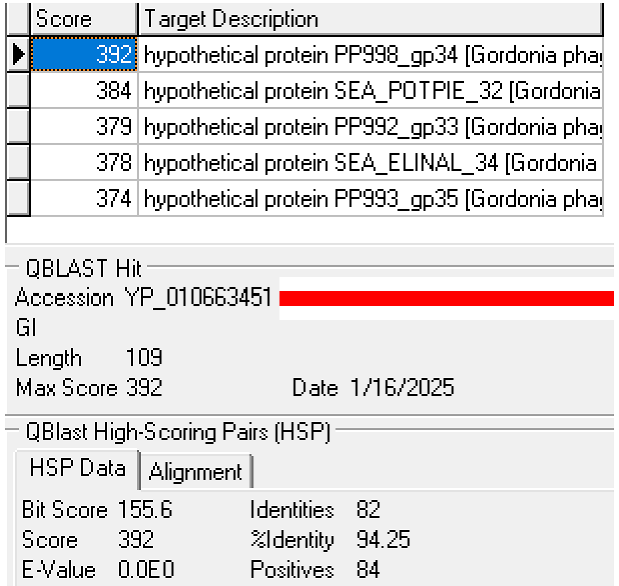
- Called by both
- Gap of 95

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- There are 2 strong peaks of coding potential separated by a weak peak in reading frame 6.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
392	hypothetical protein PP998_gp34 [Gordonia pha
384	hypothetical protein SEA_POTPIE_32 [Gordonia
379	hypothetical protein PP992_gp33 [Gordonia pha
378	hypothetical protein SEA_ELINAL_34 [Gordonia
374	hypothetical protein PP993_gp35 [Gordonia pha

QBLAST Hit

Accession YP_010663451

GI

Length 109

Max Score 392 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

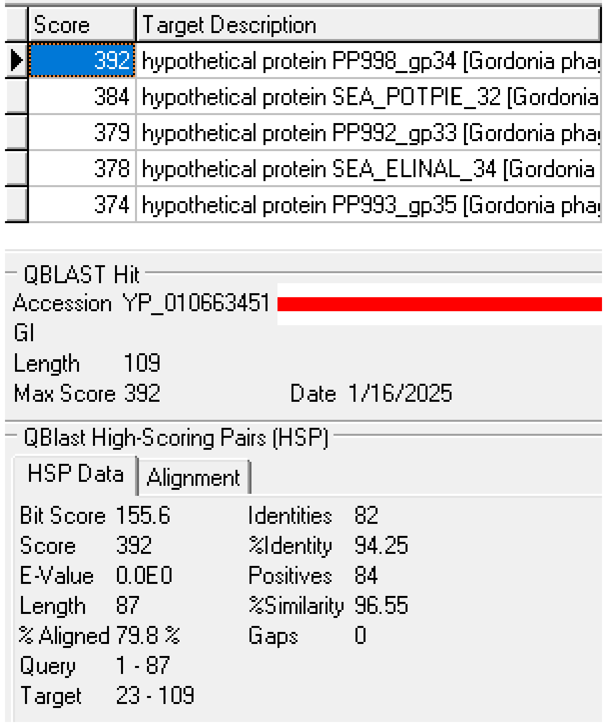
Bit Score	155.6	Identities	82
Score	392	%Identity	94.25
E-Value	0.0E0	Positives	84

- There are 8 BLAST hits with an E-value close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This is a gene. Both Glimmer and GeneMark called it. There is coding potential throughout the nucleotide sequence. There are also 8 BLAST hits with an E-value close to 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
392	hypothetical protein PP998_gp34 [Gordonia pha
384	hypothetical protein SEA_POTPIE_32 [Gordonia
379	hypothetical protein PP992_gp33 [Gordonia pha
378	hypothetical protein SEA_ELINAL_34 [Gordonia
374	hypothetical protein PP993_gp35 [Gordonia pha

QBLAST Hit

Accession YP_010663451

GI

Length 109

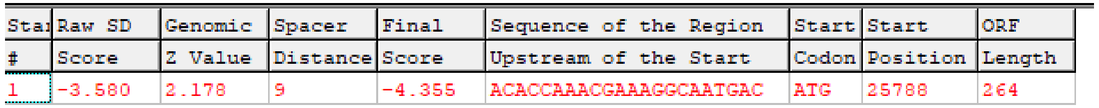
Max Score 392 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	155.6	Identities	82
Score	392	%Identity	94.25
E-Value	0.0E0	Positives	84
Length	87	%Similarity	96.55
% Aligned	79.8 %	Gaps	0
Query	1 - 87		
Target	23 - 109		

- There are 7 1:1 alignments and 1 1:23 alignment.
- No alternative starts are known at this time since Glimmer and GeneMark agree on the start site.

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

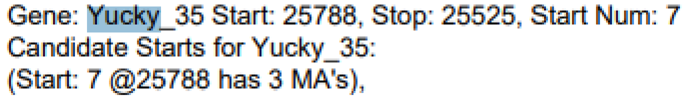


Sta #	Raw SD Score	Genomic Z Value	Spacer Distance	Final Score	Sequence of the Region Upstream of the Start	Start Codon	Start Position	ORF Length
1	-3.580	2.178	9	-4.355	ACACCAAACGAAAGGCAATGAC	ATG	25788	264

- Z-value: 2.178
- Final score: -4.355
- It is the only available start.

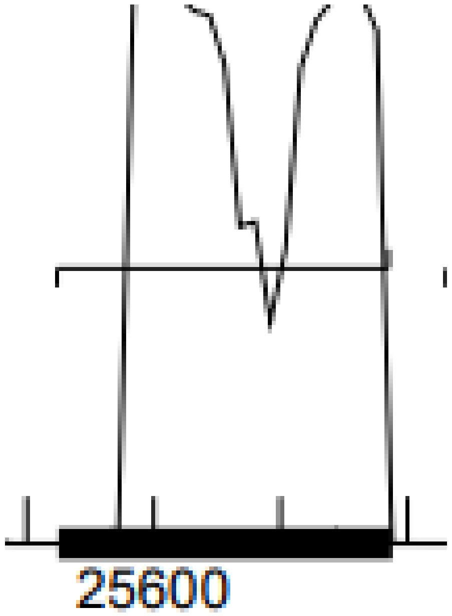
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- The proposed start site is the only one listed by Starterator, it has 3 Mas.



Gene: Yucky_35 Start: 25788, Stop: 25525, Start Num: 7
Candidate Starts for Yucky_35:
(Start: 7 @25788 has 3 MA's),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- This start site cuts off a slight bit of coding potential, particularly the start of a peak.

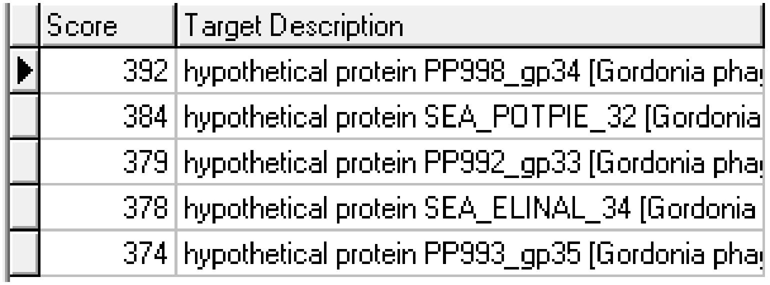
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- There is a gap of 95. This is not ideal, but it is acceptable.
- 25884-25788=96-1=95

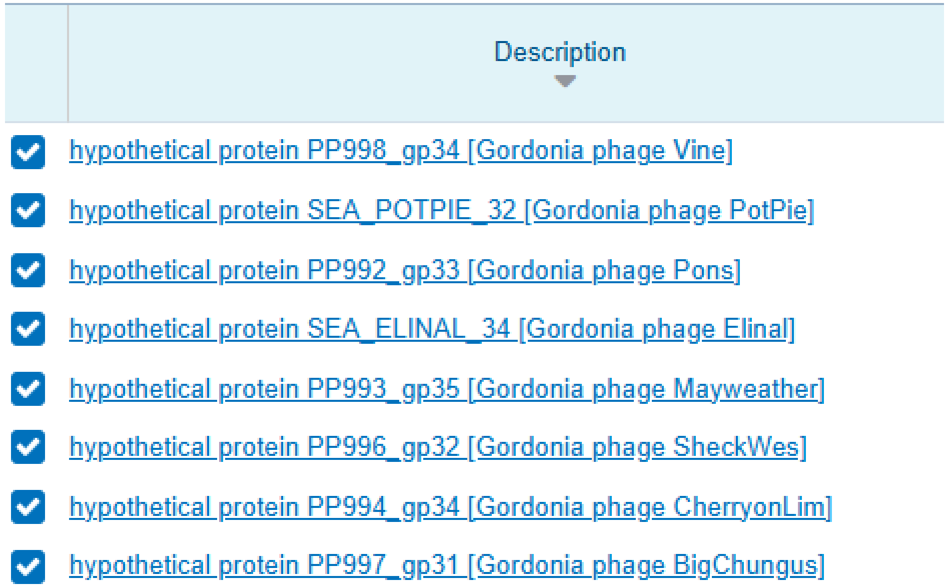
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is 25788. It has multiple 1:1 alignments. A z-value of 2.178 and a Final score of -4.355. RBS also lists it as the only available start. It has 3 MAs and Starterator lists it as the only available start as well. It cuts off minimal coding potential. Lastly, it has an acceptable gap size.

BLAST function evidence. What assigned functions do other highly similar genes have?



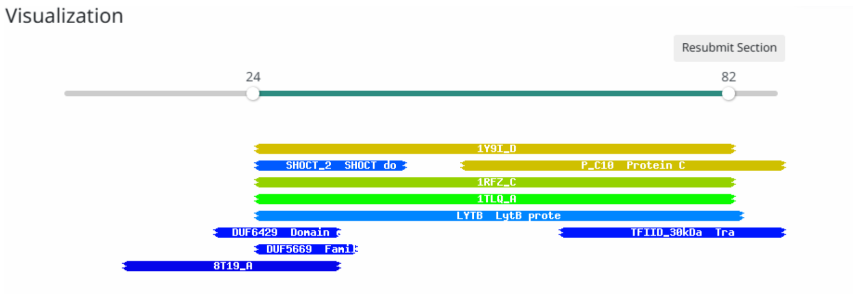
	Score	Target Description
▶	392	hypothetical protein PP998_gp34 [Gordonia pha
	384	hypothetical protein SEA_POTPIE_32 [Gordonia
	379	hypothetical protein PP992_gp33 [Gordonia pha
	378	hypothetical protein SEA_ELINAL_34 [Gordonia
	374	hypothetical protein PP993_gp35 [Gordonia pha

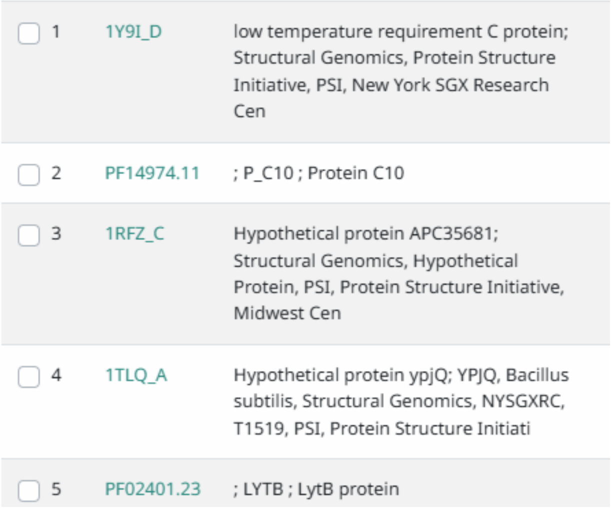


	Description
☑	hypothetical protein PP998_gp34 [Gordonia phage Vine]
☑	hypothetical protein SEA_POTPIE_32 [Gordonia phage PotPie]
☑	hypothetical protein PP992_gp33 [Gordonia phage Pons]
☑	hypothetical protein SEA_ELINAL_34 [Gordonia phage Elinal]
☑	hypothetical protein PP993_gp35 [Gordonia phage Mayweather]
☑	hypothetical protein PP996_gp32 [Gordonia phage SheckWes]
☑	hypothetical protein PP994_gp34 [Gordonia phage CherryonLim]
☑	hypothetical protein PP997_gp31 [Gordonia phage BigChungus]

- DNA master BLAST shows only 8 hits, all of which are hypothetical proteins.
- NCBI BLAST shows the same results.

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

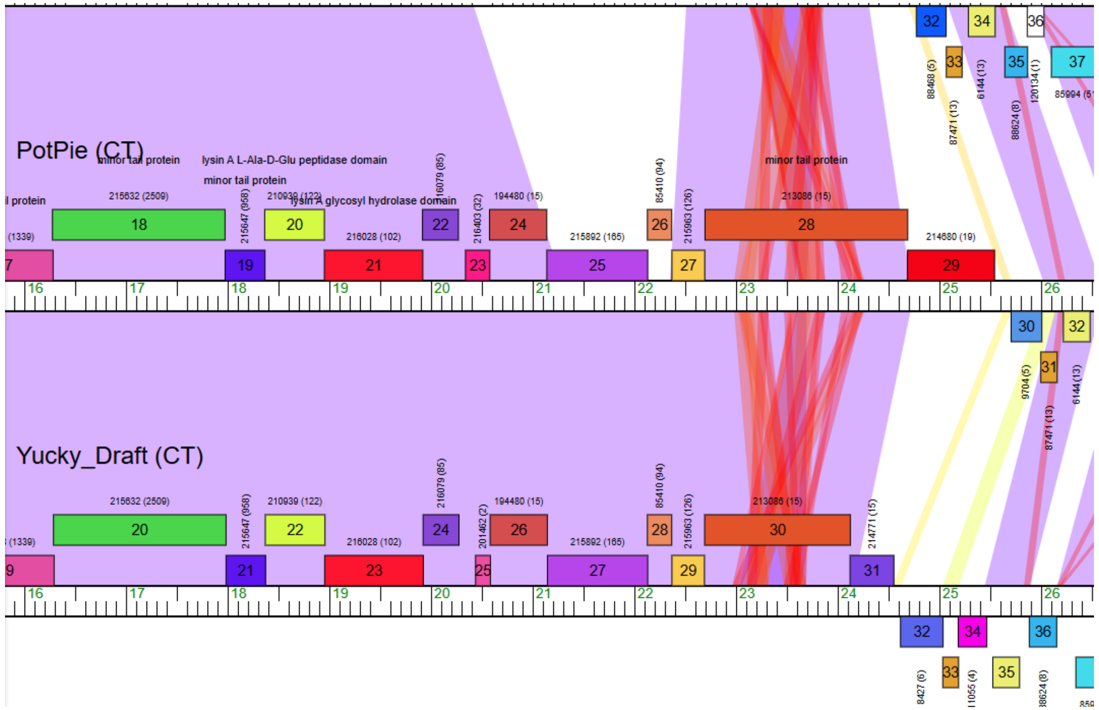




□	1	1Y9I_D	low temperature requirement C protein; Structural Genomics, Protein Structure Initiative, PSI, New York SGX Research Cen
□	2	PF14974.11	; P_C10 ; Protein C10
□	3	1RFZ_C	Hypothetical protein APC35681; Structural Genomics, Hypothetical Protein, PSI, Protein Structure Initiative, Midwest Cen
□	4	1TLQ_A	Hypothetical protein ypjQ; YPJQ, Bacillus subtilis, Structural Genomics, NYSGXRC, T1519, PSI, Protein Structure Initiati
□	5	PF02401.23	; LYTB ; LytB protein

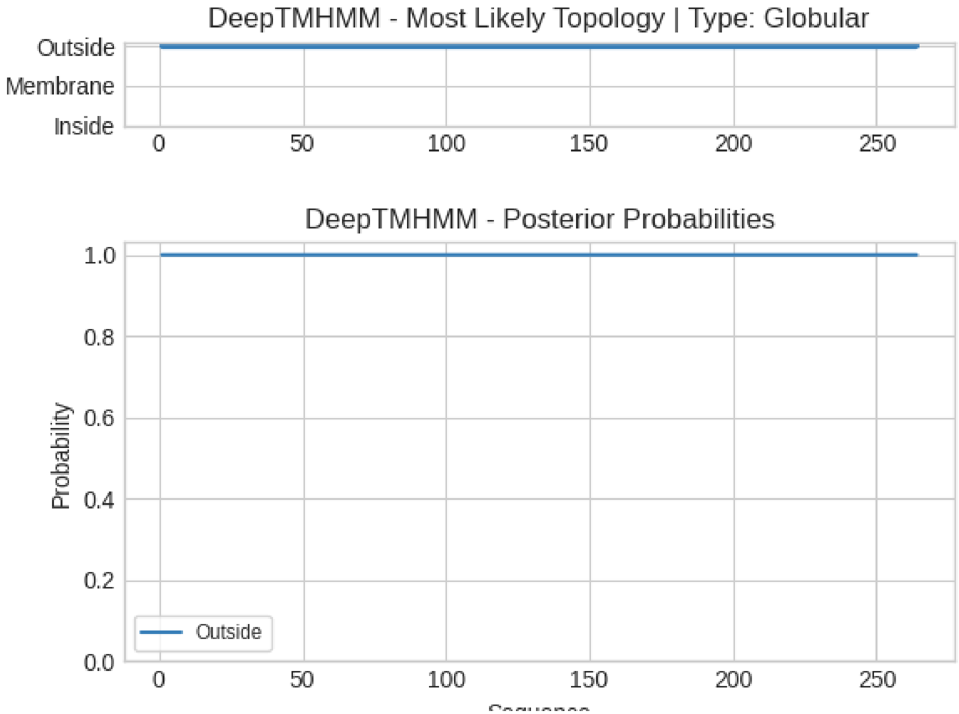
- Hhpred shows hits for a couple of true functions, but mostly hypothetical proteins. The hits are mostly homologous for just a region.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- BigChungus, PotPie, and Elinal all contain this gene, however all 3 of them have it called as a hypothetical protein.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- This is not a transmembrane protein as it never crosses the membrane.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- This is a hypothetical protein. BLAST on both DNA master and NCBI show hits for only hypothetical proteins. Hhpred shows some hits for true functions but not enough for it to be solid evidence. Lastly, Phamerator shows the 3 similar phages I have been looking at have the gene, but none of them have assigned functions.

Feature 34 – reverse – stop 25884

Glimmer/GeneMark

What feature number is this? 34

What is the stop site? 25884

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Called by both, they disagree

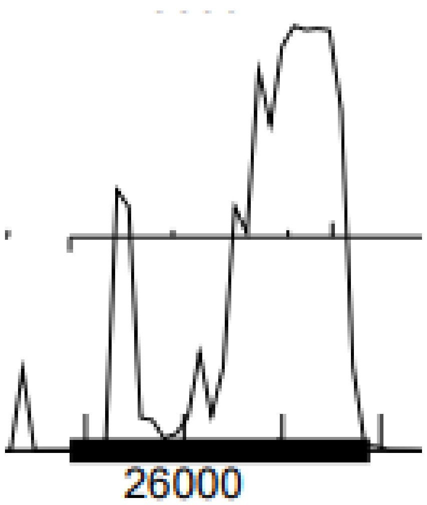
What is the autoannotated start?

Glimmer: 26153

GeneMark: 26189

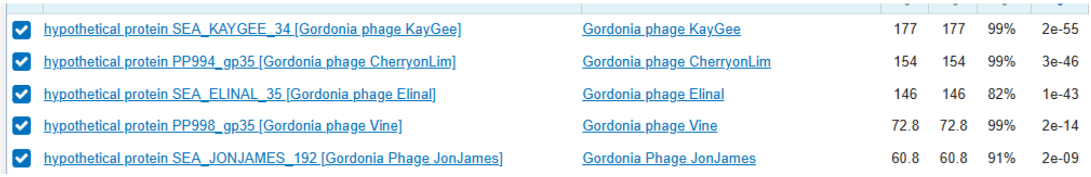
Gap: ____95____ or overlap: _______ (with gene in front of it) for the autoannotated start

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

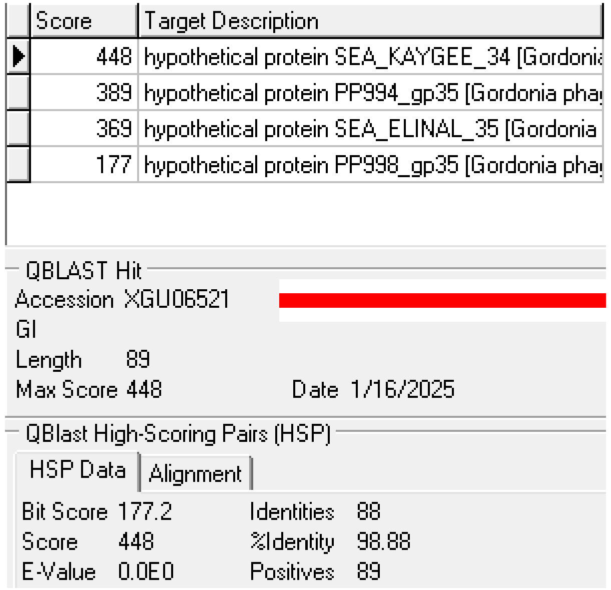


- Reading frame 5 contains a strong peak of coding potential that tapers off before peaking again.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



☑	hypothetical protein SEA_KAYGEE_34 [Gordonia phage KayGee]	Gordonia phage KayGee	177	177	99%	2e-55
☑	hypothetical protein PP994_gp35 [Gordonia phage CherryonLim]	Gordonia phage CherryonLim	154	154	99%	3e-46
☑	hypothetical protein SEA_ELINAL_35 [Gordonia phage Elinal]	Gordonia phage Elinal	146	146	82%	1e-43
☑	hypothetical protein PP998_gp35 [Gordonia phage Vine]	Gordonia phage Vine	72.8	72.8	99%	2e-14
☑	hypothetical protein SEA_JONJAMES_192 [Gordonia Phage JonJames]	Gordonia Phage JonJames	60.8	60.8	91%	2e-09



Score	Target Description
448	hypothetical protein SEA_KAYGEE_34 [Gordonia
389	hypothetical protein PP994_gp35 [Gordonia phag
369	hypothetical protein SEA_ELINAL_35 [Gordonia
177	hypothetical protein PP998_gp35 [Gordonia phag

QBLAST Hit

Accession XGU06521

GI

Length 89

Max Score 448 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	177.2	Identities	88
Score	448	%Identity	98.88
E-Value	0.0E0	Positives	89

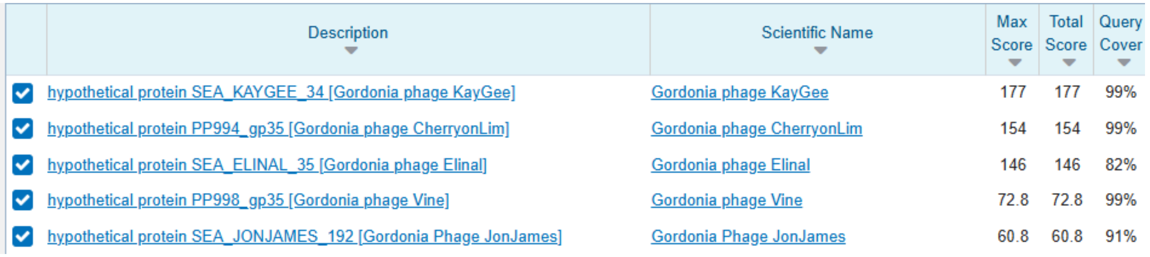
- Glimmer: all 4 BLAST hits had an E-value close to 0.
- GeneMark: all 5 BLAST hits had an E-value close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

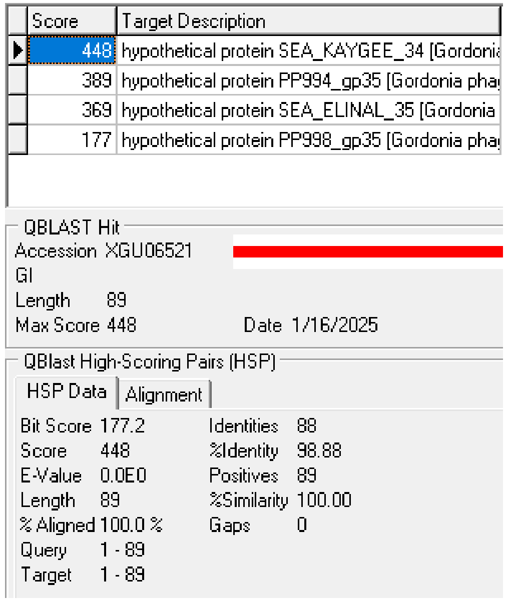
- This is a gene. It is called by both Glimmer and GeneMark, despite their disagreement on the start site, it has strong coding potential, and BLAST showed multiple hits with an E-value close to 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- **Glimmer: 3 1:1 alignments**
- **GeneMark: 5 1:1 alignments**



	Description	Scientific Name	Max Score	Total Score	Query Cover
☑	hypothetical protein SEA_KAYGEE_34 [Gordonia phage KayGee]	Gordonia phage KayGee	177	177	99%
☑	hypothetical protein PP994_gp35 [Gordonia phage CherryonLim]	Gordonia phage CherryonLim	154	154	99%
☑	hypothetical protein SEA_ELINAL_35 [Gordonia phage Elinal]	Gordonia phage Elinal	146	146	82%
☑	hypothetical protein PP998_gp35 [Gordonia phage Vine]	Gordonia phage Vine	72.8	72.8	99%
☑	hypothetical protein SEA_JONJAMES_192 [Gordonia Phage JonJames]	Gordonia Phage JonJames	60.8	60.8	91%



Score	Target Description
448	hypothetical protein SEA_KAYGEE_34 [Gordonia
389	hypothetical protein PP994_gp35 [Gordonia pha
369	hypothetical protein SEA_ELINAL_35 [Gordonia
177	hypothetical protein PP998_gp35 [Gordonia pha

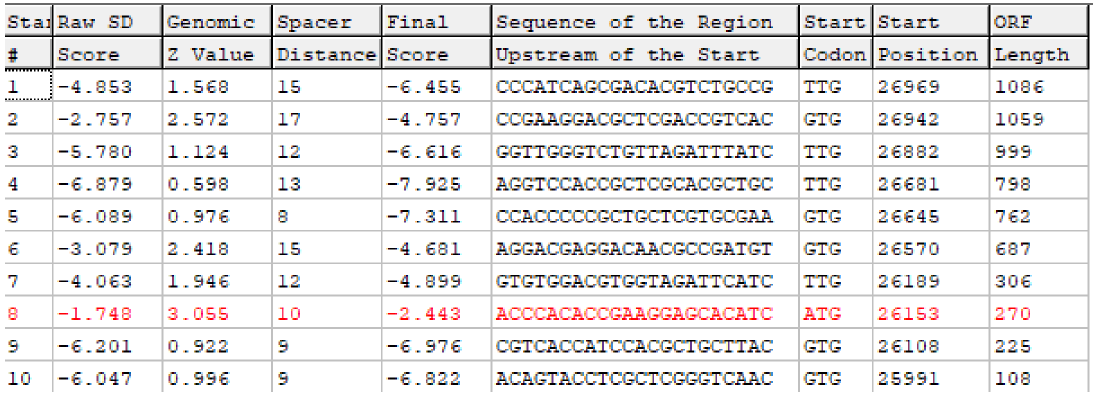
QBLAST Hit

Accession XGU06521
GI
Length 89
Max Score 448 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	177.2	Identities	88
Score	448	%Identity	98.88
E-Value	0.0E0	Positives	89
Length	89	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 89		
Target	1 - 89		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

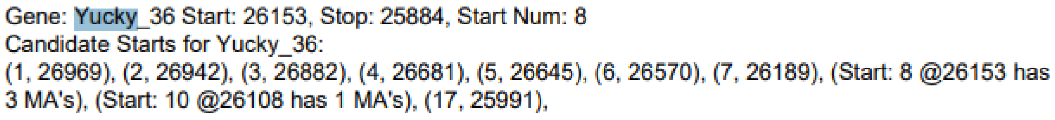


Sta	Raw SD	Genomic	Spacer	Final	Sequence of the Region	Start	Start	ORF
#	Score	Z Value	Distance	Score	Upstream of the Start	Codon	Position	Length
1	-4.853	1.568	15	-6.455	CCCATCAGCGACACGTCTGCCG	TTG	26969	1086
2	-2.757	2.572	17	-4.757	CCGAAGGACGCTCGACCGTCAC	GTG	26942	1059
3	-5.780	1.124	12	-6.616	GGTTGGGTCTGTTAGATTTATC	TTG	26882	999
4	-6.879	0.598	13	-7.925	AGGTCCACCGCTCGCACGCTGC	TTG	26681	798
5	-6.089	0.976	8	-7.311	CCACCCCCGCTGCTCGTGCGAA	GTG	26645	762
6	-3.079	2.418	15	-4.681	AGGACGAGGACAACGCCGATGT	GTG	26570	687
7	-4.063	1.946	12	-4.899	GTGTGGACGTGGTAGATTCATC	TTG	26189	306
8	-1.748	3.055	10	-2.443	ACCCACACCGAAGGAGCACATC	ATG	26153	270
9	-6.201	0.922	9	-6.976	CGTCACCATCCACGCTGCTTAC	GTG	26108	225
10	-6.047	0.996	9	-6.822	ACAGTACCTCGCTCGGGTCAAC	GTG	25991	108

- Glimmer: Z-value: 3.055, Final score: -2.443
- GeneMark: Z-value: 1.946, Final score: -4.899
- Glimmer site is stronger

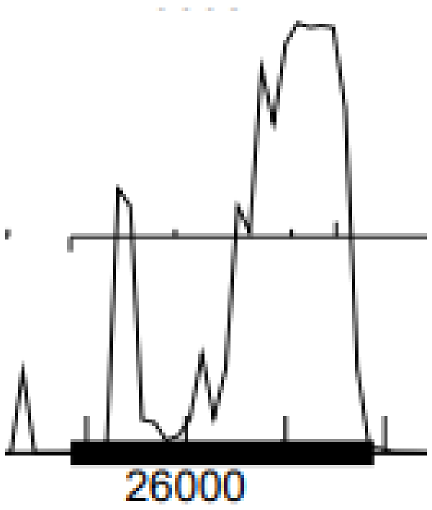
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Glimmer: 3 MA's, most of any start
- GeneMark: Never been manually annotated.



Gene: Yucky_36 Start: 26153, Stop: 25884, Start Num: 8
Candidate Starts for Yucky_36:
(1, 26969), (2, 26942), (3, 26882), (4, 26681), (5, 26645), (6, 26570), (7, 26189), (Start: 8 @26153 has 3 MA's), (Start: 10 @26108 has 1 MA's), (17, 25991),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- Glimmer: cuts off some coding potential
- GeneMark: includes all coding potential

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

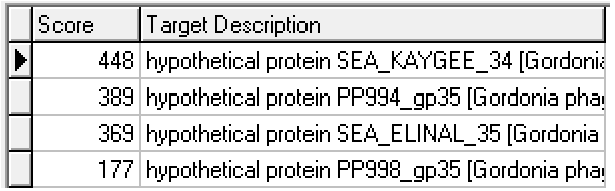
- Glimmer: 26342-26153 for gap= 188
- GeneMark: 26342-26189= 153-1 for gap= 152

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

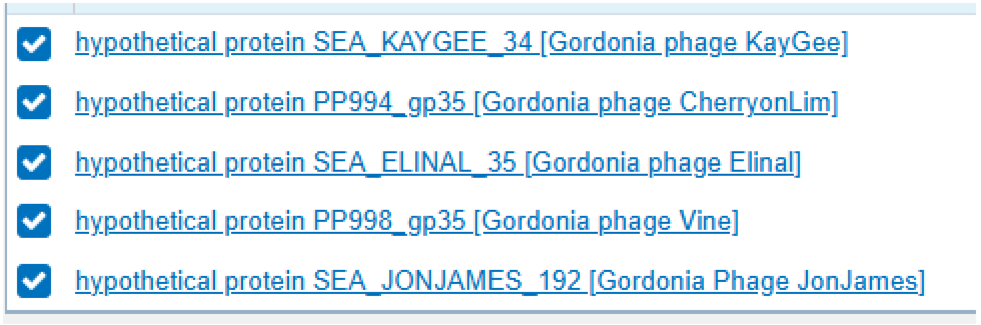
	26153	26189
BLAST	3 1:1 alignments	5 1:1 alignments
RBS	Z-value: 3.055, Final score: -2.443	Z-value: 1.946, Final score: -4.899
Starterator	3 MA's	0 MA's
Coding potential	Cuts off slight piece	Includes all
Gap/overlap	188	152

Despite there being one more piece of evidence in favor of 26189, I feel as though I can't call it that due to how bad the RBS numbers are and that it has never been manually annotated. Because of this I believe the start site to be 26153.

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
448	hypothetical protein SEA_KAYGEE_34 [Gordoni
389	hypothetical protein PP994_gp35 [Gordonia pha
369	hypothetical protein SEA_ELINAL_35 [Gordonia
177	hypothetical protein PP998_gp35 [Gordonia pha



☑ hypothetical protein SEA_KAYGEE_34 [Gordonia phage KayGee]

☑ hypothetical protein PP994_gp35 [Gordonia phage CherryonLim]

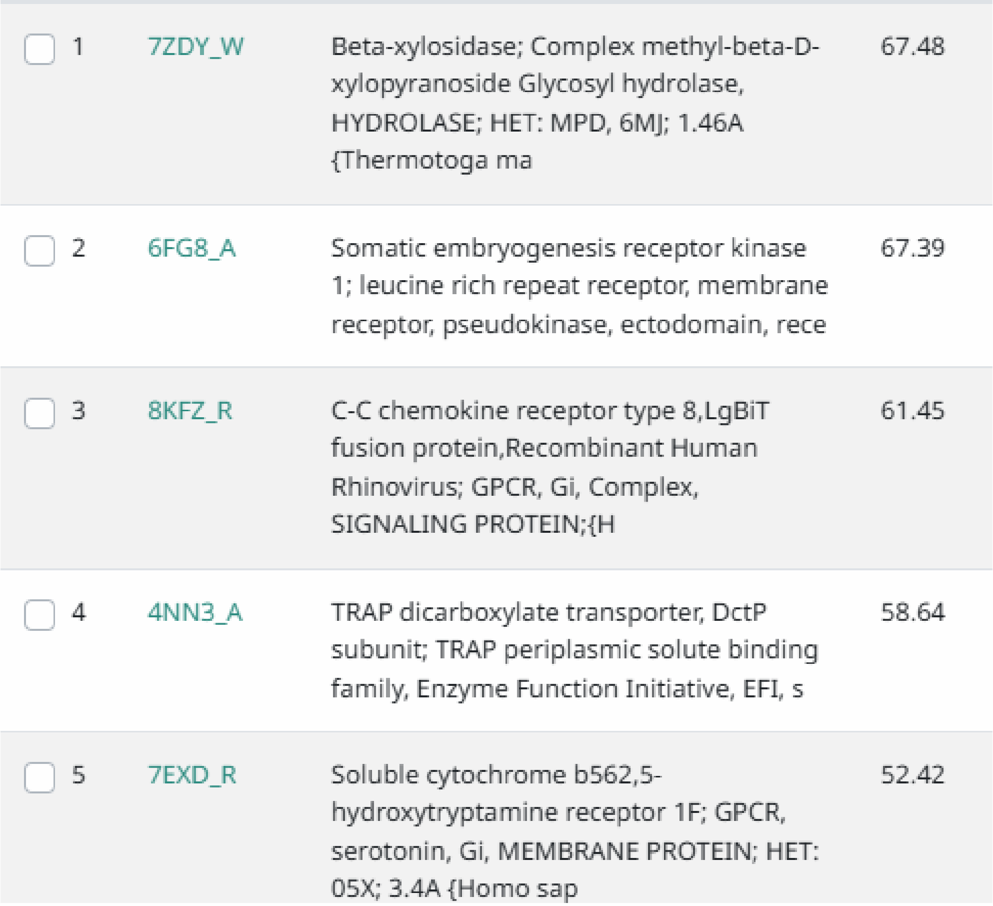
☑ hypothetical protein SEA_ELINAL_35 [Gordonia phage Elinal]

☑ hypothetical protein PP998_gp35 [Gordonia phage Vine]

☑ hypothetical protein SEA_JONJAMES_192 [Gordonia Phage JonJames]

- All 4 BLAST hits have a hypothetical protein function.
- NCBI BLAST shows 5 hits with a hypothetical protein function.

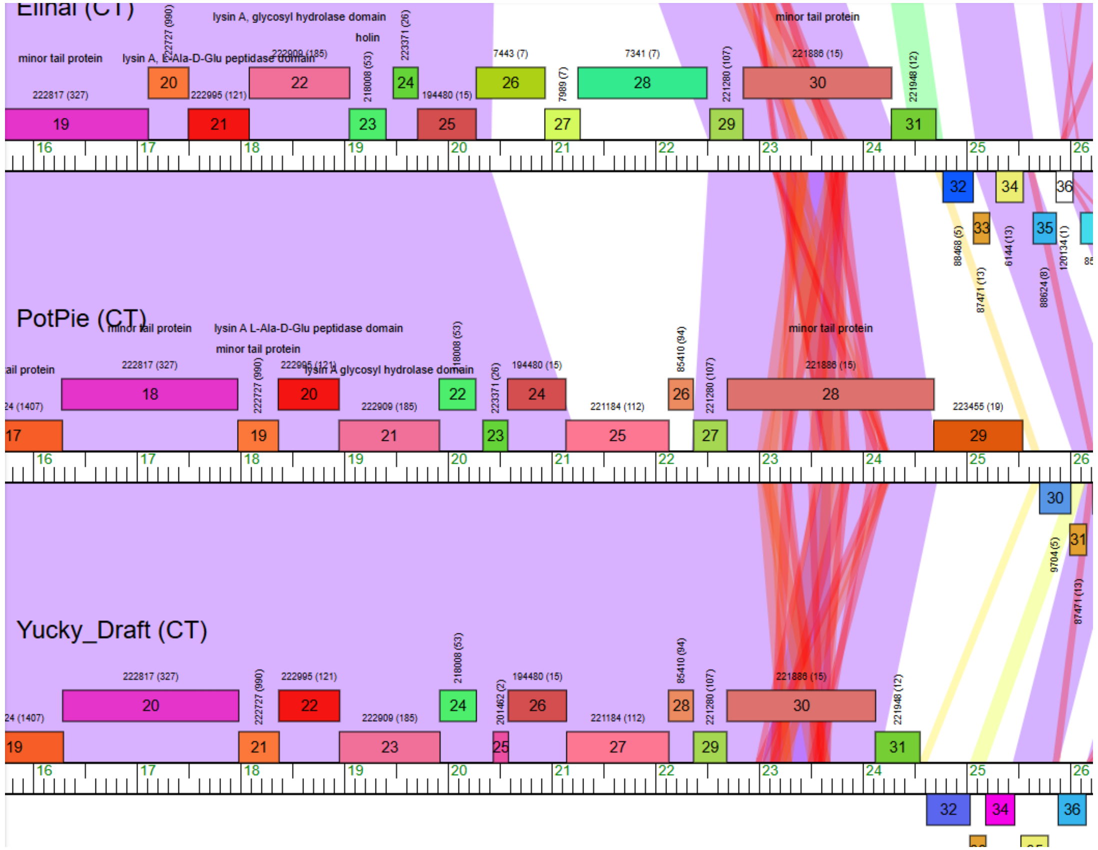
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



□	1	7ZDY_W	Beta-xylosidase; Complex methyl-beta-D-xylopyranoside Glycosyl hydrolase, HYDROLASE; HET: MPD, 6MJ; 1.46A {Thermotoga ma	67.48
□	2	6FG8_A	Somatic embryogenesis receptor kinase 1; leucine rich repeat receptor, membrane receptor, pseudokinase, ectodomain, rece	67.39
□	3	8KFZ_R	C-C chemokine receptor type 8,LgBiT fusion protein,Recombinant Human Rhinovirus; GPCR, Gi, Complex, SIGNALING PROTEIN;{H	61.45
□	4	4NN3_A	TRAP dicarboxylate transporter, DctP subunit; TRAP periplasmic solute binding family, Enzyme Function Initiative, EFI, s	58.64
□	5	7EXD_R	Soluble cytochrome b562,5-hydroxytryptamine receptor 1F; GPCR, serotonin, Gi, MEMBRANE PROTEIN; HET: 05X; 3.4A {Homo sap	52.42

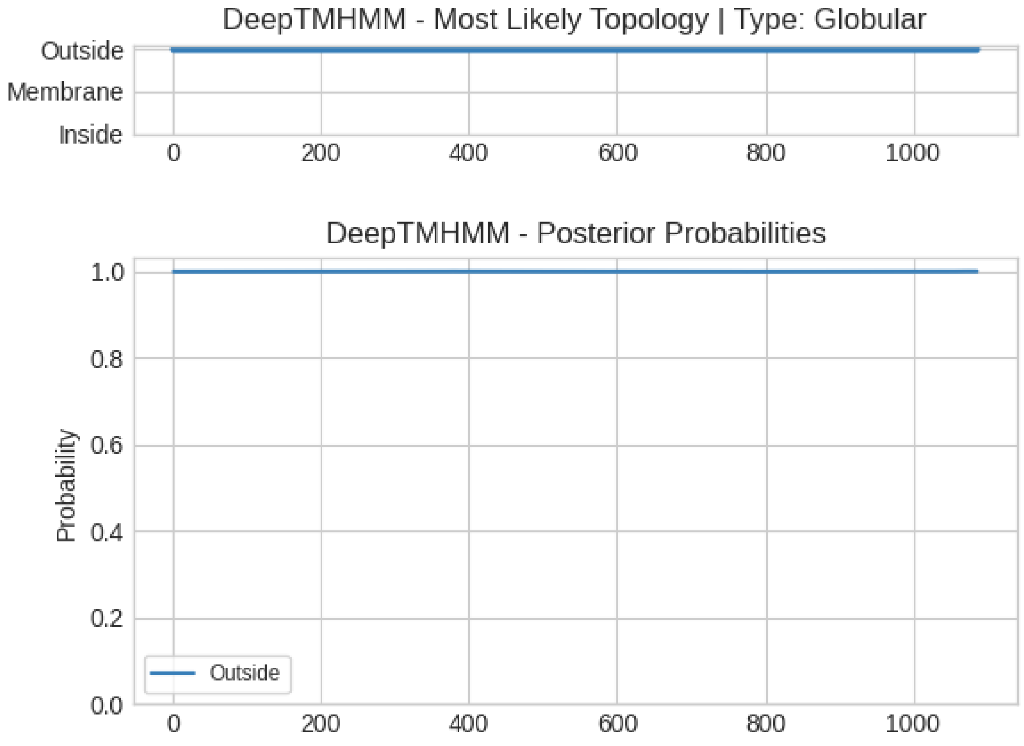
- There are no Hhpred hits with a probability above 90.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- Only Elinal has this gene and it is a hypothetical protein. No conserved domains.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- This is not an intermembrane protein and it likely functions outside of the membrane.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I am assigning this as a hypothetical protein. This is consistent with the BLAST hits on both DNA Master and NCBI. Hhpred did not have any viable hits. Phamerator showed another similar phage had this gene as a hypothetical protein. Lastly, It was determined that this was not an intermembrane protein.

Feature 35 – Reverse – Stop 26105

Glimmer/GeneMark

What feature number is this?

What is the stop site?

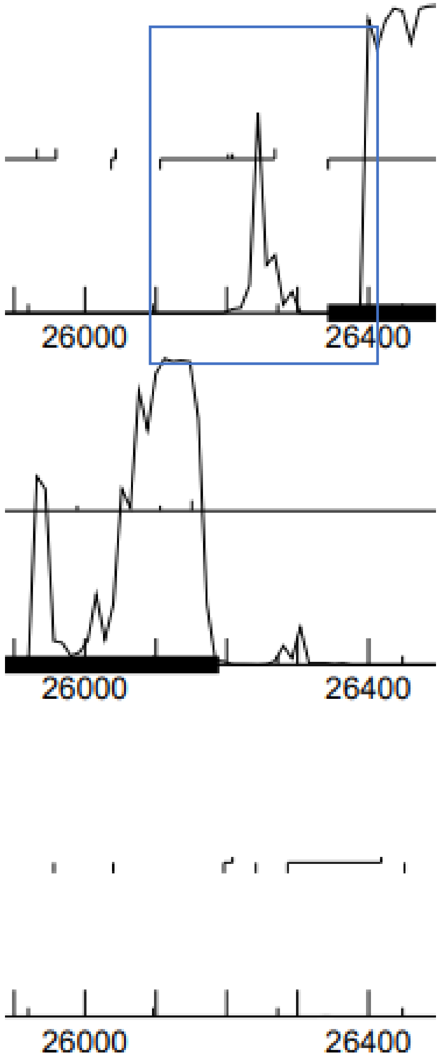
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

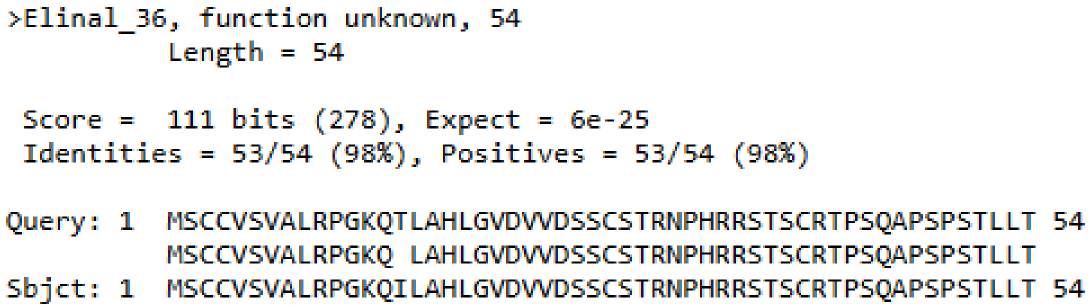
- 36_37 REVERSE
- Stop Site: 26105
- Start Site: 26269
- No Auto-annotated Start
- 72 bp gap with downstream feature 37
- 49 bp overlap with upstream feature 36 (assuming the start is the auto-annotated start of 26153)

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

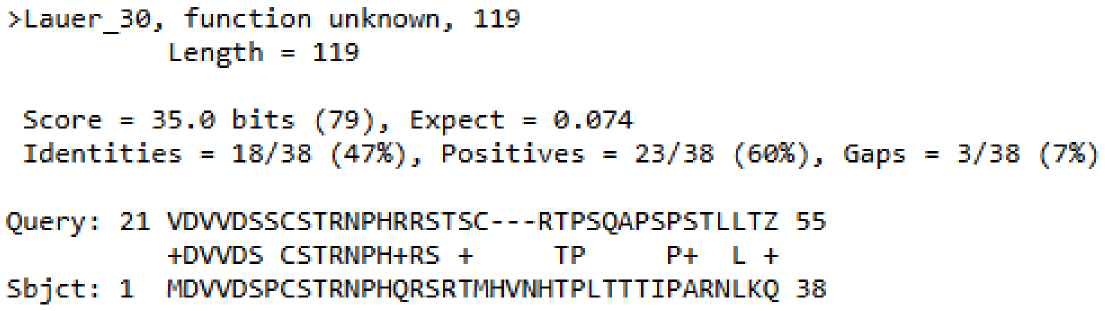


- There is some weak CP in the frame below that was not included in the auto-annotation of feature 36
- Weak to Moderate CP
- The CP only briefly spikes above the middle line.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



```
>Elinal_36, function unknown, 54
          Length = 54

 Score =  111 bits (278), Expect = 6e-25
 Identities = 53/54 (98%), Positives = 53/54 (98%)

Query: 1  MSCCVSVALRPGKQTLAHLGVDVVDSSCSTRNPHRRSTSCRTPSQAPSPSTLLT 54
          MSCCVSVALRPGKQ LAHLGVDVVDSSCSTRNPHRRSTSCRTPSQAPSPSTLLT
Sbjct: 1  MSCCVSVALRPGKQILAHLGVDVVDSSCSTRNPHRRSTSCRTPSQAPSPSTLLT 54
```



```
>Lauer_30, function unknown, 119
          Length = 119

 Score = 35.0 bits (79), Expect = 0.074
 Identities = 18/38 (47%), Positives = 23/38 (60%), Gaps = 3/38 (7%)

Query: 21 VDVVDSSCSTRNPHRRSTSC---RTPSQAPSPSTLLTZ 55
          +DVVDS CSTRNPH+RS +     TP     P+  L +
Sbjct: 1  MDVVDSPCSTRNPHQRSRTMHVNHTPLTTTIPARNLKQ 38
```

- Elinal and Lauer are both CT cluster phage

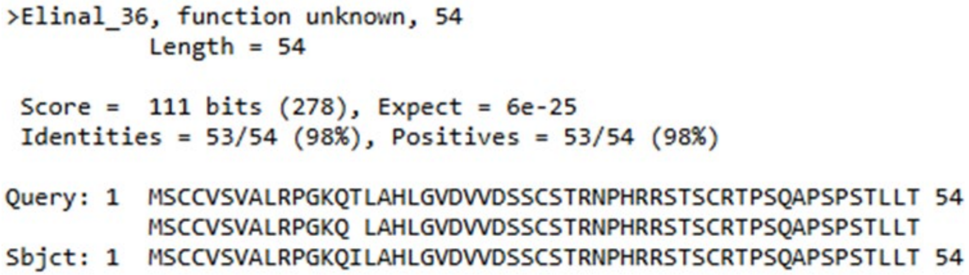
- Elinal_36 is an orpham

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes!

- It has a BLAST hit to another CT cluster phage
- Has moderate CP

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

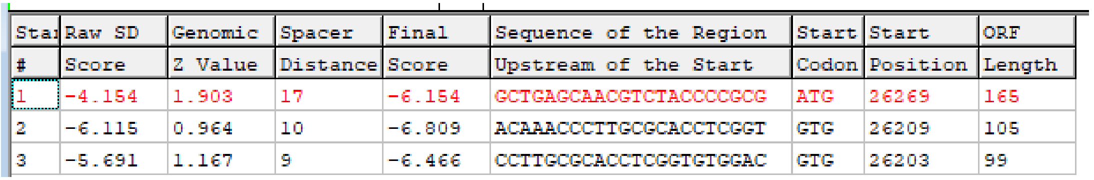


```
>Elinal_36, function unknown, 54
          Length = 54

 Score =  111 bits (278), Expect = 6e-25
 Identities = 53/54 (98%), Positives = 53/54 (98%)

Query: 1  MSCCVSVALRPGKQTLAHLGVDVVDSSCSTRNPHRRSTSCRTPSQAPSPSTLLT 54
           MSCCVSVALRPGKQ LAHLGVDVVDSSCSTRNPHRRSTSCRTPSQAPSPSTLLT
Sbjct: 1  MSCCVSVALRPGKQILAHLGVDVVDSSCSTRNPHRRSTSCRTPSQAPSPSTLLT 54
```

- 26269 start is Q1:S1 with Elinal_36 which is an orpham. Elinal is a CT cluster phage.

- Only this 1 Q1:S1 alignment hit for 26269 start
- Q1:S1 hit with Lauer for start 26209

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-4.154	1.903	17	-6.154	GCTGAGCAACGTCTACCCCGCG	ATG	26269	165
2	-6.115	0.964	10	-6.809	ACAAACCCTTGCGCACCTCGGT	GTG	26209	105
3	-5.691	1.167	9	-6.466	CCTTGCGCACCTCGGTGTGGAC	GTG	26203	99

Start 26269 has the slightly better RBS scores

Start 26269

- Z-Value: 1.903
- Final Score: -6.164

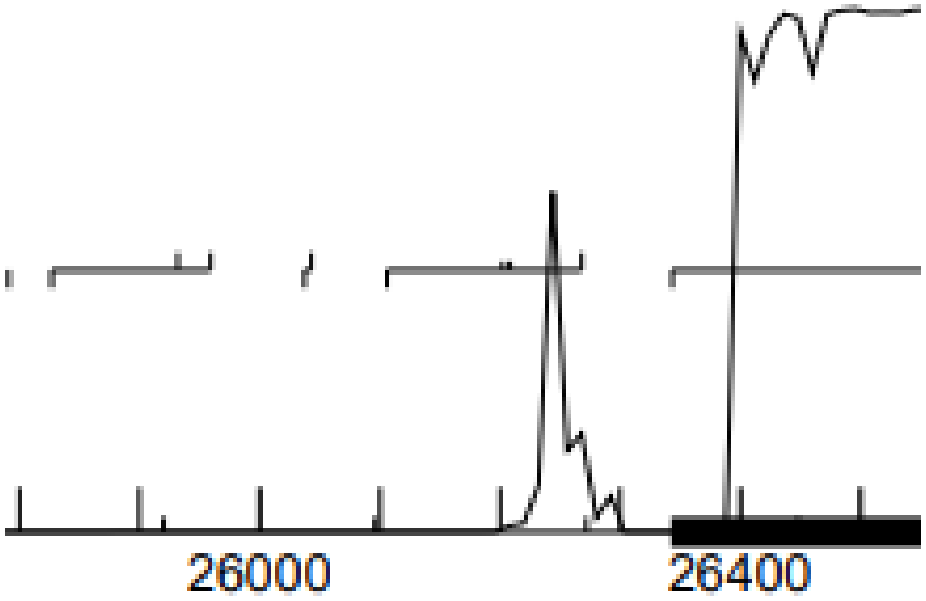
Start 26209

- Z-Value: 0.964
- Final Score: -6.466

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

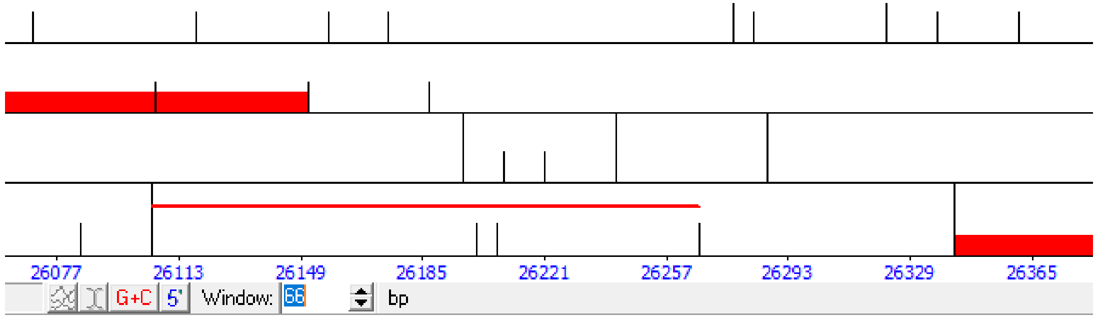
- No Starterator Evidence

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- 26269 ~30 bp of weak CP cut off
- 26209 No CP included
- 26203 No CP included

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.



- 26269 72 bp gap
- 26209 132 bp gap
- 26203 138 bp gap

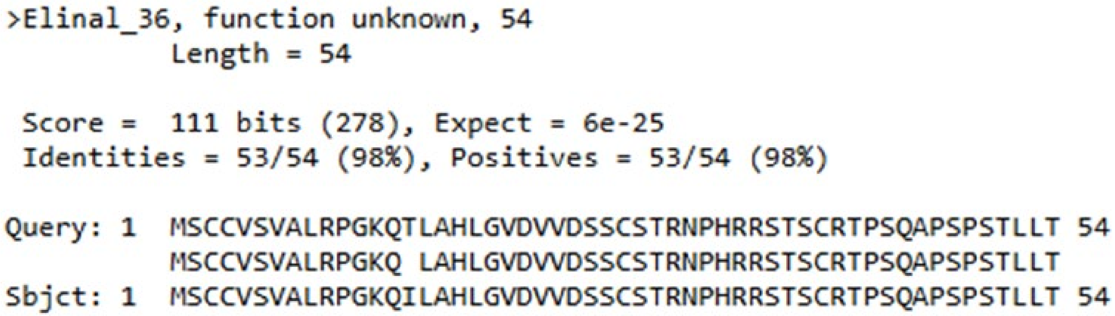
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- 26269 Start

- There was not an auto-annotated start for this feature
- This start includes most CP
- Has the better RBS scores
- Has Q1:S1 BLAST hit to CT cluster phage Elinal

BLAST function evidence. What assigned functions do other highly similar genes have?

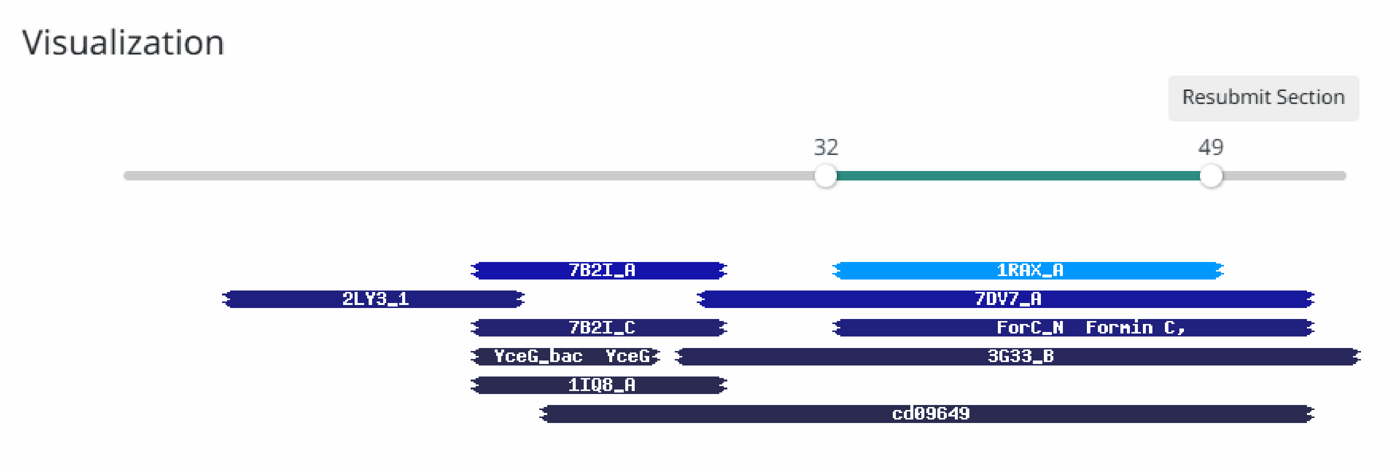
- Hypothetical Protein



```
>Elinal_36, function unknown, 54
          Length = 54

 Score =  111 bits (278), Expect = 6e-25
 Identities = 53/54 (98%), Positives = 53/54 (98%)

Query: 1  MSCCVSVALRPGKQTLAHLGVDVVDSSCSTRNPHRRSTSCRTPSQAPSPSTLLT 54
           MSCCVSVALRPGKQ LAHLGVDVVDSSCSTRNPHRRSTSCRTPSQAPSPSTLLT
Sbjct: 1  MSCCVSVALRPGKQILAHLGVDVVDSSCSTRNPHRRSTSCRTPSQAPSPSTLLT 54
```

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

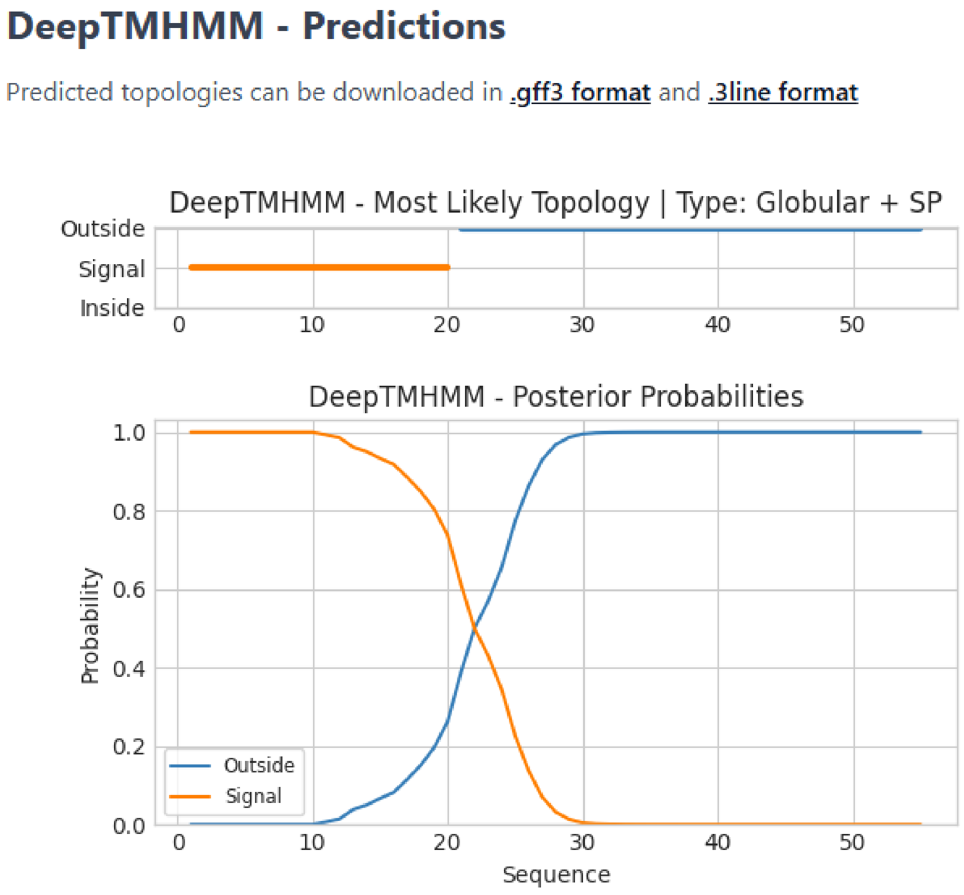
- NKF, there are no hits with a probability >90%



Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- No Phamerator Evidence

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

- No predicted TMRs

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Hypothetical Protein

- All BLAST hits were hypothetical proteins
- There were no HHPred hits with a probability >90%
- Deep TMHMM did not predict any TMRs

Feature 36 – reverse – stop 26342

Glimmer/GeneMark

What feature number is this? 36

What is the stop site? 26342

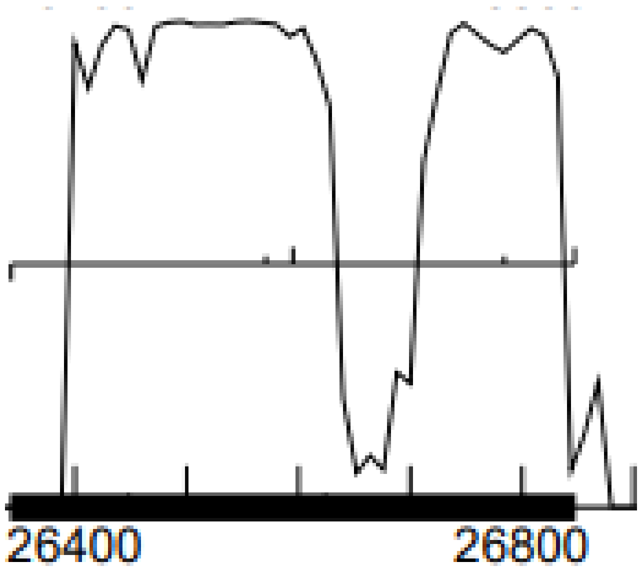
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Called by both

What is the autoannotated start? 26848

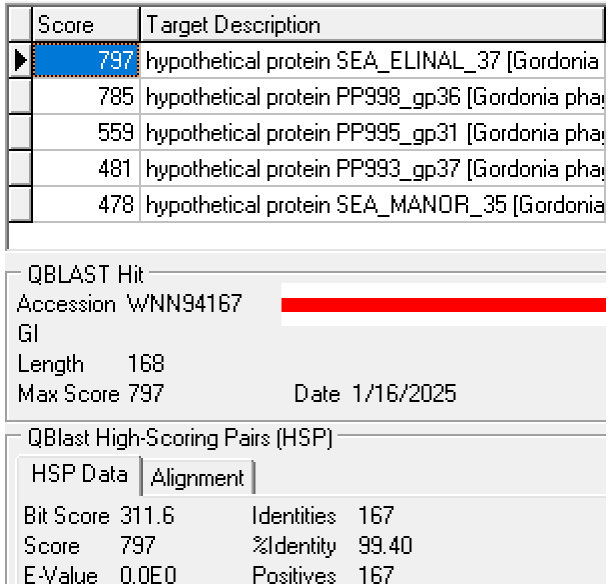
Gap: ____75____ or overlap: _______ (with gene in front of it) for the autoannotated start

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Reading frame 4 contains 2 strong peaks of coding potential separated by a weak peak.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
797	hypothetical protein SEA_ELINAL_37 [Gordonia
785	hypothetical protein PP998_gp36 [Gordonia pha
559	hypothetical protein PP995_gp31 [Gordonia pha
481	hypothetical protein PP993_gp37 [Gordonia pha
478	hypothetical protein SEA_MANOR_35 [Gordonia

QBLAST Hit
Accession WNN94167
GI
Length 168
Max Score 797
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	311.6	Identities	167
Score	797	%Identity	99.40
E-Value	0.0E0	Positives	167

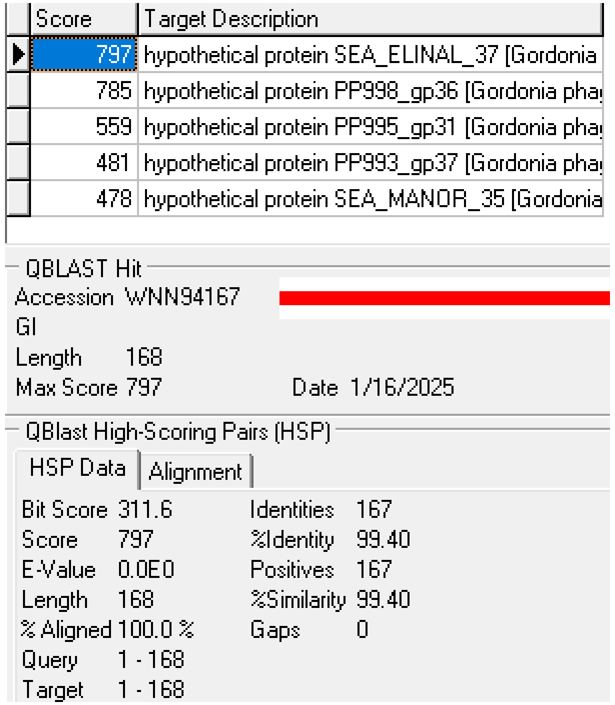
- All 25 hits have an E-value close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- I believe this to be a gene. It was called by both glimmer and GeneMark, has strong coding potential, and has at ;east 25 hits with an E-value close to 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- 7 1:1 alignments



Score	Target Description
797	hypothetical protein SEA_ELINAL_37 [Gordonia
785	hypothetical protein PP998_gp36 [Gordonia pha
559	hypothetical protein PP995_gp31 [Gordonia pha
481	hypothetical protein PP993_gp37 [Gordonia pha
478	hypothetical protein SEA_MANOR_35 [Gordonia

QBLAST Hit

Accession WNN94167

GI

Length 168

Max Score 797

Date 1/16/2025

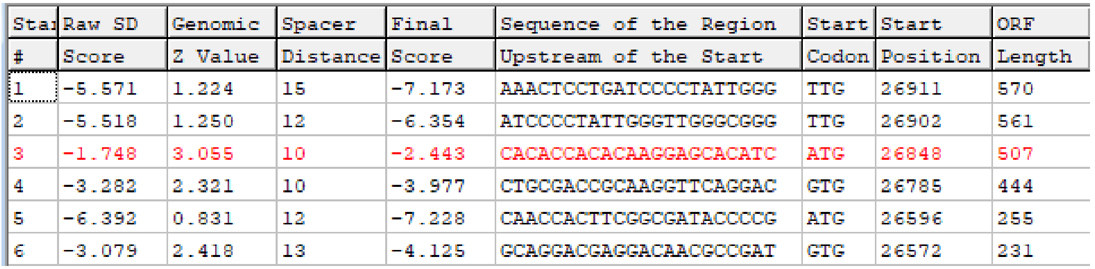
QBlast High-Scoring Pairs (HSP)

HSP Data | Alignment

Bit Score	311.6	Identities	167
Score	797	%Identity	99.40
E-Value	0.0E0	Positives	167
Length	168	%Similarity	99.40
% Aligned	100.0 %	Gaps	0
Query	1 - 168		
Target	1 - 168		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

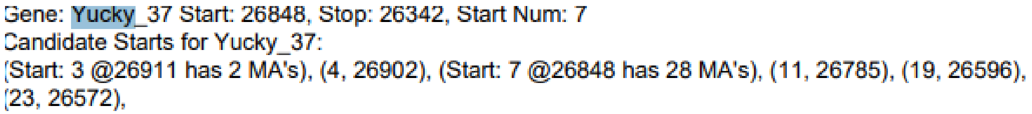
- Z-value: 3.055, Final score: -2.443



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	-5.571	1.224	15	-7.173	AAACTCCTGATCCCCTATTGGG	TTG	26911	570
2	-5.518	1.250	12	-6.354	ATCCCCTATTGGGTTGGGCGGG	TTG	26902	561
3	-1.748	3.055	10	-2.443	CACACCACACAAGGAGCACATC	ATG	26848	507
4	-3.282	2.321	10	-3.977	CTGCGACCGCAAGGTTCAGGAC	GTG	26785	444
5	-6.392	0.831	12	-7.228	CAACCACTTCGGCGATACCCCG	ATG	26596	255
6	-3.079	2.418	13	-4.125	GCAGGACGAGGACAACGCCGAT	GTG	26572	231

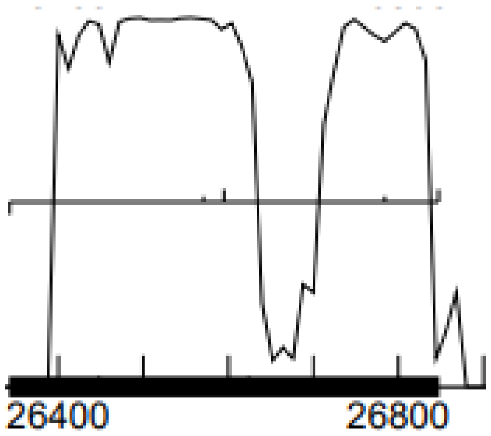
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- 28 MA's, most of any site. Another site has 3 but no other evidence points towards it so I am disregarding it.



```
Gene: Yucky_37 Start: 26848, Stop: 26342, Start Num: 7
Candidate Starts for Yucky_37:
(Start: 3 @26911 has 2 MA's), (4, 26902), (Start: 7 @26848 has 28 MA's), (11, 26785), (19, 26596),
(23, 26572),
```

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- This start cuts off a slight piece of coding potential.

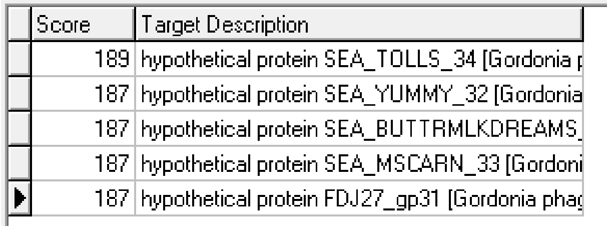
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 26924-26848= 76-1 for gap=75

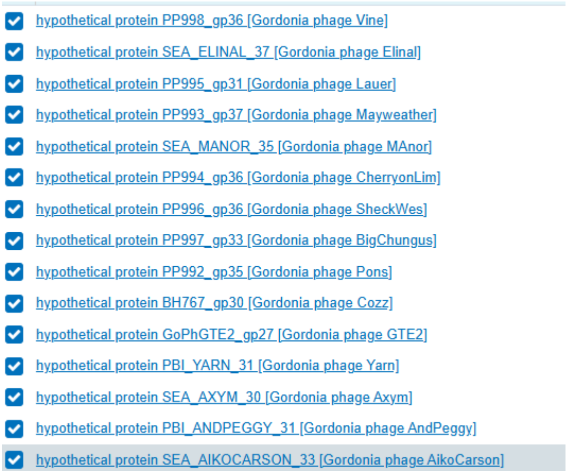
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is the automated start 26848. It has 7 1:1 alignments, the best RBS numbers, the most MA’s by a lot, only cuts off a little coding potential, and has a large, but not unacceptable gap.

BLAST function evidence. What assigned functions do other highly similar genes have?



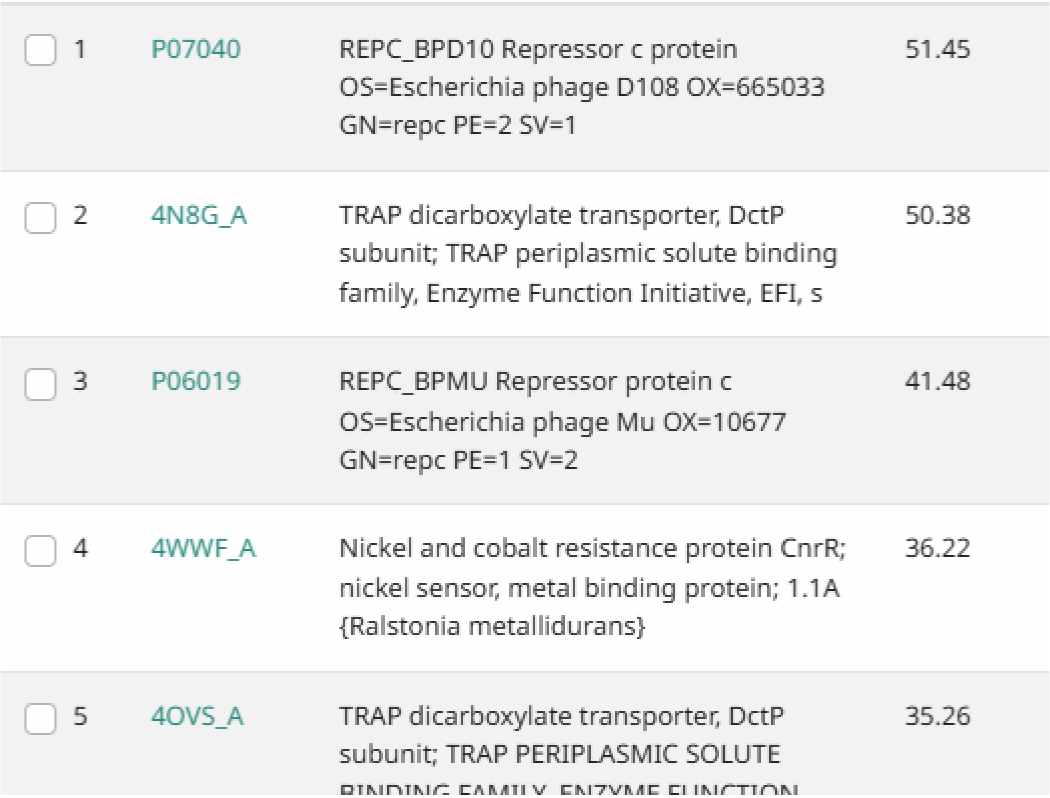
	Score	Target Description
	189	hypothetical protein SEA_TOLLS_34 [Gordonia p
	187	hypothetical protein SEA_YUMMY_32 [Gordonia
	187	hypothetical protein SEA_BUTTRMLKDREAMS_
	187	hypothetical protein SEA_MSCARN_33 [Gordoni
▶	187	hypothetical protein FDJ27_gp31 [Gordonia phag



- ☑ hypothetical protein PP998_gp36 [Gordonia phage Vine]
- ☑ hypothetical protein SEA_ELINAL_37 [Gordonia phage Elinal]
- ☑ hypothetical protein PP995_gp31 [Gordonia phage Lauer]
- ☑ hypothetical protein PP993_gp37 [Gordonia phage Mayweather]
- ☑ hypothetical protein SEA_MANOR_35 [Gordonia phage MAnor]
- ☑ hypothetical protein PP994_gp36 [Gordonia phage CherryonLim]
- ☑ hypothetical protein PP996_gp36 [Gordonia phage SheckWes]
- ☑ hypothetical protein PP997_gp33 [Gordonia phage BigChungus]
- ☑ hypothetical protein PP992_gp35 [Gordonia phage Pons]
- ☑ hypothetical protein BH767_gp30 [Gordonia phage Cozz]
- ☑ hypothetical protein GoPhGTE2_gp27 [Gordonia phage GTE2]
- ☑ hypothetical protein PBI_YARN_31 [Gordonia phage Yarn]
- ☑ hypothetical protein SEA_AXYM_30 [Gordonia phage Axym]
- ☑ hypothetical protein PBI_ANDPEGGY_31 [Gordonia phage AndPeggy]
- ☑ hypothetical protein SEA_AIKOCARSON_33 [Gordonia phage AikoCarson]

- All 25 hits are hypothetical proteins.
- All NCBI hits are also hypothetical proteins.

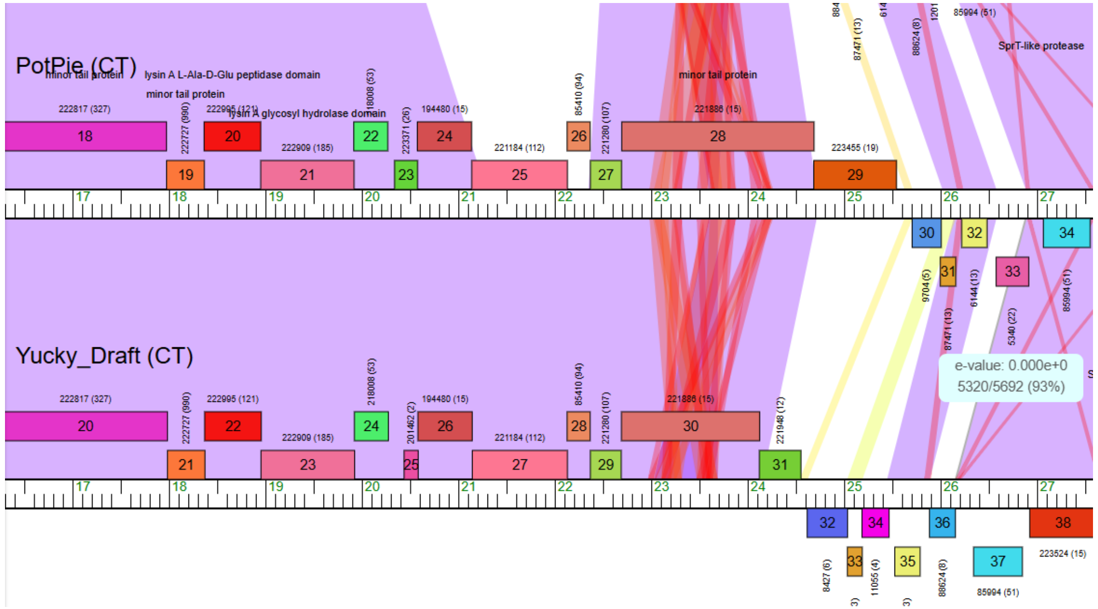
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



□	1	P07040	REPC_BPD10 Repressor c protein OS=Escherichia phage D108 OX=665033 GN=repc PE=2 SV=1	51.45
□	2	4N8G_A	TRAP dicarboxylate transporter, DctP subunit; TRAP periplasmic solute binding family, Enzyme Function Initiative, EFI, s	50.38
□	3	P06019	REPC_BPMU Repressor protein c OS=Escherichia phage Mu OX=10677 GN=repc PE=1 SV=2	41.48
□	4	4WWF_A	Nickel and cobalt resistance protein CnrR; nickel sensor, metal binding protein; 1.1A {Ralstonia metallidurans}	36.22
□	5	4OVS_A	TRAP dicarboxylate transporter, DctP subunit; TRAP PERIPLASMIC SOLUTE BINDING FAMILY, ENZYME FUNCTION	35.26

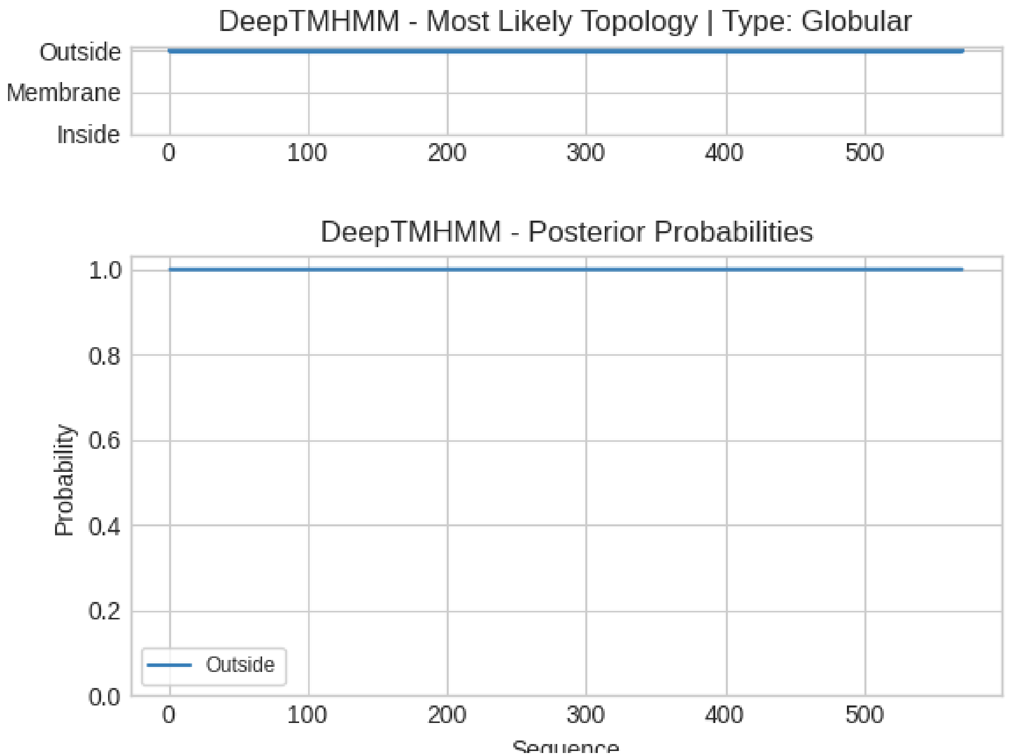
- There are no Hhpred hits with a probability of 90+.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- Elinal, PotPie, and BigChungus all contain this gene and it is labeled a hypothetical protein.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- This is not an intermembrane protein and it function outside of the membrane.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I am assigning this a hypothetical protein. All BLAST hits showed this as the function, Hhpred didn’t have any viable results, phamerator showed that all similar phages I have been looking at had this gene listed as a hypothetical protein. Lastly, it was determined that it was not an intermembrane protein.

Feature 37 – reverse – stop 26924

Glimmer/GeneMark

What feature number is this?

- 37

What is the stop site?

- 26924

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

- Both

What is the autoannotated start?

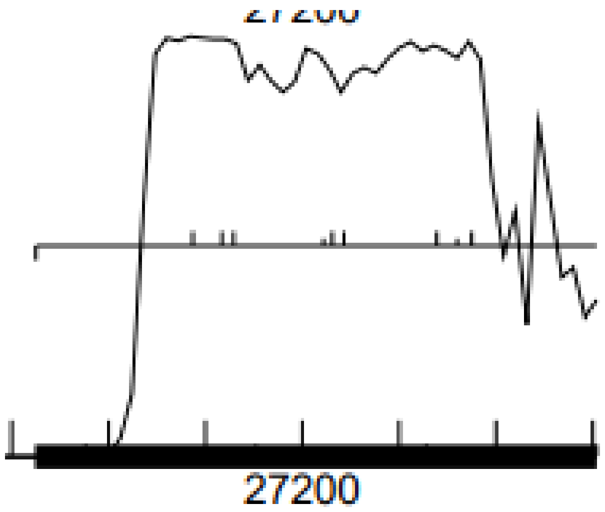
- 27610

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

- There is a gap of 218 nucleotides

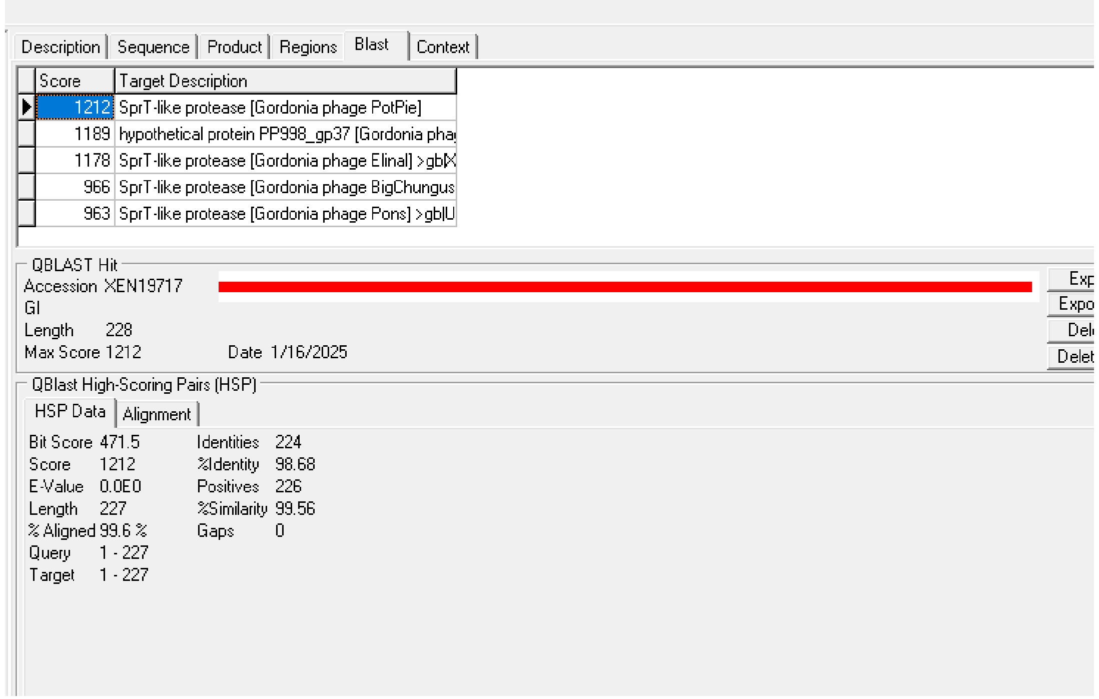
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- There is very strong coding potential for this graph with it having a very strong peak that continues for several hundred nucleotides



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are 13 blast hits that have an e-value of zero



Score	Target Description
1212	SprT-like protease [Gordonia phage PotPie]
1189	hypothetical protein PP998_gp37 [Gordonia pha
1178	SprT-like protease [Gordonia phage Elinal] >gb|X
966	SprT-like protease [Gordonia phage BigChungus
963	SprT-like protease [Gordonia phage Pons] >gb|U

QBLAST Hit
Accession XEN19717
GI
Length 228
Max Score 1212
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	471.5	Identities	224
Score	1212	%Identity	98.68
E-Value	0.0E0	Positives	226
Length	227	%Similarity	99.56
% Aligned	99.6 %	Gaps	0
Query	1 - 227		
Target	1 - 227		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

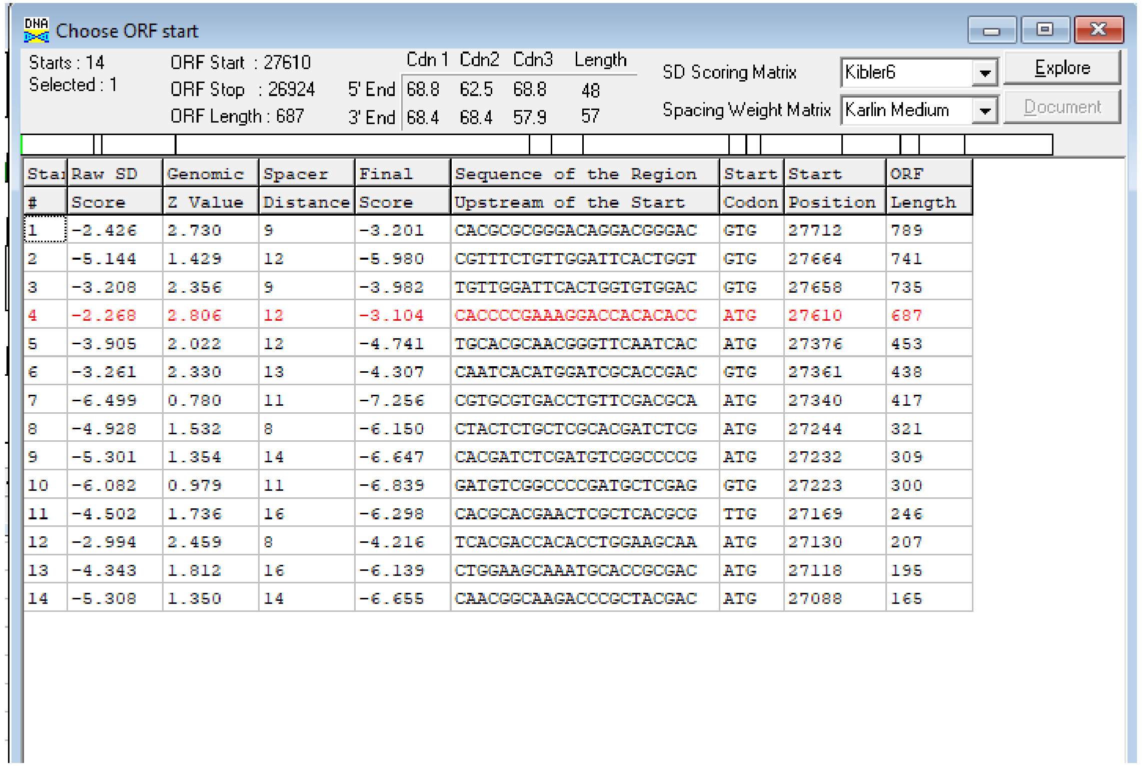
- Yes this feature is a gene because it was called by both genemark and glimmer, it has very strong coding potential, and it has 12 blast hits that have an e-value of zero.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There are 8 1:1 blast hits with other phages like PotPie and Elinal for start 27610
- There is 1 1:1 blast hit with the phage Vine for start 27712

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Start of 27610 has
- Z-value:2.806
- FS:-3.104
- These are the best rbs scores of the ones proposed on the page
- Start 27712
- Z-value:2.730
- FS:-3.201



Starts : 14
Selected : 1

ORF Start : 27610
ORF Stop : 26924
ORF Length : 687

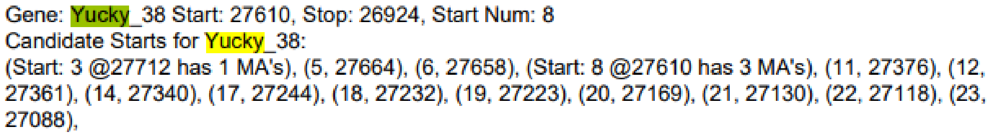
	Cdn 1	Cdn2	Cdn3	Length
5' End	68.8	62.5	68.8	48
3' End	68.4	68.4	57.9	57

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

Star	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	-2.426	2.730	9	-3.201	CACGCGCGGGACAGGACGGGAC	GTG	27712	789
2	-5.144	1.429	12	-5.980	CGTTTCTGTTGGATTCACTGGT	GTG	27664	741
3	-3.208	2.356	9	-3.982	TGTTGGATTCACTGGTGTGGAC	GTG	27658	735
4	-2.268	2.806	12	-3.104	CACCCCGAAAGGACCACACACC	ATG	27610	687
5	-3.905	2.022	12	-4.741	TGCACGCAACGGGTTCAATCAC	ATG	27376	453
6	-3.261	2.330	13	-4.307	CAATCACATGGATCGCACCGAC	GTG	27361	438
7	-6.499	0.780	11	-7.256	CGTGCGTGACCTGTTCGACGCA	ATG	27340	417
8	-4.928	1.532	8	-6.150	CTACTCTGCTCGCACGATCTCG	ATG	27244	321
9	-5.301	1.354	14	-6.647	CACGATCTCGATGTCGGCCCCG	ATG	27232	309
10	-6.082	0.979	11	-6.839	GATGTCGGCCCCGATGCTCGAG	GTG	27223	300
11	-4.502	1.736	16	-6.298	CACGCACGAACTCGCTCACGCG	TTG	27169	246
12	-2.994	2.459	8	-4.216	TCACGACCACACCTGGAAGCAA	ATG	27130	207
13	-4.343	1.812	16	-6.139	CTGGAAGCAAATGCACCGCGAC	ATG	27118	195
14	-5.308	1.350	14	-6.655	CAACGGCAAGACCCGCTACGAC	ATG	27088	165

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

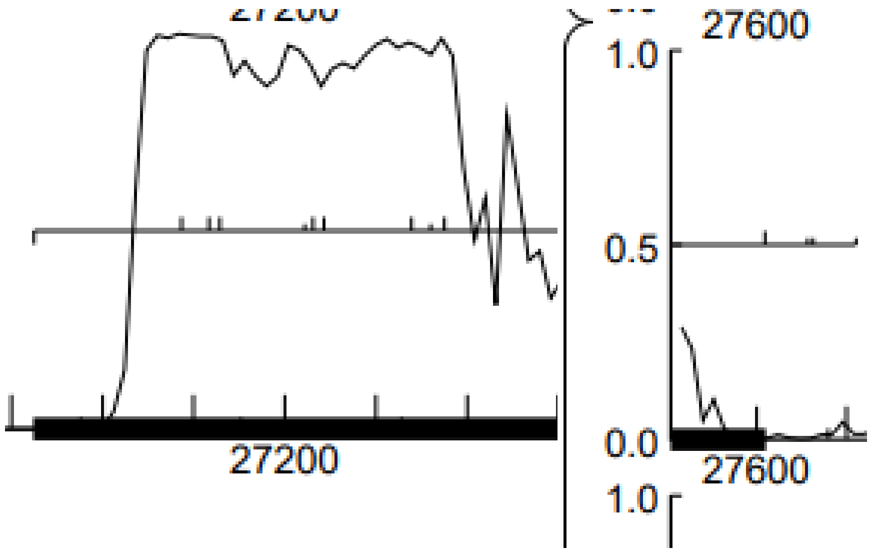
- Start of 27610 has 3 MA's and start 27712 has 1 MA making the start of 27610 the best



Gene: Yucky_38 Start: 27610, Stop: 26924, Start Num: 8
Candidate Starts for Yucky_38:
(Start: 3 @27712 has 1 MA's), (5, 27664), (6, 27658), (Start: 8 @27610 has 3 MA's), (11, 27376), (12, 27361), (14, 27340), (17, 27244), (18, 27232), (19, 27223), (20, 27169), (21, 27130), (22, 27118), (23, 27088),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- The start site of 27610 includes all of the coding potential in the feature
- The start of 27712 includes all the coding potential also but it makes for a very long area before the feature having no coding potential at all.



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- There is a gap of 218 nucleotides which isn’t great for 27610
- There is a gap of 116 nucleotides for the start of 27712 making this start the better choice here

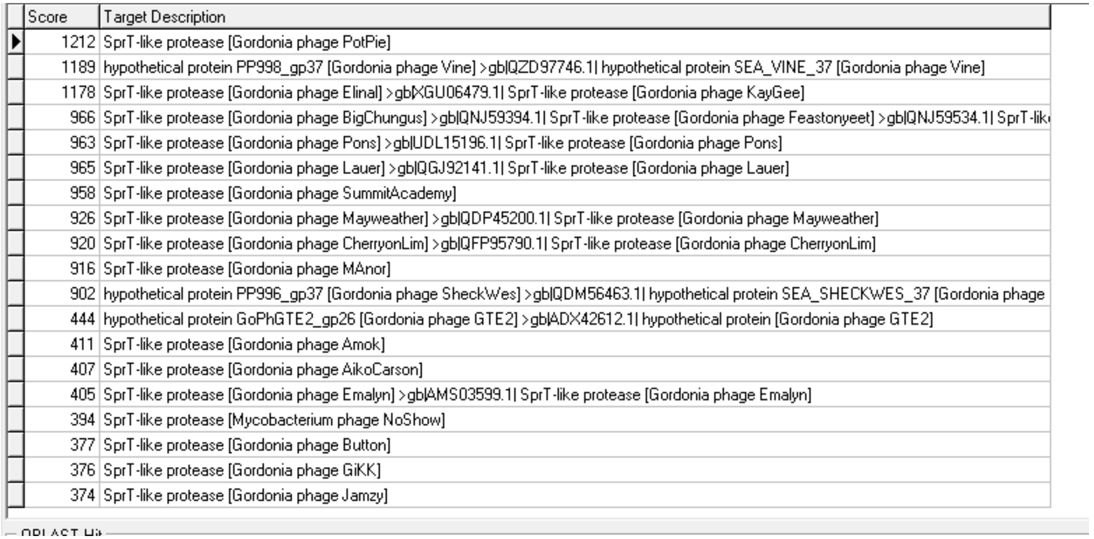
- I BLAST some potential start sites inside of the gap but there was no other proposed features that could be inserted in the gap

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	27610	27712
BLAST	8 1:1 hits	1 1:1 hit
RBS	Z-value:2.806 FS:-3.104	Z-value:2.730 FS:-3.201
Starterator	3 MA's	1 MA
Coding Potential	Includes all coding potential	Includes all coding potential
Gap/Overlap	218	116

- The start site is 27610 because although it has a bigger gap it has better blast hits, better RBS scores, and more MA's. Start 27712 also has a great chance since the gap is so big. I will have to see if I can find any genes that could be inserted between to see if it can get a clear choice.

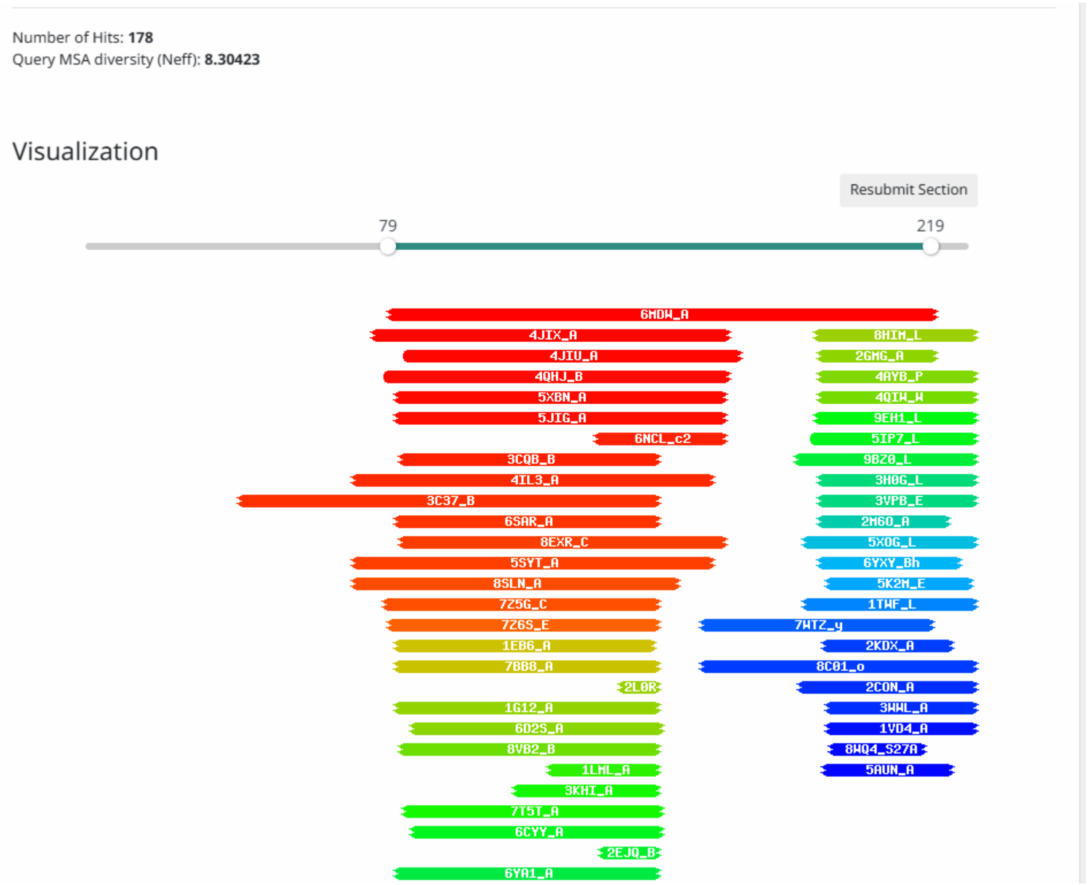
BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
1212	SprT-like protease [Gordonia phage PotPie]
1189	hypothetical protein PP998_gp37 [Gordonia phage Vine] >gb|QZD97746.1| hypothetical protein SEA_VINE_37 [Gordonia phage Vine]
1178	SprT-like protease [Gordonia phage Elinal] >gb|XGU06479.1| SprT-like protease [Gordonia phage KayGee]
966	SprT-like protease [Gordonia phage BigChungus] >gb|QNJ59394.1| SprT-like protease [Gordonia phage Feastonyeet] >gb|QNJ59534.1| SprT-lik
963	SprT-like protease [Gordonia phage Pons] >gb|UDL15196.1| SprT-like protease [Gordonia phage Pons]
965	SprT-like protease [Gordonia phage Lauer] >gb|QGJ92141.1| SprT-like protease [Gordonia phage Lauer]
958	SprT-like protease [Gordonia phage SummitAcademy]
926	SprT-like protease [Gordonia phage Mayweather] >gb|QDP45200.1| SprT-like protease [Gordonia phage Mayweather]
920	SprT-like protease [Gordonia phage CherryonLim] >gb|QFP95790.1| SprT-like protease [Gordonia phage CherryonLim]
916	SprT-like protease [Gordonia phage MAnor]
902	hypothetical protein PP996_gp37 [Gordonia phage SheckWes] >gb|QDM56463.1| hypothetical protein SEA_SHECKWES_37 [Gordonia phage
444	hypothetical protein GoPhGTE2_gp26 [Gordonia phage GTE2] >gb|ADX42612.1| hypothetical protein [Gordonia phage GTE2]
411	SprT-like protease [Gordonia phage Amok]
407	SprT-like protease [Gordonia phage AikoCarson]
405	SprT-like protease [Gordonia phage Emalyn] >gb|AMS03599.1| SprT-like protease [Gordonia phage Emalyn]
394	SprT-like protease [Mycobacterium phage NoShow]
377	SprT-like protease [Gordonia phage Button]
376	SprT-like protease [Gordonia phage GiKK]
374	SprT-like protease [Gordonia phage Jamzy]

- BLAST proposes that it may be a SprT –like protease and this has the majority of hits on the page

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

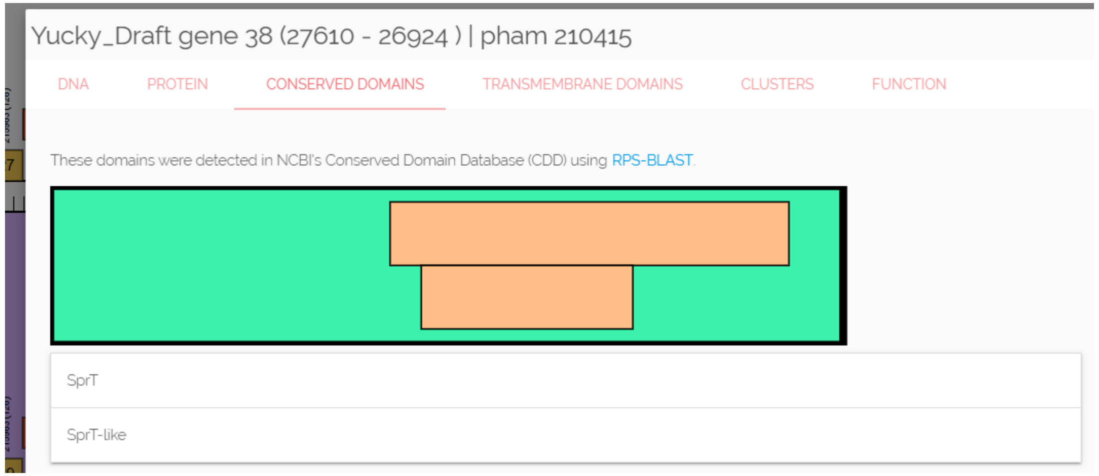


Number of Hits: **178**
Query MSA diversity (Neff): **8.30423**

Visualization

- The longest chain on the top gives evidence that this is a SprT -like protease. I’m not considering the others because their chains are so short.

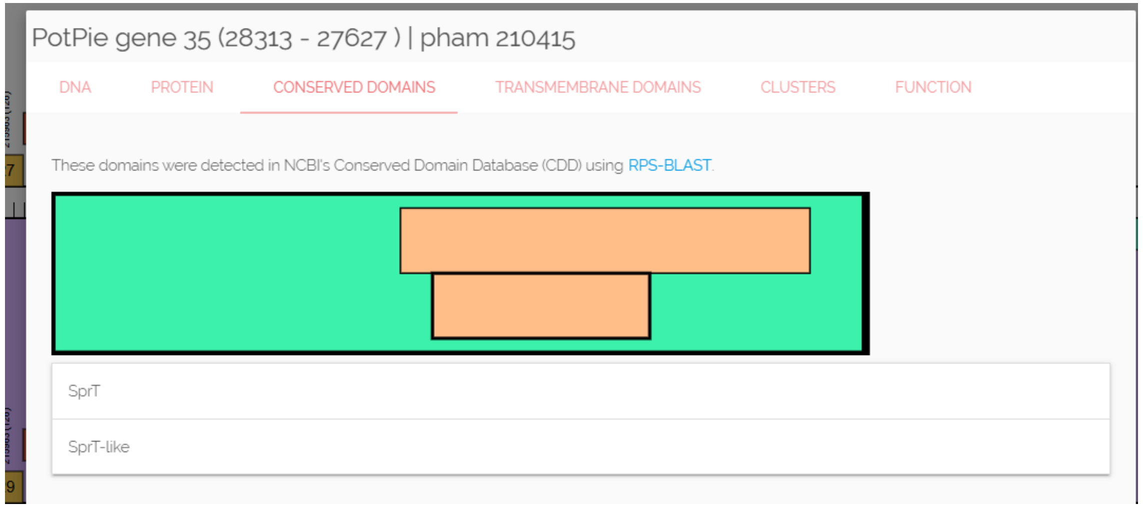
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?


Yucky_Draft gene 38 (27610 - 26924) | pham 210415

These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

SprT

SprT-like


PotPie gene 35 (28313 - 27627) | pham 210415

These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

SprT

SprT-like

- Phamerator gives evidence that this is a SprT –like protease because it matches perfectly with the same gene in PotPie

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- Since this has a possible function, Deep TMHMM not needed here.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function for this gene is a SprT –like protease due to it being called by HHPRED and having a very long chain being the strongest out of all of them called. It was called by phamerator and matched with the same gene in PotPie. It was also called numerous times in blast.

Feature 38 – Reverse – Stop 27829

Glimmer/GeneMark

What feature number is this?

What is the stop site?

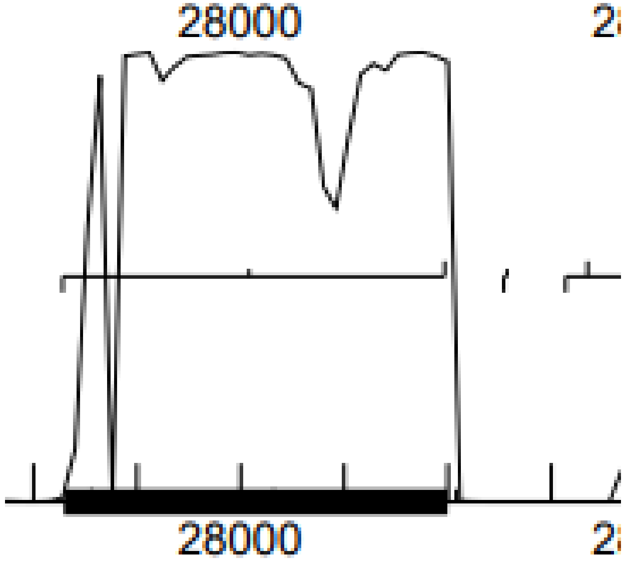
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

- 38
- 27829
- Both
- 28200
- It has a gap of 165 nucleotides

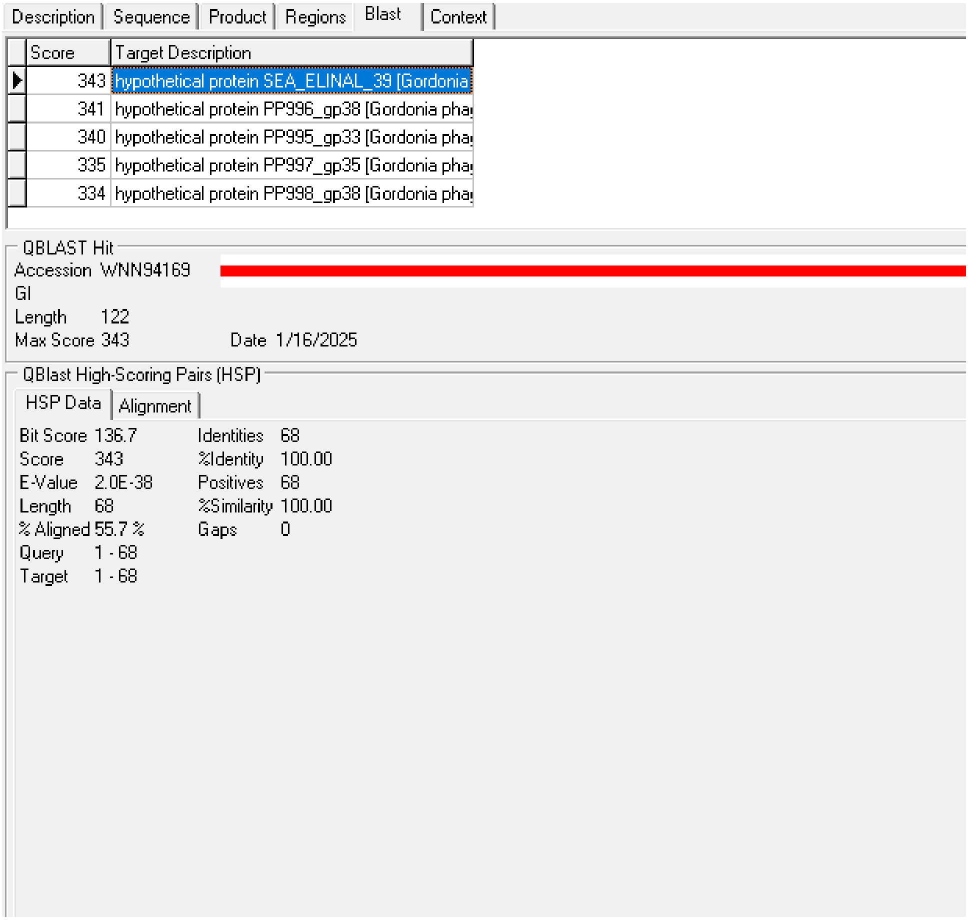
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- This graph has very great coding potential with it peaking the majority of the length of the feature

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are seven blast hits that have e-values from 10^-37 and 10^-38



Score	Target Description
343	hypothetical protein SEA_ELINAL_39 [Gordonia
341	hypothetical protein PP996_gp38 [Gordonia pha
340	hypothetical protein PP995_gp33 [Gordonia pha
335	hypothetical protein PP997_gp35 [Gordonia pha
334	hypothetical protein PP998_gp38 [Gordonia pha

QBLAST Hit

Accession WNN94169

GI

Length 122

Max Score 343

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	136.7	Identities	68
Score	343	%Identity	100.00
E-Value	2.0E-38	Positives	68
Length	68	%Similarity	100.00
% Aligned	55.7 %	Gaps	0
Query	1 - 68		
Target	1 - 68		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

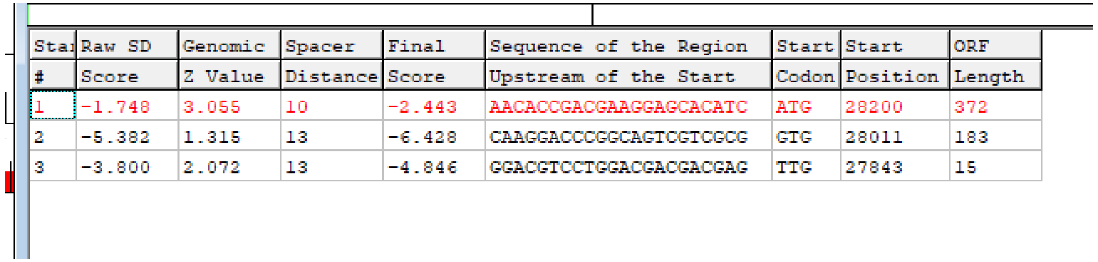
- Yes this feature is a gene because it was called by both glimmer and genemark, it has strong coding potential, and has seven blast hits that are at 10^-38.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There are 8 1:1 blast hits with phages like Elinal for start 28200

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

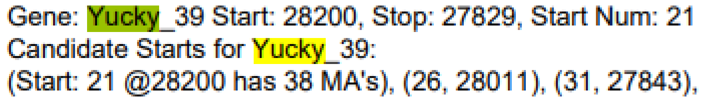
- The start of 28200 had a
- Z-value:3.055
- FS:-2.443
- This is the only start site that has decent RBS values



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	-1.748	3.055	10	-2.443	AACACCGACGAAGGAGCACATC	ATG	28200	372
2	-5.382	1.315	13	-6.428	CAAGGACCCGGCAGTCGTCGCG	GTG	28011	183
3	-3.800	2.072	13	-4.846	GGACGTCCTGGACGACGACGAG	TTG	27843	15

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

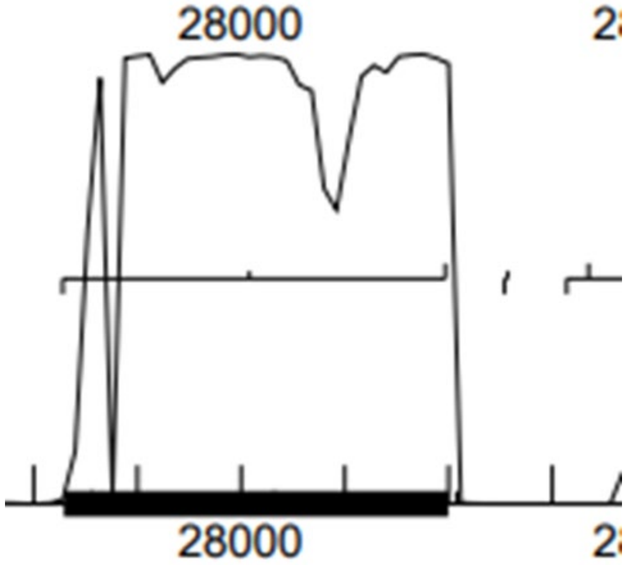
- Start 28200 has 38 MA's which is the only start that has any



Gene: Yucky_39 Start: 28200, Stop: 27829, Start Num: 21
Candidate Starts for Yucky_39:
(Start: 21 @28200 has 38 MA's), (26, 28011), (31, 27843),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- The start of 28200 includes almost all the coding potential it cuts off a tiny piece at the start



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

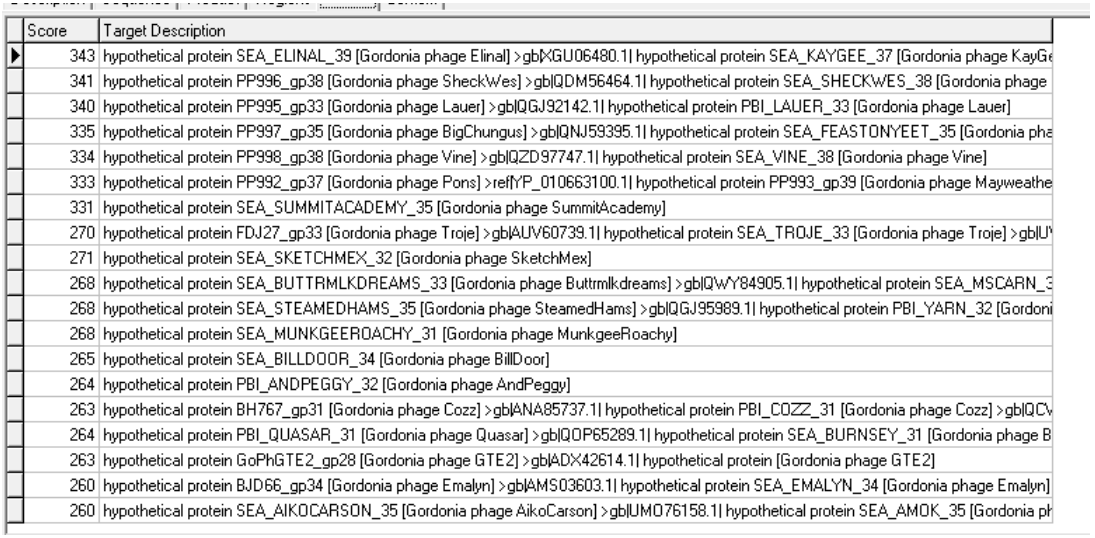
- There is a gap of 165 nucleotides

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site for feature 39 is 28200 because it is the only proposed site by genemark and glimmer, it has 38 MA’s, it has 8 1:1 alignments with phages like Elinal, and it includes the majority of coding potential.

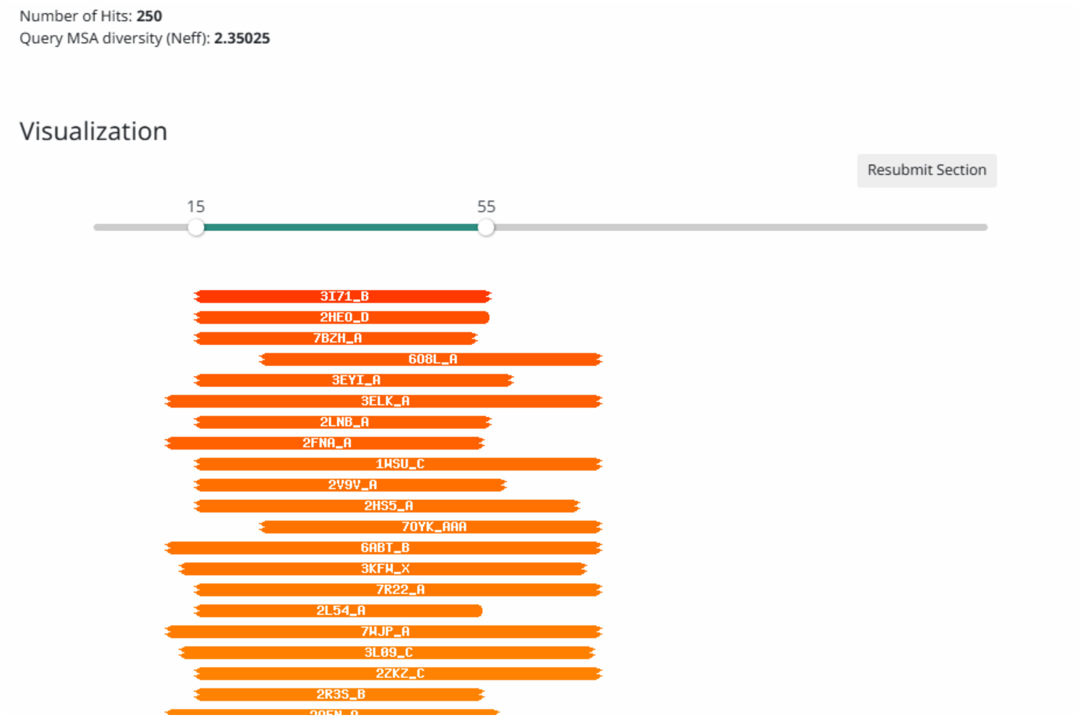
BLAST function evidence. What assigned functions do other highly similar genes have?

- It is called hypothetical protein by every single blast hit



<table>
<tr><th>Score</th><th>Target Description</th></tr>
<tr><td>343</td><td>hypothetical protein SEA_ELINAL_39 [Gordonia phage Elinal] >gb|XGU06480.1| hypothetical protein SEA_KAYGEE_37 [Gordonia phage KayGe</td></tr>
<tr><td>341</td><td>hypothetical protein PP996_gp38 [Gordonia phage SheckWes] >gb|QDM56464.1| hypothetical protein SEA_SHECKWES_38 [Gordonia phage</td></tr>
<tr><td>340</td><td>hypothetical protein PP995_gp33 [Gordonia phage Lauer] >gb|QGJ92142.1| hypothetical protein PBI_LAUER_33 [Gordonia phage Lauer]</td></tr>
<tr><td>335</td><td>hypothetical protein PP997_gp35 [Gordonia phage BigChungus] >gb|QNJ59395.1| hypothetical protein SEA_FEASTONYEET_35 [Gordonia pha</td></tr>
<tr><td>334</td><td>hypothetical protein PP998_gp38 [Gordonia phage Vine] >gb|QZD97747.1| hypothetical protein SEA_VINE_38 [Gordonia phage Vine]</td></tr>
<tr><td>333</td><td>hypothetical protein PP992_gp37 [Gordonia phage Pons] >ref|YP_010663100.1| hypothetical protein PP993_gp39 [Gordonia phage Mayweathe</td></tr>
<tr><td>331</td><td>hypothetical protein SEA_SUMMITACADEMY_35 [Gordonia phage SummitAcademy]</td></tr>
<tr><td>270</td><td>hypothetical protein FDJ27_gp33 [Gordonia phage Troje] >gb|AUV60739.1| hypothetical protein SEA_TROJE_33 [Gordonia phage Troje] >gb|U</td></tr>
<tr><td>271</td><td>hypothetical protein SEA_SKETCHMEX_32 [Gordonia phage SketchMex]</td></tr>
<tr><td>268</td><td>hypothetical protein SEA_BUTTRMLKDREAMS_33 [Gordonia phage Buttrmlkdreams] >gb|QWY84905.1| hypothetical protein SEA_MSCARN_3</td></tr>
<tr><td>268</td><td>hypothetical protein SEA_STEAMEDHAMS_35 [Gordonia phage SteamedHams] >gb|QGJ95989.1| hypothetical protein PBI_YARN_32 [Gordoni</td></tr>
<tr><td>268</td><td>hypothetical protein SEA_MUNKGEEROACHY_31 [Gordonia phage MunkgeeRoachy]</td></tr>
<tr><td>265</td><td>hypothetical protein SEA_BILLDOOR_34 [Gordonia phage BillDoor]</td></tr>
<tr><td>264</td><td>hypothetical protein PBI_ANDPEGGY_32 [Gordonia phage AndPeggy]</td></tr>
<tr><td>263</td><td>hypothetical protein BH767_gp31 [Gordonia phage Cozz] >gb|ANA85737.1| hypothetical protein PBI_COZZ_31 [Gordonia phage Cozz] >gb|QCV</td></tr>
<tr><td>264</td><td>hypothetical protein PBI_QUASAR_31 [Gordonia phage Quasar] >gb|QOP65289.1| hypothetical protein SEA_BURNSEY_31 [Gordonia phage B</td></tr>
<tr><td>263</td><td>hypothetical protein GoPhGTE2_gp28 [Gordonia phage GTE2] >gb|ADX42614.1| hypothetical protein [Gordonia phage GTE2]</td></tr>
<tr><td>260</td><td>hypothetical protein BJD66_gp34 [Gordonia phage Emalyn] >gb|AMS03603.1| hypothetical protein SEA_EMALYN_34 [Gordonia phage Emalyn]</td></tr>
<tr><td>260</td><td>hypothetical protein SEA_AIKOCARSON_35 [Gordonia phage AikoCarson] >gb|UMO76158.1| hypothetical protein SEA_AMOK_35 [Gordonia ph</td></tr>
</table>

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



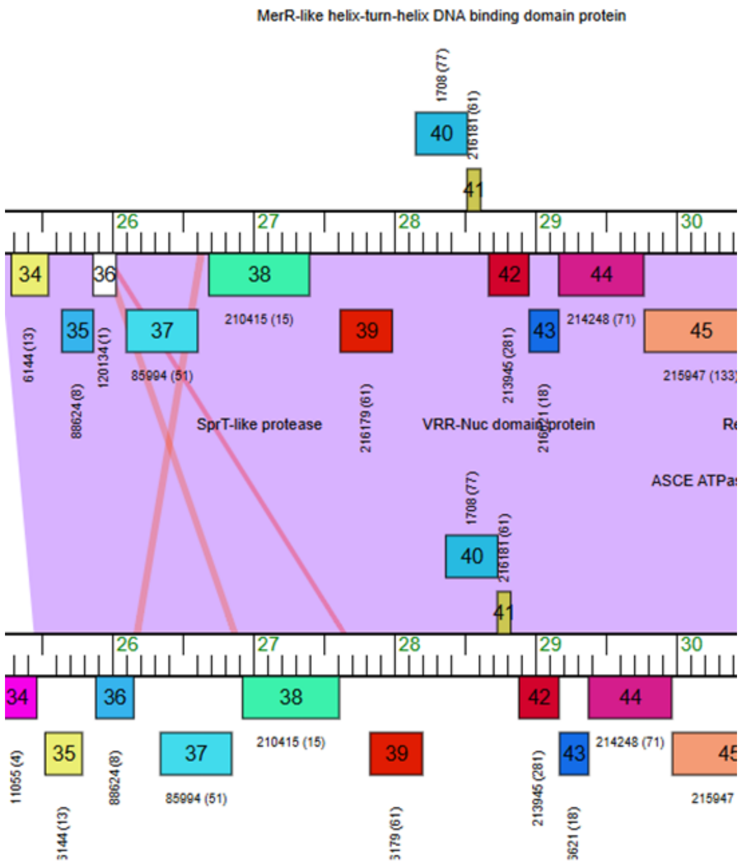
Number of Hits: **250**

Query MSA diversity (Neff): **2.35025**

Visualization

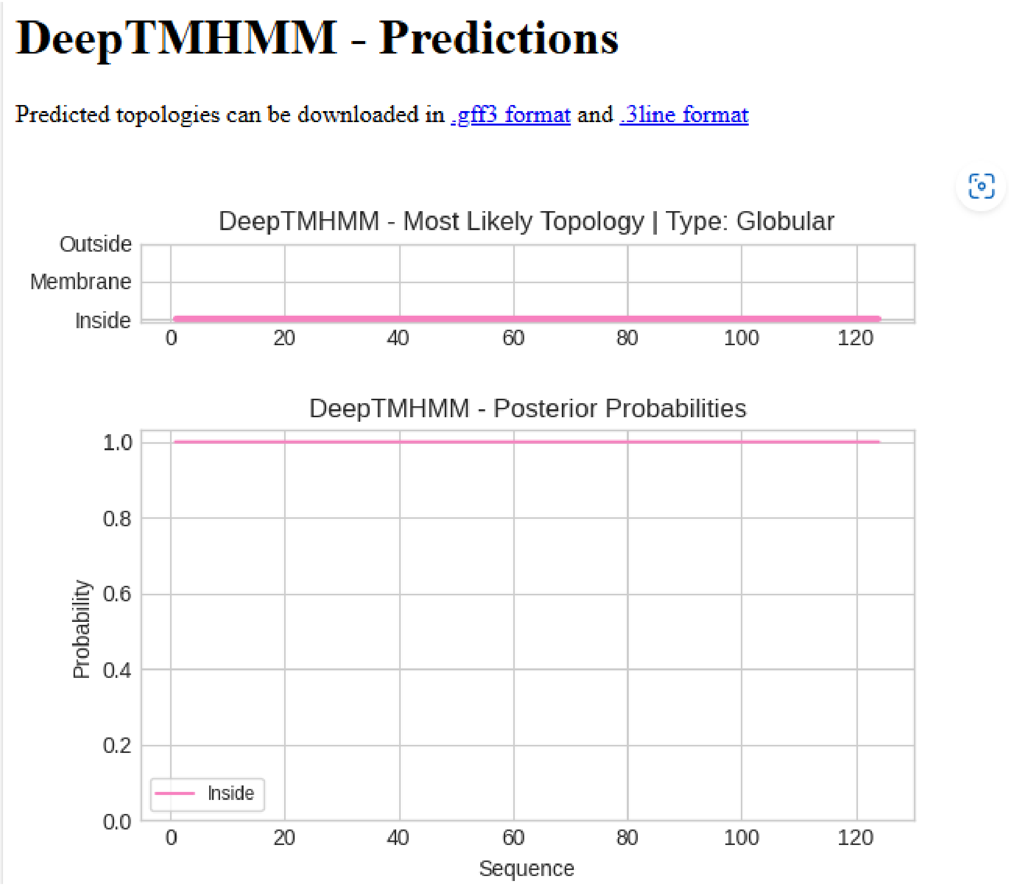
- The strongest hit in HHPRED called this to be an Ethanolamine utilization protein which is not on the official function list. It also doesn't have a probability greater than 90%

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- There is no proposed functions in other genes phamerator and phamerator didn’t call the gene anything

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

DeepTMHMM - Most Likely Topology | Type: Globular
Outside
Membrane
Inside
0
20
40
60
80
100
120
DeepTMHMM - Posterior Probabilities
1.0
0.8
0.6
0.4
0.2
0.0
Probability
Inside
Sequence

- There are no transmembrane domains called

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

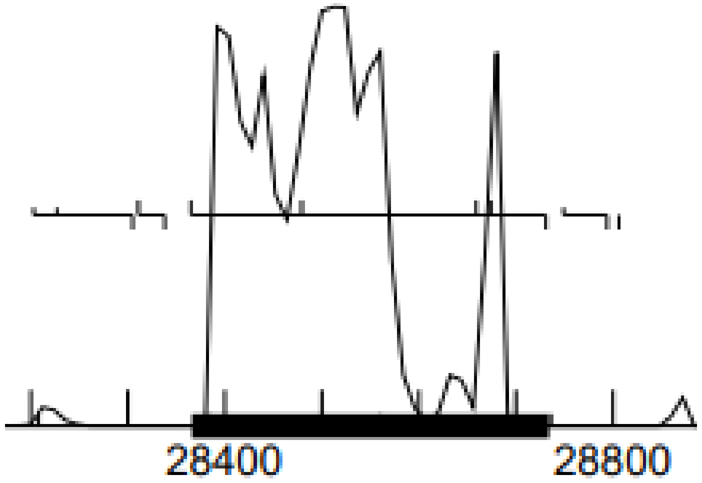
- I believe this function to be a hypothetical protein because every single blast hit said that it was. HHPRED evidence was thrown out due to the function being unknown of the function it called. Phamerator didn’t call any function and there are no transmembrane domains.

Feature 39 Stop 28734

Glimmer/GeneMark

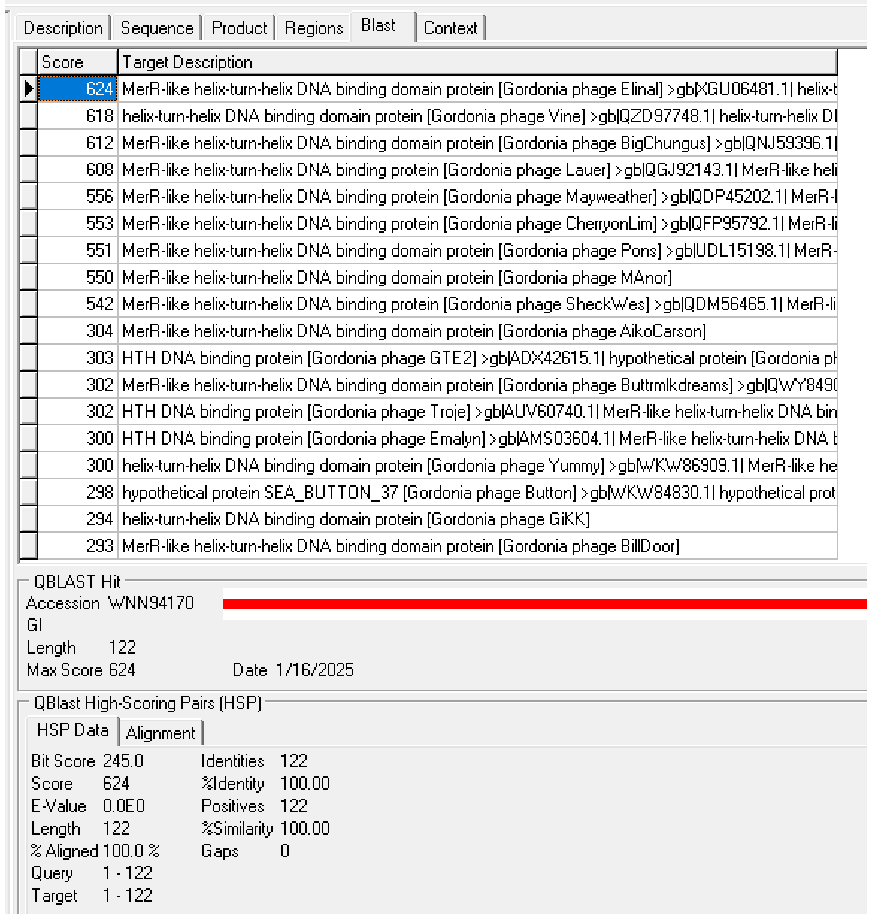
What feature number is this?	• 39
What is the stop site?	• 28734
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?	• Both
What is the autoannotated start?	• 28366
Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start	• Gap of 165

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- There is some strong coding potential in this graph. It falls off almost to zero for about 50 nucleotides, but it does have another peak near the end

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
624	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage Elinal] >gb|XGU06481.1| helix-t
618	helix-turn-helix DNA binding domain protein [Gordonia phage Vine] >gb|QZD97748.1| helix-turn-helix DI
612	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage BigChungus] >gb|QNJ59396.1|
608	MerR-like helix-turn-helix DNA binding protein [Gordonia phage Lauer] >gb|QGJ92143.1| MerR-like heli
556	MerR-like helix-turn-helix DNA binding protein [Gordonia phage Mayweather] >gb|QDP45202.1| MerR-l
553	MerR-like helix-turn-helix DNA binding protein [Gordonia phage CherryonLim] >gb|QFP95792.1| MerR-li
551	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage Pons] >gb|UDL15198.1| MerR-
550	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage MAnor]
542	MerR-like helix-turn-helix DNA binding protein [Gordonia phage SheckWes] >gb|QDM56465.1| MerR-li
304	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage AikoCarson]
303	HTH DNA binding protein [Gordonia phage GTE2] >gb|ADX42615.1| hypothetical protein [Gordonia pl
302	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage Buttrmlkdreams] >gb|QWY849(
302	HTH DNA binding protein [Gordonia phage Troje] >gb|AUV60740.1| MerR-like helix-turn-helix DNA bin
300	HTH DNA binding protein [Gordonia phage Emalyn] >gb|AMS03604.1| MerR-like helix-turn-helix DNA b
300	helix-turn-helix DNA binding domain protein [Gordonia phage Yummy] >gb|WKW86909.1| MerR-like he
298	hypothetical protein SEA_BUTTON_37 [Gordonia phage Button] >gb|WKW84830.1| hypothetical prot
294	helix-turn-helix DNA binding domain protein [Gordonia phage GiKK]
293	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage BillDoor]

QBLAST Hit

Accession WNN94170
GI
Length 122
Max Score 624
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	245.0	Identities	122
Score	624	%Identity	100.00
E-Value	0.0E0	Positives	122
Length	122	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 122		
Target	1 - 122		

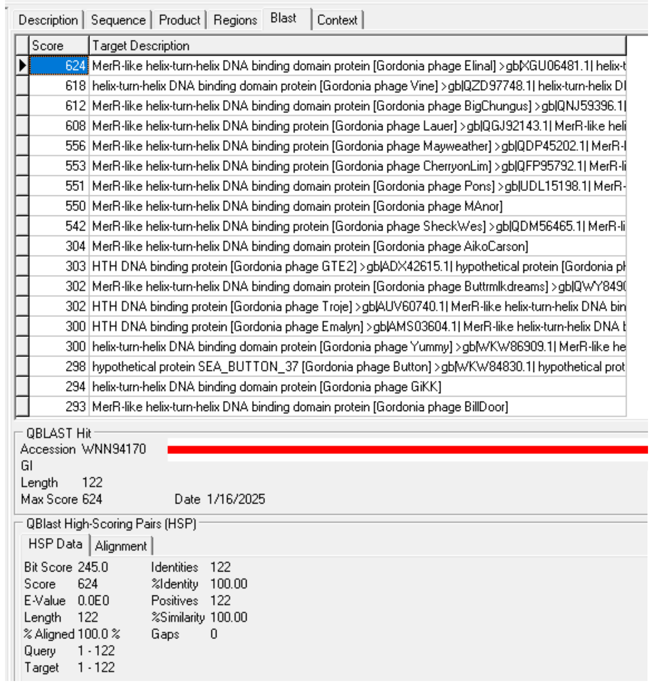
- 9 hits with an e-value of zero

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes this feature is a gene because it was called by both glimmer and genemark while also having strong coding potential and several e-values that were zero.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There's 21 1:1 hits in blast which makes this a very good start
- The second proposed start at 28558 has only alignments of 1:65

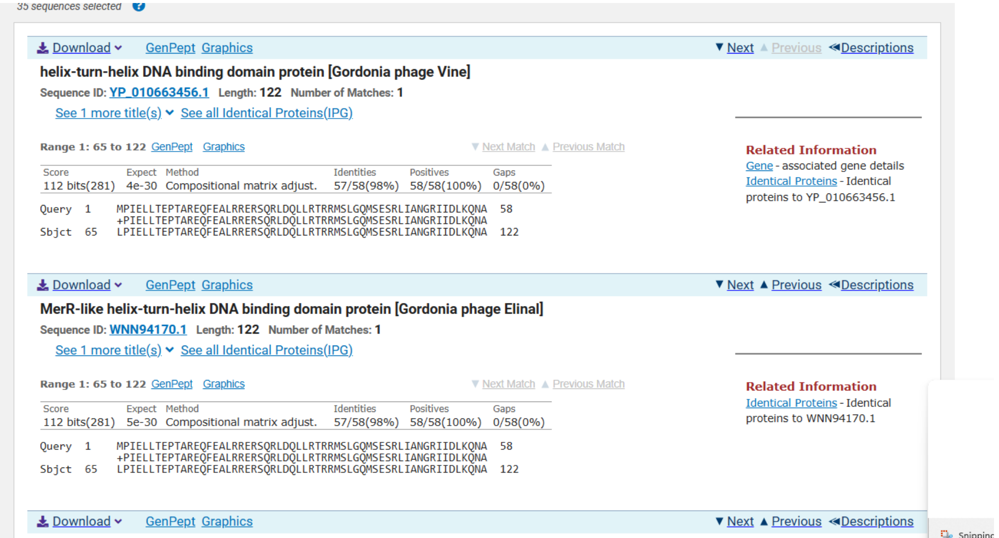


Score	Target Description
624	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage Elinal] >gb|XGU06481.1| helix-t
618	helix-turn-helix DNA binding domain protein [Gordonia phage Vine] >gb|QZD97748.1| helix-turn-helix DI
612	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage BigChungus] >gb|QNJ59396.1|
608	MerR-like helix-turn-helix DNA binding protein [Gordonia phage Lauer] >gb|QGJ92143.1| MerR-like heli
556	MerR-like helix-turn-helix DNA binding protein [Gordonia phage Mayweather] >gb|QDP45202.1| MerR-l
553	MerR-like helix-turn-helix DNA binding protein [Gordonia phage CherryonLim] >gb|QFP95792.1| MerR-li
551	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage Pons] >gb|UDL15198.1| MerR-
550	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage MAnor]
542	MerR-like helix-turn-helix DNA binding protein [Gordonia phage SheckWes] >gb|QDM56465.1| MerR-li
304	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage AikoCarson]
303	HTH DNA binding protein [Gordonia phage GTE2] >gb|ADX42615.1| hypothetical protein [Gordonia ph
302	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage Buttrmlkdreams] >gb|QWY849(
302	HTH DNA binding protein [Gordonia phage Troje] >gb|AUV60740.1| MerR-like helix-turn-helix DNA bin
300	HTH DNA binding protein [Gordonia phage Emalyn] >gb|AMS03604.1| MerR-like helix-turn-helix DNA b
300	helix-turn-helix DNA binding domain protein [Gordonia phage Yummy] >gb|WKW86909.1| MerR-like he
298	hypothetical protein SEA_BUTTON_37 [Gordonia phage Button] >gb|WKW84830.1| hypothetical prot
294	helix-turn-helix DNA binding domain protein [Gordonia phage GiKK]
293	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage BillDoor]

QBLAST Hit
Accession WNN94170
GI
Length 122
Max Score 624 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	245.0	Identities	122
Score	624	%Identity	100.00
E-Value	0.0E0	Positives	122
Length	122	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 122		
Target	1 - 122		


35 sequences selected

helix-turn-helix DNA binding domain protein [Gordonia phage Vine]

Sequence ID: YP_010663456.1 Length: 122 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 65 to 122 GenPept Graphics

```
Score          Expect  Method                        Identities   Positives     Gaps
112 bits(281)  4e-30   Compositional matrix adjust.  57/58(98%)   58/58(100%)   0/58(0%)

Query  1    MPIELLTEPTAREQFEALRRERSQRLDQLLRTRRMSLGQMSESRLIANGRIIDLKQNA  58
            +PIELLTEPTAREQFEALRRERSQRLDQLLRTRRMSLGQMSESRLIANGRIIDLKQNA
Sbjct  65   LPIELLTEPTAREQFEALRRERSQRLDQLLRTRRMSLGQMSESRLIANGRIIDLKQNA  122
```

Related Information
Gene - associated gene details
Identical Proteins - Identical proteins to YP_010663456.1

MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage Elinal]

Sequence ID: WNN94170.1 Length: 122 Number of Matches: 1

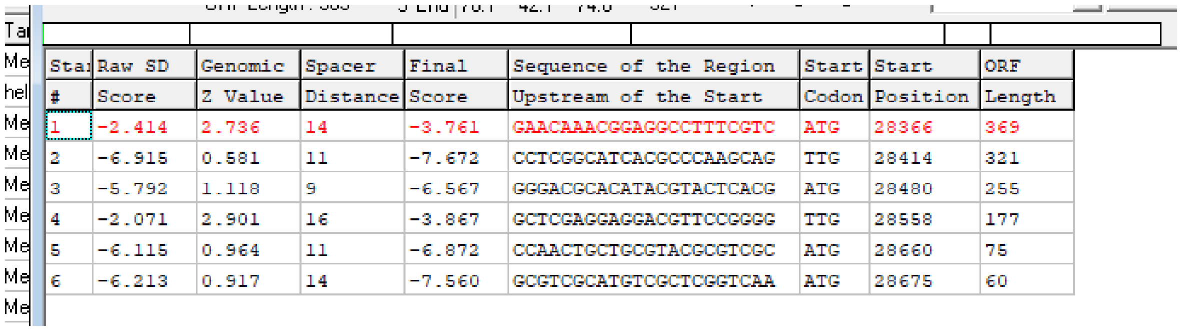
See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 65 to 122 GenPept Graphics

```
Score          Expect  Method                        Identities   Positives     Gaps
112 bits(281)  5e-30   Compositional matrix adjust.  57/58(98%)   58/58(100%)   0/58(0%)

Query  1    MPIELLTEPTAREQFEALRRERSQRLDQLLRTRRMSLGQMSESRLIANGRIIDLKQNA  58
            +PIELLTEPTAREQFEALRRERSQRLDQLLRTRRMSLGQMSESRLIANGRIIDLKQNA
Sbjct  65   LPIELLTEPTAREQFEALRRERSQRLDQLLRTRRMSLGQMSESRLIANGRIIDLKQNA  122
```

Related Information
Identical Proteins - Identical proteins to WNN94170.1

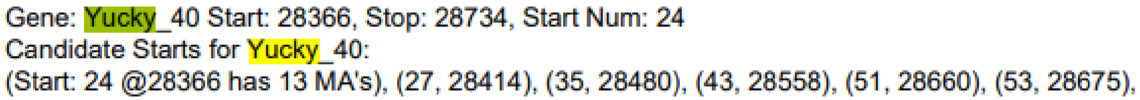
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts?   Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-2.414	2.736	14	-3.761	GAACAAACGGAGGCCTTTCGTC	ATG	28366	369
2	-6.915	0.581	11	-7.672	CCTCGGCATCACGCCCAAGCAG	TTG	28414	321
3	-5.792	1.118	9	-6.567	GGGACGCACATACGTACTCACG	ATG	28480	255
4	-2.071	2.901	16	-3.867	GCTCGAGGAGGACGTTCCGGGG	TTG	28558	177
5	-6.115	0.964	11	-6.872	CCAACTGCTGCGTACGCGTCGC	ATG	28660	75
6	-6.213	0.917	14	-7.560	GCGTCGCATGTCGCTCGGTCAA	ATG	28675	60

- 28366
- Z-value:2.736
- FS:-3.761
- 28558
- Z-value:2.901
- FS:-3.867
- The scores proposed a new start at 28558

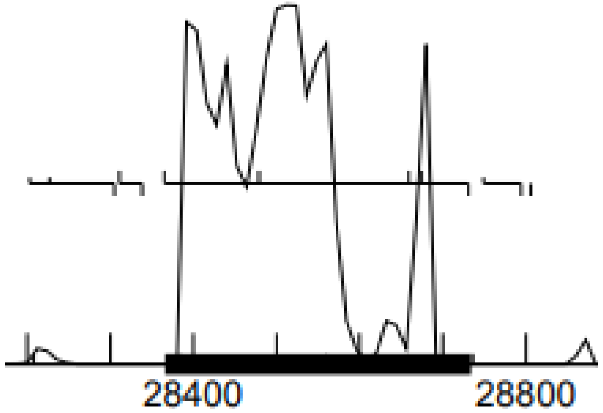
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



```
Gene: Yucky_40 Start: 28366, Stop: 28734, Start Num: 24
Candidate Starts for Yucky_40:
(Start: 24 @28366 has 13 MA's), (27, 28414), (35, 28480), (43, 28558), (51, 28660), (53, 28675),
```

- The original start has 13 manual annotations with the secondary start having none making the original start preferred

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- The original start includes all of the coding potential for this graph while the second proposed start cuts off about 100 nucleotides

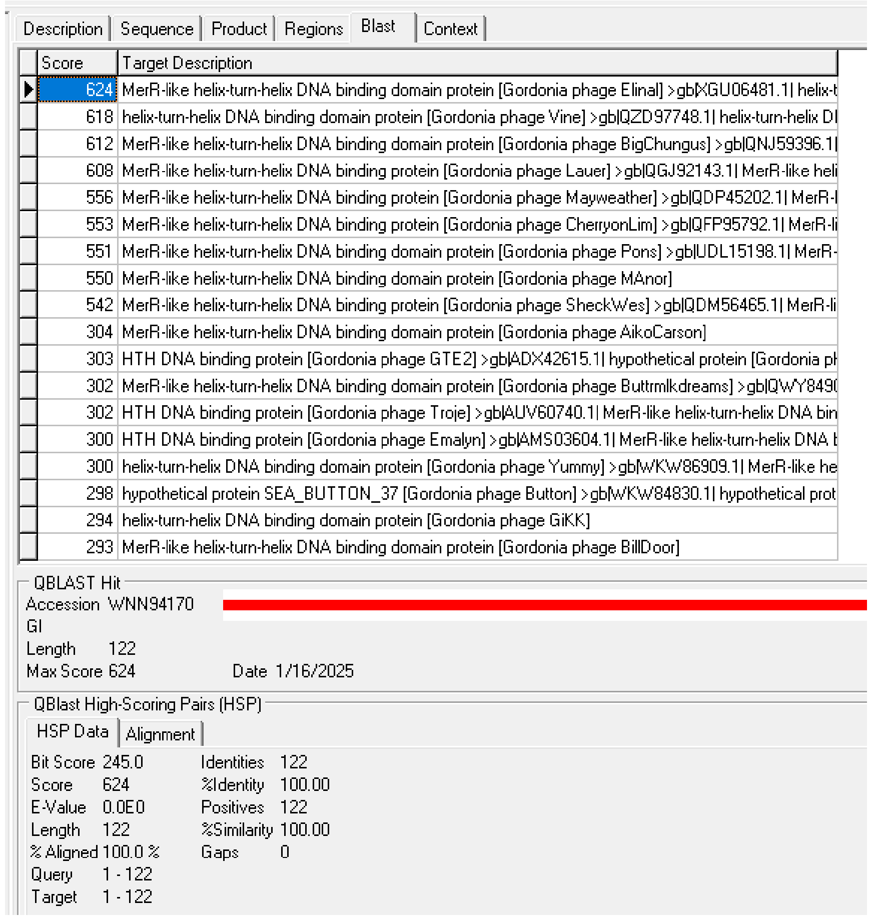
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- There is a gap of 165 with feature 39

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start is 28366 which is the original called start for this feature. It has 21 1:1 alignments, 13 manual annotations, includes all coding potential, and is called by both glimmer and genemark as the start.

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
624	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage Elinal] >gb|XGU06481.1| helix-t
618	helix-turn-helix DNA binding domain protein [Gordonia phage Vine] >gb|QZD97748.1| helix-turn-helix D
612	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage BigChungus] >gb|QNJ59396.1|
608	MerR-like helix-turn-helix DNA binding protein [Gordonia phage Lauer] >gb|QGJ92143.1| MerR-like heli
556	MerR-like helix-turn-helix DNA binding protein [Gordonia phage Mayweather] >gb|QDP45202.1| MerR-l
553	MerR-like helix-turn-helix DNA binding protein [Gordonia phage CherryonLim] >gb|QFP95792.1| MerR-li
551	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage Pons] >gb|UDL15198.1| MerR-
550	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage MAnor]
542	MerR-like helix-turn-helix DNA binding protein [Gordonia phage SheckWes] >gb|QDM56465.1| MerR-li
304	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage AikoCarson]
303	HTH DNA binding protein [Gordonia phage GTE2] >gb|ADX42615.1| hypothetical protein [Gordonia ph
302	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage Buttrmlkdreams] >gb|QWY849
302	HTH DNA binding protein [Gordonia phage Troje] >gb|AUV60740.1| MerR-like helix-turn-helix DNA bin
300	HTH DNA binding protein [Gordonia phage Emalyn] >gb|AMS03604.1| MerR-like helix-turn-helix DNA b
300	helix-turn-helix DNA binding domain protein [Gordonia phage Yummy] >gb|WKW86909.1| MerR-like he
298	hypothetical protein SEA_BUTTON_37 [Gordonia phage Button] >gb|WKW84830.1| hypothetical prot
294	helix-turn-helix DNA binding domain protein [Gordonia phage GiKK]
293	MerR-like helix-turn-helix DNA binding domain protein [Gordonia phage BillDoor]

QBLAST Hit

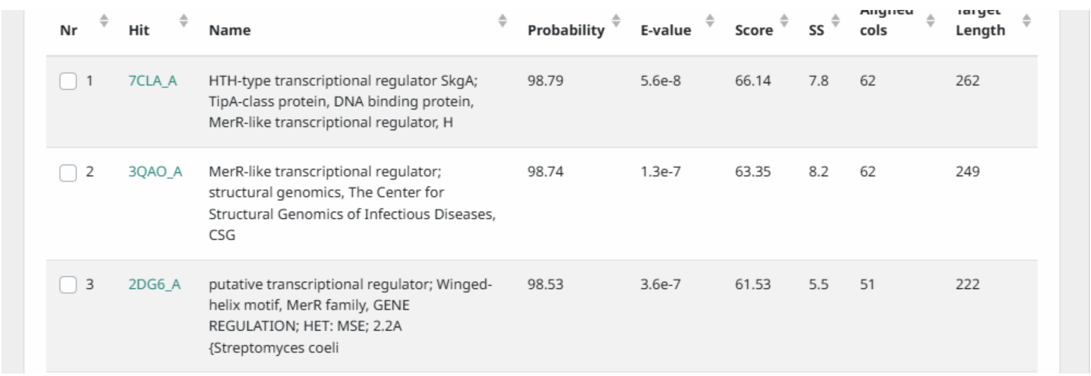
Accession WNN94170
GI
Length 122
Max Score 624 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

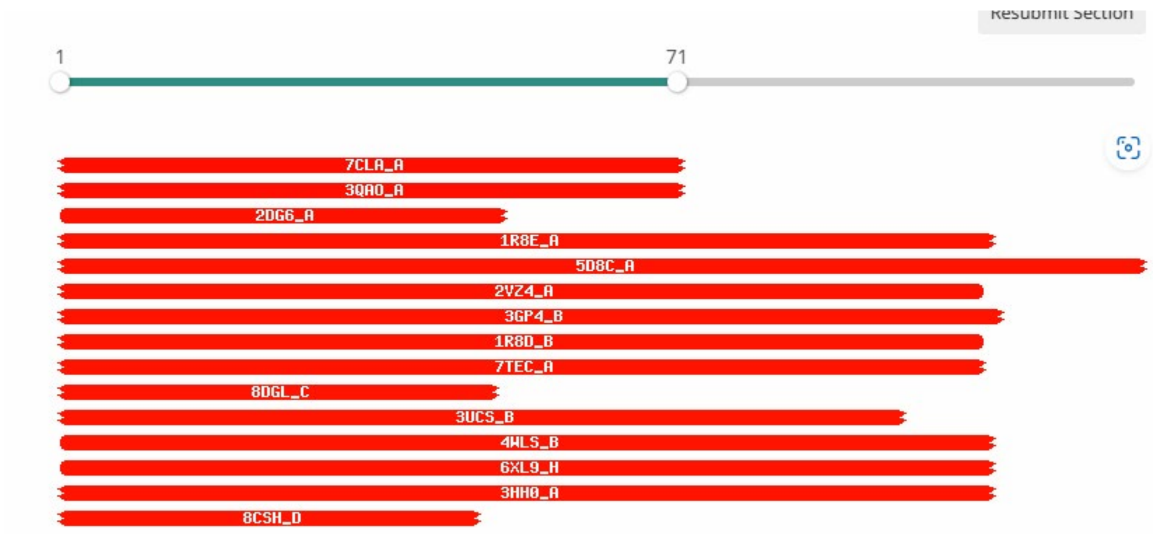
Bit Score	245.0	Identities	122
Score	624	%Identity	100.00
E-Value	0.0E0	Positives	122
Length	122	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 122		
Target	1 - 122		

- Blast proposes MerR-like helix-turn-helix as the most likely function of this gene

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

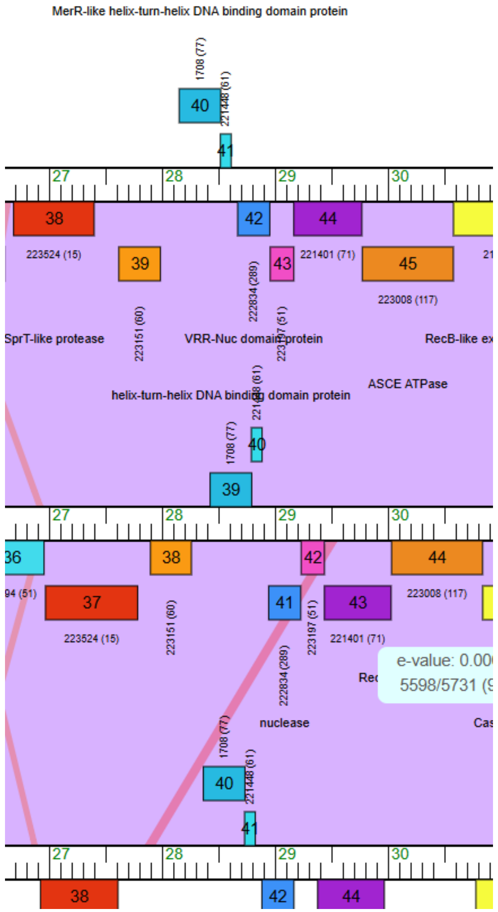


Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	7CLA_A	HTH-type transcriptional regulator SkgA; TipA-class protein, DNA binding protein, MerR-like transcriptional regulator, H	98.79	5.6e-8	66.14	7.8	62	262
□ 2	3QAO_A	MerR-like transcriptional regulator; structural genomics, The Center for Structural Genomics of Infectious Diseases, CSG	98.74	1.3e-7	63.35	8.2	62	249
□ 3	2DG6_A	putative transcriptional regulator; Winged-helix motif, MerR family, GENE REGULATION; HET: MSE; 2.2A {Streptomyces coeli	98.53	3.6e-7	61.53	5.5	51	222



- All of the HHPRED hits suggest that it is in the MerR family which makes me want to believe it is a MerR-like helix-turn-helix

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- In other phages like vine and elinal this feature was called a merR-like helix-turn-helix and just helix-turn-helix

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function of this gene is a helix-turn-helix DNA binding domain, MerR-like because in HHPRED the hot hits all suggested that it was part of the MerR family and blast and phamerator also suggested the same thing

Feature 40 – Stop 28826

Glimmer/GeneMark

What feature number is this?

What is the stop site?

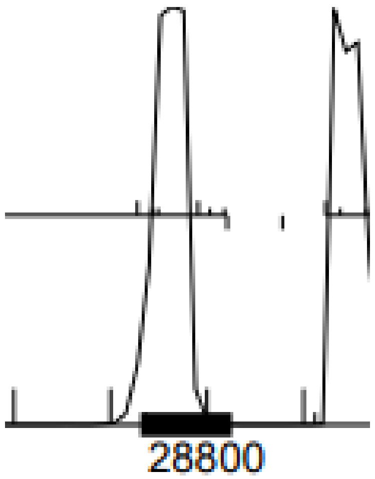
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

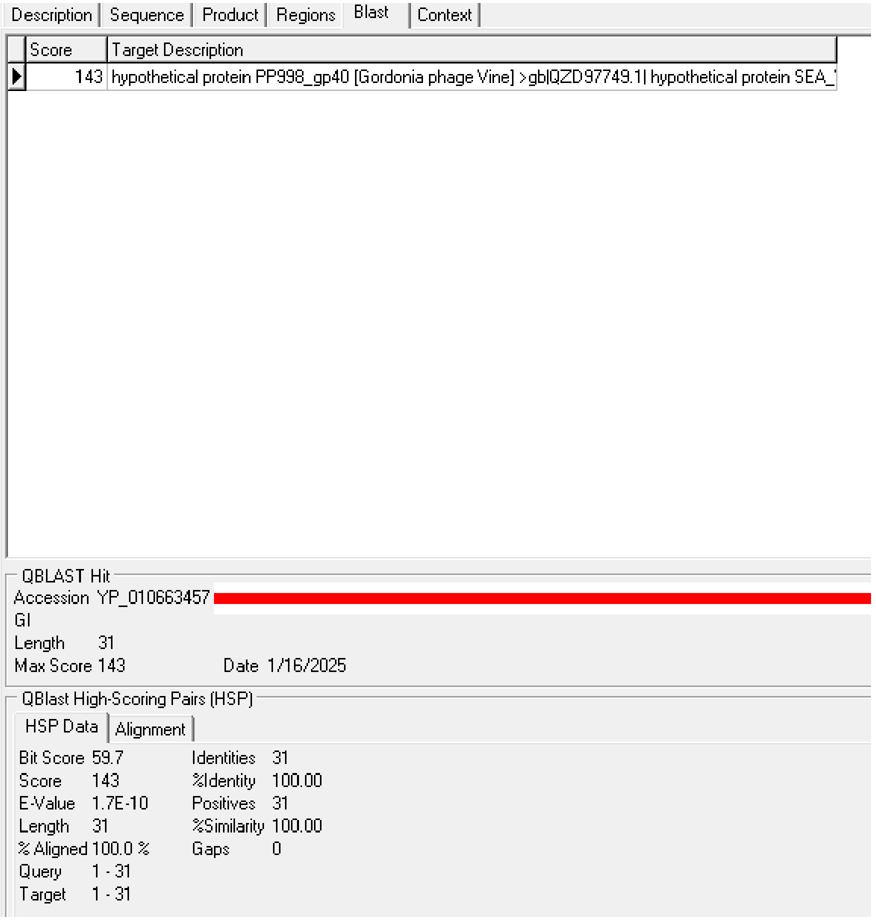
- 40
- 28826
- Genemark
- 28731
- Overlap of 4

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- There is strong coding potential for this feature, but it only includes some of it at the current start as it cuts off about a fourth of it.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
143	hypothetical protein PP998_gp40 [Gordonia phage Vine] >gb|QZD97749.1| hypothetical protein SEA_

QBLAST Hit
Accession YP_010663457
GI
Length 31
Max Score 143 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

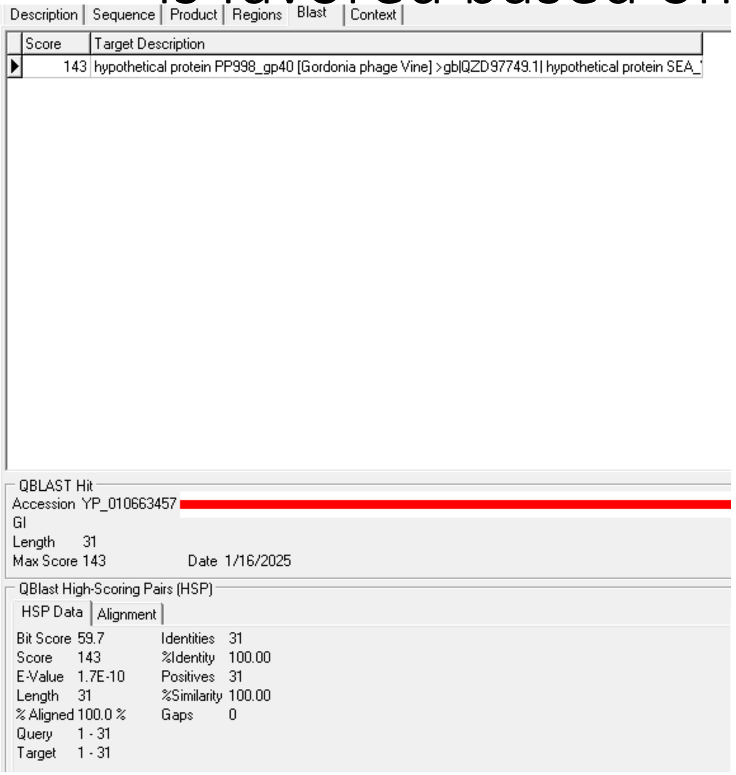
Bit Score	59.7	Identities	31
Score	143	%Identity	100.00
E-Value	1.7E-10	Positives	31
Length	31	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 31		
Target	1 - 31		

- There is only one blast hit and the e-value for it is past the acceptable amount at 1.7^-10

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- I think this feature is a gene because it is called by genemark, it has high coding potential, and it has a blast hit that is close to zero

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
143	hypothetical protein PP998_gp40 [Gordonia phage Vine] >gb|QZD97749.1| hypothetical protein SEA_

QBLAST Hit

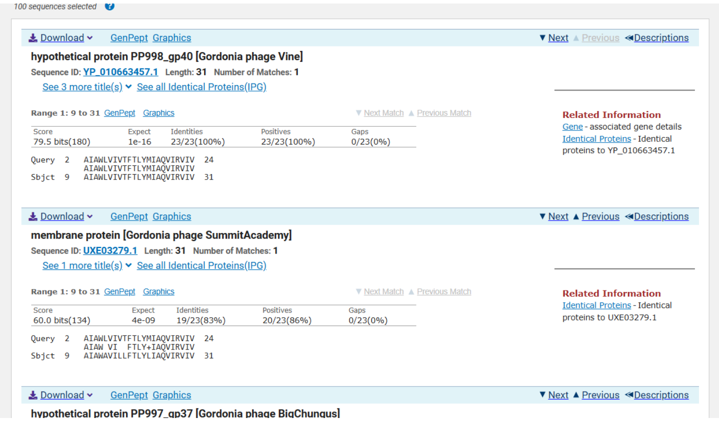
Accession YP_010663457
GI
Length 31
Max Score 143 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

HSP Data | Alignment

Bit Score	59.7	Identities	31
Score	143	%Identity	100.00
E-Value	1.7E-10	Positives	31
Length	31	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 31		
Target	1 - 31		

- There is only one blast hit for this feature and it has a blast hit of 1:1
- For 28752 there are multiple hits but they are 2:9 alignments



hypothetical protein PP998_gp40 [Gordonia phage Vine]

Sequence ID: YP_010663457.1 Length: 31 Number of Matches: 1

See 3 more title(s) See all Identical Proteins(IPG)

Range 1: 9 to 31 GenPept Graphics

Score	Expect	Identities	Positives	Gaps
79.5 bits(180)	1e-16	23/23(100%)	23/23(100%)	0/23(0%)

```
Query  2    AIAWLVIVTFTLYMIAQVIRVIV  24
            AIAWLVIVTFTLYMIAQVIRVIV
Sbjct  9    AIAWLVIVTFTLYMIAQVIRVIV  31
```

Related Information
Gene - associated gene details
Identical Proteins - Identical proteins to YP_010663457.1

membrane protein [Gordonia phage SummitAcademy]

Sequence ID: UXE03279.1 Length: 31 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 9 to 31 GenPept Graphics

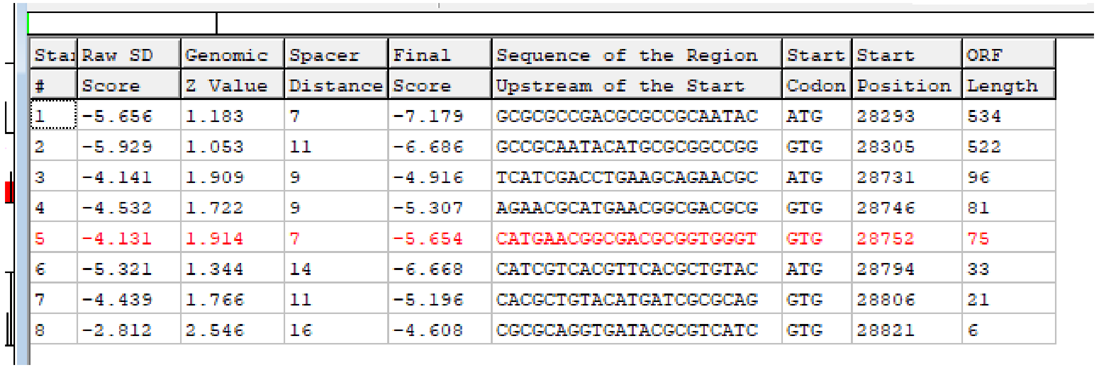
Score	Expect	Identities	Positives	Gaps
60.0 bits(134)	4e-09	19/23(83%)	20/23(86%)	0/23(0%)

```
Query  2    AIAWLVIVTFTLYMIAQVIRVIV  24
            AIAW VI  FTLY+IAQVIRVIV
Sbjct  9    AIAWAVILLFTLYLIAQVIRVIV  31
```

Related Information
Identical Proteins - Identical proteins to UXE03279.1

hypothetical protein PP997_gp37 [Gordonia phage BigChungus]

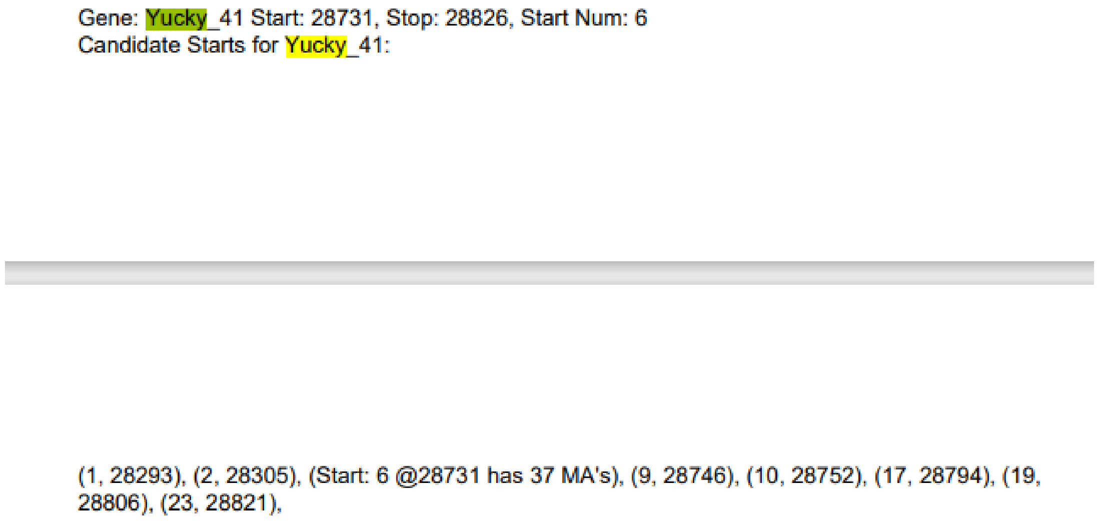
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-5.656	1.183	7	-7.179	GCGCGCCGACGCGCCGCAATAC	ATG	28293	534
2	-5.929	1.053	11	-6.686	GCCGCAATACATGCGCGGCCGG	GTG	28305	522
3	-4.141	1.909	9	-4.916	TCATCGACCTGAAGCAGAACGC	ATG	28731	96
4	-4.532	1.722	9	-5.307	AGAACGCATGAACGGCGACGCG	GTG	28746	81
5	-4.131	1.914	7	-5.654	CATGAACGGCGACGCGGTGGGT	GTG	28752	75
6	-5.321	1.344	14	-6.668	CATCGTCACGTTCACGCTGTAC	ATG	28794	33
7	-4.439	1.766	11	-5.196	CACGCTGTACATGATCGCGCAG	GTG	28806	21
8	-2.812	2.546	16	-4.608	CGCGCAGGTGATACGCGTCATC	GTG	28821	6

- RBS scores show multiple new proposed starts that this gene could have one at 28752 and one at 28821 which I'm going to automatically boot out since that would only make the feature 5 nucleotides long although it has the best scores. For now 28731 z-value:1.909 FS:-4.916 28752 z-value:1.914 FS:-5.654

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

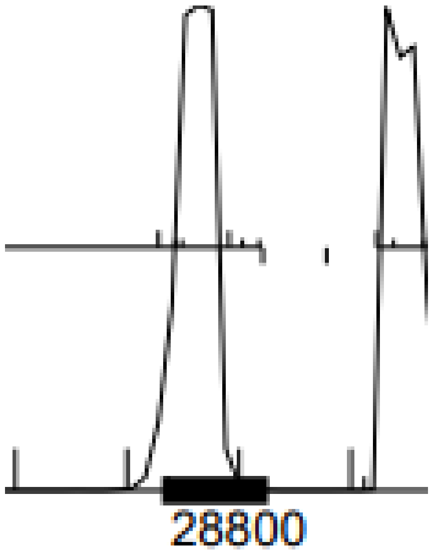


Gene: Yucky_41 Start: 28731, Stop: 28826, Start Num: 6
Candidate Starts for Yucky_41:

(1, 28293), (2, 28305), (Start: 6 @28731 has 37 MA's), (9, 28746), (10, 28752), (17, 28794), (19, 28806), (23, 28821),

- The start of 28731 has 37 manual annotations and the start of 28752 has none

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- The first start of 28731 does cut off about a fourth of the coding potential shown in the graph while the second start of 28752 cuts off over half of the coding potential and most of the strongest coding potential is included in this area.

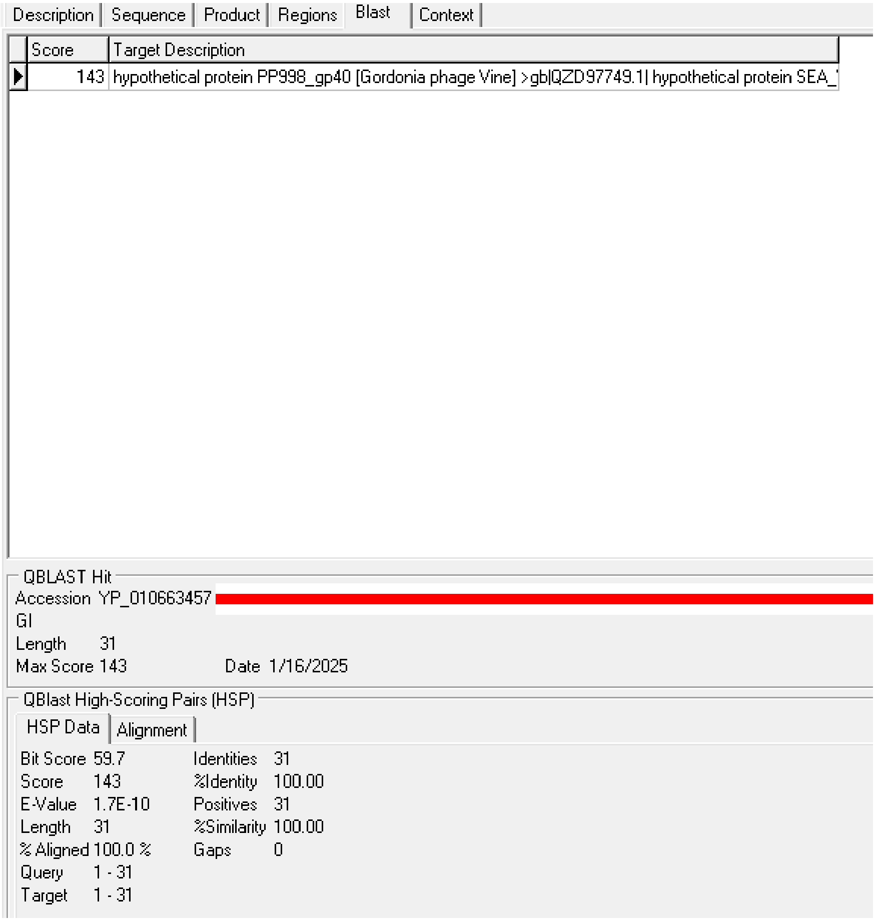
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start?   Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Overlap of 4 for 28731
- Gap of 17 for 28752

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start for this gene is 28731 because the other start of 28752 only has rbs scores going for it while the first start includes most of the coding potential, 37 manual annotations, and a 1:1 alignment in blast. The overlap is 4.

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
143	hypothetical protein PP998_gp40 [Gordonia phage Vine] >gb|QZD97749.1| hypothetical protein SEA_

QBLAST Hit

Accession YP_010663457

GI

Length 31

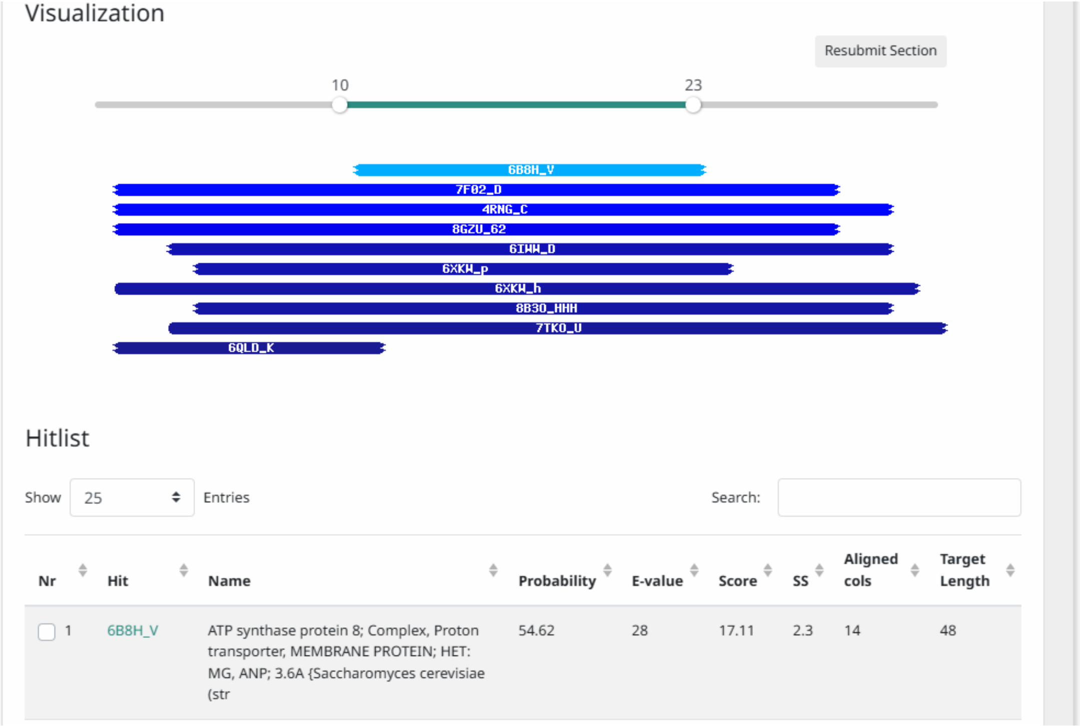
Max Score 143 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	59.7	Identities	31
Score	143	%Identity	100.00
E-Value	1.7E-10	Positives	31
Length	31	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 31		
Target	1 - 31		

- The only proposed function for this feature is a hypothetical protein

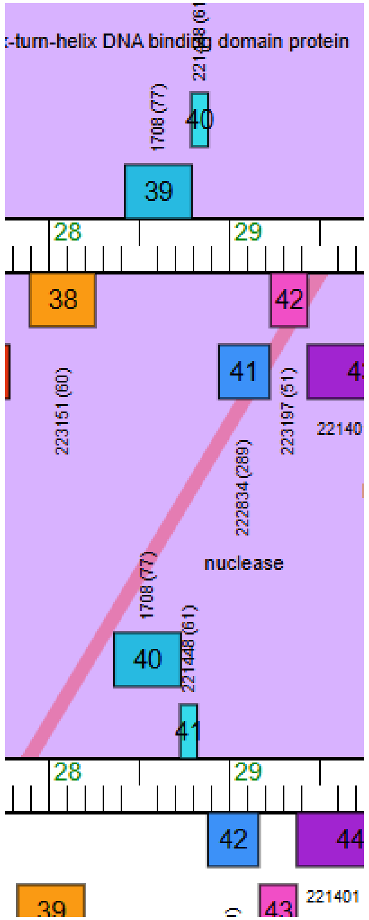
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	6B8H_V	ATP synthase protein 8; Complex, Proton transporter, MEMBRANE PROTEIN; HET: MG, ANP; 3.6A {Saccharomyces cerevisiae (str	54.62	28	17.11	2.3	14	48

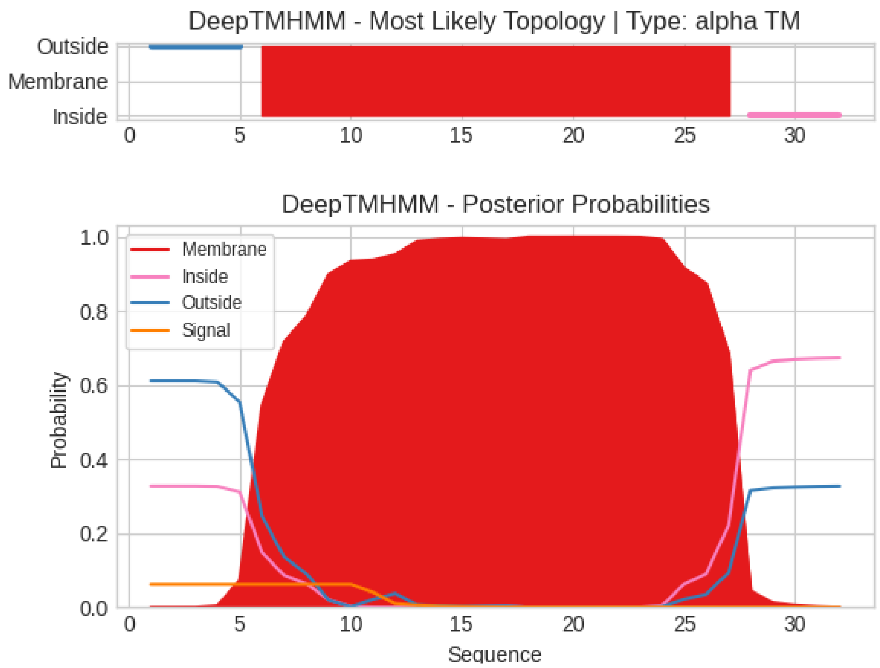
- HHPRED calls this a membrane protein which makes sense since it has a transmembrane domain. However, the probability is less than 90%.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- In Vine which is the only other phage that has an alignment there is no proposed function which could mean it’s a hypothetical protein

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- There is one transmembrane domain for this feature which could make it a membrane protein

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Even though this has what appears to be a large transmembrane domain, we are going to call this a hypothetical protein. We are unsure of the transmembrane domain as it takes up the majority of the sequence.

Feature 41 – reverse – stop 28884

Glimmer/GeneMark

What feature number is this?

- 41 Reverse Gene

What is the stop site?

- 28884

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

- Both Glimmer and GeneMark

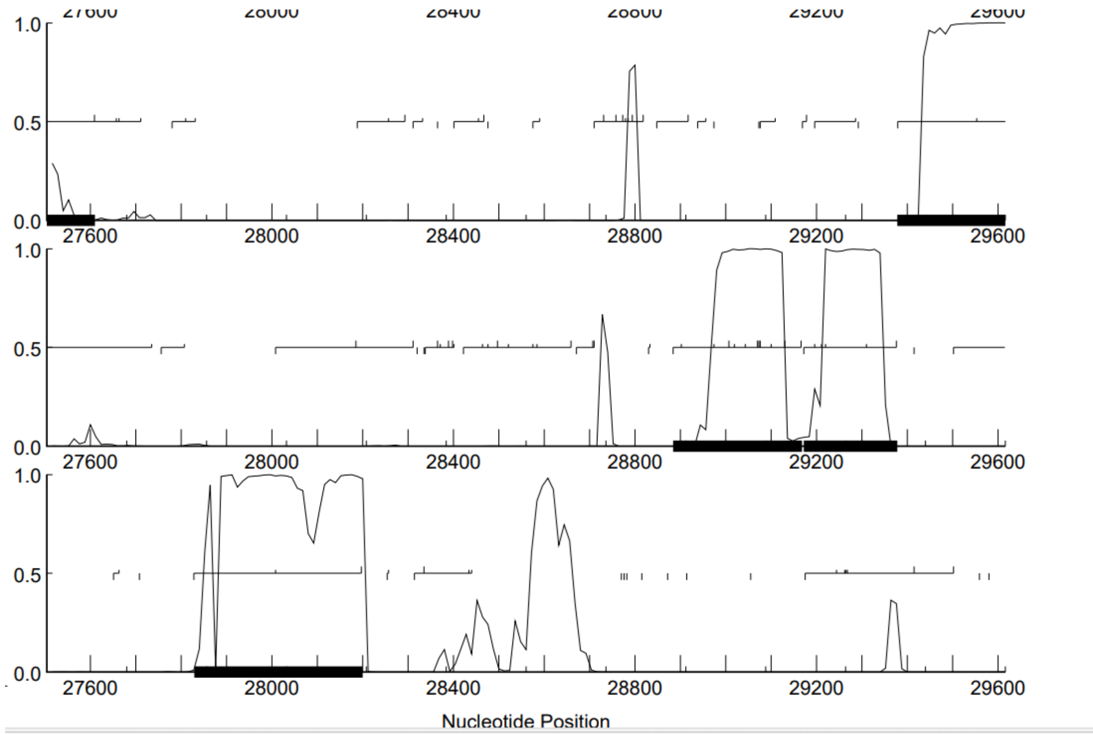
What is the autoannotated start?

- 29168
- Starterator suggested 29102.

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

- 3 gap

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

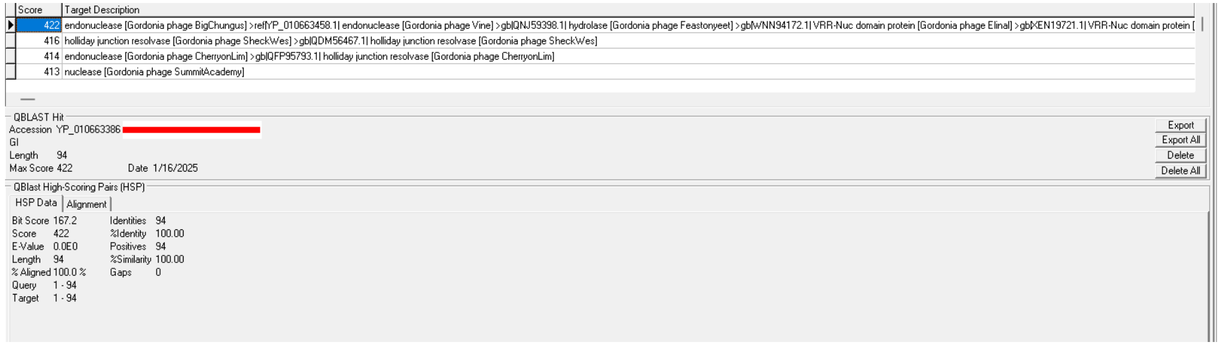


- Coding potential in reverse reading frame 2 is strong.

29102

Coding potential is strong.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
422	endonuclease [Gordonia phage BigChungus] >ref|YP_010663458.1| endonuclease [Gordonia phage Vine] >gb|QNJ59398.1| hydrolase [Gordonia phage Feastonyeet] >gb|WNN94172.1| VRR-Nuc domain protein [Gordonia phage Elinal] >gb|XEN19721.1| VRR-Nuc domain protein [
416	holliday junction resolvase [Gordonia phage SheckWes] >gb|QDM56467.1| holliday junction resolvase [Gordonia phage SheckWes]
414	endonuclease [Gordonia phage CherryonLim] >gb|QFP95793.1| holliday junction resolvase [Gordonia phage CherryonLim]
413	nuclease [Gordonia phage SummitAcademy]

QBLAST Hit

Accession YP_010663386

GI

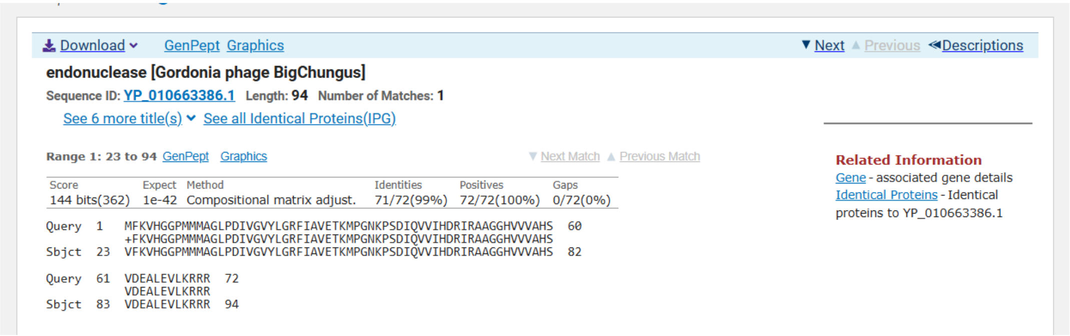
Length 94

Max Score 422 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	167.2	Identities	94
Score	422	%Identity	100.00
E-Value	0.0E0	Positives	94
Length	94	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 94		
Target	1 - 94		

- There are 25 highly similar genes with E value of 0 or less than 1x10-7.



endonuclease [Gordonia phage BigChungus]

Sequence ID: YP_010663386.1 Length: 94 Number of Matches: 1

See 6 more title(s) See all Identical Proteins(IPG)

Range 1: 23 to 94 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
144 bits(362)	1e-42	Compositional matrix adjust.	71/72(99%)	72/72(100%)	0/72(0%)

```
Query  1   MFKVHGGPMMMAGLPDIVGVYLGRFIAVETKMPGNKPSDIQVVIHDRIRAAGGHVVVAHS  60
           +FKVHGGPMMMAGLPDIVGVYLGRFIAVETKMPGNKPSDIQVVIHDRIRAAGGHVVVAHS
Sbjct  23  VFKVHGGPMMMAGLPDIVGVYLGRFIAVETKMPGNKPSDIQVVIHDRIRAAGGHVVVAHS  82

Query  61  VDEALEVLKRRR  72
           VDEALEVLKRRR
Sbjct  83  VDEALEVLKRRR  94
```

Related Information

Gene - associated gene details

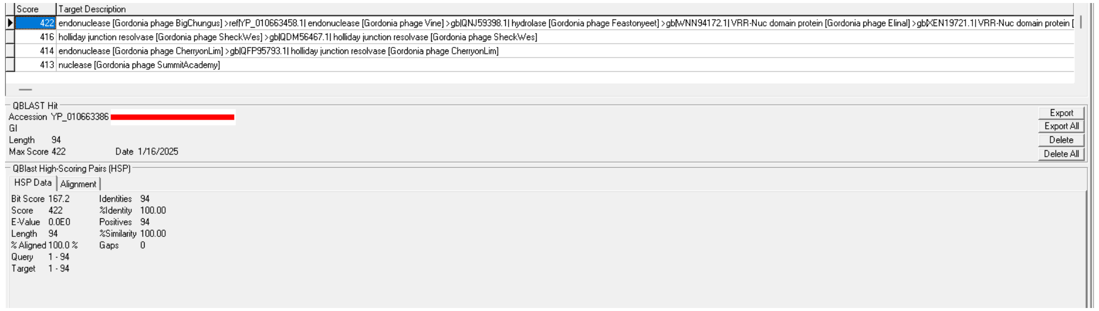
Identical Proteins - Identical proteins to YP_010663386.1

- There are many highly similar genes with E value that’s less than 1x10-7.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- It is a gene because:
- Coding potential is strong.
- Both Glimmer and GeneMark called it a gene.
- There are highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



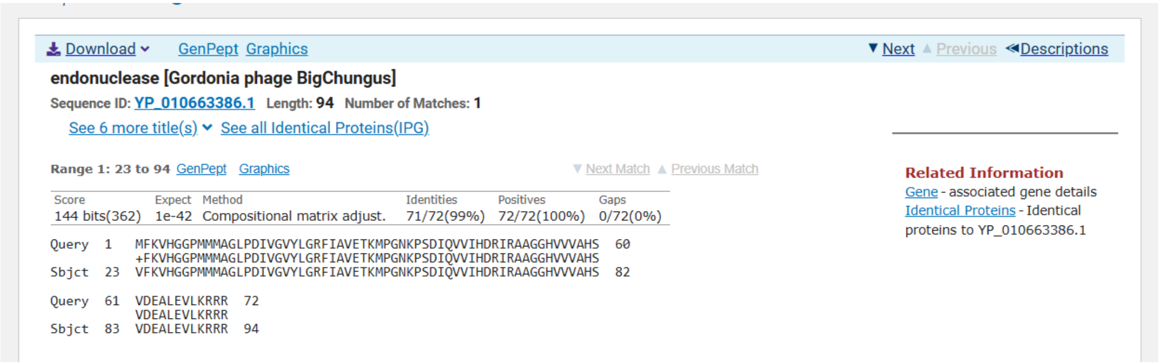
Score	Target Description
422	endonuclease [Gordonia phage BigChungus] >ref|YP_010663458.1| endonuclease [Gordonia phage Vine] >gb|QNJ59398.1| hydrolase [Gordonia phage Feastonyeet] >gb|WNN94172.1| VRR-Nuc domain protein [Gordonia phage Elinal] >gb|XEN19721.1| VRR-Nuc domain protein [
416	holliday junction resolvase [Gordonia phage SheckWes] >gb|QDM56467.1| holliday junction resolvase [Gordonia phage SheckWes]
414	endonuclease [Gordonia phage CherryonLim] >gb|QFP95793.1| holliday junction resolvase [Gordonia phage CherryonLim]
413	nuclease [Gordonia phage SummitAcademy]

QBLAST Hit
Accession YP_010663386
GI
Length 94
Max Score 422 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	167.2	Identities	94
Score	422	%Identity	100.00
E-Value	0.0E0	Positives	94
Length	94	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 94		
Target	1 - 94		

- 23 1:1 alignments.



endonuclease [Gordonia phage BigChungus]
Sequence ID: YP_010663386.1 Length: 94 Number of Matches: 1
See 6 more title(s) See all Identical Proteins(IPG)

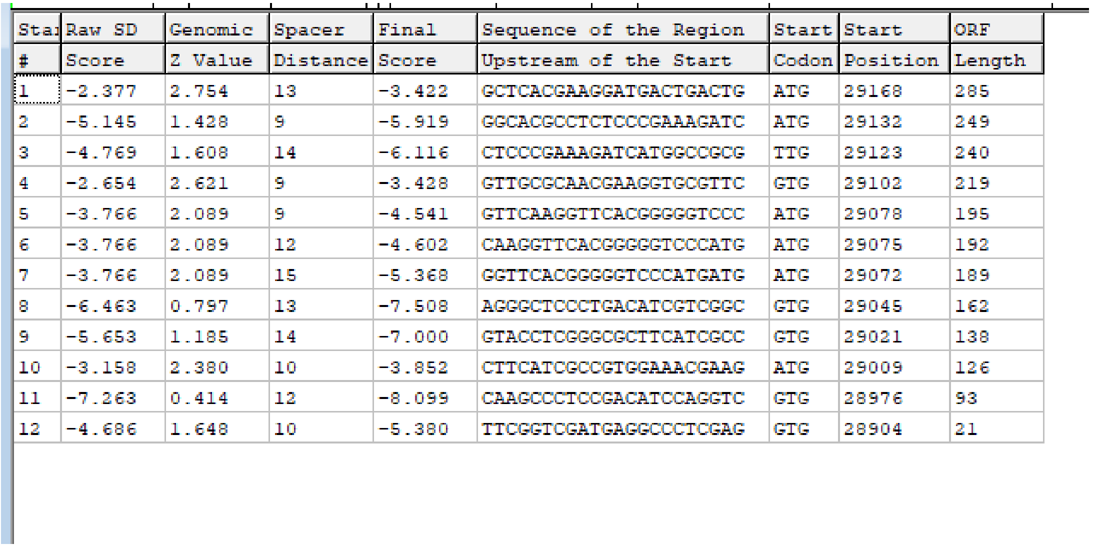
Range 1: 23 to 94 GenPept Graphics

```
Score            Expect  Method                          Identities   Positives     Gaps
144 bits(362)    1e-42   Compositional matrix adjust.    71/72(99%)   72/72(100%)   0/72(0%)

Query  1    MFKVHGGPMMMAGLPDIVGVYLGRFIAVETKMPGNKPSDIQVVIHDRIRAAGGHVVVAHS  60
            +FKVHGGPMMMAGLPDIVGVYLGRFIAVETKMPGNKPSDIQVVIHDRIRAAGGHVVVAHS
Sbjct  23   VFKVHGGPMMMAGLPDIVGVYLGRFIAVETKMPGNKPSDIQVVIHDRIRAAGGHVVVAHS  82

Query  61   VDEALEVLKRRR  72
            VDEALEVLKRRR
Sbjct  83   VDEALEVLKRRR  94
```

Related Information
Gene - associated gene details
Identical Proteins - Identical proteins to YP_010663386.1

- 29102
- One 1:1 alignment

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

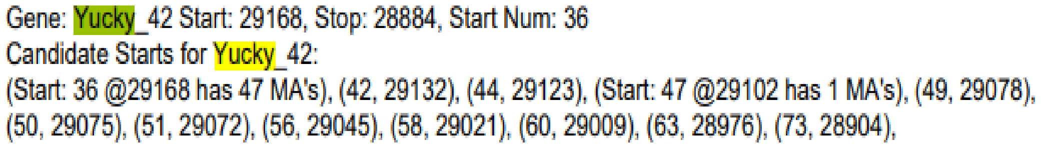


Sta:	Raw SD	Genomic	Spacer	Final	Sequence of the Region	Start	Start	ORF
#	Score	Z Value	Distance	Score	Upstream of the Start	Codon	Position	Length
1	-2.377	2.754	13	-3.422	GCTCACGAAGGATGACTGACTG	ATG	29168	285
2	-5.145	1.428	9	-5.919	GGCACGCCTCTCCCGAAAGATC	ATG	29132	249
3	-4.769	1.608	14	-6.116	CTCCCGAAAGATCATGGCCGCG	TTG	29123	240
4	-2.654	2.621	9	-3.428	GTTGCGCAACGAAGGTGCGTTC	GTG	29102	219
5	-3.766	2.089	9	-4.541	GTTCAAGGTTCACGGGGGTCCC	ATG	29078	195
6	-3.766	2.089	12	-4.602	CAAGGTTCACGGGGGTCCCATG	ATG	29075	192
7	-3.766	2.089	15	-5.368	GGTTCACGGGGGTCCCATGATG	ATG	29072	189
8	-6.463	0.797	13	-7.508	AGGGCTCCCTGACATCGTCGGC	GTG	29045	162
9	-5.653	1.185	14	-7.000	GTACCTCGGGCGCTTCATCGCC	GTG	29021	138
10	-3.158	2.380	10	-3.852	CTTCATCGCCGTGGAAACGAAG	ATG	29009	126
11	-7.263	0.414	12	-8.099	CAAGCCCTCCGACATCCAGGTC	GTG	28976	93
12	-4.686	1.648	10	-5.380	TTCGGTCGATGAGGCCCTCGAG	GTG	28904	21

- 29168
- Z value: 2.754 (Greatest)
- Final Score: -3.422 (Least negative)

- 29102
- Z value: 2.621
- Final socre: -3.428

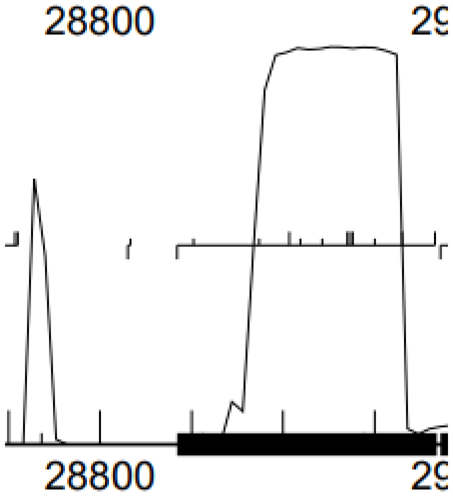
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



Gene: Yucky_42 Start: 29168, Stop: 28884, Start Num: 36
Candidate Starts for Yucky_42:
(Start: 36 @29168 has 47 MA's), (42, 29132), (44, 29123), (Start: 47 @29102 has 1 MA's), (49, 29078), (50, 29075), (51, 29072), (56, 29045), (58, 29021), (60, 29009), (63, 28976), (73, 28904),

- 47 MA’s
- But there are also 1 MA at 29102.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

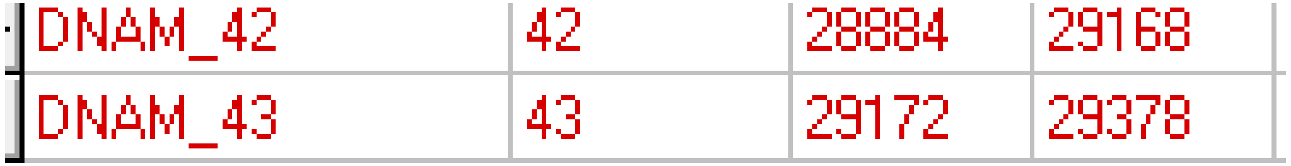


- Start site 29168 includes all coding potential.

- 29102:
- Coding potential is cut off.

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 29172-29168 = 4
- 4-1 = 3 gap



DNAM_42	42	28884	29168
DNAM_43	43	29172	29378

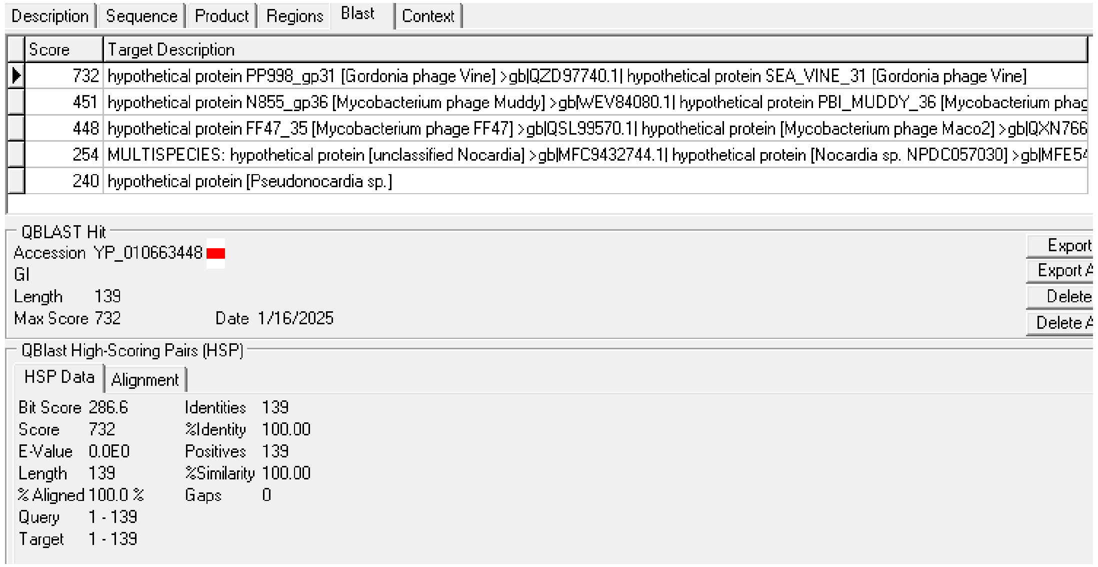
- 29102:
- 29172-29102 = 70
- 70-1=69 gap

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	29168	29102
GeneMark	Both Glimmer and GeneMark	NA
Coding potential	Included	Cut off
RBS	Z score: 2.754 Final Score:-3.422	Z score: 2.621 Final Score: -3.428
Blast	23 1:1 alignments	1 1:1 alignment
Starterator	47	1
Gap/overlap	3 gap	69 gap

All evidences support that the start site is at the nucleotide number 29168. Both Glimmer and GeneMark agree the start site. Coding potential is included as well. RBS score, number of alignment and the number of manual annotation support 29168 as a start site. 3 gap is also better than 69 gap.

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
732	hypothetical protein PP998_gp31 [Gordonia phage Vine] >gb|QZD97740.1| hypothetical protein SEA_VINE_31 [Gordonia phage Vine]
451	hypothetical protein N855_gp36 [Mycobacterium phage Muddy] >gb|WEV84080.1| hypothetical protein PBI_MUDDY_36 [Mycobacterium phag
448	hypothetical protein FF47_35 [Mycobacterium phage FF47] >gb|QSL99570.1| hypothetical protein [Mycobacterium phage Maco2] >gb|QXN766
254	MULTISPECIES: hypothetical protein [unclassified Nocardia] >gb|MFC9432744.1| hypothetical protein [Nocardia sp. NPDC057030] >gb|MFE54
240	hypothetical protein [Pseudonocardia sp.]

QBLAST Hit

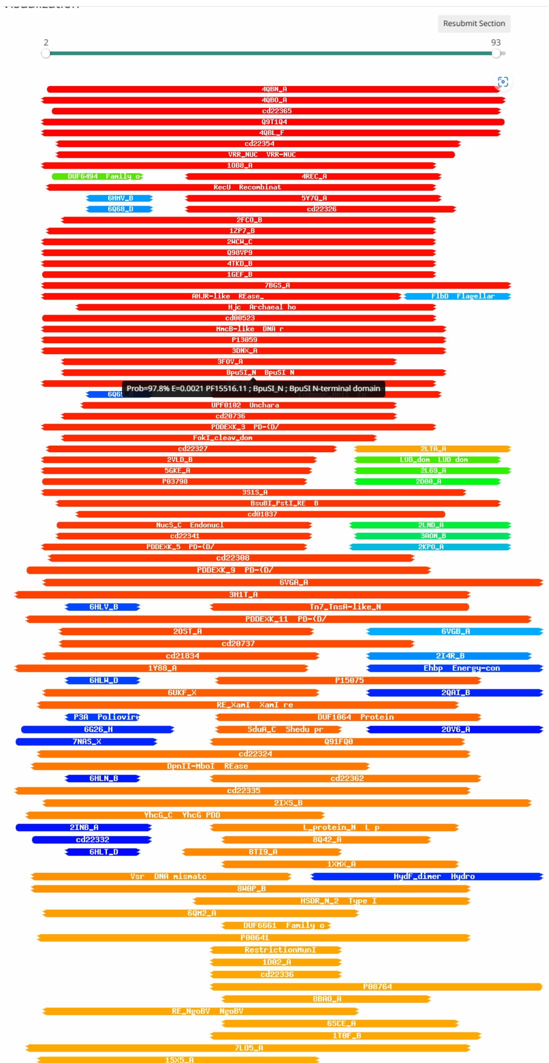
Accession YP_010663448 🟥
GI
Length 139
Max Score 732 Date 1/16/2025

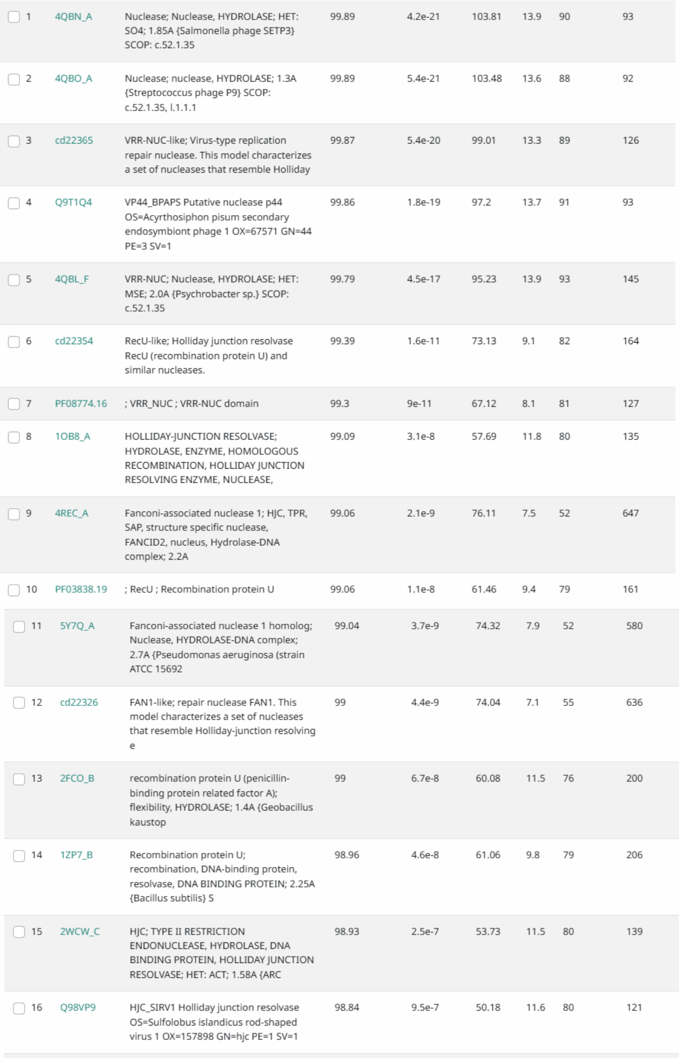
QBlast High-Scoring Pairs (HSP)

Bit Score	286.6	Identities	139
Score	732	%Identity	100.00
E-Value	0.0E0	Positives	139
Length	139	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 139		
Target	1 - 139		

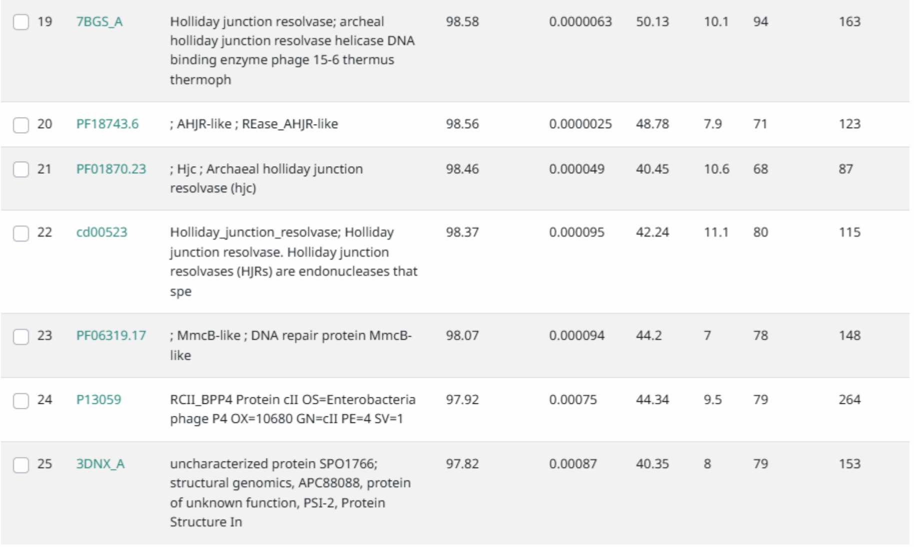
- BLAST call it an endonuclease(Vine), holliday junction resolvase(SheckWes), nuclease(SummitAcademy), hydrolase(Feastonyeet), VRR-Nuc domain protein(Elinal), and a hypothetical protein (Axym).

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.





□	1	4QBN_A	Nuclease; Nuclease, HYDROLASE; HET: SO4; 1.85A {Salmonella phage SETP3} SCOP: c.52.1.35	99.89	4.2e-21	103.81	13.9	90	93
□	2	4QBO_A	Nuclease; nuclease, HYDROLASE; 1.3A {Streptococcus phage P9} SCOP: c.52.1.35, l.1.1.1	99.89	5.4e-21	103.48	13.6	88	92
□	3	cd22365	VRR-NUC-like; Virus-type replication repair nuclease. This model characterizes a set of nucleases that resemble Holliday	99.87	5.4e-20	99.01	13.3	89	126
□	4	Q9T1Q4	VP44_BPAPS Putative nuclease p44 OS=Acyrthosiphon pisum secondary endosymbiont phage 1 OX=67571 GN=44 PE=3 SV=1	99.86	1.8e-19	97.2	13.7	91	93
□	5	4QBL_F	VRR-NUC; Nuclease, HYDROLASE; HET: MSE; 2.0A {Psychrobacter sp.} SCOP: c.52.1.35	99.79	4.5e-17	95.23	13.9	93	145
□	6	cd22354	RecU-like; Holliday junction resolvase RecU (recombination protein U) and similar nucleases.	99.39	1.6e-11	73.13	9.1	82	164
□	7	PF08774.16	; VRR_NUC ; VRR-NUC domain	99.3	9e-11	67.12	8.1	81	127
□	8	1OB8_A	HOLLIDAY-JUNCTION RESOLVASE; HYDROLASE, ENZYME, HOMOLOGOUS RECOMBINATION, HOLLIDAY JUNCTION RESOLVING ENZYME, NUCLEASE,	99.09	3.1e-8	57.69	11.8	80	135
□	9	4REC_A	Fanconi-associated nuclease 1; HJC, TPR, SAP, structure specific nuclease, FANCID2, nucleus, Hydrolase-DNA complex; 2.2A	99.06	2.1e-9	76.11	7.5	52	647
□	10	PF03838.19	; RecU ; Recombination protein U	99.06	1.1e-8	61.46	9.4	79	161
□	11	5Y7Q_A	Fanconi-associated nuclease 1 homolog; Nuclease, HYDROLASE-DNA complex; 2.7A {Pseudomonas aeruginosa (strain ATCC 15692	99.04	3.7e-9	74.32	7.9	52	580
□	12	cd22326	FAN1-like; repair nuclease FAN1. This model characterizes a set of nucleases that resemble Holliday-junction resolving e	99	4.4e-9	74.04	7.1	55	636
□	13	2FCO_B	recombination protein U (penicillin-binding protein related factor A); flexibility, HYDROLASE; 1.4A {Geobacillus kaustop	99	6.7e-8	60.08	11.5	76	200
□	14	1ZP7_B	Recombination protein U; recombination, DNA-binding protein, resolvase, DNA BINDING PROTEIN; 2.25A {Bacillus subtilis} S	98.96	4.6e-8	61.06	9.8	79	206
□	15	2WCW_C	HJC; TYPE II RESTRICTION ENDONUCLEASE, HYDROLASE, DNA BINDING PROTEIN, HOLLIDAY JUNCTION RESOLVASE; HET: ACT; 1.58A {ARC	98.93	2.5e-7	53.73	11.5	80	139
□	16	Q98VP9	HJC_SIRV1 Holliday junction resolvase OS=Sulfolobus islandicus rod-shaped virus 1 OX=157898 GN=hjc PE=1 SV=1	98.84	9.5e-7	50.18	11.6	80	121



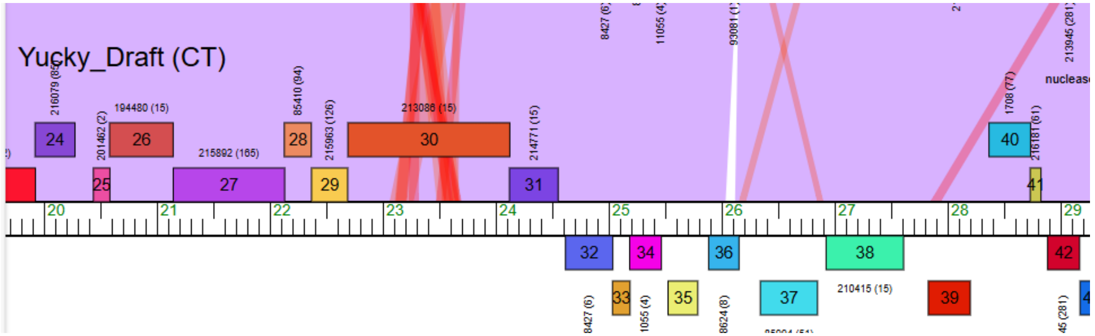
□	19	7BGS_A	Holliday junction resolvase; archeal holliday junction resolvase helicase DNA binding enzyme phage 15-6 thermus thermoph	98.58	0.0000063	50.13	10.1	94	163
□	20	PF18743.6	; AHJR-like ; REase_AHJR-like	98.56	0.0000025	48.78	7.9	71	123
□	21	PF01870.23	; Hjc ; Archaeal holliday junction resolvase (hjc)	98.46	0.000049	40.45	10.6	68	87
□	22	cd00523	Holliday_junction_resolvase; Holliday junction resolvase. Holliday junction resolvases (HJRs) are endonucleases that spe	98.37	0.000095	42.24	11.1	80	115
□	23	PF06319.17	; MmcB-like ; DNA repair protein MmcB-like	98.07	0.000094	44.2	7	78	148
□	24	P13059	RCII_BPP4 Protein cII OS=Enterobacteria phage P4 OX=10680 GN=cII PE=4 SV=1	97.92	0.00075	44.34	9.5	79	264
□	25	3DNX_A	uncharacterized protein SPO1766; structural genomics, APC88088, protein of unknown function, PSI-2, Protein Structure In	97.82	0.00087	40.35	8	79	153

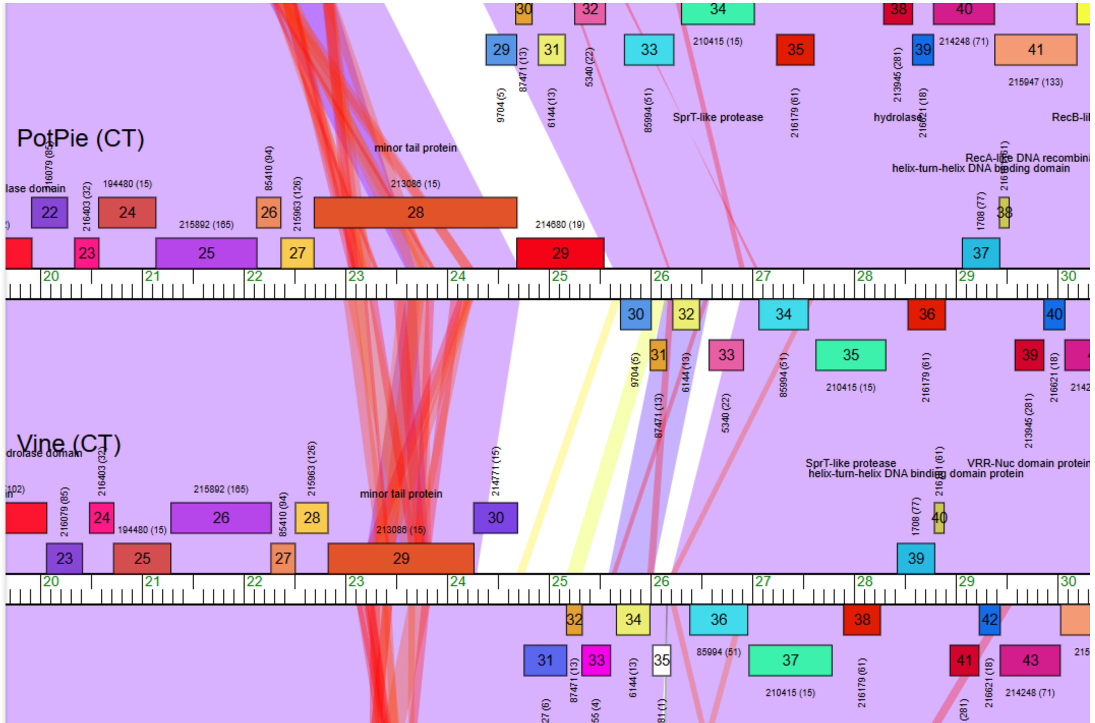
There are many hits with nuclease

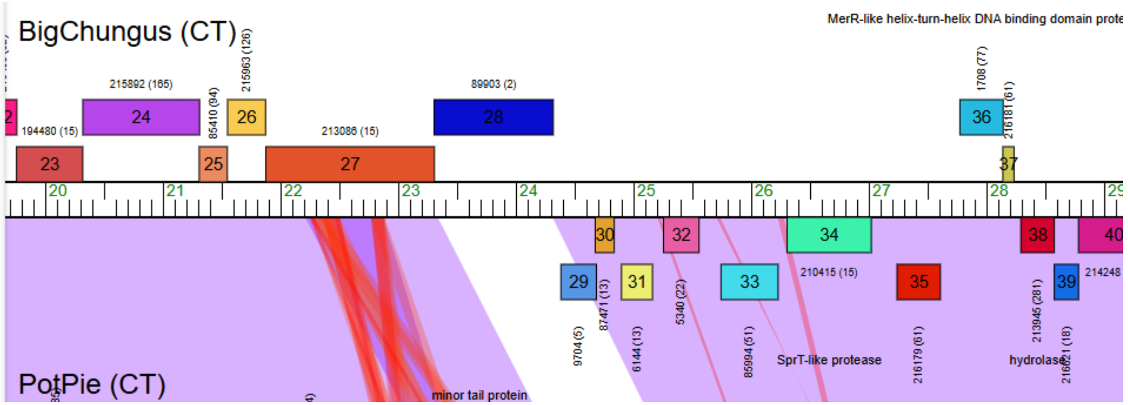
There are also many hits with Holliday Junction

There are some hits with VRR-Nuc

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?







Same genes in the same pham call it differently.

PotPie – VRR-Nuc domain protein

Vine – nuclease

BigChungus – hydrolase

It shares three conserved domains with PotPie (2 VRR-Nuc and 1 VRR Nuc like).

It shares two conserved domains with Vine and BigChungus.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- Some functions are given.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I will call it a VRR-Nuc protein because

- It is one of suggestion from BLAST
- There are some hits in Hhpred even though other functions were suggested more.
- Phamerator show that Potpie’s gene is VRR-Nuc protein, and it shares the most conserved domain with gene 42 of Yucky (This was the strongest evidence that I considered).

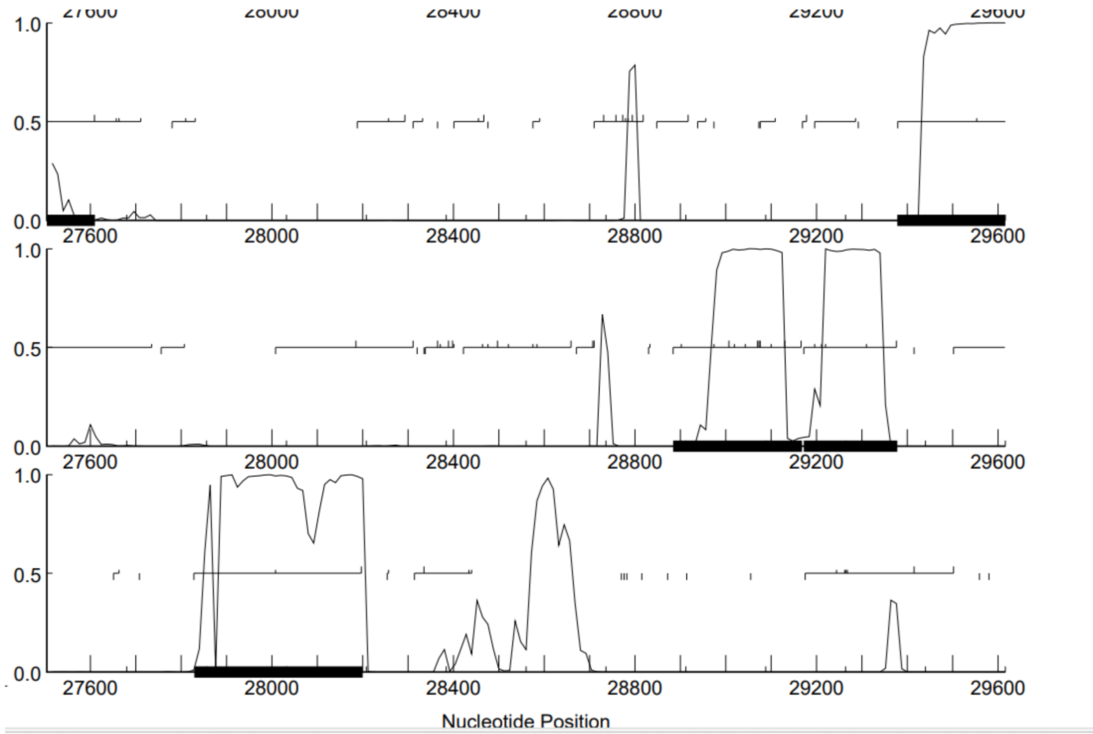
Feature 42 – reverse – stop 29172

Glimmer/GeneMark

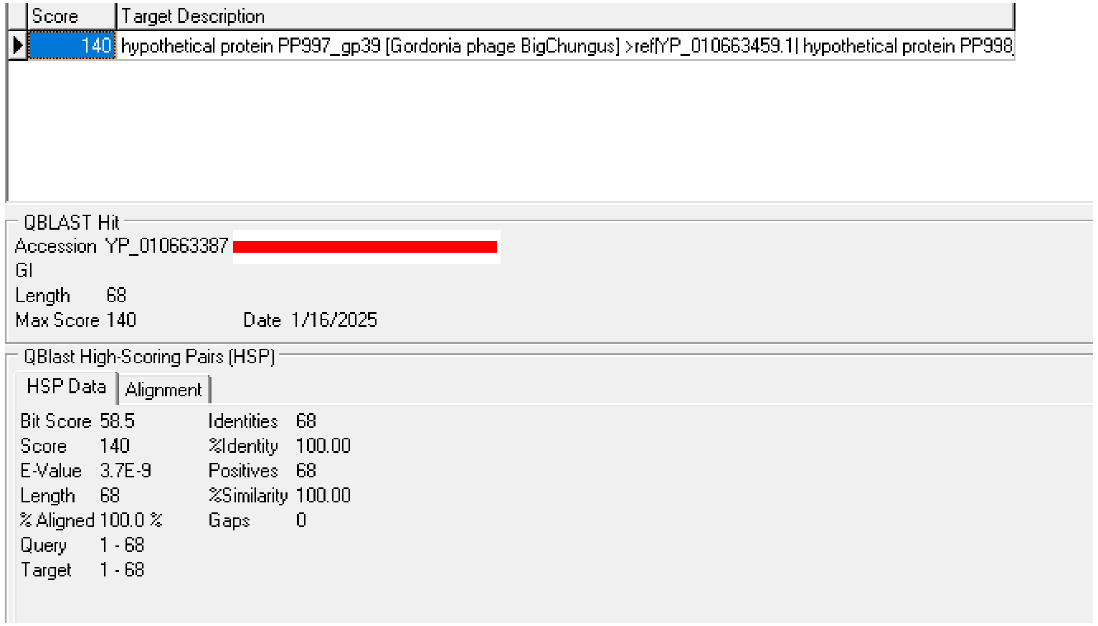
What feature number is this?	42 reverse gene
What is the stop site?	29172
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?	Both Glimmer and GeneMark
What is the autoannotated start?	29378
Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start	1 overlap

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

Strong coding potential



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
140	hypothetical protein PP997_gp39 [Gordonia phage BigChungus] >ref|YP_010663459.1| hypothetical protein PP998

QBLAST Hit

Accession YP_010663387 [redacted]

GI

Length 68

Max Score 140 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

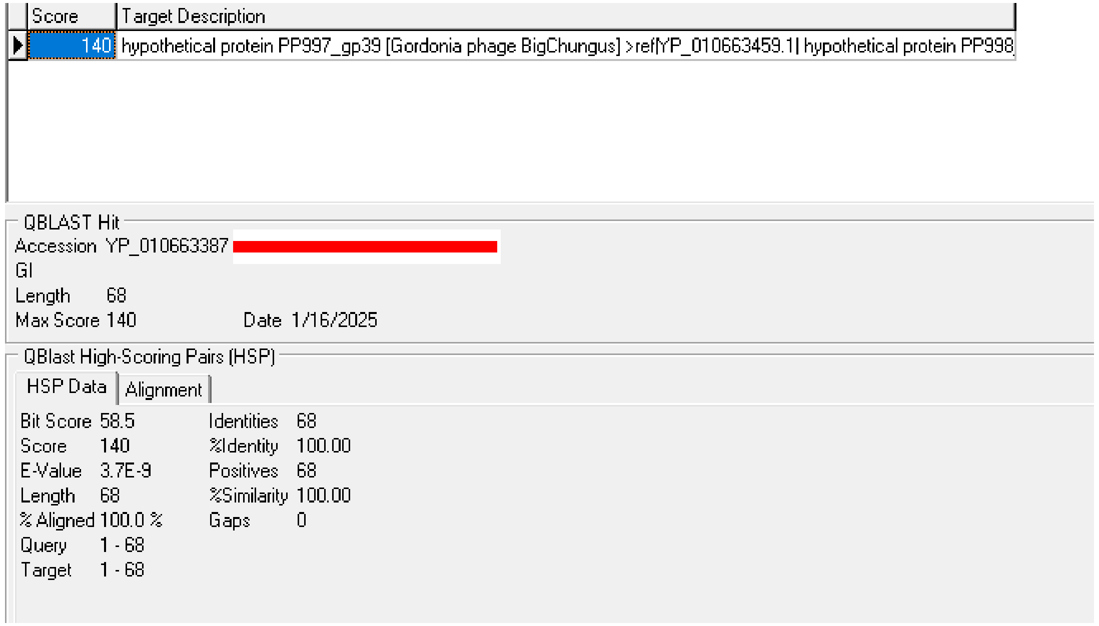
Bit Score	58.5	Identities	68
Score	140	%Identity	100.00
E-Value	3.7E-9	Positives	68
Length	68	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 68		
Target	1 - 68		

- There is only one highly similar gene with an E value of close to 0 (BigChungus).

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- It is a gene because:
- The coding potential is strong.
- There is one highly similar gene with an E value of close to 0.
- Both Glimmer and GeneMark called it a gene.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
140	hypothetical protein PP997_gp39 [Gordonia phage BigChungus] >ref|YP_010663459.1| hypothetical protein PP998

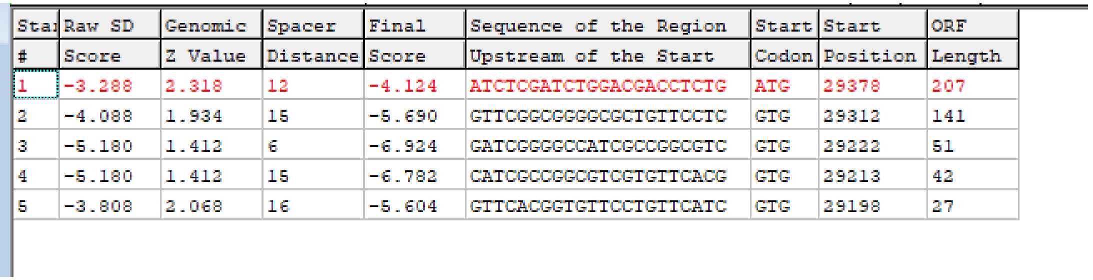
QBLAST Hit
Accession YP_010663387
GI
Length 68
Max Score 140 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	58.5	Identities	68
Score	140	%Identity	100.00
E-Value	3.7E-9	Positives	68
Length	68	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 68		
Target	1 - 68		

- There is only one 1:1 alignment (BigChungus).

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-3.288	2.318	12	-4.124	ATCTCGATCTGGACGACCTCTG	ATG	29378	207
2	-4.088	1.934	15	-5.690	GTTCGGCGGGGCGCTGTTCCTC	GTG	29312	141
3	-5.180	1.412	6	-6.924	GATCGGGGCCATCGCCGGCGTC	GTG	29222	51
4	-5.180	1.412	15	-6.782	CATCGCCGGCGTCGTGTTCACG	GTG	29213	42
5	-3.808	2.068	16	-5.604	GTTCACGGTGTTCCTGTTCATC	GTG	29198	27

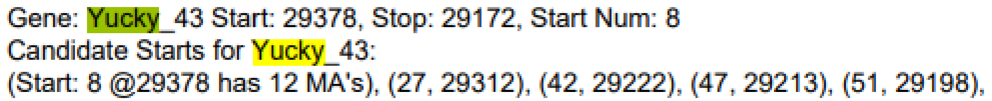
Z value: 2.318

Final score: -4.124

It is favored because the Z value is the greatest and the final score is least negative.

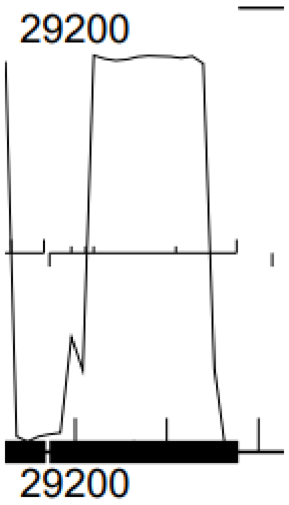
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- There are 12 MA’s.



Gene: Yucky_43 Start: 29378, Stop: 29172, Start Num: 8
Candidate Starts for Yucky_43:
(Start: 8 @29378 has 12 MA's), (27, 29312), (42, 29222), (47, 29213), (51, 29198),

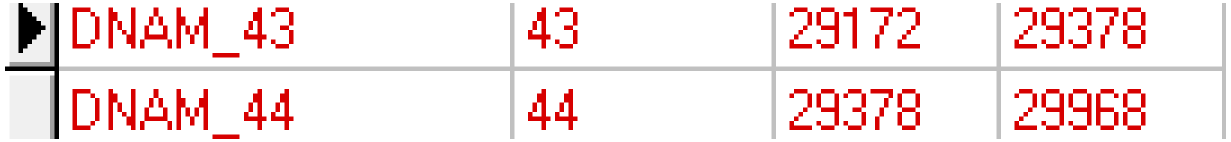
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- Coding potential is included at between the start site and stop site of feature 43.

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 29388-29388 = 0
- 0+1=1 overlap



DNAM_43	43	29172	29378
DNAM_44	44	29378	29968

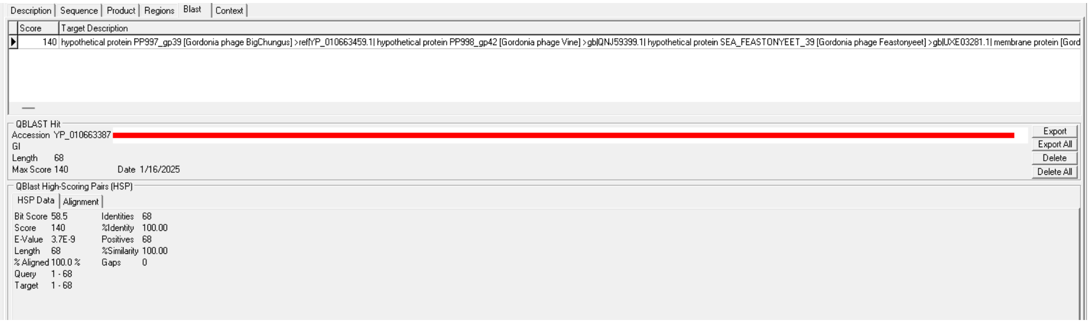
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	29378
GeneMark	Both Glimmer and GeneMark
Coding potential	Included
RBS	Z value: 2.314 Final score: -4.124
Blast	1
Starterator	12
Gap/overlap	1

29378 is a start site because all factors support it. Especially, gap of one is favored. Though, only one 1:1 alignment does not support strongly.

BLAST function evidence. What assigned functions do other highly similar genes have?

- There is only one highly gene.
- It is a hypothetical protein.

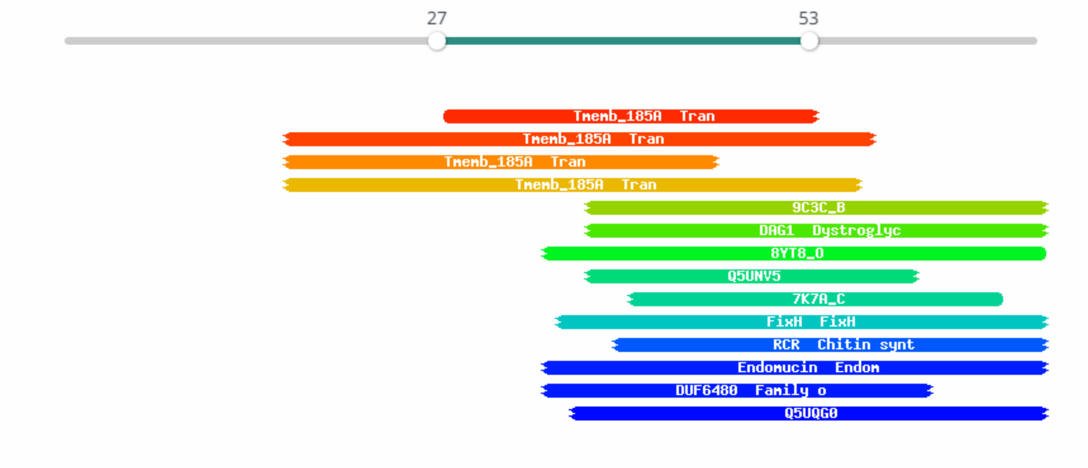


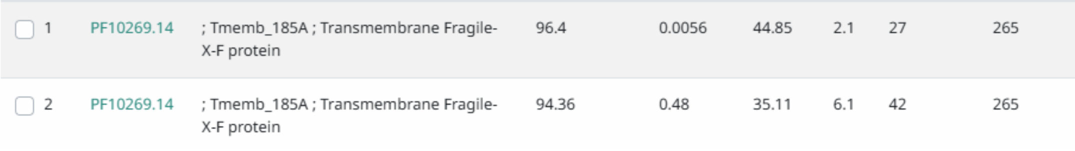
Score	Target Description
140	hypothetical protein PP997_gp39 [Gordonia phage BigChungus] >ref|YP_010663459.1| hypothetical protein PP998_gp42 [Gordonia phage Vine] >gb|QNJ59399.1| hypothetical protein SEA_FEASTONYEET_39 [Gordonia phage Feastonyeet] >gb|UXE03281.1| membrane protein [Gord

```
QBLAST Hit
Accession  YP_010663387
GI
Length     68
Max Score 140            Date  1/16/2025
```

```
QBlast High-Scoring Pairs (HSP)
HSP Data | Alignment
Bit Score  58.5       Identities  68
Score      140        %Identity   100.00
E-Value    3.7E-9     Positives   68
Length     68         %Similarity 100.00
% Aligned  100.0 %    Gaps        0
Query      1 - 68
Target     1 - 68
```

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.





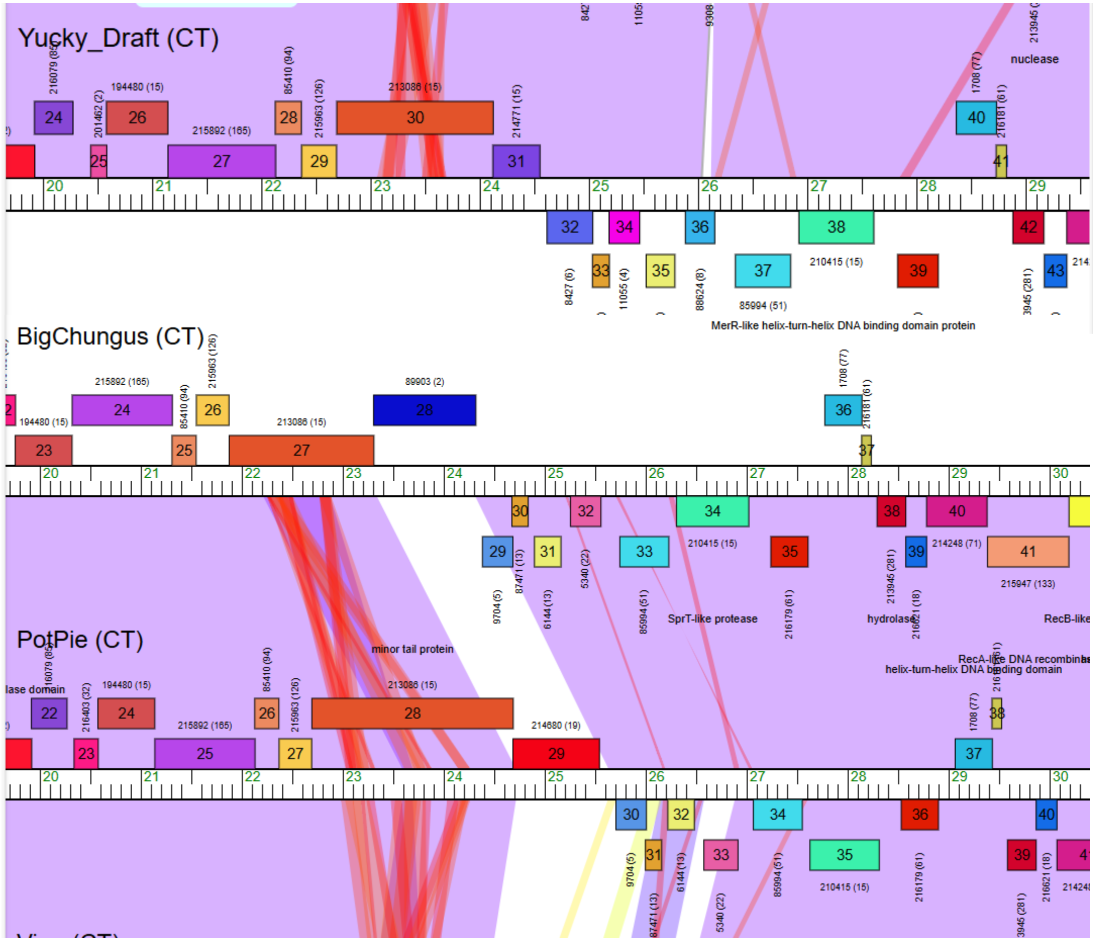
□	1	PF10269.14	; Tmemb_185A ; Transmembrane Fragile-X-F protein	96.4	0.0056	44.85	2.1	27	265
□	2	PF10269.14	; Tmemb_185A ; Transmembrane Fragile-X-F protein	94.36	0.48	35.11	6.1	42	265

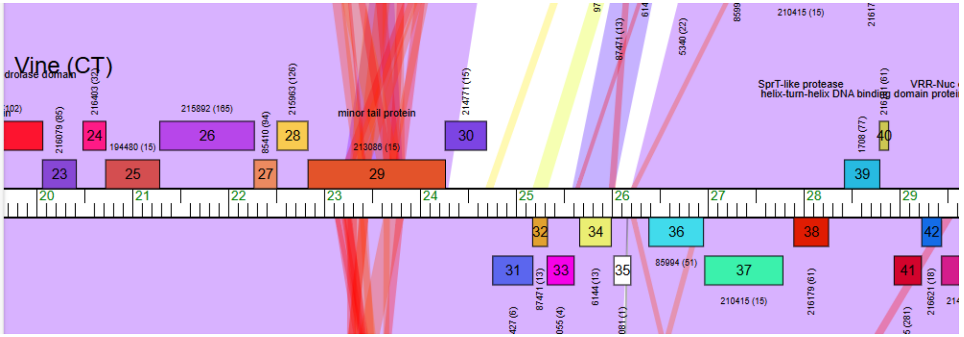
There are 2 hits with probability greater than 90.

Both call it a transmembrane protein.

Though there are no functions called transmembrane.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

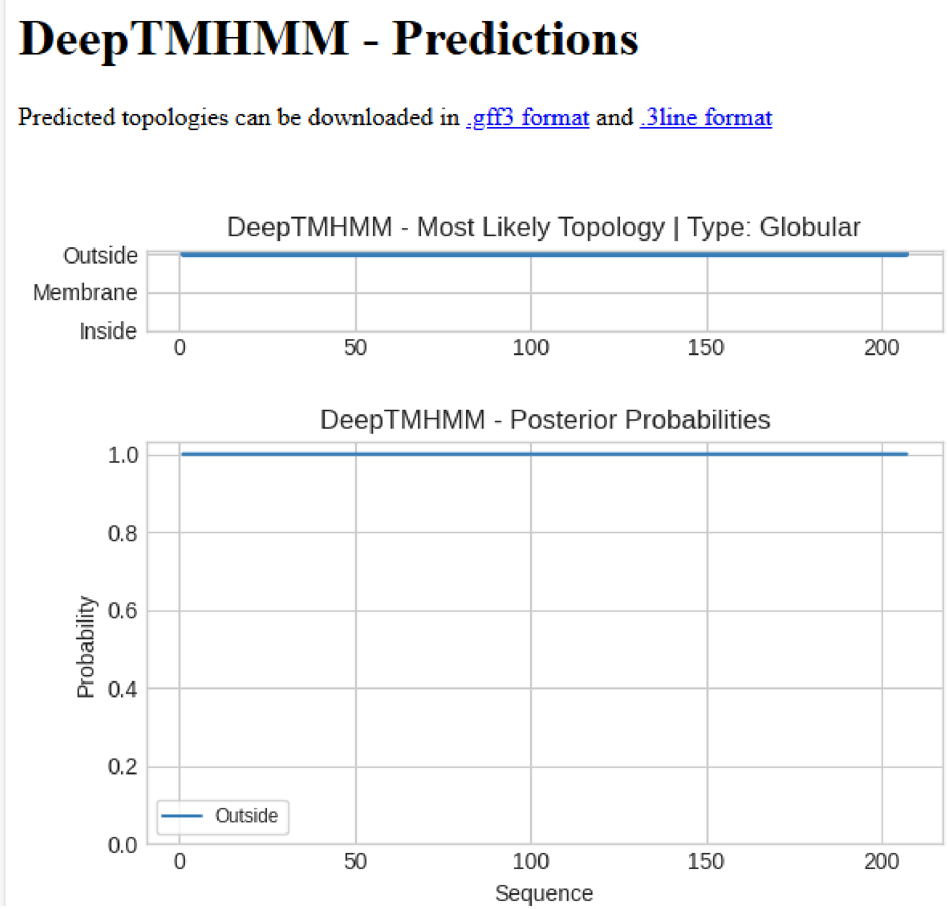




BigChungus – hypothetical protein
PotPie – hypothetical protein
Vine – hypothetical protein

There are no conserved domains.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

- I was not sure about transmembrane protein even though it was suggested by Hhpred.

- So I looked at Deep TMHMM.
- The graph shows a horizontal line on the Outside axis, meaning we cannot know its function.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I decided to call it a hypothetical protein because.

- BLAST call it a hypothetical protein.
- Hhpred gives function that does not exist in the official function list.
- Phamerator show that same genes in the same pham do not have functions as well.
- Deep THMHH gave a graph with a horizontal line on the outside axis.

Feature 43 – Reverse – Stop 29378

Glimmer/GeneMark

What feature number is this? 43

What is the stop site? 29378

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

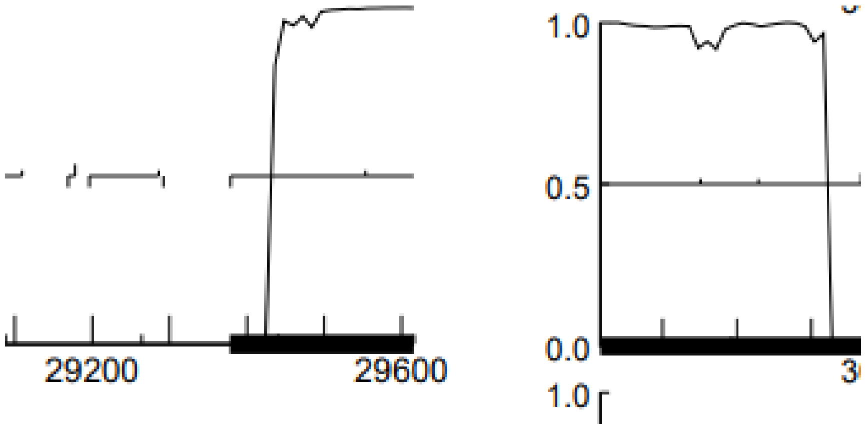
Called by both Glimmer and GeneMark

What is the autoannotated start?

29968

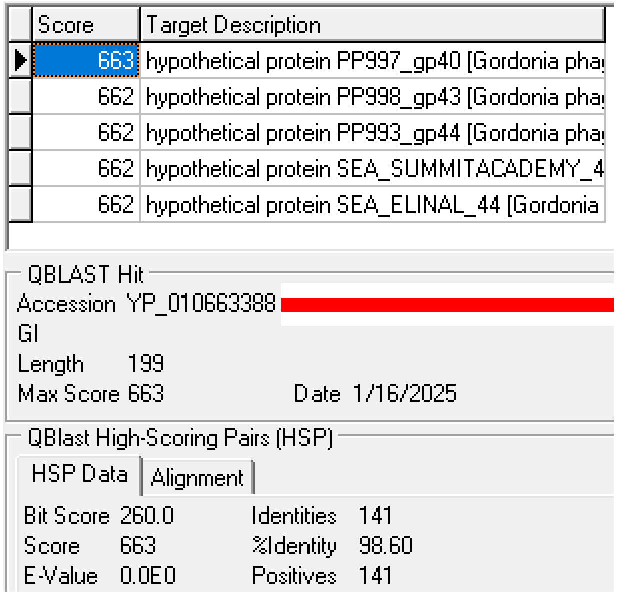
Gap: ___3_____ or overlap: _______ (with gene in front of it) for the autoannotated start

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- There is a strong peak of coding potential that persists throughout the entirety of the features sequence. Reading frame 4 is the only frame with coding potential.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
663	hypothetical protein PP997_gp40 [Gordonia pha
662	hypothetical protein PP998_gp43 [Gordonia pha
662	hypothetical protein PP993_gp44 [Gordonia pha
662	hypothetical protein SEA_SUMMITACADEMY_4
662	hypothetical protein SEA_ELINAL_44 [Gordonia

QBLAST Hit

Accession YP_010663388

GI

Length 199

Max Score 663 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 260.0 Identities 141

Score 663 %Identity 98.60

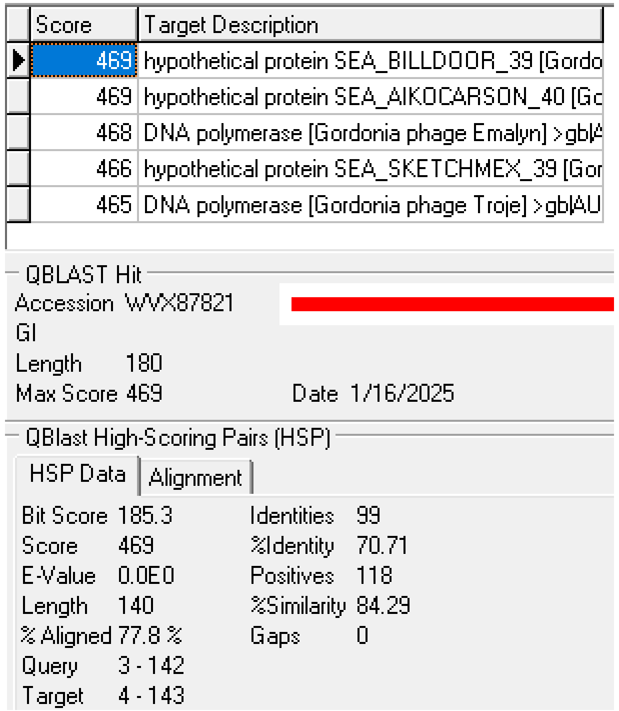
E-Value 0.0E0 Positives 141

- There are 25 BLAST hits with E-values close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This is a gene. It is called by both Glimmer and GeneMark and it has a very strong peak of coding potential throughout the feature sequence. Also, BLAST shows 25 highly similar phages with an E-value close to 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
469	hypothetical protein SEA_BILLDOOR_39 [Gordo
469	hypothetical protein SEA_AIKOCARSON_40 [Gc
468	DNA polymerase [Gordonia phage Emalyn] >gb|A
466	hypothetical protein SEA_SKETCHMEX_39 [Gor
465	DNA polymerase [Gordonia phage Troje] >gb|AU

QBLAST Hit

Accession WVX87821

GI

Length 180

Max Score 469

Date 1/16/2025

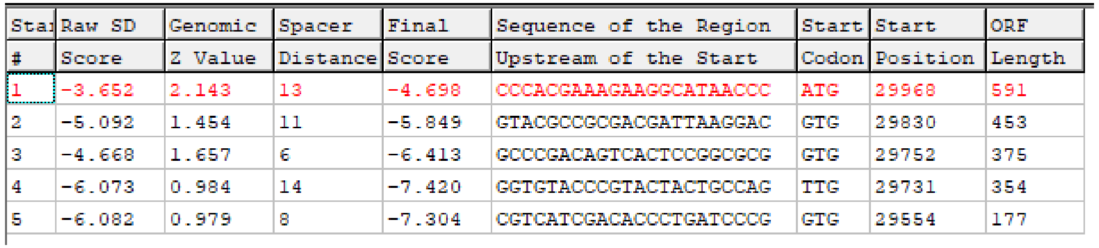
QBlast High-Scoring Pairs (HSP)

HSP Data | Alignment

Bit Score	185.3	Identities	99
Score	469	%Identity	70.71
E-Value	0.0E0	Positives	118
Length	140	%Similarity	84.29
% Aligned	77.8 %	Gaps	0
Query	3 - 142		
Target	4 - 143		

- There are 21 1:1 alignments and 4 3:4 alignments. There are no known alternate starts yet.

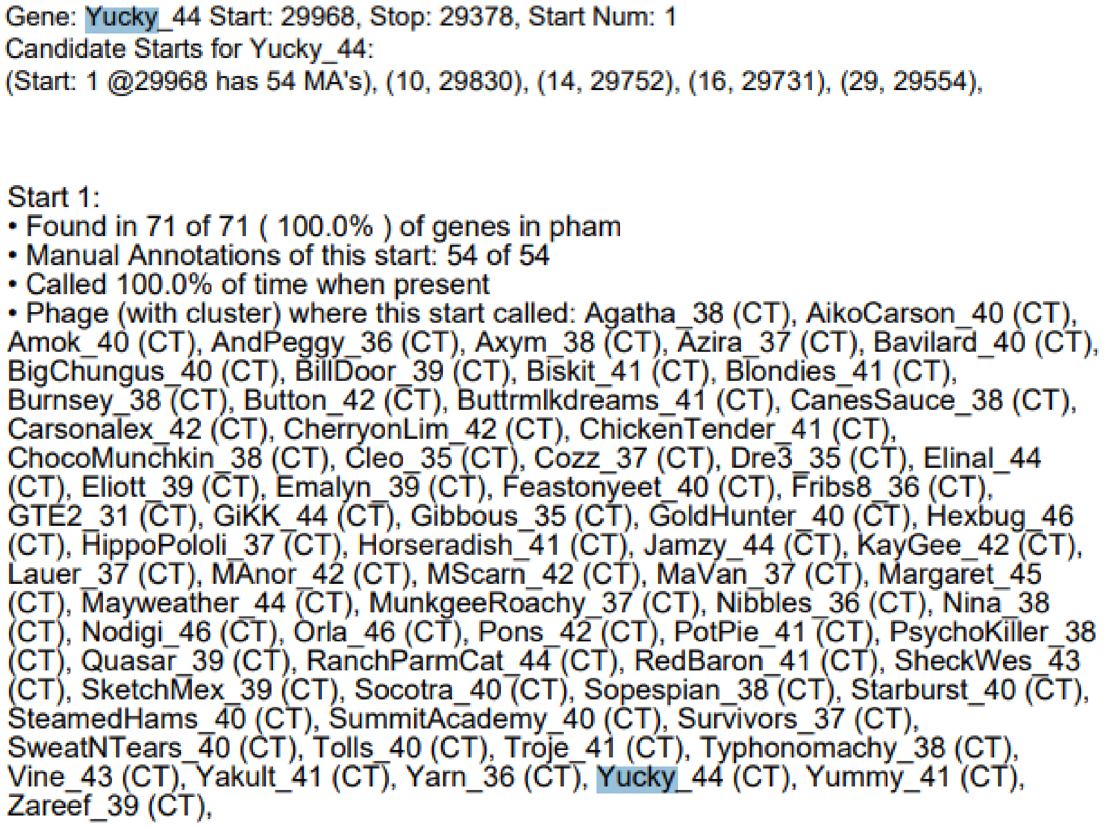
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-3.652	2.143	13	-4.698	CCCACGAAAGAAGGCATAACCC	ATG	29968	591
2	-5.092	1.454	11	-5.849	GTACGCCGCGACGATTAAGGAC	GTG	29830	453
3	-4.668	1.657	6	-6.413	GCCCGACAGTCACTCCGGCGCG	GTG	29752	375
4	-6.073	0.984	14	-7.420	GGTGTACCCGTACTACTGCCAG	TTG	29731	354
5	-6.082	0.979	8	-7.304	CGTCATCGACACCCTGATCCCG	GTG	29554	177

- The z-value is 2.143 and the final score is -4.698. These are the only good RBS numbers.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

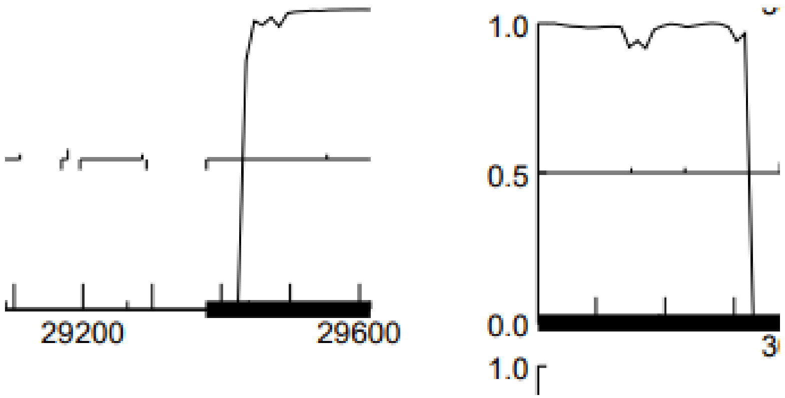


Gene: Yucky_44 Start: 29968, Stop: 29378, Start Num: 1
Candidate Starts for Yucky_44:
(Start: 1 @29968 has 54 MA's), (10, 29830), (14, 29752), (16, 29731), (29, 29554),

Start 1:
• Found in 71 of 71 (100.0%) of genes in pham
• Manual Annotations of this start: 54 of 54
• Called 100.0% of time when present
• Phage (with cluster) where this start called: Agatha_38 (CT), AikoCarson_40 (CT), Amok_40 (CT), AndPeggy_36 (CT), Axym_38 (CT), Azira_37 (CT), Bavilard_40 (CT), BigChungus_40 (CT), BillDoor_39 (CT), Biskit_41 (CT), Blondies_41 (CT), Burnsey_38 (CT), Button_42 (CT), Buttrmlkdreams_41 (CT), CanesSauce_38 (CT), Carsonalex_42 (CT), CherryonLim_42 (CT), ChickenTender_41 (CT), ChocoMunchkin_38 (CT), Cleo_35 (CT), Cozz_37 (CT), Dre3_35 (CT), Elinal_44 (CT), Eliott_39 (CT), Emalyn_39 (CT), Feastonyeet_40 (CT), Fribs8_36 (CT), GTE2_31 (CT), GiKK_44 (CT), Gibbous_35 (CT), GoldHunter_40 (CT), Hexbug_46 (CT), HippoPololi_37 (CT), Horseradish_41 (CT), Jamzy_44 (CT), KayGee_42 (CT), Lauer_37 (CT), MAnor_42 (CT), MScarn_42 (CT), MaVan_37 (CT), Margaret_45 (CT), Mayweather_44 (CT), MunkgeeRoachy_37 (CT), Nibbles_36 (CT), Nina_38 (CT), Nodigi_46 (CT), Orla_46 (CT), Pons_42 (CT), PotPie_41 (CT), PsychoKiller_38 (CT), Quasar_39 (CT), RanchParmCat_44 (CT), RedBaron_41 (CT), SheckWes_43 (CT), SketchMex_39 (CT), Socotra_40 (CT), Sopespian_38 (CT), Starburst_40 (CT), SteamedHams_40 (CT), SummitAcademy_40 (CT), Survivors_37 (CT), SweatNTears_40 (CT), Tolls_40 (CT), Troje_41 (CT), Typhonomachy_38 (CT), Vine_43 (CT), Yakult_41 (CT), Yarn_36 (CT), Yucky_44 (CT), Yummy_41 (CT), Zareef_39 (CT),

- This start site is found within 100% of the genes in the Pham and is called the manually annotated start 100% of the time when present. The autoannotated start has 54 MA’s, no other site has ever received an MA

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- The start site does not cut off any coding potential.

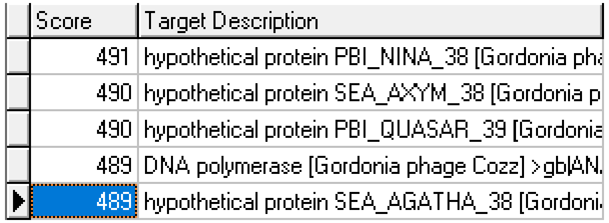
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 29972-29968= 4-1 for gap=3

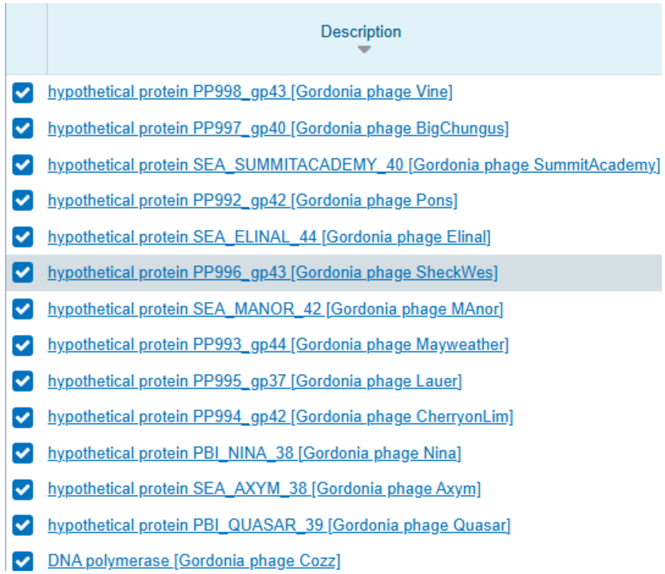
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is the manually annotated start of 29968. There are 21 1:1 BLAST alignments with other highly similar phages. It is the only start with acceptable RBS numbers. It is called 100% of the time when present and it is the only start site to ever receive MA’s. The start site does not cut off any coding potential and it has an optimal gap of 3.

BLAST function evidence. What assigned functions do other highly similar genes have?



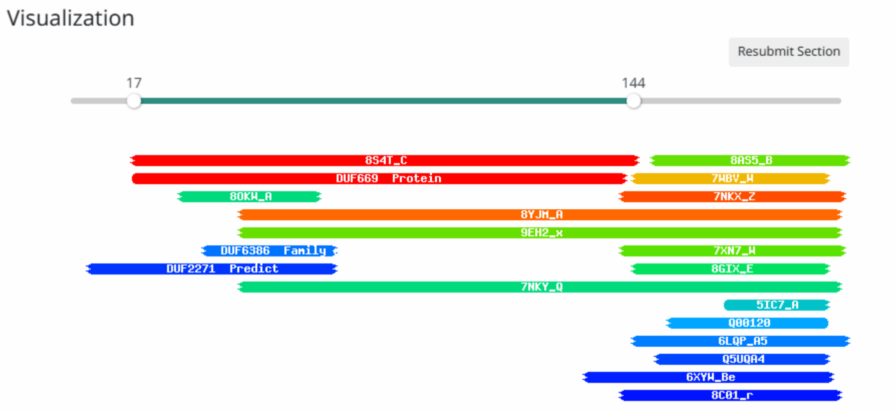
Score	Target Description
491	hypothetical protein PBI_NINA_38 [Gordonia pha
490	hypothetical protein SEA_AXYM_38 [Gordonia p
490	hypothetical protein PBI_QUASAR_39 [Gordonia
489	DNA polymerase [Gordonia phage Cozz] >gb|AN.
489	hypothetical protein SEA_AGATHA_38 [Gordoni

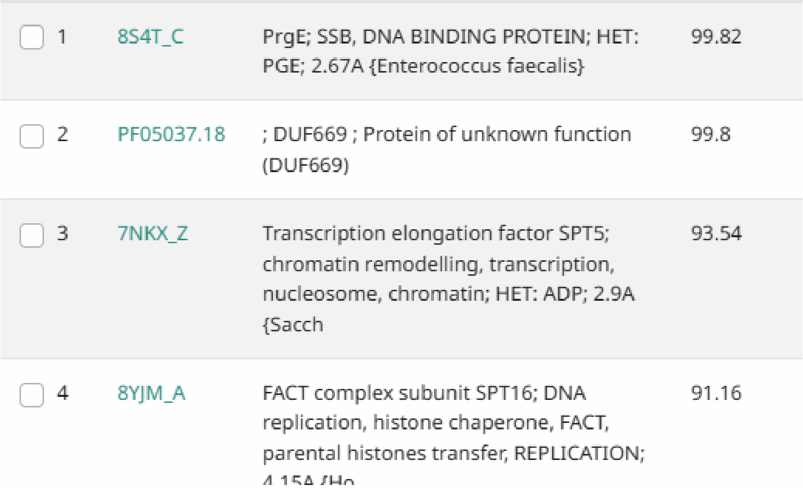


	Description
☑	hypothetical protein PP998_gp43 [Gordonia phage Vine]
☑	hypothetical protein PP997_gp40 [Gordonia phage BigChungus]
☑	hypothetical protein SEA_SUMMITACADEMY_40 [Gordonia phage SummitAcademy]
☑	hypothetical protein PP992_gp42 [Gordonia phage Pons]
☑	hypothetical protein SEA_ELINAL_44 [Gordonia phage Elinal]
☑	hypothetical protein PP996_gp43 [Gordonia phage SheckWes]
☑	hypothetical protein SEA_MANOR_42 [Gordonia phage MAnor]
☑	hypothetical protein PP993_gp44 [Gordonia phage Mayweather]
☑	hypothetical protein PP995_gp37 [Gordonia phage Lauer]
☑	hypothetical protein PP994_gp42 [Gordonia phage CherryonLim]
☑	hypothetical protein PBI_NINA_38 [Gordonia phage Nina]
☑	hypothetical protein SEA_AXYM_38 [Gordonia phage Axym]
☑	hypothetical protein PBI_QUASAR_39 [Gordonia phage Quasar]
☑	DNA polymerase [Gordonia phage Cozz]

- There are 21 BLAST hits with a function of hypothetical protein on DNA master, the other 4 are DNA polymerase.
- BLASTing on NCBI showed the top 13 hits to be a hypothetical protein. The 14th was the first DNA polymerase.

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

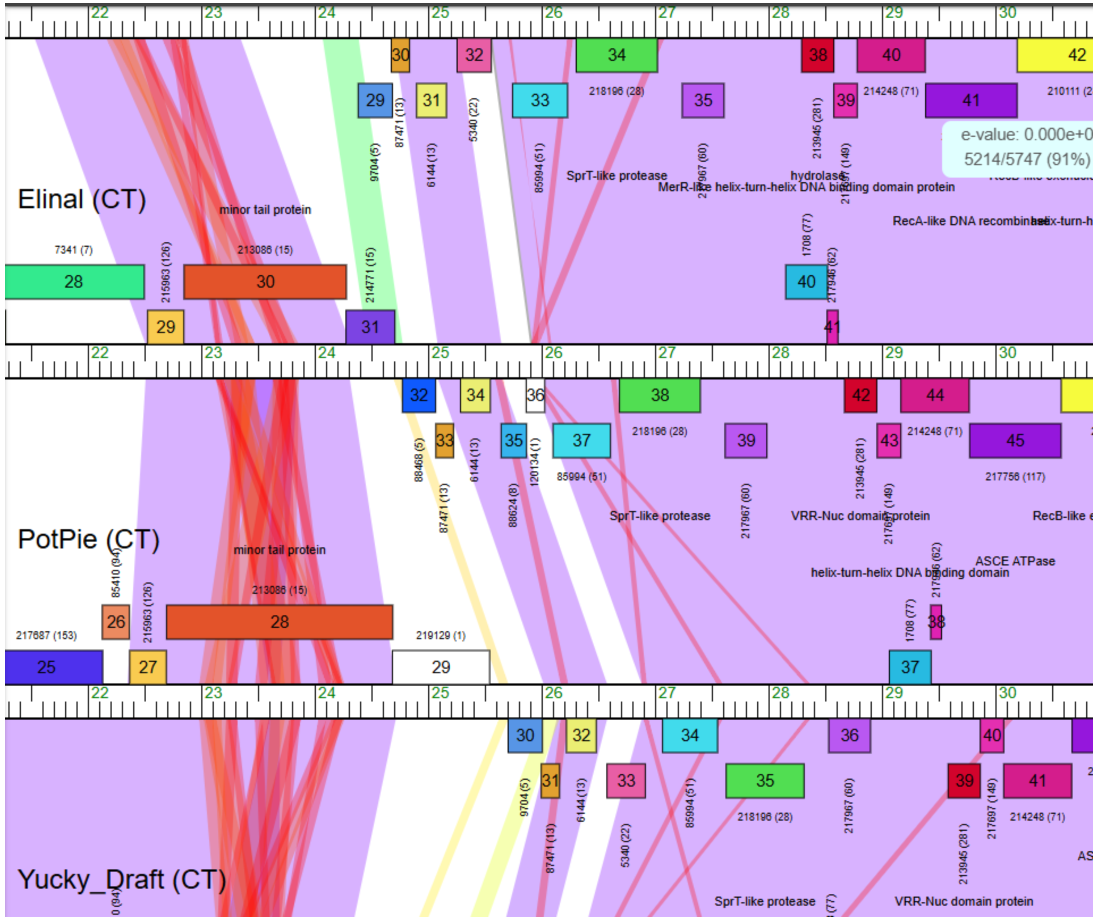




□	1	8S4T_C	PrgE; SSB, DNA BINDING PROTEIN; HET: PGE; 2.67A {Enterococcus faecalis}	99.82
□	2	PF05037.18	; DUF669 ; Protein of unknown function (DUF669)	99.8
□	3	7NKX_Z	Transcription elongation factor SPT5; chromatin remodelling, transcription, nucleosome, chromatin; HET: ADP; 2.9A {Sacch	93.54
□	4	8YJM_A	FACT complex subunit SPT16; DNA replication, histone chaperone, FACT, parental histones transfer, REPLICATION; 4.15A {Ho	91.16

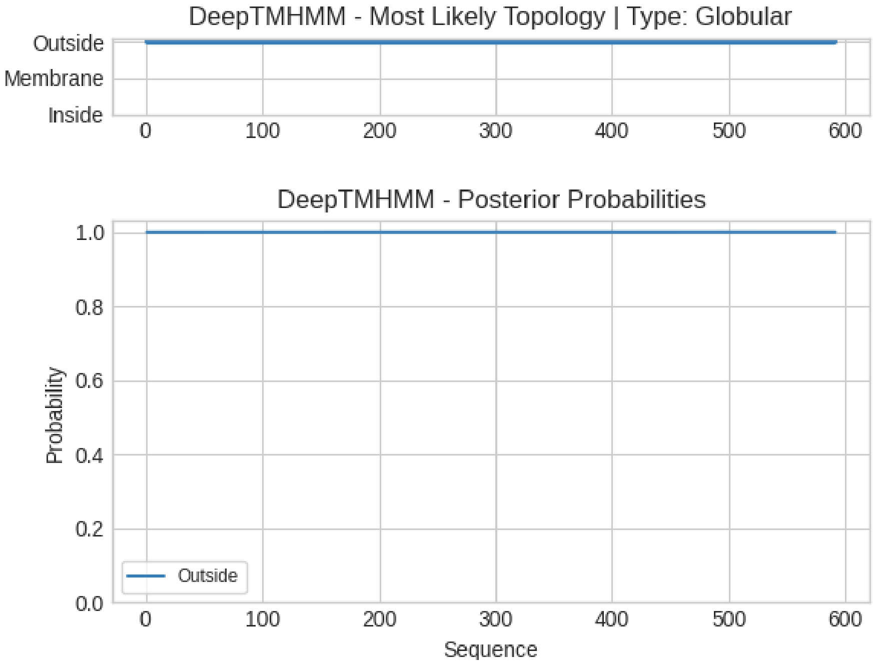
- Hhpred shows 4 hits with 90% probability or better. One hit is a hypothetical protein, the other 3 have a listed function.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- PotPie, BigChungus, and Elinal all contain this gene and have it called as a hypothetical protein. There are no conserved domains.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- This is not an intermembrane protein as it never crosses the membrane.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I am assigning this gene the function of hypothetical protein. BLAST via both NCBI and DNA master show several hits as a hypothetical protein. The Hhpred evidence is not as strong as I would prefer, but I believe it to be strong enough, showing 1 hit as a hypothetical protein. Phamerator shows that 3 highly similar phages all have the gene and call it a hypothetical protein with no conserved domains. Lastly, it was determined to not be an intermembrane protein.

Feature 44 – Reverse – Stop 29972

Glimmer/GeneMark

What feature number is this? 44

What is the stop site? 29972

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

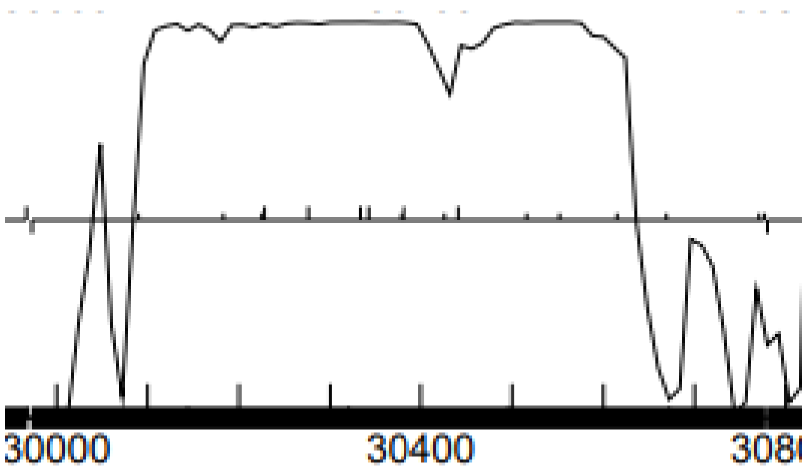
Called by both Glimmer and GeneMark

What is the autoannotated start?

30778

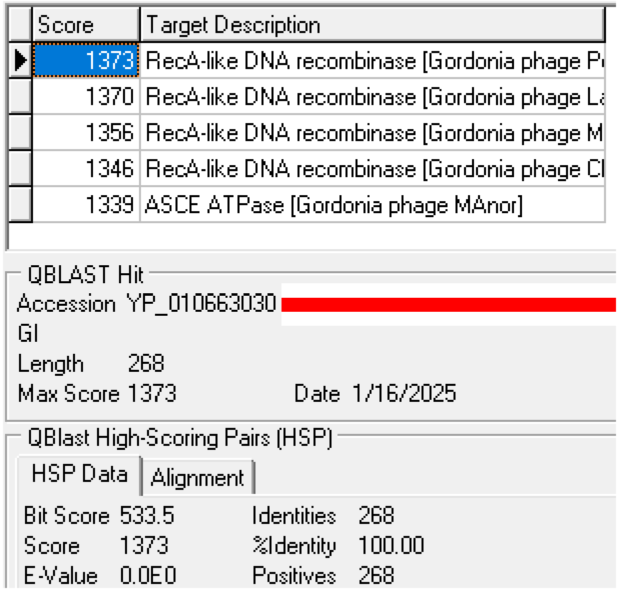
Gap: ____0____ or overlap: _______ (with gene in front of it) for the autoannotated start

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Throughout the sequence there are many strong and weak peaks of coding potential on reading frame 4. It is the only frame with coding potential.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
1373	RecA-like DNA recombinase [Gordonia phage P
1370	RecA-like DNA recombinase [Gordonia phage La
1356	RecA-like DNA recombinase [Gordonia phage M
1346	RecA-like DNA recombinase [Gordonia phage Cl
1339	ASCE ATPase [Gordonia phage MAnor]

QBLAST Hit

Accession YP_010663030

GI

Length 268

Max Score 1373 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

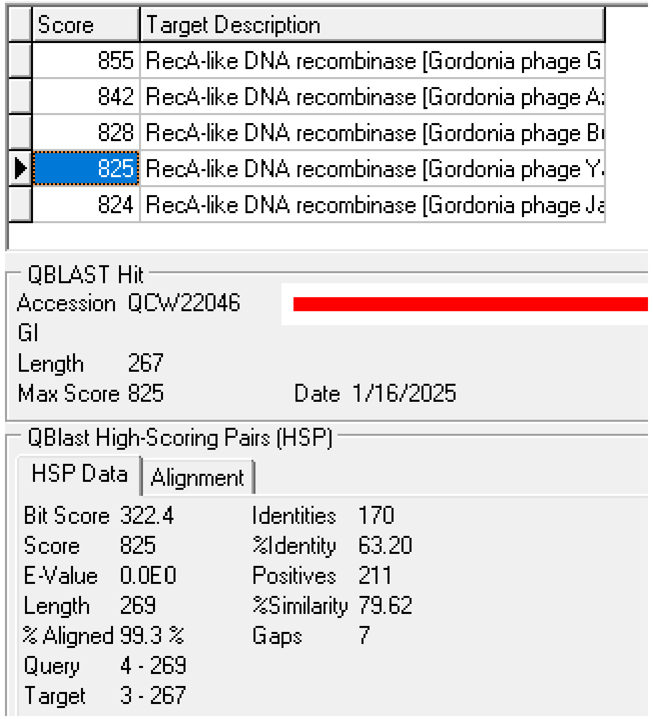
Bit Score	533.5	Identities	268
Score	1373	%Identity	100.00
E-Value	0.0E0	Positives	268

- All 25 highly similar phages have an E-value close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This is a gene. There is a lot of coding potential throughout the sequence of the gene and BLAST shows 25 highly similar phages with an E-value close to 0. It is also called by Glimmer and GeneMark.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
855	RecA-like DNA recombinase [Gordonia phage G
842	RecA-like DNA recombinase [Gordonia phage A:
828	RecA-like DNA recombinase [Gordonia phage B(
825	RecA-like DNA recombinase [Gordonia phage Y.
824	RecA-like DNA recombinase [Gordonia phage Ja

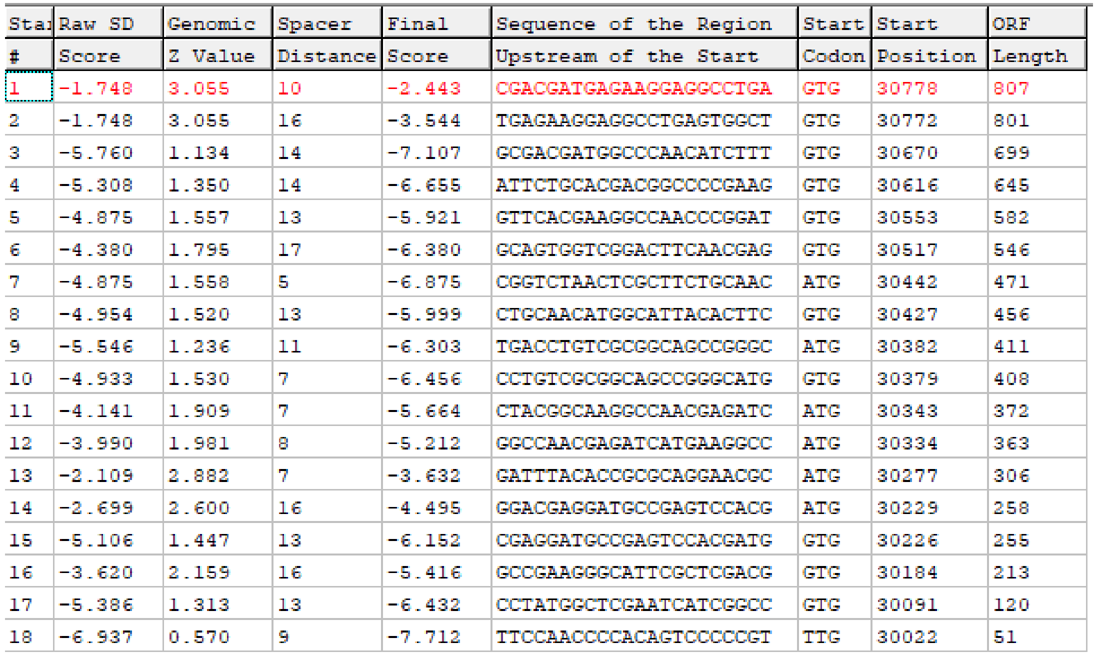
QBLAST Hit
Accession QCW22046
GI
Length 267
Max Score 825
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 322.4
Score 825
E-Value 0.0E0
Length 269
% Aligned 99.3 %
Query 4 - 269
Target 3 - 267
Identities 170
%Identity 63.20
Positives 211
%Similarity 79.62
Gaps 7

- There are 19 1:1 alignments. One 1:7 alignment. One 4:3 alignment. One 3:1 alignment. Two 2:3 alignments and one 1:2 alignment.

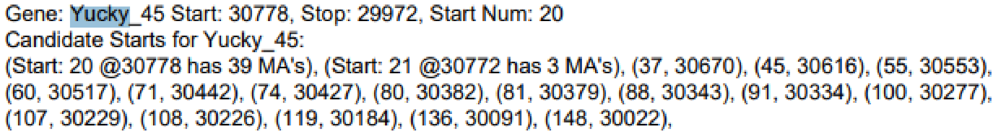
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



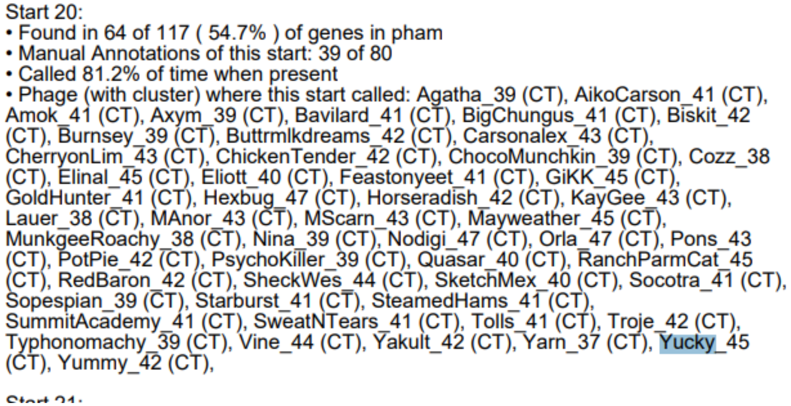
Sta#	Raw SD Score	Genomic Z Value	Spacer Distance	Final Score	Sequence of the Region Upstream of the Start	Start Codon	Start Position	ORF Length
1	-1.748	3.055	10	-2.443	CGACGATGAGAAGGAGGCCTGA	GTG	30778	807
2	-1.748	3.055	16	-3.544	TGAGAAGGAGGCCTGAGTGGCT	GTG	30772	801
3	-5.760	1.134	14	-7.107	GCGACGATGGCCCAACATCTTT	GTG	30670	699
4	-5.308	1.350	14	-6.655	ATTCTGCACGACGGCCCCGAAG	GTG	30616	645
5	-4.875	1.557	13	-5.921	GTTCACGAAGGCCAACCCGGAT	GTG	30553	582
6	-4.380	1.795	17	-6.380	GCAGTGGTCGGACTTCAACGAG	GTG	30517	546
7	-4.875	1.558	5	-6.875	CGGTCTAACTCGCTTCTGCAAC	ATG	30442	471
8	-4.954	1.520	13	-5.999	CTGCAACATGGCATTACACTTC	GTG	30427	456
9	-5.546	1.236	11	-6.303	TGACCTGTCGCGGCAGCCGGGC	ATG	30382	411
10	-4.933	1.530	7	-6.456	CCTGTCGCGGCAGCCGGGCATG	GTG	30379	408
11	-4.141	1.909	7	-5.664	CTACGGCAAGGCCAACGAGATC	ATG	30343	372
12	-3.990	1.981	8	-5.212	GGCCAACGAGATCATGAAGGCC	ATG	30334	363
13	-2.109	2.882	7	-3.632	GATTTACACCGCGCAGGAACGC	ATG	30277	306
14	-2.699	2.600	16	-4.495	GGACGAGGATGCCGAGTCCACG	ATG	30229	258
15	-5.106	1.447	13	-6.152	CGAGGATGCCGAGTCCACGATG	GTG	30226	255
16	-3.620	2.159	16	-5.416	GCCGAAGGGCATTCGCTCGACG	GTG	30184	213
17	-5.386	1.313	13	-6.432	CCTATGGCTCGAATCATCGGCC	GTG	30091	120
18	-6.937	0.570	9	-7.712	TTCCAACCCCACAGTCCCCCGT	TTG	30022	51

- The Z-value of the autoannotated start is 3.055 and the final score is -2.443. There is an alternate start with the same Z-value, but a worse final score. I will look into it in starterator. All other values are not ideal.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



Gene: Yucky_45 Start: 30778, Stop: 29972, Start Num: 20
Candidate Starts for Yucky_45:
(Start: 20 @30778 has 39 MA's), (Start: 21 @30772 has 3 MA's), (37, 30670), (45, 30616), (55, 30553), (60, 30517), (71, 30442), (74, 30427), (80, 30382), (81, 30379), (88, 30343), (91, 30334), (100, 30277), (107, 30229), (108, 30226), (119, 30184), (136, 30091), (148, 30022),

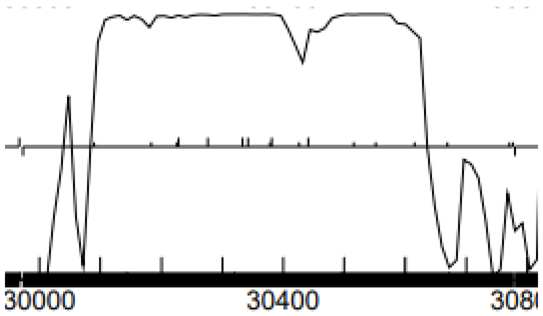


Start 20:
• Found in 64 of 117 (54.7%) of genes in pham
• Manual Annotations of this start: 39 of 80
• Called 81.2% of time when present
• Phage (with cluster) where this start called: Agatha_39 (CT), AikoCarson_41 (CT), Amok_41 (CT), Axym_39 (CT), Bavilard_41 (CT), BigChungus_41 (CT), Biskit_42 (CT), Burnsey_39 (CT), Buttrmlkdreams_42 (CT), Carsonalex_43 (CT), CherryonLim_43 (CT), ChickenTender_42 (CT), ChocoMunchkin_39 (CT), Cozz_38 (CT), Elinal_45 (CT), Eliott_40 (CT), Feastonyeet_41 (CT), GiKK_45 (CT), GoldHunter_41 (CT), Hexbug_47 (CT), Horseradish_42 (CT), KayGee_43 (CT), Lauer_38 (CT), MAnor_43 (CT), MScarn_43 (CT), Mayweather_45 (CT), MunkgeeRoachy_38 (CT), Nina_39 (CT), Nodigi_47 (CT), Orla_47 (CT), Pons_43 (CT), PotPie_42 (CT), PsychoKiller_39 (CT), Quasar_40 (CT), RanchParmCat_45 (CT), RedBaron_42 (CT), SheckWes_44 (CT), SketchMex_40 (CT), Socotra_41 (CT), Sopespian_39 (CT), Starburst_41 (CT), SteamedHams_41 (CT), SummitAcademy_41 (CT), SweatNTears_41 (CT), Tolls_41 (CT), Troje_42 (CT), Typhonomachy_39 (CT), Vine_44 (CT), Yakult_42 (CT), Yarn_37 (CT), Yucky_45 (CT), Yummy_42 (CT),

Start 21:

- Automated start: 39 MA's, called 81.2% of the time when present.
- Alternate start (30772): 3 MA's, called 15.2% of the time when present.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- Both starts cut off a slight peak of coding potential, however the autoannotated start cuts off less by about 6 nucleotides.

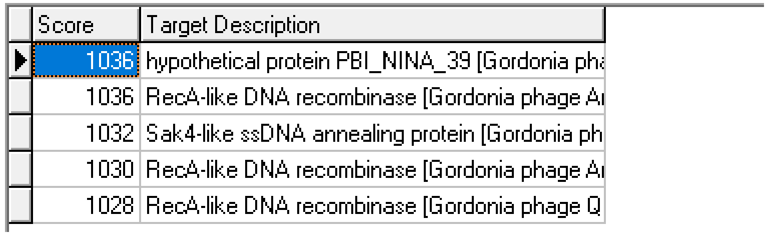
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 30779-30778=1-1 for gap= 0

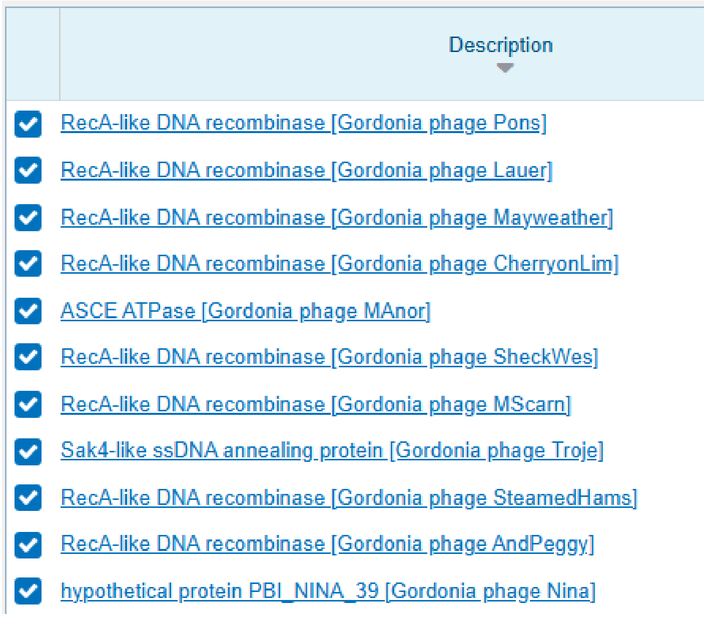
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is the autoannotated start of 30778. There are 19 1:1 alignments on BLAST. The RBS numbers showed that the automated start had the best numbers, but there was another good start so I looked into it in starterator. Starterator showed that the automated start had more manual annotations and was called more often. Both starts cut off some coding potential, but the automated start cut off less so I stopped considering the possible alternate start. The automated start also had a gap of 0, which is optimal. The start is 30778.

BLAST function evidence. What assigned functions do other highly similar genes have?



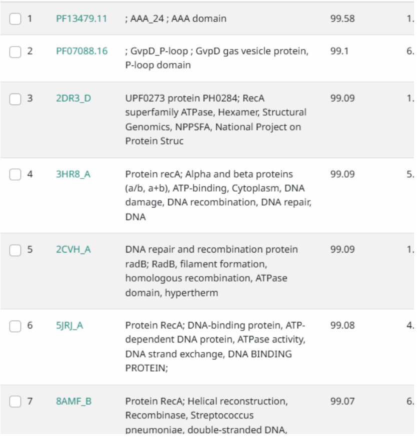
Score	Target Description
1036	hypothetical protein PBI_NINA_39 [Gordonia pha
1036	RecA-like DNA recombinase [Gordonia phage Ar
1032	Sak4-like ssDNA annealing protein [Gordonia ph
1030	RecA-like DNA recombinase [Gordonia phage Ar
1028	RecA-like DNA recombinase [Gordonia phage Q



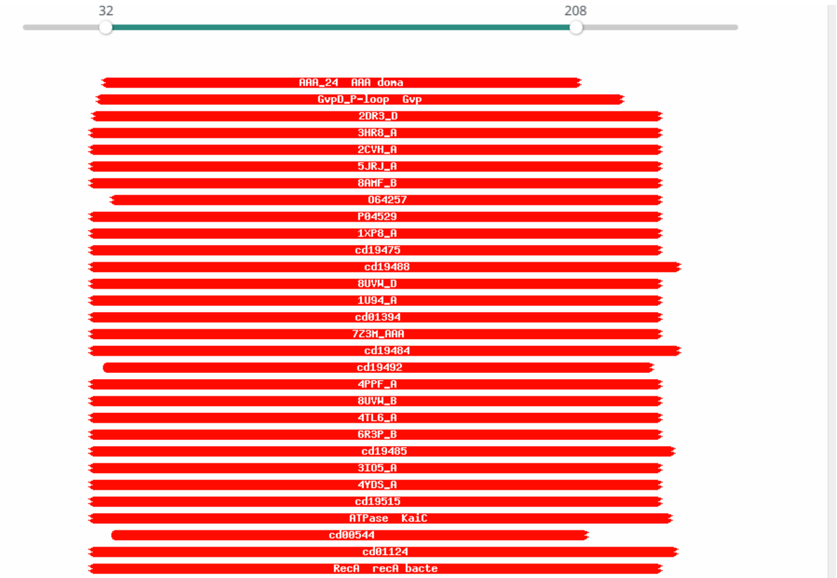
	Description
☑	RecA-like DNA recombinase [Gordonia phage Pons]
☑	RecA-like DNA recombinase [Gordonia phage Lauer]
☑	RecA-like DNA recombinase [Gordonia phage Mayweather]
☑	RecA-like DNA recombinase [Gordonia phage CherryonLim]
☑	ASCE ATPase [Gordonia phage MAnor]
☑	RecA-like DNA recombinase [Gordonia phage SheckWes]
☑	RecA-like DNA recombinase [Gordonia phage MScarn]
☑	Sak4-like ssDNA annealing protein [Gordonia phage Troje]
☑	RecA-like DNA recombinase [Gordonia phage SteamedHams]
☑	RecA-like DNA recombinase [Gordonia phage AndPeggy]
☑	hypothetical protein PBI_NINA_39 [Gordonia phage Nina]

- DNA master BLAST shows many possible functions. The most abundant is a RecA-like DNA recombinase. There are also some results for a hypothetical protein and ASCE ATPase and a SAK4-like ssDNA annealing protein.
- BLASTing on NCBI yielded results for all of the above listed functions, the most abundant being a RecA-like DNA recombinase.

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

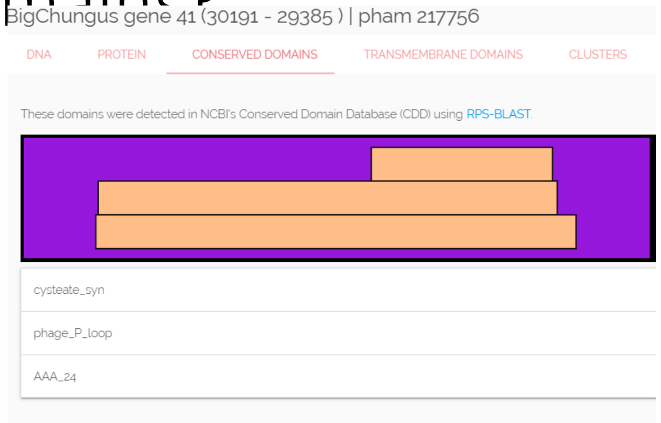


□ 1	PF13479.11	; AAA_24 ; AAA domain	99.58	1.
□ 2	PF07088.16	; GvpD_P-loop ; GvpD gas vesicle protein, P-loop domain	99.1	6.
□ 3	2DR3_D	UPF0273 protein PH0284; RecA superfamily ATPase, Hexamer, Structural Genomics, NPPSFA, National Project on Protein Struc	99.09	1.
□ 4	3HR8_A	Protein recA; Alpha and beta proteins (a/b, a+b), ATP-binding, Cytoplasm, DNA damage, DNA recombination, DNA repair, DNA	99.09	5.
□ 5	2CVH_A	DNA repair and recombination protein radB; RadB, filament formation, homologous recombination, ATPase domain, hypertherm	99.09	1.
□ 6	5JRJ_A	Protein RecA; DNA-binding protein, ATP-dependent DNA protein, ATPase activity, DNA strand exchange, DNA BINDING PROTEIN;	99.08	4.
□ 7	8AMF_B	Protein RecA; Helical reconstruction, Recombinase, Streptococcus pneumoniae, double-stranded DNA,	99.07	6.



- There are many strong hits. Primarily for a RecA like protein.
- After discussion with Dr. Rueschhoff, this gene does not meet the requirements to be called a RecA like protein.
- I was advised to call it an ASCE ATPase, but more evidence is needed, likely to call it that.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



BigChungus gene 41 (30191 - 29385) | pham 217756

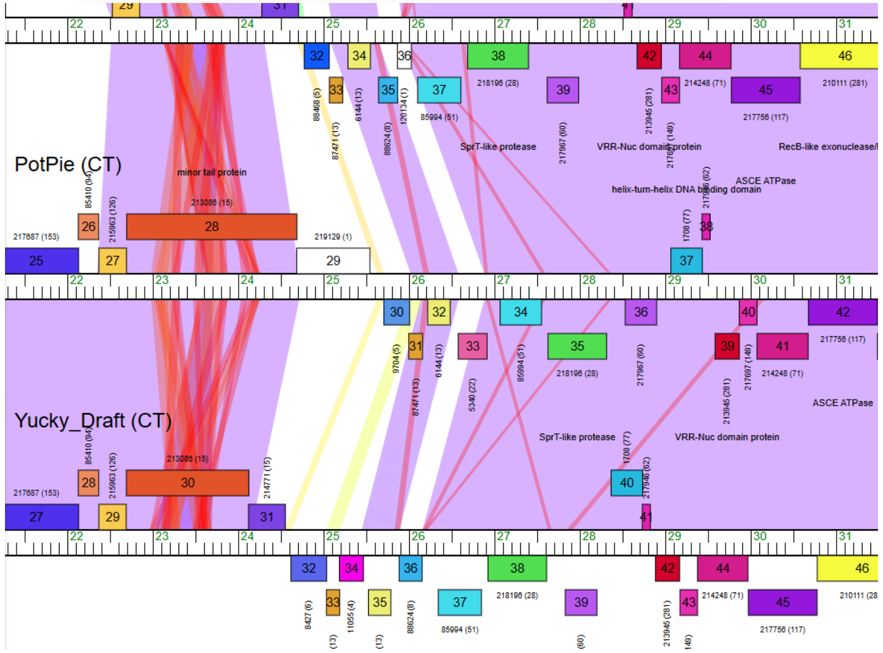
DNA PROTEIN CONSERVED DOMAINS TRANSMEMBRANE DOMAINS CLUSTERS

These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

cysteate_syn

phage_P_loop

AAA_24



- Elinal, PotPie, and BigChungus have this gene and it is called an ASCE ATPase by Elinal and PotPie.
- Called a RecA-like DNA recombinase.
- Elinal and PotPie have a AAA conserved domain.
- BigChungus has 3 conserved domains, as pictured.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- I would like to call this an ASCE ATPase.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I am calling this an ASCE ATPase. BLAST had numerous hits for this function and Hhpred showed it could not be called a RecA-like protein. HHPred also contained hits for an ASCE ATPase. Phamerator also showed that 2 of the 3 highly similar phages observed had this gene with the ASCE ATPse function.

Feature 45 – Reverse – Stop 30079

Glimmer/GeneMark

What feature number is this? 45

What is the stop site? 30779

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

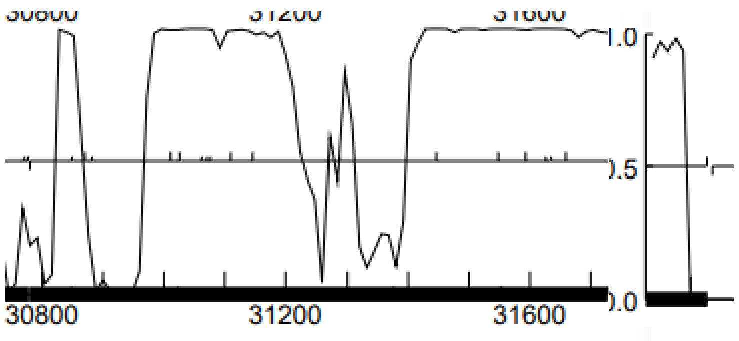
Called by both Glimmer and GeneMark

What is the autoannotated start?

31828

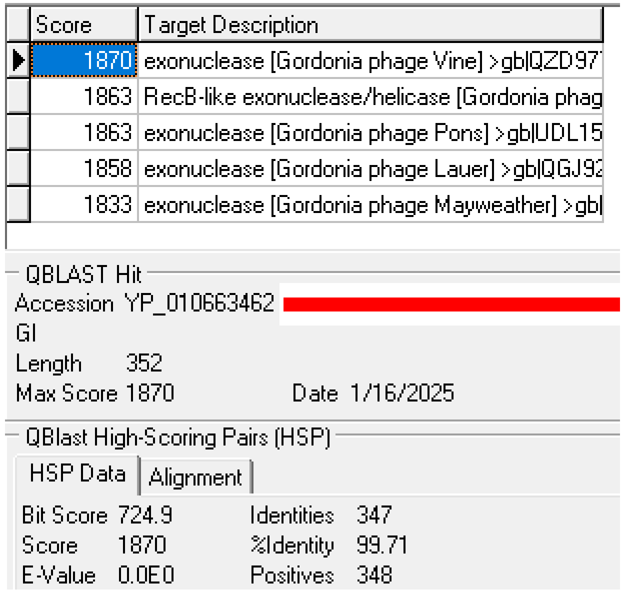
Gap: ________ or overlap: ____11___ (with gene in front of it) for the autoannotated start

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



There are many strong peaks throughout the sequence of coding potential, that taper off to weak peaks before repeaking strongly.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
1870	exonuclease [Gordonia phage Vine] >gb|QZD97
1863	RecB-like exonuclease/helicase [Gordonia phag
1863	exonuclease [Gordonia phage Pons] >gb|UDL15
1858	exonuclease [Gordonia phage Lauer] >gb|QGJ92
1833	exonuclease [Gordonia phage Mayweather] >gb|

QBLAST Hit

Accession YP_010663462

GI

Length 352

Max Score 1870 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 724.9 Identities 347

Score 1870 %Identity 99.71

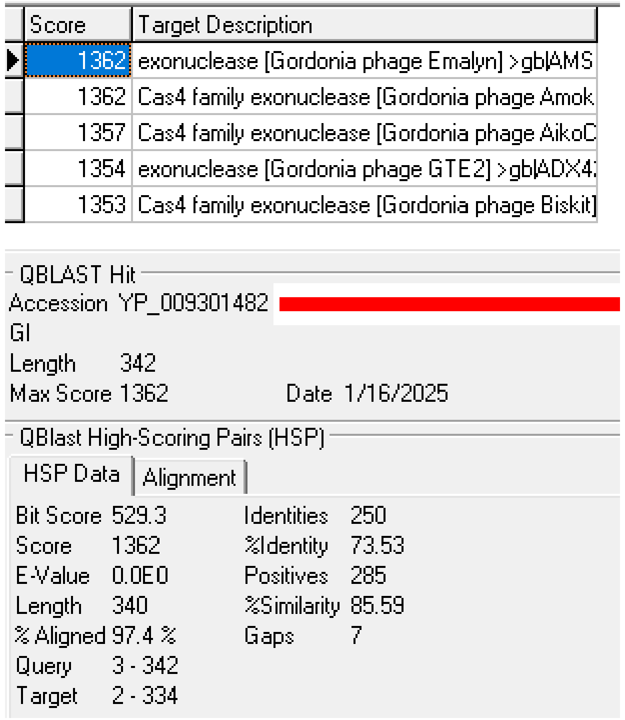
E-Value 0.0E0 Positives 348

- There are at least 25 highly similar phages with an E-value close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This is a gene. It is called by both Glimmer and GeneMark, has a lot of strong coding potential throughout the sequence, and has at least 25 highly similar phages with an E-value close to 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
1362	exonuclease [Gordonia phage Emalyn] >gb|AMS
1362	Cas4 family exonuclease [Gordonia phage Amok
1357	Cas4 family exonuclease [Gordonia phage AikoC
1354	exonuclease [Gordonia phage GTE2] >gb|ADX4:
1353	Cas4 family exonuclease [Gordonia phage Biskit]

QBLAST Hit

Accession YP_009301482

GI

Length 342

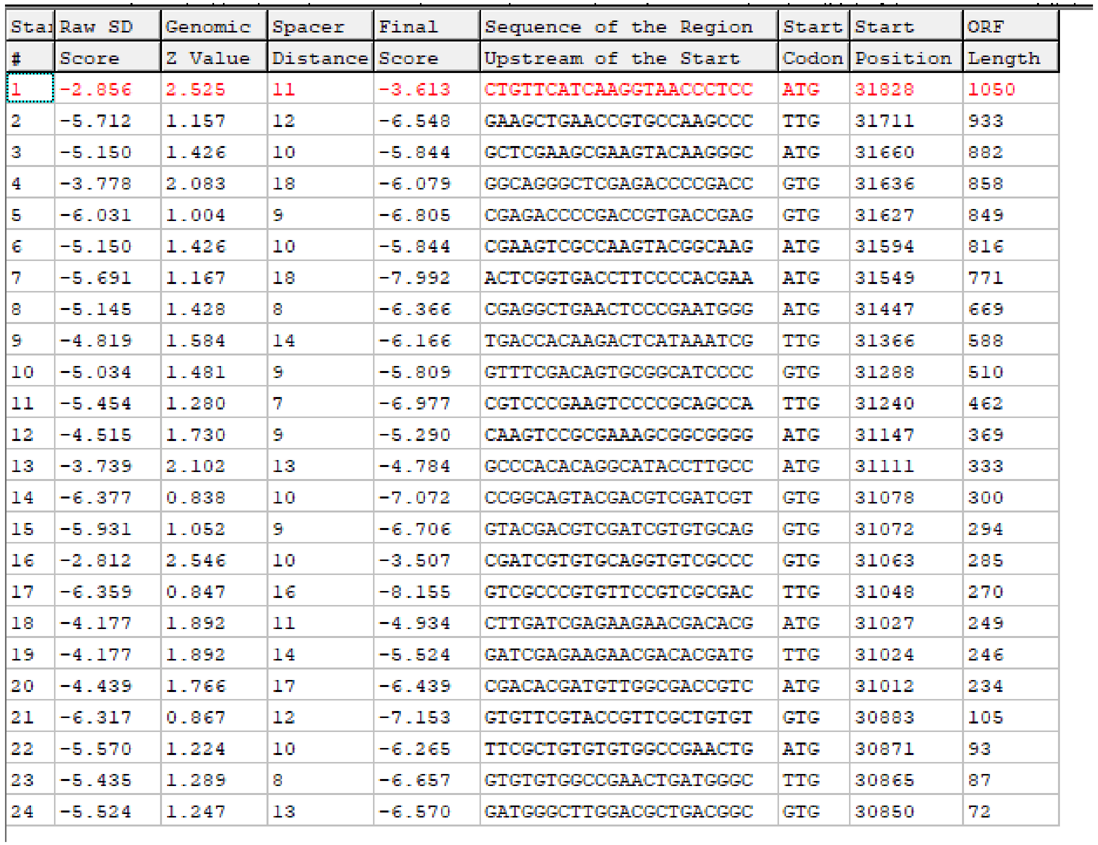
Max Score 1362 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	529.3	Identities	250
Score	1362	%Identity	73.53
E-Value	0.0E0	Positives	285
Length	340	%Similarity	85.59
% Aligned	97.4 %	Gaps	7
Query	3 - 342		
Target	2 - 334		

- There is one 1:1 alignment. There are 13 3:2 alignments and 6 2:5 alignments. There are a handful of a couple others, including a 4:7, a 2:26, and a 3:17

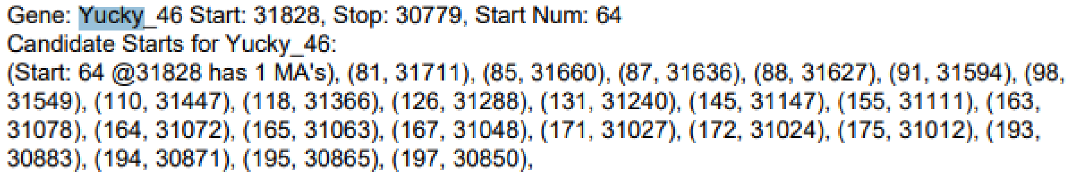
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts?   Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



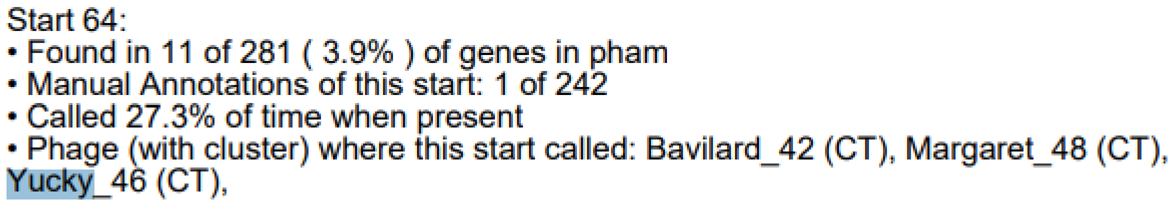
Sta#	Raw SD Score	Genomic Z Value	Spacer Distance	Final Score	Sequence of the Region Upstream of the Start	Start Codon	Start Position	ORF Length
1	-2.856	2.525	11	-3.613	CTGTTCATCAAGGTAACCCTCC	ATG	31828	1050
2	-5.712	1.157	12	-6.548	GAAGCTGAACCGTGCCAAGCCC	TTG	31711	933
3	-5.150	1.426	10	-5.844	GCTCGAAGCGAAGTACAAGGGC	ATG	31660	882
4	-3.778	2.083	18	-6.079	GGCAGGGCTCGAGACCCCGACC	GTG	31636	858
5	-6.031	1.004	9	-6.805	CGAGACCCCGACCGTGACCGAG	GTG	31627	849
6	-5.150	1.426	10	-5.844	CGAAGTCGCCAAGTACGGCAAG	ATG	31594	816
7	-5.691	1.167	18	-7.992	ACTCGGTGACCTTCCCCACGAA	ATG	31549	771
8	-5.145	1.428	8	-6.366	CGAGGCTGAACTCCCGAATGGG	ATG	31447	669
9	-4.819	1.584	14	-6.166	TGACCACAAGACTCATAAATCG	TTG	31366	588
10	-5.034	1.481	9	-5.809	GTTTCGACAGTGCGGCATCCCC	GTG	31288	510
11	-5.454	1.280	7	-6.977	CGTCCCGAAGTCCCCGCAGCCA	TTG	31240	462
12	-4.515	1.730	9	-5.290	CAAGTCCGCGAAAGCGGCGGGG	ATG	31147	369
13	-3.739	2.102	13	-4.784	GCCCACACAGGCATACCTTGCC	ATG	31111	333
14	-6.377	0.838	10	-7.072	CCGGCAGTACGACGTCGATCGT	GTG	31078	300
15	-5.931	1.052	9	-6.706	GTACGACGTCGATCGTGTGCAG	GTG	31072	294
16	-2.812	2.546	10	-3.507	CGATCGTGTGCAGGTGTCGCCC	GTG	31063	285
17	-6.359	0.847	16	-8.155	GTCGCCCGTGTTCCGTCGCGAC	TTG	31048	270
18	-4.177	1.892	11	-4.934	CTTGATCGAGAAGAACGACACG	ATG	31027	249
19	-4.177	1.892	14	-5.524	GATCGAGAAGAACGACACGATG	TTG	31024	246
20	-4.439	1.766	17	-6.439	CGACACGATGTTGGCGACCGTC	ATG	31012	234
21	-6.317	0.867	12	-7.153	GTGTTCGTACCGTTCGCTGTGT	GTG	30883	105
22	-5.570	1.224	10	-6.265	TTCGCTGTGTGTGGCCGAACTG	ATG	30871	93
23	-5.435	1.289	8	-6.657	GTGTGTGGCCGAACTGATGGGC	TTG	30865	87
24	-5.524	1.247	13	-6.570	GATGGGCTTGGACGCTGACGGC	GTG	30850	72

- The automated starts Z-value is 2.525 and the final score is -3.613. There is another start with decent numbers, but it cuts off too much CP to be considered.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



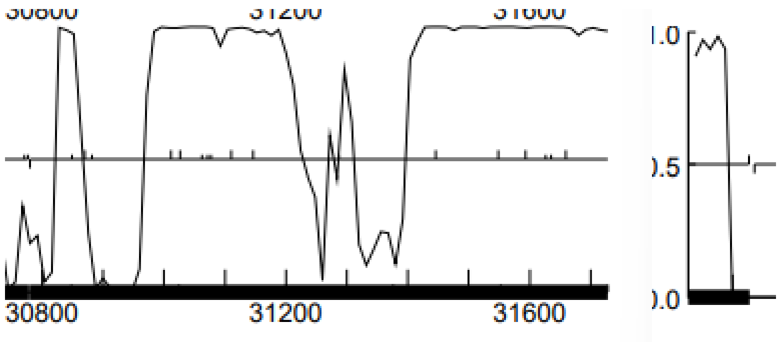
Gene: Yucky_46 Start: 31828, Stop: 30779, Start Num: 64
Candidate Starts for Yucky_46:
(Start: 64 @31828 has 1 MA's), (81, 31711), (85, 31660), (87, 31636), (88, 31627), (91, 31594), (98, 31549), (110, 31447), (118, 31366), (126, 31288), (131, 31240), (145, 31147), (155, 31111), (163, 31078), (164, 31072), (165, 31063), (167, 31048), (171, 31027), (172, 31024), (175, 31012), (193, 30883), (194, 30871), (195, 30865), (197, 30850),



Start 64:
- Found in 11 of 281 (3.9%) of genes in pham
- Manual Annotations of this start: 1 of 242
- Called 27.3% of time when present
- Phage (with cluster) where this start called: Bavilard_42 (CT), Margaret_48 (CT), Yucky_46 (CT),

- The automated start has 1 MA, however it is the only start to ever receive a manual annotation.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- The automated start includes all coding potential.

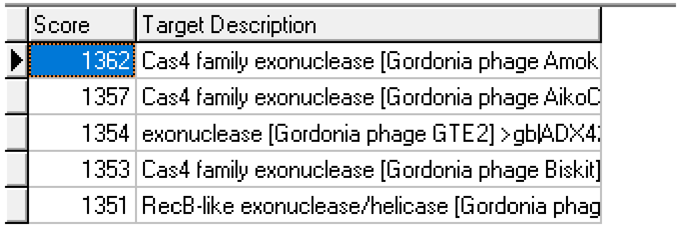
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start?   Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 31828-31818= 10+1 for overlap=11

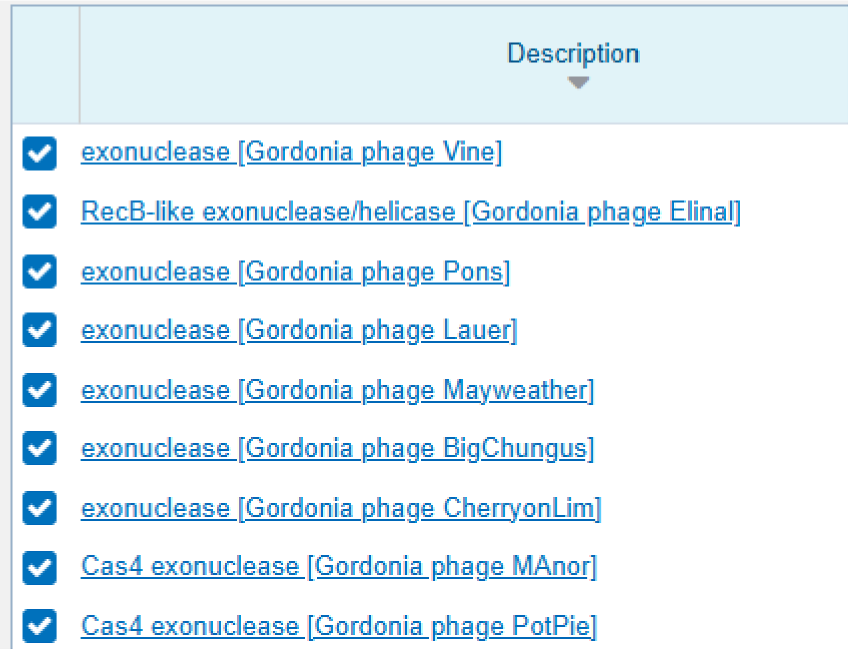
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is the automated start of 31828. It has one 1:1 alignment, and the best RBS numbers that make sense as a possible start site. The starterator evidence isn’t as good as I’d like, but it is compelling enough to call with the start site being the only possible start to receive an MA. It also cuts off no coding potential and has an acceptable gap.

BLAST function evidence. What assigned functions do other highly similar genes have?



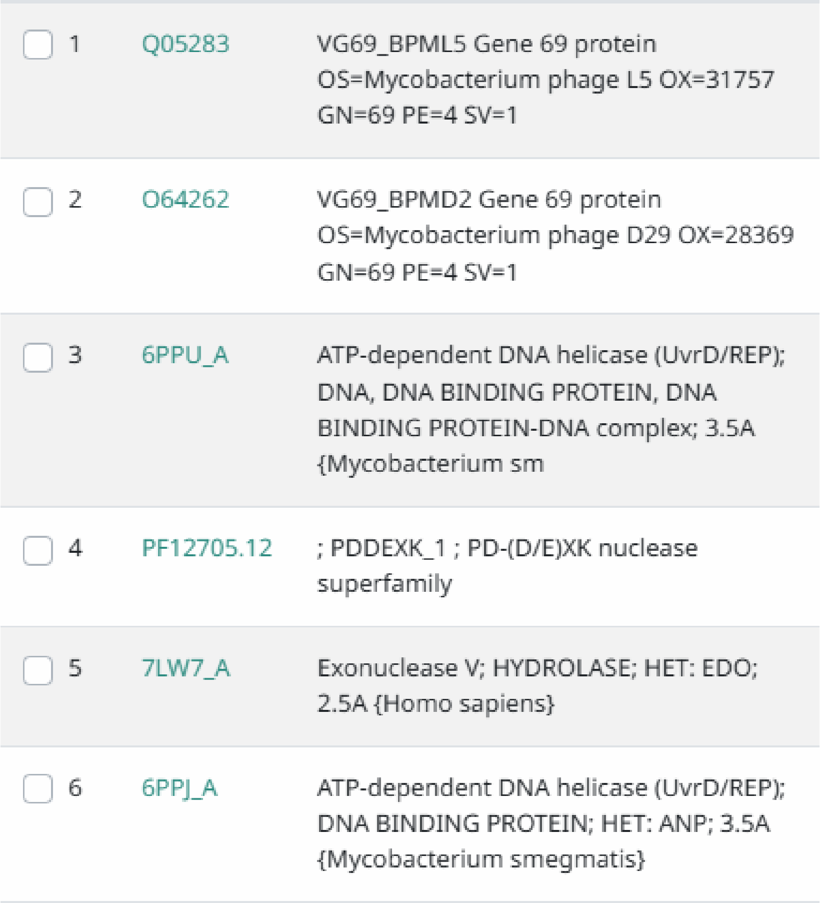
Score	Target Description
1362	Cas4 family exonuclease [Gordonia phage Amok
1357	Cas4 family exonuclease [Gordonia phage AikoC
1354	exonuclease [Gordonia phage GTE2] >gb|ADX4
1353	Cas4 family exonuclease [Gordonia phage Biskit]
1351	RecB-like exonuclease/helicase [Gordonia phag



	Description
☑	exonuclease [Gordonia phage Vine]
☑	RecB-like exonuclease/helicase [Gordonia phage Elinal]
☑	exonuclease [Gordonia phage Pons]
☑	exonuclease [Gordonia phage Lauer]
☑	exonuclease [Gordonia phage Mayweather]
☑	exonuclease [Gordonia phage BigChungus]
☑	exonuclease [Gordonia phage CherryonLim]
☑	Cas4 exonuclease [Gordonia phage MAnor]
☑	Cas4 exonuclease [Gordonia phage PotPie]

- All 25 highly similar genes shown by DNA master BLAST have an exonuclease function. After discussing with Dr. Rueschhoff, it is likely a Cas4 exonuclease.
- NCBI BLAST showed largely the same results.

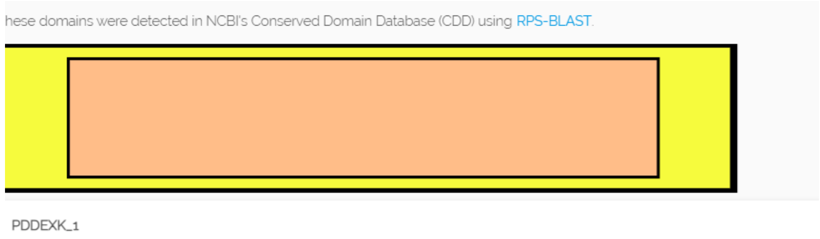
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



□	1	Q05283	VG69_BPML5 Gene 69 protein OS=Mycobacterium phage L5 OX=31757 GN=69 PE=4 SV=1
□	2	O64262	VG69_BPMD2 Gene 69 protein OS=Mycobacterium phage D29 OX=28369 GN=69 PE=4 SV=1
□	3	6PPU_A	ATP-dependent DNA helicase (UvrD/REP); DNA, DNA BINDING PROTEIN, DNA BINDING PROTEIN-DNA complex; 3.5A {Mycobacterium sm
□	4	PF12705.12	; PDDEXK_1 ; PD-(D/E)XK nuclease superfamily
□	5	7LW7_A	Exonuclease V; HYDROLASE; HET: EDO; 2.5A {Homo sapiens}
□	6	6PPJ_A	ATP-dependent DNA helicase (UvrD/REP); DNA BINDING PROTEIN; HET: ANP; 3.5A {Mycobacterium smegmatis}

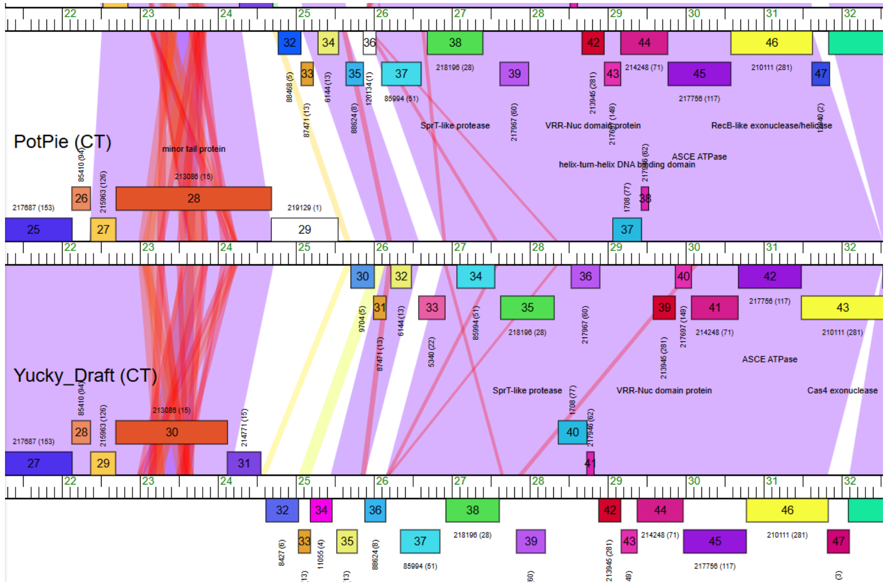
- After viewing Hhpred evidence, a helicase domain was found. This means this is likely a RecB-like exonuclease.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



hese domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

PDDEXK_1



- PotPie, BigChungus, and Elinal all have this gene. PotPie calls it a Cas4 exonuclease. BigChungus and Elinal call it a RecB-like exonuclease due to having a helicase domain.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- I would like to call this a RecB-like exonuclease.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I call this a RecB-like exonuclease/helicase. BLAST showed many hits for this function, at the time I thought it was a Cas4 exonuclease due to a lack of a helicase domain. However, multiple helicase domains were found on Hhpred and Phamerator showed that 2 of the 3 similar phages I have been looking at call a RecB-like exonuclease. Due to it having a helicase domain I believe this to be a RecB-like exonuclease.

Feature 46 – Reverse – Stop 31818

Glimmer/GeneMark

What feature number is this?

What is the stop site?

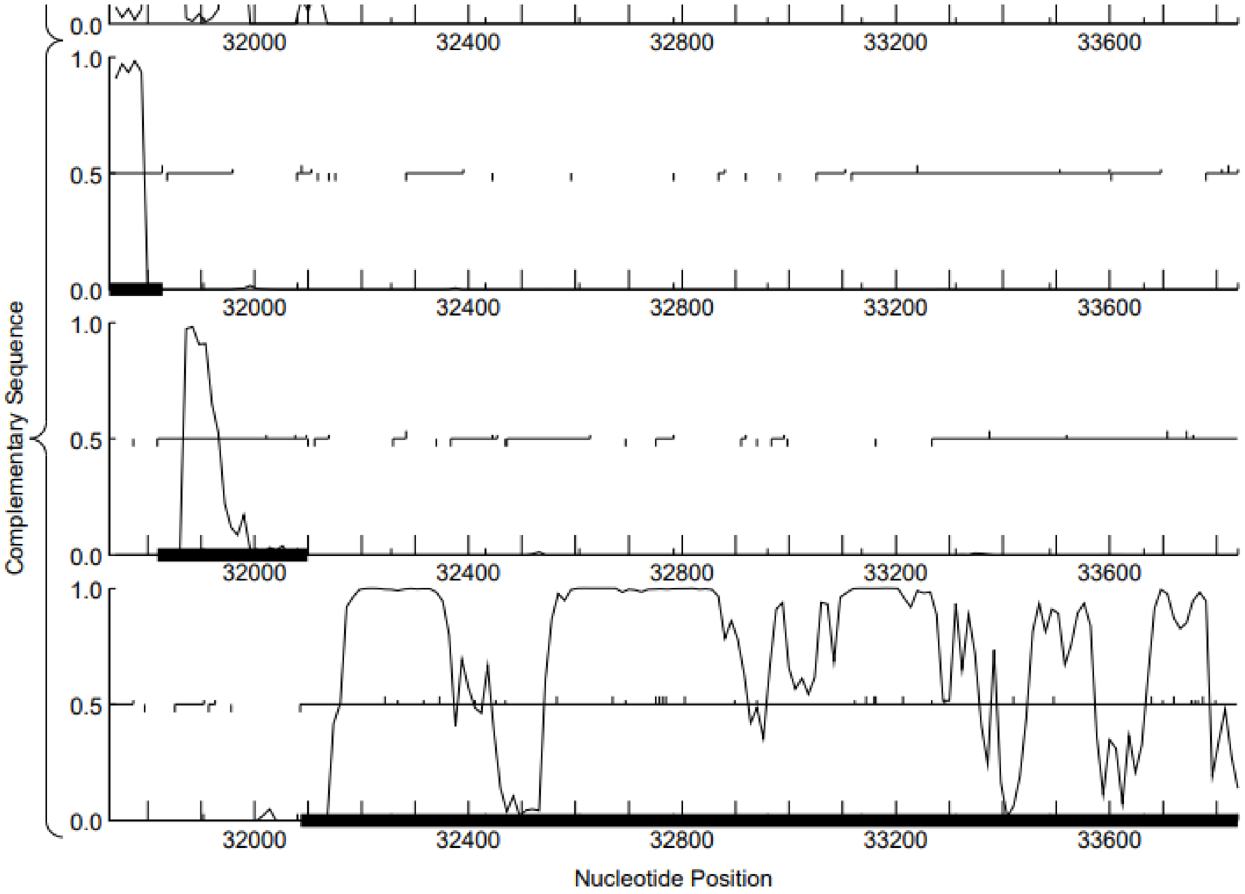
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

What is the autoannotated start?

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

- 46
- 31818
- Reverse
- Both glimmer and genemark
- 32099
- 14 overlap

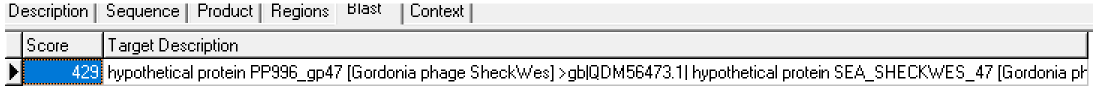
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



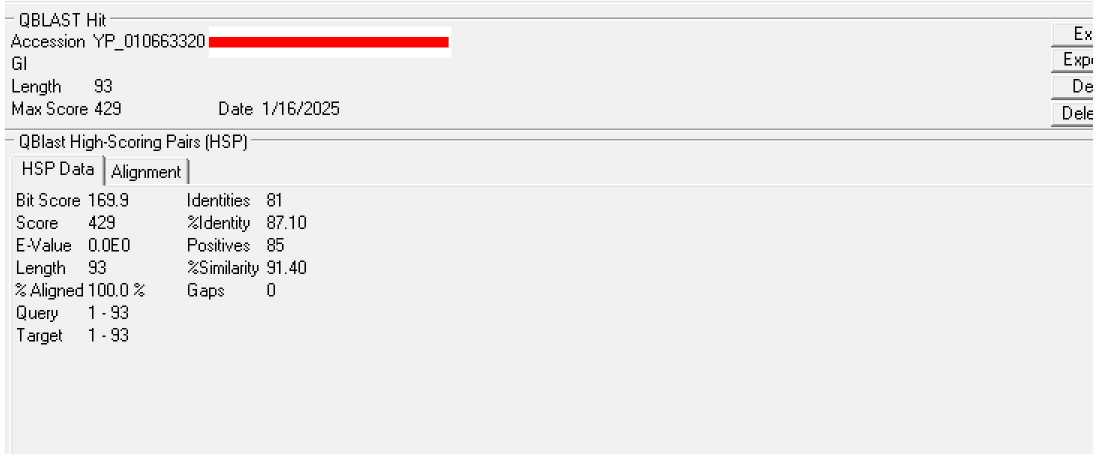
- 32099-31818

- Coding potential in reading frame 2 is

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
429	hypothetical protein PP996_gp47 [Gordonia phage SheckWes] >gb|QDM56473.1| hypothetical protein SEA_SHECKWES_47 [Gordonia ph



QBLAST Hit

Accession YP_010663320

GI

Length 93

Max Score 429 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	169.9	Identities	81
Score	429	%Identity	87.10
E-Value	0.0E0	Positives	85
Length	93	%Similarity	91.40
% Aligned	100.0 %	Gaps	0
Query	1 - 93		
Target	1 - 93		

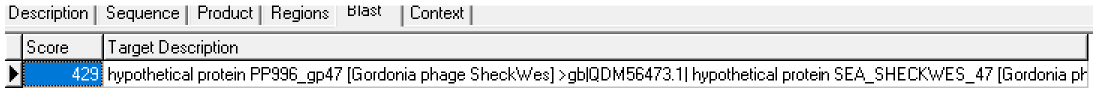
- One highly similar gene with E value of 0 (SheckWes).

Answer: Is it a gene? Give evidence why you think this is a gene or not.

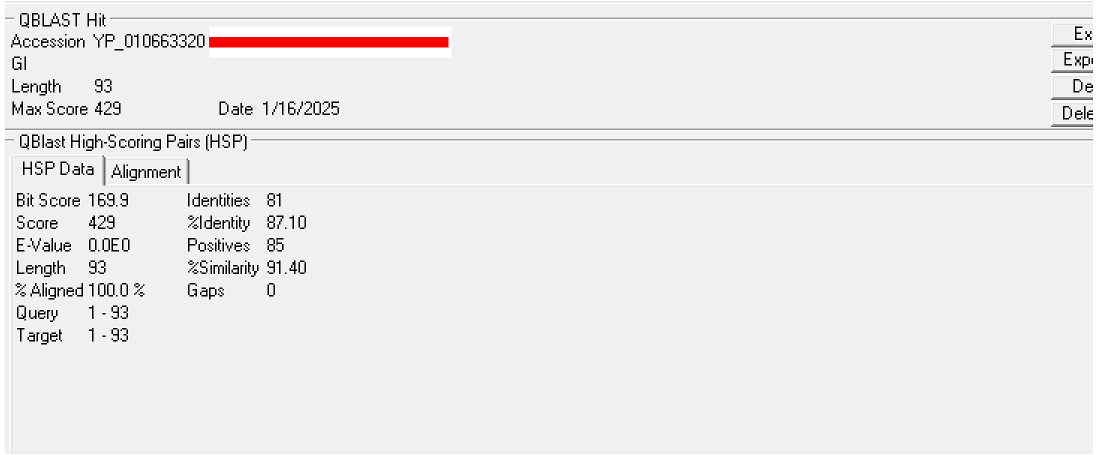
- It is a gene because:

- Both Glimmer and genemark called it a gene.
- Coding potential is strong.
- There is one highly similar gene with E value of 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
429	hypothetical protein PP996_gp47 [Gordonia phage SheckWes] >gb|QDM56473.1| hypothetical protein SEA_SHECKWES_47 [Gordonia ph



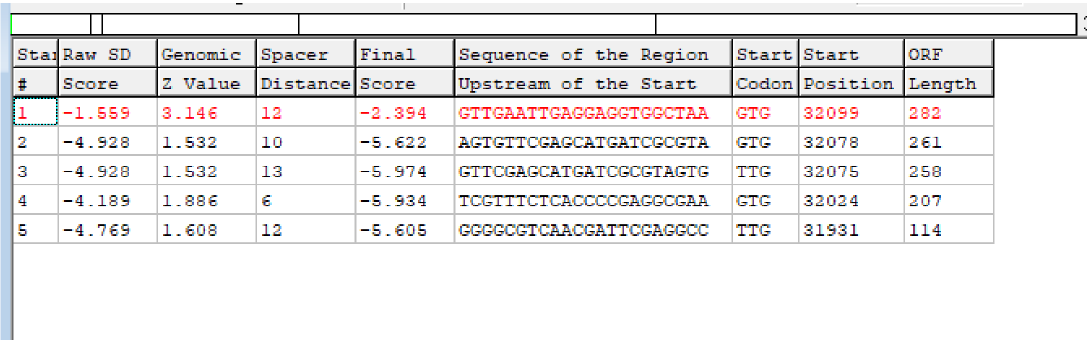
QBLAST Hit
Accession YP_010663320
GI
Length 93
Max Score 429 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	169.9	Identities	81
Score	429	%Identity	87.10
E-Value	0.0E0	Positives	85
Length	93	%Similarity	91.40
% Aligned	100.0 %	Gaps	0
Query	1 - 93		
Target	1 - 93		

- 1 1:1 alignment (SheckWes).

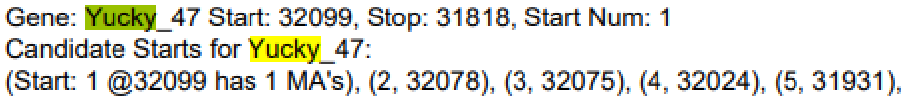
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-1.559	3.146	12	-2.394	GTTGAATTGAGGAGGTGGCTAA	GTG	32099	282
2	-4.928	1.532	10	-5.622	AGTGTTCGAGCATGATCGCGTA	GTG	32078	261
3	-4.928	1.532	13	-5.974	GTTCGAGCATGATCGCGTAGTG	TTG	32075	258
4	-4.189	1.886	6	-5.934	TCGTTTCTCACCCCGAGGCGAA	GTG	32024	207
5	-4.769	1.608	12	-5.605	GGGGCGTCAACGATTCGAGGCC	TTG	31931	114

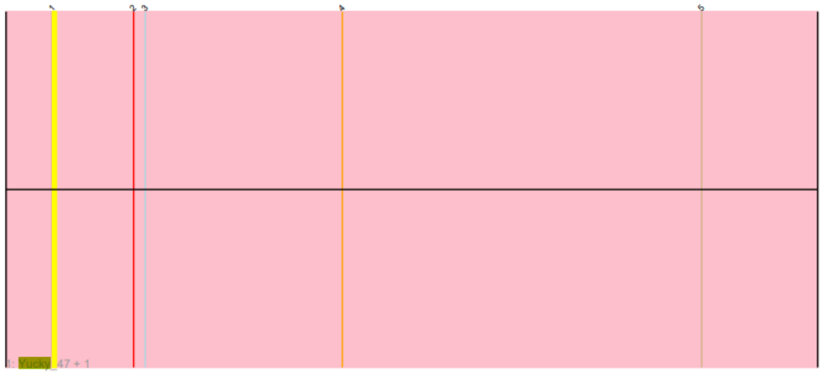
- RBS score favors.
- Z value is the greatest with 3.146
- Final score is the least negative with -2.394

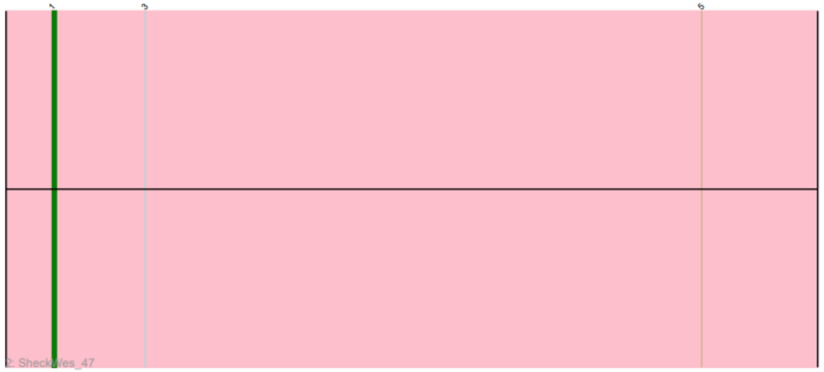
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



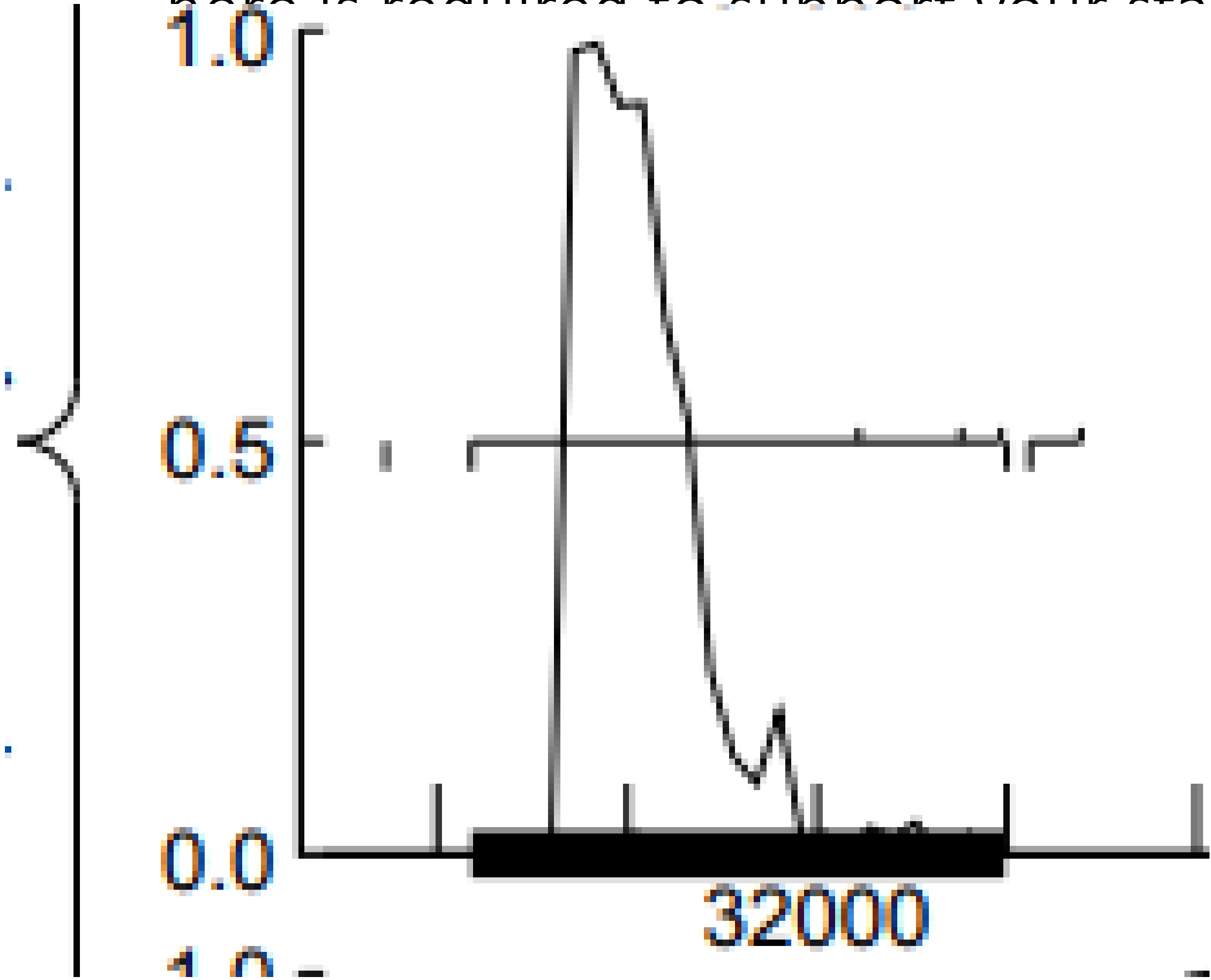
Gene: Yucky_47 Start: 32099, Stop: 31818, Start Num: 1
Candidate Starts for Yucky_47:
(Start: 1 @32099 has 1 MA's), (2, 32078), (3, 32075), (4, 32024), (5, 31931),

There is 1 MA for the autoannotated start site.





GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



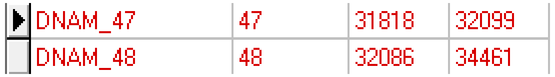
- Reading frame 2 show that coding potential is included.

- There is a really small coding potential if look close.

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

32099-32086 = 13

13+1 = 14 overlap



DNAM_47	47	31818	32099
DNAM_48	48	32086	34461

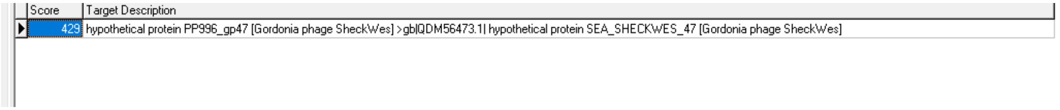
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	32099
GeneMark	Both Glimmer and GeneMark
Coding potential	Included
RBS score	Z-value: 3.146 Final score: -2.394
BLAST	1
Starterator	1
Gap/overlap	14 overlap

Nucleotide number 32099 is the start site because all of the factors favor it. Both Glimmer and GeneMark agree. Coding potential is included. RBS score favor the start site. There are 1 1:1 alignment and MA, which are better than nothing. Overlap is not huge.

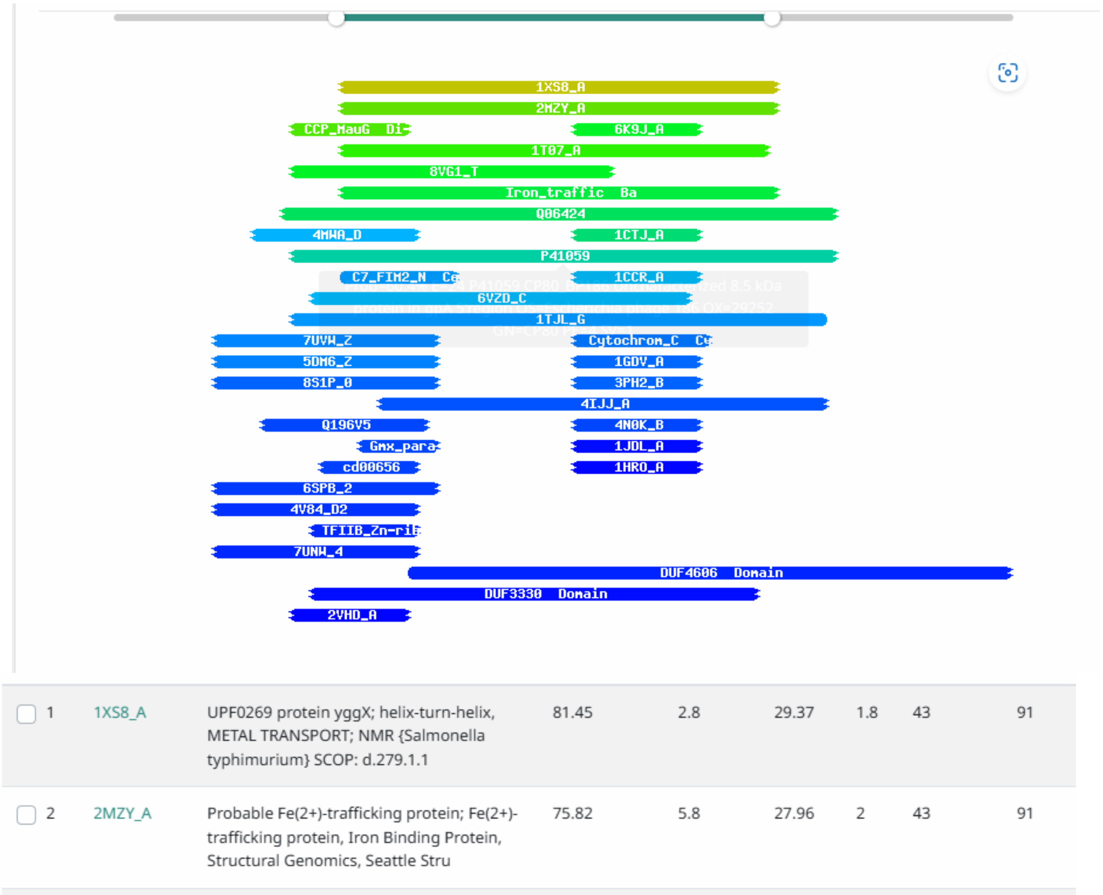
BLAST function evidence. What assigned functions do other highly similar genes have?

- Hypothetical protein.



Score	Target Description
429	hypothetical protein PP996_gp47 [Gordonia phage SheckWes] >gb|QDM56473.1| hypothetical protein SEA_SHECKWES_47 [Gordonia phage SheckWes]

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

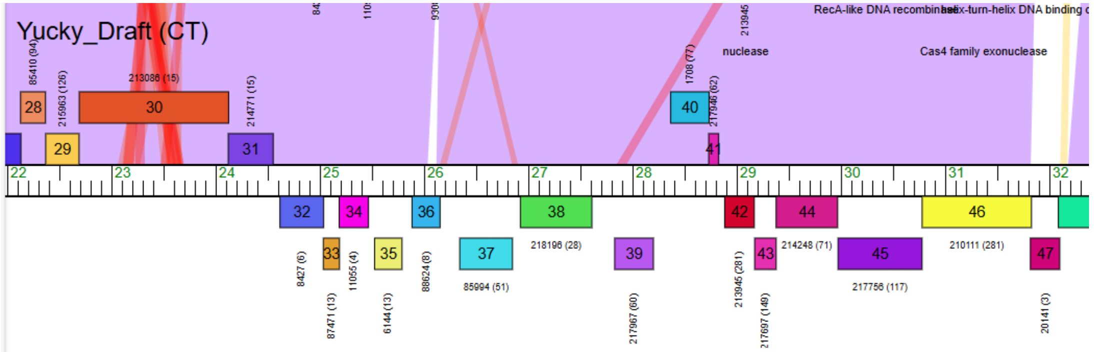


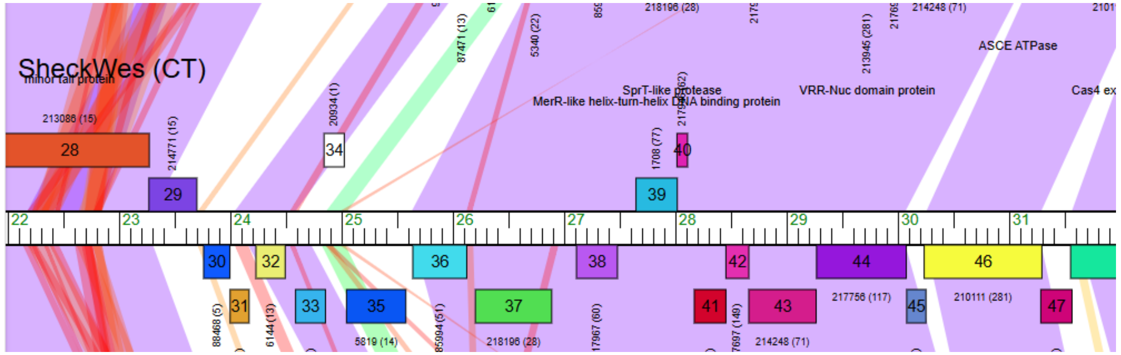
□	1	1XS8_A	UPF0269 protein yggX; helix-turn-helix, METAL TRANSPORT; NMR {Salmonella typhimurium} SCOP: d.279.1.1	81.45	2.8	29.37	1.8	43	91
□	2	2MZY_A	Probable Fe(2+)-trafficking protein; Fe(2+)-trafficking protein, Iron Binding Protein, Structural Genomics, Seattle Stru	75.82	5.8	27.96	2	43	91

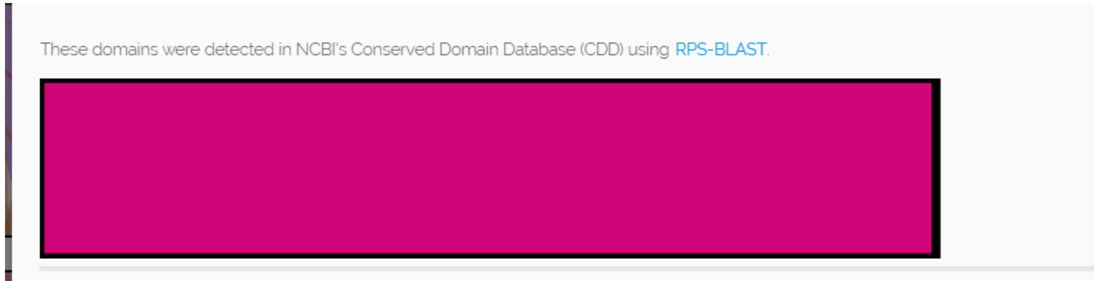
There are no hits with probability greater than 90.

Therefore, it is a hypothetical protein.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?





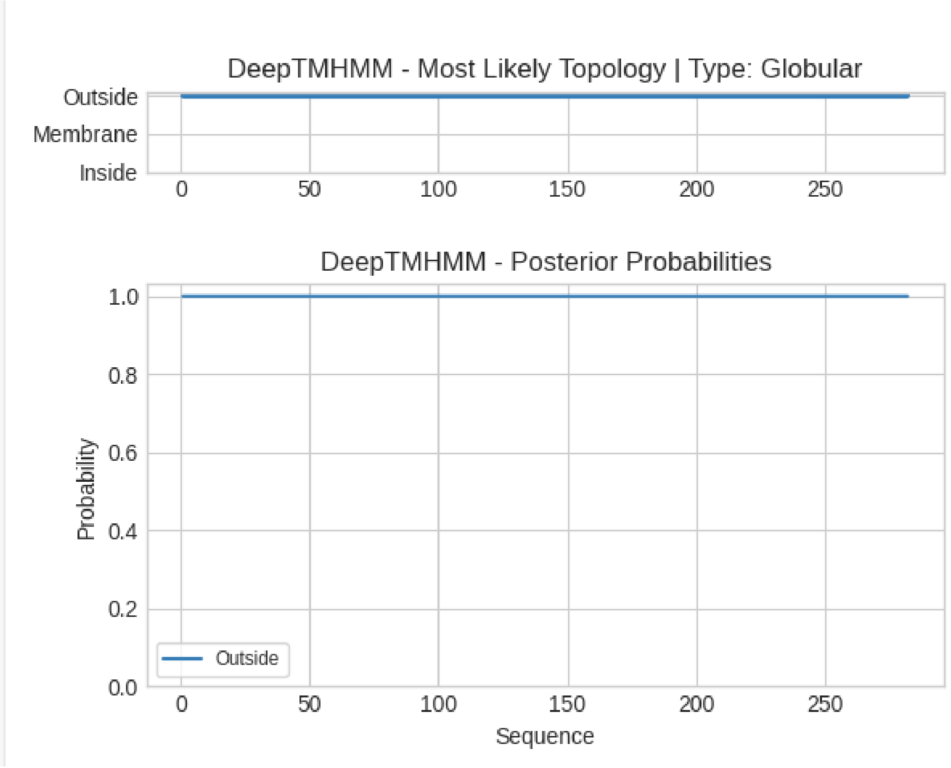


These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

There is only one highly similar gene in the same pham.

There is no assigned function either there.
There is no conserved domains.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- It is located outside the cell.
- Therefore, it is a hypothetical protein.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- It is a hypothetical protein because:

- BLAST function list shows a highly similar gene is a hypothetical protein.
- Hhpred does not show the hits with the probability greater than 90.
- Highly similar gene in the same pham does not have an assigned function.
- Deep TMHMM tells this gene is located outside the cell.

Feature 47 – Reverse – Stop 32086

Glimmer/GeneMark

What feature number is this?

What is the stop site?

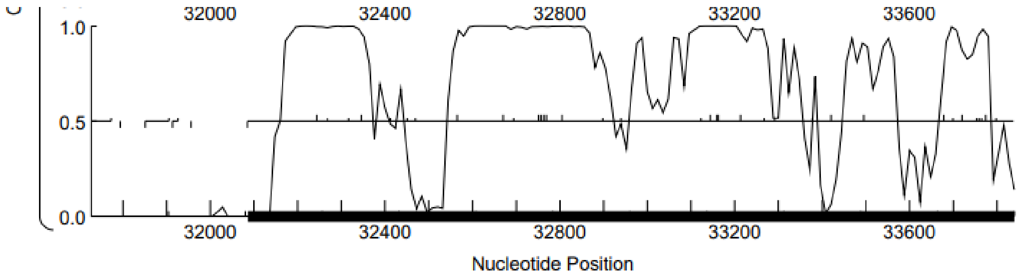
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

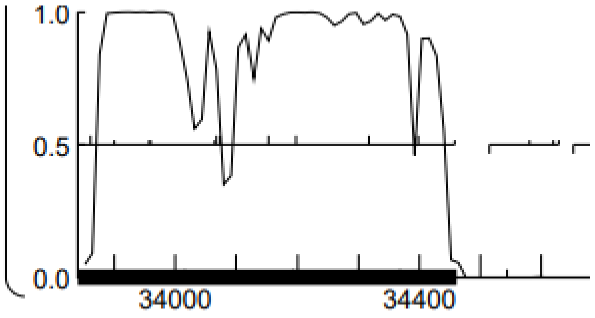
What is the autoannotated start?

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

- 47
- 32086
- Reverse
- Both Glimmer and genemark
- 34461
- 4 overlap

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

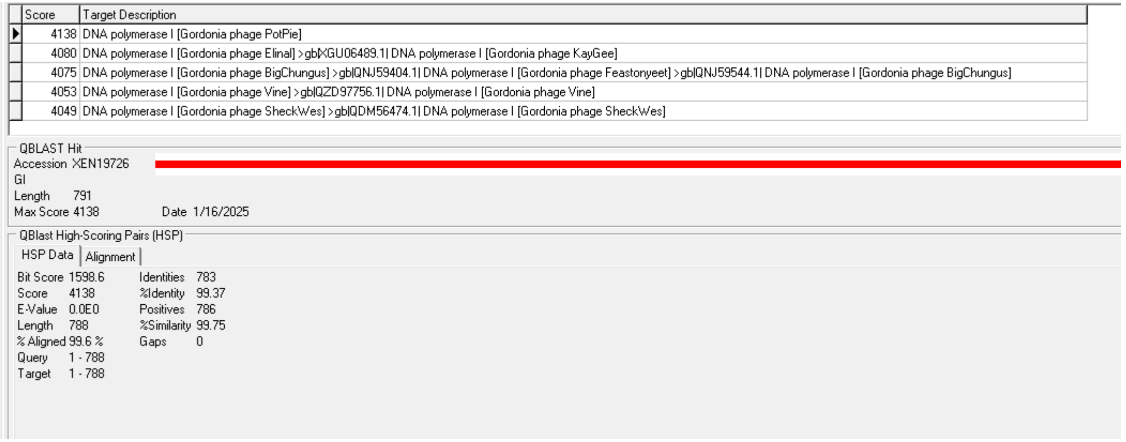




Coding potential is strong.

It starts around at 34470 and ends at 32090.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
4138	DNA polymerase I [Gordonia phage PotPie]
4080	DNA polymerase I [Gordonia phage Elinal] >gb|XGU06489.1| DNA polymerase I [Gordonia phage KayGee]
4075	DNA polymerase I [Gordonia phage BigChungus] >gb|QNJ59404.1| DNA polymerase I [Gordonia phage Feastonyeet] >gb|QNJ59544.1| DNA polymerase I [Gordonia phage BigChungus]
4053	DNA polymerase I [Gordonia phage Vine] >gb|QZD97756.1| DNA polymerase I [Gordonia phage Vine]
4049	DNA polymerase I [Gordonia phage SheckWes] >gb|QDM56474.1| DNA polymerase I [Gordonia phage SheckWes]

```
QBLAST Hit
Accession XEN19726
GI
Length      791
Max Score 4138          Date  1/16/2025

QBlast High-Scoring Pairs (HSP)
HSP Data | Alignment
Bit Score 1598.6      Identities   783
Score      4138       %Identity    99.37
E-Value    0.0E0      Positives    786
Length     788        %Similarity  99.75
% Aligned 99.6 %      Gaps         0
Query      1 - 788
Target     1 - 788
```

- There are 25 highly similar genes with E value of 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

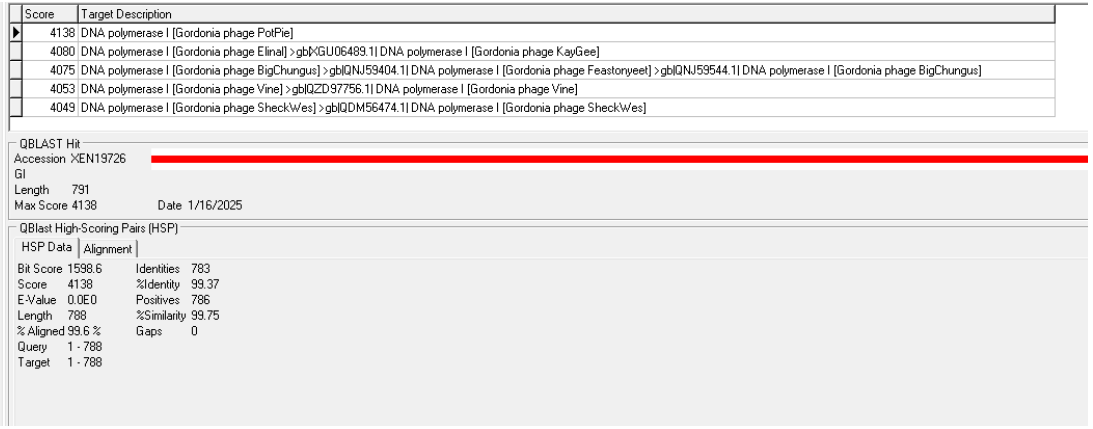
- It is a gene because:

Both glimmer and genemark called it a gene.

Coding potential is strong.

There are 25 highly similar genes with E value of 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



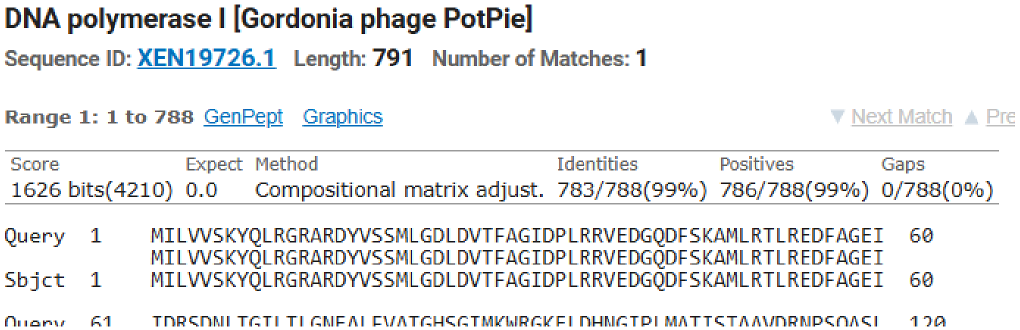
Score	Target Description
4138	DNA polymerase I [Gordonia phage PotPie]
4080	DNA polymerase I [Gordonia phage Elinal] >gb|XGU06489.1| DNA polymerase I [Gordonia phage KayGee]
4075	DNA polymerase I [Gordonia phage BigChungus] >gb|QNJ59404.1| DNA polymerase I [Gordonia phage Feastonyeet] >gb|QNJ59544.1| DNA polymerase I [Gordonia phage BigChungus]
4053	DNA polymerase I [Gordonia phage Vine] >gb|QZD97756.1| DNA polymerase I [Gordonia phage Vine]
4049	DNA polymerase I [Gordonia phage SheckWes] >gb|QDM56474.1| DNA polymerase I [Gordonia phage SheckWes]

QBLAST Hit

```
Accession XEN19726
GI
Length     791
Max Score 4138           Date 1/16/2025
```

QBlast High-Scoring Pairs (HSP)

```
Bit Score 1598.6    Identities  783
Score     4138      %Identity   99.37
E-Value   0.0E0     Positives   786
Length    788       %Similarity 99.75
% Aligned 99.6 %    Gaps        0
Query     1 - 788
Target    1 - 788
```



DNA polymerase I [Gordonia phage PotPie]

Sequence ID: XEN19726.1 Length: 791 Number of Matches: 1

Range 1: 1 to 788 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
1626 bits(4210)	0.0	Compositional matrix adjust.	783/788(99%)	786/788(99%)	0/788(0%)

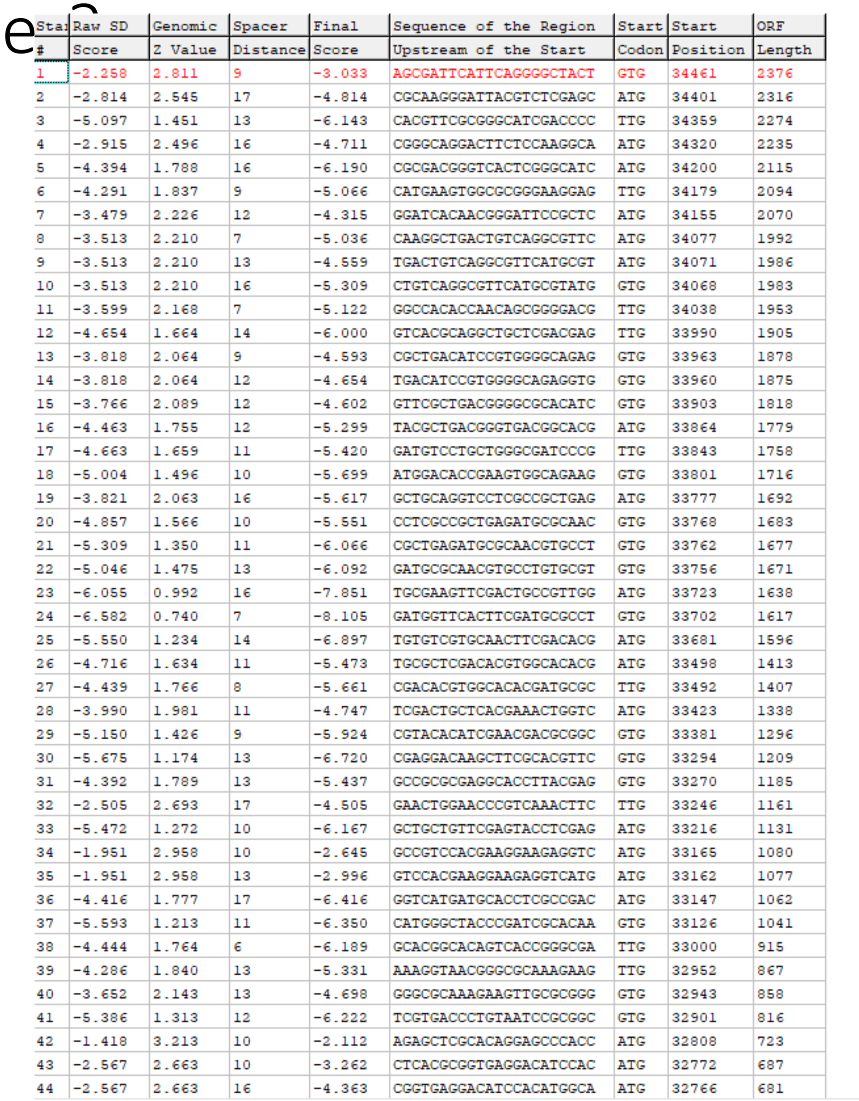
```
Query  1    MILVVSKYQLRGRARDYVSSMLGDLDVTFAGIDPLRRVEDGQDFSKAMLRTLREDFAGEI  60
            MILVVSKYQLRGRARDYVSSMLGDLDVTFAGIDPLRRVEDGQDFSKAMLRTLREDFAGEI
Sbjct  1    MILVVSKYQLRGRARDYVSSMLGDLDVTFAGIDPLRRVEDGQDFSKAMLRTLREDFAGEI  60

Query  61   IDRSDNLTGILTLGNEALEVATGHSGIMKWRGKELDHNGIPLMATTSTAAVDRNPSQASL  120
```

- There are 11 1:1 alignments.

- 33960:

- There are 11 1:1 alignments.

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts?   Screenshot RBS Values here. Answer the question: Which start is favored based on RBS value ?



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	-2.258	2.811	9	-3.033	AGCGATTCATTCAGGGGCTACT	GTG	34461	2376
2	-2.814	2.545	17	-4.814	CGCAAGGGATTACGTCTCGAGC	ATG	34401	2316
3	-5.097	1.451	13	-6.143	CACGTTCGCGGGCATCGACCCC	TTG	34359	2274
4	-2.915	2.496	16	-4.711	CGGGCAGGACTTCTCCAAGGCA	ATG	34320	2235
5	-4.394	1.788	16	-6.190	CGCGACGGGTCACTCGGGCATC	ATG	34200	2115
6	-4.291	1.837	9	-5.066	CATGAAGTGGCGCGGGAAGGAG	TTG	34179	2094
7	-3.479	2.226	12	-4.315	GGATCACAACGGGATTCCGCTC	ATG	34155	2070
8	-3.513	2.210	7	-5.036	CAAGGCTGACTGTCAGGCGTTC	ATG	34077	1992
9	-3.513	2.210	13	-4.559	TGACTGTCAGGCGTTCATGCGT	ATG	34071	1986
10	-3.513	2.210	16	-5.309	CTGTCAGGCGTTCATGCGTATG	GTG	34068	1983
11	-3.599	2.168	7	-5.122	GGCCACACCAACAGCGGGGACG	TTG	34038	1953
12	-4.654	1.664	14	-6.000	GTCACGCAGGCTGCTCGACGAG	TTG	33990	1905
13	-3.818	2.064	9	-4.593	CGCTGACATCCGTGGGGCAGAG	GTG	33963	1878
14	-3.818	2.064	12	-4.654	TGACATCCGTGGGGCAGAGGTG	GTG	33960	1875
15	-3.766	2.089	12	-4.602	GTTCGCTGACGGGGCGCACATC	GTG	33903	1818
16	-4.463	1.755	12	-5.299	TACGCTGACGGGTGACGGCACG	ATG	33864	1779
17	-4.663	1.659	11	-5.420	GATGTCCTGCTGGGCGATCCCG	TTG	33843	1758
18	-5.004	1.496	10	-5.699	ATGGACACCGAAGTGGCAGAAG	GTG	33801	1716
19	-3.821	2.063	16	-5.617	GCTGCAGGTCCTCGCCGCTGAG	ATG	33777	1692
20	-4.857	1.566	10	-5.551	CCTCGCCGCTGAGATGCGCAAC	GTG	33768	1683
21	-5.309	1.350	11	-6.066	CGCTGAGATGCGCAACGTGCCT	GTG	33762	1677
22	-5.046	1.475	13	-6.092	GATGCGCAACGTGCCTGTGCGT	GTG	33756	1671
23	-6.055	0.992	16	-7.851	TGCGAAGTTCGACTGCCGTTGG	ATG	33723	1638
24	-6.582	0.740	7	-8.105	GATGGTTCACTTCGATGCGCCT	GTG	33702	1617
25	-5.550	1.234	14	-6.897	TGTGTCGTGCAACTTCGACACG	ATG	33681	1596
26	-4.716	1.634	11	-5.473	TGCGCTCGACACGTGGCACACG	ATG	33498	1413
27	-4.439	1.766	8	-5.661	CGACACGTGGCACACGATGCGC	TTG	33492	1407
28	-3.990	1.981	11	-4.747	TCGACTGCTCACGAAACTGGTC	ATG	33423	1338
29	-5.150	1.426	9	-5.924	CGTACACATCGAACGACGCGGC	GTG	33381	1296
30	-5.675	1.174	13	-6.720	CGAGGACAAGCTTCGCACGTTC	GTG	33294	1209
31	-4.392	1.789	13	-5.437	GCCGCGCGAGGCACCTTACGAG	GTG	33270	1185
32	-2.505	2.693	17	-4.505	GAACTGGAACCCGTCAAACTTC	TTG	33246	1161
33	-5.472	1.272	10	-6.167	GCTGCTGTTCGAGTACCTCGAG	ATG	33216	1131
34	-1.951	2.958	10	-2.645	GCCGTCCACGAAGGAAGAGGTC	ATG	33165	1080
35	-1.951	2.958	13	-2.996	GTCCACGAAGGAAGAGGTCATG	ATG	33162	1077
36	-4.416	1.777	17	-6.416	GGTCATGATGCACCTCGCCGAC	ATG	33147	1062
37	-5.593	1.213	11	-6.350	CATGGGCTACCCGATCGCACAA	GTG	33126	1041
38	-4.444	1.764	6	-6.189	GCACGGCACAGTCACCGGGCGA	TTG	33000	915
39	-4.286	1.840	13	-5.331	AAAGGTAACGGGCGCAAAGAAG	TTG	32952	867
40	-3.652	2.143	13	-4.698	GGGCGCAAAGAAGTTGCGCGGG	GTG	32943	858
41	-5.386	1.313	12	-6.222	TCGTGACCCTGTAATCCGCGGC	GTG	32901	816
42	-1.418	3.213	10	-2.112	AGAGCTCGCACAGGAGCCCACC	ATG	32808	723
43	-2.567	2.663	10	-3.262	CTCACGCGGTGAGGACATCCAC	ATG	32772	687
44	-2.567	2.663	16	-4.363	CGGTGAGGACATCCACATGGCA	ATG	32766	681

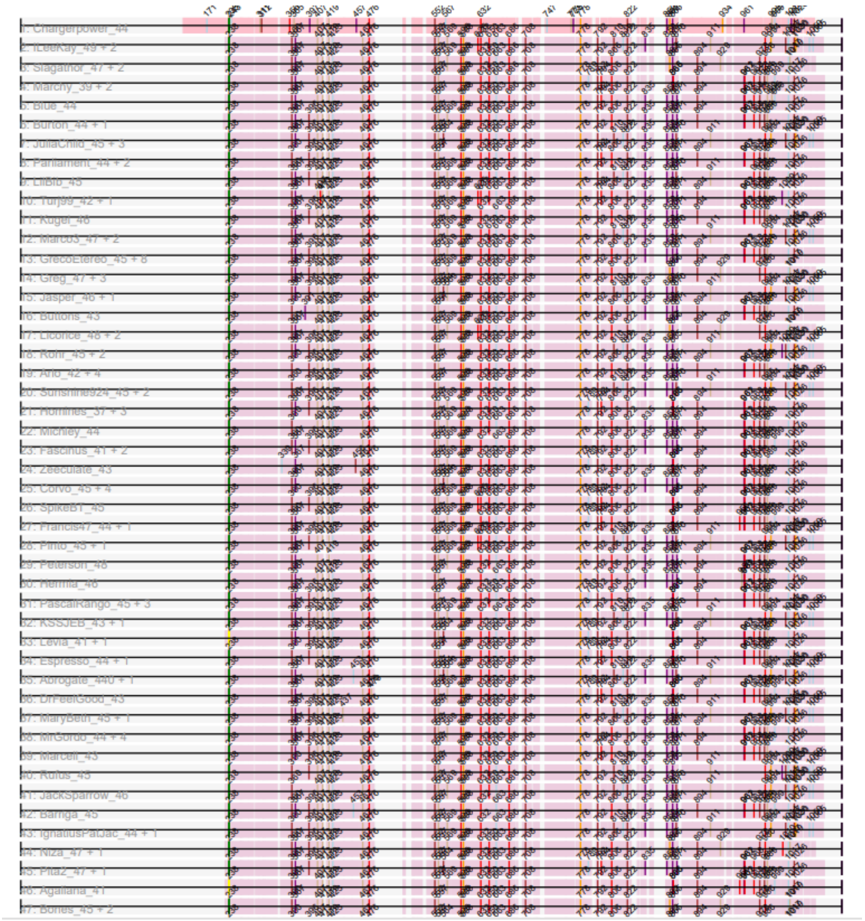
- Autoannotated start site is favored by the RBS evidence.
- Z value is the greatest with 2.811.
- Final score is the least negative with -3.033

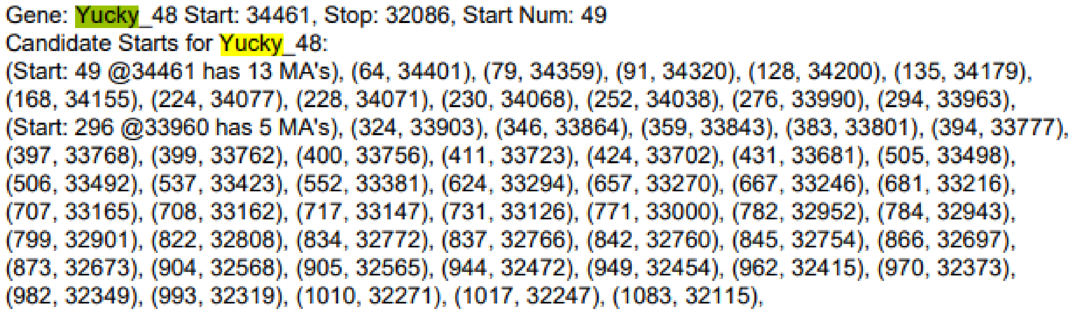
33960:

Z value of 2.064

Final score of -4.564

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

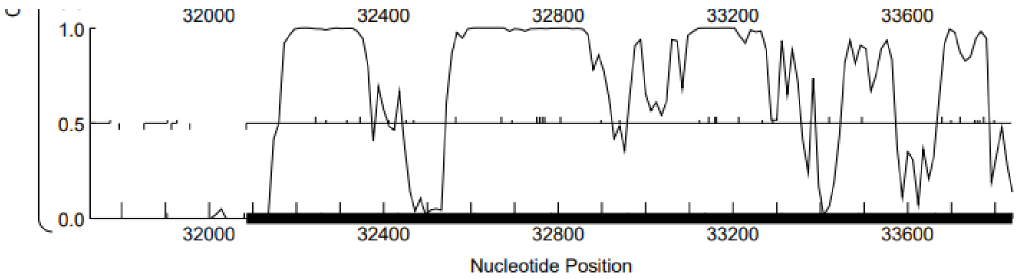


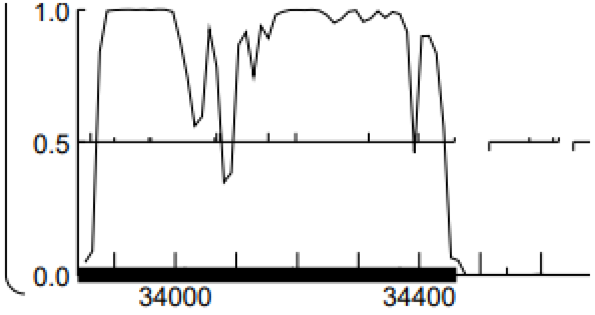

Gene: Yucky_48 Start: 34461, Stop: 32086, Start Num: 49
Candidate Starts for Yucky_48:
(Start: 49 @34461 has 13 MA's), (64, 34401), (79, 34359), (91, 34320), (128, 34200), (135, 34179), (168, 34155), (224, 34077), (228, 34071), (230, 34068), (252, 34038), (276, 33990), (294, 33963), (Start: 296 @33960 has 5 MA's), (324, 33903), (346, 33864), (359, 33843), (383, 33801), (394, 33777), (397, 33768), (399, 33762), (400, 33756), (411, 33723), (424, 33702), (431, 33681), (505, 33498), (506, 33492), (537, 33423), (552, 33381), (624, 33294), (657, 33270), (667, 33246), (681, 33216), (707, 33165), (708, 33162), (717, 33147), (731, 33126), (771, 33000), (782, 32952), (784, 32943), (799, 32901), (822, 32808), (834, 32772), (837, 32766), (842, 32760), (845, 32754), (866, 32697), (873, 32673), (904, 32568), (905, 32565), (944, 32472), (949, 32454), (962, 32415), (970, 32373), (982, 32349), (993, 32319), (1010, 32271), (1017, 32247), (1083, 32115),

There are 13 MA’s at the autoannotated start site.

Starterator proposed start site at 33960 has 5 MA’s.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

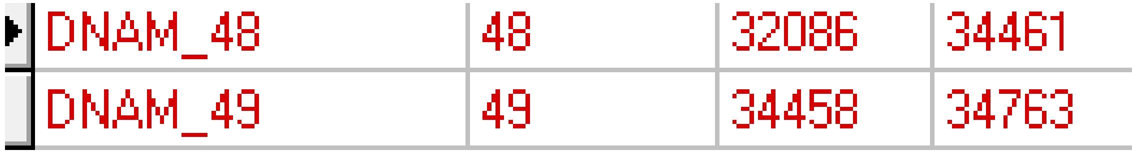




All of the coding potential is included at the autoannotated start site.

Coding potential is cut off for the starterator proposed start site.

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.



DNAM_48	48	32086	34461
DNAM_49	49	34458	34763

- Autoannotated start site:

34461-34458 = 3

3+1 = 4 overlap

33960:

34458-33960 = 498

498-1 = 497 gap

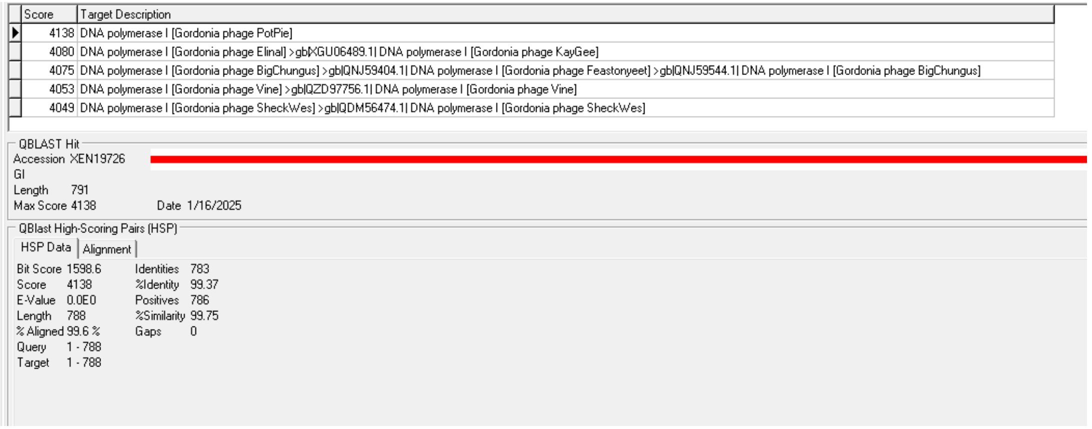
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	34461	33960
GeneMark	**Both Glimmer and GeneMark**	NA
Coding potential	**Included**	Cut off
RBS score	**Z-value: 2.811** **Final score: - 3.033**	Z-value: 2.064 Final score: -4.564
BLAST	**11**	**11**
Starterator	**13**	5
Gap/overlap	**4 overlap**	497 gap

Start site at 34461 is favored because all evidence favor it. Also, it only has 4 overlap.
33960 has too many nucleotides gap.

BLAST function evidence. What assigned functions do other highly similar genes have?

- Similar genes are DNA polymerase.



Score	Target Description
4138	DNA polymerase I [Gordonia phage PotPie]
4080	DNA polymerase I [Gordonia phage Elinal] >gb|XGU06489.1| DNA polymerase I [Gordonia phage KayGee]
4075	DNA polymerase I [Gordonia phage BigChungus] >gb|QNJ59404.1| DNA polymerase I [Gordonia phage Feastonyeet] >gb|QNJ59544.1| DNA polymerase I [Gordonia phage BigChungus]
4053	DNA polymerase I [Gordonia phage Vine] >gb|QZD97756.1| DNA polymerase I [Gordonia phage Vine]
4049	DNA polymerase I [Gordonia phage SheckWes] >gb|QDM56474.1| DNA polymerase I [Gordonia phage SheckWes]

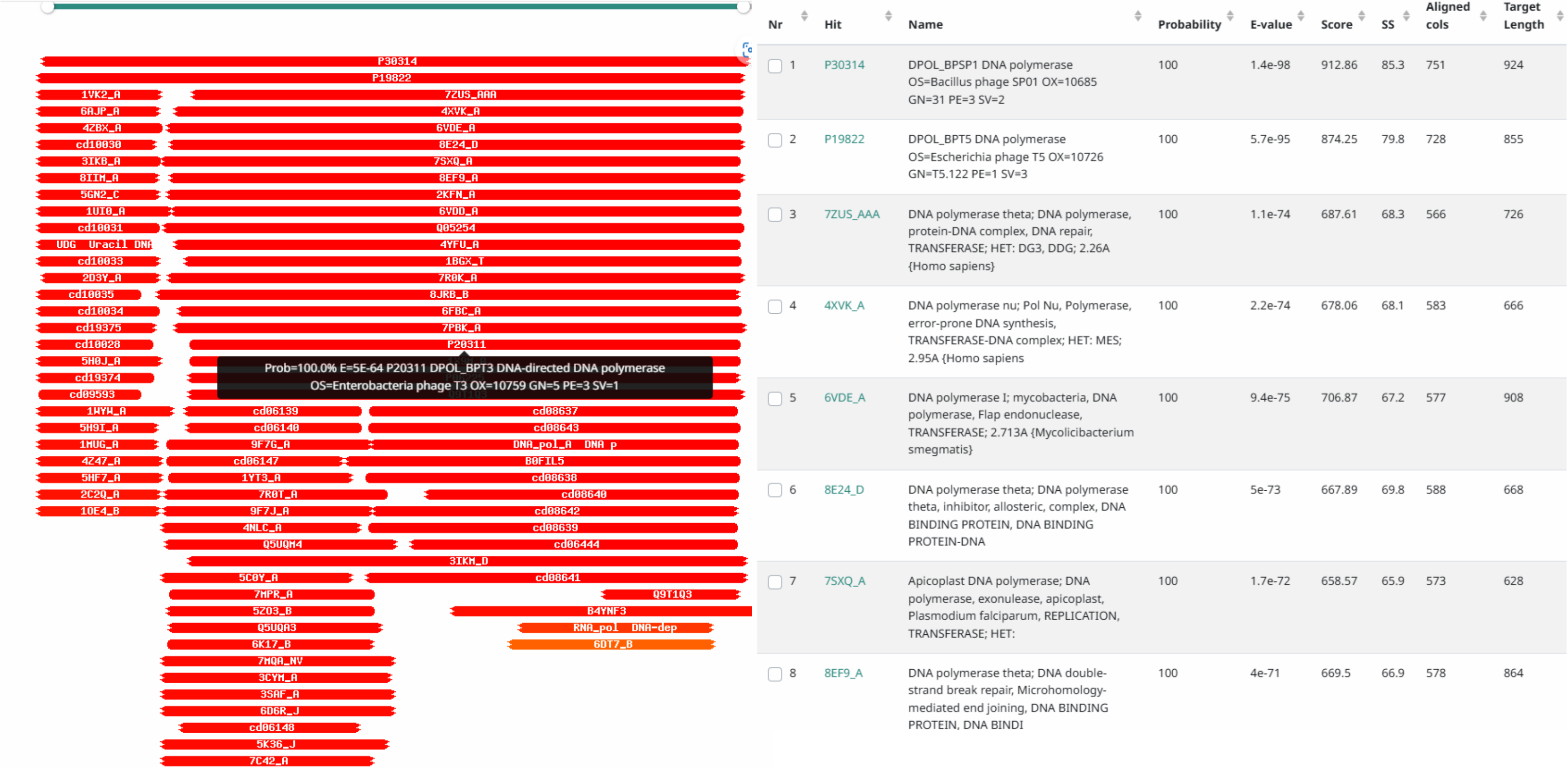
QBLAST Hit

```
Accession XEN19726
GI
Length     791
Max Score 4138          Date 1/16/2025
```

QBlast High-Scoring Pairs (HSP)

```
Bit Score 1598.6      Identities  783
Score     4138        %Identity   99.37
E-Value   0.0E0       Positives   786
Length    788         %Similarity 99.75
% Aligned 99.6 %      Gaps        0
Query     1 - 788
Target    1 - 788
```

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

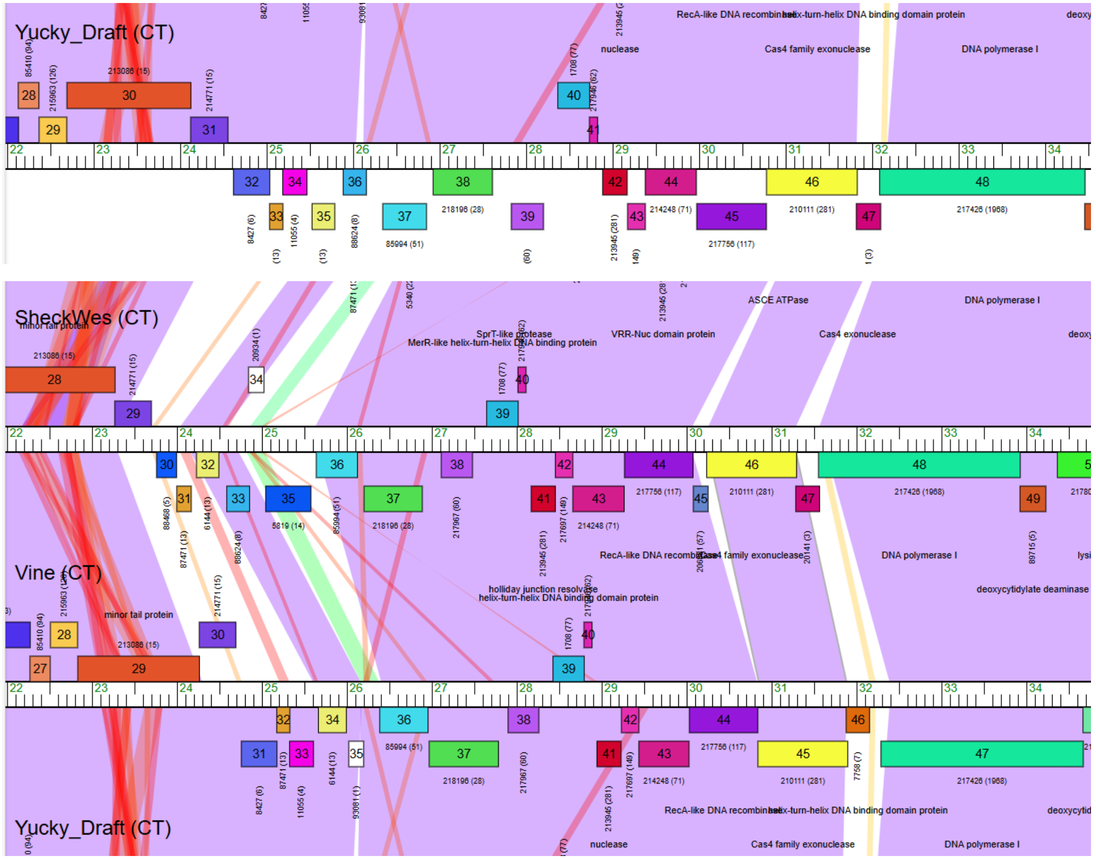


Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	P30314	DPOL_BPSP1 DNA polymerase OS=Bacillus phage SP01 OX=10685 GN=31 PE=3 SV=2	100	1.4e-98	912.86	85.3	751	924
□ 2	P19822	DPOL_BPT5 DNA polymerase OS=Escherichia phage T5 OX=10726 GN=T5.122 PE=1 SV=3	100	5.7e-95	874.25	79.8	728	855
□ 3	7ZUS_AAA	DNA polymerase theta; DNA polymerase, protein-DNA complex, DNA repair, TRANSFERASE; HET: DG3, DDG; 2.26A {Homo sapiens}	100	1.1e-74	687.61	68.3	566	726
□ 4	4XVK_A	DNA polymerase nu; Pol Nu, Polymerase, error-prone DNA synthesis, TRANSFERASE-DNA complex; HET: MES; 2.95A {Homo sapiens	100	2.2e-74	678.06	68.1	583	666
□ 5	6VDE_A	DNA polymerase I; mycobacteria, DNA polymerase, Flap endonuclease, TRANSFERASE; 2.713A {Mycolicibacterium smegmatis}	100	9.4e-75	706.87	67.2	577	908
□ 6	8E24_D	DNA polymerase theta; DNA polymerase theta, inhibitor, allosteric, complex, DNA BINDING PROTEIN, DNA BINDING PROTEIN-DNA	100	5e-73	667.89	69.8	588	668
□ 7	7SXQ_A	Apicoplast DNA polymerase; DNA polymerase, exonulease, apicoplast, Plasmodium falciparum, REPLICATION, TRANSFERASE; HET:	100	1.7e-72	658.57	65.9	573	628
□ 8	8EF9_A	DNA polymerase theta; DNA double-strand break repair, Microhomology-mediated end joining, DNA BINDING PROTEIN, DNA BINDI	100	4e-71	669.5	66.9	578	864

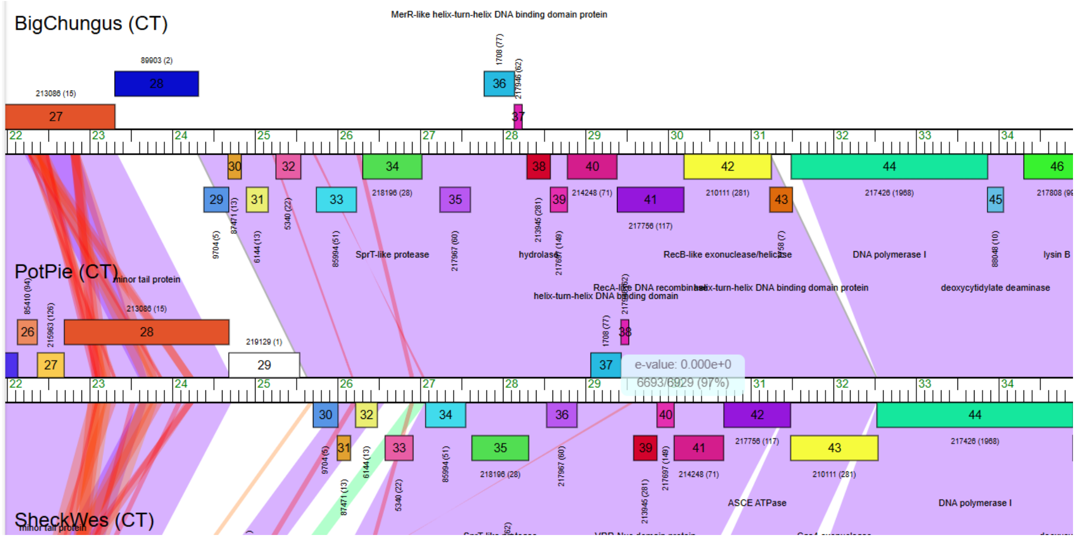
There are many hits with probability greater than 90.

Parts of hits suggest that it is a DNA polymerase I.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



Yucky_Draft (CT)
RecA-like DNA recombinase
helix-turn-helix DNA binding domain protein
nuclease
Cas4 family exonuclease
DNA polymerase I
SheckWes (CT)
minor tail protein
SprT-like protease
MerR-like helix-turn-helix DNA binding domain protein
VRR-Nuc domain protein
ASCE ATPase
Cas4 exonuclease
DNA polymerase I
Vine (CT)
minor tail protein
holliday junction resolvase
helix-turn-helix DNA binding domain protein
RecA-like DNA recombinase
Cas4 family exonuclease
DNA polymerase I
deoxycytidylate deaminase
Yucky_Draft (CT)
RecA-like DNA recombinase
helix-turn-helix DNA binding domain protein
nuclease
Cas4 family exonuclease
DNA polymerase I



BigChungus (CT)
MerR-like helix-turn-helix DNA binding domain protein
SprT-like protease
hydrolase
RecB-like exonuclease/helicase
DNA polymerase I
lysin B
PotPie (CT)
minor tail protein
RecA-like DNA recombinase
helix-turn-helix DNA binding domain protein
helix-turn-helix DNA binding domain
deoxycytidylate deaminase
e-value: 0.000e+0
6693/6929 (97%)
ASCE ATPase
DNA polymerase I
SheckWes (CT)
minor tail protein

All highly similar genes within the same pham are assigned a function of DNA polymerase I.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- No

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- It is a DNA polymerase I because:

- BLAST suggest it is a DNA polymerase, close to DNA polymerase I.
- Hhpred suggests part of the hits call it a DNA polymerase I.
- Phamerator calls other similar genes in the same pham DNA polymerase I.

Feature 48 – Reverse – Stop 34458

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

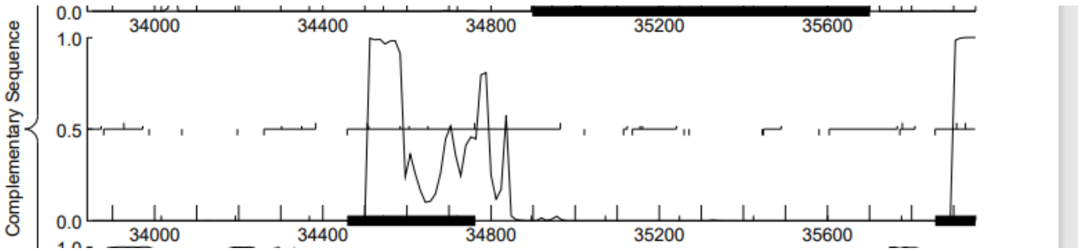
What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

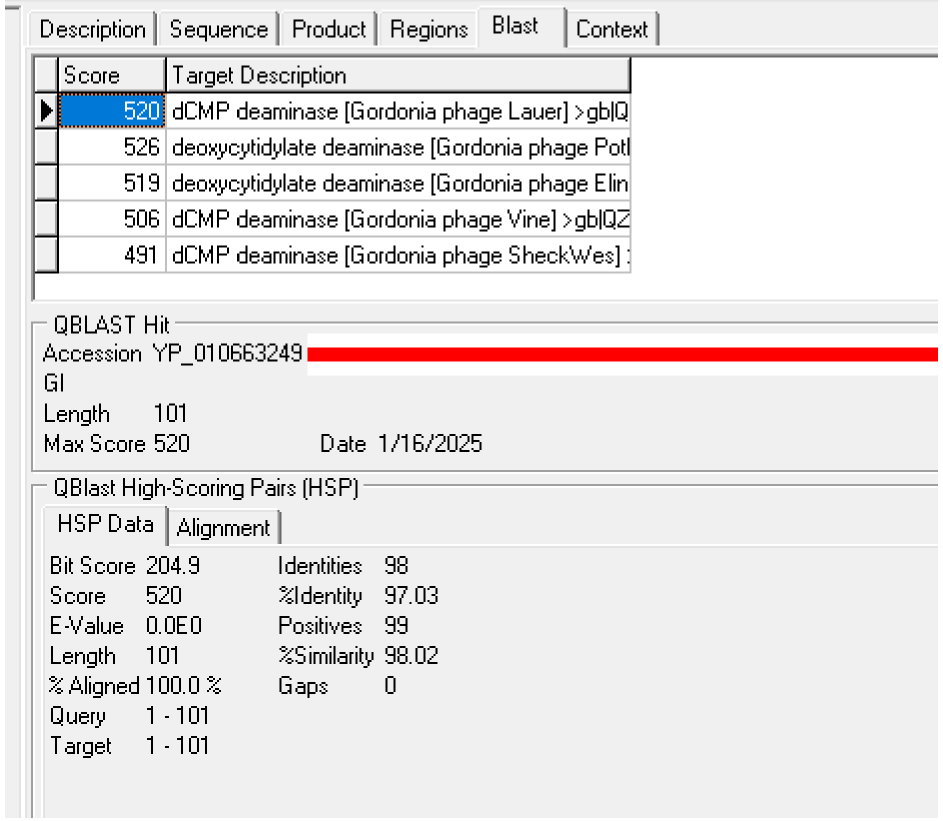
- 48
- Reverse
- 34458
- Both Glimmer and GeneMark
- 34763
- 134 gap

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Coding potential in reading frame -2 is strong.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
520	dCMP deaminase [Gordonia phage Lauer] >gb|Q
526	deoxycytidylate deaminase [Gordonia phage Potl
519	deoxycytidylate deaminase [Gordonia phage Elin
506	dCMP deaminase [Gordonia phage Vine] >gb|QZ
491	dCMP deaminase [Gordonia phage SheckWes] :

QBLAST Hit

Accession YP_010663249

GI

Length 101

Max Score 520 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	204.9	Identities	98
Score	520	%Identity	97.03
E-Value	0.0E0	Positives	99
Length	101	%Similarity	98.02
% Aligned	100.0 %	Gaps	0
Query	1 - 101		
Target	1 - 101		

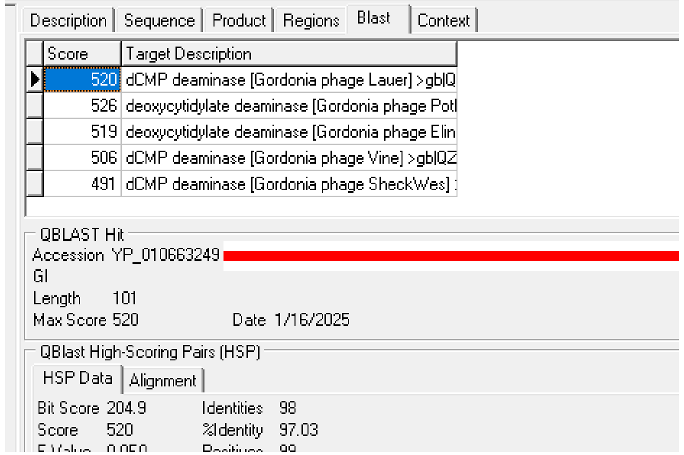
- There are 25 highly similar genes with E value of 0 or less than 10-7.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- It is a gene because:

- Both Glimmer and GeneMark call it a gene.
- Coding potential is strong.
- There are 25 highly similar genes with E value of close to 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



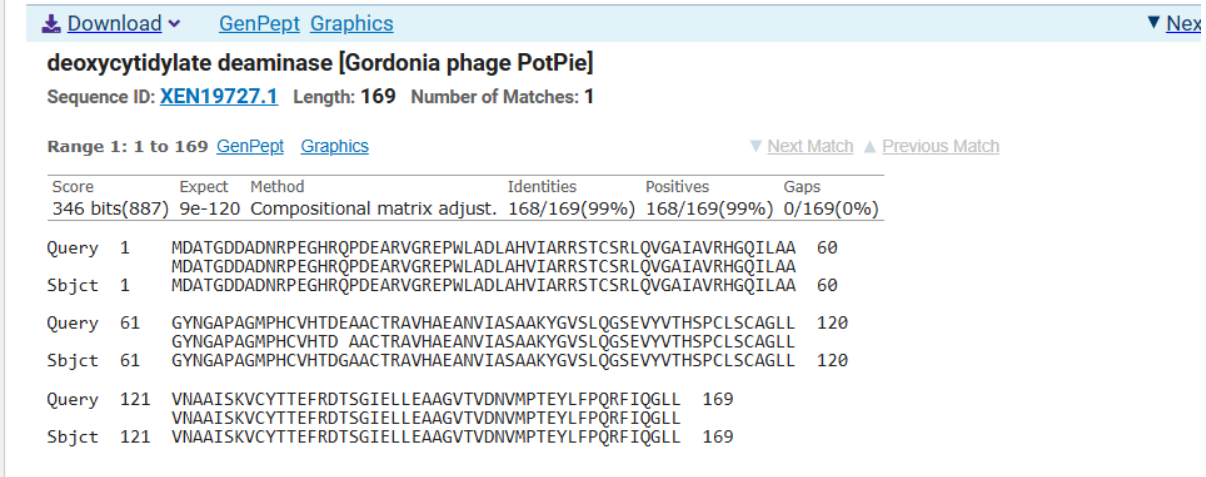
Score	Target Description
520	dCMP deaminase [Gordonia phage Lauer] >gb|Q
526	deoxycytidylate deaminase [Gordonia phage Pot|
519	deoxycytidylate deaminase [Gordonia phage Elin
506	dCMP deaminase [Gordonia phage Vine] >gb|QZ
491	dCMP deaminase [Gordonia phage SheckWes] :

QBLAST Hit
Accession YP_010663249
GI
Length 101
Max Score 520 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 204.9 Identities 98
Score 520 %Identity 97.03
E Value 0.050 Positives 99

- There are three 1:1 alignments.



deoxycytidylate deaminase [Gordonia phage PotPie]

Sequence ID: XEN19727.1 Length: 169 Number of Matches: 1

Range 1: 1 to 169 GenPept Graphics

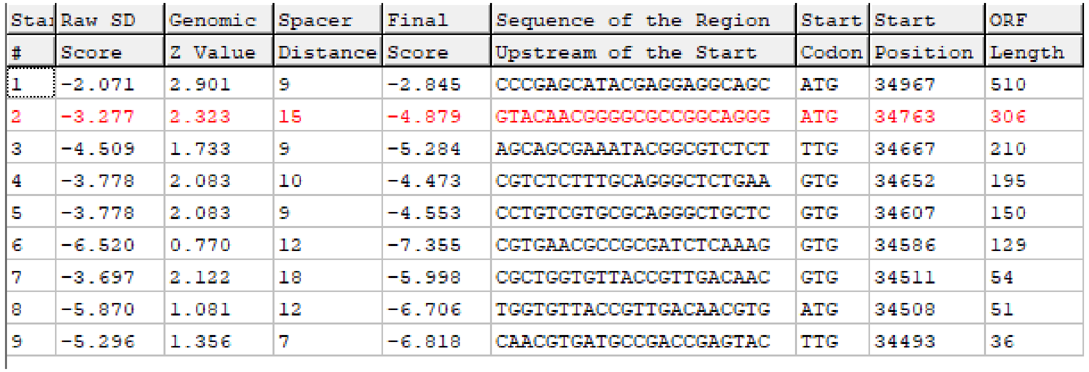
```
Score            Expect  Method                        Identities     Positives      Gaps
346 bits(887)    9e-120  Compositional matrix adjust.  168/169(99%)   168/169(99%)   0/169(0%)

Query  1    MDATGDDADNRPEGHRQPDEARVGREPWLADLAHVIARRSTCSRLQVGAIAVRHGQILAA  60
            MDATGDDADNRPEGHRQPDEARVGREPWLADLAHVIARRSTCSRLQVGAIAVRHGQILAA
Sbjct  1    MDATGDDADNRPEGHRQPDEARVGREPWLADLAHVIARRSTCSRLQVGAIAVRHGQILAA  60

Query  61   GYNGAPAGMPHCVHTDEAACTRAVHAEANVIASAAKYGVSLQGSEVYVTHSPCLSCAGLL  120
            GYNGAPAGMPHCVHTD AACTRAVHAEANVIASAAKYGVSLQGSEVYVTHSPCLSCAGLL
Sbjct  61   GYNGAPAGMPHCVHTDGAACTRAVHAEANVIASAAKYGVSLQGSEVYVTHSPCLSCAGLL  120

Query  121  VNAAISKVCYTTEFRDTSGIELLEAAGVTVDNVMPTEYLFPQRFIQGLL  169
            VNAAISKVCYTTEFRDTSGIELLEAAGVTVDNVMPTEYLFPQRFIQGLL
Sbjct  121  VNAAISKVCYTTEFRDTSGIELLEAAGVTVDNVMPTEYLFPQRFIQGLL  169
```

34967:

There are 2 1:1 alignments.

Autoannotated start site is favored

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



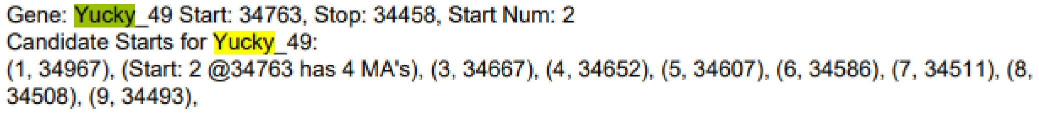
Stai	Raw SD	Genomic	Spacer	Final	Sequence of the Region	Start	Start	ORF
#	Score	Z Value	Distance	Score	Upstream of the Start	Codon	Position	Length
1	-2.071	2.901	9	-2.845	CCCGAGCATACGAGGAGGCAGC	ATG	34967	510
2	-3.277	2.323	15	-4.879	GTACAACGGGGCGCCGGCAGGG	ATG	34763	306
3	-4.509	1.733	9	-5.284	AGCAGCGAAATACGGCGTCTCT	TTG	34667	210
4	-3.778	2.083	10	-4.473	CGTCTCTTTGCAGGGCTCTGAA	GTG	34652	195
5	-3.778	2.083	9	-4.553	CCTGTCGTGCGCAGGGCTGCTC	GTG	34607	150
6	-6.520	0.770	12	-7.355	CGTGAACGCCGCGATCTCAAAG	GTG	34586	129
7	-3.697	2.122	18	-5.998	CGCTGGTGTTACCGTTGACAAC	GTG	34511	54
8	-5.870	1.081	12	-6.706	TGGTGTTACCGTTGACAACGTG	ATG	34508	51
9	-5.296	1.356	7	-6.818	CAACGTGATGCCGACCGAGTAC	TTG	34493	36

- The z value of autoannotated start site is 2.323 (second greatest) and the final score is -4.879 (not close to the least negative).

- A new start site has suggested: 34967.
- Z value is the greatest with 2.901 and the final score is the least negative with -2.845

- New start 34967 site is favored

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



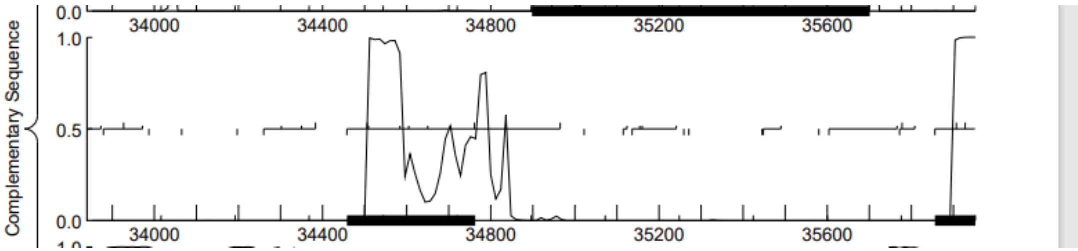
Gene: Yucky_49 Start: 34763, Stop: 34458, Start Num: 2
Candidate Starts for Yucky_49:
(1, 34967), (Start: 2 @34763 has 4 MA's), (3, 34667), (4, 34652), (5, 34607), (6, 34586), (7, 34511), (8, 34508), (9, 34493),

- There are 4 MA's for the autoannotated start site 34763

34967:

No MA

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- 34763 Coding potential is cut off.

34967:

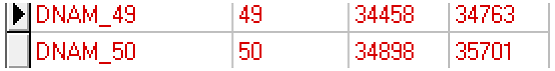
Coding potential is included.

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Autoannotated start site:

34898 – 34763 = 135

135-1 = 134 gap



DNAM_49	49	34458	34763
DNAM_50	50	34898	35701

34967:

34967-34898 = 69

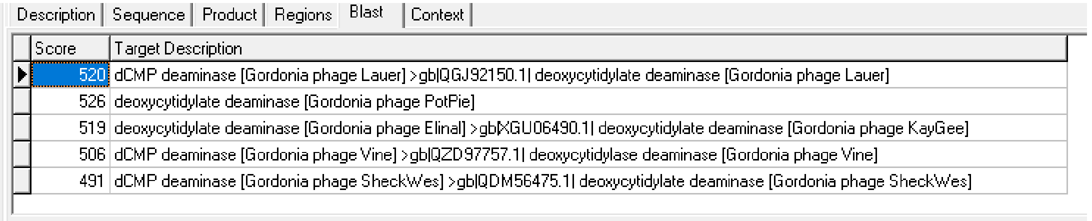
69+1 = 70 overlap

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	34763	**34967**
GeneMark	**Both Glimmer and GeneMark**	NA
Coding potential	Cut off	**Included**
RBS Score	Z-value: 2.323 Final score: -4.879	**Z-value: 2.901** **Final Score: -2.845**
BLAST	**3 1:1 alignments**	2 1:1 alignments
Starterator	**4**	0
Gap/overlap	134 gap	70 overlap

Start site at nucleotide 34967 is supported by coding potential and the RBS score evidence. Number of 1:1 alignments and the MA's of the autoannotated start site and the proposed start site are just different by small number. 70 overlap is smaller than having 134 gap. Calling 34967 is the start we are calling as we feel a 70 bp overlap is preferred.

BLAST function evidence. What assigned functions do other highly similar genes have?

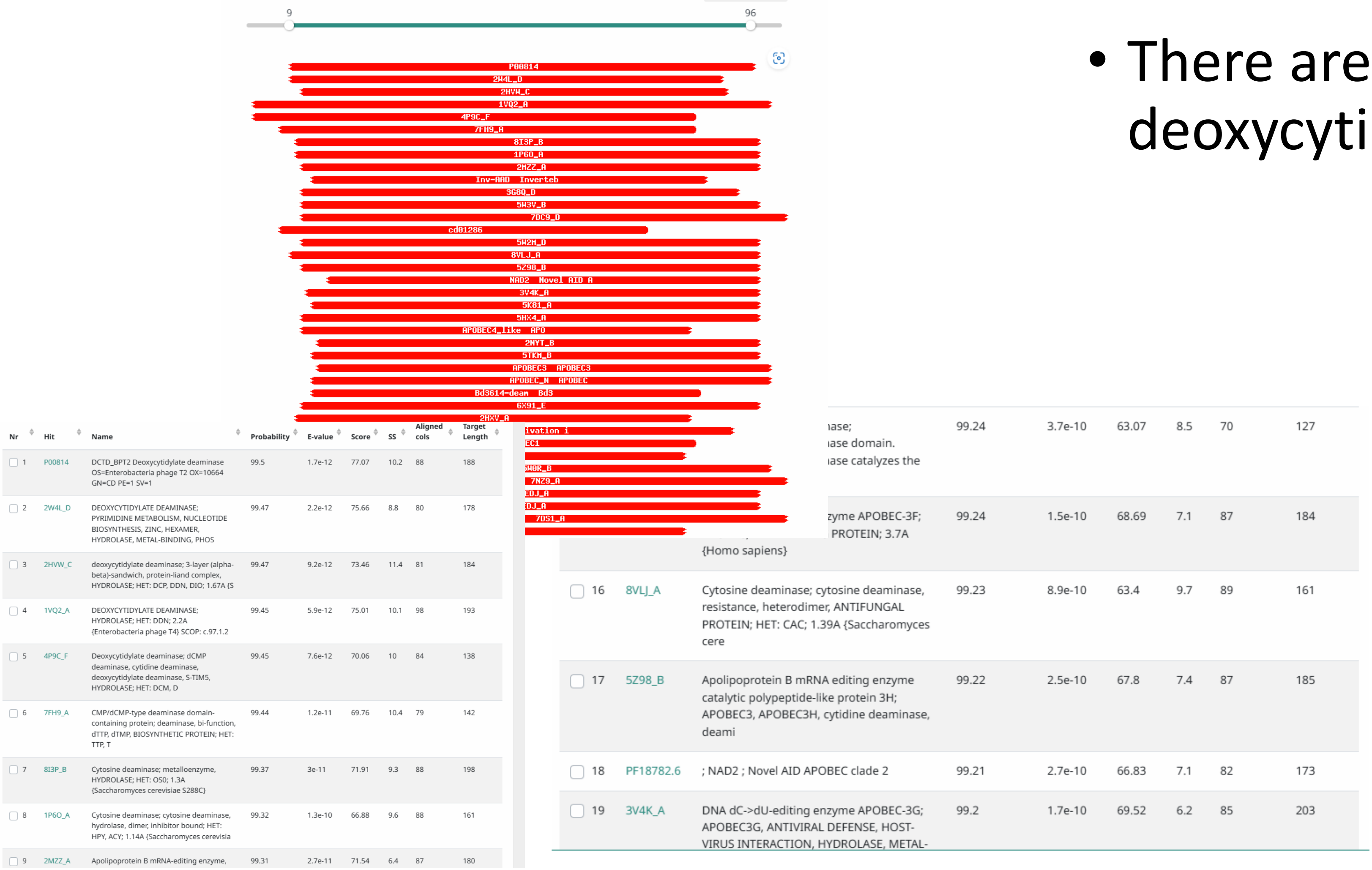


Score	Target Description
520	dCMP deaminase [Gordonia phage Lauer] >gb|QGJ92150.1| deoxycytidylate deaminase [Gordonia phage Lauer]
526	deoxycytidylate deaminase [Gordonia phage PotPie]
519	deoxycytidylate deaminase [Gordonia phage Elinal] >gb|XGU06490.1| deoxycytidylate deaminase [Gordonia phage KayGee]
506	dCMP deaminase [Gordonia phage Vine] >gb|QZD97757.1| deoxycytidylase deaminase [Gordonia phage Vine]
491	dCMP deaminase [Gordonia phage SheckWes] >gb|QDM56475.1| deoxycytidylate deaminase [Gordonia phage SheckWes]

- Other highly similar genes are dCMP deaminase, deoxycytidylate deaminase, nucleoside deaminase, hypothetical protein.

- There are deoxycytidylate deaminase.
- dCMP deaminase is the wrong name for deoxycytidylate deaminase.

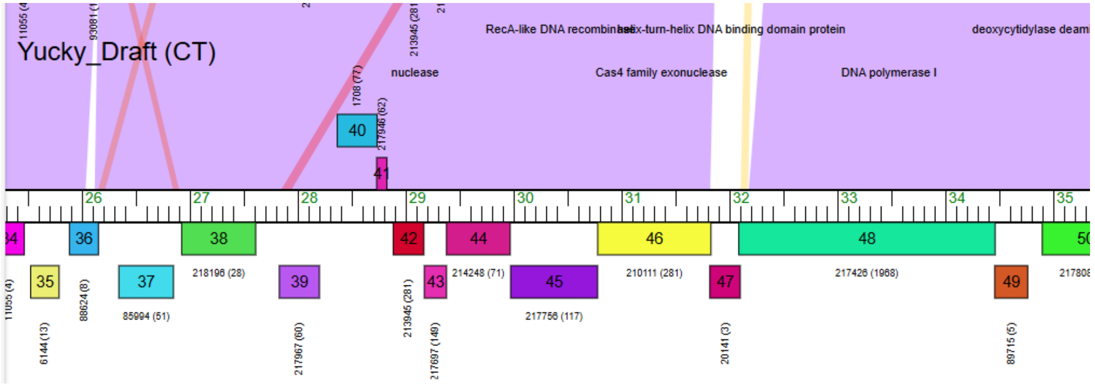
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

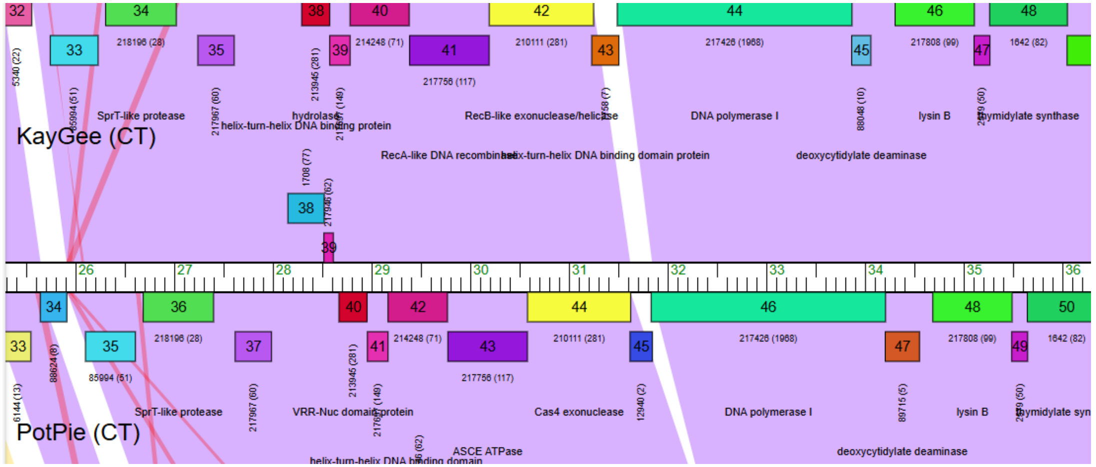


Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
☐ 1	P00814	DCTD_BPT2 Deoxycytidylate deaminase OS=Enterobacteria phage T2 OX=10664 GN=CD PE=1 SV=1	99.5	1.7e-12	77.07	10.2	88	188
☐ 2	2W4L_D	DEOXYCYTIDYLATE DEAMINASE; PYRIMIDINE METABOLISM, NUCLEOTIDE BIOSYNTHESIS, ZINC, HEXAMER, HYDROLASE, METAL-BINDING, PHOS	99.47	2.2e-12	75.66	8.8	80	178
☐ 3	2HVW_C	deoxycytidylate deaminase; 3-layer (alpha-beta)-sandwich, protein-liand complex, HYDROLASE; HET: DCP, DDN, DIO; 1.67A {S	99.47	9.2e-12	73.46	11.4	81	184
☐ 4	1VQ2_A	DEOXYCYTIDYLATE DEAMINASE; HYDROLASE; HET: DDN; 2.2A {Enterobacteria phage T4} SCOP: c.97.1.2	99.45	5.9e-12	75.01	10.1	98	193
☐ 5	4P9C_F	Deoxycytidylate deaminase; dCMP deaminase, cytidine deaminase, deoxycytidylate deaminase, S-TIM5, HYDROLASE; HET: DCM, D	99.45	7.6e-12	70.06	10	84	138
☐ 6	7FH9_A	CMP/dCMP-type deaminase domain-containing protein; deaminase, bi-function, dTTP, dTMP, BIOSYNTHETIC PROTEIN; HET: TTP, T	99.44	1.2e-11	69.76	10.4	79	142
☐ 7	8I3P_B	Cytosine deaminase; metalloenzyme, HYDROLASE; HET: OS0; 1.3A {Saccharomyces cerevisiae S288C}	99.37	3e-11	71.91	9.3	88	198
☐ 8	1P6O_A	Cytosine deaminase; cytosine deaminase, hydrolase, dimer, inhibitor bound; HET: HPY, ACY; 1.14A {Saccharomyces cerevisia	99.32	1.3e-10	66.88	9.6	88	161
☐ 9	2MZZ_A	Apolipoprotein B mRNA-editing enzyme,	99.31	2.7e-11	71.54	6.4	87	180
		ase; ase domain. ase catalyzes the	99.24	3.7e-10	63.07	8.5	70	127
		zyme APOBEC-3F; PROTEIN; 3.7A {Homo sapiens}	99.24	1.5e-10	68.69	7.1	87	184
☐ 16	8VLJ_A	Cytosine deaminase; cytosine deaminase, resistance, heterodimer, ANTIFUNGAL PROTEIN; HET: CAC; 1.39A {Saccharomyces cere	99.23	8.9e-10	63.4	9.7	89	161
☐ 17	5Z98_B	Apolipoprotein B mRNA editing enzyme catalytic polypeptide-like protein 3H; APOBEC3, APOBEC3H, cytidine deaminase, deami	99.22	2.5e-10	67.8	7.4	87	185
☐ 18	PF18782.6	; NAD2 ; Novel AID APOBEC clade 2	99.21	2.7e-10	66.83	7.1	82	173
☐ 19	3V4K_A	DNA dC->dU-editing enzyme APOBEC-3G; APOBEC3G, ANTIVIRAL DEFENSE, HOST-VIRUS INTERACTION, HYDROLASE, METAL-	99.2	1.7e-10	69.52	6.2	85	203

- There are many hits with deoxycytidylate deaminase.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?





The highly similar gene is a deoxycytidylate deaminase. They share 9 conserved domains.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- No

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- It is a deoxycytidylate deaminase because:

- Most of highly similar genes shown in BLAST were deoxycytidylate deaminase.
- There are many hits with deoxycytidylate deaminase in Hhpred.
- One highly similar gene with 9 conserved domains is a deoxycytidylate deaminase.

Feature 49 – Reverse – Stop 34898

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

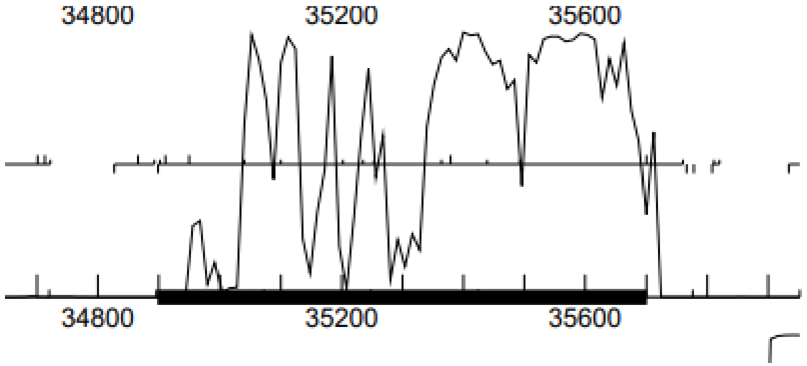
What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

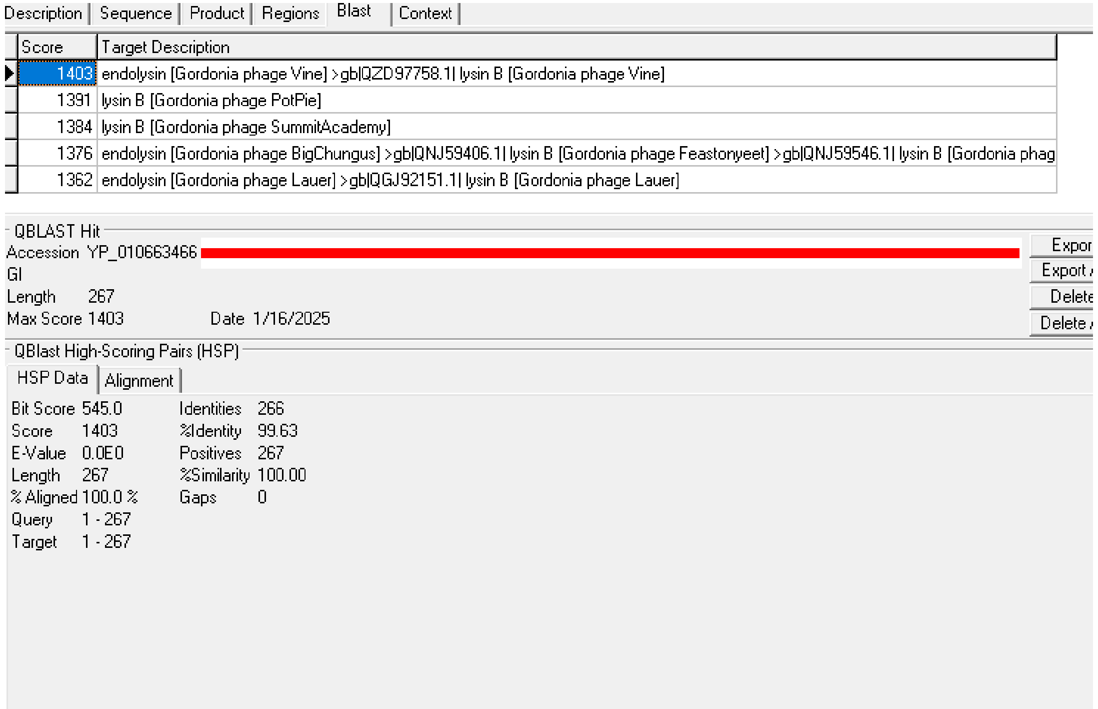
- 49
- Reverse
- 34898
- Both Glimmer and GeneMark
- 35701
- 4 overlap

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Coding potential is strong.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
1403	endolysin [Gordonia phage Vine] >gb|QZD97758.1| lysin B [Gordonia phage Vine]
1391	lysin B [Gordonia phage PotPie]
1384	lysin B [Gordonia phage SummitAcademy]
1376	endolysin [Gordonia phage BigChungus] >gb|QNJ59406.1| lysin B [Gordonia phage Feastonyeet] >gb|QNJ59546.1| lysin B [Gordonia phag
1362	endolysin [Gordonia phage Lauer] >gb|QGJ92151.1| lysin B [Gordonia phage Lauer]

QBLAST Hit

Accession YP_010663466
GI
Length 267
Max Score 1403 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

HSP Data | Alignment

Bit Score	545.0	Identities	266
Score	1403	%Identity	99.63
E-Value	0.0E0	Positives	267
Length	267	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 267		
Target	1 - 267		

- There are 25 highly similar genes with E value close to 0 (Vine).

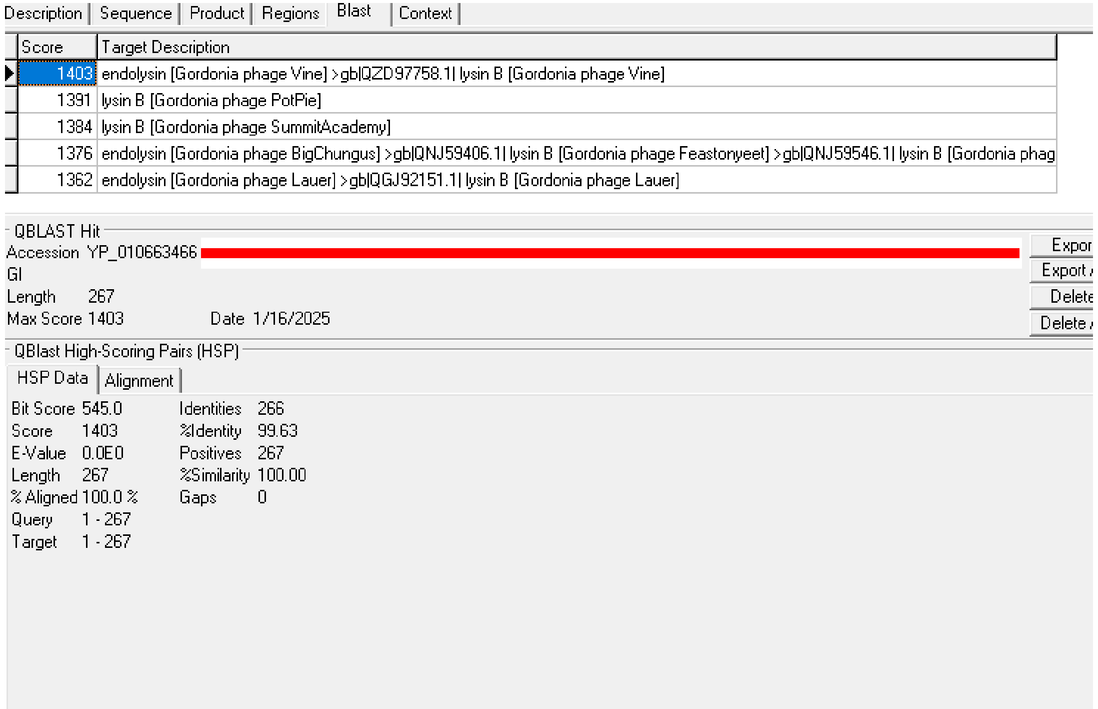
Answer: Is it a gene? Give evidence why you think this is a gene or not.

- It is a gene because:

- Both glimmer and genemark called it a gene.
- Coding potential is strong.
- There are 25 highly similar genes with E-value of 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There are 23 1:1 alignments.
- It is favored.



Score	Target Description
1403	endolysin [Gordonia phage Vine] >gb|QZD97758.1| lysin B [Gordonia phage Vine]
1391	lysin B [Gordonia phage PotPie]
1384	lysin B [Gordonia phage SummitAcademy]
1376	endolysin [Gordonia phage BigChungus] >gb|QNJ59406.1| lysin B [Gordonia phage Feastonyeet] >gb|QNJ59546.1| lysin B [Gordonia phag
1362	endolysin [Gordonia phage Lauer] >gb|QGJ92151.1| lysin B [Gordonia phage Lauer]

QBLAST Hit

Accession YP_010663466

GI

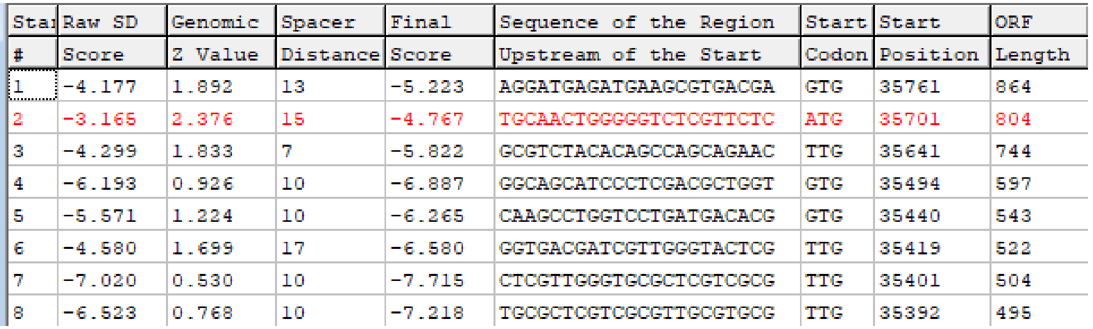
Length 267

Max Score 1403 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	545.0	Identities	266
Score	1403	%Identity	99.63
E-Value	0.0E0	Positives	267
Length	267	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 267		
Target	1 - 267		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



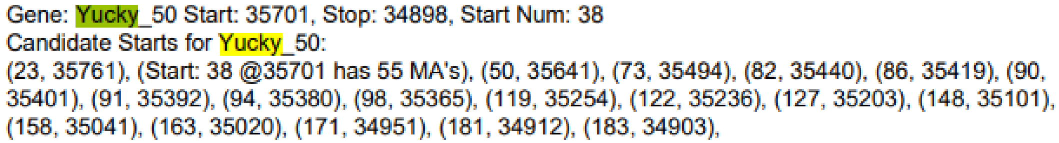
Sta: #	Raw SD Score	Genomic Z Value	Spacer Distance	Final Score	Sequence of the Region Upstream of the Start	Start Codon	Start Position	ORF Length
1	-4.177	1.892	13	-5.223	AGGATGAGATGAAGCGTGACGA	GTG	35761	864
2	-3.165	2.376	15	-4.767	TGCAACTGGGGGTCTCGTTCTC	ATG	35701	804
3	-4.299	1.833	7	-5.822	GCGTCTACACAGCCAGCAGAAC	TTG	35641	744
4	-6.193	0.926	10	-6.887	GGCAGCATCCCTCGACGCTGGT	GTG	35494	597
5	-5.571	1.224	10	-6.265	CAAGCCTGGTCCTGATGACACG	GTG	35440	543
6	-4.580	1.699	17	-6.580	GGTGACGATCGTTGGGTACTCG	TTG	35419	522
7	-7.020	0.530	10	-7.715	CTCGTTGGGTGCGCTCGTCGCG	TTG	35401	504
8	-6.523	0.768	10	-7.218	TGCGCTCGTCGCGTTGCGTGCG	TTG	35392	495

- The z-value is 2.376 (the greatest) and the final score is -4.767 (the least negative).

- It is favored.

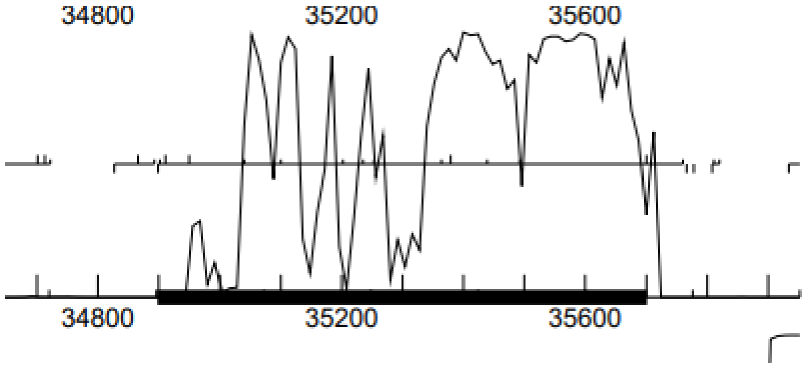
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- There are 55 manual annotation on this start site.



Gene: Yucky_50 Start: 35701, Stop: 34898, Start Num: 38
Candidate Starts for Yucky_50:
(23, 35761), (Start: 38 @35701 has 55 MA's), (50, 35641), (73, 35494), (82, 35440), (86, 35419), (90, 35401), (91, 35392), (94, 35380), (98, 35365), (119, 35254), (122, 35236), (127, 35203), (148, 35101), (158, 35041), (163, 35020), (171, 34951), (181, 34912), (183, 34903),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

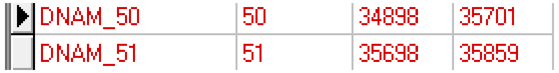


- Coding potential is a little bit cut off.

- The start site is at 35701, but GeneMark S file show the coding potential starts around at 35715.

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start?   Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 35701-35698 = 3
- 3+1 = 4 overlap



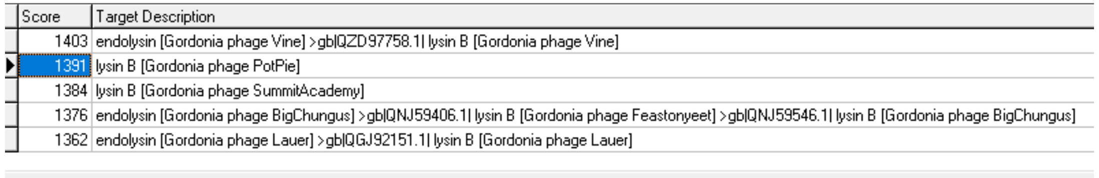
DNAM_50	50	34898	35701
DNAM_51	51	35698	35859

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	35701
GeneMark	Both Glimmer and GeneMark
Coding potential	Cut off a little bit
RBS score	Z-value: 2.376 Final score: -4.767
BLAST	23 1:1 alignments
Starterator	55 MA's
Gap/overlap	4 overlap.

This gene starts at 35701 because it is the only proposed start site with most of evidence supporting it. Both Glimmer and GeneMark agree. There are many 1:1 alignments and MA's. With overlap of 4, RBS becomes more important. Even though coding potential is cut off by a little bit, other evidence support the autoannotated start site.

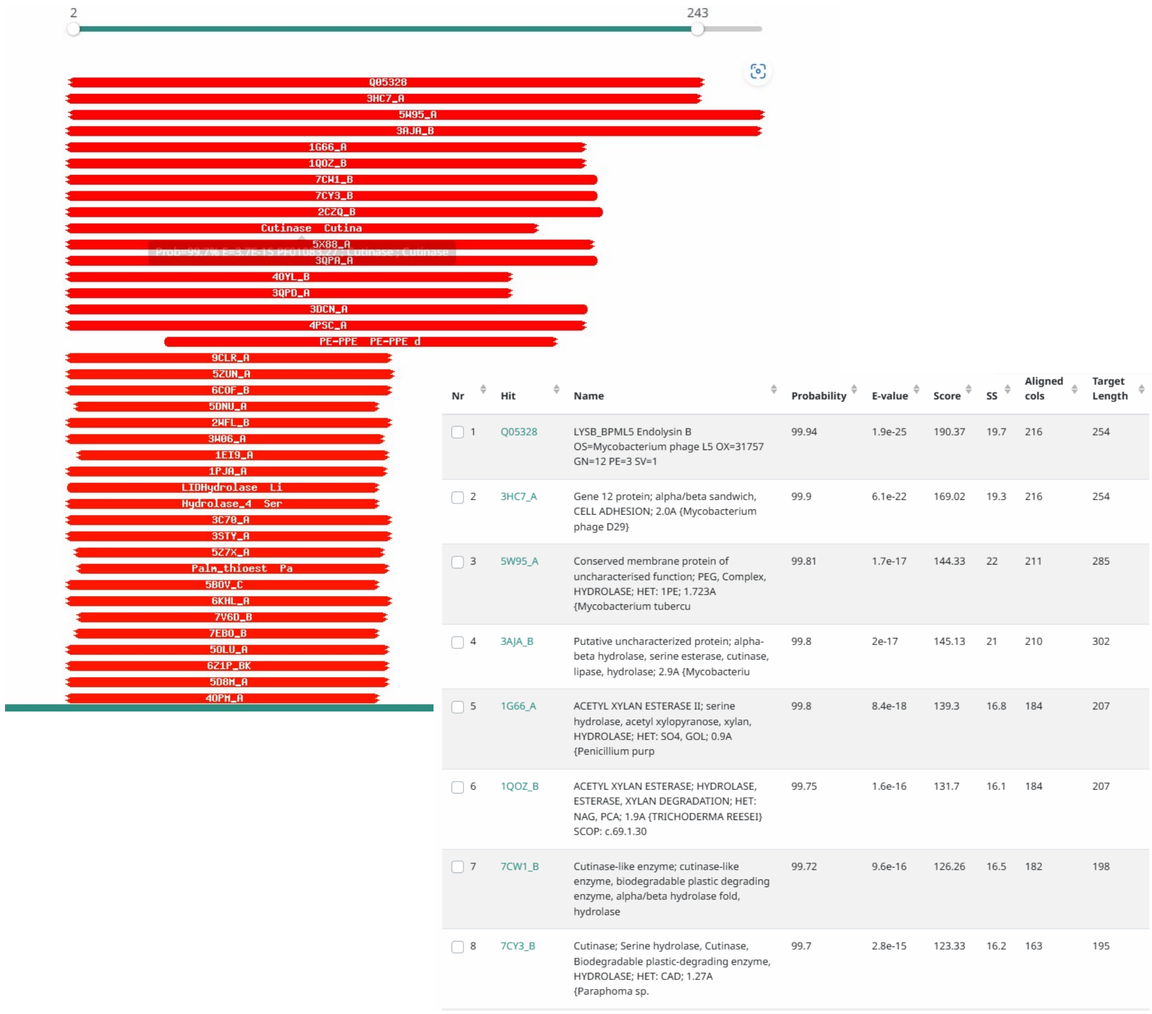
BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
1403	endolysin [Gordonia phage Vine] >gb|QZD97758.1| lysin B [Gordonia phage Vine]
1391	lysin B [Gordonia phage PotPie]
1384	lysin B [Gordonia phage SummitAcademy]
1376	endolysin [Gordonia phage BigChungus] >gb|QNJ59406.1| lysin B [Gordonia phage Feastonyeet] >gb|QNJ59546.1| lysin B [Gordonia phage BigChungus]
1362	endolysin [Gordonia phage Lauer] >gb|QGJ92151.1| lysin B [Gordonia phage Lauer]

- Other highly similar genes are endolysin, lysin B.

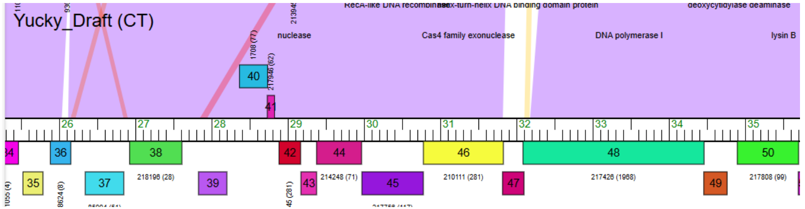
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

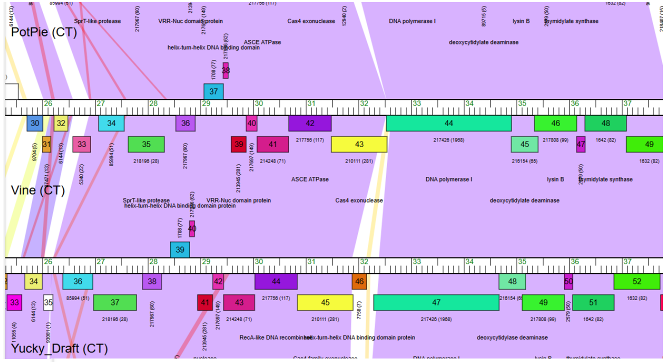


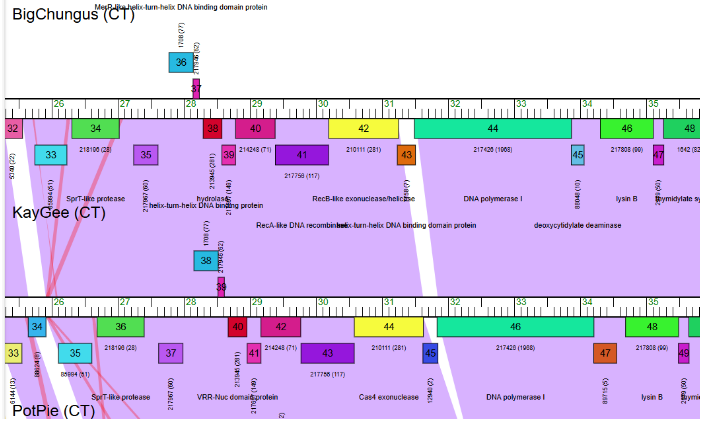
Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	Q05328	LYSB_BPML5 Endolysin B OS=Mycobacterium phage L5 OX=31757 GN=12 PE=3 SV=1	99.94	1.9e-25	190.37	19.7	216	254
□ 2	3HC7_A	Gene 12 protein; alpha/beta sandwich, CELL ADHESION; 2.0A {Mycobacterium phage D29}	99.9	6.1e-22	169.02	19.3	216	254
□ 3	5W95_A	Conserved membrane protein of uncharacterised function; PEG, Complex, HYDROLASE; HET: 1PE; 1.723A {Mycobacterium tubercu	99.81	1.7e-17	144.33	22	211	285
□ 4	3AJA_B	Putative uncharacterized protein; alpha-beta hydrolase, serine esterase, cutinase, lipase, hydrolase; 2.9A {Mycobacteriu	99.8	2e-17	145.13	21	210	302
□ 5	1G66_A	ACETYL XYLAN ESTERASE II; serine hydrolase, acetyl xylopyranose, xylan, HYDROLASE; HET: SO4, GOL; 0.9A {Penicillium purp	99.8	8.4e-18	139.3	16.8	184	207
□ 6	1QOZ_B	ACETYL XYLAN ESTERASE; HYDROLASE, ESTERASE, XYLAN DEGRADATION; HET: NAG, PCA; 1.9A {TRICHODERMA REESEI} SCOP: c.69.1.30	99.75	1.6e-16	131.7	16.1	184	207
□ 7	7CW1_B	Cutinase-like enzyme; cutinase-like enzyme, biodegradable plastic degrading enzyme, alpha/beta hydrolase fold, hydrolase	99.72	9.6e-16	126.26	16.5	182	198
□ 8	7CY3_B	Cutinase; Serine hydrolase, Cutinase, Biodegradable plastic-degrading enzyme, HYDROLASE; HET: CAD; 1.27A {Paraphoma sp.	99.7	2.8e-15	123.33	16.2	163	195

Many call it a lysin B or endolysin. Because there are lysin B, it can not be called as an endolysin.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?






BigChungus (CT)
MerR-like helix-turn-helix DNA binding domain protein
KayGee (CT)
SprT-like protease
hydrolase
helix-turn-helix DNA binding protein
RecB-like exonuclease/helicase
DNA polymerase I
lysin B
thymidylate synthase
RecA-like DNA recombinase
helix-turn-helix DNA binding domain protein
deoxycytidylate deaminase
PotPie (CT)
SprT-like protease
VRR-Nuc domain protein
Cas4 exonuclease
DNA polymerase I
lysin B

Other highly similar genes in the same pham are lysin B.

There are no conserved domains.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- No

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- This gene is a lysin B because

- Many highly genes are lysin B protein.
- There are many hits with lysin B or endolysin (it is a lysin B, more specific).
- Other highly similar genes in the same pham are lysin B.

Feature 50 – Reverse – Stop 35698

Glimmer/GeneMark

What feature number is this?

What is the stop site?

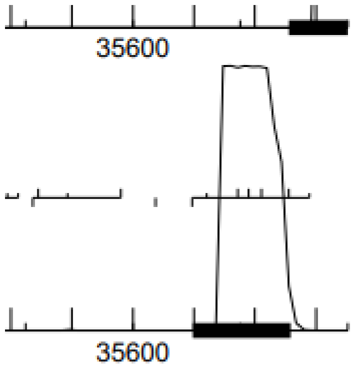
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

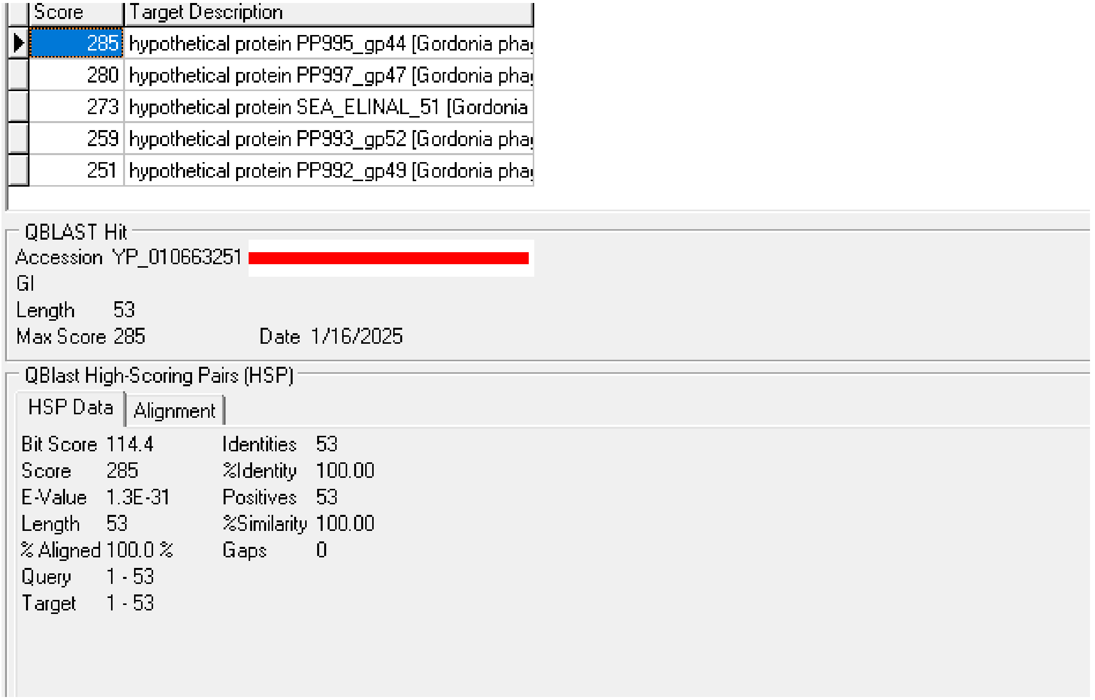
- 50
- Reverse
- 35698
- Both Glimmer and GeneMark
- 35859
- 4 overlap

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Reading frame 3 shows a strong coding potential.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
285	hypothetical protein PP995_gp44 [Gordonia pha
280	hypothetical protein PP997_gp47 [Gordonia pha
273	hypothetical protein SEA_ELINAL_51 [Gordonia
259	hypothetical protein PP993_gp52 [Gordonia pha
251	hypothetical protein PP992_gp49 [Gordonia pha

QBLAST Hit

Accession YP_010663251 [redacted]
GI
Length 53
Max Score 285 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

```
Bit Score 114.4       Identities  53
Score     285         %Identity   100.00
E-Value   1.3E-31     Positives   53
Length    53          %Similarity 100.00
% Aligned 100.0 %     Gaps        0
Query     1 - 53
Target    1 - 53
```

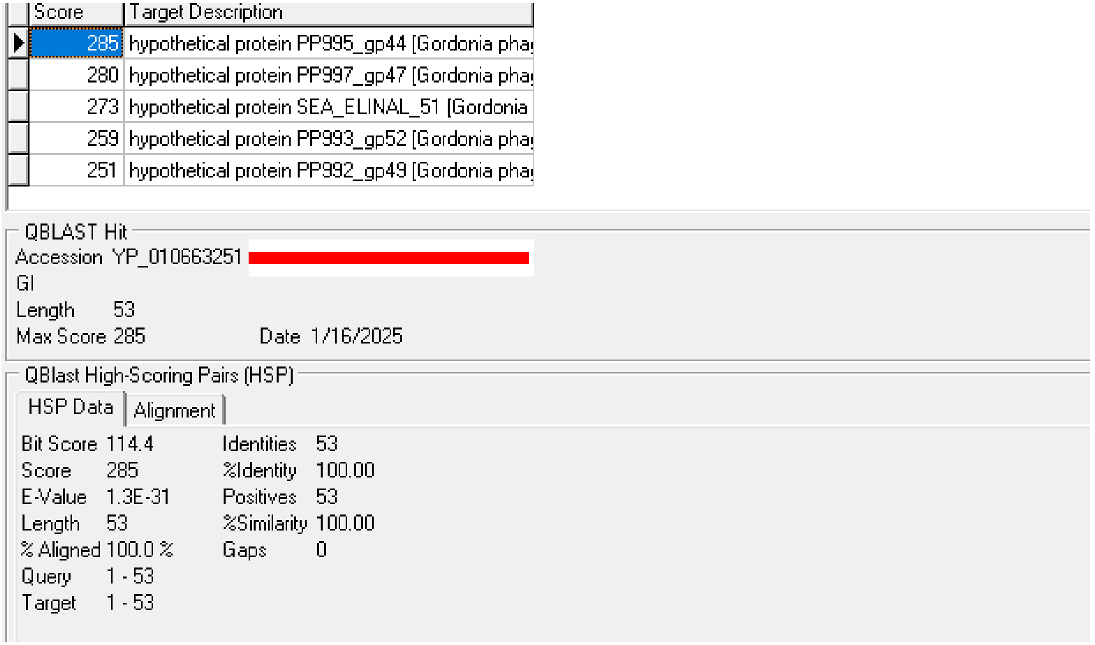
- There are 13 highly similar genes with E value of close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- It is a gene because:

- Both Glimmer and GeneMark called it a gene.
- Coding potential is strong.
- There are 13 highly similar genes with E value of close to 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
285	hypothetical protein PP995_gp44 [Gordonia pha
280	hypothetical protein PP997_gp47 [Gordonia pha
273	hypothetical protein SEA_ELINAL_51 [Gordonia
259	hypothetical protein PP993_gp52 [Gordonia pha
251	hypothetical protein PP992_gp49 [Gordonia pha

QBLAST Hit

Accession YP_010663251

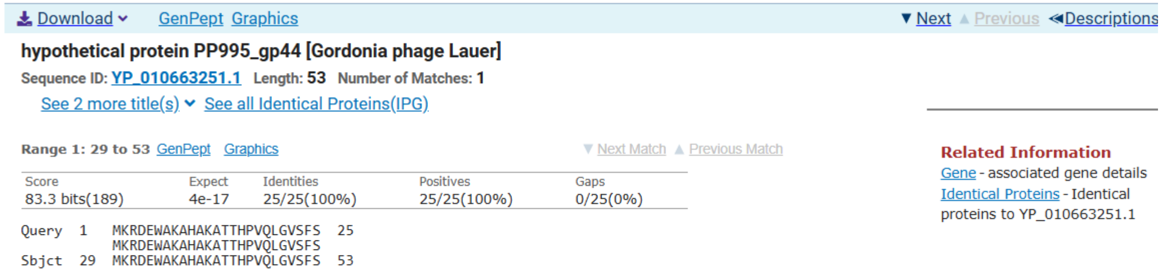
GI

Length 53

Max Score 285 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	114.4	Identities	53
Score	285	%Identity	100.00
E-Value	1.3E-31	Positives	53
Length	53	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 53		
Target	1 - 53		



hypothetical protein PP995_gp44 [Gordonia phage Lauer]

Sequence ID: YP_010663251.1 Length: 53 Number of Matches: 1

See 2 more title(s) See all Identical Proteins(IPG)

Range 1: 29 to 53 GenPept Graphics

Score	Expect	Identities	Positives	Gaps
83.3 bits(189)	4e-17	25/25(100%)	25/25(100%)	0/25(0%)

```
Query  1   MKRDEWAKAHAKATTHPVQLGVSFS  25
           MKRDEWAKAHAKATTHPVQLGVSFS
Sbjct  29  MKRDEWAKAHAKATTHPVQLGVSFS  53
```

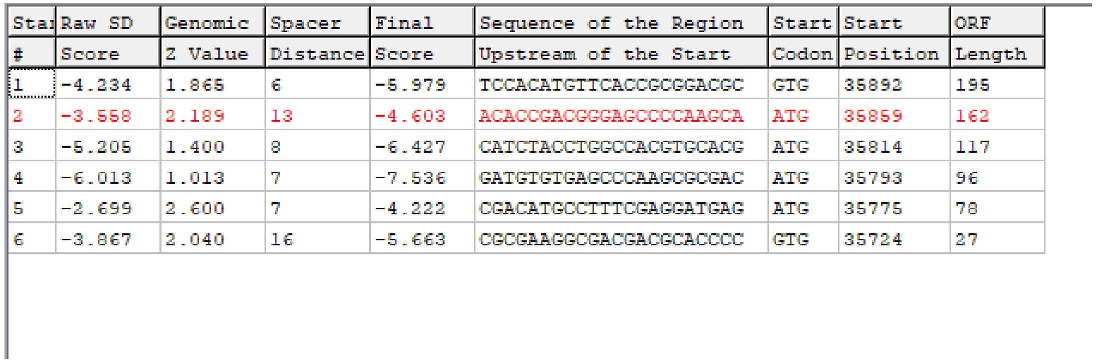
Related Information

Gene - associated gene details

Identical Proteins - Identical proteins to YP_010663251.1

- There are 6 1:1 alignments. (Elinal)
- No 1:1 alignments for RBS suggested start site.

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

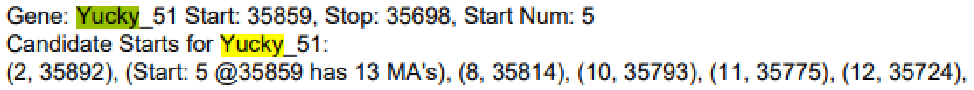


Sta	Raw SD	Genomic	Spacer	Final	Sequence of the Region	Start	Start	ORF
#	Score	Z Value	Distance	Score	Upstream of the Start	Codon	Position	Length
1	-4.234	1.865	6	-5.979	TCCACATGTTCACCGCGGACGC	GTG	35892	195
2	-3.558	2.189	13	-4.603	ACACCGACGGGAGCCCCAAGCA	ATG	35859	162
3	-5.205	1.400	8	-6.427	CATCTACCTGGCCACGTGCACG	ATG	35814	117
4	-6.013	1.013	7	-7.536	GATGTGTGAGCCCAAGCGCGAC	ATG	35793	96
5	-2.699	2.600	7	-4.222	CGACATGCCTTTCGAGGATGAG	ATG	35775	78
6	-3.867	2.040	16	-5.663	CGCGAAGGCGACGACGCACCCC	GTG	35724	27

- The z value is 2.189, not the greatest.
- Final score is -4.603, not the least negative.

- There is a better start site at 35775 with RBS value, but it has too much gap of 83.
- Z value of 2.600, the greatest.
- Final score of -4.222, the least negative.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

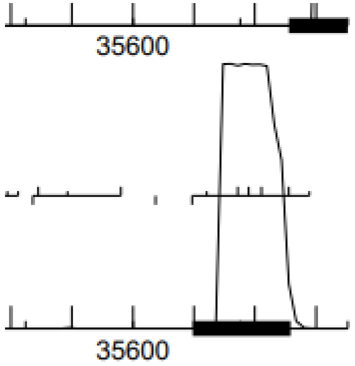


Gene: Yucky_51 Start: 35859, Stop: 35698, Start Num: 5
Candidate Starts for Yucky_51:
(2, 35892), (Start: 5 @35859 has 13 MA's), (8, 35814), (10, 35793), (11, 35775), (12, 35724),

- There are 13 MA's for the autoannotated start site.

- There are no MA's for the RBS suggested start site.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



Autoannotated start site:

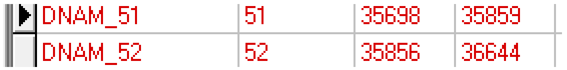
Includes all.

RBS start site:

Cuts off most of coding potential.

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 35859-35856 = 3
- 3+1 = 4 overlap for autoannotated start site.



DNAM_51	51	35698	35859
DNAM_52	52	35856	36644

- 35856 – 35775 = 84
- 84-1 = 83 gap for RBS suggested start site.

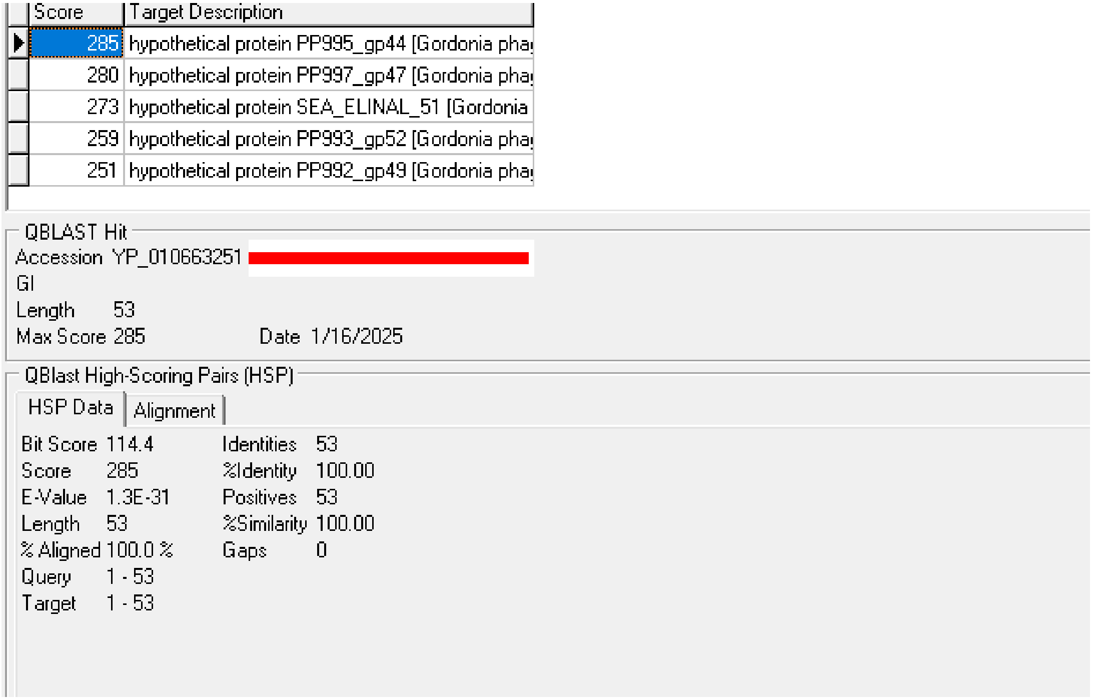
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	35859	35775
GeneMark	**Both Glimmer and Genemark**	NA
Coding potential	**Included**	Cuts off most parts
RBS score	Z-value: 2.189 Final score: -4.603	**Z value: 2.600** **Final score: -4.222**
BLAST	**6 1:1 alignments**	NA
Starterator	**13 MA's**	NA
Gap/overlap	**Overlap of 4**	Gap of 83

Gene 51 starts at 35859 because all evidence except RBS score support it. Its RBS score also slightly lower than the ones 35775. So, this difference does not affect it.

BLAST function evidence. What assigned functions do other highly similar genes have?

- Hypothetical protein (Lauer)



Score	Target Description
285	hypothetical protein PP995_gp44 [Gordonia phag
280	hypothetical protein PP997_gp47 [Gordonia phag
273	hypothetical protein SEA_ELINAL_51 [Gordonia
259	hypothetical protein PP993_gp52 [Gordonia phag
251	hypothetical protein PP992_gp49 [Gordonia phag

QBLAST Hit

Accession YP_010663251

GI

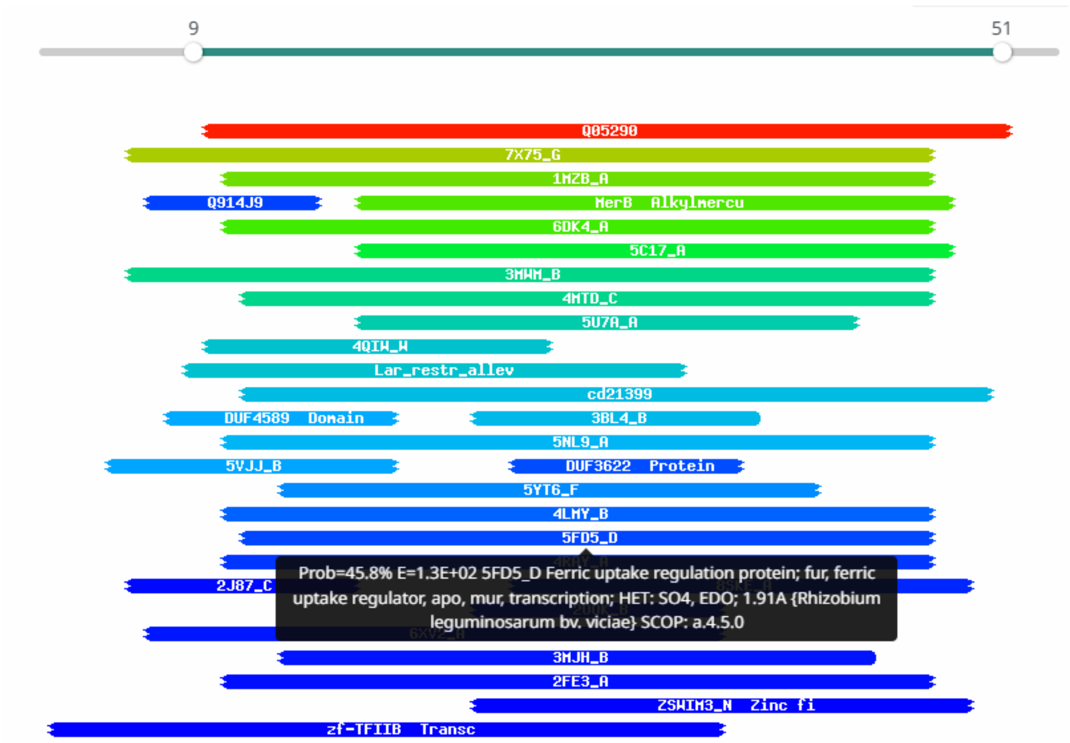
Length 53

Max Score 285 Date 1/16/2025

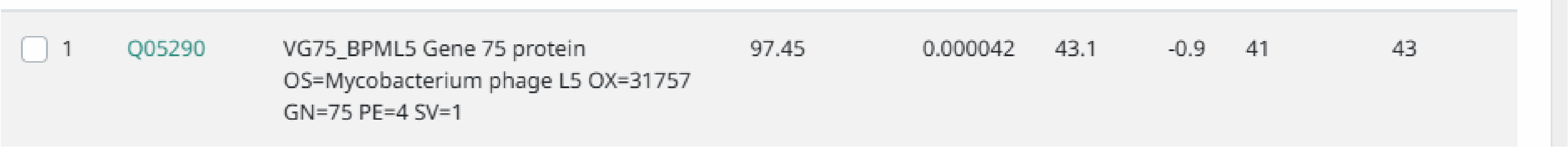
QBlast High-Scoring Pairs (HSP)

Bit Score	114.4	Identities	53
Score	285	%Identity	100.00
E-Value	1.3E-31	Positives	53
Length	53	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 53		
Target	1 - 53		

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



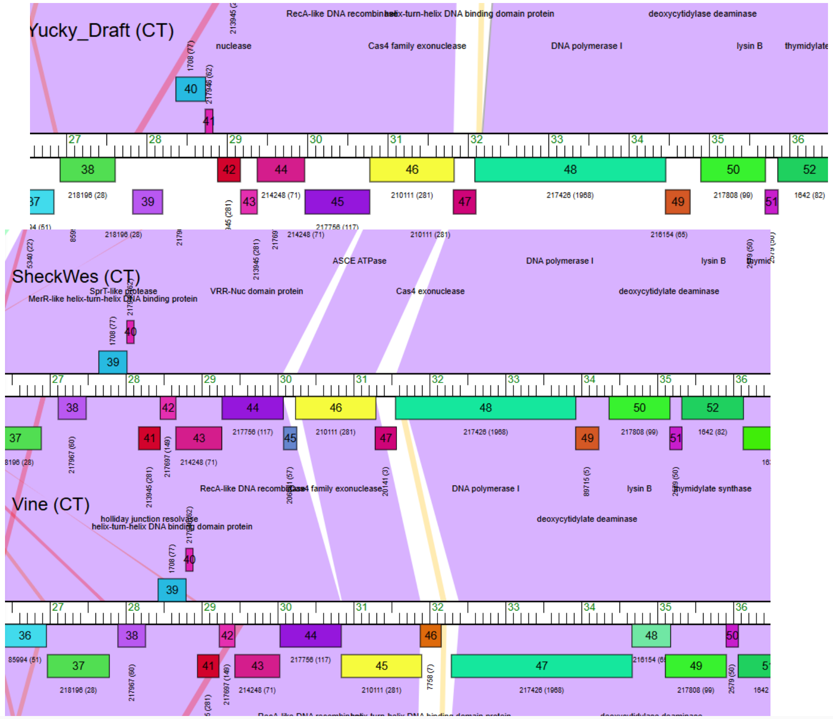
Prob=45.8% E=1.3E+02 5FD5_D Ferric uptake regulation protein; fur, ferric uptake regulator, apo, mur, transcription; HET: SO4, EDO; 1.91A {Rhizobium leguminosarum bv. viciae} SCOP: a.4.5.0



□	1	Q05290	VG75_BPML5 Gene 75 protein OS=Mycobacterium phage L5 OX=31757 GN=75 PE=4 SV=1	97.45	0.000042	43.1	-0.9	41	43

- There is only one hit with probability greater than 90.
- Though, there are no functions like that.
- So hypothetical protein.

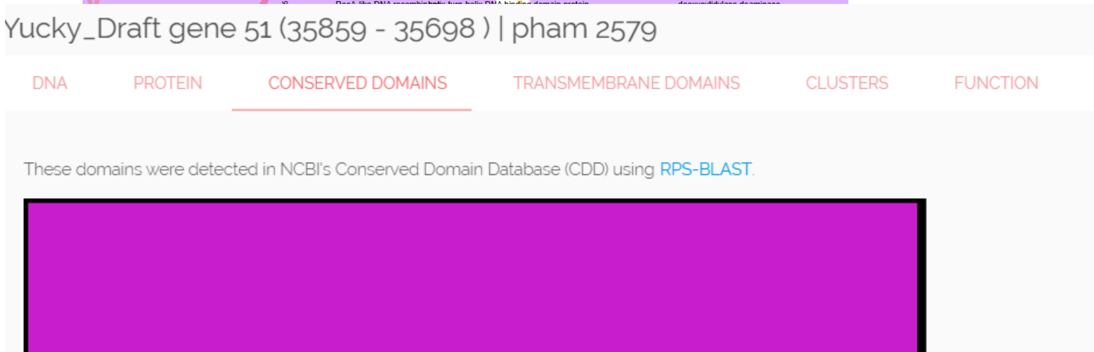
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?


Yucky_Draft (CT)
RecA-like DNA recombinase
helix-turn-helix DNA binding domain protein
deoxycytidylate deaminase
nuclease
Cas4 family exonuclease
DNA polymerase I
lysin B
thymidylate
SheckWes (CT)
SprT-like protease
MerR-like helix-turn-helix DNA binding protein
VRR-Nuc domain protein
ASCE ATPase
Cas4 exonuclease
DNA polymerase I
lysin B
deoxycytidylate deaminase
Vine (CT)
holliday junction resolvase
helix-turn-helix DNA binding domain protein
RecA-like DNA recombinase
Cas4 family exonuclease
DNA polymerase I
lysin B
thymidylate synthase
deoxycytidylate deaminase

Other highly similar genes in the same pham are not assigned a function.

So, hypothetical protein.

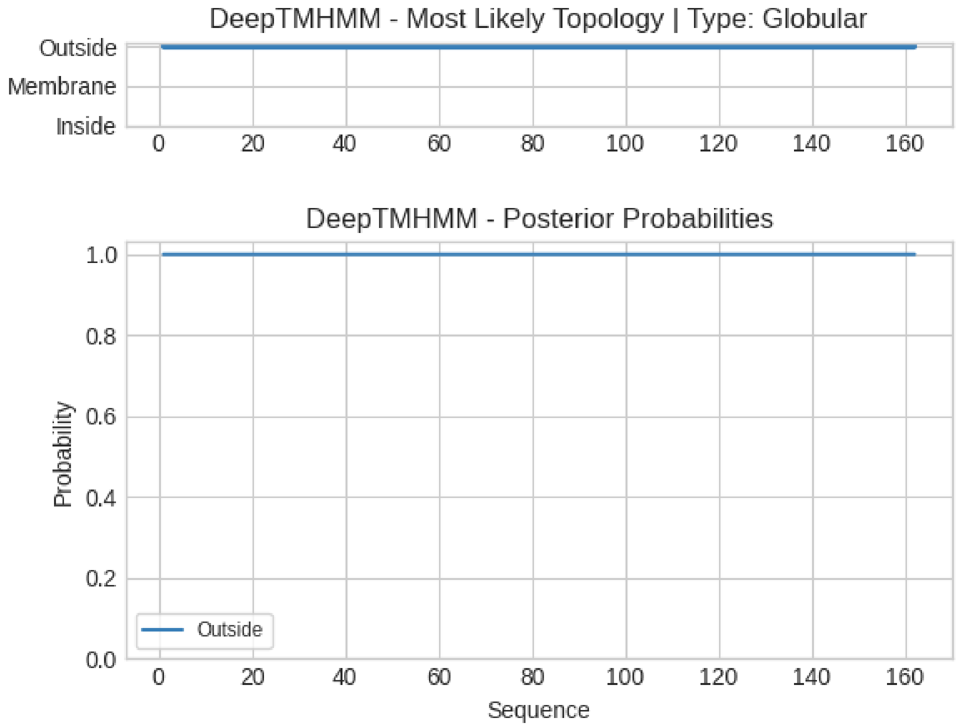
There are also no conserved domains.



Yucky_Draft gene 51 (35859 - 35698) | pham 2579

These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- It is located on the outside of the cell.

- So, there is no way to know its function
- Therefore, hypothetical potein

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Gene 51 is a hypothetical protein because:

- All highly similar genes in BLAST are hypothetical protein.
- There is only one hit with probability greater than 90, but its function is not in official function list.
- Other highly similar genes in the same pham do not have a function assigned.
- This gene is located on the outside of the cell.

Feature 51 – Reverse – Stop 35856

Glimmer/GeneMark

What feature number is this? **51**

What is the stop site? **35856**

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Called by Glimmer and GeneMark

What is the autoannotated start?

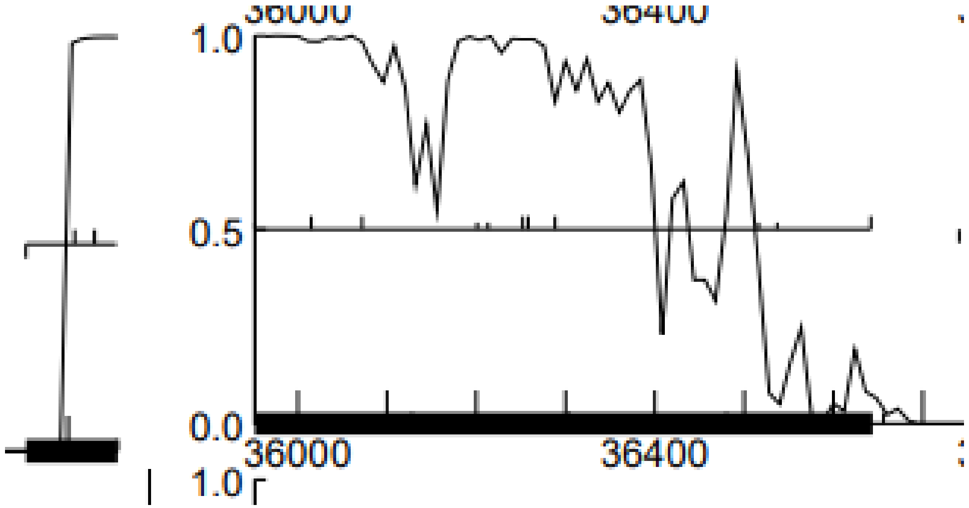
36644

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

There is an overlap of 4

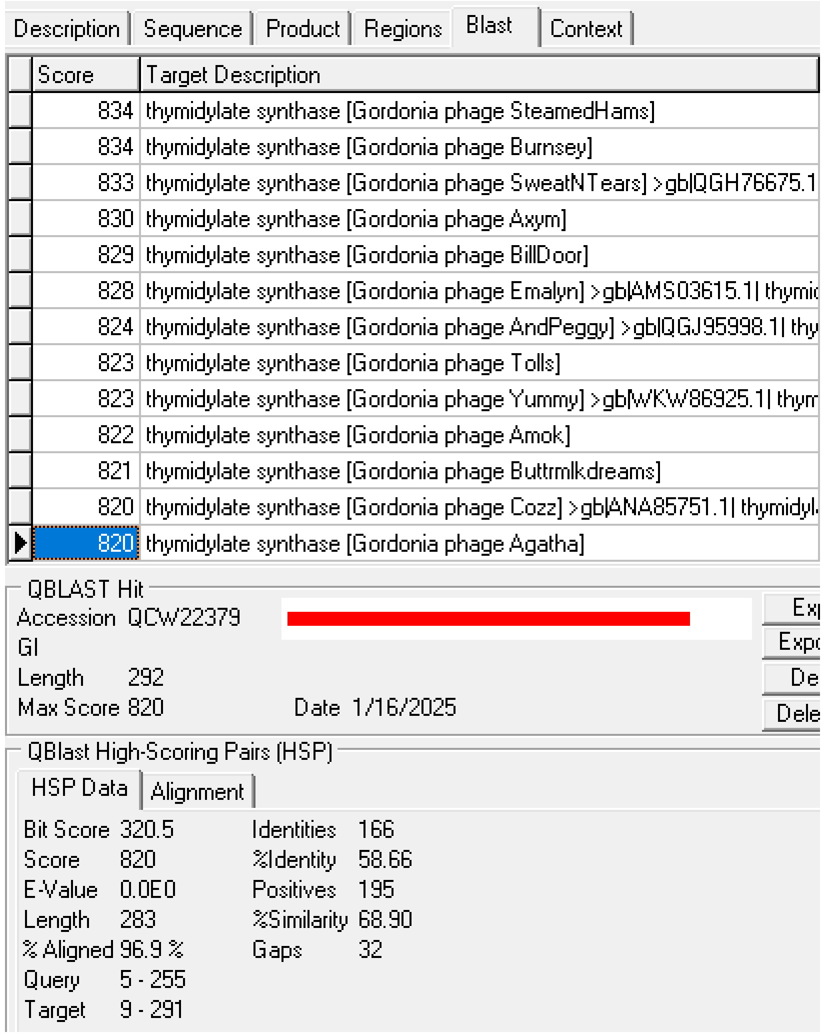
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- The coding potential starts off weak a little bit before the feature is called to start at 36644 and continues to alternate until dropping off around 35810.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are at least 25 BLAST hits of highly similar genes from other phages
- All BLAST hits have e-values extremely close to zero
- There are 3 1:1 alignments



Score	Target Description
834	thymidylate synthase [Gordonia phage SteamedHams]
834	thymidylate synthase [Gordonia phage Burnsey]
833	thymidylate synthase [Gordonia phage SweatNTears] >gb|QGH76675.1
830	thymidylate synthase [Gordonia phage Axym]
829	thymidylate synthase [Gordonia phage BillDoor]
828	thymidylate synthase [Gordonia phage Emalyn] >gb|AMS03615.1| thymic
824	thymidylate synthase [Gordonia phage AndPeggy] >gb|QGJ95998.1| thy
823	thymidylate synthase [Gordonia phage Tolls]
823	thymidylate synthase [Gordonia phage Yummy] >gb|WKW86925.1| thym
822	thymidylate synthase [Gordonia phage Amok]
821	thymidylate synthase [Gordonia phage Buttrmlkdreams]
820	thymidylate synthase [Gordonia phage Cozz] >gb|ANA85751.1| thymidyl
820	thymidylate synthase [Gordonia phage Agatha]

QBLAST Hit

Accession QCW22379
GI
Length 292
Max Score 820
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

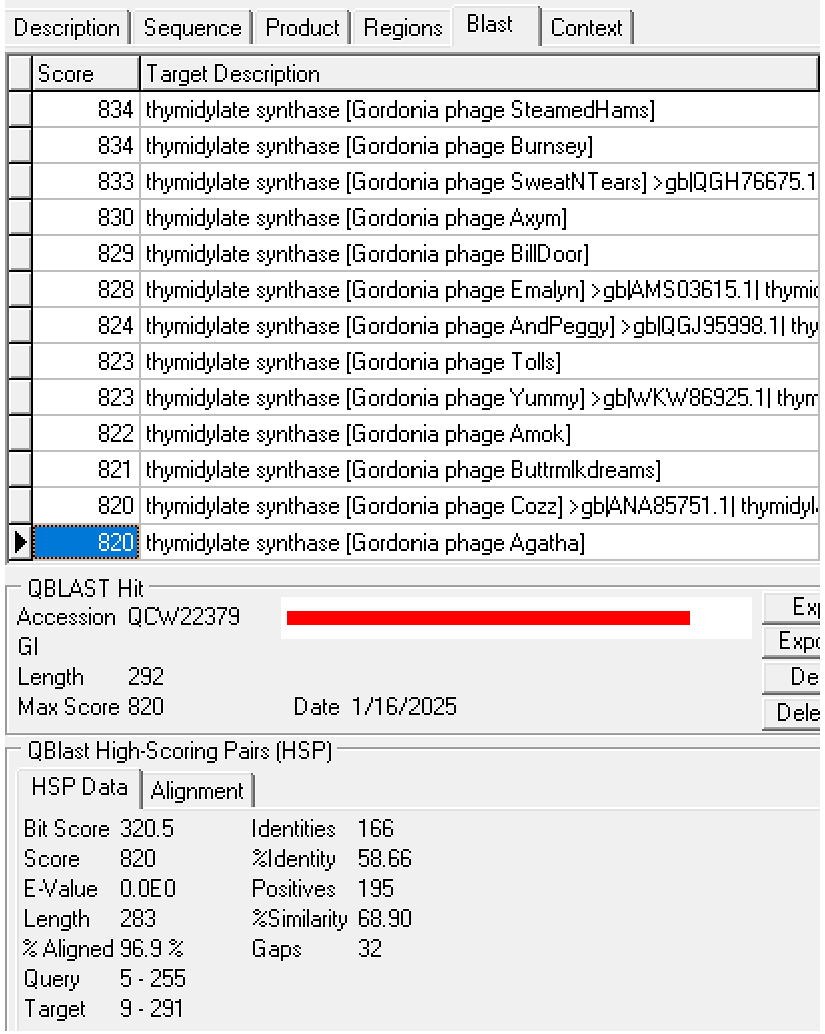
Bit Score	320.5	Identities	166
Score	820	%Identity	58.66
E-Value	0.0E0	Positives	195
Length	283	%Similarity	68.90
% Aligned	96.9 %	Gaps	32
Query	5 - 255		
Target	9 - 291		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This feature is a gene. The is strong coding potential throughout where the feature is called to run, and there are at least 25 BLAST hits of phages with highly similar genes that all have e-values extremely close to zero. Three of these hits were 1:1 alignments.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There are at least 25 BLAST hits of phages with highly similar genes that all have e-values extremely close to zero
- There are 3 1:1 alignments



Score	Target Description
834	thymidylate synthase [Gordonia phage SteamedHams]
834	thymidylate synthase [Gordonia phage Burnsey]
833	thymidylate synthase [Gordonia phage SweatNTears] >gb|QGH76675.1
830	thymidylate synthase [Gordonia phage Axym]
829	thymidylate synthase [Gordonia phage BillDoor]
828	thymidylate synthase [Gordonia phage Emalyn] >gb|AMS03615.1| thymi
824	thymidylate synthase [Gordonia phage AndPeggy] >gb|QGJ95998.1| thy
823	thymidylate synthase [Gordonia phage Tolls]
823	thymidylate synthase [Gordonia phage Yummy] >gb|WKW86925.1| thym
822	thymidylate synthase [Gordonia phage Amok]
821	thymidylate synthase [Gordonia phage Buttrmlkdreams]
820	thymidylate synthase [Gordonia phage Cozz] >gb|ANA85751.1| thymidyl
820	thymidylate synthase [Gordonia phage Agatha]

QBLAST Hit

Accession QCW22379

GI

Length 292

Max Score 820

Date 1/16/2025

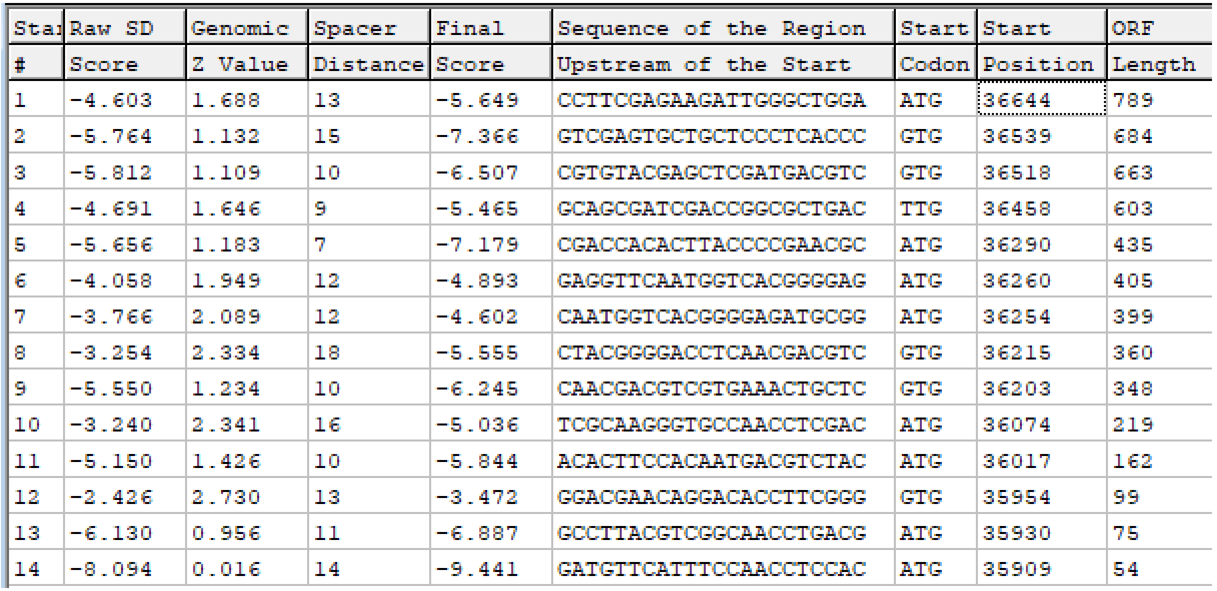
QBlast High-Scoring Pairs (HSP)

HSP Data | Alignment

Bit Score	320.5	Identities	166
Score	820	%Identity	58.66
E-Value	0.0E0	Positives	195
Length	283	%Similarity	68.90
% Aligned	96.9 %	Gaps	32
Query	5 - 255		
Target	9 - 291		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

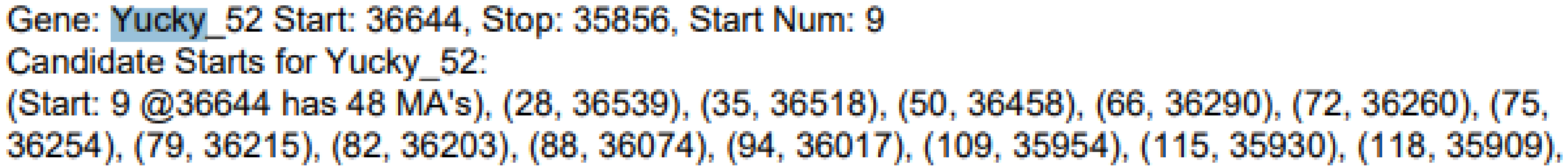
- Starting at 36644:
  - Z-value = 1.688
  - Final score = -5.649
- There are some starting sites that have better RBS scores but they cut off a much larger amount of coding potential and do not show up in starterator.



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	-4.603	1.688	13	-5.649	CCTTCGAGAAGATTGGGCTGGA	ATG	36644	789
2	-5.764	1.132	15	-7.366	GTCGAGTGCTGCTCCCTCACCC	GTG	36539	684
3	-5.812	1.109	10	-6.507	CGTGTACGAGCTCGATGACGTC	GTG	36518	663
4	-4.691	1.646	9	-5.465	GCAGCGATCGACCGGCGCTGAC	TTG	36458	603
5	-5.656	1.183	7	-7.179	CGACCACACTTACCCCGAACGC	ATG	36290	435
6	-4.058	1.949	12	-4.893	GAGGTTCAATGGTCACGGGGAG	ATG	36260	405
7	-3.766	2.089	12	-4.602	CAATGGTCACGGGGAGATGCGG	ATG	36254	399
8	-3.254	2.334	18	-5.555	CTACGGGGACCTCAACGACGTC	GTG	36215	360
9	-5.550	1.234	10	-6.245	CAACGACGTCGTGAAACTGCTC	GTG	36203	348
10	-3.240	2.341	16	-5.036	TCGCAAGGGTGCCAACCTCGAC	ATG	36074	219
11	-5.150	1.426	10	-5.844	ACACTTCCACAATGACGTCTAC	ATG	36017	162
12	-2.426	2.730	13	-3.472	GGACGAACAGGACACCTTCGGG	GTG	35954	99
13	-6.130	0.956	11	-6.887	GCCTTACGTCGGCAACCTGACG	ATG	35930	75
14	-8.094	0.016	14	-9.441	GATGTTCATTTCCAACCTCCAC	ATG	35909	54

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

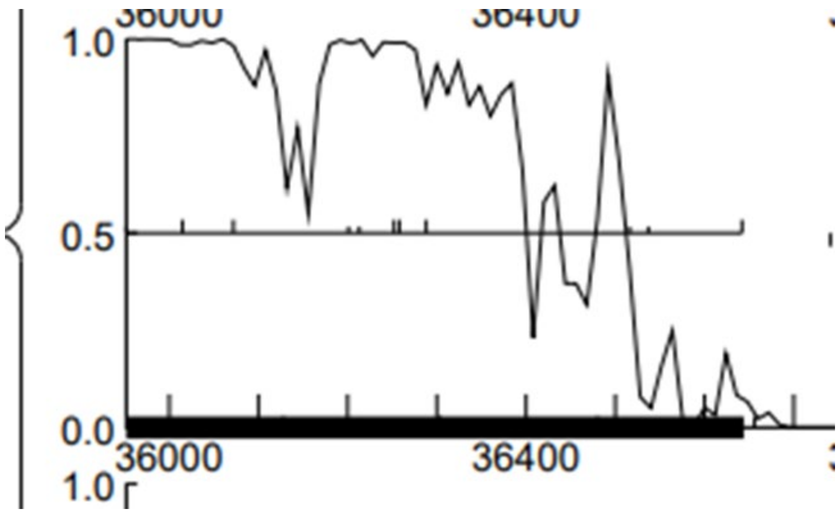
- Calls the “Most Annotated” start
- The only start site that has any manual annotation is 36644, and it has a total of 48 MA’s.



Gene: Yucky_52 Start: 36644, Stop: 35856, Start Num: 9
Candidate Starts for Yucky_52:
(Start: 9 @36644 has 48 MA's), (28, 36539), (35, 36518), (50, 36458), (66, 36290), (72, 36260), (75, 36254), (79, 36215), (82, 36203), (88, 36074), (94, 36017), (109, 35954), (115, 35930), (118, 35909),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Starting at 36644 cuts off a small amount of coding potential, but none of it is strong. A majority of the coding potential is included with this start site.



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

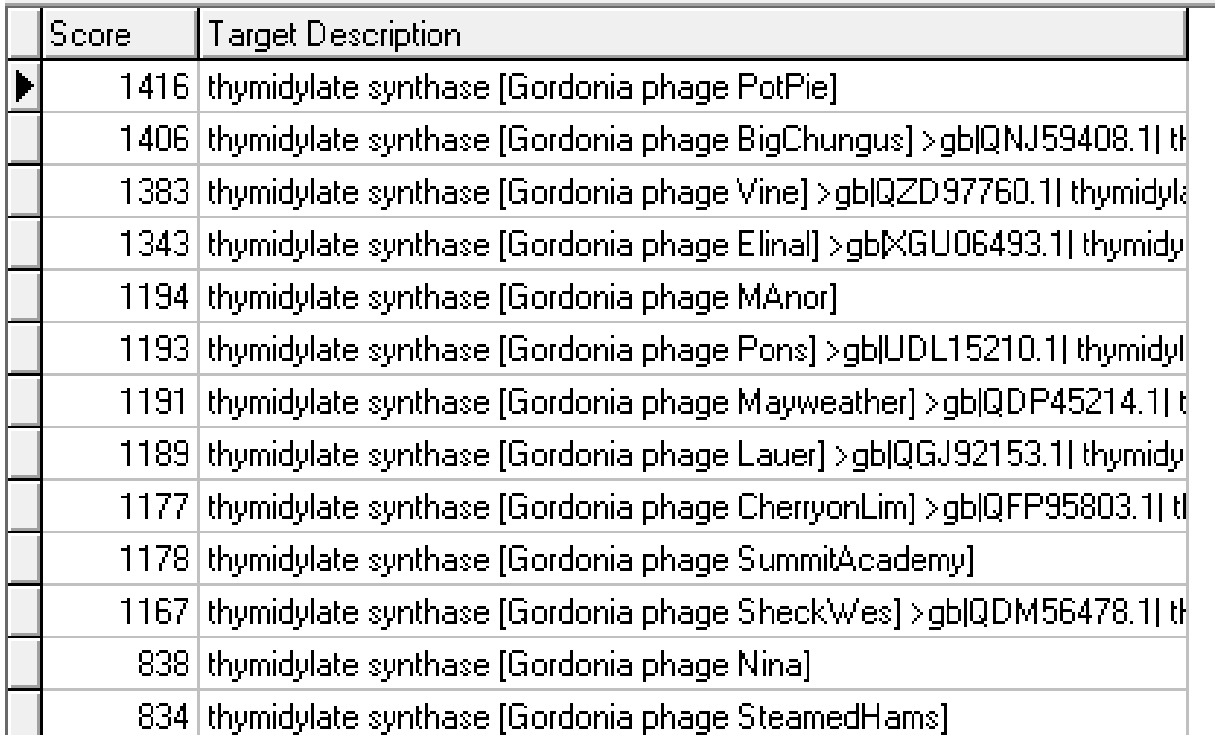
- Starting at 36644 would leave an overlap of 4 nucleotides.

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site for this gene is 36644 and it was the only proposed start site based off all the evidence. There were 3 1:1 alignments for starting at this position with highly similar genes from other phages according to BLAST. The RBS scores for starting here came to a z-value of 1.688 and a final score of -5.649. There were better scores, but they cut out a significant portion of coding potential. 36644 was the only start site that had manual annotations according to the starterator report for which it has 48. 36644 cuts off a small amount of coding potential, but a majority of it is included. Starting here would leave an overlap of 4 nucleotides with the previous reverse gene.

BLAST function evidence. What assigned functions do other highly similar genes have?

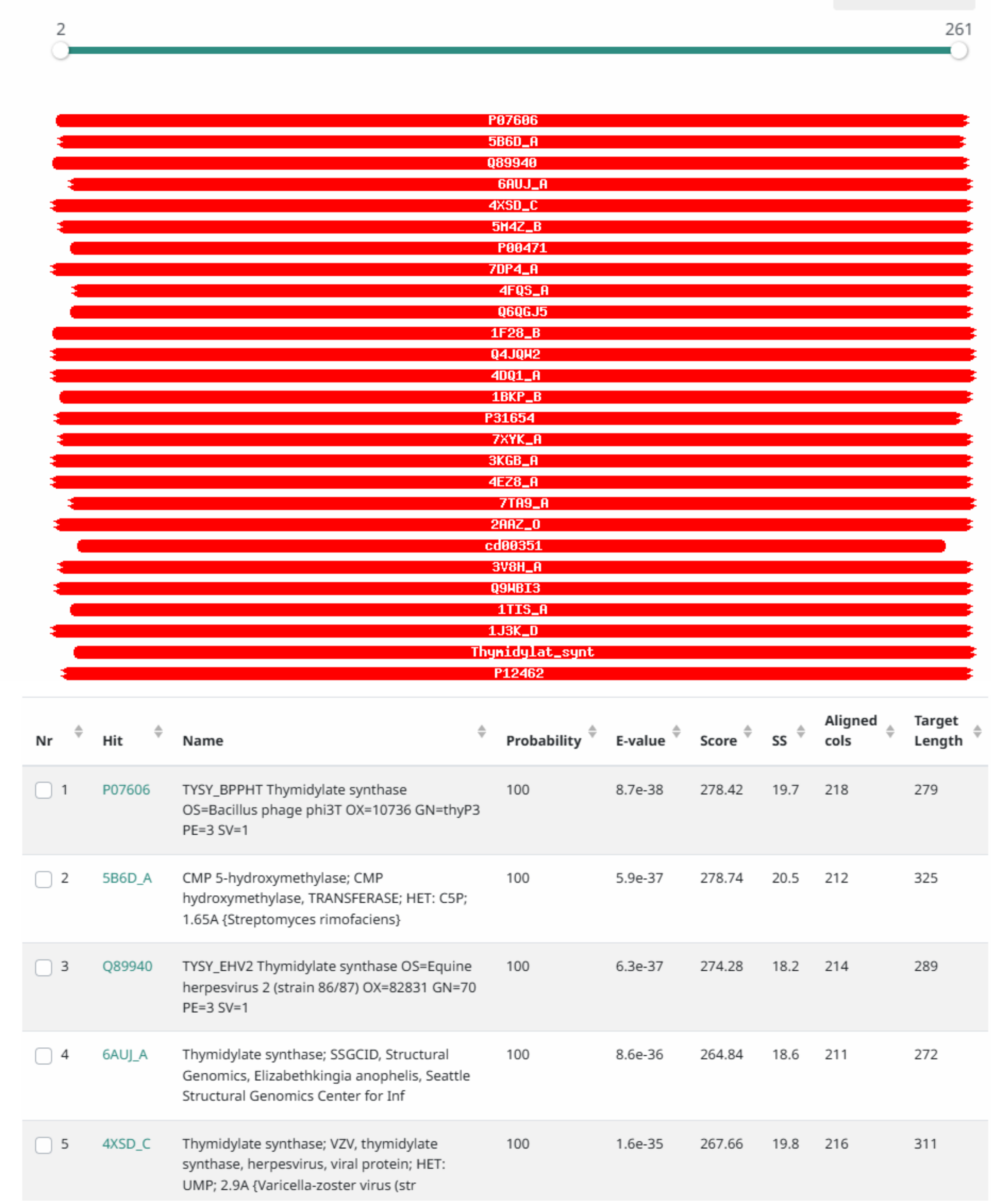
- There are at least 25 BLAST hits of highly similar genes that all have the function labeled thymidylate synthase



	Score	Target Description
▶	1416	thymidylate synthase [Gordonia phage PotPie]
	1406	thymidylate synthase [Gordonia phage BigChungus] >gb|QNJ59408.1| th
	1383	thymidylate synthase [Gordonia phage Vine] >gb|QZD97760.1| thymidyla
	1343	thymidylate synthase [Gordonia phage Elinal] >gb|XGU06493.1| thymidy
	1194	thymidylate synthase [Gordonia phage MAnor]
	1193	thymidylate synthase [Gordonia phage Pons] >gb|UDL15210.1| thymidyl
	1191	thymidylate synthase [Gordonia phage Mayweather] >gb|QDP45214.1| t
	1189	thymidylate synthase [Gordonia phage Lauer] >gb|QGJ92153.1| thymidy
	1177	thymidylate synthase [Gordonia phage CherryonLim] >gb|QFP95803.1| t
	1178	thymidylate synthase [Gordonia phage SummitAcademy]
	1167	thymidylate synthase [Gordonia phage SheckWes] >gb|QDM56478.1| th
	838	thymidylate synthase [Gordonia phage Nina]
	834	thymidylate synthase [Gordonia phage SteamedHams]

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of high the gene homologous, or just a region? A screenshot here of HHPRED results is desi

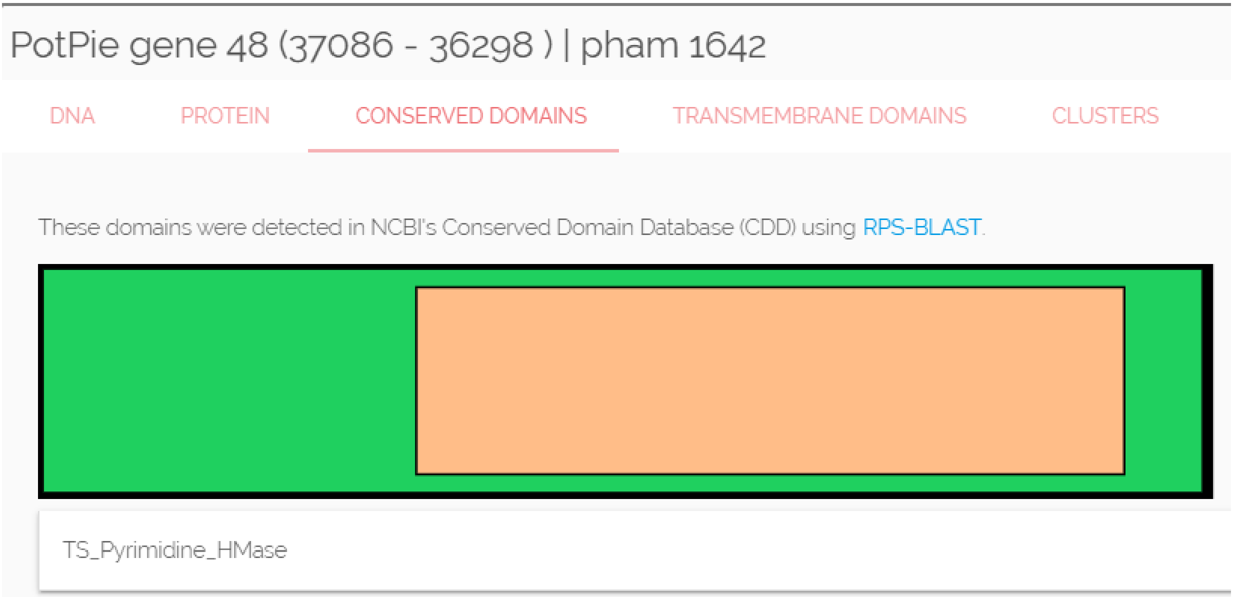
- There were several HHpred hits with probabilities of 100 that had functions labeled as thymidylate synthase and extremely small e-values.
- Some had functions labeled as hydroxymethylase.



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	P07606	TYSY_BPPHT Thymidylate synthase OS=Bacillus phage phi3T OX=10736 GN=thyP3 PE=3 SV=1	100	8.7e-38	278.42	19.7	218	279
□ 2	5B6D_A	CMP 5-hydroxymethylase; CMP hydroxymethylase, TRANSFERASE; HET: C5P; 1.65A {Streptomyces rimofaciens}	100	5.9e-37	278.74	20.5	212	325
□ 3	Q89940	TYSY_EHV2 Thymidylate synthase OS=Equine herpesvirus 2 (strain 86/87) OX=82831 GN=70 PE=3 SV=1	100	6.3e-37	274.28	18.2	214	289
□ 4	6AUJ_A	Thymidylate synthase; SSGCID, Structural Genomics, Elizabethkingia anophelis, Seattle Structural Genomics Center for Inf	100	8.6e-36	264.84	18.6	211	272
□ 5	4XSD_C	Thymidylate synthase; VZV, thymidylate synthase, herpesvirus, viral protein; HET: UMP; 2.9A {Varicella-zoster virus (str	100	1.6e-35	267.66	19.8	216	311

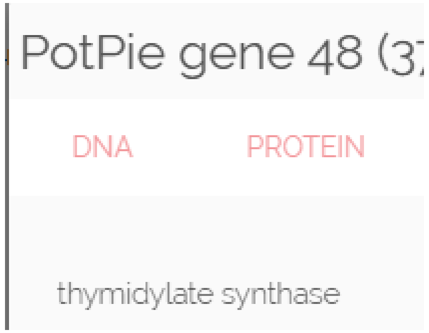
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Phamerator showed that closely related phages with genes in the same pham had functions labeled as thymidylate synthase and conserved domains labeled TS_Pyrimidine_HMase


PotPie gene 48 (37086 - 36298) | pham 1642

These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

TS_Pyrimidine_HMase


PotPie gene 48 (3

thymidylate synthase

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- **Not applicable since there is a probable function**

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Official function list assignment → thymidylate synthase
- The function of this gene should be labeled as thymidylate synthase. At least 25 BLAST hits of highly similar genes with functions labeled thymidylate synthase and had extremely small e-values that were close to zero. HHpred showed several hits with probabilities of 100 and e-values close to zero that had functions labeled as thymidylate synthase. Phamerator also shows that phages with genes in the same pham as this one have functions labeled as thymidylate synthase as well as conserved domains labeled TS_Pyrimidine_HMase.

Feature 52 – Reverse – Stop 36641

Glimmer/GeneMark

What feature number is this? 52

What is the stop site? **36641**

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Called by Glimmer and GeneMark

What is the autoannotated start?

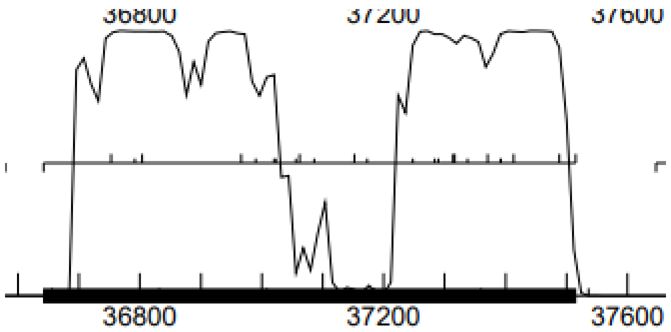
37516

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

There is a gap of 10

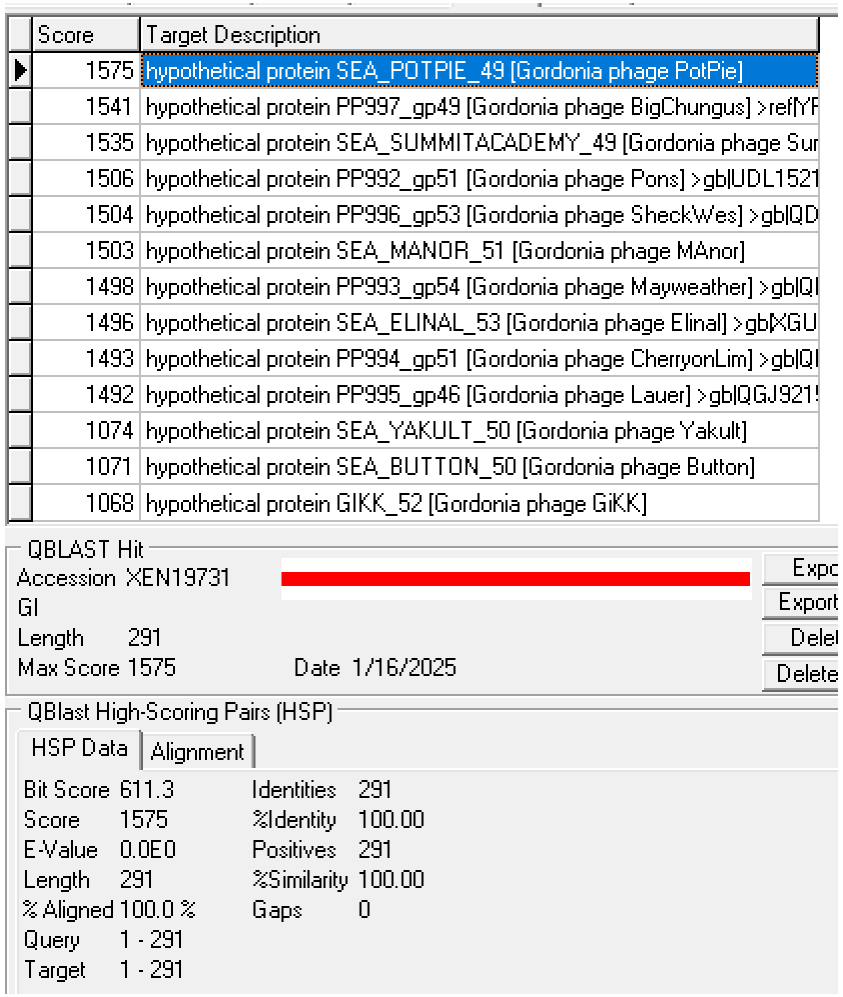
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- The coding potential for this feature starts off slightly before the feature is called to start at 37550 and peaks to strong until dropping to weak around 37210. The potential then peaks back to strong around 37050 before dropping off to nothing at 36690.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are at least 25 BLAST hits
- 9 1:1 alignments
- All hits have e-values that are extremely close to zero



Score	Target Description
1575	hypothetical protein SEA_POTPIE_49 [Gordonia phage PotPie]
1541	hypothetical protein PP997_gp49 [Gordonia phage BigChungus] >ref|YF
1535	hypothetical protein SEA_SUMMITACADEMY_49 [Gordonia phage Sur
1506	hypothetical protein PP992_gp51 [Gordonia phage Pons] >gb|UDL1521
1504	hypothetical protein PP996_gp53 [Gordonia phage SheckWes] >gb|QD
1503	hypothetical protein SEA_MANOR_51 [Gordonia phage MAnor]
1498	hypothetical protein PP993_gp54 [Gordonia phage Mayweather] >gb|QI
1496	hypothetical protein SEA_ELINAL_53 [Gordonia phage Elinal] >gb|XGU
1493	hypothetical protein PP994_gp51 [Gordonia phage CherryonLim] >gb|QI
1492	hypothetical protein PP995_gp46 [Gordonia phage Lauer] >gb|QGJ921!
1074	hypothetical protein SEA_YAKULT_50 [Gordonia phage Yakult]
1071	hypothetical protein SEA_BUTTON_50 [Gordonia phage Button]
1068	hypothetical protein GIKK_52 [Gordonia phage GiKK]

QBLAST Hit

Accession XEN19731

GI

Length 291

Max Score 1575

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	611.3	Identities	291
Score	1575	%Identity	100.00
E-Value	0.0E0	Positives	291
Length	291	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 291		
Target	1 - 291		

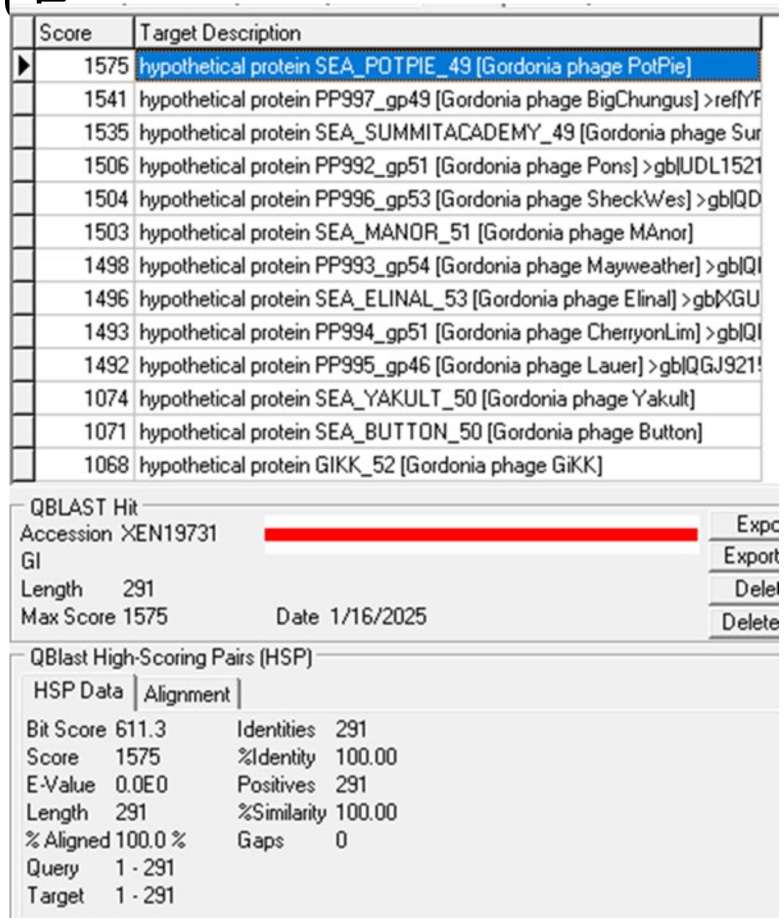
Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This feature is a gene! There is strong coding potential running throughout where the feature is called to be based off the GeneMark file, and there were at least 25 BLAST hits of phages with genes highly similar to this one that had e-values extremely close to zero. Nine of those hits were also 1:1 alignments.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence

Starting at 37516:

- There were at least 25 BLAST hits that all have e-values extremely close to zero.
- There are 9 1:1 alignments



Score	Target Description
1575	hypothetical protein SEA_POTPIE_49 [Gordonia phage PotPie]
1541	hypothetical protein PP997_gp49 [Gordonia phage BigChungus] >ref|YF
1535	hypothetical protein SEA_SUMMITACADEMY_49 [Gordonia phage Sur
1506	hypothetical protein PP992_gp51 [Gordonia phage Pons] >gb|UDL1521
1504	hypothetical protein PP996_gp53 [Gordonia phage SheckWes] >gb|QD
1503	hypothetical protein SEA_MANOR_51 [Gordonia phage MAnor]
1498	hypothetical protein PP993_gp54 [Gordonia phage Mayweather] >gb|QI
1496	hypothetical protein SEA_ELINAL_53 [Gordonia phage Elinal] >gb|XGU
1493	hypothetical protein PP994_gp51 [Gordonia phage CherryonLim] >gb|QI
1492	hypothetical protein PP995_gp46 [Gordonia phage Lauer] >gb|QGJ921!
1074	hypothetical protein SEA_YAKULT_50 [Gordonia phage Yakult]
1071	hypothetical protein SEA_BUTTON_50 [Gordonia phage Button]
1068	hypothetical protein GIKK_52 [Gordonia phage GiKK]

QBLAST Hit

Accession XEN19731

GI

Length 291

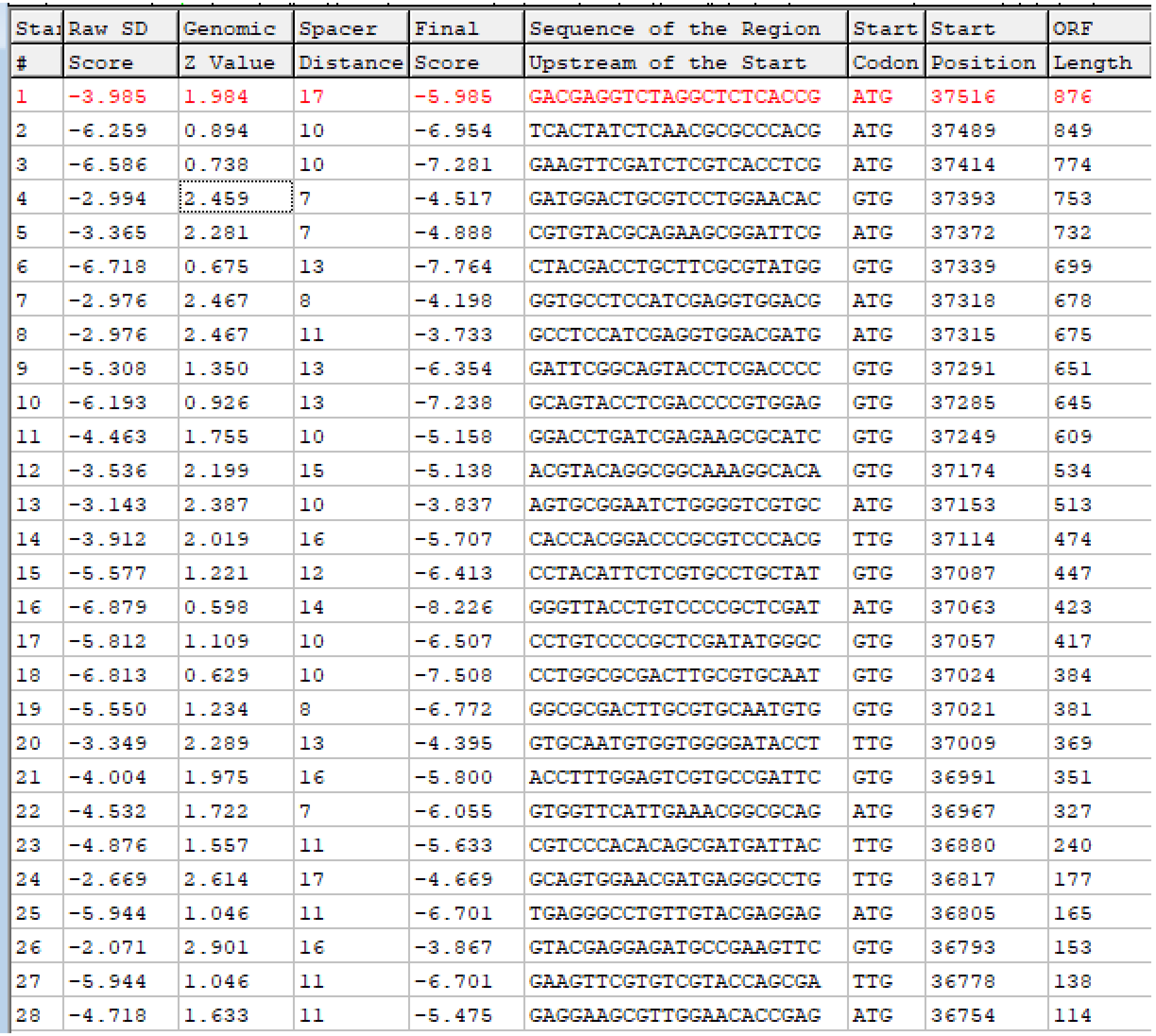
Max Score 1575 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	611.3	Identities	291
Score	1575	%Identity	100.00
E-Value	0.0E0	Positives	291
Length	291	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 291		
Target	1 - 291		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

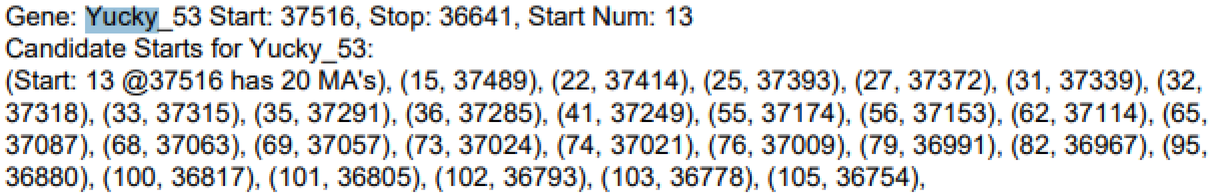
- Starting at 37156:
  - Z-value = 1.984
  - Final score = -5.985
- There were a couple other start sites that had better RBS scores, but they cut of a lot more coding potential and were not recognized by Starterator.



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	-3.985	1.984	17	-5.985	GACGAGGTCTAGGCTCTCACCG	ATG	37516	876
2	-6.259	0.894	10	-6.954	TCACTATCTCAACGCGCCCACG	ATG	37489	849
3	-6.586	0.738	10	-7.281	GAAGTTCGATCTCGTCACCTCG	ATG	37414	774
4	-2.994	2.459	7	-4.517	GATGGACTGCGTCCTGGAACAC	GTG	37393	753
5	-3.365	2.281	7	-4.888	CGTGTACGCAGAAGCGGATTCG	ATG	37372	732
6	-6.718	0.675	13	-7.764	CTACGACCTGCTTCGCGTATGG	GTG	37339	699
7	-2.976	2.467	8	-4.198	GGTGCCTCCATCGAGGTGGACG	ATG	37318	678
8	-2.976	2.467	11	-3.733	GCCTCCATCGAGGTGGACGATG	ATG	37315	675
9	-5.308	1.350	13	-6.354	GATTCGGCAGTACCTCGACCCC	GTG	37291	651
10	-6.193	0.926	13	-7.238	GCAGTACCTCGACCCCGTGGAG	GTG	37285	645
11	-4.463	1.755	10	-5.158	GGACCTGATCGAGAAGCGCATC	GTG	37249	609
12	-3.536	2.199	15	-5.138	ACGTACAGGCGGCAAAGGCACA	GTG	37174	534
13	-3.143	2.387	10	-3.837	AGTGCGGAATCTGGGGTCGTGC	ATG	37153	513
14	-3.912	2.019	16	-5.707	CACCACGGACCCGCGTCCCACG	TTG	37114	474
15	-5.577	1.221	12	-6.413	CCTACATTCTCGTGCCTGCTAT	GTG	37087	447
16	-6.879	0.598	14	-8.226	GGGTTACCTGTCCCCGCTCGAT	ATG	37063	423
17	-5.812	1.109	10	-6.507	CCTGTCCCCGCTCGATATGGGC	GTG	37057	417
18	-6.813	0.629	10	-7.508	CCTGGCGCGACTTGCGTGCAAT	GTG	37024	384
19	-5.550	1.234	8	-6.772	GGCGCGACTTGCGTGCAATGTG	GTG	37021	381
20	-3.349	2.289	13	-4.395	GTGCAATGTGGTGGGATACCT	TTG	37009	369
21	-4.004	1.975	16	-5.800	ACCTTTGGAGTCGTGCCGATTC	GTG	36991	351
22	-4.532	1.722	7	-6.055	GTGGTTCATTGAAACGGCGCAG	ATG	36967	327
23	-4.876	1.557	11	-5.633	CGTCCCACACAGCGATGATTAC	TTG	36880	240
24	-2.669	2.614	17	-4.669	GCAGTGGAACGATGAGGGCCTG	TTG	36817	177
25	-5.944	1.046	11	-6.701	TGAGGGCCTGTTGTACGAGGAG	ATG	36805	165
26	-2.071	2.901	16	-3.867	GTACGAGGAGATGCCGAAGTTC	GTG	36793	153
27	-5.944	1.046	11	-6.701	GAAGTTCGTGTCGTACCAGCGA	TTG	36778	138
28	-4.718	1.633	11	-5.475	GAGGAAGCGTTGGAACACCGAG	ATG	36754	114

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

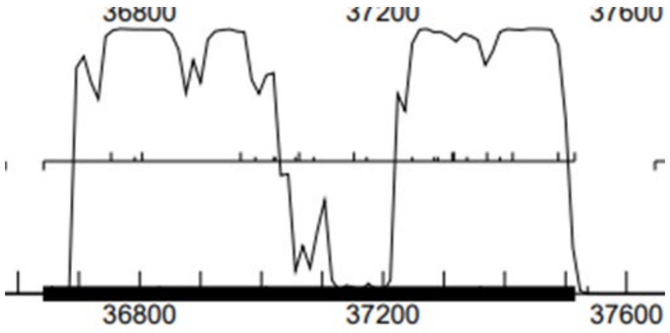
- Yucky does not have the “Most Annotated” start
- Starting at 37516, the autoannotated start, has 20 MA’s
  - It is the only start site with manual annotations



Gene: Yucky_53 Start: 37516, Stop: 36641, Start Num: 13
Candidate Starts for Yucky_53:
(Start: 13 @37516 has 20 MA's), (15, 37489), (22, 37414), (25, 37393), (27, 37372), (31, 37339), (32, 37318), (33, 37315), (35, 37291), (36, 37285), (41, 37249), (55, 37174), (56, 37153), (62, 37114), (65, 37087), (68, 37063), (69, 37057), (73, 37024), (74, 37021), (76, 37009), (79, 36991), (82, 36967), (95, 36880), (100, 36817), (101, 36805), (102, 36793), (103, 36778), (105, 36754),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Starting at 57516:
  - A small amount of coding potential is cut off by starting at this position, but it is also the earliest possible start site.



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

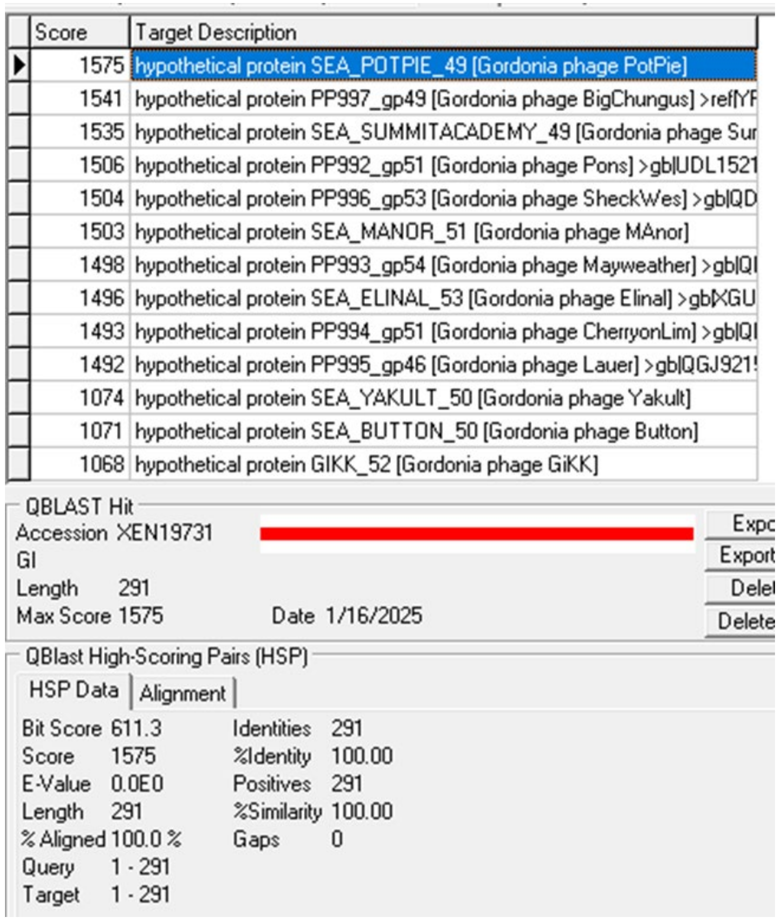
- Starting at 57516 would leave an gap of 10 with the previous gene.

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site for this gene is 57516! There was at least 25 BLAST hits of highly similar phages that all have e-values extremely close to zero. Nine of these hits were 1:1 alignments. The z-value for this start site was 1.984 and the final score was -5.585. There were some start sites with better RBS scores, but they cut off a much larger portion of coding potential. 57516 was the only start site from the starterator report that had any manual annotations for which it had 20. Starting at 57516 would leave a gap of 10 nucleotides with the previous gene.

BLAST function evidence. What assigned functions do other highly similar genes have?

- There were at least 25 BLAST hits that had functions labeled as hypothetical protein.



Score	Target Description
1575	hypothetical protein SEA_POTPIE_49 [Gordonia phage PotPie]
1541	hypothetical protein PP997_gp49 [Gordonia phage BigChungus] >ref|YF
1535	hypothetical protein SEA_SUMMITACADEMY_49 [Gordonia phage Sur
1506	hypothetical protein PP992_gp51 [Gordonia phage Pons] >gb|UDL1521
1504	hypothetical protein PP996_gp53 [Gordonia phage SheckWes] >gb|QD
1503	hypothetical protein SEA_MANOR_51 [Gordonia phage MAnor]
1498	hypothetical protein PP993_gp54 [Gordonia phage Mayweather] >gb|QI
1496	hypothetical protein SEA_ELINAL_53 [Gordonia phage Elinal] >gb|XGU
1493	hypothetical protein PP994_gp51 [Gordonia phage CherryonLim] >gb|QI
1492	hypothetical protein PP995_gp46 [Gordonia phage Lauer] >gb|QGJ921
1074	hypothetical protein SEA_YAKULT_50 [Gordonia phage Yakult]
1071	hypothetical protein SEA_BUTTON_50 [Gordonia phage Button]
1068	hypothetical protein GIKK_52 [Gordonia phage GiKK]

QBLAST Hit

Accession XEN19731

GI

Length 291

Max Score 1575

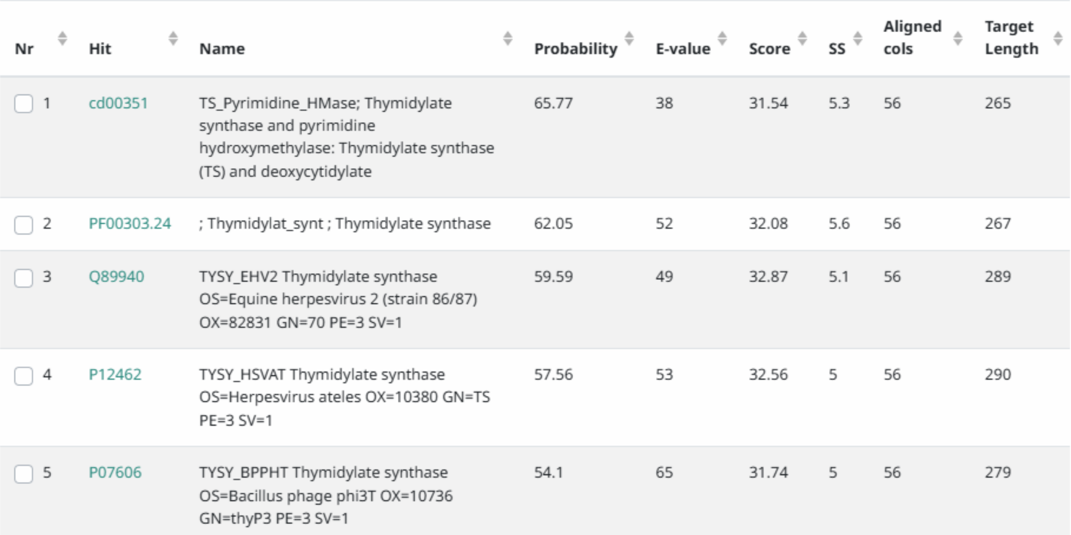
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	611.3	Identities	291
Score	1575	%Identity	100.00
E-Value	0.0E0	Positives	291
Length	291	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 291		
Target	1 - 291		

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

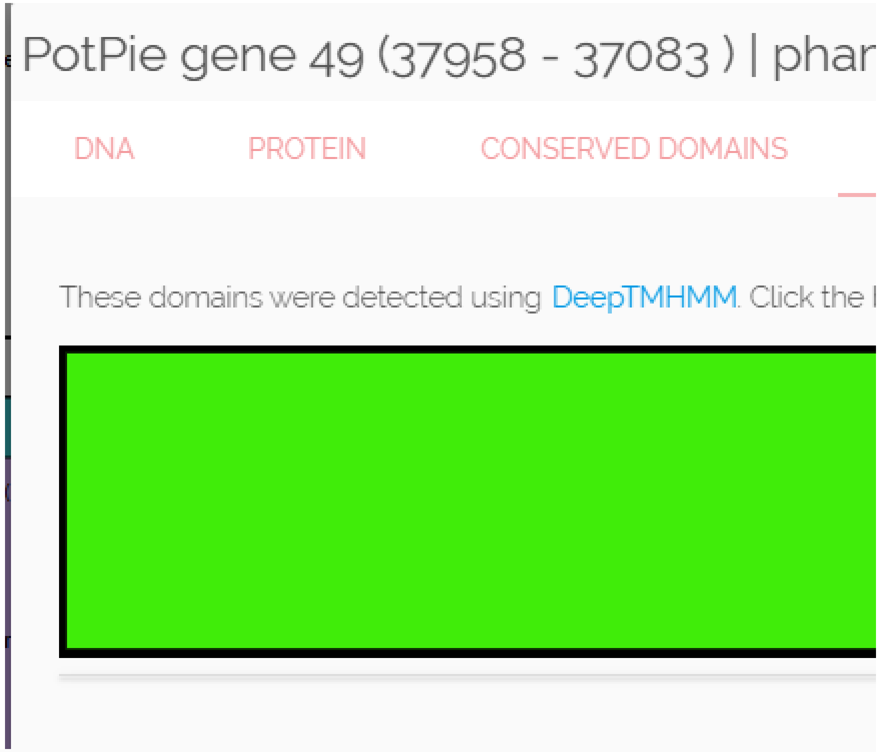
- Hhpred did not show any hits with probabilities above 90, and any hits that were there only matched with portions of the gene.
- There were no conserved domains present.



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	cd00351	TS_Pyrimidine_HMase; Thymidylate synthase and pyrimidine hydroxymethylase: Thymidylate synthase (TS) and deoxycytidylate	65.77	38	31.54	5.3	56	265
□ 2	PF00303.24	; Thymidylat_synt ; Thymidylate synthase	62.05	52	32.08	5.6	56	267
□ 3	Q89940	TYSY_EHV2 Thymidylate synthase OS=Equine herpesvirus 2 (strain 86/87) OX=82831 GN=70 PE=3 SV=1	59.59	49	32.87	5.1	56	289
□ 4	P12462	TYSY_HSVAT Thymidylate synthase OS=Herpesvirus ateles OX=10380 GN=TS PE=3 SV=1	57.56	53	32.56	5	56	290
□ 5	P07606	TYSY_BPPHT Thymidylate synthase OS=Bacillus phage phi3T OX=10736 GN=thyP3 PE=3 SV=1	54.1	65	31.74	5	56	279

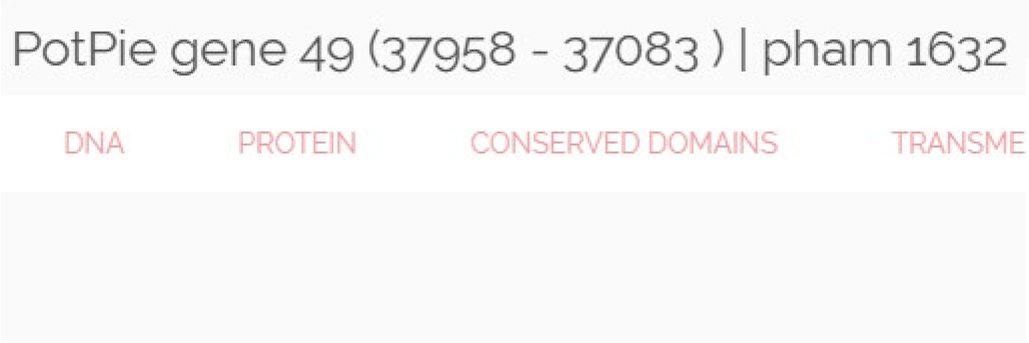
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Closely related phages with genes in the same pham as this one do not have a designate function or conserved domains.
- No evidence to support predicting a function for this gene.



PotPie gene 49 (37958 - 37083) | phar

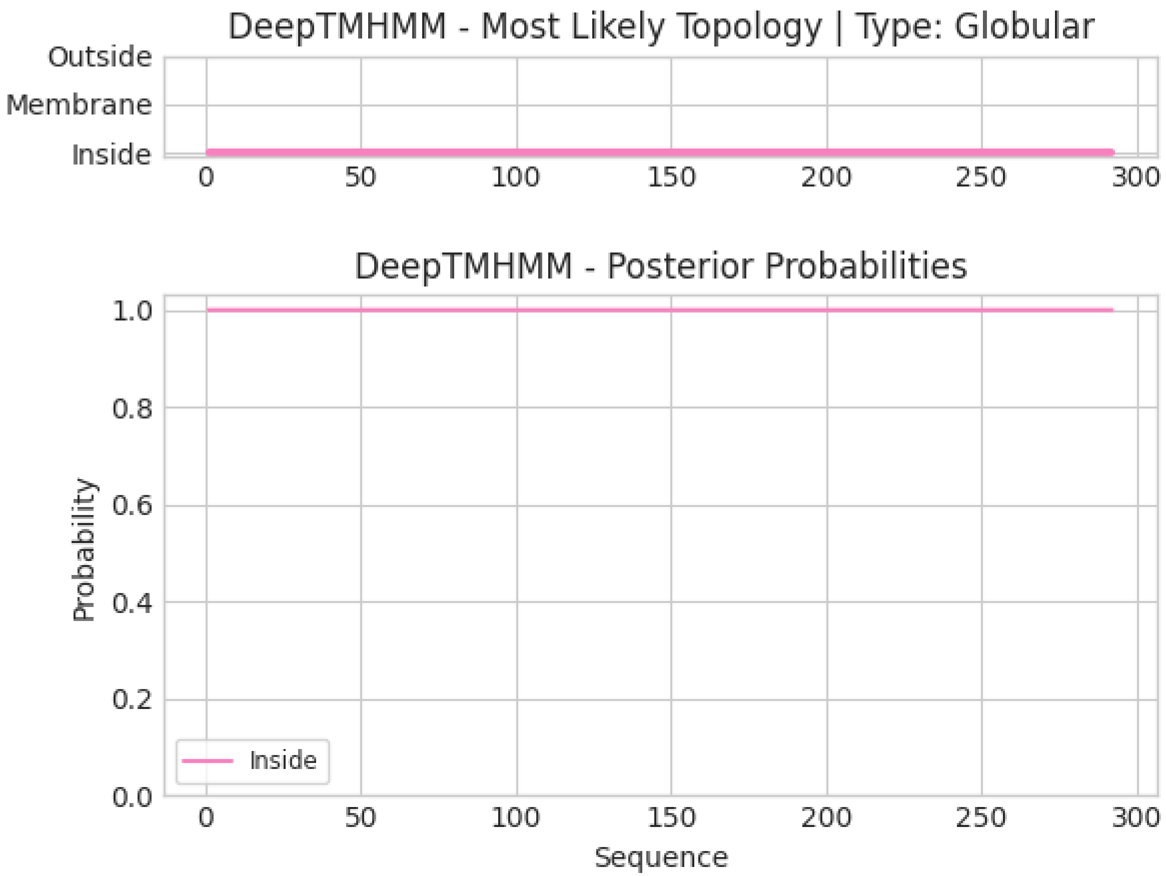
These domains were detected using DeepTMHMM. Click the l



PotPie gene 49 (37958 - 37083) | pham 1632

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- There were no transmembrane domains.



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Offical function → hypothetical protein
- The function for this gene should be labeled as a hypothetical protein. There were at least 25 BLAST hits of highly similar genes from other phages that had functions of hypothetical protein, and Hhpred did not show any hits with probabilities above 90. Phamerator also did not predict a function for this gene as phages with genes in this pham did not have assigned functions and there were no conserved domains present. Deep TMHMM did not predict any transmembrane domains, so the function cannot be labeled as a membrane protein either.

Feature 53 – Reverse – Stop 37527

Glimmer/GeneMark

What feature number is this? **53**

What is the stop site? **37527**

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Glimmer and GeneMark

What is the autoannotated start?

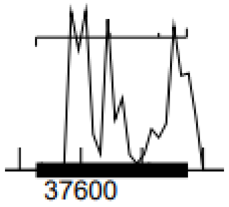
37775

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

There is an overlap of 8

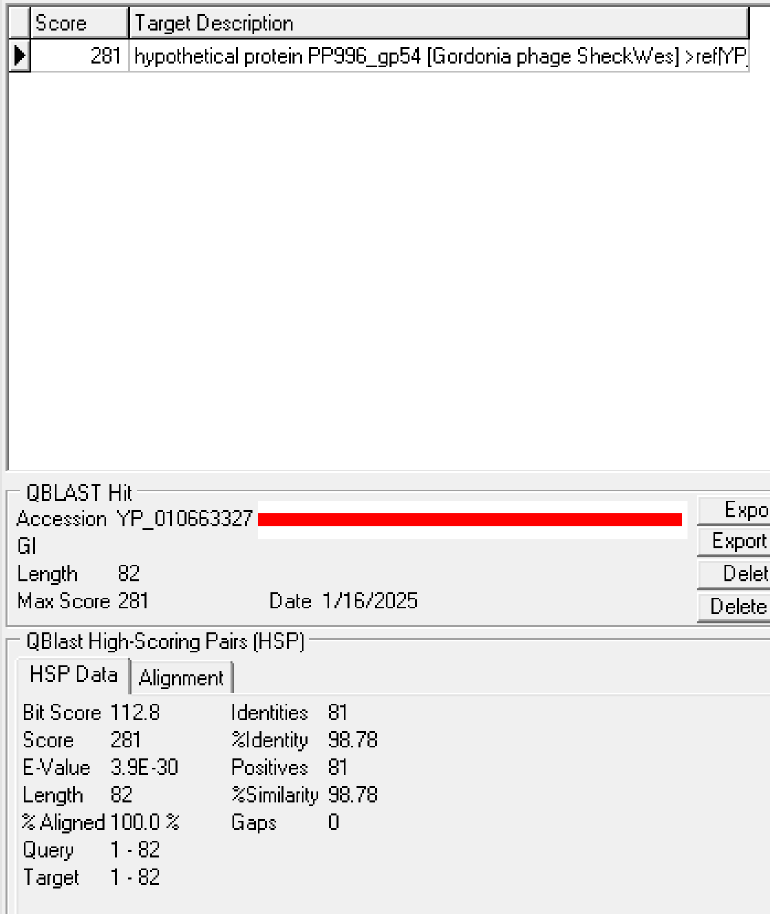
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- There is coding potential call throughout where the feature is called to be. A majority of it is weak with periodic peaks into strong coding potential.
- The earliest start site is the autoannotated start of 37775, but it does cut off part of the initial peak of coding potential for this feature.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There was only one BLAST hit for this feature, but it was 1:1 alignment and had an e-value extremely close to zero.



Score	Target Description
281	hypothetical protein PP996_gp54 [Gordonia phage SheckWes] >ref|YP_

QBLAST Hit
Accession YP_010663327
GI
Length 82
Max Score 281 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

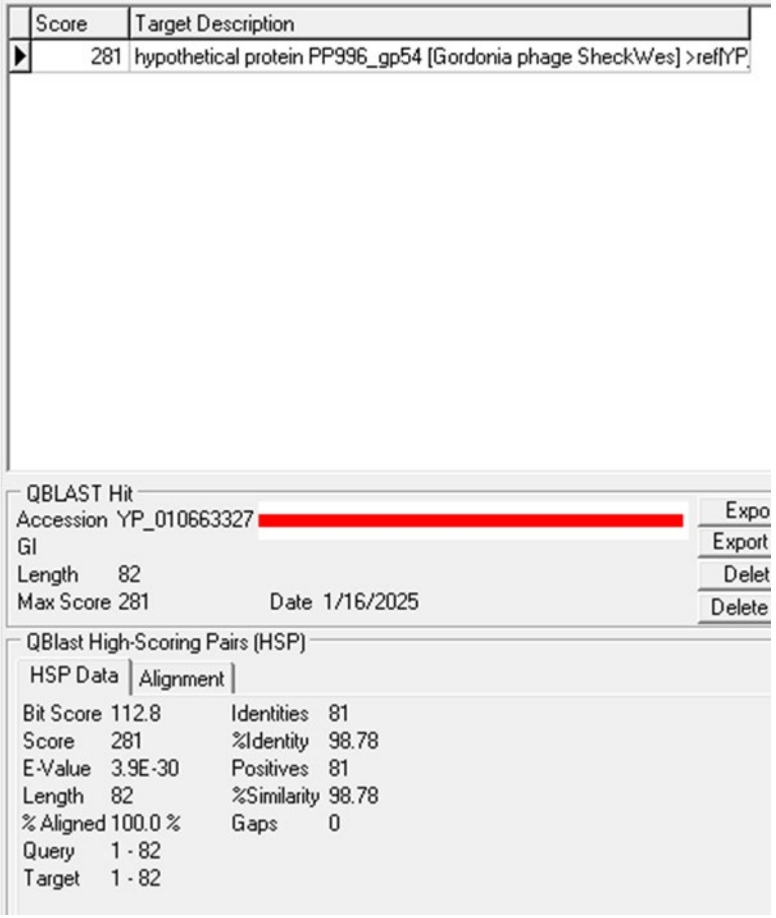
Bit Score	112.8	Identities	81
Score	281	%Identity	98.78
E-Value	3.9E-30	Positives	81
Length	82	%Similarity	98.78
% Aligned	100.0 %	Gaps	0
Query	1 - 82		
Target	1 - 82		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- I would call this feature a gene! There is coding potential running throughout where the feature is called to be alternated between strong and weak. There was also a BLAST hit with an e-value extremely close to zero that was also a 1:1 alignment.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There was only 1 BLAST hit for this gene with the phage SheckWes that had an e-value extremely close to zero. This hit was a 1:1 alignment.



Score	Target Description
281	hypothetical protein PP996_gp54 [Gordonia phage SheckWes] >ref|YP_

QBLAST Hit

Accession YP_010663327

GI

Length 82

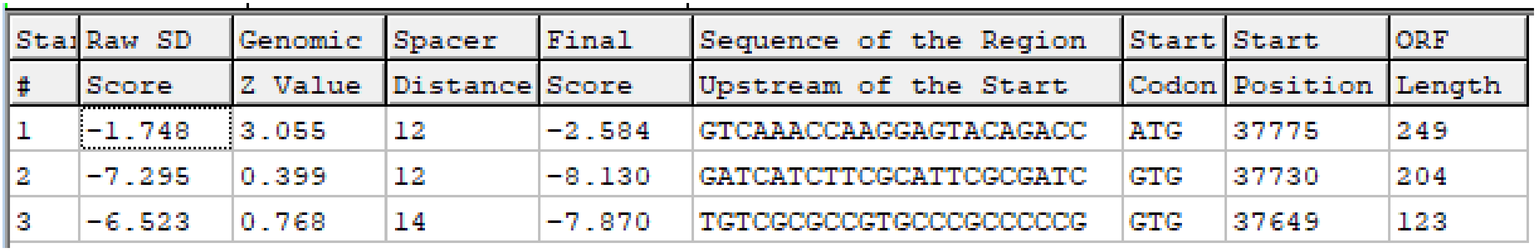
Max Score 281 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	112.8	Identities	81
Score	281	%Identity	98.78
E-Value	3.9E-30	Positives	81
Length	82	%Similarity	98.78
% Aligned	100.0 %	Gaps	0
Query	1 - 82		
Target	1 - 82		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

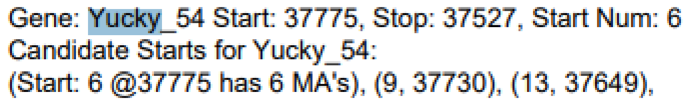
- Starting at 37775:
  - Z-value = 3.055
  - Final score = -2.584
- The autoannotated start site had the best RBS scores.



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	-1.748	3.055	12	-2.584	GTCAAACCAAGGAGTACAGACC	ATG	37775	249
2	-7.295	0.399	12	-8.130	GATCATCTTCGCATTCGCGATC	GTG	37730	204
3	-6.523	0.768	14	-7.870	TGTCGCGCCGTGCCCGCCCCCG	GTG	37649	123

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

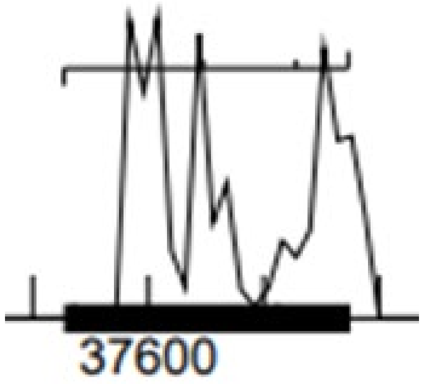
- 37775 was the only start site that had manual annotations. There were 6 MA's for this start site.



Gene: Yucky_54 Start: 37775, Stop: 37527, Start Num: 6
Candidate Starts for Yucky_54:
(Start: 6 @37775 has 6 MA's), (9, 37730), (13, 37649),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- 37775 is the earliest possible start site, but it does cut off part of the initial peak of coding potential. A majority of the coding potential is included.
- Any start site after 37775 would cut off a larger amount of coding potential.



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

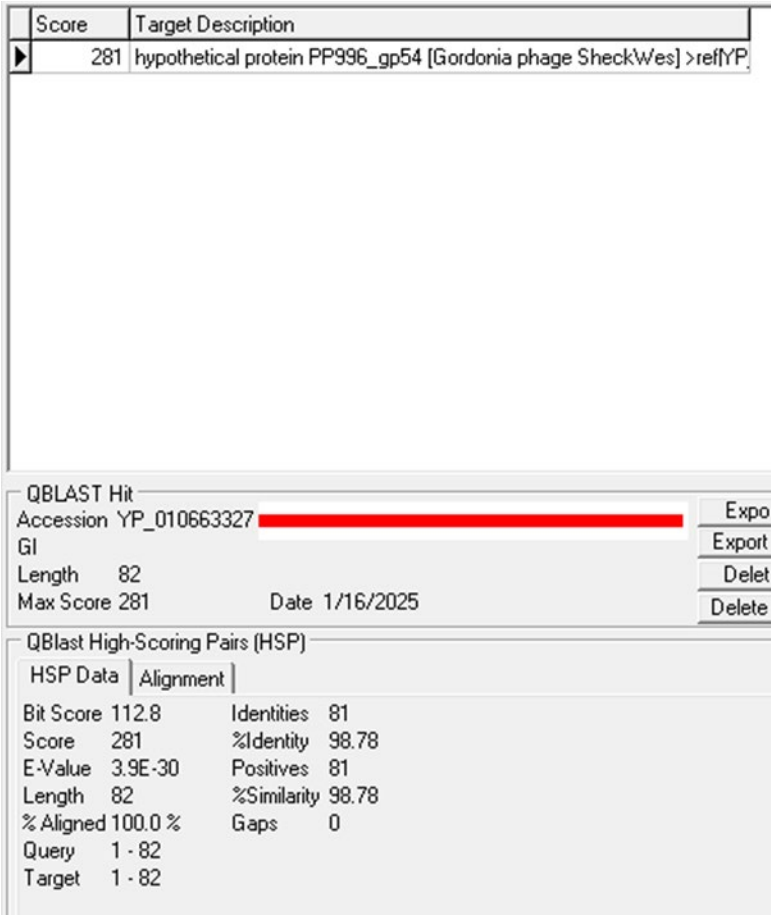
- Starting at 37775 would leave an overlap of 8 with the previous gene.

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site for this feature is 37775. There was only one BLAST hit for this start site with the phage SheckWes and it was a 1:1 alignment with an e-value extremely close to zero. This start site also had the best RBS values of all the possible start sites (z-value of 3.055 and a final score of -2.584). 37775 was the only start site that had manual annotations according to the starterator report (6 manual annotations). This start site does cut off part of the initial peak of coding potential, but a majority of it is included. There would be an overlap of 8 with the previous feature, but this is not an unfavorable condition.

BLAST function evidence. What assigned functions do other highly similar genes have?

- There was only one BLAST hit and it had hypothetical protein as the function.



Score	Target Description
281	hypothetical protein PP996_gp54 [Gordonia phage SheckWes] >ref|YP

QBLAST Hit

Accession YP_010663327

GI

Length 82

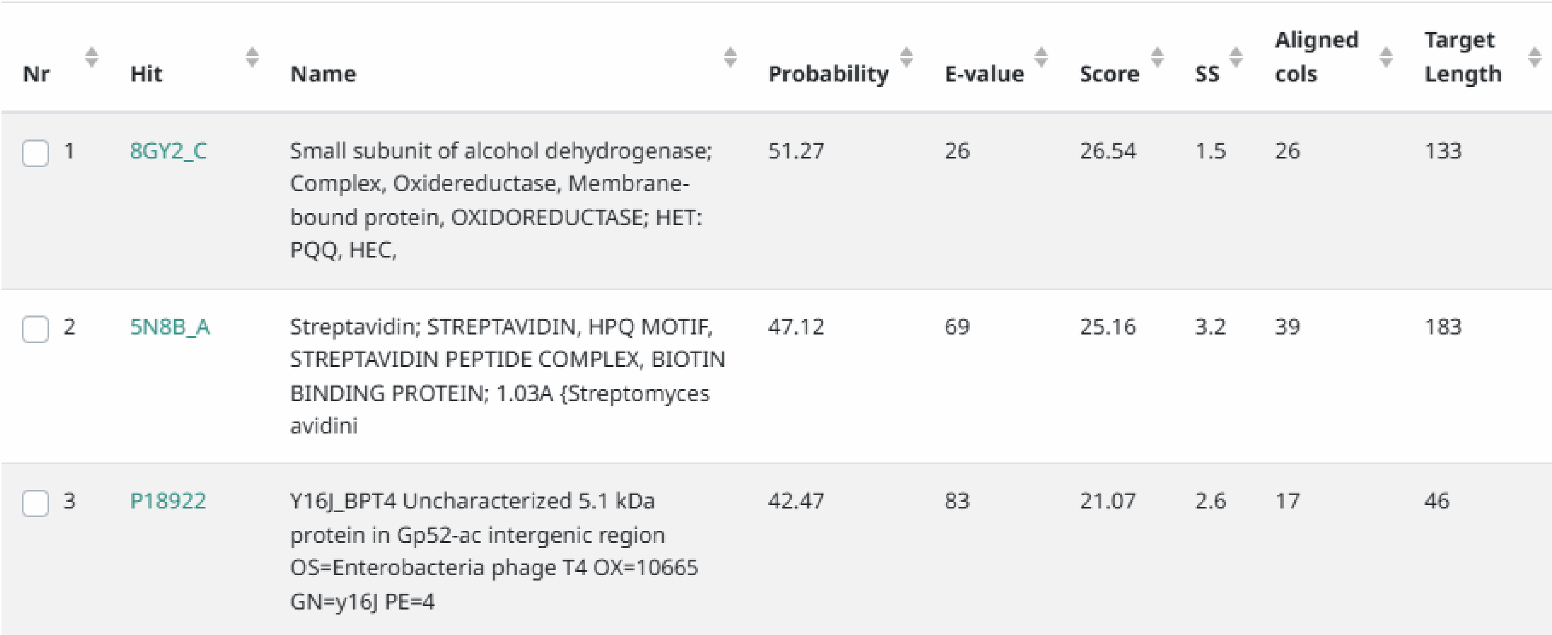
Max Score 281 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	112.8	Identities	81
Score	281	%Identity	98.78
E-Value	3.9E-30	Positives	81
Length	82	%Similarity	98.78
% Aligned	100.0 %	Gaps	0
Query	1 - 82		
Target	1 - 82		

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

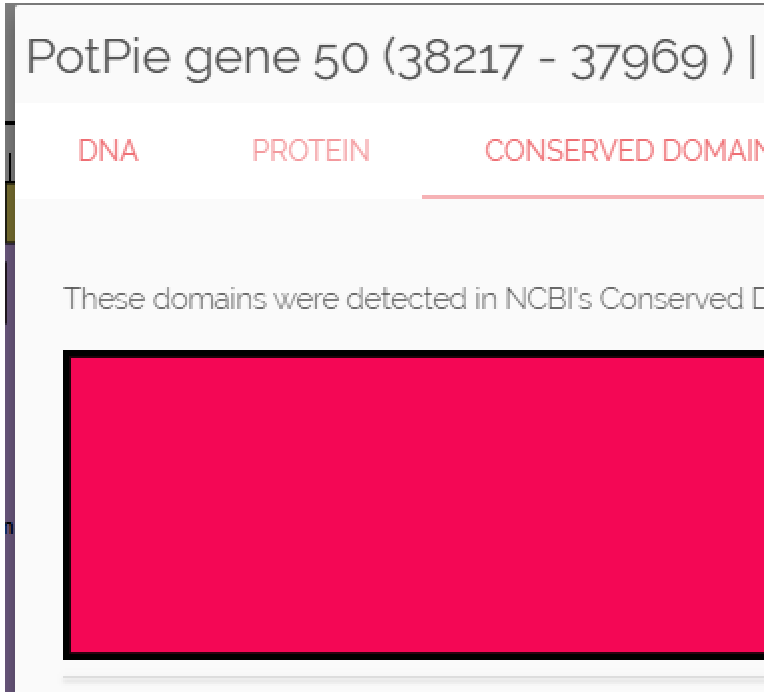
- Hhpred did not have any hits with probability above 90 (the highest was 51.27), so the results did not support the assignment of a function for this gene.



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
☐ 1	8GY2_C	Small subunit of alcohol dehydrogenase; Complex, Oxidereductase, Membrane-bound protein, OXIDOREDUCTASE; HET: PQQ, HEC,	51.27	26	26.54	1.5	26	133
☐ 2	5N8B_A	Streptavidin; STREPTAVIDIN, HPQ MOTIF, STREPTAVIDIN PEPTIDE COMPLEX, BIOTIN BINDING PROTEIN; 1.03A {Streptomyces avidini	47.12	69	25.16	3.2	39	183
☐ 3	P18922	Y16J_BPT4 Uncharacterized 5.1 kDa protein in Gp52-ac intergenic region OS=Enterobacteria phage T4 OX=10665 GN=y16J PE=4	42.47	83	21.07	2.6	17	46

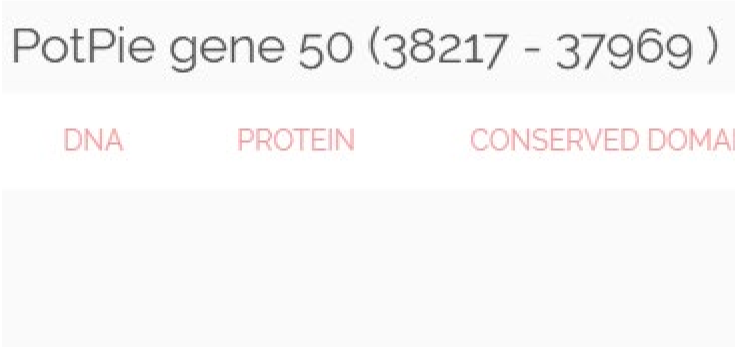
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Phages with genes in the same pham do not predict a function for this gene. There were no conserved domains or specific functions assigned to them.



PotPie gene 50 (38217 - 37969) |

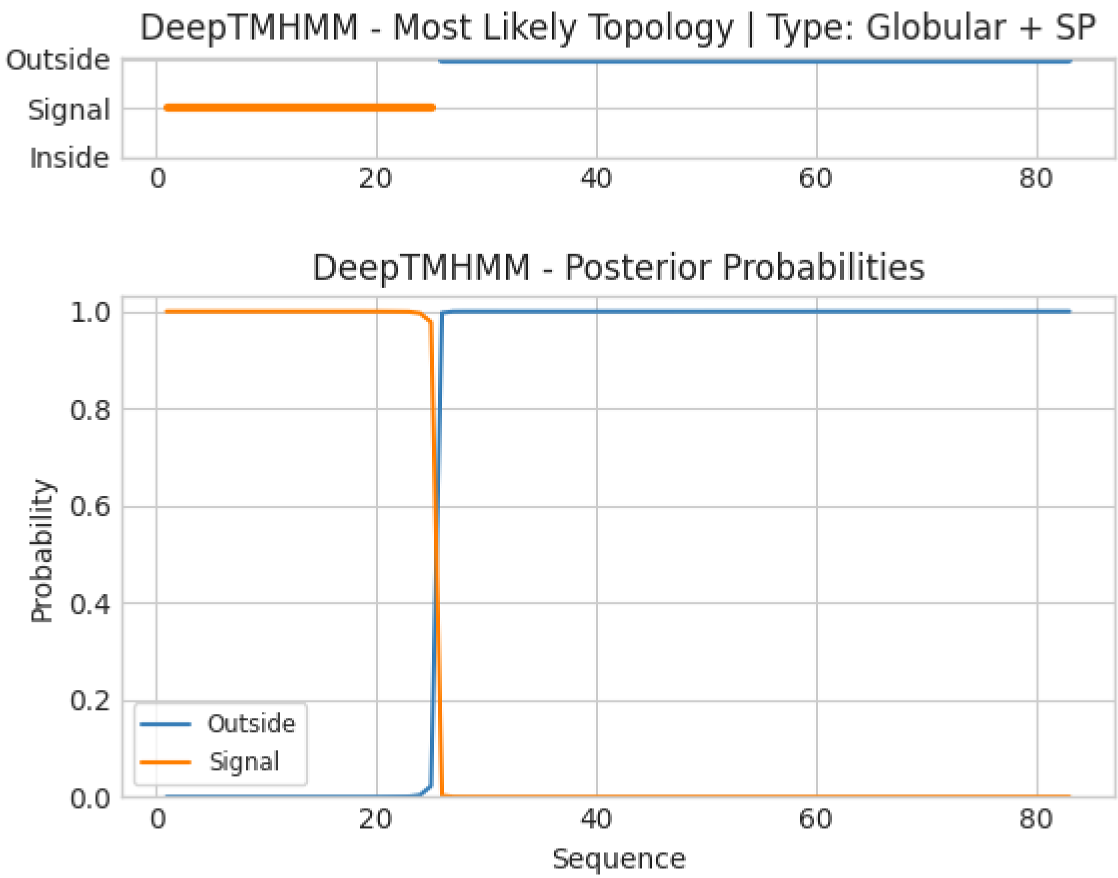
These domains were detected in NCBI's Conserved D



PotPie gene 50 (38217 - 37969)

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- Deep TMHMM showed evidence of transmembrane domains, so the function of this gene can be categorized as a membrane protein over a hypothetical protein.



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Official function → membrane protein
- There was only one BLAST hit for this gene, and it had the function of hypothetical protein. Hhpred did not show any hits with probabilities over 90, so it did not support the assignment of a specific function. Phamerator showed that phages with genes in the same pham do not have designated functions or conserved domains, so it also did not support an assignment of a specific function for this gene. The Deep TMHMM graph for this gene showed transmembrane domains, so the function should be labeled as a membrane protein.

Feature 54 – Reverse – Stop 37768

Glimmer/GeneMark

What feature number is this? **54**

What is the stop site? **37768**

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Glimmer

What is the autoannotated start?

37923

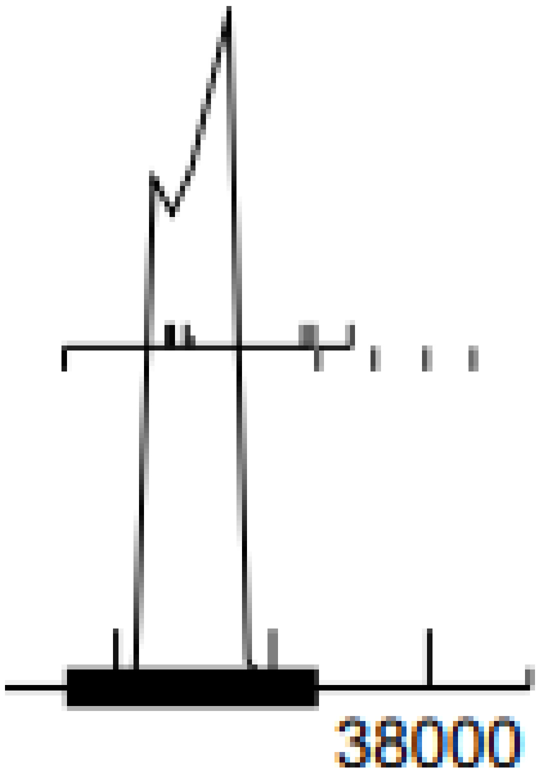
Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

There would be an overlap of 1

- Genemark called start at 37929 (there would be an overlap of 7)

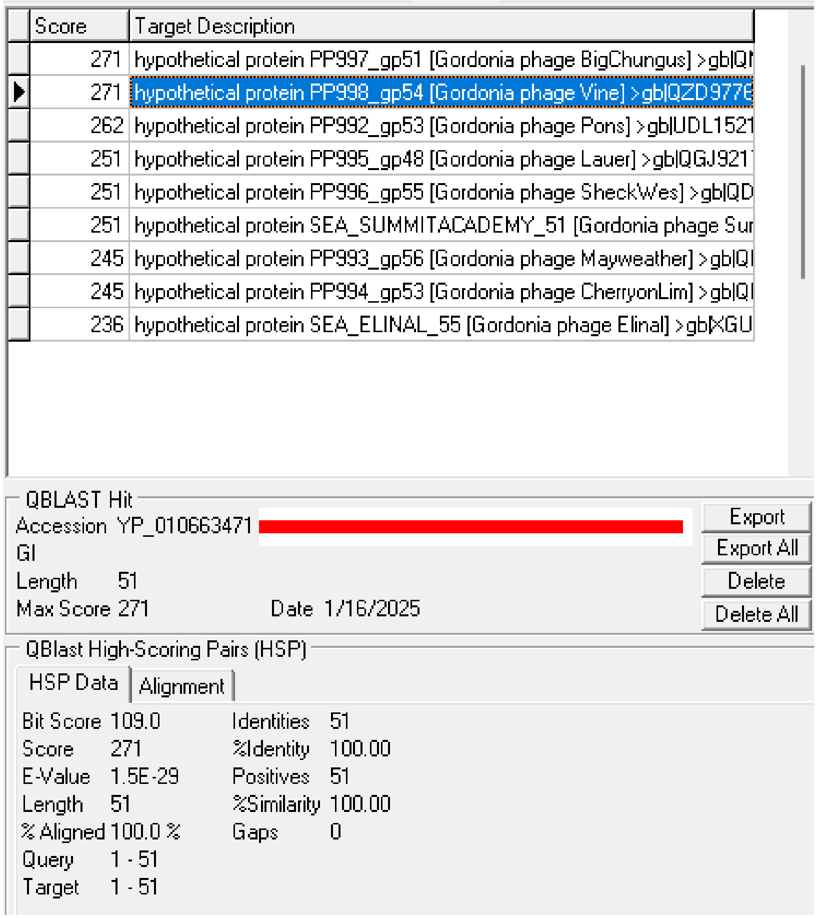
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Coding potential starts at 37900 immediately peaking to strong and staying that way until falling off 37820.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There were 9 BLAST hits for highly similar genes to this one that all have e-values extremely close to zero.
- 6 of these hits were 1:1 alignments



Score	Target Description
271	hypothetical protein PP997_gp51 [Gordonia phage BigChungus] >gb|Q
271	hypothetical protein PP998_gp54 [Gordonia phage Vine] >gb|QZD977
262	hypothetical protein PP992_gp53 [Gordonia phage Pons] >gb|UDL1521
251	hypothetical protein PP995_gp48 [Gordonia phage Lauer] >gb|QGJ921
251	hypothetical protein PP996_gp55 [Gordonia phage SheckWes] >gb|QD
251	hypothetical protein SEA_SUMMITACADEMY_51 [Gordonia phage Su
245	hypothetical protein PP993_gp56 [Gordonia phage Mayweather] >gb|Q
245	hypothetical protein PP994_gp53 [Gordonia phage CherryonLim] >gb|Q
236	hypothetical protein SEA_ELINAL_55 [Gordonia phage Elinal] >gb|XGU

QBLAST Hit

Accession YP_010663471

GI

Length 51

Max Score 271 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

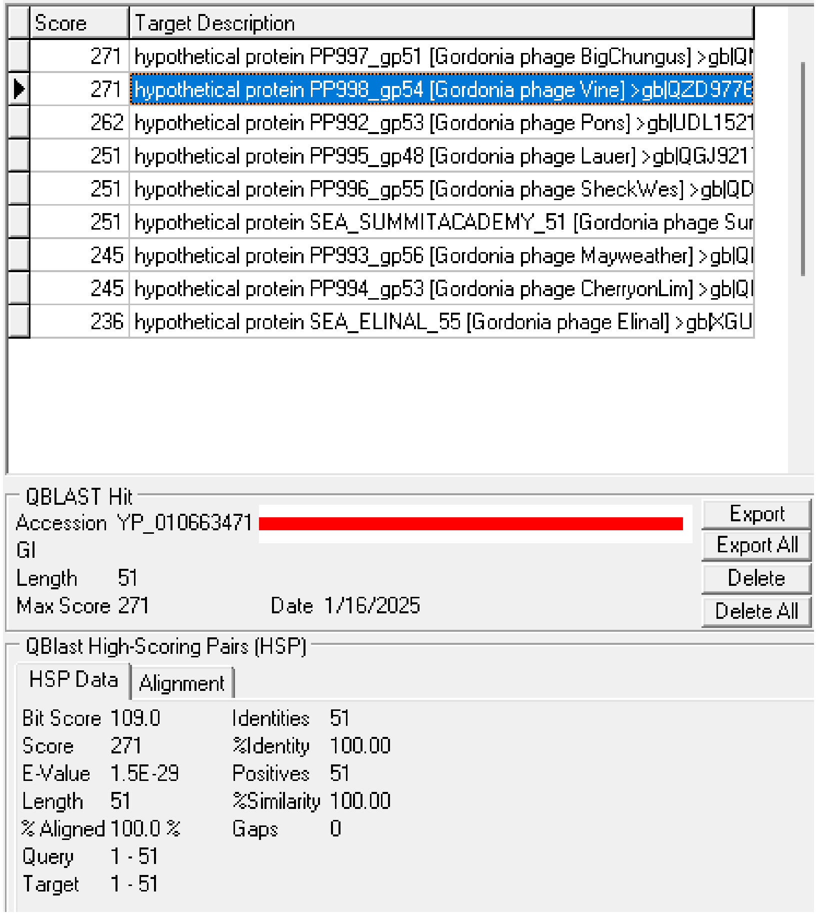
Bit Score	109.0	Identities	51
Score	271	%Identity	100.00
E-Value	1.5E-29	Positives	51
Length	51	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 51		
Target	1 - 51		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This feature is a gene! There were 9 BLAST hits of highly similar genes that have e-values extremely close to zero. Six of these hits were 1:1 alignments. There is also strong coding potential throughout where the feature is called to be.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence

- Starting at 37923:
  - There were 9 BLAST hits
  - 6 1:1 alignments
- Starting at 37929:
  - Need to look into this one



Score	Target Description
271	hypothetical protein PP997_gp51 [Gordonia phage BigChungus] >gb|Q
271	hypothetical protein PP998_gp54 [Gordonia phage Vine] >gb|QZD977
262	hypothetical protein PP992_gp53 [Gordonia phage Pons] >gb|UDL1521
251	hypothetical protein PP995_gp48 [Gordonia phage Lauer] >gb|QGJ921
251	hypothetical protein PP996_gp55 [Gordonia phage SheckWes] >gb|QD
251	hypothetical protein SEA_SUMMITACADEMY_51 [Gordonia phage Su
245	hypothetical protein PP993_gp56 [Gordonia phage Mayweather] >gb|Q
245	hypothetical protein PP994_gp53 [Gordonia phage CherryonLim] >gb|Q
236	hypothetical protein SEA_ELINAL_55 [Gordonia phage Elinal] >gb|XGU

QBLAST Hit

Accession YP_010663471
GI
Length 51
Max Score 271
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

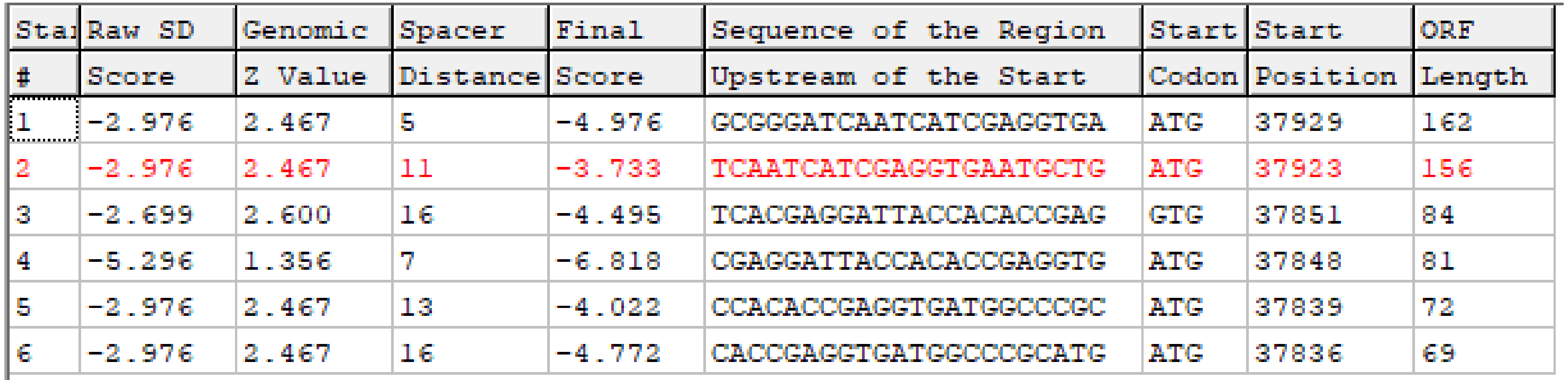
Bit Score 109.0
Score 271
E-Value 1.5E-29
Length 51
% Aligned 100.0 %
Query 1 - 51
Target 1 - 51

Identities 51
%Identity 100.00
Positives 51
%Similarity 100.00
Gaps 0

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Starting at 37923:
  - Z-value = 2.467
  - Final score = -3.733

- Starting at 37929:
  - Z-value = 2.467
  - Final score = -4.976

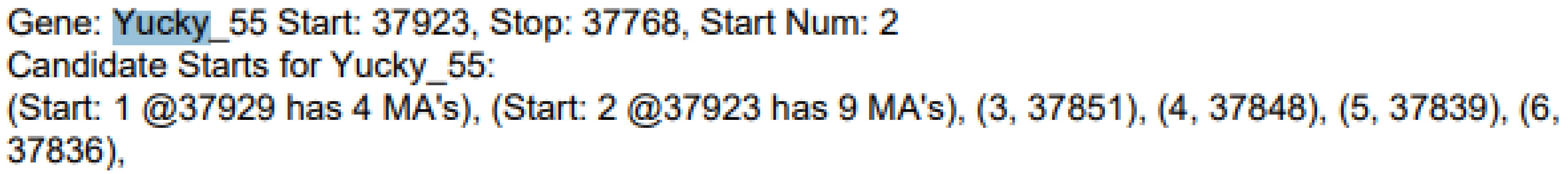


Sta	Raw SD	Genomic	Spacer	Final	Sequence of the Region	Start	Start	ORF
#	Score	Z Value	Distance	Score	Upstream of the Start	Codon	Position	Length
1	-2.976	2.467	5	-4.976	GCGGGATCAATCATCGAGGTGA	ATG	37929	162
2	-2.976	2.467	11	-3.733	TCAATCATCGAGGTGAATGCTG	ATG	37923	156
3	-2.699	2.600	16	-4.495	TCACGAGGATTACCACACCGAG	GTG	37851	84
4	-5.296	1.356	7	-6.818	CGAGGATTACCACACCGAGGTG	ATG	37848	81
5	-2.976	2.467	13	-4.022	CCACACCGAGGTGATGGCCCGC	ATG	37839	72
6	-2.976	2.467	16	-4.772	CACCGAGGTGATGGCCCGCATG	ATG	37836	69

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Starting at 37923:
  - 9 manual annotation

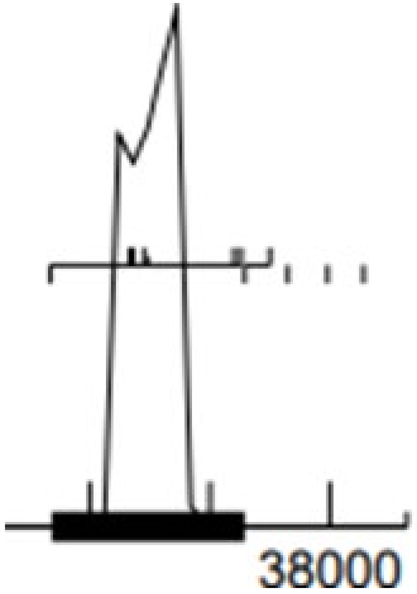
- Starting at 37929:
  - 4 manual annotations



Gene: Yucky_55 Start: 37923, Stop: 37768, Start Num: 2
Candidate Starts for Yucky_55:
(Start: 1 @37929 has 4 MA's), (Start: 2 @37923 has 9 MA's), (3, 37851), (4, 37848), (5, 37839), (6, 37836),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Starting at 37923:
  - Doesn't cut off any coding potential
- Starting at 37929:
  - Doesn't cut off any coding potential.
  - The extra few nucleotides added with this start site don't include any more coding potential than the autoannotated start site.



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Starting at 37923:
  - There would be an overlap of 1

- Starting at 37929:
  - There would be an overlap of 7

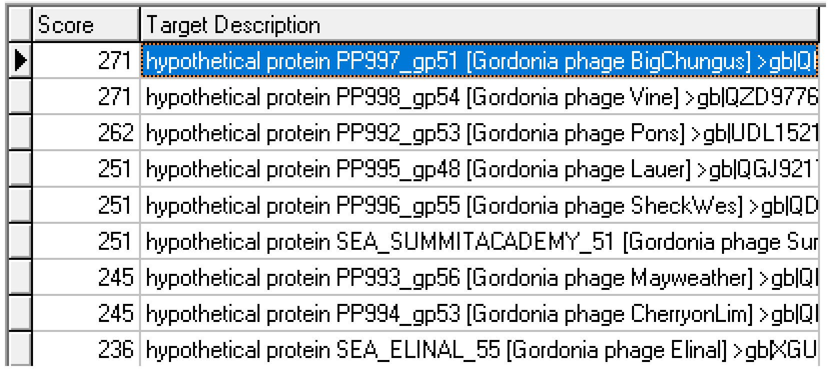
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	Starting at 37923	Starting at 37929
Glimmer/GeneMark	Glimmer	GeneMark
BLAST	6 1:1 alignments	Haven't been able to look at it
RBS scores	Z-value = 2.467 Final score = -3.733	Z-value = 2.467 Final score = -4.976
Starterator	9 MA's	4 MA's
GeneMark	All coding potential included	All coding potential included
Gap/Overlap	Overlap of 1	Overlap of 7

The start site is 37923! This start site was called by Glimmer only, and it had 9 MA's whereas the start site called by glimmer only had 4. It also has 6 1:1 alignments according to BLAST. The z-value for both start sites was 2.467, but 37923 had the better final score of -3.733. Both start site included all the possible coding potential for the gene, but 37923 had a smaller overlap of only 1 nucleotide.

BLAST function evidence. What assigned functions do other highly similar genes have?

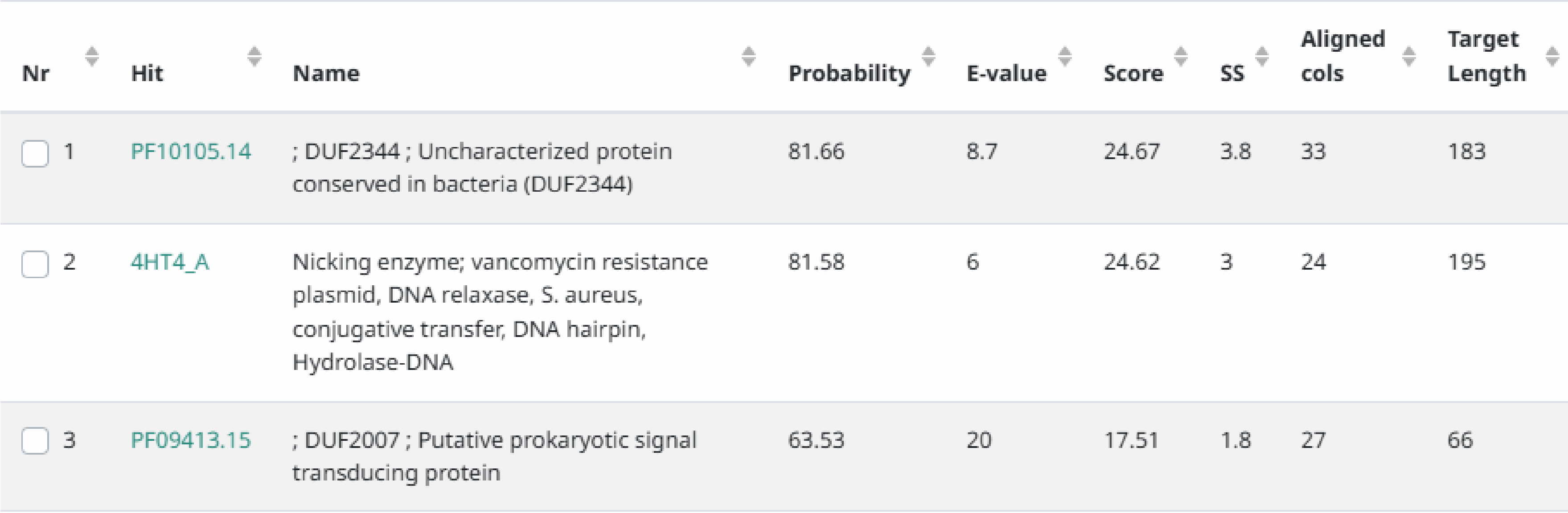
- All 9 BLAST hits showed the function of hypothetical protein.



Score	Target Description
271	hypothetical protein PP997_gp51 [Gordonia phage BigChungus] >gb|QI
271	hypothetical protein PP998_gp54 [Gordonia phage Vine] >gb|QZD9776
262	hypothetical protein PP992_gp53 [Gordonia phage Pons] >gb|UDL1521
251	hypothetical protein PP995_gp48 [Gordonia phage Lauer] >gb|QGJ921
251	hypothetical protein PP996_gp55 [Gordonia phage SheckWes] >gb|QD
251	hypothetical protein SEA_SUMMITACADEMY_51 [Gordonia phage Sur
245	hypothetical protein PP993_gp56 [Gordonia phage Mayweather] >gb|QI
245	hypothetical protein PP994_gp53 [Gordonia phage CherryonLim] >gb|QI
236	hypothetical protein SEA_ELINAL_55 [Gordonia phage Elinal] >gb|XGU

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

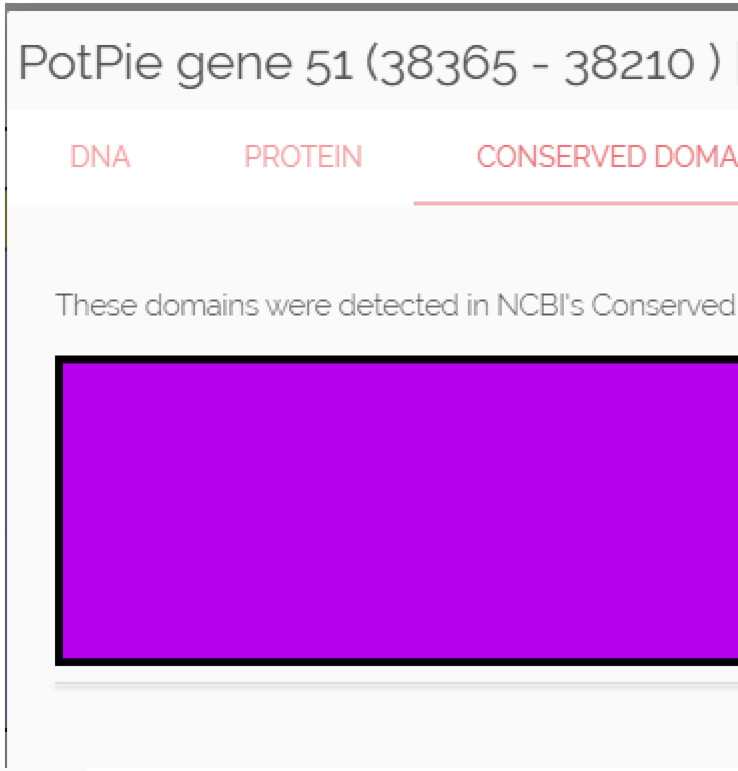
- There were no Hhpred hits with probabilities over 90, so it does not support the assignment of a function for this gene.
- There were no conserved domains present.



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	PF10105.14	; DUF2344 ; Uncharacterized protein conserved in bacteria (DUF2344)	81.66	8.7	24.67	3.8	33	183
□ 2	4HT4_A	Nicking enzyme; vancomycin resistance plasmid, DNA relaxase, S. aureus, conjugative transfer, DNA hairpin, Hydrolase-DNA	81.58	6	24.62	3	24	195
□ 3	PF09413.15	; DUF2007 ; Putative prokaryotic signal transducing protein	63.53	20	17.51	1.8	27	66

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Phages with genes in the same pham do not predict a function for this gene. They do not show assigned function or the presence of conserved domains.



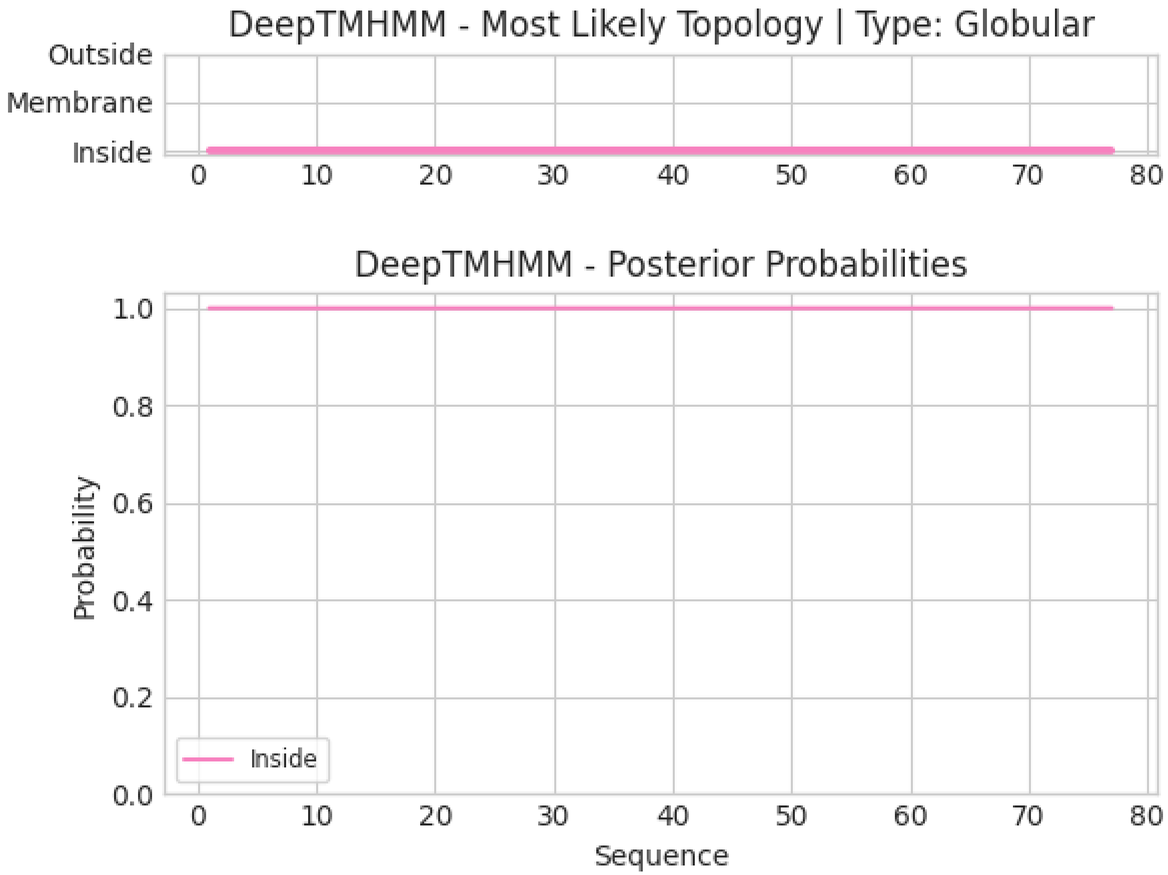
PotPie gene 51 (38365 - 38210)

These domains were detected in NCBI's Conserved

PotPie gene 51 (38365 - 38210)

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- There is no presence of transmembrane domains.



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Official function → hypothetical protein
- The function for this gene should be labeled as hypothetical protein. All 9 of the BLAST hits showed functions of hypothetical protein. Hhpred did not show any hits with probabilities over 90, so it doesn’t support the assignment of a specific function of this gene. Phamerator showed that phages with genes in the same pham do not have designated function or show the presence of conserved domains. The Deep TMHMM graph showed that there were no transmembrane domains, so it cannot be labeled as a membrane potein.

Feature 55 – Reverse – Stop 37923

Glimmer/GeneMark

What feature number is this? **55**

What is the stop site? **37923**

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Called by Glimmer and GeneMark

What is the autoannotated start?

38153

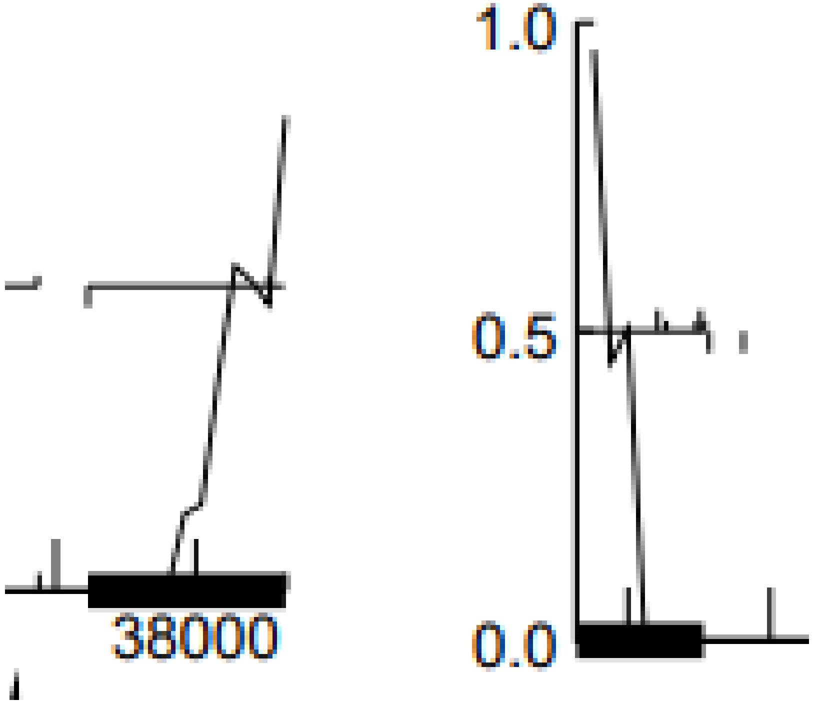
Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

There is an overlap of 1

- Previous feature end at 38153

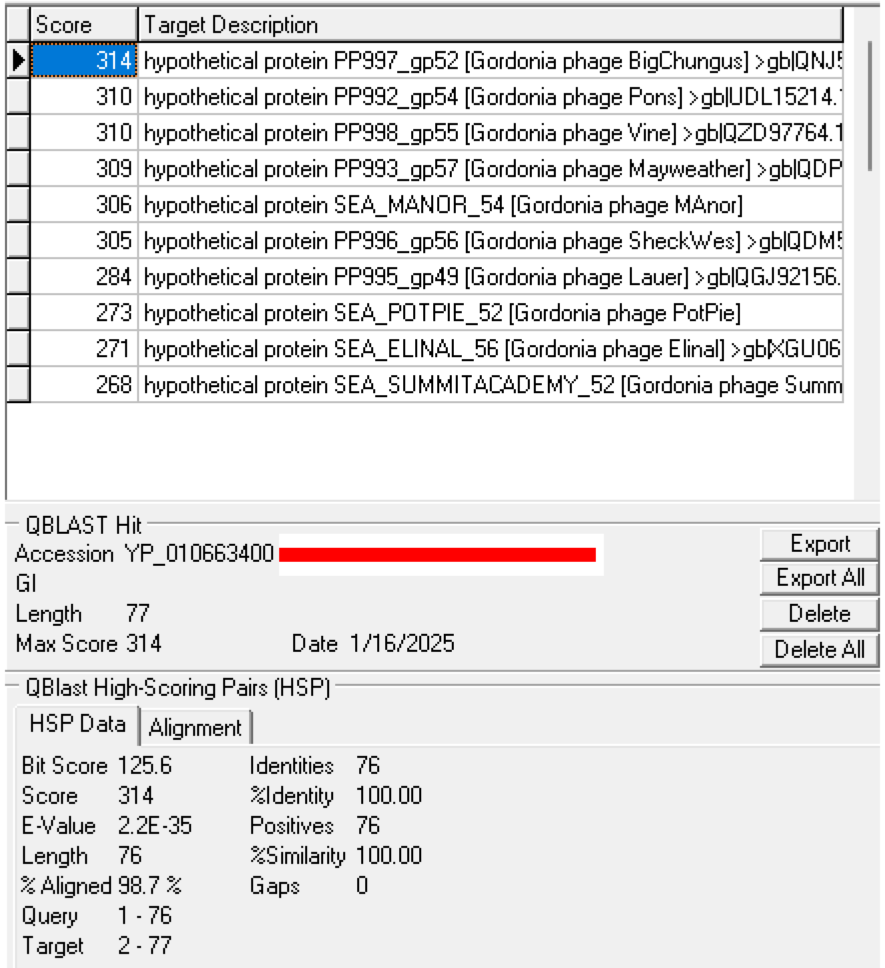
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- The coding potential starts by peaking to strong at 38110 and staying that way until it peters of to weak around 38050 before dropping off at 37980.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There were 10 BLAST hits for this feature with highly similar genes of other phages, and all hits had e-values that were extremely close to zero.
- Nine of these hits were 1:1 alignments



Score	Target Description
314	hypothetical protein PP997_gp52 [Gordonia phage BigChungus] >gb|QNJ5
310	hypothetical protein PP992_gp54 [Gordonia phage Pons] >gb|UDL15214.1
310	hypothetical protein PP998_gp55 [Gordonia phage Vine] >gb|QZD97764.1
309	hypothetical protein PP993_gp57 [Gordonia phage Mayweather] >gb|QDP
306	hypothetical protein SEA_MANOR_54 [Gordonia phage MAnor]
305	hypothetical protein PP996_gp56 [Gordonia phage SheckWes] >gb|QDM5
284	hypothetical protein PP995_gp49 [Gordonia phage Lauer] >gb|QGJ92156.
273	hypothetical protein SEA_POTPIE_52 [Gordonia phage PotPie]
271	hypothetical protein SEA_ELINAL_56 [Gordonia phage Elinal] >gb|XGU06
268	hypothetical protein SEA_SUMMITACADEMY_52 [Gordonia phage Summ

QBLAST Hit

Accession YP_010663400

GI

Length 77

Max Score 314 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

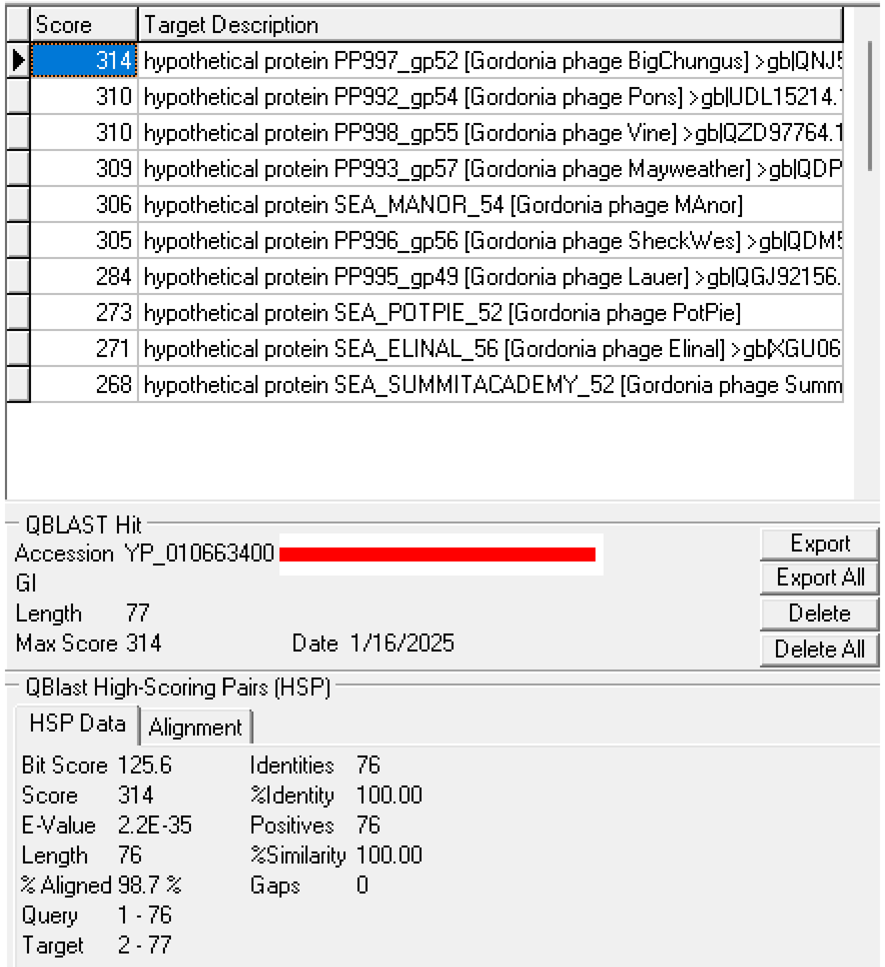
Bit Score	125.6	Identities	76
Score	314	%Identity	100.00
E-Value	2.2E-35	Positives	76
Length	76	%Similarity	100.00
% Aligned	98.7 %	Gaps	0
Query	1 - 76		
Target	2 - 77		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This feature is a gene! There were 10 BLAST hits for this feature with highly similar genes of other phages that had e-values extremely close to zero. Nine of these hits were 1:1 alignments. There is also strong coding potential running throughout where the feature is called to be.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There were 9 1:1 alignments with start at 38153



Score	Target Description
314	hypothetical protein PP997_gp52 [Gordonia phage BigChungus] >gb|QNJ5
310	hypothetical protein PP992_gp54 [Gordonia phage Pons] >gb|UDL15214.1
310	hypothetical protein PP998_gp55 [Gordonia phage Vine] >gb|QZD97764.1
309	hypothetical protein PP993_gp57 [Gordonia phage Mayweather] >gb|QDP
306	hypothetical protein SEA_MANOR_54 [Gordonia phage MAnor]
305	hypothetical protein PP996_gp56 [Gordonia phage SheckWes] >gb|QDM5
284	hypothetical protein PP995_gp49 [Gordonia phage Lauer] >gb|QGJ92156.
273	hypothetical protein SEA_POTPIE_52 [Gordonia phage PotPie]
271	hypothetical protein SEA_ELINAL_56 [Gordonia phage Elinal] >gb|XGU06
268	hypothetical protein SEA_SUMMITACADEMY_52 [Gordonia phage Summ

QBLAST Hit

Accession YP_010663400

GI

Length 77

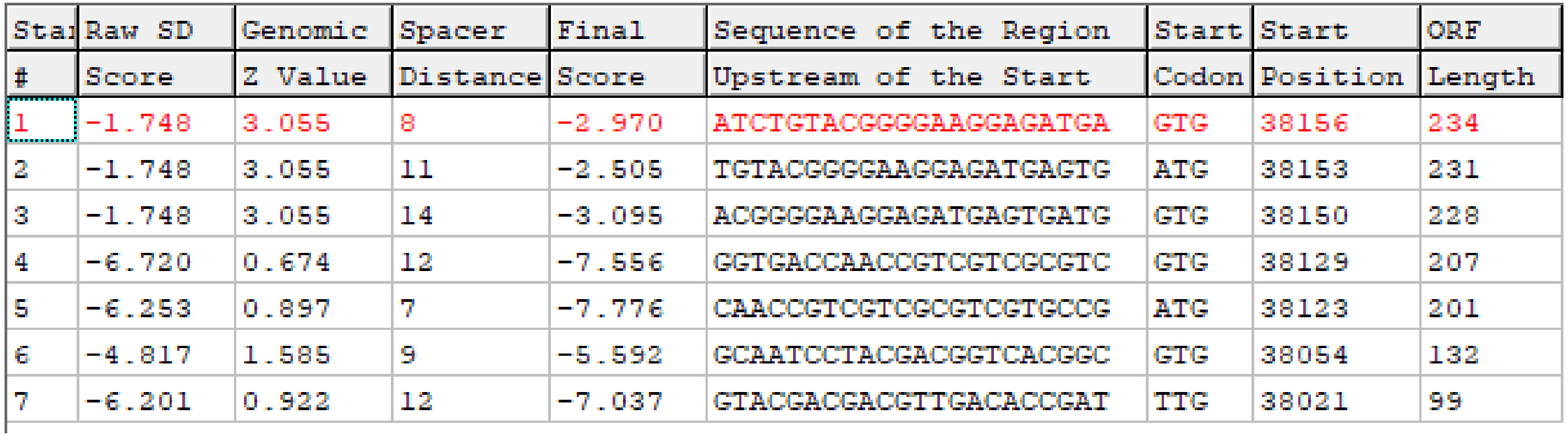
Max Score 314 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	125.6	Identities	76
Score	314	%Identity	100.00
E-Value	2.2E-35	Positives	76
Length	76	%Similarity	100.00
% Aligned	98.7 %	Gaps	0
Query	1 - 76		
Target	2 - 77		

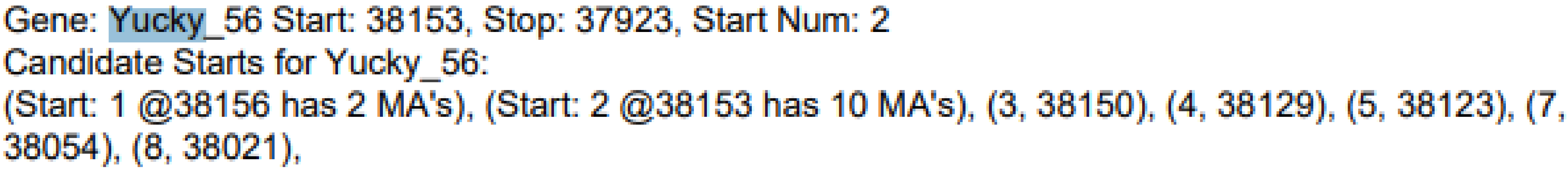
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- 38513 has a good Z value at 3.055 and the best FS at -2.505



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	-1.748	3.055	8	-2.970	ATCTGTACGGGGAAGGAGATGA	GTG	38156	234
2	-1.748	3.055	11	-2.505	TGTACGGGGAAGGAGATGAGTG	ATG	38153	231
3	-1.748	3.055	14	-3.095	ACGGGGAAGGAGATGAGTGATG	GTG	38150	228
4	-6.720	0.674	12	-7.556	GGTGACCAACCGTCGTCGCGTC	GTG	38129	207
5	-6.253	0.897	7	-7.776	CAACCGTCGTCGCGTCGTGCCG	ATG	38123	201
6	-4.817	1.585	9	-5.592	GCAATCCTACGACGGTCACGGC	GTG	38054	132
7	-6.201	0.922	12	-7.037	GTACGACGACGTTGACACCGAT	TTG	38021	99

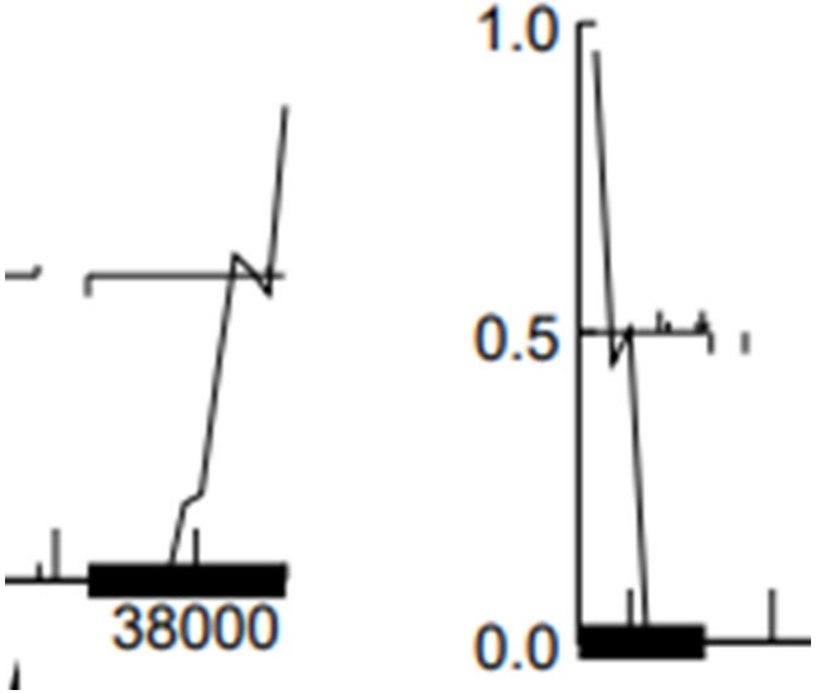
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



Gene: Yucky_56 Start: 38153, Stop: 37923, Start Num: 2
Candidate Starts for Yucky_56:
(Start: 1 @38156 has 2 MA's), (Start: 2 @38153 has 10 MA's), (3, 38150), (4, 38129), (5, 38123), (7, 38054), (8, 38021),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

Coding potential is not cut off



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

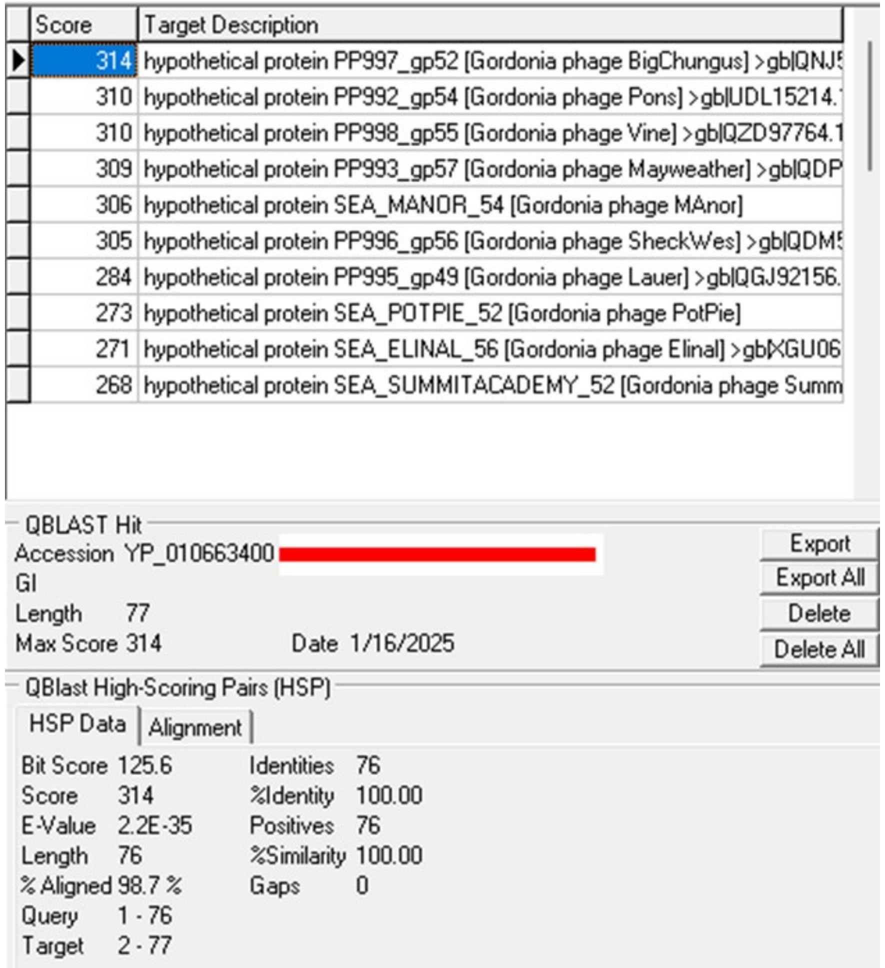
- 38153 has an overlap of 1
- 38516 has a overlap of 4

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- 38513 is the start. It is a tandem start and is the second start in the sequence. It has sufficiently good evidence.

BLAST function evidence. What assigned functions do other highly similar genes have?

- Functions annotated previously are hypothetical protein



Score	Target Description
314	hypothetical protein PP997_gp52 [Gordonia phage BigChungus] >gb|QNJ
310	hypothetical protein PP992_gp54 [Gordonia phage Pons] >gb|UDL15214.
310	hypothetical protein PP998_gp55 [Gordonia phage Vine] >gb|QZD97764.1
309	hypothetical protein PP993_gp57 [Gordonia phage Mayweather] >gb|QDP
306	hypothetical protein SEA_MANOR_54 [Gordonia phage MAnor]
305	hypothetical protein PP996_gp56 [Gordonia phage SheckWes] >gb|QDM
284	hypothetical protein PP995_gp49 [Gordonia phage Lauer] >gb|QGJ92156.
273	hypothetical protein SEA_POTPIE_52 [Gordonia phage PotPie]
271	hypothetical protein SEA_ELINAL_56 [Gordonia phage Elinal] >gb|XGU06
268	hypothetical protein SEA_SUMMITACADEMY_52 [Gordonia phage Summ

QBLAST Hit

Accession YP_010663400

GI

Length 77

Max Score 314

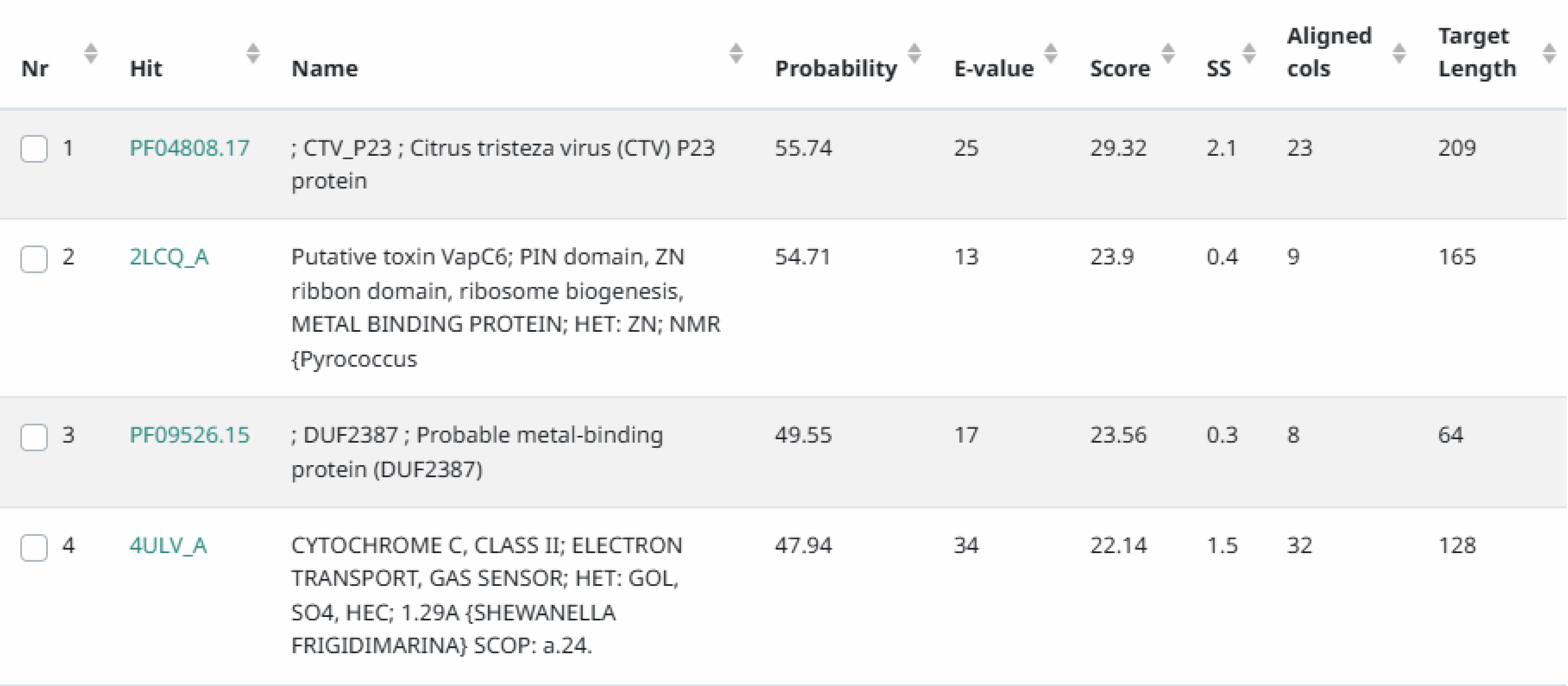
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	125.6	Identities	76
Score	314	%Identity	100.00
E-Value	2.2E-35	Positives	76
Length	76	%Similarity	100.00
% Aligned	98.7 %	Gaps	0
Query	1 - 76		
Target	2 - 77		

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

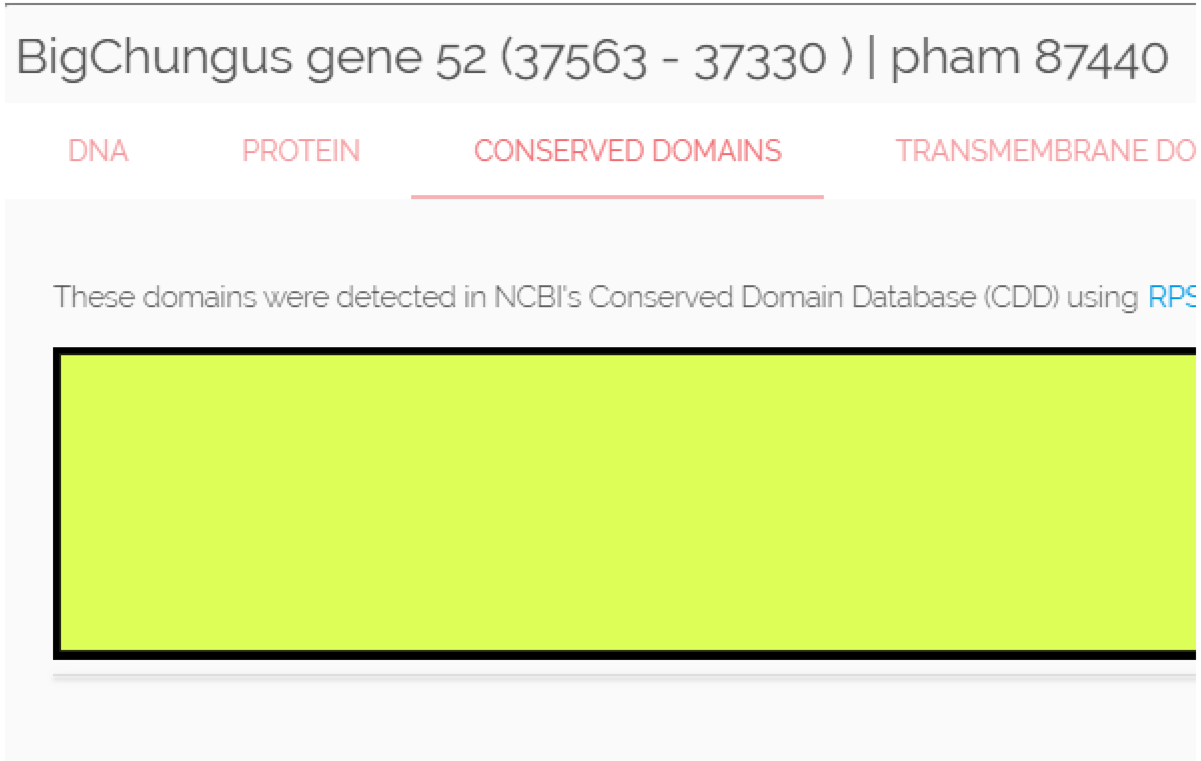
- There are no HHPRED hits above 90% probability



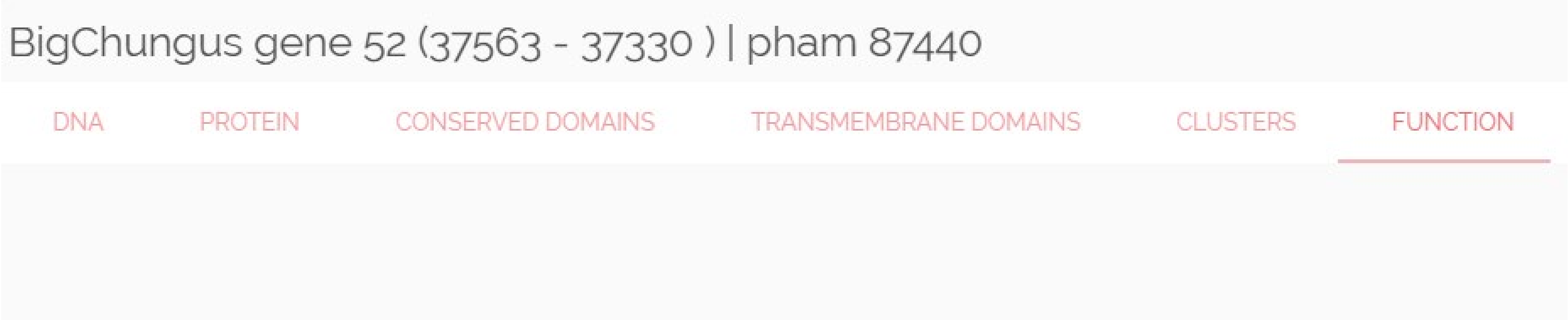
Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	PF04808.17	; CTV_P23 ; Citrus tristeza virus (CTV) P23 protein	55.74	25	29.32	2.1	23	209
□ 2	2LCQ_A	Putative toxin VapC6; PIN domain, ZN ribbon domain, ribosome biogenesis, METAL BINDING PROTEIN; HET: ZN; NMR {Pyrococcus	54.71	13	23.9	0.4	9	165
□ 3	PF09526.15	; DUF2387 ; Probable metal-binding protein (DUF2387)	49.55	17	23.56	0.3	8	64
□ 4	4ULV_A	CYTOCHROME C, CLASS II; ELECTRON TRANSPORT, GAS SENSOR; HET: GOL, SO4, HEC; 1.29A {SHEWANELLA FRIGIDIMARINA} SCOP: a.24.	47.94	34	22.14	1.5	32	128

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Closely related protein in BigChungus does not call a function and there were no conserved domains.

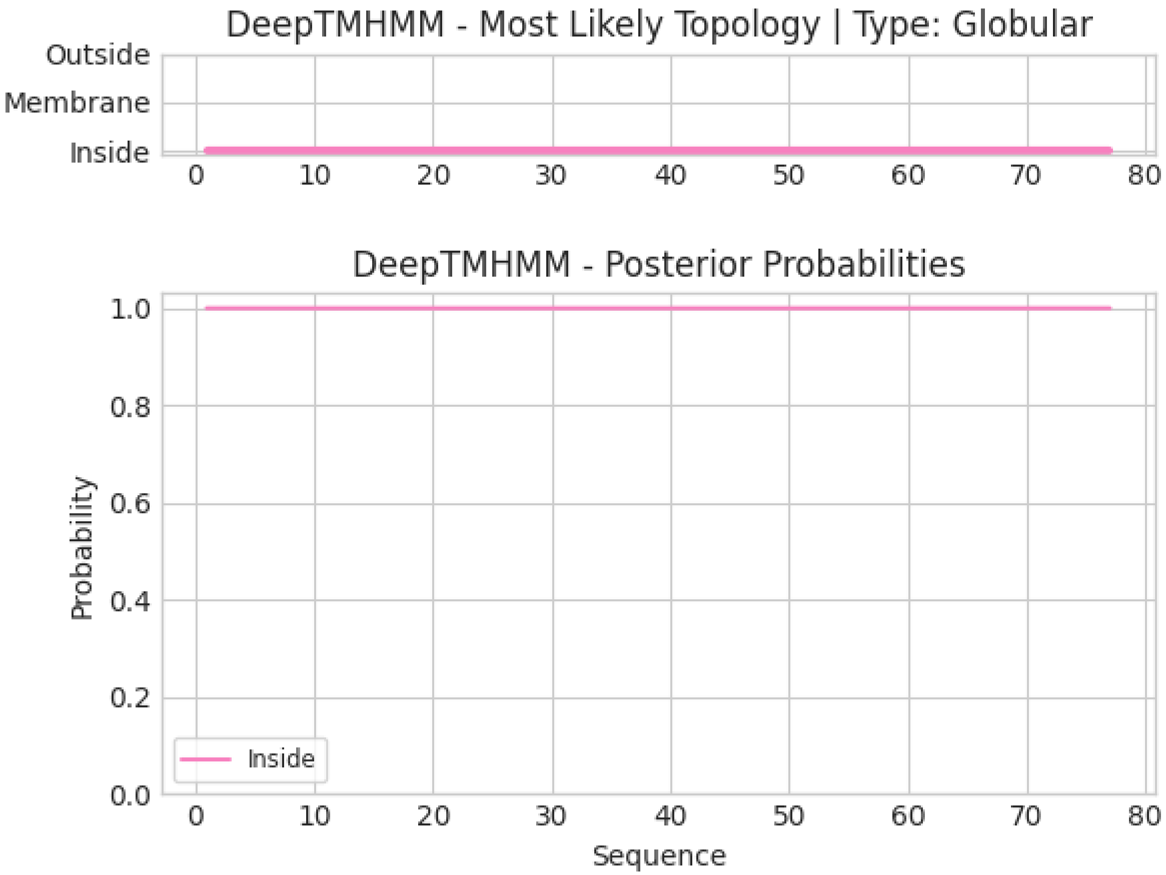

BigChungus gene 52 (37563 - 37330) | pham 87440

These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS


BigChungus gene 52 (37563 - 37330) | pham 87440

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- There are no transmembrane domains



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- This is a hypothetical protein since there is no indication of a known function and there are no transmembrane domains.

Feature 56 – Reverse – Stop 38513

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

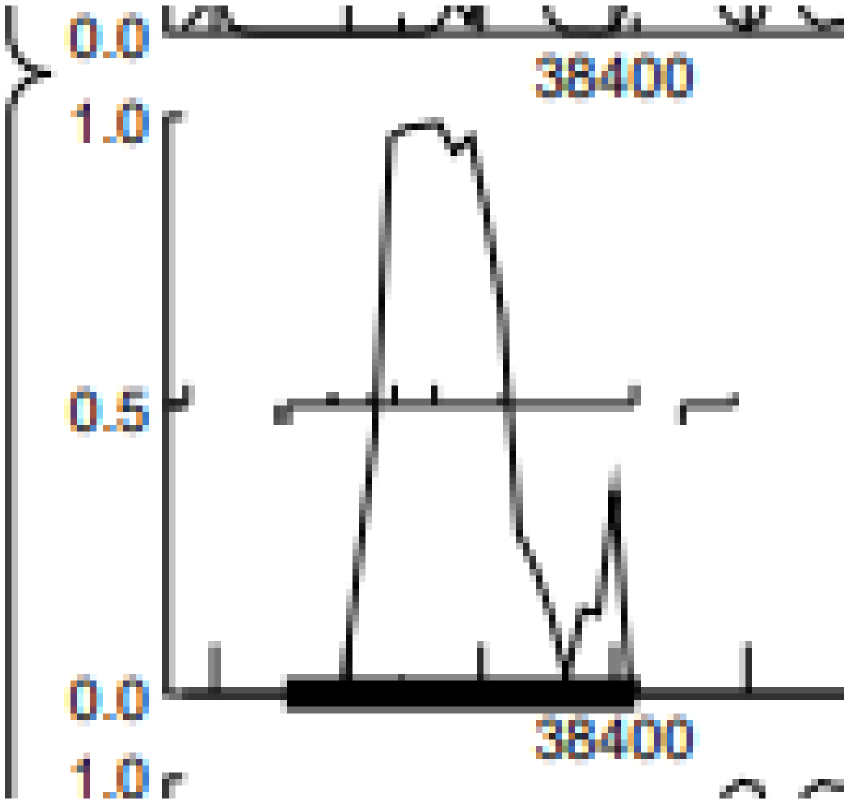
What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

- Feature: 56
- Stop site: 38153
- Called by both Glimmer and GeneMark @bp 38416
- Gap: 1

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

Reverse frame 1 includes all coding potential. It is the only reverse frame with coding potential.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- Highly similar genes:

0 highly similar genes (None have E value: 0E0)

7 1:1 alignments:

BigChungus

Pons

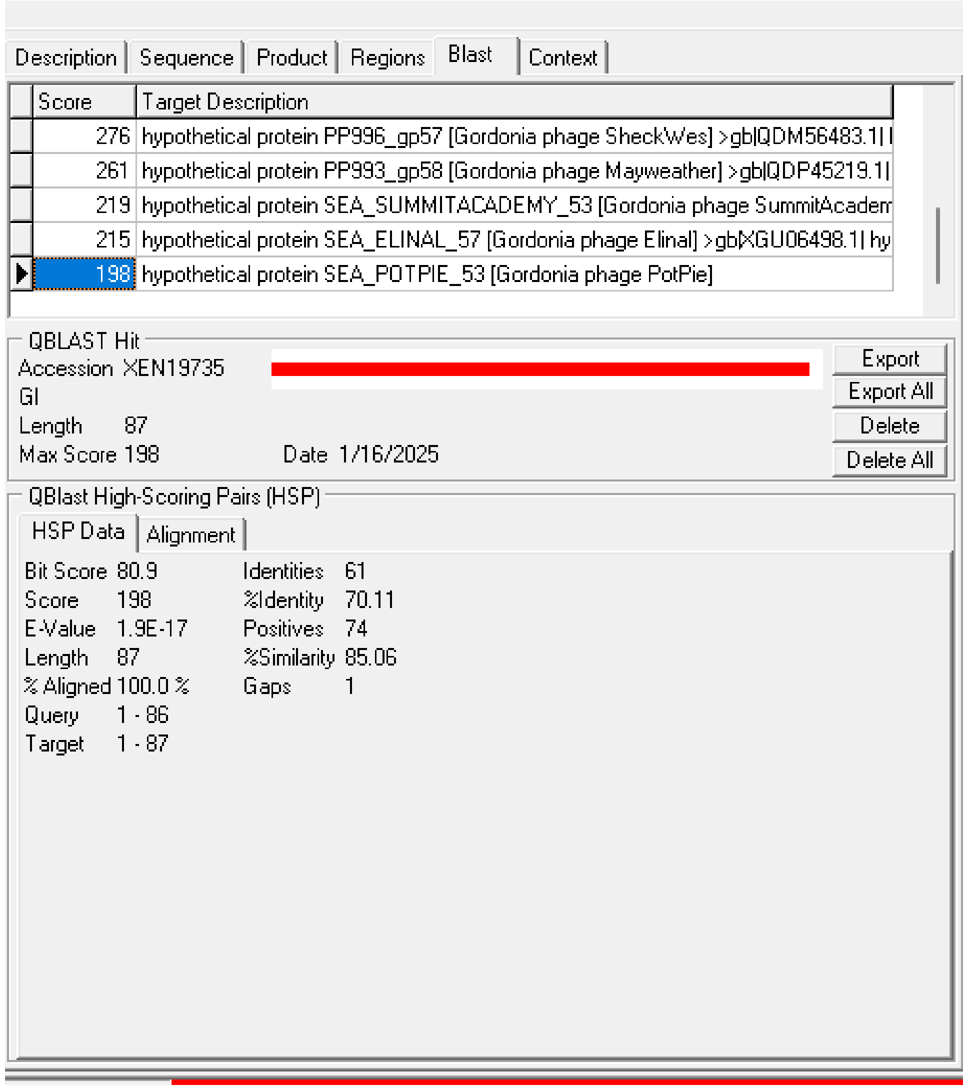
SheckWes

Mayweather

SummitAcademy

Elinal

PotPie



Score	Target Description
276	hypothetical protein PP996_gp57 [Gordonia phage SheckWes] >gb|QDM56483.1| I
261	hypothetical protein PP993_gp58 [Gordonia phage Mayweather] >gb|QDP45219.1|
219	hypothetical protein SEA_SUMMITACADEMY_53 [Gordonia phage SummitAcadem
215	hypothetical protein SEA_ELINAL_57 [Gordonia phage Elinal] >gb|XGU06498.1| hy
198	hypothetical protein SEA_POTPIE_53 [Gordonia phage PotPie]

QBLAST Hit
Accession XEN19735
GI
Length 87
Max Score 198 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	80.9	Identities	61
Score	198	%Identity	70.11
E-Value	1.9E-17	Positives	74
Length	87	%Similarity	85.06
% Aligned	100.0 %	Gaps	1
Query	1 - 86		
Target	1 - 87		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes it is a gene because both Glimmer and GeneMark call it at the same start site 38416. The start site 38416 also includes all coding potential within the reverse frame, and the gene has 1:1 alignment with 7 other genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

7 1:1 alignments:

BigChungus

Pons

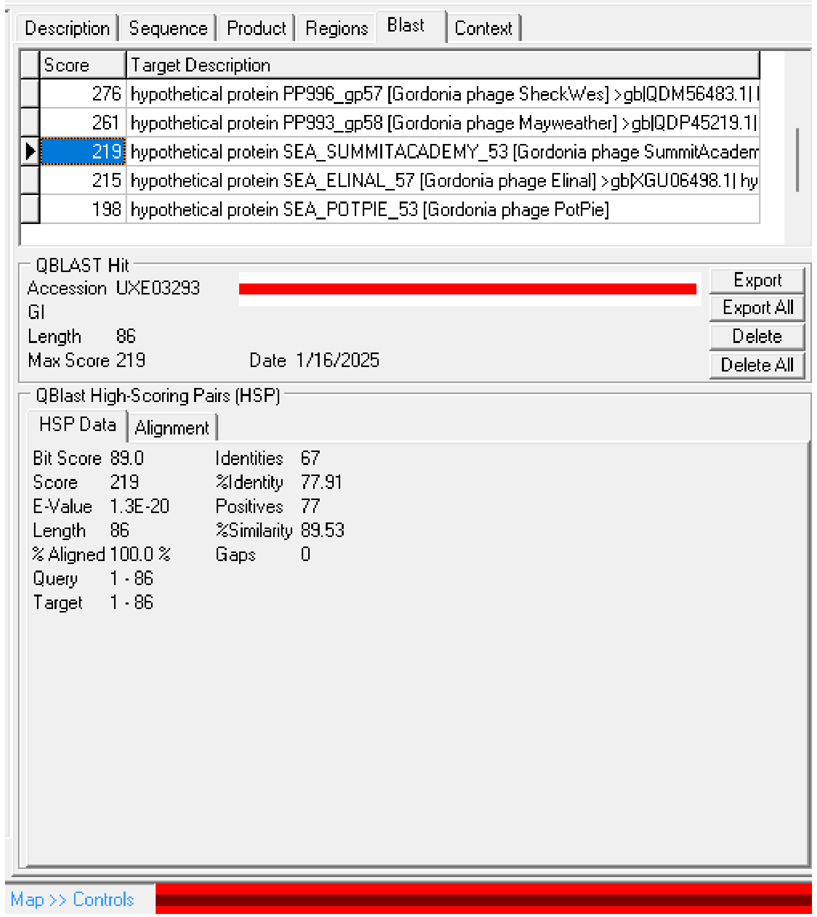
SheckWes

Mayweather

SummitAcademy

Elinal

PotPie



Score	Target Description
276	hypothetical protein PP996_gp57 [Gordonia phage SheckWes] >gb|QDM56483.1| |
261	hypothetical protein PP993_gp58 [Gordonia phage Mayweather] >gb|QDP45219.1|
219	hypothetical protein SEA_SUMMITACADEMY_53 [Gordonia phage SummitAcadem
215	hypothetical protein SEA_ELINAL_57 [Gordonia phage Elinal] >gb|XGU06498.1| hy
198	hypothetical protein SEA_POTPIE_53 [Gordonia phage PotPie]

QBLAST Hit
Accession UXE03293
GI
Length 86
Max Score 219
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

HSP Data | Alignment

Bit Score	89.0	Identities	67
Score	219	%Identity	77.91
E-Value	1.3E-20	Positives	77
Length	86	%Similarity	89.53
% Aligned	100.0 %	Gaps	0
Query	1 - 86		
Target	1 - 86		

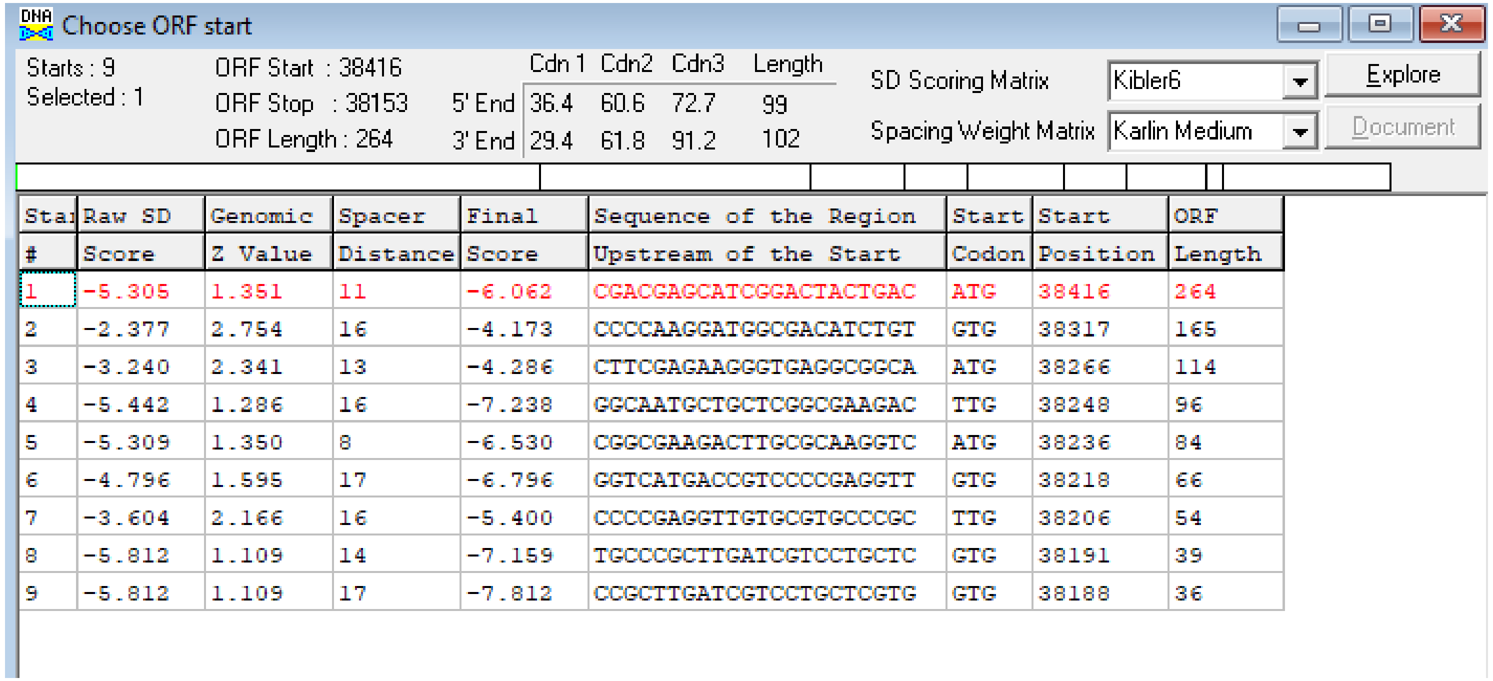
Image shows 1:1 alignment with gene SummitAcademy

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- For start site 38416

Z value = 1.351

Final score = -6.062



Choose ORF start

Starts : 9
Selected : 1

ORF Start : 38416
ORF Stop : 38153
ORF Length : 264

	Cdn 1	Cdn2	Cdn3	Length
5' End	36.4	60.6	72.7	99
3' End	29.4	61.8	91.2	102

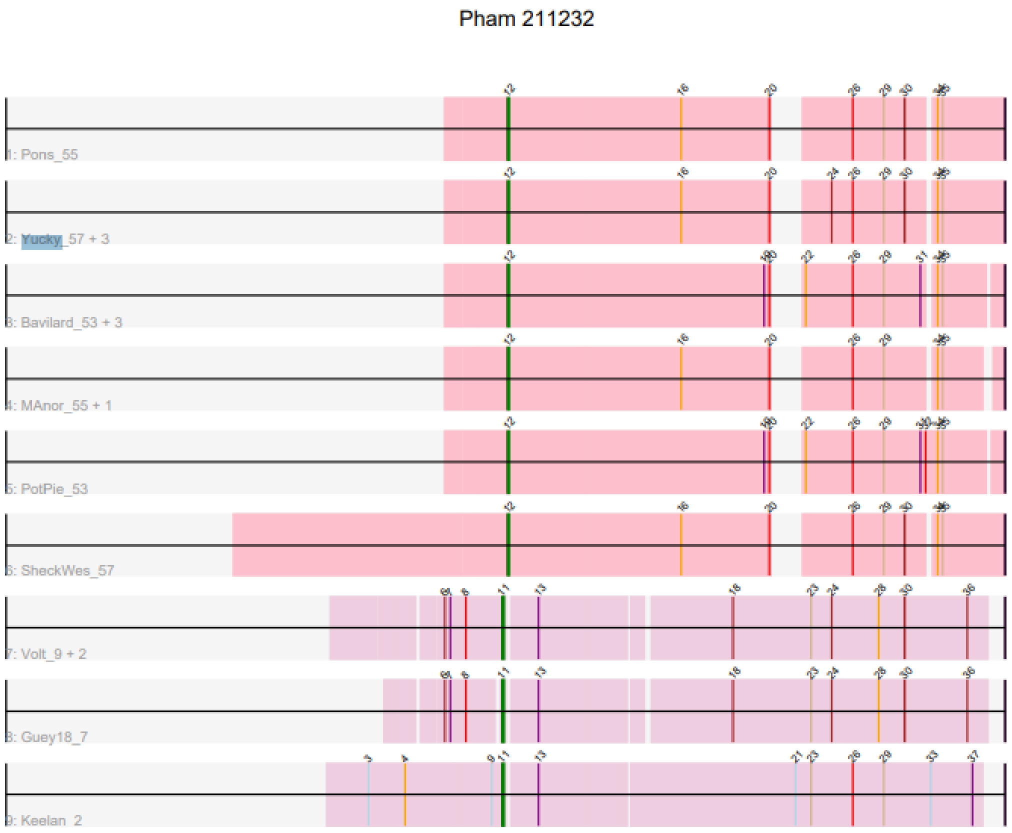
SD Scoring Matrix: Kibler6

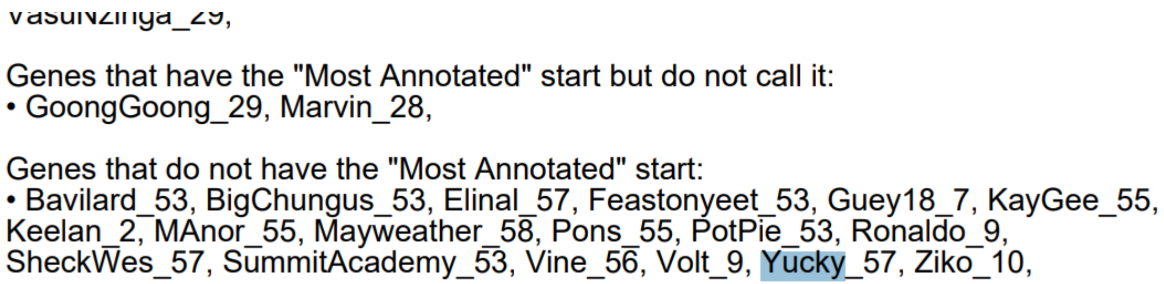
Spacing Weight Matrix: Karlin Medium

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	-5.305	1.351	11	-6.062	CGACGAGCATCGGACTACTGAC	ATG	38416	264
2	-2.377	2.754	16	-4.173	CCCCAAGGATGGCGACATCTGT	GTG	38317	165
3	-3.240	2.341	13	-4.286	CTTCGAGAAGGGTGAGGCGGCA	ATG	38266	114
4	-5.442	1.286	16	-7.238	GGCAATGCTGCTCGGCGAAGAC	TTG	38248	96
5	-5.309	1.350	8	-6.530	CGGCGAAGACTTGCGCAAGGTC	ATG	38236	84
6	-4.796	1.595	17	-6.796	GGTCATGACCGTCCCCGAGGTT	GTG	38218	66
7	-3.604	2.166	16	-5.400	CCCCGAGGTTGTGCGTGCCCGC	TTG	38206	54
8	-5.812	1.109	14	-7.159	TGCCCGCTTGATCGTCCTGCTC	GTG	38191	39
9	-5.812	1.109	17	-7.812	CCGCTTGATCGTCCTGCTCGTG	GTG	38188	36

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Start: 12 @38416 has 11 MA's



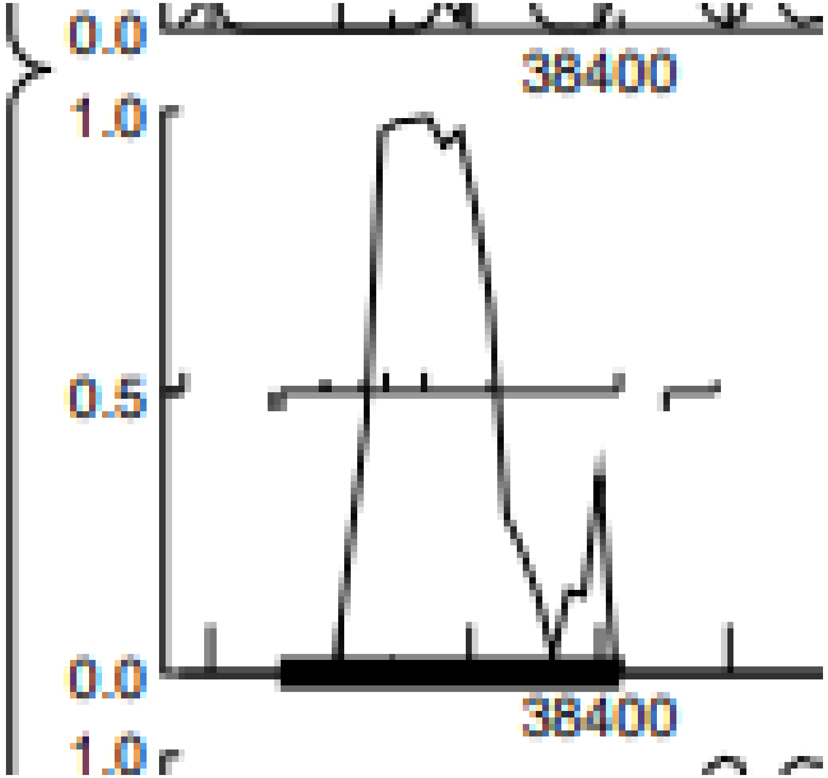

VasuNzinga_29,

Genes that have the "Most Annotated" start but do not call it:
• GoongGoong_29, Marvin_28,

Genes that do not have the "Most Annotated" start:
• Bavilard_53, BigChungus_53, Elinal_57, Feastonyeet_53, Guey18_7, KayGee_55, Keelan_2, MAnor_55, Mayweather_58, Pons_55, PotPie_53, Ronaldo_9, SheckWes_57, SummitAcademy_53, Vine_56, Volt_9, Yucky_57, Ziko_10,

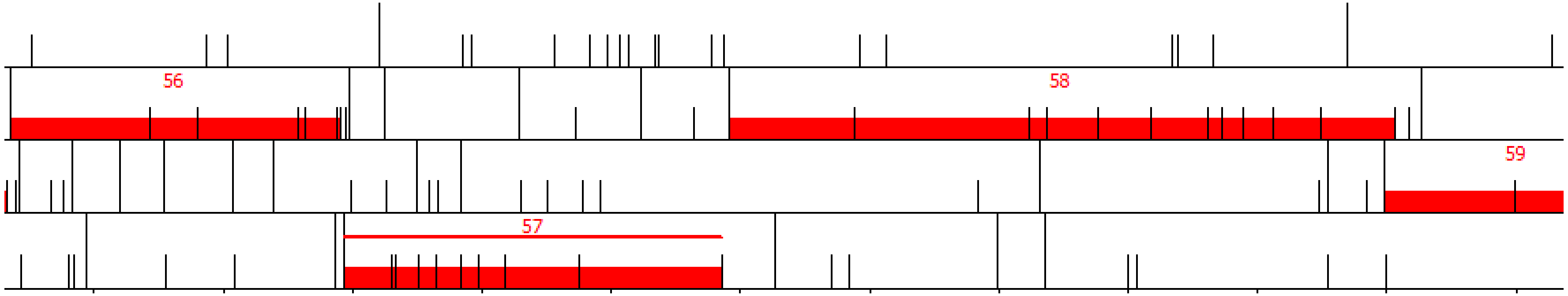
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- At start site 38416, all coding potential is included, none is cut off.
- The start site 38416 is the only start side mentioned in Starterator evidence



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Gap: 1
- 38418 – 38416 = 2 -1 = 1 gap



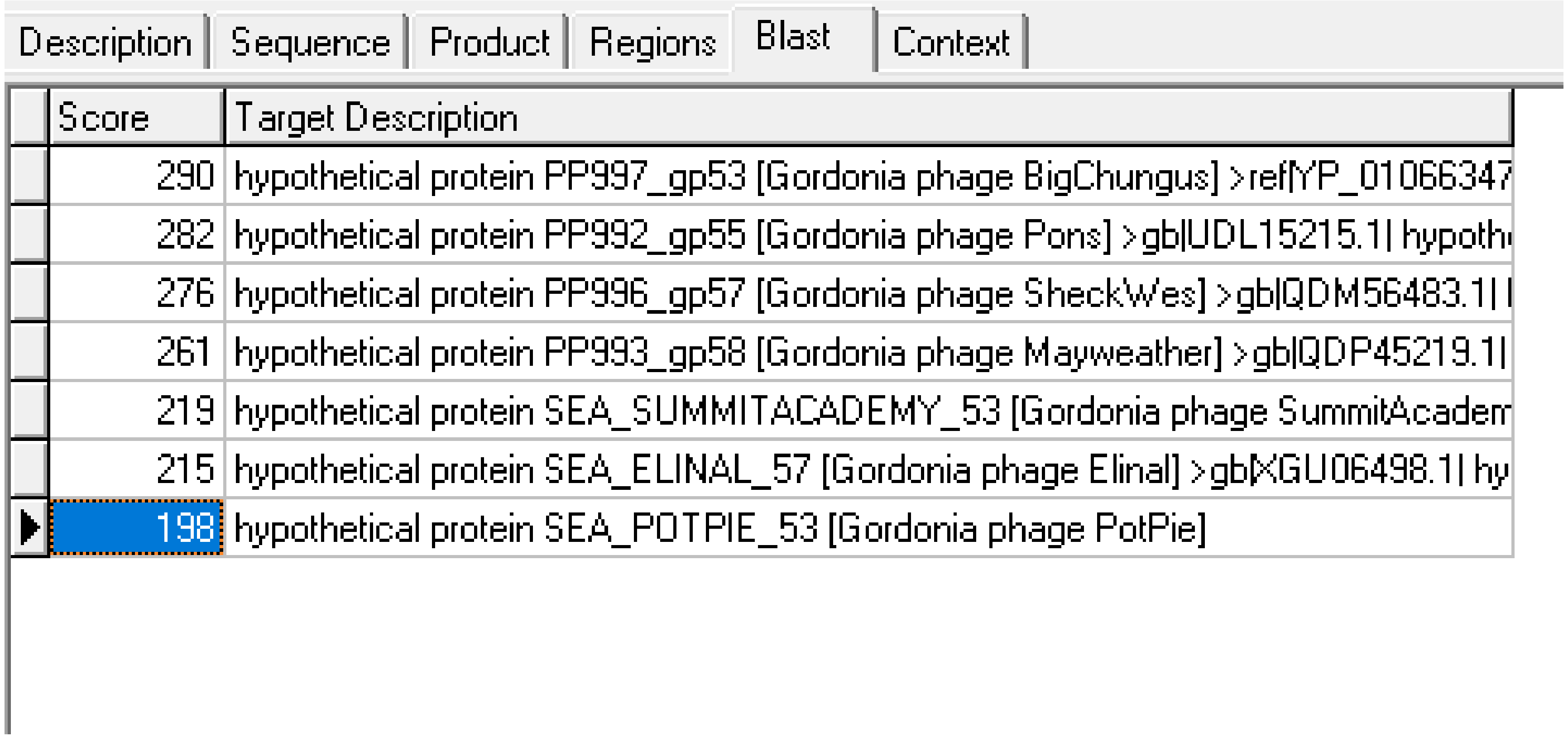
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	38416
GeneMark	Glimmer & GeneMark
Coding potential	Includes all cp
RBS	Z value = 1.351 Final score = -6.062
BLAST	7 1:1 alignments
Starterator	11 MA's
Gap	1

The start site is the auto annotated start site 38416. The reason for this is because the start site was called by both Glimmer and Genemark, the reverse frame contained all coding potential (and none of it was cut off), there are 7 1:1 alignments, and 11 MA's based on Starterator evidence. Starterator evidence also did not suggest another start site.

BLAST function evidence. What assigned functions do other highly similar genes have?

- 7 assigned function as hypothetical protein



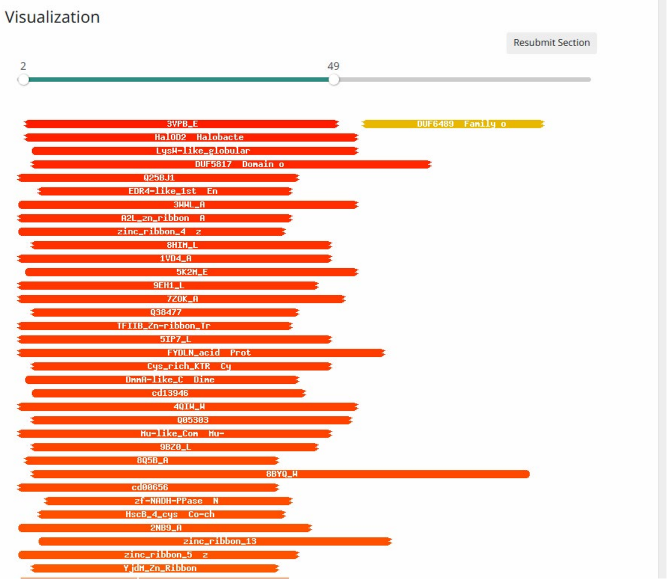
Score	Target Description
290	hypothetical protein PP997_gp53 [Gordonia phage BigChungus] >ref|YP_01066347
282	hypothetical protein PP992_gp55 [Gordonia phage Pons] >gb|UDL15215.1| hypoth
276	hypothetical protein PP996_gp57 [Gordonia phage SheckWes] >gb|QDM56483.1|
261	hypothetical protein PP993_gp58 [Gordonia phage Mayweather] >gb|QDP45219.1|
219	hypothetical protein SEA_SUMMITACADEMY_53 [Gordonia phage SummitAcadem
215	hypothetical protein SEA_ELINAL_57 [Gordonia phage Elinal] >gb|XGU06498.1| hy
198	hypothetical protein SEA_POTPIE_53 [Gordonia phage PotPie]

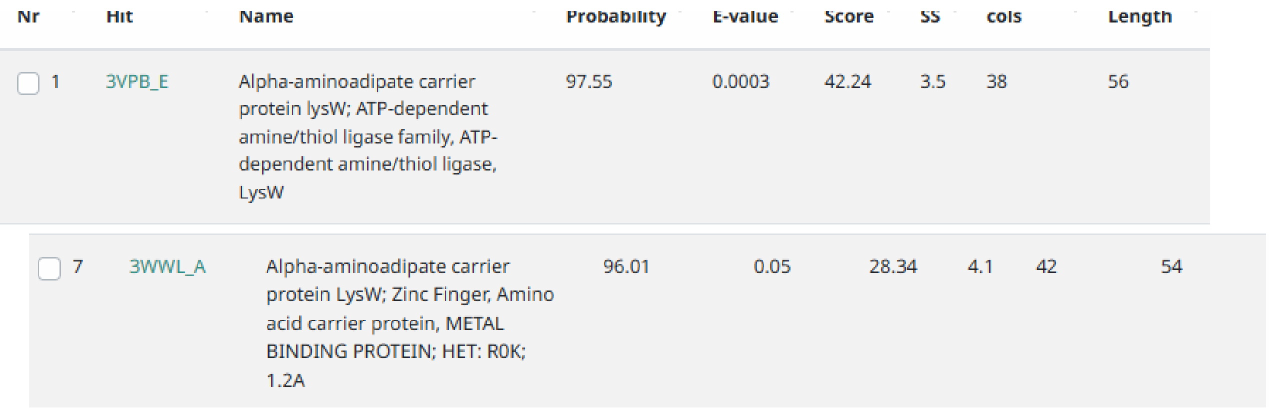
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- Hhpred evidence:

2 hits listed function as Alpha-aminoadipate carrier protein. Other hits were considered “domain of unknown”, or “uncharacterized protein”.

However, Alpha-aminoadipate carrier protein is not on the function list so we cannot call it.





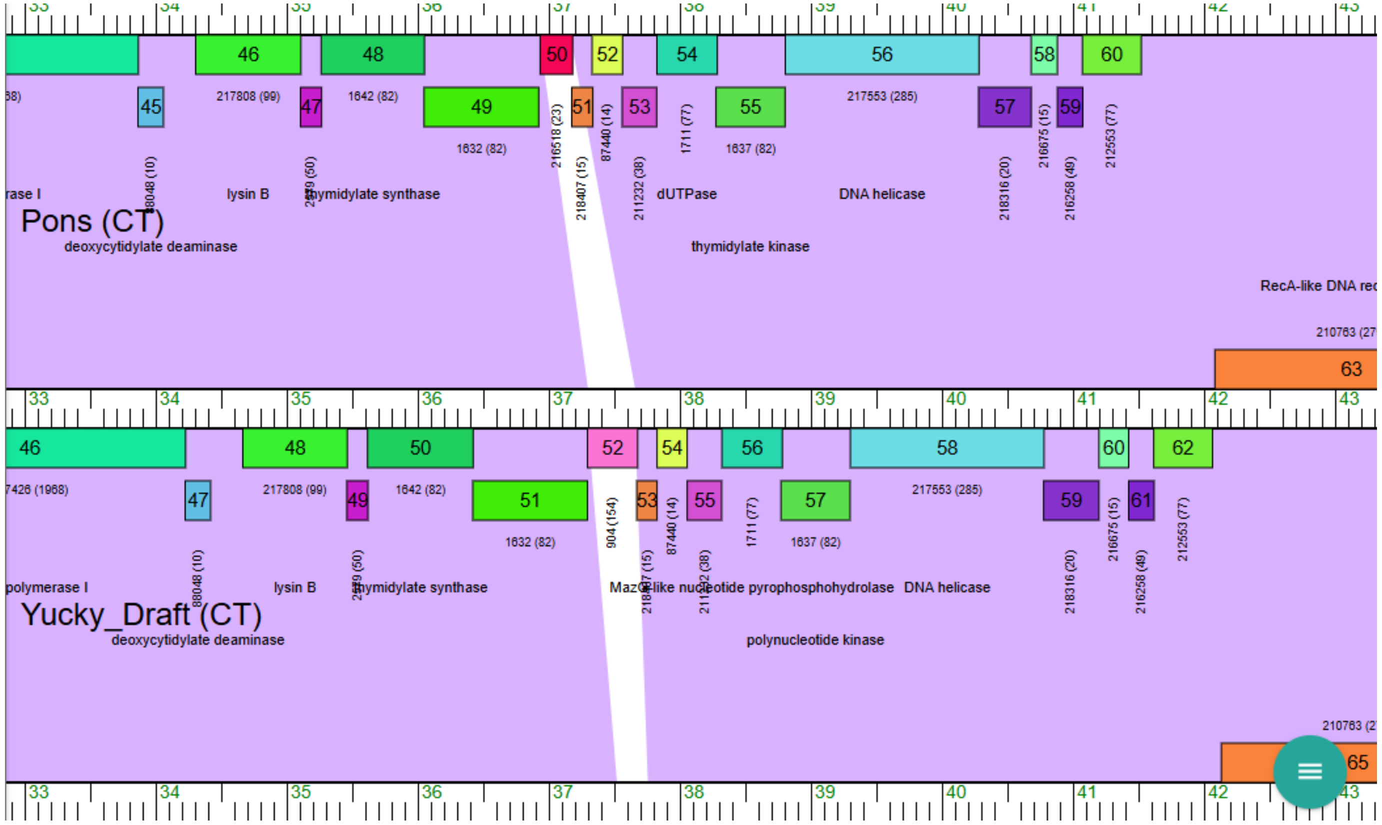
Nr	Hit	Name	Probability	E-value	Score	SS	cols	Length
□ 1	3VPB_E	Alpha-aminoadipate carrier protein lysW; ATP-dependent amine/thiol ligase family, ATP-dependent amine/thiol ligase, LysW	97.55	0.0003	42.24	3.5	38	56
□ 7	3WWL_A	Alpha-aminoadipate carrier protein LysW; Zinc Finger, Amino acid carrier protein, METAL BINDING PROTEIN; HET: R0K; 1.2A	96.01	0.05	28.34	4.1	42	54

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

Yucky feature 57 conserved domain: none function: none

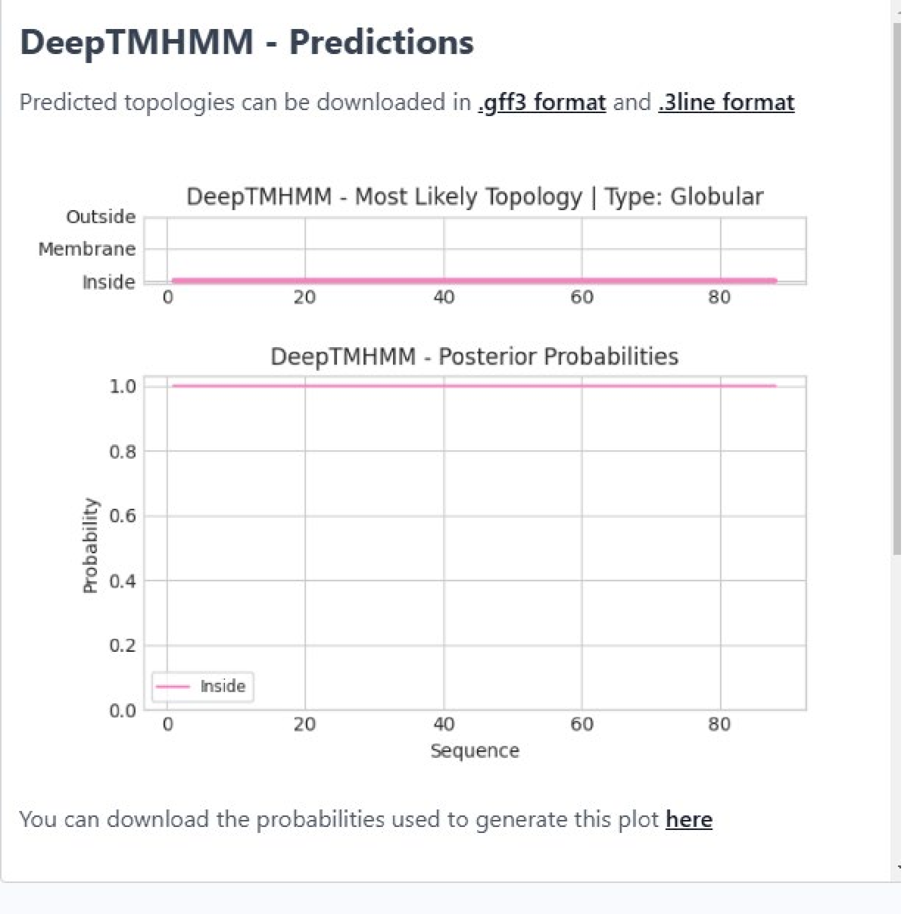
Pons feature 55 conserved domain: none function: none

BigChungus feature 53 conserved domain: none function: none



Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- Has 0 unnamed number of predicted TMRs



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

DeepTMHMM - Most Likely Topology | Type: Globular
Outside
Membrane
Inside
0
20
40
60
80
DeepTMHMM - Posterior Probabilities
1.0
0.8
0.6
0.4
0.2
0.0
Probability
Inside
0
20
40
60
80
Sequence

You can download the probabilities used to generate this plot here

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function is hypothetical protein because it has no conserved domain or function seen in Phamerator evidence. Hhpred also shows no function as the possible function that it could be (Alpha-aminoadipate carrier protein) is not on the function list. The DeepTMHMM evidence also has 0 unnamed number of predicted TMRs, so the function is automatically considered a hypothetical protein.

Feature 57 – Reverse – Stop 38418

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

What is the autoannotated start?

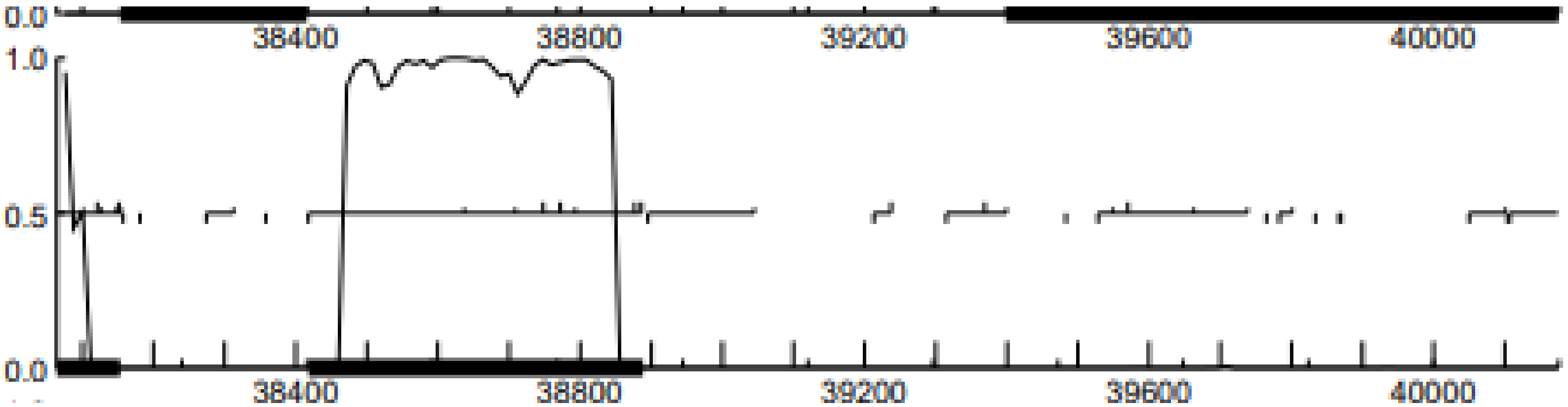
Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

- Feature: 57
- Stop site: 38418
- Called by Glimmer @bp 38879 and called by GeneMark @bp 38888
- Overlap: 11

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Start site: 38879
- Includes all cp
- Start site: 38888

Includes all cp



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 25 highly similar genes (0.0E0)

Lauer

Mayweather

SheckWes

Pons

CherryonLim

Cleo

BillDoor

SteamedHams

Survivors

HippoPololi

Tolls

Gibbous

Azira

AndPeggy

Troje

Fribs8

Emalyn

Cozz

Nina

Maargaret

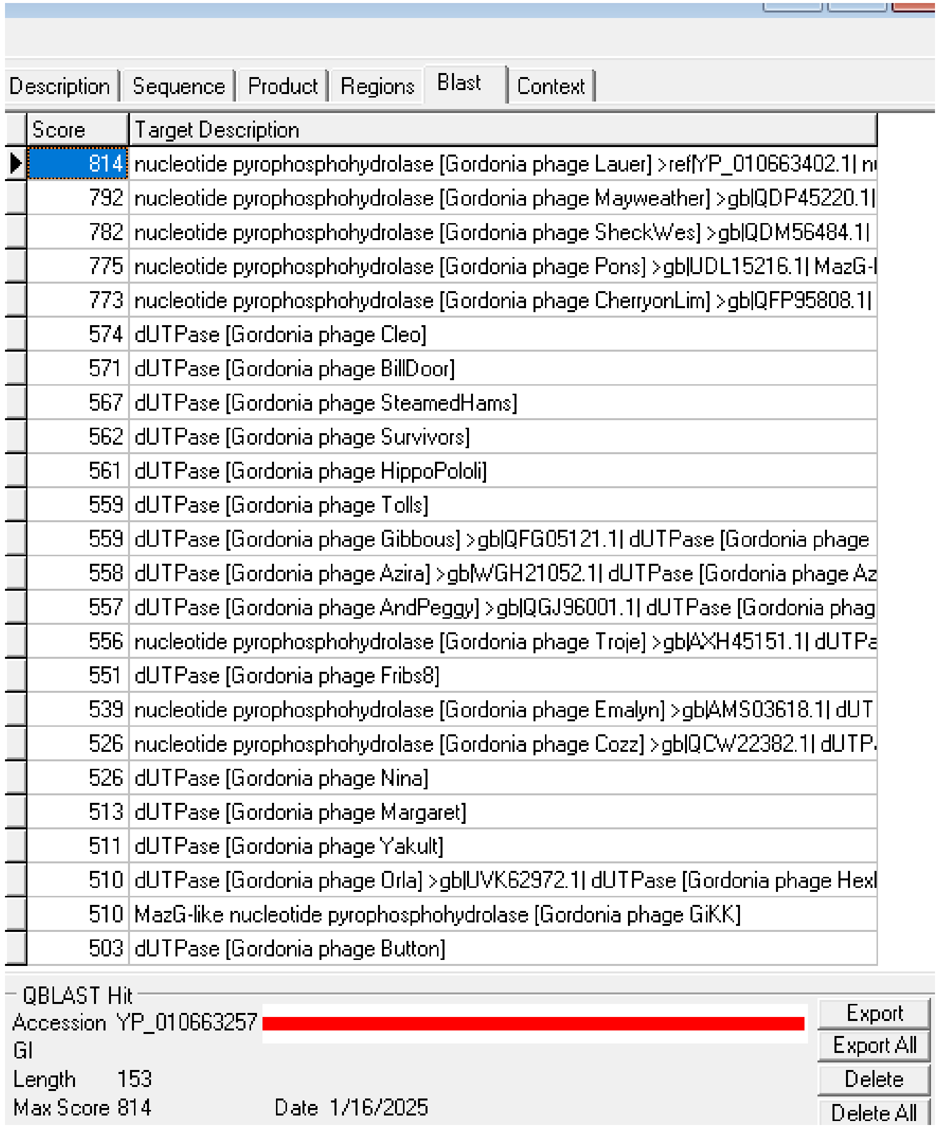
Yakult

Orla

GiKK

Button

Jamzy



Score	Target Description
814	nucleotide pyrophosphohydrolase [Gordonia phage Lauer] >ref|YP_010663402.1| n
792	nucleotide pyrophosphohydrolase [Gordonia phage Mayweather] >gb|QDP45220.1|
782	nucleotide pyrophosphohydrolase [Gordonia phage SheckWes] >gb|QDM56484.1|
775	nucleotide pyrophosphohydrolase [Gordonia phage Pons] >gb|UDL15216.1| MazG-
773	nucleotide pyrophosphohydrolase [Gordonia phage CherryonLim] >gb|QFP95808.1|
574	dUTPase [Gordonia phage Cleo]
571	dUTPase [Gordonia phage BillDoor]
567	dUTPase [Gordonia phage SteamedHams]
562	dUTPase [Gordonia phage Survivors]
561	dUTPase [Gordonia phage HippoPololi]
559	dUTPase [Gordonia phage Tolls]
559	dUTPase [Gordonia phage Gibbous] >gb|QFG05121.1| dUTPase [Gordonia phage
558	dUTPase [Gordonia phage Azira] >gb|WGH21052.1| dUTPase [Gordonia phage Az
557	dUTPase [Gordonia phage AndPeggy] >gb|QGJ96001.1| dUTPase [Gordonia phag
556	nucleotide pyrophosphohydrolase [Gordonia phage Troje] >gb|AXH45151.1| dUTPa
551	dUTPase [Gordonia phage Fribs8]
539	nucleotide pyrophosphohydrolase [Gordonia phage Emalyn] >gb|AMS03618.1| dUT
526	nucleotide pyrophosphohydrolase [Gordonia phage Cozz] >gb|QCW22382.1| dUTP
526	dUTPase [Gordonia phage Nina]
513	dUTPase [Gordonia phage Margaret]
511	dUTPase [Gordonia phage Yakult]
510	dUTPase [Gordonia phage Orla] >gb|UVK62972.1| dUTPase [Gordonia phage Hex
510	MazG-like nucleotide pyrophosphohydrolase [Gordonia phage GiKK]
503	dUTPase [Gordonia phage Button]

QBLAST Hit
Accession YP_010663257
GI
Length 153
Max Score 814 Date 1/16/2025

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes it is a gene because it is called by Glimmer and then GeneMark, both start sites include coding potential, and it has 25 highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence

- 13 1:1 alignments for start site 38879

Lauer

Mayweather

SheckWes

Pons

CherryonLim

BillDoor

SteamedHams

Survivors

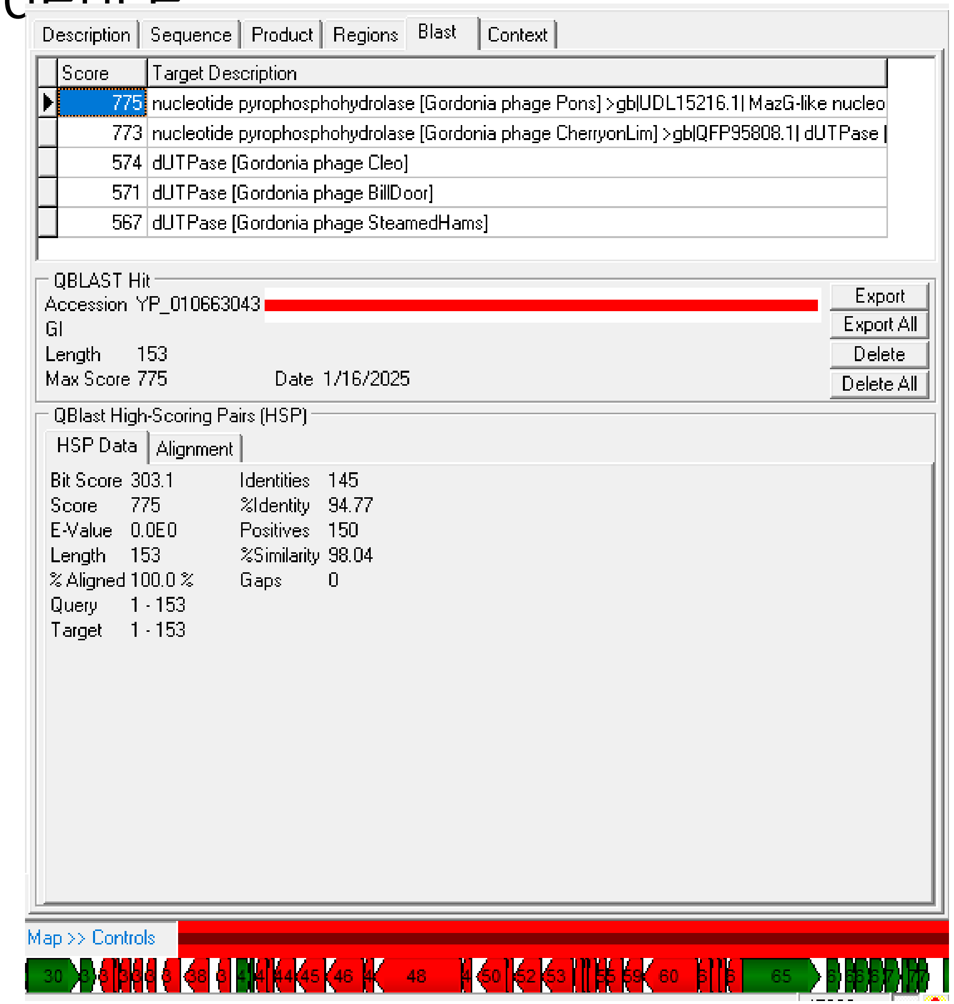
Tolls

Azira

Yarn

Troje

Orla



Score	Target Description
775	nucleotide pyrophosphohydrolase [Gordonia phage Pons] >gb|UDL15216.1| MazG-like nucleo
773	nucleotide pyrophosphohydrolase [Gordonia phage CherryonLim] >gb|QFP95808.1| dUTPase |
574	dUTPase [Gordonia phage Cleo]
571	dUTPase [Gordonia phage BillDoor]
567	dUTPase [Gordonia phage SteamedHams]

QBLAST Hit

Accession YP_010663043
GI
Length 153
Max Score 775
Date 1/16/2025

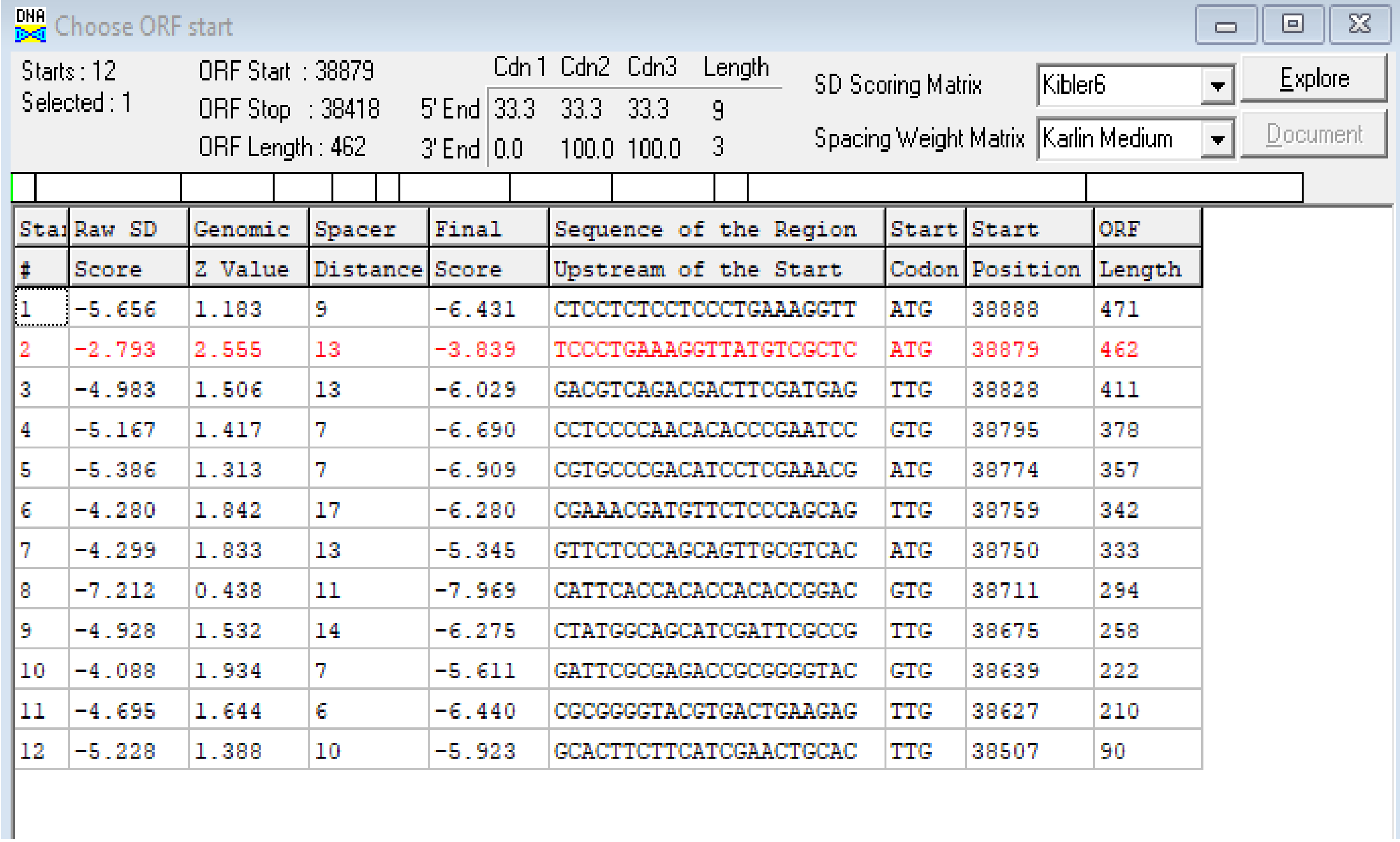
QBlast High-Scoring Pairs (HSP)

Bit Score	303.1	Identities	145
Score	775	%Identity	94.77
E-Value	0.0E0	Positives	150
Length	153	%Similarity	98.04
% Aligned	100.0 %	Gaps	0
Query	1 - 153		
Target	1 - 153		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

Start site 38879

- Z value: 2.555
- Final score: -3.839



Starts : 12
Selected : 1

ORF Start : 38879
ORF Stop : 38418
ORF Length : 462

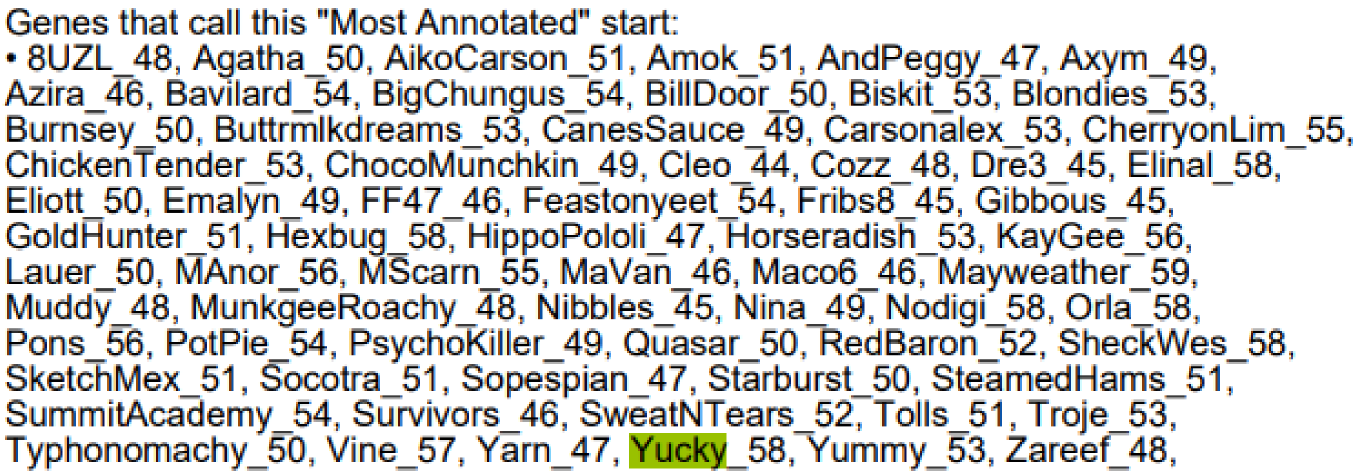
	Cdn 1	Cdn2	Cdn3	Length
5' End	33.3	33.3	33.3	9
3' End	0.0	100.0	100.0	3

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-5.656	1.183	9	-6.431	CTCCTCTCCTCCCTGAAAGGTT	ATG	38888	471
2	-2.793	2.555	13	-3.839	TCCCTGAAAGGTTATGTCGCTC	ATG	38879	462
3	-4.983	1.506	13	-6.029	GACGTCAGACGACTTCGATGAG	TTG	38828	411
4	-5.167	1.417	7	-6.690	CCTCCCCAACACACCCGAATCC	GTG	38795	378
5	-5.386	1.313	7	-6.909	CGTGCCCGACATCCTCGAAACG	ATG	38774	357
6	-4.280	1.842	17	-6.280	CGAAACGATGTTCTCCCAGCAG	TTG	38759	342
7	-4.299	1.833	13	-5.345	GTTCTCCCAGCAGTTGCGTCAC	ATG	38750	333
8	-7.212	0.438	11	-7.969	CATTCACCACACCACACCGGAC	GTG	38711	294
9	-4.928	1.532	14	-6.275	CTATGGCAGCATCGATTCGCCG	TTG	38675	258
10	-4.088	1.934	7	-5.611	GATTCGCGAGACCGCGGGGTAC	GTG	38639	222
11	-4.695	1.644	6	-6.440	CGCGGGGTACGTGACTGAAGAG	TTG	38627	210
12	-5.228	1.388	10	-5.923	GCACTTCTTCATCGAACTGCAC	TTG	38507	90

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Start: 13 @38879 has 50 MA’s



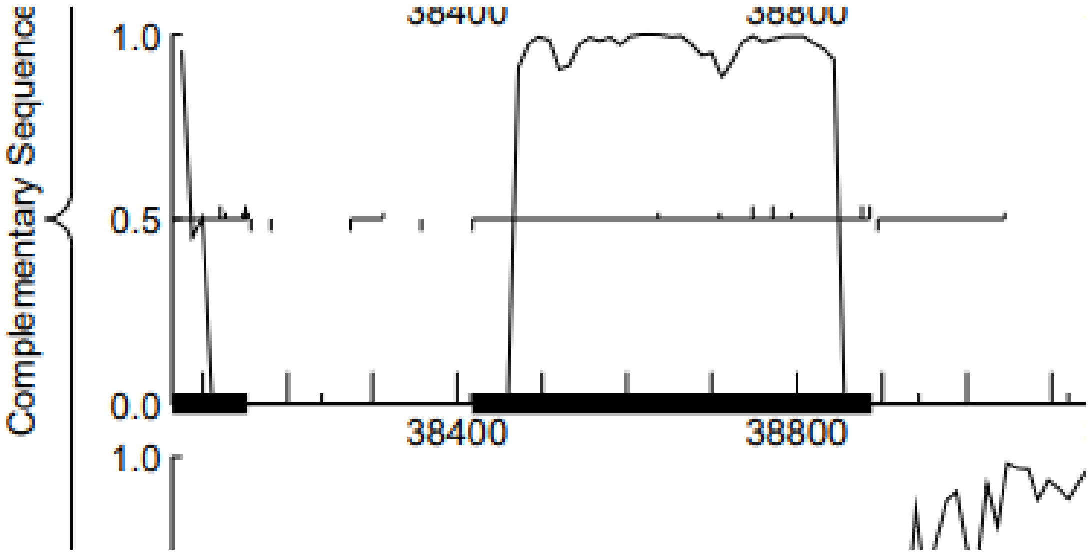
Genes that call this "Most Annotated" start:

- 8UZL_48, Agatha_50, AikoCarson_51, Amok_51, AndPeggy_47, Axym_49, Azira_46, Bavilard_54, BigChungus_54, BillDoor_50, Biskit_53, Blondies_53, Burnsey_50, Buttrmlkdreams_53, CanesSauce_49, Carsonalex_53, CherryonLim_55, ChickenTender_53, ChocoMunchkin_49, Cleo_44, Cozz_48, Dre3_45, Elinal_58, Eliott_50, Emalyn_49, FF47_46, Feastonyeet_54, Fribs8_45, Gibbous_45, GoldHunter_51, Hexbug_58, HippoPololi_47, Horseradish_53, KayGee_56, Lauer_50, MAnor_56, MScarn_55, MaVan_46, Maco6_46, Mayweather_59, Muddy_48, MunkgeeRoachy_48, Nibbles_45, Nina_49, Nodigi_58, Orla_58, Pons_56, PotPie_54, PsychoKiller_49, Quasar_50, RedBaron_52, SheckWes_58, SketchMex_51, Socotra_51, Sopespian_47, Starburst_50, SteamedHams_51, SummitAcademy_54, Survivors_46, SweatNTears_52, Tolls_51, Troje_53, Typhonomachy_50, Vine_57, Yarn_47, Yucky_58, Yummy_53, Zareef_48,

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Start site: 38879

Includes all coding potential. None of the coding potential is cut off.

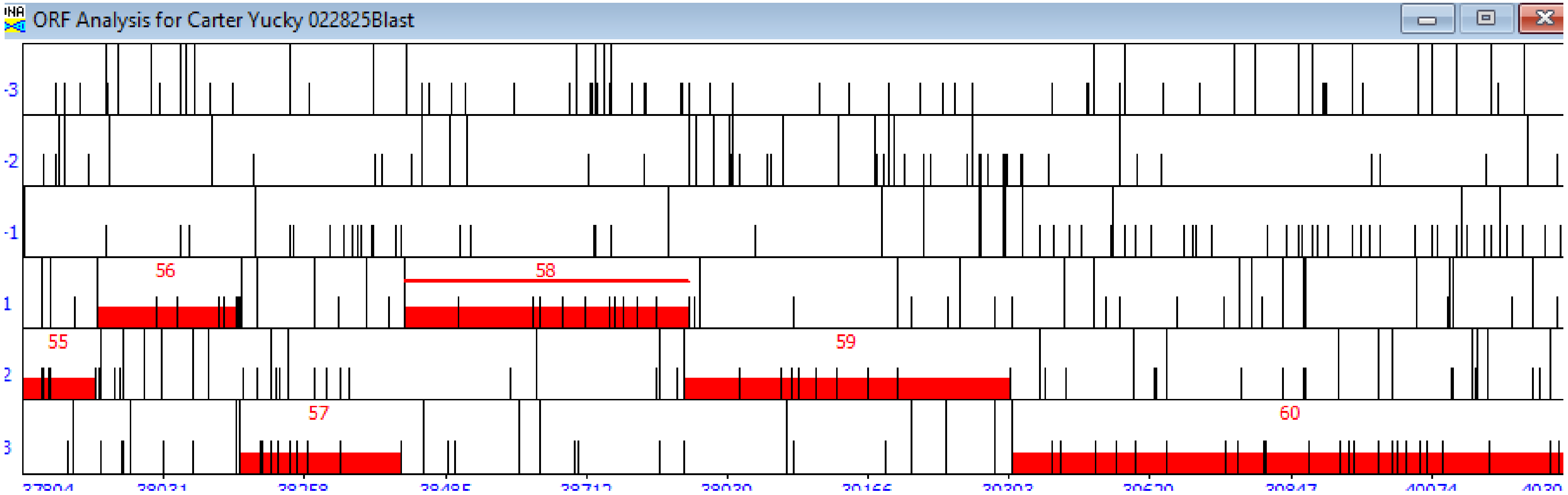


Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Start 38879:

Overlap 11

38879-38869=10 + 1= overlap 11



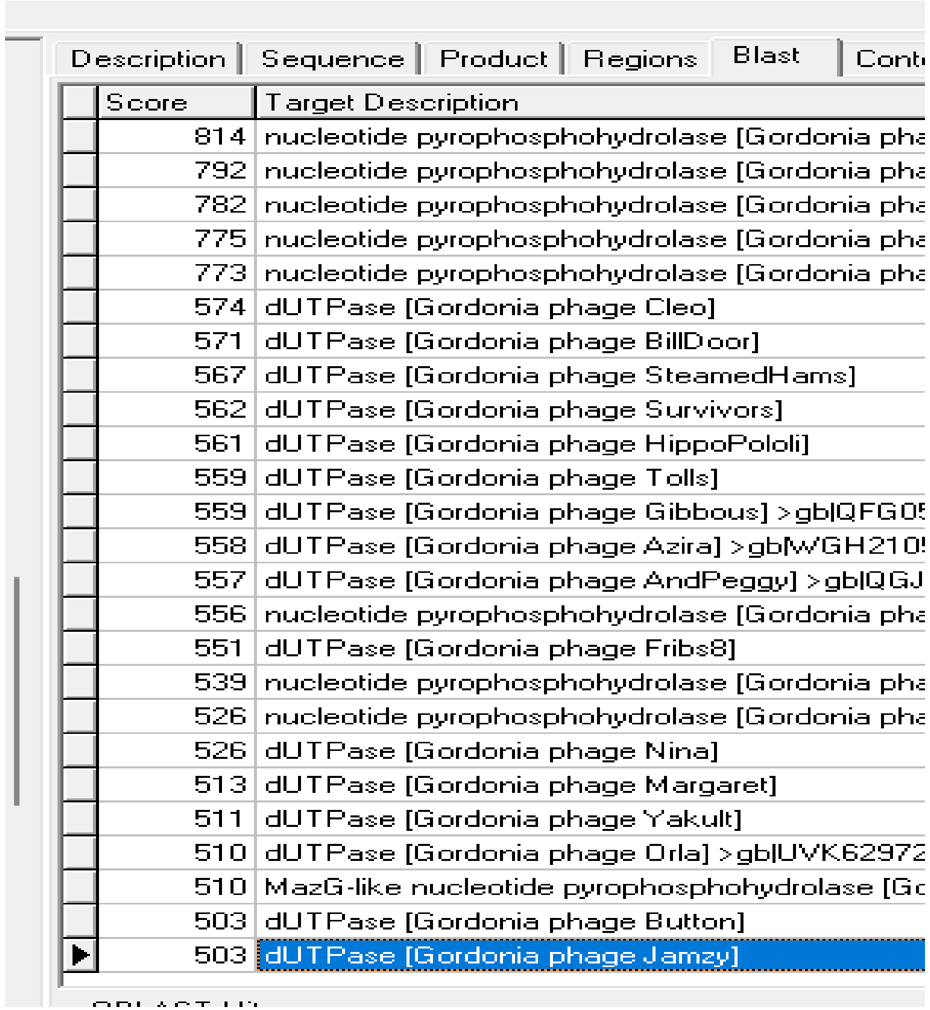
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	38879
GeneMark	Glimmer
Coding potential	All cp
RBS	Z value: 2.555 Final score: -3.839
BLAST	13 1:1 alignments
Starterator	50 MA’s
Overlap	11

The start site is 38879 because it was the only start site called by Starterator evidence. It also has strong coding potential, a z score greater than 2 (the only one on the list), 13 1:1 alignments, and 50 manual annotations.

BLAST function evidence. What assigned functions do other highly similar genes have?

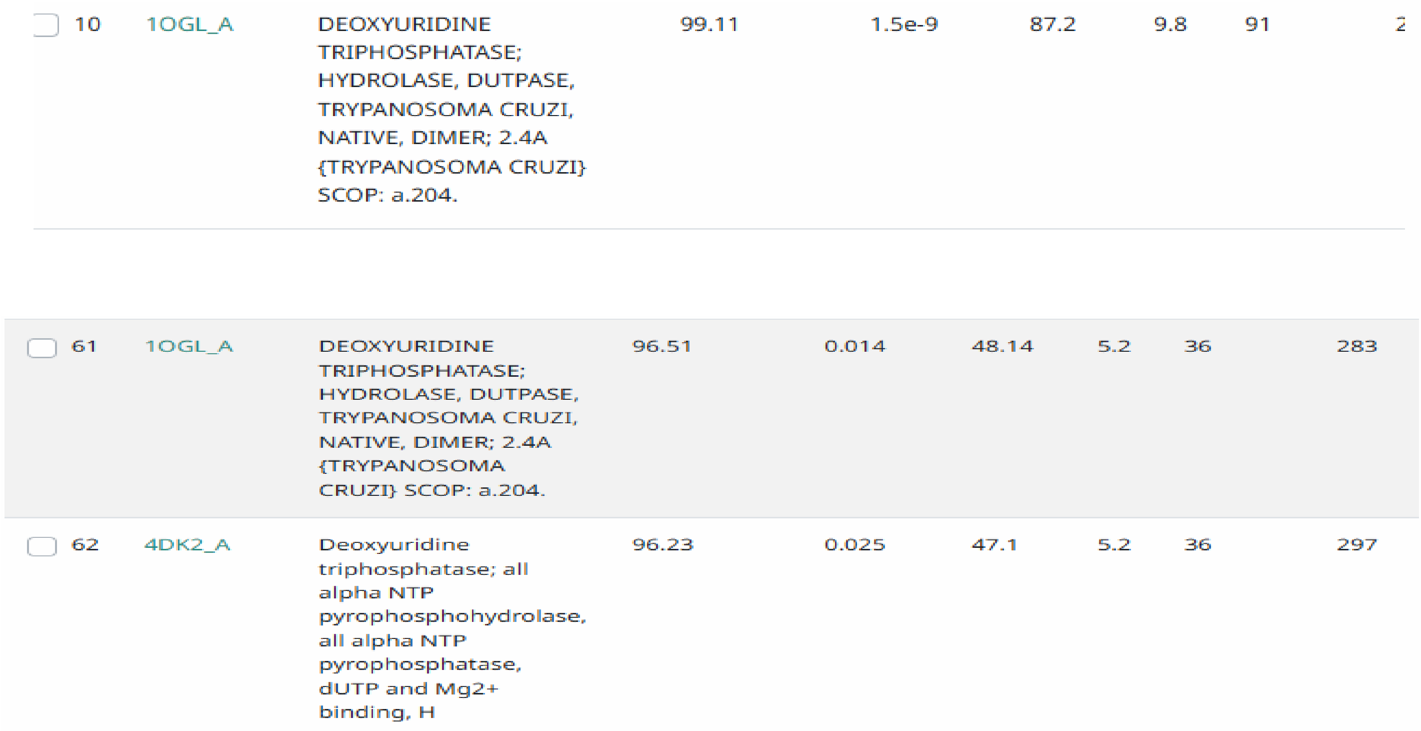
- 8 nucleotide pyrophosphohydrolase
- 16 dUTPase
- 1 MazG-like nucleotide pyrophisphohydrolase



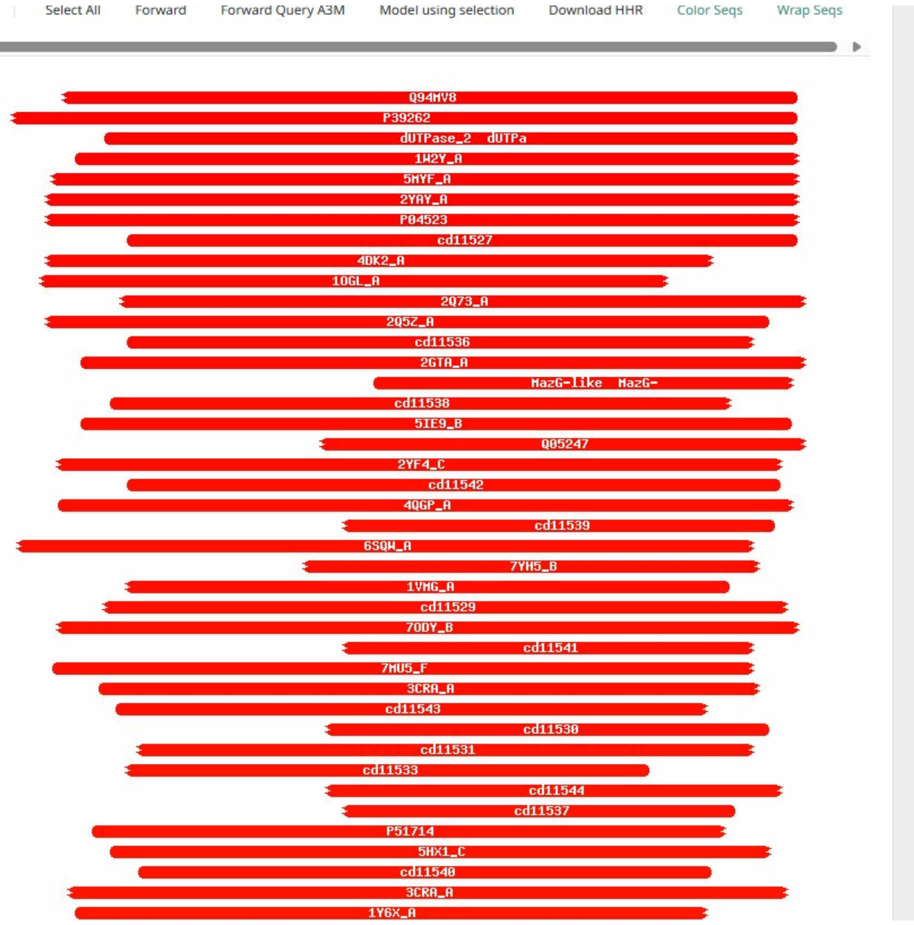
Score	Target Description
814	nucleotide pyrophosphohydrolase [Gordonia pha
792	nucleotide pyrophosphohydrolase [Gordonia pha
782	nucleotide pyrophosphohydrolase [Gordonia pha
775	nucleotide pyrophosphohydrolase [Gordonia pha
773	nucleotide pyrophosphohydrolase [Gordonia pha
574	dUTPase [Gordonia phage Cleo]
571	dUTPase [Gordonia phage BillDoor]
567	dUTPase [Gordonia phage SteamedHams]
562	dUTPase [Gordonia phage Survivors]
561	dUTPase [Gordonia phage HippoPololi]
559	dUTPase [Gordonia phage Tolls]
559	dUTPase [Gordonia phage Gibbous] >gb|QFG05
558	dUTPase [Gordonia phage Azira] >gb|WGH210
557	dUTPase [Gordonia phage AndPeggy] >gb|QGJ
556	nucleotide pyrophosphohydrolase [Gordonia pha
551	dUTPase [Gordonia phage Fribs8]
539	nucleotide pyrophosphohydrolase [Gordonia pha
526	nucleotide pyrophosphohydrolase [Gordonia pha
526	dUTPase [Gordonia phage Nina]
513	dUTPase [Gordonia phage Margaret]
511	dUTPase [Gordonia phage Yakult]
510	dUTPase [Gordonia phage Orla] >gb|UVK62972
510	MazG-like nucleotide pyrophosphohydrolase [Gc
503	dUTPase [Gordonia phage Button]
503	dUTPase [Gordonia phage Jamzy]

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- Hhpred evidence supports the function dUTPase
- On function list, dUTPase has to be deoxyuridine triphosphatase

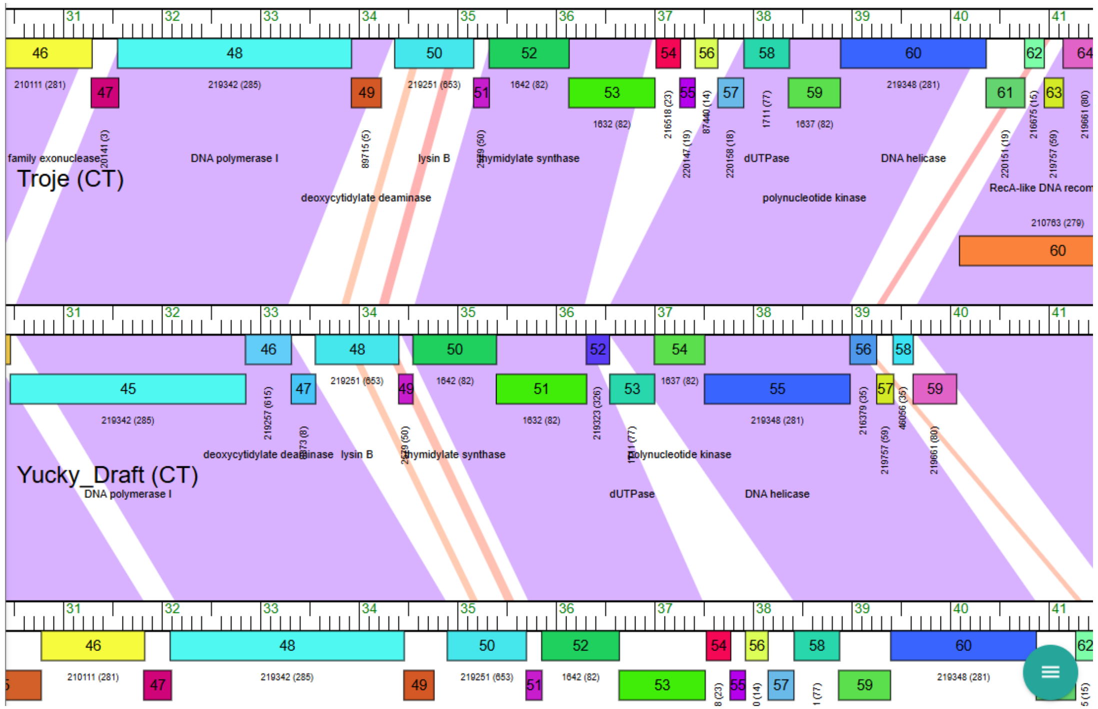


□	10	1OGL_A	DEOXYURIDINE TRIPHOSPHATASE; HYDROLASE, DUTPASE, TRYPANOSOMA CRUZI, NATIVE, DIMER; 2.4A {TRYPANOSOMA CRUZI} SCOP: a.204.	99.11	1.5e-9	87.2	9.8	91	2
□	61	1OGL_A	DEOXYURIDINE TRIPHOSPHATASE; HYDROLASE, DUTPASE, TRYPANOSOMA CRUZI, NATIVE, DIMER; 2.4A {TRYPANOSOMA CRUZI} SCOP: a.204.	96.51	0.014	48.14	5.2	36	283
□	62	4DK2_A	Deoxyuridine triphosphatase; all alpha NTP pyrophosphohydrolase, all alpha NTP pyrophosphatase, dUTP and Mg2+ binding, H	96.23	0.025	47.1	5.2	36	297



Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Yucky 58 conserved domain: 56 and NTP-PPase_dUTPase function: none
- Troje 53 conserved domain: 56 and NTP-PPase_dUTPase function: dUTPase
- SheckWes 58 conserved domain: 56 and NTP-PPase_dUTPase function: dUTPase



Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- None

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function is dUTPase, because there are 16 dUTPase functions for BLAST evidence, multiple hits of dUTPase with the requirement of deoxyuridine triphosphatase in Hhpred, and highly similar genes (Troje and SheckWes) have the function, dUTPase in Phamerator.

Feature 58 – Reverse – Stop 38869

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

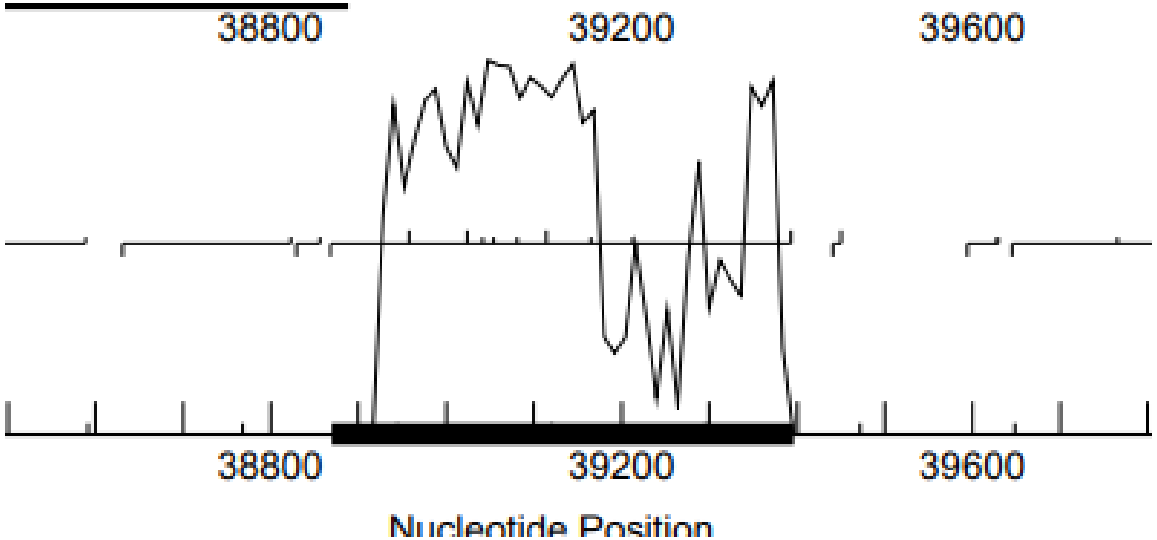
What is the autoannotated start?

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

- Feature 58
- Stop site: 38869
- Called by both Glimmer and Genemark at start site 39396
- Gap: 1

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

Reverse frame 3 includes all coding potential



It is the only reverse frame with coding potential

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 6 highly similar genes:

Vine

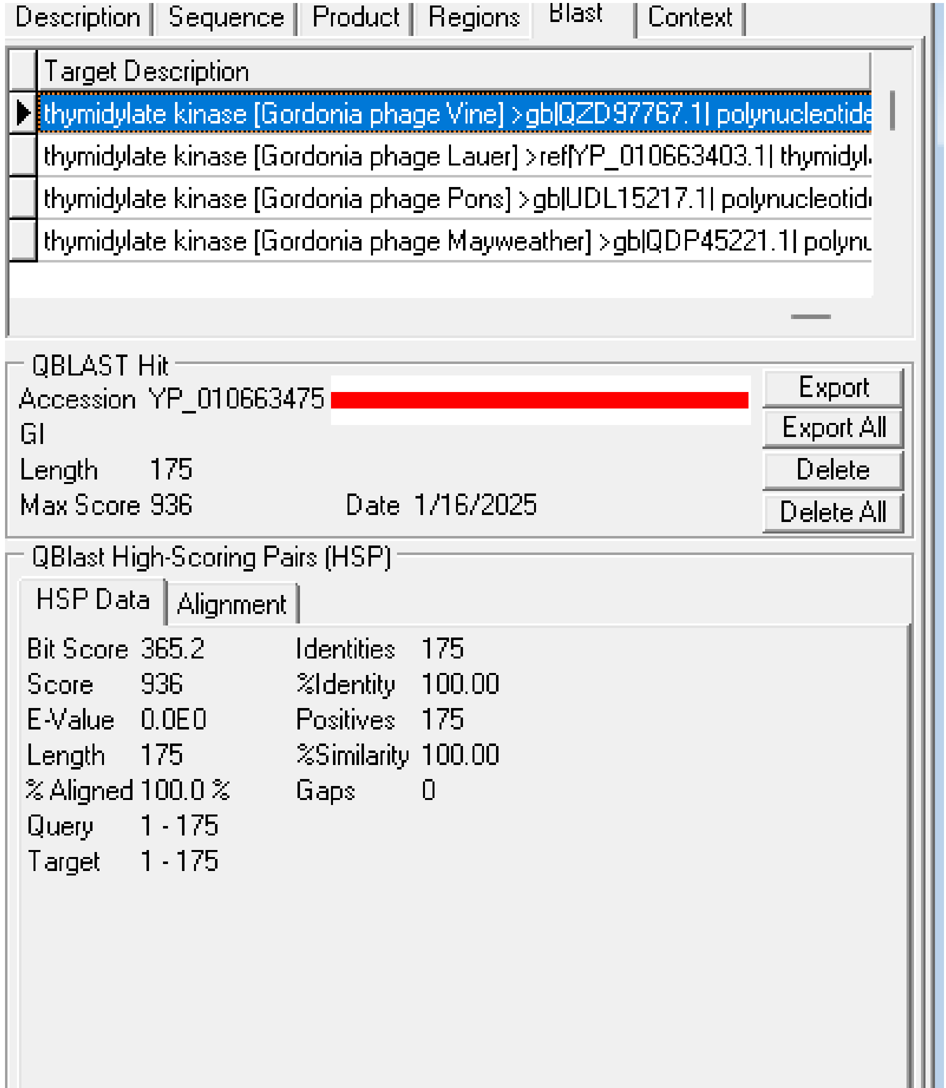
Lauer

Pons

Mayweather

CherryonLim

SheckWes



Target Description

thymidylate kinase [Gordonia phage Vine] >gb|QZD97767.1| polynucleotide

thymidylate kinase [Gordonia phage Lauer] >ref|YP_010663403.1| thymidyl

thymidylate kinase [Gordonia phage Pons] >gb|UDL15217.1| polynucleotid

thymidylate kinase [Gordonia phage Mayweather] >gb|QDP45221.1| polynu

QBLAST Hit

Accession YP_010663475

GI

Length 175

Max Score 936

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	365.2	Identities	175
Score	936	%Identity	100.00
E-Value	0.0E0	Positives	175
Length	175	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 175		
Target	1 - 175		

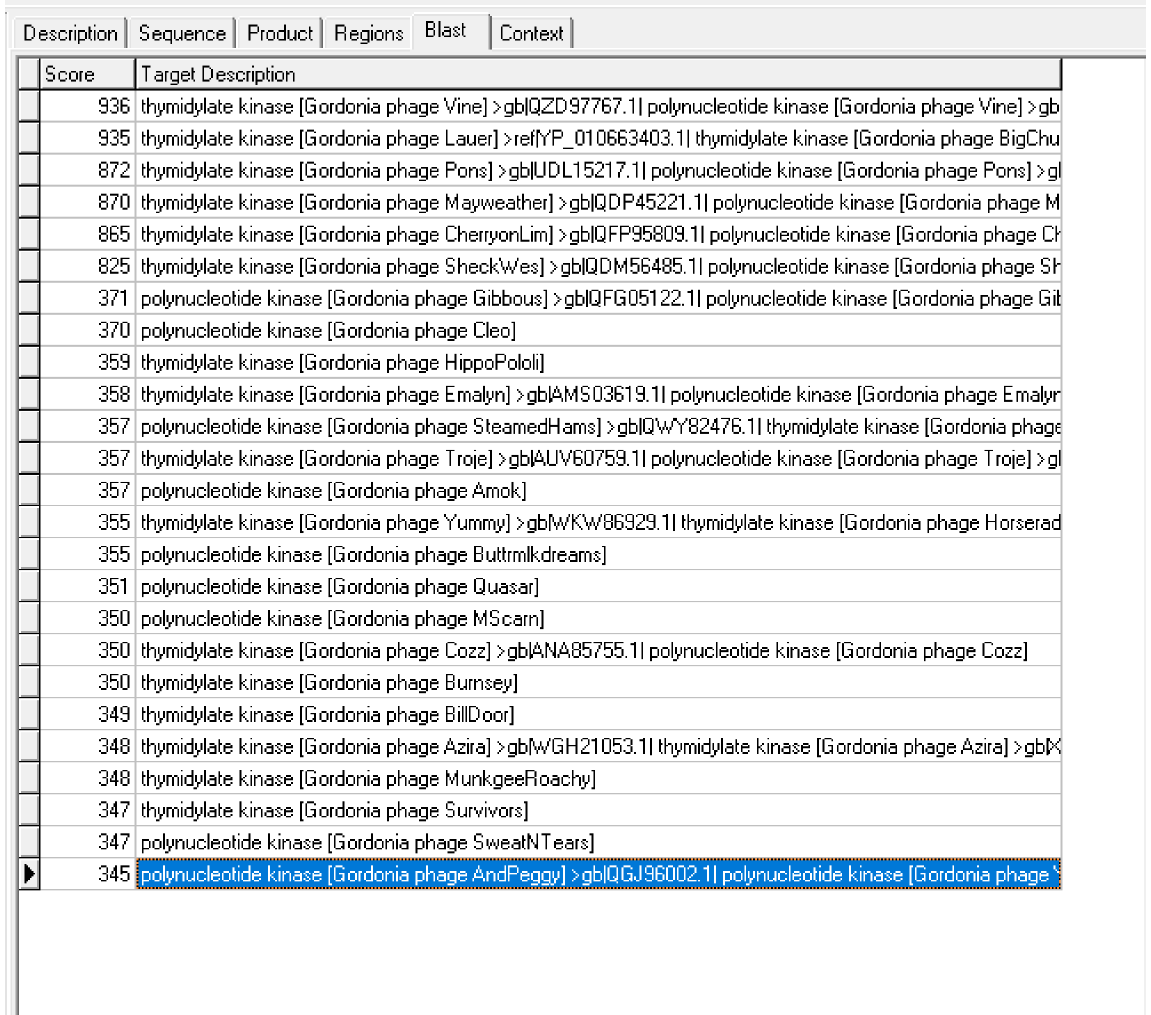
Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, it is a gene, because both Glimmer and GeneMark call it at start site 39396. The reverse frame includes all coding potential, and feature has 6 highly similar genes (0.0E0).

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- 25 1:1 alignments for start site 39396

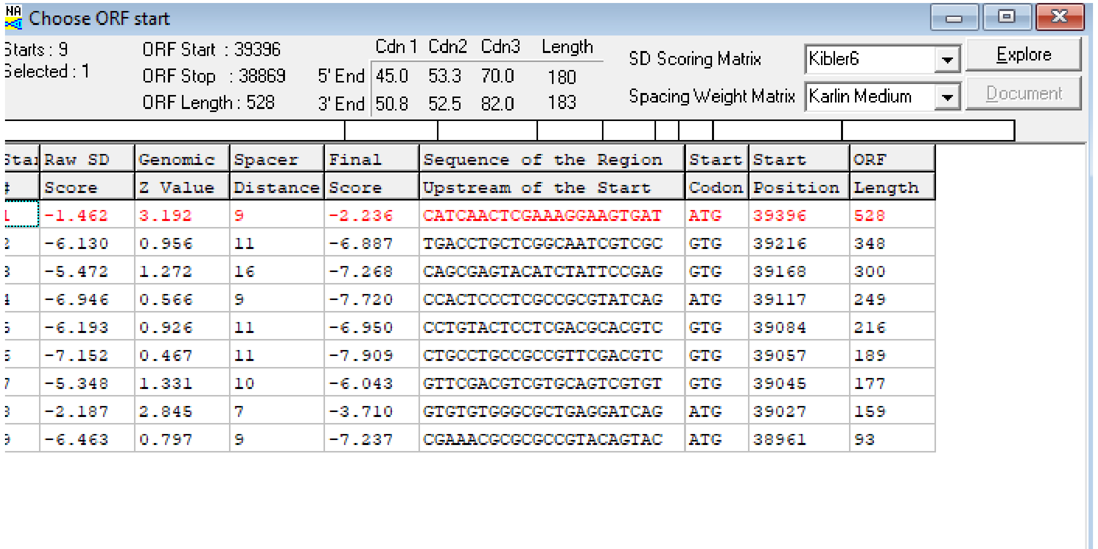
- No alternative start



Score	Target Description
936	thymidylate kinase [Gordonia phage Vine] >gb|QZD97767.1| polynucleotide kinase [Gordonia phage Vine] >gb
935	thymidylate kinase [Gordonia phage Lauer] >ref|YP_010663403.1| thymidylate kinase [Gordonia phage BigChu
872	thymidylate kinase [Gordonia phage Pons] >gb|UDL15217.1| polynucleotide kinase [Gordonia phage Pons] >gl
870	thymidylate kinase [Gordonia phage Mayweather] >gb|QDP45221.1| polynucleotide kinase [Gordonia phage M
865	thymidylate kinase [Gordonia phage CherryonLim] >gb|QFP95809.1| polynucleotide kinase [Gordonia phage Ch
825	thymidylate kinase [Gordonia phage SheckWes] >gb|QDM56485.1| polynucleotide kinase [Gordonia phage Sh
371	polynucleotide kinase [Gordonia phage Gibbous] >gb|QFG05122.1| polynucleotide kinase [Gordonia phage Gil
370	polynucleotide kinase [Gordonia phage Cleo]
359	thymidylate kinase [Gordonia phage HippoPololi]
358	thymidylate kinase [Gordonia phage Emalyn] >gb|AMS03619.1| polynucleotide kinase [Gordonia phage Emalyr
357	polynucleotide kinase [Gordonia phage SteamedHams] >gb|QWY82476.1| thymidylate kinase [Gordonia phage
357	thymidylate kinase [Gordonia phage Troje] >gb|AUV60759.1| polynucleotide kinase [Gordonia phage Troje] >gl
357	polynucleotide kinase [Gordonia phage Amok]
355	thymidylate kinase [Gordonia phage Yummy] >gb|WKW86929.1| thymidylate kinase [Gordonia phage Horserad
355	polynucleotide kinase [Gordonia phage Buttrmlkdreams]
351	polynucleotide kinase [Gordonia phage Quasar]
350	polynucleotide kinase [Gordonia phage MScarn]
350	thymidylate kinase [Gordonia phage Cozz] >gb|ANA85755.1| polynucleotide kinase [Gordonia phage Cozz]
350	thymidylate kinase [Gordonia phage Burnsey]
349	thymidylate kinase [Gordonia phage BillDoor]
348	thymidylate kinase [Gordonia phage Azira] >gb|WGH21053.1| thymidylate kinase [Gordonia phage Azira] >gb|X
348	thymidylate kinase [Gordonia phage MunkgeeRoachy]
347	thymidylate kinase [Gordonia phage Survivors]
347	polynucleotide kinase [Gordonia phage SweatNTears]
345	polynucleotide kinase [Gordonia phage AndPeggy] >gb|QGJ96002.1| polynucleotide kinase [Gordonia phage

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Start site: 39396
- Z value: 3.192
- Final score: -2.236



Starts : 9
Selected : 1

ORF Start : 39396
ORF Stop : 38869
ORF Length : 528

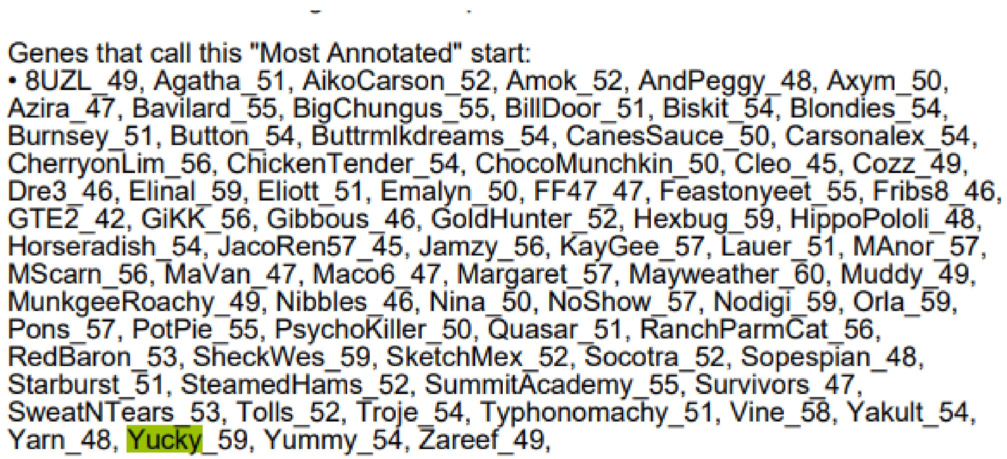
	Cdn 1	Cdn2	Cdn3	Length
5' End	45.0	53.3	70.0	180
3' End	50.8	52.5	82.0	183

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

Star #	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	-1.462	3.192	9	-2.236	CATCAACTCGAAAGGAAGTGAT	ATG	39396	528
2	-6.130	0.956	11	-6.887	TGACCTGCTCGGCAATCGTCGC	GTG	39216	348
3	-5.472	1.272	16	-7.268	CAGCGAGTACATCTATTCCGAG	GTG	39168	300
4	-6.946	0.566	9	-7.720	CCACTCCCTCGCCGCGTATCAG	ATG	39117	249
5	-6.193	0.926	11	-6.950	CCTGTACTCCTCGACGCACGTC	GTG	39084	216
6	-7.152	0.467	11	-7.909	CTGCCTGCCGCCGTTCGACGTC	GTG	39057	189
7	-5.348	1.331	10	-6.043	GTTCGACGTCGTGCAGTCGTGT	GTG	39045	177
8	-2.187	2.845	7	-3.710	GTGTGTGGGCGCTGAGGATCAG	ATG	39027	159
9	-6.463	0.797	9	-7.237	CGAAACGCGCGCCGTACAGTAC	ATG	38961	93

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Start site: 5 @39396 has 58 MA’s

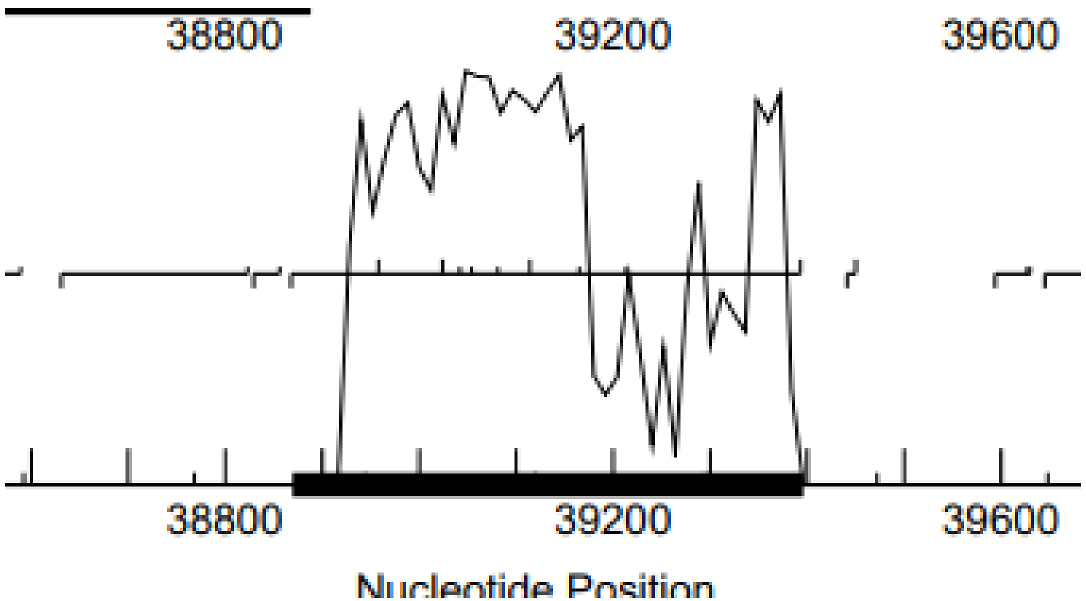


Genes that call this "Most Annotated" start:

• 8UZL_49, Agatha_51, AikoCarson_52, Amok_52, AndPeggy_48, Axym_50, Azira_47, Bavilard_55, BigChungus_55, BillDoor_51, Biskit_54, Blondies_54, Burnsey_51, Button_54, Buttrmlkdreams_54, CanesSauce_50, Carsonalex_54, CherryonLim_56, ChickenTender_54, ChocoMunchkin_50, Cleo_45, Cozz_49, Dre3_46, Elinal_59, Eliott_51, Emalyn_50, FF47_47, Feastonyeet_55, Fribs8_46, GTE2_42, GiKK_56, Gibbous_46, GoldHunter_52, Hexbug_59, HippoPololi_48, Horseradish_54, JacoRen57_45, Jamzy_56, KayGee_57, Lauer_51, MAnor_57, MScarn_56, MaVan_47, Maco6_47, Margaret_57, Mayweather_60, Muddy_49, MunkgeeRoachy_49, Nibbles_46, Nina_50, NoShow_57, Nodigi_59, Orla_59, Pons_57, PotPie_55, PsychoKiller_50, Quasar_51, RanchParmCat_56, RedBaron_53, SheckWes_59, SketchMex_52, Socotra_52, Sopespian_48, Starburst_51, SteamedHams_52, SummitAcademy_55, Survivors_47, SweatNTears_53, Tolls_52, Troje_54, Typhonomachy_51, Vine_58, Yakult_54, Yarn_48, Yucky_59, Yummy_54, Zareef_49,

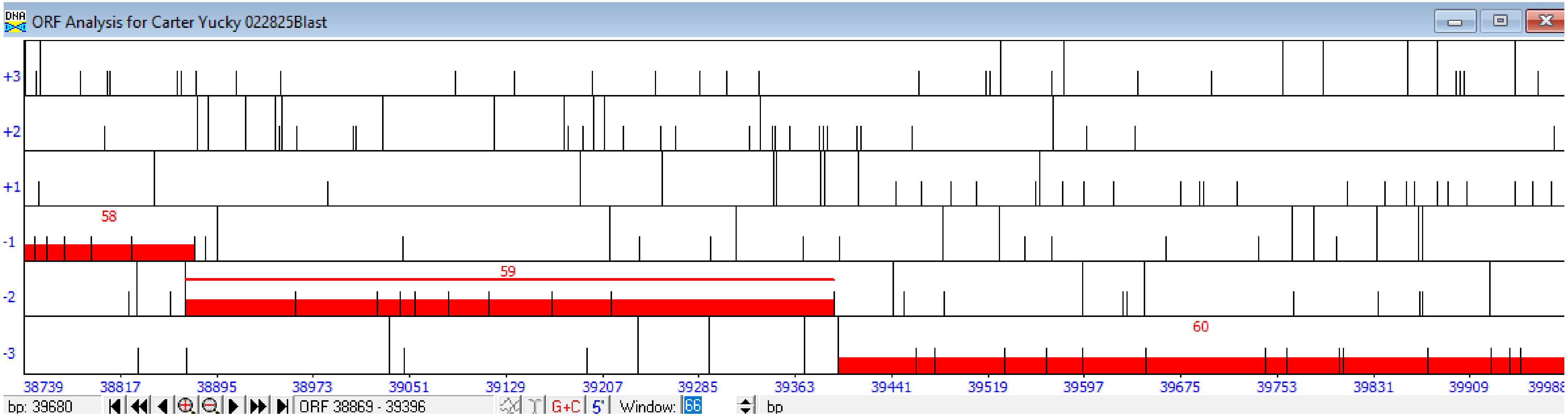
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

At start site 39396 all coding potential is included



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Gap: 1



ORF Analysis for Carter Yucky 022825Blast

bp: 39680

ORF 38869 - 39396

Window: 66 bp

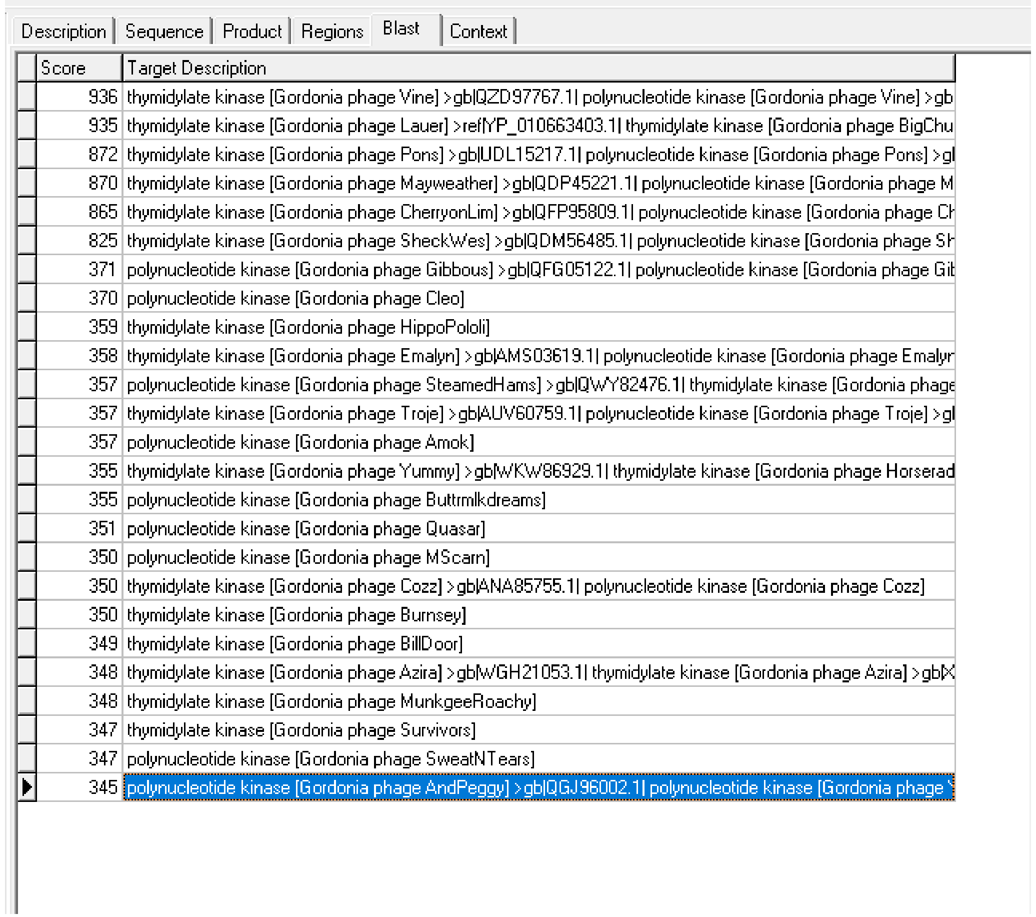
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	39396
GeneMark	Glimmer & GeneMark
Coding potential	Includes all cp
RBS	Z value: 3.192 final score: -2.236
BLAST	25 1:1
Starterator	58 MA’s
Gap	1

The start site is 39396 because both Glimmer and GeneMark call it, the reverse frame includes all coding potential, it has a z value greater than 2 and has 25 1:1 alignments.

BLAST function evidence. What assigned functions do other highly similar genes have?

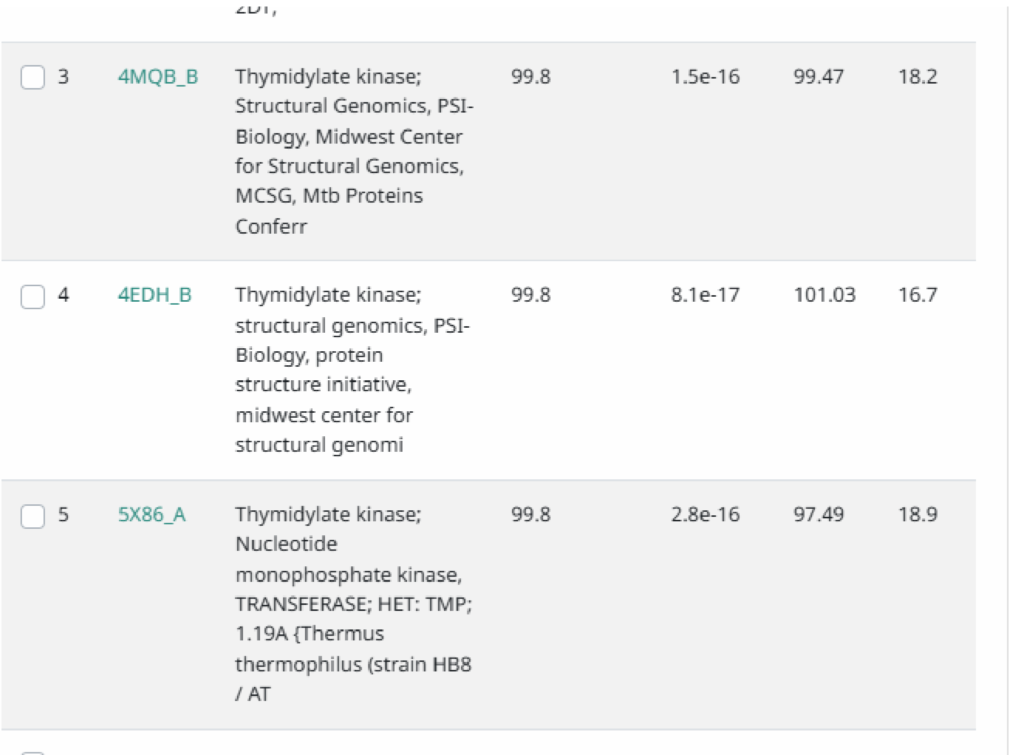
- 16 thymidylate kinase
- 9 polynucleotide kinase



Score	Target Description
936	thymidylate kinase [Gordonia phage Vine] >gb|QZD97767.1| polynucleotide kinase [Gordonia phage Vine] >gb
935	thymidylate kinase [Gordonia phage Lauer] >ref|YP_010663403.1| thymidylate kinase [Gordonia phage BigChu
872	thymidylate kinase [Gordonia phage Pons] >gb|UDL15217.1| polynucleotide kinase [Gordonia phage Pons] >gl
870	thymidylate kinase [Gordonia phage Mayweather] >gb|QDP45221.1| polynucleotide kinase [Gordonia phage M
865	thymidylate kinase [Gordonia phage CherryonLim] >gb|QFP95809.1| polynucleotide kinase [Gordonia phage Ch
825	thymidylate kinase [Gordonia phage SheckWes] >gb|QDM56485.1| polynucleotide kinase [Gordonia phage Sh
371	polynucleotide kinase [Gordonia phage Gibbous] >gb|QFG05122.1| polynucleotide kinase [Gordonia phage Gib
370	polynucleotide kinase [Gordonia phage Cleo]
359	thymidylate kinase [Gordonia phage HippoPololi]
358	thymidylate kinase [Gordonia phage Emalyn] >gb|AMS03619.1| polynucleotide kinase [Gordonia phage Emalyn
357	polynucleotide kinase [Gordonia phage SteamedHams] >gb|QWY82476.1| thymidylate kinase [Gordonia phage
357	thymidylate kinase [Gordonia phage Troje] >gb|AUV60759.1| polynucleotide kinase [Gordonia phage Troje] >gl
357	polynucleotide kinase [Gordonia phage Amok]
355	thymidylate kinase [Gordonia phage Yummy] >gb|WKW86929.1| thymidylate kinase [Gordonia phage Horserad
355	polynucleotide kinase [Gordonia phage Buttrmlkdreams]
351	polynucleotide kinase [Gordonia phage Quasar]
350	polynucleotide kinase [Gordonia phage MScarn]
350	thymidylate kinase [Gordonia phage Cozz] >gb|ANA85755.1| polynucleotide kinase [Gordonia phage Cozz]
350	thymidylate kinase [Gordonia phage Burnsey]
349	thymidylate kinase [Gordonia phage BillDoor]
348	thymidylate kinase [Gordonia phage Azira] >gb|WGH21053.1| thymidylate kinase [Gordonia phage Azira] >gb|X
348	thymidylate kinase [Gordonia phage MunkgeeRoachy]
347	thymidylate kinase [Gordonia phage Survivors]
347	polynucleotide kinase [Gordonia phage SweatNTears]
345	polynucleotide kinase [Gordonia phage AndPeggy] >gb|QGJ96002.1| polynucleotide kinase [Gordonia phage

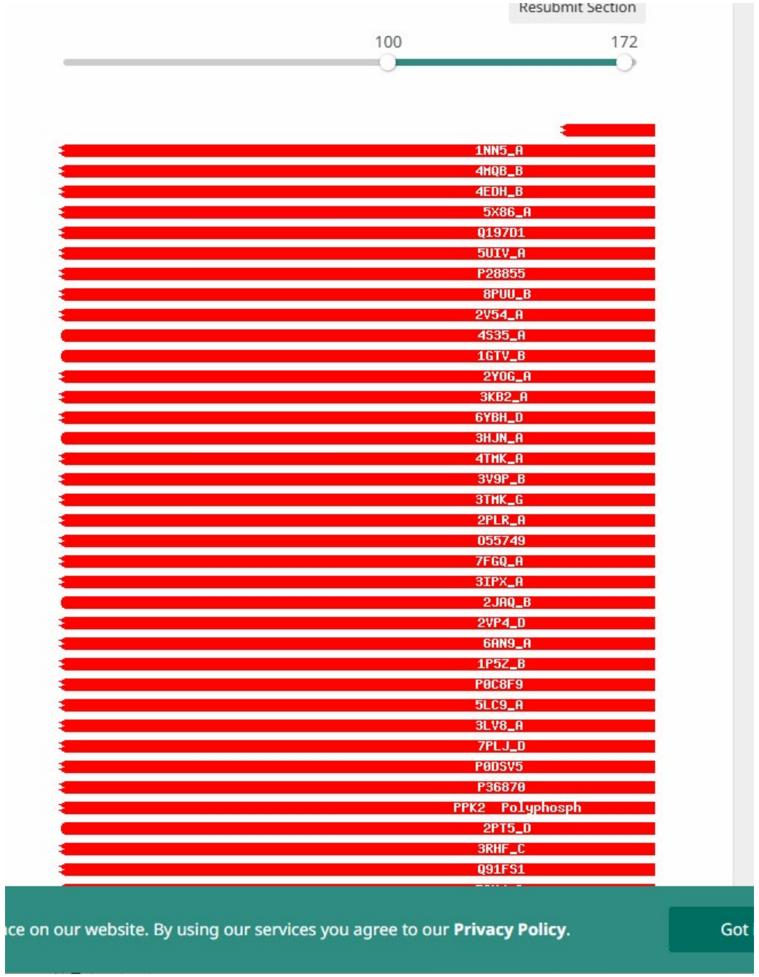
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- Multiple hits for function thymidylate kinase
- No function list requirements



2DT,

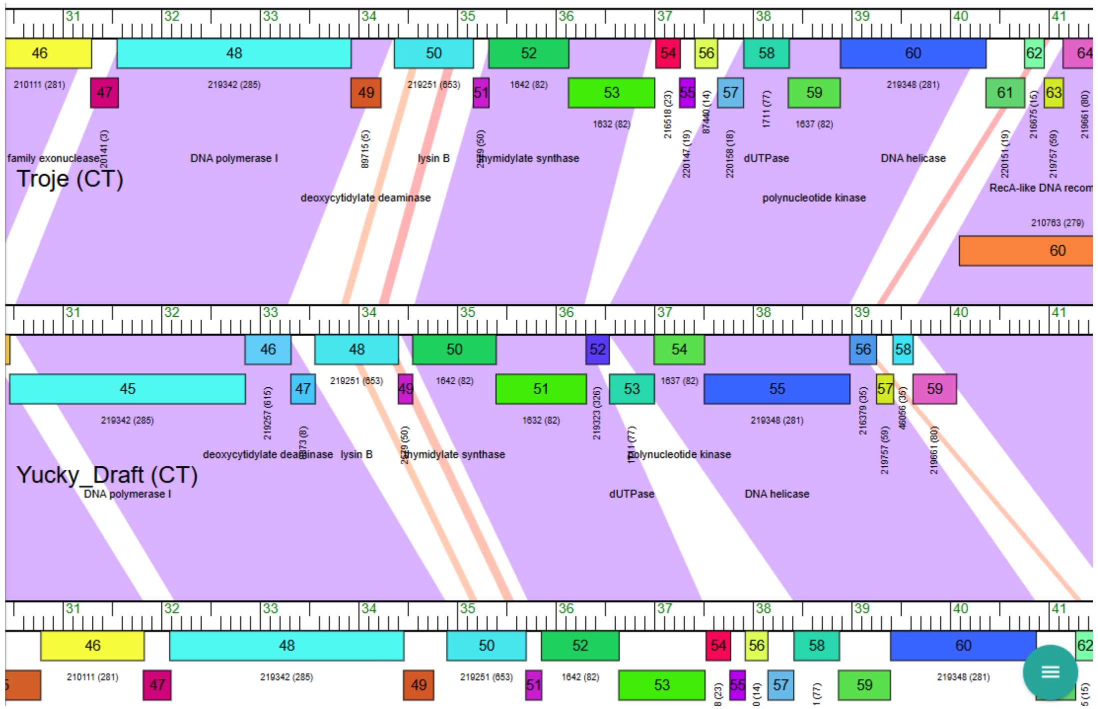
□	3	4MQB_B	Thymidylate kinase; Structural Genomics, PSI-Biology, Midwest Center for Structural Genomics, MCSG, Mtb Proteins Conferr	99.8	1.5e-16	99.47	18.2
□	4	4EDH_B	Thymidylate kinase; structural genomics, PSI-Biology, protein structure initiative, midwest center for structural genomi	99.8	8.1e-17	101.03	16.7
□	5	5X86_A	Thymidylate kinase; Nucleotide monophosphate kinase, TRANSFERASE; HET: TMP; 1.19A {Thermus thermophilus (strain HB8 / AT	99.8	2.8e-16	97.49	18.9



ce on our website. By using our services you agree to our **Privacy Policy**.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Yucky 59 conserved domain: TMPK, tmk, and Thymidylate_kin function: none
- Vine 58 conserved domain: TMPK, tmk, and Thymidylate_kin function: polynucleotide kinase
- Lauer 51 conserved domain: TMPK, tmk, and Thymidylate_kin function: polynucleotide kinase



Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- None

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

The function is thymidylate kinase because it has the highest function count for BLAST evidence, has the highest number of hits in Hhpred with 90% probability and an E value less than 1. The conserved domain for Yucky and highly similar genes is also Thymidylate_kin.

Feature 59 – Reverse 39398

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

What is the autoannotated start?

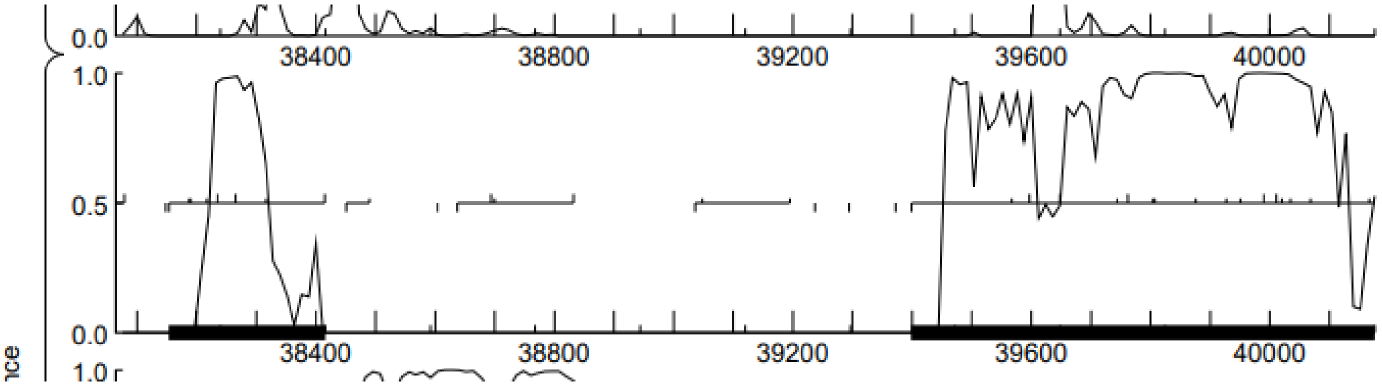
Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

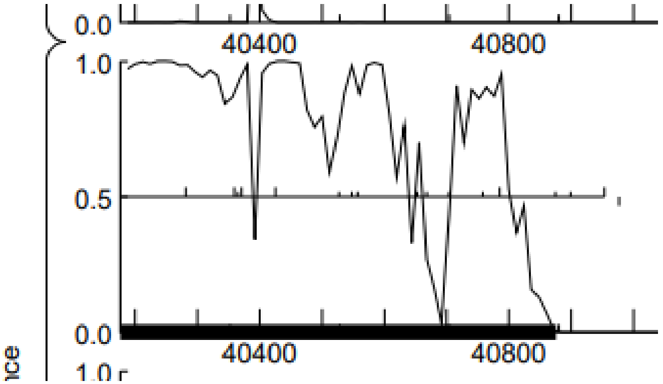
- Feature 59
- Stop site: 39398
- Called by both Glimmer and GeneMark at 40876
- Overlap: 4

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Start site 40876

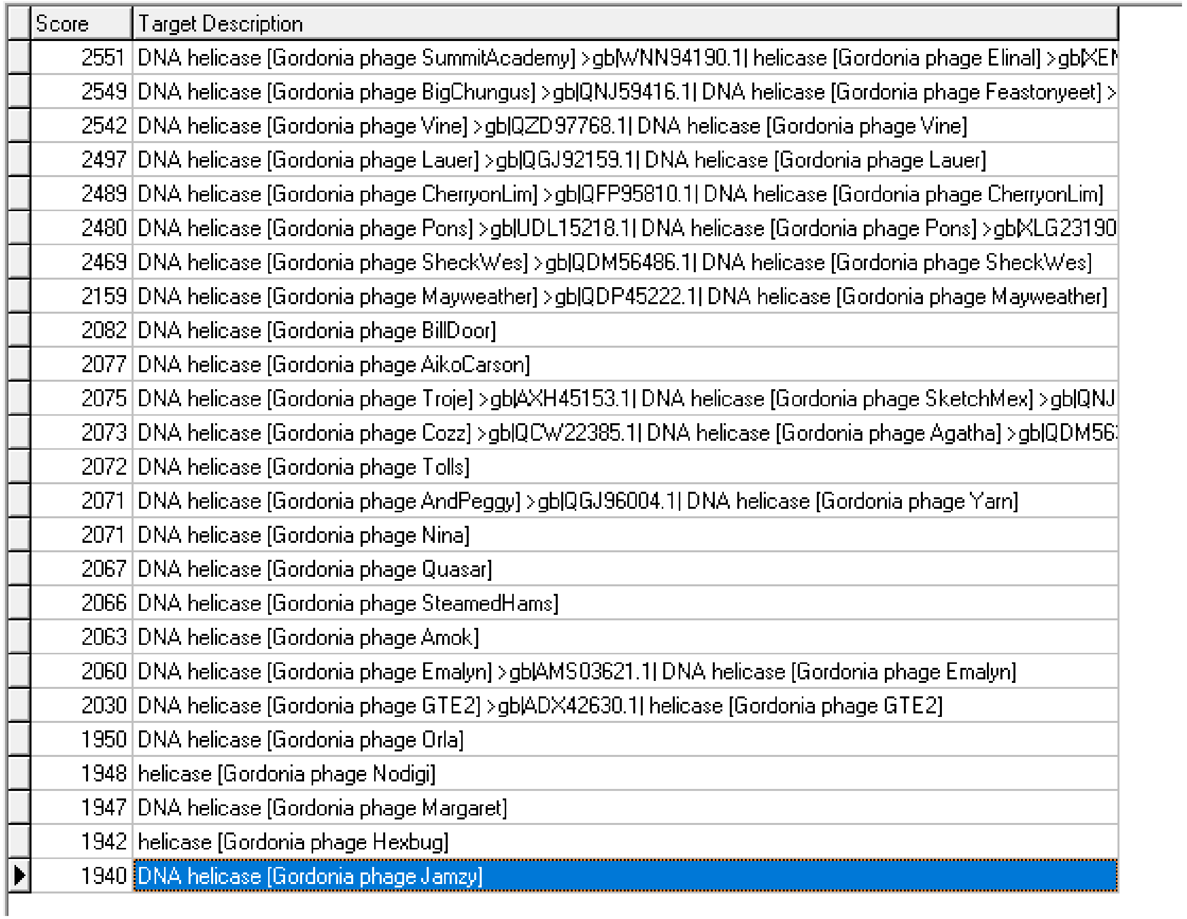
Is a continuation of coding potential in reverse frame 1 above.





BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 25 highly similar genes (0.0E0)



Score	Target Description
2551	DNA helicase [Gordonia phage SummitAcademy] >gb|WNN94190.1| helicase [Gordonia phage Elinal] >gb|XEN
2549	DNA helicase [Gordonia phage BigChungus] >gb|QNJ59416.1| DNA helicase [Gordonia phage Feastonyeet] >
2542	DNA helicase [Gordonia phage Vine] >gb|QZD97768.1| DNA helicase [Gordonia phage Vine]
2497	DNA helicase [Gordonia phage Lauer] >gb|QGJ92159.1| DNA helicase [Gordonia phage Lauer]
2489	DNA helicase [Gordonia phage CherryonLim] >gb|QFP95810.1| DNA helicase [Gordonia phage CherryonLim]
2480	DNA helicase [Gordonia phage Pons] >gb|UDL15218.1| DNA helicase [Gordonia phage Pons] >gb|XLG23190
2469	DNA helicase [Gordonia phage SheckWes] >gb|QDM56486.1| DNA helicase [Gordonia phage SheckWes]
2159	DNA helicase [Gordonia phage Mayweather] >gb|QDP45222.1| DNA helicase [Gordonia phage Mayweather]
2082	DNA helicase [Gordonia phage BillDoor]
2077	DNA helicase [Gordonia phage AikoCarson]
2075	DNA helicase [Gordonia phage Troje] >gb|AXH45153.1| DNA helicase [Gordonia phage SketchMex] >gb|QNJ
2073	DNA helicase [Gordonia phage Cozz] >gb|QCW22385.1| DNA helicase [Gordonia phage Agatha] >gb|QDM56
2072	DNA helicase [Gordonia phage Tolls]
2071	DNA helicase [Gordonia phage AndPeggy] >gb|QGJ96004.1| DNA helicase [Gordonia phage Yarn]
2071	DNA helicase [Gordonia phage Nina]
2067	DNA helicase [Gordonia phage Quasar]
2066	DNA helicase [Gordonia phage SteamedHams]
2063	DNA helicase [Gordonia phage Amok]
2060	DNA helicase [Gordonia phage Emalyn] >gb|AMS03621.1| DNA helicase [Gordonia phage Emalyn]
2030	DNA helicase [Gordonia phage GTE2] >gb|ADX42630.1| helicase [Gordonia phage GTE2]
1950	DNA helicase [Gordonia phage Orla]
1948	helicase [Gordonia phage Nodigi]
1947	DNA helicase [Gordonia phage Margaret]
1942	helicase [Gordonia phage Hexbug]
1940	DNA helicase [Gordonia phage Jamzy]

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, it is a gene because both Glimmer and GeneMark call it, the reverse frame includes a continuation of coding potential and has 25 highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- Start site 40876:

Has 12 1:1 alignments

Jamzy

Hexbug

Margaret

Nodigi

Orla

SheckWes

Pons

CherryonLim

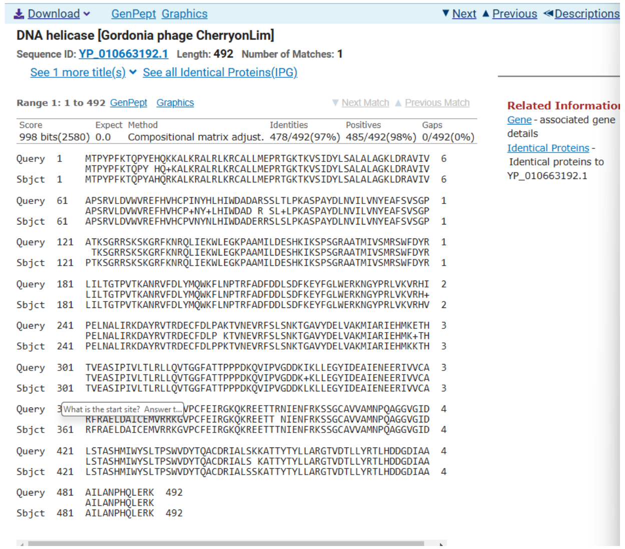
Lauer

Vine

BigChungus

SummitAcademy

- Start site 40759:
- Has 7 1:492 alignments



DNA helicase [Gordonia phage CherryonLim]

Sequence ID: YP_010663192.1 Length: 492 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 492 GenPept Graphics

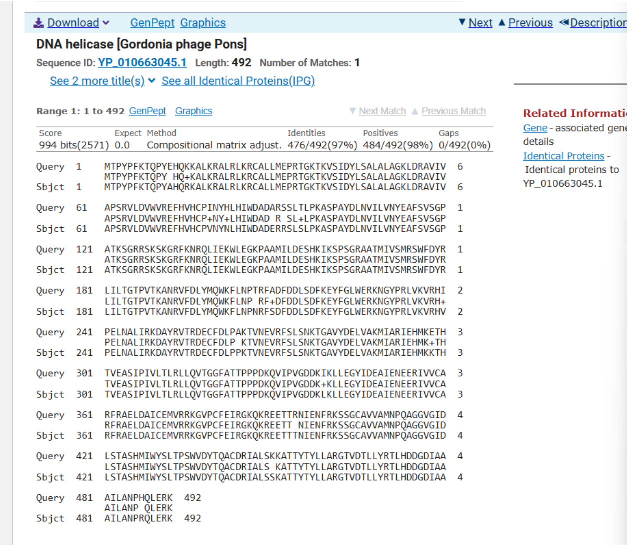
Score	Expect	Method	Identities	Positives	Gaps
998 bits(2580)	0.0	Compositional matrix adjust.	478/492(97%)	485/492(98%)	0/492(0%)

```
Query  1    MTPYPFKTQPYEHQKKALKRALRLKRCALLMEPRTGKTKVSIDYLSALALAGKLDRAVIV  6
            MTPYPFKTQPY HQ+KALKRALRLKRCALLMEPRTGKTKVSIDYLSALALAGKLDRAVIV
Sbjct  1    MTPYPFKTQPYAHQRKALKRALRLKRCALLMEPRTGKTKVSIDYLSALALAGKLDRAVIV  6

Query  61   APSRVLDVWVREFHVHCPINYHLHIWDADARSSLTLPKASPAYDLNVILVNYEAFSVSGP  1
            APSRVLDVWVREFHVHCP+NY+LHIWDAD R SL+LPKASPAYDLNVILVNYEAFSVSGP
Sbjct  61   APSRVLDVWVREFHVHCPVNYNLHIWDADERRSLSLPKASPAYDLNVILVNYEAFSVSGP  1

Query  121  ATKSGRRSKSKGRFKNRQLIEKWLEGKPAAMILDESHKIKSPSGRAATMIVSMRSWFDYR  1
             TKSGRRSKSKGRFKNRQLIEKWLEGKPAAMILDESHKIKSPSGRAATMIVSMRSWFDYR
Sbjct  121  PTKSGRRSKSKGRFKNRQLIEKWLEGKPAAMILDESHKIKSPSGRAATMIVSMRSWFDYR  1

Query  181  LILTGTPVTKANRVFDLYMQWKFLNPTRFADFDDLSDFKEYFGLWERKNGYPRLVKVRHI  2
            LILTGTPVTKANRVFDLYMQWKFLNPTRFADFDDLSDFKEYFGLWERKNGYPRLVKVRH+
Sbjct  181  LILTGTPVTKANRVFDLYMQWKFLNPTRFADFDDLSDFKEYFGLWERKNGYPRLVKVRHV  2

Query  241  PELNALIRKDAYRVTRDECFDLPAKTVNEVRFSLSNKTGAVYDELVAKMIARIEHMKETH  3
            PELNALIRKDAYRVTRDECFDLP KTVNEVRFSLSNKTGAVYDELVAKMIARIEHMK+TH
Sbjct  241  PELNALIRKDAYRVTRDECFDLPPKTVNEVRFSLSNKTGAVYDELVAKMIARIEHMKKTH  3

Query  301  TVEASIPIVLTLRLLQVTGGFATTPPPDKQVIPVGDDKIKLLEGYIDEAIENEERIVVCA  3
            TVEASIPIVLTLRLLQVTGGFATTPPPDKQVIPVGDDK+KLLEGYIDEAIENEERIVVCA
Sbjct  301  TVEASIPIVLTLRLLQVTGGFATTPPPDKQVIPVGDDKLKLLEGYIDEAIENEERIVVCA  3

Query  3What is the start site?  Answer t...VPCFEIRGKQKREETTRNIENFRKSSGCAVVAMNPQAGGVGID  4
            RFRAELDAICEMVRRKGVPCFEIRGKQKREETT NIENFRKSSGCAVVAMNPQAGGVGID
Sbjct  361  RFRAELDAICEMVRRKGVPCFEIRGKQKREETTTNIENFRKSSGCAVVAMNPQAGGVGID  4

Query  421  LSTASHMIWYSLTPSWVDYTQACDRIALSKKATTYTYLLARGTVDTLLYRTLHDDGDIAA  4
            LSTASHMIWYSLTPSWVDYTQACDRIALS KATTYTYLLARGTVDTLLYRTLHDDGDIAA
Sbjct  421  LSTASHMIWYSLTPSWVDYTQACDRIALSSKATTYTYLLARGTVDTLLYRTLHDDGDIAA  4

Query  481  AILANPHQLERK  492
            AILANPHQLERK
Sbjct  481  AILANPHQLERK  492
```

Related Informatio

Gene - associated gene details

Identical Proteins - Identical proteins to YP_010663192.1



DNA helicase [Gordonia phage Pons]

Sequence ID: YP_010663045.1 Length: 492 Number of Matches: 1

See 2 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 492 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
994 bits(2571)	0.0	Compositional matrix adjust.	476/492(97%)	484/492(98%)	0/492(0%)

```
Query  1    MTPYPFKTQPYEHQKKALKRALRLKRCALLMEPRTGKTKVSIDYLSALALAGKLDRAVIV  6
            MTPYPFKTQPY HQ+KALKRALRLKRCALLMEPRTGKTKVSIDYLSALALAGKLDRAVIV
Sbjct  1    MTPYPFKTQPYAHQRKALKRALRLKRCALLMEPRTGKTKVSIDYLSALALAGKLDRAVIV  6

Query  61   APSRVLDVWVREFHVHCPINYHLHIWDADARSSLTLPKASPAYDLNVILVNYEAFSVSGP  1
            APSRVLDVWVREFHVHCP+NY+LHIWDAD R SL+LPKASPAYDLNVILVNYEAFSVSGP
Sbjct  61   APSRVLDVWVREFHVHCPVNYNLHIWDADERRSLSLPKASPAYDLNVILVNYEAFSVSGP  1

Query  121  ATKSGRRSKSKGRFKNRQLIEKWLEGKPAAMILDESHKIKSPSGRAATMIVSMRSWFDYR  1
            ATKSGRRSKSKGRFKNRQLIEKWLEGKPAAMILDESHKIKSPSGRAATMIVSMRSWFDYR
Sbjct  121  ATKSGRRSKSKGRFKNRQLIEKWLEGKPAAMILDESHKIKSPSGRAATMIVSMRSWFDYR  1

Query  181  LILTGTPVTKANRVFDLYMQWKFLNPTRFADFDDLSDFKEYFGLWERKNGYPRLVKVRHI  2
            LILTGTPVTKANRVFDLYMQWKFLNP RF+DFDDLSDFKEYFGLWERKNGYPRLVKVRH+
Sbjct  181  LILTGTPVTKANRVFDLYMQWKFLNPNRFSDFDDLSDFKEYFGLWERKNGYPRLVKVRHV  2

Query  241  PELNALIRKDAYRVTRDECFDLPAKTVNEVRFSLSNKTGAVYDELVAKMIARIEHMKETH  3
            PELNALIRKDAYRVTRDECFDLP KTVNEVRFSLSNKTGAVYDELVAKMIARIEHMK+TH
Sbjct  241  PELNALIRKDAYRVTRDECFDLPPKTVNEVRFSLSNKTGAVYDELVAKMIARIEHMKKTH  3

Query  301  TVEASIPIVLTLRLLQVTGGFATTPPPDKQVIPVGDDKIKLLEGYIDEAIENEERIVVCA  3
            TVEASIPIVLTLRLLQVTGGFATTPPPDKQVIPVGDDK+KLLEGYIDEAIENEERIVVCA
Sbjct  301  TVEASIPIVLTLRLLQVTGGFATTPPPDKQVIPVGDDKLKLLEGYIDEAIENEERIVVCA  3

Query  361  RFRAELDAICEMVRRKGVPCFEIRGKQKREETTRNIENFRKSSGCAVVAMNPQAGGVGID  4
            RFRAELDAICEMVRRKGVPCFEIRGKQKREETT NIENFRKSSGCAVVAMNPQAGGVGID
Sbjct  361  RFRAELDAICEMVRRKGVPCFEIRGKQKREETTTNIENFRKSSGCAVVAMNPQAGGVGID  4

Query  421  LSTASHMIWYSLTPSWVDYTQACDRIALSKKATTYTYLLARGTVDTLLYRTLHDDGDIAA  4
            LSTASHMIWYSLTPSWVDYTQACDRIALS KATTYTYLLARGTVDTLLYRTLHDDGDIAA
Sbjct  421  LSTASHMIWYSLTPSWVDYTQACDRIALSSKATTYTYLLARGTVDTLLYRTLHDDGDIAA  4

Query  481  AILANPHQLERK  492
            AILANP QLERK
Sbjct  481  AILANPRQLERK  492
```

Related Informati

Gene - associated gen details

Identical Proteins - Identical proteins to YP_010663045.1

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

Start site 40876:

Z value: 2.979

Final score: -2.742

- Start site 40759:

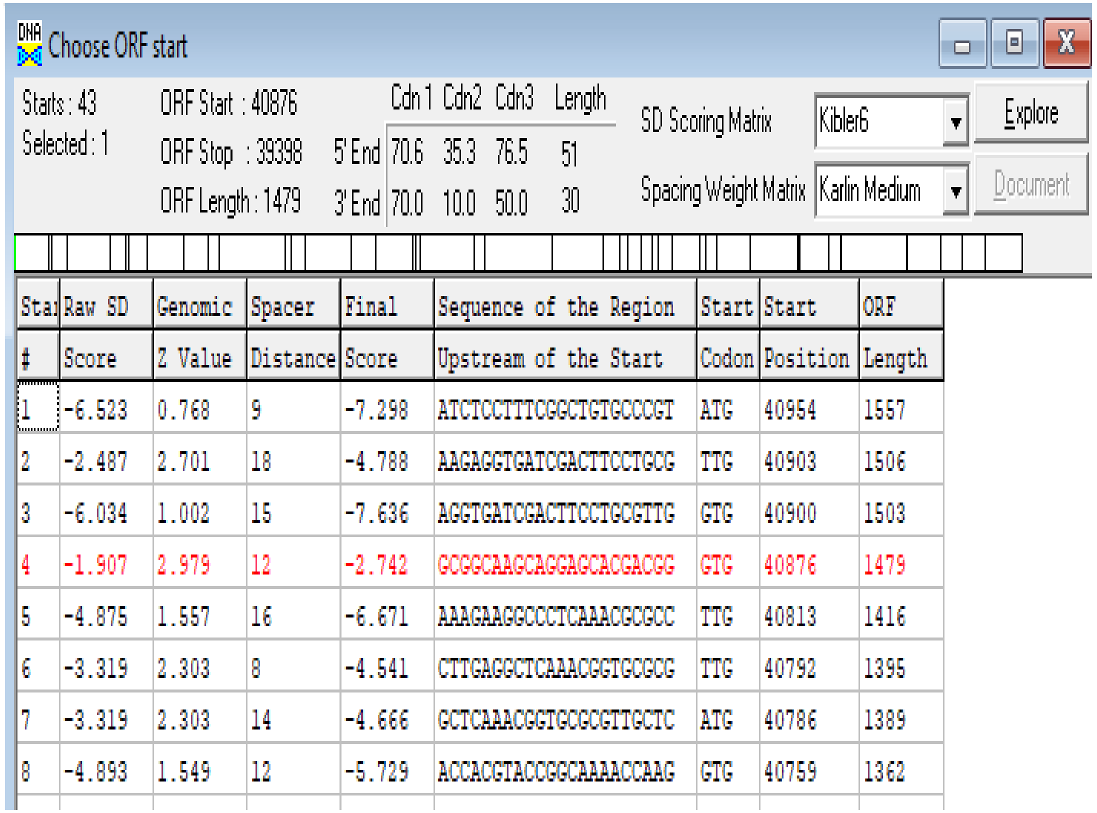
Z value: 1.549

Final score: -5.729

Start site 40759:

Z value: 1.549

Final score: -5.729



Starts : 43
Selected : 1

ORF Start : 40876
ORF Stop : 39398
ORF Length : 1479

	Cdn 1	Cdn2	Cdn3	Length
5' End	70.6	35.3	76.5	51
3' End	70.0	10.0	50.0	30

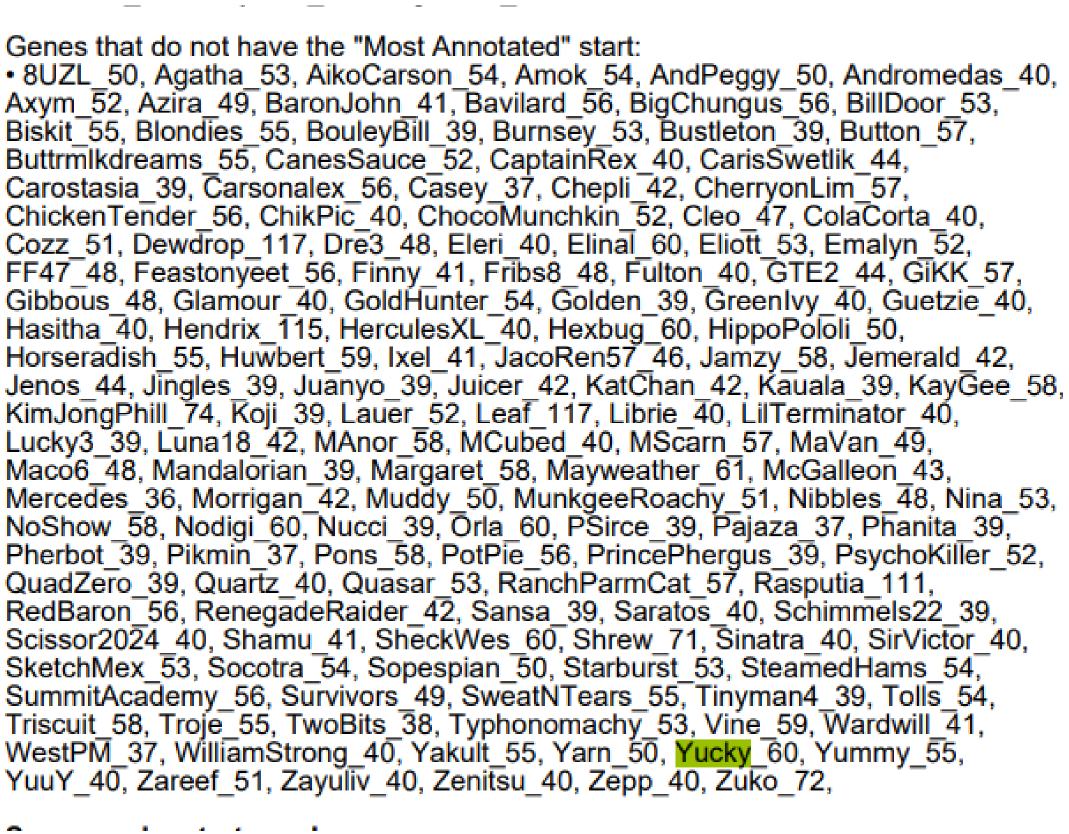
SD Scoring Matrix: Kibler6

Spacing Weight Matrix: Karlin Medium

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	-6.523	0.768	9	-7.298	ATCTCCTTTCGGCTGTGCCCGT	ATG	40954	1557
2	-2.487	2.701	18	-4.788	AAGAGGTGATCGACTTCCTGCG	TTG	40903	1506
3	-6.034	1.002	15	-7.636	AGGTGATCGACTTCCTGCGTTG	GTG	40900	1503
4	-1.907	2.979	12	-2.742	GCGGCAAGCAGGAGCACGACGG	GTG	40876	1479
5	-4.875	1.557	16	-6.671	AAAGAAGGCCCTCAAACGCGCC	TTG	40813	1416
6	-3.319	2.303	8	-4.541	CTTGAGGCTCAAACGGTGCGCG	TTG	40792	1395
7	-3.319	2.303	14	-4.666	GCTCAAACGGTGCGCGTTGCTC	ATG	40786	1389
8	-4.893	1.549	12	-5.729	ACCACGTACCGGCAAAACCAAG	GTG	40759	1362

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Start: 58 @40876 has 35 MA's
- Start: 88 @40759 has 1 MA's



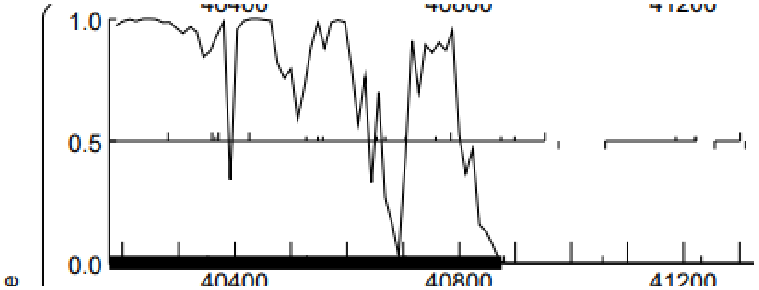
Genes that do not have the "Most Annotated" start:

• 8UZL_50, Agatha_53, AikoCarson_54, Amok_54, AndPeggy_50, Andromedas_40, Axym_52, Azira_49, BaronJohn_41, Bavilard_56, BigChungus_56, BillDoor_53, Biskit_55, Blondies_55, BouleyBill_39, Burnsey_53, Bustleton_39, Button_57, Buttrmlkdreams_55, CanesSauce_52, CaptainRex_40, CarisSwetlik_44, Carostasia_39, Carsonalex_56, Casey_37, Chepli_42, CherryonLim_57, ChickenTender_56, ChikPic_40, ChocoMunchkin_52, Cleo_47, ColaCorta_40, Cozz_51, Dewdrop_117, Dre3_48, Eleri_40, Elinal_60, Eliott_53, Emalyn_52, FF47_48, Feastonyeet_56, Finny_41, Fribs8_48, Fulton_40, GTE2_44, GiKK_57, Gibbous_48, Glamour_40, GoldHunter_54, Golden_39, GreenIvy_40, Guetzie_40, Hasitha_40, Hendrix_115, HerculesXL_40, Hexbug_60, HippoPololi_50, Horseradish_55, Huwbert_59, Ixel_41, JacoRen57_46, Jamzy_58, Jemerald_42, Jenos_44, Jingles_39, Juanyo_39, Juicer_42, KatChan_42, Kauala_39, KayGee_58, KimJongPhill_74, Koji_39, Lauer_52, Leaf_117, Librie_40, LilTerminator_40, Lucky3_39, Luna18_42, MAnor_58, MCubed_40, MScarn_57, MaVan_49, Maco6_48, Mandalorian_39, Margaret_58, Mayweather_61, McGalleon_43, Mercedes_36, Morrigan_42, Muddy_50, MunkgeeRoachy_51, Nibbles_48, Nina_53, NoShow_58, Nodigi_60, Nucci_39, Orla_60, PSirce_39, Pajaza_37, Phanita_39, Pherbot_39, Pikmin_37, Pons_58, PotPie_56, PrincePhergus_39, PsychoKiller_52, QuadZero_39, Quartz_40, Quasar_53, RanchParmCat_57, Rasputia_111, RedBaron_56, RenegadeRaider_42, Sansa_39, Saratos_40, Schimmels22_39, Scissor2024_40, Shamu_41, SheckWes_60, Shrew_71, Sinatra_40, SirVictor_40, SketchMex_53, Socotra_54, Sopespian_50, Starburst_53, SteamedHams_54, SummitAcademy_56, Survivors_49, SweatNTears_55, Tinyman4_39, Tolls_54, Triscuit_58, Troje_55, TwoBits_38, Typhonomachy_53, Vine_59, Wardwill_41, WestPM_37, WilliamStrong_40, Yakult_55, Yarn_50, Yucky_60, Yummy_55, YuuY_40, Zareef_51, Zayuliv_40, Zenitsu_40, Zepp_40, Zuko_72,

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

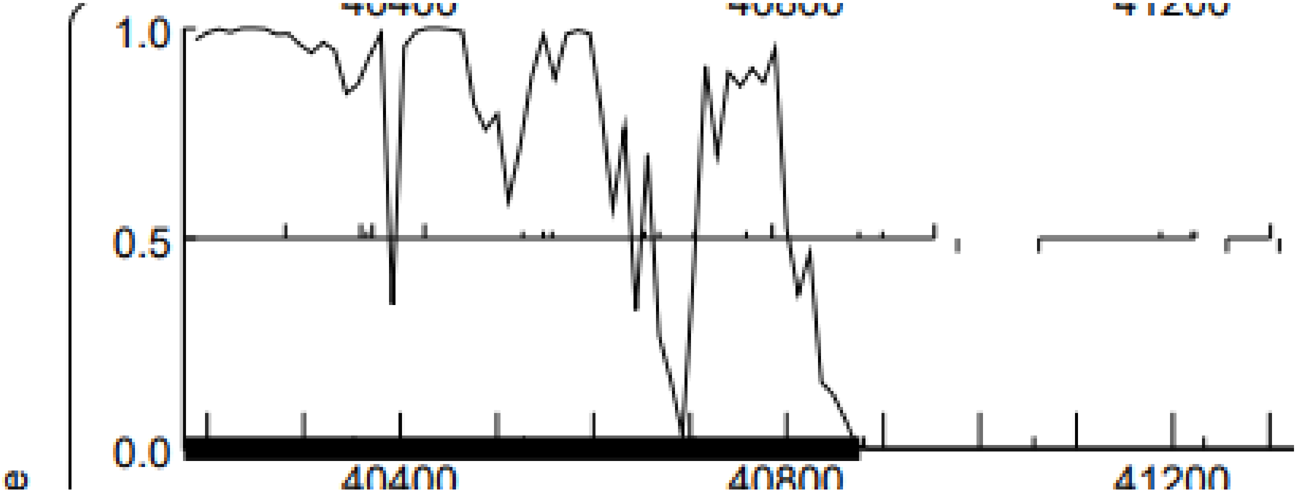
- Start: 58 @40876

Includes all coding potential



- Start: 88 @40759

Cuts off coding potential – strong peak



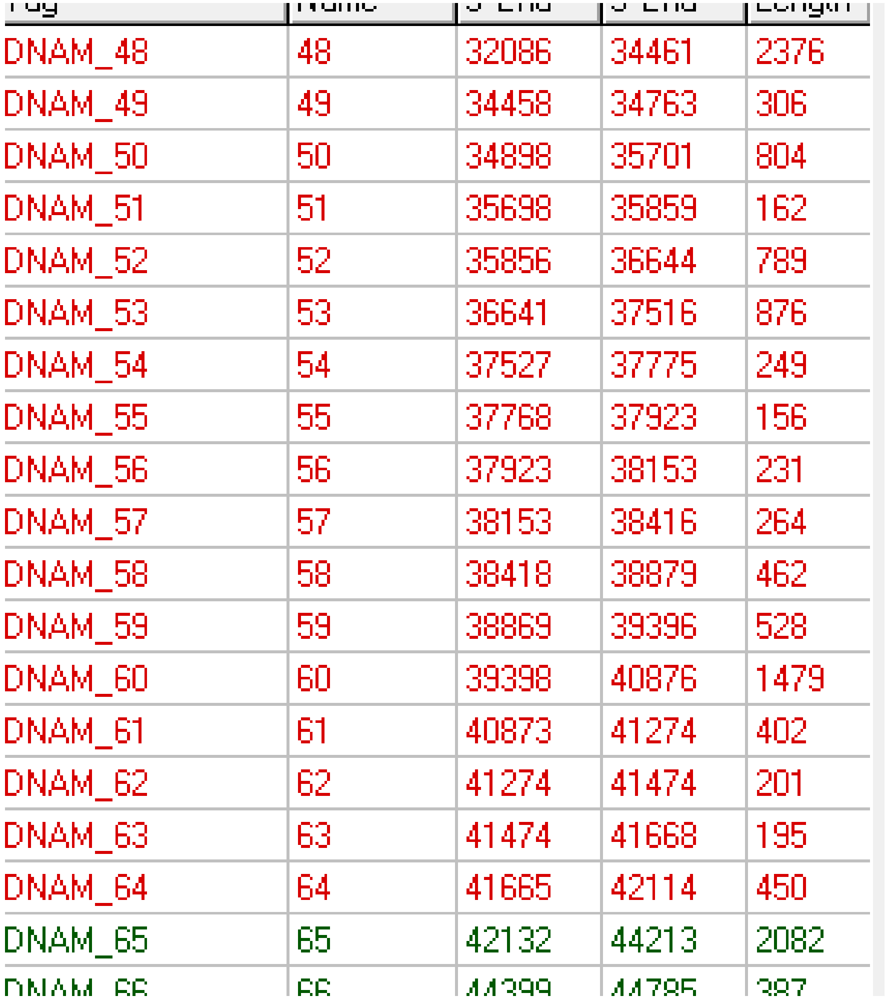
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start?   Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

Start 40876

Overlap of 4

Start 40759

Gap of 113



Tag	Name	5' End	3' End	Length
DNAM_48	48	32086	34461	2376
DNAM_49	49	34458	34763	306
DNAM_50	50	34898	35701	804
DNAM_51	51	35698	35859	162
DNAM_52	52	35856	36644	789
DNAM_53	53	36641	37516	876
DNAM_54	54	37527	37775	249
DNAM_55	55	37768	37923	156
DNAM_56	56	37923	38153	231
DNAM_57	57	38153	38416	264
DNAM_58	58	38418	38879	462
DNAM_59	59	38869	39396	528
DNAM_60	60	39398	40876	1479
DNAM_61	61	40873	41274	402
DNAM_62	62	41274	41474	201
DNAM_63	63	41474	41668	195
DNAM_64	64	41665	42114	450
DNAM_65	65	42132	44213	2082
DNAM_66	66	44399	44785	387

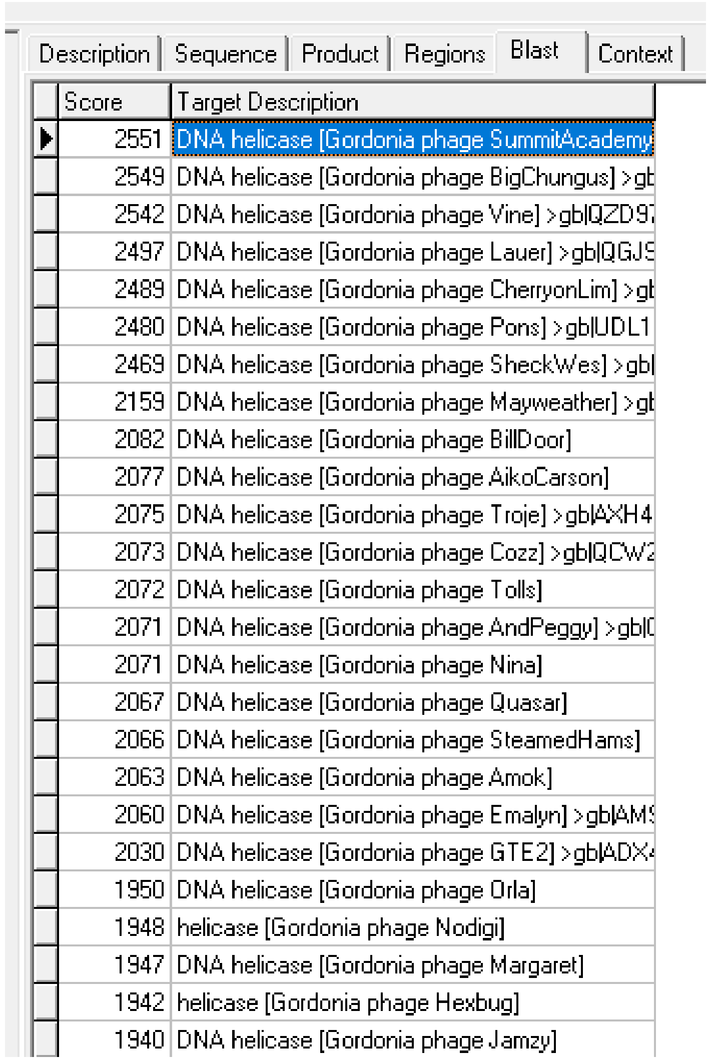
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	40876	40759
GeneMark	Glimmer and Genemark	None
Coding potential	Includes all cp	Includes all cp
RBS	Z value: 2.979 Final score: -2.742	Z value: 1.549 Final score: -5.729
BLAST	12 1:1 alignments	7 1:492 alignments
Starterator	35 MA’s	1 MA’s
Gap/overlap	Overlap of 4	Gap of 113

The best start site is 40876 because it is called by both Glimmer and Genemark. The z value is also greater than 2 and has the highest manual annotations of 35. The overlap is also 4 which is ideal.

BLAST function evidence. What assigned functions do other highly similar genes have?

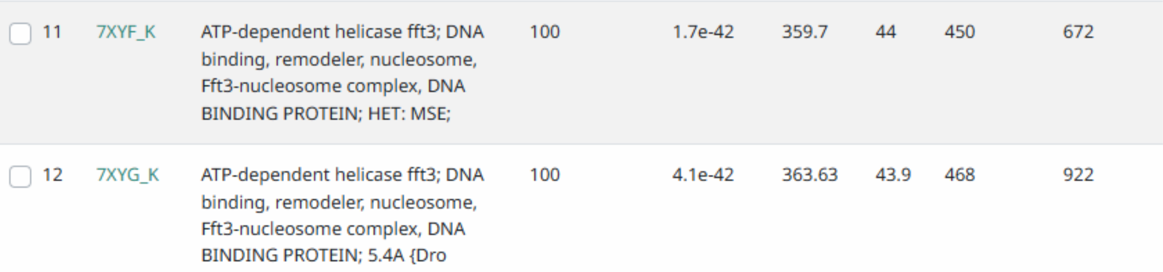
- 23 DNA helicase function
- 2 helicase function



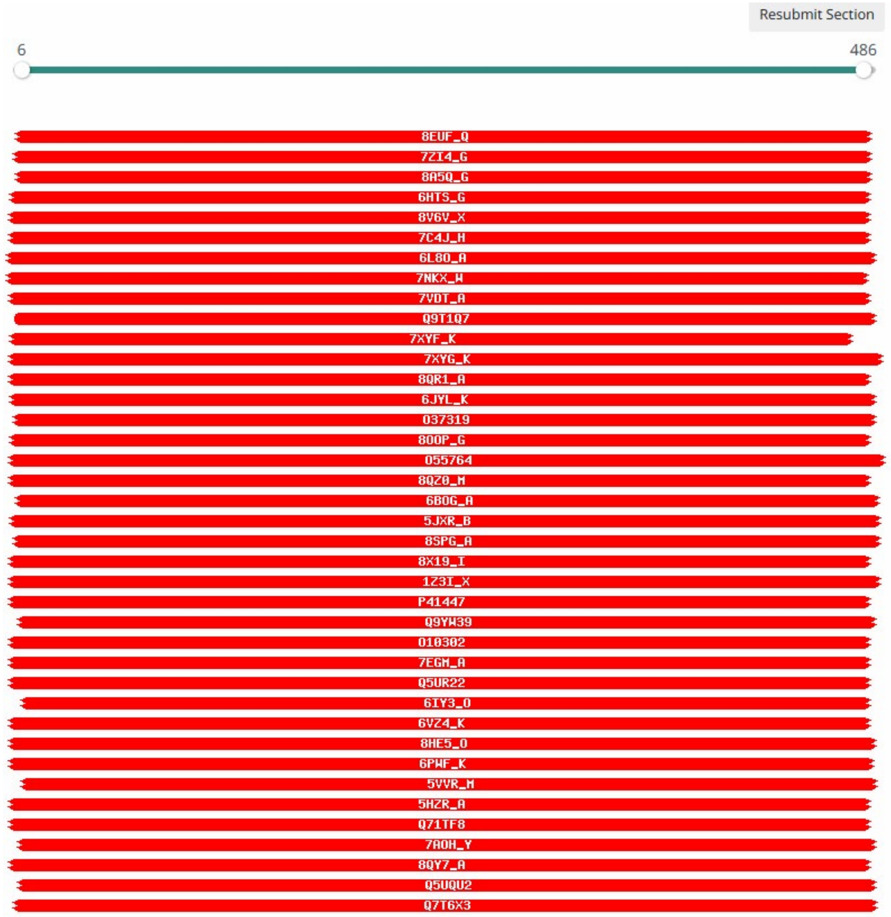
Score	Target Description
2551	DNA helicase [Gordonia phage SummitAcademy
2549	DNA helicase [Gordonia phage BigChungus] >gb
2542	DNA helicase [Gordonia phage Vine] >gb|QZD9
2497	DNA helicase [Gordonia phage Lauer] >gb|QGJ
2489	DNA helicase [Gordonia phage CherryonLim] >gb
2480	DNA helicase [Gordonia phage Pons] >gb|UDL1
2469	DNA helicase [Gordonia phage SheckWes] >gb|
2159	DNA helicase [Gordonia phage Mayweather] >gb
2082	DNA helicase [Gordonia phage BillDoor]
2077	DNA helicase [Gordonia phage AikoCarson]
2075	DNA helicase [Gordonia phage Troje] >gb|AXH4
2073	DNA helicase [Gordonia phage Cozz] >gb|QCW
2072	DNA helicase [Gordonia phage Tolls]
2071	DNA helicase [Gordonia phage AndPeggy] >gb|
2071	DNA helicase [Gordonia phage Nina]
2067	DNA helicase [Gordonia phage Quasar]
2066	DNA helicase [Gordonia phage SteamedHams]
2063	DNA helicase [Gordonia phage Amok]
2060	DNA helicase [Gordonia phage Emalyn] >gb|AM
2030	DNA helicase [Gordonia phage GTE2] >gb|ADX
1950	DNA helicase [Gordonia phage Orla]
1948	helicase [Gordonia phage Nodigi]
1947	DNA helicase [Gordonia phage Margaret]
1942	helicase [Gordonia phage Hexbug]
1940	DNA helicase [Gordonia phage Jamzy]

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- Found hits for function DNA helicase
- According to function list, had to be ATP-dependent helicase



☐	11	7XYF_K	ATP-dependent helicase fft3; DNA binding, remodeler, nucleosome, Fft3-nucleosome complex, DNA BINDING PROTEIN; HET: MSE;	100	1.7e-42	359.7	44	450	672
☐	12	7XYG_K	ATP-dependent helicase fft3; DNA binding, remodeler, nucleosome, Fft3-nucleosome complex, DNA BINDING PROTEIN; 5.4A {Dro	100	4.1e-42	363.63	43.9	468	922



Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

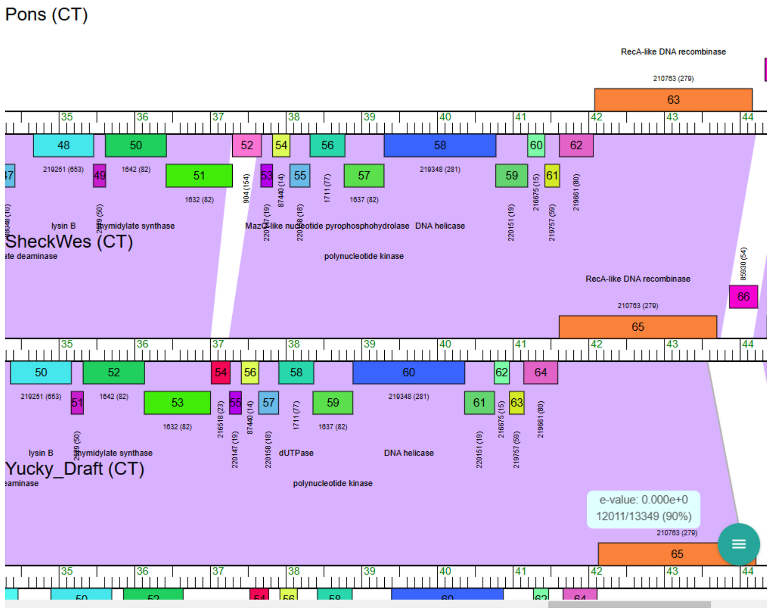
- SheckWes 60 conserved domain: Helicase_C, DEXHc_CHD3, DEXHc_CHD7

function: DNA helicase

- Pons 58 conserved domain: DEXHc_CHD6, DEXHc_CHD5, Helicase_C

function: DNA helicase

- Yucky 60 conserved domain: DEXDc, HELICc, DEXHc_ATRX-like function: none



Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- None

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function is DNA helicase, because BLAST function evidence found that there were 23 highly similar genes with function DNA helicase. Also, Hhpred evidence found hits with function at 100% probability and an E value less than 1. Highly similar genes (Pons and SheckWes) also had the function DNA helicase.

Feature 60 – Reverse – Stop 40873

Glimmer/GeneMark

What feature number is this? 60

What is the stop site? 40873

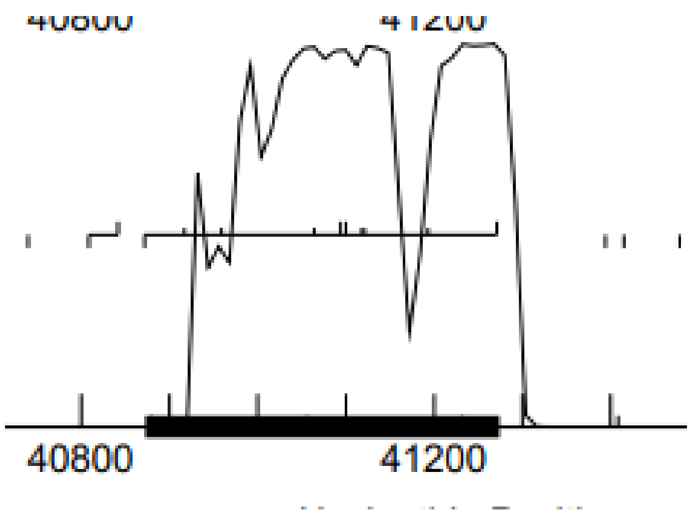
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? Glimmer called the auto-annotated start

What is the autoannotated start? 41274

Gap: ________ or overlap: ___1___ (with gene in front of it) for the autoannotated start

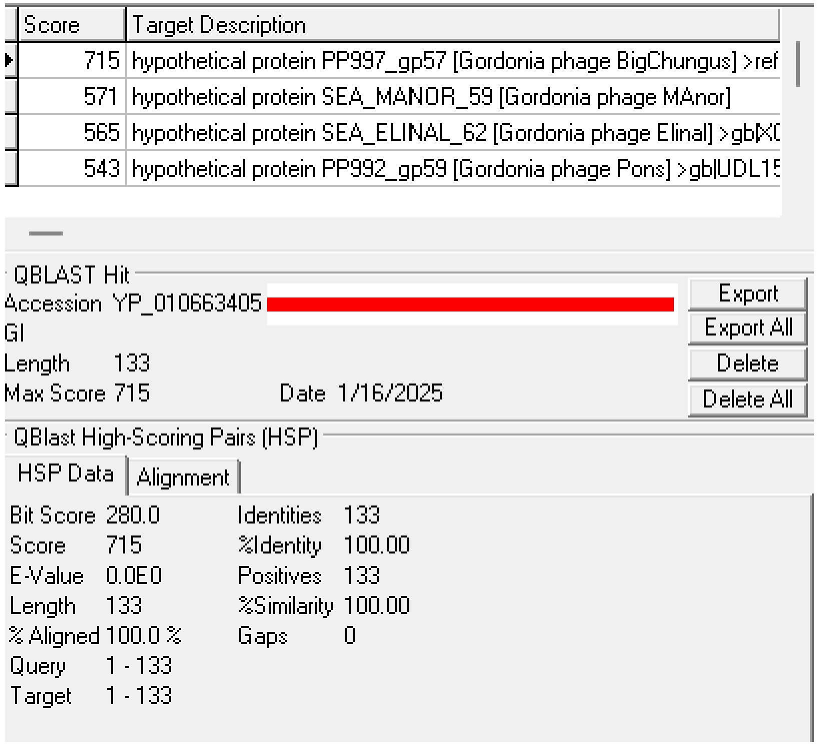
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Is it the only reading frame with cp? Frame 6 was the only one with cp.
- Describe the coding potential… is it strong or is it weak? How do you know? This is strong cp because its height is close to 1.0.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are 6 highly similar genes such as BigChungus, Manor, and Elinal. The first 4 have an E value of 0 and the 5th and 6th highly similar genes have an E value of -42.



Score	Target Description
715	hypothetical protein PP997_gp57 [Gordonia phage BigChungus] >ref
571	hypothetical protein SEA_MANOR_59 [Gordonia phage MAnor]
565	hypothetical protein SEA_ELINAL_62 [Gordonia phage Elinal] >gb|X
543	hypothetical protein PP992_gp59 [Gordonia phage Pons] >gb|UDL15

QBLAST Hit

Accession YP_010663405

GI

Length 133

Max Score 715   Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

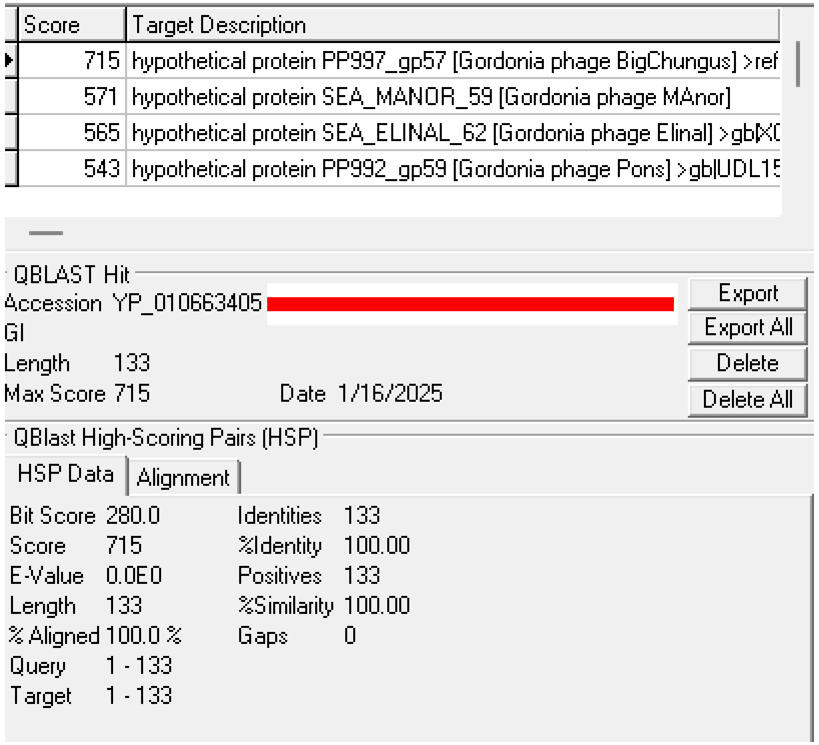
Bit Score	280.0	Identities	133
Score	715	%Identity	100.00
E-Value	0.0E0	Positives	133
Length	133	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 133		
Target	1 - 133		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This function is a gene! Both Glimmer and GeneMark call it a gene, there is strong cp, and there are 6 1:1 alignments with E values of 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There are 6 1:1 alignments with E values smaller than 10^-7. Some similar genes are Elinal, BigChungus, and Manor.



Score	Target Description
715	hypothetical protein PP997_gp57 [Gordonia phage BigChungus] >ref
571	hypothetical protein SEA_MANOR_59 [Gordonia phage MAnor]
565	hypothetical protein SEA_ELINAL_62 [Gordonia phage Elinal] >gb|X
543	hypothetical protein PP992_gp59 [Gordonia phage Pons] >gb|UDL15

QBLAST Hit

Accession YP_010663405

GI

Length 133

Max Score 715

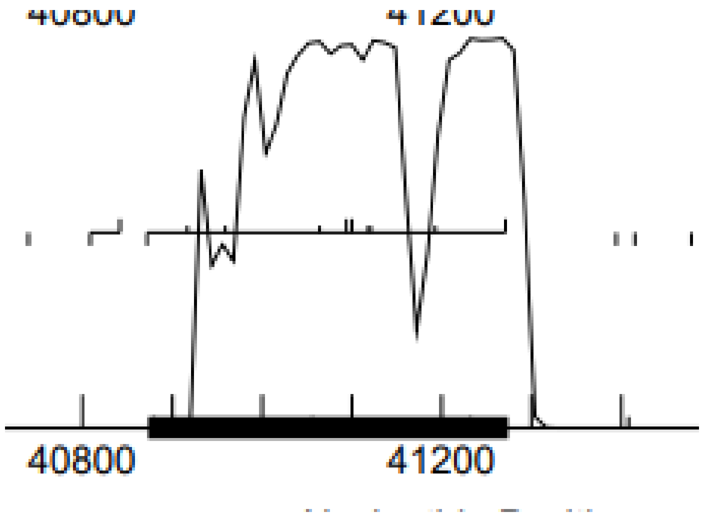
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	280.0	Identities	133
Score	715	%Identity	100.00
E-Value	0.0E0	Positives	133
Length	133	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 133		
Target	1 - 133		

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

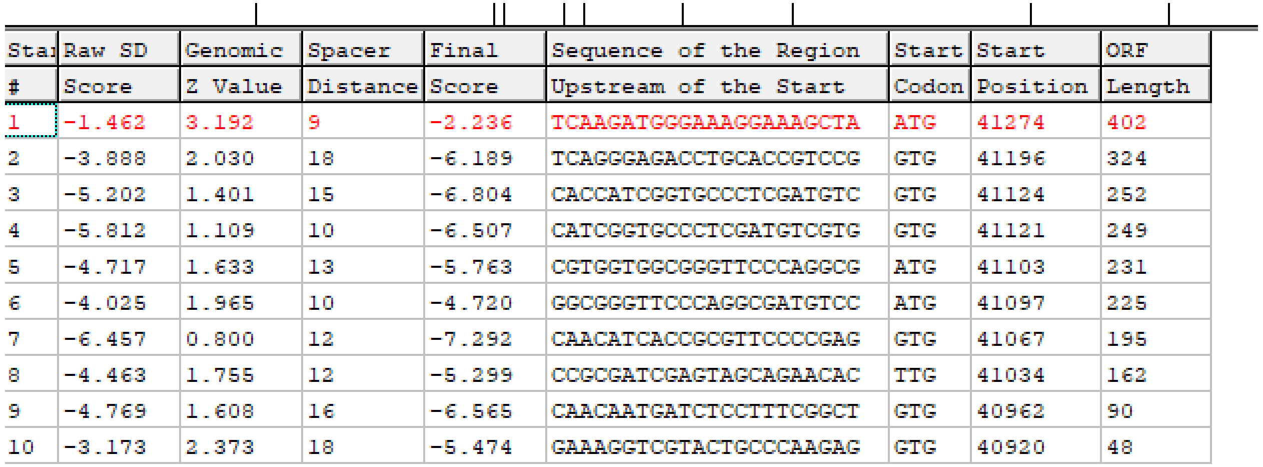
- All cp that can be included is included from 41,274-40,873. There is a decrease in cp at about 41,170 then the cp increases again.



RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- What is the z-value and final score? Z-value: 3.192 FS:-2.236
- How does the RBS compare to that of other available starts? The RBS values for start 41274 are the best RBS value that fall into the ranges we are looking for.

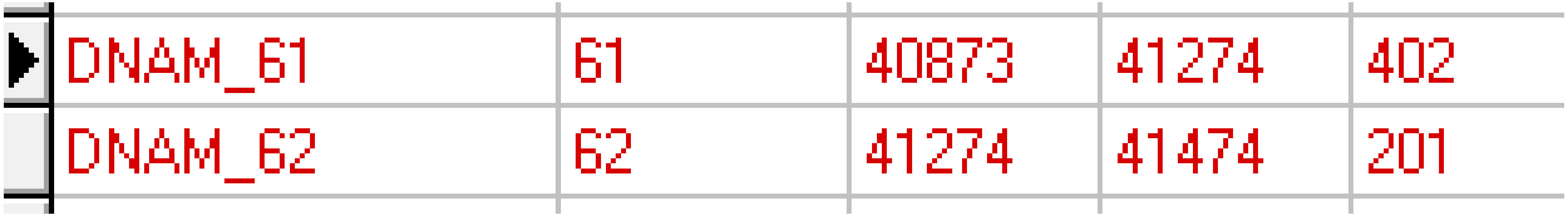
- Screenshot RBS Values here.



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	-1.462	3.192	9	-2.236	TCAAGATGGGAAAGGAAAGCTA	ATG	41274	402
2	-3.888	2.030	18	-6.189	TCAGGGAGACCTGCACCGTCCG	GTG	41196	324
3	-5.202	1.401	15	-6.804	CACCATCGGTGCCCTCGATGTC	GTG	41124	252
4	-5.812	1.109	10	-6.507	CATCGGTGCCCTCGATGTCGTG	GTG	41121	249
5	-4.717	1.633	13	-5.763	CGTGGTGGCGGGTTCCCAGGCG	ATG	41103	231
6	-4.025	1.965	10	-4.720	GGCGGGTTCCCAGGCGATGTCC	ATG	41097	225
7	-6.457	0.800	12	-7.292	CAACATCACCGCGTTCCCCGAG	GTG	41067	195
8	-4.463	1.755	12	-5.299	CCGCGATCGAGTAGCAGAACAC	TTG	41034	162
9	-4.769	1.608	16	-6.565	CAACAATGATCTCCTTTCGGCT	GTG	40962	90
10	-3.173	2.373	18	-5.474	GAAAGGTCGTACTGCCCAAGAG	GTG	40920	48

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start?Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

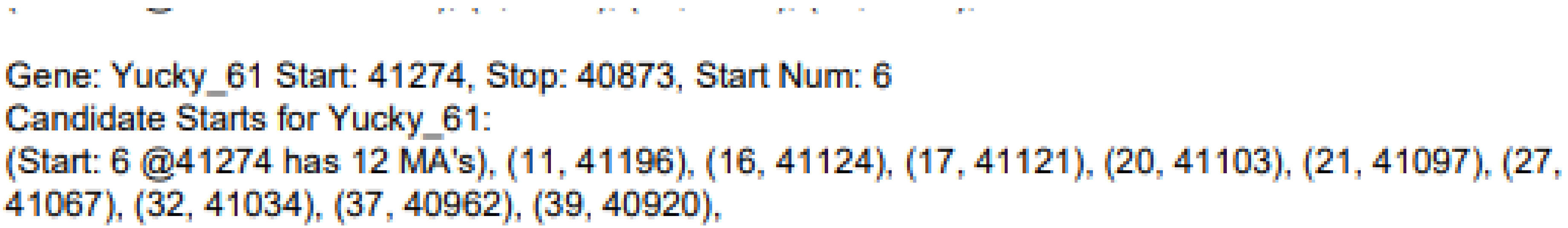
- There is an overlap of 1



▶	DNAM_61	61	40873	41274	402
	DNAM_62	62	41274	41474	201

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- There are 12 MAs for start 42,274. There are no other starts with Manual Annotations.



Gene: Yucky_61 Start: 41274, Stop: 40873, Start Num: 6
Candidate Starts for Yucky_61:
(Start: 6 @41274 has 12 MA's), (11, 41196), (16, 41124), (17, 41121), (20, 41103), (21, 41097), (27, 41067), (32, 41034), (37, 40962), (39, 40920),

Gene 61

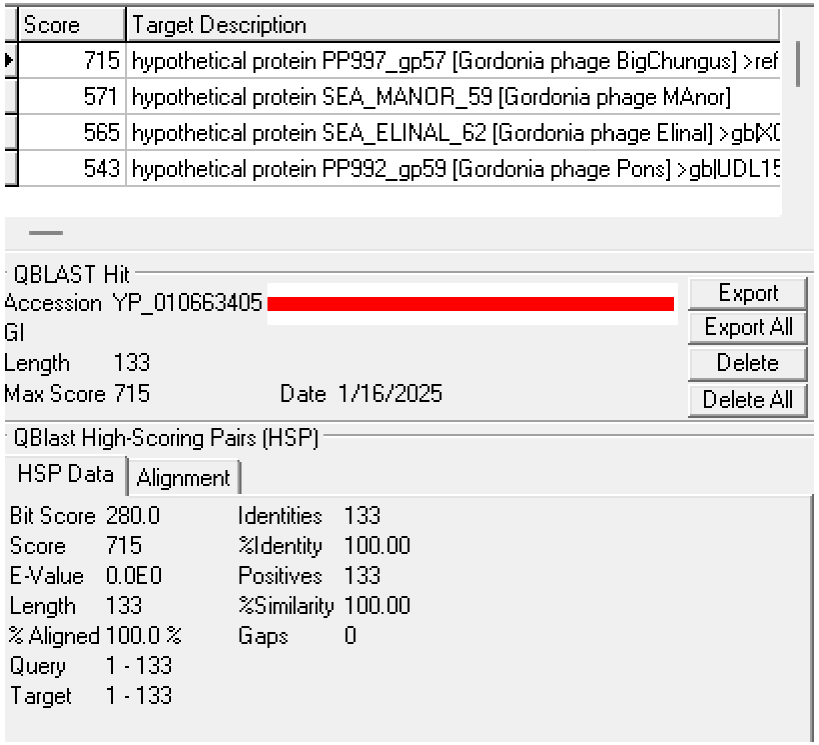
	40,274
GeneMark/Glimmer	Both call start 40,274 a gene
Coding Potential	All cp that can be included is included. Very strong. About 50 nucleotides short
RBS	
Blast	There are 6 1:1 blast alignments with an E value of less than 10^-7
Starterator	There are 12 MAs
Gap/Overlap	Overlap of 1

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is 40,274. Both Glimmer and GeneMark call this the start site, there is strong cp that is included (short about 50 nucleotides), the RBS values are Z-value: 3.192 FS:-2.236, there is an overlap of 1, there are 6 1:1 blast alignments with an E value of less than 10^-7, and there are 12 MAs for start site 40,274.

BLAST function evidence. What assigned functions do other highly similar genes have?

- There are 9 highly similar genes with the function of hypothetical proteins. Such as BigChungus, Elinal, and Manor.



Score	Target Description
715	hypothetical protein PP997_gp57 [Gordonia phage BigChungus] >ref
571	hypothetical protein SEA_MANOR_59 [Gordonia phage MAnor]
565	hypothetical protein SEA_ELINAL_62 [Gordonia phage Elinal] >gb|X
543	hypothetical protein PP992_gp59 [Gordonia phage Pons] >gb|UDL15

QBLAST Hit

Accession YP_010663405

GI

Length 133

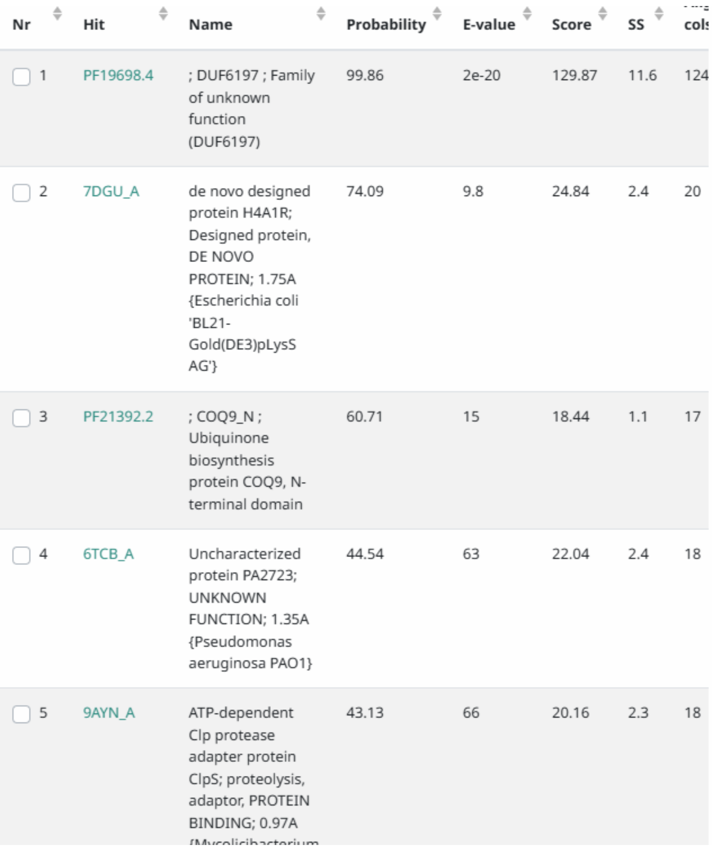
Max Score 715 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	280.0	Identities	133
Score	715	%Identity	100.00
E-Value	0.0E0	Positives	133
Length	133	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 133		
Target	1 - 133		

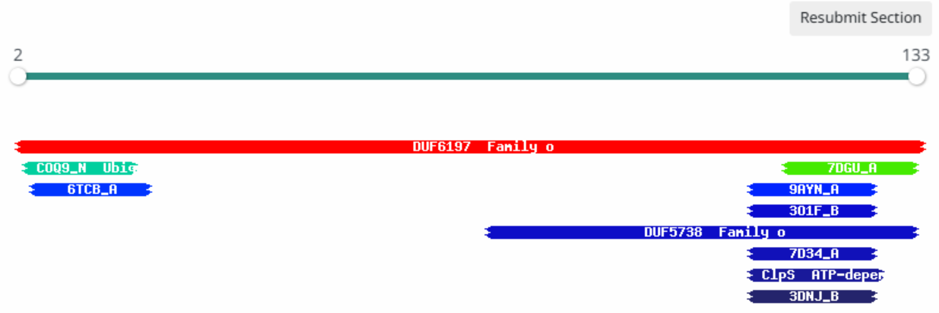
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- Hhpred assigns this to a family of unknown function. This is the only probability above 90%.



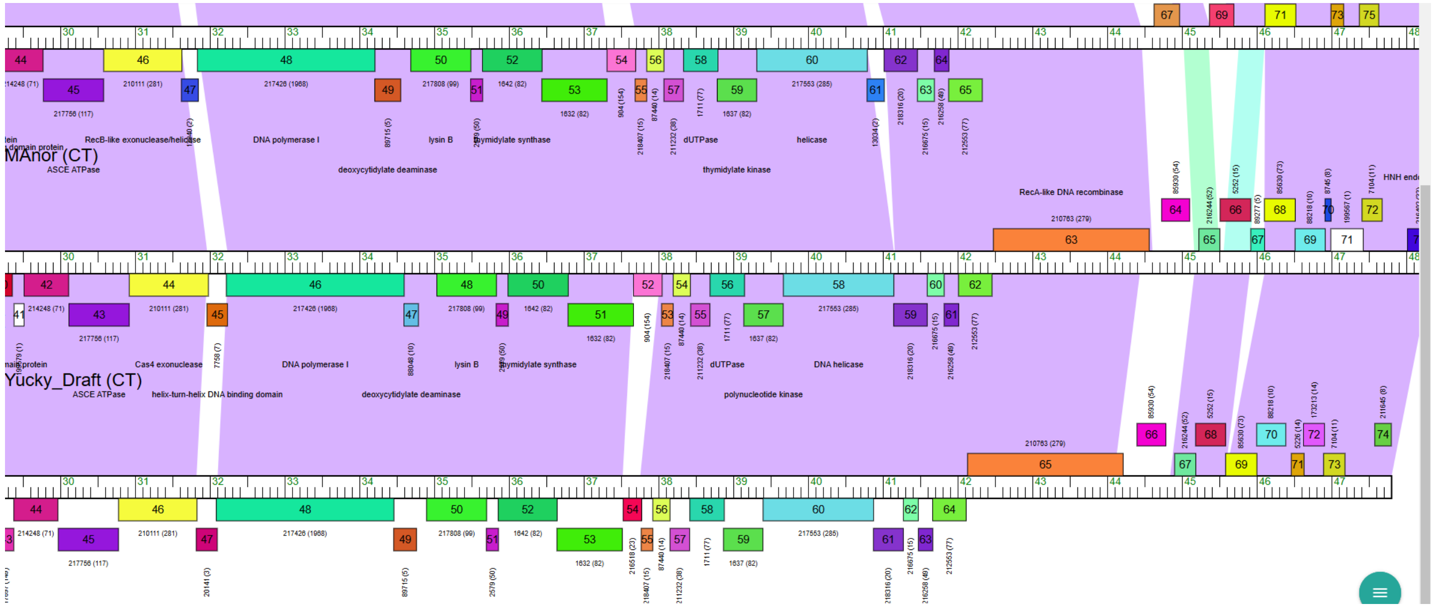
Nr	Hit	Name	Probability	E-value	Score	SS	cols
□ 1	PF19698.4	; DUF6197 ; Family of unknown function (DUF6197)	99.86	2e-20	129.87	11.6	124
□ 2	7DGU_A	de novo designed protein H4A1R; Designed protein, DE NOVO PROTEIN; 1.75A {Escherichia coli 'BL21-Gold(DE3)pLysS AG'}	74.09	9.8	24.84	2.4	20
□ 3	PF21392.2	; COQ9_N ; Ubiquinone biosynthesis protein COQ9, N-terminal domain	60.71	15	18.44	1.1	17
□ 4	6TCB_A	Uncharacterized protein PA2723; UNKNOWN FUNCTION; 1.35A {Pseudomonas aeruginosa PAO1}	44.54	63	22.04	2.4	18
□ 5	9AYN_A	ATP-dependent Clp protease adapter protein ClpS; proteolysis, adaptor, PROTEIN BINDING; 0.97A {Mycolicibacterium	43.13	66	20.16	2.3	18

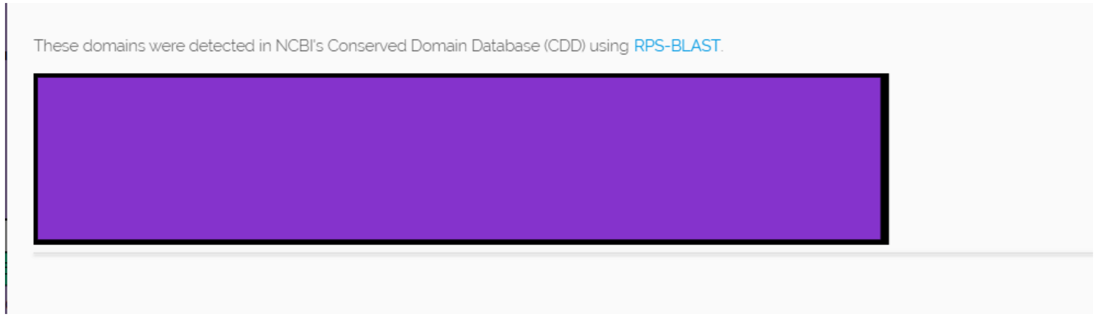
Visualization



Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

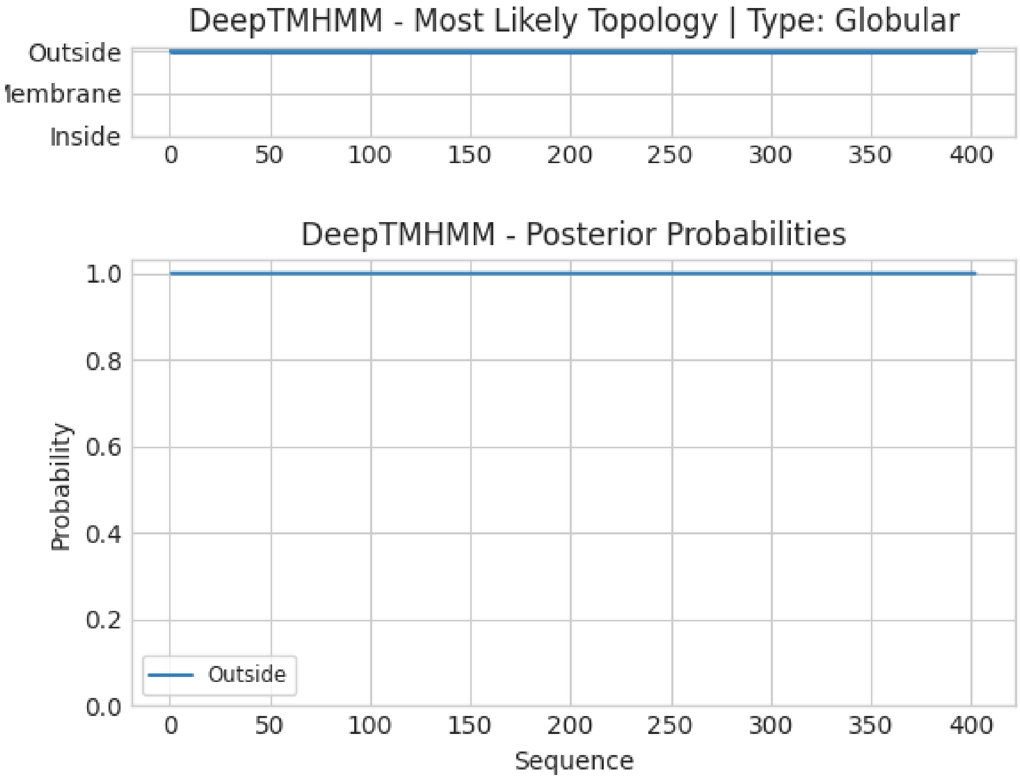
- There are no conserved domains or known functions.




These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- There are no transmembrane domains.



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function of this gene is a hypothetical protein. BLAST calls this a hypothetical protein, Hhpred assigns this to a family of unknow function, Phamerator calls no conserved domains or functions, and TMHMM shows that there are no transmembrane domains.

Feature 61 – Reverse – Stop 41274

Glimmer/GeneMark

What feature number is this? 61

What is the stop site? 41,274

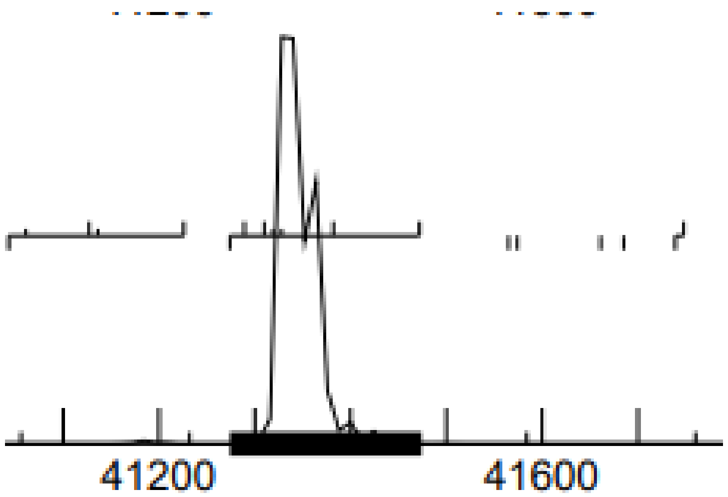
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? Glimmer called the auto-annotated start

What is the autoannotated start? 41,472

Gap: ________ or overlap: _1______ (with gene in front of it) for the autoannotated start

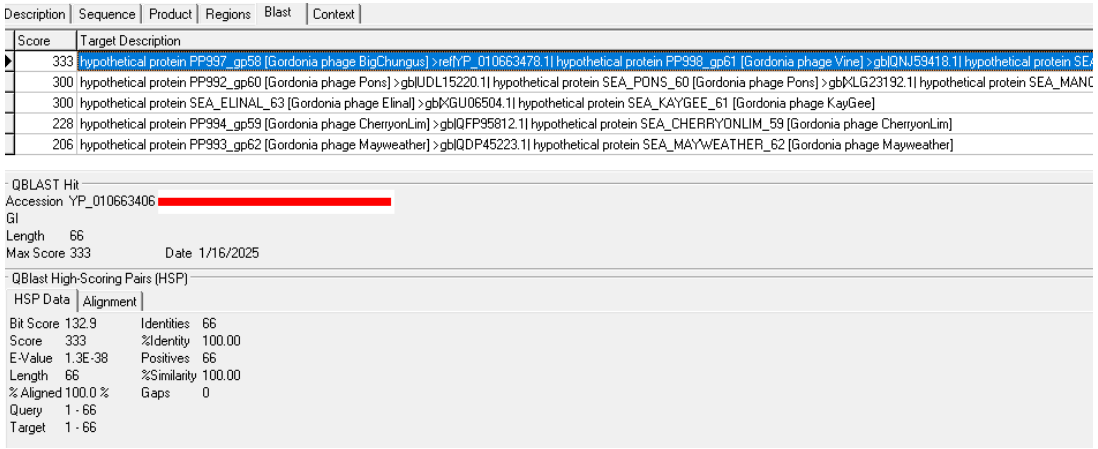
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Is it the only reading frame with cp? Reading frame 5 is the only frame with cp.
- Describe the coding potential… is it strong or is it weak? How do you know? This has strong reading potential as the height is almost 1.0.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are 8 highly similar genes with 1:1 alignments and E values smaller than 10^-7.



Score	Target Description
333	hypothetical protein PP997_gp58 [Gordonia phage BigChungus] >ref|YP_010663478.1| hypothetical protein PP998_gp61 [Gordonia phage Vine] >gb|QNJ59418.1| hypothetical protein SEA
300	hypothetical protein PP992_gp60 [Gordonia phage Pons] >gb|UDL15220.1| hypothetical protein SEA_PONS_60 [Gordonia phage Pons] >gb|XLG23192.1| hypothetical protein SEA_MANC
300	hypothetical protein SEA_ELINAL_63 [Gordonia phage Elinal] >gb|XGU06504.1| hypothetical protein SEA_KAYGEE_61 [Gordonia phage KayGee]
228	hypothetical protein PP994_gp59 [Gordonia phage CherryonLim] >gb|QFP95812.1| hypothetical protein SEA_CHERRYONLIM_59 [Gordonia phage CherryonLim]
206	hypothetical protein PP993_gp62 [Gordonia phage Mayweather] >gb|QDP45223.1| hypothetical protein SEA_MAYWEATHER_62 [Gordonia phage Mayweather]

QBLAST Hit

Accession YP_010663406
GI
Length 66
Max Score 333 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

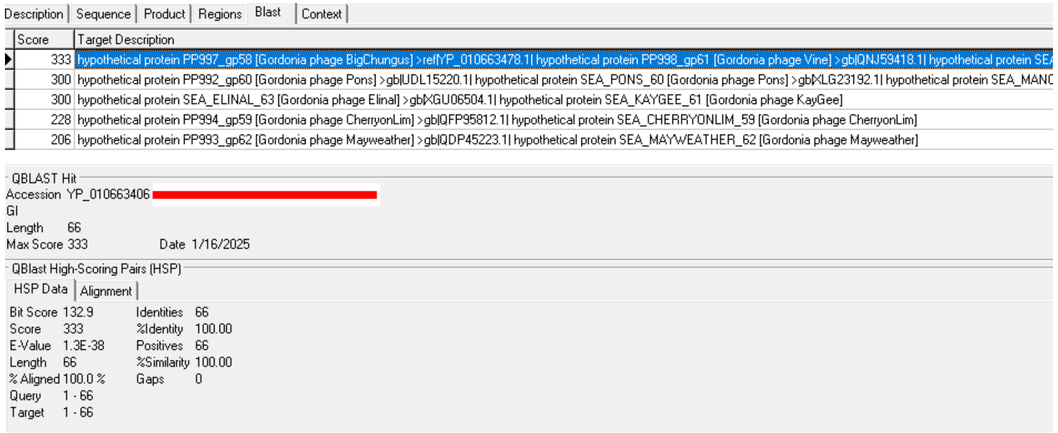
Bit Score	132.9	Identities	66
Score	333	%Identity	100.00
E-Value	1.3E-38	Positives	66
Length	66	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 66		
Target	1 - 66		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes this is a gene because both Glimmer and GeneMark call it a gene, there is strong cp, and there are multiple highly similar genes with 8 1:1 alignments.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There are 8 1:1 alignments for this start. The E values are all less than 10^-7. The highly similar genes include CherryonLim, Mayweather, and ShackWes.



Score	Target Description
333	hypothetical protein PP997_gp58 [Gordonia phage BigChungus] >ref|YP_010663478.1| hypothetical protein PP998_gp61 [Gordonia phage Vine] >gb|QNJ59418.1| hypothetical protein SEA
300	hypothetical protein PP992_gp60 [Gordonia phage Pons] >gb|UDL15220.1| hypothetical protein SEA_PONS_60 [Gordonia phage Pons] >gb|XLG23192.1| hypothetical protein SEA_MANC
300	hypothetical protein SEA_ELINAL_63 [Gordonia phage Elinal] >gb|XGU06504.1| hypothetical protein SEA_KAYGEE_61 [Gordonia phage KayGee]
228	hypothetical protein PP994_gp59 [Gordonia phage CherryonLim] >gb|QFP95812.1| hypothetical protein SEA_CHERRYONLIM_59 [Gordonia phage CherryonLim]
206	hypothetical protein PP993_gp62 [Gordonia phage Mayweather] >gb|QDP45223.1| hypothetical protein SEA_MAYWEATHER_62 [Gordonia phage Mayweather]

QBLAST Hit

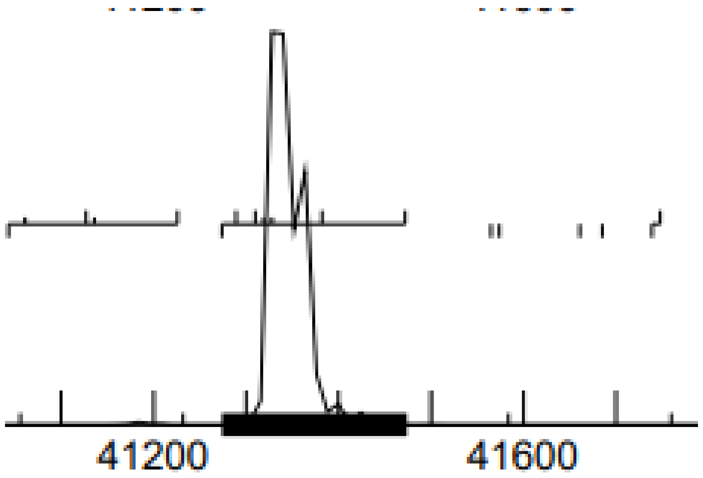
Accession YP_010663406
GI
Length 66
Max Score 333 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	132.9	Identities	66
Score	333	%Identity	100.00
E-Value	1.3E-38	Positives	66
Length	66	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 66		
Target	1 - 66		

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

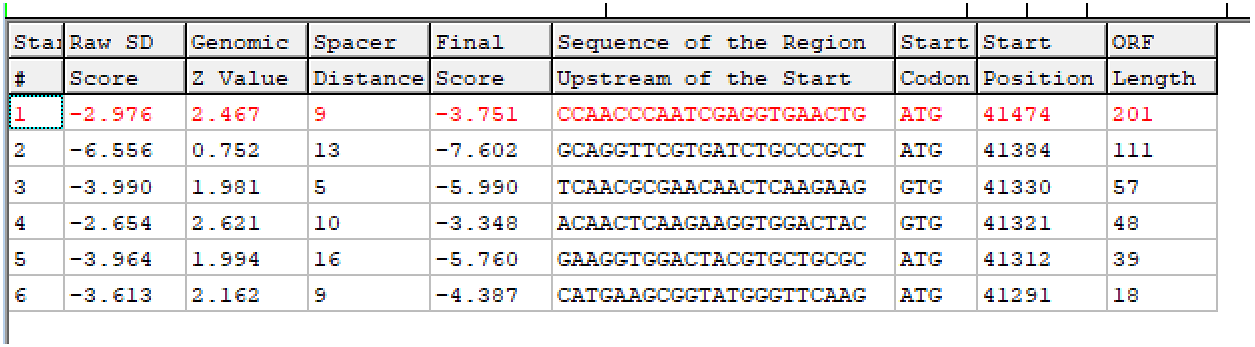
- All cp that can be included but there is no cp from 41,274-41,300 and no cp from 41,400-41,472. It is short about 30 nucleotides on the side it stops and it is short about 70 nucleotides from when it starts.



RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- What is the z-value and final score? Z Value: -2.976 FS: -3.751
- How does the RBS compare to that of other available starts? These scores are within the range we want them to be and are the best out of all the other RBS scores.

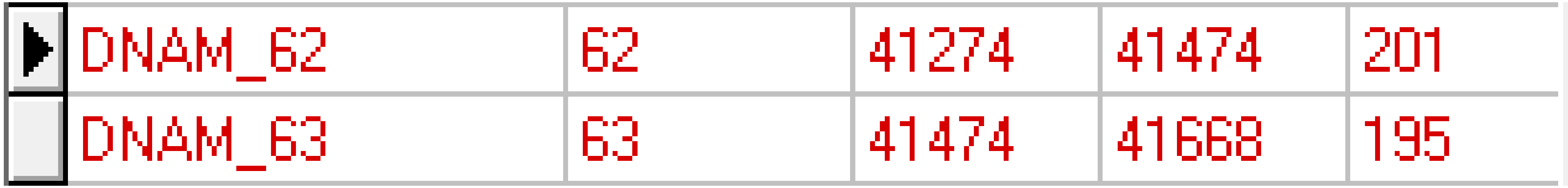
- Screenshot RBS Values here.



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	-2.976	2.467	9	-3.751	CCAACCCAATCGAGGTGAACTG	ATG	41474	201
2	-6.556	0.752	13	-7.602	GCAGGTTCGTGATCTGCCCGCT	ATG	41384	111
3	-3.990	1.981	5	-5.990	TCAACGCGAACAACTCAAGAAG	GTG	41330	57
4	-2.654	2.621	10	-3.348	ACAACTCAAGAAGGTGGACTAC	GTG	41321	48
5	-3.964	1.994	16	-5.760	GAAGGTGGACTACGTGCTGCGC	ATG	41312	39
6	-3.613	2.162	9	-4.387	CATGAAGCGGTATGGGTTCAAG	ATG	41291	18

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

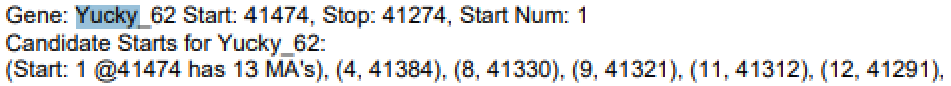
- There is an overlap of 1



▶	DNAM_62	62	41274	41474	201
	DNAM_63	63	41474	41668	195

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- There are 13 MAs for start site 41,474



Gene: Yucky_62 Start: 41474, Stop: 41274, Start Num: 1
Candidate Starts for Yucky_62:
(Start: 1 @41474 has 13 MA's), (4, 41384), (8, 41330), (9, 41321), (11, 41312), (12, 41291),

Gene 62

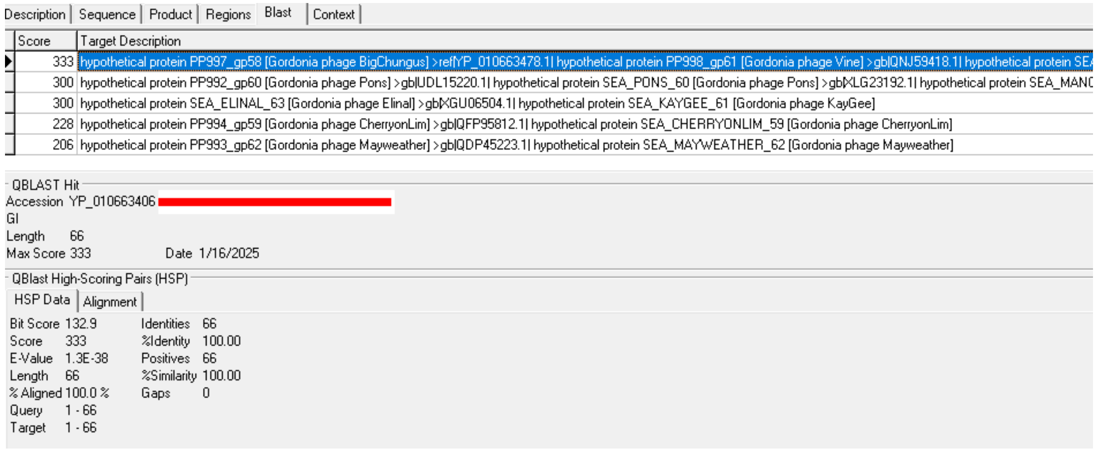
	Start Site 41,472
Glimmer/GeneMark	Both Glimmer and GeneMark call it a Gene
Coding Potential	All cp that can be included is. Short about 100 nucleotides
RBS	Z Value: -2.976 FS: -3.751
Blast	There are 8 1:1 alignments for this start. The E values are all less than 10^-7
Starterator	There are 13 MAs
Gap/Overlap	Overlap of 1

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is 41,472. Both Glimmer and GeneMark call it, all cp that can be included is, Z Value: -2.976 FS: -3.751, there are 8 1:1 alignments for this start, the E values are all less than 10^-7, there are 13 Mas and an overlap of 1

BLAST function evidence. What assigned functions do other highly similar genes have?

- There are 8 highly similar genes assigned with the function of hypothetical proteins



Score	Target Description
333	hypothetical protein PP997_gp58 [Gordonia phage BigChungus] >ref|YP_010663478.1| hypothetical protein PP998_gp61 [Gordonia phage Vine] >gb|QNJ59418.1| hypothetical protein SE
300	hypothetical protein PP992_gp60 [Gordonia phage Pons] >gb|UDL15220.1| hypothetical protein SEA_PONS_60 [Gordonia phage Pons] >gb|XLG23192.1| hypothetical protein SEA_MAN
300	hypothetical protein SEA_ELINAL_63 [Gordonia phage Elinal] >gb|XGU06504.1| hypothetical protein SEA_KAYGEE_61 [Gordonia phage KayGee]
228	hypothetical protein PP994_gp59 [Gordonia phage CherryonLim] >gb|QFP95812.1| hypothetical protein SEA_CHERRYONLIM_59 [Gordonia phage CherryonLim]
206	hypothetical protein PP993_gp62 [Gordonia phage Mayweather] >gb|QDP45223.1| hypothetical protein SEA_MAYWEATHER_62 [Gordonia phage Mayweather]

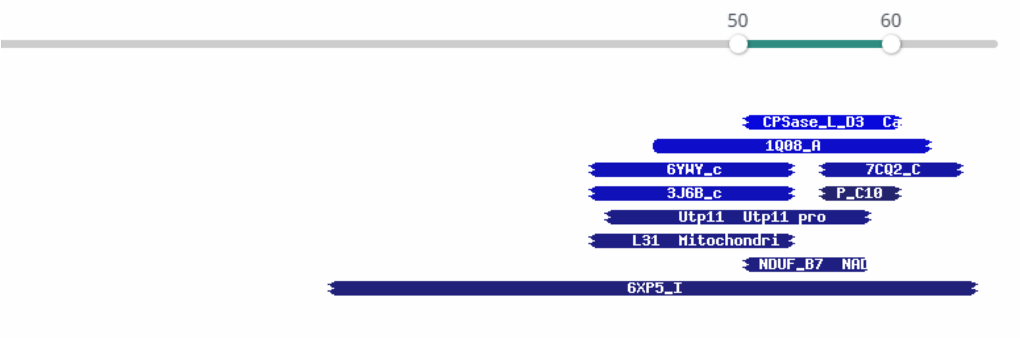
QBLAST Hit
Accession YP_010663406
GI
Length 66
Max Score 333 Date 1/16/2025

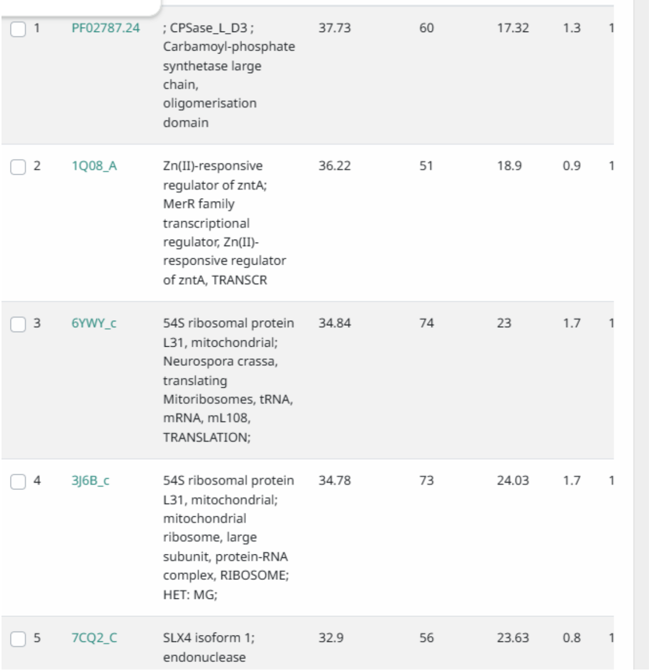
QBlast High-Scoring Pairs (HSP)

Bit Score	132.9	Identities	66
Score	333	%Identity	100.00
E-Value	1.3E-38	Positives	66
Length	66	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 66		
Target	1 - 66		

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- There are no probabilities over 90% so this evidence is conclusive.

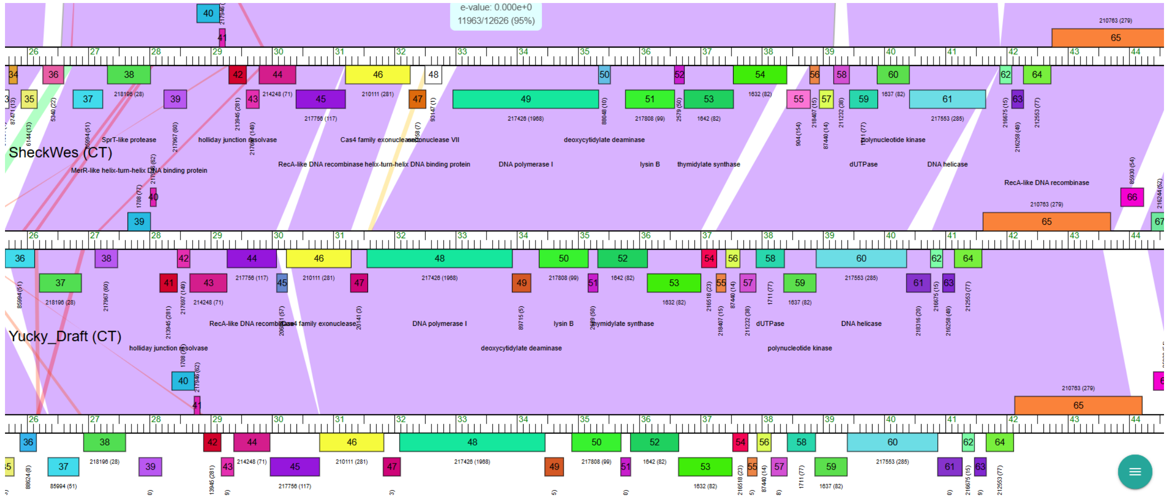


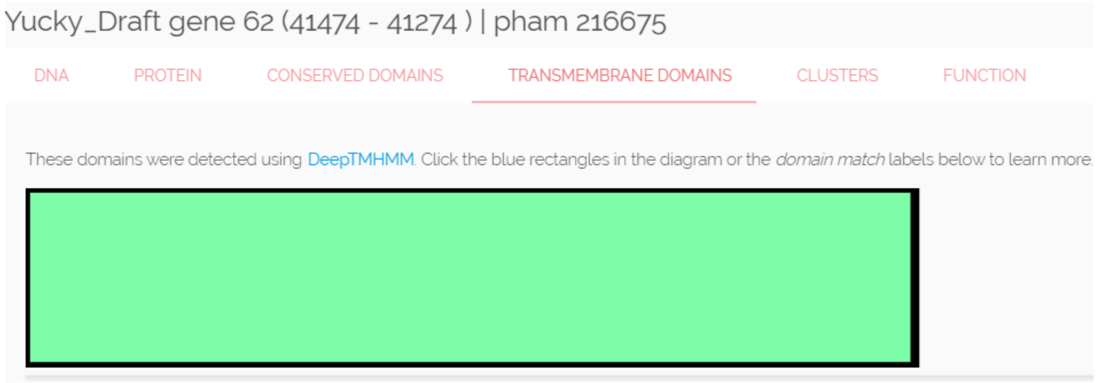


□ 1	PF02787.24	; CPSase_L_D3 ; Carbamoyl-phosphate synthetase large chain, oligomerisation domain	37.73	60	17.32	1.3	1
□ 2	1Q08_A	Zn(II)-responsive regulator of zntA; MerR family transcriptional regulator, Zn(II)-responsive regulator of zntA, TRANSCR	36.22	51	18.9	0.9	1
□ 3	6YWY_c	54S ribosomal protein L31, mitochondrial; Neurospora crassa, translating Mitoribosomes, tRNA, mRNA, mL108, TRANSLATION;	34.84	74	23	1.7	1
□ 4	3J6B_c	54S ribosomal protein L31, mitochondrial; mitochondrial ribosome, large subunit, protein-RNA complex, RIBOSOME; HET: MG;	34.78	73	24.03	1.7	1
□ 5	7CQ2_C	SLX4 isoform 1; endonuclease	32.9	56	23.63	0.8	1

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- There are no conserved domains and no known functions.



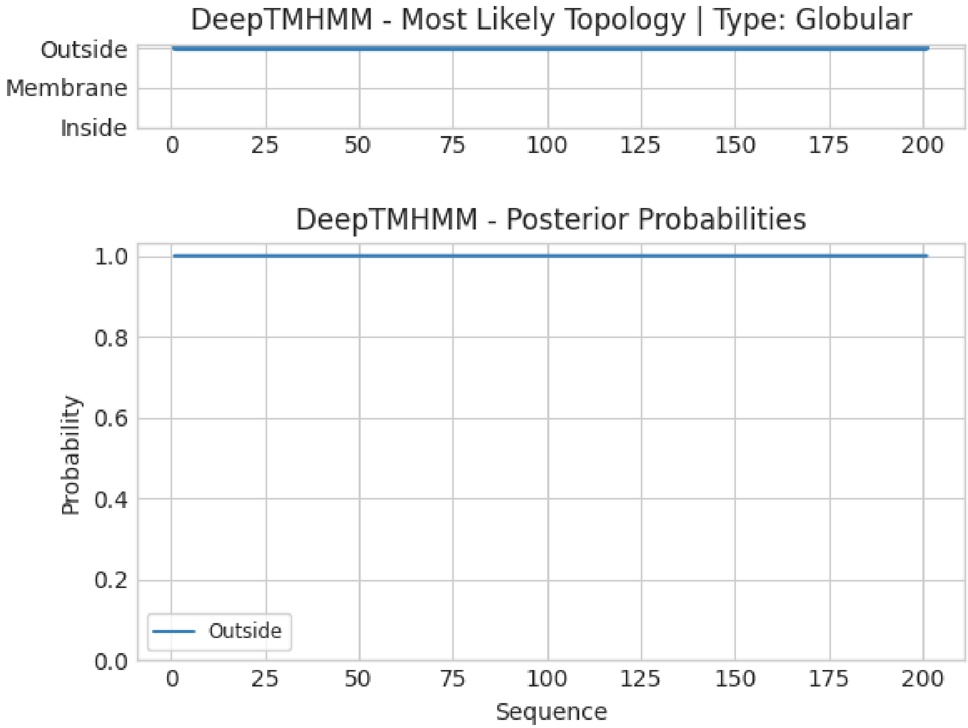

Yucky_Draft gene 62 (41474 - 41274) | pham 216675

DNA PROTEIN CONSERVED DOMAINS TRANSMEMBRANE DOMAINS CLUSTERS FUNCTION

These domains were detected using DeepTMHMM. Click the blue rectangles in the diagram or the *domain match* labels below to learn more.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- There are no transmembrane domains



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function of this protein is a hypothetical protein. Blast calls it a hypothetical protein, there is no conclusive evidence form Hhpre, there are no known functions or conserved domains, and the gene has no transmembrane domains.

Feature 62 – Reverse – Stop 41474

Glimmer/GeneMark

What feature number is this? 62

What is the stop site? 41,474

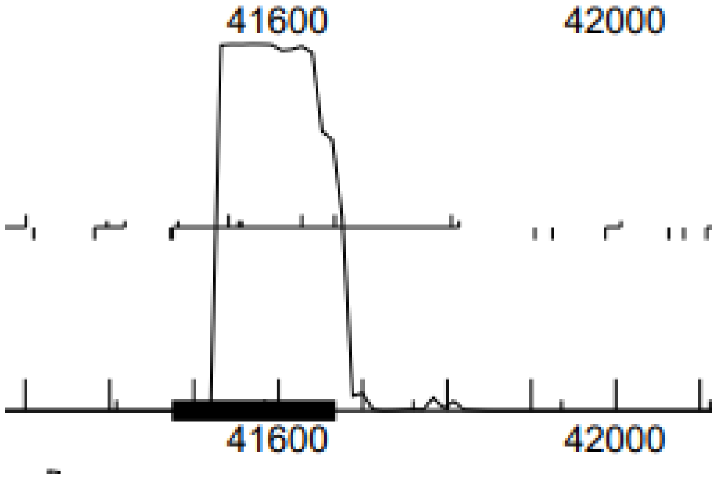
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? Glimmer called the auto-annotated start

What is the autoannotated start? 41,668

Gap: ________ or overlap: __4_____ (with gene in front of it) for the autoannotated start

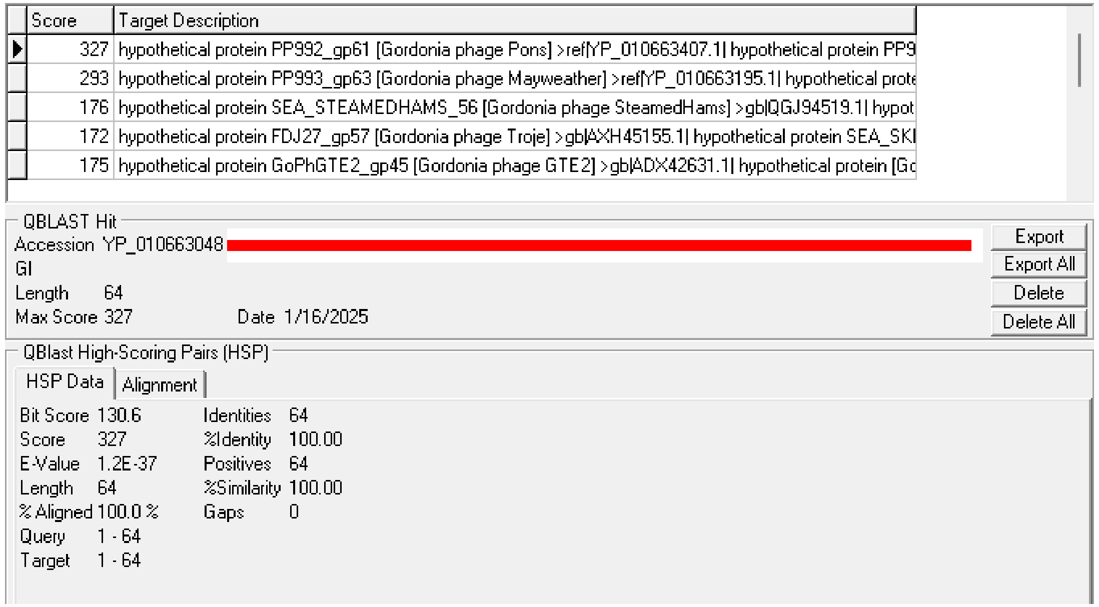
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Is it the only reading frame with cp? Yes, this is the only reading frame with cp.
- Describe the coding potential… is it strong or is it weak? How do you know? This cp is strong because the height of it is mostly a 1.0.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are 2 highly similar genes with 1:1 alignments and E-values less than 10^-7.



Score	Target Description
327	hypothetical protein PP992_gp61 [Gordonia phage Pons] >ref|YP_010663407.1| hypothetical protein PP9
293	hypothetical protein PP993_gp63 [Gordonia phage Mayweather] >ref|YP_010663195.1| hypothetical prote
176	hypothetical protein SEA_STEAMEDHAMS_56 [Gordonia phage SteamedHams] >gb|QGJ94519.1| hypot
172	hypothetical protein FDJ27_gp57 [Gordonia phage Troje] >gb|AXH45155.1| hypothetical protein SEA_SKI
175	hypothetical protein GoPhGTE2_gp45 [Gordonia phage GTE2] >gb|ADX42631.1| hypothetical protein [Go

QBLAST Hit
Accession YP_010663048
GI
Length 64
Max Score 327
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	130.6	Identities	64
Score	327	%Identity	100.00
E-Value	1.2E-37	Positives	64
Length	64	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 64		
Target	1 - 64		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

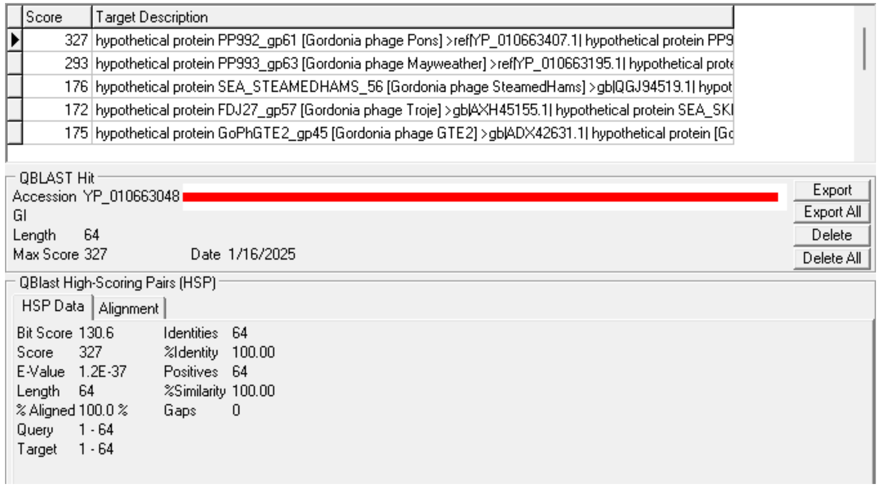
- Is there more than one feature called in this coding region? Yes, this feature is a gene as both Glimmer and GeneMark call it, there is cp, and there are highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

There are 2 1:1 alignments with E values less than 10^-7 for start 41,668.

For start 41,806 there are 0 1:1 alignments but there are 12 highly similar genes.

For start 41, 815 there are 0 1:1 alignments but there are 12 highly similar genes.



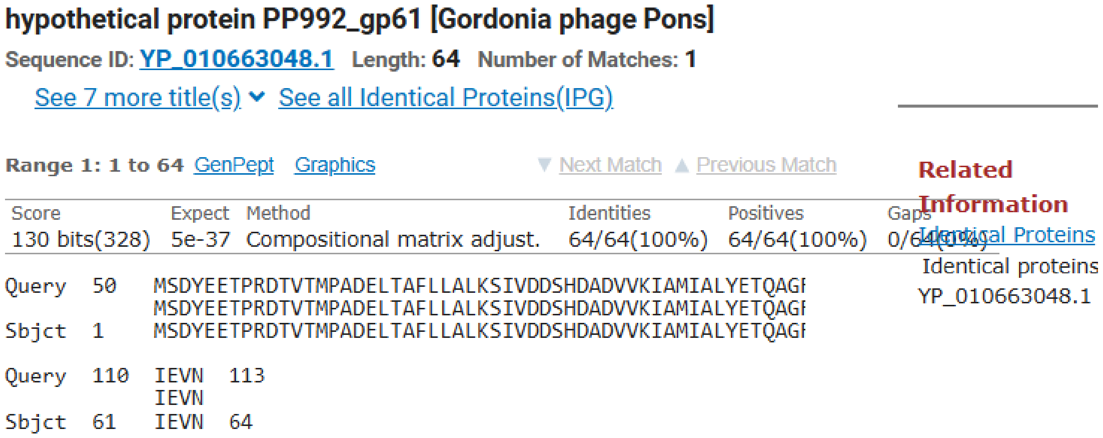
Score	Target Description
327	hypothetical protein PP992_gp61 [Gordonia phage Pons] >ref|YP_010663407.1| hypothetical protein PP9
293	hypothetical protein PP993_gp63 [Gordonia phage Mayweather] >ref|YP_010663195.1| hypothetical prote
176	hypothetical protein SEA_STEAMEDHAMS_56 [Gordonia phage SteamedHams] >gb|QGJ94519.1| hypot
172	hypothetical protein FDJ27_gp57 [Gordonia phage Troje] >gb|AXH45155.1| hypothetical protein SEA_SKI
175	hypothetical protein GoPhGTE2_gp45 [Gordonia phage GTE2] >gb|ADX42631.1| hypothetical protein [Gc

QBLAST Hit

Accession YP_010663048
GI
Length 64
Max Score 327 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	130.6	Identities	64
Score	327	%Identity	100.00
E-Value	1.2E-37	Positives	64
Length	64	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 64		
Target	1 - 64		



hypothetical protein PP992_gp61 [Gordonia phage Pons]

Sequence ID: YP_010663048.1 Length: 64 Number of Matches: 1

See 7 more title(s)
See all Identical Proteins(IPG)

Range 1: 1 to 64 GenPept Graphics

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
130 bits(328)	5e-37	Compositional matrix adjust.	64/64(100%)	64/64(100%)	0/64(0%)

```
Query  50   MSDYEETPRDTVTMPADELTAFLLALKSIVDDSHDADVVKIAMIALYETQAGF
            MSDYEETPRDTVTMPADELTAFLLALKSIVDDSHDADVVKIAMIALYETQAGF
Sbjct  1    MSDYEETPRDTVTMPADELTAFLLALKSIVDDSHDADVVKIAMIALYETQAGF

Query  110  IEVN  113
            IEVN
Sbjct  61   IEVN  64
```

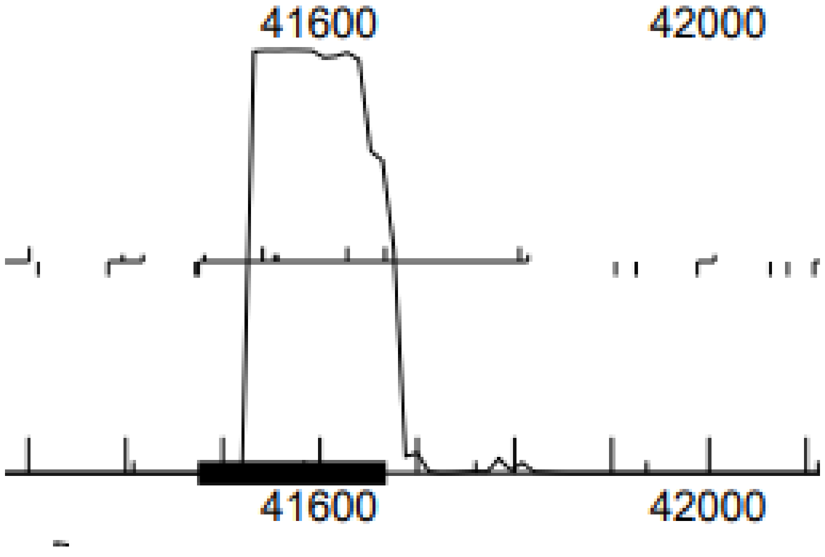
Related Information

Identical Proteins

Identical proteins YP_010663048.1

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

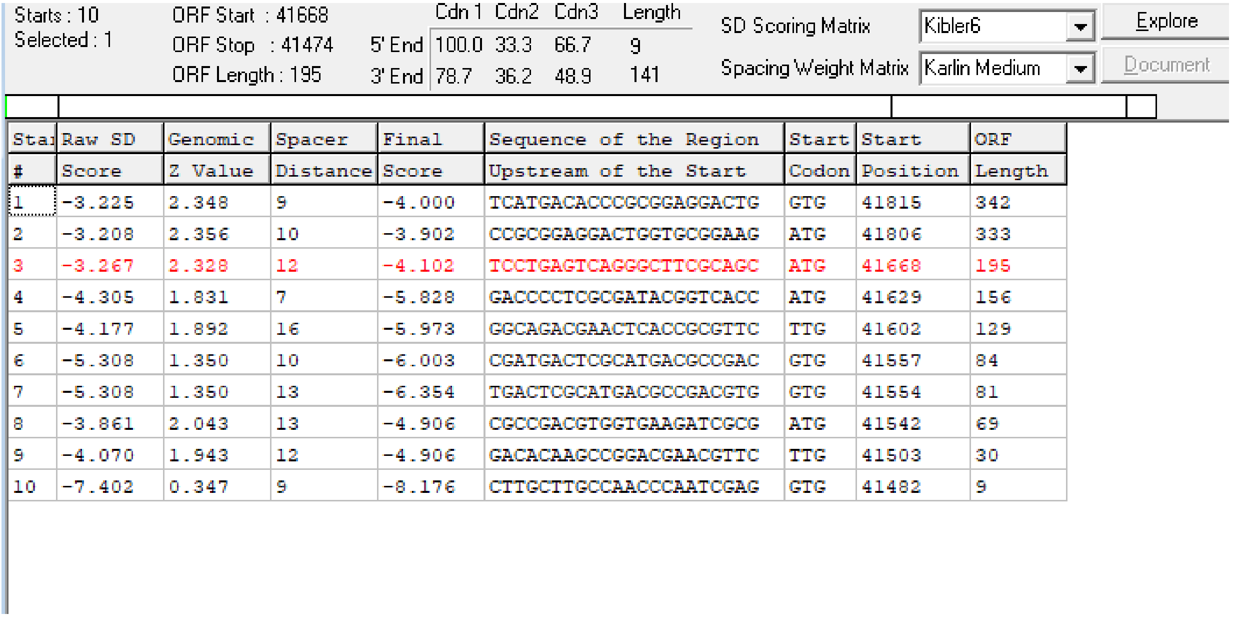
- All cp that can be included is included for start 41,668. There is no cp from 41,474-41,510. Cp is short about 150 nucleotides.
- All cp that can be included is included for starts 41,806 and 41,815 but there is very little to no cp from about 41,700-41,815.



RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- What is the z-value and final score? Z-score: 2.328 FS:-4.102
- How does the RBS compare to that of other available starts? Which start is favored based on RBS values? Other starts like 41,806 have a Z-score: 2.356 and a FS: -3.902 and start 41,815 have a Z-score: 2.348 and FS: -4.00.

- Screenshot RBS Values here.



Starts : 10
Selected : 1

ORF Start : 41668
ORF Stop : 41474
ORF Length : 195

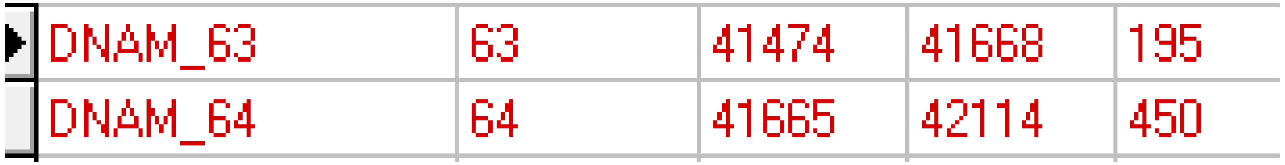
	Cdn 1	Cdn2	Cdn3	Length
5' End	100.0	33.3	66.7	9
3' End	78.7	36.2	48.9	141

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-3.225	2.348	9	-4.000	TCATGACACCCGCGGAGGACTG	GTG	41815	342
2	-3.208	2.356	10	-3.902	CCGCGGAGGACTGGTGCGGAAG	ATG	41806	333
3	-3.267	2.328	12	-4.102	TCCTGAGTCAGGGCTTCGCAGC	ATG	41668	195
4	-4.305	1.831	7	-5.828	GACCCCTCGCGATACGGTCACC	ATG	41629	156
5	-4.177	1.892	16	-5.973	GGCAGACGAACTCACCGCGTTC	TTG	41602	129
6	-5.308	1.350	10	-6.003	CGATGACTCGCATGACGCCGAC	GTG	41557	84
7	-5.308	1.350	13	-6.354	TGACTCGCATGACGCCGACGTG	GTG	41554	81
8	-3.861	2.043	13	-4.906	CGCCGACGTGGTGAAGATCGCG	ATG	41542	69
9	-4.070	1.943	12	-4.906	GACACAAGCCGGACGAACGTTC	TTG	41503	30
10	-7.402	0.347	9	-8.176	CTTGCTTGCCAACCCAATCGAG	GTG	41482	9

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

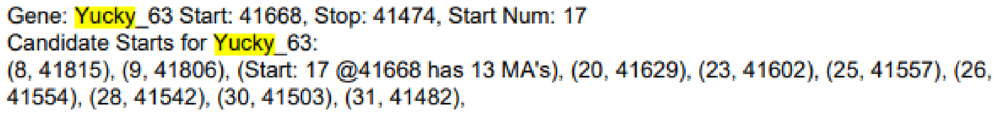
- There is an overlap of 4 for start 41,668
- There is an overlap of 151 for start 41,815
- There is an overlap of 142 for start 41,806



DNAM_63	63	41474	41668	195
DNAM_64	64	41665	42114	450

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- There are 13 MAs for start 41,668 and no MAs for starts 41,815 or 41,806



Gene: Yucky_63 Start: 41668, Stop: 41474, Start Num: 17
Candidate Starts for Yucky_63:
(8, 41815), (9, 41806), (Start: 17 @41668 has 13 MA's), (20, 41629), (23, 41602), (25, 41557), (26, 41554), (28, 41542), (30, 41503), (31, 41482),

Gene 63

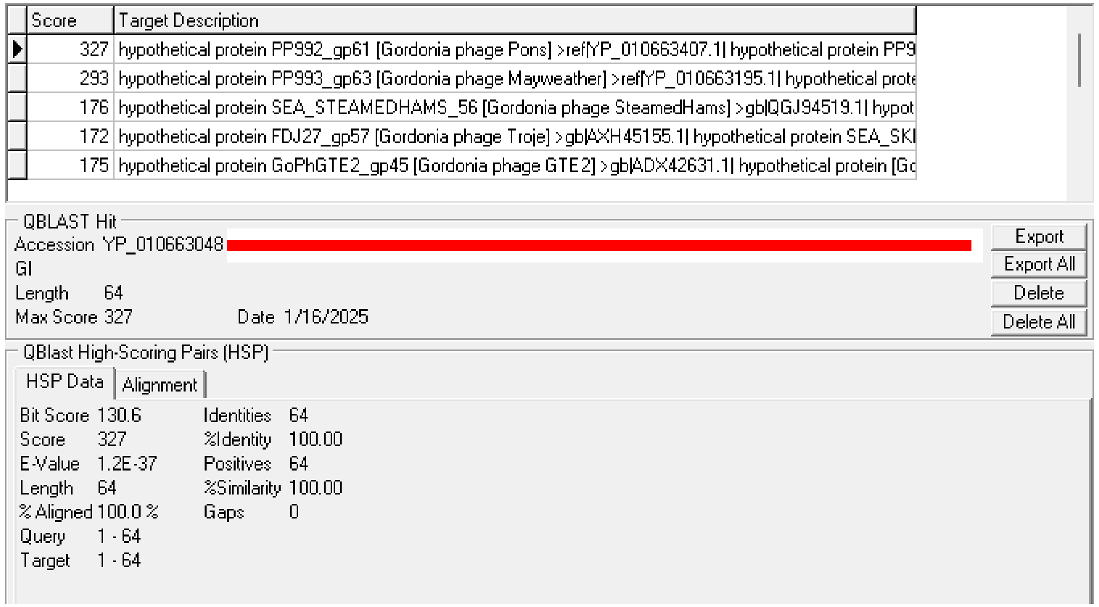
	41,668	41,815	41,506
GeneMark/Glimmer	Glimmer/GeneMark call this the start		
Coding Potential	All cp that can be included is. Short 150 nucelotides from the ending	All cp that can be included is. Little to no cp towards the start	All cp that can be included is. Little to no cp towards the start
RBS	Z-score: 2.328 FS:-4.102	Z-score: 2.356 a FS: -3.902	Z-score: 2.348 FS: -4.00.
Blast	2 1:1 alignments	0 1:1 alignments	0 1:1 alignments
Starterator	13 MAs	0 MAs	0 MAs
Gap/Overlap	Gap of 4	Overlap of 151	Overlap of 142

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- Start site is 41,668. Both Glimmer/GeneMark call this as the start site, all cp that can be included is, 2 1:1 alignments, 13 MAs, and a gap of 4. The RBS Values were best for start site 41,815 Z-score: 2.356 a FS: -3.902 compared to start 41,668 Z-score: 2.328 FS:-4.102

BLAST function evidence. What assigned functions do other highly similar genes have?

- There are 9 highly functional genes assigned the function of hypothetical protein.



Score	Target Description
327	hypothetical protein PP992_gp61 [Gordonia phage Pons] >ref|YP_010663407.1| hypothetical protein PP9
293	hypothetical protein PP993_gp63 [Gordonia phage Mayweather] >ref|YP_010663195.1| hypothetical prote
176	hypothetical protein SEA_STEAMEDHAMS_56 [Gordonia phage SteamedHams] >gb|QGJ94519.1| hypot
172	hypothetical protein FDJ27_gp57 [Gordonia phage Troje] >gb|AXH45155.1| hypothetical protein SEA_SKI
175	hypothetical protein GoPhGTE2_gp45 [Gordonia phage GTE2] >gb|ADX42631.1| hypothetical protein [Go

QBLAST Hit

Accession YP_010663048

GI

Length 64

Max Score 327

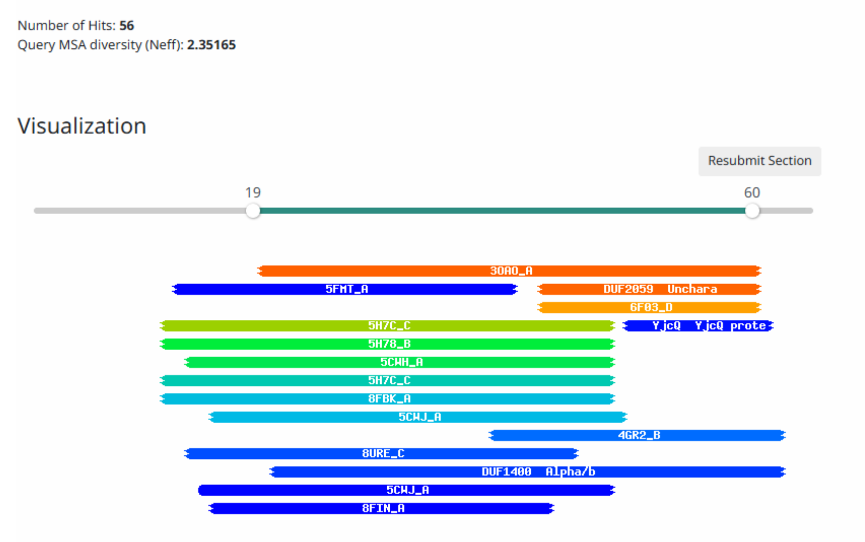
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	130.6	Identities	64
Score	327	%Identity	100.00
E-Value	1.2E-37	Positives	64
Length	64	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 64		
Target	1 - 64		

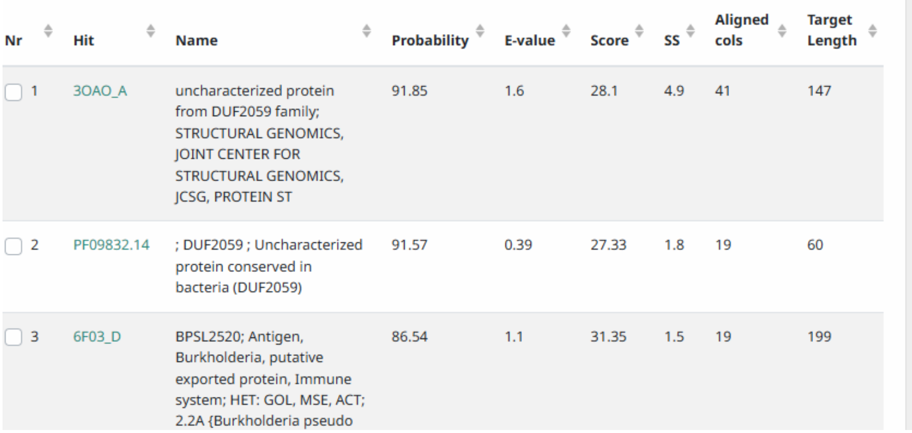
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- The two genes with probabilities over 90% are both assigned to the uncharacterized protein family DUF2059 meaning they are hypothetical proteins.



Number of Hits: **56**
Query MSA diversity (Neff): **2.35165**

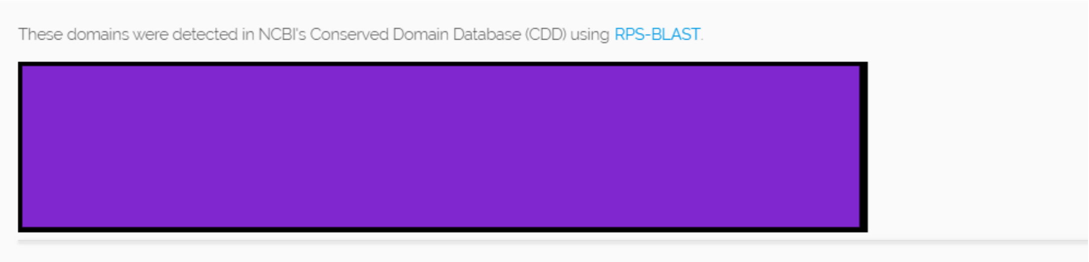
Visualization



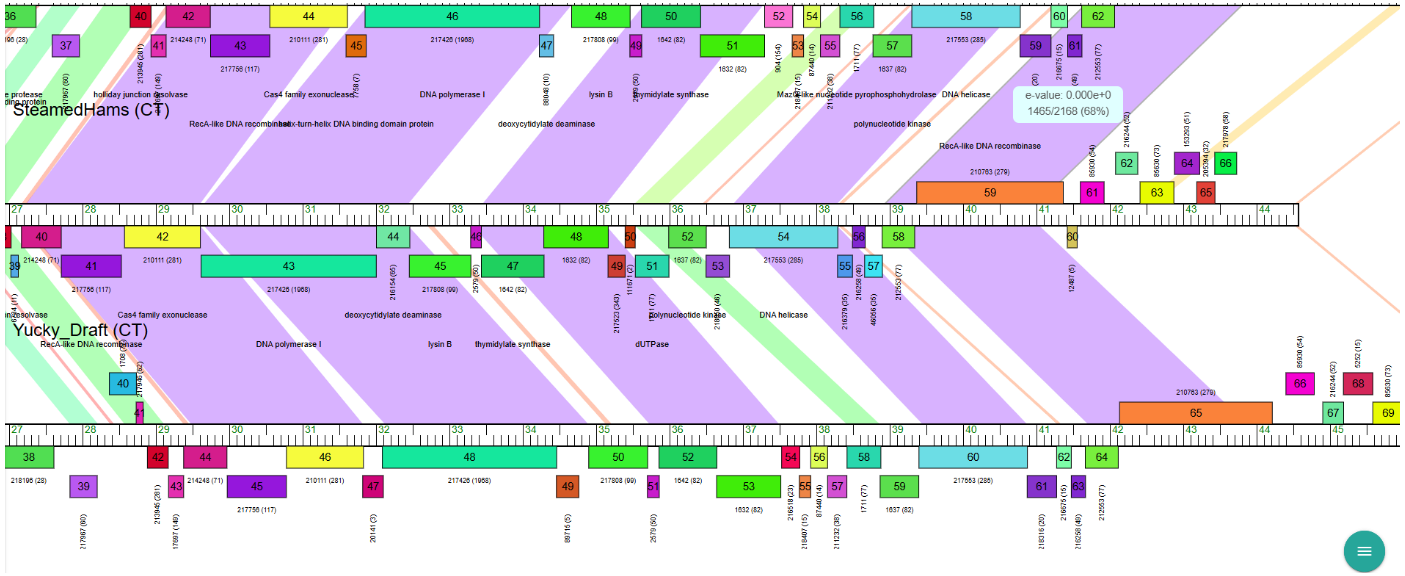
Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	3OAO_A	uncharacterized protein from DUF2059 family; STRUCTURAL GENOMICS, JOINT CENTER FOR STRUCTURAL GENOMICS, JCSG, PROTEIN ST	91.85	1.6	28.1	4.9	41	147
□ 2	PF09832.14	; DUF2059 ; Uncharacterized protein conserved in bacteria (DUF2059)	91.57	0.39	27.33	1.8	19	60
□ 3	6F03_D	BPSL2520; Antigen, Burkholderia, putative exported protein, Immune system; HET: GOL, MSE, ACT; 2.2A {Burkholderia pseudo	86.54	1.1	31.35	1.5	19	199

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- There are no conserved domains or known functions

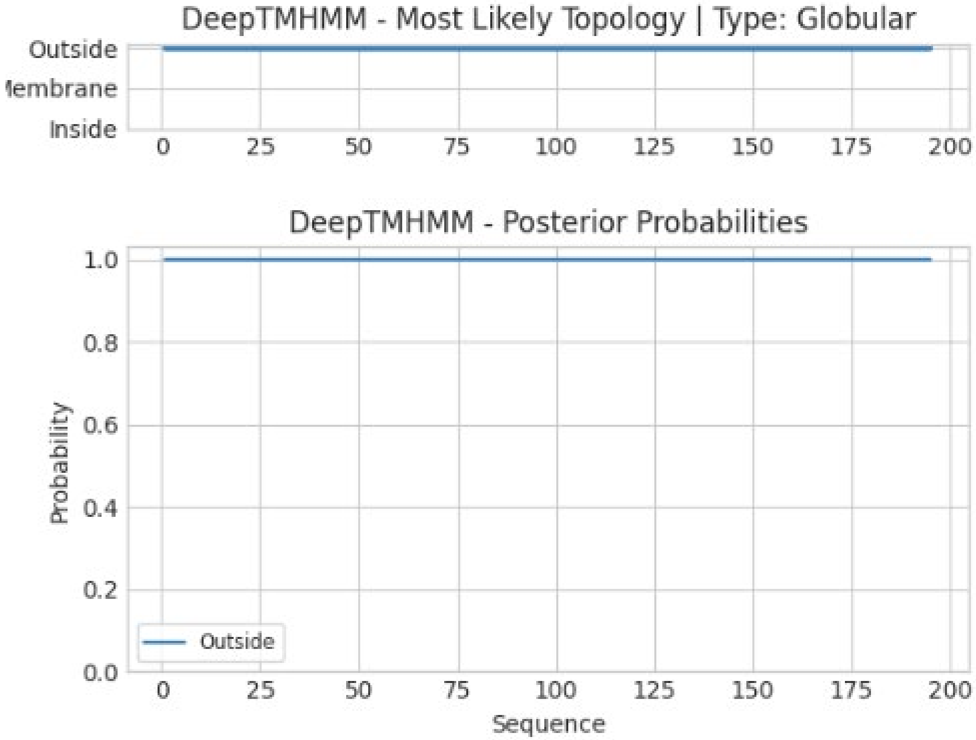


These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.



Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- The gene has no transmembrane domains.



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- This is a hypothetical protein. There are no conserved domains or functions, Hhpred assigns the two most similar genes to the uncharacterized, BLAST assigns the most highly similar genes as hypothetical proteins, and the gene has no transmembrane domains.

Feature 63 – Reverse – Stop 41665

Glimmer/GeneMark

What feature number is this? 63

What is the stop site? 41,665

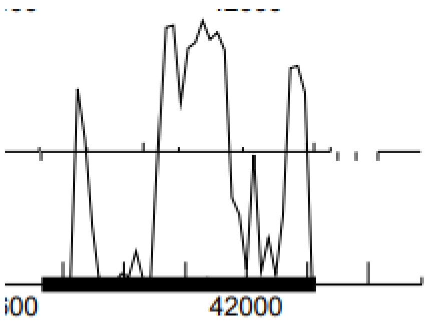
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? Glimmer and GeneMark called this the auto-annotated start

What is the autoannotated start? 42,114

Gap: __17_______ or overlap: ________ (with gene in front of it) for the autoannotated start

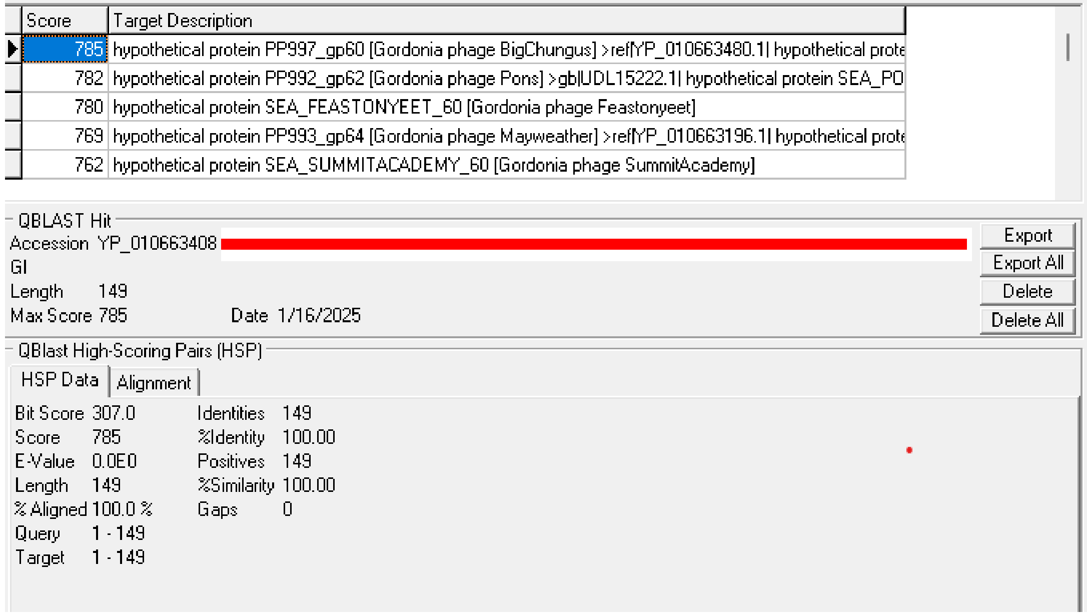
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak?   How do you know?

- Is it the only reading frame with cp? Yes this is the only reading frame with cp
- Describe the coding potential… is it strong or is it weak?   How do you know? The cp is strong as it has large peaks that reach a height of 1.0



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are more than 10 highly similar genes. There are 5 1:1 alignments with E-Values of 0



Score	Target Description
785	hypothetical protein PP997_gp60 [Gordonia phage BigChungus] >ref|YP_010663480.1| hypothetical prote
782	hypothetical protein PP992_gp62 [Gordonia phage Pons] >gb|UDL15222.1| hypothetical protein SEA_PO
780	hypothetical protein SEA_FEASTONYEET_60 [Gordonia phage Feastonyeet]
769	hypothetical protein PP993_gp64 [Gordonia phage Mayweather] >ref|YP_010663196.1| hypothetical prote
762	hypothetical protein SEA_SUMMITACADEMY_60 [Gordonia phage SummitAcademy]

QBLAST Hit

Accession YP_010663408

GI

Length 149

Max Score 785

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 307.0

Score 785

E-Value 0.0E0

Length 149

% Aligned 100.0 %

Query 1 - 149

Target 1 - 149

Identities 149

%Identity 100.00

Positives 149

%Similarity 100.00

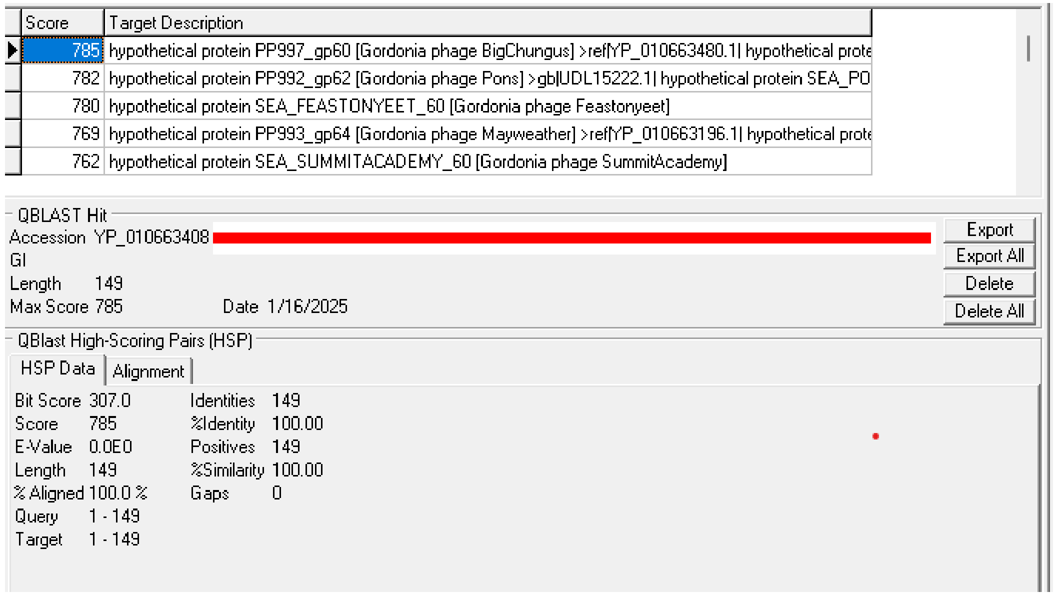
Gaps 0

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Is there more than one feature called in this coding region? Yes, this is a gene because there is cp, there are more than 10 highly similar genes, and both Glimmer and GeneMark call it a gene.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There are 5 1:1 alignments with E-values of 0. Some include BigChungus, Pons, and Feastonyeet
- There is 1 1:1 alignment for start 42,102
- There are no 1:1 alignments for start 42,197



Score	Target Description
785	hypothetical protein PP997_gp60 [Gordonia phage BigChungus] >ref|YP_010663480.1| hypothetical prote
782	hypothetical protein PP992_gp62 [Gordonia phage Pons] >gb|UDL15222.1| hypothetical protein SEA_PO
780	hypothetical protein SEA_FEASTONYEET_60 [Gordonia phage Feastonyeet]
769	hypothetical protein PP993_gp64 [Gordonia phage Mayweather] >ref|YP_010663196.1| hypothetical prote
762	hypothetical protein SEA_SUMMITACADEMY_60 [Gordonia phage SummitAcademy]

QBLAST Hit

Accession YP_010663408
GI
Length 149
Max Score 785 Date 1/16/2025

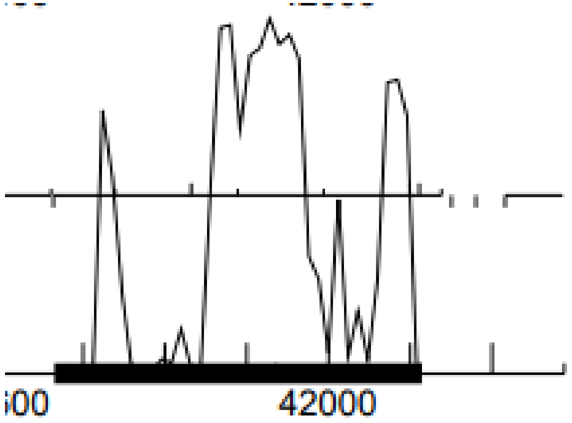
QBlast High-Scoring Pairs (HSP)

HSP Data | Alignment

```
Bit Score  307.0      Identities  149
Score      785        %Identity   100.00
E-Value    0.0E0      Positives   149
Length     149        %Similarity 100.00
% Aligned  100.0 %    Gaps        0
Query      1 - 149
Target     1 - 149
```

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

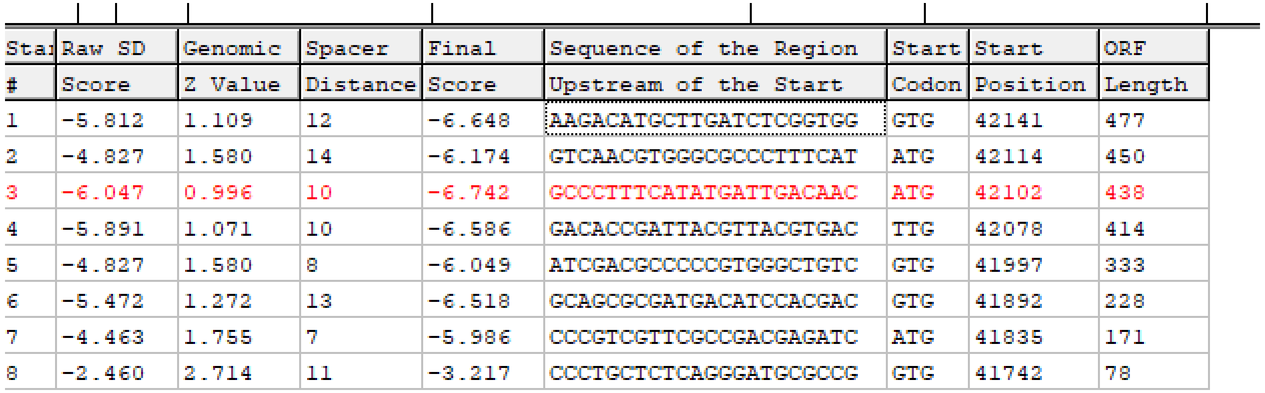
- For start site 42,114-41,665 all cp that can be included is included.It is short about 90 nucleotides at the stop.
- For start sites 42,102 and 42,197 all cp that can be included is included .



RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- What is the z-value and final score? The ZV: 1.508 FS: -6.174
- Starterator called two other start sites. 42,102 ZV: 0.996 FS:-6.742 and 42,197 ZV: 1.580 FS:-6.049

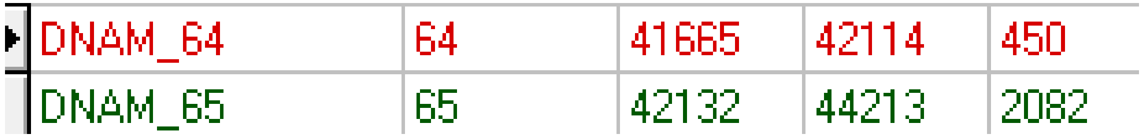
- Screenshot RBS Values here.



Star	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	-5.812	1.109	12	-6.648	AAGACATGCTTGATCTCGGTGG	GTG	42141	477
2	-4.827	1.580	14	-6.174	GTCAACGTGGGCGCCCTTTCAT	ATG	42114	450
3	-6.047	0.996	10	-6.742	GCCCTTTCATATGATTGACAAC	ATG	42102	438
4	-5.891	1.071	10	-6.586	GACACCGATTACGTTACGTGAC	TTG	42078	414
5	-4.827	1.580	8	-6.049	ATCGACGCCCCCGTGGGCTGTC	GTG	41997	333
6	-5.472	1.272	13	-6.518	GCAGCGCGATGACATCCACGAC	GTG	41892	228
7	-4.463	1.755	7	-5.986	CCCGTCGTTCGCCGACGAGATC	ATG	41835	171
8	-2.460	2.714	11	-3.217	CCCTGCTCTCAGGGATGCGCCG	GTG	41742	78

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

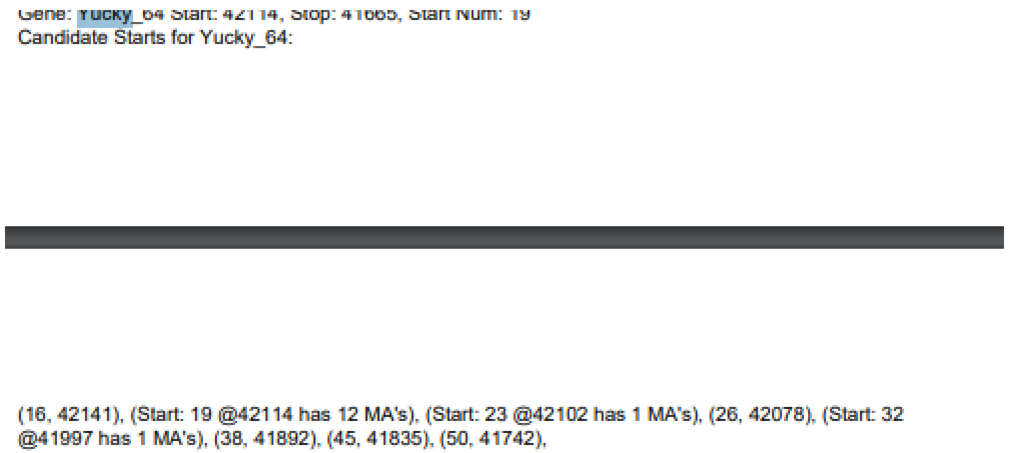
- Start site 42,114 has a gap of 17
- Start site 42,102 has a gap of 29
- Start site 42,197 has a gap 134



DNAM_64	64	41665	42114	450
DNAM_65	65	42132	44213	2082

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Start site 42,114 has 12 MAs
- Start site 42,102 has 1 MAs
- Start site 42,197 has 1 MAs



Gene: Yucky_64 Start: 42114, Stop: 41665, Start Num: 19
Candidate Starts for Yucky_64:

(16, 42141), (Start: 19 @42114 has 12 MA's), (Start: 23 @42102 has 1 MA's), (26, 42078), (Start: 32 @41997 has 1 MA's), (38, 41892), (45, 41835), (50, 41742),

Gene 64

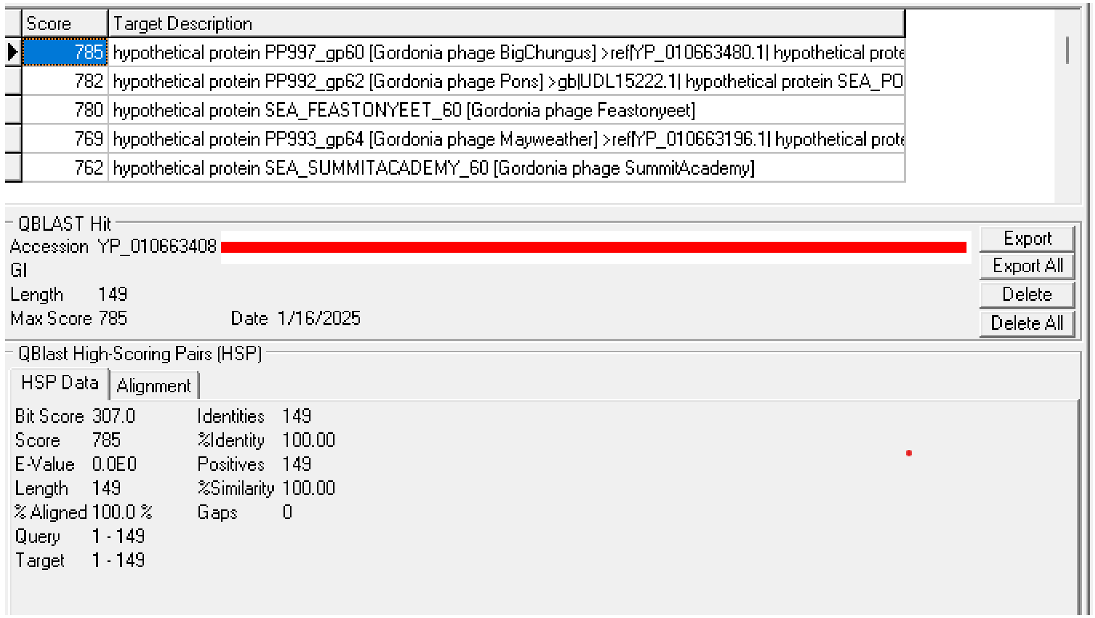
	42,114	42,102	42,197
GeneMark/Glimmer	Both Glimmer and GeneMark call this the start		
Coding Potential	All cp that can be included is included	All cp that can be included is included	All cp that can be included is included
RBS	ZV: 1.508 FS: -6.174	ZV: 0.996 FS:-6.742	ZV: 1.580 FS:-6.049
Blast	There are 5 1:1 alignments	There is 1 1:1 alignment	There are 0 1:1 alignments
Starterator	There are 12 MAs	There is 1 MAs	There is 1 MAs
Gap/Overlap	There is a gap of 17	There is a gap of 29	There is a gap of 134

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is 42,114. GeneMark and Glimmer call this as the start site, all cp that can be included is, the RBS for all 3 start sites are very out of range so they are not being considered, there are 5 1:1 alignments, 12 MAs, and a gap of 17.

BLAST function evidence. What assigned functions do other highly similar genes have?

- All highly similar genes are assigned the function of hypothetical protein.



Score	Target Description
785	hypothetical protein PP997_gp60 [Gordonia phage BigChungus] >ref|YP_010663480.1| hypothetical prote
782	hypothetical protein PP992_gp62 [Gordonia phage Pons] >gb|UDL15222.1| hypothetical protein SEA_PO
780	hypothetical protein SEA_FEASTONYEET_60 [Gordonia phage Feastonyeet]
769	hypothetical protein PP993_gp64 [Gordonia phage Mayweather] >ref|YP_010663196.1| hypothetical prote
762	hypothetical protein SEA_SUMMITACADEMY_60 [Gordonia phage SummitAcademy]

QBLAST Hit

Accession YP_010663408

GI

Length 149

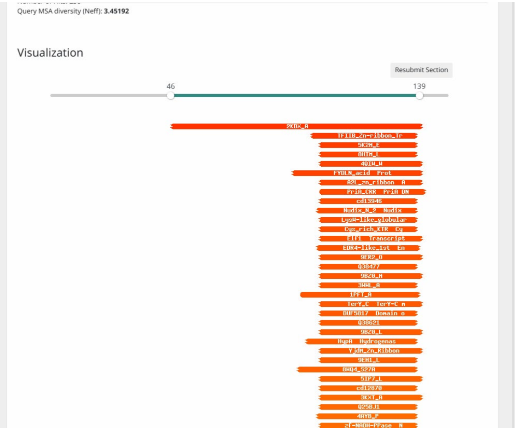
Max Score 785

Date 1/16/2025

Bit Score	307.0	Identities	149
Score	785	%Identity	100.00
E-Value	0.0E0	Positives	149
Length	149	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 149		
Target	1 - 149		

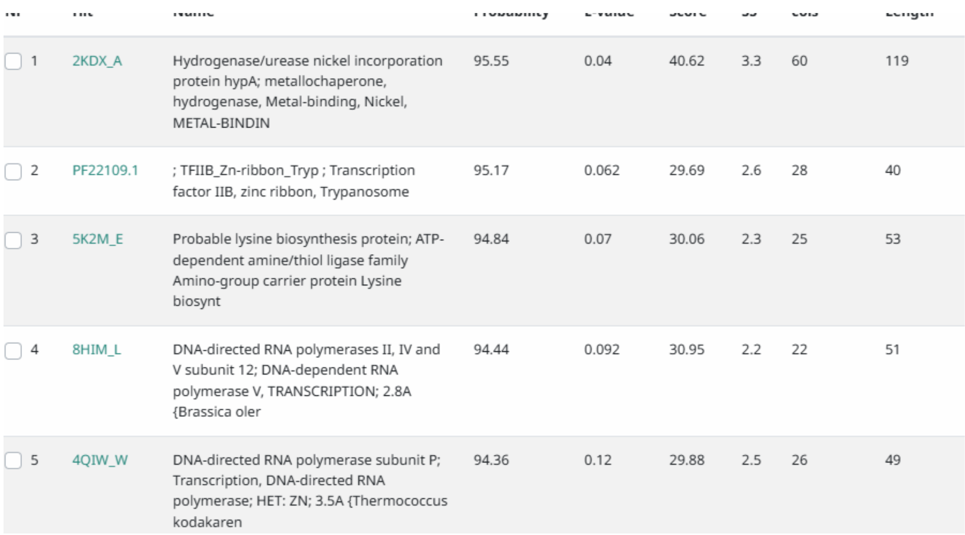
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- The highly similar matches have the assigned functions of Hydrogenase/unrease nickel incorporation protein, Trpanosome, Probable lysine biosynthesis, and DNA-directed RNA polymerase 2,4,and 5 subunit.
- There are 250 hits, 41 of which have a probability of 90 or higher.
- There are at least 1 conserved domains.
- Organism of the top function was : [Helicobacter pylori 26695]



Query MSA diversity (Neff): **3.45192**

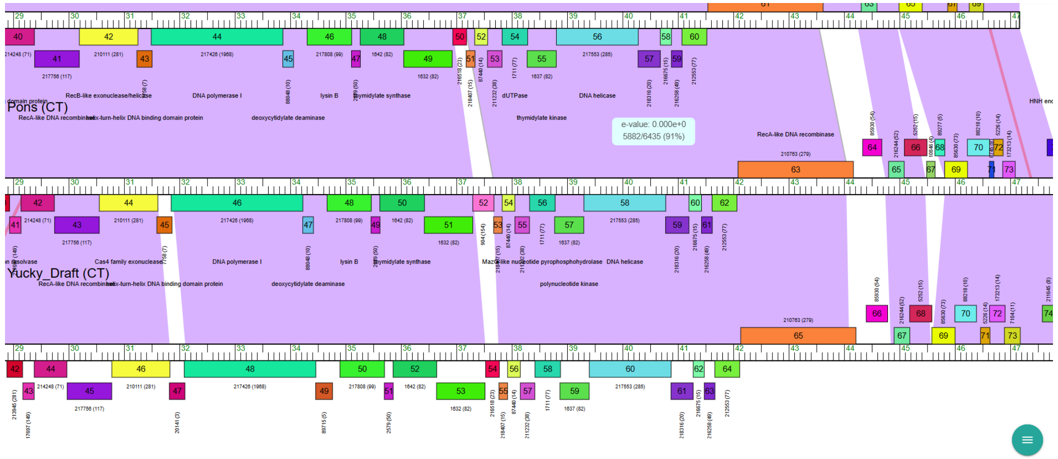
Visualization

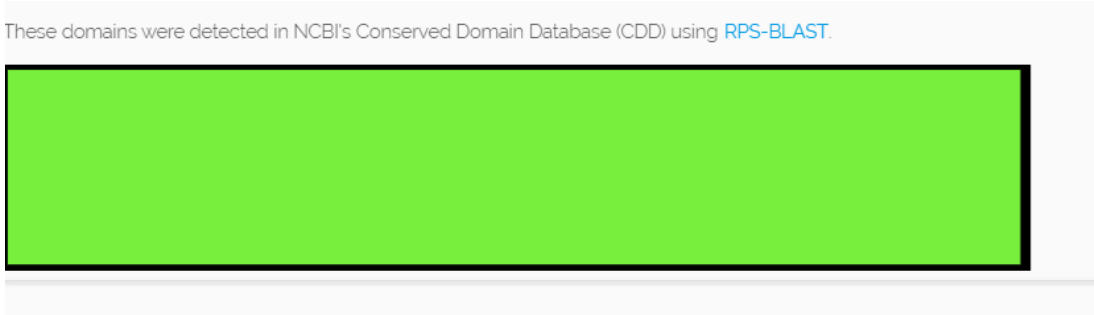


	Nr	Hit	Name	Probability	E-value	Score	SS	Cols	Length
□	1	2KDX_A	Hydrogenase/urease nickel incorporation protein hypA; metallochaperone, hydrogenase, Metal-binding, Nickel, METAL-BINDIN	95.55	0.04	40.62	3.3	60	119
□	2	PF22109.1	; TFIIB_Zn-ribbon_Tryp ; Transcription factor IIB, zinc ribbon, Trypanosome	95.17	0.062	29.69	2.6	28	40
□	3	5K2M_E	Probable lysine biosynthesis protein; ATP-dependent amine/thiol ligase family Amino-group carrier protein Lysine biosynt	94.84	0.07	30.06	2.3	25	53
□	4	8HIM_L	DNA-directed RNA polymerases II, IV and V subunit 12; DNA-dependent RNA polymerase V, TRANSCRIPTION; 2.8A {Brassica oler	94.44	0.092	30.95	2.2	22	51
□	5	4QIW_W	DNA-directed RNA polymerase subunit P; Transcription, DNA-directed RNA polymerase; HET: ZN; 3.5A {Thermococcus kodakaren	94.36	0.12	29.88	2.5	26	49

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

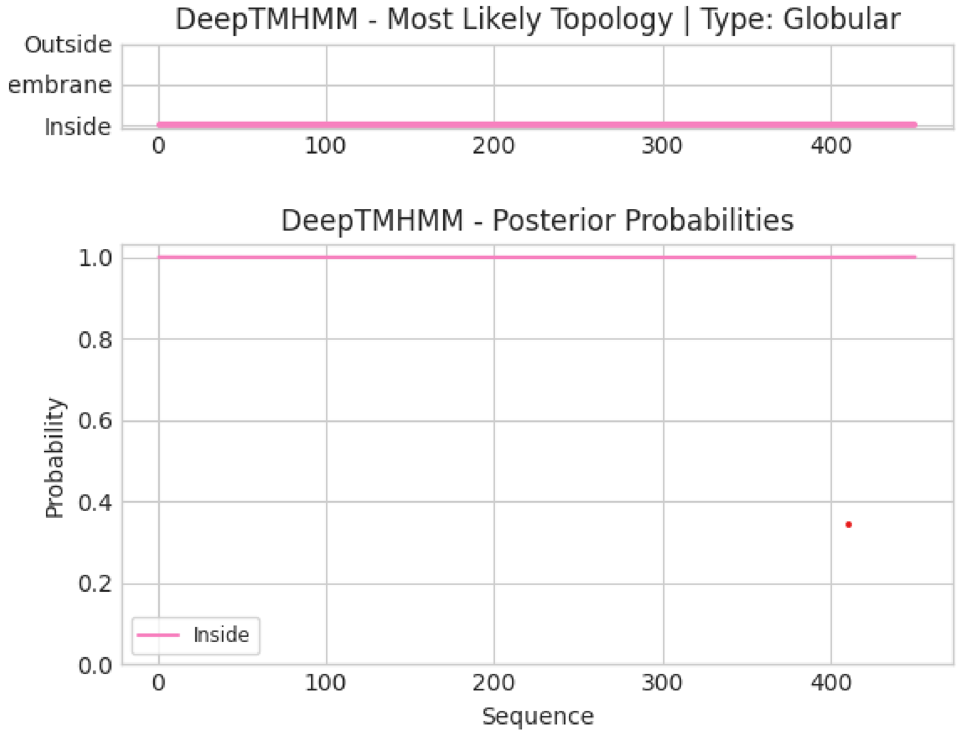
- There are no conserved domains and no known functions.


Pons (CT)
RecA-like DNA recombinase
RecB-like exonuclease/helicase
helix-turn-helix DNA binding domain protein
DNA polymerase I
deoxycytidylate deaminase
lysin B
thymidylate synthase
dUTPase
thymidylate kinase
DNA helicase
e-value: 0.000e+0
5882/6435 (91%)
RecA-like DNA recombinase
Yucky_Draft (CT)
Cas4 family exonuclease
MazG-like nucleotide pyrophosphohydrolase
polynucleotide kinase


These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- The gene has no transmembrane domains.



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function of this gene is a hypothetical protein because there is no phamerator evidence(No conserved domains and no known functions, Blast calls all highly similar genes hypothetical proteins, Hhpred does call the highly similar genes as different things, but they are not listed on the functional assignments list, and the gene has no transmembrane domains.

Feature 64 – Stop 44213

Glimmer/GeneMark

What feature number is this? 64

What is the stop site? 44213

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

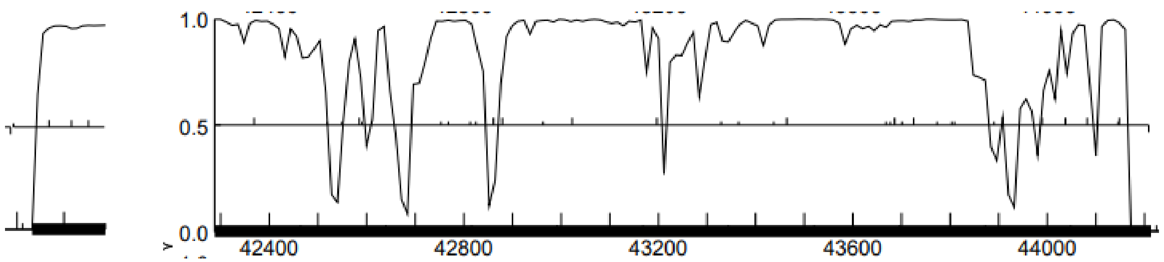
Called by both Glimmer and GeneMark.

What is the autoannotated start?

42132

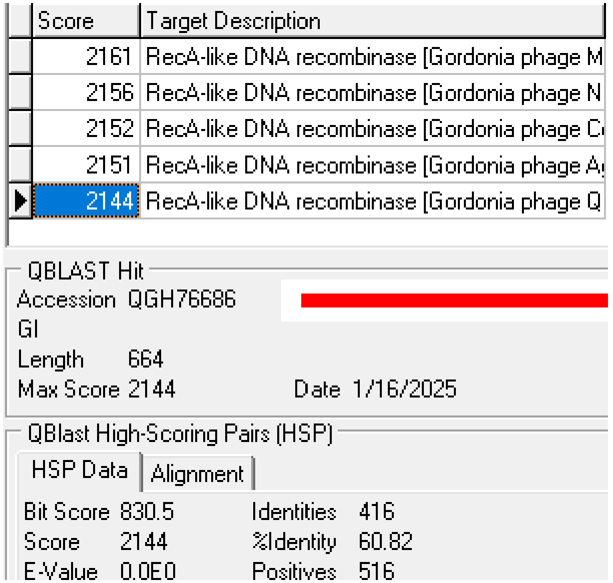
Gap: ____17____ or overlap: _______ (with gene in front of it) for the autoannotated start

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- There are many strong and weak peaks throughout the sequence, but the coding potential is uninterrupted. The potential is on frame 3. There are some small peaks in frames 4 and 6, but they are revers frames and very small peaks.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
2161	RecA-like DNA recombinase [Gordonia phage M
2156	RecA-like DNA recombinase [Gordonia phage N
2152	RecA-like DNA recombinase [Gordonia phage C
2151	RecA-like DNA recombinase [Gordonia phage A
2144	RecA-like DNA recombinase [Gordonia phage Q

Accession QGH76686
GI
Length 664
Max Score 2144 Date 1/16/2025

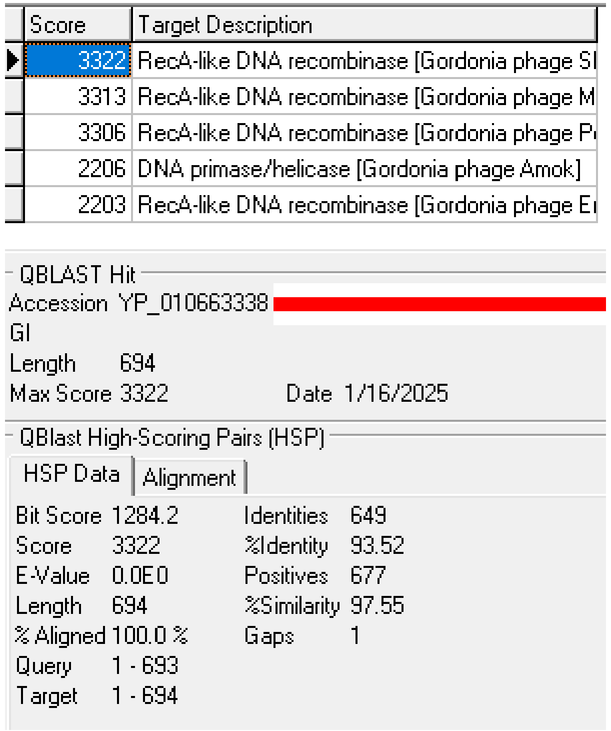
Bit Score	830.5	Identities	416
Score	2144	%Identity	60.82
E-Value	0.0E0	Positives	516

- At least 25 highly similar phages with an e-value close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- I believe this is a gene. It is called by both Glimmer and GeneMark and has consistently strong coding potential throughout the sequence of the gene. Lastly, it has at least 25 BLAST hits with an E-value close to 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
3322	RecA-like DNA recombinase [Gordonia phage Sl
3313	RecA-like DNA recombinase [Gordonia phage M
3306	RecA-like DNA recombinase [Gordonia phage P
2206	DNA primase/helicase [Gordonia phage Amok]
2203	RecA-like DNA recombinase [Gordonia phage E

QBLAST Hit

Accession YP_010663338

GI

Length 694

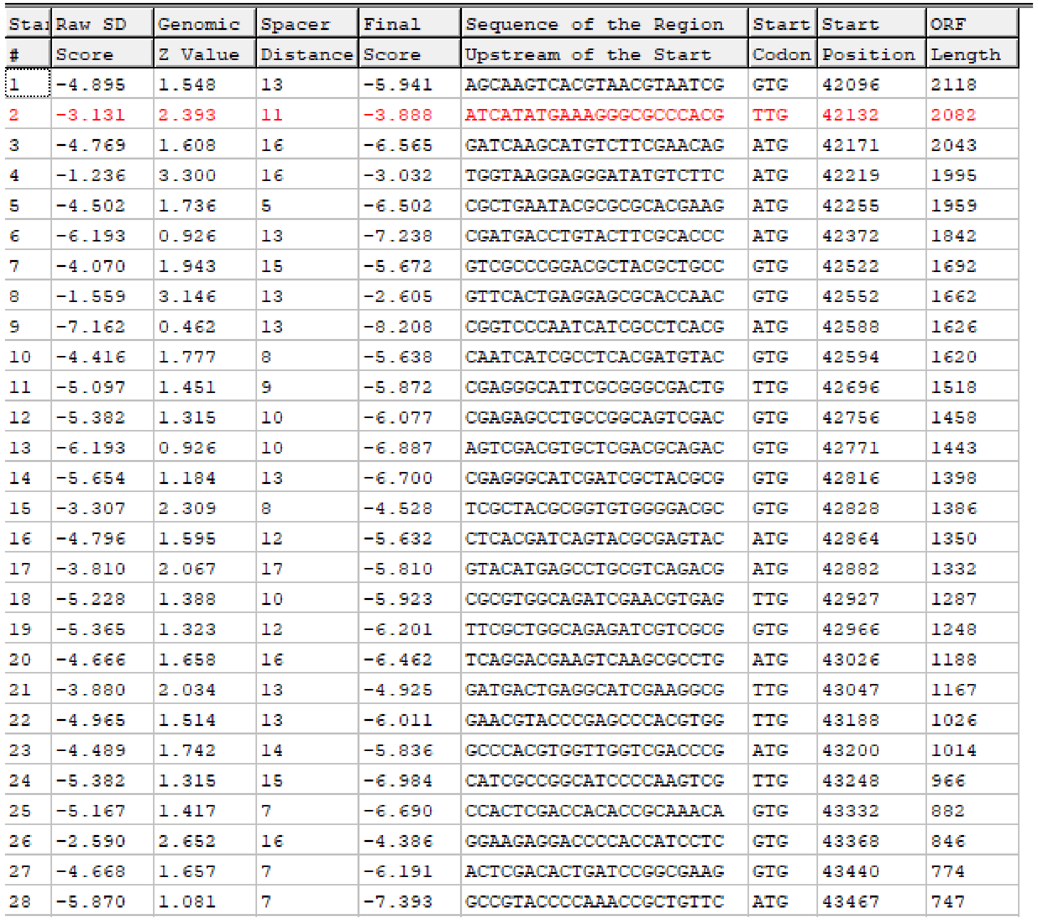
Max Score 3322 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	1284.2	Identities	649
Score	3322	%Identity	93.52
E-Value	0.0E0	Positives	677
Length	694	%Similarity	97.55
% Aligned	100.0 %	Gaps	1
Query	1 - 693		
Target	1 - 694		

- There are 10 1:1 alignments, 9 12:5 alignments, and 6 13:6 alignments. No alternate start sites are known.

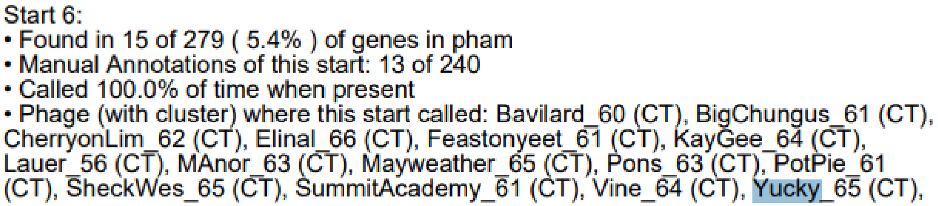
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-4.895	1.548	13	-5.941	AGCAAGTCACGTAACGTAATCG	GTG	42096	2118
2	-3.131	2.393	11	-3.888	ATCATATGAAAGGGCGCCCACG	TTG	42132	2082
3	-4.769	1.608	16	-6.565	GATCAAGCATGTCTTCGAACAG	ATG	42171	2043
4	-1.236	3.300	16	-3.032	TGGTAAGGAGGGATATGTCTTC	ATG	42219	1995
5	-4.502	1.736	5	-6.502	CGCTGAATACGCGCGCACGAAG	ATG	42255	1959
6	-6.193	0.926	13	-7.238	CGATGACCTGTACTTCGCACCC	ATG	42372	1842
7	-4.070	1.943	15	-5.672	GTCGCCCGGACGCTACGCTGCC	GTG	42522	1692
8	-1.559	3.146	13	-2.605	GTTCACTGAGGAGCGCACCAAC	GTG	42552	1662
9	-7.162	0.462	13	-8.208	CGGTCCCAATCATCGCCTCACG	ATG	42588	1626
10	-4.416	1.777	8	-5.638	CAATCATCGCCTCACGATGTAC	GTG	42594	1620
11	-5.097	1.451	9	-5.872	CGAGGGCATTCGCGGGCGACTG	TTG	42696	1518
12	-5.382	1.315	10	-6.077	CGAGAGCCTGCCGGCAGTCGAC	GTG	42756	1458
13	-6.193	0.926	10	-6.887	AGTCGACGTGCTCGACGCAGAC	GTG	42771	1443
14	-5.654	1.184	13	-6.700	CGAGGGCATCGATCGCTACGCG	GTG	42816	1398
15	-3.307	2.309	8	-4.528	TCGCTACGCGGTGTGGGGACGC	GTG	42828	1386
16	-4.796	1.595	12	-5.632	CTCACGATCAGTACGCGAGTAC	ATG	42864	1350
17	-3.810	2.067	17	-5.810	GTACATGAGCCTGCGTCAGACG	ATG	42882	1332
18	-5.228	1.388	10	-5.923	CGCGTGGCAGATCGAACGTGAG	TTG	42927	1287
19	-5.365	1.323	12	-6.201	TTCGCTGGCAGAGATCGTCGCG	GTG	42966	1248
20	-4.666	1.658	16	-6.462	TCAGGACGAAGTCAAGCGCCTG	ATG	43026	1188
21	-3.880	2.034	13	-4.925	GATGACTGAGGCATCGAAGGCG	TTG	43047	1167
22	-4.965	1.514	13	-6.011	GAACGTACCCGAGCCCACGTGG	TTG	43188	1026
23	-4.489	1.742	14	-5.836	GCCCACGTGGTTGGTCGACCCG	ATG	43200	1014
24	-5.382	1.315	15	-6.984	CATCGCCGGCATCCCCAAGTCG	TTG	43248	966
25	-5.167	1.417	7	-6.690	CCACTCGACCACACCGCAAACA	GTG	43332	882
26	-2.590	2.652	16	-4.386	GGAAGAGGACCCCACCATCCTC	GTG	43368	846
27	-4.668	1.657	7	-6.191	ACTCGACACTGATCCGGCGAAG	GTG	43440	774
28	-5.870	1.081	7	-7.393	GCCGTACCCCAAACCGCTGTTC	ATG	43467	747

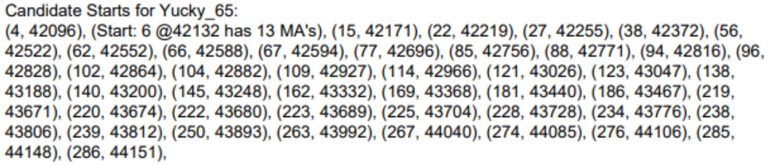
- Automated start: Z-value 2.393, Final score -3.888
- New RBS introduced start site (42219): Z-value: 3.300, Final score: -3.032.
- There is another site with good RBS numbers, but it cuts off too much coding potential.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



Start 6:
• Found in 15 of 279 (5.4%) of genes in pham
• Manual Annotations of this start: 13 of 240
• Called 100.0% of time when present
• Phage (with cluster) where this start called: Bavilard_60 (CT), BigChungus_61 (CT), CherryonLim_62 (CT), Elinal_66 (CT), Feastonyeet_61 (CT), KayGee_64 (CT), Lauer_56 (CT), MAnor_63 (CT), Mayweather_65 (CT), Pons_63 (CT), PotPie_61 (CT), SheckWes_65 (CT), SummitAcademy_61 (CT), Vine_64 (CT), Yucky_65 (CT),

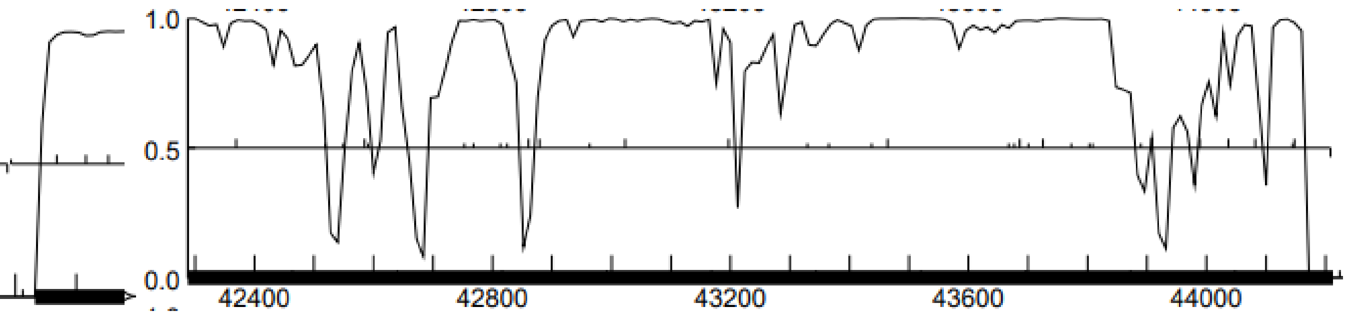
Gene: Yucky_65 Start: 42132, Stop: 44213, Start Num: 6



Candidate Starts for Yucky_65:
(4, 42096), (Start: 6 @42132 has 13 MA's), (15, 42171), (22, 42219), (27, 42255), (38, 42372), (56, 42522), (62, 42552), (66, 42588), (67, 42594), (77, 42696), (85, 42756), (88, 42771), (94, 42816), (96, 42828), (102, 42864), (104, 42882), (109, 42927), (114, 42966), (121, 43026), (123, 43047), (138, 43188), (140, 43200), (145, 43248), (162, 43332), (169, 43368), (181, 43440), (186, 43467), (219, 43671), (220, 43674), (222, 43680), (223, 43689), (225, 43704), (228, 43728), (234, 43776), (238, 43806), (239, 43812), (250, 43893), (263, 43992), (267, 44040), (274, 44085), (276, 44106), (285, 44148), (286, 44151),

- Automated start site: called 100% of the time when present, only site to ever receive an MA (13)
- Alternate start: never called, 0 MA’s

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- The automated start site cuts off no coding potential.

  Alternate start cuts off about 100 nucleotides of coding potential.

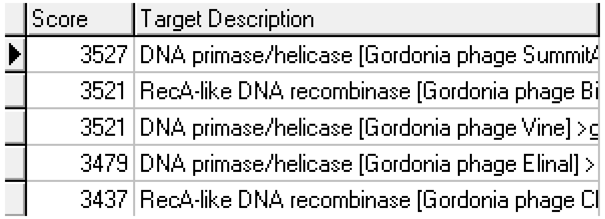
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 42132-42114= 18-1 for gap= 17
- 42219-42114=105-1 for gap= 104
- This made me decide against the alternate site.

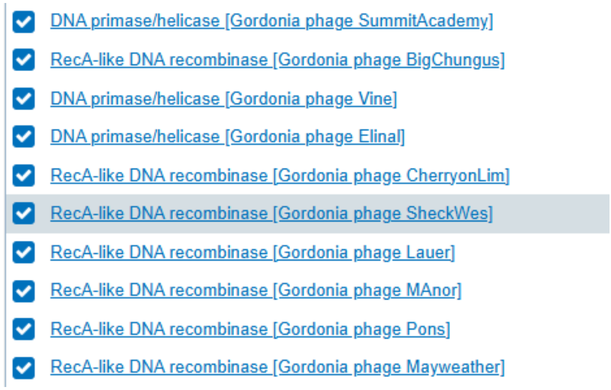
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is the automated site of 42132. It has 10 1:1 BLAST hits. It has very good RBS numbers, has more manual annotations than any other possible start and is the only site to ever receive an MA. It cuts off no coding potential and it has a much smaller gap than the potential alternate start.

BLAST function evidence. What assigned functions do other highly similar genes have?



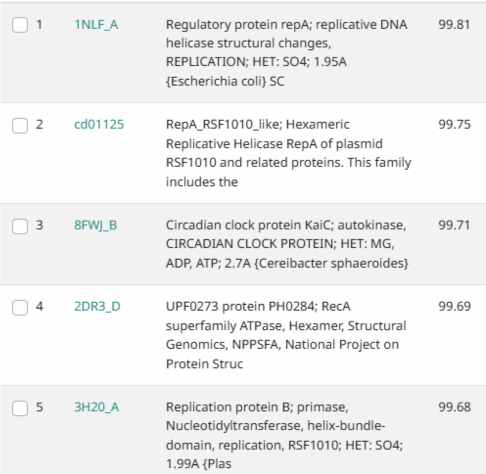
Score	Target Description
3527	DNA primase/helicase [Gordonia phage SummitA
3521	RecA-like DNA recombinase [Gordonia phage Bi
3521	DNA primase/helicase [Gordonia phage Vine] >g
3479	DNA primase/helicase [Gordonia phage Elinal] >
3437	RecA-like DNA recombinase [Gordonia phage Cl



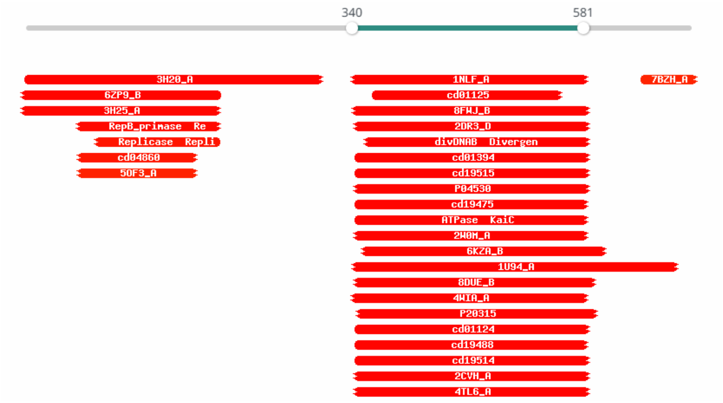
- ☑ DNA primase/helicase [Gordonia phage SummitAcademy]
- ☑ RecA-like DNA recombinase [Gordonia phage BigChungus]
- ☑ DNA primase/helicase [Gordonia phage Vine]
- ☑ DNA primase/helicase [Gordonia phage Elinal]
- ☑ RecA-like DNA recombinase [Gordonia phage CherryonLim]
- ☑ RecA-like DNA recombinase [Gordonia phage SheckWes]
- ☑ RecA-like DNA recombinase [Gordonia phage Lauer]
- ☑ RecA-like DNA recombinase [Gordonia phage MAnor]
- ☑ RecA-like DNA recombinase [Gordonia phage Pons]
- ☑ RecA-like DNA recombinase [Gordonia phage Mayweather]

- Highly similar genes on DNA master had the functions of DNA primase/helicase and RecA-like DNA recombinase.
- NCBI BLAST yielded the same 2 functions for results.

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

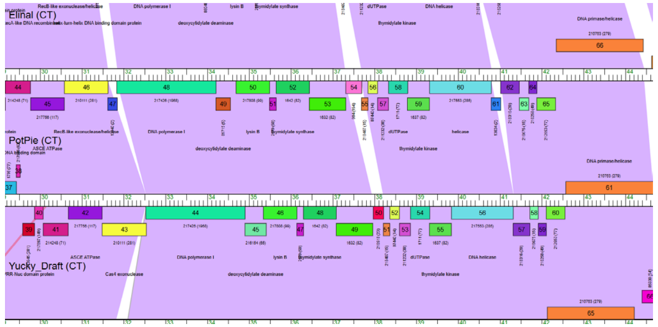


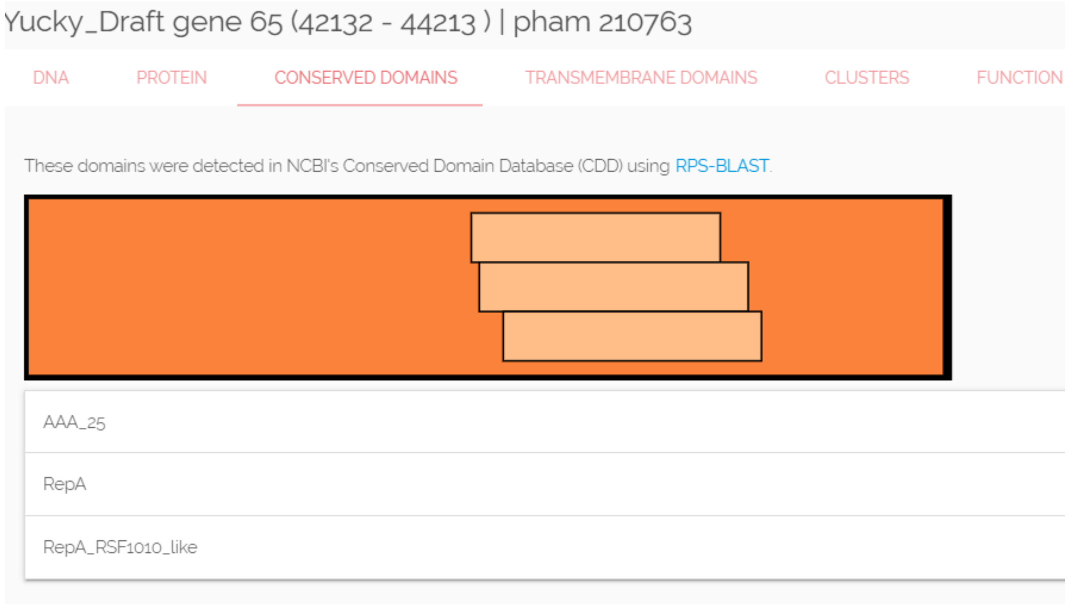
□	1	1NLF_A	Regulatory protein repA; replicative DNA helicase structural changes, REPLICATION; HET: SO4; 1.95A {Escherichia coli} SC	99.81
□	2	cd01125	RepA_RSF1010_like; Hexameric Replicative Helicase RepA of plasmid RSF1010 and related proteins. This family includes the	99.75
□	3	8FWJ_B	Circadian clock protein KaiC; autokinase, CIRCADIAN CLOCK PROTEIN; HET: MG, ADP, ATP; 2.7A {Cereibacter sphaeroides}	99.71
□	4	2DR3_D	UPF0273 protein PH0284; RecA superfamily ATPase, Hexamer, Structural Genomics, NPPSFA, National Project on Protein Struc	99.69
□	5	3H20_A	Replication protein B; primase, Nucleotidyltransferase, helix-bundle-domain, replication, RSF1010; HET: SO4; 1.99A {Plas	99.68



- HHpred shows strong hits in the N-terminal side to primase (hit #4) and strong hits in the C-terminal side to helicase (hits #1-3)

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?





Yucky_Draft gene 65 (42132 - 44213) | pham 210763

DNA PROTEIN CONSERVED DOMAINS TRANSMEMBRANE DOMAINS CLUSTERS FUNCTION

These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

AAA_25

RepA

RepA_RSF1010_like

- BigChungus, Elinal, and PotPie all have the gene. PotPie and Elinal have it called as a DNA primase/helicase, BigChungus has it called as a RecA-like recombinase.
- All 4 phages, including Yucky, have a RecA conserved domain.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- I would like to call this a DNA primase/helicase

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I am assigning this as a DNA primase/helicase. BLAST via both DNA master and NCBI show hits for this function, HHpred also shows hits for this function, including both a N-terminal side primase domain and a C-terminal side helicase domain. Phamerator also showed that similar phage had assigned it this function.

Feature 65 – Stop 44785

Glimmer/GeneMark

What feature number is this? 65

What is the stop site?

44785

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

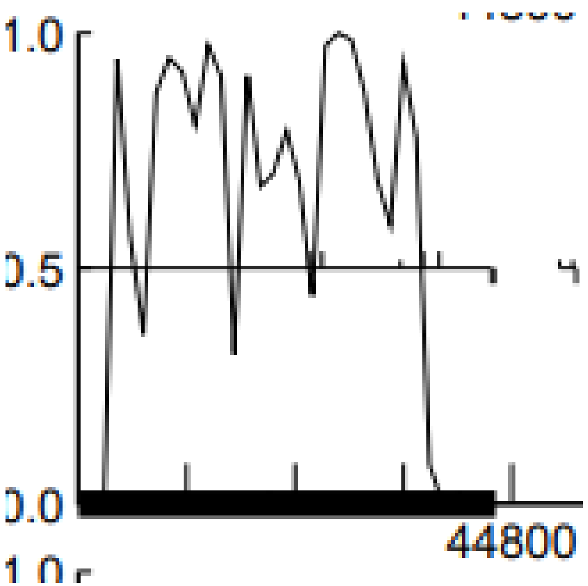
Both Glimmer and GeneMark

What is the autoannotated start?

44399

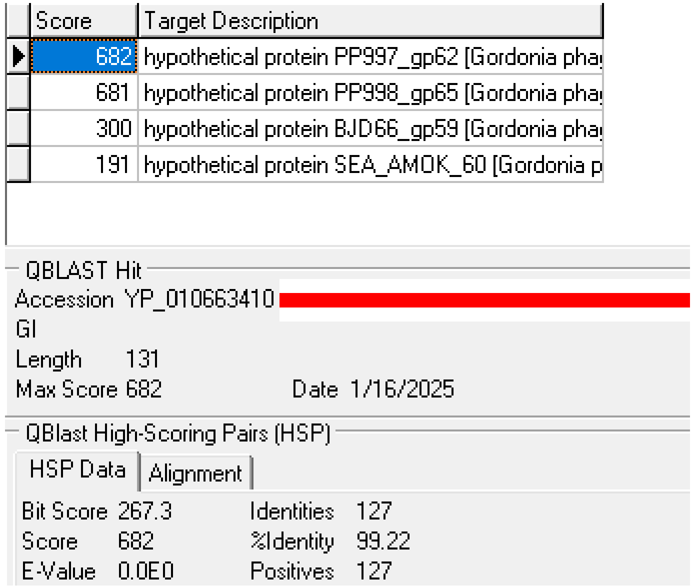
Gap: ___185_____ or overlap: _______ (with gene in front of it) for the autoannotated start

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- There are many strong and weak peaks throughout the sequence.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
682	hypothetical protein PP997_gp62 [Gordonia pha
681	hypothetical protein PP998_gp65 [Gordonia pha
300	hypothetical protein BJD66_gp59 [Gordonia pha
191	hypothetical protein SEA_AMOK_60 [Gordonia p

QBLAST Hit

Accession YP_010663410

GI

Length 131

Max Score 682 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score 267.3 Identities 127

Score 682 %Identity 99.22

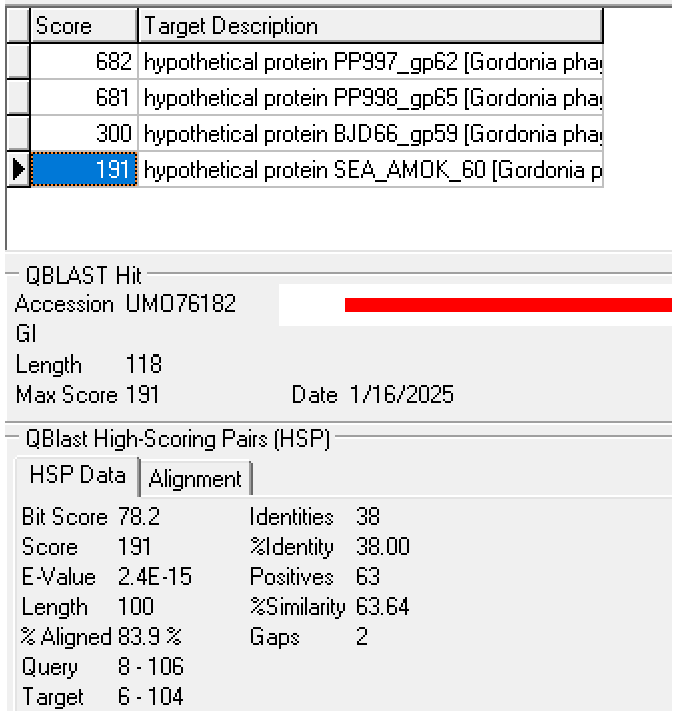
E-Value 0.0E0 Positives 127

- All 4 BLAST hits have an E-value close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- I believe this is a gene. It is called by both Glimmer and GeneMark and has fairly strong coding potential. It also has 4 BLAST hits with an E-value close to 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.



Score	Target Description
682	hypothetical protein PP997_gp62 [Gordonia pha
681	hypothetical protein PP998_gp65 [Gordonia pha
300	hypothetical protein BJD66_gp59 [Gordonia pha
191	hypothetical protein SEA_AMOK_60 [Gordonia p

QBLAST Hit

Accession UM076182

GI

Length 118

Max Score 191

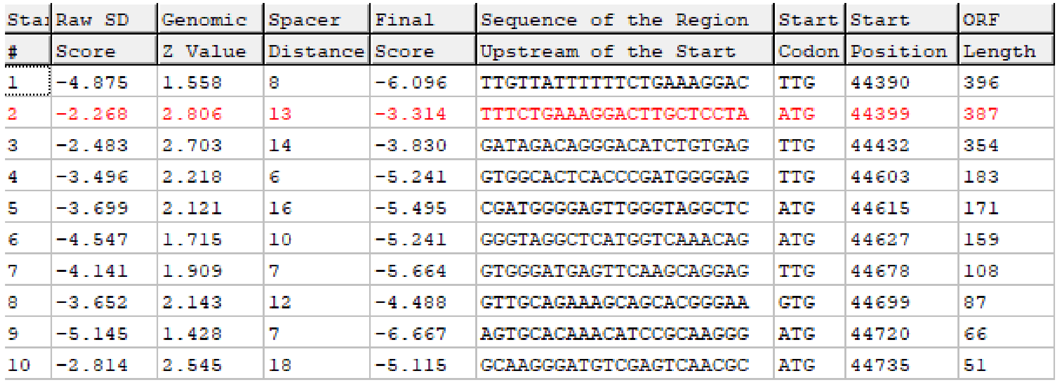
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	78.2	Identities	38
Score	191	%Identity	38.00
E-Value	2.4E-15	Positives	63
Length	100	%Similarity	63.64
% Aligned	83.9 %	Gaps	2
Query	8 - 106		
Target	6 - 104		

- There is one 1:1 hit, one 1:4 alignment, one 5:2 alignment, and one 8:6 alignment.

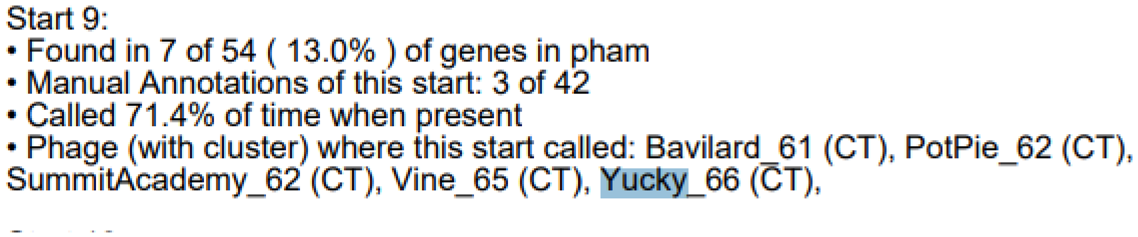
RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts?   Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



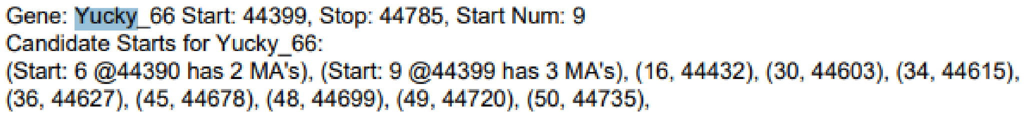
Sta #	Raw SD Score	Genomic Z Value	Spacer Distance	Final Score	Sequence of the Region Upstream of the Start	Start Codon	Start Position	ORF Length
1	-4.875	1.558	8	-6.096	TTGTTATTTTTTCTGAAAGGAC	TTG	44390	396
2	-2.268	2.806	13	-3.314	TTTCTGAAAGGACTTGCTCCTA	ATG	44399	387
3	-2.483	2.703	14	-3.830	GATAGACAGGGACATCTGTGAG	TTG	44432	354
4	-3.496	2.218	6	-5.241	GTGGCACTCACCCGATGGGGAG	TTG	44603	183
5	-3.699	2.121	16	-5.495	CGATGGGGAGTTGGGTAGGCTC	ATG	44615	171
6	-4.547	1.715	10	-5.241	GGGTAGGCTCATGGTCAAACAG	ATG	44627	159
7	-4.141	1.909	7	-5.664	GTGGGATGAGTTCAAGCAGGAG	TTG	44678	108
8	-3.652	2.143	12	-4.488	GTTGCAGAAAGCAGCACGGGAA	GTG	44699	87
9	-5.145	1.428	7	-6.667	AGTGCACAAACATCCGCAAGGG	ATG	44720	66
10	-2.814	2.545	18	-5.115	GCAAGGGATGTCGAGTCAACGC	ATG	44735	51

- Automated start: Z-value: 2.806, Final score: -3.314
- Alternate start (44432): Z-value: 2.703, Final Score: -3.830

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.



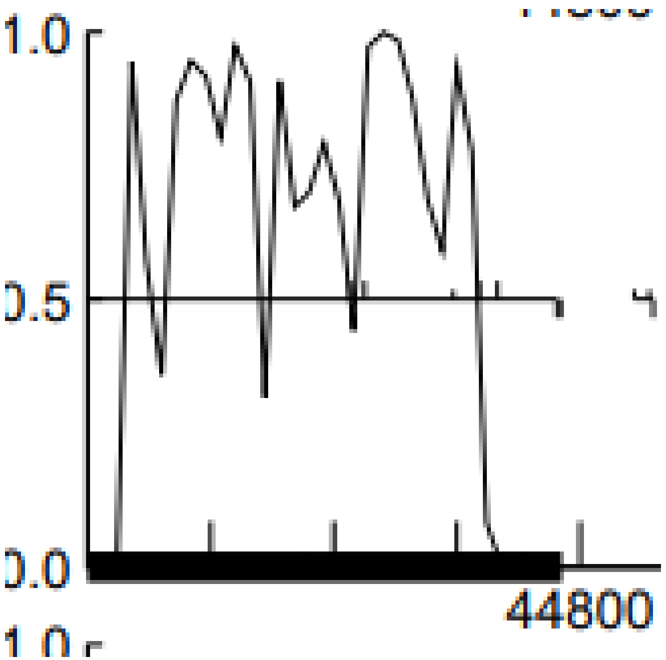
Start 9:
- Found in 7 of 54 (13.0%) of genes in pham
- Manual Annotations of this start: 3 of 42
- Called 71.4% of time when present
- Phage (with cluster) where this start called: Bavilard_61 (CT), PotPie_62 (CT), SummitAcademy_62 (CT), Vine_65 (CT), Yucky_66 (CT),



Gene: Yucky_66 Start: 44399, Stop: 44785, Start Num: 9
Candidate Starts for Yucky_66:
(Start: 6 @44390 has 2 MA's), (Start: 9 @44399 has 3 MA's), (16, 44432), (30, 44603), (34, 44615), (36, 44627), (45, 44678), (48, 44699), (49, 44720), (50, 44735),

- Alternate start 1 (44390): 2 MA’s, called 71% of the time when present.
- Automated start (44399): 3 MA’s
- Alternate start 2 (44432): 0 MA’s

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- Alternate start 1 (44390) cuts off no coding potential.
- Automated start (44399) cuts off no coding potential.
- Alternate start 2 (44432) cuts off a strong peak of coding potential. Likely not the start given current evidence.

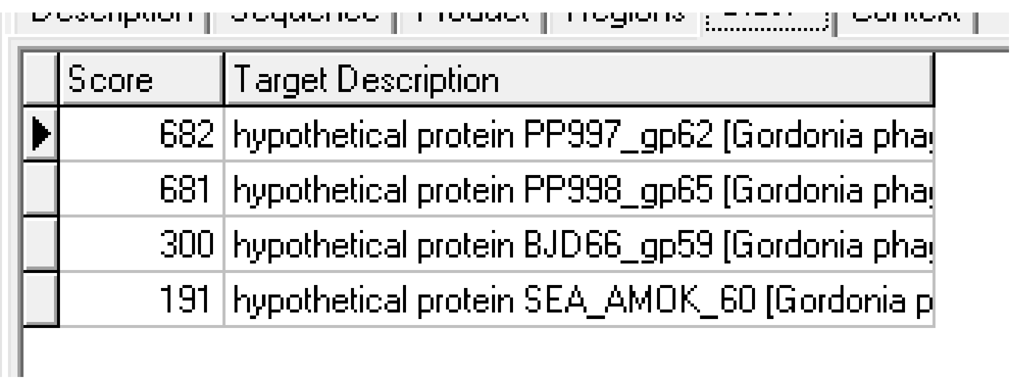
Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Alternate start 1 (44390): 44390-44213= 177-1 for gap= 176
- Automated start (44399) 44399-44213= 186-1 for gap= 185
- Alternate start 2 (44432) 44432-44213= 219-1 for gap= 218

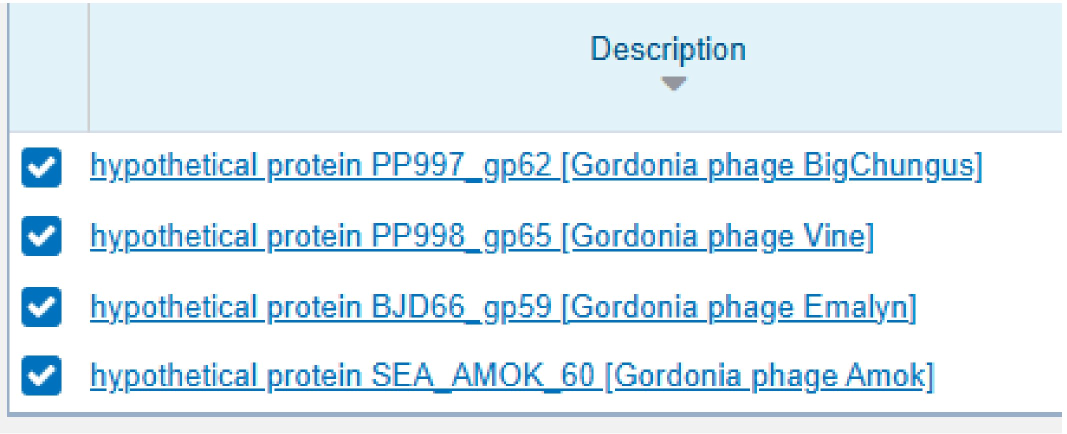
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- I believe the start site to agree with the automated start of 44399. It has a 1:1 alignment and the best RBS numbers of any possible start. It also has the strongest Starterator evidence, having the most MA's of any start. Lastly, it cuts off no coding potential and has the 2nd largest gap of possible alternate starts, but only by 9 nucleotides.

BLAST function evidence. What assigned functions do other highly similar genes have?



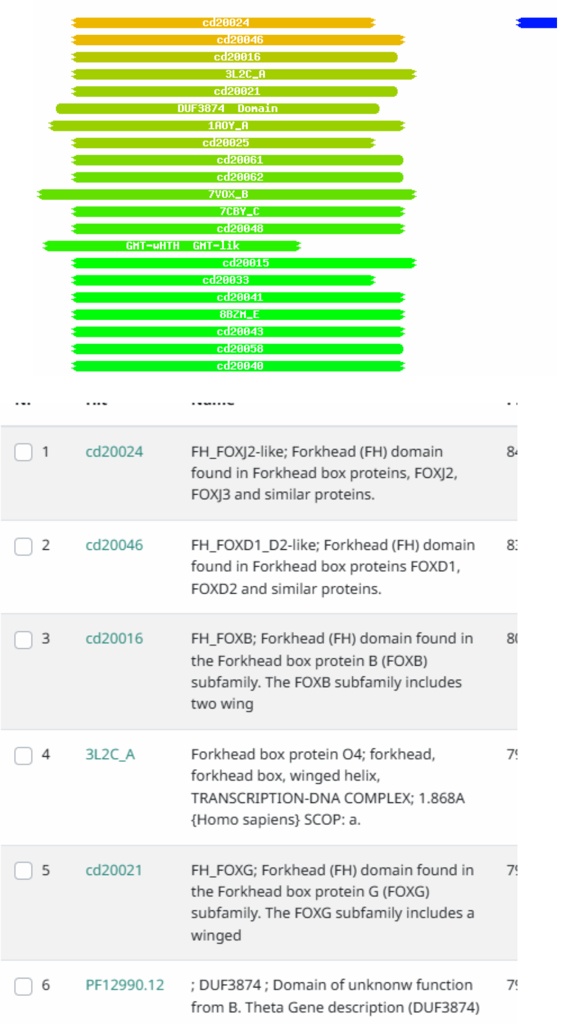
Score	Target Description
682	hypothetical protein PP997_gp62 [Gordonia pha
681	hypothetical protein PP998_gp65 [Gordonia pha
300	hypothetical protein BJD66_gp59 [Gordonia pha
191	hypothetical protein SEA_AMOK_60 [Gordonia p



	Description
☑	hypothetical protein PP997_gp62 [Gordonia phage BigChungus]
☑	hypothetical protein PP998_gp65 [Gordonia phage Vine]
☑	hypothetical protein BJD66_gp59 [Gordonia phage Emalyn]
☑	hypothetical protein SEA_AMOK_60 [Gordonia phage Amok]

- On both DNA Master and NCBI there are only 4 BLAST hits and they are all as hypothetical proteins.

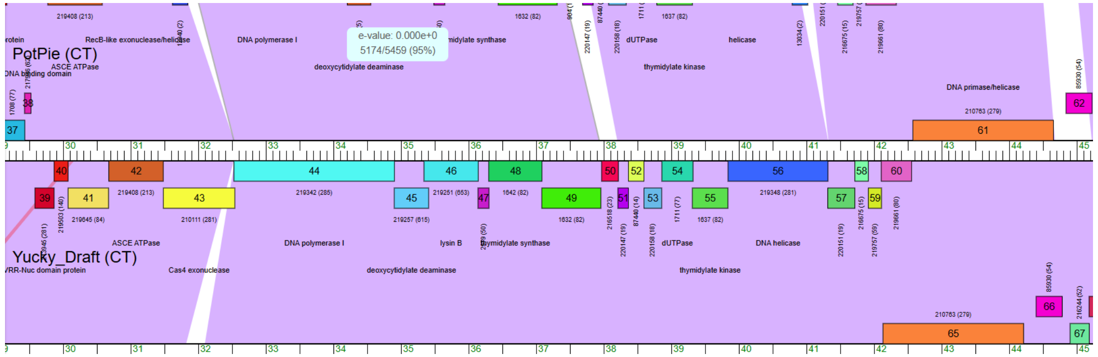
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



□	1	cd20024	FH_FOXJ2-like; Forkhead (FH) domain found in Forkhead box proteins, FOXJ2, FOXJ3 and similar proteins.	84
□	2	cd20046	FH_FOXD1_D2-like; Forkhead (FH) domain found in Forkhead box proteins FOXD1, FOXD2 and similar proteins.	83
□	3	cd20016	FH_FOXB; Forkhead (FH) domain found in the Forkhead box protein B (FOXB) subfamily. The FOXB subfamily includes two wing	80
□	4	3L2C_A	Forkhead box protein O4; forkhead, forkhead box, winged helix, TRANSCRIPTION-DNA COMPLEX; 1.868A {Homo sapiens} SCOP: a.	79
□	5	cd20021	FH_FOXG; Forkhead (FH) domain found in the Forkhead box protein G (FOXG) subfamily. The FOXG subfamily includes a winged	79
□	6	PF12990.12	; DUF3874 ; Domain of unknonw function from B. Theta Gene description (DUF3874)	79

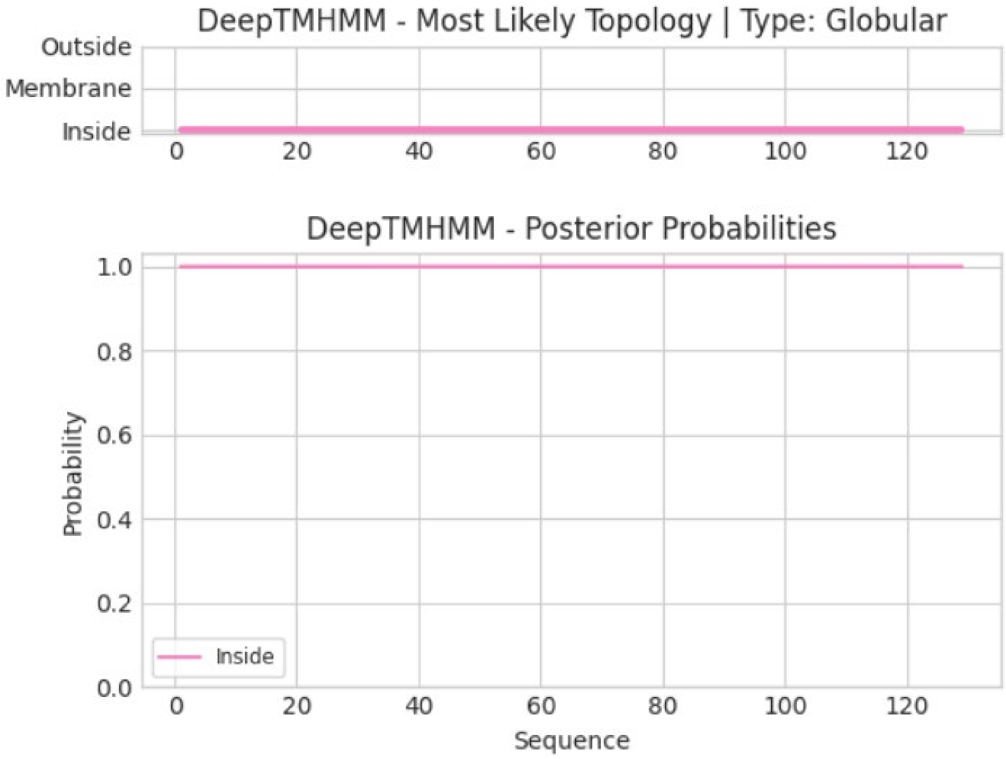
- There are 0 Hhpred hits with 90%+ probability.
- Many of these hits are called “forkhead” something, which is not in the official function list.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?



- PotPie and Big Chungus both have this gene and it is called a hypothetical protein in both. There are no conserved domains.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- There are no transmembrane domains

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- I am assigning this as a hypothetical protein. BLAST via both NCBI and DNA Master showed this as being the function. Hhpred showed no hits with a high enough probability to be considered, and phamerator showed that some similar phages contain this gene, but do not have it called as anything. There are no transmembrane domains.

Feature 66 Stop 45185

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

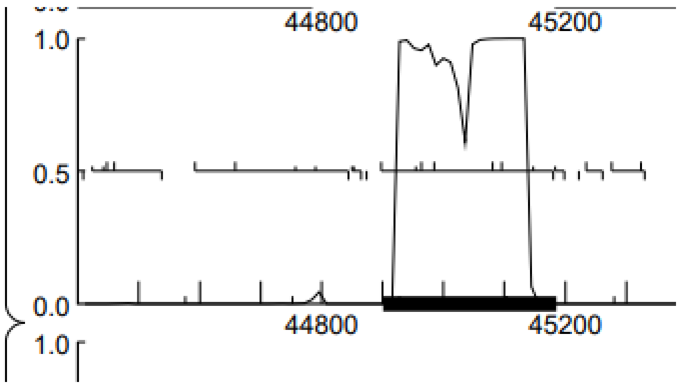
What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

- 66
- 45185
- Both Glimmer and GeneMark
- 44901
- 115 gap

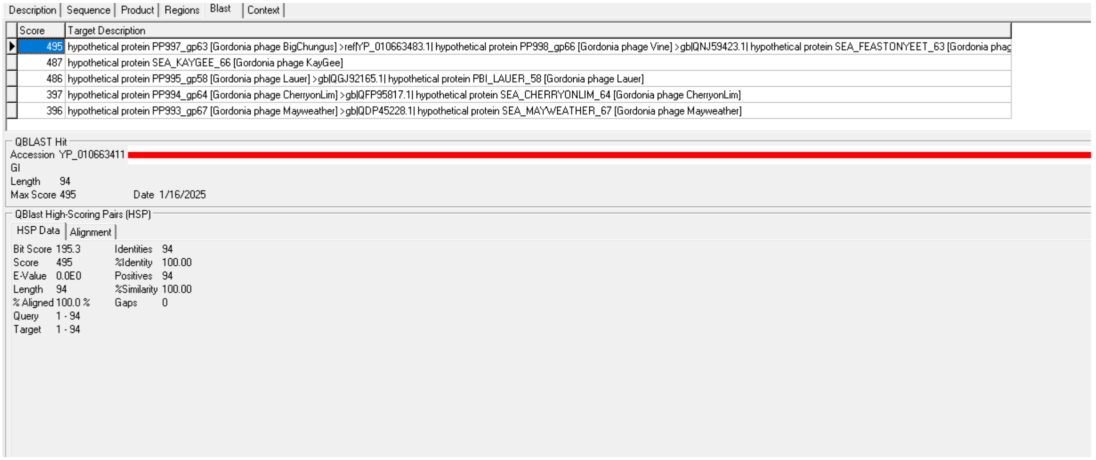
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Reading frame 3 has a strong coding potential.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are 20 highly similar genes with E value of close to 0.



Description | Sequence | Product | Regions | Blast | Context

Score	Target Description
495	hypothetical protein PP997_gp63 [Gordonia phage BigChungus] >ref|YP_010663483.1| hypothetical protein PP998_gp66 [Gordonia phage Vine] >gb|QNJ59423.1| hypothetical protein SEA_FEASTONYEET_63 [Gordonia phag
487	hypothetical protein SEA_KAYGEE_66 [Gordonia phage KayGee]
486	hypothetical protein PP995_gp58 [Gordonia phage Lauer] >gb|QGJ92165.1| hypothetical protein PBI_LAUER_58 [Gordonia phage Lauer]
397	hypothetical protein PP994_gp64 [Gordonia phage CherryonLim] >gb|QFP95817.1| hypothetical protein SEA_CHERRYONLIM_64 [Gordonia phage CherryonLim]
396	hypothetical protein PP993_gp67 [Gordonia phage Mayweather] >gb|QDP45228.1| hypothetical protein SEA_MAYWEATHER_67 [Gordonia phage Mayweather]

QBLAST Hit

```
Accession  YP_010663411
GI
Length     94
Max Score  495            Date  1/16/2025
```

QBlast High-Scoring Pairs (HSP)

```
Bit Score  195.3      Identities   94
Score      495        %Identity    100.00
E-Value    0.0E0      Positives    94
Length     94         %Similarity  100.00
% Aligned  100.0 %    Gaps         0
Query      1 - 94
Target     1 - 94
```

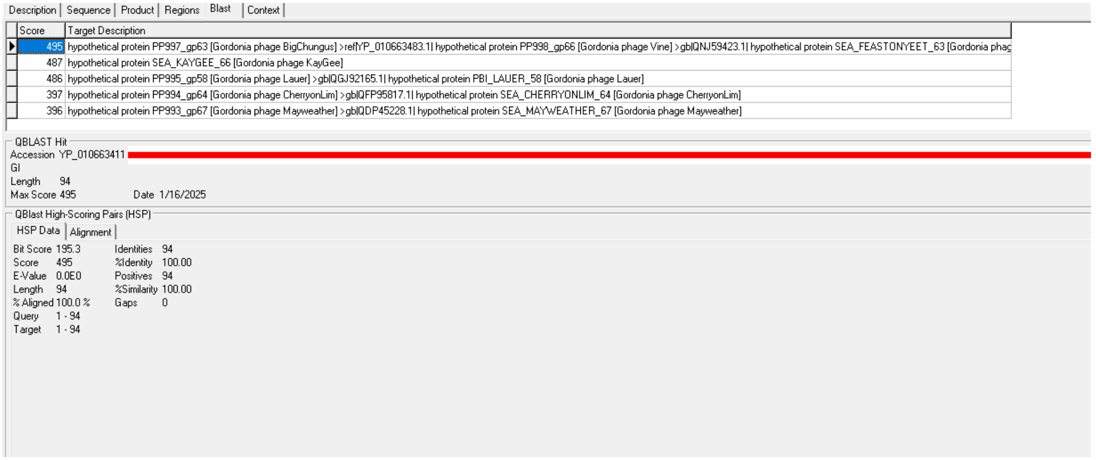
Answer: Is it a gene? Give evidence why you think this is a gene or not.

- It is a gene because:

- Both Glimmer and GeneMark call it a gene.
- Coding potential is strong.
- There are 20 highly similar genes with an E value of close to 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There are 18 1:1 alignments.



Score	Target Description
495	hypothetical protein PP997_gp63 [Gordonia phage BigChungus] >ref|YP_010663483.1| hypothetical protein PP998_gp66 [Gordonia phage Vine] >gb|QNJ59423.1| hypothetical protein SEA_FEASTONYEET_63 [Gordonia phag
487	hypothetical protein SEA_KAYGEE_66 [Gordonia phage KayGee]
486	hypothetical protein PP995_gp58 [Gordonia phage Lauer] >gb|QGJ92165.1| hypothetical protein PBI_LAUER_58 [Gordonia phage Lauer]
397	hypothetical protein PP994_gp64 [Gordonia phage CherryonLim] >gb|QFP95817.1| hypothetical protein SEA_CHERRYONLIM_64 [Gordonia phage CherryonLim]
396	hypothetical protein PP993_gp67 [Gordonia phage Mayweather] >gb|QDP45228.1| hypothetical protein SEA_MAYWEATHER_67 [Gordonia phage Mayweather]

QBLAST Hit

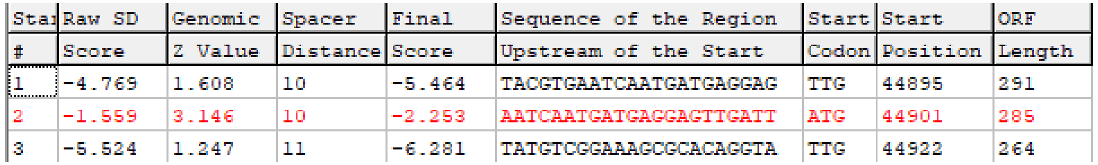
Accession YP_010663411
GI
Length 94
Max Score 495   Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	195.3	Identities	94
Score	495	%Identity	100.00
E-Value	0.0E0	Positives	94
Length	94	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 94		
Target	1 - 94		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

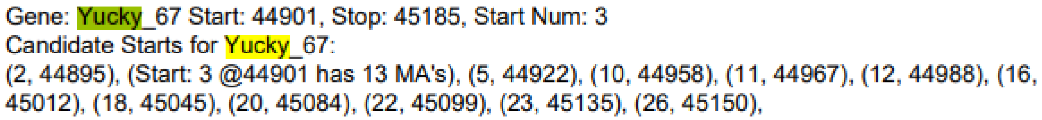
- Z value is the greatest with 3.146.
- Final score is the least negative with -2.253.



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	-4.769	1.608	10	-5.464	TACGTGAATCAATGATGAGGAG	TTG	44895	291
2	-1.559	3.146	10	-2.253	AATCAATGATGAGGAGTTGATT	ATG	44901	285
3	-5.524	1.247	11	-6.281	TATGTCGGAAAGCGCACAGGTA	TTG	44922	264

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

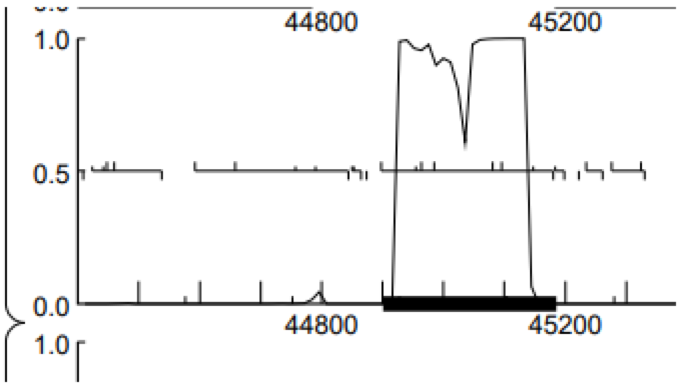
- There are 13 manual annotations.



Gene: Yucky_67 Start: 44901, Stop: 45185, Start Num: 3
Candidate Starts for Yucky_67:
(2, 44895), (Start: 3 @44901 has 13 MA's), (5, 44922), (10, 44958), (11, 44967), (12, 44988), (16, 45012), (18, 45045), (20, 45084), (22, 45099), (23, 45135), (26, 45150),

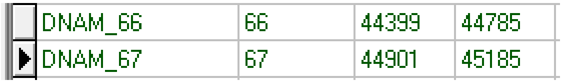
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- **Coding potential is included.**



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start?   Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- 44901-44785 = 116
- 116-1 = 115 gap



DNAM_66	66	44399	44785
DNAM_67	67	44901	45185

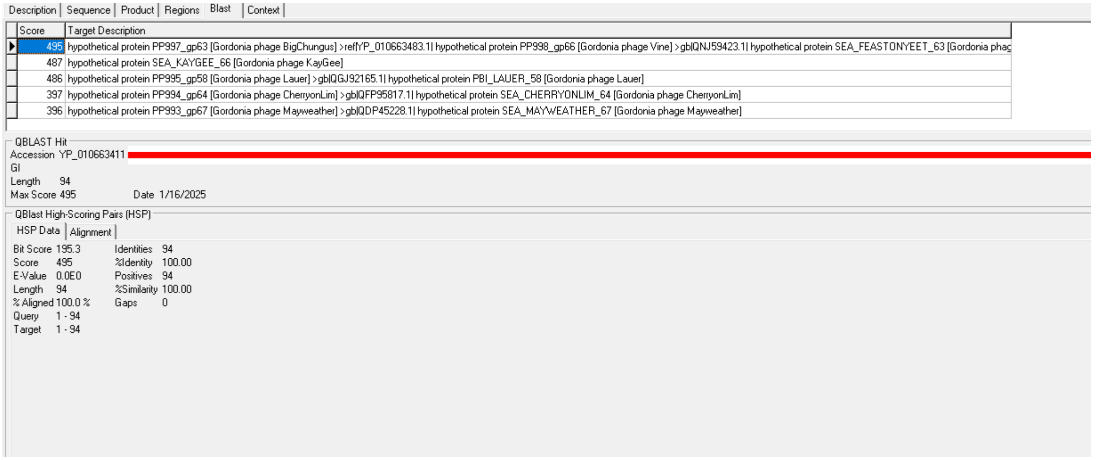
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	44901
GeneMark	Both Glimmer and GeneMark
Coding potential	Included
RBS score	Z value: 3.146 Final score: -2.253
BLAST	18 1:1 alignments
Starterator	13 MA’s
Gap/overlap	115 gap

There are no other start site suggestions. All evidence support for the autoannotated start site, except the gap. Gap is too much.

BLAST function evidence. What assigned functions do other highly similar genes have?

- They are all hypothetical protein.



Score	Target Description
495	hypothetical protein PP997_gp63 [Gordonia phage BigChungus] >ref|YP_010663483.1| hypothetical protein PP998_gp66 [Gordonia phage Vine] >gb|QNJ59423.1| hypothetical protein SEA_FEASTONYEET_63 [Gordonia phag
487	hypothetical protein SEA_KAYGEE_66 [Gordonia phage KayGee]
486	hypothetical protein PP995_gp58 [Gordonia phage Lauer] >gb|QGJ92165.1| hypothetical protein PBI_LAUER_58 [Gordonia phage Lauer]
397	hypothetical protein PP994_gp64 [Gordonia phage CherryonLim] >gb|QFP95817.1| hypothetical protein SEA_CHERRYONLIM_64 [Gordonia phage CherryonLim]
396	hypothetical protein PP993_gp67 [Gordonia phage Mayweather] >gb|QDP45228.1| hypothetical protein SEA_MAYWEATHER_67 [Gordonia phage Mayweather]

QBLAST Hit

Accession YP_010663411

GI

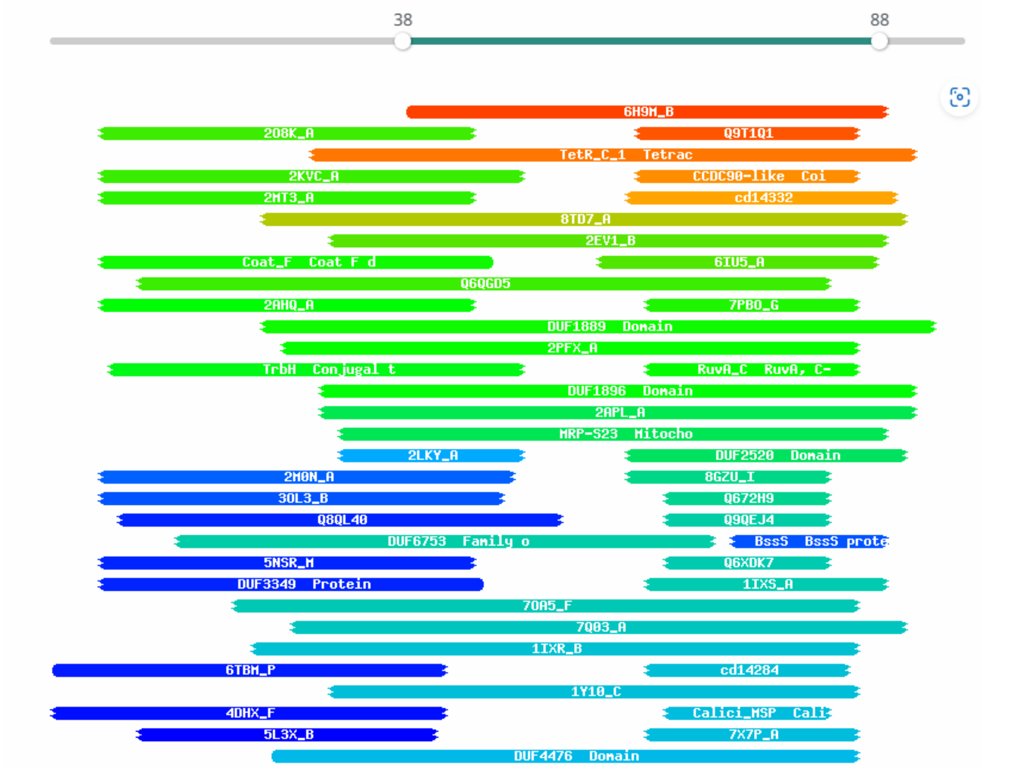
Length 94

Max Score 495 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	195.3	Identities	94
Score	495	%Identity	100.00
E-Value	0.0E0	Positives	94
Length	94	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 94		
Target	1 - 94		

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.



Nr	Hit	Name	Probability	E-value	Score	SS	cols	Length
□ 1	6H9M_B	Coiled-coil domain-containing protein 90B, mitochondrial,General control protein GCN4; coiled coil, beta-layer, mitochon	94.54	0.13	35.31	3.4	51	118
□ 2	Q9T1Q1	VP47_BPAPS Putative protein p47 OS=Acyrthosiphon pisum secondary endosymbiont phage 1 OX=67571 GN=47 PE=4 SV=1	92.79	0.48	34.34	3.8	24	190

- There are 2 hits with probability greater than 90.

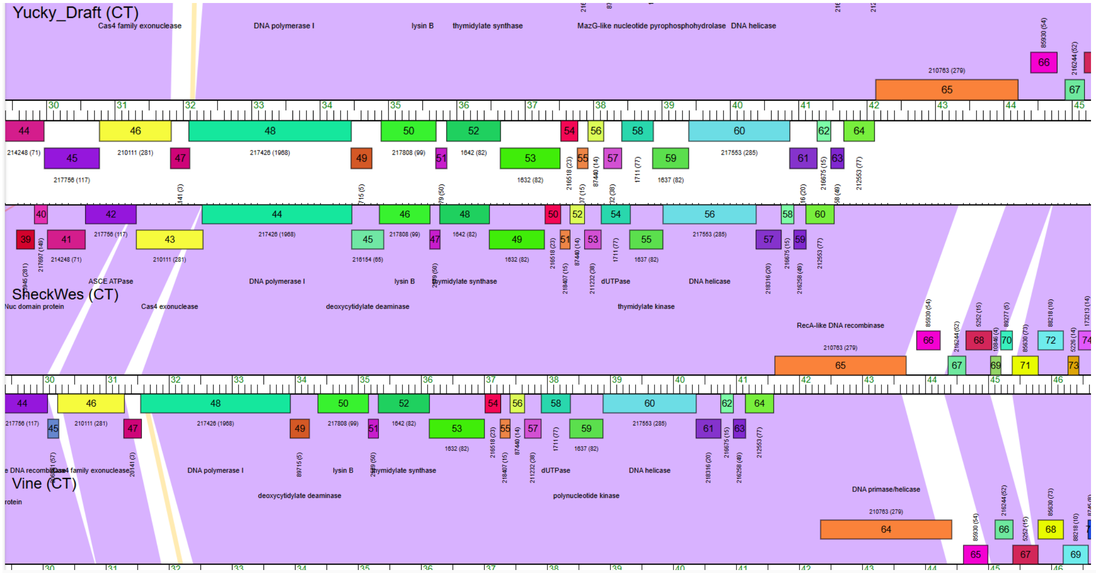
- One is a coiled-coil domain.
- One is a putative protein.

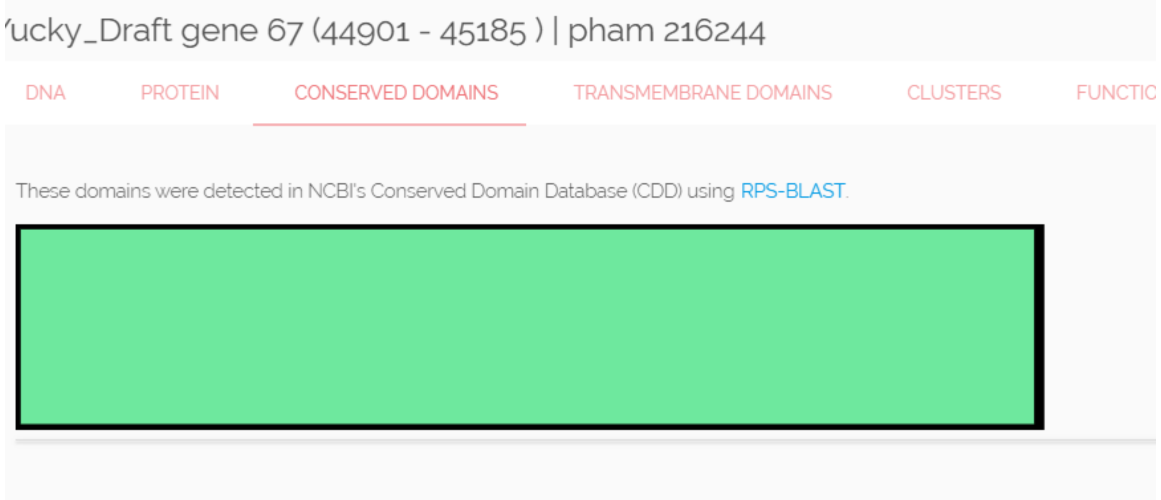
Though there are nothing called like that in the official function list.

Therefore, hypothetical protein.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

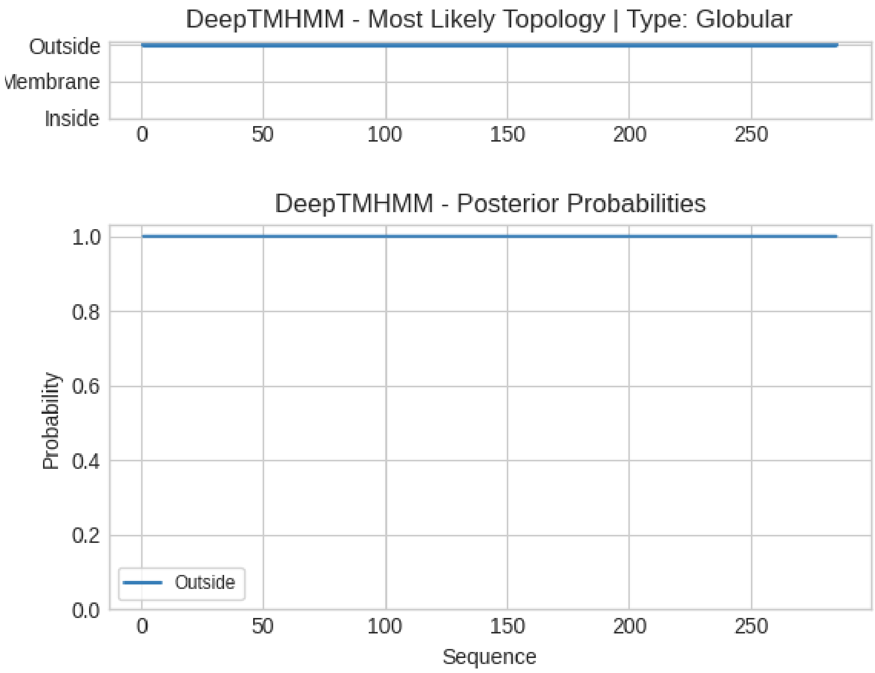
There are no functions assigned to highly similar genes in the same pham. Therefore, hypothetical protein.




ucky_Draft gene 67 (44901 - 45185) | pham 216244

These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- Gene 67 has no transmembrane domains
- So hypothetical protein.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

Gene 67 is a hypothetical protein because:

- All highly similar genes in BLAST are hypothetical protein.
- Two hits with probability greater than 90 are assigned a function, but the functions are not in the official function list.
- Highly similar genes in the same pham are not assigned a function.
- This gene has no transmembrane domains.

Feature 67 – Stop 45583

Glimmer/GeneMark

What feature number is this?

- 67

What is the stop site?

- 45583

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

- Both Glimmer and GeneMark

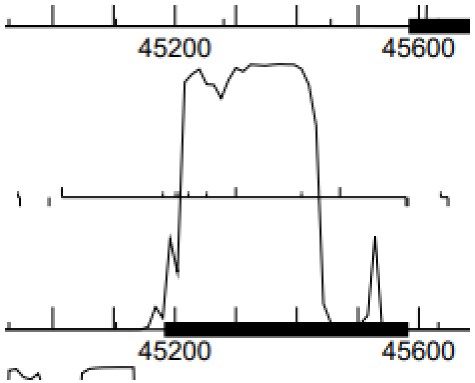
What is the autoannotated start?

- 45182

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

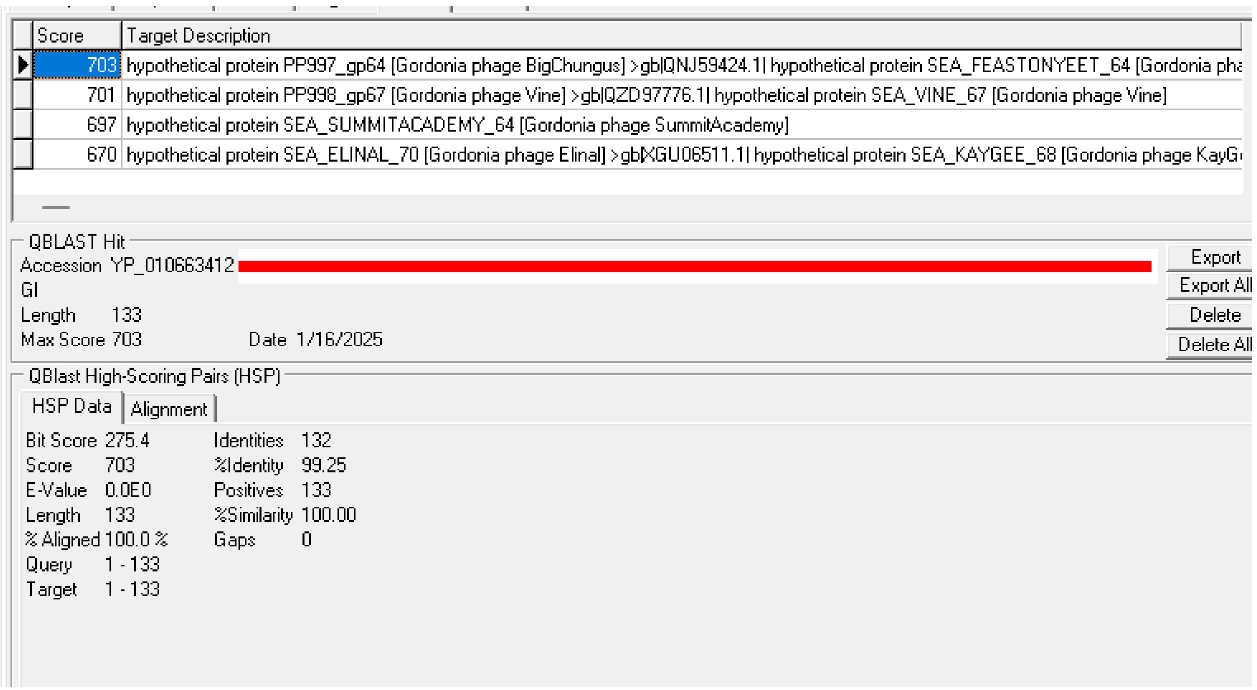
- 4 overlap.

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?



- Reading frame 2 shows a strong coding potential in the proximal of the autoannotate start site.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



Score	Target Description
703	hypothetical protein PP997_gp64 [Gordonia phage BigChungus] >gb|QNJ59424.1| hypothetical protein SEA_FEASTONYEET_64 [Gordonia pha
701	hypothetical protein PP998_gp67 [Gordonia phage Vine] >gb|QZD97776.1| hypothetical protein SEA_VINE_67 [Gordonia phage Vine]
697	hypothetical protein SEA_SUMMITACADEMY_64 [Gordonia phage SummitAcademy]
670	hypothetical protein SEA_ELINAL_70 [Gordonia phage Elinal] >gb|XGU06511.1| hypothetical protein SEA_KAYGEE_68 [Gordonia phage KayG

QBLAST Hit

Accession YP_010663412

GI

Length 133

Max Score 703 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	275.4	Identities	132
Score	703	%Identity	99.25
E-Value	0.0E0	Positives	133
Length	133	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 133		
Target	1 - 133		

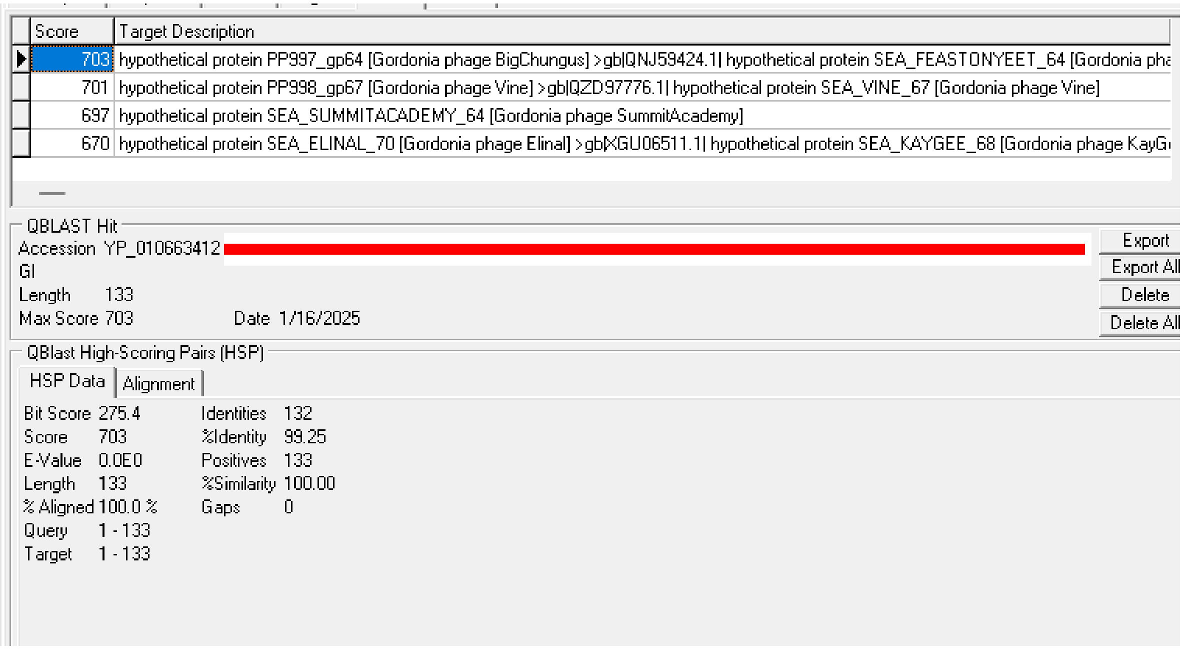
- There are 11 highly similar genes with e value of close to 0.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- It is a gene because:
- Both Glimmer and GeneMark call it a gene.
- The coding potential close to the autoannotated start site is strong.
- There are 11 highly similar genes with E value of close to 0.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- **There are 5 1:1 alignments.**



Score	Target Description
703	hypothetical protein PP997_gp64 [Gordonia phage BigChungus] >gb|QNJ59424.1| hypothetical protein SEA_FEASTONYEET_64 [Gordonia pha
701	hypothetical protein PP998_gp67 [Gordonia phage Vine] >gb|QZD97776.1| hypothetical protein SEA_VINE_67 [Gordonia phage Vine]
697	hypothetical protein SEA_SUMMITACADEMY_64 [Gordonia phage SummitAcademy]
670	hypothetical protein SEA_ELINAL_70 [Gordonia phage Elinal] >gb|XGU06511.1| hypothetical protein SEA_KAYGEE_68 [Gordonia phage KayG

QBLAST Hit

Accession YP_010663412

GI

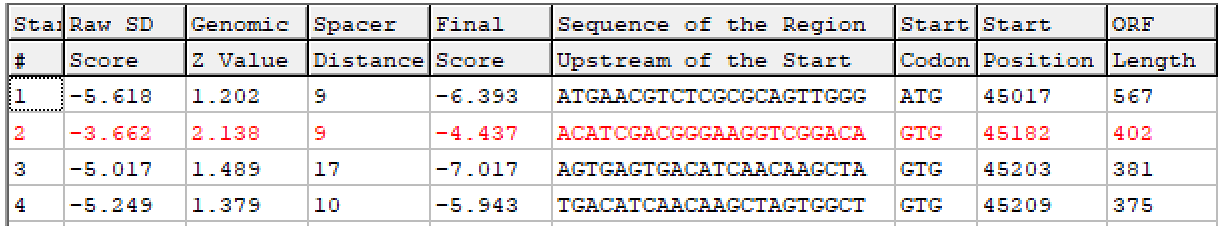
Length 133

Max Score 703 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	275.4	Identities	132
Score	703	%Identity	99.25
E-Value	0.0E0	Positives	133
Length	133	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 133		
Target	1 - 133		

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



Star #	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	-5.618	1.202	9	-6.393	ATGAACGTCTCGCGCAGTTGGG	ATG	45017	567
2	-3.662	2.138	9	-4.437	ACATCGACGGGAAGGTCGGACA	GTG	45182	402
3	-5.017	1.489	17	-7.017	AGTGAGTGACATCAACAAGCTA	GTG	45203	381
4	-5.249	1.379	10	-5.943	TGACATCAACAAGCTAGTGGCT	GTG	45209	375

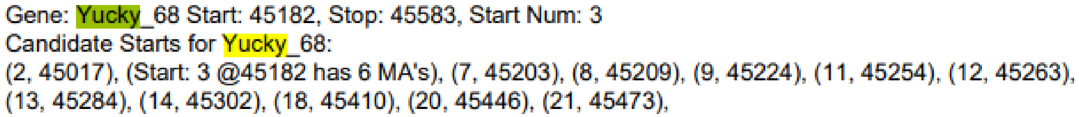
- The z value is the greatest with 2.138

The final score is the least negative with -4.437

- RBS score favors the autoannotated start site.

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- There are 6 MA’s

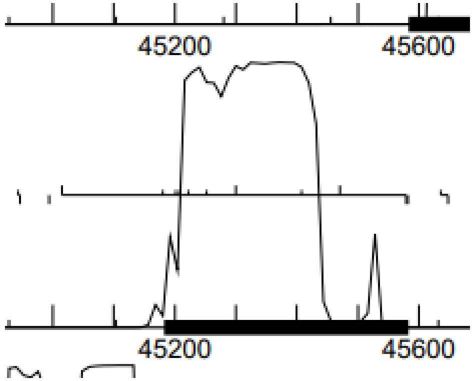


Gene: Yucky_68 Start: 45182, Stop: 45583, Start Num: 3
Candidate Starts for Yucky_68:
(2, 45017), (Start: 3 @45182 has 6 MA's), (7, 45203), (8, 45209), (9, 45224), (11, 45254), (12, 45263),
(13, 45284), (14, 45302), (18, 45410), (20, 45446), (21, 45473),

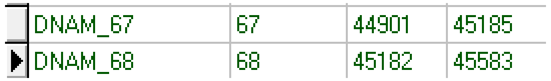
GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Coding potential is included.

- Autoannotated start site: 45182



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.



DNAM_67	67	44901	45185
DNAM_68	68	45182	45583

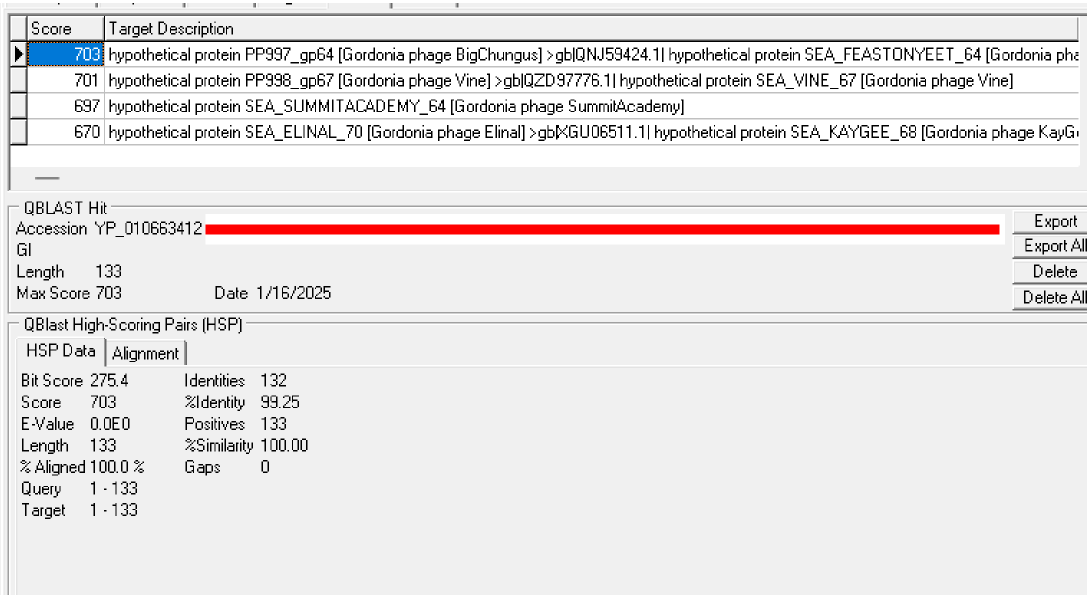
- 45185-45182 = 3
- 3+1 = 4 overlap.

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	45182
GeneMark	Both Glimmer and GeneMark
Coding potential	Included
RBS score	Z value: 2.138 Final score: -4.437
BLAST	5 1:1 alignments
Starterator	6
Gap/overlap	4 overlap

Autoannotated start site at 45182 is a start site because all evidence support it with a favorable overlap.

BLAST function evidence. What assigned functions do other highly similar genes have?



Score	Target Description
703	hypothetical protein PP997_gp64 [Gordonia phage BigChungus] >gb|QNJ59424.1| hypothetical protein SEA_FEASTONYEET_64 [Gordonia pha
701	hypothetical protein PP998_gp67 [Gordonia phage Vine] >gb|QZD97776.1| hypothetical protein SEA_VINE_67 [Gordonia phage Vine]
697	hypothetical protein SEA_SUMMITACADEMY_64 [Gordonia phage SummitAcademy]
670	hypothetical protein SEA_ELINAL_70 [Gordonia phage Elinal] >gb|XGU06511.1| hypothetical protein SEA_KAYGEE_68 [Gordonia phage KayG

QBLAST Hit

Accession YP_010663412
GI
Length 133
Max Score 703 Date 1/16/2025

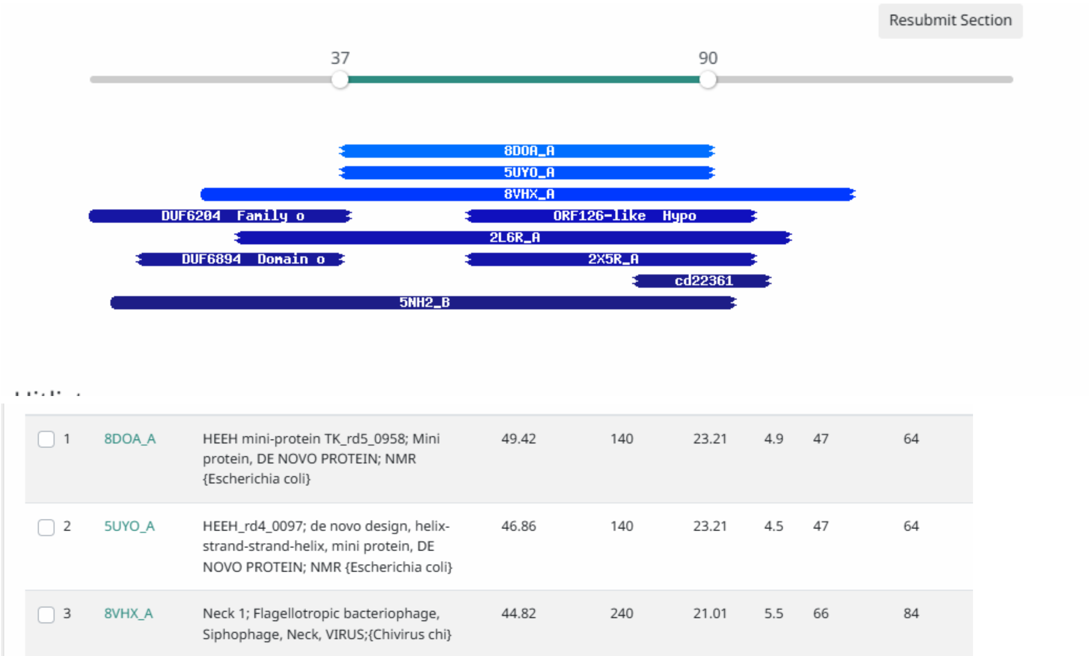
QBlast High-Scoring Pairs (HSP)

Bit Score	275.4	Identities	132
Score	703	%Identity	99.25
E-Value	0.0E0	Positives	133
Length	133	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 133		
Target	1 - 133		

- All highly similar genes ae hypothetical protein.

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

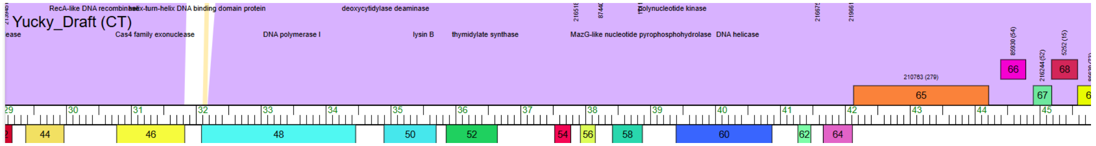
- There are no hits with probability greater than 90.

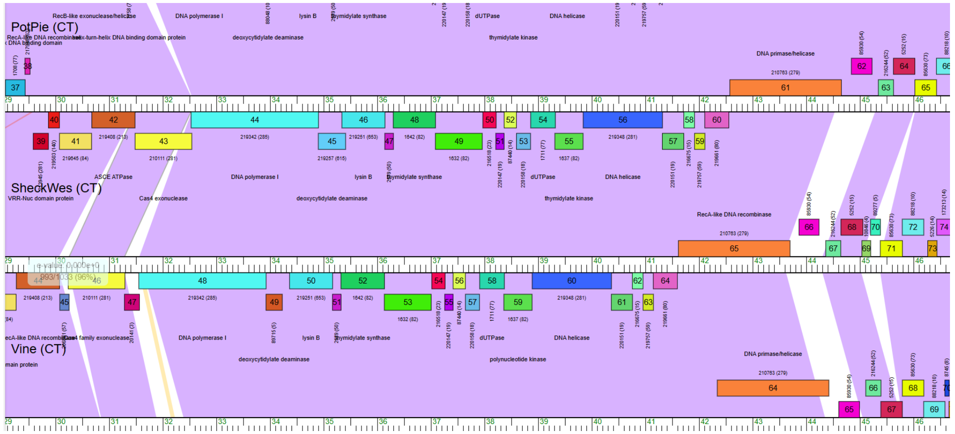


□ 1	8DOA_A	HEEH mini-protein TK_rd5_0958; Mini protein, DE NOVO PROTEIN; NMR {Escherichia coli}	49.42	140	23.21	4.9	47	64
□ 2	5UYO_A	HEEH_rd4_0097; de novo design, helix-strand-strand-helix, mini protein, DE NOVO PROTEIN; NMR {Escherichia coli}	46.86	140	23.21	4.5	47	64
□ 3	8VHX_A	Neck 1; Flagellotropic bacteriophage, Siphophage, Neck, VIRUS;{Chivirus chi}	44.82	240	21.01	5.5	66	84

- So, hypothetical protein.

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

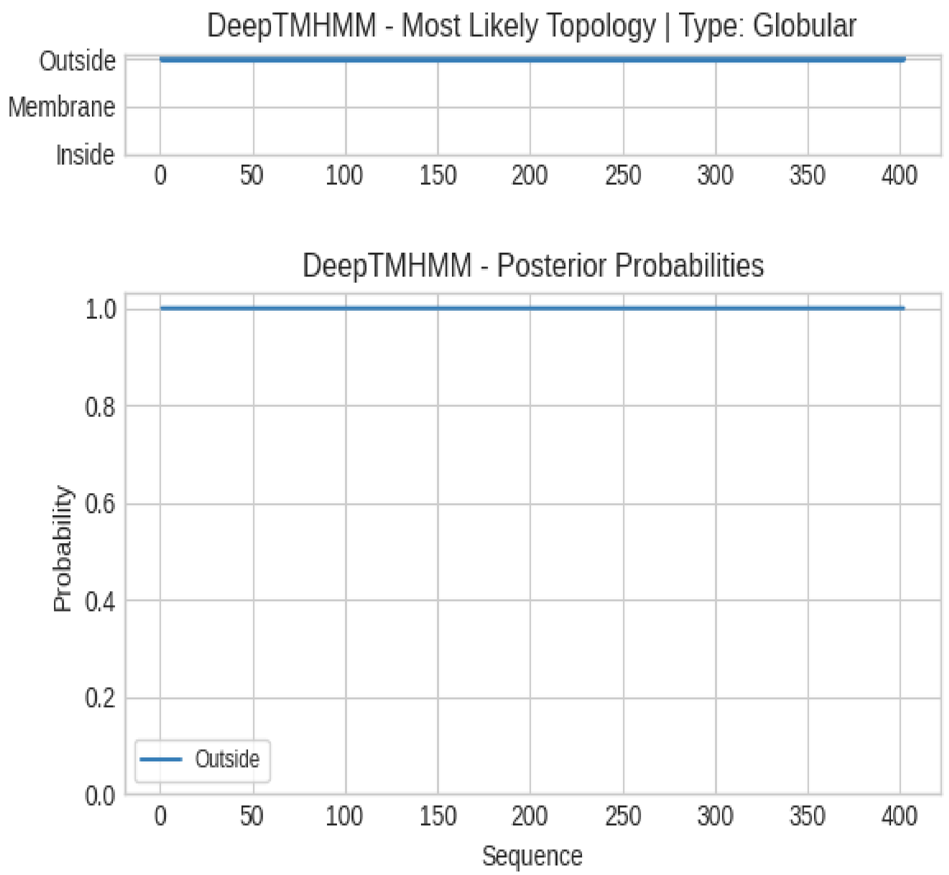




- The other highly similar genes in the same pham do not have a function assigned.

- There are no conserved domains.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



- It has no transmembrane domains.
- So hypothetical protein.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- This gene is a hypothetical protein because:

- All highly similar genes in BLAST are hypothetical gene.
- There are no hits with probability greater than 90.
- The highly similar genes in the same pham are not assigned a function.
- This protein has no transmembrane domains.

Feature 68 – Stop 46002

Glimmer/GeneMark

What feature number is this? 68

What is the stop site?

46002

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

Called by Glimmer and GeneMark

What is the autoannotated start?

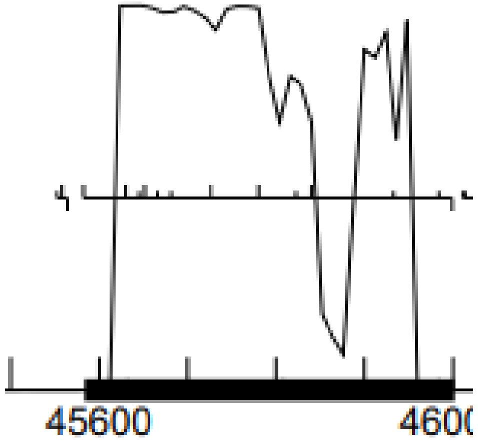
45583

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

There would be an overlap of 1

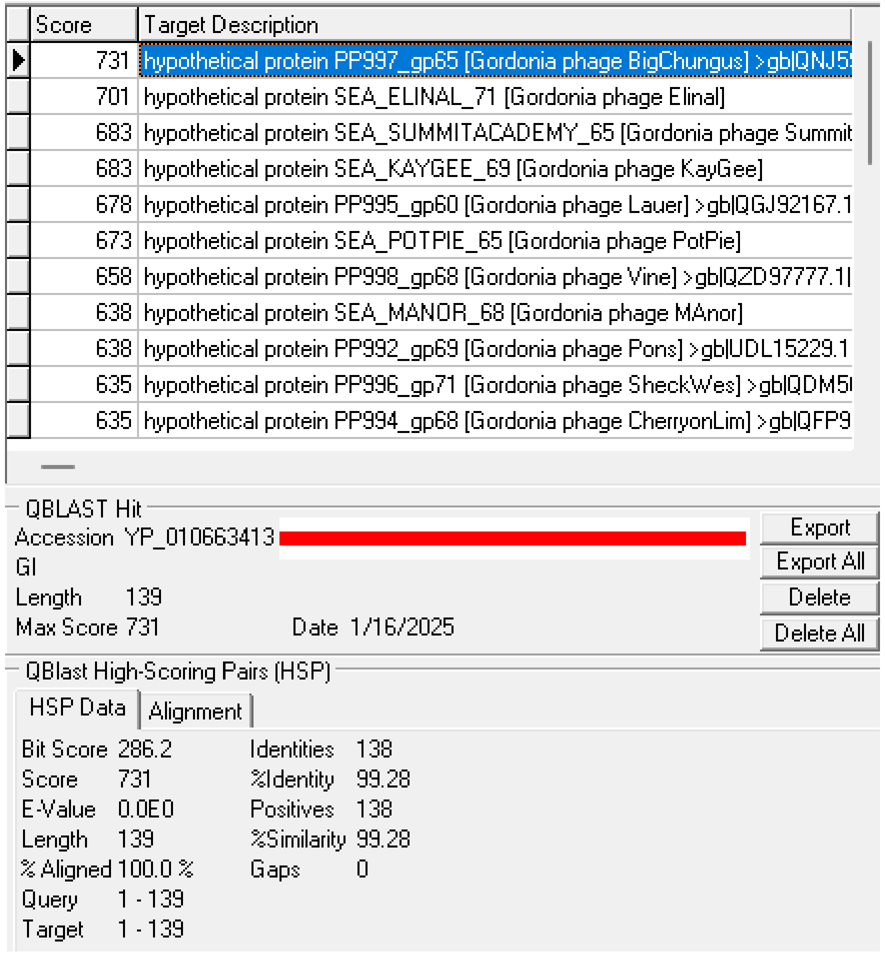
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- There is strong coding potential throughout where the feature is called to be with a dip into weak coding potential occurring between 45890 and 45840.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are at least 25 hits of phages with genes highly similar genes to this one.
- All e-values are extremely close to zero
- 4 of those hits are 1:1 alignments



Score	Target Description
731	hypothetical protein PP997_gp65 [Gordonia phage BigChungus] >gb|QNJ5
701	hypothetical protein SEA_ELINAL_71 [Gordonia phage Elinal]
683	hypothetical protein SEA_SUMMITACADEMY_65 [Gordonia phage Summit
683	hypothetical protein SEA_KAYGEE_69 [Gordonia phage KayGee]
678	hypothetical protein PP995_gp60 [Gordonia phage Lauer] >gb|QGJ92167.1
673	hypothetical protein SEA_POTPIE_65 [Gordonia phage PotPie]
658	hypothetical protein PP998_gp68 [Gordonia phage Vine] >gb|QZD97777.1|
638	hypothetical protein SEA_MANOR_68 [Gordonia phage MAnor]
638	hypothetical protein PP992_gp69 [Gordonia phage Pons] >gb|UDL15229.1
635	hypothetical protein PP996_gp71 [Gordonia phage SheckWes] >gb|QDM5
635	hypothetical protein PP994_gp68 [Gordonia phage CherryonLim] >gb|QFP9

QBLAST Hit

Accession YP_010663413

GI

Length 139

Max Score 731

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

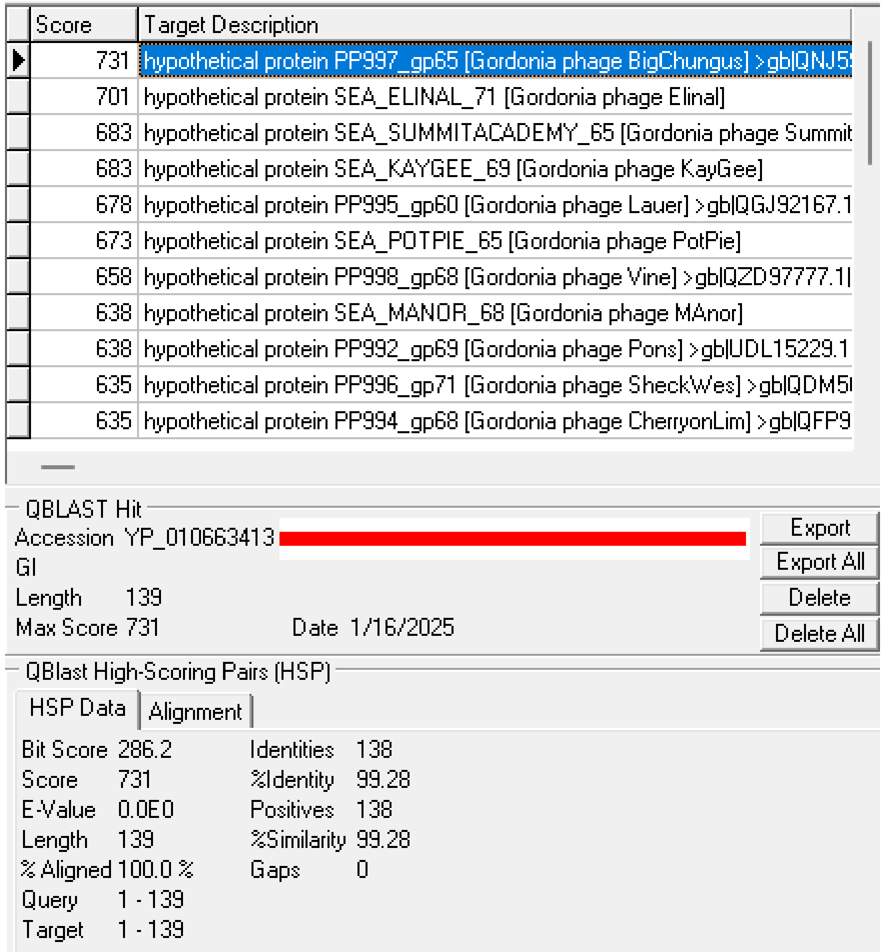
Bit Score	286.2	Identities	138
Score	731	%Identity	99.28
E-Value	0.0E0	Positives	138
Length	139	%Similarity	99.28
% Aligned	100.0 %	Gaps	0
Query	1 - 139		
Target	1 - 139		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- This feature is a gene! There is strong coding potential running throughout where the feature is called to be, and there are at least 25 BLAST hits of highly similar genes to this feature from other phages that all have e-values extremely close to zero.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There are at least 25 BLAST hits of highly similar genes from other phages that all have e-values extremely close to zero.
- There are 4 1:1 alignments for the gene starting at 45583



Score	Target Description
731	hypothetical protein PP997_gp65 [Gordonia phage BigChungus] >gb|QNJ5
701	hypothetical protein SEA_ELINAL_71 [Gordonia phage Elinal]
683	hypothetical protein SEA_SUMMITACADEMY_65 [Gordonia phage Summit
683	hypothetical protein SEA_KAYGEE_69 [Gordonia phage KayGee]
678	hypothetical protein PP995_gp60 [Gordonia phage Lauer] >gb|QGJ92167.1
673	hypothetical protein SEA_POTPIE_65 [Gordonia phage PotPie]
658	hypothetical protein PP998_gp68 [Gordonia phage Vine] >gb|QZD97777.1|
638	hypothetical protein SEA_MANOR_68 [Gordonia phage MAnor]
638	hypothetical protein PP992_gp69 [Gordonia phage Pons] >gb|UDL15229.1
635	hypothetical protein PP996_gp71 [Gordonia phage SheckWes] >gb|QDM5
635	hypothetical protein PP994_gp68 [Gordonia phage CherryonLim] >gb|QFP9

QBLAST Hit

Accession YP_010663413

GI

Length 139

Max Score 731

Date 1/16/2025

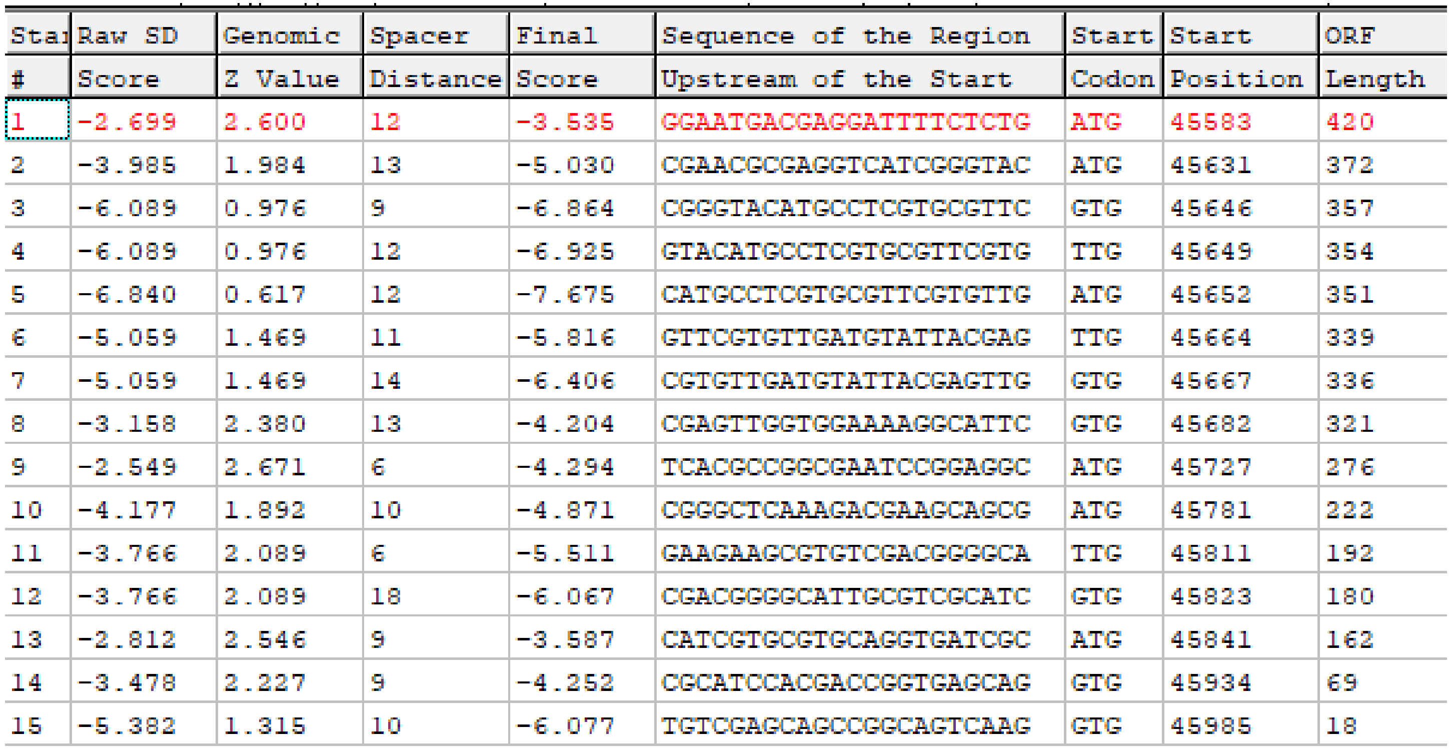
QBlast High-Scoring Pairs (HSP)

Bit Score 286.2
Score 731
E-Value 0.0E0
Length 139
% Aligned 100.0 %
Query 1 - 139
Target 1 - 139

Identities 138
%Identity 99.28
Positives 138
%Similarity 99.28
Gaps 0

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

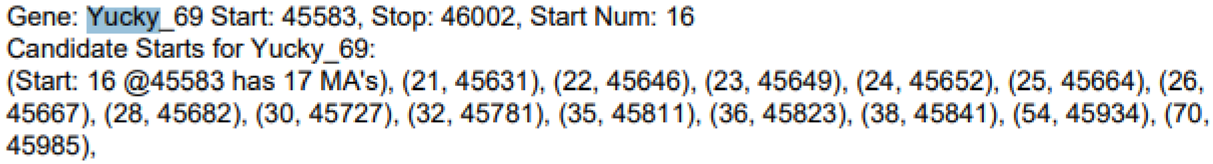
- Starting at 45583:
  - Z-value = 2.600
  - Final score = -3.535
- This start has the best RBS scores of all possible start sites.



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	-2.699	2.600	12	-3.535	GGAATGACGAGGATTTTCTCTG	ATG	45583	420
2	-3.985	1.984	13	-5.030	CGAACGCGAGGTCATCGGGTAC	ATG	45631	372
3	-6.089	0.976	9	-6.864	CGGGTACATGCCTCGTGCGTTC	GTG	45646	357
4	-6.089	0.976	12	-6.925	GTACATGCCTCGTGCGTTCGTG	TTG	45649	354
5	-6.840	0.617	12	-7.675	CATGCCTCGTGCGTTCGTGTTG	ATG	45652	351
6	-5.059	1.469	11	-5.816	GTTCGTGTTGATGTATTACGAG	TTG	45664	339
7	-5.059	1.469	14	-6.406	CGTGTTGATGTATTACGAGTTG	GTG	45667	336
8	-3.158	2.380	13	-4.204	CGAGTTGGTGGAAAAGGCATTC	GTG	45682	321
9	-2.549	2.671	6	-4.294	TCACGCCGGCGAATCCGGAGGC	ATG	45727	276
10	-4.177	1.892	10	-4.871	CGGGCTCAAAGACGAAGCAGCG	ATG	45781	222
11	-3.766	2.089	6	-5.511	GAAGAAGCGTGTCGACGGGGCA	TTG	45811	192
12	-3.766	2.089	18	-6.067	CGACGGGGCATTGCGTCGCATC	GTG	45823	180
13	-2.812	2.546	9	-3.587	CATCGTGCGTGCAGGTGATCGC	ATG	45841	162
14	-3.478	2.227	9	-4.252	CGCATCCACGACCGGTGAGCAG	GTG	45934	69
15	-5.382	1.315	10	-6.077	TGTCGAGCAGCCGGCAGTCAAG	GTG	45985	18

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

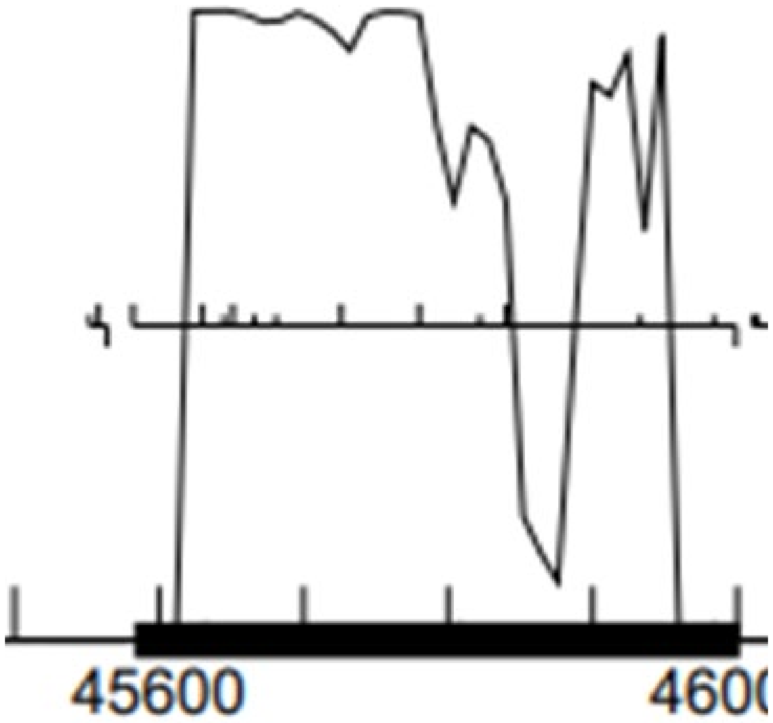
- There are 17 MA's for the gene starting at 45583.



Gene: Yucky_69 Start: 45583, Stop: 46002, Start Num: 16
Candidate Starts for Yucky_69:
(Start: 16 @45583 has 17 MA's), (21, 45631), (22, 45646), (23, 45649), (24, 45652), (25, 45664), (26, 45667), (28, 45682), (30, 45727), (32, 45781), (35, 45811), (36, 45823), (38, 45841), (54, 45934), (70, 45985),

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Starting at 45583 would include all the possible coding potential of the gene.



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

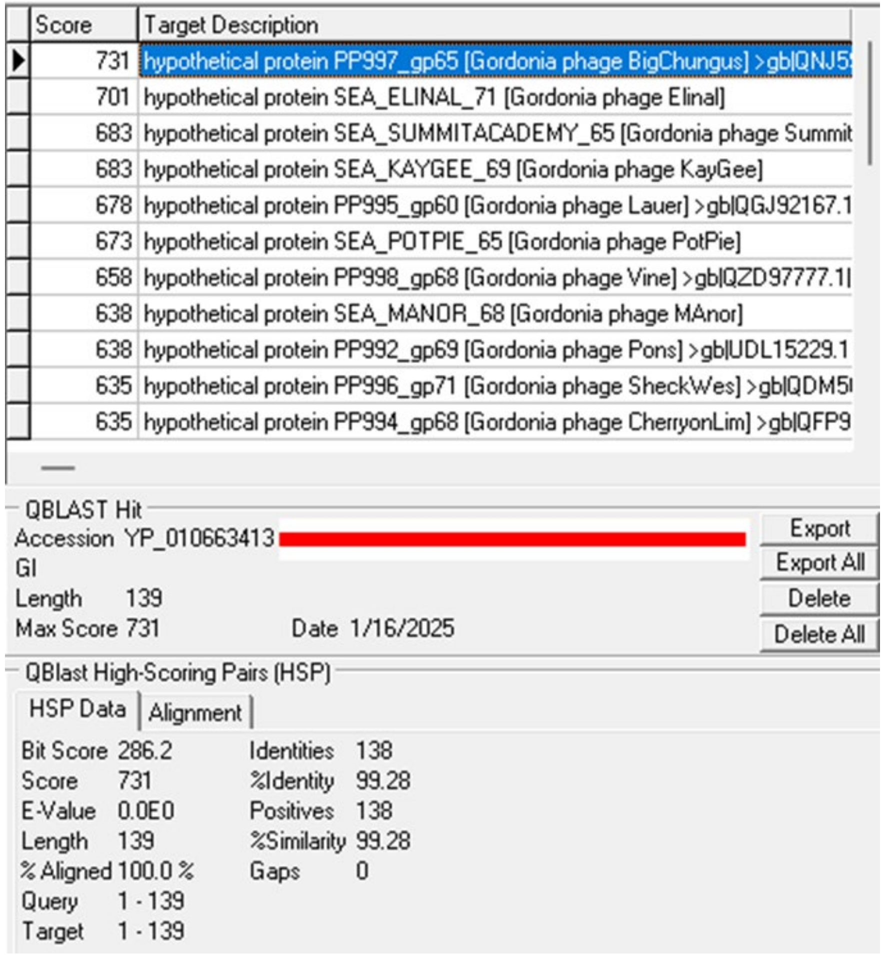
- Starting at 45583 would leave an overlap of 1 with the previous gene.

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is 45583! This was the only proposed start site based off all the evidence collected. There were 4 1:1 alignments according to BLAST with highly similar genes from other phages with this start site, and it also had the best RBS scores (z-value = 2.600 & final score = -3.353). The starterator report showed that 45583 is the only start site that had any manual annotation for which it had 17. Starting at 45583 would include all the possible coding potential of the gene, and there would by an overlap of only 1 nucleotide between this gene and the previous one which is favorable.

BLAST function evidence. What assigned functions do other highly similar genes have?

- There were at least 25 BLAST hits showing the function labeled as hypothetical protein.



Score	Target Description
731	hypothetical protein PP997_gp65 [Gordonia phage BigChungus] >gb|QNJ5
701	hypothetical protein SEA_ELINAL_71 [Gordonia phage Elinal]
683	hypothetical protein SEA_SUMMITACADEMY_65 [Gordonia phage Summit
683	hypothetical protein SEA_KAYGEE_69 [Gordonia phage KayGee]
678	hypothetical protein PP995_gp60 [Gordonia phage Lauer] >gb|QGJ92167.1
673	hypothetical protein SEA_POTPIE_65 [Gordonia phage PotPie]
658	hypothetical protein PP998_gp68 [Gordonia phage Vine] >gb|QZD97777.1|
638	hypothetical protein SEA_MANOR_68 [Gordonia phage MAnor]
638	hypothetical protein PP992_gp69 [Gordonia phage Pons] >gb|UDL15229.1
635	hypothetical protein PP996_gp71 [Gordonia phage SheckWes] >gb|QDM5
635	hypothetical protein PP994_gp68 [Gordonia phage CherryonLim] >gb|QFP9

QBLAST Hit

Accession YP_010663413

GI

Length 139

Max Score 731

Date 1/16/2025

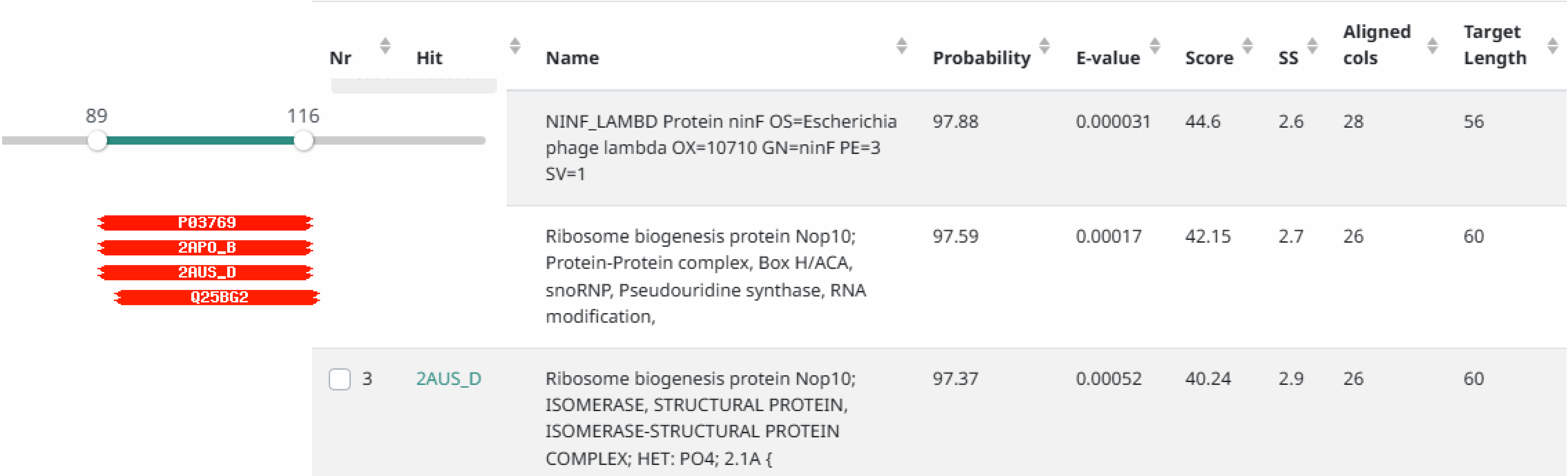
QBlast High-Scoring Pairs (HSP)

Bit Score	286.2	Identities	138
Score	731	%Identity	99.28
E-Value	0.0E0	Positives	138
Length	139	%Similarity	99.28
% Aligned	100.0 %	Gaps	0
Query	1 - 139		
Target	1 - 139		

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

There were hits with probabilities over 90 showing functions of ribosome biogenesis protein as well as another type of protein

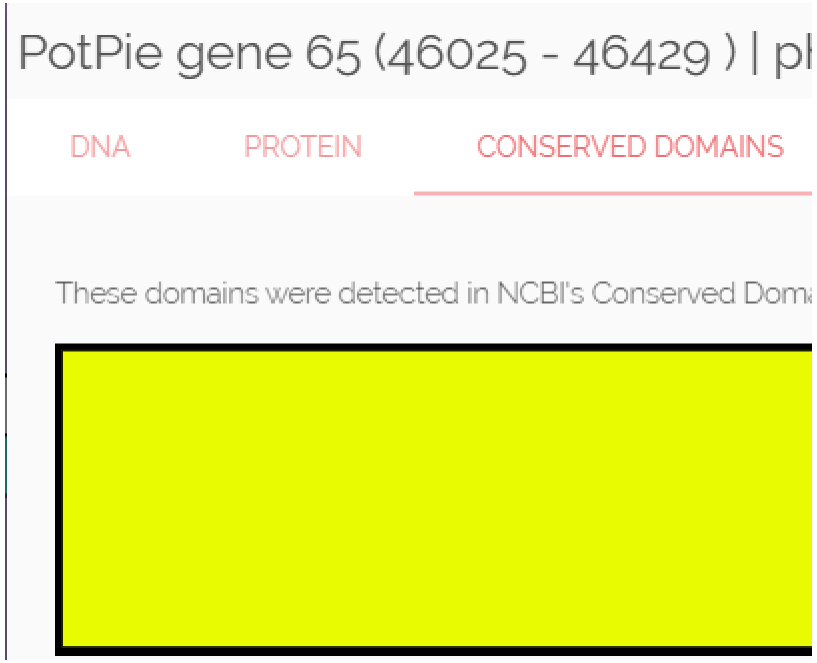
- The e-values for these hits were relatively large and they were only homologous for a small part of the gene.



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
		NINF_LAMBD Protein ninF OS=Escherichia phage lambda OX=10710 GN=ninF PE=3 SV=1	97.88	0.000031	44.6	2.6	28	56
		Ribosome biogenesis protein Nop10; Protein-Protein complex, Box H/ACA, snoRNP, Pseudouridine synthase, RNA modification,	97.59	0.00017	42.15	2.7	26	60
□ 3	2AUS_D	Ribosome biogenesis protein Nop10; ISOMERASE, STRUCTURAL PROTEIN, ISOMERASE-STRUCTURAL PROTEIN COMPLEX; HET: PO4; 2.1A {	97.37	0.00052	40.24	2.9	26	60

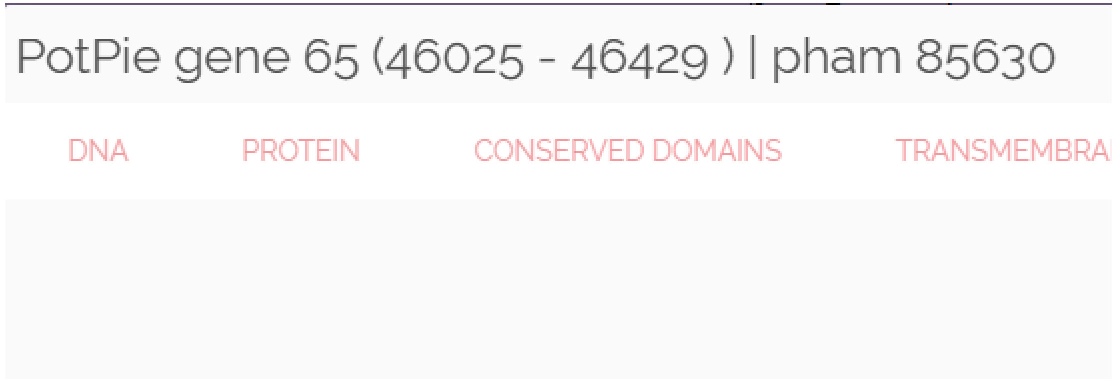
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Closely related phages with genes in the same pham did not predict a function assignment for this gene.
- They did not have assigned functions or conserved domains.



PotPie gene 65 (46025 - 46429) | pl

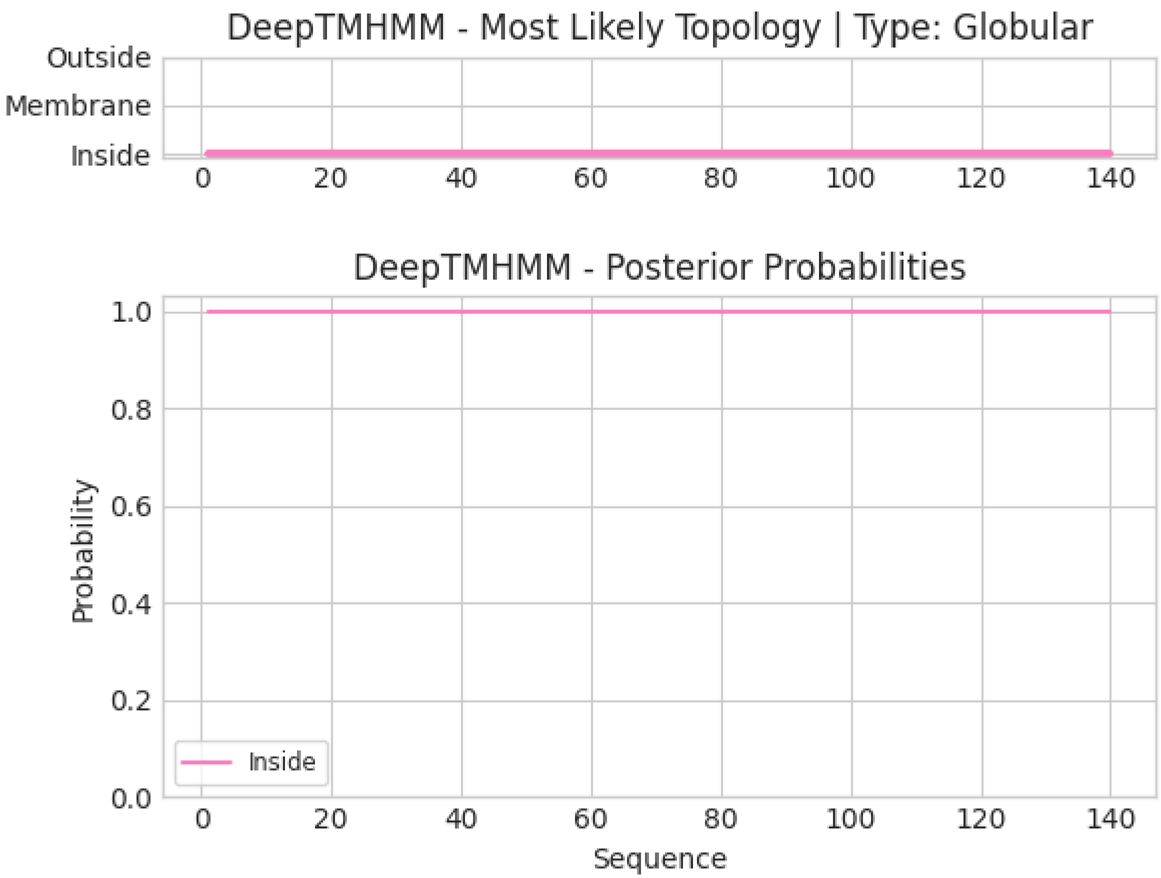
These domains were detected in NCBI's Conserved Dom



PotPie gene 65 (46025 - 46429) | pham 85630

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- The graph produced did not show any evidence of transmembrane domains.



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Official function → hypothetical protein
- The function for this gene should be labeled as hypothetical protein. There were at least 25 BLAST hits that showed functions labeled as hypothetical protein. Hhpred did show hits of suggested functions that had probabilities over 90, but they had high e-values and were only homologous for a small portion of the gene which did not justify the assignment of a specific function to the gene. Phamerator showed that phages with genes in the same pham did not have labeled functions or conserved domains which also did suggest a function to be assigned to this gene. The graph produced by Deep TMHMM did not show evidence of any transmembrane domains, so the function could not be labeled as a membrane protein either.

Feature 69 – Stop 46388

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

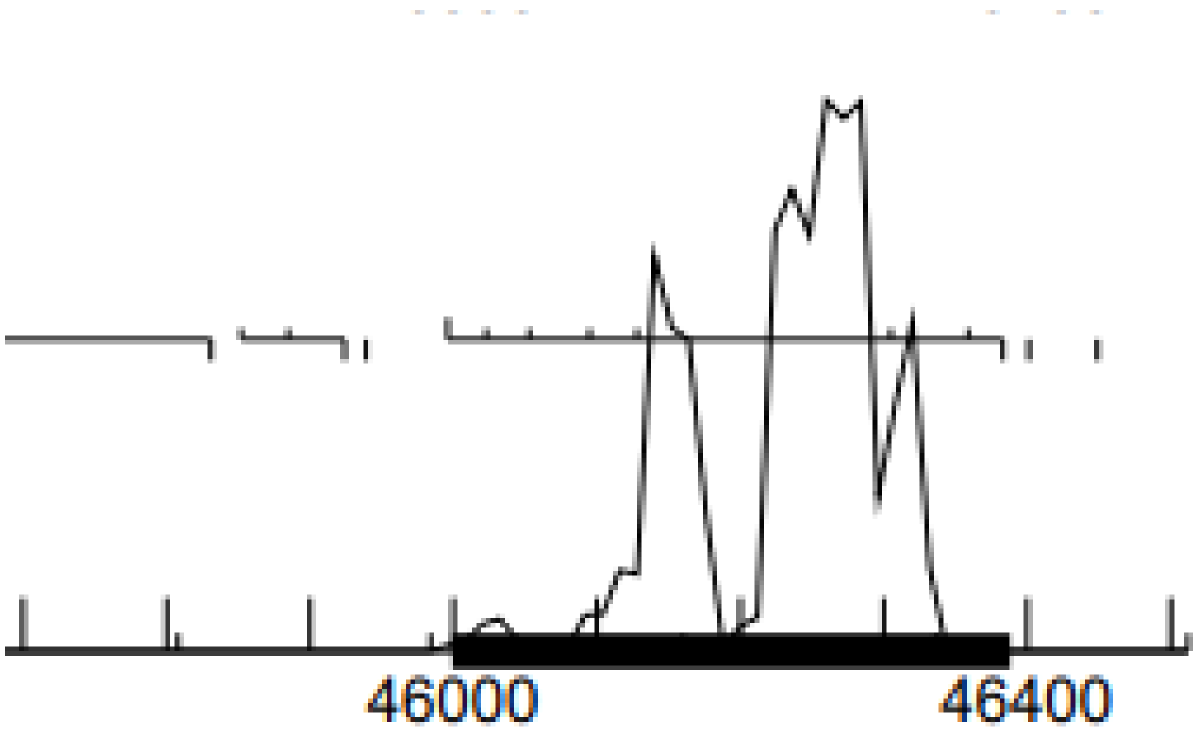
What is the autoannotated start?

Gap: ________ or overlap: _______ (with gene in front of it) for the autoannotated start

- Feature: 69
- Stop site: 46388
- Called by both Glimmer and GeneMark
- Autoannotated start: 45999
- Overlap: 4

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Start 45999
- Includes all cp
- Reading frame 3



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 9 highly similar genes

PotPie

Vine

SummitAcademy

Mayweather

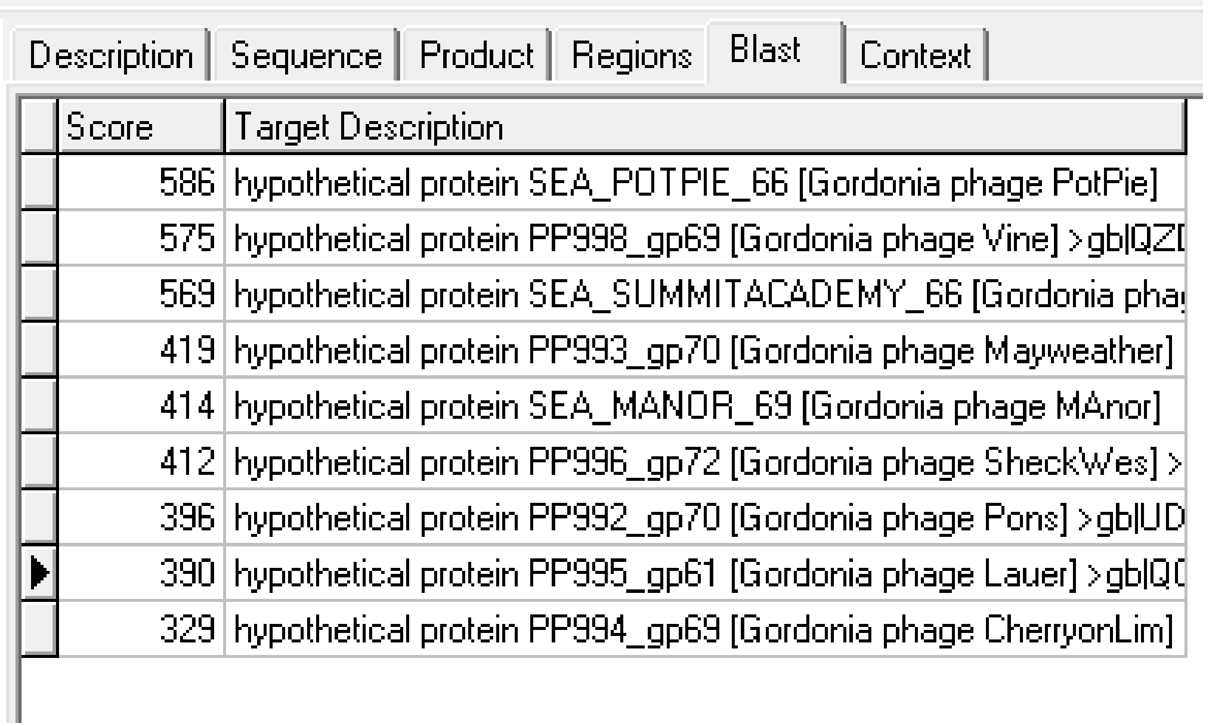
MAnor

SheckWes

Pons

Lauer

CherryonLim



	Score	Target Description
	586	hypothetical protein SEA_POTPIE_66 [Gordonia phage PotPie]
	575	hypothetical protein PP998_gp69 [Gordonia phage Vine] >gb|QZ
	569	hypothetical protein SEA_SUMMITACADEMY_66 [Gordonia pha
	419	hypothetical protein PP993_gp70 [Gordonia phage Mayweather]
	414	hypothetical protein SEA_MANOR_69 [Gordonia phage MAnor]
	412	hypothetical protein PP996_gp72 [Gordonia phage SheckWes] >
	396	hypothetical protein PP992_gp70 [Gordonia phage Pons] >gb|UD
▶	390	hypothetical protein PP995_gp61 [Gordonia phage Lauer] >gb|Q
	329	hypothetical protein PP994_gp69 [Gordonia phage CherryonLim]

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes it is a gene because both Glimmer and GeneMark call the same start site, the frame includes all coding potential and it has 9 highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

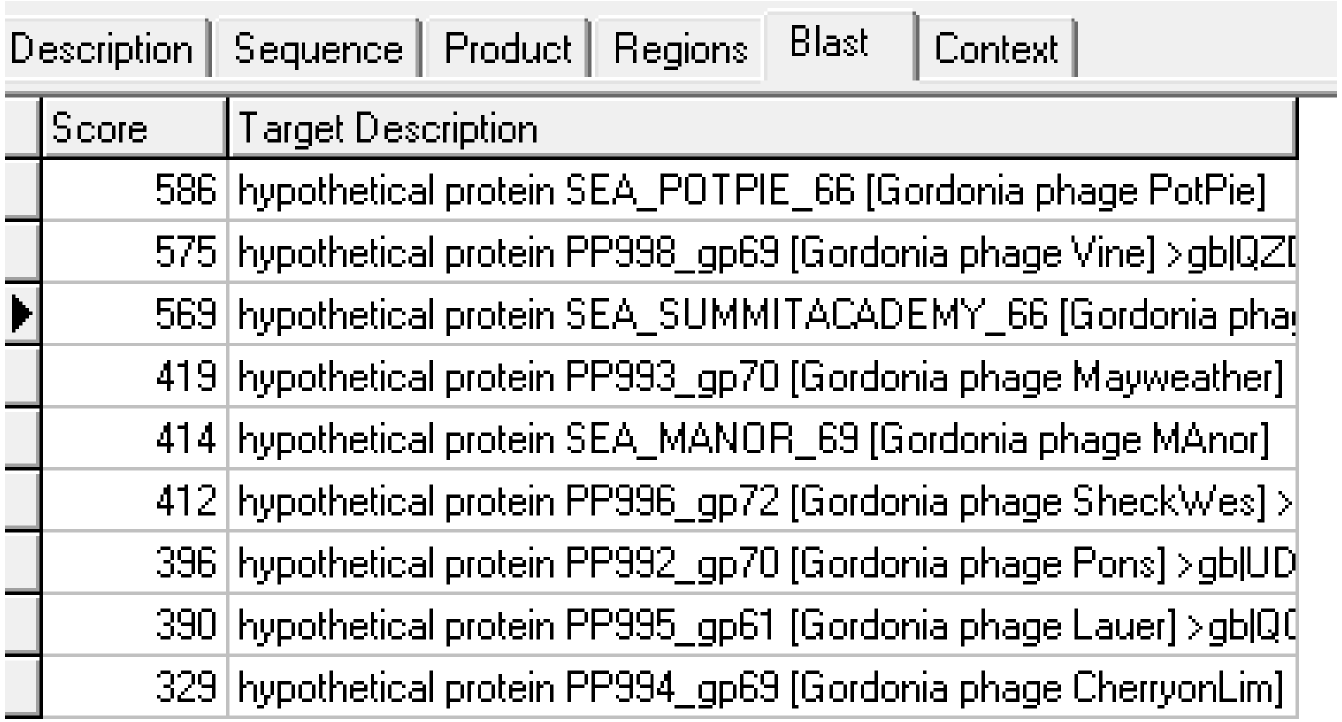
- 4 1:1 alignments

PotPie

Vine

Lauer

SummitAcademy



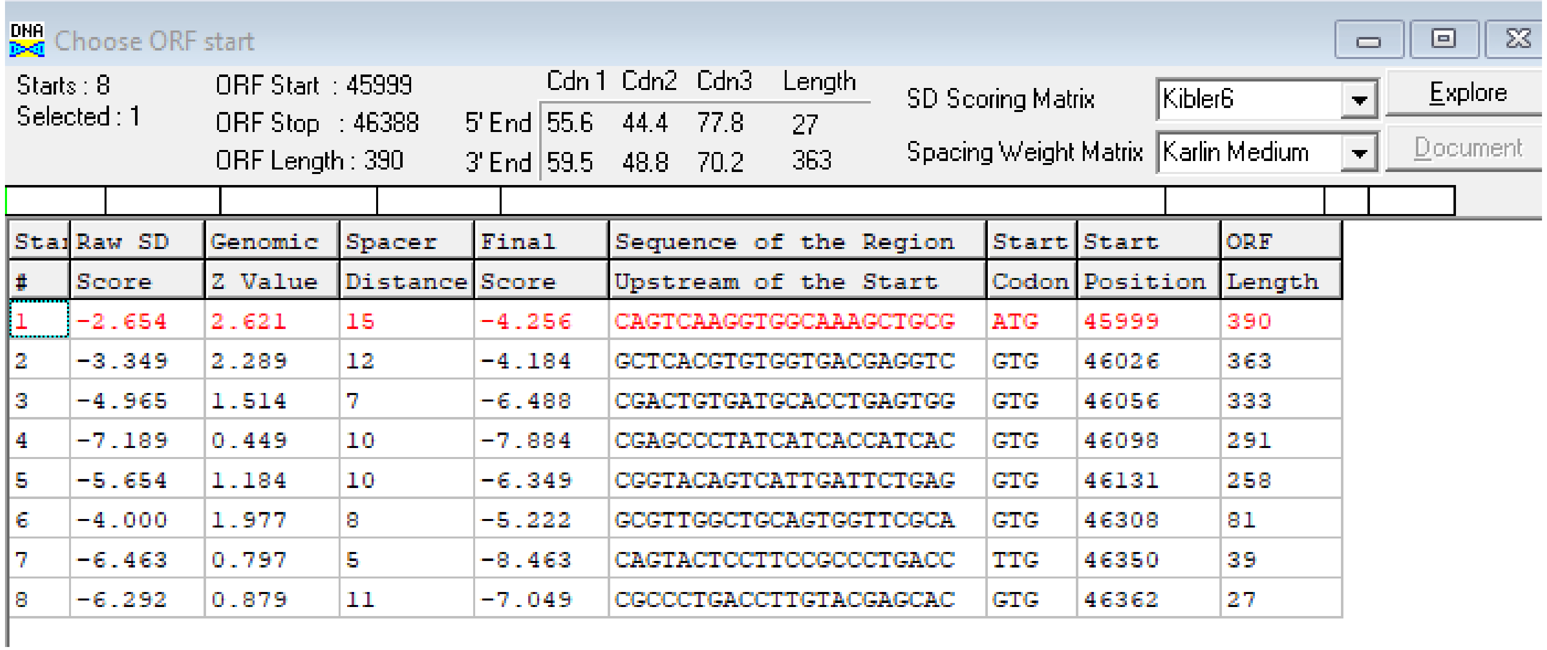
Score	Target Description
586	hypothetical protein SEA_POTPIE_66 [Gordonia phage PotPie]
575	hypothetical protein PP998_gp69 [Gordonia phage Vine] >gb|QZD
569	hypothetical protein SEA_SUMMITACADEMY_66 [Gordonia pha
419	hypothetical protein PP993_gp70 [Gordonia phage Mayweather]
414	hypothetical protein SEA_MANOR_69 [Gordonia phage MAnor]
412	hypothetical protein PP996_gp72 [Gordonia phage SheckWes] >
396	hypothetical protein PP992_gp70 [Gordonia phage Pons] >gb|UD
390	hypothetical protein PP995_gp61 [Gordonia phage Lauer] >gb|QC
329	hypothetical protein PP994_gp69 [Gordonia phage CherryonLim]

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Start 45999

Z value: 2.621

Final score: -4.256



Starts : 8
Selected : 1

ORF Start : 45999
ORF Stop : 46388
ORF Length : 390

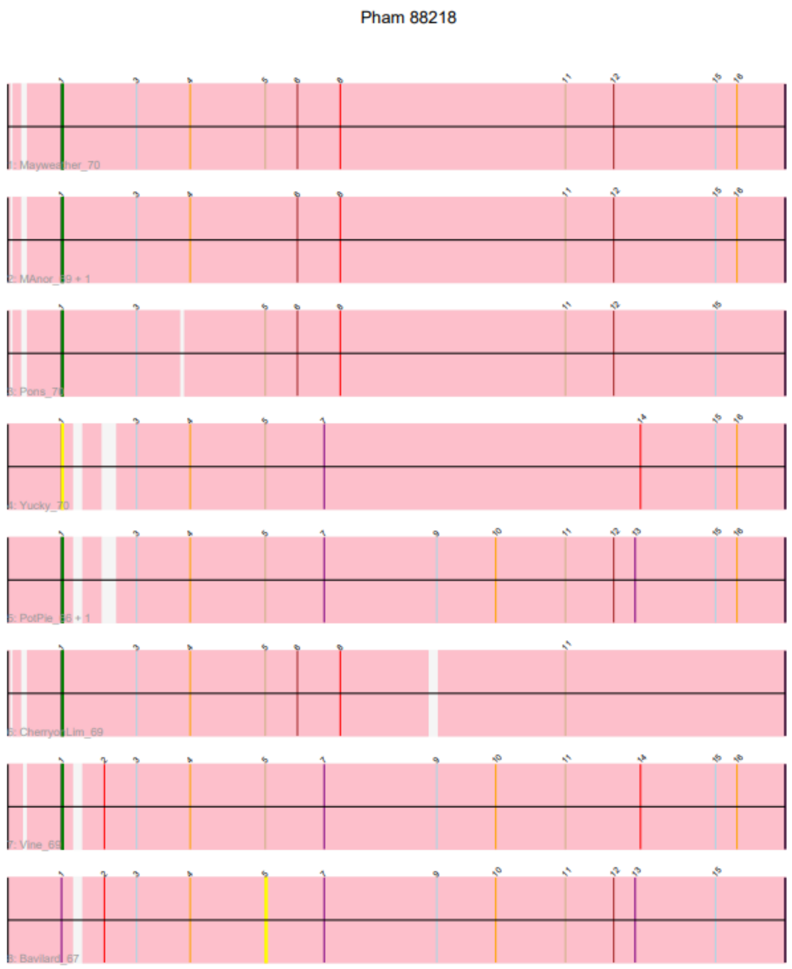
	Cdn 1	Cdn2	Cdn3	Length
5' End	55.6	44.4	77.8	27
3' End	59.5	48.8	70.2	363

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-2.654	2.621	15	-4.256	CAGTCAAGGTGGCAAAGCTGCG	ATG	45999	390
2	-3.349	2.289	12	-4.184	GCTCACGTGTGGTGACGAGGTC	GTG	46026	363
3	-4.965	1.514	7	-6.488	CGACTGTGATGCACCTGAGTGG	GTG	46056	333
4	-7.189	0.449	10	-7.884	CGAGCCCTATCATCACCATCAC	GTG	46098	291
5	-5.654	1.184	10	-6.349	CGGTACAGTCATTGATTCTGAG	GTG	46131	258
6	-4.000	1.977	8	-5.222	GCGTTGGCTGCAGTGGTTCGCA	GTG	46308	81
7	-6.463	0.797	5	-8.463	CAGTACTCCTTCCGCCCTGACC	TTG	46350	39
8	-6.292	0.879	11	-7.049	CGCCCTGACCTTGTACGAGCAC	GTG	46362	27

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

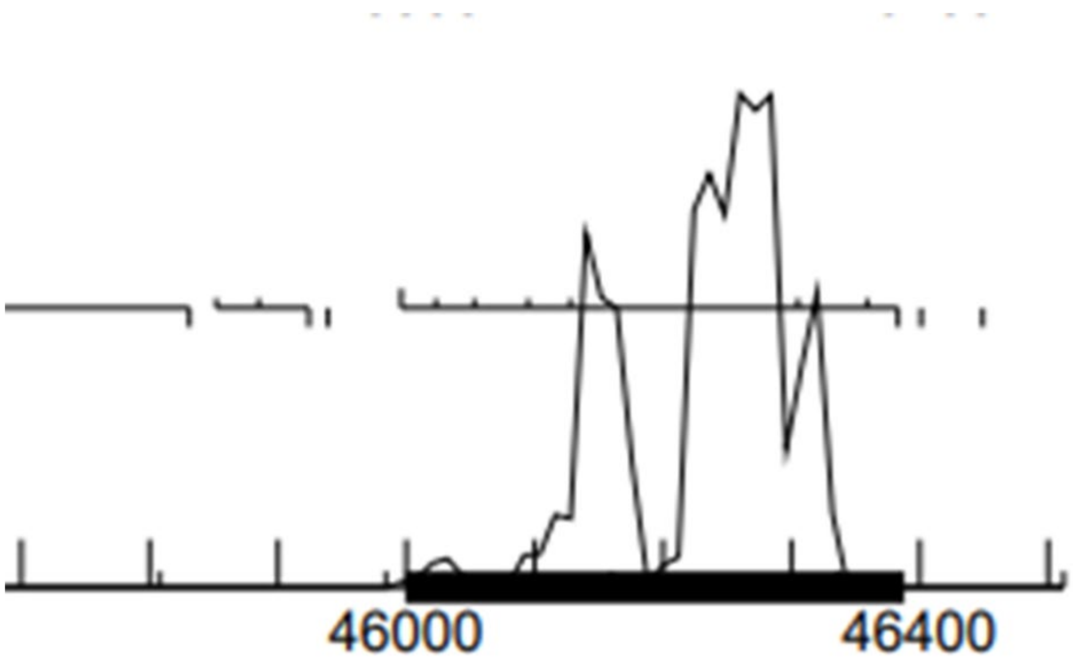
- Start: 1 @45999 has 8 MAs



GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Start 45999

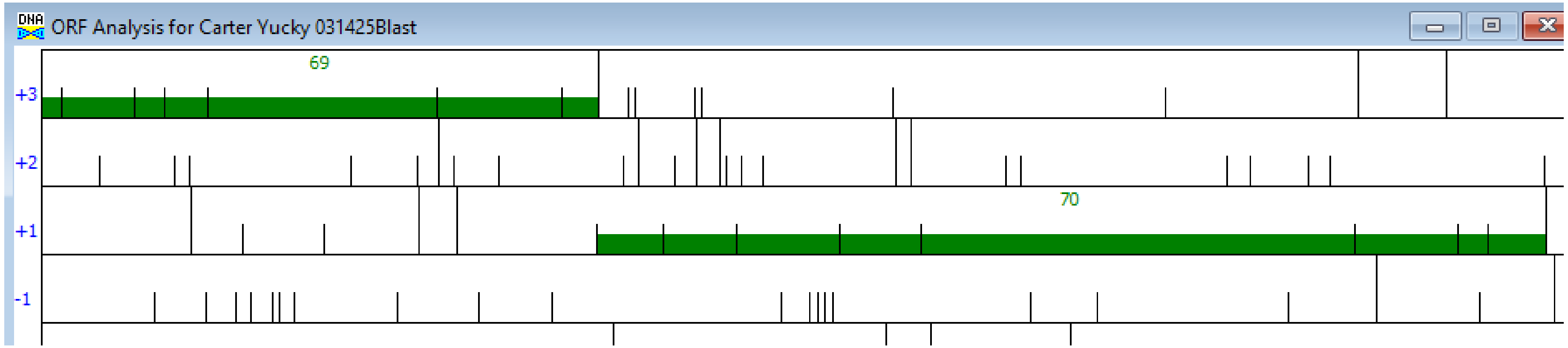
Includes all cp



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

Start 45999 – Previous gene ends at 46002

Overlap: 4



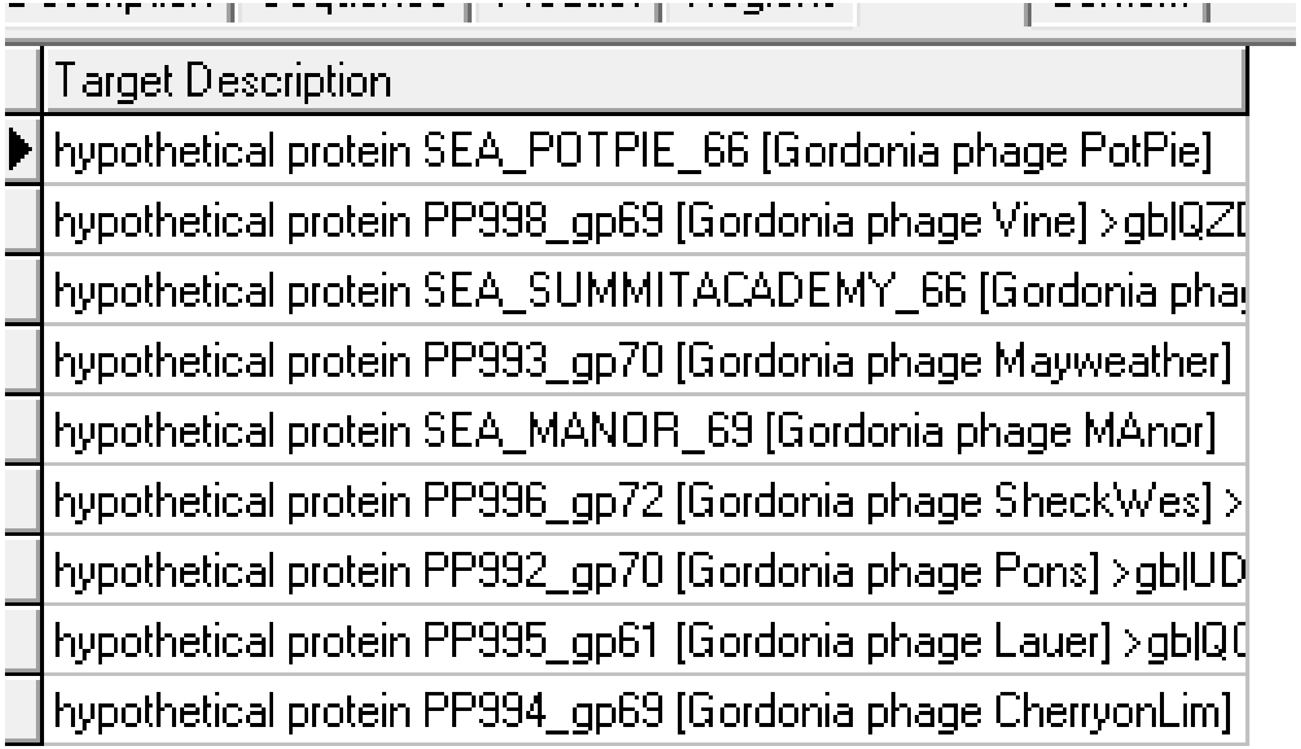
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	45999
Genemark	Glimmer and GeneMark
Coding potential	Includes all cp
RBS	Z value: 2.621 Final score: -4.256
BLAST	4 1:1 alignments
Starterator	8 MAs
Overlap	4

The start site is 45999 because it is called by both Glimmer and GeneMark, the frame includes all coding potential, the Z value is greater than 2, and the overlap is 4 which is ideal.

BLAST function evidence. What assigned functions do other highly similar genes have?

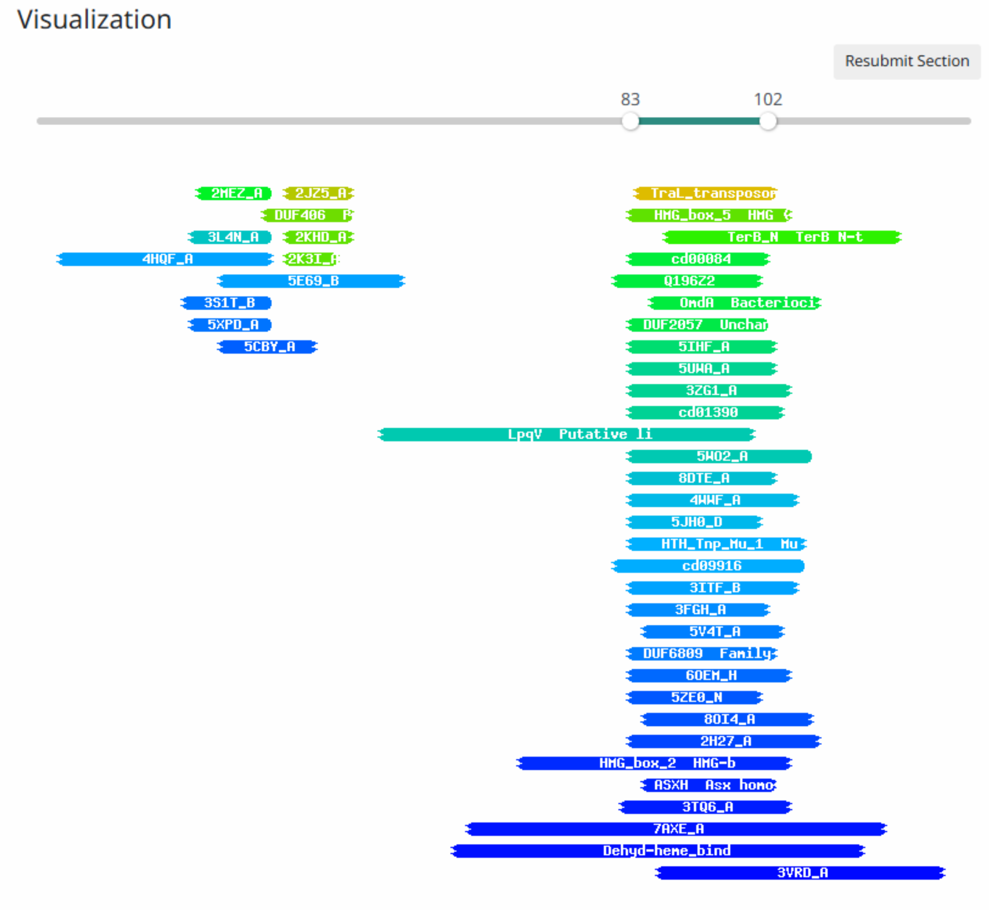
- 9 hypothetical protein



Target Description
hypothetical protein SEA_POTPIE_66 [Gordonia phage PotPie]
hypothetical protein PP998_gp69 [Gordonia phage Vine] >gb|QZ
hypothetical protein SEA_SUMMITACADEMY_66 [Gordonia pha
hypothetical protein PP993_gp70 [Gordonia phage Mayweather]
hypothetical protein SEA_MANOR_69 [Gordonia phage MAnor]
hypothetical protein PP996_gp72 [Gordonia phage SheckWes] >
hypothetical protein PP992_gp70 [Gordonia phage Pons] >gb|UD
hypothetical protein PP995_gp61 [Gordonia phage Lauer] >gb|Q
hypothetical protein PP994_gp69 [Gordonia phage CherryonLim]

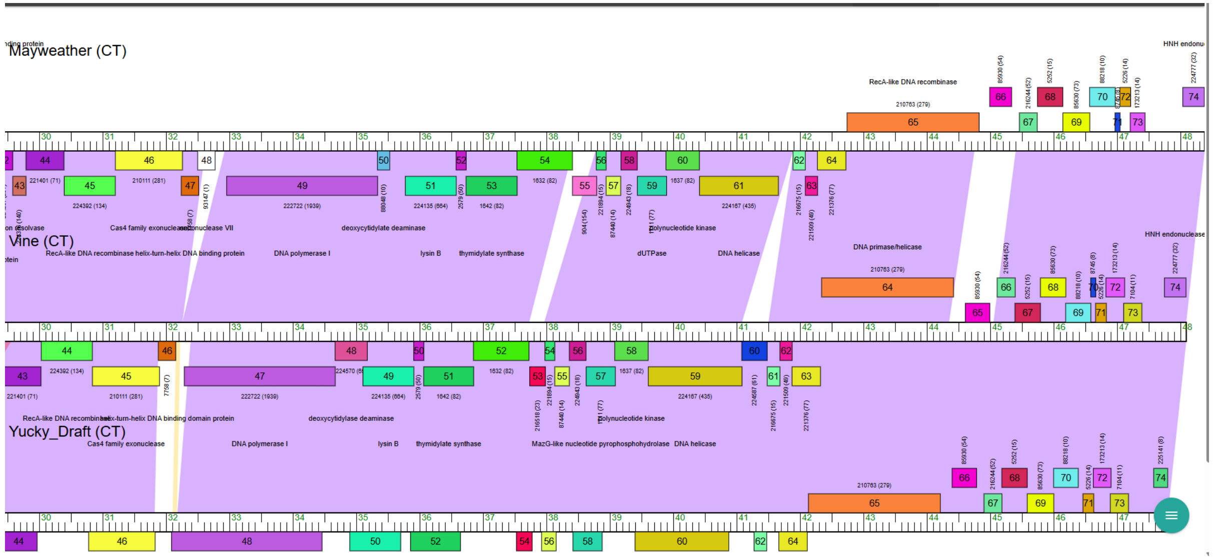
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- No hits as no probabilities are greater than 90%



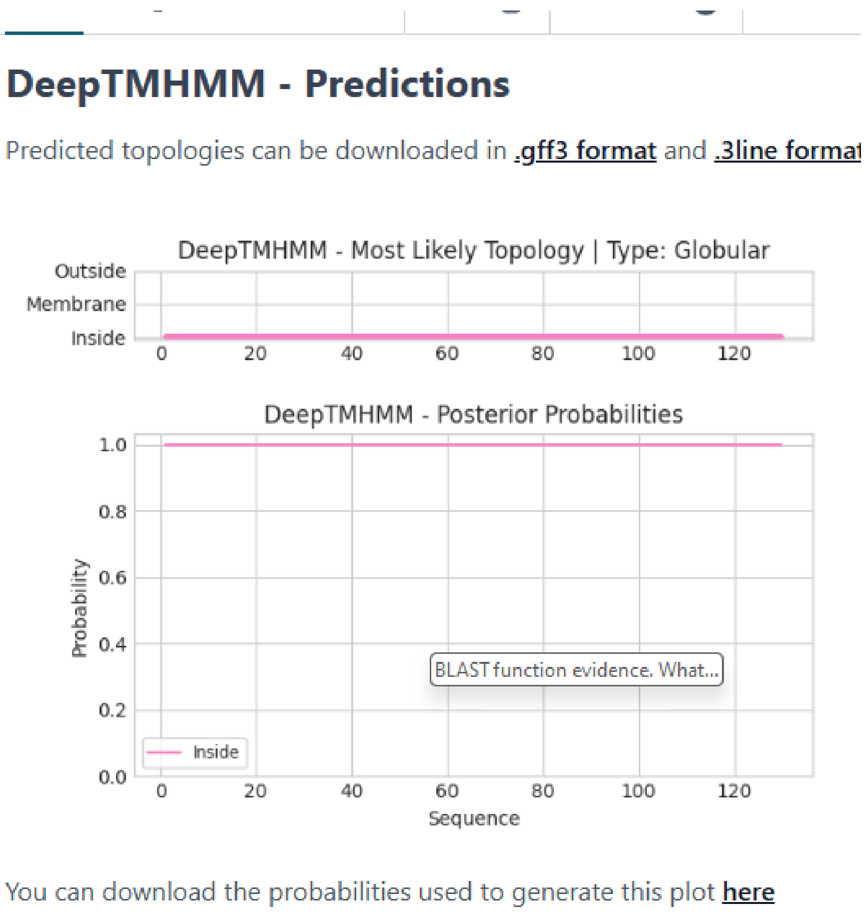
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Yucky feature 70 conserved domain: none function: none
- Vine feature 69 conserved domain: none function: none
- Mayweather feature 70 conserved domain: none function: none



Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- # of predicted TMRs: 0



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

DeepTMHMM - Most Likely Topology | Type: Globular
Outside
Membrane
Inside
0
20
40
60
80
100
120
DeepTMHMM - Posterior Probabilities
1.0
0.8
0.6
0.4
0.2
0.0
Probability
Inside
BLAST function evidence. What...
Sequence

You can download the probabilities used to generate this plot here

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function is hypothetical protein because it is the only function listed in BLAST, there is no function labeled for highly similar genes in Phamerator, no hits in Hhpred, and 0 predicted TMRs for DeepTMHMM evidence.

Feature 70 Stop 46465

Glimmer/GeneMark

What feature number is this?

What is the stop site?

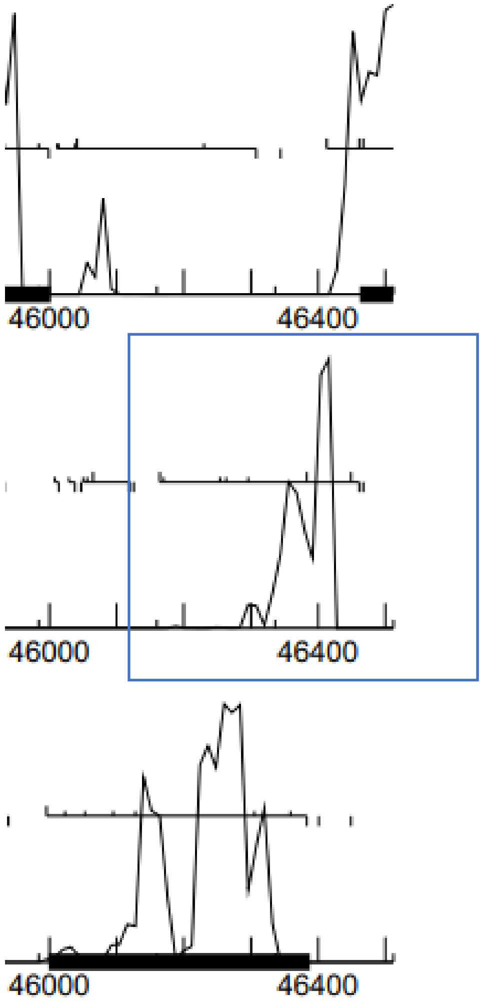
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

What is the autoannotated start?

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

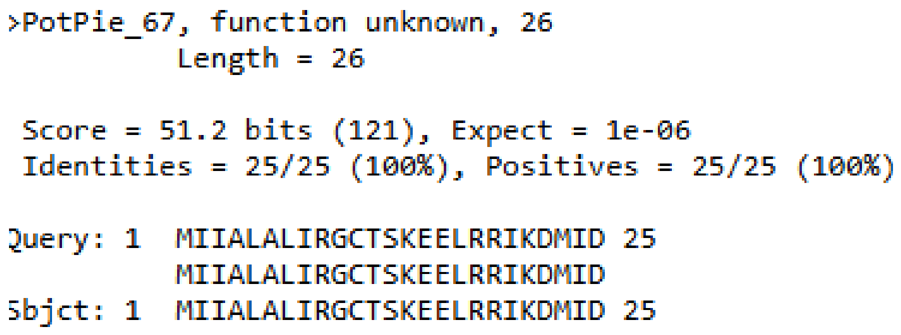
- 70
- Stop site: 46465
- Start Site: 46385
- Not an auto-annotated start
- It would have a 4 bp overlap with both the adjacent upstream and downstream genes.

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak?   How do you know?

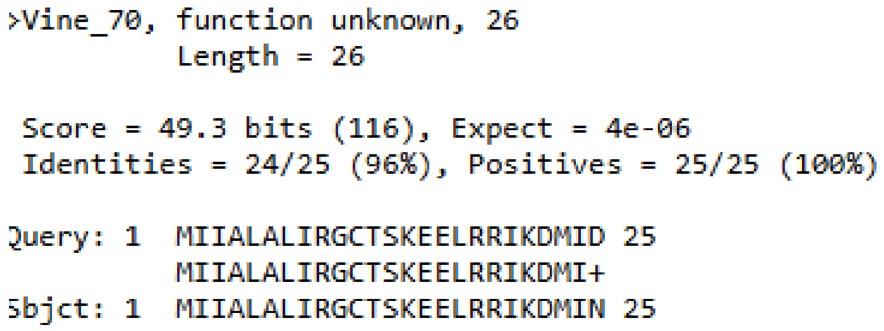


- Moderate to strong CP
- Some of the weaker CP is cut off at the start where it overlaps with the adjacent upstream gene.

BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.



```
>PotPie_67, function unknown, 26
          Length = 26

 Score = 51.2 bits (121), Expect = 1e-06
 Identities = 25/25 (100%), Positives = 25/25 (100%)

Query: 1  MIIALALIRGCTSKEELRRIKDMID 25
           MIIALALIRGCTSKEELRRIKDMID
Sbjct: 1  MIIALALIRGCTSKEELRRIKDMID 25
```



```
>Vine_70, function unknown, 26
          Length = 26

 Score = 49.3 bits (116), Expect = 4e-06
 Identities = 24/25 (96%), Positives = 25/25 (100%)

Query: 1  MIIALALIRGCTSKEELRRIKDMID 25
           MIIALALIRGCTSKEELRRIKDMI+
Sbjct: 1  MIIALALIRGCTSKEELRRIKDMIN 25
```

- Hits to several other CT cluster phage including SummitAcademy, PotPie, Vine, Feastonyeet, BigChungus, Mayweather, Pons, and MAnor.

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes!

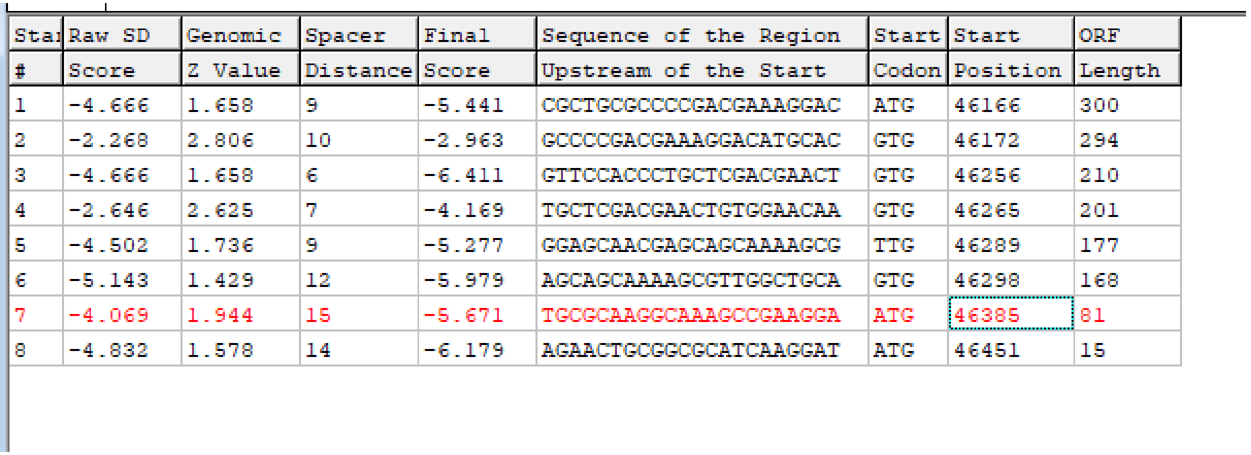
- It has coding potential and BLAST evidence

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- 46385 start is favored

- There are 8 Q1:S1 alignments for the 46385 start. All are CT cluster phage.
- There is a Q1:S1 start with the phage Lauer at bp 46298 but that start would create a 91 bp overlap with upstream feature 70.

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?



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	-4.666	1.658	9	-5.441	CGCTGCGCCCCGACGAAAGGAC	ATG	46166	300
2	-2.268	2.806	10	-2.963	GCCCCGACGAAAGGACATGCAC	GTG	46172	294
3	-4.666	1.658	6	-6.411	GTTCCACCCTGCTCGACGAACT	GTG	46256	210
4	-2.646	2.625	7	-4.169	TGCTCGACGAACTGTGGAACAA	GTG	46265	201
5	-4.502	1.736	9	-5.277	GGAGCAACGAGCAGCAAAAGCG	TTG	46289	177
6	-5.143	1.429	12	-5.979	AGCAGCAAAAGCGTTGGCTGCA	GTG	46298	168
7	-4.069	1.944	15	-5.671	TGCGCAAGGCAAAGCCGAAGGA	ATG	46385	81
8	-4.832	1.578	14	-6.179	AGAACTGCGGCGCATCAAGGAT	ATG	46451	15

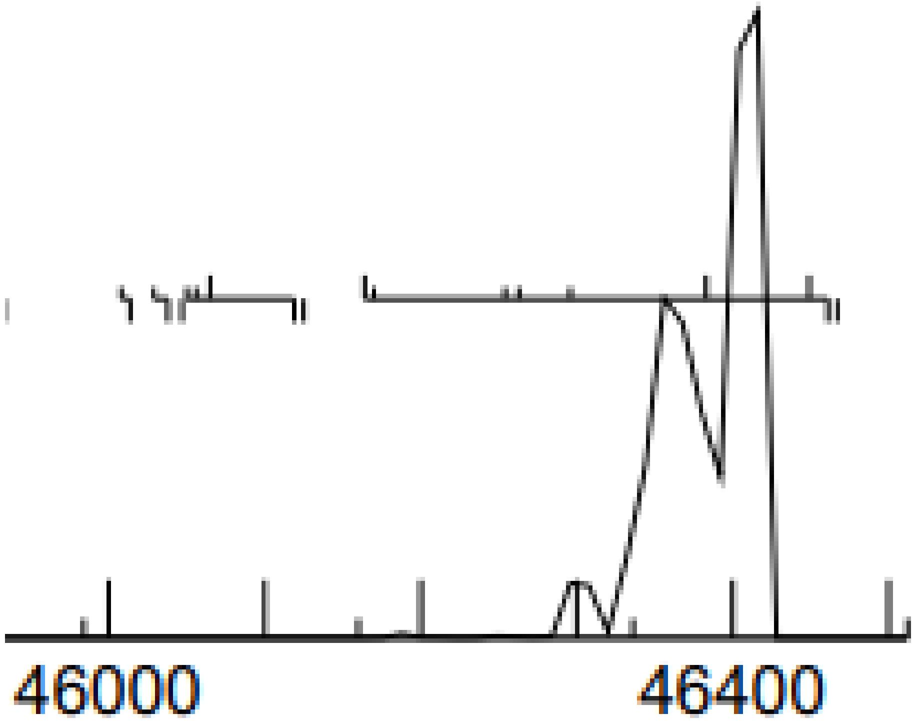
- RBS data is okay, but this gene would have 4 bp overlap on both sides. It is likely in an operon which does not necessarily need to exhibit excellent RBS scores.

- Z-Value: 1.944
- Final Score: -5.671

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

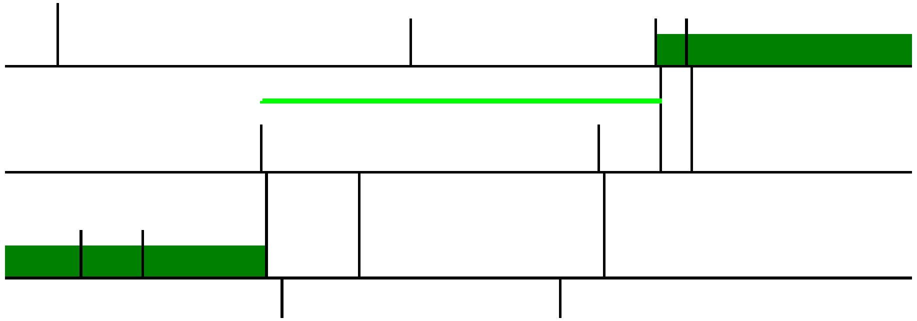
- No Starterator evidence since this wasn’t an auto-annotated gene.

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.



- 46166 Includes all CP
- 46172 Includes all CP
- 46256 Includes all CP
- 46265 Includes all CP
- 46289 Includes all CP
- 46298 Cuts off just a few bp
- 46385 Cuts off ~ 100 bp of CP
- 46451 Doesn’t include any CP

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.



The 46385 starts would have a 4 bp overlap on both the upstream and downstream sides.

- 46166 223 bp overlap
- 46172 217 bp overlap
- 46256 133 bp overlap
- 46265 124 bp overlap
- 46289 100 bp overlap
- 46298 9 bp overlap
- 46385 4 bp overlap
- 46451 62 bp gap

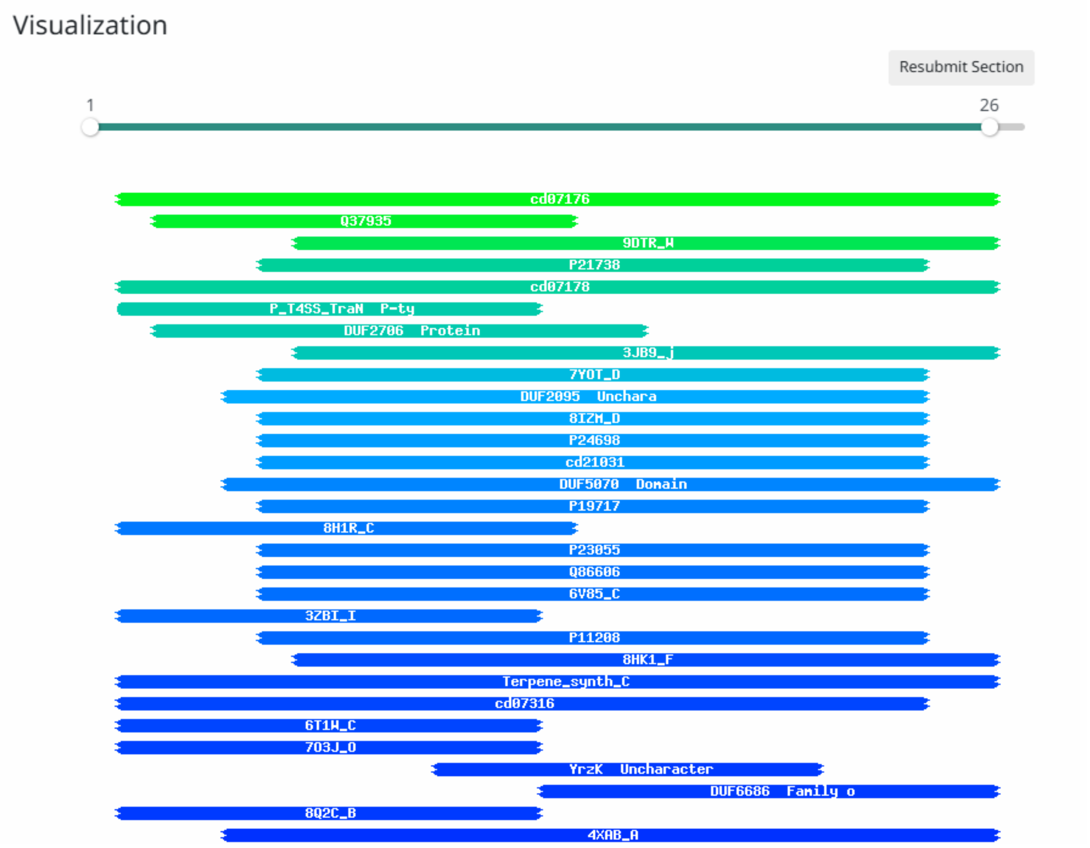
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- Start site is 46385
- There was not an auto-annotated start for this gene
- This start site has 8 Q1:S1 BLAST hits with other CT cluster phage
- This start site has a 4 bp overlap with the upstream feature

BLAST function evidence. What assigned functions do other highly similar genes have?

- All other highly similar genes have a function of Hypothetical protein

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

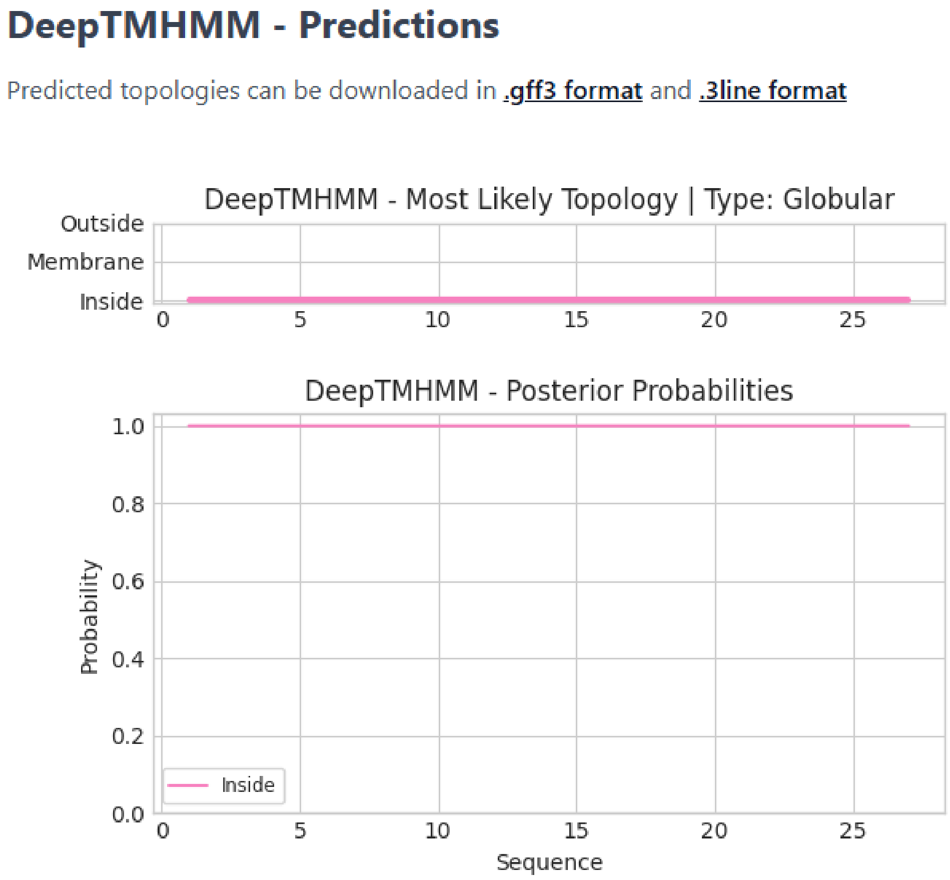


- NKF, no hits with a probability >90%

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- No Phamerator Data due to it not being an auto-annotated feature

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

DeepTMHMM - Most Likely Topology | Type: Globular
Outside
Membrane
Inside
0
5
10
15
20
25
DeepTMHMM - Posterior Probabilities
1.0
0.8
0.6
0.4
0.2
0.0
Probability
Inside
0
5
10
15
20
25
Sequence

- No predicted TMRs

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- Hypothetical Protein

- All BLAST hits were hypothetical proteins
- There were no HHPred hits with a probability >90%
- Deep TMHMM did not predict any TMRs

Feature 71 – Stop 46632

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

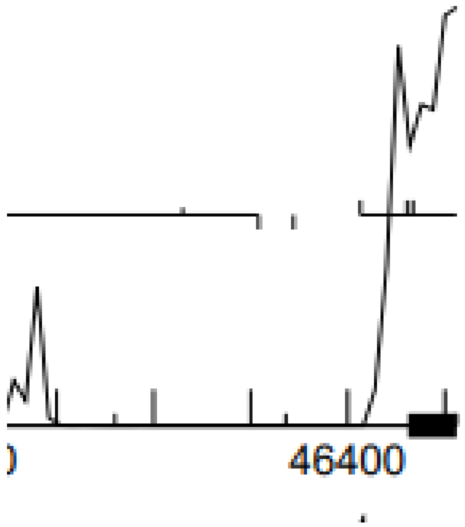
What is the autoannotated start?

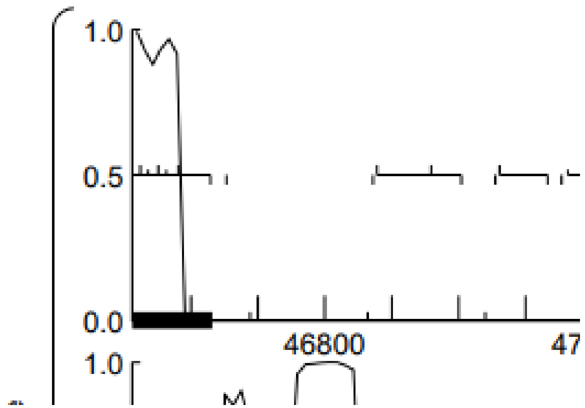
Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

- Feature 71
- Stop site: 46632
- Called by both Glimmer and GeneMark
- Autoannotated start is 46462
- Overlap of 4

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

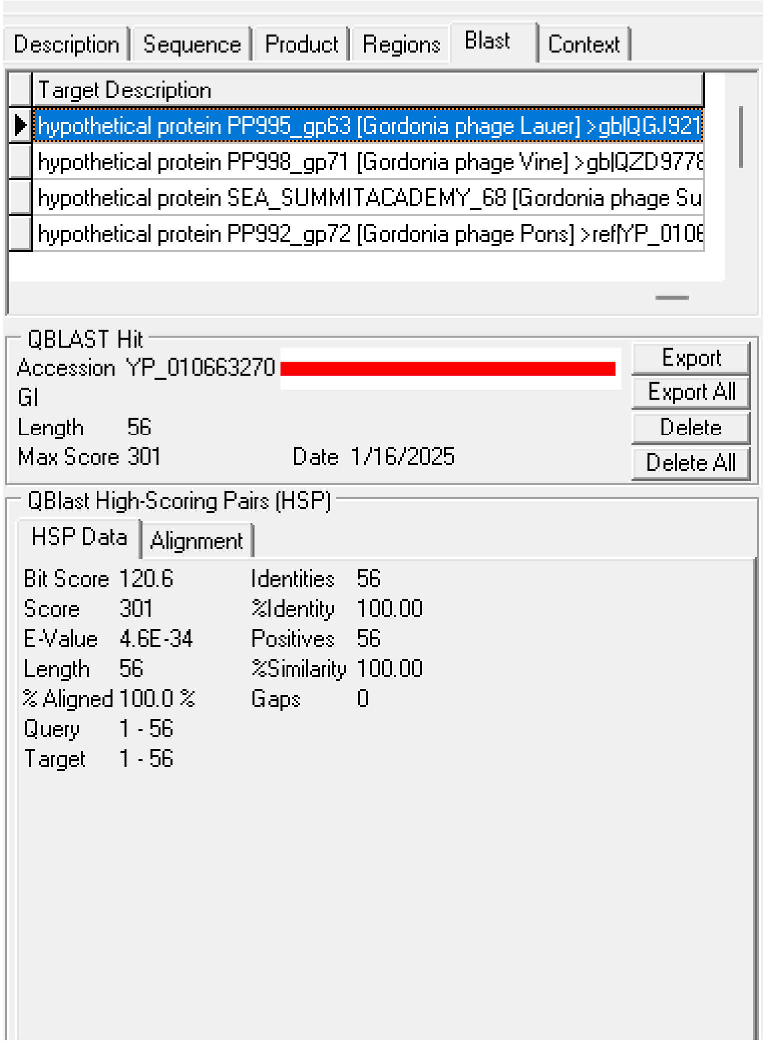
- Coding potential is cut off at start site 46462 and goes onto next page
- However, all coding potential is included at stop site 46632 and stops before 46632





BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 4 highly similar genes:
- Lauer
- Vine
- Pons
- MAnor



Target Description
hypothetical protein PP995_gp63 [Gordonia phage Lauer] >gb|QGJ921
hypothetical protein PP998_gp71 [Gordonia phage Vine] >gb|QZD977
hypothetical protein SEA_SUMMITACADEMY_68 [Gordonia phage Su
hypothetical protein PP992_gp72 [Gordonia phage Pons] >ref|YP_010

QBLAST Hit
Accession YP_010663270
GI
Length 56
Max Score 301
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	120.6	Identities	56
Score	301	%Identity	100.00
E-Value	4.6E-34	Positives	56
Length	56	%Similarity	100.00
% Aligned	100.0 %	Gaps	0
Query	1 - 56		
Target	1 - 56		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, it is a gene because both Glimmer and GeneMark agree at the start site. There are also 4 highly similar genes based on DNAM file BLAST evidence, and there are two frames that include coding potential for this feature.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

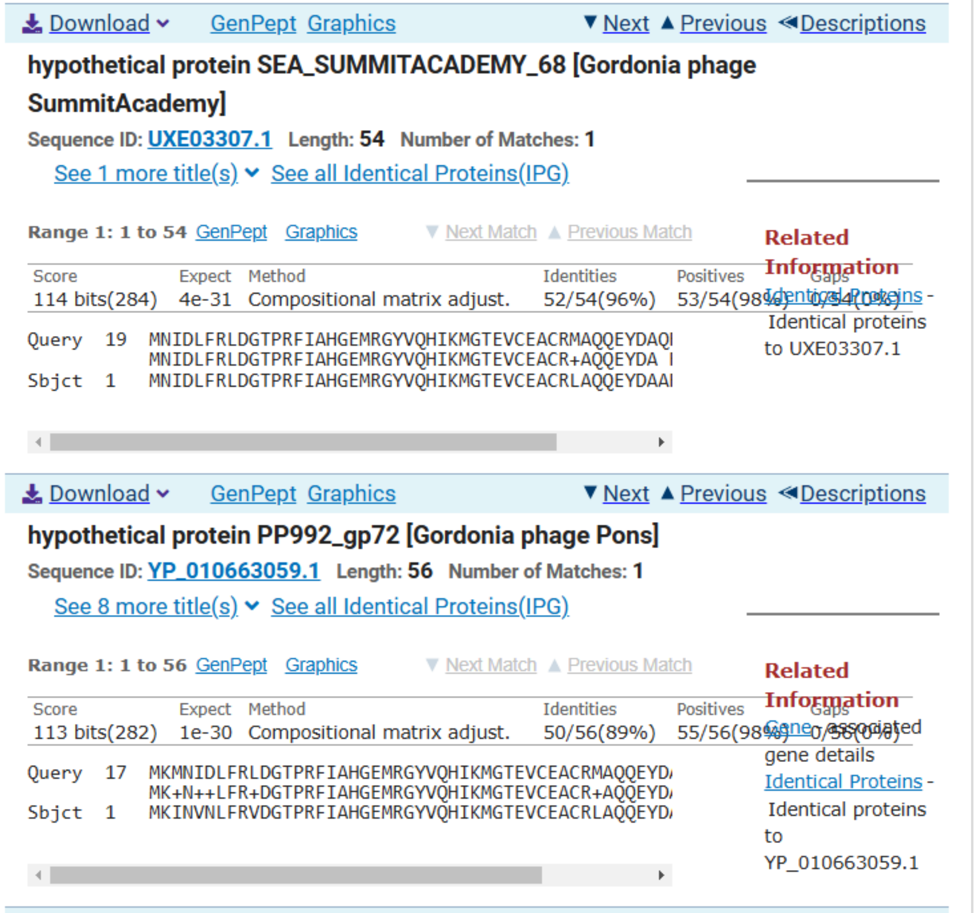
- Alignments for start site: 46462

4 1:1 alignments

- Alignments for start site: 46468

4 1:17 alignments

1 1:19 alignments



hypothetical protein SEA_SUMMITACADEMY_68 [Gordonia phage SummitAcademy]

Sequence ID: UXE03307.1 Length: 54 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 54 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
114 bits(284)	4e-31	Compositional matrix adjust.	52/54(96%)	53/54(98%)	0/54(0%)

```
Query  19   MNIDLFRLDGTPRFIAHGEMRGYVQHIKMGTEVCEACRMAQQEYDAQI
            MNIDLFRLDGTPRFIAHGEMRGYVQHIKMGTEVCEACR+AQQEYDA I
Sbjct  1    MNIDLFRLDGTPRFIAHGEMRGYVQHIKMGTEVCEACRLAQQEYDAAI
```

Related Information

Identical Proteins - Identical proteins to UXE03307.1

hypothetical protein PP992_gp72 [Gordonia phage Pons]

Sequence ID: YP_010663059.1 Length: 56 Number of Matches: 1

See 8 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 56 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
113 bits(282)	1e-30	Compositional matrix adjust.	50/56(89%)	55/56(98%)	0/56(0%)

```
Query  17   MKMNIDLFRLDGTPRFIAHGEMRGYVQHIKMGTEVCEACRMAQQEYDA
            MK+N++LFR+DGTPRFIAHGEMRGYVQHIKMGTEVCEACR+AQQEYDA
Sbjct  1    MKINVNLFRVDGTPRFIAHGEMRGYVQHIKMGTEVCEACRLAQQEYDA
```

Related Information

Gene - associated gene details

Identical Proteins - Identical proteins to YP_010663059.1

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Start: 2 @46462

Z value: 2.754

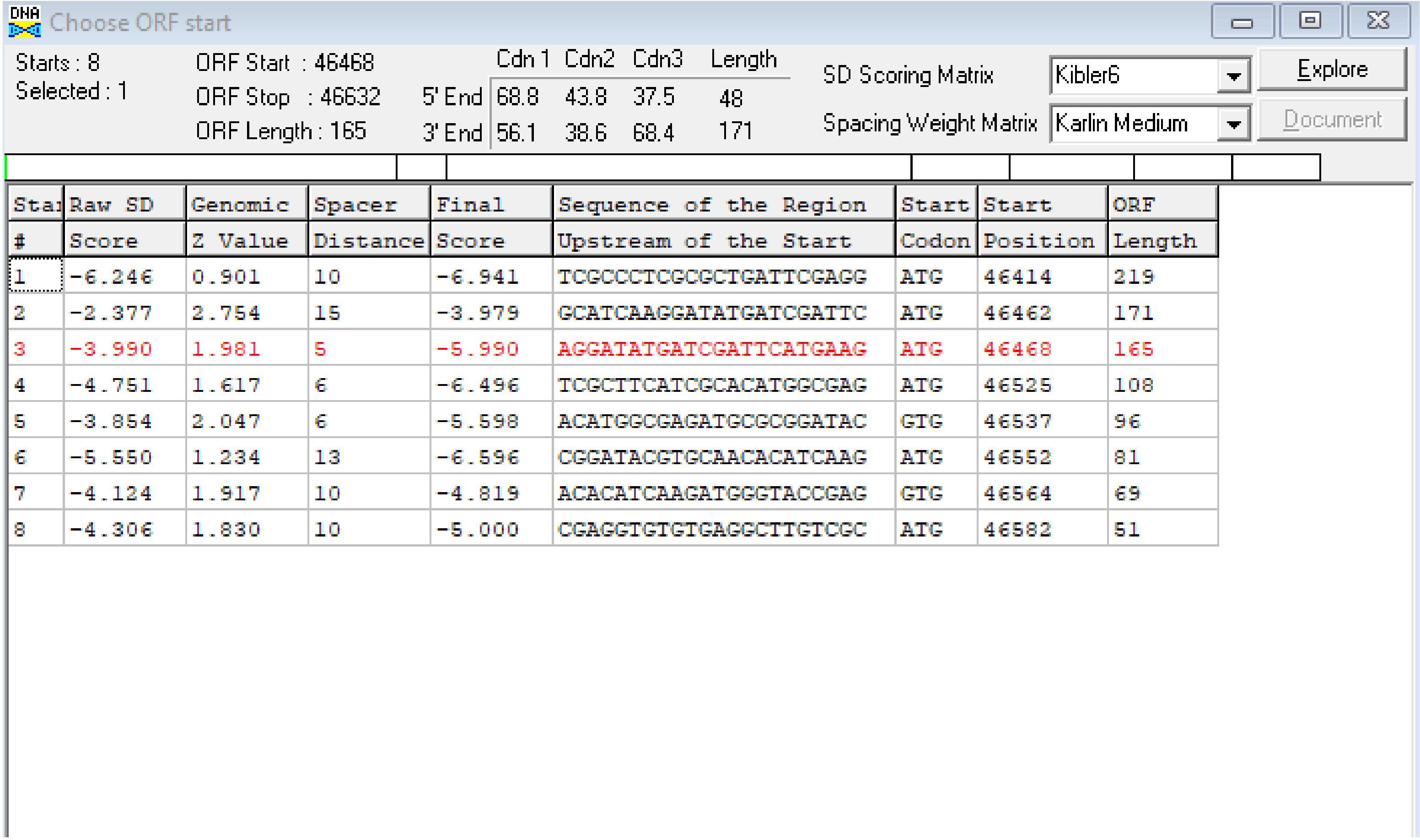
Final Score: -3.979

46462 Favored

- Start: 3 @46468

Z value: 1.981

Final Score: -5.990



Starts : 8
Selected : 1

ORF Start : 46468
ORF Stop : 46632
ORF Length : 165

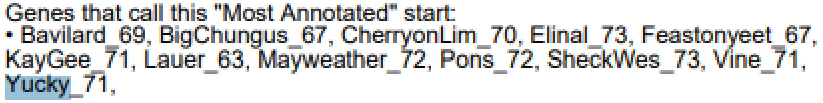
	Cdn 1	Cdn2	Cdn3	Length
5' End	68.8	43.8	37.5	48
3' End	56.1	38.6	68.4	171

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

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	-6.246	0.901	10	-6.941	TCGCCCTCGCGCTGATTCGAGG	ATG	46414	219
2	-2.377	2.754	15	-3.979	GCATCAAGGATATGATCGATTC	ATG	46462	171
3	-3.990	1.981	5	-5.990	AGGATATGATCGATTCATGAAG	ATG	46468	165
4	-4.751	1.617	6	-6.496	TCGCTTCATCGCACATGGCGAG	ATG	46525	108
5	-3.854	2.047	6	-5.598	ACATGGCGAGATGCGCGGATAC	GTG	46537	96
6	-5.550	1.234	13	-6.596	CGGATACGTGCAACACATCAAG	ATG	46552	81
7	-4.124	1.917	10	-4.819	ACACATCAAGATGGGTACCGAG	GTG	46564	69
8	-4.306	1.830	10	-5.000	CGAGGTGTGTGAGGCTTGTCGC	ATG	46582	51

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- Start: 2 @46462 has 10 MA’s
- Start: 3 @46468 has 2 MA’s



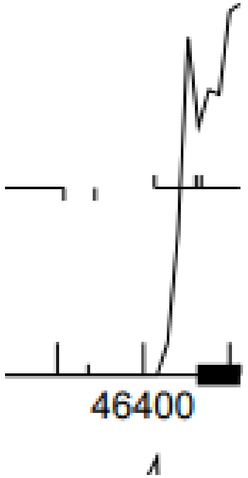
Genes that call this "Most Annotated" start:

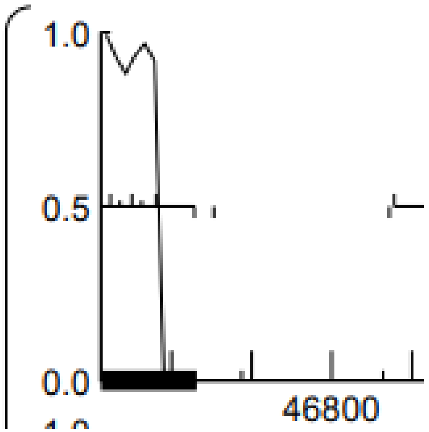
- Bavilard_69, BigChungus_67, CherryonLim_70, Elinal_73, Feastonyeet_67, KayGee_71, Lauer_63, Mayweather_72, Pons_72, SheckWes_73, Vine_71, Yucky_71,

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

- Start: 2 @46462

Cuts coding potential





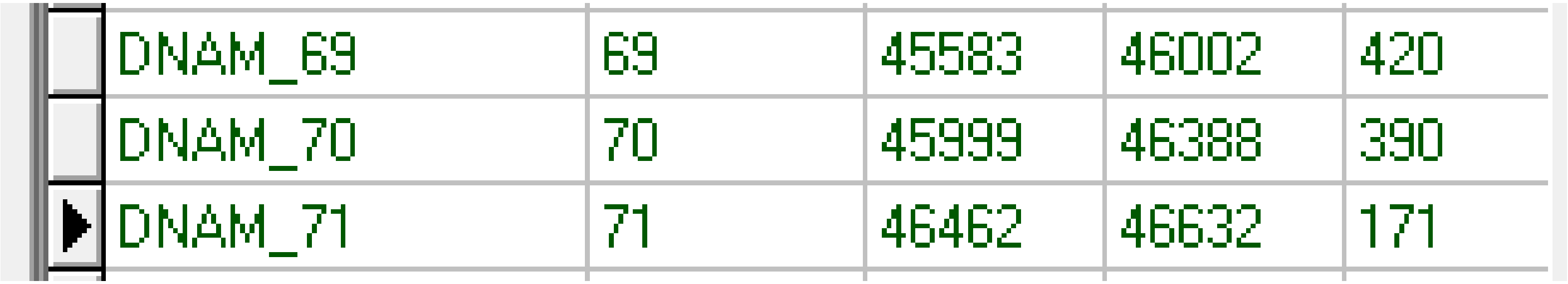
- Start: 3 @46468

Cuts coding potential

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start?   Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Start: 2 @46462

Gap of 73



	DNAM_69	69	45583	46002	420
	DNAM_70	70	45999	46388	390
▶	DNAM_71	71	46462	46632	171

- Start: 3 @46468

Gap of 79

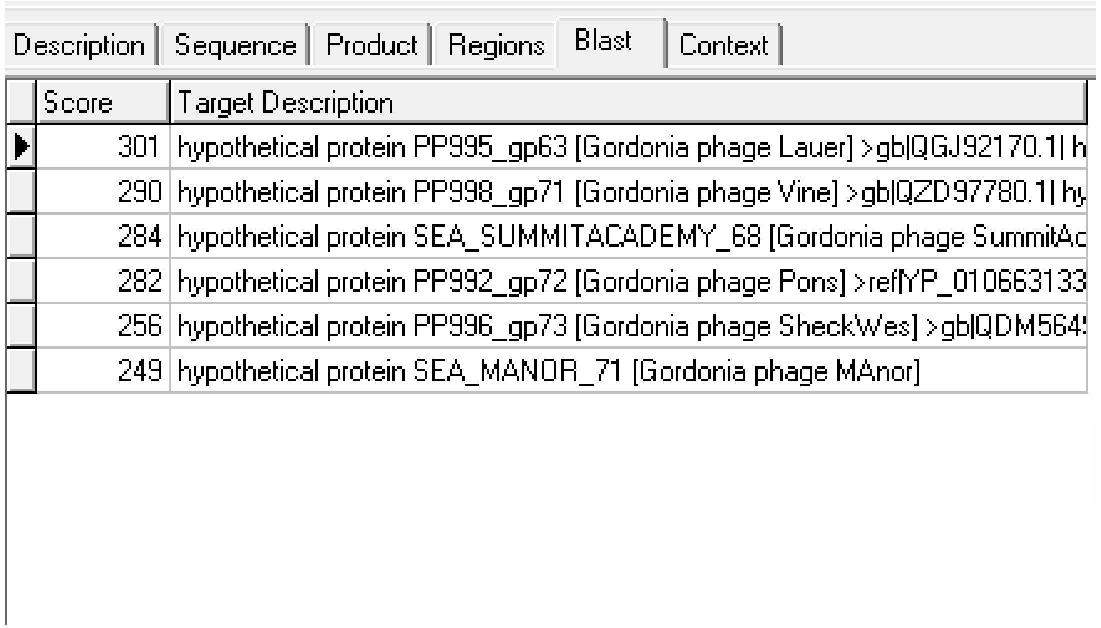
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	46462	46468
Glimmer	Called by both Gimmer & GeneMark	None
Coding potential	Cut off	Cut off
RBS	Z value: 2.754 Final Score: -3.979	Z value: 1.981 Final Score: -5.990
BLAST	4 1:1 alignment	4 1:17 alignments 1 1:19 alignments
Starterator	10 MA's	2 MA's
Gap	Overlap of 4	Gap of 79

The start site is 46462 because both Glimmer and GeneMark call it the start site, the z value is greater than 2, has 10 manual annotations, and its gap is the lowest of 73.

BLAST function evidence. What assigned functions do other highly similar genes have?

- 6 genes list its function as hypothetical protein

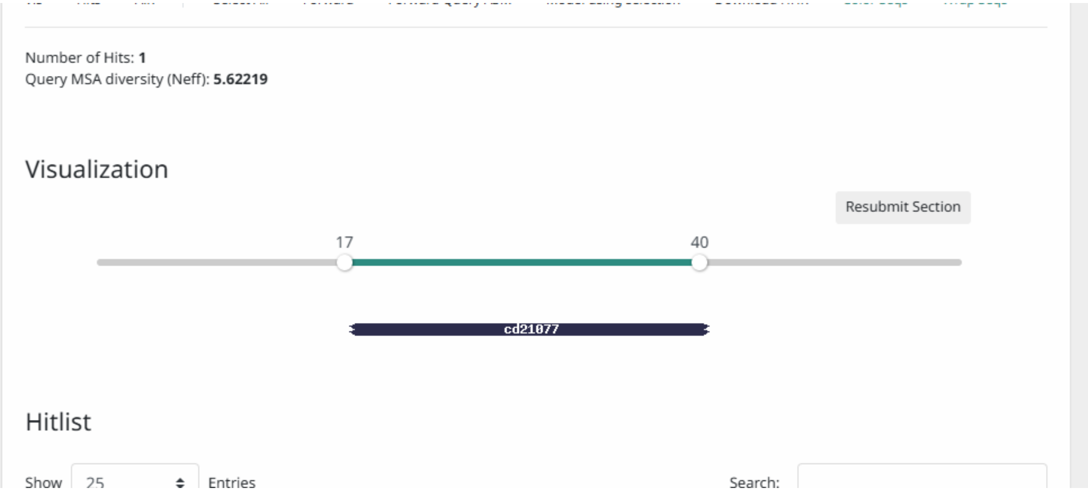


	Score	Target Description
▶	301	hypothetical protein PP995_gp63 [Gordonia phage Lauer] >gb|QGJ92170.1| h
	290	hypothetical protein PP998_gp71 [Gordonia phage Vine] >gb|QZD97780.1| hy
	284	hypothetical protein SEA_SUMMITACADEMY_68 [Gordonia phage SummitAc
	282	hypothetical protein PP992_gp72 [Gordonia phage Pons] >ref|YP_010663133
	256	hypothetical protein PP996_gp73 [Gordonia phage SheckWes] >gb|QDM564
	249	hypothetical protein SEA_MANOR_71 [Gordonia phage MAnor]

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- 1 hit

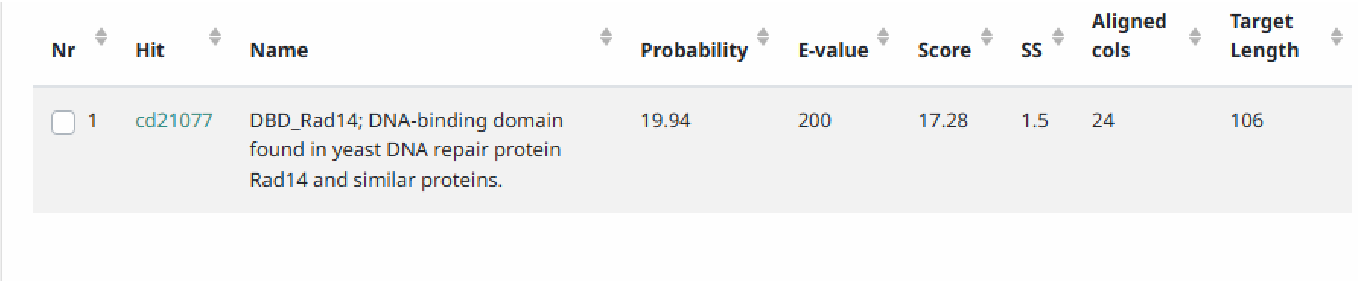
- Would not consider hit because it has less than 90% probability and an E-value that is greater than 1



Number of Hits: **1**
Query MSA diversity (Neff): **5.62219**

Visualization

Hitlist



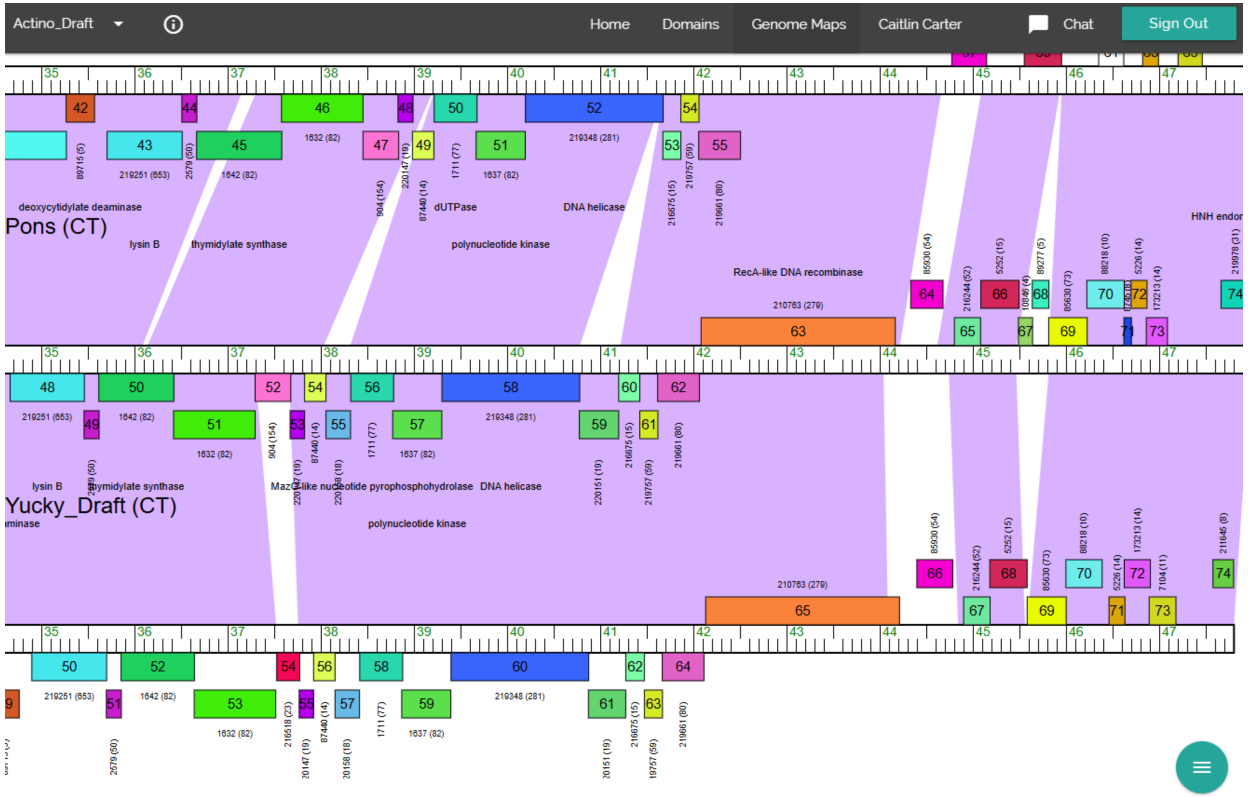
Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	cd21077	DBD_Rad14; DNA-binding domain found in yeast DNA repair protein Rad14 and similar proteins.	19.94	200	17.28	1.5	24	106

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Yucky feature 71 conserved domain: none function: none

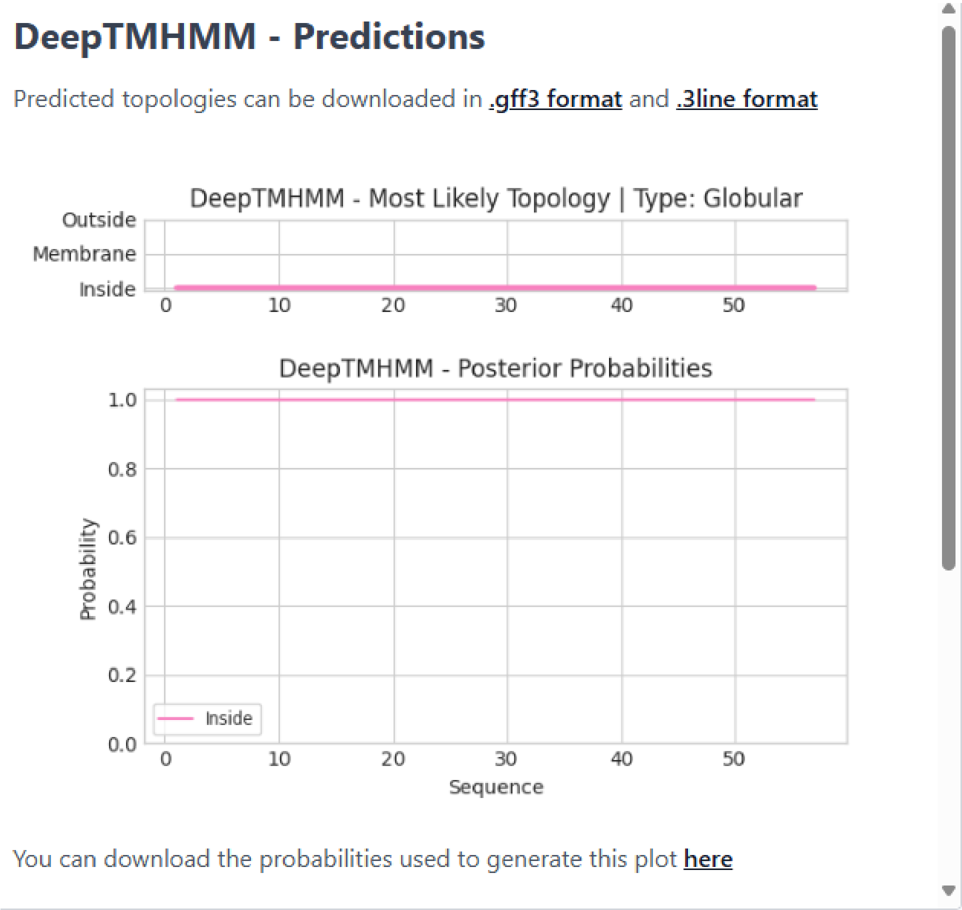
- Pons feature 72 conserved domain: none function: none

- Lauer feature 63 conserved domain: none function: none



Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- # Unnamed Number of predicted TMRs: 0



DeepTMHMM - Predictions

Predicted topologies can be downloaded in .gff3 format and .3line format

You can download the probabilities used to generate this plot here

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function is hypothetical protein, because all 6 genes showed function as hypothetical protein, the Hhpred evidence had 1 insufficient hit, Phamerator evidence showed no similar genes with a conserved domain or function, and Deep TMHMM evidence had 0 unnamed number of predicted TMRs.

Feature 72 – Stop 46906

Glimmer/GeneMark

What feature number is this?

What is the stop site?

Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither?

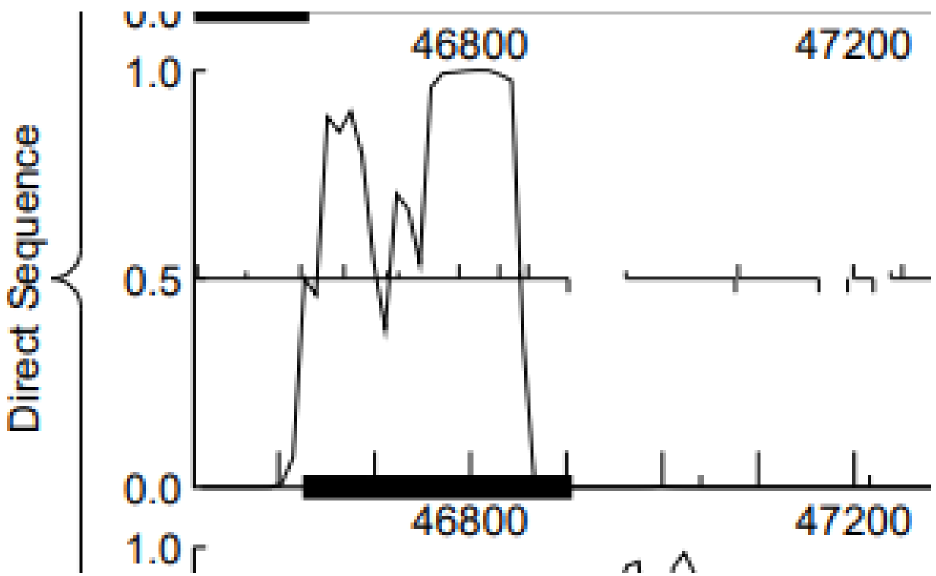
What is the autoannotated start?

Gap: _________ or overlap: ________ (with gene in front of it) for the autoannotated start

- Feature: 72
- Stop site: 46906
- Called by both Glimmer and GeneMark
- Autoannotated start: 46625
- Overlap: 8

GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

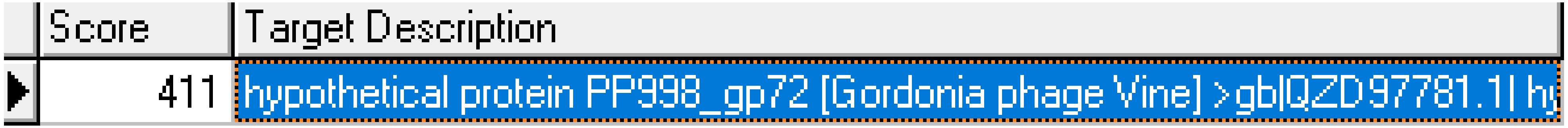
- Start site 46625
- Some of the coding potential is cut off at the start site
- In forward reading frame 2
- No other cp in other frames



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- 1 highly similar gene

Vine 1:1 alignment E-value: 0.0E0



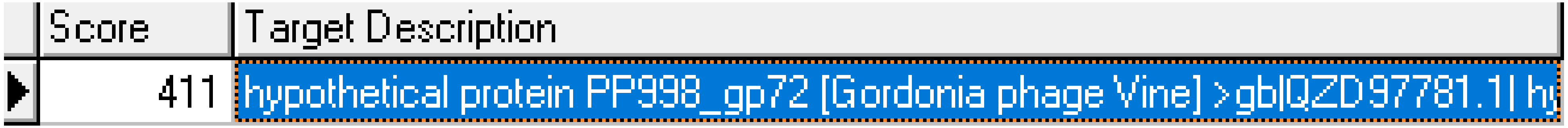
Score	Target Description
411	hypothetical protein PP998_gp72 [Gordonia phage Vine] >gb|QZD97781.1| hy

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Yes, it is a gene because both Glimmer and GeneMark call it. It includes coding potential even though it cuts some of it off and it has 1 highly similar gene (Vine) and it has 1 1:1 alignment

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- 1:1 alignment with Vine



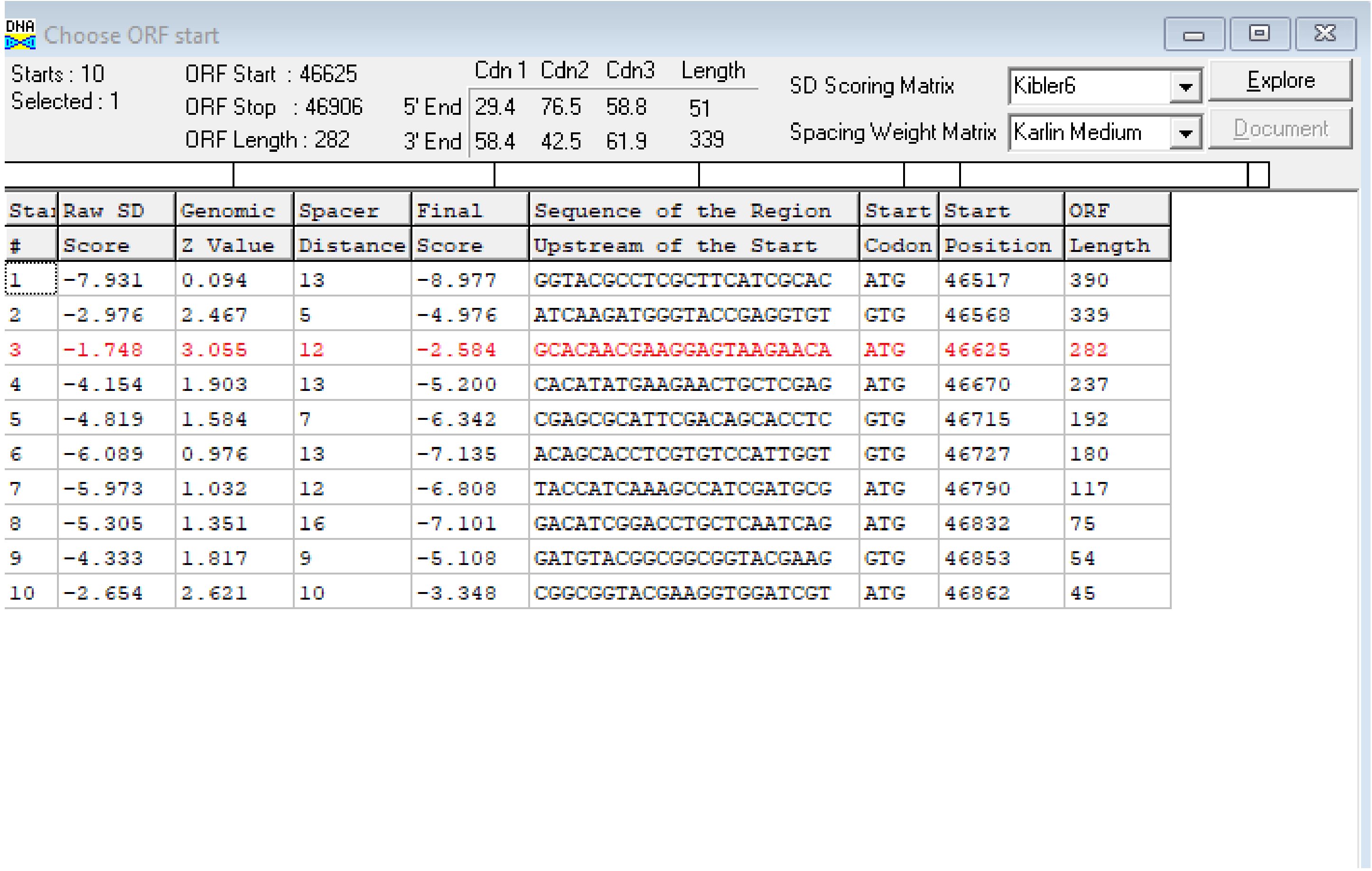
	Score	Target Description
▶	411	hypothetical protein PP998_gp72 [Gordonia phage Vine] >gb|QZD97781.1| hy

RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- Start 46625:

Z value: 3.055

Final score: -2.584



Starts : 10
Selected : 1

ORF Start : 46625
ORF Stop : 46906
ORF Length : 282

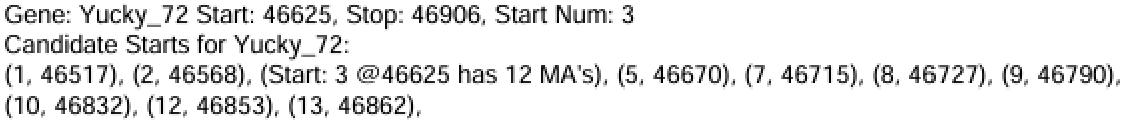
	Cdn 1	Cdn2	Cdn3	Length
5' End	29.4	76.5	58.8	51
3' End	58.4	42.5	61.9	339

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

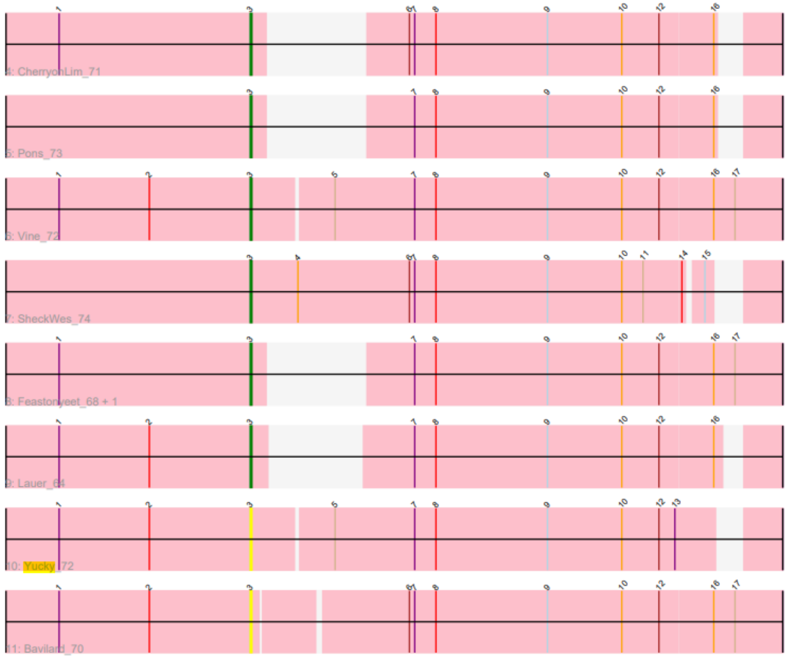
Sta	Raw SD	Genomic	Spacer	Final	Sequence of the Region	Start	Start	ORF
#	Score	Z Value	Distance	Score	Upstream of the Start	Codon	Position	Length
1	-7.931	0.094	13	-8.977	GGTACGCCTCGCTTCATCGCAC	ATG	46517	390
2	-2.976	2.467	5	-4.976	ATCAAGATGGGTACCGAGGTGT	GTG	46568	339
3	-1.748	3.055	12	-2.584	GCACAACGAAGGAGTAAGAACA	ATG	46625	282
4	-4.154	1.903	13	-5.200	CACATATGAAGAACTGCTCGAG	ATG	46670	237
5	-4.819	1.584	7	-6.342	CGAGCGCATTCGACAGCACCTC	GTG	46715	192
6	-6.089	0.976	13	-7.135	ACAGCACCTCGTGTCCATTGGT	GTG	46727	180
7	-5.973	1.032	12	-6.808	TACCATCAAAGCCATCGATGCG	ATG	46790	117
8	-5.305	1.351	16	-7.101	GACATCGGACCTGCTCAATCAG	ATG	46832	75
9	-4.333	1.817	9	-5.108	GATGTACGGCGGCGGTACGAAG	GTG	46853	54
10	-2.654	2.621	10	-3.348	CGGCGGTACGAAGGTGGATCGT	ATG	46862	45

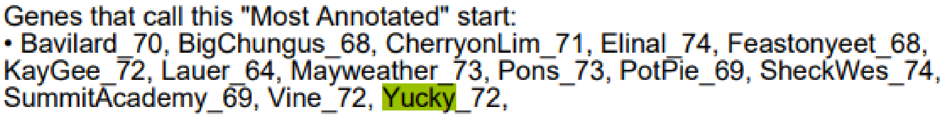
Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

Start 3: @46625 has 12 MA's - there are no other manual annotations



Gene: Yucky_72 Start: 46625, Stop: 46906, Start Num: 3
Candidate Starts for Yucky_72:
(1, 46517), (2, 46568), (Start: 3 @46625 has 12 MA's), (5, 46670), (7, 46715), (8, 46727), (9, 46790), (10, 46832), (12, 46853), (13, 46862),

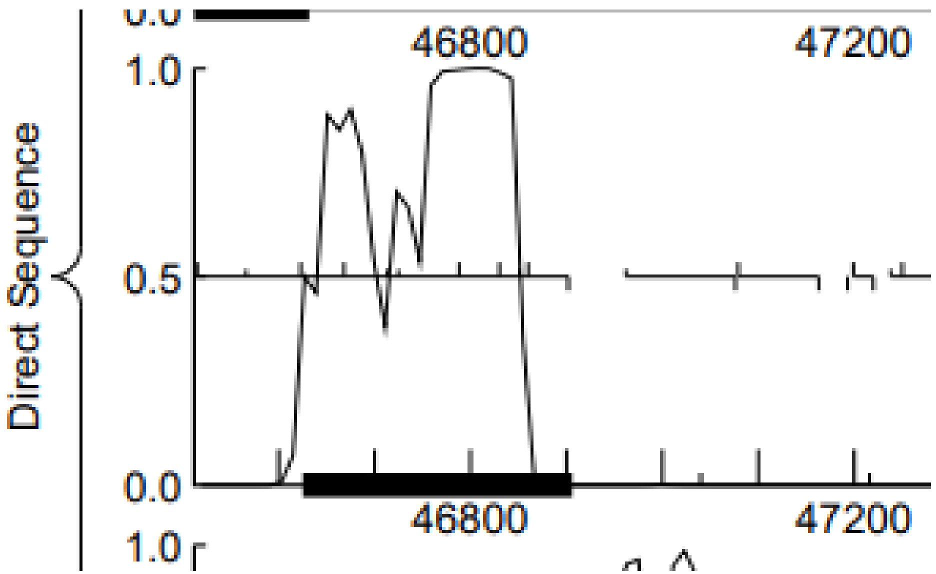




Genes that call this "Most Annotated" start:
- Bavilard_70, BigChungus_68, CherryonLim_71, Elinal_74, Feastonyeet_68, KayGee_72, Lauer_64, Mayweather_73, Pons_73, PotPie_69, SheckWes_74, SummitAcademy_69, Vine_72, Yucky_72,

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

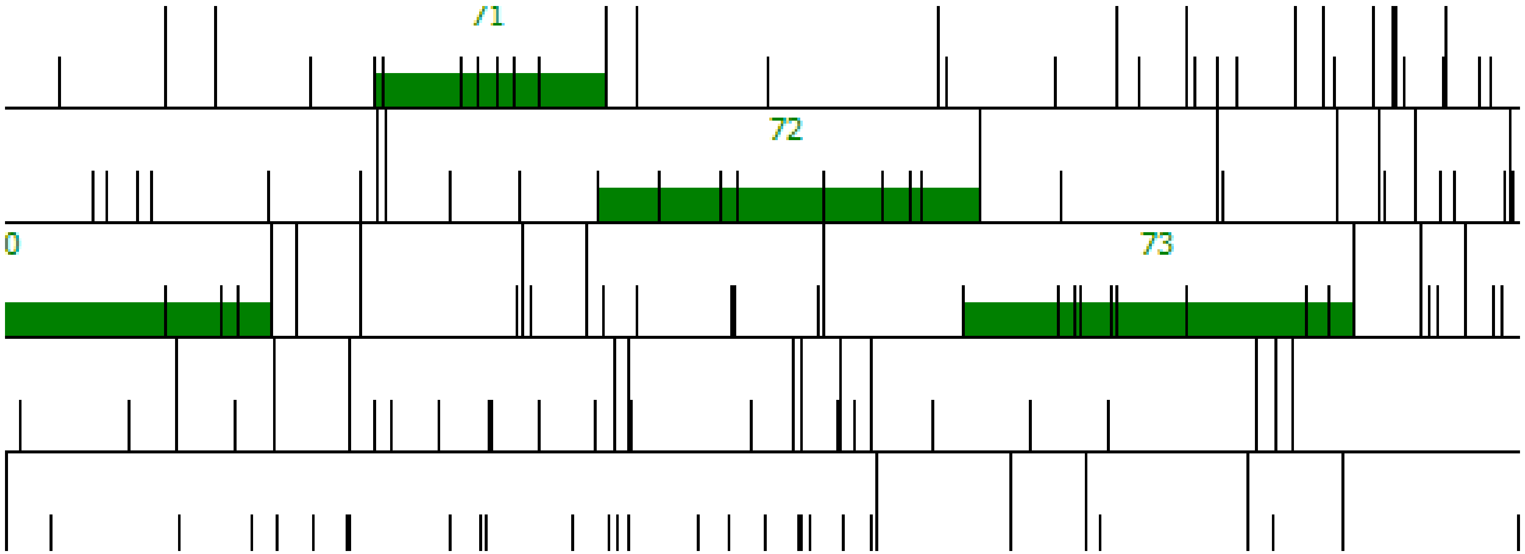
- Start 46625:
- Cuts off some coding potential before start site



Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

- Start 46625 (previous gene stop is 46632)

Overlap: 8



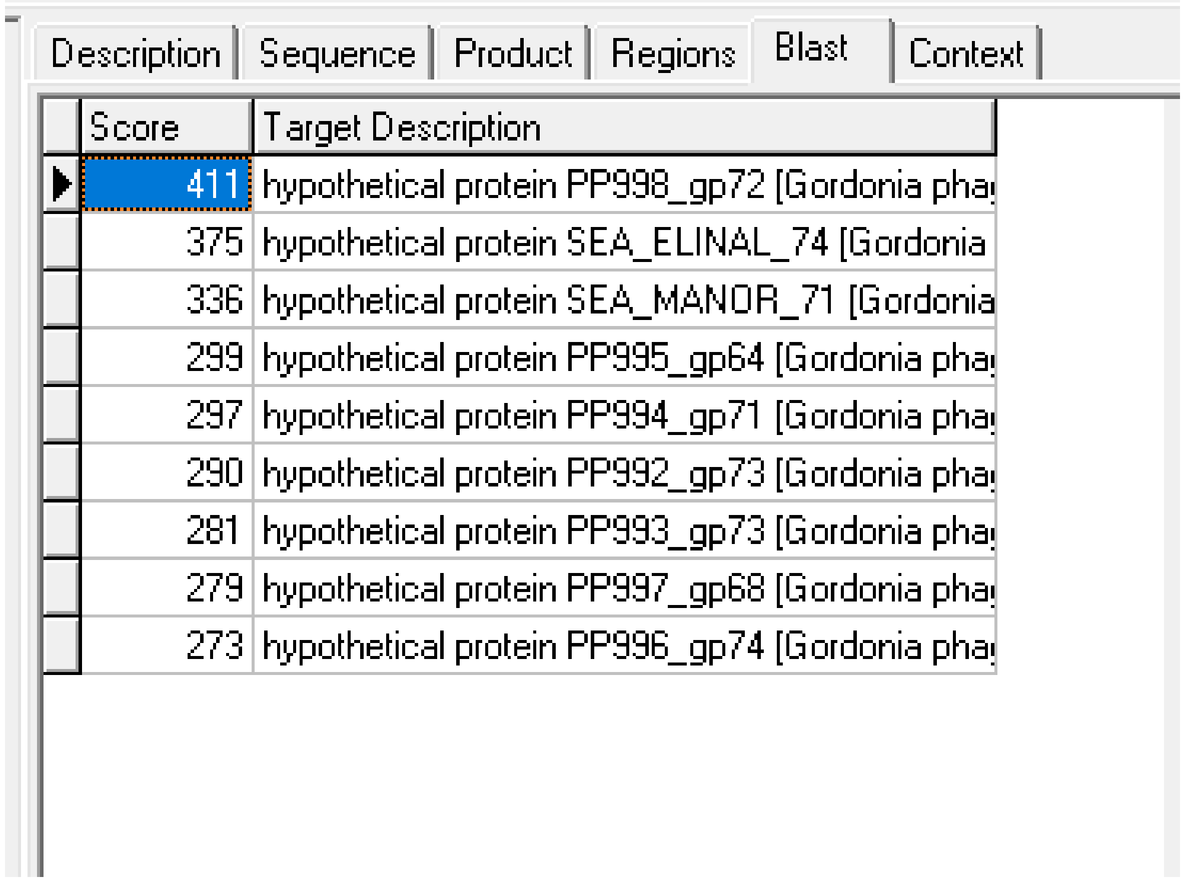
What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

	46625
GeneMark	Glimmer & GeneMark
Coding potential	Cuts off some cp
RBS	Z value: 3.055 Final score: -2.584
BLAST	1 1:1 alignment
Starterator	12 MA's
Overlap	8

Start site is 46625 because it is called by both Glimmer and GeneMark. It includes some coding potential. It has a high z value score (greater than 2 is ideal). And it has 12 manual annotations based on starterator evidence.

BLAST function evidence. What assigned functions do other highly similar genes have?

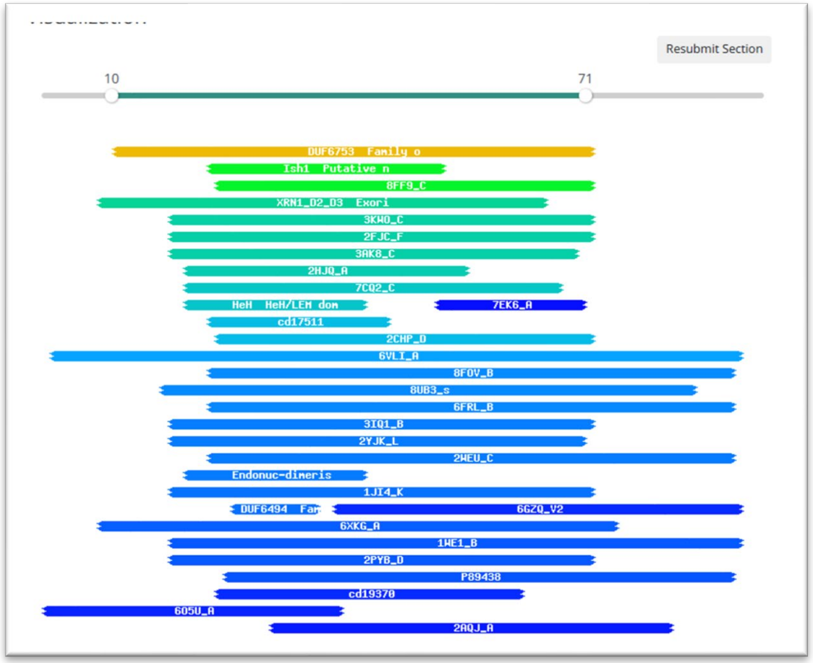
- 9 genes assign function as hypothetical protein

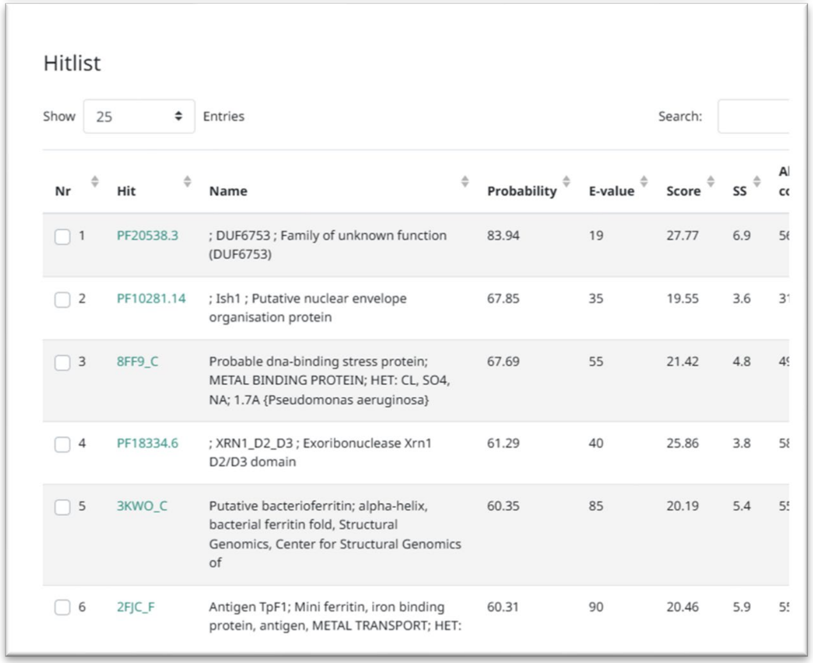


Score	Target Description
411	hypothetical protein PP998_gp72 [Gordonia pha
375	hypothetical protein SEA_ELINAL_74 [Gordonia
336	hypothetical protein SEA_MANOR_71 [Gordonia
299	hypothetical protein PP995_gp64 [Gordonia pha
297	hypothetical protein PP994_gp71 [Gordonia pha
290	hypothetical protein PP992_gp73 [Gordonia pha
281	hypothetical protein PP993_gp73 [Gordonia pha
279	hypothetical protein PP997_gp68 [Gordonia pha
273	hypothetical protein PP996_gp74 [Gordonia pha

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- No hits found with probabilities greater than 90 or E value less than 1
- No determined function





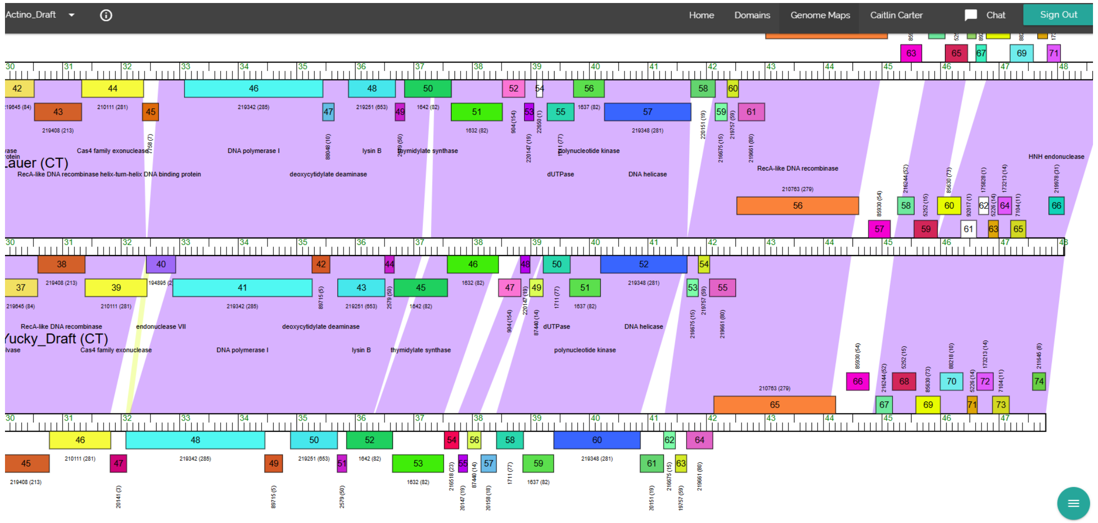
Hitlist

Show 25 Entries Search:

Nr	Hit	Name	Probability	E-value	Score	SS	Al cc
□ 1	PF20538.3	; DUF6753 ; Family of unknown function (DUF6753)	83.94	19	27.77	6.9	56
□ 2	PF10281.14	; Ish1 ; Putative nuclear envelope organisation protein	67.85	35	19.55	3.6	31
□ 3	8FF9_C	Probable dna-binding stress protein; METAL BINDING PROTEIN; HET: CL, SO4, NA; 1.7A {Pseudomonas aeruginosa}	67.69	55	21.42	4.8	49
□ 4	PF18334.6	; XRN1_D2_D3 ; Exoribonuclease Xrn1 D2/D3 domain	61.29	40	25.86	3.8	58
□ 5	3KWO_C	Putative bacterioferritin; alpha-helix, bacterial ferritin fold, Structural Genomics, Center for Structural Genomics of	60.35	85	20.19	5.4	55
□ 6	2FJC_F	Antigen TpF1; Mini ferritin, iron binding protein, antigen, METAL TRANSPORT; HET:	60.31	90	20.46	5.9	55

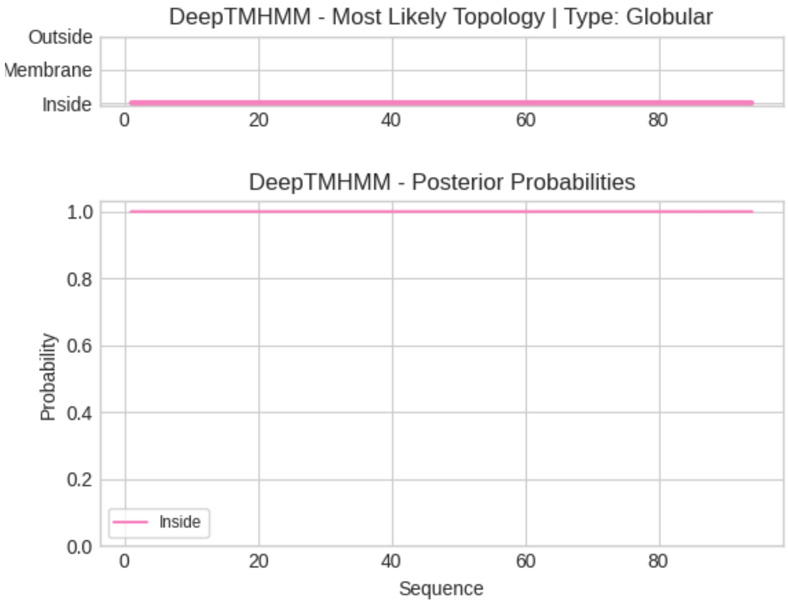
Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- Yucky feature 72 conserved domain: none function: none
- Lauer feature 64 conserved domain: none function: none
- CherryonLim feature 71 conserved domain: none function: none



Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

# sequence Number of predicted
TMRS: 0



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function is hypothetical protein because all 9 genes in DNAM file list function as hypothetical protein. There are no hits in Hhpred evidence with probability greater than 90 or E value less than 1, and Phamerator evidence assigns no function and no conserved domain for genes Lauer and CherryonLim. DeepTMHMM evidence also has 0 sequence number of predicted TMRs.

Feature 73 – Stop 47180

Glimmer/GeneMark

What feature number is this? 73

What is the stop site? 47,180

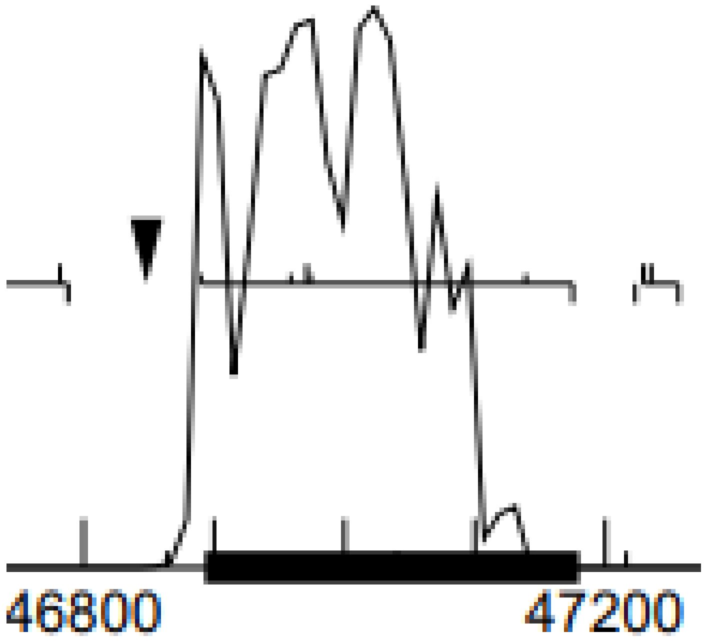
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer only, GeneMark only, or neither? Only GeneMark called this as the start site.

What is the autoannotated start? 46,893

Gap: ________ or overlap: _14____ (with gene in front of it) for the autoannotated start

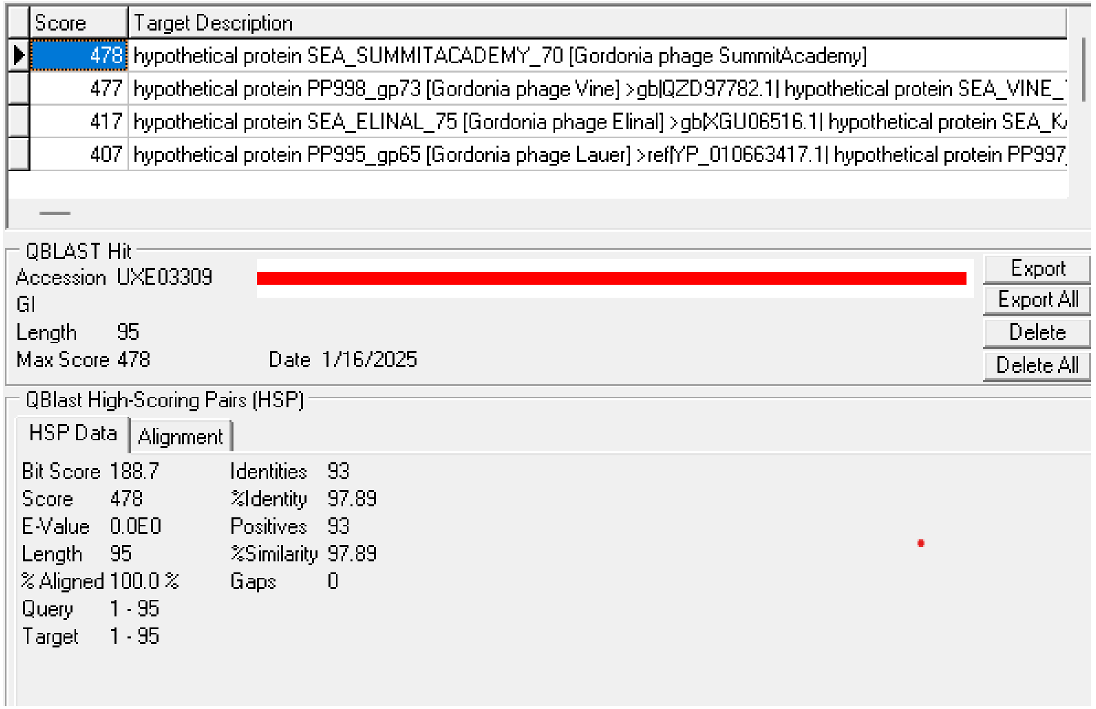
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Is it the only reading frame with cp? This is the reading frame with the most reading potential. Frame 2 has very little potential.
- Describe the coding potential… is it strong or is it weak? How do you know? This cp is strong as it has mostly has a height of 1.0.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are 5 highly similar genes. They all have 1:1 alignments with E values less than 10^-7.



Score	Target Description
478	hypothetical protein SEA_SUMMITACADEMY_70 [Gordonia phage SummitAcademy]
477	hypothetical protein PP998_gp73 [Gordonia phage Vine] >gb|QZD97782.1| hypothetical protein SEA_VINE_
417	hypothetical protein SEA_ELINAL_75 [Gordonia phage Elinal] >gb|XGU06516.1| hypothetical protein SEA_K
407	hypothetical protein PP995_gp65 [Gordonia phage Lauer] >ref|YP_010663417.1| hypothetical protein PP997

QBLAST Hit

Accession UXE03309

GI

Length 95

Max Score 478

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

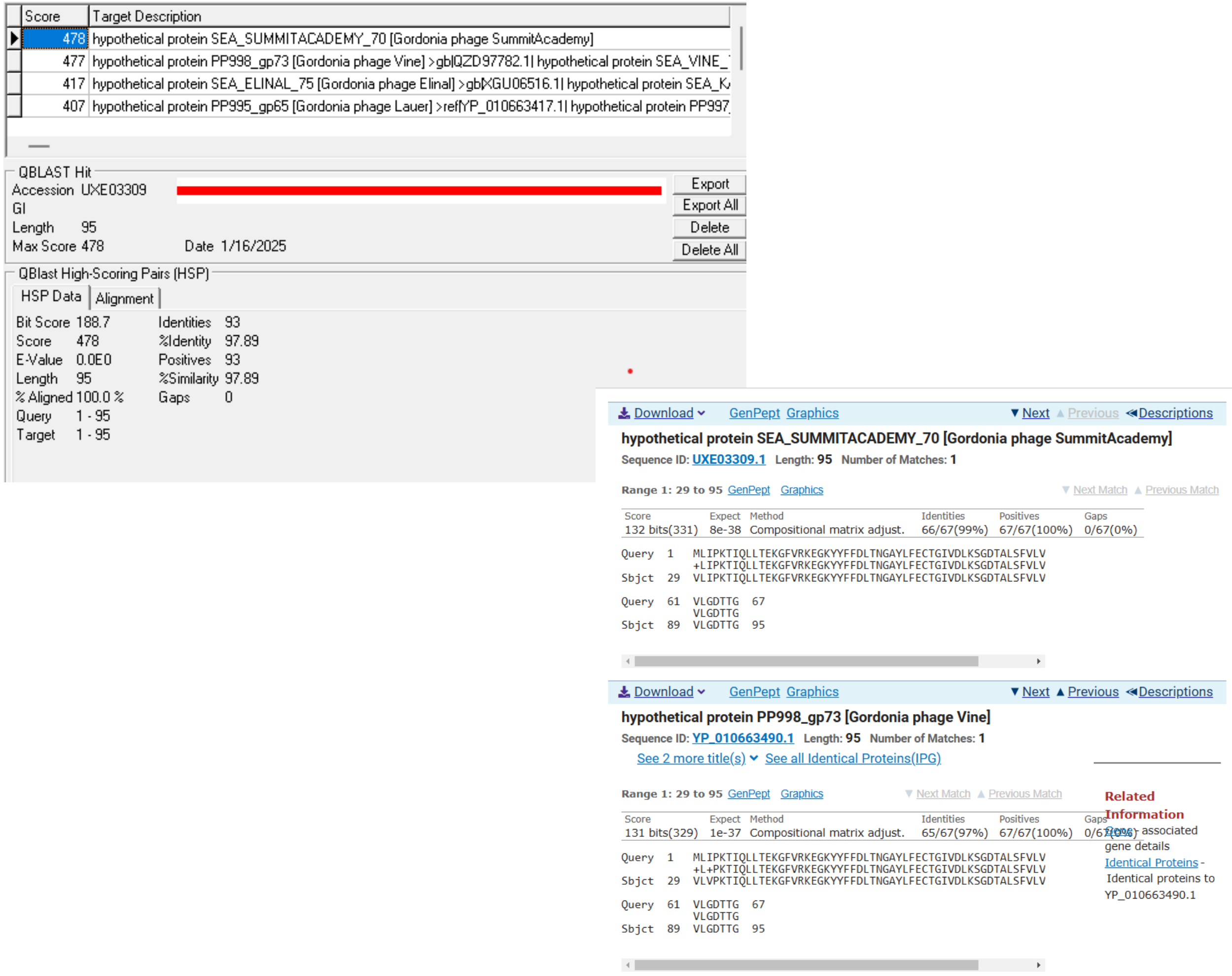
Bit Score	188.7	Identities	93
Score	478	%Identity	97.89
E-Value	0.0E0	Positives	93
Length	95	%Similarity	97.89
% Aligned	100.0 %	Gaps	0
Query	1 - 95		
Target	1 - 95		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Is there more than one feature called in this coding region?. Ye function 73 is a gene because GeneMark calls it a gene, there is cp, and there are 5 highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There are 5 highly similar genes. They all have 1:1 alignments with E values less than 10^-7. Some include SummitAcademy, Vine and Elinal.
- For start 46,977 there are no 1:1 alignments but 5 highly similar genes which include: SummitAcademy, Vine and Elinal.



Score	Target Description
478	hypothetical protein SEA_SUMMITACADEMY_70 [Gordonia phage SummitAcademy]
477	hypothetical protein PP998_gp73 [Gordonia phage Vine] >gb|QZD97782.1| hypothetical protein SEA_VINE_
417	hypothetical protein SEA_ELINAL_75 [Gordonia phage Elinal] >gb|XGU06516.1| hypothetical protein SEA_K/
407	hypothetical protein PP995_gp65 [Gordonia phage Lauer] >ref|YP_010663417.1| hypothetical protein PP997

```
QBLAST Hit
Accession  UXE03309
GI
Length     95
Max Score  478              Date  1/16/2025
```

```
QBlast High-Scoring Pairs (HSP)
HSP Data | Alignment
Bit Score  188.7        Identities   93
Score      478          %Identity    97.89
E-Value    0.0E0        Positives    93
Length     95           %Similarity  97.89
% Aligned  100.0 %      Gaps         0
Query      1 - 95
Target     1 - 95
```

hypothetical protein SEA_SUMMITACADEMY_70 [Gordonia phage SummitAcademy]

Sequence ID: UXE03309.1 Length: 95 Number of Matches: 1

Range 1: 29 to 95 GenPept Graphics

```
Score          Expect  Method                          Identities    Positives     Gaps
132 bits(331)  8e-38   Compositional matrix adjust.    66/67(99%)    67/67(100%)   0/67(0%)

Query  1   MLIPKTIQLLTEKGFVRKEGKYYFFDLTNGAYLFECTGIVDLKSGDTALSFVLV
           +LIPKTIQLLTEKGFVRKEGKYYFFDLTNGAYLFECTGIVDLKSGDTALSFVLV
Sbjct  29  VLIPKTIQLLTEKGFVRKEGKYYFFDLTNGAYLFECTGIVDLKSGDTALSFVLV

Query  61  VLGDTTG  67
           VLGDTTG
Sbjct  89  VLGDTTG  95
```

hypothetical protein PP998_gp73 [Gordonia phage Vine]

Sequence ID: YP_010663490.1 Length: 95 Number of Matches: 1

See 2 more title(s) See all Identical Proteins(IPG)

Range 1: 29 to 95 GenPept Graphics

```
Score          Expect  Method                          Identities    Positives     Gaps
131 bits(329)  1e-37   Compositional matrix adjust.    65/67(97%)    67/67(100%)   0/67(0%)

Query  1   MLIPKTIQLLTEKGFVRKEGKYYFFDLTNGAYLFECTGIVDLKSGDTALSFVLV
           +L+PKTIQLLTEKGFVRKEGKYYFFDLTNGAYLFECTGIVDLKSGDTALSFVLV
Sbjct  29  VLVPKTIQLLTEKGFVRKEGKYYFFDLTNGAYLFECTGIVDLKSGDTALSFVLV

Query  61  VLGDTTG  67
           VLGDTTG
Sbjct  89  VLGDTTG  95
```

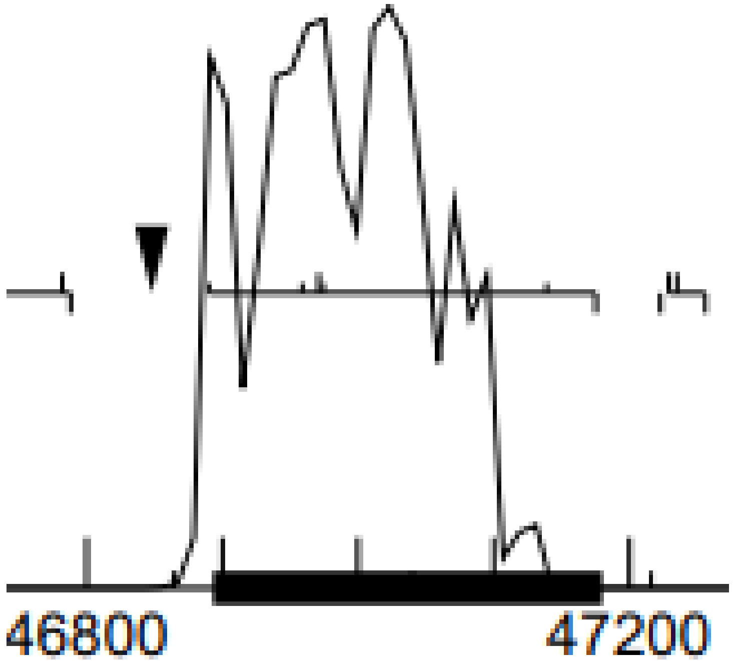
Related Information

Gene - associated gene details

Identical Proteins - Identical proteins to YP_010663490.1

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

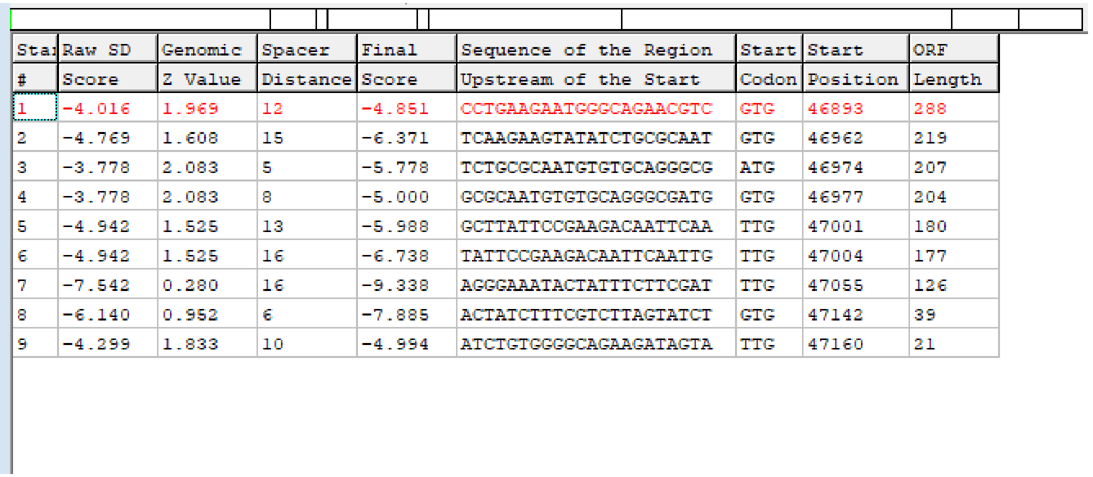
- For start site 46,893 all cp that can be included is included.
- For start 46,977 all cp that can be included is included.



RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- What is the z-value and final score? ZV: 1.969 FS: -4.851
- How does the RBS compare to that of other available starts? The RBS values are not the best ones but there are worse scores. Start 46,977 has a better ZV: 2.083 but a worse FS:-5.000
- Which start is favored based on RBS values? This is a toss up and I would rely on other information to make this call.

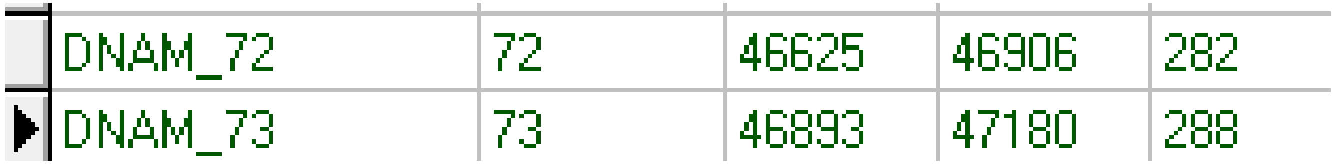
- Screenshot RBS Values here.



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	-4.016	1.969	12	-4.851	CCTGAAGAATGGGCAGAACGTC	GTG	46893	288
2	-4.769	1.608	15	-6.371	TCAAGAAGTATATCTGCGCAAT	GTG	46962	219
3	-3.778	2.083	5	-5.778	TCTGCGCAATGTGTGCAGGGCG	ATG	46974	207
4	-3.778	2.083	8	-5.000	GCGCAATGTGTGCAGGGCGATG	GTG	46977	204
5	-4.942	1.525	13	-5.988	GCTTATTCCGAAGACAATTCAA	TTG	47001	180
6	-4.942	1.525	16	-6.738	TATTCCGAAGACAATTCAATTG	TTG	47004	177
7	-7.542	0.280	16	-9.338	AGGGAAATACTATTTCTTCGAT	TTG	47055	126
8	-6.140	0.952	6	-7.885	ACTATCTTTCGTCTTAGTATCT	GTG	47142	39
9	-4.299	1.833	10	-4.994	ATCTGTGGGGCAGAAGATAGTA	TTG	47160	21

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

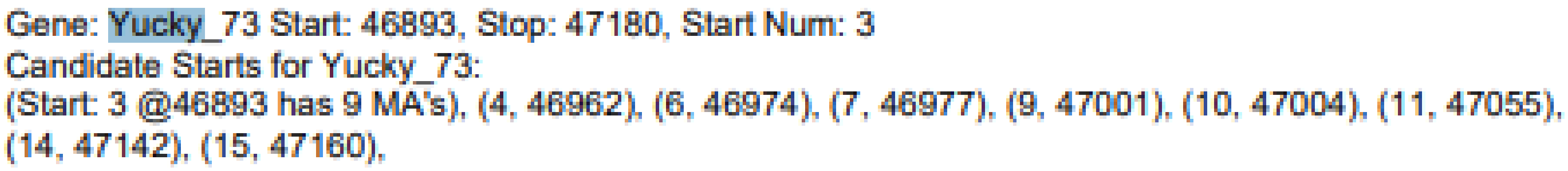
- There is an overlap of 14
- Start 46,997 has an gap of 72



DNAM_72	72	46625	46906	282
DNAM_73	73	46893	47180	288

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- For start 46,893 there are 9 MAs.
- For start 46,997 there are no MAs



Gene: Yucky_73 Start: 46893, Stop: 47180, Start Num: 3
Candidate Starts for Yucky_73:
(Start: 3 @46893 has 9 MA's), (4, 46962), (6, 46974), (7, 46977), (9, 47001), (10, 47004), (11, 47055), (14, 47142), (15, 47160),

Gene 73

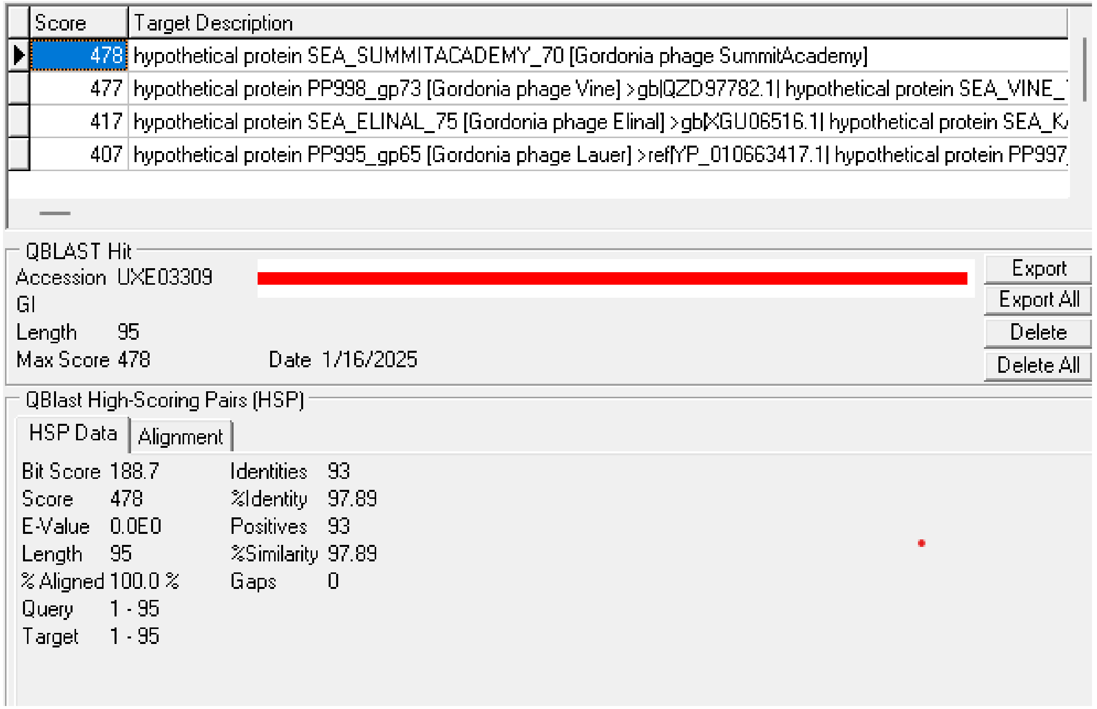
	46,893	46,997
GeneMark/Glimmer	GeneMark calls this the start	N/A
Coding Potential	There is strong cp and all cp that can be included is	There is strong cp and all cp that can be included is included.
RBS	ZV: 1.969 FS: -4.851	ZV: 2.083 FS:-5.000
Blast	There are 5 highly similar genes. They all have 1:1 alignments with E values less than 10^-7.	There are 5 highly similar genes but no 1:1 alignments.
Starterator	9 MA	N/A
Gap/Overlap	There is an overlap of 14	Gap of 72

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- The start site is 46, 893 because GeneMark calls it as the start, there is strong cp and all cp that can be included is, there are 5 highly similar genes that have 1 1:1 alignments, 9 MAs and an overlap of 14. The RBS scores are ZV: 1.969 FS: -4.851 which are not great but there are no other starts that have a better FS. Start 46,977 has a better ZValue: 2.083 but a worse Final Score:-5.000.

BLAST function evidence. What assigned functions do other highly similar genes have?

- All 5 highly similar genes are assigned the function of hypothetical protein.



Score	Target Description
478	hypothetical protein SEA_SUMMITACADEMY_70 [Gordonia phage SummitAcademy]
477	hypothetical protein PP998_gp73 [Gordonia phage Vine] >gb|QZD97782.1| hypothetical protein SEA_VINE_
417	hypothetical protein SEA_ELINAL_75 [Gordonia phage Elinal] >gb|XGU06516.1| hypothetical protein SEA_K
407	hypothetical protein PP995_gp65 [Gordonia phage Lauer] >ref|YP_010663417.1| hypothetical protein PP997

QBLAST Hit

Accession UXE03309

GI

Length 95

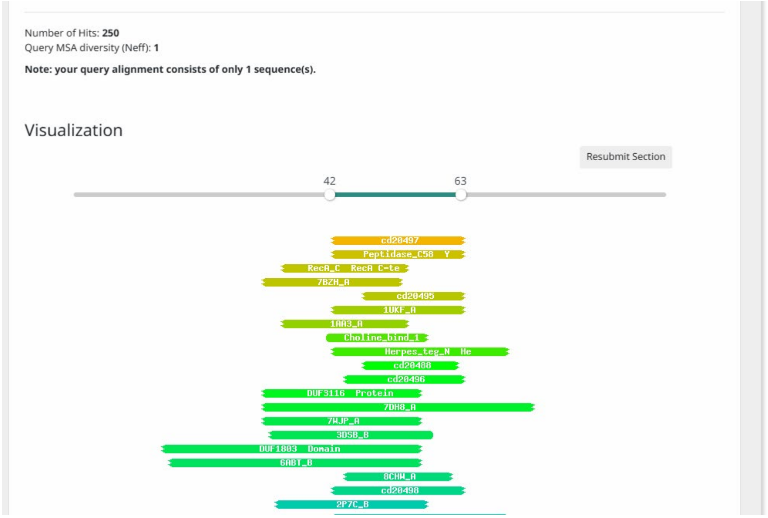
Max Score 478 Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	188.7	Identities	93
Score	478	%Identity	97.89
E-Value	0.0E0	Positives	93
Length	95	%Similarity	97.89
% Aligned	100.0 %	Gaps	0
Query	1 - 95		
Target	1 - 95		

HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

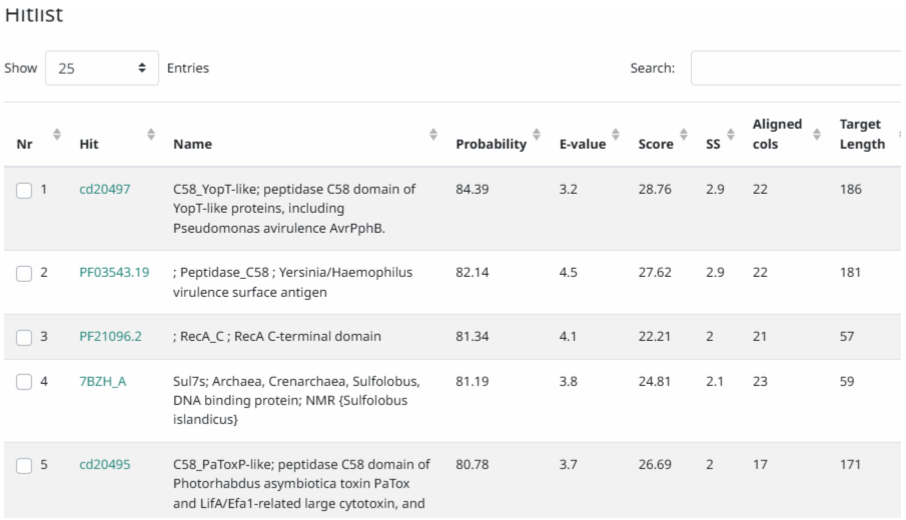
- There are no probabilities above 90% therefore the Hhpred evidence is N/A.



Number of Hits: **250**
Query MSA diversity (Neff): **1**

Note: your query alignment consists of only 1 sequence(s).

Visualization



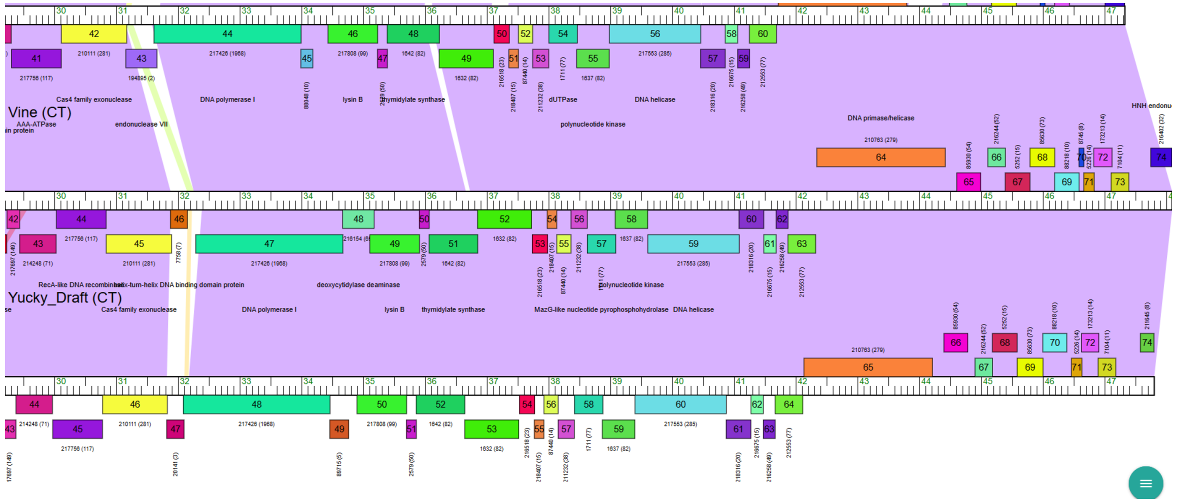
Hitlist

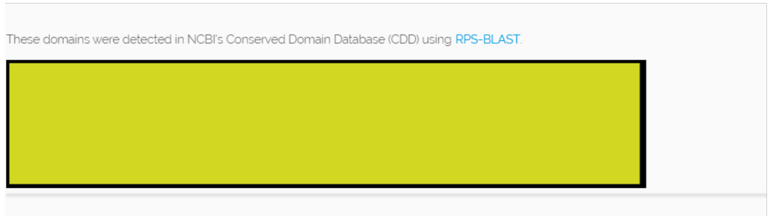
Show 25 Entries Search:

Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	cd20497	C58_YopT-like; peptidase C58 domain of YopT-like proteins, including Pseudomonas avirulence AvrPphB.	84.39	3.2	28.76	2.9	22	186
□ 2	PF03543.19	; Peptidase_C58 ; Yersinia/Haemophilus virulence surface antigen	82.14	4.5	27.62	2.9	22	181
□ 3	PF21096.2	; RecA_C ; RecA C-terminal domain	81.34	4.1	22.21	2	21	57
□ 4	7BZH_A	Sul7s; Archaea, Crenarchaea, Sulfolobus, DNA binding protein; NMR {Sulfolobus islandicus}	81.19	3.8	24.81	2.1	23	59
□ 5	cd20495	C58_PaToxP-like; peptidase C58 domain of Photorhabdus asymbiotica toxin PaTox and LifA/Efa1-related large cytotoxin, and	80.78	3.7	26.69	2	17	171

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

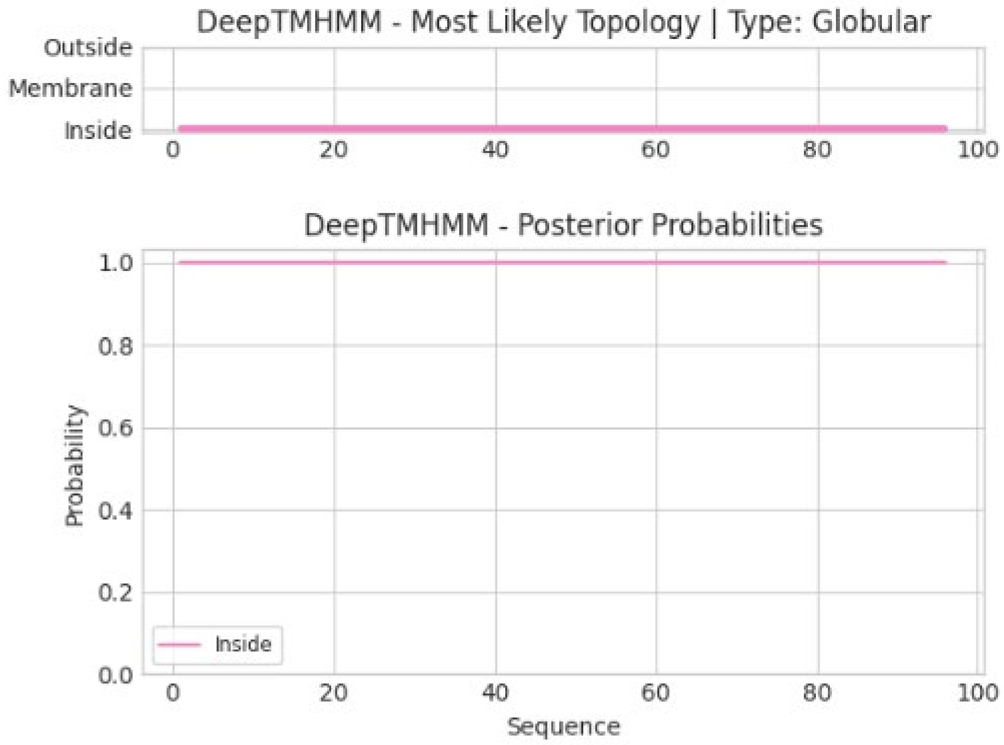
- There are no conserved domains or known functions




These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- The gene has no transmembrane domains



What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function of gene 73 is a hypothetical protein because BLAST assigned all 5 highly similar genes are assigned the hypothetical protein function, Hhpred has no probabilities above 90%, Phamerator had no conserved domains or known functions, and the protein has no transmembrane domains.

Feature 74 – Stop 47799

Glimmer/GeneMark

What feature number is this? 74

What is the stop site? 47,799

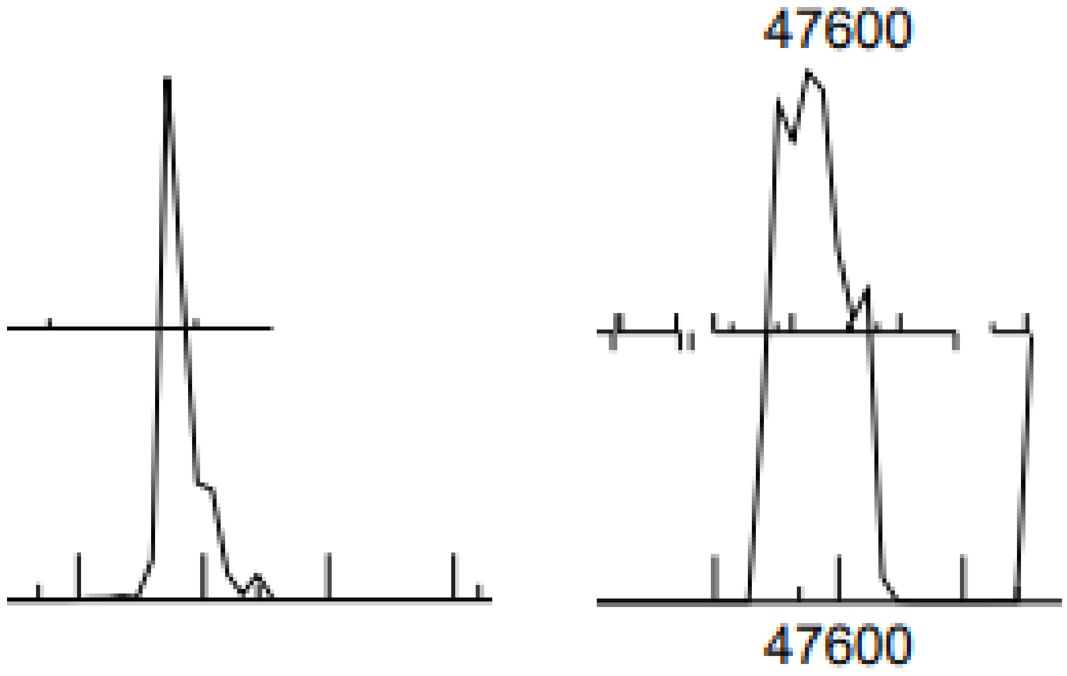
Is auto-annotated start called by both Glimmer and GeneMark, Glimmer? Only Glimmer called the start site.

What is the autoannotated start? 47,578

Gap: __397_______ or overlap: ________ (with gene in front of it) for the autoannotated start

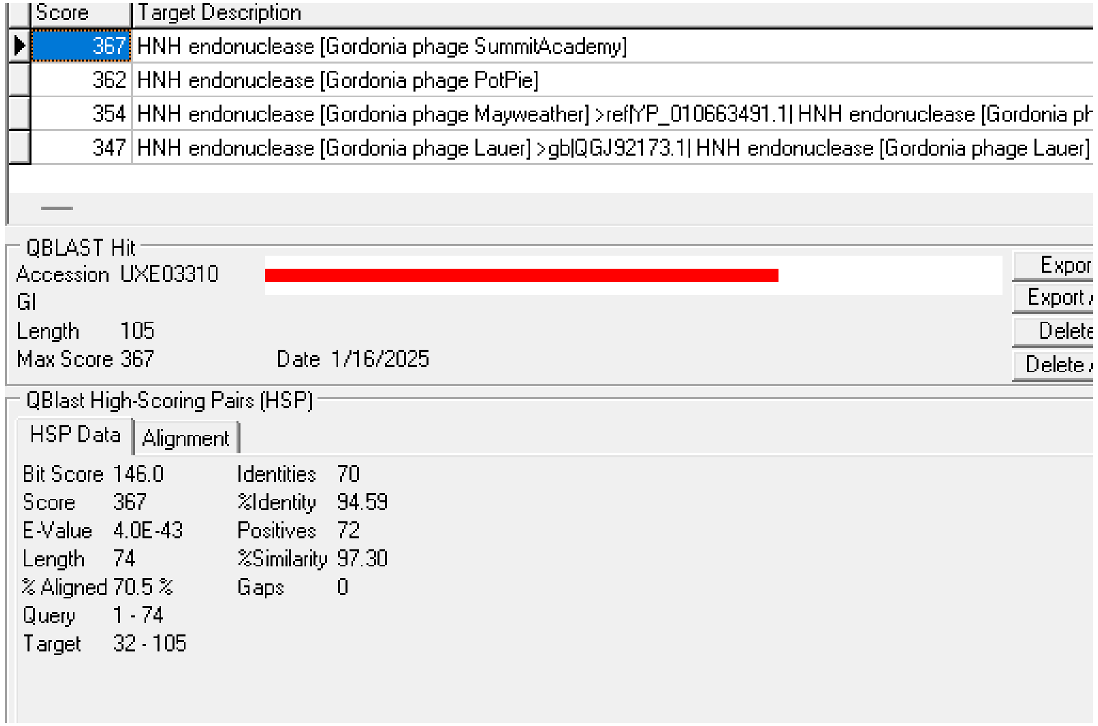
GeneMark evidence. Screenshot the coding potential graph for the predicted ORF. Answer these questions: Is it the only reading frame with cp? Describe the coding potential… is it strong or is it weak? How do you know?

- Is it the only reading frame with cp? This is the reading frame with the most cp. Frame 1 has very little cp but is very strong but goes to the stop so this frame will be used. Frame 2 has more cp and is very strong.
- Describe the coding potential… is it strong or is it weak? How do you know? The cp is strong as it has a height of 1.0.



BLAST conservation evidence. Are there other highly similar genes? How many? Screenshot BLAST evidence from DNA Master.

- There are more than 10 highly similar genes but there are no 1:1 alignments for start 46,578.



Score	Target Description
367	HNH endonuclease [Gordonia phage SummitAcademy]
362	HNH endonuclease [Gordonia phage PotPie]
354	HNH endonuclease [Gordonia phage Mayweather] >ref|YP_010663491.1| HNH endonuclease [Gordonia ph
347	HNH endonuclease [Gordonia phage Lauer] >gb|QGJ92173.1| HNH endonuclease [Gordonia phage Lauer]

QBLAST Hit

Accession UXE03310

GI

Length 105

Max Score 367

Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

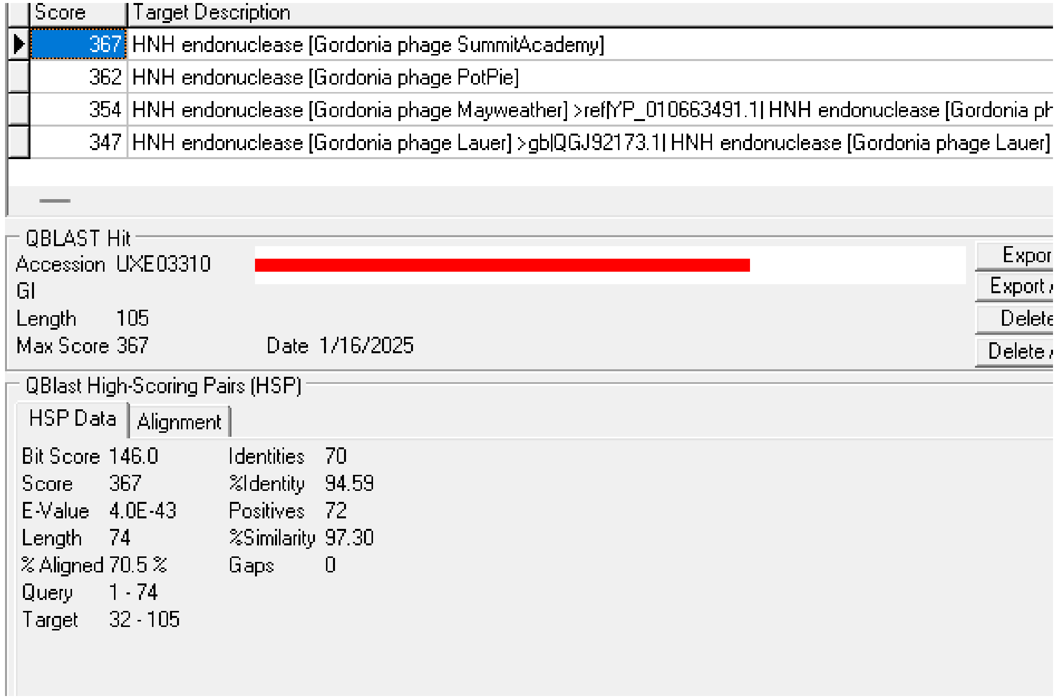
Bit Score	146.0	Identities	70
Score	367	%Identity	94.59
E-Value	4.0E-43	Positives	72
Length	74	%Similarity	97.30
% Aligned	70.5 %	Gaps	0
Query	1 - 74		
Target	32 - 105		

Answer: Is it a gene? Give evidence why you think this is a gene or not.

- Is there more than one feature called in this coding region? Yes function 74 is a gene because Glimmer calls it a gene, there is cp, and there are more than 10 highly similar genes.

BLAST alignment evidence. How does the start of this predicted gene align with the start of other highly similar genes? How many 1:1 Alignments are there for the predicted start? How many 1:1 Alignments are there for any alternative starts? Answer the question: Which start is favored based on BLAST alignment evidence.

- There are more than 10 highly similar genes but there are no 1:1 alignments. SummitAcademy and PotPie have a 1:34 alignment and Mayweather has a 1:41 alignment.
- For start 47,485 there are 6 1:1 alignments such as PotPie, SummitAademy, and Mayweather
- For start 47,509 there are no 1:1 alignments but more than 10 highly similar genes. PotPie and SummitAcademy have a 1:9 alignment and Mayweather has a 23:40 alignment.



Score	Target Description
367	HNH endonuclease [Gordonia phage SummitAcademy]
362	HNH endonuclease [Gordonia phage PotPie]
354	HNH endonuclease [Gordonia phage Mayweather] >ref|YP_010663491.1| HNH endonuclease [Gordonia ph
347	HNH endonuclease [Gordonia phage Lauer] >gb|QGJ92173.1| HNH endonuclease [Gordonia phage Lauer]

QBLAST Hit

Accession UXE03310

GI

Length 105

Max Score 367

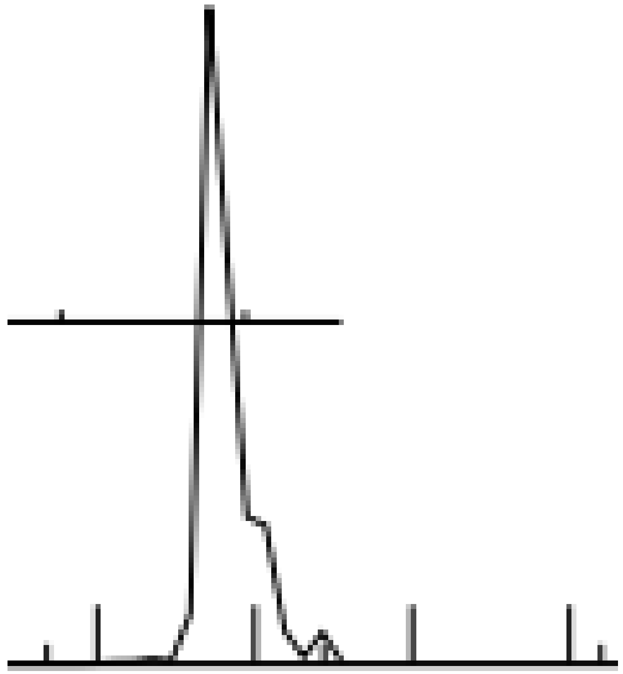
Date 1/16/2025

QBlast High-Scoring Pairs (HSP)

Bit Score	146.0	Identities	70
Score	367	%Identity	94.59
E-Value	4.0E-43	Positives	72
Length	74	%Similarity	97.30
% Aligned	70.5 %	Gaps	0
Query	1 - 74		
Target	32 - 105		

GeneMark evidence: Comment on each start site and whether or not coding potential is included or cut off. A screenshot of coding potential here is required to support your statement.

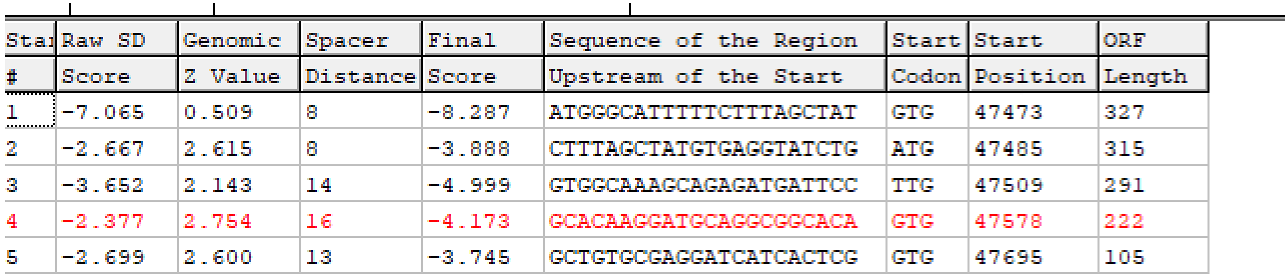
- For start sites 47,578, 47,485, and 47,509 all cp that can be included is.



RBS evidence. What is the z-value and final score? How does the RBS compare to that of other available starts? Screenshot RBS Values here. Answer the question: Which start is favored based on RBS values?

- What is the z-value and final score? ZV: is 2.754 FS: -4.173
- How does the RBS compare to that of other available starts? Two other starts have better RBS scores. Start 47,485 ZV: 2.615 FS:-3.888 and start 47,509 ZV: 2.143 FS:-4.999
- Which start is favored based on RBS values? Start 47,485 would be favored based on RBS scores.

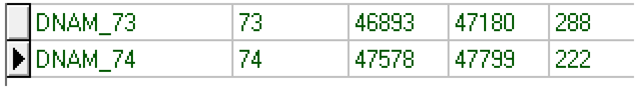
- Screenshot RBS Values here.



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	-7.065	0.509	8	-8.287	ATGGGCATTTTTCTTTAGCTAT	GTG	47473	327
2	-2.667	2.615	8	-3.888	CTTTAGCTATGTGAGGTATCTG	ATG	47485	315
3	-3.652	2.143	14	-4.999	GTGGCAAAGCAGAGATGATTCC	TTG	47509	291
4	-2.377	2.754	16	-4.173	GCACAAGGATGCAGGCGGCACA	GTG	47578	222
5	-2.699	2.600	13	-3.745	GCTGTGCGAGGATCATCACTCG	GTG	47695	105

Gap/overlap evidence. What is the gap or overlap between the start of this gene and the start or stop of the gene closest to this start? Calculate either the gap or the overlap for all proposed starts. Indicate whether the value is a gap or overlap.

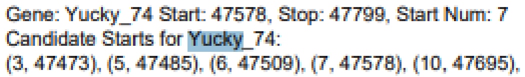
- Start 47,578 has a gap of 397
- Start 47,485 has a gap of 304
- Start 47,509 has a gap of 328



	DNAM_73	73	46893	47180	288
▶	DNAM_74	74	47578	47799	222

Starterator evidence. How many manual annotations (MAs) are there for the proposed start? How does the proposed start align with starts for other pham members? Comment on data for all proposed starts.

- There are no MAs for any of the potential starts.



Gene: Yucky_74 Start: 47578, Stop: 47799, Start Num: 7
Candidate Starts for Yucky_74:
(3, 47473), (5, 47485), (6, 47509), (7, 47578), (10, 47695),

Gene 74

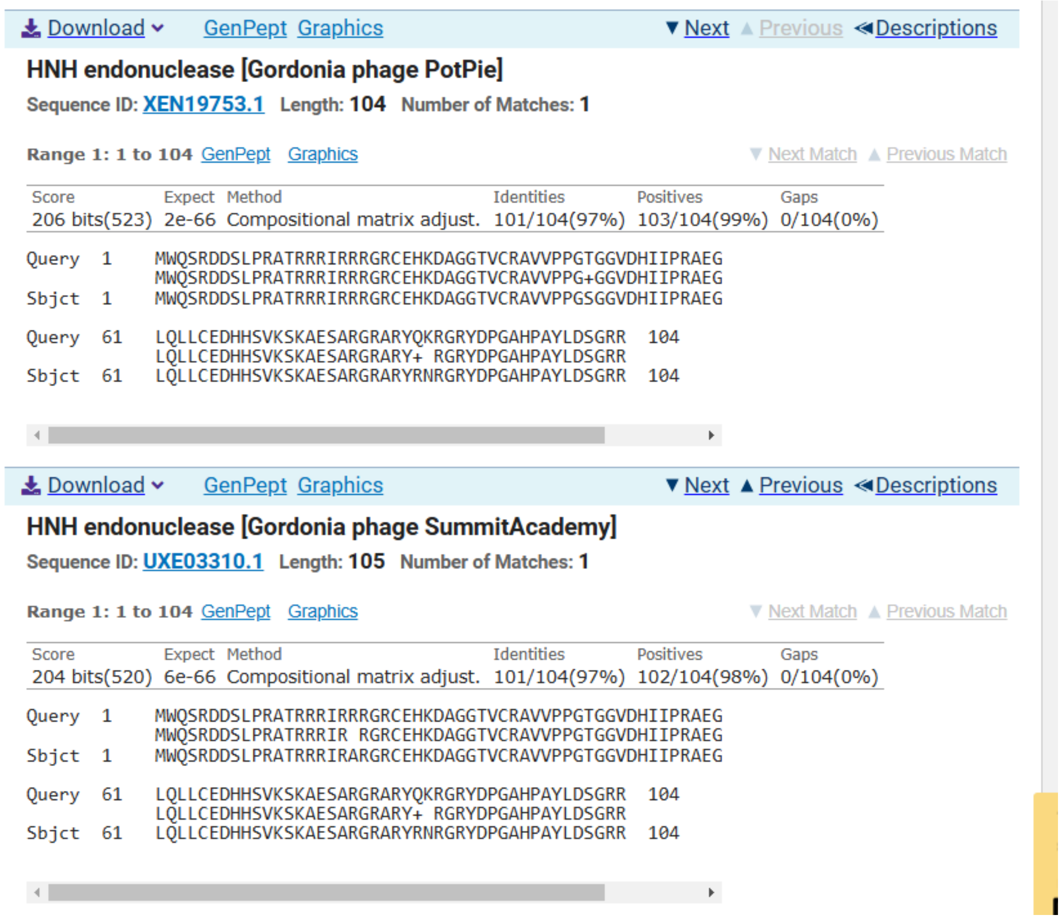
	47,578	47,485	47,509
Glimmer/GeneMark	Only Glimmer called this the start	N/A	N/A
Coding Potential	all cp that can be included is included.	All cp that can be included.	All cp that can be included is included.
RBS	ZV:2.754 FS: -4.173	ZV: 2.615 FS:-3.888	ZV: 2.143 FS:-4.999
Blast	There are no 1:1 alignments but 10 highly similar genes	There are 6 1:1 alignments	There are no 1:1 alignments but more than 10 highly similar genes
Gap/Overlap	Gap of 397	Gap of 304	Gap of 328
Starterator	No MA	No MA	No MA

What is the start site? Answer the following questions: What are you calling the start site. Does it agree with the automated start site. What evidence do you have to support the manually annotated start site?

- Start site is 47,485 because all cp that can be included is included, the zv: 2.615 and FS:-3.888, 6 1:1 blast alignments, and a gap of 304.

BLAST function evidence. What assigned functions do other highly similar genes have?

- All highly similar genes gave the function as HNH endonuclease such as PotPie, SummitAcademy, and Mayweather. They all have 1:1 alignments.



HNH endonuclease [Gordonia phage PotPie]

Sequence ID: XEN19753.1 Length: 104 Number of Matches: 1

```
Range 1: 1 to 104

Score          Expect  Method                        Identities     Positives      Gaps
206 bits(523)  2e-66   Compositional matrix adjust.  101/104(97%)   103/104(99%)   0/104(0%)

Query  1    MWQSRDDSLPRATRRRIRRRGRCEHKDAGGTVCRAVVPPGTGGVDHIIPRAEG
            MWQSRDDSLPRATRRRIRRRGRCEHKDAGGTVCRAVVPPG+GGVDHIIPRAEG
Sbjct  1    MWQSRDDSLPRATRRRIRRRGRCEHKDAGGTVCRAVVPPGSGGVDHIIPRAEG

Query  61   LQLLCEDHHSVKSKAESARGRARYQKRGRYDPGAHPAYLDSGRR  104
            LQLLCEDHHSVKSKAESARGRARY+ RGRYDPGAHPAYLDSGRR
Sbjct  61   LQLLCEDHHSVKSKAESARGRARYRNRGRYDPGAHPAYLDSGRR  104
```

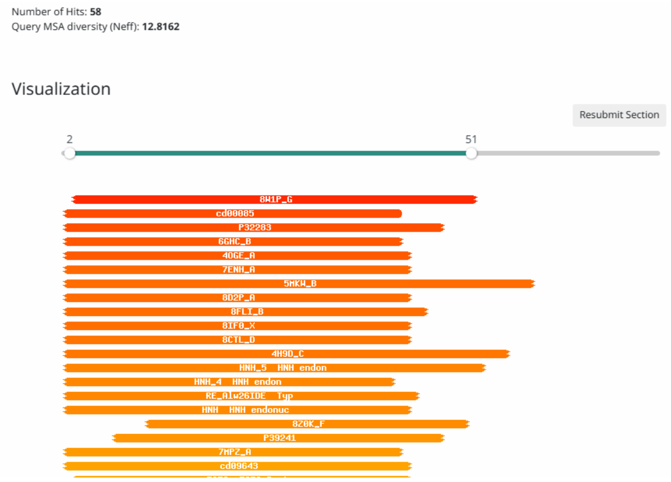
HNH endonuclease [Gordonia phage SummitAcademy]

Sequence ID: UXE03310.1 Length: 105 Number of Matches: 1

```
Range 1: 1 to 104

Score          Expect  Method                        Identities     Positives      Gaps
204 bits(520)  6e-66   Compositional matrix adjust.  101/104(97%)   102/104(98%)   0/104(0%)

Query  1    MWQSRDDSLPRATRRRIRRRGRCEHKDAGGTVCRAVVPPGTGGVDHIIPRAEG
            MWQSRDDSLPRATRRRIR RGRCEHKDAGGTVCRAVVPPGTGGVDHIIPRAEG
Sbjct  1    MWQSRDDSLPRATRRRIRARGRCEHKDAGGTVCRAVVPPGTGGVDHIIPRAEG

Query  61   LQLLCEDHHSVKSKAESARGRARYQKRGRYDPGAHPAYLDSGRR  104
            LQLLCEDHHSVKSKAESARGRARY+ RGRYDPGAHPAYLDSGRR
Sbjct  61   LQLLCEDHHSVKSKAESARGRARYRNRGRYDPGAHPAYLDSGRR  104
```

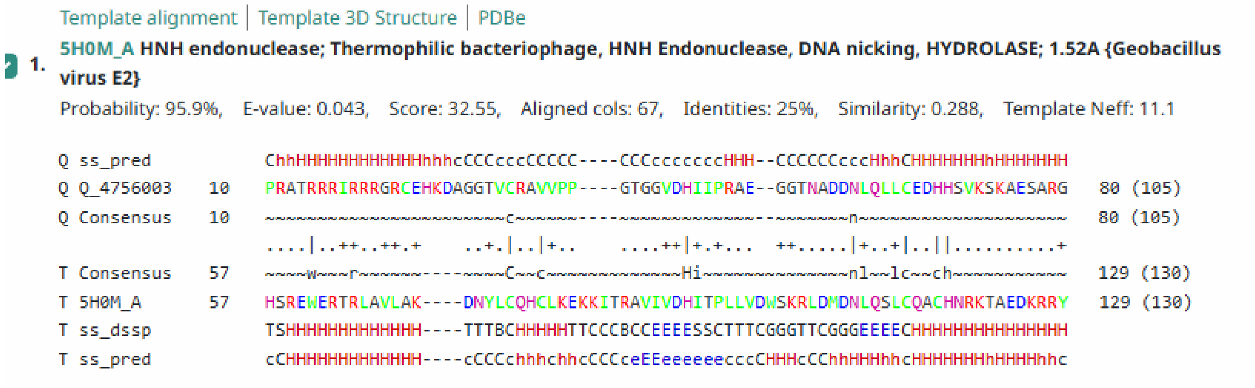
HHpred evidence. Does HHpred data support a function assignment for this gene? Describe the functions of highly similar matches. Is most of the gene homologous, or just a region? Are there conserved domains? A screenshot here of HHPRED results is desired.

- The most similar match has the assigned function of HNH endonuclease. This has a probability of 95.9%. For this to be an endonuclease it must have H-N-H over a 30 aa span in which it does. Pointing to this gene being an HNH endonuclease.



Number of Hits: **58**
Query MSA diversity (Neff): **12.8162**

Visualization



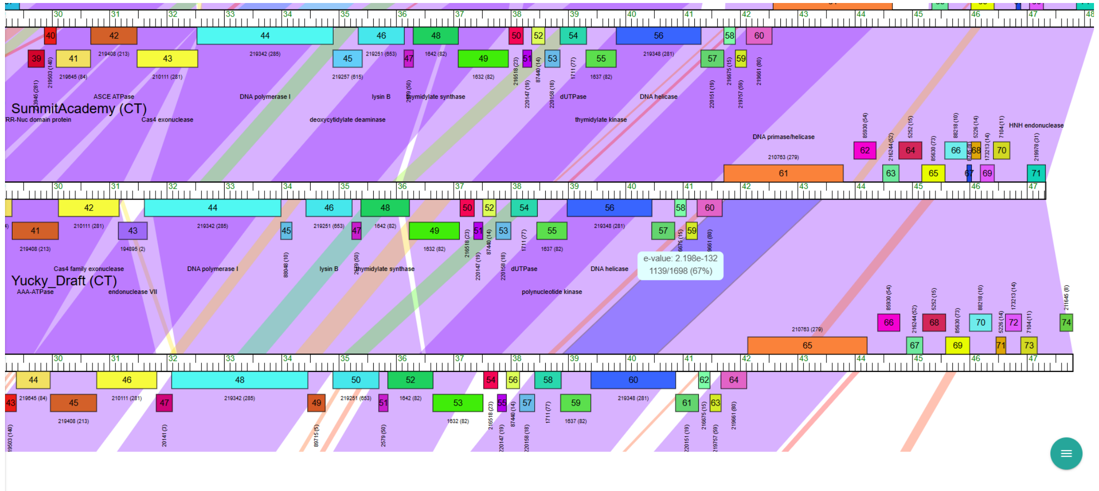
1. **5H0M_A HNH endonuclease; Thermophilic bacteriophage, HNH Endonuclease, DNA nicking, HYDROLASE; 1.52A {Geobacillus virus E2}**

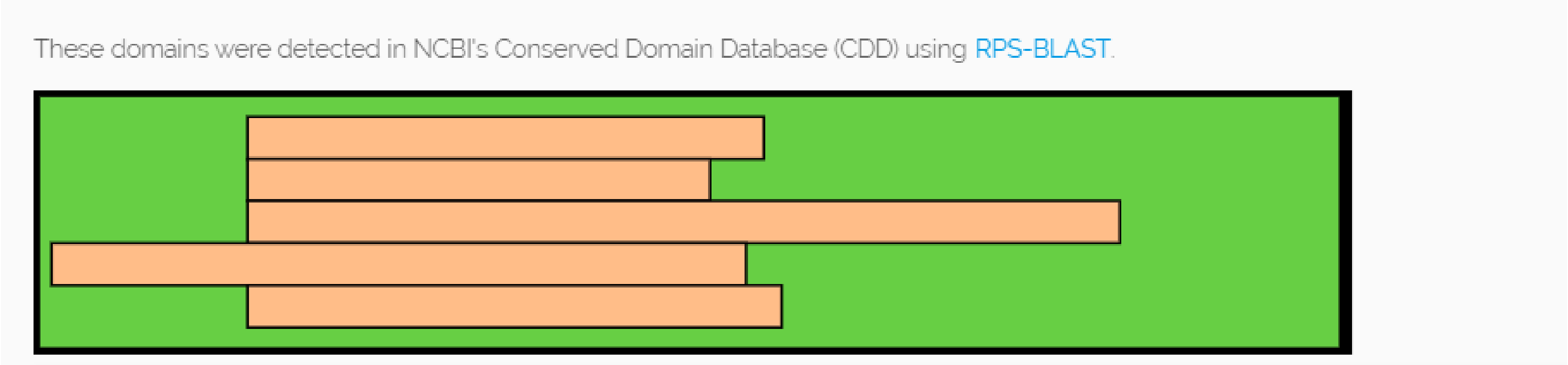
Probability: 95.9%, E-value: 0.043, Score: 32.55, Aligned cols: 67, Identities: 25%, Similarity: 0.288, Template Neff: 11.1

```
Q ss_pred              ChhHHHHHHHHHHHHhhhcCCCcccCCCCC----CCCcccccccHHH--CCCCCCcccHhhCHHHHHHHhHHHHHHH
Q Q_4756003     10    PRATRRRIRRRGRCEHKDAGGTVCRAVVPP----GTGGVDHIIPRAE--GGTNADDNLQLLCEDHHSVKSKAESARG    80 (105)
Q Consensus     10    ~~~~~~~~~~~~~~~~~~~~~~~c~~~~~~----~~~~~~~~~~~~~--~~~~~~~n~~~~~~~~~~~~~~~~~~    80 (105)
                       ....|..++..++.+    ..+.|..|+..    ....++|+.+...  ++.....|+..+|..||..........+
T Consensus     57    ~~~~w~~~r~~~~~~----~~~~C~~c~~~~~~~~~~~~~Hi~~~~~~~~~~~~~~nl~~lc~~ch~~~~~~~~~~~   129 (130)
T 5H0M_A        57    HSREWERTRLAVLAK----DNYLCQHCLKEKKITRAVIVDHITPLLVDWSKRLDMDNLQSLCQACHNRKTAEDKRRY   129 (130)
T ss_dssp             TSHHHHHHHHHHHHH----TTTBCHHHHHTTCCCBCCEEEESSCTTTCGGGTTCGGGEEEECHHHHHHHHHHHHHHH
T ss_pred              cCHHHHHHHHHHHHH----cCCCchhhchhcCCCceEEeeeeeecccCHHHcCChhHHHhhcHHHHHHHhHHHHhhc
```

Phamerator evidence. Do closely related phages with genes in the same pham predict a function for this gene? Are there conserved domains?

- There are 5 conserved domains such as 2 HNHc (nucleases), McrA (restriction endonuclease), HNH_5 (endonuclease), and HNH (endonuclease). No known functions.





These domains were detected in NCBI's Conserved Domain Database (CDD) using RPS-BLAST.

Deep TMHMM evidence. If there is no known function for the gene, search for transmembrane domains. Screenshot and describe DeepTMHMM evidence below.

- N/A as this is not a hypothetical protein.

What is the function? What official SEA-PHAGES function are you assigning to this gene? Justify your rationale for this function choice (include BLAST, Hhpred, and Phamerator evidence)

- The function of Gene 74 is a HNH Endonuclease because blast calls all highly similar genes HNH Endonucleases, Hhpred calls highly similar genes endonucleases and the 1st similar gene is an HNH endonuclease, and Phamerator has 5 conserved domains which are mainly HNH endonucleases.