## Positional and functional annotation 1-15

*\*In case of two entries for one gene, the chosen gene is highlighted in green and the unchosen in red\**

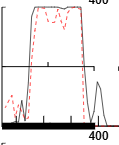
| **Gene #** | **Start/Stop coordinates** | **CP: Does the prediction cover all the coding potential?** | **SCS: Is this start position predicted by both Genemark and Glimmer?** | **Blast: Best blastp match** | **Gap: Gap/overlap** | **LO: Is this gene the longest?** | **RBS: scores, spacing, for start you chose** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TomBrady #1 | SSC: 44-388 | CP: yes | SCS: Glimmer | Blast: best match is Angel Gene 1 with a 1:2 match | Gap: none | Yes | Spacer  10  Zscore  2.202  Final Score  -4.549 |
| TomBrady #1 | SSC: 47 - 388 | CP:no | SCS: neither | Blast best match | Gap: none | No | Spacer  9  Zscore  2.447  Final Score  -4.152 |
| TomBrady #2 | SSC: 360-1892 | CP: Yes | SCS: neither | Blast best match is Schiebel Gene 2 w/ 1:1 match | Gap: 27 bp overlap | Yes | Spacer  11  Zscore  2.007  Final Score  -4.991 |
| TomBrady #3 | SSC: 1889 - 3404 | CP: Not necessarily.  The CP of the bacteria extends a little bit past the end of the predicted end codon | SCS: Glimmer and GeneMark | Blast best match is Schlebeil gene 3 w/ a 1 to 1 match | Gap: 4 bp overlap w/ previous gene | No | Spacer  12  Zscore  2.281  Final Score  -4.536 |
| TomBrady #4 | SSC: 3404-6193 | CP: yes | SCS: Glimmer and GeneMark | Best blast match is 1:1 match w/ ANGEL\_4 | Gap: 0 | Yes | Spacer  10  Zscore  1.776  Final Score  -5.400 |
| TomBrady #5 | SSC: 6190 - 6396 | CP: yes | SCS: Glimmer and GeneMark | Best blast match is a 1:1 match w/ ANGEL\_5 | Overlap: 4 | Yes | Spacer  13  Zscore  1.901  Final Score  -5.552 |
| TomBrady #6 | SSC: 6515 - 7057 | CP: yes | SCS: Glimmer and GeneMark | Best blast match is a 1:1 match w/ Scaffold protein in ANGEL | Gap: 118 | Yes | Spacer  13  Zscore  2.313  Final Score  -4.684 |
| TomBrady #9 | SSC: 8315-8812 | CP: Yes | SCS: Glimmer and GeneMark | Best blast match a 21:21 match w/ head-to-tail adapter protein | Gap:11 | Yes | Spacer  15  Zscore  1.974  Final Score  -5.029 |
| TomBrady #11 | SSC: 9155-6418 | CP: Yes | SCS: Glimmer and GeneMark | Best blast match is 1:1 w/ hypothetical protein Angel | Overlap: 14 | Yes | Spacer  12  Zscore  3.38  Final Score  -2.394 |
| TomBrady #12 | SSC: 9430 - 9858 | CP: No | SCS: GeneMark | Best blast match is 1:6 w/ the tail terminator protein in Angel | Gap: 11 | No | Spacer  9  Zscore  1.115  Final Score  -6.748 |
| TomBrady #13 | SSC: 9855 - 10469 | CP: No | SCS: Glimmer and GeneMark | Best blast match is 1:1 w/ major tail subunit protein in Angel | Overlap: 4 | No | Spacer  11  Zscore  2.736  Final Score  -3.571 |
| TomBrady #14 | SSC: 10570 - 11079 | CP: No | SCS: Glimmer and GeneMark | Best blast match is w/ a 1:1 w/ tail assembly chaperone protein in Angel | Gap: 100 | No | Spacer  12  Zscore  3.201  Final Score  -2.742 |
| TomBrady #15 | SSC: 11088 - 11465 | CP: No | SCS: Glimmer and GeneMark | Best blast match is 1:1 w/ hypothetical protein M695\_gp15 in Phage Leo | Gap: 8 | No | Spacer  8  Zscore  2.458  Final Score  -4.577 |
| TomBrady # 7 | SSC: 7104-8039 | CP: Yes | SCS: Glimmer and GeneMark | Best blast match is a 1:! Match w/ capsid protein in Angel | Gap: 46 | Yes | Spacer  12  Zscore  2.517  Final Score  -4.075 |
| TomBrady # 8 | SSC: 8076 - 8303 | CP: Yes | SCS: Glimmer and GeneMark | Best blast match is 1:1 w/ gene 8 of Angel | Gap:36 | Yes | Spacer  10  Zscore  3.091  Final Score  -2.817 |
| TomBrady # 10 | SSC: 8812-9168 | CP: Yes | SCS: Glimmer and GeneMark | Best blast match 1:1 match w/ head-to-tail stopper protein | Overlap: 1 | No | Spacer  11  Zscore  2.154  Final Score  -4.705 |

## **Image Documentation:**

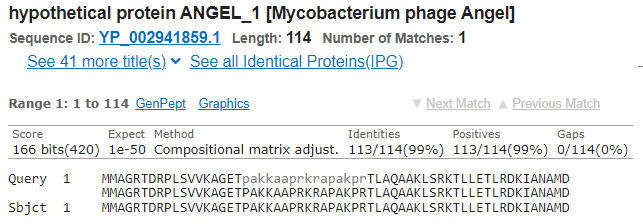
*This section contains snips of the GeneMark map as well as the closest BlastP match to the gene selected*

*Tombrady Gene #1*

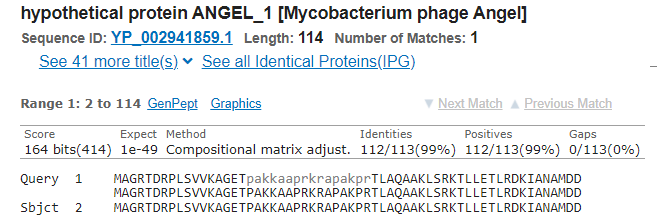
GeneMark:



BlastP Closest Match - Start BP 44 :



BlastP Closest Match - Start BP 47



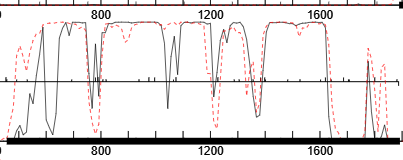
Reasoning:

This one was tricky and I ended up selecting the gene with a start codon at bp 44 instead of bp 47. This was a gene that had very similar scores with **no overlap or gaps**. The scores were both quite similar with no significant difference between the two. The main difference is that there was a 1 codon difference between the two genes. It brings up a interesting argument because that would result in two start codons consecutively back to back. The main reason why the gene with a **start codon at bp 44 was chosen** was due to the 1**:1 match w/ bacteriophage ANGEL\_1 instead of a 1:2 match whilst starting at bp 47 with the same phage.**

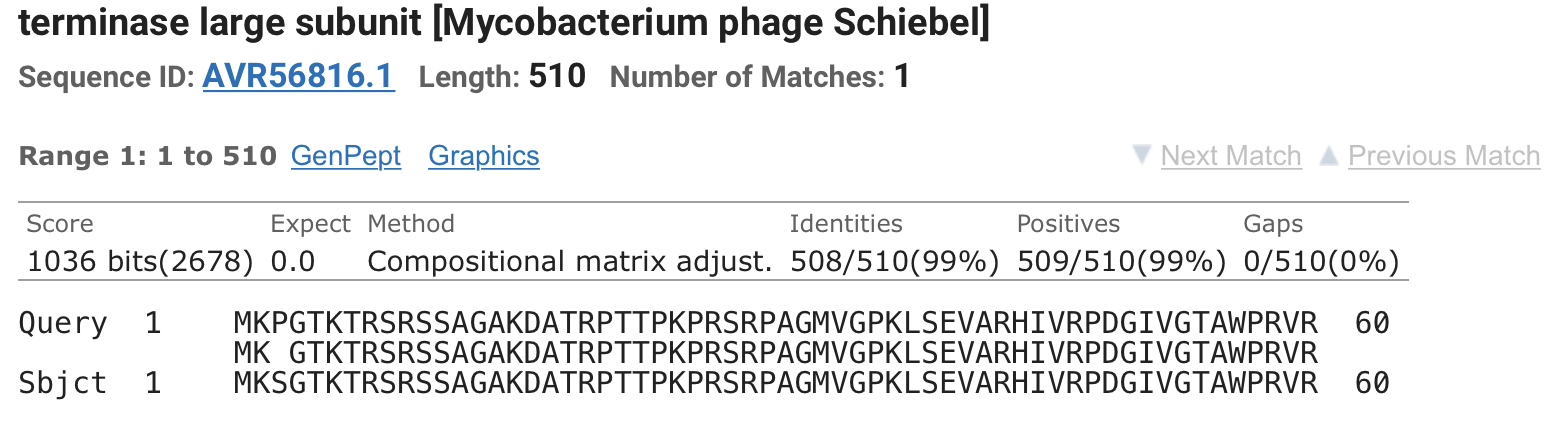
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #2*

GeneMark:



BlastP Closest Match



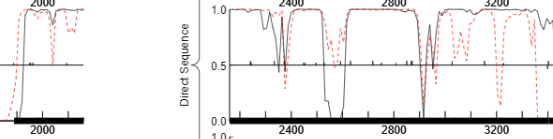
**Rationale**:

There was a decision that had to be made between **the predicted start and stop by both GeneMark/Glimmer and the ORF**. The predicted values for both Glimmer and GeneMark had admirable scores, however the main reason why this SCS frame was chosen was due to the fact that the predicted start and stop had a gap between the previous gene that was nearly 2x the length of an overlap between the frame chosen and the previous gene; this would leave **a nearly 100bp difference b**etween the Glimmer/GeneMark predicted start and the one I chose. I came to the conclusion that generally gaps are preferred to overlaps, however w/ a 100 nucleotide exclusion if the predicted was chosen I didn’t trust that no genes woulda been left out. **The Fasta GeneMark map also supported this decision.**

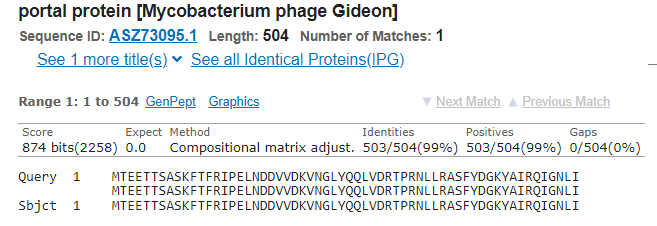
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

TomBrady Gene #3

GeneMark:



BlastP Closest Match:

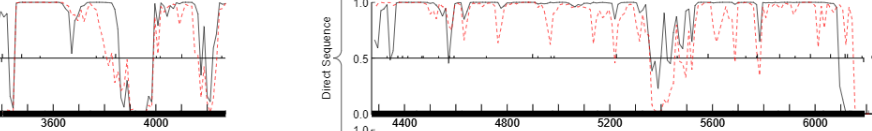


**Rationale:** The start codon chosen was **supported by both Glimmer and GeneMark** predicted start codon, as well as visually in the GeneMark map. The scores were overall the best, however the gene chosen **was not the longest Open Reading Frame.** The longest ORF had an overlap of 571 codons whereas the gene I chose only one of 4; it was almost an instant decision to discount the longest ORF and continue forth with the one I chose. This codon also has a **1:1 Blastp match w/ Mycobacterium phage Gideon.**

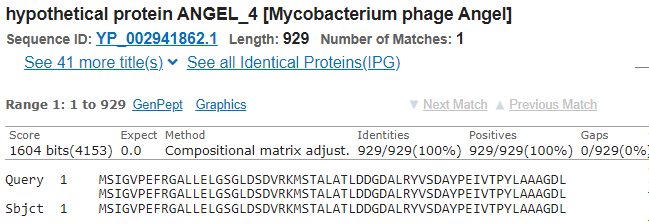
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #4*

GeneMark



BlastP Closest Match

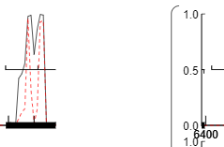


**Rationale:** This choice was somewhat of a no brainer. The first thing I noticed was **a gap of 0 bps**. This left me with zero possibility for loss of coding potential **(LORF)** w/ both Glimmer and GeneMark agreeing. The scores were also decently compared to the rest, although, I think they were terrible. The other main reason I went with this option was that when entered into the Blastp database, I had a **1:1 match w/ Angel gene #4.**

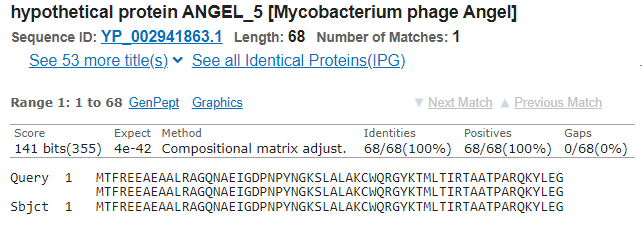
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #5*

GeneMark:



BlastP Closest Match

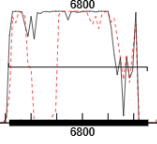


**Rationale:** Glimmer and Genemark only gave me 2 options for genes to pick from. This resulted in me choosing the option that had the lowest gap/overlap **(overlap of 4 bps)**. It had less than appreciable scores, but they were better than the rest. It was a relatively small gene and had a **1:1 match with Angel gene 5.** Visually speaking, the GeneMark map matched with what the predicted values were.

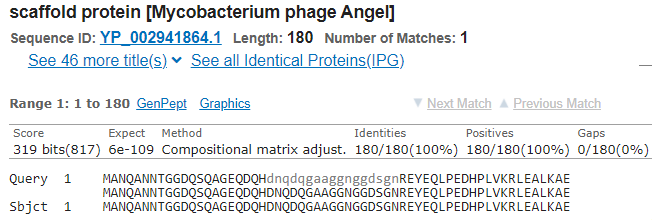
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #6*

GeneMark



BlastP Closest Match



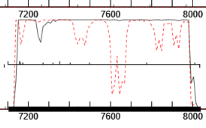
**Rationale:** The predicted values for the SSC codon were the same across Glimmer and GeneMark w/ a visual analysis revealing the start and stop containing the entire gene, although there might be a section that isn't included. The gene chosen was the **longest open reading frame**. The gap between this gene and the previous one is **a gap of 118 bp**, however that is the smallest gap out of the options provided and **visually matches with the GeneMark graph**. There is a **1:1 match between this gene and the scaffold protein of Angel.**

*Is it possible for a gene to exist on the reverse strand that is the gap.*

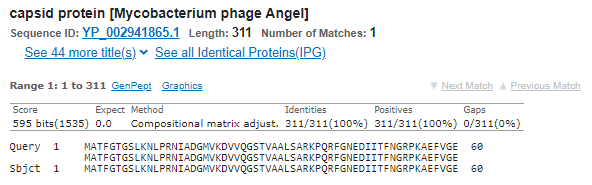
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #7*

GeneMark



BlastP Closest Match

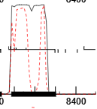


**Rationale:** This gene was chosen, with consideration from both GeneMark and Glimmer. Both of the predicted starts were the same, which gave me a vague idea as to where I thought the start codon would be. In addition, the SSC chosen had the **smallest gap** between the end of the previous gene and the start of this one. This meant that it would have the **Longest Open Reading Frame,** which was confirmed by Pecaan, and result in the least amount of potential gene loss. This decision was backed up by a **1:1 match w/ the capsid protein of Angel**

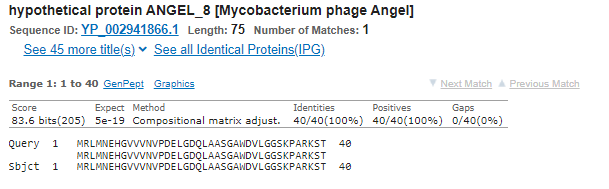
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #8*

GeneMark



BlastP Closest Match

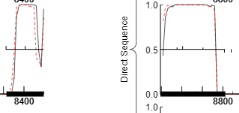


**Rationale:** The main reason why this gene was chosen was because there was a 36 bp gap between this and previous gene; this is a value that is dangerously close to the max 30 bp gap for highest accuracy. This gap was the smallest and also had the largest open reading frame. There was a direct 1:1 match w/ the phage Angel and the scores were the best, when compared to the others.

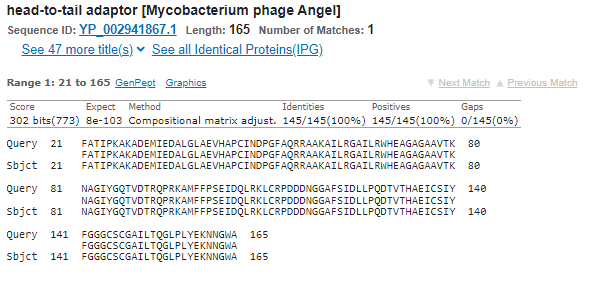
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #9*

GeneMark



BlastP Closest Match

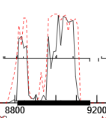


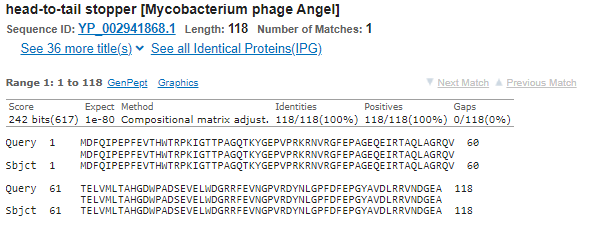
**Rationale:** The other possible gene SSC was one that started 30+ bp after the one chosen. This gene also as a 21:21 match to the head-to-tail adapter gene in the Angel phage. Scores for this specific gene weren't the best, and there were better options available, however the decision was made to stick to this one due to the nature of this frame being the largest and containing the most coding potential.

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #10*

GeneMark

BlastP Closest Match

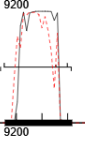


**Rationale:** This gene was chosen mostly due to the fact that it had such a small overlap/gap between the end of the last one and the start of its own gene. With an overlap of 1 bp, the chosen was pretty obvious, and this combined with the visual validation from the GeneMark map starting at the same spot. Finally, the 1:1 match with the head-to-tail stopper gene of Angel.

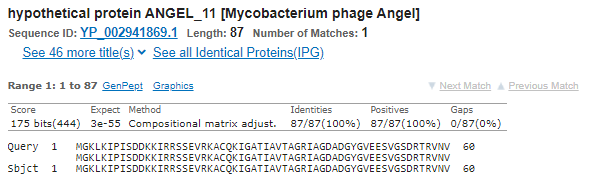
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #11*

GeneMark



BlastP Closest Match

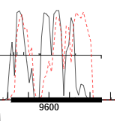


**Rationale:** Has the smallest gap/overlap . The rest of the over predicted start/stop markers had gaps over 200 between the end of the previous gene and the start of this one. It was something that allowed me to pretty much cross the other predicted start/stops out; the 1:1 match w/ the Angel protein also allowed me to determine that this was the best position for the start.

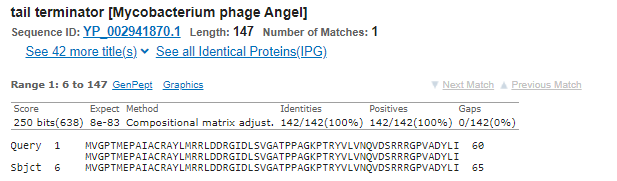
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #12*

GeneMark



BlastP Closest Match

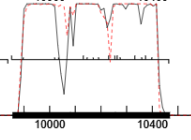


**Rationale:** I was able to narrow the choices down to 2 potential SSCs. One had an larger overlap and one had a slight gap. I ended up choosing the gene with a slight overlap. This was mostly because the other overlaps were too large for me to confidently choose them. This decision was validated by both **GeneMark**. The scores were less that optimal, however they were the both when all factors were taken into account.

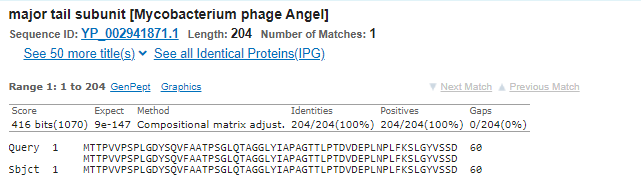
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #13*

GeneMark:



BlastP Closest Match

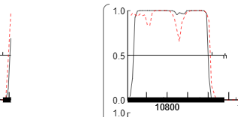


**Rationale:** This was one that required me to reference the GeneMark map and visually attempt to validate my choice. All the scores were relatively similar, so looking for the relatively starting point of the shrug of the gene allowed me to determine the beginning. There was a 1:1 match w/ the major tail subunit protein from Angel which allowed me confirm that this was the best choice for length of the gene. It was not the largest ORF.

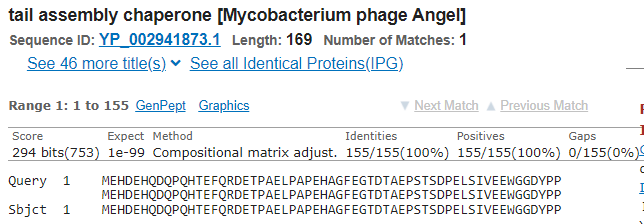
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #14*

GeneMark



BlastP Closest Match



**Rationale:** The scores were the most balanced when compared to others and both GeneMark and Glimmer predicted the same start bp. This is not the largest ORF, but the one that is designated the larged ORF has an insane overlap. There was another choice to pick which had a smaller gap, but when referring to the GeneMark map, this choice is the one that made the most sense.

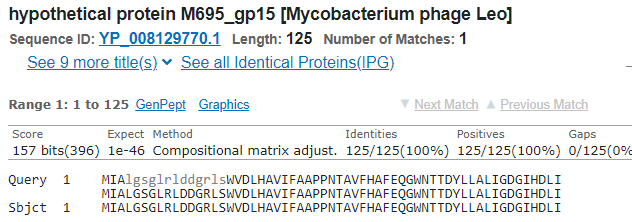
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*TomBrady Gene #15*

GeneMark



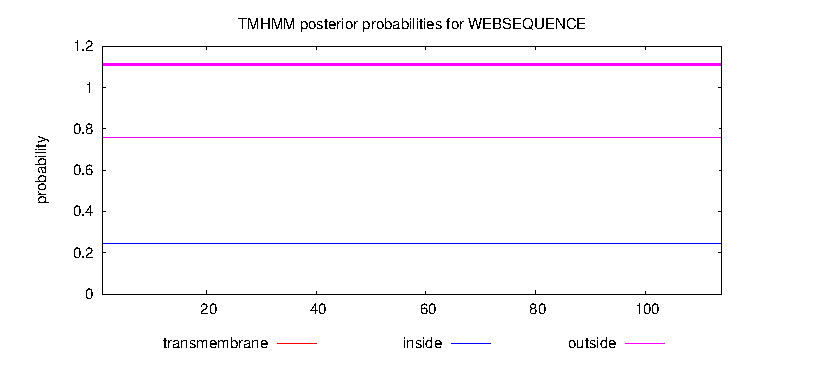
BlastP Closest Match



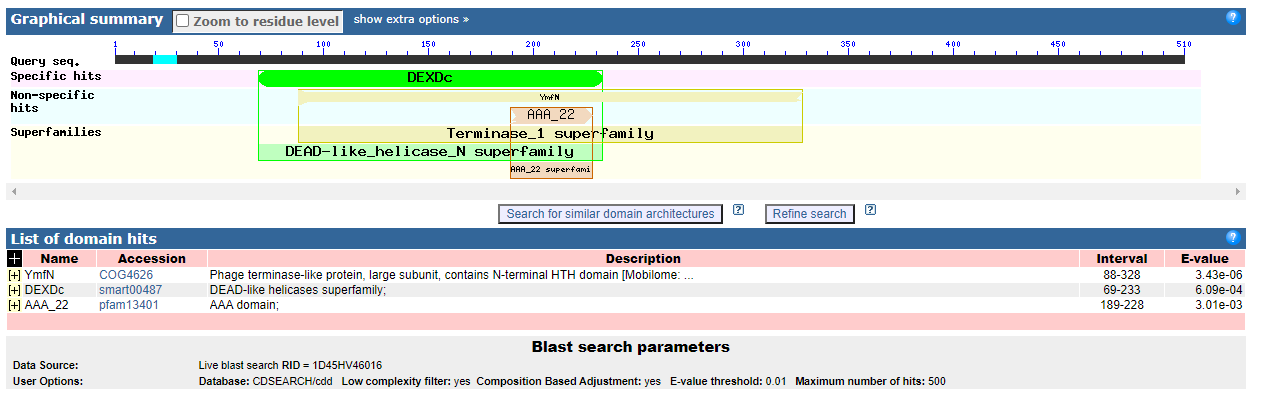
**Rationale:** This gene was one that fairly easy to determine. The scores were not great, but with both Glimmer and GeneMark predicting the same start/stop, as well as a 1:1 match to a protein in phage Leo, as well as a visual semi-confirmation, it was the best choice. The gap was also quite low compared to some of the other ones w/ a spacer under 14, both of which only bolster this SSC.

## Functional Analysis of Genes

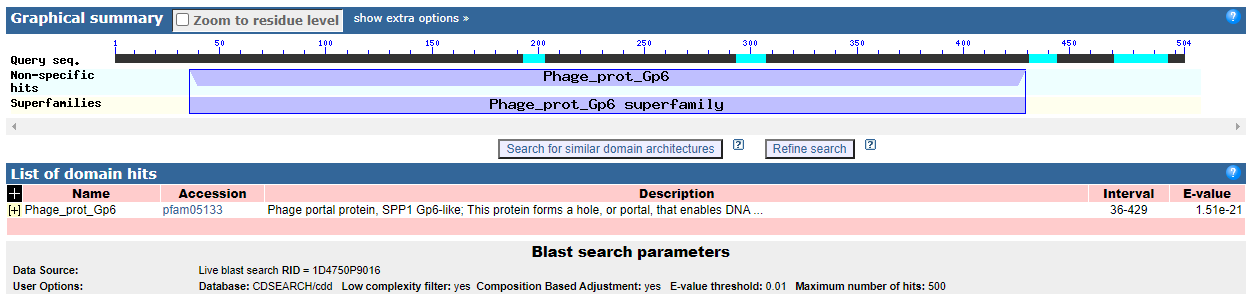
***TomBrady Gene 1 -No protein previously identified***

******

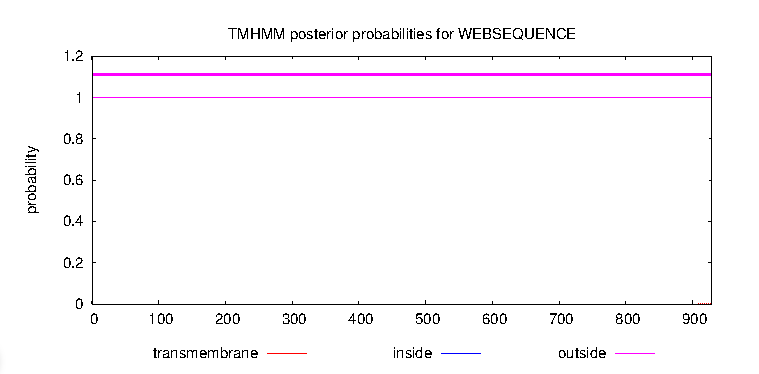
***TomBrady Gene 2 -Terminase-Like Protein***

******

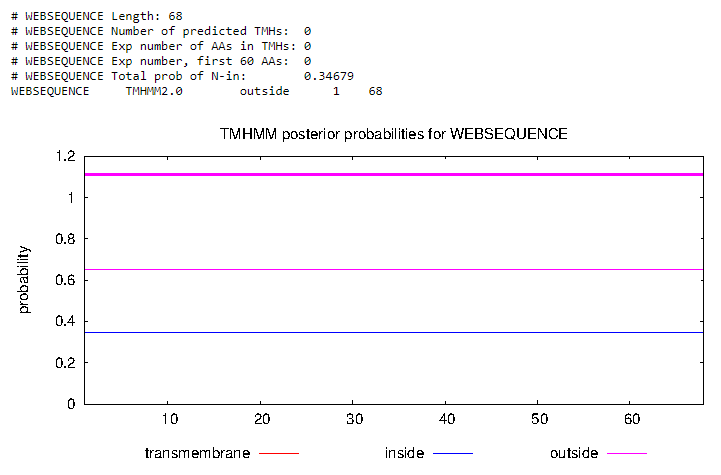
***TomBrady Gene 3 - Portal Protein***

******

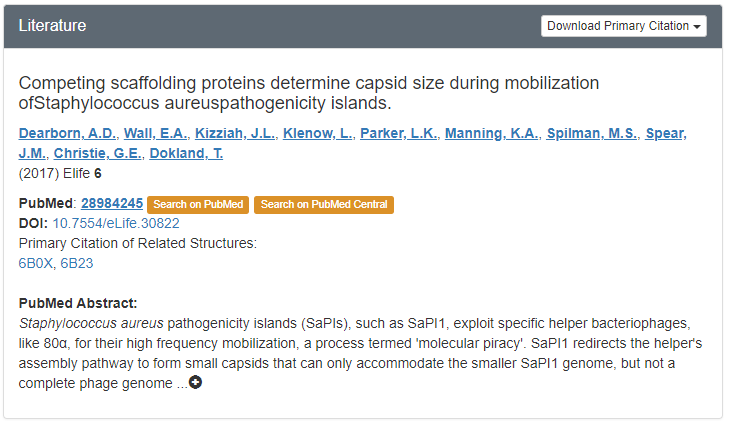
***TomBrady Gene 4 -No protein previously identified***

******

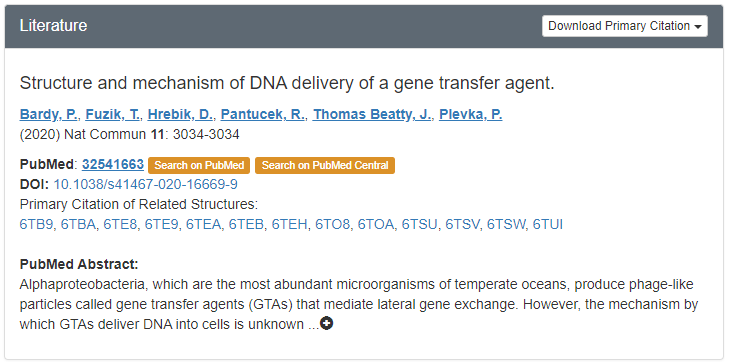
***TomBrady Gene 5 - No previous Protein Found***

******

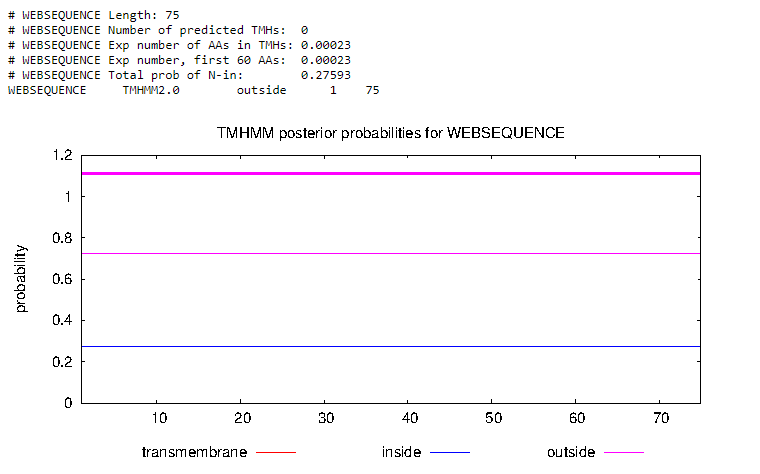
***TomBrady Gene 6 - Scaffolding Protein***

******

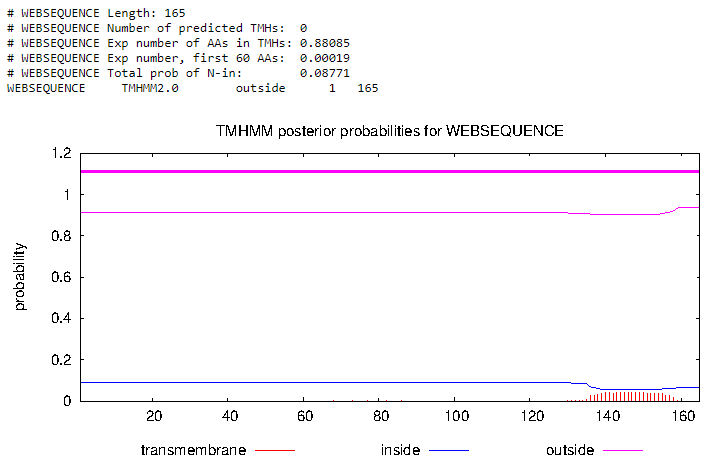
***TomBrady Gene 7 - Major Capsid Protein***

******

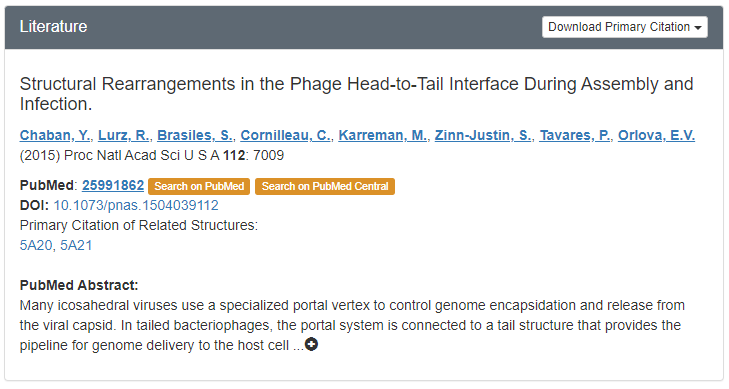
***TomBrady Gene 8 -No protein previously identified***

******

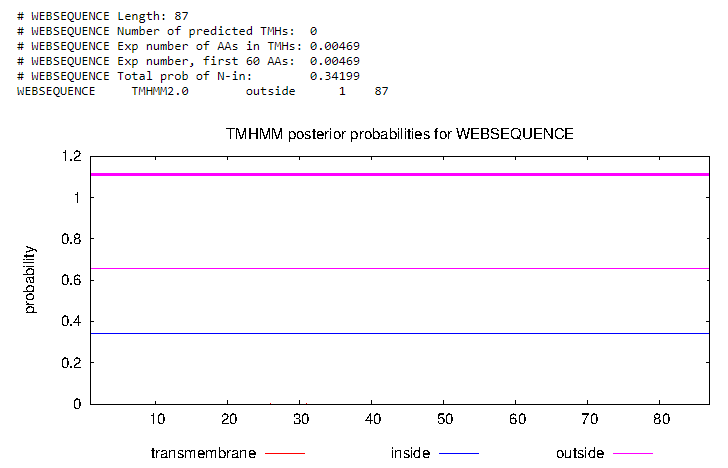
***TomBrady Gene 9 -No protein previously identified***

******

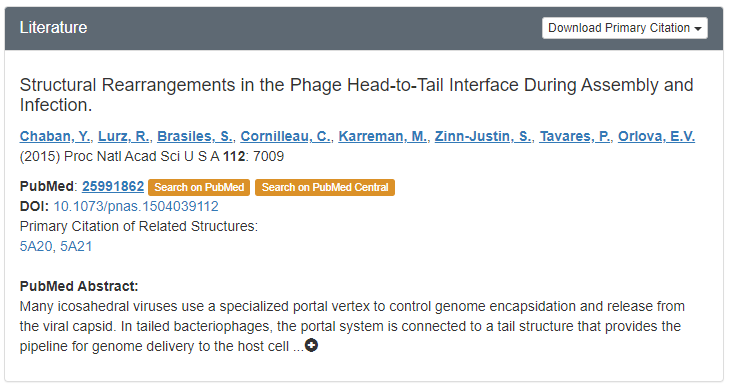
***TomBrady Gene 10 - Head Completion Protein***

******

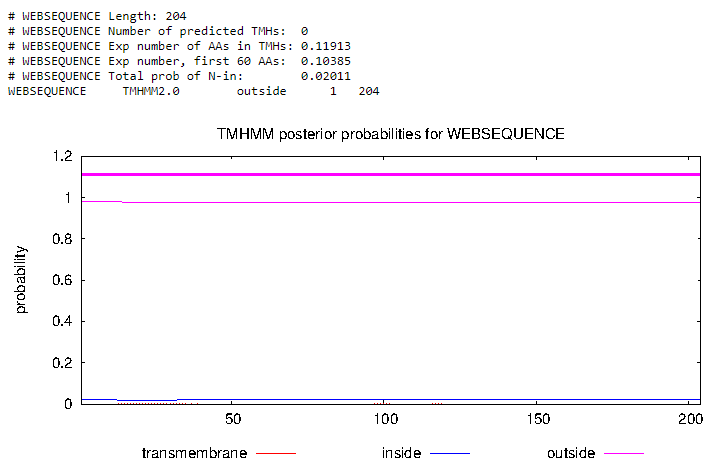
***TomBrady Gene 11 -No protein previously identified***

******

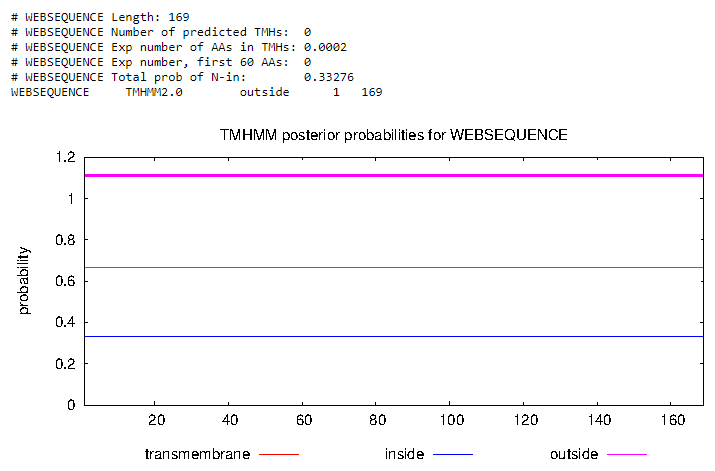
***TomBrady Gene 12 - Head to Tail Protein***

******

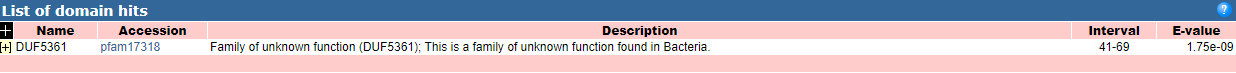
***TomBrady Gene 13- No previous function identified***

******

***TomBrady Gene 14- No previous function identified***

******

***TomBrady Gene 15 - No previous function identified***

******

## 

## Positional annotation 16-10

| **Gene #** | **Start/Stop coordinates** | **CP: Does the prediction cover all the coding potential?** | **SCS: Is this start position predicted by both Genemark and Glimmer?****(Start Choice Sequence)** | **Blast: Best blastp match** | **Gap: Gap/overlap** | **LO: Is this gene the longest?** | **RBS: scores, spacing, for start you chose** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TomBrady #16 | SSC: 11465- 15478 | CP: yes | SCS: yes | Blast:best match is Angel , has 1:1 match | Gap: -1 bp overlap | LORF:Yes | RBS:15 spacer,2.281 z-score, -5.302 final score; not the best |
| TomBrady #17 | SSC: 15478- 16650 | CP: yes | SCS: yes | Blast:best match is Angel (hypothetical protein Angel\_17); 1:1 match | Gap: -1 bp overlap | LORF:No(explain) | RBS:9 spacer, 2.496 z-score, -4.057 final score; best final score |
| TomBrady #18 | SSC: 16650- 18404 | CP: yes | SCS: yes | Blast:best match is Angel (minor tail protein Angel); 1:1 match | Gap: -1 bp overlap | LORF:No | RBS:16 spacer, 1.9 z-score, -6.240 final score; not the best |
| TomBrady #19 | SSC: 18404-18865 | CP: yes | SCS: yes | Blast:best match is Angel (hypothetical protein Angel\_19); 1:1 match | Gap: -1 bp overlap | LORF: No | RBS:14 spacer, 2.917 z-score, -3.807 final score; not the best but it is better than the others |
| TomBrady #20 | SSC: 18862- 19998 | CP: yes | SCS: yes | Blast:best match is BPs (minor tail protein BPs); 1:1 match | Gap: -4 bp overlap | LORF: No | RBS:9 spacer, 2.281 z-score, -4.475 final score; one of the better candidates because the gap is the most ideal and the final score is good |
| TomBrady #21 | SSC: 20009-20443 | CP: yes | SCS: yes | Blast: best match is Angel (hypothetical protein Angel\_21); 1:1 match | Gap: 10 bp | LORF: Yes | RBS:10 spacer, 3.102 z-score, -2.794 final score; best candidate based on final score |
| TomBrady #22 | SSC: 20443-22836 | CP: yes | SCS: yes | Blast: best match is Angel (minor tail protein Angel); 1:1 match | Gap: -1 bp overlap | LORF: Yes | RBS:11 spacer,0.698 z-score. -5.937 final score; not ideal final score |
| TomBrady #23 | SSC: 22851-23171 | CP: yes | SCS: yes | Blast:best match is Angel (hypothetical protein Angel\_23); 1:1 match | Gap: 14 bp | LORF: Yes | RBS:12 spacer, 2.57 z-score, -3.972 final score; best candidate in the list |
| TomBrady #24 | SSC: 23168-23356 | CP: yes | SCS: yes | Blast:best match is Phreak (hypothetical protein SEA\_PHREAK\_24); 1:1 match | Gap: -4 bp overlap | LORF: No | RBS:8 spacer, 2.84 z-score, -3.833 final score; this gene was chosen instead because it has more ideal values than the original gene chosen. |
| TomBrady #25 | SSC: 23356-23631 | CP: yes | SCS: yes | Blast:best match is Chy2 (hypothetical protein Chy2\_0024); 32:32 match | Gap:-1 bp overlap | LORF: No | RBS:12 spacer, 2.03 z-score, -5.025 final score; not the best final score but the rest of the values are ideal. |
| TomBrady #26 | SSC: 23631-23825 | CP: yes | SCS: yes | Blast:best match is Angel (hypothetical protein Angel\_26); 1:1 match | Gap: -1 bp overlap | LORF: Yes | RBS:15 spacer, 2.241 z-score, -5.381 final score; not the most ideal |
| TomBrady #27 | SSC: 23886-25202 | CP: yes | SCS: yes | Blast:best match is Angel (lysin A); 1:1 match | Gap: 60 bp | LORF: No | RBS:13 spacer, 2.341 z-score, -4.629 final score; not the most ideal |
| TomBrady #28 | SSC: 25202-26410 | CP: yes | SCS: yes | Blast:best match is Gideon (lysin B); 1:1 match | Gap: -1 bp overlap | LORF: No | RBS:9 spacer, 2.283 z-score, -4.471 final score; ideal gene candidate |
| TomBrady #29 | SSC: 26436-26786 | CP: yes | SCS: yes | Blast:best match is Angel (hypothetical protein Angel\_29); 1:1 match | Gap: 25 bp | LORF: No | RBS:13 spacer, 2.736 z-score, -3.860 final score; ideal gene candidate |
| TomBrady #30 | SSC: 26791-27111 | CP: yes | SCS: no | Blast:best match is Angel (hypothetical protein Angel\_30); 1:1 match | Gap: 4 bp | LORF: Yes | RBS:12 spacer, 1.89 z-score, -5.299 final score; chose this candidate instead of the automatically chosen one because the gap and the final score of this candidate is much better than the one chosen before. |

## **Tom Brady #16:**

## 

## Above is the top gene candidates for the TomBrady gene #16. The first candidate was the top choice. The gene is also the longest in TomBrady.

## 

## The phage gene that is the most similar to the TomBrady gene #16 is the tape measure protein of Mycobacterium phage Angel. There is a one to one match between the query and the subject.

## 

## The coding potential above shows the start and stop of the TomBrady gene #16

## **Tom Brady #17:**

## 

## The top gene candidate for Tom Brady #17 was the second option. This option had the best final score compared to the other candidates.

## 

## The best phage match to TomBrady #17 was the hypothetical protein Angel, also mycobacterium phage Angel. The identities matched 100% for up to 360 genes. There is a one to one match between the query and the subject.

## 

## This is the coding potential for TomBrady #17. This gene is pretty short compared to TomBrady #16.

## **Tom Brady #18:**

## 

## The top gene candidate for Tom Brady #18 was also the second option. The final score was one of the worse out of all of the candidates, but it had the least amount of gap.

## 

## The best phage match to TomBrady #18 was the minor tail protein Angel, also mycobacterium phage Angel. The identities matched 100% for up to 584 genes. There is a one to one match between the query and the subject.

## 

## Above is the coding potential for TomBrady #18. It was a bit cut off because the gene was stretched out on two pages of the document.

## **TomBrady #19:**

## 

## The start position 18404-18865 was chosen because it was predicted by both tools, has a small overlap (better to include rather than exclude), covers all the coding potential, has many 1:1 blastp matches, has some great RBS scores.

## 

## In the screenshot above, the best phage match to TomBrady #19 was the hypothetical protein Angel, also mycobacterium phage Angel. The identities matched 100% for up to 153 genes. There is a one to one match between the query and the subject.

## 

## Above is the coding potential for TomBrady #19. It is a pretty short sequence compared to others.

## **TomBrady #20**

## 

## The start position 18862-19998 was chosen because it was predicted by both tools, has a small overlap (better to include rather than exclude), covers all the coding potential, has many 1:1 blastp matches, has some great RBS scores.

## 

## In the screenshot above, the best phage match to TomBrady #20 was the minor tail protein BPs, also mycobacterium phage BPs. The identities matched 100% for up to 378 genes. There is a one to one match between the query and the subject.

## 

## 

## Above is the coding potential for TomBrady #20. It is a pretty short sequence compared to others.

## **TomBrady #21:**

## 

## The start position 20009-20443 was chosen because it was predicted by both tools, is the longest ORF, has a small gap, covers all the coding potential, has many 1:1 blastp matches, has the best RBS scores.

## 

## In the screenshot above, the best phage match to TomBrady #21 was the hypothetical protein Angel, also mycobacterium phage Angel. The identities matched 100% for up to 144 genes. There is a one to one match between the query and the subject.

## 

## The coding potential above shows the start and stop of the TomBrady gene #21. It is pretty short compared to other sequences.

## **TomBrady #22:**

## 

## The start position 20443-22836 was chosen because it was predicted by both tools, is the longest ORF, has a slight overlap, covers all the coding potential, has many 1:1 blastp matches, has the best RBS scores.

## 

## In the screenshot above, the best phage match to TomBrady #22 was the minor tail protein Angel, also mycobacterium phage Angel. The identities matched 100% for up to 797 genes. There is a one to one match between the query and the subject.

## 

## The coding potential above shows the start and stop of the TomBrady gene #22.

## **TomBrady #23**

## 

## The start position 22851-23171 was chosen because it was predicted by both tools, is the longest ORF, has a small gap, covers all the coding potential, has many 1:1 blastp matches, has the best RBS scores.

## 

## In the screenshot above, the best phage match to TomBrady #23 was the hypothetical protein Angel, also mycobacterium phage Angel. The identities matched 100% for up to 106 genes. There is a one to one match between the query and the subject.

## 

## The coding potential above shows the start and stop of the TomBrady gene #23. It is pretty short compared to other sequences.

## **TomBrady #24**

## 

## The start position 23168-23356 was chosen because it was predicted by both tools, has a small overlap, covers all the coding potential, has many 1:1 blastp matches, has the best RBS scores.

## 

## In the screenshot above, the best phage match to TomBrady #24 was the hypothetical protein Sea Phreak, also mycobacterium phage Phreak. The identities matched 100% for up to 62 genes. There is a one to one match between the query and the subject.

## 

## The coding potential above shows the start and stop of the TomBrady gene #24. It is pretty short compared to other sequences.

## **TomBrady #25**

## 

## The start position 23356-23631 was chosen because it was predicted by both tools, i, has a small gap, covers all the coding potential, has many 1:1 blastp matches, has the best RBS scores.

## 

## In the screenshot above, the best phage match to TomBrady #25 was the hypothetical protein Chy2\_0024, also mycobacterium phage Chy2. The identities matched 100% for up to 26 genes.

## 

## The coding potential above shows the start and stop of the TomBrady gene #25. It is pretty short compared to other sequences.

## **TomBrady #26**

## 

## The start position 23631-23825 was chosen because it was predicted by both tools, is the longest ORF, has a very small overlap, covers all the coding potential, has many 1:1 blastp matches, has the best RBS scores.

## 

## In the screenshot above, the best phage match to TomBrady #26 was the hypothetical protein Angel, also mycobacterium phage Angel. The identities matched 100% for up to 64 genes. There is a one to one match between the query and the subject.

## 

## The coding potential above shows the start and stop of the TomBrady gene #26. It is very short compared to other sequences.

## **TomBrady #27**

## 

## The start position 23886-25202 was chosen because it was predicted by both tools, has a small gap, covers all the coding potential, has many 1:1 blastp matches, has the best RBS scores.

## 

## In the screenshot above, the best phage match to TomBrady #27 was the lysin A in Angel, also mycobacterium phage Angel. The identities matched 100% for up to 438 genes. There is a one to one match between the query and the subject.

## 

## The coding potential above shows the start and stop of the TomBrady gene #27. It is longer compared to other sequences.

## **TomBrady #28**

## 

## The start position 25202-26410 was chosen because it was predicted by both tools, has a very small overlap, covers all the coding potential, has many 1:1 blastp matches, has the best RBS scores.

## 

## In the screenshot above, the best phage match to TomBrady #28 was the lysin B in Angel, also mycobacterium phage Gideon. The identities matched 99% for up to 401 genes. There is a one to one match between the query and the subject.

## 

## The coding potential above shows the start and stop of the TomBrady gene #28. It is pretty long compared to other sequences.

## **TomBrady #29**

## 

## The start position 26436-26786 was chosen because it was predicted by both tools, has a small gap, covers all the coding potential, has many 1:1 blastp matches, has the best RBS scores.

## 

## In the screenshot above, the best phage match to TomBrady #29 was the hypothetical protein Angel, also mycobacterium phage Angel. The identities matched 100% for up to 116 genes. There is a one to one match between the query and the subject.

## 

## The coding potential above shows the start and stop of the TomBrady gene #29. It is pretty short compared to other sequences.

## **TomBrady #30**

## 

## The start position 26791-27111 was chosen because it was predicted by both tools, is the longest ORF, has a small gap, covers all the coding potential, has many 1:1 blastp matches, has the best RBS scores.

## 

## In the screenshot above, the best phage match to TomBrady #30 was the hypothetical protein Angel, also mycobacterium phage Angel. The identities matched 100% for up to 106 genes. There is a one to one match between the query and the subject.

## 

## 

## The coding potential above shows the start and stop of the TomBrady gene #30. It is pretty short compared to other sequences.

## 

## Functional annotation 16-30

## **TomBrady #16**

## BLAST: Insufficient blastp information

## HHpred:

## 

## 

## 

## TMHMM: 4 predicted TMHs

## 

## **TomBrady #17**

## BLAST: Insufficient blastp information

## HHpred:

## 

## 

## 

## **TomBrady #18**

## BLAST: Insufficient blastp information

## HHpred:

## 

## (No PubMed publications, thus this sequence has no functional prediction)

## TMHMM:

## 

## **TomBrady #19**

## BLAST: Insufficient blastp information

## HHpred: Insufficient information

## TMHMM:

## 

## **TomBrady #20**

## BLAST: Insufficient blastp information

## HHpred: Insufficient information

## TMHMM:

## 

## **TomBrady #21**

## BLAST: Insufficient blastp information

## HHpred: Insufficient information

## TMHMM:

## 

## **TomBrady #22**

## BLAST: Insufficient blastp information

## HHpred: Insufficient information

## TMHMM:

## 

## **TomBrady #23**

## BLAST: Insufficient blastp information

## HHpred: : Insufficient information

## TMHMM: No predicted TMHs

## **TomBrady #24**

## BLAST: Insufficient blastp information

## HHpred: : Insufficient information

## TMHMM: No predicted TMHs

## **TomBrady #25**

## BLAST: Insufficient blastp information

## HHpred: Insufficient information (E-value too large)

## TMHMM: No predicted TMHs

## **TomBrady #26**

## BLAST: Insufficient blastp information

## HHpred: Insufficient information (E-value too large)

## TMHMM: No predicted TMHs

## **TomBrady #27**

## BLAST: Insufficient blastp information

## HHpred:

## 

## 

## TMHMM: No predicted TMHs

## **TomBrady #28**

## BLAST: Insufficient blastp information

## HHpred:

## 

## 

## 

## TMHMM: No predicted TMHs

## **TomBrady #29**

## BLAST: Insufficient blastp information

## HHpred: Insufficient information

## TMHMM: 2 predicted TMHs

## 

## **TomBrady #30**

## BLAST: Insufficient blastp information

## HHpred: Insufficient information

## TMHMM: 2 predicted TMHs

## 

## 

## 31-45 Positional Annotation

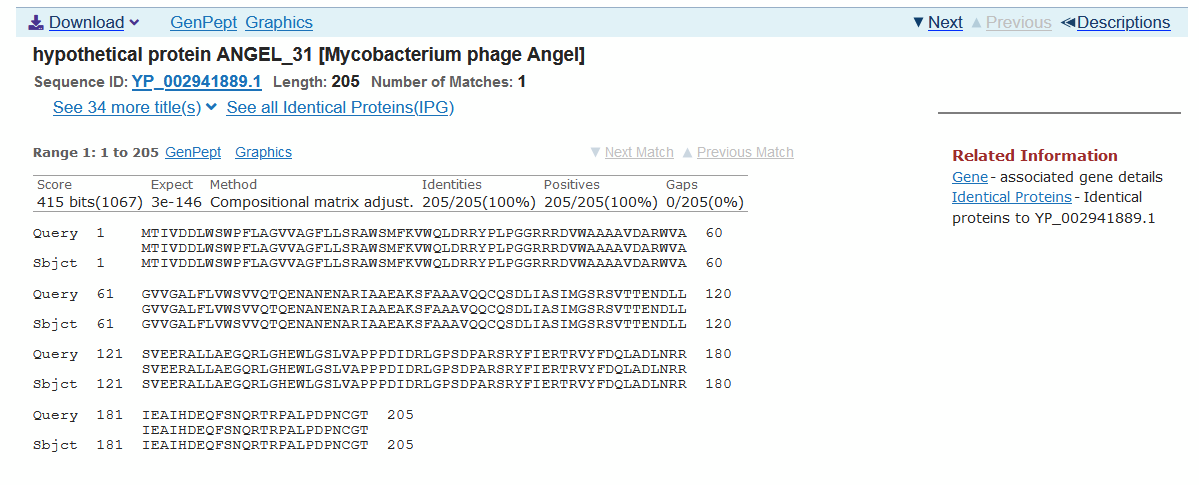
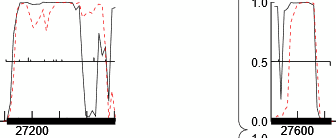
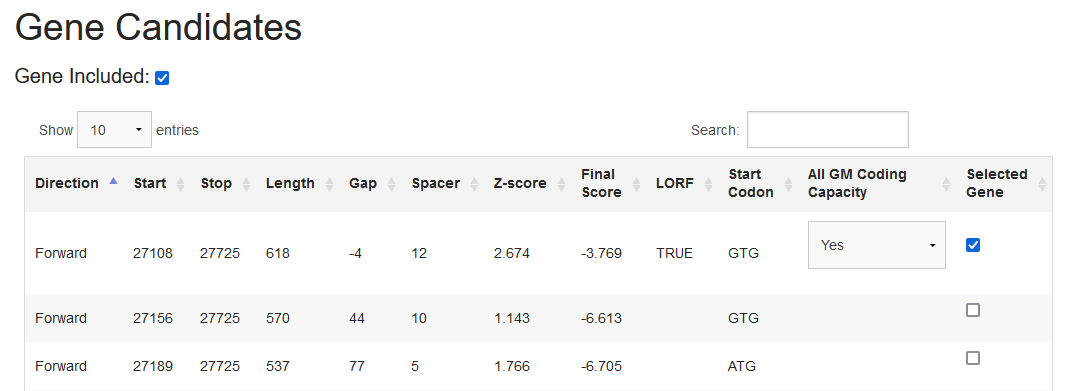
Goal: Annotation of assigned section of TomBrady genome (PECAAN genes #31-45)

| **Gene #** | **SSC: Start/Stop Coordinates** | **CP: Does the prediction cover all coding potential?** | **SCS: Start choice source: Predicted by both GeneMark and Glimmer?** | **Blast: best Blastp match** | **Gap/Overlap** | **LO: longest ORF: Is this gene the longest?** | **RBS: spacing matrix, Z score, final score, is it the best?** | **Date** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [TomBrady #31](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#31) | SSC: 27108-27725 | CP: yes | SCS: both | Blast: best match is Angel gene #31; 1:1 match | Gap: 4 bp overlap | LORF: Yes | 12 spacer, 2.674 Z-score, -3.769 final score; yes | 2/9/2022 |
| [TomBrady #32](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#32) | SSC: 28918-27722 | CP: yes | SCS: neither | Blast: best match is Angel integrase; 1:1 match | Gap: 4 bp overlap | LORF: Yes | 12 spacer, 2.425 Z-score, -4.255 final score; yes | 2/9/2022 |
| [TomBrady #33](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#33) | SSC: 29490-28915 | CP: yes | SCS: Glimmer only | Blast: best match is Phish immunity repressor; 1:1 match | Gap: 8 bp gap | LORF: Yes | 10 spacer, 0.486 Z-score, -7.894 final score; no | 2/9/2022 |
| [TomBrady #34](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#34) | SSC: 29499-29756 | CP: yes | SCS: both | Blast: best match is Angel RDF protein; 1:1 match | Gap: 8 bp gap | LORF: Yes | 14 spacer, 1.623 Z-score, -6.330 final score; yes | 2/9/2022 |
| [TomBrady #35](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#35) | SSC: 29753-30151 | CP: yes | SCS: both | Blast: best match is Leo gene #34; 1:1 match\* | Gap: 4 bp overlap | LORF: Yes | 15 spacer, 1.733 Z-score, -6.371 final score; yes | 2/16/2022 |
| [TomBrady #36](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#36) | SSC: 30214-30750 | CP: yes | SCS: Glimmer only | Blast: best match is Leo gene #35; 1:1 match | Gap: 62 bp gap | LORF: No | 12 spacer, 3.179 Z-score, -2.786 final score; yes | 2/16/2022 |
| [TomBrady #37](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#37) | SSC: 30750-30932 | CP: yes | SCS: both | Blast: best match is Leo gene #36; 1:1 match | Gap: 1 bp overlap | LORF: Yes | 14 spacer, 2.092 Z-score, -5.417 final score; yes | 2/16/2022 |
| [TomBrady #38](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#38) | SSC: 30929-31279 | CP: yes | SCS: both | Blast: best match is Liefie gene #38; 1:1 match | Gap: 4 bp overlap | LORF: No | 13 spacer, 2.567 Z-score, -4.188 final score; yes | 2/16/2022 |
| [TomBrady #39](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#39)† | SSC: 31276-31425 | CP: yes | SCS: both | Blast: best match is Angel gene #39; 1:1 match\* | Gap: 4 bp overlap | LORF: Yes | 10 spacer, 3.276 Z-score, -2.456 final score; yes | 2/16/2022 |
| [TomBrady #40](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#40) | SSC: 31425-31775 | CP: yes | SCS: GeneMark only | Blast: best match is Angel gene #40; 1:1 match | Gap: 1 bp overlap | LORF: No | 10 spacer, 2.270 Z-score, -4.417 final score | 2/16/2022 |
| [TomBrady #41](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#41) | SSC: 31759-31911 | CP: yes | SCS: both | Blast: best match is Angel gene #41; 1:1 match | Gap: 17 bp overlap | LORF: No | 8 spacer, 3.452 Z-score, -2.640 final score | 2/21/2022 |
| [TomBrady #42](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#42) | SSC: 31908-32993 | CP: yes | SCS: both | Blast: best match is Angel RecE; 1:1 match | Gap: 4 bp overlap | LORF: Yes | 11 spacer, 1.981 Z-score, -5.043 final score | 2/21/2022 |
| [TomBrady #43](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#43) | SSC: 33011-34399 | CP: yes | SCS: Glimmer only | Blast: best match is Angel RecT; 1:1 match | Gap: 17 bp gap | LORF: No | 17 spacer, 3.276 Z-score, -3.761 final score | 2/21/2022 |
| [TomBrady #44](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#44)\* | SSC: 34396-34950 | CP: yes | SCS: Glimmer only | Blast: best match is Gomashi gene #44; 1:1 match | Gap: 4 bp overlap | LORF: Yes | 8 spacer, 2.736 Z-score, -4.036 final score | 2/21/2022 |
| [TomBrady #45](https://wiki.vcu.edu/display/phagelab/Cassidy+Coates+Spring+2022#CassidyCoatesSpring2022-Gene#45)\* | SSC: 35090-35521 | CP: yes | SCS: Glimmer only | Blast: best match is Gomashi gene #45; 1:1 match | Gap: 139 bp gap | LORF: Yes | 9 spacer, 3.276 Z-score, -2.536 final score | 2/21/2022 |

\*See relevant discussion on this gene below.

### **Gene #31**

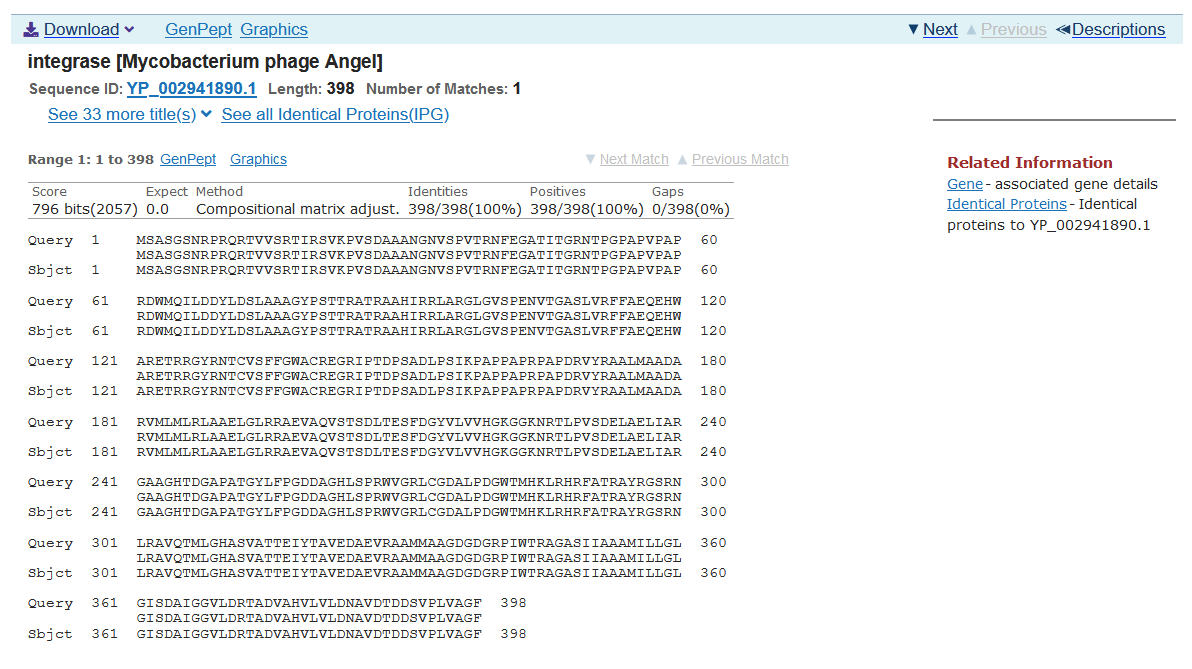
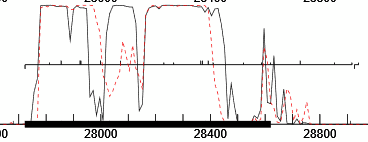
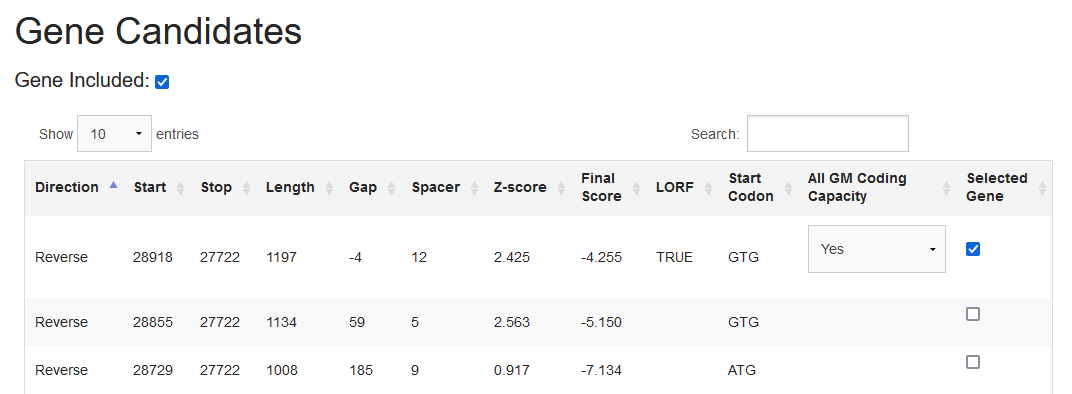
SSC: 27108-27725



Chosen start contains the longest ORF and has the best Z-score and final score. Chosen start covers all coding potential with 4 bp overlap. BLASTp indicates a 1:1 match with Angel gene #31 (100% query coverage) and 10 others.

### **Gene #32**

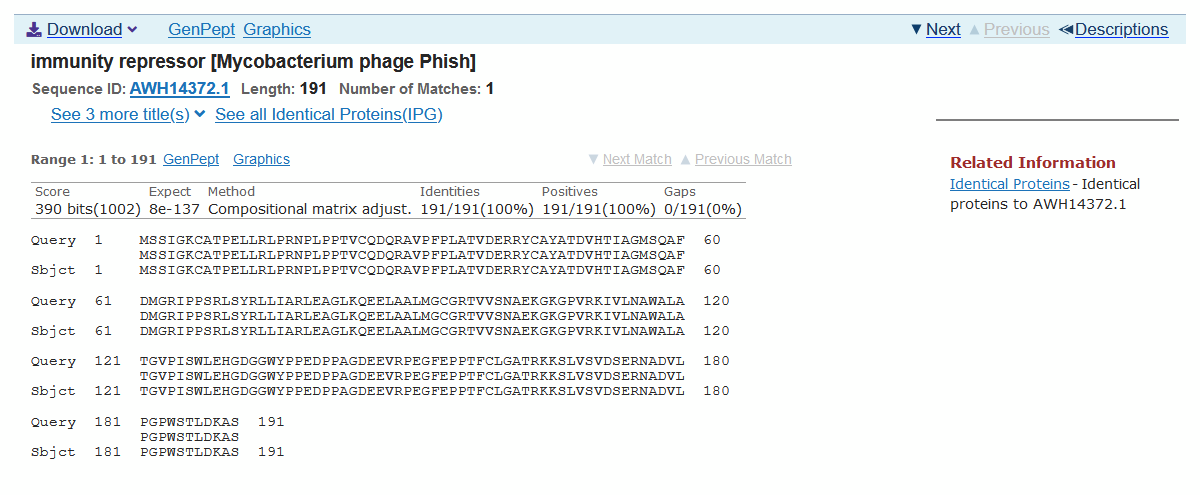
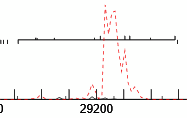
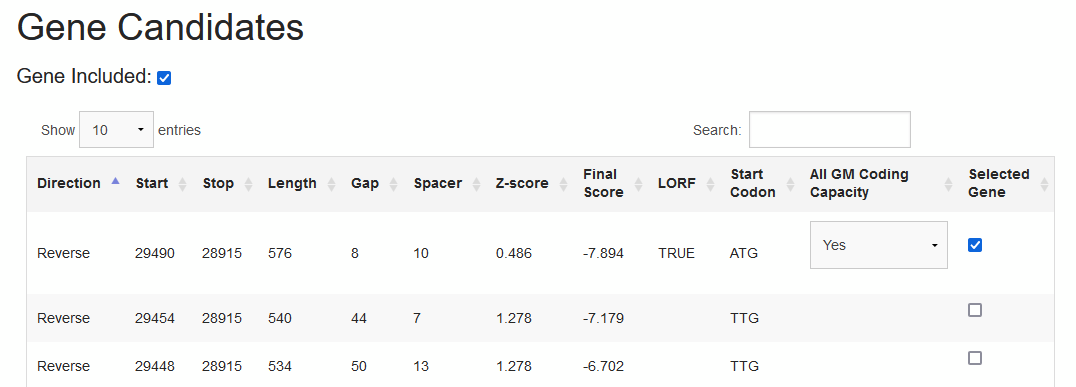
SSC: 28918-27722



Chosen start is not recognized by either GeneMark or Glimmer, but contains the longest ORF and has the best Z-score and final score. Chosen start covers all coding potential with 4 bp overlap. BLASTp indicates a 1:1 match with Angel integrase (100% query coverage) and 3 others.

### **Gene #33**

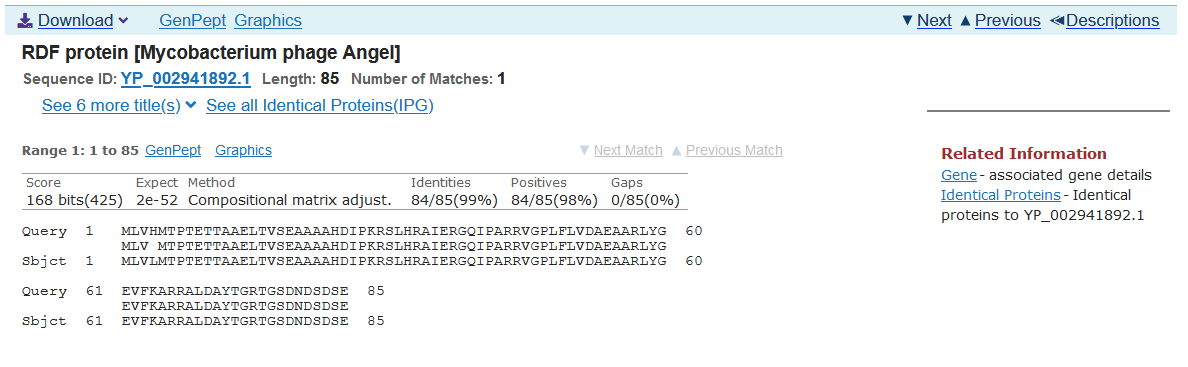
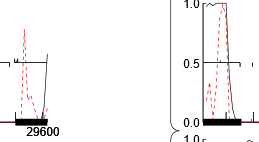
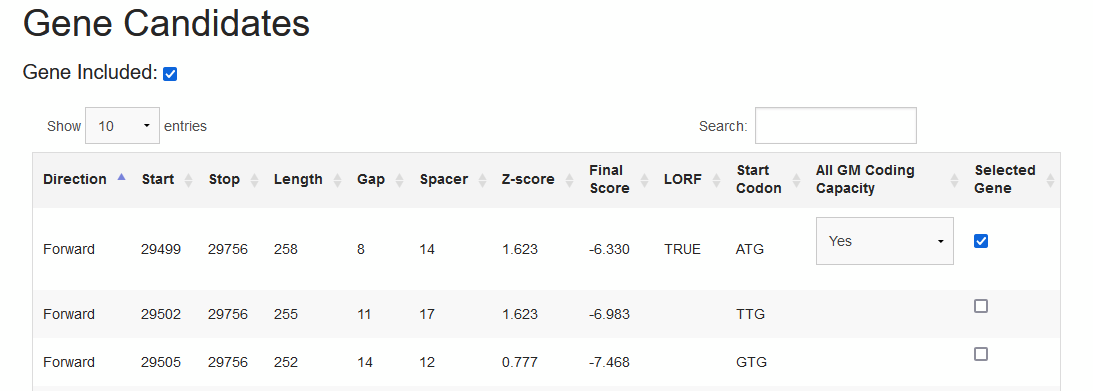
SSC: 29490-28915



Chosen start contains the longest ORF and minimizes gap, but does not have the best Z-score or final score. Chosen start covers all coding potential, although GeneMark does not recognize much coding potential in the ORF. BLASTp indicates a 1:1 match with Phish immunity repressor (100% query coverage) and 1 other.

### **Gene #34**

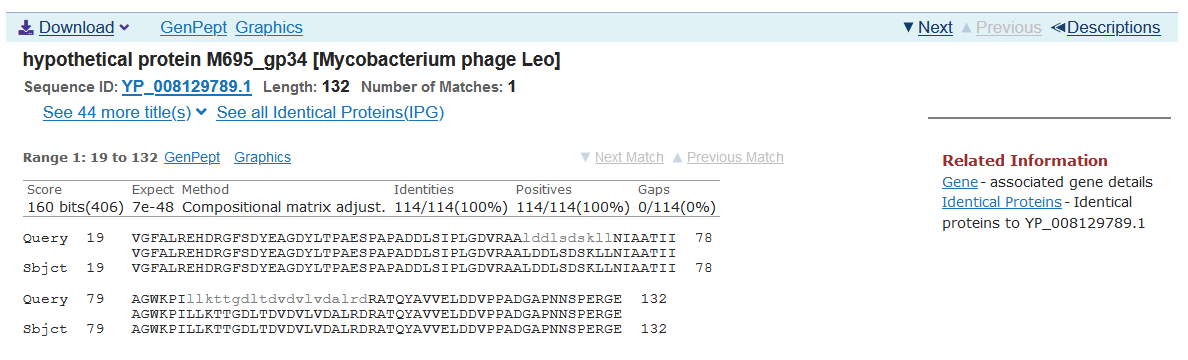
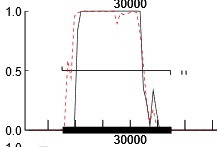
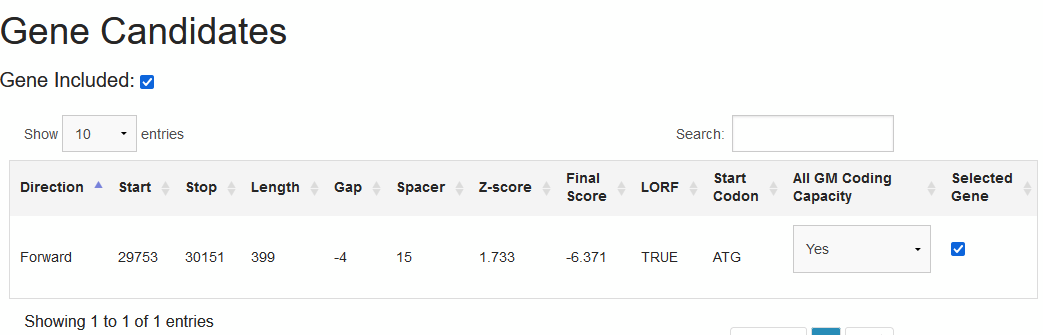
SSC: 29499-29756



Chosen start contains the longest ORF and has the best Z-score and final score. Chosen start covers all coding potential. BLASTp indicates a 1:1 match with Angel RDF protein (100% query coverage) and 3 others.

### **Gene #35**

SSC: 29753-30151

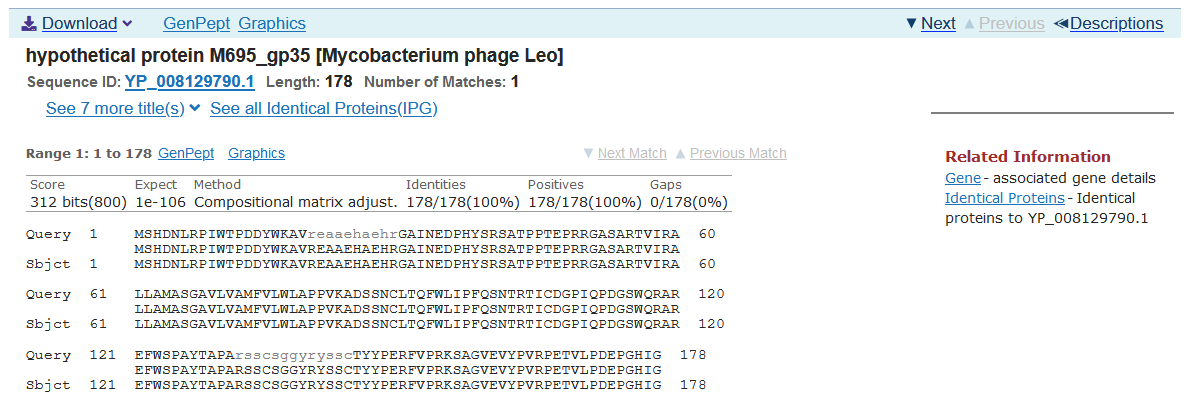
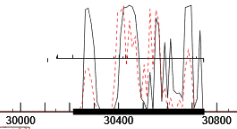
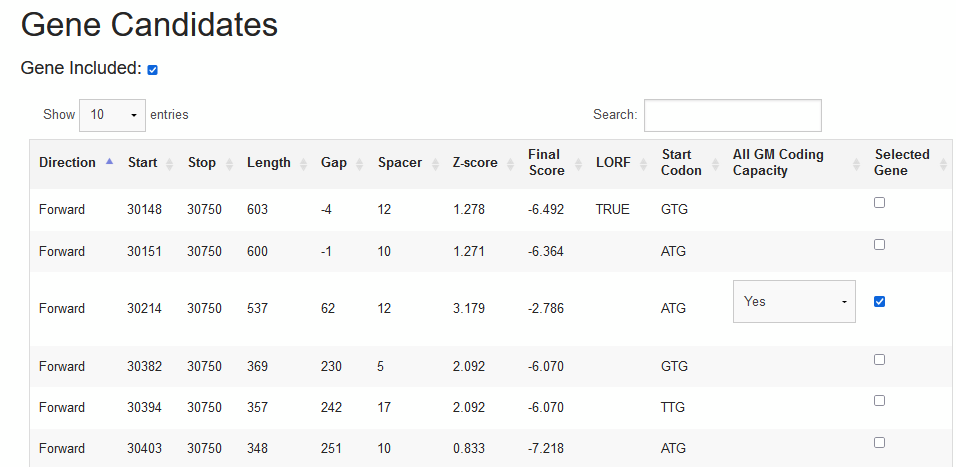


Chosen start is the only candidate start codon recognized by both Glimmer and GeneMark. It therefore contains the longest ORF and has the best Z-score and final score. Chosen start covers all coding potential.

\*Web BLASTp indicates the best match is with Leo gene #34 (86% query coverage, 100% identity), starting at bp 19; however, BLASTp run through PECAAN indicates a 1:1 match with Leo gene #34 and 8 others. The reason for this difference in results is unclear.

### **Gene #36**

SSC: 30214-30750

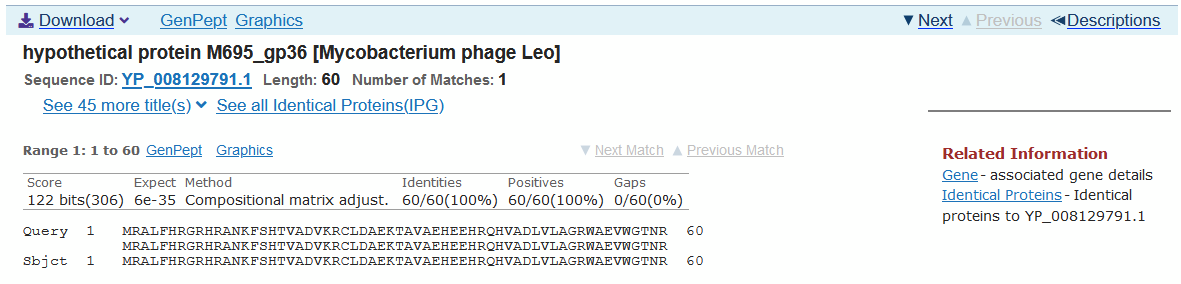
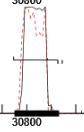
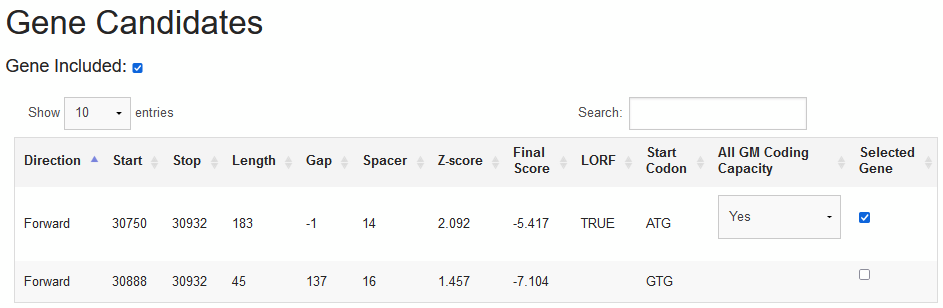


Chosen start is recognized by Glimmer only (PECAAN GeneMark prefers start = 30403, which seems unlikely given the vastly shortened ORF, and also does not match the graphical indication on the GeneMark printout). Chosen start does not contain the longest ORF and has a substantial gap, but it has a substantially better Z-score and final score than the two longer ORFs. Chosen start covers all coding potential. BLASTp indicates a 1:1 match with Leo gene #35 (100% query coverage).

Start = 30148 (the longest ORF) is arguable, but has a much worse Z-score and final score; BLASTp gives a 1:1 match with Remy19 gene #36 (100% query coverage), suggesting past analysis of similar phages is inconclusive, but chosen start still covers all coding potential and has significantly better RBS score.

### **Gene #37**

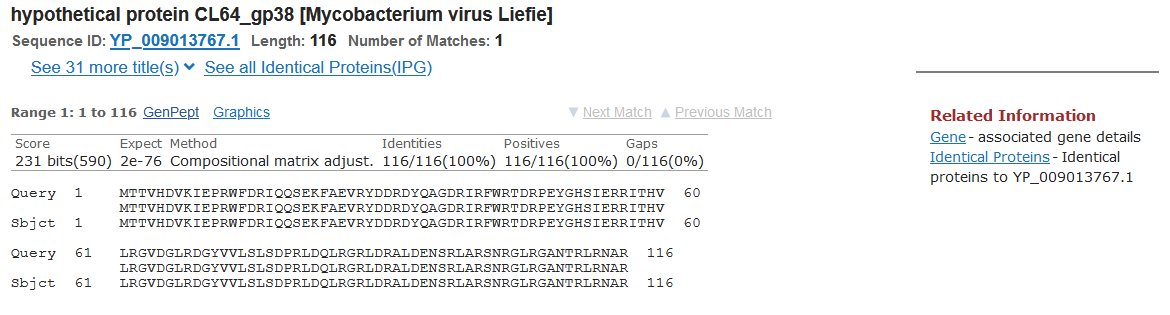
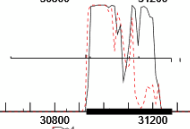
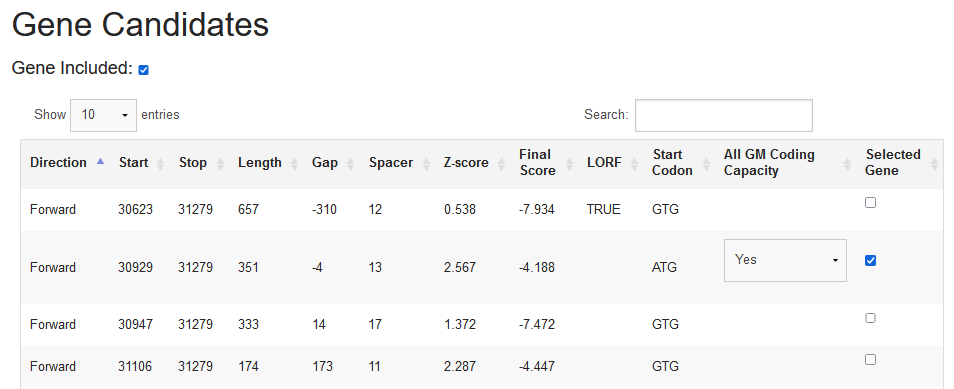
SSC: 30750-30932



Chosen start contains the longest ORF and has the best Z-score and final score. Chosen start covers all coding potential. BLASTp indicates a 1:1 match with Leo gene #36 (100% query coverage).

### **Gene #38**

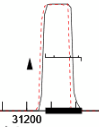
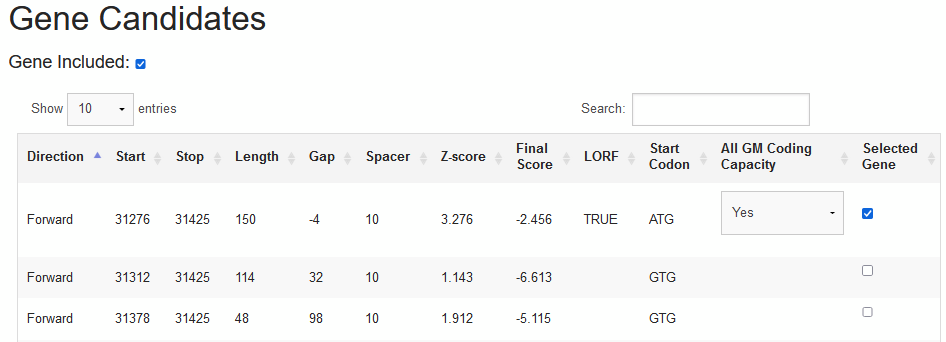
SSC: 30929-31279



Chosen start does not contain the longest ORF, but substantially minimizes overlap and has the best Z-score and final score. Chosen start covers all coding potential. BLASTp indicates a 1:1 match with Liefie gene #38 (100% query coverage).

### **Gene #39**

SSC: 31276-31425



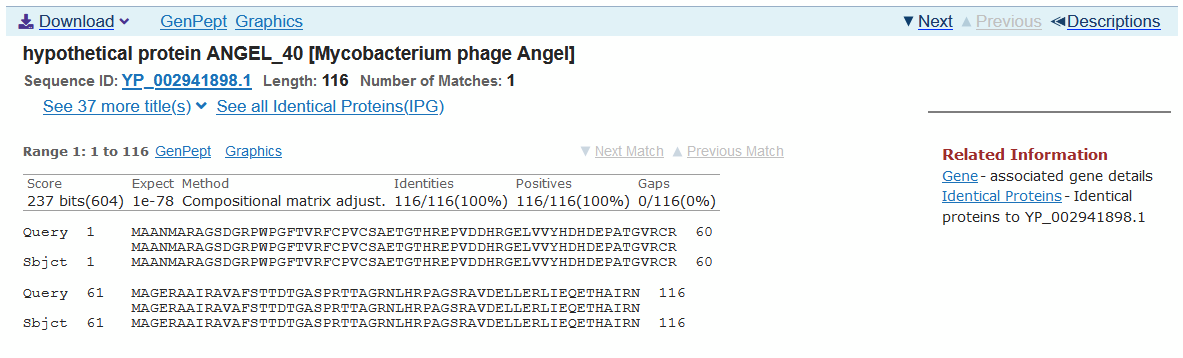
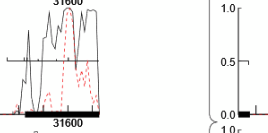
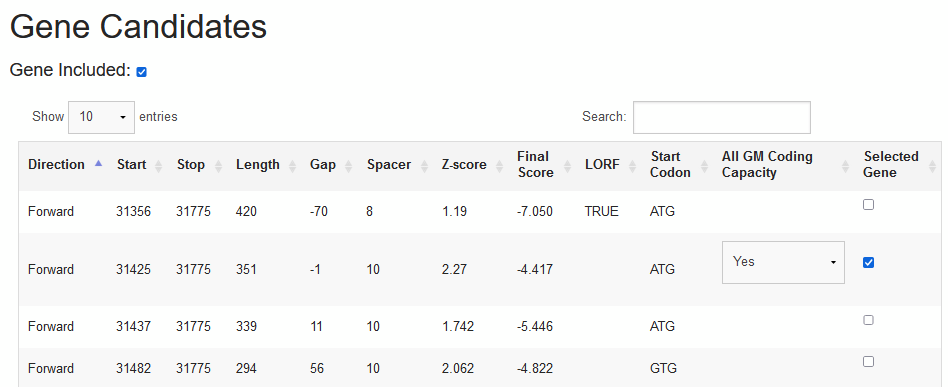
Chosen start contains the longest ORF and has the best Z-score and final score. Chosen start covers all coding potential.

\*Web BLASTp is unable to identify significant similarity and suggests the protein is too short; however, BLASTp run through PECAAN indicates a 1:1 match with Angel gene #39 (100% query coverage) and 9 others.

†GeneMark suggests that there may be a possible frame shift present.

### **Gene #40**

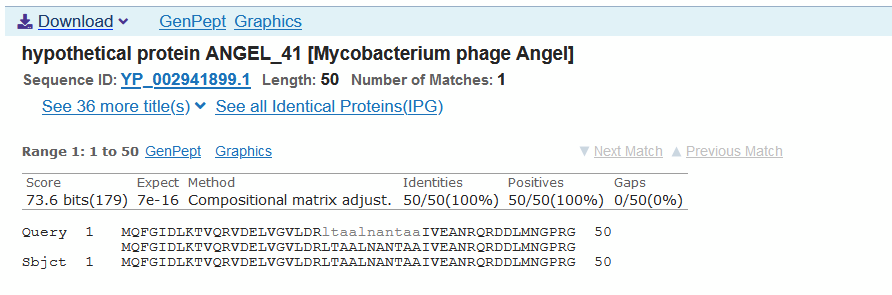
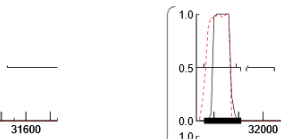
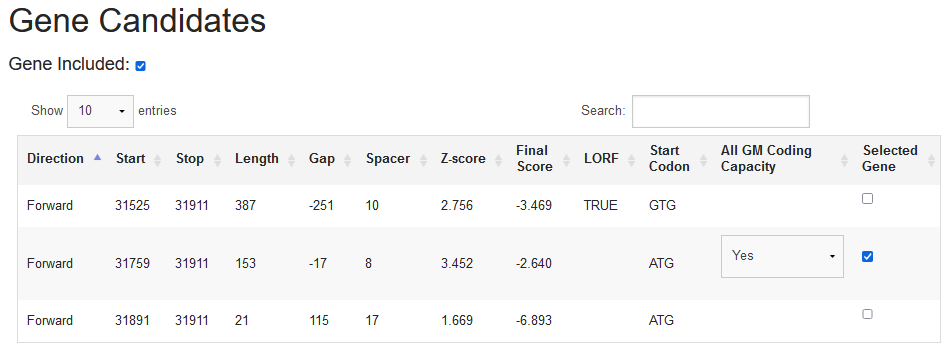
SSC: 31425-31775



Chosen start is predicted by GeneMark only (Glimmer prefers start = 31437). Chosen start does not contain the longest ORF (LORF has a 70 bp overlap), but has the best Z-score and final score. Furthermore, start = 31437 seems unlikely as it is only 12 bp downstream from the chosen start, which significantly restricts the possible space for the gene's ribosome binding site. Chosen start covers all coding potential. BLASTp indicates a 1:1 match with Angel gene #40 (100% query coverage).

### **Gene #41**

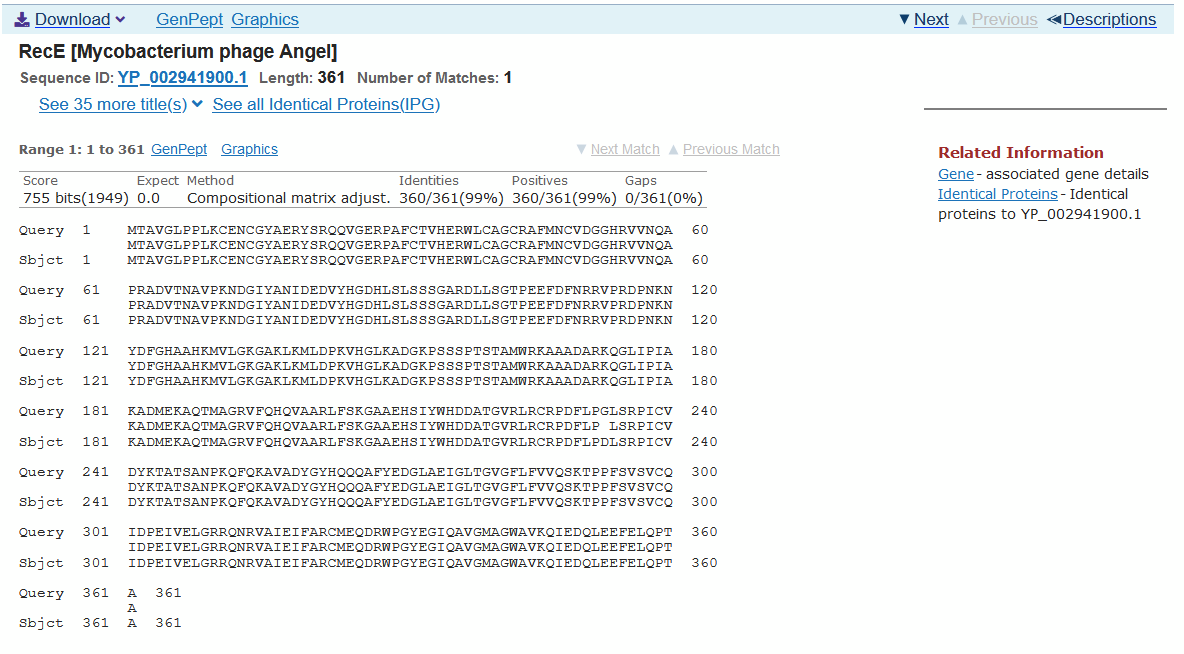
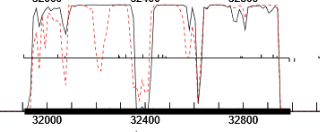
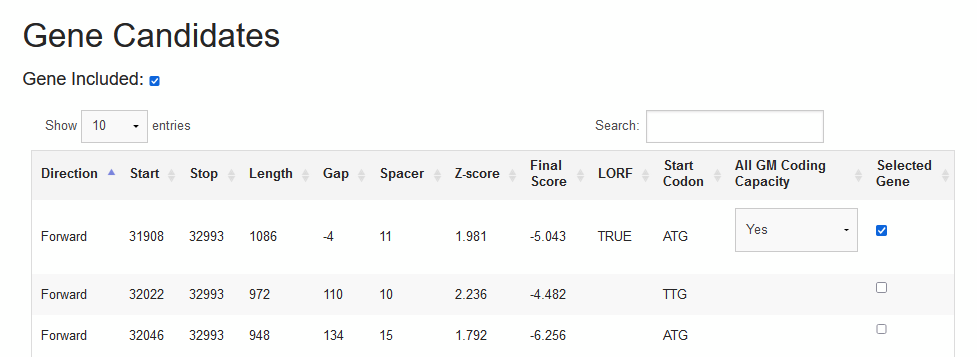
SSC: 31759-31911



Chosen start does not contain the longest ORF, but has the best Z-score and final score. Chosen start covers all coding potential with a large but acceptable 17 bp overlap. BLASTp indicates a 1:1 match with Angel gene #41 (100% query coverage).

### **Gene #42**

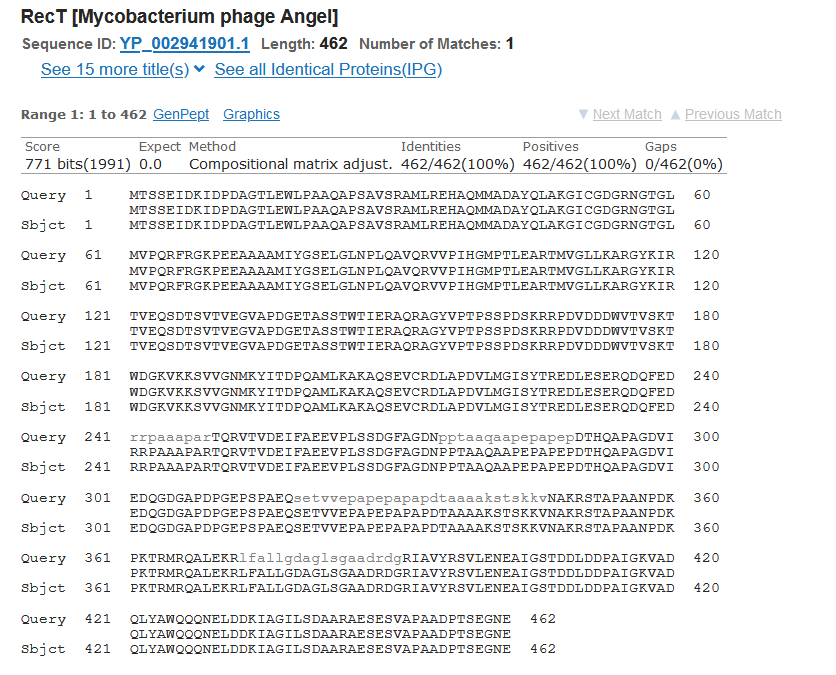
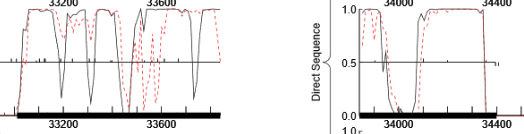
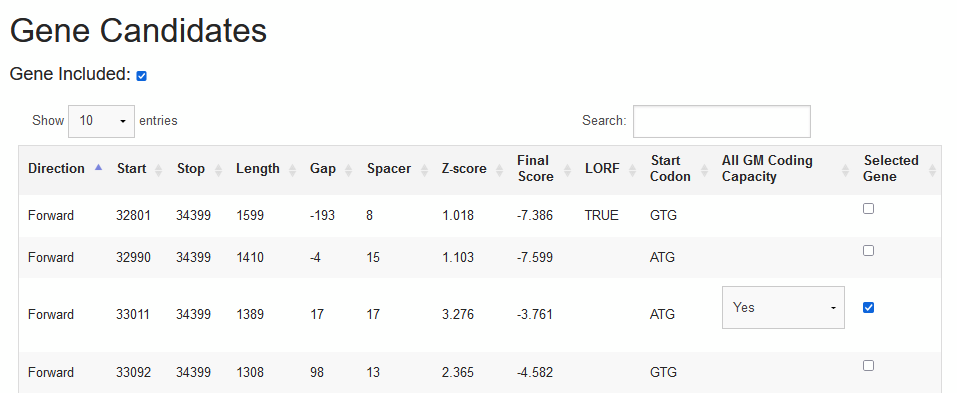
SSC: 31908-32993



Chosen start does not have especially good RBS scores, but contains the longest ORF by over 100 bp. Chosen start covers all coding potential. BLASTp indicates a 1:1 match with Angel RecE (100% query coverage) and 6 others, though even the best match (Angel) differs by one amino acid.

### **Gene #43**

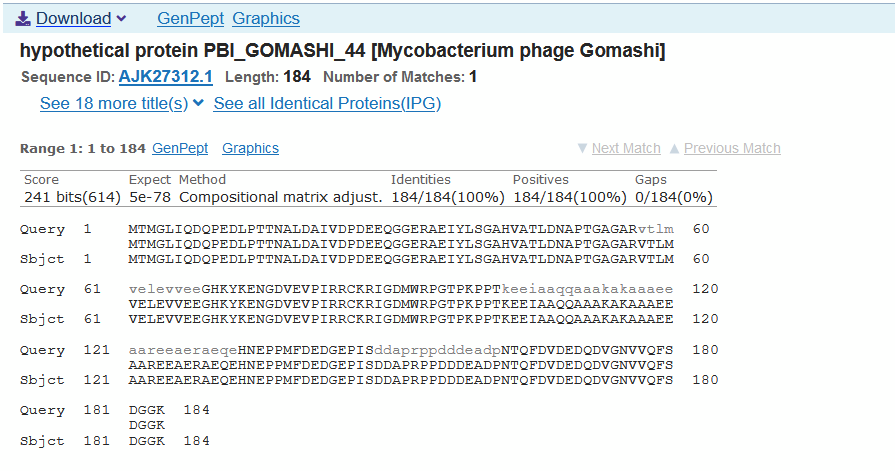
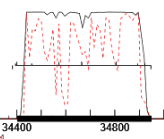
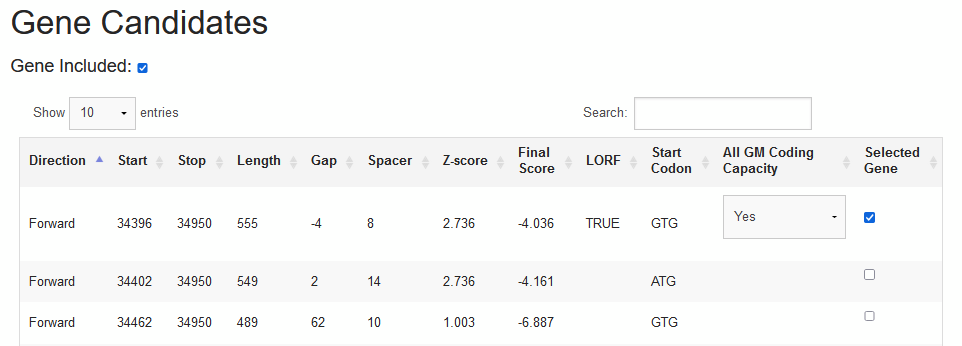
SSC: 33011-34399



Chosen start is not the LORF, but has the best Z-score and final score. Chosen start covers all coding potential. BLASTp indicates a 1:1 match with Angel RecT (100% query coverage, 100% identity).

### **Gene #44**

SSC: 34396-34950

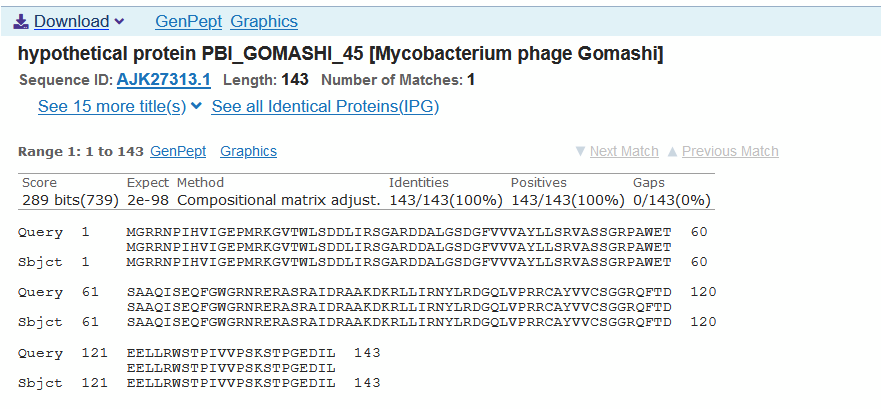
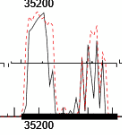
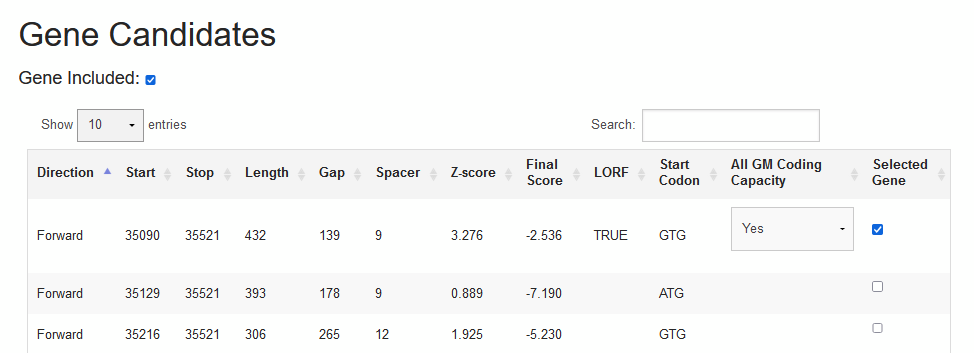


Chosen start contains the longest ORF and has the best Z-score and final score. Chosen start covers all coding potential.

\*Chosen start, which is predicted by Glimmer, is nearly identical to start = 34402, which is predicted by GeneMark - differing by only 6 bp in length. Chosen start was selected to maximize the ORF and for its slightly shorter spacer and marginally better final score. BLASTp suggests previous analysis of similar phages is inconclusive, as it indicates 1:1 matches for both start positions; chosen start has a 1:1 match with Gomashi gene #44 (100% query coverage) and 1 other, while start = 34402 has a 1:1 match with Angel gene #44.

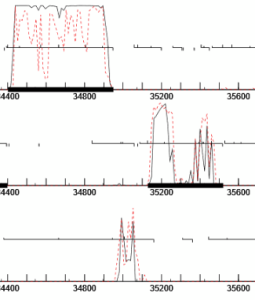
### **Gene #45**

SSC: 35090-35521



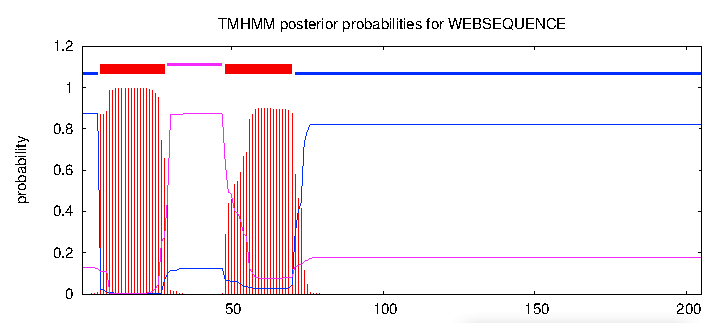
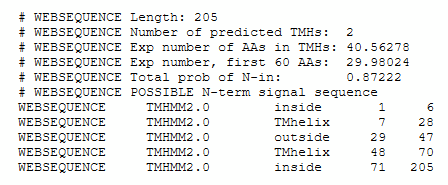
Chosen start contains the longest ORF and has the best Z-score and final score. Chosen start covers all coding potential. BLASTp indicates a 1:1 match with Gomashi gene #45 (100% query coverage).

\*This gene's fairly long gap (139 bp) is suggestive of the possibility of the presence of a gene not predicted by Glimmer or GeneMark. The relevant section of GeneMark data is shown below; as can be seen, there is a possible ORF between genes #44 and #45, which contains some coding potential. Further analysis is necessary to determine if this is actually a gene.



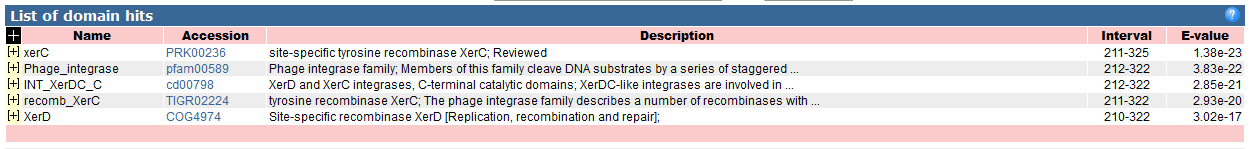
### **Gene #31: membrane protein**

BLASTp detected no putative conserved domains. No HHPred result with E-value < 10-7 was found. TMHMM prediction as follows:



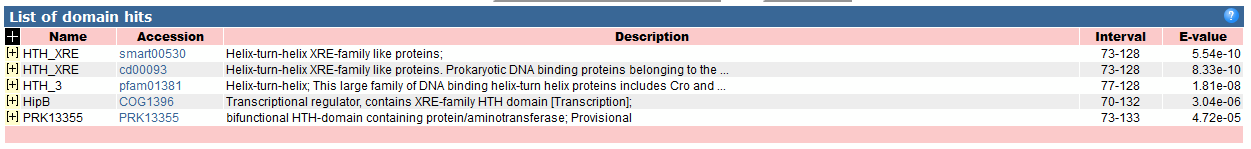
### **Gene #32: tyrosine integrase**

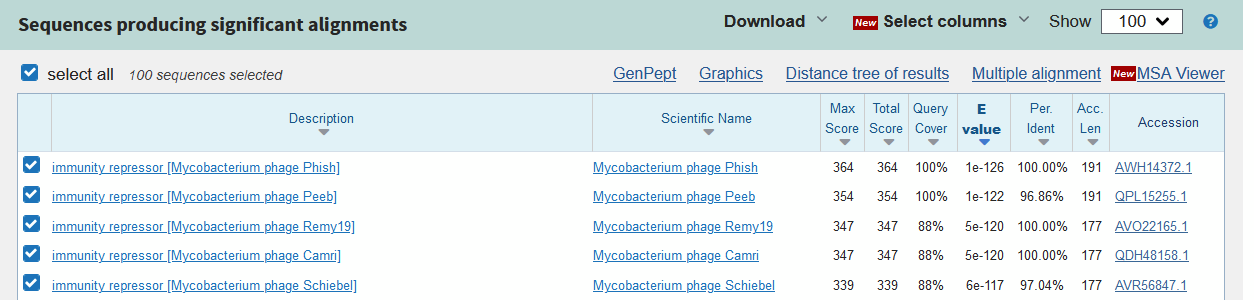
BLASTp indicated putative conserved domains:



### **Gene #33: immunity repressor**

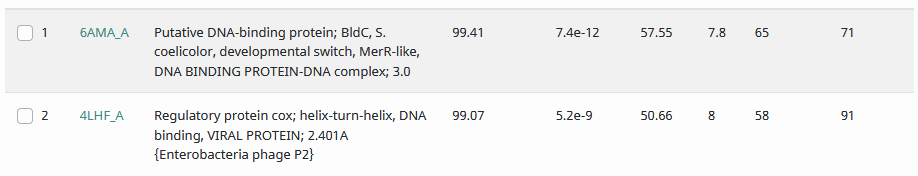
BLASTp indicated putative conserved domains:





### **Gene #34: helix-turn-helix DNA binding domain**

BLASTp detected no putative conserved domains. Several published HHPred results with E-value < 10-7 were found:

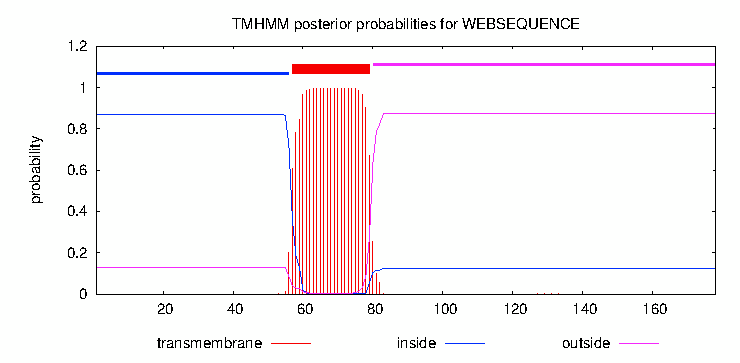
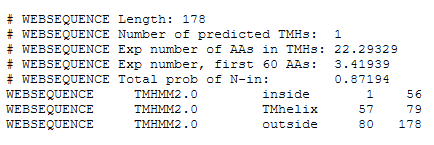


### **Gene #35: hypothetical protein**

BLASTp detected no putative conserved domains. No HHPred result with E-value < 10-7 was found. No TMHs were predicted by TMHMM.

### **Gene #36: membrane protein**

BLASTp detected no putative conserved domains. No HHPred result with E-value < 10-7 was found. TMHMM prediction as follows:

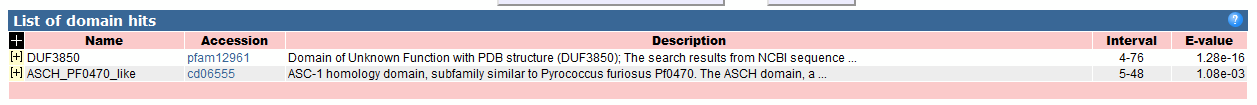


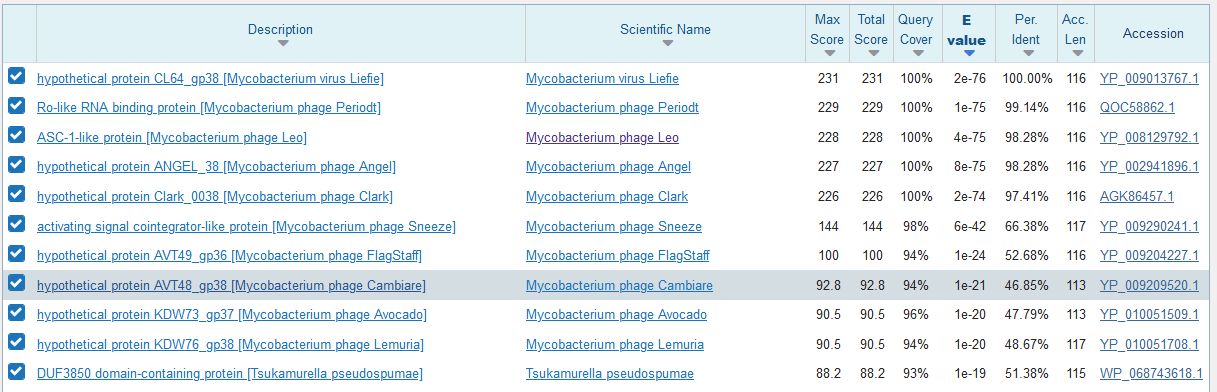
### **Gene #37: hypothetical protein**

BLASTp detected no putative conserved domains. No HHPred result with E-value < 10-7 was found. No TMHs were predicted by TMHMM.

### **Gene #38: Ro-like RNA binding protein**

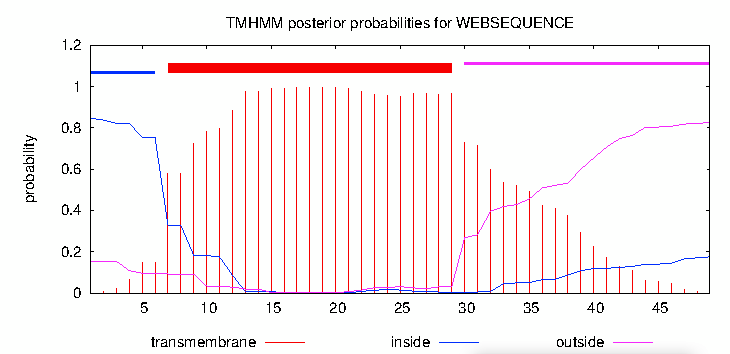
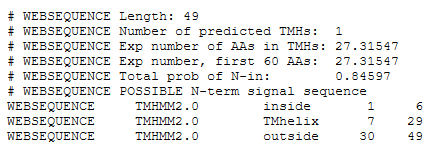
BLASTp indicated putative conserved domains:





### **\*Gene #39: membrane protein**

BLASTp detected no putative conserved domains and indicated the query was too short. No HHPred result with E-value < 10-7 was found. TMHMM prediction as follows:



### **Gene #40: hypothetical protein**

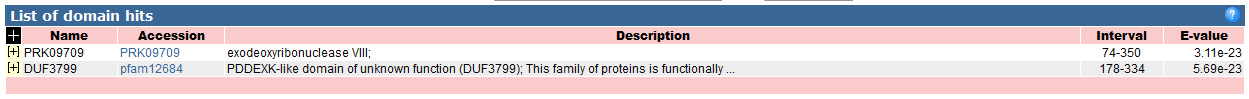
BLASTp detected no putative conserved domains. No HHPred result with E-value < 10-7 was found. No TMHs were predicted by TMHMM.

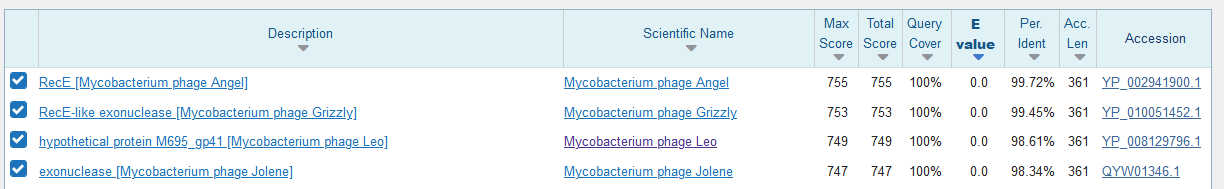
### **Gene #41: hypothetical protein**

BLASTp detected no putative conserved domains. No HHPred result with E-value < 10-7 was found. No TMHs were predicted by TMHMM.

### **Gene #42: RecE-like exonuclease**

BLASTp indicated putative conserved domains:





### **Gene #43: hypothetical protein**

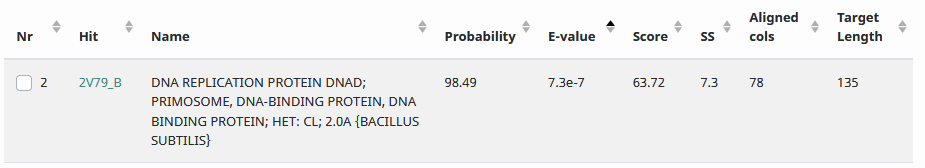
BLASTp detected no putative conserved domains. No HHPred result with E-value < 10-7 was found. No TMHs were predicted by TMHMM.

### **Gene #44: hypothetical protein**

BLASTp detected no putative conserved domains. No HHPred result with E-value < 10-7 was found. No TMHs were predicted by TMHMM.

### **Gene #45: DNA binding protein**

BLASTp detected no putative conserved domains. One published HHPred result was found; E-value is 7.3x10-7, very close to the threshold.



Positional and Functional Annotation 44-62

| **Gene #** | **SSC: Start/Stop coordinates** | **CP: Does the prediction cover all the coding potential?** | **SCS: Is this start position predicted by both Genemark and Glimmer?** | **Blast: Best blastp match** | **Gap: Gap/overlap** | **LO: Is this gene the longest?** | **RBS: scores, spacing, for start you chose** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Tom Brady **#44.5** inserted gene 1 | SSC: 34947-35165 | CP: Yes | Neither  (see notes) | Blast: best match is ANGEL gene 45  1:1 match  28 others | Gap: 4 bp overlap | LORF: No, but minimizes overlap | RBS:  Spacer: 8  Z-Score: 1.579  Final Score: -6.145 |
| #44.5 Notes: This was a manual entry using DNA master to assist with start and end coordinates and protein sequence. Thank you, Dr. Johnson! There were 28 1:1 matches.  Protein sequence from Glimmer: **MSTADAGHSAHWISECAACFAMFRDSPETAGVCPRCGHENYYSDEESQWDGEILSTSSASRCEKASRGYPTTZ**  **DNA master gene data 44.5**  **Blastp match 44.5** | | | | | | | |
| #44.5 Functional prediction:    [RCSB PDB - 2CON: Solution Structure of RSGI RUH-035, a Zn-ribbon module in Mouse cDNA](https://www.rcsb.org/structure/2CON) | | | | | | | |
| Tom Brady **#45** | SSC: 35090-35521 | CP: yes | SCS: Glimmer | Blast: best match is Gomashi gene 45  1:1 match  15 others | Gap: 139 bp gap | LORF: Yes | RBS:  Spacer: 9  Z-score: 3.276  Final Score: -2.536  These were the best scores |
| **#45(2)** | SSC: 35129-35521 | CP: yes | SCS: Genemark | Blast: best match is HALO gene 46  1:1 match  22 others | Gap: 178 bp gap | LORF: no | RBS:  Spacer:9  Z-score: 0.889  Final score: -7.190  These were the worst scores |
| #45 Note: Glimmer predicted start was chosen because it had a smaller gap, better z-score and final score, and it was the LORF. However, Genemark's predicted start had more 1:1 matches, and if we insert gene 44.5 the genemark start would cause less overlap.  **Glimmer:**  **Blastp glimmer match 45**  **Genemark:**  **Blastp genemark match 45** | | | | | | | |
| **#45 Functional prediction:****DNA binding protein** | | | | | | | |
| Tom Brady **#45.5** Insterted gene 2 | SSC: 35518-35907 | CP: there seems to be no coding potential in common with the host, and very little phage coding potential. | SCS: neither | Blast: best match is ANGEL gene 47  1:1 match  32 others | Gap: 4 bp overlap | LORF: No, but minimizes overlap. | RBS:  Spacer: 9  Z-score: 2.647  Final score: -3.973 |
| #45.5 Notes: This was a manual entry using DNA Master to assist with start and end coordinates and protein sequence. Thank you, Dr. Johnson and Kyle Stoecker! It has homology 32 other sequences.  Protein sequence: **MKAHHARIPCNIARSVMHENRATSAVHTLHGNRATSATFETHNLARFFVQHHRPESCTAAVQDVMHGFSATKDNPPTYVGKDNPIPSCTACKTAVLPYRNLTANHLDQTSHRGNQLTARAERGVSPCQHZ**  **DNA master graph 45.5DNA master gene data 45.5**  **blastp match 45.5Genemark Graph 45.5** | | | | | | | |
| #45.5 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady **#46** | SSC: 35895-36140 | CP: yes | SCS: Chosen: Glimmer  (see note) | Blast: best match is ANGEL gene 48  1:1 match  39 others | Gap: 373 bp gap | LORF:  Yes | RBS:  Spacer: 11  Z-Score: 2.674  Final Score: -3.691  These were the best scores |
| #46 Note: This start position was chosen because it had minimal overlap, covered all coding potential, the best z-score and final score, and had a 1:1 match on blastp. Genemark prediction missing on PECAAN but visible on the genemark map; they appear identical  blastp match 46genemark graph 46 | | | | | | | |
| #46 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady **#47** | SSC: 36137-36481 | CP: yes | SCS: Both | Blast: best match is ANGEL gene 49  1:1 match  39 others | Gap: 4 bp overlap | LORF: N/A (<10bp gap) | RBS:  Spacer: 8  Z-Score: 2.343  Final Score: -4.802  These were not the best scores. (see note) |
| #47 Note: This starting position was chosen as it was predicted by both tools and had minimal overlap. The Z-score and final score were not the best, however, a longer ORF was not needed as all coding potential was covered  blastp match 47 | | | | | | | |
| #47 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady **#48** | SSC: 36478-36918 | CP: yes | SCS: Both | Blast: best match is ANGEL gene 50  1:1 match | Gap: 4 bp overlap | LORF: N/A (<10bp gap) | RBS:  Spacer: 12  Z-Score: 3.201  Final Score: -2.742  These were the best scores |
| #48 Note: This start position was chosen because it was predicted by both tools, had minimal overlap, covered all coding potential, the best z-score and final score, and had a 1:1 match on blastp.  blastp match 48 | | | | | | | |
| #48 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady **#49** | SSC: 36915-37505 | CP: Maybe, hard to tell on genemark profle. | SCS: Both | Blast: best match is ANGEL gene RuvC  1:1 match | Gap: 4 bp overlap | LORF: N/A (<10bp gap) | RBS:  Spacer: 12  Z-Score: 2.736  Final Score: -3.649  These were the best scores |
| #49 Note: This start position was chosen because it was predicted by both tools, had minimal overlap, the best z-score and final score, and had a 1:1 match on blastp.  blastp match 49 | | | | | | | |
| **#49 functional prediction:****RUVC-like resolvase** [RCSB PDB - 4EP4: Thermus thermophilus RuvC structure](https://www.rcsb.org/structure/4EP4) | | | | | | | |
| Tom Brady **#50** | SSC: 37502-37612 | CP: yes | SCS: Chosen: Genemark.  Glimmer predicted 37478 | Blast: best match is Cadesite gene 52  1:1 match  20 others | Gap: 4 bp overlap | LORF: no | RBS:  Spacer: 9  Z-Score: 1.548  Final Score: -5.904  These were second best scores |
| #50 Note: This start position was chosen because it had minimal overlap, the second-best z-score and final score, covered all coding potential, and had a 1:1 match on blastp with 20 homologs. Glimmer predicted start point had worse Z-score and final score, a larger overlap, and an oversized spacer of 18 bp.  Extra note: the Bar for the open reading frame is not on the genemark profile in spite of the genemark predicted start being visible on PECAAN as shown below.  **Genemark:genemark graph 50PECAAN:PECAAN genemark/glimmer predicted starts 50**  **Genemark:**  **blastp match genemark 50**  **Glimmer:**  **blastp match glimmer 50**this match only has 13 others. | | | | | | | |
| **#50 functional prediction:****Transcription initiation factor**   [RCSB PDB - 7NVU: RNA polymerase II core pre-initiation complex with open promoter DNA](https://www.rcsb.org/structure/7NVU) | | | | | | | |
| Tom Brady **#51** | SSC: 37609-38052 | CP: Yes | SCS: Chosen: Neither  Both genemark and glimmer predict 37633  (see note) | Blast: best match is Liefie gene 51  1:1 match  22 others | Gap: 4 bp overlap | LORF: N/A (<10bp gap) | RBS:  Spacer: 10  Z-Score: 3.211  Final Score: -2.582  These were the best scores |
| #51 Note: Chosen start position gives a longer ORF, and the 4 overlap doesn't cover any of the coding potential of the previous gene, while covering all of its own coding potential. Additionally, this start position has a better Z-score and final score. I also ensured this starting position has a 1:1 match.  **Chosen:**  **blastp match chosen 51**This match has 22 more homologs  **Glimmer/Genemark start:**  **blastp match genemark/glimmer 51**this match has only 4 others | | | | | | | |
| #51 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady **#52** | SSC: 38049-38207 | CP: No? Hard to tell on genemark profile | SCS: Both | Blast: best match is Angel gene 53  1:1 match  48 others | Gap: 4 bp overlap | LORF: N/A (<10bp gap) | RBS:  Spacer: 11  Z-Score: 2.84  Final Score: -3.368  These were the best scores |
| #52 Note: This start position was chosen because it was predicted by both tools, had minimal overlap, the best z-score and final score, and had a 1:1 match on blastp. It does not cover all coding potential but there is no viable earlier start.  blastp match 52 | | | | | | | |
| #52 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady **#53** | SSC: 38204-38590 | CP: yes | SCS: Glimmer | Blast: Best match is Cadadesite gene 55  1:1 match  2 others | Gap: 4 bp overlap | LORF: no, but minimizes overlap | RBS:  Spacer: 6  Z-Score: 2.447  Final Score: -5.122  These were the second best scores |
| **#53(2)** | SSC: 38207-38950 | CP: no | SCS: Genemark | Blast: Best match is Cadadesite gene 55  1:2 match | Gap: 1 bp overlap | LORF: no | RBS:  Spacer: 9  Z-Score: 2.447  Final Score: -4.152  These were the best scores |
| #53 Note: Although genemark's predicted start has better RBS scores, glimmer's start has a slightly longer ORF **and** 3 1:1 matches.  blastp match 53 | | | | | | | |
| #53 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady  Reverse **#54** | SSC: 38693-38580 | CP: yes? | SCS: Glimmer | Blast: Best match is Dylan gene 48  1:1 match  2 others | Gap: 440 | LORF: no | RBS:  Spacer: 12  Z-Score: 0.801  Final Score: -7.422 |
| #54 Note: **To be removed in favor of inserted gene 3 (54.5)**  **blastp match 54** | | | | | | | |
| #54 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady **#54.5** inserted gene 3 | SSC: 38644-39015 | CP: yes | SCS: Neither  (see note) | Blast: Best match is Che8 gene 89  1:1 match  97 others | Gap: 53 bp gap | LORF: Yes | RBS:  Spacer: 8  Z-Score: 1.235  Final Score: -6.819 |
| #54.5 Notes: This was a manual entry using DNA Master to assist with start and end coordinates and protein sequence. Thank you, Dr. Johnson! There were 97 1:1 matches.  Protein sequence: **MFPITDTRREMTTMPTTEHGSDVQHLSPEHRDRAWRDRFNARWHYDYGGWIRTRPQDEASTFALIPTKHYGPFTEDHSCPACLVVHPPEDCPVLSGNTDMLVVFDYDTSPNKAQADTADDDPR**  **DNA master gene data 54.5**  **blastp match 54.5** | | | | | | | |
| #54.5 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady **#55** | SSC: 39134-39418 | CP: yes | SCS: neither | Blast: best blast match is BPs gene 55  1:1 match  26 others | Gap: 440 bp gap (calculated from 54, when calculated from 54.5 it is a 119 bp gap) | LORF: Yes | RBS:  Spacer: 10  Z-Score: 2.303  Final Score: -4.352  These were the best scores |
| #55 Note: Both glimmer and genemark chose a start that had worse RBS scores, and cut off some of the phage coding potential (although not the host coding potential). Genemark's predicted start also had a spacer of 18 (too long) My chosen start position also had 27 homologs in the blastp database.  blastp match 55 | | | | | | | |
| #55 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady **#56** | SSC: 39415-40212 | CP: No | SCS: neither  (see note) | Blast: best blast match Halo gene 59  1:1 match  19 others | Gap: 4 bp overlap | LORF: no, but minimizes overlap | RBS:  Spacer: 8  Z-Score: 1.981  Final Score: -5.507  These were the best scores within this region or start options. (there were better scores far earlier with major overlap, or far later that sacrificed hundreds of bps) |
| #56 Note: Glimmer's chosen start was late, cutting off the majority of the gene. Genemark's start was close, however my start had slightly better Final score, and an 8 bp spacer as opposed to 14 on genemark's. In addition the blastp results showed genemark's start position had only 1 99% match, whereas mine had 20 homologs. The 4 bp overlap is acceptable. None of these options cover ALL of the coding potential, but my chosen start comes closest.  **CHOSEN:**  **blastp match 56 chosen**This match has 19 others  **Genemark:**  **blastp match genemark 56** | | | | | | | |
| #56 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady  reverse **#57** | SSC: 39951-39463 | CP: Yes | SCS: Glimmer (see note) | Blast: Best blastp match is DNAIII gene 55  1:1 match  1 other | Gap: 260 bp gap | LORF: no | RBS:  Spacer: 15  Z-Score: 2.082  Final Score: -5.690 |
| #57 Note: ***Assuming this gene is coded (it overlaps completely with gene 56)-***genemark had no prediction for this gene. My initial start position choice was based on my incorrect reading of the genemark map, I have updated my start position accordingly, and it is the Glimmer predicted start.  genemark graph 57  **Glimmer match:**  **blastp match glimmer 57** | | | | | | | |
| #57 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady **#58** | SSC: 40212-40394 | CP: Yes (see note) | SCS: both | Blast: best blastp match is ANGEL gene 57  1:1 match  30 more | Gap: 113 | LORF: Yes | RBS:  Spacer: 12  Z-Score: 2.534  Final Score: -4.043  These were the best scores |
| #58 Note: This start covers all host coding potential, but loses a small shoulder of phage coding potential. This does not seem significant, and there is no earlier start for this reading frame.  blastp match 58genemark graph 58 | | | | | | | |
| #58 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady **#59** | SSC: 40426-40761 | CP: yes | SCS: genemark | Blast: Best blastp match is LEO gene 56  1:1 match  22 others | Gap: 31 | LORF: yes | RBS:  Spacer: 6  Z-Score: 2.829  Final Score: -4.378  These were the second best scores |
| #59 Note: I chose the genemark start as it covered more of the coding potential, had better RBS scores, and 23 homologs in blastp. The glimmer start only had fewer homologs and a 15 spacer.  **Genemark:**  **blastp match genemark 59**  **Glimmer:**  **blastp match glimmer 59** | | | | | | | |
| **#59 functional prediction:****membrane protein** | | | | | | | |
| Tom Brady **#60** | SSC: 40860-41102 | CP: yes | SCS: both | Blast: Best blastp match is ANGEL gene 59  1:1 match  45 others | Gap: 98 bp gap | LORF: No | RBS:  Spacer: 14  Z-Score: 3.201  Final Score: -3.254  These were the best scores |
| #60 Note: Although the "best blastp match" was technically Chance64 gene 61 with a score of 411 (vs. 409 for ANGEL) that match was a 1:22 match, and had only 1 other. I chose to go with this start position because it visually covered all of the coding potential with no waste, had more 1:1 matches, with ANGEL being one of them and highly curated, genemark and glimmer consensus, and it having better RBS scores  blastp match 60 | | | | | | | |
| #60 functional prediction:  hypothetical protein | | | | | | | |
| Tom Brady **#61** | SSC: 41126-41452 | CP: Yes | SCS: both | Blast: Best blastp match is ANGEL endonuclease  1:1 match  44 others | Gap: 23 | LORF: Yes | RBS:  Spacer: 15  Z-Score: 2.287  Final Score: -5.292  These were the best scores |
| #61 Note:  blastp match 61 | | | | | | | |
| **#61 functional prediction:****HNH**   [RCSB PDB - 6J9N: NmeHNH+AcrIIC3](https://www.rcsb.org/structure/6J9N) | | | | | | | |
| Tom Brady **#62** | SSC: 41546-41767 | CP: Yes | SCS: both (see note) | Blast: Best blast match is ANGEL gene 61  1:1 match  30 others | Gap: 93 | LORF: Yes | RBS:  Spacer: 6  Z-Score: 3.013  Final Score: -4.018  These were the best scores |
| #62 Note: Genemark start prediction is missing on PECAAN but is present on the graph. They appear to be the same starts  blastp match 62 | | | | | | | |
| #62 functional prediction:  hypothetical protein | | | | | | | |