Things about this genome:

• the location of the slippery sequence is not known for phages in cluster EA1. The tail assembly chaperone frameshift could not be annotated.

Vispistious\_1 pham 675

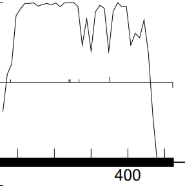
1. auto-annotated start/stop coordinates: 1 → 525; forward transcribed

2. length = 525 bp

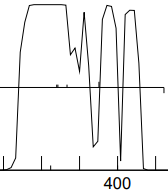
3. start data and coding potential

a. Glimmer @ 1 has strength = 15.26

b. GeneMarkS shows decent coding potential:



GeneMark mfolio also shows good coding potential:



c. start sequence = TTG, stop sequence = TGA

d. downstream N/A; upstream 4 bp GTGA overlap with vis\_2

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 1 | 1.781 | -5.135 | 525 |
| 7 | 1.643 | -5.370 | 519 |
| 82 | 2.004 | -5.617 | 444 |

all of the rbs scores in this gene are low, perhaps because the TTG start codon is rare

f. Starterator: 1 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 1: a) includes all coding potential, b) most annotated start, c) longest ORF.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream genes

c. expected downstream streams

2. BlastP (phagesDB) many hits with great alignment.

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Winzigespinne\_1 | function unknown | 3e-92 |

All non-draft annotations on phagesDB blast are specifying vis\_1 = function unknown

3. BlastP (NCBI) 51 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| PuppyEggo | 1 | hypothetical protein | 2e-121 | 100 | 100 |

All hits on NCBI blastp for vis\_1 are for a hypothetical protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred 0 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| no hits for this gene on HHpred > 90% probability |  |  |  |

6. Transmembrane finders

a. deep TMHMM: vis\_1 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp1 is a soluble protein

conclusion: the function of vis\_1 is unknown.

Vispistious\_2 pham 318

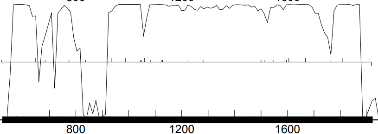
1. auto-annotated start/stop coordinates: 522 → 1922; forward transcribed

2. length = 1401 bp

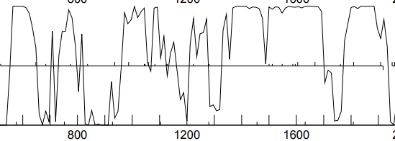
3. start data and coding potential

a. Glimmer @ 522 has strength = 10.90

b. GeneMarkS shows decent coding potential:



GeneMark mfolio also shows good coding potential:



c. start sequence = GTG, stop sequence = TAG

d. downstream 4 bp GTGA overlap with vis\_1; upstream 30 bp gap with vis\_3

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 522 | 1.928 | -4.758 | 1401 |
| 651 | 1.539 | -6.613 | 1272 |
| 687 | 0.735 | -8.067 | 1236 |

again, the rbs scores are not strong, perhaps because this gene is part of an operon with vis\_1. The auto-annotated start@522 is the only start that includes all coding potential

f. Starterator: 522 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 1: a) includes all coding potential, b) most annotated start, c) longest ORF.

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: gene 1

c. expected downstream genes: vis\_2 should be a terminase

2. BlastP (phagesDB) many hits with great alignment.

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Winzigespinne\_2 | terminase | 0e-0 |

All non-draft annotations on phagesDB blast are specifying vis\_2 = terminase

3. BlastP (NCBI) 100 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat | 2 | terminase | 0e-0 | 100 | 100 |

All hits on NCBI blastp for vis\_2 are for either terminase or terminase, large subunit

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| 56-264 | protein with unknown function | 5.55e-21 |
| 247-455 | terminase large subunit | 1.88 e-17 |

45% alignment on the right side of the gene with a conserved domain of terminase large subunit

5. HHPred 250 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| terminase, large subunit a | 100 | 1.7e-38 | 93 |
| terminase, large subunit | 100 | 2.6e-38 | 93 |



above: nice

a Gao et al 2022. Prokaryotic innate immunity through pattern recognition of conserved viral proteins. Science 377 (6607) …super-interesting: bacteria have proteins that recognize highly conserved domains such as the terminase

28 hits on HHpred are *p* = 100% for terminase; most of those specifying the large subunit

6. Transmembrane finders

a. deep TMHMM: vis\_2 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp2 is a soluble protein

conclusion: since there is no obvious small subunit of a terminase in this genome, the function of vis\_2 = terminase.

Vispistious\_3 pham 633

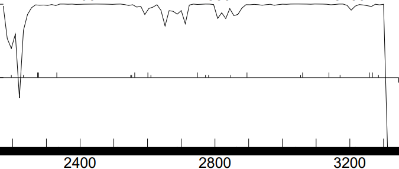
1. auto-annotated start/stop coordinates: 1953 → 3350; forward transcribed

2. length = 1398 bp

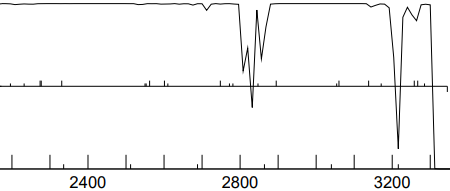
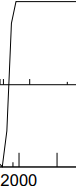
3. start data and coding potential

a. Glimmer @ 1953 has strength = 16.07

b. GeneMarkS shows consistent coding potential:



GeneMark mfolio also shows strong coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream 31 bp gap with vis\_2; upstream 5 bp gap with vis\_4

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 1953 | 2.905 | -2.582 | 1398 |
| 1962 | 1.677 | -5.568 | 1389 |
| 2031 | 1.578 | -6.479 | 2031 |

1953 is a clear winner among competing starts

f. Starterator: 1953 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 1: a) includes all coding potential, b) most annotated start, c) longest ORF, d) strong rbs score.

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: terminase

c. expected downstream genes: after the terminase comes the portal protein

2. BlastP (phagesDB) many hits with great alignment.

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | PuppyEggo\_3 | portal protein | 0e-0 |

All non-draft annotations on phagesDB blast are specifying vis\_3 = portal protein

3. BlastP (NCBI) 100 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| PuppyEggo | 3 | portal protein | 0e-0 | 100 | 100 |

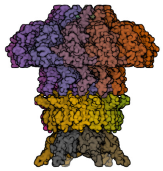
All hits on NCBI blastp for vis\_3 are for portal protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred 57 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| portal protein, SPP1 a | 99.9 | 1.6e-21 | 82 |



above: portal protein

the top 15 hits on HHpred are high-likelihood (> *p* = 98%) and id vis\_3 as a portal protein

6. Transmembrane finders

a. deep TMHMM: vis\_3 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp3 is a soluble protein

conclusion: gp3 is a portal protein

Vispistious\_4 pham 698

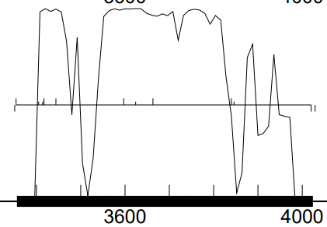
1. auto-annotated start/stop coordinates: 3356 → 4024; forward transcribed

2. length = 669 bp

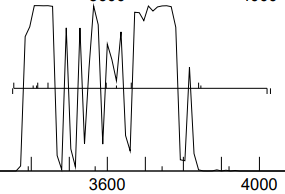
3. start data and coding potential

a. Glimmer @ 3356 has strength = 11.50

b. GeneMarkS shows extensive coding potential:



GeneMark mfolio shows spikey coding potential without some the coding potential found on the right side of gene by GenemarkS:



c. start sequence = ATG, stop sequence = TAG

d. downstream 5 bp gap vis\_3; upstream 83 bp gap with vis\_5.

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 3356 | 2.107 | -4.645 | 669 |
| 3407 | 1.311 | -6.066 | 618 |
| 1.058 | 1.058 | -6.624 | 609 |

3356 has an OK rbs score (but not the best) out of 12 starts in this gene and includes all of the coding potential

f. Starterator: 3356 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 3356: a) includes all coding potential, b) longest ORF, c) most annotated start, d) fairly strong rbs score

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: vis\_2 = terminase, vis\_3 = portal protein.

c. ⸫ vis\_4 should be capsid maturation protease

2. BlastP (phagesDB) many hits with great alignment.

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Zada\_4 | MuF-like minor capsid protease | 1e-128 |
|  | WildNOut\_4 | MuF-like minor capsid protease | 1e-128 |
|  | Velene\_4 | capsid maturation protease | 1e-128 |

All non-draft annotations are specifying functions for vis\_4 as 1) MuF-like minor capsid protease, 2)capsid maturation protease, or 3) both (Shee\_4)

3. BlastP (NCBI) 96 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat | 4 | head morphogenesis | 5e-162 | 100 | 100 |
| Superfresh | 4 | MuF-like minor capsid protein | 1e-161 | 100 | 99.55 |
| Shee | 4 | capsid maturation protease and MuF-like fusion protein | 1e-161 | 100 | 99.55 |

Hits on NCBI blastp for gp4 are zeroing in on the general area of capsid maturation protease.

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| 86-181 | Mu F-like protein | 3.99e-08 |

The Mu F-like protein domain is present

5. HHPred 13 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| capsid maturation protease | 97.6 | 0.021 | 76 |
| structural protein | 99.6 | 7.7e-14 | 82 |
| phage Mu F-like protein | 99.7 | 2.9e-15 | 43 |

HHpred also identified the Mu F-like protein domain, but it overlaps completely with the capsid maturation protease domain. Conclusion = the Mu F-like protein domain is an integral part of capsid maturation protease

6. Transmembrane finders

a. deep TMHMM: vis\_4 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp4 is a soluble protein

conclusion: gp4 is a capsid maturation protease

Vispistious\_5 pham 67102

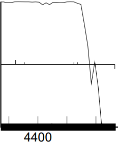
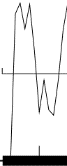
1. auto-annotated start/stop coordinates: 4108 → 4668; forward transcribed

2. length = 561 bp

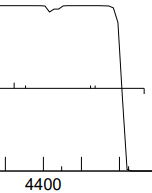
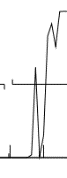
3. start data and coding potential

a. Glimmer @ 4108 has strength = 12.07

b. GeneMarkS shows good coding potential:



GeneMark mfolio also shows strong coding potential:



c. start sequence = ATG, stop sequence = TAG

d. downstream 83 bp gap with vis\_4; upstream 38 bp gap with vis\_6.

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 4108 | 3.030 | -3.222 | 561 |
| 4324 | 1.127 | -6.923 | 345 |
| 4354 | 1.118 | -6.944 | 315 |

4108 has the best rbs score and includes all of the coding potential

f. Starterator: 4108 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 4108: a) includes all coding potential, b) longest ORF, c) most annotated start, d) fairly strong rbs score

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: vis\_2 = terminase, vis\_3 = portal protein, vis\_4 = capsid maturation protease.

c. ⸫ vis\_5 should be a scaffolding protein

2. BlastP (phagesDB) many hits with great alignment.

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNout\_5 | scaffolding protein | 1e-103 |

All non-draft annotations are specifying functions for vis\_5 as scaffolding protein without dissent

3. BlastP (NCBI) 79 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat | 5 | scaffolding protein | 1e-130 | 100 | 100 |
| Peep | 5 | scaffolding protein | 3e-130 | 100 | 99.46 |
| Gargoyle | 5 | scaffolding protein | 4e-130 | 100 | 99.46 |

Hits on NCBI blastp for gp5 = scaffolding protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred 250 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| chaperone | 92.81 | 6.5 | 76 |
| scaffolding protein | 92.43 | 11 | 83 |
| scaffolding protein (phrog\_4080) | 100 | 5.7e-83 | 98 |

A scaffolding protein would have a chaperone-like function…the two are not mutually exclusive. Phrog is zeroing in on function = scaffolding protein

6. Transmembrane finders

a. deep TMHMM: vis\_5 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp5 is a soluble protein

conclusion: gp5 is a scaffolding protein

Vispistious\_6 pham 70123

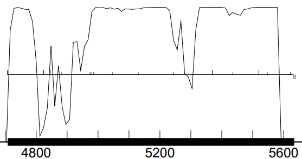
1. auto-annotated start/stop coordinates: 4708 → 5634; forward transcribed

2. length = 927 bp

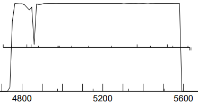
3. start data and coding potential

a. Glimmer @ 4708 has strength = 17.89

b. GeneMarkS shows good coding potential:



GeneMark mfolio also shows strong coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream 38 bp gap with vis\_5; upstream 4 bp GTGA overlap with vis\_7.

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 4708 | 3.198 | -2.305 | 927 |
| 4825 | 0.589 | -7.613 | 810 |
| 4846 | 1.815 | -5.063 | 789 |

4708 has the best rbs score and includes all of the coding potential

f. Starterator: 4708 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 4708: a) includes all coding potential, b) longest ORF, c) most annotated start, d) best rbs score

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: vis\_2 = terminase, vis\_3 = portal protein, vis\_4 = capsid maturation protease, vis\_5 is a scaffolding protein

c. ⸫ vis\_6 should be a major capsid protein

2. BlastP (phagesDB) many hits with great alignment.

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Winzigespinne\_6 | major capsid protein | 1e-173 |

All non-draft annotations on phagesDB blast are specifying functions for vis\_6 as a major capsid protein without dissent

3. BlastP (NCBI) 100 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| MonChoix | 6 | major capsid protein | 0e-0 | 99 | 100 |
| Ilzat | 6 | major capsid protein | 0e-0 | 99 | 99.68 |
| TinSulphur | 6 | major capsdid protein | 0e-0 | 99 | 99.68 |

Hits on NCBI blastp for gp6 = major capsid protein

4. BlastP (Conserved domains)

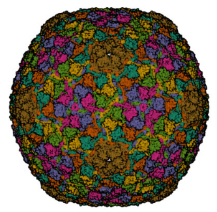
|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| 41-282 | phage capsid protein | 2.51e-15 |

79% of vis\_6 aligns with this protein

5. HHPred 103 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| major capsid protein | 99.97 | 8.1e-29 | 91 |

An unusual number of high-probability hits on HHpred converge on a major capsid protein for vis\_6



above: capsid protein structure

6. Transmembrane finders

a. deep TMHMM: vis\_6 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp6 is a soluble protein

conclusion: gp6 is a major capsid protein

Vispistious\_7 pham 70330

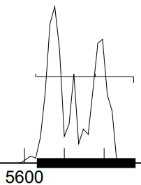
1. auto-annotated start/stop coordinates: 5631→ 5879; forward transcribed

2. length = 249 bp

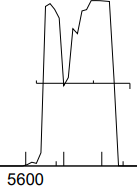
3. start data and coding potential

a. Glimmer @ 5631 has strength = 7.99

b. GeneMarkS shows moderate coding potential:



GeneMark mfolio shows more robust coding potential:



c. start sequence = GTG, stop sequence = TGA

d. downstream 4 bp GTGA overlap with vis\_6; upstream 70 bp gap to vis\_8.

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 5631 | 1.888 | -4.906 | 249 |
| 5781 | 1.466 | -5.750 | 99 |

5631 is the only viable start for this gene

f. Starterator: pham report not available.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 5631: a) includes all coding potential, b) longest ORF, c) best rbs score (#1/2), d) only viable start

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: vis\_2 = terminase, vis\_3 = portal protein, vis\_4 = capsid maturation protease, vis\_5 is a scaffolding protein, vis\_6 is a major capsid protein

c. next in line would be tail assembly chaperones, although the 4 bp overlap with the structural capsid protein would hint at some structural function for vis\_7 related to the capsid

2. BlastP (phagesDB) many hits with great alignment.

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | PuppyEggo\_7 | function unknown | 9e40 |

All non-draft annotations on phagesDB blast are specifying vis\_7 = function unknown

3. BlastP (NCBI) 19 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| PuppyEggo | 7 | hypothetical protein | 1e-47 | 98 | 100 |

all hits on NCBI blastp for gp7 = hypothetical protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred 0 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| no hits > 90% probability |  |  |  |

6. Transmembrane finders

a. deep TMHMM: vis\_7 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp7 is a soluble protein

conclusion: the function of vis\_7 is unknown

Vispistious\_8 pham 1681

1. auto-annotated start/stop coordinates: 5951→ 6085; forward transcribed

2. length = 135 bp…barely log enough to be a gene

3. start data and coding potential

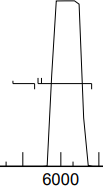
a. Glimmer @ 5631 has strength = 16.22

b. GeneMark calls the start at 5942

c. GeneMarkS shows a sharp bump of coding potential:



GeneMark mfolio shows the same potential:



d. start sequence = ATG, stop sequence = TGA

e. downstream 70 gap with vis\_7; upstream 28 bp gap to vis\_9.

f. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 5942 | 1.503 | -5.652 | 144 |
| 5951 | 2.970 | -2.794 | 135 |

there are two starts for this gene, both encompass all of the narrow region of coding potential

g. Starterator: 5942 is the most annotated start, sort of (27/50).

h. longest possible ORF? the start at 5942 would include an extra 9bp

i. conclusion: this is a gene with start at 5951: a) includes all coding potential and b) strong rbs score. 5942 is the most annotated start, but could find no compelling reason not go with 5951. Both starts include all coding potential. The deciding factor is the rbs score.

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: vis\_2 = terminase, vis\_3 = portal protein, vis\_4 = capsid maturation protease, vis\_5 is a scaffolding protein, vis\_6 is a major capsid protein

c. next in line would be tail assembly chaperones, although vis\_8 is perhaps too small to be a minor tail protein.

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 58 | WildNout\_8 | function unknown | 2e-21 |

All non-draft annotations on phagesDB blast are specifying vis\_8 = function unknown

3. BlastP (NCBI) 4 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Teagan | 8 | hypothetical protein | 6e-23 | 97 | 100 |

all hits on NCBI blastp for gp8 = hypothetical protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred 0 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| no hits > 90% probability |  |  |  |

6. Transmembrane finders

a. deep TMHMM: vis\_8 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp8 is a soluble protein

conclusion: the function of vis\_8 is unknown

Vispistious\_9 pham 643

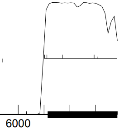
1. auto-annotated start/stop coordinates: 6114 → 6527; forward transcribed

2. length = 414 bp

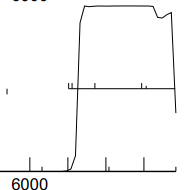
3. start data and coding potential

a. Glimmer @ 6114 has strength = 10.21

b. GeneMarkS shows good coding potential:



GeneMark mfolio also shows good coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream 28 bp gap to vis\_8; upstream 28 bp overlap with vis\_10...that’s a large overlap

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 6105 | 1.439 | -6.381 | 423 |
| 6114 | 2.970 | -2.794 | 414 |
| 6174 | 2.432 | -4.426 | 354 |

6114 is looking good vs. the other two potential starts

f. Starterator: 6114 is the most annotated start.

g. longest possible ORF? 6114 includes all coding potential

h. conclusion: this is a gene with start at 6114: a) includes all coding potential, b) best rbs score, c) most annotated start.

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: vis\_2 = terminase, vis\_3 = portal protein, vis\_4 = capsid maturation protease, vis\_5 is a scaffolding protein, vis\_6 is a major capsid protein

c. next in line would be tail assembly chaperones?

2. BlastP (phagesDB) many hits with great alignment.

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Zada\_9 | function unknown | 6e-76 |

All non-draft annotations on phagesDB blast are specifying vis\_9 = function unknown

3. BlastP (NCBI) 19 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Teagan | 9 | hypothetical protein | 8e-96 | 99 | 100 |

all hits on NCBI blastp for gp9 = hypothetical protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred 0 hits

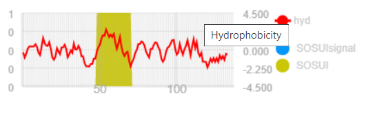
|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| no hits > 90% probability |  |  |  |

6. Transmembrane finders

a. deep TMHMM: vis\_9 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp9 is a membrane protein, with residues 49-71 in a membrane environment



conclusion: the function of vis\_9 is unknown. Cannot call membrane protein without evidence from TOPCON or TMHMM

Vispistious\_10 pham 632

1. auto-annotated start/stop coordinates: 6493 → 6906; forward transcribed

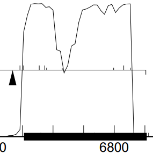
2. length = 414 bp

3. start data and coding potential

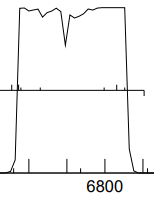
a. Glimmer @ 6493 has strength = 12.00

b. GeneMark calls the start at 6505

c. GeneMarkS shows good coding potential:



GeneMark mfolio also shows good coding potential:



d. start sequence = ATG, stop sequence = TGA

e. downstream 28 bp overlap with vis\_9; upstream 4 bp ATGA overlap with vis\_11

f. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 6397 | 1.00 | -7.915 | 540 |
| 6424 | 2.628 | -4.006 | 483 |
| 6493 | 2.677 | -3.071 | 414 |
| 6505 | 1.843 | -4.861 | 402 |

g. Starterator: 6493 is the most annotated start.

h. longest possible ORF? 6493 includes all coding potential

i. conclusion: this is a gene with start at 6493: a) includes all coding potential, b) best rbs score, c) most annotated start.

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: vis\_2 = terminase, vis\_3 = portal protein, vis\_4 = capsid maturation protease, vis\_5 is a scaffolding protein, vis\_6 is a major capsid protein

c. next in line would be tail assembly chaperones?

2. BlastP (phagesDB) many hits with great alignment but no function

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut\_10 | function unknown | 2e-69 |

All non-draft annotations on phagesDB blast are specifying that the function of vis\_10 is unknown

3. BlastP (NCBI) 66 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 10 | hypothetical protein | 2e-91 | 99 | 100 |

all hits on NCBI blastp for gp10 = hypothetical protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred 0 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| a pf 19586.2 family of unknown function | 96.56 | 0.017 | 66 |
| unknown function; pyruvate kinase beta-barrel-like | 100 | 7.5e-77 | 98 |
| unknown function; putative phage tail protein | 98.7 | 7.5e-08 | 86 |
| function unknown; putative phage tail protein | 97.37 | 0.00056 | 88 |

a biosynthetic gene cluster?

a few good hits on HHpred linking to protein of unknown function; predicted that these are tail proteins

6. Transmembrane finders

a. deep TMHMM: vis\_10 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp10 is not a membrane protein

conclusion: the function of vis\_10 is unknown.

Vispistious\_11 pham 64518

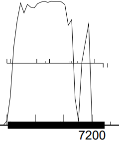
1. auto-annotated start/stop coordinates: 6903 → 7244; forward transcribed

2. length = 342 bp

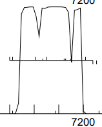
3. start data and coding potential

a. Glimmer @ 6903 has strength = 14.11

b. GeneMarkS shows decent coding potential:



GeneMark mfolio also shows good coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream 4 bp ATGA overlap with vis\_10; upstream 2 bp gap to vis\_12

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 6903 | 2.970 | -2.584 | 342 |
| 6915 | 1.761 | -5.056 | 330 |
| 7008 | 2.432 | -5.105 | 237 |

6903 has the best rbs score of 9 starts in this gene

f. Starterator: 6903 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 6903: a) includes all coding potential, b) best rbs score, c) most annotated start, d) longest ORF.

Gene function

1. Synteny

a. is synteny informative for this gene? maybe

b. upstream genes: vis\_2 = terminase, vis\_3 = portal protein, vis\_4 = capsid maturation protease, vis\_5 is a scaffolding protein, vis\_6 is a major capsid protein

c. next in line would be tail assembly chaperones?

2. BlastP (phagesDB) many hits with great alignment.

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut1 | function unknown | 8e-58 |

All non-draft annotations on phagesDB blast are specifying vis\_11 = function unknown

3. BlastP (NCBI) 80 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Elen | 11 | tail completion or Neck1 protein | 5e-39 | 99 | 58.3 |
| HerculesXL |  | minor capsid protein | 9e-39 | 99 | 61.7 |

the highest probability hits on NCBI blastp for vis\_11 are for a hypothetical protein; lower on the list are tail completion protein/minor capsid protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred 20 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| minor capsid protein a | 99.5 | 1.9e-13 | 71 |
| putative tail component | 99.2 | 1.5e-10 | 61 |
| putative capsid assembly protein | 99 | 3.6e-9 | 58 |
| tail completion protein | 98.4 | 6.7e-6 | 97 |
| TIGR01725 b putative tail component | 99.8 | 1.7e-17 | 94 |

a genomes with this protein usually have three adjacent genes that code for similar proteins; vis\_11 would code for minor capsid protein #2

b this protein may assist in tail morphogenesis

Conflicting lines of evidence in HHpred. All of these hits are high likelihood.

6. Transmembrane finders

a. deep TMHMM: vis\_11 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp11 is a soluble protein

conclusion: the function of vis\_11 is unknown. This protein likely is either a minor capsid protein or has some role in tail morphogenesis. Conclusive evidence is lacking to make a call either way.

Vispistious\_12 pham 640

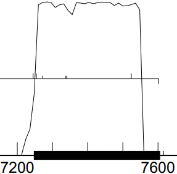
1. auto-annotated start/stop coordinates: 7247 → 7606; forward transcribed

2. length = 360 bp

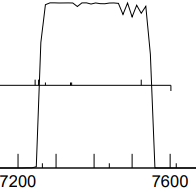
3. start data and coding potential

a. Glimmer @ 7247 has strength = 16.27

b. GeneMarkS shows decent coding potential:



GeneMark mfolio also shows solid coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream 2 bp gap to vis\_11; upstream 4 bp ATGA overlap with vis\_13

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 7247 | 2.320 | -3.978 | 360 |
| 7256 | 1.666 | -6.070 | 351 |
| 7274 | 1.925 | -5.990 | 333 |

7247 has the top rbs score

f. Starterator: 7247 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 7240: a) includes all coding potential, b) best rbs score, c) most annotated start, d) longest ORF.

Gene function

1. Synteny

a. is synteny informative for this gene? maybe

b. upstream genes: vis\_2 = terminase, vis\_3 = portal protein, vis\_4 = capsid maturation protease, vis\_5 is a scaffolding protein, vis\_6 is a major capsid protein

c. next in line would be tail assembly chaperones?

2. BlastP (phagesDB) many hits with excellent alignment.

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut\_12 | tail terminator | 2e-62 |

All non-draft annotations on phagesDB blast are specifying vis\_12 = tail terminator

3. BlastP (NCBI) 54 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 12 | tail terminator | 2e-79 | 100 | 100 |
| Alyxandracam | 12 | tail terminator | 9e-39 | 100 | 99.2 |

all hits on NCBI blastp are pointing towards a tail terminator protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred 0 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| 5A21\_G tail terminator a | 98.2 | 2.7e-5 | 70 |

a Chaban et al 2015 Structural rearrangements in the phage head-to-tail interface during assembly and infection. PNAS 112(22) 7009-7014



above: structure of gp\_12, or something like it

6. Transmembrane finders

a. deep TMHMM: vis\_12 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp12 is a soluble protein

conclusion: the function of vis\_12 = tail terminator. The SEAPHAGES case study was super-helpful

Vispistious\_13 pham 67103

1. auto-annotated start/stop coordinates: 7603 → 7839; forward transcribed

2. length = 237 bp

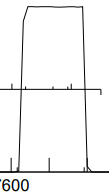
3. start data and coding potential

a. Glimmer @ 7603 has strength = 14.71

b. GeneMarkS shows good coding potential:



GeneMark mfolio also shows strong coding potential:



c. start sequence = ATG, stop sequence = TAA

d. downstream 4 bp ATGA overlap with vis\_12; upstream 14 bp gap to vis\_14

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 7603 | 3.125 | -2.175 | 237 |
| 7639 | 1.350 | -5.998 | 201 |
| 7711 | 0.963 | -7.101 | 129 |

7603 has the top rbs score

f. Starterator: 7603 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 7603: a) includes all coding potential, b) best rbs score, c) most annotated start, d) longest ORF.

Gene function

1. Synteny

a. is synteny informative for this gene? maybe

b. upstream genes: vis\_2 = terminase, vis\_3 = portal protein, vis\_4 = capsid maturation protease, vis\_5 is a scaffolding protein, vis\_6 is a major capsid protein

c. next in line would be tail assembly chaperones?

2. BlastP (phagesDB) many hits with excellent alignment.

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut\_13 | function unknown | 2e-69 |

All non-draft annotations on phagesDB blast are specifying vis\_13 = function unknown

3. BlastP (NCBI) 54 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 13 | hypothetical protein | 7e-49 | 100 | 100 |

all hits on NCBI blastp for vis\_13 = hypothetical protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred 0 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| no HHpred hits > 90% probability for this protein |  |  |  |

6. Transmembrane finders

a. deep TMHMM: vis\_13 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp13 is a soluble protein

conclusion: the function of vis\_13 = function unknown

Vispistious\_14 pham 647

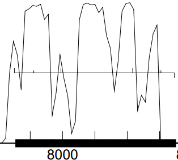
1. auto-annotated start/stop coordinates: 7854 → 8351; forward transcribed

2. length = 498 bp

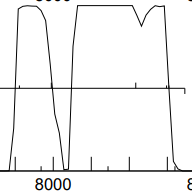
3. start data and coding potential

a. Glimmer @ 7854 has strength = 14.92

b. GeneMarkS shows spikey coding potential:



GeneMark mfolio shows two lumps of coding potential:



c. start sequence = ATG, stop sequence = TAA

d. downstream 14 bp gap to vis\_13; upstream 30 bp gap to vis\_15

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 7854 | 2.876 | -2.645 | 458 |
| 7914 | 1.191 | -6.612 | 438 |
| 7998 | 1.076 | -6.569 | 354 |

7854 has the best rbs score

f. Starterator: 7854 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 7854: a) includes all coding potential, b) best rbs score, c) most annotated start, d) longest ORF.

Gene function

1. Synteny

a. is synteny informative for this gene? maybe

b. upstream genes: vis\_2 = terminase, vis\_3 = portal protein, vis\_4 = capsid maturation protease, vis\_5 is a scaffolding protein, vis\_6 is a major capsid protein

c. waiting for those tail assembly chaperones

2. BlastP (phagesDB) many hits with excellent alignment.

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut\_14 | major tail protein | 2e-91 |

All non-draft annotations on phagesDB blast are specifying vis\_14 = major tail protein

3. BlastP (NCBI) 56 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 14 | major tail protein | 7e-49 | 100 | 100 |

the consensus on NCBI blastp for vis\_14 = major tail protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred 0 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| phage tail tube protein, lambda-like | 88.6 | 16 | 67 |
| major tail protein (phrog\_2491) | 100 | 2.5e-35 | 92 |
| major tail protein (phrog\_9522) | 95.4 | 1 | 86 |

phrog is convinced that this is a major tail protein

6. Transmembrane finders

a. deep TMHMM: vis\_14 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp14 is a soluble protein

conclusion: the function of vis\_14 = major tail protein

Vispistious\_15 pham 634

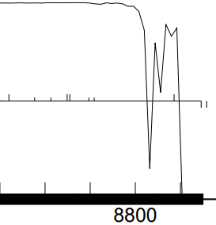
1. auto-annotated start/stop coordinates: 8382 → 8951; forward transcribed

2. length = 570 bp

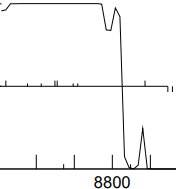
3. start data and coding potential

a. Glimmer @ 8382 has strength = 13.82

b. GeneMarkS shows extensive coding potential:



GeneMark mfolio shows the same coding potential:



c. start sequence = ATG, stop sequence = TAG

d. downstream 18 base gap with vis\_14; upstream 14 bp gap to vis\_16

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 8382 | 2.970 | -2.794 | 570 |
| 8409 | 1.555 | -5.558 | 543 |
| 8439 | 2.820 | -3.867 | 513 |

8382 has the best rbs score and includes all of the coding potential

f. Starterator: 8382 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 8382: a) includes all coding potential, b) longest ORF, c) best rbs score of competing starts, d) most annotated start

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: vis\_14 = major tail protein

c. downstream gene: vis\_15 is likely to be a tail assembly chaperone; vis\_16 is going to be the tape measure gene

d. ⸫ vis\_15 is likely to be a tail assembly chaperone

2. BlastP (phagesDB) many hits with great alignment. All non-draft genes are annotated as tail assembly chaperone

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Thorongil\_15 | tail assembly chaperone | e-106 |

3. BlastP (NCBI) 76 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| PuppyEggo | 15 | tail assembly chaperone | 3e-135 | 99 | 100 |
| AxiPup | 15 | tail assembly chaperone | 6e-135 | 99 | 99.5 |
| Peep | 15 | tail assembly chaperone | 2e-134 | 99 | 98.9 |

Hits on NCBI blastp for gp15 are function = tail assembly chaperone.

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred (only for genes > 90% probable matches)

78 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| phrog 4129 tail assembly chaperone | 100 | 3e-110 | 99 |
| phrog 18843 tail assembly chaperone | 100 | 7.7e-83 | 91 |
| phrog\_20192 tail assembly chaperone | 95.9 | 0.051 | 79 |

phrog is producing high probability hits = tail assembly chaperone

6. Transmembrane finders

a. deep TMHMM: vis\_15 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp15 is a soluble protein

conclusion: synteny, HHpred and phamerator evidence → vis\_15 is a tail assembly chaperone

Vispistious\_16 pham 885

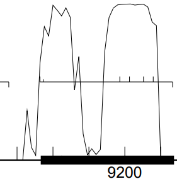
1. auto-annotated start/stop coordinates: 8966 → 9337; forward transcribed

2. length = 372 bp

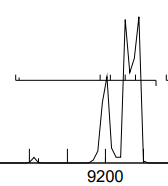
3. start data and coding potential

a. Glimmer @ 8966 has strength = 1.81…super-weak Glimmer score likely because there is a frameshift somewhere in here that cannot be annotated without wet bench evidence

b. GeneMarkS shows two peaks of coding potential:



GeneMark mfolio shows weak coding potential:



c. start sequence = ATG, stop sequence = TAA

d. downstream 14 base gap with vis\_15; upstream 23 bp gap to vis\_17. Note: there’s an undiscovered frameshift in here somewhere

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 8894 | 0.882 | -6.923 | 444 |
| 8966 | 0.852 | -7.815 | 372 |
| 8975 | 1.277 | -6.903 | 363 |

8966 has a super-weak rbs score, but includes all of the coding potential. Again, there is a frameshift lurking in here somewhere…the ribosome is likely not initiating a read at 8966

f. Starterator: 8966 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 8966: a) includes all coding potential, b) longest ORF, c) most annotated start. There’s a missing frameshift that cannot be annotated.

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: vis\_14 = major tail protein; vis\_15 = tail assembly chaperone.

c. downstream gene: vis\_17 is going to be the tape measure gene

d. ⸫ vis\_16 is likely to be a tail assembly chaperone

2. BlastP (phagesDB) many hits with great alignment. All non-draft genes are annotated as tail assembly chaperone

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Zada\_16 | tail assembly chaperone | 2e-71 |

3. BlastP (NCBI) 52 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat | 16 | tail assembly chaperone | 8e-86 | 100 | 100 |
| Endor | 16 | tail assembly chaperone | 2e-85 | 100 | 98.4 |
| Baines | 16 | tail assembly chaperone | 3e-85 | 100 | 99.2 |

Hits on NCBI blastp for gp16 are function = tail assembly chaperone.

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| phrog 4537 tail assembly chaperone | 100 | 3.3e-42 | 99 |
| phrog 18843 tail assembly chaperone | 100 | 4e-34 | 99 |
| phrog 20192 tail assembly chaperone | 99.9 | 2.6e-22 | 82 |

again phrog is liking those tail assembly chaperones

6. Transmembrane finders

a. deep TMHMM: vis\_16 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp16 is a soluble protein

conclusion: synteny and phamerator evidence → vis\_16 is a tail assembly chaperone. The frameshift cannot be annotated at this time.

Vispistious\_17 pham 59270

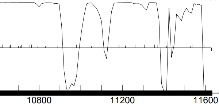
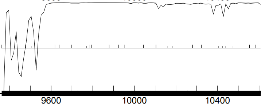
1. auto-annotated start/stop coordinates: 8966 → 9337; forward transcribed

2. length = 2271 bp…one of the largest genes in this genome

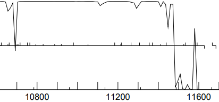
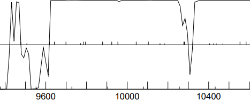
3. start data and coding potential

a. Glimmer @ 9362 has strength = 20.21

b. GeneMarkS shows extensive coding potential:



GeneMark mfolio shows the same coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream 23 base gap with vis\_16; upstream 4 bp ATGA overlap with vis\_18

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 9362 | 3.198 | -1.954 | 2271 |
| 9479 | 1.433 | -5.741 | 2154 |
| 9509 | 2.275 | -3.934 | 2124 |

9362 has a super-strong rbs score, the best out of the 51 starts in this region.

f. Starterator: 9362 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 9362: a) includes all coding potential, b) longest ORF, c) most annotated star, d) #1/51 starts.

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: vis\_14 = major tail protein; vis\_15 and vis\_16 = tail assembly chaperones.

c. downstream gene: vis\_18 and vis\_19 are minor tail proteins.

d. ⸫ vis\_17 is likely to be the tape measure protein, in part identifiable by its length

2. BlastP (phagesDB) many hits with great alignment. All non-draft genes are annotated as tape measure protein

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNout\_17 | tape measure protein | 0.0 |

3. BlastP (NCBI) 100 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 17 | tape measure protein | 0e-0 | 100 | 100 |
| Alyxandracam | 17 | tape measure protein | 0e-0 | 100 | 99.74 |
| Peep | 17 | tape measure protein | 0e-0 | 100 | 99.2 |

mega hits on NCBI blastp for gp17 are function = tape measure protein.

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| 150-679 | phage-related protein | 1.14e-29 |

hit on CDD, but not informative re. function

5. HHPred (only for genes > 90% probable matches)

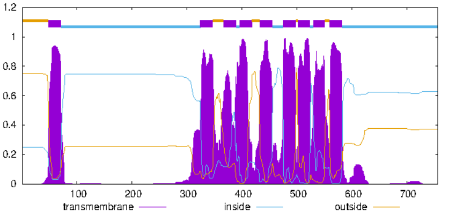
133 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| tail tape measure protein | 100 | 4.6e-19 | 97 |
| tape measure protein | 100 | 1.8-20 | 91 |
| tape measure protein | 99.9 | 2.2e-17 | 96 |

HHpred is confident that this is a tape measure protein

6. Transmembrane finders

a. deep TMHMM: vis\_17 has 9 transmembrane domains



b. TOPCONs: many transmembrane proteins detected



c. SOSUI: gp17 has 6 transmembrane domains



conclusion: vis\_17 is a tape measure protein.

Vispistious\_18 pham 736

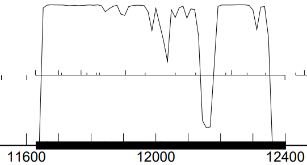
1. auto-annotated start/stop coordinates: 11629 → 12402; forward transcribed

2. length = 774 bp

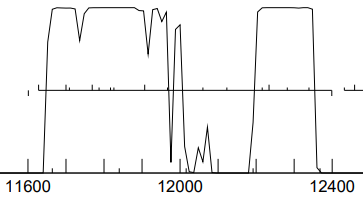
3. start data and coding potential

a. Glimmer @ 11629 has strength = 15.58

b. GeneMarkS shows fairly extensive coding potential:



GeneMark mfolio shows peaky coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream 4 base ATGA overlap with vis\_17; upstream 1 base A overlap with vis\_19.

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 11629 | 2.400 | -3.808 | 774 |
| 11701 | 1.317 | -5.990 | 702 |
| 11710 | 1.178 | -6.940 | 693 |

11629 has the best rbs score out of all 16 starts in this gene

f. Starterator: 11629 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 11629: a) includes all coding potential, b) longest ORF, c) most annotated start, d) strongest rbs score

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: vis\_14 = major tail protein; vis\_15 and 16 = tail assembly chaperones, vis\_17 = tape measure protein.

c. downstream gene: region of minor tail proteins

d. ⸫ vis\_18 could be a minor tail protein

2. BlastP (phagesDB) many hits with great alignment. All non-draft genes are annotated as minor tail protein

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | PuppyEggo\_18 | minor tail protein | 1e-146 |

3. BlastP (NCBI) 70 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 18 | minor tail protein | 0e-0 | 100 | 100 |
| Thompsone | 18 | minor tail protein | 0e-0 | 100 | 99.22 |
| Martin | 18 | minor tail protein | 0e-0 | 100 | 98.44 |

Hits on NCBI blastp for gp18 are function = minor tail protein.

4. BlastP (Conserved domains)

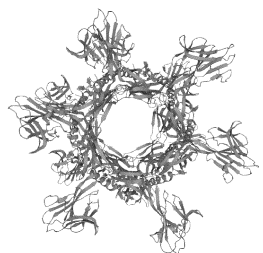
|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred (only for genes > 90% probable matches)

29 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| Distal tail protein a | 99.96 | 2e-26 | 89 |
| distal tail protein | 99.96 | 8.5e-26 | 89 |
| distal tail protein/phage baseplate | 99.93 | 7.2e-23 | 86 |

a Veesler et al 2010. Crystal structure of bacteriophage SPP1 distal tail protein (gp19.1): a baseplate hub paradigm in gram-positive infecting phages. J Biol Chem 285: 36666-36673



6. Transmembrane finders

a. deep TMHMM: vis\_18 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp18 is a soluble protein

conclusion: (tube-shaped) minor tail protein, likely located in the distal part of the tail, forming some sort of baseplate.

Vispistious\_19 pham 742

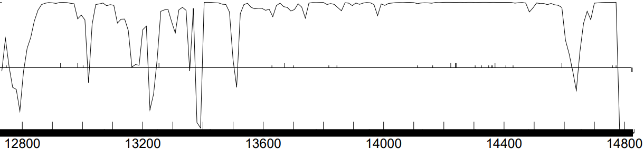
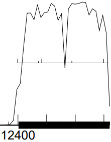
1. auto-annotated start/stop coordinates: 12402 → 14828; forward transcribed

2. length = 2427 bp…surpasses the tape measure protein in length

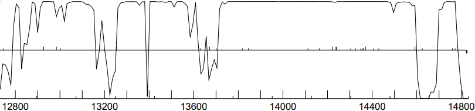
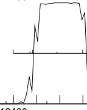
3. start data and coding potential

a. Glimmer @ 11629 has strength = 16.32

b. GeneMarkS shows extensive coding potential:



GeneMark mfolio shows extensive peaky coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream 1 base A overlap with vis\_17; upstream 1 base A overlap with vis\_20.

e. start scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 12402 | 2.820 | -2.765 | 2427 |
| 12474 | 2.065 | -5.486 | 2355 |
| 12486 | 2.713 | -3.346 | 2238 |

12402 has the best rbs score out of 34 starts in this gene and includes all of the coding potential

f. Starterator: 12402 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 12402: a) includes all coding potential, b) longest ORF, c) most annotated start, d) strongest rbs score

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: vis\_14 = major tail protein; vis\_15 and 16 = tail assembly chaperones, vis\_17 = tape measure protein, vis\_18 = minor tail protein.

c. downstream gene: region of minor tail proteins

d. ⸫ vis\_19 could be another minor tail protein

2. BlastP (phagesDB) many hits with great alignment. All non-draft genes are annotated as minor tail protein

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Velene\_19 | minor tail protein | 0e-0 |

3. BlastP (NCBI) 100 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 19 | minor tail protein | 0e-0 | 100 | 99.88 |
| Thorongil | 19 | minor tail protein | 0e-0 | 100 | 99.75 |
| Oxtober96 | 19 | minor tail protein | 0e-0 | 100 | 99.75 |

Hits on NCBI blastp for gp19 are function = minor tail protein.

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| 458-695 | prophage tail with endopeptidase activity | 6.89e-33 |
| 102-333 | LamG super family…Ca2+-activated receptor that in this case may have an adhesion function | 1.97e-03 |

5. HHPred (only for genes > 90% probable matches)

29 hits

two domains pop up on HHpred: 1) some sort of hydrolase/toxin (approx. residues 20-240) and 2) phage tail protein (approx. residues 370-700)

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| hydrolase/toxin complex | 98.07 | 0.00016 | 26 |
| prophage tail protein | 99.93 | 2.2e-23 | 39 |

6. Transmembrane finders

a. deep TMHMM: vis\_19 is not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp19 is a soluble protein

conclusion: minor tail protein

Vispistious\_20 pham 735

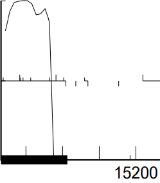
1. auto-annotated start/stop coordinates: 14828 → 15013; forward transcribed

2. length = 186 bp

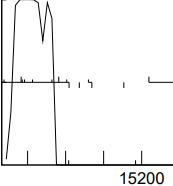
3. start data and coding potential

a. Glimmer @ 14828 has strength = 12.61

b. GeneMarkS shows strong coding potential:



GeneMark mfolio also shows strong coding potential:



d. start sequence = ATG, stop sequence = TGA

e. downstream 1 base overlap with vis\_19; upstream 4 bp ATGA overlap with vis\_21

f. there are at least three competing starts:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 14825 | 2.484 | -4.536 | 189 |
| 14828 | 2.484 | -3.566 | 186 |
| 14840 | 1.650 | -5.337 | 174 |
| 14885 | 1.855 | -5.937 | 129 |

the rbs scores create a toss-up between the auto-annotated start @14828 and a nearby start at 14825. The start at 14825 would establish a four-base GTGA overlap with vis\_19.

g. BLAST-start

100 hits

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val |
| 14828 | Benjalauren + 11 | 20 | NCBI | q 1-186  t 1-186 | 100 | 3e-90 |
| 14825 | Benjalauren + 11 | 20 | NCBI | q 1-189  t 1-189 | 100 | 6e-92 |

the start @14825 produces more complete alignment with annotated genomes.

h. Starterator: 14828 is the most annotated start.

i. longest possible ORF? yes if the start is at 14825

j. conclusion: this is a gene with start at 14825: a) includes all coding potential, b) longest ORF, c) same z-score as the Glimmer-called start at 14828, d) excellent blast alignment with annotated genomes, e) establishes a 4-base overlap with vis\_19, potentially linking genes 19, 20 and 21 as an operon system. Not sure about the significance/plausibility of two consecutive methionines, but will roll with it.

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream gene: gene 19 is annotated as a minor tail protein in all members of this cluster

c. downstream gene: gene 21 is annotated as a minor tail protein in **some** members of this cluster

d. this is the right spot for a minor tail protein

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Velene\_20 + 24 | function unknown | 9e-30 |

100% alignments; the majority of hits on phagesDB = function unknown; four have function = minor tail protein. All hits on phagesdb agree closely with respect to alignment and identity: this is essentially the same protein in each case

3. BlastP (NCBI) 100nhits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat | 20 | function unknown | 4e-36 | 98 | 100 |

all hits on NCBI blastp are function = hypothetical protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| domain | function | e-val |
| no conserved domains identified |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| tail fiber protein (UniProt) | 95.8 | 0.034 | 87 |
| tail protein (Phrog) | 99.8 | 4.6e-21 | 71 |
| galactose-binding domain-like (Phrog) | 96.3 | 0.018 | 81 |
| galactose-binding domain-like/halyuronate (Phrog) | 95.7 | 0.059 | 82 |

This is interesting. This is the correct spot for a structural protein associated with the tail. 19 hits on HHpred (phrog) with ≥ 90% probability for tail protein/minor structural protein/galactose-binding domain. Followed protein family links and found a publication linked to ncbi reference sequence YP\_009227016: Yuan et al 2015. Effects of actin-like proteins encoded by two *Bacillus pumilus* phages on unstable lysogeny, revealed by genomic analysis. Appl. Environ Microbiol 81(1) 339-350.

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp20 is a soluble protein

conclusion: the function of vis\_20 = minor tail protein: 1) synteny, 2) 4-base overlaps with upstream and downstream genes, 3) many hits for a structural tail protein/galactose-binding domain on HHpred.

Vispistious\_21 pham 734

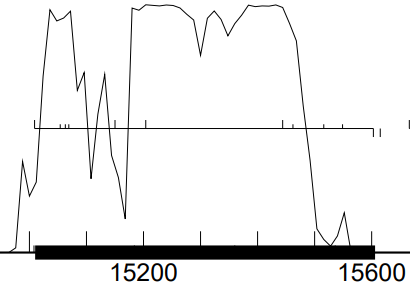
1. auto-annotated start/stop coordinates: 15010 → 15606; forward transcribed

2. length = 597 bp

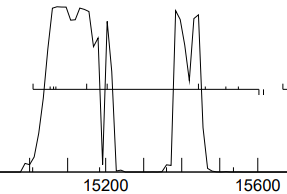
3. start data and coding potential

a. Glimmer @ 15010 has strength = 6.80…Glimmer is not exuding confidence re. this start

b. GeneMarkS shows good coding potential:



GeneMark mfolio shows uneven coding potential, easily interpretable as two separate genes if there were a stop located near the center of the gene:



c. start sequence = ATG, stop sequence = TAA

d. downstream 4 base overlap with vis\_20; upstream 1 base A overlap with vis\_22

e. there are at least three competing starts:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 15010 | 2.030 | -4.602 | 597 |
| 15055 | 1.369 | -6.980 | 552 |
| 15064 | 1.936 | -4.661 | 543 |
| 15070 | 1.936 | -5.763 | 537 |

the auto-annotated start at 15010 has the best (albeit mediocre) rbs score of competing starts. The strongest rbs score (-3.703) would be for a start @15310, which would exclude nearly half of the coding potential indicated by GeneMark

f. BLAST-start

100 hits

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val |
| 15010 | PuppyEggo + AxiPup | 21 | NCBI | q 1-597  t 1-597 | 100 | 0e-0 |

the start @15010 gives more complete alignment with annotated genomes.

g. Starterator: 15010 is the most annotated start.

h. longest possible ORF? yes if the start is at 15010

i. conclusion: this is a gene with start at 15010: a) includes all coding potential, b) longest ORF, c) best rbs score of competing starts, d) blast alignment with annotated genomes, e) most annotated start

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream gene: gene 20 is annotated as a minor tail protein in some members of this cluster

c. downstream gene: gene 22 is annotated as a minor tail protein in some members of this cluster

d. gene 23 is annotated as a minor tail protein in all members of this cluster

e. ⸫ this is the right spot for a minor tail protein

2. BlastP (phagesDB) many hits with great alignment. The majority = function unknown, but some with good alignment and identity are minor tail proteins

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | LovelyUnicorn\_21 | minor tail protein | 1e-114 |
|  | KannH\_21 | minor tail protein | 1e-114 |
|  | Chako\_21 | minor tail protein | 1e-114 |
|  | Figueroism\_21 | minor tail protein | 1e-112 |
|  | TatarkaPM\_21 | minor tail protein | 1e-112 |

Similar to vis\_20: hits on phagesDB agree closely with respect to alignment and identity. Genes identified as minor tail proteins are not super different than genes without assigned function

3. BlastP (NCBI) 36 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 21 | function unknown | 4e-141 | 100 | 100 |
| Figueroism | 21 | minor tail protein | 7e-139 | 100 | 98 |

most hits on NCBI blastp are function = hypothetical protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| domain | function | e-val |
| no conserved domains identified |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| tail protein (Phrog) | 99.96 | 1.9e-28 | 49 |
| tail fiber protein (UniProt) a | 92.5 | 0.4 | 25 |

a Uchiyama et al 2011 examine the structure of a related tail fiber protein in PLOS: https://doi.org/10.1317/journal.pone.0026648

HHpred is providing evidence that this is another minor tail protein

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp21 is a soluble protein

conclusion: the function of vis\_21 = minor tail protein: 1) synteny, 2) 4-base overlap with vis\_20, 3) high probability hits for tail protein on HHpred.

Vispistious\_22 pham 743

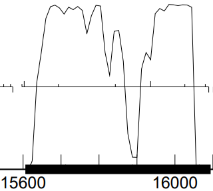
1. auto-annotated start/stop coordinates: 15606 → 16094; forward transcribed

2. length = 489 bp

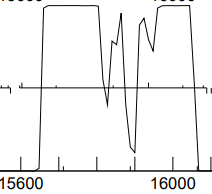
3. start data and coding potential

a. Glimmer @ 15606 has strength = 13.53

b. GeneMarkS shows good coding potential:



GeneMark mfolio shows the same coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream 4 base overlap with vis\_20; upstream 35 bp gap to vis\_22

e. there are at least two competing starts:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 15600 | 2.558 | -4.633 | 495 |
| 15606 | 2.558 | -3.390 | 489 |
| 15696 | 1.589 | -5.487 | 399 |

the auto-annotated start at 15606 has the best rbs score of competing starts.

f. BLAST-start

100 hits

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val |
| 15606 | PuppyEggo, AxiPup and Thorongil | 22 | NCBI | q 1-489  t 1-489 | 100 | 0e-0 |

the start @15606 gives complete alignment with annotated genomes.

g. Starterator: 15606 is the most annotated start.

h. longest possible ORF? nearly

i. conclusion: this is a gene with start at 15606: a) includes all coding potential, b) long ORF, c) best rbs score of competing starts, d) blast alignment with annotated genomes, e) most annotated start

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream gene: gene 20 is annotated as a minor tail protein in some members of this cluster

c. downstream gene: gene 22 is annotated as a minor tail protein in some members of this cluster

d. gene 23 is annotated as a minor tail protein in all members of this cluster

e. ⸫ this is the right spot for a minor tail protein

2. BlastP (phagesDB) many hits with great alignment. The majority = function unknown, but 5 with good alignment and identity are minor tail proteins

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Thorongil\_22 + 2 | function unknown | 1e-90 |
|  | LovelyUnicorn\_22 + 4 | minor tail protein | 3e-82 |

Similar to vis\_21: hits on phagesDB agree closely with respect to alignment and identity. Genes identified as minor tail proteins are not super different than genes without assigned function

3. BlastP (NCBI) 46 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 22 | function unknown | 6e-113 | 100 | 100 |

most hits on NCBI blastp for gp22 are function = hypothetical protein. One hit for a membrane protein (endor\_22), which there is no evidence for

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| domain | function | e-val |
| no conserved domains identified |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| receptor binding protein | 99.27 | 4.4e-10 | 67 |
| receptor binding protein of tail base-plate Siphoviridaea | 96.5 | 0.2 | 67 |
| phrog produces all kinds of high-probability hits, mostly clustered around baseplate protein/attachment protein |  |  |  |

a Sciara et al 2010. Structure of latococcal phage p2 baseplate and its mechanism of activation. Proc Natl Acad Sci 107: 6852-6857

HHpred seems to be zeroing in on a protein that: 1) is located in the tail, 2) is part of the baseplate, and 3) is associated with recognizing/attaching to the host. Not seeing a function of this type described in the official function list

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp22 is a soluble protein

conclusion: the function of vis\_22 is related to the tail. Most likely it functions in recognition and or attachment to the host. On the official function list there is reference to a N-terminal tip binding domain, which is descriptive of this protein, but “tip binding protein” is not an official function. Therefore function = unknown.

Vispistious\_23 pham 741

1. auto-annotated start/stop coordinates: 16130 → 18226; forward transcribed

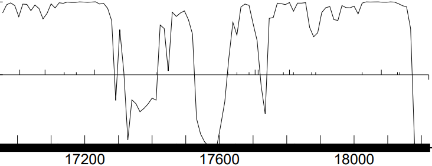
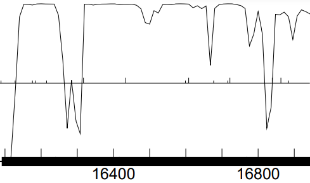
2. length = 2097 bp…one of the larger genes in this genome

3. start data and coding potential

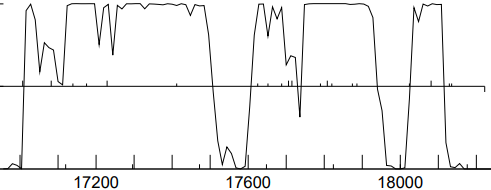
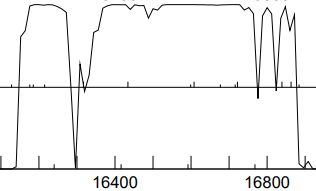
a. Glimmer @ 16130 has strength = 10.02

b. GeneMark calls the start at 16091

c. GeneMarkS shows extensive coding potential:



GeneMark mfolio shows the same coding potential:



d. start sequence = GTG, stop sequence = TAA

e. downstream 35 base gap with vis\_22; upstream 28 bp gap to vis\_24

f. there are at least two competing starts:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 15839 | 2.004 | -4.656 | 2388 |
| 16091 | 2.215 | -4.414 | 2136 |
| 16130 | 1.915 | -4.770 | 2097 |

comparison between the auto-annotated start at 16130 and the Genemark start at 16091: slight edge to the Genemark start

g. BLAST-start

100 hits

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val |
| 16091 | PuppyEggo and AxiPup | 23 | NCBI | q 1-1369  t 1-1369 | 100 | 0e-0 |
| 16130 | PuppyEggo and AxiPup | 23 | NCBI |  | 100 | 0e-0 |

the start @16091 gives more complete alignment with annotated genomes. This start (16091) also establishes a four bp GTGA overlap with vis\_22

h. Starterator: 16091 is the most annotated start.

i. longest possible ORF? yes if start = 16091

j. conclusion: this is a gene with start at 16091: a) includes all coding potential, b) longest ORF, c) best rbs score of competing starts, d) blast alignment with annotated genomes, e) most annotated start, f) establishes a 4 bp overlap with vis\_22

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: when assigned a function, genes 18-22 of this cluster are described as minor tail proteins

c. downstream gene: gene 24 is annotated as lysin A or endolysin in all members of this cluster

d. ⸫ this is the right spot for a minor tail protein

2. BlastP (phagesDB) many hits with great alignment. All non-draft genes are annotated as minor tail protein

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | PuppyEggo\_23 | minor tail protein | 0e-0 |

3. BlastP (NCBI) 100 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 23 | minor tail protein | 0e-0 | 99 | 100 |
| family 18 protein |  | glycosyl hydrolase | 5e-140 | 97 | 39 |

all SEAPHAGE-related hits NCBI blastp for gp23 are function = minor tail protein. The non-SEAPHAGE hits = glycosyl hydrolase

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| 19-223 | glycosyl hydrolase family 18 | 6.2e-25 |

A domain of vis\_23 = glycosyl hydrolase. At least part of Vis\_23 has a lysin-like function

5. HHPred (only for genes > 90% probable matches)

250 hits!

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| glycosyl hydrolase region 2-357; many similar hits in this region | 99.96 | 7.3e-28 | 45 |
| levanase/hydrolase region 357-556 | 99.2 | 6.1e-9 | 21 |
| minor tail protein with cell-wall hydrolyzing domains (phrog) | 100 | 5.4e-91 | 98 |

HHpred seems to be identifying a tail protein with hydrolytic functions; likely helps to breach the cell wall of the host. This protein has been uniformly annotated on phagesDB as function = minor tail protein, but if this protein hydrolyses cell walls, should it not be better characterized as a lysin? Specifically endolysin, glycosyl hydrolase domain? Phrog conveniently matches the entire sequence with great alignment to a minor tail protein.

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp23 is a soluble protein

conclusion: vis\_23 is a minor tail protein with hydrolytic functions.

Vispistious\_24 pham 62106

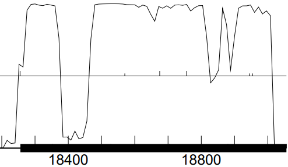
1. auto-annotated start/stop coordinates: 18256 → 19062; forward transcribed

2. length = 807 bp

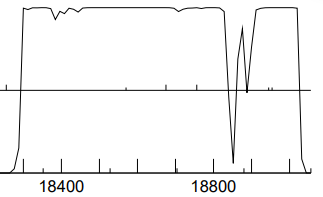
3. start data and coding potential

a. Glimmer @ 18256 has strength = 12.97

b. GeneMarkS shows extensive coding potential:



GeneMark mfolio shows the same coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream 28 base gap with vis\_23; upstream 35 bp gap to vis\_25

e. there is one start that makes sense for this gene:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 18139 | 2.062 | -4.454 | 924 |
| 18256 | 2.970 | -2.505 | 807 |
| 18571 | 1.683 | -5.345 | 492 |

the start at 18139 overlaps massively with vis\_24. The start at 18571 excludes coding potential. Neither have an rbs score that can compare to the auto-annotated start@18256.

f. Starterator: 18256 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 18256: a) includes all coding potential, b) longest ORF, c) best rbs score of competing starts, d) most annotated start

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: when assigned a function, genes 18-22 of this cluster are described as minor tail proteins

c. downstream gene: gene 24 is annotated as lysin A or endolysin in all members of this cluster

d. ⸫ this is the right spot for a minor tail protein

2. BlastP (phagesDB) many hits with great alignment. All non-draft genes are annotated as either lysin A or endolysin

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Thorongil\_24 | lysin A | e-162 |

3. BlastP (NCBI) 100 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 24 | lysin A | 0e-0 | 99 | 100 |
| Ilzat | 24 | endolysin | 5e-140 | 97 | 39 |
| Microbacterium sp XT11 a |  | peptidoglycan-binding protein | 2e-49 | 55 | 43 |

a Briers et al 2007. Muralytic activity and modular structure of the endolysins of *Pseudomonas aeruginosa* bacteriophages phiKZ and EL. Mol Microbiol 65(5): 1334-1344

all SEAPHAGE-related hits NCBI blastp for gp24 are function = lysin A or endolysin. The non-SEAPHAGE hits = peptidoglycan-binding protein…not inconsistent with function = lysin

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| 217-266 | peptidoglycan binding domain | 2.12e-08 |
| 135-266 | peptidoglycan binding domain of peptidoglycan hydrolases | 2.12e-08 |

these binding domains would be expected for a cell wall-dissolving enzyme

5. HHPred (only for genes > 90% probable matches)

58 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| D,D-dipeptidase/D,D carboypeptidase | 99.11 | 1.3e-9 | 38 |
| peptidoglycan-binding protein | 96.6 | 0.19 | 49 |

HHpred is identifying two principal domains: the first is a hydrolytic enzyme that can destroy peptidoglycan, the second is a domain that binds to the peptidoglycan substrate. Together these two domains account for 87% of vis\_24

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp24 is a soluble protein

conclusion: vis\_24 is an endolysin (no lysin B in this genome). However, if vis\_23 is also a lysin, then vis\_23 would be a lysin A and vis\_24 lysin B. Going with function = endolysin.

Vispistious\_25 pham 737

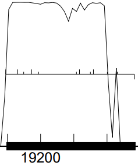
1. auto-annotated start/stop coordinates: 19097 → 19486; forward transcribed

2. length = 390 bp

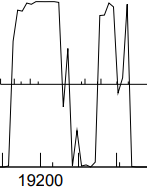
3. start data and coding potential

a. Glimmer @ 19097 has strength = 11.84

b. GeneMarkS shows strong coding potential:



GeneMark mfolio shows more uneven coding potential:



c. start sequence = ATG, stop sequence = TAG

d. downstream 35 base gap with vis\_24; upstream 4 bp AAAC gap to vis\_26

e. there is one start that makes sense for this gene:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 19097 | 2.970 | -2.970 | 390 |
| 19133 | 1.640 | -6.603 | 354 |
| 19139 | 1.175 | -7.123 | 348 |

the auto-annotated start at 19097 is the clear winner with respect to coding potential inclusion and rbs score

f. Starterator: 19097 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 19097: a) includes all coding potential, b) longest ORF, c) best rbs score, d) most annotated start

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: when assigned a function, genes 18-22 of this cluster are described as minor tail proteins; vis\_24 is an endolysin

c. this is the right location for a cell-wall degrading enzyme

2. BlastP (phagesDB) many hits with great alignment. All non-draft genes are annotated as function unknown

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Zada\_25 | function unknown | 3e-67 |

3. BlastP (NCBI) 29 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat | 25 | membrane protein | 4e-88 | 99 | 100 |
| Figueroism | 25 | membrane protein | 1e-86 | 99 | 98.45 |
| Bonno | 25 | membrane protein | 3e-86 | 99 | 98.45 |

interesting: of the 29 hits on NCBI blastp, all but two are membrane proteins

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred (only for genes > 90% probable matches)

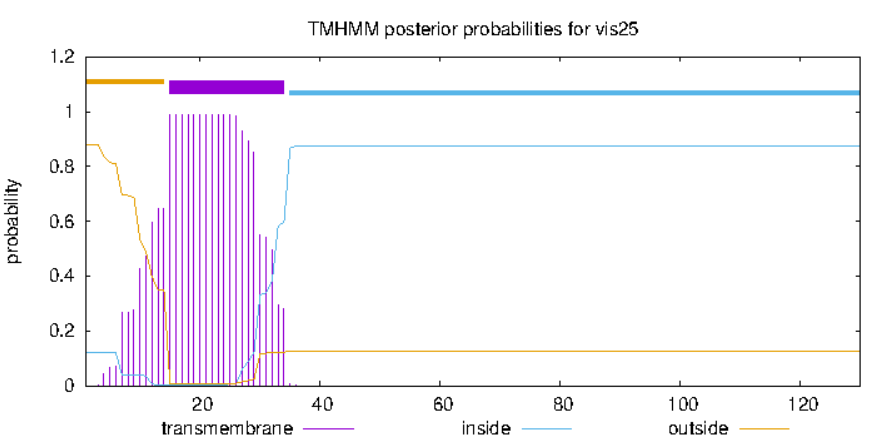
58 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| gene 30 Mycobacterium phage D29 | 96 | 0.089 | 55 |
| probable spanin, inner membrane subunit | 95.6 | 1.1 | 87 |
| phage shock protein | 95.4 | 0.68 | 55 |

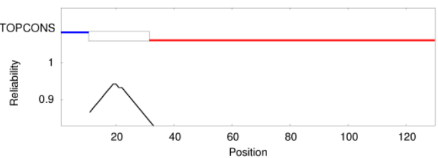
Spanin is an interesting function for this gene, fitting well in the neighborhood of lytic warriors

6. Transmembrane finders

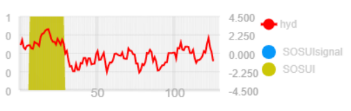
a. deep TMHMM: residues 15-34 are a transmembrane helix



b. TOPCONs: residues 11-31 are a transmembrane domain



c. SOSUI: gp25 is a membrane protein



conclusion: vis\_25 is a membrane protein. More specifically, it may be a spanin.

Vispistious\_26 pham 68719

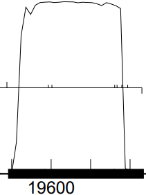
1. auto-annotated start/stop coordinates: 19491 → 19835; forward transcribed

2. length = 345 bp

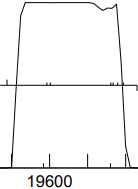
3. start data and coding potential

a. Glimmer @ 19491 has strength = 18.86

b. GeneMarkS shows strong coding potential:



GeneMark mfolio shows the same coding potential:



c. start sequence = ATG, stop sequence = TAA

d. downstream 4 bp AAAC gap with vis\_25; upstream 58 bp gap to vis\_27

e. there is one start that makes sense for this gene:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 19374 | 2.098 | -5.620 | 462 |
| 19491 | 2.083 | -4.874 | 345 |
| 19596 | -0.09 | -9.834 | 240 |

the auto-annotated start at 19491 avoids a large overlap with vis\_25, has the best rbs score of potential starts, and includes all of the coding potential.

f. Starterator: 19491 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 19491: a) includes all coding potential, b) longest ORF, c) best rbs score, d) most annotated start

Gene function

1. Synteny

a. is synteny informative for this gene? yes

b. upstream genes: when assigned a function, genes 18-22 of this cluster are described as minor tail proteins; vis\_24 is an endolysin and vis\_25 may be a spanin. These genes are involved with disrupting the host cell wall

c. this is the right location for a cell-wall degrading enzyme

2. BlastP (phagesDB) many hits with great alignment. All non-draft genes are annotated as holins

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | PuppyEggo\_26 | holin | 5e-55 |

3. BlastP (NCBI) 100 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Puppy Eggo | 26 | holin | 4e-69 | 100 | 100 |
| AxiPup | 26 | holin | 2e-68 | 100 | 99.12 |
| Robinson | 26 | holin | 9e-68 | 100 | 98.25 |
| Caudovirales a | residues 1-92 | holin | 2e-12 | 60 | 43 |

a Tsza and Buck 2021. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic disease. Proc Natl acad Sci 118(23)

nearly all hits on NCBI blastp are holins, with a few annotated as membrane proteins

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred (only for genes > 90% probable matches)

58 hits

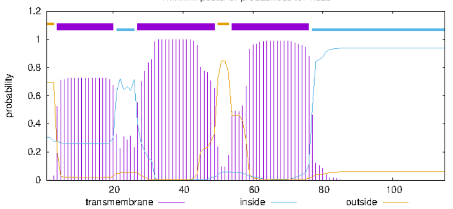
|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| phage holin a+ | 99.7 | 7e-16 | 70 |
| holin b | 99.7 | 2.e-15 | 70 |
| phage holin | 99.5 | 2.4e-13 | 57 |

a Grunding et al 2001. Holins kill without warning. Proc Natl Acad Sci 98: 9348-52

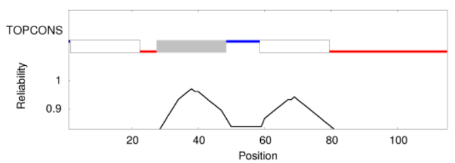
b Wendlinger and Scherer 1995. Mol Microbiol 16: 1231-1241

6. Transmembrane finders

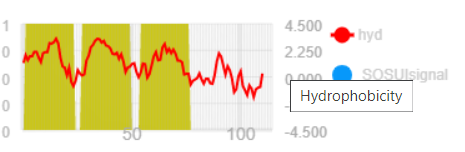
a. deep TMHMM: residues 4-20, 27-49 and 54-76 are transmembrane helices



b. TOPCONs: residues 2-22, 28-48 and 58-79 are a transmembrane domains



c. SOSUI: gp26 is a membrane protein with 3 hydrophobic domains



conclusion: vis\_26 is a holin

Vispistious\_27 pham 67263

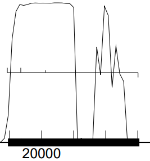
1. auto-annotated start/stop coordinates: 19895 → 20305; forward transcribed…the last forward transcribed gene before the direction of transcription reverses

2. length = 411 bp

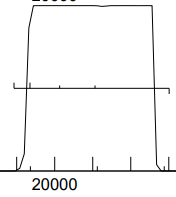
3. start data and coding potential

a. Glimmer @ 19895 has strength = 12.34

b. GeneMarkS shows two strong subregions of coding potential:



GeneMark mfolio shows more uniform coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream 58 bp gap with vis\_26; upstream 67 bp gap to reverse-transcribed vis\_28

e. there is one start that makes sense for this gene:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 19895 | 2.083 | -4.397 | 411 |
| 19937 | 0.783 | -7.196 | 369 |
| 20015 | 1.683 | -5.005 | 291 |

the auto-annotated start at 19895 the best rbs score of potential starts, and includes all of the coding potential.

f. Starterator: 19895 is the most annotated start.

g. longest possible ORF? yes

h. conclusion: this is a gene with start at 19895: a) includes all coding potential, b) longest ORF, c) best rbs score, d) most annotated start

Gene function

1. Synteny

a. is synteny informative for this gene? maybe

b. upstream genes: when assigned a function, genes 18-22 of this cluster are described as minor tail proteins; vis\_24 is an endolysin, vis\_25 may be a spanin, and vis\_26 is a holin. These genes are involved with disrupting the host cell wall

c. this gene may also have a function associated with host lysis

2. BlastP (phagesDB) many hits with great alignment. The other gp27’s in this cluster are annotated as function unknown. There are lower likelihood hits on phagesDB that suggest that gp27 may be a portal protein or a DNA primase

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Thorongil\_27 | function unknown | 4e-79 |
|  | Rachaly\_16 | portal protein | 1.5 |
|  | Zenon\_93 | DNA primase/polymerase | 5.6 |

3. BlastP (NCBI) 10 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 27 | hypothetical protein | 4e-94 | 100 | 100 |
| Teagan | 26 | hypothetical protein | 5e-93 | 100 | 98.5 |

relatively few hits hits on NCBI blastp; all hypothetical protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| gene region | function | e-val |
| no CDD hits for this gene |  |  |

5. HHPred (only for genes > 90% probable matches)

58 hits

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| no hits ≥ *p* = 90% |  |  |  |

6. Transmembrane finders

a. deep TMHMM: no transmembrane helices detected

b. TOPCONs: no homologous transmembrane domains detected

c. SOSUI: gp26 is a soluble protein

conclusion: vis\_27 = function unknown

Vispistious gp28 pham 686

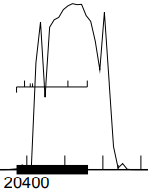
1. auto-annotated start/stop coordinates: 20373 ← 20561; reverse transcribed

2. length = 189 bp

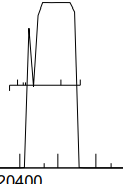
3. start data and coding potential

a. Glimmer @ 38456 has strength = 18.88

b. GeneMark shows strong coding potential that extends beyond the right end of the auto-annotated genecoding potential:



GeneMark mfolio shows more restricted coding potential that stays within the limits of the auto-annotation::



d. start sequence = ATG, stop sequence = TAG

e. 67 bp gap to forward-transcribed gp27; 70 bp gap to gp29

f. ribosomal binding site (scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 20561 | 2.970 | -2.443 | #1/5 |

20561 is the strongest start available in this region of coding potential

g. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 20561 | PuppyEggo | 28 | NCBI | q 1-189  t 1-189 | 100 | 6e-92 | 100 |

g. Starterator: 20561 is the most annotated start

h. longest possible ORF? yes

i. conclusion: this is a gene with start at 20561: a) strong coding potential, b) good alignment scores on NCBI blast, c) good rbs score, d) longest ORF.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Zada\_28 | function unknown | 3e-30 |

3. BlastP (NCBI) 37 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat | 27 | unknown | 1e-36 | 100 | 100 |

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| no CDD hits |  |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| description | domain | *p* | e-val | % align |
| no hits > 90% *p* |  |  |  |  |

4. Transmembrane finders

a. deep TMHMM: gp28 is a globular non-membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp28 is a soluble protein

conclusion: the function of vis\_28 is unknown

Vispistious gp29 pham 858

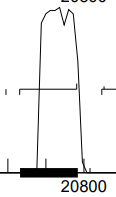
1. auto-annotated start/stop coordinates: 20632 ← 20784; reverse transcribed

2. length = 153 bp

3. start data and coding potential

a. Glimmer @ 20784 has strength = 14.76

b. GeneMark shows a small region of coding potential:



GeneMark mfolio shows more restricted coding potential that stays within the limits of the auto-annotation::



c. start sequence = ATG, stop sequence = TGA

d. upstream 70 bp gap to gp28; 14 bp downstream gap to gp30

e. 2 possible starts: 20784 and 20802

ribosomal binding scores (scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 20784 | 1.849 | -5.501 | #1/2 |
| 20802 | 1.510 | -7.181 | #2/2 |

20784 has a weak RBS score. @20802 is worse

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 20784 | Lovelyunicorn | 29 | NCBI | q 1-153  t 1-153 | 100 | 2e-70 | 99.35 |
| 20802 | Lovelyunicorn | 29 | NCBI | q 1-171  t 1-171 | 100 | 3e-80 | 99.42 |

the start at 20802 yields more complete alignment

g. Starterator: 20802 is the most annotated start

h. longest possible ORF? yes if the start is at 20802

i. conclusion: this is a gene with start at 20802: a) within region of coding potential, b) improved alignment scores on NCBI blast, c) both starts have weak rbs score, d) adding 18 bp would create a 4-base TTGA overlap with gp30…not an ATGA overlap, but the next best thing and could explain the weak rbs score, e) longest ORF.

⸫ change start from 20784 to 20802

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Zada\_29 | function unknown | 5e-27 |

3. BlastP (NCBI) 32 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 29 | unknown | 3e-33 | 100 | 98.21 |

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| no CDD hits |  |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| description | domain | *p* | e-val | % align |
| no hits > 90% *p* |  |  |  |  |

6. Transmembrane finders

a. deep TMHMM: gp29 is a globular non-membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp29 is a soluble protein

conclusion: the function of vis\_29 is unknown

Vispistious gp30 pham 909

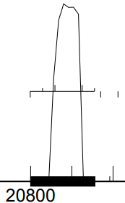
1. auto-annotated start/stop coordinates: 20799 ← 20957; reverse transcribed

2. length = 159 bp

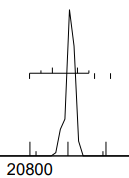
3. start data and coding potential

a. Glimmer @ 20957 has strength = 7.44

b. GeneMark shows a small region of coding potential:



GeneMark mfolio shows restricted coding:



c. start sequence = GTG, stop sequence = TGA

d. upstream 4 bp TTGA overlap with gp29; 4 bp GTGA overlap with gp31…operon

e. starts are looking straight-forward: the auto-annotated start has the best score and also yields the longest ORF

ribosomal binding scores (scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 20957 | 2.514 | -3.564 | #1/5 |
| 20927 | 0.889 | -6.908 | #3/5 |
| 20861 | 1.049 | -7.053 | #5/5 |

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 20957 | Lovelyunicorn | 30 | NCBI | q 1-159  t 1-159 | 100 | 2e-75 | 100 |

g. Starterator: 20957 is the most annotated start

h. longest possible ORF? yes if the start is at 20957

i. conclusion: this is a gene with start at 20957: a) within region of coding potential, b) best rbs score, c) 4-base GTGA overlap with gp31, d) longest ORF.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut | function unknown | 2e-22 |

3. BlastP (NCBI) 32 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| PuppyEggo | 30 | unknown | 3e-28 | 100 | 100 |

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| no CDD hits |  |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| description | domain | *p* | e-val | % align |
| 93% similarity to protein AF1548featuring alpha-hairpins that interact with DNA and proteins; yet exact function unknown |  |  |  |  |

6. Transmembrane finders

a. deep TMHMM: gp30 is a globular non-membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp30 is a soluble protein

conclusions: function unknown

Vispistious gp31 pham 67104

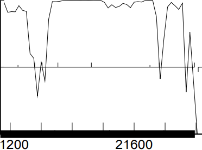
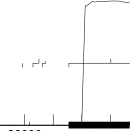
1. auto-annotated start/stop coordinates: 20954 ← 21976; reverse transcribed

2. length = 843 bp

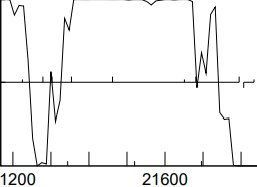
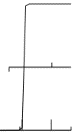
3. start data and coding potential

a. Glimmer @ 21796 has strength = 16.96

b. GeneMark shows a large region of coding potential:



GeneMark mfolio shows a wide region of coding potential:



c. start sequence = ATG, stop sequence = TGA

d. upstream 4 bp GTGA overlap with gp29; 44 bp GTGA gap to gp32

e. none of the starts have a strong rbs score, but only one (21796) captures the ample coding potential. Weak rbs scores seem to be a thing with gp 29/30/31

ribosomal binding scores (scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 21796 | 1.503 | -6.242 | #6/9 |
| 21682 | 0.882 | -6.986 | #8/9 |
| 21652 | 1.049 | -6.627 | #5/9 |

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 21796 | PuppyEggo | 31 | NCBI | q 1-843  t 1-843 | 100 | 0e-0 | 100 |

g. Starterator: 21976 is the most annotated start

h. longest possible ORF? yes if the start is at 21976

i. conclusion: this is a gene with start at 21976: a) strong coding potential, b) most annotated start, c) longest ORF, d) excellent alignment with other members of the pham.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut\_31 | function unknown | 1e-162 |

3. BlastP (NCBI) 32 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| GardenState | 30 | ssDNA binding protein | 1e-08 | 96 | 27 |
| AmGroot |  | ssDNA binding protein | 2e-08 | 94 | 27 |

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| no CDD hits |  |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| description | domain | *p* | e-val | % align |
| transcription antitermination complex | 25-52 | 94% |  |  |
| antitermination protein | 24-66 | 94% |  |  |
| tryptophan RNA binding attenuation protein | 21-70 | 94% |  |  |

6. Transmembrane finders

a. deep TMHMM: gp31 is a globular non-membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp31 is a soluble protein

conclusion: function unknown. It is possible that 1) domains of gp31 bind to either DNA or RNA and 2) gp31 may help to control transcription

Vispistious gp32 pham 1612

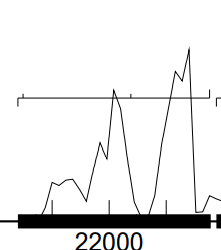
1. auto-annotated start/stop coordinates: 21840 ← 22178; reverse transcribed

2. length = 339 bp

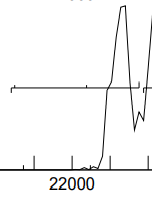
3. start data and coding potential

a. Glimmer @ 22178 has strength = 3.84

b. GeneMarkS shows irregular coding potential:



GeneMark mfolio shows relatively weak coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream: 44 bp GTGA gap to gp31; upstream gap 9 bp GGAATCATC

e. One start (22178) includes the coding potential shown by GeneMarks

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 22178 | 3.115 | -2.214 | #1/3 |
| 22040 | 0.578 | -7.925 | #3/3 |
| 21851 | 1.769 | -6.121 | #2/3 |

22178 is looking like the obvious start

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 22178 | PuppyEggo | 32 | NCBI | q 1-339  t 1-339 | 100 | 5e-175 | 100 |

g. Starterator: 22178 is the most annotated start

h. longest possible ORF? yes if the start is at 22178

i. conclusion: this is a gene with start at 22178: a) enough coding potential, b) most annotated start, c) longest ORF, d) excellent alignment with other members of the pham, e) strong rbs score

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 60 | PuppyEggo\_32 | function unknown | 3e-64 |

hits on phages DB with the best agreement are all function unknown; some of the lesser matches with poor e-values list function as lesser tail protein.

3. BlastP (NCBI) 32 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| WildNOut | 32 | membrane protein | 8e-74 | 99 | 99.11 |
| Figueroism | 32 | membrane | 1e-72 | 99 | 96.43 |
| Microbacterium testaceum |  | LysM peptidoglycan-binding domain protein | 6e-20 | 59 | 65 |
| Microbacterium BH-3-3-3 |  | LysM peptidoglycan-binding domain protein | 6e-19 | 91 | 45 |
| Microbacterium sp. P-26 |  | LysM peptidoglycan-binding domain protein | 3e-17 | 46 | 69 |
| Microbacterium proteolyticum |  | LysM peptidoglycan-binding domain protein | 2e-16 | 46 | 69 |
| Microbacterium sp Leaf151 |  | LysM peptidoglycan-binding domain protein | 2e-16 | 46 | 69 |

gp32 is a membrane protein. Specifically it may be a LysM peptidoglycan-binding domain protein, or at least have a shared domain with a protein of this type.

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| no CDD hits |  |  |  |

5. HHPred (only for genes > 90% probable matches)

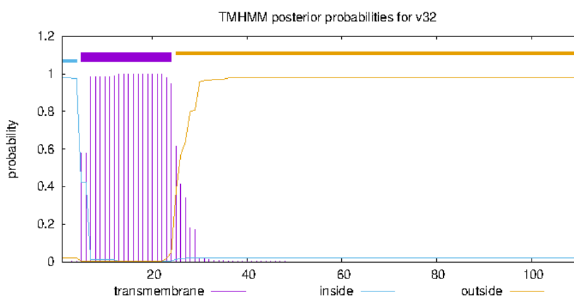
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| description | domain | *p* | e-val | % align | citation |
| inner membrane lipoprotein YiaD |  | 94% | 0.54 | 42 | Sheedlo et al. 2021 |
| Putative B-type lectin protein |  | 92% | 2.5 | 65 |  |
| Putative PAN domain-containing protein |  | 92% | 1.3 | 45 |  |

Sheedlo et al 2021. Cryo-EM reveals new species-specific proteins and symmetry elements in the *Legionella pneumophila* Dot/Icm T4SS.

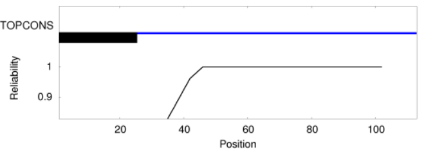
HHPred supports the idea that gp32 is a membrane protein

4. Transmembrane finders

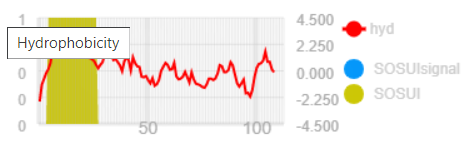
a. deep TMHMM: residues 5-24 are a transmembrane helix



b. TOPCONs: residues 1-25 are a transmembrane helix



c. SOSUI: gp32 is a membrane protein



conclusion: gp32 is a membrane protein. Parts of gp32 have strong resemblance to a LysM peptidoglycan-binding domain

Vispistious gp33 pham 931

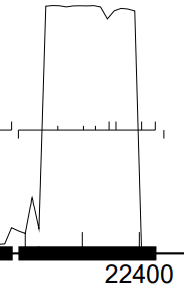
1. auto-annotated start/stop coordinates: 22188 ← 22430; reverse transcribed

2. length = 243 bp

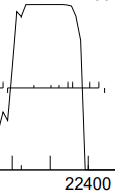
3. start data and coding potential

a. Glimmer @ 22430 has strength = 16.51

b. GeneMarkS shows strong uniform coding potential:



GeneMark mfolio shows strong coding potential:



c. start sequence = ATG, stop sequence = TAA

d. downstream: 9 bp GGAATCATC gap to gp31; upstream 4 bp overlap ATGA with bp33…operon

e. The auto-annotated start (22430) includes all of the coding potential shown by GeneMark and also has the best rbs score:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 22430 | 3.059 | -2.394 | #1/7 |

22430 is looking like the start

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 22430 | PuppyEggo | 33 | NCBI | q 1-243  t 1-243 | 100 | 8e-122 | 100 |

g. Starterator: 22430 is the most annotated start

h. longest possible ORF? yes if the start is at 22430

i. conclusion: this is a gene with start at 22430: a) strong, coding potential, b) most annotated start, c) longest ORF, d) excellent alignment with other members of the pham, e) strong rbs score, f) 4 bp apparent operon overlap with gp34

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut\_33 | function unknown | 3e-41 |

3. BlastP (NCBI) 32 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 33 | membrane protein | 2e-49 | 98 | 100 |
| Teagan | 33 | membrane protein | 3e-49 | 98 | 98.75 |
| Greys | 33 | membrane protein | 3e-49 | 98 | 98.75 |

good evidence that gp33 is a membrane protein.

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| no CDD hits |  |  |  |

5. HHPred (only for genes > 90% probable matches)

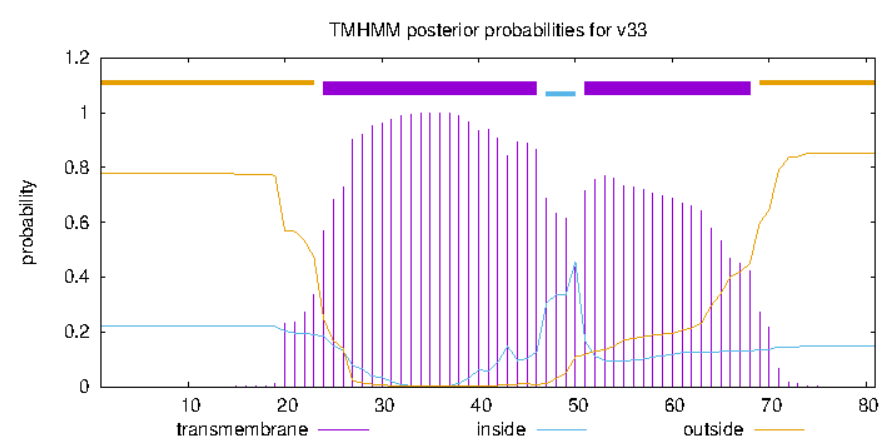
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| description | domain | *p* | e-val | % align | citation |
| trans membrane helical complex |  | 85% | 0.54 | 27 | Jean-Francois et al 2014. |

#### Jean-Francois et al 2014. Binding of MgtR, a Salmonella Transmembrane Regulatory Peptide, to MgtC, a Mycobacterium tuberculosis Virulence Factor: A Structural Study.  J Mol Biol ****426****: 436-446

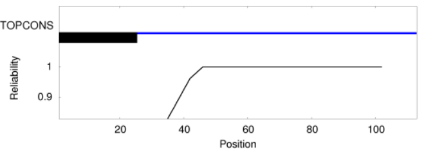
HHPred supports the idea that gp33 is a membrane protein

4. Transmembrane finders

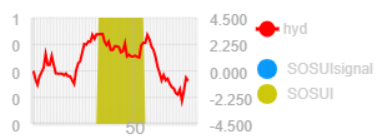
a. deep TMHMM: residues 24-46 and 51-68 are transmembrane helices



b. TOPCONs: residues 27-47 are a transmembrane helix



c. SOSUI: gp33 is a membrane protein



conclusion: gp33 is a membrane protein.

Vispistious gp34 pham 56382

1. auto-annotated start/stop coordinates: 22427 ← 24028; reverse transcribed

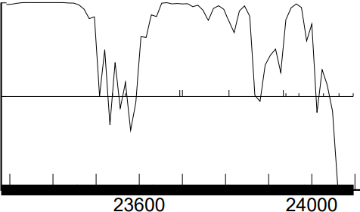
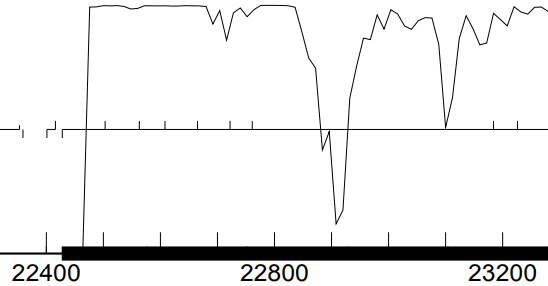
2. length = 1602 bp

3. start data and coding potential

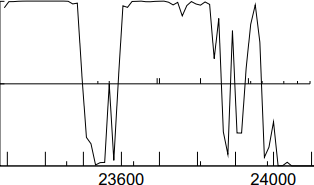
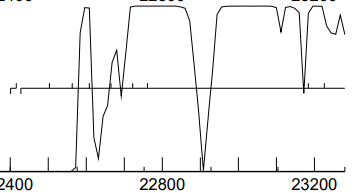
a. Glimmer @ 24028 has strength = 11.76

b. GeneMarkS calls the start at 24097

c. GeneMarkS shows extensive coding potential:



GeneMark mfolio shows strong coding potential:



d. start sequence = GTG, stop sequence = TGA

e. upstream: 4 bp ATGA overlap with gp33; downstream 42 bp gap with bp35…operon

f. The auto-annotated start (22430) includes all of the coding potential shown by GeneMark and also has the best rbs score:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 24097 | 3.041 | -2.433 | #1/22 |
| 24064 | 0.873 | -7.003 | #20/22 |
| 24028 | 1.889 | -4.804 | #6/22 |

24097 has a substantially stronger rbs score

g. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 24028 | PuppyEggo | 34 | NCBI | q 1-1394  t 1-1394 | 100 | 0e-0 | 100 |
| 24097 | PuppyEggo | 34 | NCBI | q 1-1671  t 1-1671 | 100 | 0e-0 | 100 |

h. Starterator: 24097 is the most annotated start

i. longest possible ORF? yes if the start is at 24097

j. conclusion: this is a gene with start at 24097: a) strong, coding potential, b) most annotated start, c) longest ORF, d) excellent alignment with other members of the pham, e) best rbs score

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut\_34 | RepA-like helicase | 0e-0 |
|  | PuppyEggo | DNA recombinase | 0e-0 |
|  | AxiPup | DNA recombinase | 0e-0 |
|  | TinSulphur | Rec-A like DNA recombinase | 0e-0 |

gp34 is identical to these proteins…check function list for correct name

3. BlastP (NCBI) 32 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 33 | membrane protein | 2e-49 | 98 | 100 |
| Teagan | 33 | membrane protein | 3e-49 | 98 | 98.75 |
| Greys | 33 | membrane protein | 3e-49 | 98 | 98.75 |

good evidence that gp33 is a membrane protein.

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| Helicase RepA | couples the hydrolysis of NTPs to the unwinding of DNA | 319-520 | 2.13e-27 |
| RepB DNA primase | recognize specific DNA sequences and synthesize primers | 99-195 | 3.3e-05 |

Geibel et al 2009. Structure and function of primase RepB' encoded by broad-host-range plasmid RSF1010 that replicates exclusively in leading-strand mode. Proc Natl Acad Sci 106(19): 7810-5.

* gp34 has 1) a helicase component, and 2) a primase component. This makes it a DNA primase/helicase. Does it have a polymerase function as well?

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| description | residues | *p* | e-val | % align | citation |
| RepB DNA primase | 3-293 | 99.7% | 9.6e-15 | 75 | Wang et al 2008 |
| RepA DNA helicase | 290-546 | 99.9% | 6.8e-19 | 87 | Xu et al 2003 |

# Wang et al. 2008. The structure of a DnaB-family replicative helicase and its interactions with primase. [Nature Structural & Molecular Biology](https://www.nature.com/nsmb) 15**:** 94–100

### Xu et al. 2003. [Structure of DNA helicase RepA in complex with sulfate at 1.95 Å resolution implicates structural changes to an `open' form](https://journals.iucr.org/d/issues/2003/05/00/wd0003/index.html). [Acta Cryst*.*](https://journals.iucr.org/d) (2003). [D59](https://journals.iucr.org/d/contents/backissues.html), 815-822

HHPred identifies two domains in gp34: a primase and a helicase. gp34 does not have a polymerase ⸫ function = DNA primase/helicase

the consensus from phagesDB seems to be that gp34 is a RecA-like DNA recombinase. Spud\_205 is the example of this protein in the official function list. Schubert\_31 is the example of a DNA primase/helicase…although its annotated function = DNA recombinase, which is not an option? Compare gp34 with these two other protein sequences using the global alignment tool:

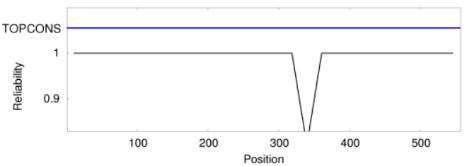
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| phage | gene# | pham | function ex official function list | n residues | %align | % identity |
| Spud | 205 | 103 | RecA-like DNA recombinase | 362 | 26 | 15 |
| Schubert | 31 | 56382 | DNA primase/helicase | 557 | 80 | 60 |

gp34 is not a RecA-like DNA recombinase

6. Transmembrane finders

a. deep TMHMM: gp34 is not a membrane protein

b. TOPCONs: SPOCTOPUS predicts that residues 330-350 create a trans-membrane helix



c. SOSUI: gp34 is a soluble protein

conclusions: 1) gp34 is a DNA primase/helicase, 2) a small portion of this protein may embed in a membrane

Vispistious gp35 pham 65611

1. auto-annotated start/stop coordinates: 24073 ← 24339; reverse transcribed

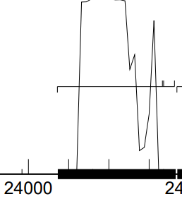
2. length = 267 bp

3. start data and coding potential

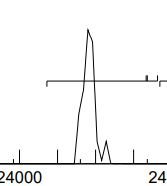
a. Glimmer @ 22430 has strength = 3.56

b. GeneMark calls start at 24366

c. GeneMarkS shows good coding potential:



GeneMark mfolio shows mediocre coding potential:



d. start sequence = GTG, stop sequence = TGA

e. downstream: 29 bp gap with gp36; upstream 25 bp overlap with gp34

f. There are three potential starts; none have good rbs scores:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 24366 | 1.160 | -6.678 | #2/3 |
| 24339 | 1.481 | -6.942 | #3/3 |
| 24336 | 1.481 | -6.164 | #1/3 |

g. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val | % id |
| 24339 | ParleG | 35 | NCBI | q 1-267  t 1-267 | 100 | 4e-135 | 100 |
| 24366 | ParleG | 35 | NCBI | q 1-294  t 1-294 | 100 | 4e-150 | 100 |

the start @24366 produces more complete alignment with annotated genomes

h. Starterator: 24366 is the most annotated start

i. longest possible ORF? yes if the start is at 24366

j. conclusion: this is a gene with start at 24366: a) strong coding potential, b) most annotated start, c) longest ORF, d) alignment with other members of the pham.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene: NA

c. downstream gene: NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Zada\_35 | Holliday junction resolvase | 1e-51 |
|  | Winzigespinne\_34 | nuclease | 1e-51 |
|  | WildNOut\_35 | nuclease | 1e-51 |
|  | Velene\_35 | nuclease | 1e-51 |

The consensus from phagesDB is that gp35 is a nuclease…maybe a Holliday junction resolvase?

3. BlastP (NCBI) 32 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat | 35 | endonuclease | 1e-65 | 98 | 100 |
| PhredFlintston | 35 | nuclease | 6e-65 | 98 | 98.97 |
| McGalleon | 35 | endonuclease | 6e-42 | 98 | 94.85 |

good evidence that gp35 is a nuclease.

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| no CDD hits |  |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| description | domain | *p* | e-val | % align | citation |
| nuclease |  | 99.9 | 1.7e-19 | 96 | Pennell et al 2014 |

#### Pennell et al 2014. FAN1 activity on asymmetric repair intermediates is mediated by an atypical monomeric virus-type replication-repair nuclease domain. Cell Reports…upshot is that this nuclease is structurally and functionally similar to a Holliday junction resolvase, but is not an HJR

#### 

#### above left: top HHPred hit hydrolase(nuclease); above right: Holliday junction resolvase…the two molecules are related but have different structures. gp35 is not a Holliday junction resolvase.

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp35 is a soluble protein

conclusion: gp35 is a nuclease with structural and functional similarity to a Holliday junction resolvase.

Vispistious gp36 pham 68708

1. auto-annotated start/stop coordinates: 24370 ← 25212; reverse transcribed

2. length = 843 bp

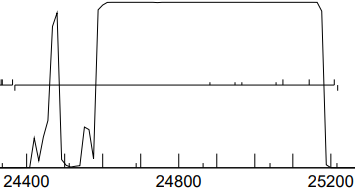
3. start data and coding potential

a. Glimmer @ 25212 has strength = 17.13

b. GeneMarkS shows strong coding potential:



GeneMark mfolio also shows strong coding potential:



c. start sequence = ATG, stop sequence = TGA

d. downstream: 32 bp gap with gp37; upstream 3 bp gap with gp35

e. There are many potential starts; of these, the auto-annotated start@25212 combines the longest ORF with a good rbs score:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 25212 | 2.677 | -3.212 | #1/7 |
| 25146 | 2.390 | -4.341 | #2/7 |
| 25077 | 1.350 | -5.998 | #5/7 |

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val | % id |
| 24339 | PuppyEggo | 35 | NCBI | q 1-843  t 1-843 | 100 | 0e-0 | 100 |

the start @25212 produces excellent alignment with annotated genomes

g. Starterator: 25212 is the most annotated start

h. longest possible ORF? yes if the start is at 25212

i. conclusion: this is a gene with start at 25212: a) strong coding potential, b) most annotated start, c) longest ORF, d) alignment with other members of the pham, e) best rbs score

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Thorongil\_36 | DNA polymerase | 1e-162 |
|  | Thompsone\_36 + 3 others with identical score | function unknown | 1e-162 |

The best hit on phagesDB (Thorongil\_36) is a 100% match for gp\_36. Thorongil\_36 is a DNA polymerase(?). If the function of Thorongil\_36 was correctly determined, then gp36 should also be a DNA polymerase. There are no other function matches for gp\_36 on phagesDB.

3. BlastP (NCBI) 87 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| PuppyEggo | 36 | hypothetical protein | 0e-0 | 100 | 100 |
| Gargoyle | 36 | hypothetical protein | 0e-0 | 100 | 99.3 |
| Figueroism | 36 | hypothetical protein | 0e-0 | 98 | 98.9 |
| Mycobacterium phage FF47 |  | DNA polymerase | 3e-16 | 87 | 40.2 |
| Mycobacterium phage Muddy |  | DNA polymerase | 1e-15 | 85.5 | 41.6 |

Many very similar proteins with unknown functions. Two DNA polymerase matches partially align with gp\_36.

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| no CDD hits |  |  |  |

If gp36 is a DNA polymerase, then at least portions of the gene should find CDD matches. The lack of a match here does not support the DNA polymerase hypothesis.

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| predicted domain | *p* | e-val | % align | % identities |
| LEM/SAP HeH motif | 100 | 1.1e-63 | 99 | 99 |
| methyl transferase | 99.9 | 4.9e-23 | 40 | 13 |
| DNA polymerase | 99.9 | 2e-22 | 59 | 32 |
| Antigen MPT63/MPB63 | 99.9 | 5e-22 | 80 | 23 |
| c-terminal domain of DNA repair helicase | 99.8 | 4.8e-20 | 43 | 17 |
| single-strand DNA binding protein | 99.8 | 2.5e-17 | 44 | 14 |

#### The standard HHpred databases used by SEAPHAGES do not work with gp36…error messages. However, PHROG and COGs produce hits without consensus. gp36 may be a protein that interacts in some way with DNA.

4. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp36 is a soluble protein

conclusion: there is not sufficient evidence to designate the function of gp36 as a DNA polymerase. Not sure what evidence was used to make that call for Thorongil\_36. Possible that gp36 interacts with DNA but can go no farther than that. Function unknown.

Vispistious gp37 pham 52746

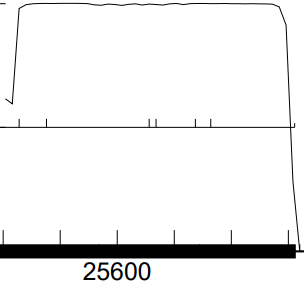
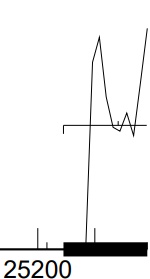
1. auto-annotated start/stop coordinates: 25245 ← 25913; reverse transcribed

2. length = 669 bp

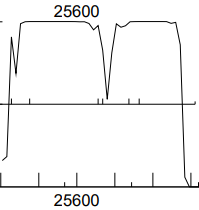
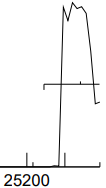
3. start data and coding potential

a. Glimmer @ 25913 has strength = 17.93

b. GeneMarkS shows strong coding potential:



GeneMark mfolio shows strong coding potential:



c. start sequence = GTG, stop sequence = TAG

d. downstream: 4 bp GTGA overlap with gp38; upstream 32 bp gap with gp36

e. There are many potential starts; among them, the auto-annotated start has a decent rbs score and includes all of the coding potential:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 25913 | 2.098 | -4.456 | #2/10 |
| 25766 | 2.726 | -3.796 | #1/10 |
| 25739 | 1.687 | -6.502 | #6/10 |

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val | % id |
| 25913 | PuppyEggo + 3 | 37 | NCBI | q 1-669  t 1-669 | 100 | 0e-0 | 100 |

the start @25913 produces complete alignment with annotated genomes

g. Starterator: 25913 is the most annotated start

h. longest possible ORF? yes if the start is at 25913

i. conclusion: this is a gene with start at 25913: a) includes all coding potential, b) most annotated start, c) longest ORF, d) alignment with other members of the pham, e) good rbs score

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Zada\_37 | AAA-ATPase | e-123 |
|  | Winzigespinne\_37 | AAA-ATPase | e-123 |
|  | WildNOut\_37 | RecA-like DNA recombinase | e-123 |

The consensus from phagesDB is that gp37 is a AAA-ATPase. WildNOut\_37 and BigRedClifford\_37 go farther, both suggesting that gp37 is a RecA-like DNA recombinase.

3. BlastP (NCBI) 32 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat | 37 | AAA-ATPase | 8e-162 | 100 | 100 |
| Knox | 37 | AAA-ATPase | 1e-161 | 100 | 99.55 |
| KannH | 37 | AAA-ATPase | 4e-161 | 100 | 99.55 |
| Peep | 37 | AAA-ATPase | 6e-161 | 100 | 99.55 |
| AlexAdler | 37 | AAA-ATPase | 8e-160 | 100 | 99.1 |
| MonChoix | 37 | AAA-ATPase | 6e-155 | 100 | 95.05 |
| McGalleon | 37 | AAA-ATPase | 5e-148 | 100 | 90.09 |

Consensus function for gp37 on NCBI blastp is AAA-ATPase. These hits are all direct submissions…no links to literature

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| CRISPR/CAS system-associated Cas3 |  | 3-199 | 1.56e-05 |

This is interesting. 88% of gp37 aligns with a CRISPR system protein. The actual hit was with pfam 13479 = AAA domain. Members of this pfam = AAA family ATPase.

5. HHPred (only for genes > 90% probable matches)

Too many hits with probability > 90% to list them all. Top hits with best alignments:

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| pfam 13479.9…AAA domain | 99.2 | 2e-9 | 79 |
| RecA ATP ase DNA binding protein | 99 | 5.7e-8 | 74 |
| RecA ATP ase DNA binding protein | 99 | 6.3e-8 | 74 |

#### Consensus from HHpred = similar to a recA-like ATPase. Nests within family of AAA ATPase proteins. However, RecA-like ATPases (like Spud\_205) trend around 350 amino acids in length. They do have the ring structure of an AAA ATPase.

4. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp37 is a soluble protein

conclusion: gp37 is an AAA ATPase. It likely combines with DNA to function in repair/recombination. There is not sufficient evidence to label this protein as a RecA-like ATPase.

Vispistious gp38 pham 65541

1. auto-annotated start/stop coordinates: 25910 ← 27085; reverse transcribed

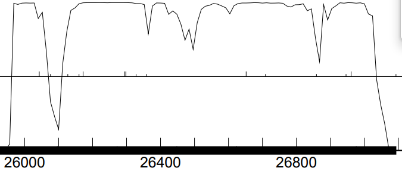
2. length = 1176 bp

3. start data and coding potential

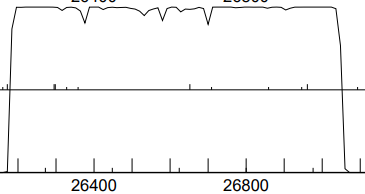
a. Glimmer @ 27085 has strength = 11.16

b. GeneMark calls the start at 27094

c. GeneMarkS shows strong coding potential:



GeneMark mfolio also shows strong coding potential:



d. start sequence = TTG, stop sequence = TGA

e. upstream: 4 bp GTGA overlap with gp37; downstream 14 bp overlap with gp39

f. There are many potential starts:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 27205 | 1.000 | -7.019 | #18/19 |
| 27154 | 1.738 | -5.148 | #5/19 |
| 27094 | 1.860 | -4.966 | #3/19 |
| 27085 | 2.398 | -3.671 | #1/19 |

the auto-annotated start @2708 has the best rbs score; genemark’s rbs score is weaker

g. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val | % id |
| 27085 | PuppyEggo + AxiPup | 37 | NCBI | q 1-1176  t 1-1176 | 100 | 0e-0 | 100 |
| 27094 | PuppyEggo + AxiPup | 37 | NCBI | q 1-1185  t 1-  1185 | 100 | 0e-0 | 100 |

the start @27094 produces the most complete alignment with annotated genomes. The start also changes from TTG to the more frequently encountered GTG

h. Starterator: 27094 is the most annotated start

i. longest possible ORF? the ORF is longer if the start is at 27094. This start also includes all of the GenMark coding potential

j. conclusion: this is a gene with start at 27094: a) includes all coding potential, b) most annotated start, c) longest ORF, d) alignment with other members of the pham, e) good although not best rbs score

Gene function

1. Synteny

a. is synteny informative for this gene? somewhat…there is a 4-base overlap with gp37, which is a AAA ATPase with potential DNA-binding function. Not unreasonable to think that gp38 has a related function if these two genes are transcribed as an operon

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Greys\_38 | exonuclease | 0e-0 |
|  | WildNOut\_38 | RecB-like exonuclease/helicase | 0e-0 |

The consensus from phagesDB is that gp38 is an exonuclease. More specifically, this protein may be a RecB-like exonuclease/helicase.

3. BlastP (NCBI) 100+ hits. Top 3:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Greys | 38 | exonuclease | 0e-0 | 100 | 99.75 |
| Teagan | 38 | exonuclease | 0e-0 | 98 | 99.74 |
| Alyxandracam | 38 | Cas4 exonuclease | 0e-0 | 98 | 99.5 |

Consensus function for gp38 on NCBI blastp is an exonuclease. Could be a Cas4 exonuclease, or also a RecB-like exonuclease/helicase [Strathdee\_38 98% alignment and 99.2% identity]

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| CRISPR/CAS system-associated Cas4 |  | 41-182 | 2.14e-10 |

Similar to gp37, a chunk of gp38 aligns with a CRISPR system protein. Cas4 is a recB-like nuclease.

5. HHPred (only for genes > 90% probable matches)

Many hits with probability > 90% . Top hits with best alignments:

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| ATP-dependent DNA helicase | 99.8 | 6.5e-18 | 62 |
| ATP-dependent DNA helicase | 99.8 | 1.3e-16 | 63 |
| ATP-dependent DNA helicase/deoxyribonuclease | 99.8 | 4.3e-16 | 64 |
| exonuclease | 99.6 | 9.6e-14 | 67 |

#### Consensus from HHpred: 1) there is a helicase domain, 2) there is an exonuclease function.

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp38 is a soluble protein

conclusion: gp38 is a RecB-like exonuclease/helicase. This protein has a helicase domain, so it should not be characterized as a Cas4 exonuclease.

Vispistious gp39 pham 68543; one of the larger genes in the genome of this virus

1. auto-annotated start/stop coordinates: 27072 ← 28937; reverse transcribed

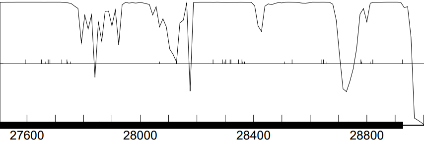
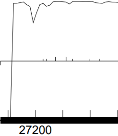
2. length = 1866 bp

3. start data and coding potential

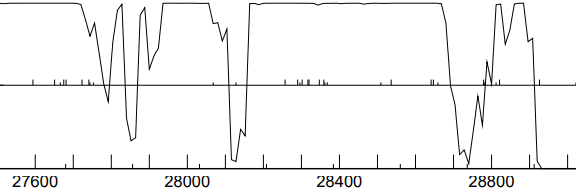
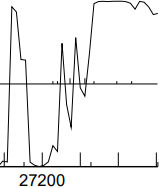
a. Glimmer @ 28937 has strength = 13.27

b. GeneMark calls the start at 28928

c. GeneMarkS shows strong coding potential throughout the region of this large gene:



GeneMark mfolio also shows strong coding potential, although GeneMark shows more coding potential on the right side of the gene:



d. start sequence = TTG, stop sequence = TAA

e. upstream: 23 bp overlap with gp39; downstream big 199 bp gap to gp40

f. This is a big gene with many starts. The first four should be evaluated:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 29024 | 2.264 | -5.060 | #6/44 |
| 29021 | 2.238 | -5.064 | #7/44 |
| 28937 | 2.897 | -2.601 | #1/44 |
| 28928 | 0.515 | -8.238 | #43/44 |

the auto-annotated start @28937 has the best rbs score; genemark’s start rbs score is one of the weakest

g. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val | % id |
| 28937 | PuppyEggo + AxiPup | 39 | NCBI | q 1-1866  t 1-1866 | 100 | 0e-0 | 100 |
| 28928 | PuppyEggo + AxiPup | 39 | NCBI | q 1-1857  t 1-  1857 | 100 | 0e-0 | 100 |

the start @28937 produces the most complete alignment with annotated genomes.

h. Starterator: 68543 is a big pfam with many members. Starterator is not superinformative…the most annotated start is not present in gp39. Among other members of cluster EA1, start 28937 is the start most often called when present.

i. longest possible ORF? the ORF is longer if the start is at 28937. This start also includes all of the GenMark coding potential

j. conclusion: this is a gene with start at 28937: a) includes all coding potential, b) longest ORF, c) best rbs score among competing starts, and d) blast alignment with annotated genomes

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Zada\_39 | DNA polymerase I | 0e-0 |
|  | WildNOut\_39 | DNA polymerase I | 0e-0 |

The clear consensus from phagesDB is that gp39 is a DNA polymerase I

3. BlastP (NCBI) 100+ hits. Top 3:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Chako | 39 | DNA polymerase I | 0e-0 | 100 | 99.8 |
| Ilzat | 39 | DNA polymerase I | 0e-0 | 100 | 100 |
| Chamuel | 39 | DNA polymerase I | 0e-0 | 100 | 99.7 |

The consensus function for gp39 on NCBI blastp is DNA polymerase I.

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| DNA polymerase I |  | 20-614 = 95% | 7.68e-62 |

# Really good evidence here that gp39 is a DNA polymerase I. Many references in the literature…this is a well-characterized protein. Joyce et al 1982. Nucleotide sequence of the Escherichia coli polA gene and primary structure of DNA polymerase I. J Biol Chem 257(4): 1958-64.

5. HHPred (only for genes > 90% probable matches)

Many hits with probability > 90% . Top hits with best alignments:

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| DNA polymerase I | 100 | 2.1e-67 | 94 |
| DNA polymerase | 100 | 7.8e-66 | 92 |
| Klenow fragment of DNA polymerase I | 100 | 8.8e-63 | 94 |
| DNA polymerase | 100 | 3.9e-62 | 95 |

#### Consensus from HHpred: DNA polymerase I

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp39 is a soluble protein

conclusion: the function of gp39 is DNA polymerase I.

Vispistious gp40 pham 652

1. auto-annotated start/stop coordinates: 29137 ← 29562; reverse transcribed

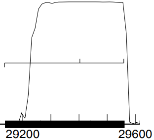
2. length = 426 bp

3. start data and coding potential

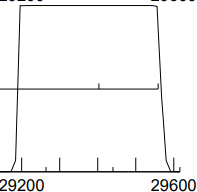
a. Glimmer @ 29562 has strength = 18.66

b. GeneMark calls the start at 28928

c. GeneMarkS shows strong uniform coding potential:



GeneMark mfolio also shows strong uniform coding potential:



d. start sequence = ATG, stop sequence = TGA

e. downstream large 199 bp gap to gp39; upstream 2bp AT overlap with gp41

f. only two starts in this small gene. Of these, only 29562 would include all of the coding potential:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 29562 | 2.970 | -2.443 | #1/2 |
| 29406 | 1.838 | -6.177 | #2/2 |

the auto-annotated start @29562 has the best rbs score

g. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val | % id |
| 29562 | LovelyUnicorn + 28 others | 40 | NCBI | q 1-426  t 1-426 | 100 | 0e-0 | 100 |

the start @29562 produces the most complete alignment with annotated genomes.

h. Starterator: 29562 is the most annotated start.

i. longest possible ORF? yes. This start also includes all of the GenMark coding potential

j. conclusion: this is a gene with start at 29562: a) includes all coding potential, b) longest ORF, c) best rbs score among competing starts, and d) good blast alignment with annotated genomes, e) most annotated start.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Winzigespinne\_40 | function unknown | 3e-74 |

The consensus from phagesDB is that the function of gp40 is unknown

3. BlastP (NCBI) 56 hits. Top 3:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat | 39 | hypothetical protein | 3e-91 | 100 | 100 |
| Microbacterium sp. HM58-2 |  | ABC transporter | 3e-26 | 45 | 71 |

The consensus function for gp40 on NCBI blastp is unknown. One hit out of 56 for an ABC transporter domain…but there is no evidence that gp40 is a membrane protein.

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| no CDD hits |  |  |  |

5. HHPred (only for genes > 90% probable matches)

Many hits with probability > 90% . Top hits with best alignments:

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| no hits > p = 90% |  |  |  |

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp40 is a soluble protein

conclusion: the function of gp40 is not known

Vispistious gp41 pham 6982

the members of this pfam are all draft annotations

1. auto-annotated start/stop coordinates: 29561 → 29677; forward transcribed

2. length = 117 bp

3. start data and coding potential

a. Glimmer @ 29562 has strength = 18.66

b. GeneMark calls the start at 28928

c. GeneMarkS shows a narrow peak of coding potential:



GeneMark mfolio does not show coding potential in this region:



d. start sequence = ATG, stop sequence = TGA

e. downstream 2 bp gap with gp40; upstream 54bp gap with gp42

f. This small gene does not have many viable positions for a start. The gene is forward-transcribed, so no overlap with gp40 or gp42 is possible (?). Alternate starts exist, but would shrink the ORF below the 120 bp threshold for a gene:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF length |
| 29561 | 1.986 | -5.384 | 117 |
| 29564 | 1.986 | -4.555 | 114 |
| 29573 | 1.488 | -5.685 | 105 |
| 28579 | 1.288 | -6.832 | 99 |

none of these starts are strong

g. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val | % id |
| 29561 | ParleG + 82 others | complete genome | NCBI | q 1-117  t 1-117 | 100 | 4e-52 | 100 |

the start @29561 aligns with an identical sequence from a large number of annotated complete genomes.

h. Starterator: 29561 is the most annotated start...only draft genomes are in this pfam.

i. longest possible ORF? yes.

j. blastp, etc. to see is this sequence matches with any known genes

phagesDB blastp: draft putative gene sequences only

blastp NCBI: no similarity to other genes found

hhpred: no hits > 90% probability

blastp CDD: no conserved domains found

membrane protein searchers: not a membrane protein

k. conclusion: draft\_41 is not a gene: a) genemark mfolio shows no coding potential, b) at 117 bp, the longest ORF is slightly below the threshold length for a gene (120 bp), c) poor rbs scores for all starts, d) no matches with known genes/domains

conclusion: delete Vispistious\_draft\_41

Vispistious gp41 (aka draft 42) pham 67046

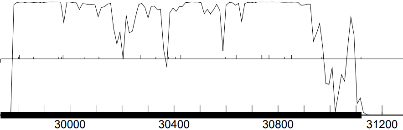
1. auto-annotated start/stop coordinates: 29733 ← 31121; reverse transcribed

2. length = 1389 bp

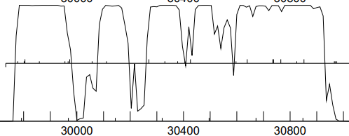
3. start data and coding potential

a. Glimmer @ 31121 has strength = 10.69

b. GeneMarkS shows strong coding potential:



GeneMark mfolio also shows good coding potential:



c. start sequence = GTG, stop sequence = TAA

d. downstream 1 bp overlap with gp42; upstream there is a big 596 bp gap to gp40. There is a lack of viable starts in this region + coding potential in this region.

e. there are many starts, but only the first @31121 includes all of the coding potential. The next few starts after 31121 are included for comparison:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 31121 | 2.078 | -4.498 | 1389 |
| 30977 | 1.311 | -7.014 | 1245 |
| 30971 | 0.947 | -6.864 | 1239 |
| 30968 | 0.947 | -6.925 | 1236 |
| 30854 | 1.385 | -6.945 | 1122 |

the auto-annotated start @31121 has the best rbs score of these starts, and includes all of the coding potential found by GeneMark

f. BLAST-start

100 hits

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val | % id |
| 31121 | Greys + 2 others | 41 | NCBI | q 1-1389  t 1-1389 | 100 | 0e-0 | 100 |

the start @31121 produces complete alignment with annotated genomes.

g. Starterator: 31121 is the most annotated start.

h. longest possible ORF? yes.

i. conclusion: this is a gene with start at 31121: a) includes all coding potential, b) longest ORF, c) best rbs score among viable competing starts, and d) good blast alignment with annotated genomes, e) most annotated start.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNout\_41 + all other non-draft genomes | DNA helicase | 0e-0 |

The consensus from phagesDB is that gp\_41 is a DNA helicase

3. BlastP (NCBI) 500 hits. Top 3:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat (best hit) | 41 | DNA helicase | 0e-0 | 99 | 99.8 |

The consensus function for gp41 on NCBI blastp = DNA helicase.

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| domain | function | e-val |
| 5-184 | N-terminal helicase of the DEAD-box superfamily | 2.9e-13 |
| 323-412 | C-terminal helicase of the DEAD-like superfamily | 1.69e-10 |

gp\_41 is a helicase

5. HHPred (only for genes > 90% probable matches)

More than 25 hits with probability = 100% . Top hits with best alignments:

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| Nuclear protein/motor protein | 100 | 7.7e-48 | 92 |
| ATP-dependent chromatin remodeler | 100 | 9.3e-46 | 91 |
| ISWI chromatin-remodeling complex ATPase ISW1 | 100 | 2.5e-46 | 92 |

Baker et al 2021. Structural insights into assembly and function of the RSC remodeling complex.Nat Struct Mol Biol 28: 71-80

The hits on Hhpred are consistent with the idea that gp41 is a DNA helicase that requires ATP for its function.

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp41 is a soluble protein

conclusion: Vispitious\_41 is a DNA helicase

Vispistious\_42 (aka draft 43) pham 998

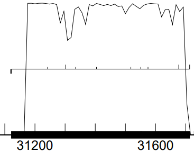
1. auto-annotated start/stop coordinates: 31121 ← 31714; reverse transcribed

2. length = 594 bp

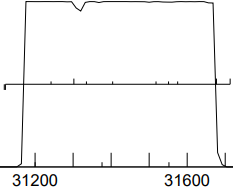
3. start data and coding potential

a. Glimmer @ 31714 has strength = 13.95

b. GeneMarkS shows strong coding potential:



GeneMark mfolio also shows strong and uniform coding potential:



c. start sequence = ATG, stop sequence = TAG

d. upstream 1 bp overlap with vis\_41; downstream there is a 4 bp ATGA overlap with vis\_43.

e. there are competing starts, but only the first @31714 includes all of the coding potential. The next few starts after 31772 are included for comparison:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 31714 | 1.966 | -4.738 | 594 |
| 31678 | 2.820 | -3.594 | 558 |
| 31576 | 1.350 | -6.826 | 456 |
| 31552 | 1.442 | -6.071 | 432 |
| 31519 | 1.555 | -5.558 | 339 |

the auto-annotated start @31714 has the 2nd best rbs score of these starts, and includes all of the coding potential found by GeneMark

f. BLAST-start

100 hits

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val | % id |
| 31714 | PuppyEggo + 12 oters | 42 | NCBI | q 1-594  t 1-594 | 100 | 0e-0 | 100 |

the start @31714 produces complete alignment with annotated genomes.

g. Starterator: 31714 is the most annotated start.

h. longest possible ORF? yes.

i. conclusion: this is a gene with start at 31714: a) includes all coding potential, b) longest ORF, c) OK rbs score among viable competing starts, and d) excellent blast alignment with annotated genomes, e) most annotated start.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Strathdee\_42 + most other non-draft genomes | phosphoesterase | 0e-0 |

The consensus from phagesDB is that gp\_42 is a phosphoesterase; RNA ligases on phagesDB matching gp42 have lower but still good e-values, alignment and identity.

3. BlastP (NCBI) 500 hits. Top 3:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| PuppyEggo (best hit) | 42 | phosphoesterase | 1e-140 | 99 | 100 |
| WildNout | 42 | phosphoesterase | 8e-140 | 99 | 99.5 |
| Ilzat | 42 | RNA ligase | 1e-139 | 99 | 99.5 |
| AxiPup | 42 | phosphoesterase | 1e-139 | 99 | 99.5 |
| Teagan | 42 | RNA ligase | 3e-139 | 99 | 99.5 |

Two competing hypotheses for the function of gp42 on NCBI blastp: 1) phosphoesterase or 2) RNA ligase. Essentially the same molecules here being given different names.

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| domain | function | e-val |
| 75-190 | 2’to 5’ RNA ligase superfamily | 4.44e-04 |

gp\_42 has a RNA ligase domain, aligning with 58% of the gene. Phosphoesterase and RNA ligase may not be mutually exclusive

5. HHPred (only for genes > 90% probable matches)

23 hits with probability > 90% + > 70% alignment. Top hits with best alignments:

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| cyclic nucleotide phosphodieasterase | 99.72 | 2.8e-15 | 76 |
| cyclic phosphodiesterase RNA processing | 99.7 | 4.2e-15 | 76 |
| cyclic phosphodieasterase | 99.7 | 6.4e-15 | 77 |
| 2’-5’ RNA ligase | 99.7 | 2e-14 | 76 |
| cyclic phosphodiesterase | 99.64 | 3.2e-14 | 76 |
| RNA ligase superfamily | 99.63 | 5e-14 | 72 |
| RNA 2’3’ cyclic phosphodiesterase | 99.63 | 1.5e-14 | 76 |
| RNA 2’3’ cyclic phosphodiesterase | 99.6 | 7.9e-15 | 76 |
| 2’-5’ RNA ligase | 99.61 | 2.7e-14 | 74 |
| 2’-5’ RNA ligase | 99.6 | 6.4e-14 | 72 |
| phosphoesterase | 99.58 | 1.3e-13 | 76 |

The hits on Hhpred are consistent with the idea that gp42 is a phosphoesterase or perhaps more specifically a RNA ligase.

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp42 is a soluble protein

conclusion: Vispitious\_42 is a RNA ligase: 1) RNA ligases are phosphoesterases), 2) 2’-5’ RNA ligase is a conserved domain match, 3) the more specific RNA ligase matches on HHPred have similar alignment and e-value as the more general phosphoesterase matches, 4) no HHPred match specifies a DNA phosphoesterase/ligase.

Vispistious\_43 (aka draft 44) pham 886

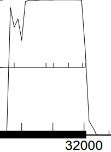
1. auto-annotated start/stop coordinates: 31711 ← 32007; reverse transcribed

2. length = 297 bp

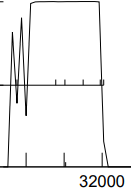
3. start data and coding potential

a. Glimmer @ 32007 has strength = 15.59

b. GeneMarkS shows strong coding potential:



GeneMark mfolio also shows strong coding potential:



c. start sequence = ATG, stop sequence = TGA

d. upstream 4 bp ATGA overlap with vis\_42; downstream there is a 1 base overlap with vis\_44.

e. there are competing starts, but only the first @32007 includes all of the coding potential. The next few starts after 32007 are included for comparison:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 32007 | 2.970 | -2.523 | 297 |
| 31998 | 2.970 | -4.049 | 288 |
| 31953 | 1.727 | -5.638 | 243 |

the auto-annotated start @32007 has the best rbs score of these starts, and includes all of the coding potential found by GeneMark

f. BLAST-start

100 hits

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val | % id |
| 32007 | PuppyEggo and WildNout | 43 | NCBI | q 1-297  t 1-297 | 100 | 1e-151 | 100 |

the start @32007 produces complete alignment with annotated genomes.

g. Starterator: 32007 is the most annotated start.

h. longest possible ORF? yes.

i. conclusion: this is a gene with start at 32007: a) includes all coding potential, b) longest ORF, c) best rbs score among viable competing starts, and d) excellent blast alignment with annotated genomes, e) most annotated start.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (PhagesDB) 500 hits. Top 3:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| WildNout + 3 | 43 | function unknown | 5e-48 | 100 | 100 |

good matches on phagesDB; none with identified functions

3. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| domain | function | e-val |
| no conserved domains identified |  |  |

4. HHPred (only for genes > 90% probable matches)

23 hits with probability > 90% + > 70% alignment. Top hits with best alignments:

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| no HHpred hits > 90% probability |  |  |  |

5. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp43 is a soluble protein

conclusion: function unknown

Vispistious\_44 (aka draft 45) pham 63710

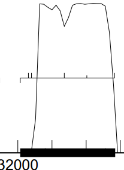
1. auto-annotated start/stop coordinates: 32007 ← 32285; reverse transcribed

2. length = 279 bp

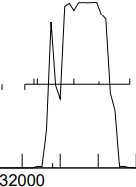
3. start data and coding potential

a. Glimmer @ 32285 has strength = 12.02

b. GeneMarkS shows strong coding potential:



GeneMark mfolio also shows strong coding potential:



c. start sequence = ATG, stop sequence = TGA

d. upstream 1 bp overlap with vis\_43; downstream 1 base overlap with vis\_45.

e. there are competing starts, but only the first @32285 includes all of the coding potential. The next few starts after 32285 are included for comparison:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 32285 | 2.004 | -4.595 | 279 |
| 32213 | 2.098 | -5.416 | 207 |
| 32204 | 1.607 | -5.718 | 198 |

the auto-annotated start @32285 has the best rbs score of the six potential starts, and includes all of the coding potential found by GeneMark

f. BLAST-start

100 hits

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val | % id |
| 32285 | PuppyEggo and WildNout | 44 | NCBI | q 1-279  t 1-279 | 100 | 9e-142 | 100 |

the start @32285 produces complete alignment with annotated genomes.

g. Starterator: 32285 is the most annotated start.

h. longest possible ORF? yes.

i. conclusion: this is a gene with start at 32285: a) includes all coding potential, b) longest ORF, c) best rbs score among viable competing starts, and d) excellent blast alignment with annotated genomes, e) most annotated start.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNout\_44 | function unknown | 1e-48 |

good alignment; all hits on phagesDB = function unknown

3. BlastP (NCBI)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| WildNout + 14 | 44 | function unknown | 1e-48 | 100 | 100 |

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| domain | function | e-val |
| no conserved domains identified |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| no HHpred hits > 90% probability |  |  |  |

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp44 is a soluble protein

conclusion: the function of vis\_44 is unknown

Vispistious\_45 (aka draft vis\_46) pham 65692

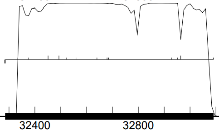
1. auto-annotated start/stop coordinates: 32285 ← 33091; reverse transcribed

2. length = 807 bp

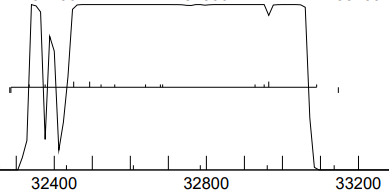
3. start data and coding potential

a. Glimmer @ 33091 has strength = 18.73

b. GeneMarkS shows strong coding potential:



GeneMark mfolio also shows strong coding potential:



c. start sequence = GTG, stop sequence = TGA

d. upstream 1 bp overlap with vis\_44; downstream 4bp GTGA overlap with vis\_46.

e. as with other genes in this region there are competing starts, but only the first @33901 includes all of the coding potential. The next few starts after 33901 are included for comparison:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 33901 | 2.702 | -3.160 | 807 |
| 32965 | 2.010 | -4.644 | 681 |
| 32593 | 1.305 | -6.078 | 669 |
| 32929 | 2.358 | -4.408 | 645 |
| 32686 | 1.011 | -6.708 | 402 |

the auto-annotated start @33901 has the best rbs score of the 13 potential starts, and includes all of the coding potential found by GeneMark

f. BLAST-start

100 hits

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val | % id |
| 33901 | PuppyEggo | 45 | NCBI | q 1-807  t 1-807 | 100 | 0e-0 | 100 |
| 33901 | 83 other phage genes have 100% alignment | 45 | NCBI | q 1-807  t 1-807 | 100 | 0e-0 | ≥90 |

the start @33901 produces complete alignment with annotated genomes.

g. Starterator: 33901 is the most annotated start.

h. longest possible ORF? yes.

i. conclusion: this is a gene with start at 33901: a) includes all coding potential, b) longest ORF, c) best rbs score among viable competing starts, and d) excellent blast alignment with annotated genomes, e) most annotated start.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNout\_44 | MazG-like nucleotide pyrophosphohydrolase | 1e-157 |

all hits on phagesDB are for either MazG-like nucleotide pyrophosphohydrolase, or nucleotide pyrophosphohydrolase

3. BlastP (NCBI) 35 hits.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Teagan | 45 | nucleotide pyrophosphohydrolase | 0e-0 | 99 | 100 |
| Oxtober | 96 | MazG-like nucleotide pyrophosphohydrolase | 0e-0 | 99 | 99.63 |
| Figueroism |  | MazG-like nucleotide pyrophosphohydrolase | 0e-0 | 99 | 98.88 |
| Pelagibacterales bacterium SAG-MED50 a |  | nucleoside triphosphate pyrophosphohydrolase family | 4e-19 | 51 | 39.6 |

a Haro-Moreno et al 2020. Ecogenomics of the SAR11 clade. Environ. Microbiol. 22(5): 1748-1763

All matches on NCBI point towards a nucleotide pyrophosphohydrolase…only question = is it a MazG-like nucleotide pyrophosphohydrolase

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| domain | function | e-val |
| 114-216 | nucleoside triphosphate pyrophosphohydrolase MazG-like domain a | 2.65e-16 |

* a [Structural and functional insights into DR2231 protein, the MazG-like nucleoside triphosphate pyrophosphohydrolase from Deinococcus radiodurans.](https://www.ncbi.nlm.nih.gov/pubmed/21733847)J Biol Chem 2011 Sep 02; 286(35):30691-30705

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| MazG-like nucleoside triphosphate pyrophosphohydrolase | 99.9 | 6.6e-24 | 48 |
| alpha helix metal binding protein | 99.97 | 3.8e-30 | 51 |

strong evidence that gp45 is a MazG-like nucleoside triphosphate pyrophosphohydrolase

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp45 is a soluble protein

conclusion: the function of gp45 is a MazG-like nucleotide pyrophosphohydrolase

Vispistious\_46 (aka draft 47) pham 67105

1. auto-annotated start/stop coordinates: 33088 ← 33744; reverse transcribed

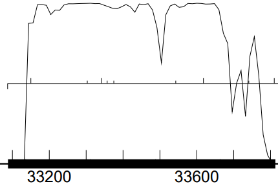
2. length = 657 bp

3. start data and coding potential

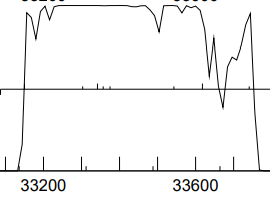
a. Glimmer @ 33744 has strength = 10.36

b. Genemark calls the start at 33813

c. GeneMarkS shows strong coding potential:



GeneMark mfolio also shows strong coding potential:



d. start sequence = GTG, stop sequence = TGA

e. upstream 4 bp GTGA overlap with vis\_45; downstream 45 base gap to vis\_47.

f. there are at least two competing starts:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 33900 | 1.127 | -7.701 | 813 |
| 33813 | 1.283 | -6.126 | 726 |
| 33744 | 2.078 | -5.185 | 657 |
| 33621 | 3.104 | -2.156 | 534 |

the auto-annotated start @33744 has a better rbs score than the GeneMark-specified start at 33813. However, the Glimmer start excludes moderate coding potential on the right side of this gene.

g. BLAST-start

100 hits

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val |
| 33744 | PuppyEggo + 82 | 46 | NCBI | q 1-657  t 1-657 | 100 | 0e-0 |
| 33813 | PuppyEggo + 85 | 46 | NCBI | q 1-843  t 1-843 | 100 | 0e-0 |

the start @33813 produces more complete alignment with annotated genomes.

h. Starterator: 33813 is the most annotated start.

i. longest possible ORF? yes if the start is at 33813.

j. conclusion: this is a gene with start at 33813: a) includes all coding potential, b) longest ORF, c) rbs score not great, but neither is the rbs score for the Glimmer start, d) excellent blast alignment with annotated genomes, e) most annotated start.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | PuppyEggo\_46 | function unknown | e-135 |

good alignments; all hits on phagesDB = function unknown

3. BlastP (NCBI)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| PuppyEggo + 57 | 46 | function unknown | 1e-97 | 99 | 100 |

all hits on NCBI blastp are function = hypothetic protein

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| domain | function | e-val |
| no conserved domains identified |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| no HHpred hits > 90% probability |  |  |  |

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp46 is a soluble protein

conclusion: the function of vis\_46 is unknown

Vispistious\_47 (aka draft 48) pham 657

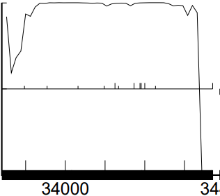
1. auto-annotated start/stop coordinates: 33791 ← 34372; reverse transcribed

2. length = 582 bp

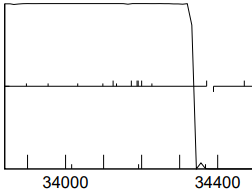
3. start data and coding potential

a. Glimmer @ 34372 has strength = 13.78

b. GeneMarkS shows good coding potential:



GeneMark mfolio shows the same strong coding potential:



c. start sequence = ATG, stop sequence = TGA

d. upstream 23 bp overlap with vis\_46; downstream 15 bp gap to vis\_48.

e. the auto-annotated start looks like the best option for this gene:

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 34372 | 2.970 | -2.584 | 582 |
| 34228 | 2.358 | -4.584 | 438 |
| 34201 | 1.178 | -6.350 | 411 |
| 34192 | 2.419 | -3.688 | 402 |

the auto-annotated start @34372 has the best rbs score and includes all of the coding potential.

f. BLAST-start

100 hits

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val |
| 33744 | LovelyUnicorn + 15 | 47 | NCBI | q 1-582  t 1-582 | 100 | 0e-0 |

the start @34732 produces more complete alignment with annotated genomes.

g. Starterator: 34732 is the most annotated start.

h. longest possible ORF? yes if the start is at 34732.

i. conclusion: this is a gene with start at 34372: a) includes all coding potential, b) longest ORF, c) best rbs score, d) excellent blast alignment with annotated genomes, e) most annotated start.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Zada\_48 | function unknown | e-112 |
|  | WildNOut\_48 | thymidylate kinase | e-112 |
|  | Velene\_47 | function unknown | e-112 |
|  | TinSulphur\_48 | function unknown | e-112 |
|  | Thorongil\_47 | AAA-ATPase | e-112 |

good alignments; the function of the same molecule is described on phagesDB as: 1) unknown, 2) AAA-ATPase, 3) thymidylate kinase.

3. BlastP (NCBI)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Teagan | 46 | thymidylate kinase | 5e-142 | 99 | 100 |
| BeauPeep30 |  | AAA-ATPase | 3e-141 | 99 | 99.48 |
| MonChoix |  | thymidylate kinase | 4e-121 | 99 | 85.57 |

NCBI blastp is identifying the function of this gene as either a thymidylate kinase or an AAA-ATPase. Both are likely correct.

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| domain | function | e-val |
| no conserved domains identified |  |  |

5. HHPred (only for genes > 90% probable matches)

popular gene on HHPred with more than 250 hits. Top hits with the best coverage:

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| deoxynucleoside kinase a | 99.4 | 4.8e-11 | 71 |
| thymidylate kinase b | 99.4 | 4.8e-12 | 84 |
| thymidylate kinase c | 99.3 | 1.4e-11 | 85 |

# a Johansson 1997 Cloning of the cDNA and chromosome localization of the gene for human thymidine kinase. J Biol Chem 272: 8454-8458

# b Sinha and Rule. 2017. The structure of Thymidylate Kinase from *Candida albicans* Reveals a Unique Structural Element. Biochemistry 56 (33): 4360-4370.

# b Huaqing et al 2012. Synthesis and Evaluation of α-Thymidine Analogues as Novel Antimalarials. J Med Chem 55: 10948-10957

Vis\_47 may be perhaps best described as an AAA-ATPAase with a thymidylate kinase domain.

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp47 is a soluble protein

conclusion: AAA-ATPase is the larger group within which this protein resides. There is enough evidence from HHPred to go farther and id the function of vis\_47 as thymidylate kinase. Conclusion: vis\_47 is a thymidylate kinase.

Vispistious\_48 (aka draft 49) pham 60726

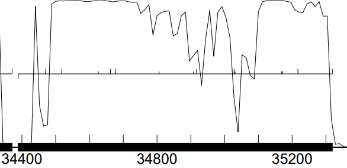
1. auto-annotated start/stop coordinates: 34388 ← 35320; reverse transcribed

2. length = 933 bp

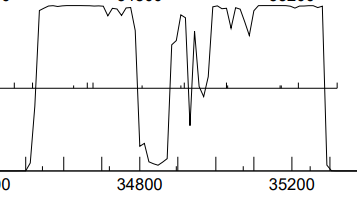
3. start data and coding potential

a. Glimmer @ 35230 has strength = 14.62

b. GeneMarkS shows good coding potential:



GeneMark mfolio shows the same strong coding potential:



c. start sequence = ATG, stop sequence = TAA

d. upstream 15 bp gap to vis\_47; downstream 8 bp overlap with vis\_49.

e. none of the most likely starts have strong rbs scores::

(scoring = Kibler6, spacing = Karlin Med.)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | ORF  length |
| 35320 | 1.987 | -6.158 | 933 |
| 35218 | 1.888 | -5.865 | 831 |
| 35173 | 1.358 | -7.002 | 786 |
| 35170 | 0.621 | -8.134 | 783 |
| 35032 | 0.977 | -6.718 | 645 |

the auto-annotated start @34372 has the best z-score and 2nd best rbs score of the viable starts for this gene, and includes all of the coding potential.

f. BLAST-start

100 hits

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene | d-base | query: subject | % align | e-val |
| 35230 | LovelyUnicorn + AxiPup | 48 | NCBI | q 1-933  t 1-933 | 100 | 0e-0 |

the start @35230 produces complete alignment with annotated genomes.

g. Starterator: 35230 is the most annotated start.

h. longest possible ORF? yes if the start is at 35230.

i. conclusion: this is a gene with start at 35230: a) includes all coding potential, b) longest ORF, c) OK rbs score, d) excellent blast alignment with annotated genomes, e) most annotated start.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut\_48 | glycosyltransferase | 0e-0 |
|  | Velene\_48 | glycosyltransferase | 0e-0 |
|  | Thorongil\_48 | glycosyltransferase | 0e-0 |
|  | Thompsone\_48 | glycosyltransferase | 0e-0 |
|  | Strathdee\_48 | glycosyltransferase | 0e-0 |

excellent alignments; the uniform function phagesDB consensus among annotated is that gp48 is glycosyltransferase

3. BlastP (NCBI) 56 hits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat | 46 | glycosyltransferase | 0e-0 | 99 | 100 |
| Ludgate |  | glycosyltransferase | 0e-0 | 99 | 99.7 |
| PuppyEggo |  | glycosyltransferase | 0e-0 | 99 | 99.7 |

NCBI blastp function = glycosyltransferase

4. BlastP (Conserved domains)

|  |  |  |
| --- | --- | --- |
| domain | function | e-val |
| no conserved domains identified |  |  |

5. HHPred (only for genes > 90% probable matches)

popular gene on HHPred with more than 250 hits. Top hits with the best coverage:

|  |  |  |  |
| --- | --- | --- | --- |
| description | *p* | e-val | % align |
| glucosyl transferase | 100 | 1.3e-27 | 82 |
| acetylgalactosaminyltransferasea | 99.5 | 3.4e-12 | 73 |
| glycosyl transferase b | 99.5 | 1e-12 | 70 |
| glycosyl transferase c | 99.4 | 3.5e-11 | 72 |

a de las Rivas et al 2020. Nature Chemical Biology 16: 351-360

b Enroth and Strid 2007. Biochimica et Biophysica Acta-Protein and Proteomics 1784 (2): 379-384

c 1996. Organization of the e-coli K-12 gene cluster responsible for production of extracellular polysaccharide colonic acid. J Bacteriol 178(16): 4885-93

Good evidence from HHpred that Vis\_48 = glycosyltransferase.

6. Transmembrane finders

a. deep TMHMM: not a membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp48 is a soluble protein

conclusion: the function of vis\_48 = glycosyltransferase

Vispistious\_49 pham 68729

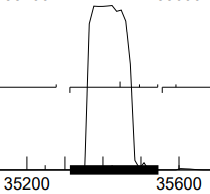
1. auto-annotated start/stop coordinates: 35313 ← 35546; reverse transcribed

2. length = 234 bp

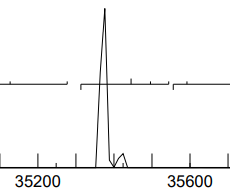
3. start data and coding potential

a. Glimmer @ 35546 has strength = 7.57

b. GeneMark shows coding potential in this small region of the genome:

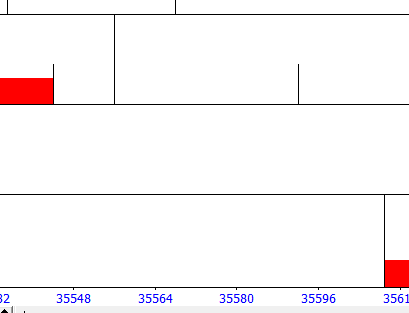


by GeneMark mfolio indicates a spike of coding in this region:



c. start sequence = GTG, stop sequence = TAA

d. downstream overlap = 8 bp ATGAATAA with gp48; 62 bp gap with upstream gp50. That’s a large gap. A stop near 35560 in this gap interrupts the potential coding region...no hits when the gap sequence is blasted. The start at 35546 is the first available start for this gene



e. Starterator: 35546 is the most annotated start

f. BLAST-start

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id | % sim |
| 35546 | Teagan | 49 | NCBI | q 1-77  t 1-77 | 100 | 2x10-51 | 100 | 100 |

g. ribosomal binding site

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| start | scoring matrix | spacing matrix | z score | final score | best score? |
| 35546 | Kibler6 | Karlin Medium | 2.093 | -4.326 | #1/5 |

h. longest possible ORF? yes

conclusion: this is a gene with start at 35546

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. phagesDB

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut\_49 | unknown | 7e-43 |

3. NCBI or both

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene function | % alignment | e-value |
| Teaggan | hypothetical protein | 98 | 2 x 10-51 |
| 9 others with 98% alignment and e-value < 2 x 10-50 | | | |

4. CDD

no hits

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene number | % alignment | probability |
| no hits > 90% likelihood |  |  |  |

note: % alignment = length of alignment/total gene length

4. Transmembrane finders

a. deep TMHMM: gp49 is a globular non-membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp49 is a soluble non-membrane protein

conclusion: the function of gp49 is not known

Vispistious\_50 pham 67173

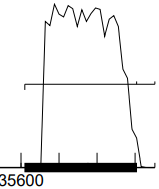
1. auto-annotated start/stop coordinates: 35609 ← 35546; reverse transcribed

2. length = 297 bp

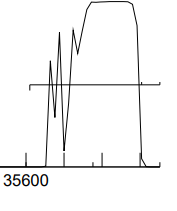
3. start data and coding potential

a. Glimmer @ 35905 has strength = 10.09

b. GeneMark shows coding potential in this small region of the genome:



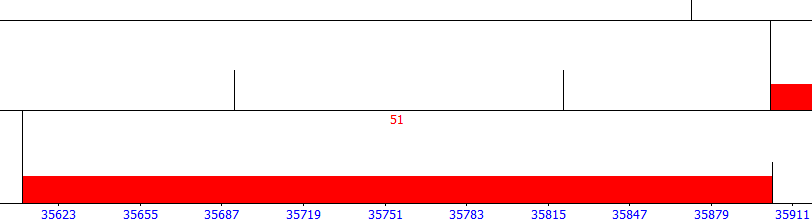
by GeneMark mfolio also finds strong coding potential in this region:



c. start sequence = GTG, stop sequence = TAA

d. downstream overlap = 62 bp with gp49, but no discernable coding potential in the gap. Upstream overlap = 4 bp GTGA with gp51 (two-gene operon?)

e. Starterator: 35905 is the suggested start. In fact 35905 is the only start



f. BLAST-start

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id | % sim |
| 35905 | LovelyUnicorn | 49 | NCBI | q 1-77  t 1-77 | 100 | 4 x 10-150 | 99.66 | 100 |
| 12 others w/ 100% alignment and e-value = 4 x 10-150 | | | | | | | | |

g. ribosomal binding site

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| start | scoring matrix | spacing matrix | z score | final score | best score? |
| 35546 | Kibler6 | Karlin Medium | 3.125 | -2.253 | #1/2 |

h. longest possible ORF? yes

conclusion: this is a gene with start at 35905

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. phagesDB

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Zada\_51 | unknown | 3e-53 |

3. NCBI

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene function | % alignment | e-value |
| Winzigespinne + 5 others | hypothetical protein | 98 | 3 x 10-62 |
| Endor | membrane protein | 97 | 6 x 10-37 |
| Jenos | membrane protein | 91 | 3 x 10-25 |
| McGalleon | membrane protein | 88 | 3 x 10-15 |

4. CDD: no hits

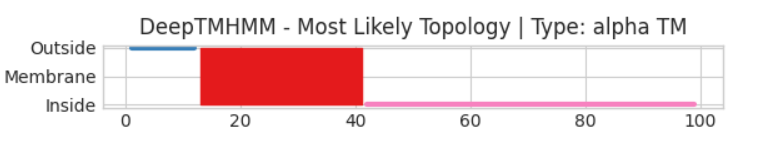
5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene number | % alignment | probability |
| no hits > 90% likelihood |  |  |  |

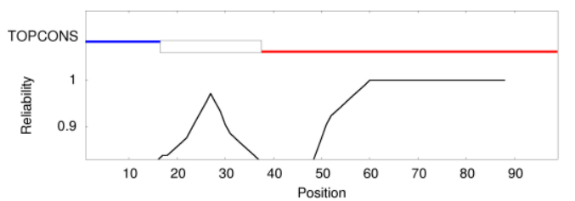
note: % alignment = length of alignment/total gene length

4. Transmembrane finders

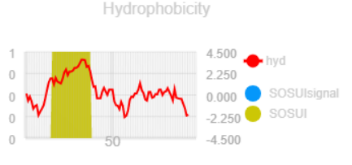
a. deep TMHMM: gp50 is an alpha transmembrane protein



b. TOPCONs: gp50 is a membrane protein, with residues 17-37 in a membrane environment



c. SOSUI: gp50 is a membrane protein, with residues 16-38 in a membrane domain



function conclusion: gp50 is a membrane protein

Vispistious\_51 pham 654

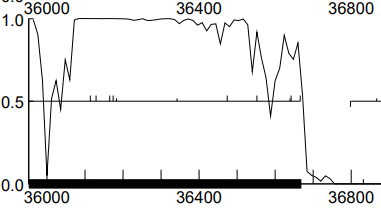
1. auto-annotated start/stop coordinates: 35902 ← 36669; reverse transcribed

2. length = 768 bp

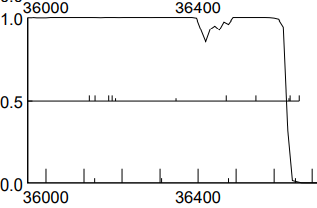
3. start data and coding potential

a. Glimmer @ 35905 has strength = 14.18

b. GeneMark shows good coding potential in this region:



GeneMark mfolio also indicates strong coding potential in this region:



c. start sequence = ATG, stop sequence = TGA

d. downstream overlap = 4 GTGA bp with gp50. Upstream gap = 19 bp sequence TCCGTCCGGAGGGTCCCGG between gp51 and gp52

e. Starterator: 36669 is the suggested start, called in 100% of genes in this pham when present.

f. BLAST-start

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id | % sim |
| 36669 | PuppyEggo | 51 | NCBI | q 1-768  t 1-768 | 100 | 0e0.0 | 100 | 100 |
| 67 others w/ 100% alignment and e-value 0.0e0 | | | | | | | | |

g. ribosomal binding site

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| start | scoring matrix | spacing matrix | z score | final score | best score? |
| 36669 | Kibler6 | Karlin Medium | 2.970 | -2.523 | #1/12 |

There is a potential upstream start which would increase the length of the ORF, but this start has a low z-value (0.947), the lowest final score of 13 possible starts and is well outside of regions of coding potential.

h. longest possible ORF? no

i. conclusion: this is a gene with start at 36669

Gene function

1. Synteny

a. is synteny informative for this gene? hypothesis: gp50 is a membrane protein and has an operon-like overlap with gp51. Does gp50 have a transmembrane transport function for the product of gp51?

b. upstream gene NA

c. downstream gene NA

2. phagesDB

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| n hits | best hit | function | e-value | % alignment | %  identity |
| 100 | WildNOut\_51 | thymidylate synthase | 1e-150 | 100 | 100 |

3. NCBI

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene function | % alignment | e-value |
| Aubergine + 10 others | thymidylate synthase | 100 | 0.0e0 |

4. CDD: no hits

3. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene number | % alignment | probability |
| Thymidylate synthase from Staphylococcus aureus (ex PDB) Wang et al 2013 J Am Chem Soc 135: 7583-7592 |  | 86 | 100 |

4. Transmembrane finders

a. deep TMHMM: gp51 is a globular non-membrane protein

b. TOPCONs: not a membrane protein

c. SOSUI: gp51 is a soluble protein

conclusion: the function of gp51 = thymidylate synthase

Vispistious\_52 pham 949

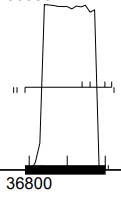
1. auto-annotated start/stop coordinates: 36789 ← 37001; reverse transcribed

2. length = 213 bp

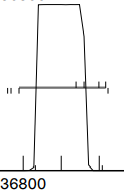
3. start data and coding potential

a. Glimmer @ 37001 has strength = 10.31

b. GeneMark shows good coding potential in this region:



GeneMark mfolio also indicates strong coding potential in this region:



c. start sequence = ATG, stop sequence = TGA

d. downstream gap = 120 bp to gp51: CACAGGAGGAACCTGCTAACAGCCCGGCCCCGACCGGGACCCACATTCAAATCGCGTATCTAAGAAGAACTCGTTGGTGCTCCCCTCCGGATGGGCGCGATCCGTCCGGAGGGTCCCGG. No blast hits on this nucleotide sequence

Upstream gap = 78 bp sequence between gp52 and gp53.

e. BLAST-start

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id | % sim |
| 37001 | LovelyUnicorn + 93 others | 52 | NCBI | q 1-212  t 1-212 | 100 | 0e0.0 | 100 | 100 |
| 37019 | LovelyUnicorn + 93 others | 52 | NCBI | q 1-231  t 1-231 | 100 | 1x10-105 | 99.54 | 100 |

f. ribosomal binding site

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| start | scoring matrix | spacing matrix | z score | final score | best score? |
| 37019 | Kibler6 | Karlin Medium | 1.607 | -6.672 | #1/4 |
| 37001 | Kibler6 | Karlin Medium | 1.014 | -7.291 | #2/4 |

There is a potential upstream start which would increase the length of the ORF with a superior score (yet still low).

g. Starterator: 37019 is the suggested start, called in 100% of genes in this pham when present.

h. longest possible ORF? yes with start at 37019

i. conclusion: this is a gene with start at 37019: a) highest ribosomal start score, b) excellent alignment scores on NCBI blast, c) longest ORF, d) most annotated start on Starterator

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. PhagesDB

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut\_52 | unknown | 5e-44 |

3. NCBI or both

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene function | % alignment | e-value |
| WildNOut | function unknown | 100 | 7x10-40 |

4. CDD: No hits

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene number | % alignment | *p* |
| no hits > 90% likelihood |  |  |  |

4. Transmembrane finders

a. deep TMHMM: gp52 is a soluble protein

b. TOPCONs: not a membrane protein

c. SOSUI: gp52 is a soluble protein

conclusion: the function of gp52 is unknown

Vispistious\_53 pham 750

1. auto-annotated start/stop coordinates: 37080 ← 37304; reverse transcribed

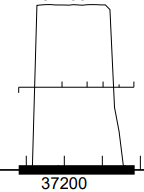
2. length = 225 bp

3. start data and coding potential

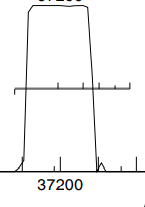
a. Glimmer @ 37304 has strength = 20.04

b. GeneMark calls start at 37385

c. GeneMark shows good coding potential in this region:



GeneMark mfolio also indicates strong coding potential in this region:



d. start sequence = ATG, stop sequence = TAG

e. downstream gap = 67 bp to gp52

Upstream gap = 62 bp sequence between gp53 and gp54.

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 37304 | ParleG + 18 others | 53 | NCBI | q 1-225  t 1-225 | 100 | 1e-111 | 100 |
| 37346 | ParleG + 18 others | 53 | NCBI | q1-267  t1-267 | 100 | 8e-135 | 100 |

g. ribosomal binding site

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| start | scoring matrix | spacing matrix | z score | final score | best score? |
| 37304 | Kibler6 | Karlin Medium | 3.125 | -2.253 | #1/6 |
| 37385 | Kibler6 | Karlin Medium | 1.687 | -5.277 | #4/6 |

37385 is the start called by GenMark. Its score is weak.

h. Starterator: 37304 is the suggested start, called in 100% of genes in this pham when present.

i. longest possible ORF? no if start @37304; yes if start at 37385

j. conclusion: this is a gene with start at 37304: a) highest ribosomal start score, b) excellent alignment scores on NCBI blast, c) most annotated start on Starterator.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut\_53 | unknown | 1e-35 |

3. NCBI

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene function | % alignment | e-value |
| Peep | function unknown | 100 | 1e-58 |

4. CDD: No hits

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene number | % alignment | *p* |
| no hits > 90% likelihood |  |  |  |

note: % alignment = length of alignment/total gene length

6. Transmembrane finders

a. deep TMHMM: gp53 is a globular non-membrane protein

b. TOPCONs: not a membrane protein

c. SOSUI: gp53 is a soluble protein

conclusion: the function of Vispistious\_53 is unknown

Vispistious\_54 pham 1009

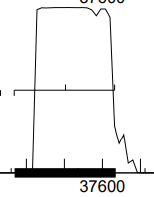
1. auto-annotated start/stop coordinates: 37369 ← 37635; reverse transcribed

2. length = 267 bp

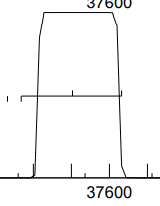
3. start data and coding potential

a. Glimmer @ 37635 has strength = 18.82

b. GeneMark shows good coding potential in this region:



GeneMark mfolio also indicates strong coding potential in this region:



c. start sequence = ATG, stop sequence = TAG

d. downstream gap = 62 bp to gp53

Upstream gap = 70 bp sequence between gp54 and gp55. Why do these small genes in this region consistently have large 60-70 bp gaps?

e. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 37635 | Lovelyunicorn + 16 others | 54 | NCBI | q 1-267  t 1-267 | 100 | 4e-135 | 100 |

f. ribosomal binding site

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| start | scoring matrix | spacing matrix | z score | final score | best score? |
| 37635 | Kibler6 | Karlin Medium | 3.125 | -2.112 | #1/2 |
| 37506 | Kibler6 | Karlin Medium | 2.897 | -3.703 | #2/2 |

g. Starterator: 37635 is the suggested start, called in 100% of genes in this pham when present.

h. longest possible ORF? yes

i. conclusion: this is a gene with start at 37635: a) highest ribosomal start score, b) excellent alignment scores on NCBI blast, c) most annotated start on Starterator, d) longest ORF

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut\_54 | unknown | 6e-44 |

3. NCBI 15 hits

|  |  |  |  |
| --- | --- | --- | --- |
| best hit | gene function | % alignment | e-value |
| Ilzat | function unknown | 100 | 4e-53 |

d. was hit identified at the conserved domain database? No

4. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene number | % alignment | *p* |
| no hits > 90% likelihood |  |  |  |

note: % alignment = length of alignment/total gene length

4. Transmembrane finders

a. deep TMHMM: gp54 is a globular non-membrane protein

b. TOPCONs: not a membrane protein

c. SOSUI: gp54 is a soluble protein

conclusion: the function of Vispistious\_54 is not known

Vispistious\_55 pham 60741

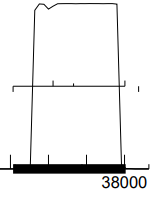
1. auto-annotated start/stop coordinates: 37707 ← 38003; reverse transcribed

2. length = 297 bp

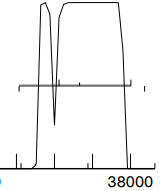
3. start data and coding potential

a. Glimmer @ 38003 has strength = 17.02

b. GeneMark shows good coding potential in this region:



GeneMark mfolio also indicates strong coding potential in this region:



c. start sequence = ATG, stop sequence = TAG

d. downstream gap = 70 bp to gp54

Upstream gap = 100 bp sequence between gp55 and gp56…another large gap between these small genes of the right arm of the genome. Is there a function?

e. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 38003 | HanSolo + 4 others | 55 | NCBI | q 1-297  t 1-297 | 100 | 1e-151 | 100 |

f. ribosomal binding site

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| start | scoring matrix | spacing matrix | z score | final score | best score? |
| 38003 | Kibler6 | Karlin Medium | 3.059 | -2.253 | #1/3 |

The start at 38003 is the most viable option for this gene. There is no start available in the upstream 100 bp gap.

g. Starterator: 38003 is the most annotated start.

h. longest possible ORF? yes

i. conclusion: this is a gene with start at 38003: a) highest ribosomal start score, b) excellent alignment scores on NCBI blast, c) most annotated start on Starterator, d) longest ORF

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene # | function | e-value |
| Strathdee | 55 | function unknown | 3e-52 |

3. BlastP (NCBI)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| AxiPup | 55 | unknown | 5e-64 | 98 | 100 |

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| COG1645 | Zn-finger containing protein | 38-98 | 3.23e-3 |

Evidence that gp55 has a zinc finger that is used to do something

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| description | domain | *p* | e-val | % align |
| NinF protein | 75-98 | 98.2 | 1.5e-6 | 42 |
| ribosome biogenesis protein | 75-98 | 97.8 | 5.5e-5 | 42 |
| zinc ribbon-containing domain | 75-98 | 97.8 | 5.1e-5 | 23 |

many hits with *p* > 0.9 for residues 75-98 on HHPred. Many of these refer to a zinc finger or zinc ribbon, and ribosome biogenesis. Not enough evidence to call a function for gp55. Given evidence from the CDD, it is likely that gp55 has a zinc-containing structure that is a key to its function.

4. Transmembrane finders

a. deep TMHMM: gp55 is a globular non-membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp55 is a soluble protein

conclusion: residues 75-98 of gp55 form a zinc finger/zinc ribbon domain that is likely an important part of the function of this protein. However, the function of Vispistious\_55 remains unknown

Vispistious\_56 pham 65745

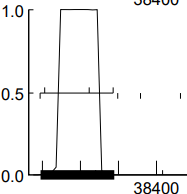
1. auto-annotated start/stop coordinates: 38095 ← 38289; reverse transcribed

2. length = 195 bp

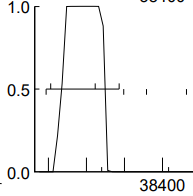
3. start data and coding potential

a. Glimmer @ 38289 has strength = 12.39

b. GeneMark shows good coding potential in this region:



GeneMark mfolio also indicates strong coding potential in this region:



c. start sequence = ATG, stop sequence = TAA

d. downstream gap = 100 bp to gp55

Upstream gap = 90 bp sequence to gp57…another sizeable gap between small genes of the right arm.

e. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 38289 | ParleG + 30 others | 56 | NCBI | q 1-195  t 1-195 | 100 | 3e-95 | 100 |

f. ribosomal binding site

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| start | scoring matrix | spacing matrix | z score | final score | best score? |
| 38289 | Kibler6 | Karlin Medium | 1.695 | -5.831 | #2/3 |

The start at 38289 is the most viable option for this gene. There is no start available in the upstream 90 bp gap. Starts at 38226 and 38109 exclude coding potential and capitate the ORF to < 120 bp.

g. Starterator: 38289 is the most annotated start.

h. longest possible ORF? yes

i. conclusion: this is a gene with start at 38289: a) strong alignment scores on NCBI blast, b) most annotated start on Starterator, c) longest ORF, d) called by Glimmer and GeneMark

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene # | function | e-value |
| Winzigespinne | 56 | function unknown | 4e-33 |

3. BlastP (NCBI)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Ilzat | 56 | unknown | 5e-39 | 98 | 100 |

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| no CDD hits |  |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| description | domain | *p* | e-val | % align |
| > 90% *p*  of match with domains of several eukaryotic proteins; no viral proteins |  | 92 | 1.3 | 23 |

4. Transmembrane finders

a. deep TMHMM: gp56 is a globular non-membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp56 is a soluble protein

conclusion: the function of Vispistious\_56 is unknown

Vispistious\_57 pham 7157

1. auto-annotated start/stop coordinates: 38328 ← 38456; reverse transcribed

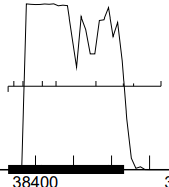
2. length = 129 bp

3. start data and coding potential

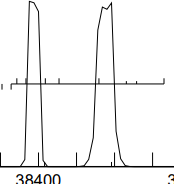
a. Glimmer @ 38456 has strength = 15.63

b. GeneMark calls the start at 38633

c. GeneMark shows good coding potential in this region:



GeneMark mfolio shows two spikes coding potential in this region:



d. start sequence = ATG, stop sequence = TAA

e. downstream gap = 90 bp to gp56

Upstream gap = 173 bp sequence to gp58. A portion of this gap may be part of gp58

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 38456 | ParleG + 51 others | 56 | NCBI | q 1-129  t 1-129 | 100 | 8e-59 | 100 |
| 38633 | Etta + 6 others |  | NCBI | q 1-306  t 1-306 | 100 | 1e-156 | 100 |

f. ribosomal binding site

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| start | scoring matrix | spacing matrix | z score | final score | best score? |
| 38633 | Kibler6 | Karlin Medium | 1.844 | -5.386 | #2/6 |
| 38561 | Kibler6 | Karlin Medium | 1.486 | -5.690 | #3/6 |
| 38456 | Kibler6 | Karlin Medium | 1.925 | -5.337 | #1/6 |

g. Starterator: NA

h. longest possible ORF? yes, if start called at 38633

i. conclusion: this is a gene with start at 38633: a) GeneMarkS shows good coding potential in the region between 38400 and 38650, b) strong alignment scores on NCBI blast, c) the start at 38633 yields the longest ORF and includes all of the coding potential, d) 38633 was called by GeneMarkS; where GeneMarkS and Glimmer have disagreed with starts in this genome GeneMarkS has usually been making better calls, e) the start at 38633 creates a four base overlap GTGA with gp59.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene # | function | e-value |
| Winzigespinne | 57 | function unknown | 2e-54 |

3. BlastP (NCBI)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| Papafritta | 57 | unknown | 6e-67 | 99 | 99 |

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| no CDD hits |  |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| description | domain | *p* | e-val | % align |
| no hits > 90% *p* |  |  |  |  |

4. Transmembrane finders

a. deep TMHMM: gp57 is a globular non-membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp57 is a soluble protein

conclusion: the function of Vispistious\_57 is not known

Vispistious\_58 pham 807

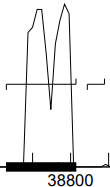
1. auto-annotated start/stop coordinates: 38630 ← 38815; reverse transcribed

2. length = 186 bp

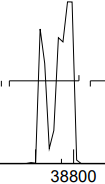
3. start data and coding potential

a. Glimmer @ 38456 has strength = 11.66

b. GeneMark shows two spikes of coding potential:



GeneMark mfolio shows two spikes of coding potential in this region:



d. start sequence = ATG, stop sequence = TGA

e. downstream overlap = 4 bp GTGA with gp57

Upstream gap = 1 base (G) sequence to gp59.

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id |
| 38815 | PuppyEggo | 58 | NCBI | q 1-186  t 1-129 | 100 | 3e-90 | 100 |

f. ribosomal binding site

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| start | scoring matrix | spacing matrix | z score | final score | best score? |
| 38815 | Kibler6 | Karlin Medium | 1.442 | -5.763 | #1/1 |

38815 is the only start available in this region of coding potential

g. Starterator: 38815 is the most annotated start

h. longest possible ORF? yes, if start called at 38815

i. conclusion: this is a gene with start at 38815: a) decent potential, b) good alignment scores on NCBI blast, c) the start at 38815 is the only available start.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. BlastP (phagesDB)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene # | function | e-value |
| WildNOut | 58 | function unknown | 3e-30 |

3. BlastP (NCBI)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phage name | gene # | function | e-val | % align | % id |
| MonChoix | 58 | unknown | 3e-34 | 96 | 93 |

4. BlastP (Conserved domains)

|  |  |  |  |
| --- | --- | --- | --- |
| domain hit | function | aligned residues | e-val |
| no CDD hits |  |  |  |

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| description | domain | *p* | e-val | % align |
| no hits > 90% *p* |  |  |  |  |

4. Transmembrane finders

a. deep TMHMM: gp58 is a globular non-membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp58 is a soluble protein

conclusion: the function if Vispistious\_58 is not known

Vispistious\_59 pham 58291

1. auto-annotated start/stop coordinates: 38817 ← 39548; reverse transcribed

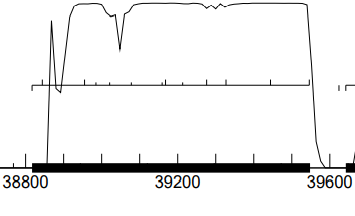
2. length = 732

2. coding potential

3. start data

a. Glimmer @ 39548 has strength 13.46

b. GeneMark



there is one start (39548) that includes all regions of coding potential

c. start sequence = ATG; stop sequence = TGA

d. upstream gap = 92 bases; downstream gap =1 base

e. Starterator: the start at 39548 is called 98% of the time when present

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| phage name | gene number | d-base | query: subject | % alignment | e-value | % identity | % similarity |
| Teagan | 59 | NCBI | 1-243 | 100 | 0.0E0 | 100 | 100 |

g. ribosomal binding site

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| scoring matrix | spacing matrix | z score | final score | best score? |
| Kibler6 | Karlin Medium | 3.059 | -2.253 | #1/12 |

h. longest possible ORF? yes

i. Vispistious\_59 is a gene with start at 39548

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. Blast results

a. phagesDB

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-val |
| 100 | WildNOut | unknown | 1e-134 |

b. NCBI

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene function | % alignment | e-value |
| Teagan | hypothetical protein | 100 | 0.0E0 |

d. was hit identified at the conserved domain database?

c. HHPred

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene number | % alignment | probability |
| no hits > 90% *p* |  |  |  |

note: % alignment = length of alignment/total gene length

4. Transmembrane finders

a. TMHMM: no transmembrane helices

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp59 is a soluble non-membrane protein

conclusion: the function of Vispistious\_59 is not known

Vispistious\_60 pham 881

1. auto-annotated start/stop coordinates: 39642 ← 39980; reverse transcribed

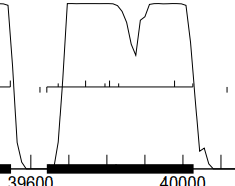
2. length = 387 bp

2. coding potential

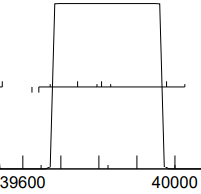
3. start data

a. Glimmer @ 39980 has strength = 19.94

b. GeneMark calls the start at 40028



GeneMarkS shows coding potential that includes the 40028 start



GeneMarkMfolio does not show this coding potential. The start at 40028 (TGA) would establish a 4-base overlap with upstream gp59.

c. start sequence = ATG; stop sequence = TAG (39980)/TGA(40028)

d. upstream gap = 54 bases (39980) or a 4 base overlap (40028); downstream gap = 6 bp base.

e. Starterator: the start at 40028 is the most annotated start for this pham

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| phage name | gene number | d-base | query: subject | % alignment | e-value | % identity | % similarity |
| Gargoyle | 60 | NCBI | 1-128 | 100 | 0.0E0 | 100 | 99.22 |

g. ribosomal binding site

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| start | scoring matrix | spacing matrix | z score | final score | best score? |
| 40028 | Kibler6 | Karlin Medium | 2.426 | -3.962 | #3/7 |
| 39980 | Kibler6 | Karlin Medium | 3.059 | -3.594 | #1/7 |

h. longest possible ORF? yes if start changed to 40028

i. Start conclusion: based on the available evidence the start was changed to bp 40028.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. phagesDB

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | WildNOut | unknown | 1e-69 |

3. NCBI

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-val |
| 35 | Teagan | unknown | 6e-88 |

4. CCD

no hits

3. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene number | % alignment | probability |
| no matches > 90% *p* |  |  |  |

note: % alignment = length of alignment/total gene length

4. Transmembrane finders

a. TMHMM: no transmembrane helices

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp60 is a soluble non-membrane protein

conclusion: the function of Vispistious\_60 is not known

Vispistious\_61 pham 731

1. auto-annotated start/stop coordinates: 40025 ← 40384; reverse transcribed

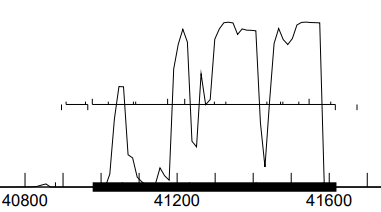
2. length = 360 bp

2. coding potential

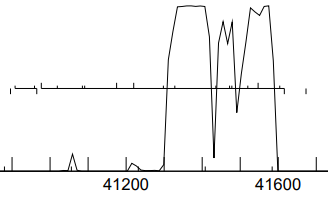
3. start data

a. Glimmer @ 40384 has strength = 3.654

b. GeneMark



GeneMarkS shows coding potential in the region 41000 → 41300 not indicated by GeneMark mfolio



GeneMarkMfolio does not show this coding potential.

c. start sequence = GTG; stop sequence = TGA (40025)/TGA(40384)

d. upstream overlap = 4 bases(GTGA); downstream overlap = 4 bp (GTGA).

e. Starterator: 40384 is the most annotated start for this pham

f. BLAST-start

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| phage name | gene number | d-base | query: subject | % alignment | e-value | % identity | % similarity |
| Ilzat | 61 | NCBI | 1-119 | 100 | 0.0E0 | 100 | 100 |

g. ribosomal binding site (scoring = Kibler6, spacing = Karlin medium)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 40384 | 2.053 | -5.239 | #4/8 |
| 40354 | 1.989 | -5.201 | #3/8 |

h. longest possible ORF? yes if start at 40384

i. conclusion: gp61 is a gene with start @ 40384

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. phagesDB

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | Zada | function unknown | 3e-60 |

3. NCBI

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene function | % alignment | e-value |
| Ilzat | hypothetical protein | 100 | 0.0E0 |

4. CCD

no hits

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene number | % alignment | probability |
| no hits > 90% likelihood |  |  |  |

4. Transmembrane finders

a. TMHMM: no transmembrane helices

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp61 is a soluble non-membrane protein

conclusion: the function of Vispistious\_61 is not known

Vispistious\_62 pham 930

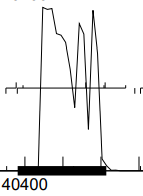
1. auto-annotated start/stop coordinates: 40381 ← 40563; reverse transcribed

2. length = 183 bp

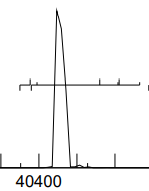
3. start data and coding potential

a. Glimmer @ 40563 has strength = 5.48

b. GeneMark calls the start at 40614



GeneMarkS shows coding potential not found by GeneMark mfolio



c. start sequence = GTG (40563)/ATG (40614) stop sequence = TGA

d. gap to gp63 = 613 bases(40563)/562 bases (40614); upstream overlap with gp61 = 4 bp (GTGA).

e. Starterator: 40614 is the most annotated start in this pham

f. BLAST-start

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene number | d-base | query: subject | % alignment | e-val | % id | % sim |
| 40563 | PuppyEggo | 62 | NCBI | q 1-60  t 18-77 | 77.9 | 8.8 x 10-37 | 100 | 100 |
| 40614 | AxiPup | 62 | NCBI | q 1-77  t 1-77 | 100 | 0.0E0 | 100 | 100 |

the start at 40614 is a superior fit with other phages in the NCBI database

g. ribosomal binding site (scoring = Kibler6, spacing = Karlin medium)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 40563 | 3.125 | -2.112 | #1/5 |
| 40611 | 1.915 | -5.615 | #4/5 |
| 40614 | 1.915 | -4.848 | #2/5 |

h. longest possible ORF? 40614 does not provide the longest ORF, but it does include all of the coding potential found by GeneMarkS

i. conclusion: gp62 is a gene with start at 40614

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. phagesDB

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | AxiPup | function unknown | 2e-40 |

3. NCBI

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene function | % alignment | e-value |
| Superfresh | hypothetical protein | 100 | 1 x 10-98 |
| 49 others with 100% alignment and e-value < 1 x 10-7 | | | |

4. CDD

no hits

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene number | % alignment | probability |
| no hits > 90% likelihood |  |  |  |

4. Transmembrane finders

a. TMHMM: no transmembrane helices

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp62 is a soluble non-membrane protein

conclusion: the function of Vispistious\_62 is not known

Explored the large gap between gp62 and gp63. Did not find evidence of a gene:

1) lack of coding potential

2) despite the large gap, there are not many places to put a gene here due to the numerous stops in this region

3) no hits on NCBI blastx in this region forwards or backwards.

Vispistious\_63 pham 67482

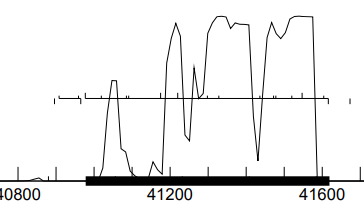
1. auto-annotated start/stop coordinates: 41176 → 41619; forward transcribed

2. length = 444 bp

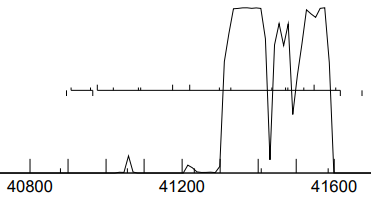
3. start data and coding potential

a. Glimmer @ 41176 has strength = 13.56

b. GeneMark calls the start at 40978



GeneMarkS shows coding potential at the downstream end of this gene not found by GeneMark mfolio. The start at 40978 would include coding potential on the left side of this gene excluded by a start at 41176.



c. start sequence = ATG (41176)/ATG (40978) stop sequence = TGA

d. upstream gap = 183 bp to the end of the genome sequence; downstream overlap with gp62 = 613 bp.

e. Starterator: 41176 is the most annotated start, but there is not clear convergence amongst the members of this diverse pham.

f. BLAST-start

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| start | phage name | gene # | d-base | query: subject | % align | e-val | % id | % sim |
| 41176 | PuppyEggo | 63 | NCBI | q 1-147  t 61-207 | 71 | 0.0E0 | 100 | 100 |
| 40978 | Joannes | 63 | NCBI | q 1-213  t 1-213 | 100 | 0.0E0 | 99.1 | 99.1 |

40978 provides superior alignment

g. ribosomal binding site (scoring = Kibler6, spacing = Karlin medium)

|  |  |  |  |
| --- | --- | --- | --- |
| start | z score | final score | best score? |
| 40978 | 2.004 | -5.168 | #8/15 |
| 40996 | 1.148 | -6.413 | #14/15 |
| 41020 | 2.604 | -4.535 | #6/15 |
| 41086 | 2.082 | -5.450 | #10/15 |
| 41092 | 1.735 | -5.173 | #9/15 |
| 41176 | 2.970 | -2.443 | #1/15 |

41176 is the clear rbs winner

h. Conclusion: Vis\_63 is a gene, but the location of its start is unclear. Going to go here with 40978 to include all of the coding potential shown by genemarkS and produce the longest possible ORF.

Gene function

1. Synteny

a. is synteny informative for this gene? no

b. upstream gene NA

c. downstream gene NA

2. phagesDB

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 100 | PuppyEggo | function unknown | 1e-78 |

3. NCBI

|  |  |  |  |
| --- | --- | --- | --- |
| n hits | best hit | function | e-value |
| 76 | PuppyEggo | function unknown | 2e-99 |

4. CDD

no hits

5. HHPred (only for genes > 90% probable matches)

|  |  |  |  |
| --- | --- | --- | --- |
| phage name | gene number | % alignment | probability |
| no hits > 90% likelihood |  |  |  |

4. Transmembrane finders

a. TMHMM: gp63 is a globular non-membrane protein

b. TOPCONs: no homologous transmembrane proteins detected

c. SOSUI: gp63 is a soluble non-membrane protein

conclusion: the function of Vispistious\_63 is not known

tRNA search

1) Aragorn found no evidence of tRNA sequences in the Vispistious genome

2) tRNAscan-SE found no evidence of tRNA sequences in the Vispistious genome

3) conclusion: Vispistious does not have tRNA’s encoded in its genome