Gingerbug phage. Oregon Coast Community College. Hope T. Poet, Chyanna G. Blackburn, Matthew R. Fisher.

**Gene 1**

-Original Glimmer call @bp 1 has strength 14.78

Start 1, Stop 447, Length 447 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Quite a few significant results (e values more negative than E-4) on phagedb & NCBI. Gene is in a pham w/ 45 members.

Starterator report: Does not have the most annotated start codon, but 6 of the 45 pham members call the same start codon. Start contains full coding potential and has best RBS score by far. Glimmer and GeneMark agree on start.

Function: helix-turn-helix DNA binding domain. HHPred: Lists as either unknown function, small terminase subunit, or helix-turn-helix with significant probability (higher than 90%). Blastp: helix-turn-helix DNA binding domain and terminase small subunit.

**Gene 2**

-Original Glimmer call @bp 444 has strength 1.97

Start 444, stop 707, Length 264 bp.

It is a gene: GeneMarkS indicates high coding potential in about half of the ORF. Blastp: No significant results on phagedb & NCBI. Gene is an orpham. If the gene were not there, the gap between gene 1 and gene 3 would be almost 300 nucleotides in length which would indicate that there would likely be a gene located in that gap.

Starterator report: No report as gene is an orpham. GeneMark and Glimmer agree on start. Start contains full coding potential, has the best RBS score, and gives gene common 4 bp overlap with upstream gene.

Function: Hypothetical Protein. HHPred: no significant results. Blastp: no significant results. SOSUI: soluble protein.

**Gene 3**

-Original Glimmer call @bp 704 has strength 14.67; GeneMark calls start at 689

Start 704, stop 853, Length 150 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: no significant results on phagedb & NCBI. Gene is an orpham. There would be a very large gap between genes 2 and 4 if the gene was not present.

Starterator report: No report as gene is an orpham. GeneMark and Glimmer do not agree on start. Glimmer calls start at 704 while GeneMark calls start at 689. With GeneMark start, overlap with gene 2 would be 19 bp. Overlap for the Glimmer start is 4 bp. 704 Glimmer start has a much better RBS score.

Function: Hypothetical Protein. HHPred: no significant results. Blastp: no significant results. SOSUI: soluble protein.

**Gene 4**

-Original Glimmer call @bp 1048 has strength 12.25

Start 1048, stop 1347, Length 300 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: significantly related to phages in cluster on phagedb & NCBI. Gene is in a pham w/ 3 members.

Starterator report: All three members of pham call this start. GeneMark and Glimmer agree on start. Start contains full coding potential. 195 bp gap between gene 4 start and gene 3 stop, however, if a new gene were there in that gap, it would have a very large overlap with gene 4.

Function: Hypothetical Protein. HHPred results: No significant results (none greater than or equal to 90%). Blastp: Related most to its clusters, whose protein function is unknown. SOSUI: is a soluble protein.

**Gene 5**

-Original Glimmer call @bp 1344 has strength 16.67

Start 1344, Stop 2693, Length 1350 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Many significant results with cluster members having e values of 0 on phagedb and NCBI. Gene is in pham w/ 48 members.

Starterator report: Does not have the most annotated start codon; start codon that is called is shared by 26.1% of genes in the pham which is the second highest percentage (behind start 28 which Gingerbug does not have). GeneMark and Glimmer agree on start. Gene gives 4 bp with gene 4. Start contains full coding potential and has 100% alignment with DelaGarza in its pham. The predicted start is the longest without excessive overlap, if the start codon before it was called the overlap would be 64 nucleotides which is well over the maximum 30 bp overlap.

Function: terminase. HHPred: many significant results (over 90%) for terminase. Blastp: results for terminase with low e values.

**Gene 6**

-Original Glimmer call @bp 2784 has strength 13.80

Start 2784, Stop 4424, Length 1641 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Many significant results with pham members having e values of 0 in phagedb & NCBI. Gene in a pham w/ 48 members.

Starterator report: Does not have the most annotated start codon but six other genes in the pham call the same start. GeneMark and Glimmer agree on start. It is the first start codon after the stop codon, so a different codon would have to be called behind it which would cause the start to not contain all of the coding potential. Start has 96% alignment with pham and cluster member DelaGarza and has the best RBS score.

Function: Portal Protein. Blastp: portal protein (e values close to or at 0). HHPred: many significant results for portal protein. The probability is 100% that the function is a portal protein.

**Gene 7**

-Original Glimmer call @bp 4424 has strength 11.83

Start 4424, Stop 5527, Length 1104 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Many significant results with many e values close or at 0 on phagedb & NCBI. Gene in a pham w/ 18 members.

Starterator report: Calls the most annotated start, almost 40% of genes in the pham call this start. GeneMark and Glimmer agree on start. Start contains full coding potential, has 100% alignment to pham and cluster member DelaGarza. The called start codon does not have the best RBS score, it still has a low score and the differences to the other score are very small (4424 RBS: -4.372 to start 10 (4931): -4.335), choosing start 10 would lose coding potential.

Function: minor capsid protein.HHPred: 100% probability for minor capsid protein. Other high probability is for a Mu protein that could be a minor head protein. Blastp: Many significant results in both phage database and NCBI that function is minor capsid protein. Also, a good portion of significant results list function as capsid maturation protease.

**Gene 8**

-Original Glimmer call @bp 5806 has strength 11.46

Start 5806, stop 6582, Length 777 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Many significant results. Gene is in a pham w/ 9 other members.

Starterator report: Does not have the most annotated start, but 3 out of 9 members call the same start. GeneMark and Glimmer agree on start. 280 bp gap between gene 7 and gene 8, but there is no coding potential in this gap. Called start contains full coding potential, has about a 71% alignment with cluster and pham member DelaGarza, and has best RBS score.

Function: scaffolding protein**.** HHPred: 96.53% probability that the function is minor structural, two other high results for scaffolding protein and domain of unknown function. Blastp: scaffolding protein. SOSUI: is a soluble protein.

**Gene 9**

-Original Glimmer call @bp 6620 has strength 20.46

Start 6620, Stop 7483, Length 864 bp.

It is a gene: GeneMarkS indictes high coding potential. Blastp: Many significant results on phagedb blast, e values for the genes of DelaGarza and Teng both e-144, many significant results on NCBI blast, e values at 0. Gene in pham w/ 31 members.

Starterator report: Calls the most annotated start as start, at the time of the report 2/14/22, 30 out 30 members call the start. GeneMark and Glimmer agree on start. Start contains full coding potential, has 100% alignment to pham and cluster member DelaGarza, and has best RBS score.

Function: Major capsid protein. HHPred: Over 99% probability for many results that function is major capsid protein. Blastp: major capsid protein.

**Gene 10**

-Original Glimmer call @bp 7487 has strength 11.68

Start 7487, Stop 7969, Length 483 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Gene is only significantly related to genes in cluster, DelaGarza and Teng. Gene in pham w/ 3 members.

Starterator report calls same start codon as 100% of the members of its pham. GeneMark and Glimmer agree on start. Start contains full coding potential and has 100% alignment with DelaGarza. Any other start would drastically shorten the gene and create a large gap.

Function: Hypothetical protein. HHPred: No significant results. Blastp: NCBI and database results that are significant show function as function unknown. SOSUI: soluble protein.

**Gene 11**

-Original Glimmer call @bp 8045 has strength 17.17

Start 8,045, Stop 8,794, Length 750 bp.

It is a gene: GeneMarkS shows high coding potential. Pham has 3 members. There is a gap of 86 base pairs between the stop of gene 10 and the start of gene 11. No other genes could fill this so we stick with the start codon provided.

Function: head-to-tail adaptor. HHPred: shows top 5 results with a >90% probability with the top probability showing function for Putative head-tail connector protein; has matches to SPP1 15 and Gp6, as stipulated in official function list. BlastP shows an E Value of 0.0 and an alignment of 1 to 1 to the DelaGarza gene showing that a head to tail adaptor is the main function.

**Gene 12**

-Original Glimmer call @bp 8787 has strength 10.87

Start 8,787, Stop 9,137, Length 351 bp.

It is a gene: GeneMarkS shows high coding potential. Pham has 8 members. Gene 12 has an 8 base pair overlap with the stop on gene 11. Start contains all coding potential, any other start would exclude coding potential.

Function: Head-to-tail stopper. HHPred: minor capsid protein, head-tail adaptor, stopper protein, head-tail joining protein, head completion protein, etc; match to SPP1 GP 16, as called out in the official gene function list, which is needed to justify head to tail stopper. BlastP: shows an E-Value of 0.0 and an alignment of 1 to 1 to the DelaGarza gene showing that a head to tail stopper is the main function.

**Gene 13**

-Original Glimmer call @bp 9134 has strength 16.48; GeneMark calls start at 9140

Start 9,134, Stop 9,493, Length 360 bp.

It is a gene: GeneMarkS shows high coding potential. Pham has 123 members. There is a 4 base pair overlap with the stop on gene #12 and the start codon on gene #13. The start contains all coding potential.

Function: minor capsid protein. HHPred: minor capsid protein is the top result, but other 2 results above 90% probability are “putative tail component” and “phage\_tail”. Has a match to GP 17, which is support for tail terminator protein. BlastP: shows an E-Value of 0.0 and an alignment of 1 to 1 to the DelaGarza gene showing that a minor capsid protein is the main function.

**Gene 14**

-Original Glimmer call @bp 9495 has strength 11.04

Start 9,495, Stop 9,923, Length 429 bp.

It is a gene: GenemarkS shows high coding potential. Pham has 94 members. Start includes most coding potential, extending the upstream start would produce huge overlap with gene 13 and would include large regions with no coding potential.

Function: tail terminator. HHPred: top results is for minor capsid protein but rest of top results (>90% probability) are for tail proteins; match to SPP1 GP17, which function list says is needed to justify tail terminator. BlastP: tail terminator.

**Gene 15**

-Original Glimmer call @bp 9935 has strength 14.90

Start 9,935, Stop 10,438, Length 504 bp.

It is a gene: Pham has 113 members. Start includes all coding potential, is the first start in ORF, moving it downstream would exclude coding potential.

Function: Major tail protein. HHPred: shows top 12 results with a 90% or higher probability with the top result showing function as major tail protein. BlastP: several matches for major tail protein.

**Gene 16**

-Original Glimmer call @bp 10503 has strength 10.82

Start 10,503, Stop 10,961, length 459 bp.

It is a gene: GeneMarkS shows high coding potential. Pham has 133 members. Start includes all coding potential. Moving it upstream would include no additional coding potential and would substantially overlap with gene 15. Moving start downstream would exclude areas of coding potential.

Function: tail assembly chaperone. HHPred: no results over 90% probability. 58% match to tail assembly chap, which is top result. BlastP: multiple matches to tail assembly chaperone.

**Gene 17**

-Original Glimmer call @bp 11042 has strength 3.95 \*\* not called by GeneMark

Start 10,503, Stop 11,179, length bp: 678

I determined that this has 2 regions: 10503-10874 and 10874-11179. -1 frameshift at 10874. Has GGGAAAA sequence, frameshift at the first A at 10874 results in GK to GK shift.

I used closely related Appa phage as reference. GGGAAAA sequence was found at all of the following highlighted residues, which is where frameshift occurs in Appa:

>Appa gp19 (TAC 1)- pham 3816

MKHQEDEMTDLGQMLDFMEDEDGFKSPPMPSKAHPEGRQYQVASPDAKTGLRLNALADIMVKRENKVPVSDLDVARLRLDGDAEQEFVRSVLGAERNDGGNALDEMIADGCKWEHIKRMSQYAFTYFGVSPEAATEAAKNGLLSGKLVAPNRAARRKKPKKK

>Appa gp20 (TAC 2)

MKHQEDEMTDLGQMLDFMEDEDGFKSPPMPSKAHPEGRQYQVASPDAKTGLRLNALADIMVKRENKVPVSDLDVARLRLDGDAEQEFVRSVLGAERNDGGNALDEMIADGCKWEHIKRMSQYAFTYFGVSPEAATEAAKNGLLSGKTRGAEPGGATQEAQEEVIPGVQVNDWMPRNVTDPRSMKSRGWIGPMLDNWGLVELDFQQVYGIDLSTPGLLQSRTWRWMTMRMRGLLSTEGRIQRVLNPTPEQKKADAKRRA

>Gingerbug\_Draft gp16 – pham 3816

MPFVDLSEFLTENDLVMQGLGPRDYVVPAPDVDTGLRYTALSNVAIKVNKGEAVSPEEMASLKLEGAEERAFVDQILSEEVVRQMADDGLSWPAVVRCANYAFTHFAVSPEAAKKAFEAGVFSGKALAPTNRAARRGTPSAPRASRASKNPR

GeneMarkS shows good coding potential where the frameshift would occur.

Function: tail assembly chaperone. HHPred: 95% probably “Phage\_Gp15 ; Bacteriophage Gp15 protein”. Next result is tail assembly chaperone at 52%. BlastP: several significant results for tail assembly chaperone.

**Gene 18**

-Original Glimmer call @bp 11192 has strength 15.15

Start 11,192, Codon 13,474, length 2,283 bp.

It is a gene: GeneMarkS shows high coding potential. Pham has 18 members. This start is the only one that includes all coding potential. GeneMark and Glimmer agree on start.

Function: tape measure protein. HHPred: shows top 4 results with a 90% or higher probability, with the top result showing the tape measure protein as the function. BlastP: multiple results for tape measure.

**Gene 19**

-Original Glimmer call @bp 13474 has strength 17.17

Start 13474, Stop 14337, Length 864 bp.

It is a gene: GeneMarkS indicates high coding potential. GeneMark and Glimmer agree on start. Pham: 190 members. 1 bp overlap with gene 18.

Starterator report: Has the most annotated start codon but does not call it, 26.7% of pham members have the start and those that have it call it 98% of the time, most annotated start is at Gingerbug start 13471. All coding potential contained by 13474 start codon. 100% alignment with DelaGarza. The RBS score is better for 13474.

Function: minor tail protein. HHPred: Most significant result for distal tail protein. BlastP: Many significant results for minor tail protein.

**Gene 20**

-Original Glimmer call @bp 14337 has strength 15.19

Start 14337, Stop 16679, Length 2343 bp.

It is a gene: GeneMarkS indicates high coding potential. GeneMark and Glimmer agree on start. Pham: 185 members. Gives 1 bp overlap with upstream gene.

Starterator report: Does not have most annotated start, Gingerbug is the only pham member that has this start. Start codon contains all coding potential. Glimmer and gene mark agree on start codon. Start is only one to include all coding potential.

Function: minor tail protein. HHPred: a significant result for tail protein. Blastp: significant results for being a minor tail protein on phage database and NCBI.

**Gene 21**

-Original Glimmer call @bp 16692 has strength 6.68

Start 16692, Stop 16979, Length 288 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Multiple results greater than 3e-13 on phage database blast, NCBI blast shows significant results as well with e values close to 0. Pham: 21 members.

Starterator report: Calls the most annotated start. Start codon contains all of the coding potential. Glimmer and GeneMark agree on start. Start has the best RBS score.

Function: Hypothetical protein. HHPred: no significant results. Blastp: All significant results for database listed as function unknown, hypothetical protein according to NCBI blast. SOSUI: Is a soluble protein.

**Gene 22**

-Original Glimmer call @bp 16985 has strength 14.56

Start 16985, Stop 17878, Length 894 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Many significant results according to phagedb blast (multiple e values greater than 6e-35), NCBI blast also shows significant results with e values essentially 0. Pham: 98 members.

Starterator Report: calls most annotated start codon which is called 100% of the time when it is present. Start codon contains all of the coding potential. Glimmer and GeneMark agree on start. Called start is the most upstream, it does not have the best RBS score, but picking a further back start codon would lose coding potential.

Function: minor tail protein. HHPred: receptor binding protein of phage tail base plate. Blastp: minor tail protein. Based on HHpred, the most specific function should be a base-plate head protein, but that’s not on the official function list (yet?).

**Gene 23**

-Original Glimmer call @bp 17954 has strength 9.27

Start 17888, Stop 18721, Length 834 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Many results on phagedb blast less than 5e-10, NCBI blast also shows significantly related results. Pham: 93 members.

Starterator report: has the most annotated start but does not call it, only Gingerbug has the called start. Start codon does contain all regions of high coding potential but is missing some minor coding potential. Glimmer and GeneMark agree on start. However, the RBS score is 100 times better for the first start codon after the stop codon, 17888 (a start at 17888 would narrow the gap from 76 to 11 bp), choosing 17888 would give Gingerbug the most annotated start.

Function: minor tail protein. HHPred: Many, many results with > 90% probably, many stating cell adhesion molecules, fibronectin domain. If fibronectin is “collagen -like” then that meets the criteria on the official function list for minor tail protein. Blastp: mostly minor tail protein.

**Gene 24**

-Original Glimmer call @bp 18714 has strength 10.71

Start: 18714, Stop: 18956, Length: 243 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: No results in phage database, one result for the NCBI blast but the e value is 4.35 and is not significant. Is an orpham. Gives 8 bp overlap with gene 23.

No starterator report as is an orpham. Called start codon is the first start after the stop and contains all the coding potential, the next possible start at 18726 would also contain all of the coding potential and would cause a 6 bp gap between gene 24 and gene 23 but it has a much lower RBS score. Glimmer and GeneMark agree on start. Has the best RBS score.

Function: hypothetical protein. HHPred: no significant results (90% probability or higher). Blastp: no results that are significant. SOSUI: is a soluble protein.

**Gene 25**

-Original Glimmer call @bp 18946 has strength 15.98; GeneMark calls start at 18958

Start 18946, Stop 19185, Length: 240 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Many significant results.Pham: 62 members. Gives 11 base pair overlap with gene 24.

Starterator report: Calls most annotated start codon, 61.3% of genes have the start with it being called 86.8% when it is present. Start codon contains all of the coding potential. Glimmer and GeneMark do not agree on start, GeneMark calls start at 18958. Start 18958 is only present in 3 other phage genes according to starterator report and is not called by any of them. The RBS score for the 18946 start is 1000 times better than the start at 18958, Starting the gene at 18958 would cause a 3 base pair gap and would not cause any loss of coding potential.

Function: hypothetical protein. HHPred: only significant probability (90 percent or higher) for family of unknown protein Blastp: hypothetical protein. SOSUI: soluble.

**Gene 26**

-Original Glimmer call @bp 19185 has strength 13.65

Start 19185, Stop 20060, Length 876 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: many significant results. Pham: 117 members.

Starterator report: Has most annotated start, 42.7% of genes in the pham have this start; when the genes have the start, they call it 98.0% of the time. According to GeneMarkS, start codon contains all of the coding potential. GeneMark and Glimmer agree on start. RBS score is about 10000 times better than the next start at 19230.

Function: endolysin. HHPred: peptidases and endolysins. Blastp: endolysin, lysin A. From official function list, “some arthrobacter and streptomyces phages have a single endolysin with domains not found in the Mycobacteriophages (like the CHAP domain). use "endolysin" rather than lysin A if the phage does not infect Mycobacterium and no lysin b can be identified.” I think this means that we would use endolysin because our host bacteria are *Microbacteria* and we found no lysin B.

**Gene 27**

-Original Glimmer call @bp 20129 has strength 11.40; GeneMark calls start at 20141

Start 20129, Stop 20503, Length 375 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: multiple significant results Pham: 15 members.

Starterator report: does not have the most annotated start, has a start that is found in 3 genes, start is called 100% of the time when it is present. Start codon contains all of the coding potential. GeneMark calls start at 20141. RBS score for 20129 is 1000 times better than GeneMark’s call; GeneMark start would lose small amount of the coding potential and would cause an even larger gap between gene 27 and gene 26.

Function: membrane protein. HHPred: highest probability 96.76% for membrane protein, second highest probability 95.49% for protein of unknown function. Blastp: all significant results (multiple results greater than 1e-09) show function as hypothetical protein, NCBI top results shows DelaGarza function as membrane protein, others as hypothetical protein. SOSUI: membrane protein (1 domain). TMHMM: membrane protein (1 domain).

**Gene 28**

-Original Glimmer call @bp 20500 has strength 12.41

Start 20500, Stop 21243, Length 744 bp.

It is a gene: GeneMarkS indicates high coding potential, mostly. Strangely, there is a ~200 bp region within the gene with no coding potential – in this exact spot in a reverse frame, there is high coding potential – weird. Blastp: several significant results on phage database blast (many results less than 6e-18. Pham: 12 members.

Starterator report: Does not have the most annotated start codon, 25% of the pham genes have this start, and it is called 100% of the time when it is present. GeneMark and Glimmer agree on start. Start codon gives the preferred 4 bp pair overlap.

Function: Hypothetical protein. HHPred: Only significant result (90% or higher) shows function as a structural protein, contractile protein. Blastp: Most significant results show hypothetical protein, but other close results show lysin B, others show minor tail protein; NCBI lists DelaGarza as a membrane protein although the phage database lists the function of DelaGarza as function unknown. SOSUI: soluble. TMHMM: 1 transmembrane domain.

**Gene 29**

-Original Glimmer call @bp 21256 has strength 15.02

Start 21256, Stop 21753, Length 498 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: A few results significant results. In a pham of 9 members.

Starterator report: Does not have most annotated start, calls start present in 33% of genes in the pham, this start is called 100% of the time that it is present. GeneMark and Glimmer agree on start. Start contains full coding potential. Called start is first start and has the best RBS score.

Function: holin. HHPred: 96% and 95% matches to phage holin. Blastp: phage db shows hypothetical protein or holin, NCBI shows membrane protein or hypothetical protein. SOSUI: membrane protein. Holin is an official function of the list but to be a holin protein it “must be adjacent to lysins and have multiple transmembrane domains.” Endolysin was at gene 26, so it’s nearby and SOSUI & TMHMM: 4 transmembrane domains.

**Gene 30**

-Original Glimmer call @bp 22165 has strength 14.00

Start 22165, Stop 21857, Length: 309 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Two significant results with DelaGarza e value 1e-23 and Teng e value 3e-23, NCBI shows one significant result DelaGarza (Teng is still a draft). In a pham with two other members.

Starterator report: Does not have most annotated start called by the only other members of the pham. Called start contains full coding potential. GeneMark and Glimmer agree on start. Called start has highest RBS score The next start would be 22123 and would cause a 44 bp gap between gene 30 and gene 31 would exclude a tiny amount of coding potential.

Function: hypothetical protein. HHPred: Lists no significant results (90% probability or higher). Blastp: only significant results for hypothetical protein. SOSUI: soluble protein.

**Gene 31**

-Original Glimmer call @bp 22564 has strength 8.17

Start 22564, Stop 22166, Length 399 bp.

It is a gene: GeneMark S indicates high coding potential, Blastp: Many significant results. Pham: 366 members.

Starterator report: Does not have most annotated start, 6.6% of genes in the pham have the start called by Gingerbug and this start is called 12.5% of the time that it is present. GeneMark and Glimmer agree on start. Start codon contains full coding potential. Called start has best RBS score and gives 1 bp overlap with gene 32. The start before the called start would cause a larger overlap with gene 32

Function: HNH endonuclease. HHPred: Many significant results (90% and higher) for HNH endonuclease. Blastp: HNH endonuclease.

**Gene 32**

-Original Glimmer call @bp 22860 has strength 8.84; GeneMark calls start at 22893

Start 22893, Stop 22564, Length 330 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Two significant results. Pham: 3 members.

GeneMark calls at 22893, Glimmer calls at 22860. Starterator report: The two other members call start 2 while Gingerbug calls start 3; start 2 is at the GeneMark call 22893. RBS score is better for the start at 22860. A start at 22860 would cause a gap of 31 bp, a start at 22893 would cause an overlap of 4 bp with the downstream gene (gene 33)! Start at 22860 does contain all but a tiny portion of weak coding potential.

Function: membrane protein. HHPred: no significant results, Blastp: hypothetical protein, SOSUI: membrane protein (one domain) called. TmHmm: membrane protein (1 domain)

**Gene 33**

-Original Glimmer call @bp 23156 has strength 19.26

Start 23156, Stop 22890, Length: 267 bp.

It is a gene: GeneMarkS indicates high coding potential Blastp: several significant results. Pham: 4 members.

Starterator report: calls most annotated start which is found in 3 out of 4 pham members and is called 100% of the time that it is present. GeneMark and Glimmer agree on start. Overlap/gap with downstream gene: 87 bp gap with downstream gene (gene 34), changing the start codon to 23234 would decrease the gap. The RBS score is -2.645 for the start at 23156 and is -6.351 for the start at 23234. Start contains full coding potential.

Function: Hypothetical protein. HHPred: no significant results (90% or higher). Blastp: hypothetical protein. SOSUI: soluble protein.

**Gene 34**

-Original Glimmer call @bp 23484 has strength 15.99

Start: 23,484 Stop: 23,242, length 243 bp.

This is a gene: Glimmer calls bp at 23,484. GenemarkS agrees. GenemarkS shows high coding potential. Pham w/ 51 members. 156 base pair gap between the stop codon on gene 34 and the start codon on

gene 35. No other genes could fill this gap. The start’s RBS score is 1000x better than any other start that would extend the gene and therefore reduce the gap – these starts do not encode any extra coding potential.

Function: hypothetical protein. HHPred shows no significant probability. All results under 90%. BlastP: Results were insignificant. SOSUI: soluble.

**Gene 35**

-Original Glimmer call @bp 24160 has strength 7.02

Start: 24,160, Stop: 23,630, length 531 bp.

This is a gene: GenemarkS shows high coding potential. Pham has 12 members. 4 bp overlap with gene 36. Start contains all coding potential.

Function: winged helix-turn-helix DNA binding domain. HHPred: 99% probability “MarR family regulatory protein; Winged helix-turn-helix, DNA binding protein”. It often refers to a function for a transcriptional

regulator. BlastP: hypothetical protein, MarR fam transcript. regulator, & DNA binding protein.

**Gene 36**

Original Glimmer call @bp 24786 has strength 13.90

Start 24,786, Stop: 24,157, length 630 bp.

This is a gene: GenemarkS shows high coding potential. Pham w/ 38 members. There is a 1 base pair overlap gene 37. The start contains all of the coding potential.

Function: hypothetical protein. HHPred shows no significant results for known functions. BlastP: hypothetical protein. SOSUI: soluble.

**Gene 37**

-Original Glimmer call @bp 25076 has strength 15.99

Start 25,076, Stop 24,786, length 291 bp.

This is a gene: GenemarkS shows high coding potential. Blastp: a couple significant results. Pham: 62 members. The start is the only one that contains all coding potential, and it has one of the best RBS scores.

Function: hypothetical protein. HHPred shows no significant results for main function. BlastP: hypothetical protein. SOSUI: soluble.

**Gene 38**

-Original Glimmer call @bp 25270 has strength 11.95; GeneMark calls start at 25282

Start: 25,282, Stop: 25,091, length 192 bp.

This is a gene: GenemarkS shows high coding potential. Pham has 3 members. Blastp NCBI: significant similarity to DelaGarza.

Glimmer calls start 25,270. GenemarkS calls base pairs at 25,282. 25,282 has a worse RBS score but it creates a 4 bp overlap and includes a small amount of extra coding potential. The other 2 genes in this pham use the 25282 start.

Function: hypothetical protein. HHPred shows no significant results. BlastP: hypothetical protein. SOSUI & TMHMM: soluble.

**Gene 39**

-Original Glimmer call @bp 25506 has strength 11.76

Start: 25,506, Stop: 25,279, length: 228 bp.

This is a gene: GenemarkS shows high coding potential. Orpham. The start has best RBS score and contains all coding potential within the ORF.

Function: hypothetical protein. HHPred shows no significant results. BlastP: shows no significant results. SOSUI & TMHMM: soluble.

**Gene 40**

-Original Glimmer call @bp 25627 has strength 7.61

Start: 25,627, Stop: 25,851, length: 225 bp.

This is a gene: GenemarkS shows high coding potential. Pham has 2 members. The start is the first in the ORF and the only to include all coding potential.

Function: helix-turn-helix DNA binding domain. HHPred shows many results with a probability of 90% or higher, functions include DNA binding protein, HTH, terminase small subunit, recombinase, etc. HTH seems the most common. blastP: no significant results. SOSUI: soluble.

**Gene 41**

-Original Glimmer call @bp 26445 has strength 11.21

Start: 26,445, Stop 26,023, length 423 bp.

This is a gene: GenemarkS shows high coding potential. Pham has 346 members. The start is the first and it is the only that includes all coding potential.

There is a 172 base pair gap between the stop codon on gene 40 and the start codon on gene 41. There is some high coding potential in the -3 frame but that would overlap too much with gene 40 (~100 bp).

Function: RusA-like resolvase (endonuclease). HHPred: 99% RusA endodeoyribonuclease. BlastP: RusA-like resolvase.

**Gene 42**

-Original Glimmer call @bp 27000 has strength 11.42

Start: 27,000, Stop: 26,503, length 498 bp.

This is a gene: GenemarkS shows high coding potential. Pham: 451 members. Start includes all coding potential and has the best RBS score.

Function: DNA binding protein. HHPred: DNA binding protein. BlastP: single-stranded DNA binding protein.

**Gene 43**

-Original Glimmer call @bp 27395 has strength 10.33

Start: 27,395, Stop: 27,042, length: 354 bp.

This is a gene: GenemarkS shows high coding potential. Pham: 17 members. Start contains all coding potential and gives a 4 bp overlap with gene 44.

Function: helicase loader. HHPred shows top 4 results with 90% or higher probability with the main

function as a helicase loader. BlastP: e-value of 0.0 and an alignment of 1:1 with the DelaGarza gene for the main function as helicase loader.

**Gene 44**

-Original Glimmer call @bp 28129 has strength 6.09

Start 28129, Stop 27392, Length 738 bp.

It is a gene: GeneMarkS shows good coding potential. Pham: 3 members.

Starterator report: Calls most annotated start which is present and called 100% of the time when present. GeneMark and Glimmer agree on start. Start is the first in the ORF and contains all of the coding potential.

Function: helix-turn-helix DNA binding domain. HHPred: shows DNA replication protein and helix-turn-helix. Blastp: top results on phage db for helix-turn-helix DNA binding domain protein, NCBI shows multiple results for helix-turn-helix domain-containing protein.

**Gene 45**

-Original Glimmer call @bp 28588 has strength 13.36; GeneMark calls start at 28594

Start: 28588, Stop: 28232, Length 357 bp.

It is a gene: GeneMarkS shows high coding potential. Blastp: only two significant results phagedb, NCBI shows one significant result DelaGarza (Teng is still a draft.). Pham: 3 members.

Starterator report: Calls most annotated start which is called by all 3 members of the pham. Glimmer calls start at 28588, GeneMark calls start at 28594. Having the start at 28594 would make the gap smaller between gene 45 and 46 (95 bp gap instead of 101 bp gap). Start codon contains all coding potential, except perhaps a tiny sliver of low coding potential. The RBS score for the start at 28588 is better than the RBS score of 28594 (RBS 28588: -4.275; RBS 28594: -5.888). 28588 gives 1:1 alignment with DelaGarza & Teng. The called start is the best start.

Function: hypothetical protein. HHPred: Top significant results over 99% probability show function as hypothetical protein. Blastp: hypothetical protein. SOSUI: Is a soluble protein.

**Gene 46**

-Original Glimmer call @bp 28689 has strength 17.18; GeneMark calls start at 28638

Start 28689, Stop 29009, Length 321 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Only two other significant, e values are 1e-58 and 4e-58. Pham: three members.

Starterator report: Calls the most annotated start; all 3 genes in pham call the start. Glimmer calls start 28689, GeneMark calls start 28638. Called start 28689 contains all regions of high coding potential. RBS score for 28689 start is -3.514, RBS score for 28638 is -6.978, this start contains some additional areas of very low coding potential. The called start is the best start.

Function: hypothetical protein. HHPred: No significant results (90% probability or higher). Blastp: hypothetical protein. SOSUI: Is a soluble protein.

**Gene 47**

-Original Glimmer call @bp 29006 has strength 14.99

Start 29006, Stop 29167, Length: 162 bps.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Two significant results, e values of 9e-21. Pham: 3 members.

Starterator report: Calls most annotated start, 100% of phage genes in the pham call this start. Glimmer and GeneMark agree on start. Start codon contains all coding potential. 4 bp overlap with gene 46.

Function: membrane protein. HHPred: only significant result (90% or higher) is for a domain of transmembrane proteins. Blastp: Only significant results show function unknown. SOSUI: membrane protein (2 domains). TMHMM: membrane protein (2 domains).

**Gene 48**

-Original Glimmer call @bp 29164 has strength 13.82

Start 29164, Stop 29382, Length 219 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Two significant results, e values for both are 2e-35. Pham: 3 members.

Starterator report: calls most annotated start, 100% of pham members call this start. Glimmer and GeneMark agree on start. Start codon is the only that contains all coding potential. 4 bp overlap with gene 47.

Function: hypothetical protein. HHPred: no significant results (90% or higher probability. Blastp: only significant results show function unknown (hypothetical protein). SOSUI: is a soluble protein.

**Gene 49**

-Original Glimmer call @bp 29379 has strength 18.61

Start 29379, Stop 30539, Length 1161 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Multiple results less than 9e-05. Pham: 3 members.

Starterator report: Calls most annotated start, called by 100% of phage genes. Glimmer and GeneMark agree on start. Start codon contains full coding potential and gives a 4 bp overlap with gene 48.

Function: exonuclease. HHPred: 99.5% probability lambda exonuclease. Blastp: 100% aligned and 94% similarity with DelaGarza, whose function is exonuclease.

**Gene 50**

-Original Glimmer call @bp 30544 has strength 17.75

Start 30544, Stop 31479, Length 936 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Multiple results less than 6e-12. Pham: 3 members.

Starterator report: Calls most annotated gene, 100% of phage genes call this start. Glimmer and GeneMark agree on start. Start codon is the first in the ORF, contains full coding potential. While there is a start codon that has a better RBS score it would not include a good portion of the coding potential.

Function: RecA-like DNA recombinase. HHPred: high probability for Rad52 recombinase, ssDNA binding, Rad52\_Rad22 family protein, etc. Blastp: top significant result DelaGarza shows function as RecA-like DNA recombinase, other high results are ssDNA binding protein, hypothetical protein, and DNA recombinase. NCBI also lists Rad52/Rad22 family DNA repair protein. Rad52 seems to be related to RecA-like, so listing that as function.

**Gene 51**

-Original Glimmer call @bp 31479 has strength 12.93

Start 31476, Stop 31919, Length 444 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: A few significant results less than 8e-04. Pham: 3 members.

Starterator report: Calls most annotated start as does 100% of the genes in its small pham. GeneMark and Glimmer agree on start. Start codon contains all of the coding potential. BUT there is a start at 31476 that has a slightly better RBS score than the start at 31479 and this start would create an overlap of 4 bps with gene 50, which is preferred. 31476 is therefore chosen.

Function: hypothetical protein. HHPred: No significant results (90% or higher probability). Blastp: All significant results list as function unknown (hypothetical protein). SOSUI: soluble protein.

**Gene 52**

-Original Glimmer call @bp 31916 has strength 15.70; GeneMark calls start at 31973

Start 31916, Stop 32314, Length 399 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Two significant results. Pham: 3 members.

Starterator report: calls most annotated start, 100% of genes in the pham call this start. GeneMark calls start at 31973, Glimmer calls at 31916. Start codon 31916 contains most but not all coding potential and produces 4 bp overlap with gene 52. Start 31973 would exclude lots of coding potential.

Function: Hypothetical protein. HHPred: no significant results. Blastp: only significant results show function unknown (hypothetical protein). SOSUI: soluble protein.

**Gene 53**

-Original Glimmer call @bp 32316 has strength 6.37

Start 32316, Stop 32486, Length 171 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: a few significant results. Pham: 22 members.

Starterator report: Does not have the most annotated start, only Gingerbug calls this start. Start is the first in the ORF and contains all coding potential. RBS is better than the next downstream start, which excludes some coding potential.

Function: hypothetical protein. HHPred: Several significant results (90% or higher) for the function being related to zinc binding, also functions such as RNA polymerase, NADH pyrophosphatase, etc. Blastp: hypothetical protein. SOSUI: soluble protein.

**Gene 54**

-not called by Gimmer or GeneMark

Start 32483, Stop 32674, length 192 bp.

This is a gene: GeneMarkS shows some moderate and high coding potential. Blastp: no significant similarities. Likely an orpham. This fills what would otherwise be a 233 bp gap between 53 and 55.

Start contains all coding potential and has a good RBS score: -3.49. 4 bp overlap with gene 53.

Function: hypothetical protein. HHPred and blastp: no significant results. SOSUI & TMHMM: soluble

**Gene 55**

-Original Glimmer call @bp 32718 has strength 16.49

Start 32718, Stop 32927, Length 210 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: One significant result on phagesdb, e value of 1e-31, no significant results on NCBI blast. Pham: 2 members.

Starterator report: Calls same start as Teng. GeneMark and Glimmer agree on start. Start codon contains full coding potential. Has the best RBS score, at least a thousand times better.

Function: hypothetical protein. HHPred: no significant related results (90% or higher). Blastp: hypothetical protein. SOSUI: soluble protein.

**Gene 56**

-Original Glimmer call @bp 32979 has strength 14.00

Start 32979, Stop 33773, Length 795 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: a few significant results. Pham: 3 members.

Starterator report: Calls same start as 100% of pham. Start contains full coding potential. Calls first start after stop. Has best RBS score.

Function: DNA polymerase III sliding clamp (Beta). HHPred: Many significant results (90% or greater) for DNA polymerase III beta subunit sliding clamp. Blastp: hypothetical protein. SOSUI: Is soluble.

**Gene 57**

-Original Glimmer call @bp 33770 has strength 9.34

Start 33770, Stop 34138, Length 369 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: multiple significant results. Pham: 4 members.

Starterator report: calls same start as one other member of the pham (50% of the pham call this start). Start contains full coding potential and gives a 4 bp overlap.

Function: HNH endonuclease. HHPred: significant results (90% or greater probability) for HNH endonuclease. Blastp: Many significant results for HNH endonuclease. SOSUI: Soluble protein.

**Gene 58**

-Original Glimmer call @bp 34226 has strength 12.72; GeneMark calls start at 34208

Start 34196, Stop 34741, Length 546 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Multiple significant results. Pham: Is an orpham.

Starterator report: it is an orpham so no report. RBS score is most significant for Glimmer-called start. GeneMark calls at 34208, Glimmer calls at 34226. Start at 34226 has RBS of -3.514, start at 34208 has RBS of -6.006. There is a start at 34196 with RBS of -5.611. 34196 is only one that contains all coding potential – plus if gives the biggest gene and reduces the gap with the upstream gene, that’s why it was chosen.

Function: hypothetical protein. HHPred: no significant results (90% probability or greater). Blastp: results show hypothetical protein. SOSUI: soluble protein.

**Gene 59**

-Original Glimmer call @bp 35799 has strength 12.87

Start 35799, Stop 34804, Length 996 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: Several significant results. Pham: 4 members.

Starterator report: Calls same start as 100% of genes in the pham. Start codon contains full coding potential except a tiny bit of weak potential. Extending the start out would create too much overlap with gene 59, already there is a 23 bp overlap between gene 59 and gene 60.

Function: hypothetical protein. HHPred: no significant results (90% probability or greater). Blastp: hypothetical protein. SOSUI: soluble protein.

**Gene 60**

-Original Glimmer call @bp 36358 has strength 12.39

Start: 36,358 Stop: 35,777, length 582 bp.

It is a gene: GeneMarkS shows good coding potential. Pham: 107 members.

Starterator shows that the original start codon was called 100% when present among other genes. This start has the best RBS score among starts that contain all coding potential.

Function: ParB-like nuclease domain. HHPred: Many results for ParB-like nuclease/binding protein. blastP: many results for ParB-like-nuclease domain.

**Gene 61**

-Original Glimmer call @bp 36451 has strength 15.41

Start: 36,388, Stop: 37,182, length 795 bp.

This is a gene: GenemarkS shows high coding potential. Pham: 408 members. Blastp: several significant results.

Starterator: Link says “page/report not found. Error 404” as of June 12 2022. The called start excludes about a 50 bp region of high coding potential. The start at 36388 contains all coding potential and it reduces what was initially a 94 bp gap between this gene and the upstream gene. The called start at 36451 has RBS -3.5 and at 36388 is -4.5. 36388 is picked for the reasons discussed.

Function: hypothetical protein. HHPred: hypothetical protein. BlastP: hypothetical protein. SOSUI & TMHMM: soluble.

**Gene 62**

-Original Glimmer call @bp 37179 has strength 9.01

Start:37,179, Stop: 37,517, length: 339 bp.

It is a gene: GeneMarkS indicates high coding potential. Blastp: a few significant results. Pham: 3 members.

Start has the best RBS score, is the only start to include all coding potential, and gives a 4 base pair overlap gene 61.

Function: hypothetical protein. HHPred: Showed no significant probability. BlastP: hypothetical protein. SOSUI & TMHMM: soluble.

**Gene 63**

-Original Glimmer call @bp 37507 has strength 6.57

Start: 37,507, Stop: 37,764, length 258 bp.

It is a gene: GenemarkS shows high coding potential in about 60% of ORF. Pham: 3 members. Blastp: a couple significant results.

Starterator shows that the start was called 100% when present in other genes. RBS score is best for this start and is the only start to contain all coding potential. There is a 11 bp overlap with gene 62.

Function: hypothetical protein. HHPred: No significant probability over 90%. BlastP: hypothetical protein. SOSUI & TMHMM: soluble.

**Gene 64**

-Original Glimmer call @bp 37749 has strength 12.47

Start: 37,749, Stop: 38,012, length 264 bp.

It is a gene: GenemarkS shows high coding potential. Pham: 3 members. Blastp: a couple significant results.

Starterator shows that the original start codon was called 100% when present in other genes. RBS score of this start is the best and is only to contain all coding potential. There is a 16 base pair overlap between gene 63 and gene 64.

Function: hypothetical protein. HHPred: No significant probability over 90%. BlastP: hypothetical protein. SOSUI & TMHMM: soluble.

**Gene 65**

-Original Glimmer call @bp 38017 has strength 14.52

Start: 38,017, Stop: 38,376, length 360 bp.

It is a gene: GenemarkS shows high coding potential. Blastp: a couple significant results. Pham: 3 members.

Startorator shows that the original start codon was called 100% when present in other genes. RBS score is best. Start gives 6 bp overlap with gene 64. Moving the start upstream would increase overlap another 45+ bp.

Function: hypothetical protein. HHPred: No significant probability above 90%. BlastP: hypothetical protein. SOSUI & TMHMM: soluble.

**Gene 66**

-Original Glimmer call @bp 38380 has strength 11.42

Start: 38,380, Stop: 38,544, length: 165 bp.

It is a gene: GenemarkS shows high to moderate coding potential in about half of ORF. Blastp: a couple of significant results. Pham: 17 members.

Starterator shows that the original start codon was called 100% when present in other genes. Only 2 starts present, the called start has the best RBS, the other produces a gene only 84 bp long.

Function: hypothetical protein. HHPred: No significant probability above 90%. BlastP: hypothetical protein. SOSUI & TMHMM: soluble.

**Gene 67**

-Original Glimmer call @bp 38613 has strength 11.18; GeneMark calls start at 38541

Start: 38,541, Stop: 39,062, length 522 bp.

It is a gene: GeneMarkS indicates good high coding potential. Blastp: a couple significant results. Pham: 3 members.

Glimmer calls start at 38613, this does not include all coding potential. GenemarkS calls start at 38451, which has a better RBS score, includes all coding potential, gives 4 bp overlap with gene 66, and improves alignment with top blast result – this is the start we chose.

Function: hypothetical protein. HHPred: No significant probability above 90%. BlastP: hypothetical protein. SOSUI & TMHMM: soluble.

**Gene 68**

-Original Glimmer call @bp 39059 has strength 13.95

Start: 39,059, Stop: 39,430, length 372 bp.

It is a gene: GenemarkS shows high coding potential. Pham 90,779: 3 members. Blastp: a couple significant results.

Start includes all coding potential and creates a 4 base pair overlap with gene 67.

Function: hypothetical protein. HHPred: No significant probability above 90%. BlastP: hypothetical protein. SOSUI & TMHMM: soluble.

**Gene 69**

-Original Glimmer call @bp 39427 has strength 10.31

Start: 39,427, Stop: 39,660, length: 234 bp.

It is a gene: GenemarkS shows high coding potential. Orpham. Blastp: no significant results.

Start contains all coding potential and creates a 4 base pair overlap gene 68.

Function: hypothetical protein. HHPred and BlastP show insignificant results. SOSUI & TMHMM: soluble.