**Student Gene Annotation Worksheet**

|  |  |
| --- | --- |
| **Basic Phage Information** | |
| **Phage Name** | **Wardwill** |
| **Gene #** | **10** |
| **Stop Coordinate** | **6632** |
| **Direction (For/Rev)** | **For** |
| **Gap (Overlap) with Previous Gene** | **Gap with gene 9 3 bp** |
| **Selected Start Coordinate** | **6342** |
| **Selected Function** | **NKF** |

**Annotation Decision #1: Is this a Gene?**

|  |  |
| --- | --- |
| **Gathering Evidence** | **Explain Your Rationale** |
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Yes, both* |
| Is there evidence for coding potential? | *Both GeneMarkS and/or GeneMark-host trained coding potential maps show coding potential for gene 10.* |
| Is this gene present in other annotated genomes? | *BLAST hit Zepp\_10 e value 0.0*  *Phamerator (May 5th)*  *Wardwill\_10 Pham 228701*  *Zepp\_10 Pham 228701* |
| Does the gene violate any major guiding principles? | *No significant overlap with other genes. There is a gap of 3 bp with the previous gene.*  *It is long enough (291 bp, 96 aa)*  *Genes before and after are in the same direction.* |
| **DECISION:** | *Yes* |

**Annotation Decision #2: What is the best possible start site for this gene?**

|  |  |
| --- | --- |
| **Gathering Evidence** | **Explain Your Rationale** |
| What start site do Glimmer and GeneMark suggest? | *Glimmer Start Coordinate 6342*  *GeneMark Start Coordinate 6342* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *The start site has more than just one possible associated Ribosome Binding site with a high score.*  *RBS 1 (Start site at 6336)*  *Raw SD Score -3,422*  *Z Value 2,208*  *Spacer Distance 6*  *Final Score -5,167*  *RBS 2 (Start site at 6342)*  *Raw SD Score -3,422*  *Z Value 2,208*  *Spacer Distance 12*  *Final Score -4,258*  *Although both start sites have the same Raw SD Score, RBS 2 has the less negative Final Score.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *For RBS 1: The predicted start codon is the longest ORF, the length is 297 bp. This proposed start site has a 3 bp overlap with the nearest upstream gene but it does not violate the Guiding Principles.*  *For RBS 2: The predicted start codon is not the longest ORF, the length is 291 bp. This proposed start site has a 3 bp gap with the nearest upstream gene.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *According to Starterator:*  *Candidate start sites for Wardwill\_10*  *Start 9 6336 bp*  *Start 10 6342 bp*  *The candidate start sites are conserved in other phage genomes.*  *Start 9 was called 26.7 % of time when present. It was called in 4 of the 35 non-draft genes in the pham.*  *Start 10 was called 82.4 % of time when present. It was called in 20 of the 35 non-draft genes in the pham.*  *The start number called the most often in the published annotations is 10, it was called in 20 of the 35 non-draft genes in the pham.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *BlastP match (best hit)*  *GreenIvy E value 2e-54 Accession Number* [*XEN16661.1*](https://www.ncbi.nlm.nih.gov/protein/XEN16661.1?report=genbank&log$=prottop&blast_rank=1&RID=2BKNGNZ7013)  *Alignment with GreenIvy*  *Q#1 MTDTQEQPTG*  *S#3 MTDTQEQPTG* |
| **DECISION:** | ***Keep DNA Master predicted start site at 6342. This decision was made because it is the start site called by both Glimmer and GeneMark and according to Starterator it is the most called start number.*** |

**Annotation Decision #3: What is the Function of the Putative Protein?**

|  |  |
| --- | --- |
| **Gathering Evidence** | **Explain Your Rationale** |
| Does this protein align with a protein having a functional assignment in BlastP (phagesDB and/or GenBank) with an alignment of 10-4 or smaller with appropriate coverage? | *The most informative BlastP match from each source*  *PhagesDB: hypothetical protein from Zepp*  *E value 2e-45*  *Q#1 MTDTQEQPTGA*  *S#1 MTDTQEQPTGA*  *NCBI: hypothetical protein from Zepp*  *E value 4,5e-32*  *Q#1 MTDTQEQPTGA*  *S#1 MTDTQEQPTGA*  *DNA Master: hypothetical protein from Zepp*  *E value 4,5e-32*  *Q#1 MTDTQEQPTGA*  *S#1 MTDTQEQPTGA*  *This protein aligns with an hypothetical protein from Zepp.* |
| Does this protein align with a protein having a functional assignment in the PDB or other database in HHPred with a probability of 90% or greater with appropriate coverage? | *Selected databases: PDB\_mmCIF70\_30\_Mar, Pfam-A\_v37, UniProt-SwissProt-viral70\_3\_Nov\_2021, NCBI\_Conserved\_Domais(CD)\_v3.19*  *There is not a quality match from HHPred*    *HIT 6*[*ZQB\_UN*](http://www.rcsb.org/pdb/explore/explore.do?structureId=6ZQB)  *NAME U3 small nucleolar RNA-associated protein 14,U3 small nucleolar RNA-associated protein 14; ribosome, 90S pre-ribosome, 4*  *PROBABILITY 8,21*  *E VALUE 1500*  *SCORE 17,51*  *This protein does not align with a protein having a functional assignment in the databases of HHPred with a 90% probability.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *This gene is not located adjacent to genes of known functions (just a major capsid protein upstream).This gene is in a region of the genome that shows high conservation of gene order.*  *According to Phamerator, this gene is flanked upstream by the major capsid protein and downstream by a hypothetical protein belonging to a pham family conserved among related phages* |
| Is this gene a possible transmembrane protein? | *This gene is not a possible transmembrane protein.*  *According to DeepTMHMM there is not a transmembrane domain in this protein.*  *According to SOSUI this protein is a soluble protein.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Non known function* |
| **DECISION:** | ***Hypothetical Protein*** |