**Student Gene Annotation Worksheet**

|  |  |
| --- | --- |
| **Basic Phage Information** | |
| **Phage Name** | **Wardwill** |
| **Gene #** | **11** |
| **Stop Coordinate** | **7133** |
| **Direction (For/Rev)** | **For** |
| **Gap (Overlap) with Previous Gene** | **Gap with gene 10 76 bp** |
| **Selected Start Coordinate** | **6708** |
| **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 11.* |
| Is this gene present in other annotated genomes? | *BLAST hit QuadZero\_10 e value e-77*  *Phamerator (May 5th)*  *Wardwill\_11 Pham 643*  *QuadZero\_10 Pham 643* |
| Does the gene violate any major guiding principles? | *No significant overlap with other genes. There is a gap of 76 bp with the previous gene.*  *It is long enough (426 bp, 141 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 6708*  *GeneMark Start Coordinate 6708* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *The start site has just one possible associated Ribosome Binding site with a high score.*  *RBS 1 (Start site at 6708)*  *Raw SD Score -1,95*  *Z Value 2,900*  *Spacer Distance 11*  *Final Score -2,708* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *The predicted start codon is the longest ORF, the length is 426 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\_11*  *Start 3 6708 bp*  *Start 5 6717 bp*  *The candidate start sites are conserved in other phage genomes.*  *Start 3 was called 40.4 % of time when present. It was called in 53 of the 182 non-draft genes in the pham.*  *Start 5 was called 72.8 % of time when present. It was called in 119 of the 182 non-draft genes in the pham.*  *The start number called the most often in the published annotations is 5, it was called in 119 of the 182 non-draft genes in the pham* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Best BlastP hit*  *QuadZero e value 1e-95 Accession Number* [*QWY84456.1*](https://www.ncbi.nlm.nih.gov/protein/QWY84456.1?report=genbank&log$=prottop&blast_rank=1&RID=2BM4809K013)  *Q#1 MTAMDDLNGL*  *S#1 MTAMDDLNGL*  *The start site is conserved at 6708 bp.* |
| **DECISION:** | ***Keep DNA Master predicted start site at 6708*** |

**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 QuadZero*  *E value 1e-77*  *Q#1 MTAMDDLNGLLA*  *S#1 MTAMDDLNGLLA*  *NCBI: hypothetical protein from QuadZero*  *E value 1e-95*  *Q#1 MTAMDDLNGLLA*  *S#1 MTAMDDLNGLLA*  *DNA Master: hypothetical protein from QuadZero*  *E value 0.0*  *Q#1 MTAMDDLNGLLA*  *S#1 MTAMDDLNGLLA*  *This protein aligns with an hypothetical protein from QuadZero.* |
| 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 a quality match from HHPred*    *HIT* [*9GB7\_U*](http://www.rcsb.org/pdb/explore/explore.do?structureId=9GB7)  *NAME gp50 - Portal adaptor protein; Bacteriophage, virus, needle, baseplate; 3.4A {Clostridioides difficile}*  *PROBABILITY 98,13*  *E VALUE 0.00003*  *SCORE 58,96*  *This protein aligns with a protein having a functional assignment in the databases of HHPred with a 90% probability. This protein is a portal adaptor protein.* |
| 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.This gene is not in a region of the genome that shows high conservation of gene order.*  *According to Phamerator, this gene is flanked by hypothetical proteins from a pham family conserved across 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*** |