**Student Gene Annotation Worksheet**

|  |  |
| --- | --- |
| **Basic Phage Information** | |
| **Phage Name** | **Wardwill** |
| **Gene #** | **1** |
| **Stop Coordinate** | **519** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **No Overlap or Gap** |
| **Selected Start Coordinate** | **1** |
| **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 1.* |
| Is this gene present in other annotated genomes? | *BLAST hit Zepp\_1 e value 8e-92.*  *Phamerator (May 5th)*  *Wardwill\_1 Pham 84998.*  *Zepp\_1 Pham 84998.* |
| Does the gene violate any major guiding principles? | *No significant overlap with other genes.*  *It is long enough (519 bp, 172 aa)*  *The gene after is in the same direction. If we take gene 63 as the gene before, it is also in the forward 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 1*  *GeneMark Start Coordinate 1* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *The start site has an associated Ribosome Binding site with a high score.*  *RBS 1 (Start site at 1 bp)*  *Raw SD Score -4,299*  *Genomic Z Value 1,796*  *Spacer distance 40 020*  *Final Score -6,600*  *RBS 2 (Start site at 163 bp)*  *Raw SD Score -2,109*  *Genomic Z Value 2,825*  *Spacer distance 17*  *Final Score -4,109* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *FOR RBS 1: Yes, the predicted start codon is the longest ORF. The length of the ORF with the predicted start is 519 bp. The proposed start site does not have a gap or overlap because there is no upstream gene.*  *FOR RBS 2: The predicted start codon is the longest ORF. The length of the ORF with the predicted start is 357 bp. The proposed start site does not have a gap or overlap because there is no upstream gene.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *According to Starterator:*  *Candidate start sites for Wardwill\_1*  *Start 11 1 bp*  *The candidate start sites are conserved in other phage genomes. The start called by Glimmer and GeneMark matches the start predicted by Starterator.*  *Start 11 was called 12.2 % of time when present. It was called in 13 of the 179 non-draft genes in the pham.*  *The start number called the most often in the published annotations is 10, it was called in 130 of the 179 non-draft genes in the pham.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *BlastP match (best hit)*  *Zepp E value 2e-119 Accession Number* [*QIG58091.1*](https://www.ncbi.nlm.nih.gov/protein/QIG58091.1?report=genbank&log$=prottop&blast_rank=1&RID=24ZGG45D013)  *Alignment with Zepp*  *Q#1 MCKSHGGGTA*  *S#1 MCKSHGGGTA*  *The start site at 1 is conserved in other phage genomes as indicated by BlastP.* |
| **DECISION:** | ***Keep DNA Master predicted start site at 1*** |

**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\_1*  *E value 0.0*  *Q#1 MCKSHGGGTAA*  *S#1 MCKSHGGGTAA*  *NCBI: hypothetical protein from Zepp\_1*  *E value 0.0*  *Q#1 MCKSHGGGTAA*  *S#1 MCKSHGGGTAA*  *DNA Master: hypothetical protein from Zepp\_1*  *E value 0.0*  *Q#1 MCKSHGGGTAA*  *S#1 MCKSHGGGTAA*  *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 HHPred match.*  *HIT:* [*PF19775.4*](https://www.ebi.ac.uk/interpro/entry/pfam/PF19775)  *NAME: ; DUF6261 ; Family of unknown function (DUF6261)*  *PROBABILITY 78.3*  *E VALUE 39*  *SCORE 25,73* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *The gene is located adjacent to genes of known function (terminase large subunit downstream). The gene is located in a region of the genome that shows high conservation of gene order.*  *According to Phamerator there are no conserved domains in this gene.* |
| 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 gene is a soluble protein.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *No known function* |
| **DECISION:** | **Hypothetical Protein** |