**Student Gene Annotation Worksheet**

|  |  |
| --- | --- |
| **Basic Phage Information** | |
| **Phage Name** | **Wardwill** |
| **Gene #** | **3** |
| **Stop Coordinate** | **2109** |
| **Direction (For/Rev)** | **For** |
| **Gap (Overlap) with Previous Gene** | **Gap with gene 2 of 3 bp** |
| **Selected Start Coordinate** | **1963** |
| **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 3.* |
| Is this gene present in other annotated genomes? | *BLAST hit Zepp\_3 e value 3e-22*  *Phamerator (May 5th)*  *Wardwill\_3 Pham 233544*  *Zepp\_3 Pham 233544* |
| Does the gene violate any major guiding principles? | *No significant overlap with other genes.*  *It is long enough (147 bp, 48 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 1963*  *GeneMark Start Coordinate 1963* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *The start site has one possible associated Ribosome Binding site with a high score.*  *RBS 1 (Start position 1963)*  *Raw SD Score -2,984*  *Genomic Z Value 2,414*  *Spacer distance 12*  *Final Score -3,820* |
| 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 147 bp. This proposed start site has a 3 bp gap with the nearest upstream gene but it does not violate the Guiding Principles.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *According to Starterator:*  *Candidate start sites for Wardwill\_3*  *Start 24 1963 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 24 was called 100.0 % of time when present. It was called in 21 of the 64 non-draft genes in the pham.*  *The start number called the most often in the published annotations is 26, it was called in 33 of the 64 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 6e-25 Accession Number QIG58092*  *Alignment with Zepp*  *Q#1 MTVTLYDAQG*  *S#1 MTVTLYDAQG*  *The start site at 1963 is conserved in other phage genomes as indicated by BlastP.* |
| **DECISION:** | ***Keep DNA Master predicted start site at 1963*** |

**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 of Zepp*  *E value 3e-22*  *Q#1 MTVTLYDAQ*  *S#1 MTVTLYDAQ*  *NCBI: hypothetical protein of Zepp*  *E value 6e-25*  *Q#1 MTVTLYDAQ*  *S#1 MTVTLYDAQ*  *DNA Master: hypothetical protein of Zepp*  *E value 6,4e-25*  *Q#1 MTVTLYDAQ*  *S#1 MTVTLYDAQ*  *This protein aligns with a 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 P22499*  *NAME BOF\_BPP1 Modulator protein OS=Escherichia phage P1 OX=10678 GN=bof PE=1 SV=2*  *PROBABILITY 78,29*  *E VALUE 7,4*  *SCORE 24,91*  *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 located adjacent to genes of known functions (terminase large subunit and portal protein). This gene is in a region of the genome that shows high conservation of gene order.* |
| 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? | *Not known function* |
| **DECISION:** | ***Hypothetical Protein*** |