**Student Gene Annotation Worksheet**

|  |  |
| --- | --- |
| **Basic Phage Information** | |
| **Phage Name** | **Wardwill** |
| **Gene #** | **2** |
| **Stop Coordinate** | **1960** |
| **Direction (For/Rev)** | **For** |
| **Gap (Overlap) with Previous Gene** | **Overlap with gene 1 7 bp** |
| **Selected Start Coordinate** | **512** |
| **Selected Function** | **TERMINASE, LARGE SUBUNIT.** |

**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 2* |
| Is this gene present in other annotated genomes? | *BLAST hit Zepp\_2 e value 0.0*  *Phamerator (May 3rd)*  *Wardwill\_2 Pham 84823*  *Zepp\_2 Pham 84823.* |
| Does the gene violate any major guiding principles? | *No significant overlap with other genes. 7 bp overlap with gene 1.*  *It is long enough (1449 bp, 482 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 512*  *GeneMark start coordinate 521* |
| 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 position 512)*  *Raw SD Score -2,976*  *Genomic Z Value 2,418*  *Spacer distance 11*  *Final Score -3,733*  *RBS 2 (Start position 521)*  *Raw SD Score -3,240*  *Genomic Z Value 2,294*  *Spacer distance 11*  *Final Score -3,997*  *While RBS 2 has the highest Z value, RBS 1 has the highest final score. It is not possible to make a decision considering just this information.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *According to RBS 1 with a start site at 512, the length of the ORF is 1449. This proposed start site does have an overlap of 7 bp with the nearest upstream gene but it does not violate the Guiding Principles.*  *According to RBS 2 with a start site at 521, the length of the ORF is 1440. This proposed start site does not overlap with the nearest upstream gene. 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\_2*  *Start 54 461 bp*  *Start 68 512 bp*  *Start 73 521 bp*  *The candidate start sites are conserved in other phage genomes. The start called by Glimmer matches the start predicted by Starterator (68) and the start called by GeneMark matches the start predicted by Starterator (73)*  *Start 54 was called 75.5% of time when present. It was called in 36 of 243 non draft genes in the pham.*  *Start 68 was called 80 % of time when present. It was called in 8 of the 243 non-draft genes in the pham.*  *Start 73 was called 7.1 % of time when present. It was called in 1 of the 243 non-draft genes in the pham.*  *The start number called the most often in the published annotations is 80, it was called in 125 of the 243 non-draft genes in the pham. The start site according to this would be 522 bp* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *BlastP match (best hit)*  *Zepp E value 0 Accession Number QIG58092*  *Alignment with Zepp*  *Q#1 MLDMLDVDTA*  *S#1 MLDMLDVDTA* |
| **DECISION:** | *The proposed start site by Starterator at 461 bp is not correct because it has a gap that violates the Guiding Principles. The decision is* ***Keep DNA Master predicted site at 512 bp.*** *This decision was made because this start was called in 8 of the 243 non draft genes in the pham, while Start 73 )521 bp) was called just one time.* |

**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: terminase of Zepp*  *E value 0.0*  *Q#1 MLDMLDVDTA*  *S#1 MLDMLDVDTA*  *NCBI: terminase of Zepp*  *E value 0.0*  *Q#1 MLDMLDVDTA*  *S#1 MLDMLDVDTA*  *DNA Master: terminase of Zepp*  *E value 0.0*  *Q#1 MLDMLDVDTA*  *S#1 MLDMLDVDTA*  *This protein aligns with a terminase 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*  *Most informative HHPred match:*  *HIT* [*P10310*](http://www.uniprot.org/uniprot/P10310)  *NAME TERL\_BPT3 Terminase, large subunit OS=Enterobacteria phage T3 OX=10759 GN=19 PE=1 SV=1*  *PROBABILITY 100*  *E VALUE 5e-38*  *SCORE 315.12*  *This protein does align with a protein having a functional assignment in the databases used in HHPred with a probability of 90%. This protein with a functional assignment is a terminase (large subunit).* |
| 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 in a region of the genome that shows high conservation of gene order.*  *According to Phamerator, this gene has conserved domains: Terminase 6C, COG5362, COG5410, psiM2ORF9.* |
| 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? | *Yes* |
| **DECISION:** | ***terminase*** |