**Student Gene Annotation Worksheet**

|  |  |
| --- | --- |
| **Basic Phage Information** | |
| **Phage Name** | **Wardwill** |
| **Gene #** | **5** |
| **Stop Coordinate** | **3705** |
| **Direction (For/Rev)** | **For** |
| **Gap (Overlap) with Previous Gene** | **Gap with gene 4 of 4 bp** |
| **Selected Start Coordinate** | **3547** |
| **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 5. Esta confuso.* |
| Is this gene present in other annotated genomes? | *BLAST hit Pikmin\_42 e value 3e-7*  *Phamerator (May 5th)*  *Wardwill\_5 Pham 89731*  *Pikmin\_42 Pham 89731. Not in the same EA5 Cluster but in EA Cluster.* |
| Does the gene violate any major guiding principles? | *No significant overlap with other genes. There is a gap of 4 bp with the previous gene.*  *It is long enough (159 bp, 52 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: 3547*  *GeneMark Start Coordinate: 3547* |
| 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 3547)*  *Raw SD Score -3,631*  *Genomic Z Value 2,110*  *Spacer distance 11*  *Final Score -4,388*  *RBS 2 (Start position 3601)*  *Raw SD Score -2,210*  *Genomic Z Value 2,778*  *Spacer Distance 13*  *Final Score -3,256*  *RBS 3 (Start position 3634)*  *Raw SD Score -3,240*  *Genomic Z Value 2,294*  *Spacer Distance 13*  *Final Score -4,286*  *According to this information, RBS 2 has the highest Z value and the highest (less negative) final score. RBS 2 would be a better option than the start site predicted by both Glimmer and GeneMark.* |
| 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 159 bp. This proposed start site has a 4 bp gap 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 105 bp. This violates one of the Guiding Principles that says \*Protein coding genes are generally at least 120 bp\*. This proposed start site has a 58 bp gap with the nearest upstream gene.*  *For RBS 3: the predicted start codon is not the longest ORF, the length is 72 bp. This violates one of the Guiding Principles that says \*Protein coding genes are generally at least 120\*. This proposed start site has a 91 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\_5*  *Start 1 3547 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 1 was called 100.0 % of time when present. It was called in 3 of the 3 non-draft genes in the pham.*  *The start number called the most often in the published annotations is 1, it was called in 3 of the 3 non-draft genes in the pham.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *BlastP match (best hit)*  *Pikmin E value 7e-07 Accession Number* [*YP\_009624297*](https://www.ncbi.nlm.nih.gov/protein/YP_009624297.1?report=genbank&log$=prottop&blast_rank=1&RID=24YRD58E016)  *Alignment with Pikmin*  *Q#1 MDAGRIVINY*  *S#1 MRAGSITVTY*  *The start site at 3547 is conserved in other phage genomes as indicated by BlastP.* |
| **DECISION:** | ***Keep DNA Master predicted start site at 3547*** |

**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 FDJ44\_gp42 from Pikmin*  *E value 3e-7*  *Q#1 MDAGRIVINY*  *S#1 MRAGSITVT*  *NCBI:hypothetical protein FDJ44\_gp42 from Pikmin*  *E value 7e-7*  *Q#1 MDAGRIVINY*  *S#1 MRAGSITVT*  *DNA Master: hypothetical protein FDJ44\_gp42 from Pikmin*  *E value 7,2e-7*  *Q#1 MDAGRIVINY*  *S#1 MRAGSITVT*  *This protein aligns with an hypothetical protein from Pikmin.* |
| 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* [*PF11419.13*](https://www.ebi.ac.uk/interpro/entry/pfam/PF11419)  *NAME ; DUF3194 ; Protein of unknown function (DUF3194)*  *PROBABILITY 53,86*  *E VALUE 97*  *SCORE 19,6*  *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 (a portal protein and a minor capsid 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? | *No known function* |
| **DECISION:** | ***Hypothetical Protein*** |