# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **1** |
| **Stop Coordinate** | **522 bp** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | |
| **Selected Start Coordinate** | **1 bp** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, GenemarkS does show coding potential.* |
| Is this gene present in other annotated genomes? | *There were several phages from the same cluster (EA1) that observed this same gene. This start site was annotated by 100 phages in cluster EA1, with only 4 phages annotating another start site.* |
| Does the gene violate any major guiding principles? | *Discuss if there are any significant violations of the Guiding Principles of Genome Annotation with the gene call. Do you see significant overlap with other genes? Is it long enough? Are the genes before and after this gene in the same direction?* *There is no significant overlap with other genes, the ORF is long enough (522 bp), and the genes before and after this one are both forward.* |
| **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? | *Both Glimmer and GeneMark say start is at 1 bp* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *Final RBS Score: -5.135*<br>*Z-score: 1.774*<br>*Z-score isn't the lowest z-score on there.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *Yes, it is the longest ORF with a length of 522 bp.*<br>*No, it does not overlap to the nearest stop codon of the upstream ORF.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Yes, starterator said that this start site was called in 111 of 155 non-draft genes.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The gene should start at bp 1 due to the following reasons:*<br>• *Glimmer and Genemark agree it starts here*<br>• *No overlap with nearest stop codon*<br>• *Not the lowest z-score*<br>• *Starterator showed other phages calling this the start site for this gene.* |

## 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 | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:* |

| | |
|---|---|
| alignment of $10^{-4}$ or smaller with appropriate coverage? | *Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **2** |
| **Stop Coordinate** | **1922** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **Yes, there is overlap with previous Gene** |
| **Selected Start Coordinate** | **519** |
| **Selected Function** | **terminase** |

## 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 Glimmer and Gene Mark* |
| Is there evidence for coding potential? | *Yes, there does seem to be coding potential for this gene.* |
| Is this gene present in other annotated genomes? | *Yes, the gene is present in other annotated genomes in the EA1 cluster.* |
| Does the gene violate any major guiding principles? | *There is overlap of the genes and they both are heading the same direction as eachother.* |
| **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? | *They both agreed on 519 and starterator also agrees.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *2.626 was the z value and the final score is –4.066. These were the best score that were avaliable.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *Yes, it does seem to be, and it does result in some gene overlap.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Yes, the start point does match what starterator predicted.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Record where you think the gene should start here and briefly explain your rationale.* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB: Many e values of 0.0*<br>*NCBI: Also many e values of 0.0*<br>*DNA Master: Many e values of 0.0*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *There are a few results that have over 95% match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *No, since having no evidence to support this possibility.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.*<br>*Yes, it is supported.* |
| **DECISION:** | *Terminase* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | 3 |
| **Stop Coordinate** | 3334 bp |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | 37 |
| **Selected Start Coordinate** | 1958 bp |
| **Selected Function** | Portal protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | Yes, GenemarkS does show coding potential. |
| Is this gene present in other annotated genomes? | This gene occurred in 104 annotated phages from cluster EA1. |
| Does the gene violate any major guiding principles? | There are no significant violations. There is no overlap with other genes, the gene is long enough, and the genes before and after this one 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? | *Both suggest start site of 1958 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final RBS score: -2.156<br>Z-score: 3.092<br>This was the highest Z-score. |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *Length of ORF is 1377 bp, and there is a gap of 36 bp between the start codon of this gene and the stop codon of the upstream ORF.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This start site was called in 116 of 158 non-draft genes. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Yes, this should be the starting site for the gene due to:*<br>• *Glimmer and GeneMark both agree on this site*<br>• *Highest z-score*<br>• *No gene overlap*<br>• *Start site called in 116 of 158 non-draft genes* |

## 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? | *PhagesDB: several e values of 0.0 that say portal protein*<br>*NCBI:* several e values of 0.0 thay say portal protein<br>*HHpred: 1.4x10$^{-25}$ that says portal protein* |
| 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? | HHPred: 99.95% probability that says portal protein<br>Ncbi: 2 with 90% identical that both say portal 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? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *no* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *yes* |
| **DECISION:** | *Portal protein* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | 4 |
| **Stop Coordinate** | 3437 bp |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | 3 |
| **Selected Start Coordinate** | 3336 bp |
| **Selected Function** | |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | GenemarkS shows coding potential. |
| Is this gene present in other annotated genomes? | There are no other annotated phages that have observed this same gene. |
| Does the gene violate any major guiding principles? | No significant overlap, it is not longer than 120 bp, and the genes near this one are in the same direction. |
| **DECISION:** | No, it's super short and hasn't been annotated in any annotated phages. |

## 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? | *Genemark says start site of 3336 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final RBS score: -4.541<br>Z-score: 2.406<br>Highest z-score |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *Indicate the length of the ORF is with the predicted start and the gap/overlap to the nearest stop codon of the upstream ORF. Does the proposed start site have a gap/overlap with the nearest upstream gene that does not violate the Guiding Principles?*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | Starterator not helpful. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Record where you think the gene should start here and briefly explain your rationale.* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | 5 |
| **Stop Coordinate** | 4102 |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | -1 |
| **Selected Start Coordinate** | 3437 |
| **Selected Function** | Hypothetical protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | There is coding potential in GeneMarkS. |
| Is this gene present in other annotated genomes? | This gene is present in many other cluster EA1 phages. |
| Does the gene violate any major guiding principles? | There are no significant violations of the major guiding principles. No significant overlap, the gene is long enough, and the genes before and after this one are both forward. |
| **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 and GeneMark both have start site at 3437 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final RBS Score: -3.161<br>Z-score: 3.047<br>This is the highest z-score. |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | Length of ORF is 666, and there is overlap of 1 bp with the stop codon of the upstream gene. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | Starterator not useful: this gene doesn't contain the most often annotated starting site, meaning it is not in consensus with the majority. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 3437 bp because:*<br>• *Glimmer and GeneMark agree on it*<br>• *This start site has the highest z-score* |

## 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 | *PhagesDB: several e values of e-123 that says MuF-like minor capsid protein* |

| | |
|---|---|
| functional assignment in BlastP (phagesDB and/or GenBank) with an alignment of $10^{-4}$ or smaller with appropriate coverage? | *NCBI: e value that of 6e-157 that says MuF-like minor capsid protein*<br>*HHpred: e value of 14 that says Transferase* |
| 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? | *HHPred: 99.95% probabilty match MuF-like minor capsid protein*<br>*NCBI: 65.68 % probabilty match with Transferase* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *NO* |
| Is the proposed function found on the SEA-PHAGES approved function list? | Yes..... |
| **DECISION:** | Hypothetical protein |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **6** |
| **Stop Coordinate** | **4748** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **-82** |
| **Selected Start Coordinate** | **4185** |
| **Selected Function** | **Scaffolding protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *There is coding potential* |
| Is this gene present in other annotated genomes? | *104 times for cluster EA1* |
| Does the gene violate any major guiding principles? | *No, genes do not overlap and the gene is long enough* |
| **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? | *Both agree @bp 4185* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *Z value: 2.333*<br>*RBS: -4.533* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *723 is the longest* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Found in 186 of 186 (100.0%)* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *It should start at 4185 because*<br>- *glimmer and genemark both agree on it*<br>- *No major overlap* |

## 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? | *NCBI:4e-94; scaffolding protein*<br>*Phagesdb: 2e-104; scaffolding protein*<br>*HHPred: no good e values* |
| 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? | *NCBI: 94.51% identical; scaffolding protein*<br>*HHPred: 93.41 probablility; Coronin-1A* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *no* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *Scaffolding protein* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **7** |
| **Stop Coordinate** | **5819** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **79** |
| **Selected Start Coordinate** | **4827** |
| **Selected Function** | **Major capsid protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, there is coding potiental* |
| Is this gene present in other annotated genomes? | *Start number 8 was manually annotated 94 times for cluster EA1.* |
| Does the gene violate any major guiding principles? | *No significant violations, no extreme gene overlap, long enough ORF, and the genes around it are both forward* |
| **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 and Genemark agree on 4827 bp* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS score= -2.095*<br>*Z-score= 3.186*<br>*This was the highest Z score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF length is 993 bp and this is the longest length.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Found in 424 of 453 ( 93.6% ) of genes in pham* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *I think the start site should be at 4827 bp because:*<br>• *Genemark and Glimmer agree*<br>• *Highest z-score*<br>• *It is found in 93.6% of genes* |

# 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? | *PhagesDB: an e value of 0.0 that says major capsid protein*<br>*NCBI: several e values of 0.0 that say major capsid protein*<br>*HHPred: e value of 3.7e-30 that says major capsid protein* |
| 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? | *HHPred: 100% probability match with major capsid protein*<br>*NCBI: 94.48% match that says major capsid 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? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *Major capsid protein* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **8** |
| **Stop Coordinate** | **6115** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **Yes, there is gene overlap with the previous gene** |
| **Selected Start Coordinate** | **5822** |
| **Selected Function** | **Hypothetical Protein** |

## 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 Glimmer and GeneMark* |
| Is there evidence for coding potential? | *No, they do not seem to have some coding potential.* |
| Is this gene present in other annotated genomes? | *Yes, this gene is present within other geromes in the EA1 cluster.* |
| Does the gene violate any major guiding principles? | *Yes, there is an overlap with this gene and the previous one, but they are heading the same direction.* |
| **DECISION:** | *No* |

## 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? | *They both agreed on 5822 as the start site.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *A z- value of 1.954 and had a final score of –4.747. They were not the longest ones and they do overlap with other genes.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *They were not the longest ones and they do overlap with other genes.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Yes, phages dB, Starterator, and DNA Master all say the same thing it is 5822.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Record where you think the gene should start here and briefly explain your rationale.* |

## 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? | *PhagesDB:*<br>*NCBI:*<br>*HHPred:* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 6371 |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | 74 |
| **Selected Start Coordinate** | 6189 |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | GenemarkS does show coding potential. |
| Is this gene present in other annotated genomes? | 3 annotated phages from cluster EA1 also observed this same game. |
| Does the gene violate any major guiding principles? | There are no significant violations. No overlap, it is longer than 120 bp, and the genes around it also go forward. |
| **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 and Genemark agree on 6189 start site.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final RBS Score: -2.505<br>Z-Score: 2.959<br>This is the highest z-score. |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 183 bp long, and there is no overlap with stop codon of upstream ORF. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | Starterator not helpful: this gene doesn't contain the consensus start site. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 6189 bp because:*<br>• *Glimmer and Genemark agree on it*<br>• *Z-score is highest out of options* |

## 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 | *List the most informative BlastP match from each source*<br>*PhagesDB: unknown* |

| | |
|---|---|
| functional assignment in BlastP (phagesDB and/or GenBank) with an alignment of $10^{-4}$ or smaller with appropriate coverage? | *NCBI: hypothetical*<br>*DNA Master: hypothetical* |
| 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? | *HIghest HHPred was 68%* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | Hypothetical Protein |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 6800 bp |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | 16 bp |
| **Selected Start Coordinate** | 6387 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes GenemarkS shows coding potential.* |
| Is this gene present in other annotated genomes? | This gene was observed in 104 annotated phages in the cluster EA1. |
| Does the gene violate any major guiding principles? | *No significant violations, it is long enough, there is no gene overlap, and the genes nearby are also forward.* |
| **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 and Genemark agree on this start coordinate at 6387 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final RBS Score: -2.443<br>Z-score: 2.959<br>Z-score is the highest score on the list. |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 414 bp long, and there is a gap of 16 bp with the nearest upstream gene. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This start site is the most annotated start site, being manually annotated in 102 of 158 annotated phages. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Yes, I think the start site is at 6387:*<br>• *Genemark and Glimmer agree*<br>• *Most annotated start on starterator*<br>• *Highest z-score* |

## 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 | *List the most informative BlastP match from each source*<br>*PhagesDB: function unknown*<br>*NCBI:* Hypothetical Protein |

| | |
|---|---|
| BlastP (phagesDB and/or GenBank) with an alignment of $10^{-4}$ or smaller with appropriate coverage? | *DNA Master:* Hypothetical Protein<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *Highest match was 75* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | Hypothetical Protein |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **7179** |
| **Stop Coordinate** | **7179** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **Yes, there is overlap -28** |
| **Selected Start Coordinate** | **6772** |
| **Selected Function** | **Hypothetical Protein** |

## 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 do* |
| Is there evidence for coding potential? | *Yes, it does show coding potential.* |
| Is this gene present in other annotated genomes? | *Yes, it is present within other genomes in the cluster.* |
| Does the gene violate any major guiding principles? | *No, it does not violate any guiding principles.* |
| **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 (6772)*<br>*GeneMark Start Coordinate (6778)* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *6772- final score –4.407 & z 2.558*<br>*6778- final score –3.497 & z 2.358*<br>*Either of them was the largest score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *6772- ORF 408*<br>*6778- ORF 402*<br>*Yes, both have an overlap and do not seem to break any guidelines.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *6772 was what starterator presented* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *6772- is conserved and presented in other genomes within the cluster*<br>*6778- is not on starterator nor phagesdb* |
| **DECISION:** | *Starterator, glimmer, and Phagedb all argee that start is 6772* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB: unknown*<br>*NCBI:* Hypothetical Protein<br>*DNA Master:* Hypothetical Protein<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *98% to* Hypothetical 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? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | Hypothetical Protein |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **12** |
| **Stop Coordinate** | **7517** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **-3** |
| **Selected Start Coordinate** | **7176** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes there is coding potiental* |
| Is this gene present in other annotated genomes? | *Start number 2 was manually annotated 104 times for cluster EA1.* |
| Does the gene violate any major guiding principles? | *There are no significant violations. There is no overlap with other genes, the gene is long enough, and the genes before and after this one 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? | *They suggest the start site is at 7176 bp* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS score= -2.584*<br>*Z-score= 2.959*<br>*This was the highest Z score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *Length of ORF is 342 bp, this was the longest ORF length.*<br>*There is a gap of -3 bp between the start codon of this gene and the stop codon of the upstream ORF.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *• Found in 184 of 186 ( 98.9% ) of genes in pham*<br>*• Manual Annotations of this start: 157 of 159*<br>*• Called 100.0% of time when present* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Yes, this should be the starting site for the gene due to:*<br>• *Glimmer and GeneMark both agree on this site*<br>• *Highest z-score*<br>• *Start site called in 184 of 186 non-draft genes*<br>• *Called 100% of the 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? | *List the most informative BlastP match from each source*<br>*PhagesDB: unknown function*<br>*NCBI:* Hypothetical Protein<br>*DNA Master:* Hypothetical Protein<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *99 minor tIL 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? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | Hypothetical Protein |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 7876 bp |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | 0 bp |
| **Selected Start Coordinate** | 7517 bp |
| **Selected Function** | Tail Terminator |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GeneMarkS shows coding potential.* |
| Is this gene present in other annotated genomes? | *This gene was manually annotated in 104 other phages from cluster EA1.* |
| Does the gene violate any major guiding principles? | *No significant violations, no extreme gene overlap, long enough ORF, and the genes around it are both forward.* |
| **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 and Genemark agree on 7517 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *Final RBS score: -3.917*<br>*Z-score: 2.311*<br>*Tied for highest z-score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 360 bp long, and the start site starts exactly where the upstream gene ends. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This gene doesn't contain the consensus start site on starterator, but this start site I chose was manually annotated in 44 of 158 phages with this gene. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *I think the start site should be at 7517 bp because:*<br>• *Genemark and Glimmer agree*<br>• *Highest z-score*<br>• *Not consensus start site, but still annotated by others* |

## 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 | *PhagesDB: e of 3e-46, tail terminator*<br>*NCBI: e of 2e-56, tail terminator* |

| | |
|---|---|
| functional assignment in BlastP (phagesDB and/or GenBank) with an alignment of $10^{-4}$ or smaller with appropriate coverage? | *HHpred: no e values below $10^{-4}$* |
| 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? | HHPred: 97.84 probability with a gene for a tail terminator 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? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *Tail Terminator Protein* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | 14 |
| **Stop Coordinate** | 8103 bp |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | -4 bp |
| **Selected Start Coordinate** | 7873 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GeneMarkS shows coding potential.* |
| Is this gene present in other annotated genomes? | *Gene manually annotated in 104 other phages in cluster EA1.* |
| Does the gene violate any major guiding principles? | *No significant violations, no significant gene overlap, long enough ORF, and the genes surrounding are both forward* |
| **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 and GenemarkS agree on start site 7873 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final RBS score: -2.175<br>Z-score: 3.112<br>Highest z-score |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 231 bp, and there is overlap of 4 bp with the upstream gene. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This start site is the most annotated start site on starterator, called in 132 of 137 annotated genomes. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *This gene start site should be at 7873 bp because:*<br>• *Highest z-score*<br>• *Glimmer and Genemark agree*<br>• *Starterator calls this most annotated start site* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB: unknown*<br>*NCBI: hypothetical*<br>*DNA Master: hypothetical*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *Highest match was 68%* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | Hypothetical Protein |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **15** |
| **Stop Coordinate** | **8615** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **Yes, there is a gap 15 bp** |
| **Selected Start Coordinate** | **8118** |
| **Selected Function** | **Major tail protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, there seems to be coding potenial* |
| Is this gene present in other annotated genomes? | *Yes, similar genes have been present in other genomes.* |
| Does the gene violate any major guiding principles? | *No, it does not seem to violate any major guidelines* |
| **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? | *Both point to 8118* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *8118- Final score –2.786 & ORF 498*<br>*Not the highest score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *The gap is 14 bp*<br>*Yes, it does result in gene overlap* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Yes, the start site seems to be the same in multpile other genes.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *8118- Yes, it is present within other genomes.* |
| **DECISION:** | *8118- starts here because startetator, glimmer, genemark, and phagesdb all say the same thing.* |

## 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? | *List the most informative BlastP match from each source PhagesDB:* `4e-90`<br>*NCBI: 6e-112*<br>*HHPred: No e-values below $10^{-4}$* |
| 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? | *Yes, in NCBI 97.58% align with a function.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *Major tail protein* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | 16 |
| **Stop Coordinate** | 9215 |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | 30 bp |
| **Selected Start Coordinate** | 8646 bp |
| **Selected Function** | Tail Assembly Chaperone |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | GenemarkS shows coding potential |
| Is this gene present in other annotated genomes? | *Gene observed in 104 annotated phages from cluster EA1.* |
| Does the gene violate any major guiding principles? | *No significant violations, not significant overlap, long enough reading frame, and genes surrounding are both forward* |
| **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 and Genemark agree on 8646* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *List the final RBS score and Z-score of the currently predicted start site using the Kibler6/Karlin Medium scoring table. Indicate in your response if this is the best score or not.*<br>*Final RBS score: -2.794*<br>*Z-score: 2.959*<br>*Highest z-score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 570 bp long, with there being a 30 bp gap with the nearest upstream gene. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This is the most annotated start site in starterator, being called in 140 of 158 annotated genes. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 8646 bp because:*<br>• *Highest z-score*<br>• *Glimmer and Genemark agree*<br>• *Starterator says this is most annotated start site* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB: 2e-97 for a gene assigned tail assembly chaperone*<br>*NCBI: 1e-124 for a gene assigned tail assembly chaperone* |
| 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? | No HHPred matches with above 90 probability<br>NCBI: 91.49% identical gene that is for a tail assembly chaperone |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *Tail assembly chaperone* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 9604 bp |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | 15 bp |
| **Selected Start Coordinate** | 9230 bp |
| **Selected Function** | Tail Assembly Chaperone |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Genemark S and Genemark-host shows coding potential* |
| Is this gene present in other annotated genomes? | *This gene was manually annotated in 106 other cluster EA1 phage genomes.* |
| Does the gene violate any major guiding principles? | There are no significant violations, it is long enough, there is no overlap with the upstream gene, and the genes around go 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 and Genemark agree on 9230 bp.* |
| Does the start site have an associated Ribosome | Final RBS: -7.612<br>Z-score: .943 |

| | |
|---|---|
| Binding Site with a high score? | These aren't very good scores. |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | The ORF is 375 bp with a gap of 15 bp between the upstream gene.<br><br>The longest ORF has 171 bp overlap, so not an option. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This is the consensus start site, called in 110 of 121 non-draft genes. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *I think the start site should be 9230 bp because:*<br>• *Glimmer and GeneMark agree*<br>• *Consensus start site on starterator* |

## 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? | *PhagesDB: 2e-68; tail assembly chaperone*<br>*NCBI: 3e-83; tail assembly chaperone*<br>*HHpred: no low e value* |
| 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 | *NCBI: 95.97% identical; tail assembly chaperone*<br>*HHPred: 90.66 probability; metal binding protein* |

| | |
|---|---|
| greater with appropriate coverage? | |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *Tail assembly chaperone* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **18** |
| **Stop Coordinate** | **11892** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **-24** |
| **Selected Start Coordinate** | **9628** |
| **Selected Function** | **Tape Measure Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *There is coding potential for both GeneMarkS and GeneMark-host.* |
| Is this gene present in other annotated genomes? | *Start number 4 was manually annotated 3 times for cluster EA1.* |
| Does the gene violate any major guiding principles? | *No significant violations, not significant overlap, long enough reading frame, and genes surrounding are both forward* |
| **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? | *They both suggest a start at 9628 bp* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *final RBS score- (-2.095)*<br>*Z-score-( 3.186)*<br>*This was the highest Z score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF is 2265 bp long, with there being a -24 bp gap with the nearest upstream gene* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *• Found in 69 of 186 ( 37.1% ) of genes in pham*<br>*• Manual Annotations of this start: 58 of 159*<br>*• Called 100.0% of time when present* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 9628 bp because:*<br>• *Highest z-score*<br>• *Glimmer and Genemark agree*<br>• *Starterator says this is most annotated start site* |

# 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? | *List the most informative BlastP match from each source*<br>*PhagesDB: 0.0, tape measure protein*<br>*NCBI: e- 0.0*<br>*HHPred: 1.6e-18*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *HHPred has 99.94% with gene probability of tape measure 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? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *Tape Measure Protein* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **19** |
| **Stop Coordinate** | **12662** |
| **Direction (For/Rev)** | **Foward** |
| **Gap (Overlap) with Previous Gene** | **3** |
| **Selected Start Coordinate** | **11889** |
| **Selected Function** | **Minor tail protein** |

## 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? | *GeneMark and GeneMarkS both show coding potential* |
| Is this gene present in other annotated genomes? | *104 times for cluster EA1* |
| Does the gene violate any major guiding principles? | *End and start do not overlap. Gene is long enoughF* |
| **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 11889*<br>*GeneMark Start Coordinate 11889* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS score: -3.319*<br>*Z value: 2.618* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF: 774*<br>*Longest ORF in gene overlap* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *It is found in 160 out of 160 (100.0%)* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Start should be at 11889*<br>*Genemark and glimmer show this start*<br>*Does not overlap* |

## 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? | *PhagesDB: an e value of e –138 that says minor tail protien*<br>*NCBI: an e value of 6e-176 that says minor tail protien*<br>*HHPred: e value of 2.3e-25 that says tail protien* |
| 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? | *HHPred: 98.19% probabilty match with*<br>*NCBI: 94.94% probabilty match with minor tail protien* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *NO* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *Minor tail protein* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **15085** |
| **Stop Coordinate** | **15085** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **No** |
| **Selected Start Coordinate** | **12662** |
| **Selected Function** | **Minor tail protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, it shows coding potential.* |
| Is this gene present in other annotated genomes? | *Yes, it is present within other genomes in the same cluster.* |
| Does the gene violate any major guiding principles? | *It doesn't seem to be violating any guide lines.* |
| **DECISION:** | *Respond here with YES or NO after reviewing the evidence gathered above.* |

## 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? | *They both point to 12662 and so does phagesdb.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *List the final RBS score and Z-score of the currently predicted start site using the Kibler6/Karlin Medium scoring table. Indicate in your response if this is the best score or not.*<br>*Final score: -2.443 and Z-score: 2.959 these are indeed the best scores shown.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *Indicate the length of the ORF is with the predicted start and the gap/overlap to the nearest stop codon of the upstream ORF. Does the proposed start site have a gap/overlap with the nearest upstream gene that does not violate the Guiding Principles?*<br><br>*ORF length is 2424 and it seems to start where it stopped.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Yes, other genomes have the same start site within this cluster.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Yes, there are some other gemones that do start the same .* |
| **DECISION:** | *12662 because Dna master, Glimmer, and Startetator all say the same.* |

## 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? | *PhagesDB: 0.0; minor tail protein*<br>*NCBI: 0.0; minor tail protein*<br>*HHPred: 2.5e-22; phage tail, tail tip, tape measure protein* |
| 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? | *NCBI: 97.89% identical; minor tail protein*<br>*HHPred: 99.93 probablity; prophage tail 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? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *Minor tail protein* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | 22 |
| **Stop Coordinate** | 15270 bp |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | -1 bp |
| **Selected Start Coordinate** | 15085 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host show coding potential* |
| Is this gene present in other annotated genomes? | This gene was observed in 104 other cluster EA1 phages. |
| Does the gene violate any major guiding principles? | *Discuss if there are any significant violations of the Guiding Principles of Genome Annotation with the gene call. Do you see significant overlap with other genes? Is it long enough? Are the genes before and after this gene in the same direction?*<br>No serious violations: no signficant overlap, it is long enough, and the genes surrounding are also forward. |
| **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 and Genemark agree on the start site 15085* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final RBS score: -3.566<br>Z-score: 2.474<br>Tied for highest z-score. |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | The ORF would be 186 bp long with an overlap of 1 bp with the nearest upstream gene. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This matches the consensus start site on starterator, being called in 74 of the 138 non-draft genes. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Reasons for 15085 bp as the start site:*<br>• *Glimmer and Genemark agree on it*<br>• *Tied for highest z-score*<br>• *Starterator says this is consensus start site* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and* |

| | |
|---|---|
| | *e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **23** |
| **Stop Coordinate** | **15863** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **overlap** |
| **Selected Start Coordinate** | **15267** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, there is some evidence of coding potential* |
| Is this gene present in other annotated genomes? | *Yes, the gene is present in other annotated genomes.* |
| Does the gene violate any major guiding principles? | *It doesn't seem to violate any guide lines* |
| **DECISION:** | *Respond here with YES or NO after reviewing the evidence gathered above.* |

## 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? | *Both of them say it's 15267* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *The final RBS score is –4.602 and Z-score 2.022, both are the best score shown.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *The length of the ORF is 597 and it does have an overlap that doesn't seem to break any guidelines.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Yes, the start site is reserved in other genomes.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Yes, other genomes have the same start site.* |
| **DECISION:** | *15267 Startetator, Phagesdb,Glimmer, and Gene mark all say the same.* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **16348** |
| **Stop Coordinate** | **16348bp** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **0** |
| **Selected Start Coordinate** | **15863bp** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, there is coding potiental* |
| Is this gene present in other annotated genomes? | *Start number 11 was manually annotated 103 times for cluster EA1.* |
| Does the gene violate any major guiding principles? | *There are no significant violations, it is long enough, there is no overlap with the upstream gene, and the genes around go 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 and Genemark agree on 15863 bp* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *Final RBS: -3.390*<br>*Z-score: 2.548*<br>*This was the highest Z score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *The ORF is 486 bp with a gap of 0 bp between the upstream gene.*<br><br>*This was the highest ORF length* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Found in 123 of 160 ( 76.9% ) of genes in pham*<br>*Manual Annotations of this start: 103 of 138* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *I think the start site should be 9230 bp because:*<br>• *Glimmer and GeneMark agree*<br>• *Has the longest ORF length*<br>• *Has the highest Z score* |

# 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | 24 |
| **Stop Coordinate** | 18477 bp |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | -1 bp |
| **Selected Start Coordinate** | 16348 bp |
| **Selected Function** | Minor tail protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host both show coding potential.* |
| Is this gene present in other annotated genomes? | This gene was called in 104 other cluster EA1 phages. |
| Does the gene violate any major guiding principles? | No significant violations: no significant overlap, it is long enough, and the surrounding genes do go 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 and Genemark say start site is at 16348 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final RBS score: -3.837<br>Z-score: 2.474<br>Highest z-score |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 2130 bp long, with an overlap of 1 bp with the nearest upstream gene. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This is the consensus start site on starterator, called in 114 of 138 non-draft genes. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site is at 16348 bp because:*<br>• *Glimmer and Genemark agree*<br>• *Highest z-score*<br>• *Consensus start site on starterator* |

## 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 | *PhagesDB: 0.0; minor tail protein*<br>*NCBI: 0.0; minor tail protein* |

| | |
|---|---|
| functional assignment in BlastP (phagesDB and/or GenBank) with an alignment of $10^{-4}$ or smaller with appropriate coverage? | *HHpred: 5.6e-25; chitinase* |
| 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? | *NCBI: 96.33% identical; minor tail protein*<br>*HHPred: 99.94 probability; chitinase* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *no* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *Minor tail protein* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **25** |
| **Stop Coordinate** | **19330** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **0** |
| **Selected Start Coordinate** | **18509** |
| **Selected Function** | Lysin A |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Both show coding potential* |
| Is this gene present in other annotated genomes? | *Start number 24 was manually annotated 2 times for cluster EA1* |
| Does the gene violate any major guiding principles? | *Long enough, does not overlap.* |
| **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 18509*<br>*GeneMark Start Coordinate 18509* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS: -2.443*<br>*Z score: 2.959* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *822 is the longest ORG with 0 bp gap* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | • Found in 4 of 179 ( 2.2% ) of genes in pham<br>• Manual Annotations of this start: 2 of 153<br>• Called 100.0% of time when present |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start should be at 18509 because Genemark and Glimmer agree.* |

# 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:* e-153<br>*NCBI:* 4e-177<br>*HHPred:* 1.3e-8<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *NCBI: has a* 93.41% *identical match to lysin A* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *Yes* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *NCBI: calls* 93.41% for Lysin A |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **26** |
| **Stop Coordinate** | **19776** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **gap** |
| **Selected Start Coordinate** | **19363** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, there does seem to have some coding potenial.* |
| Is this gene present in other annotated genomes? | *Yes, it is present in other genomes.* |
| Does the gene violate any major guiding principles? | *It does not seem to violate any guide lines.* |
| **DECISION:** | *Respond here with YES or NO after reviewing the evidence gathered above.* |

## 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? | *Both say 19363 as the start site.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS score- -1.954*<br>*Z-Score- 3.186*<br>*Yes, these scores are the highest that are shown.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF Length- 414*<br>*Everything seems to line up and the gap does not seem to violate any guidelines.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *No, other genomes don't have it conserved.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *No, the start site is not conserved in other genomes.* |
| **DECISION:** | *19363 GeneMark, Glimmer, Phages DB, and Startetator all say the same start site.* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | 27 |
| **Stop Coordinate** | 20120 bp |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | 7 bp |
| **Selected Start Coordinate** | 19782 bp |
| **Selected Function** | Holin |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host both show coding potential.* |
| Is this gene present in other annotated genomes? | This gene was observed in 104 other Cluster EA1 phages. |
| Does the gene violate any major guiding principles? | No significant violations: no significant overlap, it is long enough, and the genes surrounding are also foward. |
| **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 and Genemark both suggest 19782 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final RBS Score: -4.245<br>Z-score: 2.728<br>2nd highest z-score and final score, but the highest z-score start site has ORF that's 48 bp long. |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 339 bp long, with a gap of 7 bp to the nearest upstream gene. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This is the most annotated start site according to starterator, being called in 112 of 127 non-draft genes. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Start site should be at 19782 because:*<br>• *Genemark and Glimmer agree*<br>• *Starterator says this is consensus*<br>• *High z-score and final score* |

## 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 | *List the most informative BlastP match from each source*<br>*PhagesDB:* 8e-51<br>*NCBI:* 1e-62<br>*HHPred:* 3.6e-16 |

| | |
|---|---|
| smaller with appropriate coverage? | |
| 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? | *HHPred calls it Holin with a score of* 99.72% that *is identical.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *No* |
| Is this gene a possible transmembrane protein? | *Yes*<br>Query MSA diversity (Neff): **7.42017**<br>Detected sequence features: ▪**Transmembrane segment(s)** |
| Is the proposed function found on the SEA-PHAGES approved function list? | Yes, it is on the approved list of functions. |
| **DECISION:** | *NCBI calls it with a 91.96% match and HHPred calls it with a 99.72% match both calling holin.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | 29 |
| **Stop Coordinate** | 20654 bp |
| **Direction (For/Rev)** | Reverse |
| **Gap (Overlap) with Previous Gene** | 61 bp |
| **Selected Start Coordinate** | 20845 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host both show coding potential.* |
| Is this gene present in other annotated genomes? | This gene was manually annotated in 104 other cluster EA1 genes according to starterator. |
| Does the gene violate any major guiding principles? | There are no significant violations: there is no overlap, it is long enough, and the gene before this one is the opposite direction (but there is a big enough gap between). |
| **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 and Genemark say the start site is 20845 bp. |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final RBS score: -2.443<br>Z-Score: 2.959<br>These are the highest z-score and final score. |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *The ORF is 192 bp long, and there is a 61 bp gap. Although, this gap is between two opposite direction genes, so it doesn't violate any guidelines.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This is the most annotated start site on starterator, being called in 147 of 151 non-draft genes. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Start site should be at 20845 bp because:*<br>• *Glimmer and Genemark*<br>• *Final and Z scores are the highest*<br>• *Most annotated start site on starterator* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |

| | |
|---|---|
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **21070** |
| **Stop Coordinate** | **21070** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **There is a Gap** |
| **Selected Start Coordinate** | **21228** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, there is some coding potential.* |
| Is this gene present in other annotated genomes? | *Yes, the gene is shown in other annotated genomes.* |
| Does the gene violate any major guiding principles? | *I believe it may violate due to going past over 50* |
| **DECISION:** | *Respond here with YES or NO after reviewing the evidence gathered above.* |

## 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? | *They both point to* **21228** |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS score = -4.075*<br>*Z-score = 2.266*<br>*RBS score is best scoring, but the Z-score is the second best.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF- 159 is the longest one and does exceed the amount needed.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *No, Starterator does not show that the start site is conserved in other genomes* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *No, the start site is not conserved by other genomes* |
| **DECISION:** | *Yes, because Starterator, Glimmer, Gene Mark, and Phages db say the same.* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | 21225 |
| **Stop Coordinate** | 21225 bp |
| **Direction (For/Rev)** | Reverse |
| **Gap (Overlap) with Previous Gene** | -4 |
| **Selected Start Coordinate** | 22079 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host both show coding potential.* |
| Is this gene present in other annotated genomes? | The gene was annotated in 104 other cluster EA1 genes. |
| Does the gene violate any major guiding principles? | There is no significant violations: no significant overlap, it is long enough, and the genes surrounding are also reverse. |
| **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 and Genemark both say 22079 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final: -6.275<br>Z-score: 1.925<br>Middle z-score and lowest final score. Not very good |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | The ORF is 855 bp long, and there is a small 4 bp overlap with the nearest upstream gene. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This is the most annotated start site on starterator, being called in 129 of 156 non-draft genes. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 22079 bp because:*<br>• *Glimmer and genemark agree*<br>• *Consensus start site on starterator* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| Does this protein align with a protein having a functional assignment in | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.* |

| | |
|---|---|
| the PDB or other database in HHPred with a probability of 90% or greater with appropriate coverage? | *Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | 22189 |
| **Stop Coordinate** | 22189 bp |
| **Direction (For/Rev)** | Reverse |
| **Gap (Overlap) with Previous Gene** | -11 bp |
| **Selected Start Coordinate** | 22545 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | GenemarkS and Genemark-host both show coding potential. |
| Is this gene present in other annotated genomes? | *There is a single cluster EA1 phage that manually annotated this gene.* |
| Does the gene violate any major guiding principles? | There are no significant violations: no significant overlap, it is long enough, and the surrounding genes are also reverse. |
| **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 and Genemark say 22545 bp as start site.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final: -2.214<br>z-score: 3.103<br>Highest z and final scores. |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 357 bp long, and there is an 11 bp overlap with the nearest upstream gene. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This is the consensus start site, but when that only pulls from a pool of a single annotated gene, that isn't very helpful. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site is 22545 bp because:*<br>• *Glimmer and Genemark agree*<br>• *Highest z and final scores* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| Does this protein align with a protein having a functional assignment in | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.* |

| | |
|---|---|
| the PDB or other database in HHPred with a probability of 90% or greater with appropriate coverage? | *Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **22555** |
| **Stop Coordinate** | **22555** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **Gap** |
| **Selected Start Coordinate** | **22806** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, it does seem to have some coding potential.* |
| Is this gene present in other annotated genomes? | *No, it does not seem to be* |
| Does the gene violate any major guiding principles? | *No, it does not seem to* |
| **DECISION:** | *Respond here with YES or NO after reviewing the evidence gathered above.* |

## 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? | *Both call 22806* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS- -3.262*<br>*Z- score- 2.616*<br>*They are the best scores shown* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF- 252*<br>*Yes, it has a gap and it doesn't violate any guidelines* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *No, it does not seem to have any other genomes that are similar.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *No, it does not seem to be conserved in other genomes.* |
| **DECISION:** | *22806 is the start because glimmer, Genemark, starterator, and phagesdb call as that.* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 22806 bp |
| **Direction (For/Rev)** | Forward |
| **Gap (Overlap) with Previous Gene** | -1 bp |
| **Selected Start Coordinate** | 24488 bp |
| **Selected Function** | *RecA-like DNA Recombinase* |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host both show coding potential.* |
| Is this gene present in other annotated genomes? | This gene was observed in 104 annotated phages in cluster EA1. |
| Does the gene violate any major guiding principles? | There are no significant violations: no significant overlap, it is long enough, and the surrounding genes are also reverse. |
| **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 says: 24488 bp*<br>Genemark says: 24419 |
| Does the start site have an associated Ribosome Binding Site with a high score? | 24488:<br>Z: 3.029<br>Final: -2.433<br>Highest z score and final score<br><br>24419:<br>Z: 2.179<br>Final: -4.263<br>Average scores |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *24488: ORF is 1683 bp long with a 1 bp overlap*<br><br>*24419: ORF is 1614 bp long with a 1 bp overlap.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | 24488 bp is the consensus start site according to starterator, being called in 144 of 211 non-draft genes. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 24488 bp because:*<br>• *Glimmer say so*<br>• *Highest z and final scores*<br>• *Consensus start site on starterator* |

## 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 | *PhagesDB: 0.0; DNA Recombinase*<br>*NCBI: 0.0; DNA Recombinase*<br>*HHpred: 1.3e-18;* Regulatory protein repA |

| | |
|---|---|
| smaller with appropriate coverage? | |
| 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? | NCBI: 95.36% identical; DNA Recombinase<br>HHPred: Regulatory protein repA |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Sort of; no DNA Recombinase, but RecA-like DNA Recombinase is* |
| **DECISION:** | *RecA-like DNA Recombinase* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **24759** |
| **Stop Coordinate** | **24759** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **Overlap** |
| **Selected Start Coordinate** | **25604** |
| **Selected Function** | **Hypothetical** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, there is coding potential* |
| Is this gene present in other annotated genomes? | *It does not seem to be present in other genomes* |
| Does the gene violate any major guiding principles? | *No, it does not seem to* |
| **DECISION:** | *Respond here with YES or NO after reviewing the evidence gathered above.* |

## 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? | *Both suggest 25604* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS= -4.871*<br>*Z-score = 1.831*<br>*Both are not the highest scoring* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF- 846*<br>*It is not the longest and the other do seem to result in overlap* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Yes, it is conserved in the cluster* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *No, it does not seem to be* |
| **DECISION:** | *Starterator, glimmer, Gene mark, and Phages db all suggest the same start.* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 25637 bp |
| **Direction (For/Rev)** | Reverse |
| **Gap (Overlap) with Previous Gene** | 34 bp |
| **Selected Start Coordinate** | 26305 bp |
| **Selected Function** | AAA-ATPase |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host both show coding potential.* |
| Is this gene present in other annotated genomes? | Gene observed in 104 annotated genomes in cluster EA1. |
| Does the gene violate any major guiding principles? | *Discuss if there are any significant violations of the Guiding Principles of Genome Annotation with the gene call. Do you see significant overlap with other genes? Is it long enough? Are the genes before and after this gene in the same direction?*<br>There is no significant violations: no overlap, it is long enough, and the surrounding genes are also reverse. |
| **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 and Genemark both say 26305 bp as start site.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final: -4.395<br>Z: 2.090<br>2nd highest scores |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 669 bp long, with a 34 bp gap. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This is consensus start site, being called in 103 of 158 non-draft genes. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Start site should be at 26305 bp because:*<br>• *Glimmer and Genemark agree*<br>• *Pretty high z and final scores*<br>• *Consensus start site* |

## 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 | *PhagesDB: e-118; AAA-ATPase*<br>*NCBI: 5e-155; AAA-ATPase*<br>*HHPred: 6.7e-9; RecA superfamily ATPase* |

| | |
|---|---|
| alignment of $10^{-4}$ or smaller with appropriate coverage? | |
| 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? | *NCBI: 95.95% identical; AAA-ATPase*<br>HHPred: 99.13; Reca superfamily ATPase |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *AAA-ATPase* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 26302 bp |
| **Direction (For/Rev)** | Reverse |
| **Gap (Overlap) with Previous Gene** | -3 bp |
| **Selected Start Coordinate** | 27483 bp |
| **Selected Function** | exonuclease |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | GeneMark S and Genemark-host both show coding potential. |
| Is this gene present in other annotated genomes? | This gene was manually annotated in 104 other cluster EA1 phages. |
| Does the gene violate any major guiding principles? | There are no significant violations. No significant overlap, gene is long enough, and surrounding genes are reverse. |
| **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: 27483 bp*<br>*Genemark: 27492 bp* |
| Does the start site have an associated Ribosome Binding Site with a high score? | 27483 bp:<br>Final: -3.671<br>Z-Score: 2.389<br>Second highest final and z scores<br><br>27492:<br>Final: -4.802<br>Z-Score: 1.929<br>3rd highest scores |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | With a start site at 27483 bp, the ORF is 1182 bp long, and with a start site of 27492 bp, the ORF is 1191 bp long. Neither would lead to excessive overlap. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This gene does contain the consensus start site, being called in 113 of 383 non-draft genes. This start site is at 27483 bp. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Start site should be at 27483 bp because:*<br>• *Glimmer says*<br>• *Second highest z-score and final RBS score*<br>• *Consensus start site on starterator* |

## 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 | *PhagesDB: 0.0; exonuclease*<br>*NCBI: 0.0; exonuclease*<br>*HHPred: 3.2e-19; Restriction endonuclease-like* |

| | |
|---|---|
| alignment of $10^{-4}$ or smaller with appropriate coverage? | |
| 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? | *NCBI: 97.46% identical; exonuclease*<br>*HHPred: 99.87 probability; Restriction endonuclease-like* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *Exonuclease* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **40** |
| **Stop Coordinate** | **27470** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **-13** |
| **Selected Start Coordinate** | **29335** |
| **Selected Function** | *DNA Polymerase* |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host both show coding potential.* |
| Is this gene present in other annotated genomes? | (Start: 277 @29335 has 129 MA's<br>•Start number 277 was manually annotated 95 times for cluster EA1.<br>• Found in 190 of 1729 ( 11.0% ) of genes in pham<br><br>• Manual Annotations of this start: 129 of 1532 • Called 78.4% of time when present |
| Does the gene violate any major guiding principles? | There is no significant violations: no overlap, it is long enough, and the surrounding genes are also reverse |
| **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:29335 bp*<br>*GeneMark Start Coordinate:29326* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final: -2.664<br>Z: 2.885<br>This was the highest Z score |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 1866 bp long, with a -13 bp gap. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | • Found in 190 of 1729 ( 11.0% ) of genes in pham<br>• Manual Annotations of this start: 129 of 1532<br>• Called 78.4% of time when present |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The selected start site is at 29335 bp*<br>• *Has the highest Z score*<br>• *Has more manual annotation then the other start site* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB: 0.0 e value*<br>*NCBI: 0.0 e value*<br>*HHPred:* 4.6e-66 e value |
| 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? | *HHPred:* DNA polymerase with a 100% |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *Yes* |
| Is this gene a possible transmembrane protein? | *Yes,* **Coiled coil segment** |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *HHPred: 100% was called for DNA Polymerase* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **29535** |
| **Stop Coordinate** | **29535** |
| **Direction (For/Rev)** | **Rev** |
| **Gap (Overlap) with Previous Gene** | **20** |
| **Selected Start Coordinate** | **29966** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Both Glimmer and GeneMark indicate 29966as the Start.* |
| Is there evidence for coding potential? | *GeneMark shows coding potential* |
| Is this gene present in other annotated genomes? | *104 annotated phages from cluster EA1 have observed this gene.* |
| Does the gene violate any major guiding principles? | *No major violations* |
| **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? | *Both indicate a start of bp@ 29966* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *Highest RBS Score: -2.443*<br>*Z value: 2.959*<br>*These are the highest scores* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF is 432 bp long, and this causes no overlap. This is not the longest ORF.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *This is the consensus start site, called in 133 of 158 annotated phages.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *I think it should start at 29966*<br>- *Both Glimmer and GeneMark indicate this*<br>- *Highest z and final rbs scores*<br>- *Consensus start site* |

# 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **30141** |
| **Stop Coordinate** | **30141** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **Gap** |
| **Selected Start Coordinate** | **31547** |
| **Selected Function** | **DNA Helicase** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, it does show some coding potenial.* |
| Is this gene present in other annotated genomes? | *It does not to be in the same pham track* |
| Does the gene violate any major guiding principles? | *No it does not seem to violate any guiding principles.* |
| **DECISION:** | *Respond here with YES or NO after reviewing the evidence gathered above.* |

## 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? | *Both say 31547* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS score: -3.199*<br>*Z-Score: 2.608*<br>*Both are the best scores shown.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF Length: 1407*<br>*It was not the longest ORF results and yes it does have excessive overlap.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *No, it does not seem to be conserved in other phage genomes.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *No, it does not seem to be conserved in other phage genomes.* |
| **DECISION:** | *31547 Glimmer,Genemark,and phagesdb all call it.* |

## 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 | *PhagesDB: several e values of 0.0 that say DNA Helicase*<br>*NCBI:* several e values of 0.0 thay say DNA Helicase |

| | |
|---|---|
| functional assignment in BlastP (phagesDB and/or GenBank) with an alignment of $10^{-4}$ or smaller with appropriate coverage? | *HHpred:* 7.9e-39 *that says DNA Binding protein* |
| 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? | HHPred: 100% probability that says DNA Binding protein<br>Ncbi: 99% probabilty that says DNA Helicase |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *NO* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *YES* |
| **DECISION:** | *DNA Helicase* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 31541 bp |
| **Direction (For/Rev)** | Reverse |
| **Gap (Overlap) with Previous Gene** | -6 bp |
| **Selected Start Coordinate** | 31657 bp |
| **Selected Function** | Hypothetical protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host show coding potential.* |
| Is this gene present in other annotated genomes? | 6 annotated phages in cluster EA1 observed the same gene. |
| Does the gene violate any major guiding principles? | The only real violation is that the ORF is below 120 bp long. The genes surrounding are also reverse, and there is no significant overlap. |
| **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? | *They both suggest 31657 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *Final RBS score: -3.649*<br>*Z-score: 2.464*<br>Highest z-score and Final RBS score |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | The ORF is 117 bp long, with an overlap of 7 bp. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This is the consensus start site, which isn't too helpful as only 6 annotated phages have observed this gene, but it is notable that all 6 of them agreed on this start site. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Start site should be at 31657 bp because:*<br>• *Glimmer and Genemark agree*<br>• *Highest z and final rbs scores*<br>• *Consensus start site*<br>• *The only other 2 options would have had ORF's of 87 bp and 75 bp length, so this is only choice* |

## 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? | *PhagesDB: several e values over 0 that say function unkown*<br>*NCBI:* e value of 3e-12 that says hypothetical protien<br>*HHpred: several e values over 0* |

| | |
|---|---|
| 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? | HHPred: 92.22% probability that says unkown<br>Ncbi: 100% probability that says hypothetica protien |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *NO* |
| Is the proposed function found on the SEA-PHAGES approved function list? | NO |
| **DECISION:** | *Hypothetical protien* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | 31654 |
| **Stop Coordinate** | 31654 bp |
| **Direction (For/Rev)** | Reverse |
| **Gap (Overlap) with Previous Gene** | -4 bp |
| **Selected Start Coordinate** | 31857 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host show coding potential* |
| Is this gene present in other annotated genomes? | This gene was observed in 12 annotated phages of cluster EA1. |
| Does the gene violate any major guiding principles? | There is no significant overlap, the ORF is long enough, and the surrounding genes are also reversed. |
| **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 and Genemark both suggest 31857 bp as start.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final RBS score: -6.527<br>Z-score: 1.128<br>These scores are very low, and of the 4 potential start sites, these are the second lowest scores. |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 204 bp long, and there is a 4 bp overlap with the upstream gene. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This is the consensus start site, which doesn't mean much when there's only 12 other phages that have observed this gene. All 12 of those share this site, though. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 31857 bp because:*<br>• *Glimmer and Genemark both say*<br>• *Consensus start site* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |

| | |
|---|---|
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **31854** |
| **Stop Coordinate** | **31854** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **Overlap** |
| **Selected Start Coordinate** | **32150** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, it does seem to have some coding potenial.* |
| Is this gene present in other annotated genomes? | *No, it does seem to be presented in other annotated genomes.* |
| Does the gene violate any major guiding principles? | *It does not seem to be violating any major guidelines.* |
| **DECISION:** | *Respond here with YES or NO after reviewing the evidence gathered above.* |

## 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? | *Both call 31250* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS score- -2.523*<br>*Z-score- 2.959*<br>*Both are the best scores shown* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF= 297*<br>*It's the longest on and has does not result in a gene overlap.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Yes, the start site is conserved in other phage genomes.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Yes, it does seem to be conserved in other genomes.* |
| **DECISION:** | *PhagesDB, Glimmer, GeneMark, and starterator all can the same thing.* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 33262 bp |
| **Direction (For/Rev)** | Reversed |
| **Gap (Overlap) with Previous Gene** | -4 |
| **Selected Start Coordinate** | 33795 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GeneMarkS and GeneMark-host both show coding potential.* |
| Is this gene present in other annotated genomes? | *There are no annotated genomes with this observed gene, only 2 other cluster EA1 draft genomes.* |
| Does the gene violate any major guiding principles? | No significant violations, it is long enough, the surrounding genes are also reversed, and there is no significant overlap |
| **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: 33795 bp*<br>*Genemark: 33936 bp* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *33795 bp*<br>Final RBS: -2.443<br>Z-score: 2.959<br>Best scores<br><br>33936 bp:<br>Final RBS: -6.887<br>Z-score: .895<br>2nd lowest scores |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *For 33705, the ORF is 534 bp long, and for 33936 bp, the ORF is 675 bp long. Neither result in excessive gene overlap.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | No annotated genomes has this gene on starterator. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 33795 bp because:*<br>• *Glimmer says so*<br>• *Best z-score and RBS final score*<br>• *The Genemark one isn't a very good option.* |

## 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 | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:* |

| | |
|---|---|
| and/or GenBank) with an alignment of $10^{-4}$ or smaller with appropriate coverage? | *Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **33953** |
| **Stop Coordinate** | **33953** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **Gap** |
| **Selected Start Coordinate** | **34534** |
| **Selected Function** | **thymidylate kinase** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, there is some evidence of coding potenial.* |
| Is this gene present in other annotated genomes? | *No, it not in other annotated genomes.* |
| Does the gene violate any major guiding principles? | *It may have a bigger gap than allowed.* |
| **DECISION:** | *Respond here with YES or NO after reviewing the evidence gathered above.* |

## 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? | *Both call 34534 as the startsite* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS Score- -2.253*<br>*Z Score- 3.112*<br>*They are both the best scores.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF- 582*<br>*Yes, it does have a gap that does not violate any guidelines* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *No, it does not seem to be.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *No, it does not seem to be conserved in other genomes.* |
| **DECISION:** | *34534 Glimmer, GeneMark, PhagesDB, and Starterator all call the same.* |

## 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? | *PhagesDB: e-100; thymidylate kinase*<br>*NCBI: 4e-127; thymidylate kinase*<br>*HHpred: 2.1e-15; thymidylate kinase* |
| 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? | *NCBI: no above 90% scores*<br>*HHPred: 99.73 probability; Thymidylate kinase* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *no* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *thymidylate kinase* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **50** |
| **Stop Coordinate** | **34550** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **16** |
| **Selected Start Coordinate** | **35485** |
| **Selected Function** | *glycosyltransferase* |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | GeneMark S and Genemark-host both show coding potential |
| Is this gene present in other annotated genomes? | *It is not annotated in any other genomes for cluster EA1* |
| Does the gene violate any major guiding principles? | There are no significant violations. No significant overlap, gene is long enough, and surrounding genes are reverse. |
| **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? | *They both suggest 35485 bp as the start site* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS: -3.837*<br>*Z score: 2.474*<br>*This was the highest Z score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *The ORF length is 936 bp with a gap of 16. This was not the longest ORF but the longest had too much of an overlap.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | • Found in 2 of 175 ( 1.1% ) of genes in pham<br>• Manual Annotations of this start: 1 of 146<br>• Called 100.0% of time when present |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 35485 bp because it was called 100% of the time and its Z score and ORF length were good.* |

## 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? | *PhagesDB: an e value of e-177 that says glycosyltransferase*<br>*NCBI: an e vlaue of 0 that says glycosyltransferase*<br>*HHPred: e value of 3.7e-7 that says glycosyltransferase* |
| 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? | *HHPred: 98.9% probabilty match with glycosyltransferase*<br>*NCBI: 100% probabilty match with glycosyltransferase* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *YES* |
| **DECISION:** | *glycosyltransferase* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 35482 bp |
| **Direction (For/Rev)** | Reverse |
| **Gap (Overlap) with Previous Gene** | -4 bp |
| **Selected Start Coordinate** | 35700 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | Both Genemark-S and Genemark-host show coding potential |
| Is this gene present in other annotated genomes? | *88 other annotated phages of Cluster EA1 have observed this gene.* |
| Does the gene violate any major guiding principles? | No significant violations of any guiding principles. |
| **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 and Genemark say 35700 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | There's only one start site, so best scores. |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 219 bp long. No significant gap with upstream gene. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This start site is the consensus start site, with 107 of 128 annotated phages choosing this start site for this gene. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 35700 bp:*<br>• *Glimmer and Genemark agree*<br>• *Only option for start site*<br>• *Consensus start site on starterator* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |

| | |
|---|---|
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 35763 bp |
| **Direction (For/Rev)** | Reverse |
| **Gap (Overlap) with Previous Gene** | 64 bp |
| **Selected Start Coordinate** | 36041 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host both show coding potential.* |
| Is this gene present in other annotated genomes? | 104 annotated phages of cluster EA1 have observed this gene. |
| Does the gene violate any major guiding principles? | There are no significant violations. |
| **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 and Genemark agree on 36041 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | RBS final: -2.253<br>Z: 3.112<br>Highest scores |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF is 279 bp long, and there is no significant overlap with this start site.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This is the consensus start site on starterator, chosen in 104 of 126 annotated phages. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 36041 bp because:*<br>• *Glimmer and Genemark say*<br>• *Best scores*<br>• *Consensus start site* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |

| | |
|---|---|
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **36038** |
| **Stop Coordinate** | **36038** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **Gap** |
| **Selected Start Coordinate** | **36805** |
| **Selected Function** | **thymidylate synthase** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, it does seem to have some coding potential.* |
| Is this gene present in other annotated genomes? | *It does seem to be present in other annotated genomes.* |
| Does the gene violate any major guiding principles? | *No it does not violate any guidelines.* |
| **DECISION:** | *Respond here with YES or NO after reviewing the evidence gathered above.* |

## 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 36805*<br>*GeneMark Start Coordinate 36883* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *36805*<br>*RBS: -2.443*<br>*Z-score: 2.959*<br>*36883*<br>*RBS: -7.044*<br>*Z-Score: 1.125*<br>*36805 is the highest scoring* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *36805 ORF- 768*<br>*36883 ORF- 846*<br>*Either are the longest* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *You will also need to provide the following information from Starterator: does the start match the consensus start site predicted from Starterator? If no, is the consensus start site not found in this ORF? If no, is there a better option for the consensus start site instead of the one predicted by Starterator? If Starterator doesn't reveal a consensus start site, you can record that Starterator was not informative.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *36805:*<br>• *Glimmer say so* |

| | |
|---|---|
| | • *Higgest z and final rbs scores* |

## 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? | *PhagesDB: e-146; thymidylate synthase*<br>*NCBI: 0.0; thymidylate synthase*<br>*HHpred: 5.3e-39; CMP hydroxymethylase* |
| 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? | *NCBI: 97.25% identical: Thymidylate synthase*<br>*HHPred: several 100 probabilities; some "CMP hydroxymethylase" and some "thymidylate synthase"* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA- | *yes* |

| | |
|---|---|
| PHAGES approved function list? | |
| **DECISION:** | *thymidylate synthase* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **54** |
| **Stop Coordinate** | **36935** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **360** |
| **Selected Start Coordinate** | **37147** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | GeneMark S and Genemark-host both show coding potential |
| Is this gene present in other annotated genomes? | Start number 2 was manually annotated 62 times for cluster EA1. |
| Does the gene violate any major guiding principles? | The only real violation is that overlap is over 30 bp. The genes surrounding are also reverse, and the gene is long enough |
| **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? | *They both suggest 37147 bp as the start site however start 37165 is a better choice.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS: -6.152*<br>*Z score: 1.399*<br>*This was the highest Z score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *The ORF length is 231 bp with a gap of 360. This was the longest ORF length.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | • Found in 101 of 127 ( 79.5% ) of genes in pham<br>• Manual Annotations of this start: 62 of 110<br>• Called 61.4% of time when present |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 37165 bp.*<br>• *Has the longest ORF length*<br>• *Highest Z score*<br>• *Has more manual annotations*<br>• *Is called 61.4 % of the 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **55** |
| **Stop Coordinate** | **37224** |
| **Direction (For/Rev)** | **Rev** |
| **Gap (Overlap) with Previous Gene** | **77** |
| **Selected Start Coordinate** | **37463** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes* |
| Is this gene present in other annotated genomes? | *Annotated104 times in cluster EA1* |
| Does the gene violate any major guiding principles? | *No major violations* |
| **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? | *Both suggest bp @ 37463* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS Highest Score: -2.443*<br>*Z value: 2.959* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF is 240 bp long, this is the longest ORF, and this doesn't result in excessive overlap.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *This phage doesn't have the consensus start site.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Start should be at 37463*<br>- *Both Glimmer and GeneMark indicate this* |

# 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 37555 bp |
| **Direction (For/Rev)** | Reverse |
| **Gap (Overlap) with Previous Gene** | 93 bp |
| **Selected Start Coordinate** | 37827 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | GenemarkS and Genemark-host both show coding potential. |
| Is this gene present in other annotated genomes? | 101 annotated phages from cluster EA1 have observed this gene. |
| Does the gene violate any major guiding principles? | No significant violations. |
| **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 and Genemark both say 37827 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | RBS final: -2.253<br>z-score: 3.047<br>Best Scores |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF with this start site is 273 bp long. There is no significant overlap with this start site. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | This is the consensus start site, called in 110 of 128 non-draft genes. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 37827 bp:*<br>• *Glimmer and Genemark*<br>• *Scores are the highest*<br>• *Consensus start site* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |

| | |
|---|---|
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **57** |
| **Stop Coordinate** | **37897** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **70** |
| **Selected Start Coordinate** | **38163** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | GeneMark S and Genemark-host both show coding potential |
| Is this gene present in other annotated genomes? | •Start number 8 was manually annotated 95 times for cluster EA1. |
| Does the gene violate any major guiding principles? | The only real violation is that overlap is over 30 bp. The genes surrounding are also reverse, and the gene is long enough |
| **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? | *They both suggest 38163 bp as the start site* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS: -2.584*<br>*Z score: 2.959*<br>*This was the highest Z score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *The ORF length is 267 bp with a gap of 70. This was the longest ORF length.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | • Found in 114 of 114 ( 100.0% ) of genes in pham<br>• Manual Annotations of this start: 96 of 96<br>• Called 100.0% of time when present |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 38163 bp.*<br>• *Has the longest ORF length*<br>• *Highest Z score*<br>• *Has more manual annotations*<br>• *Is called 100.0 % of the 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **38220** |
| **Stop Coordinate** | **38220** |
| **Direction (For/Rev)** | **reverse** |
| **Gap (Overlap) with Previous Gene** | **Gap** |
| **Selected Start Coordinate** | **38714** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, there seems to be some coding potential.* |
| Is this gene present in other annotated genomes? | *Yes, it does seem to be present in other annotated genomes.* |
| Does the gene violate any major guiding principles? | *It might have too big of a gap between it and the earlier gene.* |
| **DECISION:** | *Respond here with YES or NO after reviewing the evidence gathered above.* |

## 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? | *Both call 38714* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS- -4.271*<br>*Z-Score- 2.110*<br>*They are not the best scores shown.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF- 495*<br>*It has a large gap between it and the previous gene.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *No, it does not seem to be conserved in other genomes.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *No, it does not seem to be conserved in other phage genomes.* |
| **DECISION:** | *Glimmer, genemark, and phagesdb all call 38714* |

## 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 | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:* |

| | |
|---|---|
| and/or GenBank) with an alignment of $10^{-4}$ or smaller with appropriate coverage? | *Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **59** |
| **Stop Coordinate** | **38711** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **-3** |
| **Selected Start Coordinate** | **38890** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | Both Genemark-S and Genemark-host show coding potential |
| Is this gene present in other annotated genomes? | *•Start number 20 was manually annotated 2 times for cluster EA1.* |
| Does the gene violate any major guiding principles? | No significant violations of any guiding principles. |
| **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 and Genemark say 38890 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS Score: -5.783*<br>*Z Score: 1.437* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF length is 180 bp* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *• Found in 15 of 141 ( 10.6% ) of genes in pham*<br>*• Manual Annotations of this start: 11 of 119*<br>*• Called 100.0% of time when present* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site is at 38890 bp*<br>• *Was called 100% of the 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 38892 bp |
| **Direction (For/Rev)** | Reverse |
| **Gap (Overlap) with Previous Gene** | 3 bp |
| **Selected Start Coordinate** | 39620 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host both show coding potential.* |
| Is this gene present in other annotated genomes? | *This gene was observed in 104 annotated phages from cluster EA1.* |
| Does the gene violate any major guiding principles? | There are no significant violations of any of the major guiding principles. |
| **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 and Genemark both say start site at 39620 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Z-score: 3.047<br>Final RBS: -2.253<br>Highest scores available |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 729 bp long, and there would be no significant overlap with this start site. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | The consensus start site isn't found in this ORF, but this start site was chosen in 156 of the 426 annotated genomes (the consensus has 218 of 426). |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 39620 bp because:*<br>• *Glimmer and genemark agree*<br>• *Highest z score and final RBS score*<br>• *Not consensus, but start site still has many manual annotations* |

## 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and* |

| | |
|---|---|
| | *e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **39716** |
| **Stop Coordinate** | **39716** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **Gap** |
| **Selected Start Coordinate** | **40051** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, it does show coding potential.* |
| Is this gene present in other annotated genomes? | *It does not seem to be present in other genomes.* |
| Does the gene violate any major guiding principles? | *No, it does not seem to violate any guidelines.* |
| **DECISION:** | *Respond here with YES or NO after reviewing the evidence gathered above.* |

## 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? | *Both call 40051* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS- -2.253*<br>*Z-Score- 3.047*<br>*Both are the best scores shown.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF- 336*<br>*The longest score was 918 and it results in major gene overlap.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *No, the start site does not seem to be conserved in other genomes.* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *No, the site does not seem to be conserved in other genomes.* |
| **DECISION:** | *40051 Glimmer, Phagesdb, and startaretor all call it.* |

## 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 | *List the most informative BlastP match from each source*<br>*PhagesDB:* |

| | |
|---|---|
| functional assignment in BlastP (phagesDB and/or GenBank) with an alignment of $10^{-4}$ or smaller with appropriate coverage? | *NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | Acosta |
| **Gene #** | |
| **Stop Coordinate** | 40096 bp |
| **Direction (For/Rev)** | Reverse |
| **Gap (Overlap) with Previous Gene** | 46 bp |
| **Selected Start Coordinate** | 40437 bp |
| **Selected Function** | Hypothetical Protein |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host show coding potential.* |
| Is this gene present in other annotated genomes? | This gene has been observed in 104 annotated phages of cluster EA1. |
| Does the gene violate any major guiding principles? | There are no significant violations of the major guiding principles. |
| **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: 40437 bp*<br>*Genemark: 40455 bp* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *40437:*<br>*Z: 1.981*<br>*Final RBS: -4.548*<br>*Average scores*<br><br>*40455:*<br>*Z: 2.002*<br>*Final RBS: -4.644*<br>Average scores |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *40437: 342 bp*<br>*40455: 360 bp*<br>*Neither result in excessive overlap* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | 40455 bp is the consensus start site, having 134 of 138 non-draft genes choosing this start.<br>40437 bp only has 2 annotated genomes with this start site. |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site should be at 40455 bp because:*<br>• *Genemark says this site*<br>• *The z-score and final RBS score isn't the lowest*<br>• *This is the consensus start site.* |

## 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 | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:* |

| | |
|---|---|
| and/or GenBank) with an alignment of $10^{-4}$ or smaller with appropriate coverage? | *Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **63** |
| **Stop Coordinate** | **40452** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **15** |
| **Selected Start Coordinate** | **40652** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | Both Genemark-S and Genemark-host show coding potential |
| Is this gene present in other annotated genomes? | *•Start number 17 was manually annotated 20 times for cluster EA1.* |
| Does the gene violate any major guiding principles? | No significant violations of any guiding principles. |
| **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 and Genemark say 40652 bp.* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *RBS Score: -4.138*<br>*Z Score: 2.474* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF length is 174 bp. This was not the highest ORF length but the highest on had an overlap over 30bp.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Found in 124 of 124 ( 100.0% ) of genes in pham*<br>*Manual Annotations of this start: 20 of 104*<br>*Called 29.8% of time when present* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The start site is at 40652 bp*<br>• *Has the only manual annotations*<br>• *Good orf length*<br>• *Good overlap length* |

# 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **64** |
| **Stop Coordinate** | **41746** |
| **Direction (For/Rev)** | **For** |
| **Gap (Overlap) with Previous Gene** | **528** |
| **Selected Start Coordinate** | **41180** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *YES* |
| Is this gene present in other annotated genomes? | *Start 28 is annotated 1 time for cluster EA1* |
| Does the gene violate any major guiding principles? | *Overlap is over <30 bp, but there is no overlap* |
| **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? | *Both indicate 41180* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *Highest RBS Score: -4.820*<br>*Z value: 1.920* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *ORF Length: 567*<br>*Not the longest and does result in excessive gene overlap* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Found in 3 of 62 ( 4.8% ) of genes in pham* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Start should be bp@ 41180 because*<br>- *Glimmer and GeneMark indicate*<br>- *No overlap* |

# 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | |
| **Stop Coordinate** | **20593bp** |
| **Direction (For/Rev)** | **Forward** |
| **Gap (Overlap) with Previous Gene** | **-60** |
| **Selected Start Coordinate** | **20180 bp** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, there is coding potiental* |
| Is this gene present in other annotated genomes? | *Start number 3 was manually annotated 2 times for cluster EA1.* |
| Does the gene violate any major guiding principles? | *There are no significant violations, it is long enough, there is no overlap with the upstream gene however, the gene afterward goes in reverse 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 and Genemark agree on 20180 bp* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *Final RBS: -4.395*<br>*Z-score: 2.297*<br>*This was the highest Z score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *The ORF is 414 bp with a gap of -60 bp between the upstream gene.*<br><br>*This was the highest ORF length* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Found in 3 of 71 ( 4.2% ) of genes in pham*<br>*Manual Annotations of this start: 2 of 60*<br>*Called 100.0% of time when present* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *I think the start site should be 20180 bp because:*<br>• *Glimmer and GeneMark agree*<br>• *Has the longest ORF length*<br>• *Has the highest Z score*<br>• *Called in 100% when present* |

# 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **30** |
| **Stop Coordinate** | **20903** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **58** |
| **Selected Start Coordinate** | **21073** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, there is coding potential* |
| Is this gene present in other annotated genomes? | *Start number 13 was manually annotated 77 times for cluster EA1* |
| Does the gene violate any major guiding principles? | *There are no significant violations, it is long enough, there is no overlap with the upstream gene, and the genes around go 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 and Genemark agree on 20903 bp however the best possible start site will be 21073 bp* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *21073 bp*<br>*Final RBS Score- (-3.692)*<br>*Final Z Score- 2.417*<br><br>*This was the highest Z score* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *21703 bp had an ORF length of 171 and this was the longest length.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Found in 123 of 139 ( 88.5% ) of genes in pham*<br>*Manual Annotations of this start: 77 of 119*<br>*Called 62.6% of time when present* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *I think the starts site should be 21073 bp because it has the highest Z score and ORF length. On Genemark it had the most coding potential.* |

# 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **36** |
| **Stop Coordinate** | **24464** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **-24** |
| **Selected Start Coordinate** | **24757** |
| **Selected Function** | *Nuclease* |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *Yes, there is coding potential* |
| Is this gene present in other annotated genomes? | *Start number 36 was manually annotated 103 times for cluster EA1* |
| Does the gene violate any major guiding principles? | *There are no significant violations, it is long enough, there is no overlap with the upstream gene, and the genes around go 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 and Genemark agree on 24757 bp* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *Final RBS Score- (-6.742)*<br>*Final Z Score- 1.028*<br><br>*This was the highest RBS score.* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *24757 bp had an ORF length of 294 and this was not the longest length. The longest length had an overlap greater than 30 bp.* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Found in 182 of 255 ( 71.4% ) of genes in pham*<br>*Manual Annotations of this start: 152 of 213*<br>*Called 95.6% of time when present* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *I think this should be the start site because it has the highest RBS score and both glimmer and genemark agree on this start. This start was also called in 95.6 % of the time when present.* |

# 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:* 2e-51<br>*NCBI:* 2e-65<br>*DNA Master:* 3.5e-19 |
| 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? | *HHPred: Nuclease has 99.85% that matches.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *Yes* |
| Is this gene a possible transmembrane protein? | *No* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *HHPred shows 99.85% match to Nuclease* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **46** |
| **Stop Coordinate** | **32150** |
| **Direction (For/Rev)** | **Reverse** |
| **Gap (Overlap) with Previous Gene** | **0** |
| **Selected Start Coordinate** | **32413** |
| **Selected Function** | **Hypothetical Protein** |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Examine the data for the auto-annotation programs and indicate (YES BOTH, YES Glimmer only, YES GeneMark only, Neither)* |
| Is there evidence for coding potential? | *GenemarkS and Genemark-host both show coding potential.* |
| Is this gene present in other annotated genomes? | •Start number 16 was manually annotated 7 times for cluster EA1. |
| Does the gene violate any major guiding principles? | The only violation is that the ORF length is below 120bp. It is long enough, and the surrounding genes are also reverse |
| **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? | *They both suggest start at 32413 bp* |
| Does the start site have an associated Ribosome Binding Site with a high score? | Final: -5.833<br>Z: 1.872<br>This was not the highest Z score |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | ORF is 39 bp long, with a 0 bp gap. |
| Is this start site conserved in other phage genomes as indicated by Starterator? | • Found in 22 of 148 ( 14.9% ) of genes in pham<br>• Manual Annotations of this start: 8 of 125<br>• Called 45.5% of time when present |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *The selected start site is at 32413 bp*<br>• *Has more manual annotation then the other start sites* |

# 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? | *List the most informative BlastP match from each source*<br>*PhagesDB:*<br>*NCBI:*<br>*DNA Master:*<br><br>*Hint: you may have already found this information from annotation decision #2. Provide the alignment (q#:s#) and e-value. It is only necessary to provide one match from each database.* |
| 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? | *List the most informative HHPred match, including database source and probability score. It is only necessary to provide the best match.*<br><br>*Note: If you believe there is not a quality HHPred match, type No Quality Match and list the data for the best match available to affirm the poor quality of the result and to document that HHPred was considered.* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *If the answer is YES, indicate supporting data from at least 2 different transmembrane prediction programs.* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Indicate a response with a Yes or No response.*<br>*Once you have arrived at a functional decision, check the SEA-PHAGES Official Function List to ensure that you are following the guidelines for function naming. Functions that are not present on the approved list must be carefully vetted for approval.* |
| **DECISION:** | *If you believe this gene should be assigned, please write the name of the function here. If the evidence does not support a functional call, record "NKF" for no known function. 50-70% of phage genes fall into the NKF category.* |

# Student Gene Annotation Worksheet

| Basic Phage Information | |
|---|---|
| **Phage Name** | **Acosta** |
| **Gene #** | **47** |
| **Stop Coordinate** | **33265** |
| **Direction (For/Rev)** | **Rev** |
| **Gap (Overlap) with Previous Gene** | **4** |
| **Selected Start Coordinate** | **32417** |
| **Selected Function** | *MazG-like nucleotide pyrophosphohydrolase* |

## Annotation Decision #1: Is this a Gene?

| Gathering Evidence | Explain Your Rationale |
|---|---|
| Was the gene called by an auto-annotation program (Glimmer, GeneMark)? | *Both Glimmer and GeneMark* |
| Is there evidence for coding potential? | *Coding potential in GeneMark* |
| Is this gene present in other annotated genomes? | *104 times for cluster EA1* |
| Does the gene violate any major guiding principles? | *No major violations. An overlap of 4* |
| **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 and GeneMark both indicate 32417* |
| Does the start site have an associated Ribosome Binding Site with a high score? | *Highest RBS Score: -7.128*<br>*Z score: 0.946* |
| Is the predicted start codon the longest ORF? If not, does the longest ORF result in excessive gene overlap (>30bp)? | *Overlap of 4*<br>*Longest ORF length: 42* |
| Is this start site conserved in other phage genomes as indicated by Starterator? | *Found in 125 of 161 (77.6%)* |
| Is this start site conserved in other phage genomes as indicated by BlastP? | *Provide the best BlastP match from NCBI, PhagesDB, and DNA Master with alignment in the format of (Q#:S#), where Q (query) is the sequence you are analyzing and S (subject) is the database match. List the e-value and alignment of the best match for all three BlastP sources.*<br><br>*Note: if you are considering more than 1 start site, provide the same information for each proposed start site.* |
| **DECISION:** | *Start should be at 32417 bp because*<br>- *Both Glimmer and GeneMark Indicate this*<br>- *No major overlap* |

## 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? | *PhagesDB: e-121; MazG-like nucleotide pyrophosphohydrolase*<br>*NCBI: 1e-152; nucleotide pyrophosphohydrolase*<br>*HHpred: 8.3e-31; putative NTP pyrophosphohydrolase* |
| 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? | *NCBI: no high % identical*<br>*HHPred: 99.97 probability; putative NTP pyrophosphohydrolase* |
| Is this gene located adjacent to genes of known function and in a region of the genome that shows high conservation of gene order? | *If the answer is YES, evaluate the proposed function in the gene order. Examine the adjacent genes found in the most closely related annotated phage (hint: use Phamerator) and record the function of the genes found on each side of the gene in the same pham in the most closely related phage. If the answer is NO, enter No Synteny Observed.* |
| Is this gene a possible transmembrane protein? | *no* |
| Is the proposed function found on the SEA-PHAGES approved function list? | *Yes* |
| **DECISION:** | *MazG-like nucleotide pyrophosphohydrolase* |