**How do I identify a frameshift mutation?**

Programmed Translational Frameshift (PTFS) mutations occur in many, but not all, bacteriophage genomes and require special attention when annotating. But how do you know your genome has a PTFS? If your genome has a PTFS, how do you determine which genes are involved? This activity is designed to teach you how to detect a PTFS in a genome, and how to identify which genes are involved using 3 pieces of evidence: synteny, BLAST, and Phamerator.

1. **Synteny – where do frameshifts occur in relation to other genes?**

The first piece of evidence we can use to detect a frameshift mutation is synteny. Synteny refers to the order of genes. Usually, gene order is conserved in bacteriophages, so you will find the same gene/gene functions in the same sequence across similar phages. Although not all bacteriophage genomes contain a PTFS, if they do have one it is found in the structural gene region of the genome, usually the two genes prior to the tape measure protein. These genes are also usually associated with tail assembly, and are given the function of tail assembly chaperone. As you annotate you will assign functions, if possible, to the genes that you verify. The genes prior to a PTFS may include major and minor tail proteins and tail assembly chaperones and the genes after may include the tape measure protein and minor tail proteins. Usually, *but not always,* the genes involved in the PTFS are the two directly upstream of the tape measure protein. As you annotate this region you should be aware that a PTFS may be present. You can use additional pieces of evidence to detect and verify the presence of a PTFS. Complete the activity below using synteny to answer the questions, then continue on to learn about other evidence you can use to identify a PTFS.

**Activity 1:**

Analyze each scenario using synteny and determine if a PTFS is likely.

1. A student believes they are dealing with a PTFS in genes 63 and 64. Look at the functions assigned to the surrounding genes to determine if this is likely.

|  |  |
| --- | --- |
| **Gene #** | **Function assigned** |
| 60 | HTH DNA binding protein |
| 61 | Phosphoesterase |
| 62 | No Known Function |
| 63 | No Known Function |
| 64 | No Known Function |
| 65 | DNA primase |

Does synteny support this possibility? Why or why not?

1. A student believes they are dealing with a PTFS in genes 10 and 11. Look at the functions assigned to the surrounding genes to determine if this is likely.

|  |  |
| --- | --- |
| **Gene #** | **Function assigned** |
| 9 | Minor tail protein |
| 10 | Tail assembly chaperone |
| 11 | Tail assembly chaperone |
| 12 | Tape measure protein |
| 13 | Minor tail protein |

Does synteny support this possibility? Why or why not?

1. Given the assigned functions of these genes, which two genes are good candidates for containing a PTFS?

|  |  |
| --- | --- |
| **Gene #** | **Function assigned** |
| 22 | Major tail protein |
| 23 | Tail assembly chaperone |
| 24 | Tail assembly chaperone |
| 25 | Tape measure protein |
| 26 | Minor tail protein |
| 27 | Minor tail protein |

1. The tables below show the gene assignments for 3 different bacteriophages. Answer the following 3 questions about these phages.

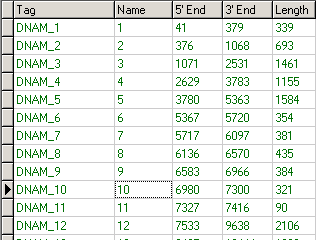
|  |  |  |  |
| --- | --- | --- | --- |
| Gene # | Phage #1 | Phage #2 | Phage #3 |
| 22 | Major tail protein | Major tail protein | No Known Function |
| 23 | Tail assembly chaperone | Tail assembly chaperone | Minor tail protein |
| 24 | Tail assembly chaperone | Tail assembly chaperone | Lysin A |
| 25 | Tape measure protein | Tape measure protein | No Known Function |
| 26 | Minor tail protein | Minor tail protein | Holin |
| 27 | Minor tail protein | Minor tail protein | No Known Function |

1. In each bacteriophage where is the frameshift likely to occur (what gene numbers)?
2. Does bacteriophage #3 have a PTFS? Explain your answer.
3. Based on this set of genes – do you think these 3 phages are from the same cluster? Why or why not?
4. **BLAST – What are the signs that you have a frameshift?**

BLAST takes the sequence of your gene and compares it to a database of other sequences. This tells you how similar the gene is to other genes from other phages, as well as how your gene characteristics (its start location, function, etc.) compare to other annotated genes from different bacteriophages. Let’s look at what BLAST results look like if a PTFS has occurred.

A PTFS will be called as two separate genes by gene prediction software programs like Glimmer and GeneMark. Usually this results in one ‘normal’ sized gene, and one very small sized, or short, gene. This is because the first gene is usually the only one transcribed, and the infrequently the ribosome slips and shifts frame to produce a larger gene (including the short ‘second’ gene that was originally predicted). This means that as you annotate, the first gene will appear normal when you compare it against the NCBI and phagesdb databases, but the second gene will appear odd. Notably, the second gene will not have the same start as what was predicted. You should see that the predicted start is much farther downstream of the starts in non-draft genome annotations. Let’s work through an example together. Our example is Sippinontea – a cluster EE bacteriophage. Sippinontea’s tape measure protein gene is gene #12, and based on synteny it appears that gene #’s 10 and 11 could contain a PTFS.

If we look at our original DNA Master auto-annotation gene predictions we see that gene 10 has a length of 321 bp and gene 11 is only 90 bp in length.



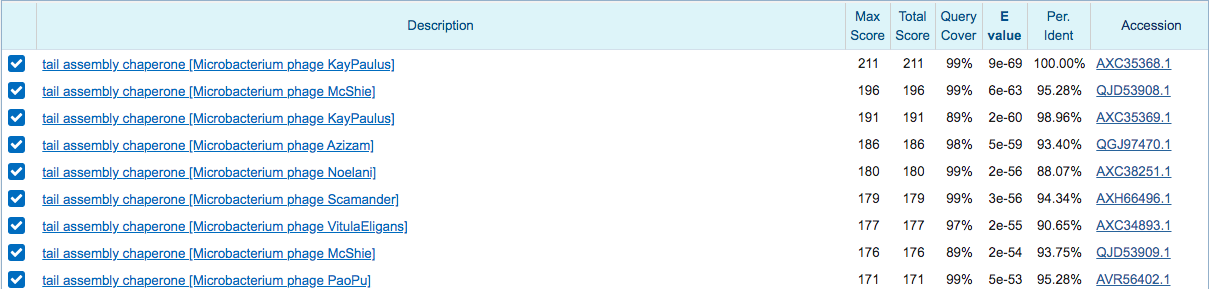
Here are the amino acid sequences for these two predicted gene products:

>Sippinontea Draft gene 10

METTKKKFDLDQLTLGEVAAIEDLSGVAIGSVSESTPQGKFLAALYMVAKRRDGQPTFTFNAALQASMSEAQSFLGFDAPDAGDEADEAESSAEGNGDSSPENAPAZ

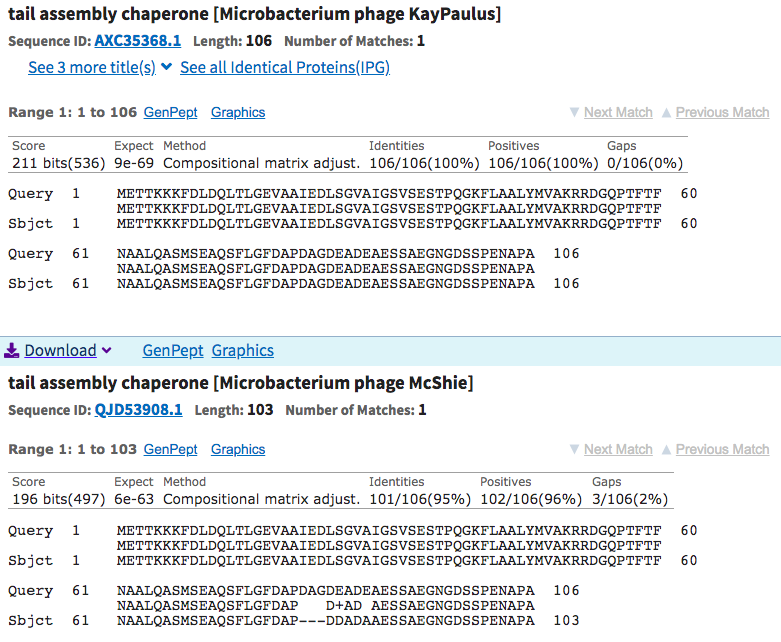
>Sippinontea\_Draft gene 11

MDPAAYERLTIAERDAIVRELNKRNSRRRZ

If we blast gene 10’s product (blastp)against the NCBI database we get the following results:

A glance over the descriptions tells us this gene matches with other bacteriophage genes, and that these genes have been assigned the function ‘tail assembly chaperone’.

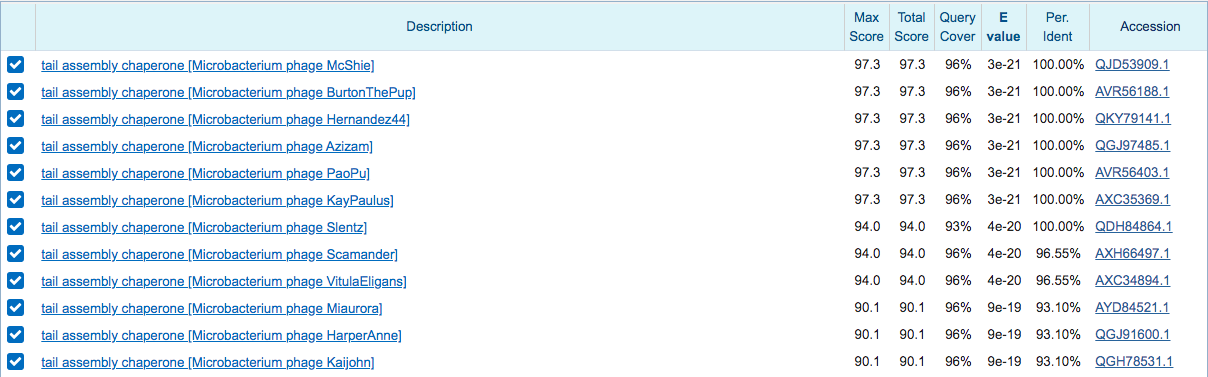
If we look at the alignments we see that these matches align 1:1 with our query sequence. This tells us that the predicted start is the same start that other annotations have called.

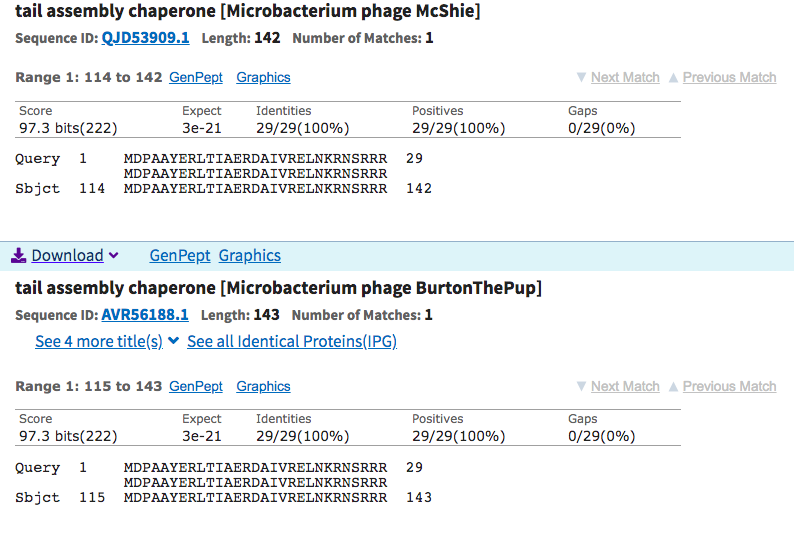


These results suggest that gene 10 functions normally as a gene. There is no sign of a frameshift mutation or of anything unusual. This is because the first gene in the PTFS of phages functions as a normal gene most of the time. It produces a product that is needed in a higher quantity than the longer version of the gene (which occurs when the ribosome slips and shifts frames).

Now let’s look at the next gene, gene 11.

When we BLAST the animo acid sequence of gene 11 using a protein blast we get these results:



As you can see, gene 11 aligns well with other phage genes, and also has the assigned function of tail assembly chaperone. Now let’s look at the alignments:

These alignments are NOT aligning 1:1 with the other genes in the database. Instead, you can see that the first amino acid of Sippinontea’s gene 11 (the Query) is aligning with the 114-116th amino acid in the other phage sequences to which it has matched. This tells us that gene 11’s start has been annotated upstream of the start that we have from our auto-annotation. In other words, our start does not seem to be correct. If we do the math (3 x 114) we can determine that the start called in these other genomes is anywhere from 342- 348 bp upstream of where ours is called. This is a large difference in start positions – and if we look at where that would put us in our genome, it would mean starting gene 11 at about the same start position as the previous gene (gene 10)!

While bacteriophage genomes are renowned for being tightly packed full of genes, one thing that does not usually happen is a large overlap of multiple genes in the same position. Each gene tends to occupy its own space, with minimal overlap at the ends. A significant overlap in genes – like what is suggested by these BLAST results, is another piece of evidence that you are looking at a PTFS. If synteny tells you that the region of the genome could contain a frameshift, and your blast results suggest that there is complete overlap with another gene, you may have identified a PTFS.

Another way to investigate the start of the second gene involved in a PTFS is to look at the starterator report (found on phagesdb). Starterator will either be missing a report for this second gene (because it only exists as a PTFS) or it will have a report containing few phages and only those with draft (only auto) annotations. This is another sign that your gene has something unusual going on – and you may be dealing with the PTFS.

Complete this activity to get practice identifying a possible PTFS with Blast. Then we’ll look at a third piece of evidence you can use to tell that you have a PTFS.

**Activity 2:** For each of the following phages there are several gene sequences. Analyze each of these gene sequences using BLAST to assess whether a PTFS may have occurred, and identify which of the two genes are involved.

1. **Bacteriophage: Scout**

>Scout\_Draft gp22

MSKIFTLDSFREEVEKEFAPVKIEVDADNSVVLRNLLRIPKGAREEIFGLLERMDKMSEGKSEDEMTVEELEATAGIALRMIELVADTPAGGRILVESLEDDLALTLKVFEAWMEATNPGEAPRSHD

>Scout\_Draft gp23

MDGGHESGGSAALARLIDDYGDAVAADLMETYGVDLRDLFVPESRLTPKWVLVLIKELPVGSRFYSEKRGGPQFRGWDESRYTLAAIVNAVRALQHTYLAAHMKSTPKPPEPYPTPDRNTRKKNNNKPNSFASIAAQMIAAKRAKKARKAAQE

>Scout\_Draft gp24

MAGGGAGGTEVGRISIRVVPNLDNFYRELKTKLEAIEKQLRGNVPIDIDLNARGTRAKMAALMAGLKAQAAQGVDVPVDVNNKGLGAAWREFRAGLADFGRLGKQAAQGVKSYRDEVNRLTLEQQRQRPLLNHTYAWWRSNNIMARRGATILRDFTDALRTQQQWLRQQDRTLTANQARWKSWAMAIRDANVNATNGFRRFRASLQALRGGGGGDDGDGFSRIFGSLGRFGNEAEKAGSQVEHVGKKFLGLTRMGWLVTGVFLAAAPAIALVSGLLAGLPSLIGAFGAGIGAVALGMDGIKAAAEVLMPAFEQMKTAVSSTFQQALVPQFQQLLGLMPMIQTGMQGVAQGMSSMFQGVTDALSKGAGPAQIENLLANTKTFFEQLQPAANQFTQSFLTLASSGSDAFGYLSGSLNTFSTQFNDMVNRVSQNGVMDGAMKGLSQTLDGVTNLFTRLMESGLQAMSQLGGPMNTFLTGIGDLAVALMPALTSLSGLFGNVAGTLGTALAPIVTALTPAFTTLADTLGSLLVPNIQTLGNILTPVATMIGTTLTTALQQIQPMIPGLVESFAQLGSTLVSQLAPHIPALATAMGQMAGAVIKLAPMLISQLVPAFIDLIPSITQLLPHVVSLAESFARMMPTIVPLVSIIFSLIAAFAQAAATIGGVVLGAISSLIGVISEVVAKISEWVSSFAQGVSDIAAKAAELPGMVKSALGDLGSFLVSSGKALVQGFINGIKSMVGAVADAARSVVQAARDFFPFSPAKKGPFSGSGWVDASGQSVGEAFADGLAGTQGKIVETARAIMQAAKDVFGDAANIAFNFNFGQMQSQMASVASSAGDLQRSMSRTVSQSTGSGKIDDETRQMLDQISIRKDELELERQRLQAEKNALDTKDKAGRAALQQRIDELNIQKDQLELQREQLSYQSKYTDSVAQTGAQYDEMFNKLTRMPYDFATANANQFLSDIGISGDGALSQALKEGLKFGEQFIFNVGSMDEAVQGQQTIQNKKSLQFDRR

* + 1. Blast each of these genes and record your results (Does the gene align well with other phages, are the start calls the same?).
    2. Is there a frameshift in this gene sequence?
    3. If yes - which two genes are involved? How do you know? If no, what made you reach this conclusion?

2. **Bacteriophage Pheobe**

>Phoebe\_Draft gp24

MALNDDAVLTAAVGYVYTAPVGTAAPTPAQLKTLNLTDTGLWTPTGWDSIGHTSRGDMPEFGFDGGDTEVRGSWQKKKLREVTTEDPVDYLTLFLHQFDEQAFELYYGANASTTPGVFGVSAASGDPTEKAFLVVIVDGDERVGFHAHKASVRRDDAIQLPTDDFAALPVRATFLQHNNELLFSWINEDLFNVEEEEE

>Phoebe\_Draft gp25

MSNVFTLDSFREEADREFAPVKLELGGDDAVVLRNVLRIQKTRREEVFQLLEKLDSIAKDDEGKQREEDDLDASEMEAMGDIALRMIELVADNDALGSRLVDELRDDLALTLKVFEAWMNATQPGEAERSPA

>Phoebe\_Draft gp26

MDLRDIYLPESRLSPKLALVLIKELPVGSRFYAEKRGGKQFRGWDESRYALVAIVNAVRALQYTYVAAHSKSKPKPPDPFPTPQRTKARQIRKAGSFAWMAAKQIAAARKRKAQT

1. Blast each of these genes and record your results (Does the gene align well with other phages, are the start calls the same?).
2. Is there a frameshift in this gene sequence?
3. If yes - which two genes are involved? How do you know? If no, what made you reach this conclusion?
4. **Bacteriophage Quartz**

>Quartz\_Draft gp13

MPLTQWNPATQISRGNVAVGVAPAVVSLDAPALSELTTGIGLDCSITTMNGTSSTDSESIDWLCDPASEQLPGSTTHAMDDLVIKGTGQDDADLIAGLNIGDVVYVWRRDGIPHDTAPAAGQFVWVWKVIITSIDPLEANNTFVGITAHITVLARSKTAVAIAA

>Quartz\_Draft gp14

MSFSSYEELKAAVDERRKDILTIEVDLGARYSQDHEDAKKELAQAEAIQKLAGGGQGFLNDNLEQLKARVAETKPEANSIWLRYGRLQLAEWSMLTKATGLTPIDQYEKVLPQTFKGVYGVDPTAEDEEGNLLHPDAEPLTTDARAVSSRSEETVLPGAMLHVVVNAFMTWQNSSGEISIRPTKSGRV

>Quartz\_Draft gp15

MQAIAIGQDMWKTANKTAIESHQKGNGPDPGMGVYWLSQGEGEVLPTPET

1. Blast each of these genes and record your results (Does the gene align well with other phages, are the start calls the same?).
2. Is there a frameshift in this gene sequence?
3. If yes - which two genes are involved? How do you know? If no, what made you reach this conclusion?

**4. Bacteriophage Phegasus**

>Phegasus\_Draft gp18

MTQDKQTFPPGPATLLGQELMLEHTDPLIMLTTADRKVTFYLSGGLAAWPRHQDGVNLVEITTPTPEFRNLRAQGARQDGGQTRDTVYDPMQIDAVLLASATTPEGLSRVVSEWIAANDPEQLCRLEWFTFEGGLWWCDVRLEKRFIDRLQQSPRRVKKQVLSTVWMNDLAFWQSVDSTCTWAFSYQTMLDTFKYDTSASKDLGENWPQYRYDGEGGGYWYANGDRAVWRDDPEDPLLTDGVSVLCGPYKDFETATDYQVIDFVIGSFQEITFPDGAENHAWGRLNRDEDGEWAGDGIRASVGPTSAVLHRFNDFEKTRIGLPVPLFPPPFIGEKFRLIIGYQGNPRKYRLLRALTDRSAGVPVLTVTEQGTGSAIGPDHRGIGFGGRAGAALLTQATPASVRKVAAGDNRTETQEGFLTLTNIGERDGWPQIVFEGPGLLEIANGPGSTDMIKFGPLEDGQRVLISTHPRYRAIVDLTQGQVGQQLDGGQKLIDTIVKLLSLGQVPPALQWFESVFGIKPPQGPLYSLLDGRFTRPIPGVRQPRDATTSRIAIRVRDGNANTKVTASVTPMRRWPEAVHD

>Phegasus\_Draft gp19

MPRVENIGDYLDLAEIQRKLLSHNPHEVMDAARQVAEVDATPSGEVTCTVRTNTYKLAGEASNRKSLQVSWPRLAVPTGKLVLDGDDTLADVVLNCHETVVPVVVDCGPLRWSGRVDVAHDKFGDPNEPDTIECELIHDKVWLTRVVAFPWWFMPLQWQGPPTRGVAFGNAISVIKYLFTSQFMRIQLGLWELVNNLLSLNLDWRSYFSTLLMQNPGEELELRDIVQMATTPVYVVPSAGWNDTSPFISLNWRMDELLQLVTKTCEDNGLTIEVYLWEPGMPQPDPFAEATNLLRVPTVVVDVKDRMQVTGITGTAFDGLQRTFVDLLGSMFGEALKPFLDPNNEAAYAPDGVNIAPALGVHAIKPWCVFNADHPRSGVRGVVSHHHPIAWRTITGGKSPAWLNSLVDATLAWLIDMITIVLGVTGVPGTILDGAFHDIAFAFQQTDNFDRRLKLGPYGLPEVFIPTGSGSYTLEAFFQQKSAQYDTRGYVSGQLVVDNCFPYELGRDTFPGALATFIRRGRVVTDFIENATLIETRGEPTQVQFQIGDGKAEEAPAAKLQRRFGDLQAGVNIALMAS

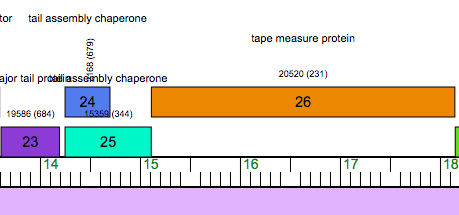
>Phegasus\_Draft gp20

MAIIVDEDKGTISFTECTVTFPYGFSVSSGVGTIVITPAGGVASFPLAIQGASGLPPNITMAFHVIGPDDPLPDPNPEMTVIDEGGPGEAAHYHYDCYVQKGDKGDAASFNFLDADDLEDGEDLAEGDVGTNGYVLSYAYEGTGTPGIRFIPQKTGDIRGPSAIAATAWSNTAIRLLCAVTLEAKPFPRKVLPTGSVVVTGSADTRVDLVAYLGDPDDGGVEIGRAFGQAGAAPPPLVMAGGPPATTAGGNANYAIVPAGQSATVYFRAEQKASSSNNWATAGAPDGARAGVVVVAV

1. Blast each of these genes and record your results (Does the gene align well with other phages, are the start calls the same?).
2. Is there a frameshift in this gene sequence?
3. If yes - which two genes are involved? How do you know? If no, what made you reach this conclusion?
4. **Phamerator – Do other similar phages have a PTFS?**

To answer this question, we’ll use the program phamerator. Phamerator can be found here: <https://phamerator.org/phages>. Instructions about phamerator and links for how to use phamerator can be found here: <https://seaphagesbioinformatics.helpdocsonline.com/article-15>.

Phamerator allows us to see bacteriophage genomes next to one another so we can easily compare them. While draft genome annotations do not include gene functions or final start sites, the database on phamerator allows you to view genomes that have been manually annotated and finalized. These genomes have assigned functions and the final decisions on gene start sites. In finalized genomes a PTFS is shown as two genes, one on top of another, both starting at the same start site and one extending beyond the other. A visual of this is below, with genes 24 and 25 containing the PTFS:



**Activity 3:** For each phage cluster, assess whether a PTFS is likely, and where in the genome it may occur, based on comparisons across phage genomes in that cluster.

1. Choose 5 genomes from the A1 cluster, including 1 draft genome.
   1. Which genomes did you choose?
   2. Examine the genomes – does it look like a PTFS occurs in these phage genomes? How do you know?
   3. If a PTFS occurs – where does it occur?
   4. If a PTFS occurs - On your draft genome, which two genes are involved in the frameshift but have not yet been annotated correctly?
2. Choose 5 genomes from the EE cluster, including 1 draft genome.
   1. Which genomes did you choose?
   2. Examine the genomes – does it look like a PTFS occurs in these phage genomes? How do you know?
   3. If a PTFS occurs – where does it occur?
   4. If a PTFS occurs - On your draft genome, which two genes are involved in the frameshift but have not yet been annotated correctly?