



**PECAAN**

***Phage Evidence Collection And Annotation Network***

Claire A. Rinehart, Bobby L. Gaffney, Jason R. Smith and James Dexter Wood  
Western Kentucky University Bioinformatics and Information Science Center, *BISC*.

## **User Guide**

### **Index**

1. Overview of PECAAN
2. User types and permissions.
3. Files needed to input a genome into PECAAN.
4. Setting up a new genome
5. Setting up new student accounts
6. Overview of buttons, checkboxes and functions
7. Annotation workflow
  - a. Choosing a start site
  - b. Changing a start site
  - c. Choosing support for a function
  - d. Entering a function
  - e. Entering comments into the Notes section
  - f. Saving Changes
  - g. The Change Log
  - h. Calling tRNAs
  - i. Frame Shifts
  - j. Adding a gene
  - k. Deleting a gene
8. Exporting Reports
  - a. Export CDS Function
  - b. Export CDS Full Annotation
  - c. Export CDS Notes
  - d. Export Author.csv file
  - e. Export Change Log
9. Acknowledgements

## Content

### 1. Overview of PECAAN

PECAAN is an annotation workflow that functions between DNA Master and the final Phagesdb annotation submittal portal for SEA PHAGES (Figure 1).

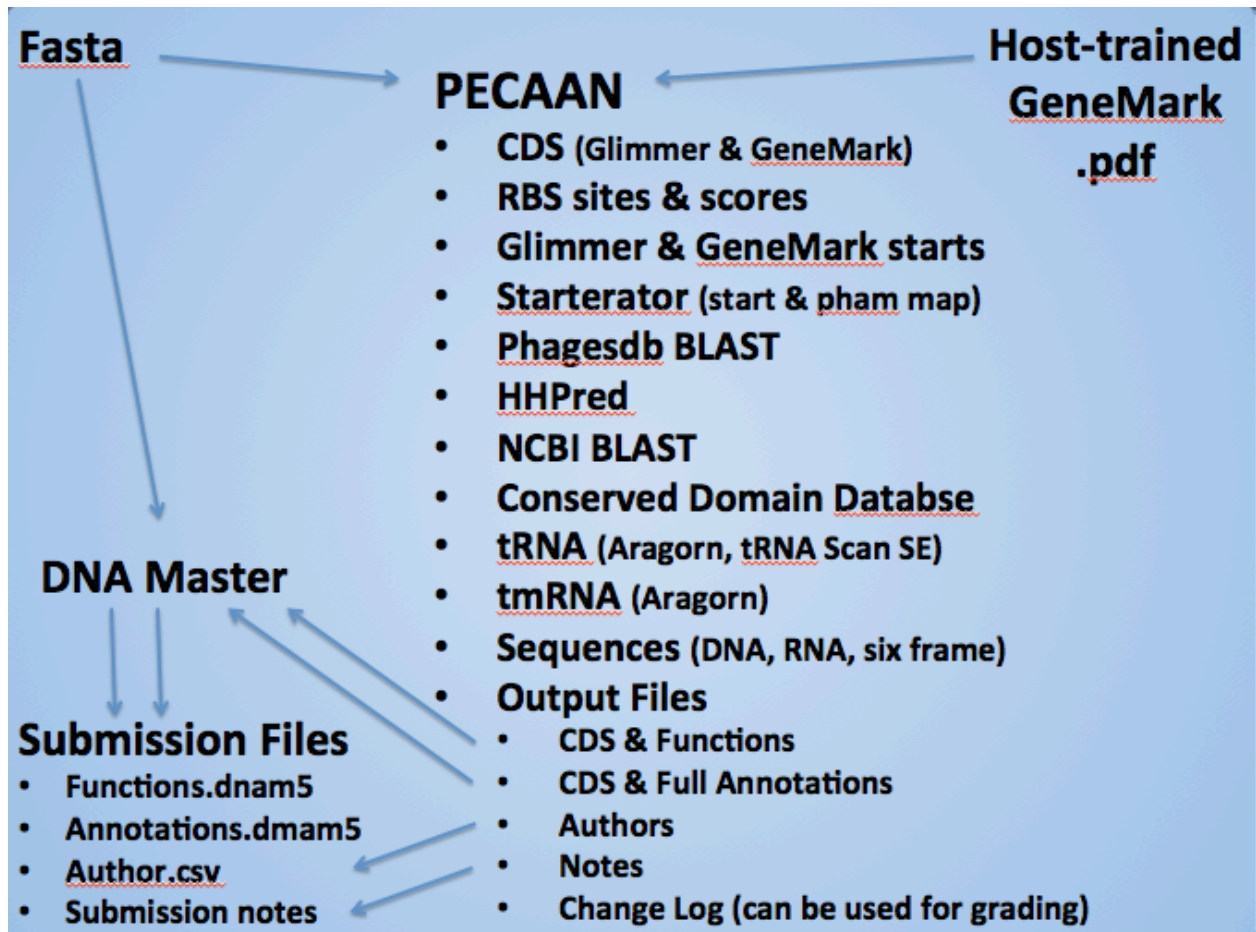


Figure 1. PECAAN and the SEA PHAGES workflow.

PECAAN stores all evidence in a database so that annotation choices and evidence can be viewed and quality checked by others. All changes are recorded in a viewable change log.

PECAAN displays gene start information from Glimmer, GeneMark, Starterator, and host-trained GeneMark. The Shine-Dalgarno ribosomal binding sites predicted by Karlin6 with spacing from Keibler Medium are selectable as checkboxes.

PECAAN also generates a protein for the start-defined gene and searches the Phagesdb, NCBI, HHPred and Conserved Domain databases for matches to the protein. The top 100 hits from each database are available for display and the

matches can be selected as functional support by checking the box next to evidence.

The function is entered by typing into the function name into the Function box. If standard function names are available they will be displayed in a drop-down window under the Function field as you type and you can select the standard function or continue typing if a function not listed. If a function is unknown, then type NKF (No Known Function) into the Function field.

Notes explaining choices for the gene can also be recorded in the NOTES field. All changes can be saved by clicking the SAVE button. Saved gene changes will be recorded in the Change Log along with the name of the annotator and a time/date stamp.

The tRNA and tmRNA evidence from Aragorn and tRNA Scan SE is also available for review and modification. The tRNA 5' and 3' end locations can be modified to bring the annotation into alignment with best evidence. Changes will be recorded following a click on the "Change tRNA Data" button. If following your review requires no changes you can enter your name in the review log by clicking on the "Checked with no change" button.

Function files and annotation files can be exported and integrated into DNA Master under the Export menu. Export of just the functions or the whole annotation evidence into DNA Master is helpful in the generation of submission files to the Phagesdb/annotations portal. The author and notes files can also be exported from PECAAN and submitted to the Phagesdb/annotation portal.

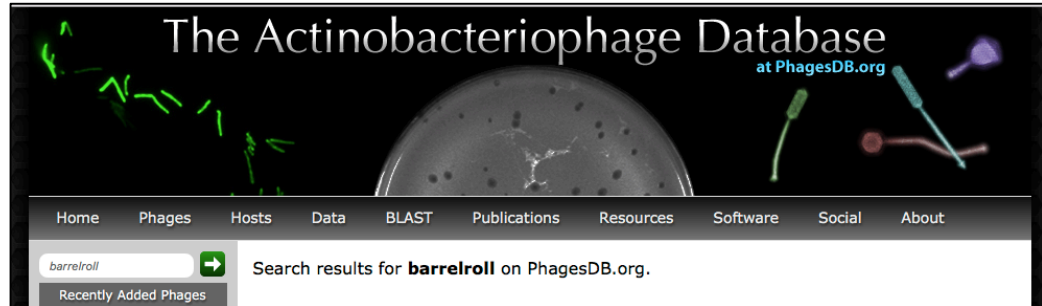
## **2. User types and permissions.**

When a new university site is established, a faculty member will be set up as an administrator (Admin) for their site. They will be able to set up User and other Admin level log-ins for their site. Admin level users will also be able to load new genomes into PECAAN for annotation. There are three types of PECAAN users: User and Admin level users, which are limited in scope to their own site, and SuperAdmin users. The SuperAdmin team will be able to edit all sites and give quality control feedback on all genomes.

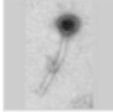
## **3. Files needed to input a genome into PECAAN.**

### **Phage.Fasta file**

A fasta formatted file containing the gene sequence for your phage should be available from [Phagesdb.org](http://Phagesdb.org). Put the phage name into the upper left search box (for example barrelroll) and click on the green arrow:



Select the phage link:



[The Actinobacteriophage Database | Phage BarrelRoll](#)  
 Discovery Notes, **BarrelRoll** was found underneath a layer of pine needles and pine cones. The site was 167 from the case of a pine tree. It was 6-8 cm below in ...  
[phagesdb.org/phages/BarrelRoll/](http://phagesdb.org/phages/BarrelRoll/)

And then scroll down to and click on the fasta download link:

Sequencing Information	
Sequencing Complete?	<b>Yes</b>
Date Sequencing Completed	<b>Jan 30, 2011</b>
Genome length (bp)	<b>59672</b>
Character of genome ends	<b>3' Sticky Overhang</b>
Overhang Length	<b>11 bp</b>
Overhang Sequence	<b>CTCGTAGGCAT</b>
GC Content	<b>66.6%</b>
Fasta file available?	<b>Yes: <a href="#">Download fasta file</a></b>
Characterization	

This will download a phagename.fasta file to your computer.

### Host-trained GeneMark file

Go to the [GeneMark.hmm prokaryotic](#) website to generate a map of your phages coding capacity.

Input sequence and Select species

Enter sequence (FASTA or multi FASTA format)

or, upload file:  Baehexic.fasta **1**

**2** Select species: Mycobacterium\_tuberculosis\_H37Rv

Action:   **4**

Options

Output format for gene prediction	Output options	Optional: results by E-mail
<input checked="" type="radio"/> LST <input type="radio"/> GFF	<input type="checkbox"/> Protein sequence <input type="checkbox"/> Gene nucleotide sequence Coding potential graph (not for multi FASTA) <input checked="" type="checkbox"/> PDF <b>3</b> <input type="checkbox"/> PostScript	E-mail Subject GeneMark.hmm prokaryoti <input type="checkbox"/> Compress files

Advanced options

Switch off gene start related motif(s)

Click on the Choose File button (1 above) and locate your phagename.fasta file.

Select a host species (2 above) to train the prediction model. For mycobacteria this will usually be the tuberculosis or smegmatis hosts.

Select the PDF output option (3 above) and then click on the “Start GeneMark hmm” button (4 above). This will generate the window below.

Gene Prediction Results

output	link
Coordinates of predicted genes	<a href="#">gmhmp.out</a>
PDF	<a href="#">gmhmp.out.pdf</a> <b>1</b>

Click on the PDF link (1 above) and then save the file as Phagename\_GM\_host.pdf so that you will recognize it.

#### 4. Setting up a new genome

Open PECAAN and log in.

You will need Admin privileges. If you have Admin privileges you will be able to see and open the Admin menu (1 below).

PECAAN - Phage Evidence Collection And Annotation Network

Summary Genes Sequence Export Admin **1**

Users

**2** Phages

Click on the Phages line (2 above).

On the Phages form (shown above), do the following:

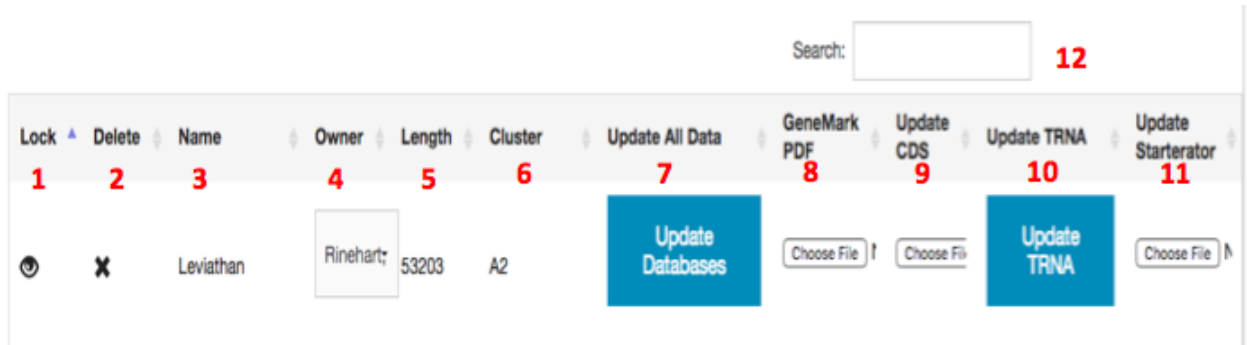
- Enter the Phage Name (1 above).
- Enter the Phage Cluster (2 above).
- Choose the fasta formatted sequence file for the phage (3 above).
- Select the host-trained Genemark PDF file that you saved as:  
Phagename\_GM\_host.pdf (4 above).
- Click the Add Phage button to generate the new phage (5 above).

When you click the Add Phage button PECAAN will set up the phage genome for editing by the all Users and Admins from your institution. Users from other institutions will not be able to edit your genome. If you enter a phage for someone from another institution you will need to refresh the page and scroll down to the bottom the phage list and change the Owner (4 below) to someone from the annotating institution.

It takes time for PECAAN to generate the NCBI and HHPred matches to the gene products and the time is dependent on the number of phages in the que. Give it 30 - 90 minutes. If you view a gene and find that NCBI or HHPred searches have not completed, return later or refresh the page every few minutes until the data appears.

If you scroll down the Phages page you will see a list of phages (3 below), their length (5 below) and their cluster (6 below). If the eye symbol at the left of the phage line (1 below) does not have a / through it then the phage can be edited by everyone in your university work group. If you click on this symbol it will put a / through the symbol indicating that the phage is locked for those with User privileges. Admin users can continue to edit and modify the genome. This allows Admins to lock phages and prevent student users from changing files while they are doing the Quality Check in preparation for submission of the finished genome to the Phagesdb/annotation portal. SuperAdmin users will also be able to edit the

annotations and communicate through the notes to Admins during the final Quality Checks.



Within each phage line, PECAAN can be prompted to “Update All Data” for that phage (7 above). **Beware!** This will reset all of the functional evidence check marks for all genes as it searches for the latest data. Updates can be performed separately for each gene and database in annotation page under the Genes menu and this is usually a better way to get an update for a few genes.

The “GeneMark PDF” button (8 above) allows alternative host-trained GeneMark files to be added after the phage has been set up.

The “Update Existing CDS Data” button (9 above) will regenerate the gene calls and update the Glimmer and GeneMark headers this is **not recommended** and is available as we continue to change PECAAN and need to update the database.

The Update tRNA button (10 above) can be used on older phages that were entered into PECAAN before the tRNA and tmRNA searches were implemented. New phage entries should not use this because it is done automatically when the phage is loaded.

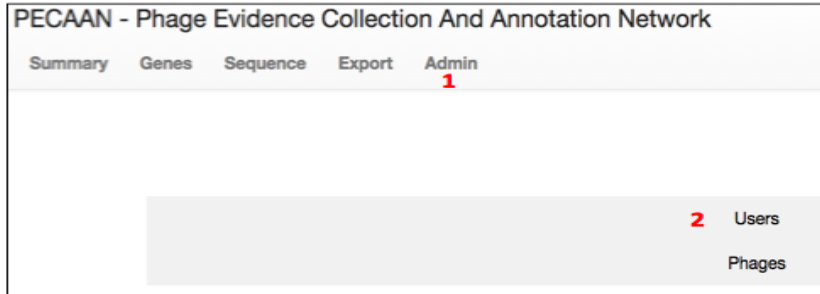
The Update Starterator (11 above) is another old link that **should not be used** and is temporarily available for maintenance purposes. We now have direct links to the Starterator outputs for the Pham associated with each gene in the annotation window and updates will be made automatically.

Since there are a large number of phages, use the Search box (12 above) to find your phage name quickly. Just start typing and the list of displayed phages will automatically be reduced to reflect the search parameters. This also works if you want to limit display of phages by cluster. It doesn't seem to work to search for owners.

## 5. Setting up new student accounts

Open PECAAN and log in.

You will need Admin privileges. If you have Admin privileges you will be able to see and open the Admin menu (1 below).



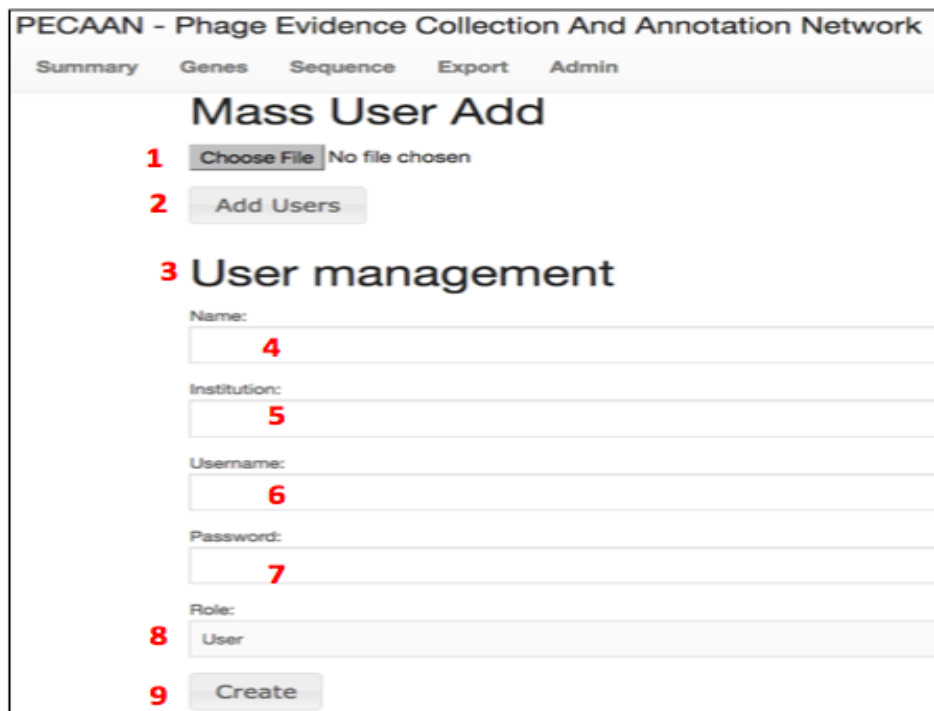
Click on the Users line (2 above).

Users can be added individually (3-9 below) or en Mass (1 & 2 below).

To add an individual user (3 below):

In the Name field (4 below), enter their complete name: LastName, Firstname MI (no period). This will allow a properly formatted Author Report to be generated for each phage with all the users that annotated or checked the phage.

In the Institution field (5 below), make sure that you use same institution name as the owner of the phage that user will be annotating. You can see the owner in the Admin/phage menu and look up the institution name for the owner in the list of users below the "User management" (3 below) section of this page. We need consistency because the scope of your working group and phage editing access will be determined by this institution name of the owner and the users.



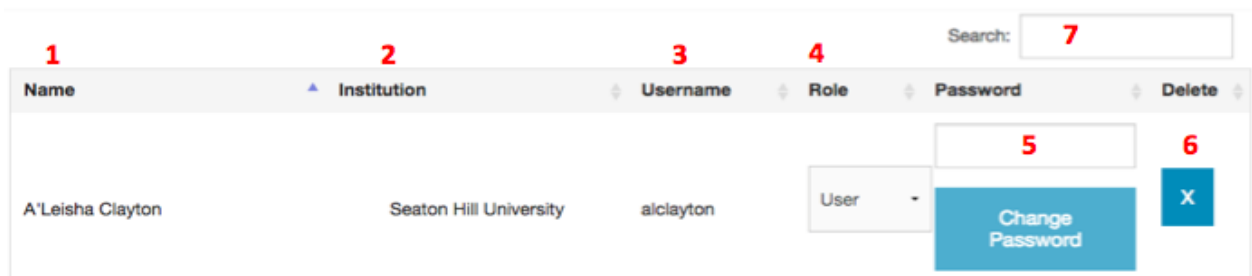


For the Username (6 above) use the first letter of the firstname and middle initial, followed by the lastname. For example if the Username was: Rinehart, Claire A, then the username would be carinehart.

\*\*\*Passwords (7 above) must be sufficiently long and complex that the strength indicator reaches 4/4. Passwords should include lower and uppercase letters as well as at least one number and at least one punctuation (!@#%&). Try to make the password something that students will remember but that is not obvious. Do not include weak words like phage, dna or other biologically relevant words. You may consider making a few different passwords for your class and then randomly distribute them. Consider using the Mass User Add button (1 & 2 above, see below for instructions) where strong passwords will be generated for each user.

For students, the Role (8 above) should be set to User. If other faculty need to load pages, add students or do Quality Checking / Editing then you should give them Admin privileges.

After entering the information as described above, press the Create button (9 above) to add a new user. If you refresh your page and scroll down you should see the new user (1 & 3 below). If at any time they need a new password, you can change it here (5 below). Clicking the X at the right side of the line will delete the user (6 below).



There is also a drop down to change the Role of the user (4 above) to User or Admin. The Search box (7 above) lets you easily filter the users by name or institution.

PECAAN - Phage Evidence Collection And Annotation Network

Summary Genes Sequence Export Admin

## Mass User Add

1 Choose File No file chosen

2 Add Users

For the “Mass User Add” (shown above) you will need to open a spreadsheet and enter your class data into three columns without headers: Lastname, Firstname, Middle Initial (no period) in the first column, the user name which consists of the initials of the first and middle names followed by the lastname in the second column, and finally the correct Institution name as described above in the third column. Click the ? next to the “Mass User Add” line to see an example (shown below).

	A	B	C	D
1	Cook, Becky, J	bjcook	Morehead State University	
2	Carrington, Jordan, A	jacarrington	Morehead State University	
3	Cyrus, Cassidy, R	krcyrus	Morehead State University	
4	Wright, Allison, V	avwright	Morehead State University	
5	Rose, Emma, N	enrose	Morehead State University	
6	Frommeyer, Joshua	jfrommeyer	Morehead State University	
7				

Save the file as a xxx.csv (comma separated file). Click the Choose File button (under the “Mass User Add” (1 above) and enter the xxx.csv file that you saved above. When you press the “Add Users” button (2 above), a new filename will show up in the bottom left panel of your screen that will contain the passwords for the users that you just entered. You may need to move the mouse to the bottom left of the window to see it. Double-click this file to open it. You can use this output file to distribute passwords to your new users.

NOTE: If you use a text editor to create the xxx.csv file then you will need to put quotes around all of the text in the first column as illustrated below:

```
"Cook, Becky, J",bjcook,Morehead State University
"Carrington, Jordan, A",jacarrington,Morehead State University
"Cyrus, Cassidy, R",krcyrus,Morehead State University
"Wright, Allison, V",avwright,Morehead State University
"Rose, Emma, N",enrose,Morehead State University
"Frommeyer, Joshua",jfrommeyer,Morehead State University
```

## 6. Overview of buttons, checkboxes and functions

### a. The Summary page

PECAAN - Phage Evidence Collection And Annotation Network

Summary Genes tRNA tmRNA Sequence Export Admin - Welcome, caradmin. 5 Logout

Phage: Leviathan Actinobacteriophage Database

Length: 53203 Number of Genes: 96 %GC: 63.3

Map (Kbp)

Codon Usage

GGG GGC GGU GGA

1. To select a genome, click on the Phage dropdown box and start typing the name and a list box below the entry field will display all of the matching phage names.
2. Clicking on the Actinobacteriophage Database button will open a tab in the browser that is connected to the phagesdb.org.
3. Scrolling the mouse button when the cursor is over the Map will compress/expand the map. Drag right and left on the map to move the map within the viewing window. Blue bands are forward genes and red bands are reverse genes. Bright blue and red regions show areas of overlap between two genes. Black areas are intergenic.
4. The Codon Usage map shows the relative frequency of each codon.
5. When you are finished, press the Logout button to exit PECAAN.

### b. The Genes Page

#### i. The Header

PECAAN - Phage Evidence Collection And Annotation Network

Summary Genes Sequence Export Admin

Add Gene Save Host-Trained GeneMark 1 Start: 342 Forward

Phage: Leviathan Cluster: A2 User: caradmin Institution: Western Kentucky University

1. Click to add a gene
2. Click to save changes
3. Click to open the host-trained GeneMark pdf file associated with this phage.

4. Click to select the genes. The current start site, stop site and direction of the genes are displayed because the gene numbers change as new genes are inserted.
- 5., 6., 7., & 8., display the Phage, Cluster, User and Institution information respectively.

## ii. Gene Candidates

Glimmer Start: Glimmer Score: GeneMark Start: Starterator Start: Pham  
**1** 654 2.88 654 654 **2** 1131

### Gene Candidates

Gene Included:  **3**

Show  entries **4b** Search:

<b>5</b> Direction ▲	Start	Stop	Length	Gap	Spacer	Z-score	LORF	Start Codon	All GM Coding Capacity	Selected Gene
Forward	654	1415	762	195	14	2.7019	TRUE	GTG	<input type="button" value="Select"/> <b>7</b>	<input checked="" type="checkbox"/> <b>8</b>
Forward	723	1415	693	264	16	1.0106		ATG		<input type="checkbox"/>
Forward	762	1415	654	303	13	2.0626		TTG		<input type="checkbox"/>

Showing 1 to 10 of 12 entries **4a**

**4c** Previous   Next

1. Glimmer and GeneMark start calls. Also Starterator start call if available.
2. If you click on the Pham number the Starterator output for the pham will be displayed in a separate window. This allows you to look at all of the starts in all of the phage genes in the pham.
3. If the Gene Included is unchecked the gene will not be included in the export files and will not be used in the Gap between gene calculation.
4. A. shows how many items are displayed out of the total number of entries.  
 B. Drop down allow you to display 10, 25, 50 or 100 entries at once.  
 C. Allow you to move between pages.
5. Each of the headings have active sort functions associated with them. Clicking on a header will sort the table based on the information in that column. An upward triangle indicates that the table is sorted in ascending order with the smallest value at the top of the list. A downward triangle indicates that the table is sorted in descending order with the largest value at the top of the list.
6. The Search box allows you to search all of the text in the list for corresponding values. Only the lines that contain matches will be displayed.

7. After viewing the Host-trained GeneMark, this dropdown box allows you to select either Yes or No to indicate if the start/stop limits contain all of the coding capacity for the gene.
8. The checkbox indicates the current start site selection for the gene. You can select alternative start sites by clicking on a box next to another start site. Changing the start site will not automatically initiate a BLAST search in Phagesdb.org, HHPred, the Conserved Domain Database and on the NCBI site. There are Rerun buttons next to each of these headers that will allow updating the entries from these databases. It may take 2-8 minutes for these searches to complete. You may refresh the page to see if these databases have completed their search.

### iii. **Function and Notes**

The image shows a web form with two main sections. The first section is labeled 'Function:' and contains a text input field with the text 'Terminase Small Subunit' and a red number '1' to its right. The second section is labeled 'Notes:' and contains an empty text area with a red number '2' to its right.

1. The Function field (1 above) is used to enter the function of the protein. Just start typing the function and a list of matching standard function values will be displayed below the entry box. Select a value from the displayed values or continue typing to enter a new function not found in the dropdown list.
2. The Notes field (2 above) is used to enter notes or justifications that you feel are important to be communicated to those that review your annotations for the gene being displayed. When the Save button (2 in the Header figure) at the top of the Genes page is pressed, these notes are saved in the change log at the bottom of the genes page. During export, which is initiated from the Export menu button, the currently displayed Function and Notes are included in the Export files. This is useful in the final submission of an annotated genome to the phagesdb.org/annotation/ portal.

### iv. **Phagesdb BLAST**

Phagesdb BLAST provides matches to proteins that may have functions associated with them which may provide evidence supporting a function choice.

1. Click on the box next to the BLAST match that you want to serve as functional evidence.

Phagesdb BLAST Rerun **6**

Last Updated:  
2/29/2016, 4:12:54 PM

**5** Show 10 entries **3** **4** Search: **2**

Evidence	Name	Protein Number	Function	Sequence Length	Score	e-value
<input type="checkbox"/>	D29	14	predicted 53.5 kD protein	485	893	0
<input checked="" type="checkbox"/> <b>1</b>	Trixie	16	Portal.	486	923	0
<input type="checkbox"/>	NaSiaTailie	13	portal protein	486	981	0

- If you want to see more functions of this type, type the function into the search box.
- You can sort the list by function by clicking on the Function header. When there is a “function unknown”, at the top of the list, clicking on the header will sort and then show you the bottom of the list. If the function is still “function unknown” then click again on the Function header to look at the top of the list. If both the top and bottom are “function unknown” then you can easily conclude that there are no other functions. This is also a great way to group all functions together.
- Clicking on the Score header sorts the list by the score. This is a great way to bring high scores to the top after you have sorted by some other criteria, like function.
- Clicking once or twice on the Evidence header will bring the checked evidence to the top of the list.
- If you have changed the start site in the Gene Candidates section described above, make sure that you Rerun the analysis to get the latest data **before** you select your evidence.

#### v. HHPred

- The functional evidence from HHPred can also be check marked (1).
- The probability column (2a) in combination with the E-value (2b) provide confidence in the matching evidence. A probability above 98% will often be related to a function. The lower the E-value the better. Shorter proteins can have significant E-values as high as 0.1. For longer proteins, expect significant matches to have much lower values.

HHPRED

Last Updated:  
2/29/2016, 4:14:35 PM

Show 10 entries **2a** **3** **4a** **4b** Search:

Evidence	Hit	Description	Probability	% Coverage	Target From	Target To	Query From	Query To	E-value	<b>2b</b>
<input type="checkbox"/>	<a href="#">TIGR01538</a>	portal_SPP1 phage po	100	81.6406	6	412	33	451	0	
<input type="checkbox"/>	<a href="#">2jes_A</a>	Portal protein; DNA tra	100	90.625	32	500	33	497	0	
<input checked="" type="checkbox"/>	<b>1</b> <a href="#">pfam05133</a>	Phage_prot_Gp6 Phage	100	83.9844	2	425	32	462	0	


**5**

- HHPred often matches proteins with conserved domain, therefore the % Coverage can help in deciding whether to call the match as a domain or a function.
- The Target From and Target To (4a) shows the range of the protein alignment in the matching protein from the database. The Query From and Query To (4b) shows the range of the protein alignment in your query protein.
- Clicking on a specific Hit link will take you to the NCBI database hits. You can then choose Conserved Domains, Protein or Protein Clusters.

Proteins		
<b>Conserved Domains</b>	1	conserved protein domains
<b>Protein</b>	12	protein sequences
<b>Protein Clusters</b>	2	sequence similarity-based protein clusters

Clicking on the Conserved Domains gives you additional information about the domain.

pfam05133: **Phage\_prot\_Gp6** ?



**Phage portal protein, SPP1 Gp6-like**  
This protein forms a hole, or portal, that enables DNA passage during packaging and ejection. It also forms the junction between the phage head (capsid) and the tail proteins. During SPP1 morphogenesis, Gp6 participates in the procapsid assembly reaction. This family also includes the old Pfam family Phage\_min\_cap (PF05126).

#### vi. NCBI BLAST

- The NCBI BLAST list also has a checkbox that can be marked and saved as functional support.
- The % Identity (2a) and the number of Positives (2b) measures the similarity of the query protein to the database matches.
- The lower the E-value the better. Shorter proteins can have significant E-values as high as 0.1. For longer proteins, expect significant matches to have much lower values.
- The Target From and Target To (4a) shows the range of the protein alignment in the matching protein from the database. The Query From and Query To (4b) shows the range of the protein

alignment in your query protein. Comparing the Target From and the Query From values to see if they are 1:1 is further evidence that the correct start site was selected.

Evidence	Accession	Region	Creation Date	CDS Note	Description	% Identity	% Aligned	% Coverage	Positives	Target From	Target To	Query From	Query To	Gaps	E-value
1	YP_009198227	Yes	2016-01-05	Portal	Mycobacterium phage Pro94	100	100	99.7942	486	1	486	1	486	0	0
	YP_009205388	Yes	2016-01-05	Portal	Mycobacterium phage Larenn	95.6879	97.9486	99.7942	477	1	486	1	486	0	0
	YP_009017147	Yes	2014-03-11		Mycobacterium phage Trixie	93.2099	95.8848	99.7942	466	1	486	1	486	0	0

- The Gaps column shows the number of gaps introduced in the BLAST alignment.
- The % Aligned and % Coverage give an indication of how much of the protein is aligned to the target protein.
- Clicking on the specific Accessions will take you to the NCBI site for that protein and usually inserts the link as a new tab in your browser so that it is easy to navigate between PECAAN and the Accession page.

GenPept ▾

**portal protein [Mycobacterium phage Kratio]**

NCBI Reference Sequence: YP\_009212752.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

---

Go to: ☑

LOCUS YP\_009212752 512 aa linear PHG 05-JAN-2016

DEFINITION portal protein [Mycobacterium phage Kratio].

ACCESSION YP\_009212752

- If there is a Region or Site section in the Accession's annotation there will be a Yes in the Region column. This is useful information that links to HHPred evidence and to the Conserved Domain Database evidence to the NCBI results. Clicking on the Yes in this column will bring up a separate window that contains all of the Region and Site annotations for the accession.

https://discover.kbrinsgd.org/evidence/genes/region/7566

2016-01-05

Region 32..466

/region\_name="Phage\_prot\_Gp6"

/note="Phage portal protein, SPP1 Gp6-like; pfam05133"

/db\_xref="CDD:253031"

The pfam, TIGR, COG, and PK ... references can be useful in



selecting significant evidence to check mark in the HHPred and the Conserved Domain Database lists. For example in the figure above pfam05133 is associated with region and is also found as a significant match in HHPred and in the Conserved Domain Database lists.

9. The CDS Note, like the region description, is also taken from the Features of the Accession. These notes often contain function descriptions and their display here saves opening the accession and searching for these values.
10. The creation date is useful in identifying newly annotated vs old phage annotations.
11. The Description usually contains the Phage name. If you want the gene number, go to the Phagesdb display to find the associated gene number.

### vii. Conserved Domain Database

Conserved Domain Database

Last Updated:  
2/29/2016, 4:06:07 PM

Show  entries

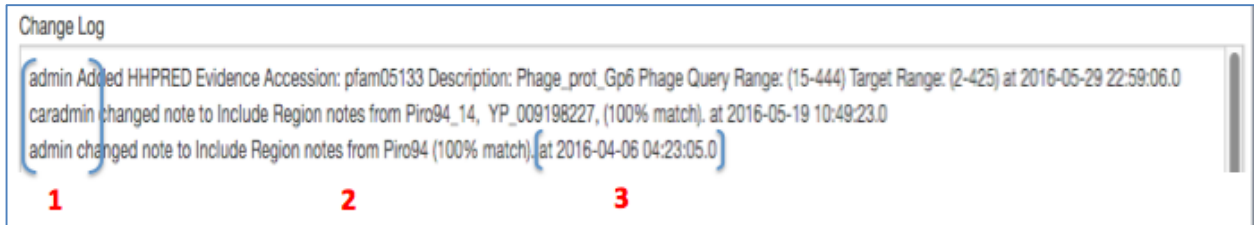
Search:

1 Evidence	Accession	Description	% Identity	% Aligned	% Coverage	Positives	Target From	Target To	Query From	Query To	E-Value
3	pfam05133	Phage portal protein, SPP1 Gp6-like. This protein forms a hole, or portal, that enables DNA passage during packaging and ejection. It also forms the junction between the phage head (capsid) and the tail proteins. During SPP1 morphogenesis, Gp6 participates in the procapsid assembly reaction. This	27.1264	40.6897	88.2716	177	6	427	19	448	0

1. The column headers for the CDD have the same functions and meanings as those in the NCBI BLAST display.
2. Searches filter the text in all of the columns and accessions.
3. Evidence is marked by clicking on the evidence box, just the same as in the other analysis sections above.

viii. **The Change Log**

The Change log window displays all of the changes that have been made and saved to the individual gene.



1. The user name is recorded for each entry.
2. A brief description of the change is recorded.
3. A date and time stamp is appended to each change.

c. **The tRNA Page**

The screenshot shows the 'PECAAN - Phage Evidence Collection And Annotation Network' interface for the tRNA page. Red annotations are placed as follows: **1** is above the 'tRNA' menu item; **2** is above the 'tRNA Included:' label; **3** is above the 'Start: 30120' label; **4** is above the 'Stop: 30203' label; **5** is above the 'Structure:' label; **6** is above the 'Sequence:' label; **7** is above the 'Aragorn' analysis header; **8** is above the 'tRNA Scan SE' analysis header; **9** is above the 'Notes' section; **10** is above the 'Change tRNA Data' button; **11** is above the 'Checked with no change' button; and **12** is above the dropdown menu showing 'TRNA: 1 (30,120-30,203) Forward'.

1. The tRNA menu page displays the Aragorn (7) and tRNA Scan SE (8) results.

2. If the tRNA Include check box is marked the tRNA is used as output. To exclude the called tRNA, uncheck the box.
3. The Start and Stop positions of the tRNA gene is displayed and the position of the 5' start can be changed by dragging the red divider (3) right or left.
4. The Stop, or 3' end of the tRNA gene can be changed by dragging the red divider (4) right or left.
5. The ( ) symbols in Aragorn (7) and the < > symbols in tRNA Scan SE (8) represent the locations in the sequence, to which they align, where the bases are hydrogen bonded. tRNAs typically have a 7-9 bp stem where the 3' and 5' ends hybridize (5). The 3' end then contains an additional base and then a CCA sequence. The amino acid is attached to the 3' terminal A.
6. In defining the 3' end of the tRNA we typically include as much of the -XCCA sequence that is encoded by the gene. For this example only the -GC (6) is included.
7. Aragorn evidence.
8. tRNA Scan SE Cove Score evidence should be greater than 20. Therefore, this gene would not be included and the tRNA Include box should be unchecked.
9. Notes section where notes can be recorded. Below the notes section a Change Log box also records changes with a similar format to the Genes Change Log described above.
10. Press this button to record changes.
11. Press this button to record that you reviewed the tRNA gene and don't need to make any changes.
12. Drop down box used to view and select other tRNA genes.

d. **The tmRNA Page**

tmRNA combines the features of tRNA and mRNA. It recycles stalled ribosomes by adding a [proteolysis](#)-inducing tag to the unfinished [polypeptide](#), and then facilitates the degradation of the aberrant mRNA. (Wikipedia)

1. Clicking on the tmRNA menu displays the tmRNA page.
2. Individual tmRNA genes are selected from the drop down box.
3. If the tmRNA Included check box is checked, the tmRNA will be included in the function and full-annotation outputs. If unchecked, it will not be included in these reports.
4. tmRNA is predicted by the Aragorn program.
5. The start and stop positions of the tmRNA are defined and not adjustable like the tRNA ends.
6. The sequence tag of the tmRNA is displayed.
7. The change log displays changes in the tmRNA Included checkbox which is the only change allowed on this page.

e. **The Sequence Page**

- i. The Sequence page is accessed from the top menu (1). If you have not selected a phage it will send you to the Summary page. If you have not selected a gene, it will send you to the Genes page first. If you have selected a gene in the Genes page and click the Sequence menu it will display both the DNA Sequence (2) and the Protein Sequence (3). The Phage (4a), Gene Start (4b) and Orientation (4c) are also shown.

PECAAN - Phage Evidence Collection And Annotation Network

Summary Genes **Sequence** 1 Export Admin -

Phage: AlleyCat 4a

Start: 55 Orientation: Forward  
4b 4c

DNA Sequence 2

```
ATGGCGCAAGCCGACAGCGGGACAGTAACGGTGTGGTCGAGCCCGATTGTAATCACTTCAAGGAGTCGGTTCAGGCGCTCGTAACCTGGCTTCAGCCACCCCGTACCGCCGCGGCCCGATTCCGGAGCTT
CGACTGGAAGCGCTGCAGCTCGCGGTCCGCGCGCACAGTTCGGCGACACAGGACCGGGACGCAACGATCGTCGAGACGGCGCGGCTTCGCTGACTTCATCATCGACGGTGGCGCGTGA
```

Protein Sequence 3

```
MAQADSGTVTLVPEDLNHFKEVQRVWVGLQPPVPPRDPSELRLLEALQLAVRAHSSGDQDRDATIVETARAFADFIIDGGA
```

- ii. The bottom of the sequence page shows the six-frame map with the position of the left end of the visible sequence or the position of the cursor displayed in the upper left corner (1). Clicking on the Move To Sequence button (2) will scroll the six frames to the current gene shown at the top of the Sequence page. (Still fixing alignment of the frames.)

Move To Sequence 2

Position: -1 1

```
A G F S P R H F F P K W I G D M E H G A S R
G G F L S P A F F S K V D R R H G T W R K P
R R V S L P G I F F Q S G * E T W N M A Q A
AGGCGGGTTTCTCTCCCCGGCATTTTTTTCCAAAGTGGATAGGAGACATGGAACATGGCGCAAGCC
TCCGCCCAAAGAGAGGGGCGTAAAAAAGGTTTCACCTATCCTCTGTACCTTGTACCGCGTTCCG
P P N R E G A N K E L T S L L C P V H R
R T E R G P M K K W L P Y S V H F M A C
A P K E G R C K K G F H I P S M S C P A
```

f. The Export Page

- i. The Export page is used to export five different files.

PECAAN - Phage Evidence Collection And Annotation Network

Summary Genes tRNA tmRNA Sequence **Export** Admin -

1

2 3 4 5 6

Export CDS Function Export CDS Full Annotation Export Gene Changelog Export CDS Notes tRNA Changelog Export

tmRNA Changelog Export Author Export

7 8

- 1. Clicking on the Export menu displays the Export page.
- 2. The Export CDS Function creates a file from the PECAAN Function fields that has all of the function information needed to create the DNA Master file that only displays the annotated functions in the Notes field.

```

CDS 55 - 303
  /gene="1"
  /product="gp1"
  /locus tag="AlleyCat_1"

CDS 300 - 545
  /gene="2"
  /product="gp2"
  /locus tag="AlleyCat_2"

CDS 542 - 772
  /gene="3"
  /product="gp3"
  /locus tag="AlleyCat_3"

CDS 997 - 1356
  /gene="4"
  /product="gp4"
  /locus tag="AlleyCat_4"
  /note=Terminase Small Subunit

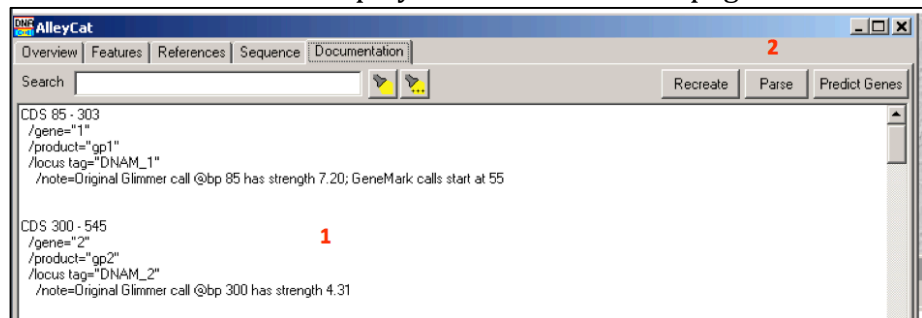
CDS 1349 - 2776
  /gene="5"
  /product="gp5"
  /locus tag="AlleyCat_5"
  /note=Terminase

```

Select all of the text in this file and copy it to your clipboard.

Open the DNA Master file for the phage corresponding to the PECAAN phage used to make the export.

Go to DNA Master and display the Documentation page.



Right-click on the body of the text (1) and click on the Select All option in the pop-up box. Right-click on the body of the selected text (1) and then click on the Paste option in the pop-up box. The selected text should have been replaced with the text from the Phagename\_CDSfunction.txt file. Verify the replacement. Next click on the Parse button (2) and approve the over-writing of the database. You should now be able to select the Features menu option and see the functions displayed in the notes field of the DNA Master file. Save the DNA Master file.

3. The Export CDS Full Annotation button creates a file from the PECAAN fields and checkboxes that contains a fully formatted and annotated genome. This file can be used to create the DNA Master file that contains the full annotations in the Notes field. The process of transferring the full annotations to DNA master is the same as in 1 above: Click on the Export CDS Full Annotation button in PECAAN and open the file. Select all of the text and copy it to

your clipboard. Open the corresponding DNA Master genome file and navigate to the Documentation menu page. Right-click on the body of the text and click on the Select All option. Right-click on the body of the selected text and click on the Paste option. Click on the Parse button to rewrite the new data to the DNA Master database. Navigate to the Features menu page and verify that the full annotations are displayed in the Notes field.

4. The Export Gene Changelog file displays all of the changes made to each gene along with the name and time/date stamp for when the change was made. This log keeps track of who touched the genome and what changes were made. This can be useful in reviewing student participation and the quality of their annotations and checking.
5. The Export CDS Notes file captures all of the currently visible text from the PECAAN Notes field. This is useful because it allows one to quickly scan all of the Notes in a genome. If students put questions into their notes section when they have questions or problems, it allows an instructor to quickly scan all of the genes and focus on specific problem genes. When an instructor is doing a quality check before submission of the annotated genome, it is a nice place to put notes that need to come to the attention of the SMART team. These notes can easily be copied from the exported file and inserted into the bottom field of the submission form.
6. The tRNA Changelog Export file displays all of the changes made for each tRNA along with the name and time/date stamp for when the change was made. This log keeps track of who touched the tRNAs and the changes that were made. This can be useful in reviewing student participation and the quality of their annotations and checking.
7. The tmRNA Changelog Export file displays all of the changes made for each tmRNA along with the name and time/date stamp for when the change was made.
8. The Author Export file will output a phagename\_authors.csv file containing all of the names of students that have annotated or check the phage genome annotations (assuming that they have made changes and clicked the Save button in the Genes, tRNA or tmRNA pages). This page can be further edited and used to submit the authors during the Genome submission process. The exported author.csv file will be in the correct format for submission if student account names are set up correctly in the beginning as Lastname, Firstname MI.

**g. The Admin Page**

A detailed description of the Admin pages and their use is given in sections:

4. Setting up a New Genome
5. Setting up a New User Account

## 7. Annotation workflow

There are only three basic questions that need to be answered for each gene during annotation: 1) should we include the gene or add a gene? 2) what is the best start site?, and 3) what function, if any, can we assign to this gene? If there is an assignable function, then the task becomes choosing the best evidence that supports that function.

### a. Should the Gene Be Included?

Use the Host-Trained GeneMark plot to determine if there is good coding capacity for the gene of interest. If a gene has not been called then consider adding a gene. The GeneMark plot can also be used to determine if the gene of interest greatly overlaps another valid gene, if it does then it maybe it should not be included. Checking other evidence, such as Phamerator alignments and function alignments, can help in making decisions to add or eliminate genes. Check the Gene Included box (10 in figure below) to include the gene and uncheck to not include the gene in the Export reports.

### b. Choosing a start site.

Look for the start site with the greatest weight of supporting evidence. This does not just mean the number of supporting elements (1 below) but also their relative weight: (Starterator = 3, A -4 Gap = 4, Glimmer = 2, GeneMark = 1, All GeneMark Coding Capacity = 3, 1:1 Target:Query = 2).

The current start site is marked with a blue check box on the right of the Gene Candidates list (4 below). Alternate start sites can be selected by checking a box on another line and confirming your choice.

Phage: Baehexic Cluster: A2 User: caradmin Institution: wku

Glimmer Start: 1 331 Glimmer Score: 13.32 GeneMark Start: 2 331 Starterator Start: 3 331 Pham 5 [7828](#)

### Gene Candidates

Gene Included:  10

Show  entries 7 6 8 Search:

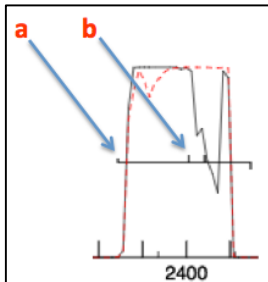
Direction	Start	Stop	Length	Gap	Spacer	Z-score	LORF	Start Codon	All GM Coding Capacity	Selected Gene
Forward	331	447	117	0	10	2.6287	TRUE	ATG	Yes <span style="color: red;">9</span>	<input checked="" type="checkbox"/> <span style="color: red;">4</span>
Forward	337	447	111	0	16	2.6287		ATG		<input type="checkbox"/>
Forward	385	447	63	0	11	0.7426		TTG		<input type="checkbox"/>

Showing 1 to 3 of 3 entries

Previous  Next



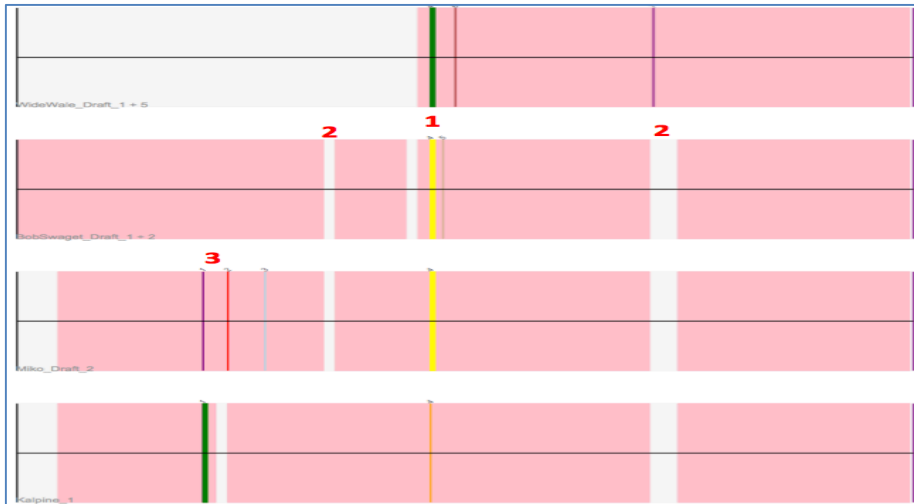
- i. The goal of start site selection is to encompass all of the coding capacity defined by the Host-Trained GeneMark. Selecting site “a” below would contain all of the coding capacity where selection of site “b” would not.



If the selected start site contains all of the coding capacity, use the drop down box under the All GM Coding Capacity (9 in top figure) to select the Yes option. If it does not contain all of the coding capacity, use the drop down box to mark the No option. If you change start sites the Yes or No choices previously made are erased and the Select prompt is displayed. You will need to re-mark this response.

- ii. Begin by comparing the start sites suggested by Glimmer, GeneMark and Starterator (1, 2, & 3 in top figure). If they all agree then there is a good chance that the start site that they suggest is correct. The relative weight for the three is Starterator > Glimmer > GeneMark. Look for suggest start sites that contain all the coding capacity defined by the Host-Trained GeneMark.
- iii. Good start sites usually have a Z-score (6 in top figure) around 2 or greater. The exception is when there is a -4 Gap, here the ribosome terminates translation and then can reinitiate translation without disassembling and reassembling. Z-scores as low as 1- 1.6 can be found at these types of valid start sites. You can use the Spacer information as a tie breaker, given that all else is equal, between two start sites. Extremely large spacers, => 17 tend to have lower translation initiation frequencies than those sites with spacers 8-14.
- iv. The Gap (7 above) indicates the number of bases from the 5' end of a gene to the gene that lies directly upstream. A 0 base Gap would end at one base and begin the next gene at the next base, for example --TGA ATG—where TGA ends the upstream gene and ATG begins the next gene. A one base overlap (Gap = -1) might look like this: TG A TG—where the A is used both in the stop (TGA) and the start (ATG) codons. A Gap of -4 might look like: -A TG A—where the TG is used by the stop codon TGA and is also used in the start codon, ATG for the next rightward gene. Gaps of 0, -1 or -4 give the ribosomes the opportunity to terminate translation of one gene and then slide either forward or back a few bases and then reinitiate translation of the next adjacent gene. The abundance of this type of overlap/ adjacency is probably not random, but has been selected as an efficient mechanism for expressing protein from a polycistronic transcript. Therefore, more weight is given to starts sites with a Gap of -4, -1, or 0.

- v. Two good ways to view the relative positions of the start sites called in other similar genes, is to look at the Starterator alignments for the Pham to which your gene belongs. Display the Starterator map for the pham associated with your gene by clicking on the blue Pham text (5 above). As illustrated in the alignment for pham 7828, below, all of the genes in in pham 7828 fall into four tracks. Each track has a unique pattern of start sites shown by the colored vertical lines (1 & 3 below). The blank regions at positions 2 represent gaps. All of the tracks have a start at position 1.



The Pham Report for 7828 is shown below (1 below). It shows the Phage\_genes associated with each of the four tracks (2 below). Notice that the position of the Most Called Start (3 below) is at start 4 (counted left to right) which corresponds to label 1 on the map above. There are lists for Genes that have the most called site but did not call it (4 below). There is also a list for genes that do not have the most called start site (5 below). Finally, there is a list of suggested starts for each gene (6 below). Notice that Baehexic\_1 has a suggested start at position 4 that corresponds to base 331 which is what is shown under the Starterator Start column in PECAAN. Note that the suggested start for Kalpine\_1 is start 4 but it was called at start 4. This may be a gene in which the start site needs to be revised.

**1 Pham 7828 Report**

**2** • Track 1 : WideWale\_Draft\_1, Equemioh13\_1, Updawg\_Draft\_1, Flare16\_Draft\_1, NaSiaTalie\_1, Baehexic\_1  
 • Track 2 : BobSwaget\_Draft\_1, Rachaly\_Draft\_1, Lokk\_Draft\_1  
 • Track 3 : Miko\_Draft\_2  
 • Track 4 : Kalpine\_1

**3** Most Called Start: 4 (number based on diagram)  
 Percent with start called: 90.9091%  
 Genes that call the most called start:  
 •WideWale\_Draft\_1, BobSwaget\_Draft\_1, Equemioh13\_1, Updawg\_Draft\_1, Miko\_Draft\_2, Flare16\_Draft\_1, NaSiaTalie\_1, Baehexic\_1, Lokk\_Draft\_1, Rachaly\_Draft\_1,

**4** Genes that have the most called start but do not call it:  
 •Kalpine\_1,

**5** Genes that do not have the most called start:  
 •

Other Starts Called:  
 •1 Kalpine\_1,  
 Percent with start called: 9.0909%  
 Suggested Starts:

WideWale\_Draft\_1, (4, 330)  
 BobSwaget\_Draft\_1, (4, 306)  
 Equemioh13\_1, (4, 330)  
 Updawg\_Draft\_1, (4, 330)  
 Miko\_Draft\_2, (4, 283)  
 Flare16\_Draft\_1, (4, 330)  
 Kalpine\_1, (4, 334)  
 NaSiaTalie\_1, (4, 330)  
 Rachaly\_Draft\_1, (4, 306)  
 Lokk\_Draft\_1, (4, 306)

**6** Baehexic\_1, (4, 331)

- vi. Another similar source of evidence can be seen by looking at the NCBI BLAST results displayed lower in the gene page. If you scan to the right of the BLAST list you will see the columns labeled Target From, Target To, Query From, and Query To (see figure below). By comparing the Target From to the Query From columns you can see if the start site in the Target gene that was aligned to your Query gene is the same. If you see 1:1 matches then the two start sites are in the same relative position. If you see a match of 1:32 or 64:1 in a BLAST match with a good e-value ( $< 10^{-20}$ ) then you may want to reconsider your start site or the Target start site may need to be re-evaluated. Again, the Starterator map and report probably gives a better analysis.

Target From	Target To	Query From	Query To
1	98	1	98
1	98	1	98
1	98	1	98
1	98	1	98
1	98	1	98
1	98	1	98

c. **Changing a start site**

As noted in point 7a above, start sites can be selected by clicking on the box at the right end of a line in the Gene Candidates list. When a new gene start site is selected the All GM Coding Capacity box is reset to Select. The function support checkboxes in Phagesdb BLAST, HHPred, NCBI BLAST and the Conserved Domain are not unchecked until new queries are sent to each of these databases by clicking on the Rerun button. When start sites are changed and the function support queries are Rerun, it may take up to five minutes for the HHPred and NCBI BLAST results to be returned. Refresh the page occasionally to see if these results have been returned or go to another gene and work on it while waiting. If you select the start site that had been previously selected, the information for the four databases is instantly available because these results not changed until they are Rerun, therefore, don't hesitate to explore alternative start sites if there is a reason to do so, because returning to the original start is quite simple.

d. **Choosing support for a function**

Before choosing a function, explore the matches to your protein query from the four databases, Phagesdb BLAST, NCBI BLAST, HHPred and the Conserved Domain Database, to determine if there is sufficient support for a function. The narratives for each database below will help you weigh the evidence supporting a function.

i. Phagesdb BLAST

The Phagesdb protein BLAST is found at <http://phagesdb.org/blast/>. This an excellent place to start the function search since it readily displays the annotated functions of matching phage genes. If you only see “function unknown” click on the Function column header, once or twice, to sort the matches by function (1).

Phagesdb BLAST

Last Updated:  
6/26/2016, 4:41:19 PM

Show  entries

Search:

<b>4</b>	<b>1</b>	<b>2</b>	<b>3</b>			
Evidence	Name	Protein Number	Function	Sequence Length	Score	e-value
<input type="checkbox"/>	Whabigall7_Draft	5	function unknown	98	214	0
<input checked="" type="checkbox"/>	Turbido	5	HNH endonuclease	98	214	0
<input type="checkbox"/>	Loser	4	HNH endonuclease	98	214	0

This is a quick way to look above and below the “function unknown” to see if there are other functions. Look at the relative score (3) and e-values to determine significance of matches with functions. To see all phages with one particular function, type part or all of the function description into the upper right Search box (2) and only those matches that correspond to the search text will be displayed. Once

you have identified a potential function, click on the Score header, once or twice, to bring the top scores, with the correct function, to the top of the list. Mark the Evidence box for the phage that best supports the function. Finally, in order to display your selected phage in the context of the other top matches, delete any text in the Search box (2), click on the Score header (3) to bring the top score matches to the top and then click the Evidence column header (4) to bring the selected phage evidence to the top.

ii. NCBI BLAST

Once a match has been identified in Phagesdb you can jump down to NCBI BLAST and view additional information such as % Identity (1), % Coverage (2), the range of BLAST alignment (3), and Region (4) /annotation information for each matching Accession (5). If the matching phage from Phagesdb is not visible in the NCBI BLAST results, type the phage name into the Search field (6) and then checkmark it's Evidence box (7).

( NOTE: If you still do not see the Phagesdb phage match in the NCBI BLAST list then click on the accession number of some of your top NCBI matches and click on the "Identical Proteins" button at the top left of the accession screen. NCBI BLAST lists only return one example of identical proteins and therefore you may only be able to find a Phagesdb match in the NCBI BLAS Identical Proteins list.)

To reset the display of your evidence relative to the % Identity, delete any text in the Search box (6) and click on the % Identity column header (1) to bring the highest identities to the top of the list and then click the Evidence column (7) header to bring the checked evidence to the top so that you can see it relative to the top identity hits.

NCBI BLAST															
Last Updated: 6/26/2016, 4:41:51 PM															
Show 10 entries															
Search: <input type="text"/>															
7 Evidence	5 Accession	4 Region	Creation Date	Description	1 % Identity	% Aligned	2 % Coverage	Positives	Target From	3 Target To	Query From	Query To	Gaps	E-value	
<input checked="" type="checkbox"/>	<a href="#">YP_007677413</a>	Yes	2013-04-14	Mycobacterium phage First	100	100	98.9796	98	1	98	1	98	0	0	
<input type="checkbox"/>	<a href="#">YP_009004454</a>	Yes	2014-03-04	Mycobacterium phage Echild	98.9796	100	98.9796	98	1	98	1	98	0	0	

Functional evidence (5a below) and conserved domain information labeled as Region (5b below) and Site (5c below) can be seen for each match by clicking on the highlighted Accession link (5 above) and scanning the FEATURES section, shown below.

FEATURES	Location/Qualifiers
source	1..98 /organism="Mycobacterium phage First" /host="Mycobacterium smegmatis mc2155" /db_xref="taxon:1245814"
<a href="#">Protein</a>	1..98 <b>5a</b> /product="HNH endonuclease" /calculated_mol_wt=11327
<b>5b</b> <a href="#">Region</a>	<45..72 /region_name="HNHc" /note="HNH nucleases; HNH endonuclease signature which is found in viral, prokaryotic, and eukaryotic proteins. The alignment includes members of the large group of homing endonucleases, yeast intron 1 protein, MutS, as well as bacterial colicins, pyocins, and...; cd00085" /db_xref="CDD:238038" <b>4a</b>
<b>5c</b> <a href="#">Site</a>	order(45,47..49,59..60,64..65,68,72) /site_type="active" /db_xref="CDD:238038"
<a href="#">CDS</a>	1..98 /locus_tag="First_004" /coded_by="NC_020876.1:2483..2779" /note="gp4" /transl_table=11 /db_xref="GeneID:15041000"

The Region column (4 top) displays a Yes for those Accession matches that contain Region or Site notes (5b, 5c). The Region and Site notes (5b, 5c) can be useful in selecting HHPred and Conserved Domain Database evidence that may be relevant to function, for example in the example above, cd00085 (4a) is also found as a matching reference in both HHPred and the Conserved Domain Database. Clicking on the Yes link (4) for an Accession will display the Region/Site information in a separate floating window for easy viewing and reference. Reference number header types that are found in the Conserved Domain Database include: COG, pfam, smart, and cd. HHPred reference types include PRK, PHA, pfam, cd, and TIGR. Search for these reference number header types in the Region and Site notes before looking for evidence in HHPred and the Conserved Domain Database.

### iii. HHPred

HHPred predicts the secondary structure of the query protein and then selects matches with known structures that correspond to the predicted structure.

The probability column (1) in combination with the E-value (2) provide confidence in the matching evidence. A probability above 98% will often be related to a function. Significant functions can be found with probabilities as low as 90%. For E-values, the lower the E-value the better, with a value of 0 being optimal. Look for E-values that are less than 1E-5. However, shorter proteins can have significant E-values as high as 0.1.

**HHPRED**  
Last Updated:  
6/26/2016, 4:43:21 PM

Show 10 entries

3 Search: [ ]

Evidence	5 Hit	4 Description	1 Probability	6 % Coverage	Target From	Target To	7 Query From	Query To	2 E-value
<input checked="" type="checkbox"/>	<a href="#">cd00085</a>	HNHc HNH nucleases; HN	99.3	54.0816	2	57	19	72	9.5e-12
<input type="checkbox"/>	<a href="#">4h9d_A</a>	HNH endonuclease; struc	99.3	68.3673	21	90	14	81	1e-12
<input type="checkbox"/>	<a href="#">pfam13395</a>	HNH_4 HNH endonuclea	99.1	44.898	1	53	32	76	1.2e-10

If you have previously identified a function from Phagesdb or a conserved domain from the NCBI BLAST Regions (such as the cd00085 domain discussed in the NCBI BLAST section) you can easily search for it using the Search box (3). Domain reference numbers may be found in the Hit column (5). Searching for functions in the Description column (4) can be tricky and sometime mislead you to think that there is no supporting HHPred evidence. This is because the Description field is truncated and the functional descriptors are not always consistent or fully representative between the target entries. To find the best evidence you may need to click on several Hit links (5) and read the fuller descriptions found in the Conserved Domains (1 below) or Protein (2 below) links at the right, under the Proteins header, of the NCBI search results page (shown below).

**Search NCBI databases**

cd00085 Search

Results found in 6 databases for "cd00085"

Literature		Genes	
Books	0 books and reports	EST	0 expressed sequence tag sequences
MeSH	0 ontology used for PubMed indexing	Gene	2,554 collected information about gene loci
NLM Catalog	0 books, journals and more in the NLM Collections	GEO DataSets	0 functional genomics studies
PubMed	0 scientific & medical abstracts/citations	GEO Profiles	0 gene expression and molecular abundance profiles
PubMed Central	2 full-text journal articles	HomoloGene	0 homologous gene sets for selected organisms
<b>Health</b>		PopSet	0 sequence sets from phylogenetic and population studies
ClinVar	0 human variations of clinical significance	UniGene	0 clusters of expressed transcripts
dbGaP	0 genotype/phenotype interaction studies	<b>Proteins</b>	
GTR	0 genetic testing registry	1 Conserved Domains	1 conserved protein domains
MedGen	0 medical genetics literature and links	29 Protein	29 protein sequences
OMIM	0 online mendelian inheritance in man	19 Protein Clusters	19 sequence similarity-based protein clusters
PubMed Health	0 clinical effectiveness, disease and drug reports	Structure	0 experimentally-determined biomolecular structures
<b>Genomes</b>		<b>Chemicals</b>	
Assembly	0 genome assembly information		

The % Coverage (6) and the Target/Query ranges (7) are helpful in deciding if this evidence points toward a local conserved domain or the actual protein function. Strong matches to short regions, 5-30% coverage, may be more indicative of a conserved domain. Coverages greater than 50% may support whole protein functions. Note that



HHPred may have much longer (or shorter) coverage for the same domains reported in the Conserved Domain Database because it is matching by predicted structure where the Conserved Domain Database is matching by amino acid homology.

#### iv. Conserved Domain Database

The Conserved Domain Database has a list of Accessions that match the Query protein sequence.

1. First look for Accession (1) matches to Region domains found in the NCBI BLAST Region or in HHPred hits.
2. Next look for matches with the highest % Identity (2a) and the highest % Coverage (2b) that have acceptable E-values (2c). The % Identity will be a fraction of the % Coverage.
3. Look to see if there is more than one distinct Query range (3) in the list of Accessions, if so, then mark the best representative Accession within each range that have acceptable e-values.
4. The Descriptions (4) are usually verbose enough that you should not need to click on the Accession link, never the less, the link is available if you would like to view more detail on the Accession.

Conserved Domain Database

Last Updated:  
6/26/2016, 4:41:20 PM

Show 10 4 entries Search:

Evidence	Accession	Description	% Identity	% Aligned	% Coverage	Positives	Target From	Target To	Query From	Query To	E-Value
<input checked="" type="checkbox"/>	<a href="#">cd00085</a>	HNH nucleases; HNH endonuclease signature which is found in viral, prokaryotic, and eukaryotic proteins. The alignment includes members of the large group of homing endonucleases, yeast intron 1 protein, MutS, as well as bacterial colicins, pyocins, and anaredoxins.	22.807	29.8246	27.551	17	27	57	45	72	0.000736605
<input type="checkbox"/>	<a href="#">pfam01844</a>	HNH endonuclease.	29.7872	38.2979	37.7551	18	6	46	36	73	0.0000143739

#### e. Entering a function

- i. Enter NKF for No Known Function. Proper entry for no function is important because during export of this field PECAAN excludes NKF tagged fields when building certain reports.
- ii. When entering functions, you need to use standard gene notation that is found in the [Function Assignments](#) table. When you start typing text into the Function field a drop down box will appear under the Function field with a list of standard gene notations that match your



text. Select the appropriate function or continue typing if no matches are found. If you find a function that is not shown in the standard gene notation list, first look at the function field in Phagesdb BLAST to see what is the most common function notation for strong matches, and use that notation. Clicking on the Function header in Phagesdb BLAST will sort and display the functions as groups. You can also check the Creation Date in the NCBI BLAST to see which Accessions have the latest creation date. Usually the Accessions with the most recent date will have the most standard notation.

- f. Entering comments into the Notes section
  - i. Student and faculty comments
 

During the annotation process, students and faculty should enter comments into the Notes field so people that check the annotations will be able to identify areas that the annotator is not confident in or that required extended investigation to derive the function or start site location.
  - ii. Faculty QC comments
 

When faculty do a final review of the genome before submission to the [Phagesdb Annotation portal](#), they need to generate a notes file to attach to the submission form. Text in the Notes field can be exported at any time using the Export CDS Notes option under the Export menu. It is therefore very useful to erase all student notes from the Notes field and enter only those notes that you want to submit to the SMART QC reviewers. Remember, you will need to press the Save button after entry of new notes.
  - iii. SMART QC Reviewer comments.
 

When the SMART QC reviewers do a final review of the genome, before sending your results back to the submitter for a final look, the Notes field can be used to easily gather comments and attach them to their associated CDS ranges. Since gene numbers sometimes can be shifted during the final file preparation, attaching the notes to a CDS range will avoid confusion. These notes can be exported using the Export CDS Notes option under the Export menu.
  - iv. Exporting the Notes
 

A text file containing only the notes, and their associated CDS region, can be exported at any time using the Export CDS Notes option under the Export menu.
  - v. Saving Changes
 

All changes in Functions, Notes and check-marked evidence is recorded in the Change Log whenever the Save button is pushed. If you go to another gene before saving your changes they will be lost. If the save button is pushed and no changes were made, then nothing new will be added to the Change Log. If you have reviewed a gene and find that no changes are necessary, make a statement, in the Notes field, that no changes are necessary and press the Save button. Your name and the note comment will be recorded in the Change Log. This

will allow other checkers know that you have reviewed the information.

g. The Change Log

Whenever the Save button is pushed in a gene record, all fields and checkboxes, that are changed, will be recorded in the Change Log field along with the Name of the User that made the change(s) and a date / time stamp.

i. Viewing the Change Log

The Change Log for each gene is located at the bottom of the Genes window. The Change Log field is a scrolling field, therefore all changes may not be visible in the initial view. Place your cursor over the field and scroll down and up to review all of the recorded changes.

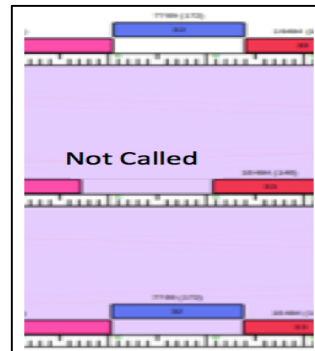
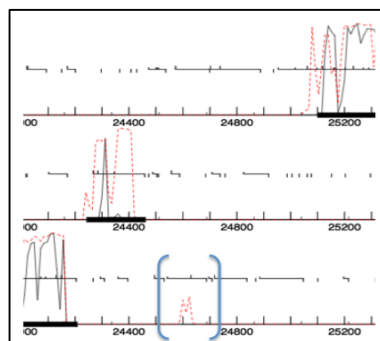
ii. Exporting the Change Log

All changes recorded in the Change Logs for all of the genes can be exported into a composite text file by clicking on the Export Changelog button on the Export menu page. This is useful for instructors that would like to review changes and comments left by students. This complete text file can be easily searched in most text viewers and is a fast way to find a gene with a particular function or note that you are looking for (much easier than scrolling through each gene's changelog).

h. Adding a gene

i. Reasons for adding a gene

1. If coding capacity is detected in the Host-Trained GeneMark for a region that is not encompassed by a gene (brackets below in left figure) you may want to add a gene and use the PECAAN evidence tools to determine it's efficacy.



2. Maps that align and compare the genome of interest to other published genomes, such as Phamerator maps (right above) or DNA Master alignments can be helpful in identifying regions that contain genes in other genomes that have not been initially called in the genome of interest. (A PECAAN genomic alignment and six reading frame map is being developed under the Sequences menu that will allow you to compare the query genome with the top closely related genomes.)

ii. How to Add a Gene

To add a gene into PECAAN click on the Add Gene button at the top of the Genes screen. A pop-up box will appear where you can enter the coordinates for the Start and Stop sites. A new gene will then be generated. PECAAN will then generate a Gene Candidate table and search the databases for matches to the predicted protein. If the start and stop do not correlate to the same reading frame or if the reading frame is not completely open between the start and stop sites then an error message will be generated indicating the problem with your coordinates. Start site coordinate is defined as the first base of the start codon (ATG, GTG or TTG) and the stop site coordinate is defined as the last base of the termination codon (TAA, TGA, of TAG).

i. Deleting a gene

i. Reasons for deleting a gene.

1. If coding regions greatly overlap or you have both forward and reverse genes covering the same region, you may choose to delete one of the genes.
2. If the gene is less than 40 amino acids or the reading frame is significantly less than 120 bp then you may consider deleting the gene.

ii. Deleting a gene.

To delete a gene in PECAAN, simply uncheck the Gene Included check box under the Gene Candidates header and then click the Save button. The gene will remain visible but will not be exported by any of the options under the Export menu.

Direction	Start	Stop	Length	Gap
Forward	411	1172	762	0

iii. Adding a gene back after a deletion.

To add a gene back after it has been deleted, simply recheck the Gene Included checkbox and press the Save button. The checked gene will then be included in all exported reports.

j. Calling tRNAs

- i. When viewing a tRNA window (below) look for the tRNA Scan SE Cove Score (1 below). In HyRo, which contains about 30 tRNAs, I found that real tRNAs have Cove Scores greater than 30 and non-real tRNA calls had Cove Scores less than 20. Unmark the tRNA Include box if the Cove Scores are less than 20.
- ii. The Aragorn tRNA calls have the most accurate Start and Stop predictions. If you need to adjust the 5' end, move the vertical marker,

at label 5 below, to the right or left to adjust the Start site. If you need to adjust the 3' end, move the vertical marker, at label 6 below, to the right or left to adjust the Stop site.

- iii. The acceptor arm of the tRNA is a double-stranded stem region of 7-9 bp made up by hydrogen bonding the 5' end of the tRNA with bases near the 3' end of the tRNA. In the figure above the seven bases, GGGTCTG, at the 5' end correlating with the (((((( bonding designators (region 2 above) are bonded to the CAGGCCT corresponding to the )))))) bonding designators (region 3 above) at the 3' end:

5'-G : TGCCA-3'  
 G : C  
 G : C  
 T : G  
 C : G  
 T : A  
 G : C  
 . . .

Yes, G can base pair with T in tRNAs. Notice that the 3' end has four unpaired bases GCCA that correspond to a -XCCA pattern found at all

tRNA 3' ends. Not all tRNAs have the CCA portion of this pattern coded in the gene and these bases can be added to the 3' end of the tRNA by cellular enzymes. When calling the 3' end of a potential tRNA gene, include all of the CCA pattern that is encoded by the gene sequence but no more.

- iv. When you have finished reviewing or editing the tRNA, click on the Change tRNA Data button (7 above) if you have made changes to the tRNA Include box or the Start/Stop positions. Click the Checked with no change button if you are satisfied with the current parameters but want to indicate that you have done a thorough review (8 above).
- v. Click the drop down box at label 7 (above) to select other tRNA genes.

k. Frame Shifts

- i. When encountering potential frameshifts two adjacent genes will align to the same BLAST target. Usually one of the query genes will align with one end of the target and the other query gene will align with the other end of the same target. This may indicate that a frameshift is taking place during translation or it may indicate that the gene has recently incurred a mutation in which a base has been inserted or deleted from the coding region. Frameshifts are commonly found in the Tail Chaperone Assembly protein. An example is shown below.
- ii. Open, in GenBank, the Accession link for phage JHC117 gene 26, and scroll down to the end of the annotation and amino acid sequence as shown below.

```

Protein      1..265
              /product="gp26"
              1 /name="tail assembly chaperone (G/T -1 frameshift)"
CDS          1..265
              /gene="26"
              /locus_tag="JHC117_26"
              3 /coded_by="join(JF704098.1:15181..15555, 4a
              JF704098.1:15555..15977)" 4b
              2 /ribosomal_slippage
              /transl_table=11
ORIGIN      5 (15555 - 15181 + 1) / 3 = 125 amino acids
           1 msnvftldsf reeadrefap vklelggdda vvlrnvlriq ktrreevfql lekldsiakd
           61 degkqreedd ldasemeamg dialrmielv adndalgsrl vdelrddlal tlkvfeawmn
           6 121 atqpggsral arlideygdg lvadlwetyg vdlrdiy lpe srlspklalv likelpvgsr
           181 fyaekrggkq frgwdesrya lvaiavnava lqytyvaahs kskpkppdpf ptpqrtkarq
           241 irkagsfawm aakqiaaark rkaqt

```

- iii. Notice that this gene is a “tail assembly chaperone” that has a frameshift (1 above). The frameshift is a result of “ribosomal\_slippage” (2 above). A frameshift occurs when the ribosome moves from one reading frame to another reading frame

before the termination codon in the first reading frame is reached. The protein is derived from the translation of two separate nucleotide sequences that are designated by the “join” annotation (3 above). The range of the first translated nucleotide range is shown in 4a above: 15181..15555. We can calculate the number of amino acids that are coded in this region by subtracting 15181 from 15555 (add 1 to the difference) and divide by 3 nucleotides/amino acid (shown in 5 above). The first 125 amino acids are coded for by the first nucleotide segment (ends at “..atqpg” shown in 6 above). Note that the last base of the first coding segment, 15555, is also used as the first base of the next coding segment. This is typical of a -1 frameshift (1 above).

- iv. In the Etude phage genes 19 and 20 both have matching BLAST hits to the Tail Chaperone Assembly protein and gene 20 is similar to JHC117 gene 26 (above). Gene 19 starts at base 14116 and gene 20 at base 14628. Open Etude and go to gene 19. Click on the top “Sequence” menu in PECAAN and scroll down the page until the six-frame translation is viewable (or open Etude in another six frame viewer if PECAAN does not display properly in your browser. This section is under construction). Drag the location bar to the right until you see the gene beginning at 14116. Notice that this gene is in the bottom forward reading frame. Scroll to the right to the beginning of gene 20, base 14628. Notice the reading frame is the top forward reading frame. Scroll to the left in the top reading frame until you find a stop codon. The reading frame shift must occur in gene 19 between this stop location and the end of gene 19. Look for possible slippery sequences (CCCC or GGGG...) in this region. You may see something like the figure below. The last five amino acids before the frameshift in phage JHC117 are ..ATQPG (6 above). Look along the green line below to locate this sequence. The final G in this sequence is located at the end of arrow 1 below.
- v. The G at the end of arrow 2 (below) is the first base of the codon (GGG) that codes for the Glycine (arrow 1 below). When the ribosome shifts -1 the anticodon of the tRNA now occupies the GGG codon that begins at the end of arrow 3 (below). Note that the next amino acid will also be a Glycine (arrow 4 below) that begins its GGA codon at the blue shaded G in the figure below. Since the blue shaded G is the last base of first coding region and it is also the first base of the second coding region it is called the slippery base. Note that the first five amino acids in the second coding region beginning at 4 below, GSRAL., are also the same sequence of amino acids found at the beginning of the JGC117 sequence shown after the 6 underline in the figure above, thus verifying the proper frameshift location.

Position: 14553

L R G V D E R D P A G G S R A L A R L I D E Y G  
 S R R G \* T R P S R G K P S A R P P D \* \* I R  
 F E A W M N A T Q P G E A E R S P A \* L M N T i  
 :TTCGAGGCGTGGATGAACGCGACCCAGCCGGGGGAAGCCGAGCGCTCGCCCGCCTGATTGATGAATACG  
 :AAGCTCCGCACCTACTTGGCGCTGGGTTCGGCCCCCTTCGGCTCGCGAGCGGGCGGACTAACTACTTATGC  
 R R P T S S R S G A P P L R A S A R R I S S Y  
 D E L R P H V R G L R P F G L A R G G S Q H I  
 K S A H I F A V W G P S A S R E G A Q N I F V

- vi. The position of the blue highlighted slippery base is 14553 (upper left above) and therefore we can construct an annotation note to go into the “Notes” field:

ribosomal slippage

CDS join(14116..14553;14553..14975)

by looking up the start for gene 19, 14116 and then putting the slippery base location, 14553 at the end of the first coding range and at the beginning of the second coding range and then finally putting the stop base for gene 20 at the end of the second range. Since none of the start sites listed under the gene candidates is correct for this gene, just leave the Selected Gene start at 14628. Set the “All GM Coding Capacity” to “Yes”. (This section of PECAAN will be modified to allow the proper entry and export of the joined regions representing the frameshift.)

## 8. Exporting Reports

A description for each of the Export buttons on the Export page is found above in section 6d.

- a. Export CDS Function

The Export CDS Function outputs a file in a format that can be imported into DNA Master, where the gene function is displayed in the Notes field on the Features page.

- i. Export from PECAAN

```
CDS 803 - 1234
  /gene="1"
  /product="gp1"
  /locus tag="TinaFeyge_1"
  /note=Terminase, small subunit

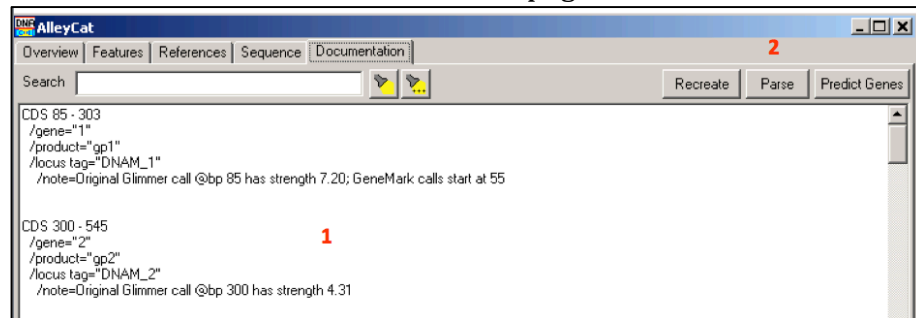
CDS 1326 - 1640
  /gene="2"
  /product="gp2"
  /locus tag="TinaFeyge_2"
  /note=minor tail protein
```



The figure above shows the format of the Export CDS Function file output. It includes the sequence range for each CDS as well as gene, product and locus tag information. If there is no /note line then no function will be displayed for the associated gene.

- ii. Searching and editing notes  
This output is a good format for quickly scanning the functions of genes or for searching for specific genes associated with a function. It can be easily searched by most text editing programs.
- iii. Importing the CDS Function file into DNA Master  
Open the CDS Function file in a text editor and copy all of the contents. Open the DNA Master file for the phage corresponding to PECAAN phage used to make the export.

Go to the DNA Master Documentation page.



Right-click on the body of the text (1) and click on the Select All option in the pop-up box. Right-click again on the body of the selected text (1) and then click on the Paste option in the pop-up box. The selected text should have been replaced with the text that you copied from the Phagename\_CDSfunction.txt file. Verify the replacement.

Next click on the Parse button (2) and approve the over-writing of the database. You should now be able to select the Features menu option and see the functions displayed in the notes field of the DNA Master file. Save the DNA Master file.

b. Export CDS Full Annotation

i. Export from PECAAN

The Export CDS Full Annotation button outputs a file with the following gene format:

```
CDS 1645 - 2718
/gene="3"
/product="gp3"
/locus tag="TinaFeyge_3"
/note=Original Glimmer call @bp 1645 has strength 1.44;
Genemark calls start at 1645
```



/note=SSC: Start = 1645, Stop = 2718. (Forward). CP: Does contain all GeneMarkSmeg capacity. SD: ZScore 2.7792 is the highest start score. SCS: Start is called by Glimmer and is called by Genemark. LO: 1074 bp is not the longest possible ORF. GAP: 4 bp. ST: SS=1645. F: NKF. FS: PHDBLAST= PhageName= Sabertooth, ProteinNumber= 4, Function= collagen-like, EValue= 0.0. NCBI BLAST= PhageName= Mycobacterium phage Peaches, Coverage= 99.7199, SubjectRange= 1:357, QueryRange= 1:357, EValue= 0.0. HHPRED= Accession= 3hqv\_B, Description= Collagen alpha-2(I) cha , Probability= 99.6. Coverage= 73.3894, SubjectRange= 544:802, QueryRange= 544:328. CDD= Accession= pfam01391, Coverage= 16.2465, SubjectRange= 1:59, QueryRange= 1:250, EValue= 3.47833E-4.  
 /note=Collagen-like region present.

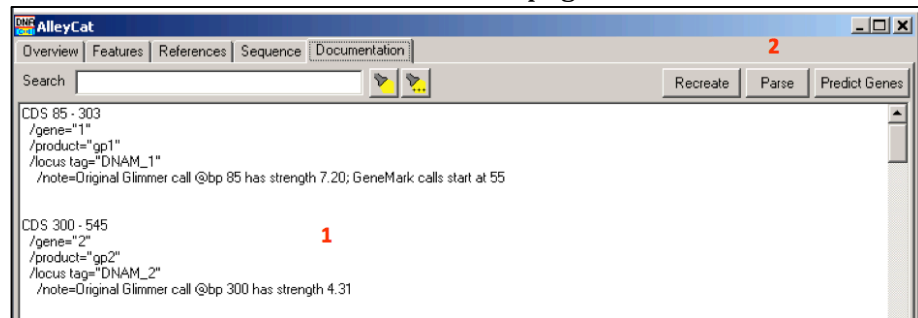
ii. Importing into DNA Master

The format shown above is designed for input into DNA Master such that the extended gene annotation is displayed in the Notes field on the Features page.

Open the CDS Full Annotation file in a text editor and copy all of the contents.

Open the DNA Master file for the phage corresponding to PECAAN phage used to make the export.

Go to the DNA Master Documentation page.



Right-click on the body of the text (1) and click on the Select All option in the pop-up box. Right-click again on the body of the selected text (1) and then click on the Paste option in the pop-up box. The selected text should have been replaced with the text from the Phagename\_CDSfunction.txt file. Verify the replacement.

Next click on the Parse button (2) and approve the over-writing of the database. You should now be able to select the Features menu option and see the functions displayed in the notes field of the DNA Master file. Save the DNA Master file.

c. Export CDS Notes

The Export CDS Notes button is used to export all of the currently displayed

text from the Notes field into a text file. This is very helpful for checkers that are preparing a genome for submission to the Phagesdb.org annotation portal. As a checker reviews the genome they can erase student notes and enter notes that they want the SMART quality control team to view. Potential problem genes or the rationale for gene/function decisions can be recorded in the Notes field. All of these notes are output together with the Export CDS Notes button.

i. Export from PECAAN

The format of the CDS Notes is shown below:

CDS 13769- 14571

/note=Ribosome Slippage: Join[13769..14125, 14125..14571]

where the sequence range is shown and the text from the Note field is attached below in the /note= annotation.

ii. Import into the Cover Page

The text from the Notes can be copied and pasted into field 9 of the Genome Annotation Submission Cover Sheet:

<p>9. Describe any issues or specific genes that you were unable to satisfactorily resolve, and warrant further inspection in the Quality Control review.</p> <p>CDS 643 - 1380 /note=Even though this gene is NKF, HHPRED and CDD match to short DUF4417 domain also annotated in Piro94. Include domain annotation.</p> <p>CDS 11218 - 12069 /note=Called start has very poor Z-score but matched Piro94 and would not contain all coding capacity at alternative starts. Is Gap of 29 sufficiently close for translation re-initiation?</p>
--

d. Export Author.csv file

The Author Export button will export a phagename\_authors.csv file as shown in the following format:

	A	B	C
1	Carroll	Amber	N
2	Khan	Sherafghan	.
3	Elliott	Davis	L
4	Brickeen	Xavier	K
5	McDavid	Jacob	K
6	Rinehart	Claire	A
7	King	Rodney	A
8	Staples	Amanda	K
9	Rowland	Naomi	S
10	Gaffney	Bobby	L
11			

e. Export Change Log

The Export Changelog is a handy output that contains all of the changes for each gene, including: the names of those who made the changes and the time/date stamp of when the changes were made. This text file is searchable and is a quick way to follow the course of the annotation development. An

example of the output is shown below:

```
Gene Number: 4
admin changed the gene function to structural at 2016-06-02 16:44:05.0
bgaffney changed the gene function to structural protein at 2016-05-31 10:48:05.0
bgaffney changed note to at 2016-05-31 10:48:05.0
jmbiddle changed note to There was reasonable evidence from Phagesdb to suggest this function
JMB at 2016-04-26 14:39:33.0
mkdaniels Added Phages DB Evidence Phage Name: Saintus Function: structural at 2016-04-19 15:13:34.0
mkdaniels changed the gene function to structural (minor tail) protein at 2016-04-19 15:13:34.0
mkdaniels changed note to There was reasonable evidence from Phagesdb to suggest this function at 2016-04-19 15:13:34.0
mkdaniels changed the coding capacity to yes at 2016-04-19 15:13:34.0
```

f. tRNA Changelog Export

The tRNA annotations are exported in the Export CDS Function and the Export CDS Full Annotation reports above but the tRNA Changelog Export button allows you to view the changes and the annotators that have checked the tRNA annotations.

g. tmRNA Changelog Export

The tmRNA annotations are exported in the Export CDS Function and the Export CDS Full Annotation reports above but the tmRNA Changelog Export button allows you to view the changes and the annotators that have checked the tmRNA annotations.

## 9. Acknowledgements

a. Authors

Claire A. Rinehart<sup>1,2,3,4</sup>, Bobby L. Gaffney<sup>1,2</sup>, James Dexter Wood<sup>1,3</sup>, and Jason R. Smith<sup>1,3</sup>.

1. BISC, Bioinformatics and Information Science Center, Western Kentucky University
2. Kentucky Biomedical Infrastructure Network (KBRIN) Small Genome Discovery (SGD) program. Research reported in this publication was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number **5P20GM103436**.
3. High Performance Computing Center (HPCC) at Western Kentucky University.
4. SEA-PHAGES (Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science)