PSAT
http://nwrce.org/psat

We have developed the Prokaryotic Sequence homology Analysis Tool (PSAT) to provide researchers an easy and efficient method for investigating sequence homologs among multiple genomes and examining their genomic neighborhoods. Exploration of genomic neighborhoods can help researchers identify potential regions of conserved gene order, or synteny, between genomes. Synteny can provide support for gene orthology or help identify a potential functional gene cluster.

Self tour:

For this tutorial, you will investigate the genomic neighborhood surrounding the pilQ gene in *Francisella novicida*

1. Finding Genes

    a. Click on **Explore Homologies** in the top menu to set Analysis Options.

    b. Select *Francisella_tularensis_novicida_U112* as the **Reference Genome**. You may leave the **Comparison Genomes** as the default of 'All genomes'. Later we will perform a more specific synteny analysis by selecting only a subset of these genomes.

    c. Enter **'pil'** as the **Gene name** to search for. The tool will find all genes with names that contain this string. You can also find genes by specifying a locus tag, gene product description, or genome location. These fields may be helpful, for example, if you are interested in investigating syntenic regions for genes with a particular function or genes that fall within a specific region of the genome. For this part of the tutorial, you can leave these other fields in this section blank.

    d. Leave the default **BLAST alignment score thresholds** setting to the default option (e-value<0.1, bit score > 200, % identity > 30%). The tool will look for BLAST hits that meet all of these cutoff values when determining what results to display. Later we will look at how the results change when choosing different threshold scores.

    e. Leave the display option to **Display only first 25 hits**. When there are a large number of hits, the tool can take a long time to query the database and display all the hits. Often only about the first 25 hits will be of interest, so this default option was introduced to improve efficiency. If you would like to view more hits, you may change this setting.

    f. Leave the display option set to the default value to **show all hits for each comparison genome**. Genomes may have duplicated regions and therefore have multiple hits of interest. An option to **show only top hits for each comparison genome** is also available.

g. Leave the display option set to the default value to **only show hits with a homolog clustering score of at least 2**. Since the homolog clustering score indicates the number of consecutive homologs in a region, a score of 1 essentially indicates an isolated homolog that is not likely to be part of a gene cluster. The default option therefore is to only show results with a score of at least 2. Users can still view isolated homologs by specifying a value of 1 in this field, if desired, and we will utilize this option later in the tutorial. For a more stringent filter based on the homolog clustering score, users can also specify a larger value in this field.

h. Check the two options under **List hit count in gene list** to **Display number of BLAST hits** and **Display number of genomes with hits**. These options are unchecked by default to help improve efficiency of loading the gene list page displayed once the form is submitted. Listing the hit count may provide a useful summary of BLAST results, so you may choose to wait a little longer to obtain these additional details.

i. Select the **Submit** button to submit the form.

2. Browsing List of Genes

a. Review the table at the top of the page. Verify that the settings we set in the input page match what is displayed here. Note that you can select the **Edit Settings** link to make any modifications you see necessary.

b. Note that 13 genes were found that matched our criteria for the gene name to contain the string 'pil'

c. The **Gene #** column indicates the position order index of a gene within the genome. Quickly scanning this column, you may note that several of the genes are found directly adjacent to each other (genes #s 1112-1116).

d. Note details such as **protein lengths** and **product** descriptions which can also be interesting.

e. Note the **number of BLAST hits** for each gene. The number displayed is only the number of hits found that meet all of our specified score thresholds in our selected comparison genomes (see grey section in the table at the top of the page to review our settings). Changing these setting values can change the number of hits indicated here.

The number of homolog hits found is specified and the number genomes in which hits were found. If the number of hits is more than the number of genomes, then some genomes may have duplicated regions. Also note that these numbers only reflect the number of hits with a homology clustering score of at least 2.

f. Find the *pilQ* gene. Mouseover this row to highlight it yellow, and click to bring up the Genomic Neighborhoods page.

3. Inspecting Gene Homologs and Homolog Clustering Scores

    a. The **number of BLAST hits** that met our specified BLAST score threshold values in our specified comparison genomes is displayed in the table header (see current settings in italics below this number). Since we used the default setting to display all hits, the number may include multiple hits from the same genome. For our current query, the number of total hits and number of genomes is the same, indicating there are likely no duplicated regions that meet the specified threshold values.

    b. The details for the selected gene *pilQ* is shown in the top row of the table, colored beige. The listed details, such as position, strand, length, name, locus tag and description, have been taken directly from the .ptt file published on NCBI. The gene# was determined based on the published position of all genes in the genome

    c. The homologs that meet the specified criteria are listed in the rows below the selected gene. Details about each homolog from NCBI published data (.ptt files) are listed directly below that of the query gene to facilitate comparison. The BLAST score values for the hit are shown in the columns colored gray.

       Note that for many of these homologs, the annotated gene name is also *pilQ* or the product description is very similar to that of the gene in *F. novicida*.

       You may collapse and expand the list of homologs by clicking the  to collapse and the  to expand.

    d. The homologs are by default ordered by the homolog clustering scores that are pre-calculated by the tool. Higher scores predict stronger synteny, so homologs with the largest scores are shown first. The score is defined as the number of consecutive, uninterrupted gene homologs neighboring a particular homolog pair.

    e. Reorder the results based on another value such as a BLAST hit score, or by Genome. Select the column headings that are links to see how the results are reordered. Selecting e-value will sort by lowest score first, and selecting any of the other BLAST score fields will sort by highest score first.

    f. You can also select which homologs to display genomic neighborhoods for in the graphic by checking and unchecking the checkboxes at the left of each row. For now, leave all homologs checked.

    g. The graphic should currently display the genome region surrounding the query *pilQ* gene in *F. novicida*. To view the genomic neighborhoods surrounding the homologs in the comparison genomes, click the  expand button to the left of the graphic or click on the **Update Graphic** button at the bottom of the table of homologs. Next you will explore this expanded graphic to analyze potential syntenic regions.

4. Analyzing Genomic Neighborhoods and Potential Gene Clusters

a. Note that a region of the reference genome *F. novicida* is drawn at the top of the graphic, with the selected query gene *pilQ* drawn at the center of this region. All genes in the reference genome are arbitrarily assigned a color. Genes found on the forward strand are drawn above the horizontal axis, and those found on the reverse strand are drawn below. The coordinates of each gene can be estimated using the ruler drawn at the very top of the image. The gene name or locus tag is displayed above each gene. All syntenic regions drawn below are aligned to this reference genome region

b. Scroll through the graphic displaying the genomic neighborhoods for homologs using the inset scroll bar. The scrolling feature allows you to easily align each potential syntenic region with the reference genome. Compare the regions shown in the graphic and the results in the table of homologs. The results should appear in the same order which can be most easily verified by comparing the bacterial species indicated in the first column of the table and on the right side of the graphic.

   The coordinates of the genome region displayed for each comparison genome are drawn in gray at both ends. An arrow at the left end indicates the direction in which the genome is drawn, an arrow pointing right indicating the same direction as the reference genome, and an arrow pointing left indicating the opposite direction (drawn reversed and inverted). For example, the *pilQ* gene is found on the reverse strand in *F. novicida* and its ortholog on the forward strand in *F. holarctica*. The arrow therefore points left to indicate that the region has been drawn reversed and inverted such that gene order in the region corresponds. Note that the coordinates of the displayed region also indicate this reversal (the coordinate on the left is greater than the one on the right).

   The color of the genes in each region indicate which gene in the reference genome it is homologous with. If a gene is gray, it indicates that it is not a homolog with any of the genes displayed in the reference genome (this does not mean it does not have a homolog in another region of the reference genome, but not in the immediate neighborhood). Because any gene that is a homolog to a gene displayed in the reference genome graphic is highlighted, PSAT can help researchers identify potential gene clusters in which the order of genes are slightly scrambled.

c. Mouseover one of the genes. A popup should appear. At the top of the popup are some details about the gene you are mousing over, including gene name, locus tag and description. If this gene is a homolog to a gene in the reference genome, details are also displayed about the query gene and BLAST hit score values for the gene pair. If multiple hits are found in the displayed genomic region, information about each hit is displayed (however, the gene color is determined by the top hit). The mouseovers can help with the analysis of each potential syntenic region by allowing you to inspect the ordering of homologs, assess the strength of each alignment, and compare product descriptions, if any.

5. Display Options
   a. The **Zoom** tool allows you to modify the size of the region drawn (in nucleotide bases). 25K bases is the default region size. Select 10K to view a smaller region size.

b. Now select 50K to view a larger region size. Since there are a large number of genes in this region, they may appear crowded in the graphic. You can increase the size of the image width to be able to view the genes more clearly. Select **++** under the **Image Width** option to increase the width. Now decrease the width slightly by selecting **-**.

c. Note the following output options:
    i. **Download file:** Download the graphic as a .pdf file
    ii. **Printer friendly:** Open a printer friendly version of the graphic in a new web browser window. This version does not include links, mouseovers, and does not have the scrollable bar for scrolling through the genomes to compare next to the reference genome.
    iii. **Tab delimited:** A tab delimited text format of the homologs in the region. This lists the genomes displayed in the graphic with the locus tag of the homologs in the genomic neighborhood. Homologs that do not meet the specified BLAST threshold score values are shown in parentheses.

6. Utilizing Tool Options
    a. Modifying BLAST Alignment Score Thresholds
        i. Scroll to the top of the Genomic Neighborhoods page and select the **Edit Settings** link in the header of the homologs table.
        ii. You should now be back to the input page, and the fields should be pre-populated with the current settings.
        iii. Select the option in the first column for **BLAST alignment score thresholds** (e-value < 1e-172, bit score > 600, % identity > 50) and resubmit the form.
        iv. The Genomic Neighborhoods page should reappear with different results. Since you selected to use stricter BLAST scores, the query returned fewer results. As you might expect, homologs were only found in the other *Francisella tularensis* genomes when using the stricter criteria

    b. Specifying a Subset of Comparison Genomes
        i. Select to **Edit Settings** again.
        ii. Edit the **Comparison Genomes** to include the *Salmonella*, *Shewanella*, and *Vibrio* genomes. Find these organisms in the top selection box. When you click on one, an option in the second selection box to select **All** genomes within the genus is available. Click on this option to select all and note that the third selection box indicates the selection.
        iii. As expected, the Genomic Neighborhoods page shows that no matches were found for these comparison genomes. Select to edit the settings again so that we can try using more relaxed criteria.
        iv. Select the radio button in the third column for **BLAST alignment score thresholds** (e-value < 10, bit score > 20, % identity > 10) and resubmit the form.
        v. The Genomic Neighborhoods page should reappear again. Since you selected to use less strict BLAST scores, the query returned results this time. Note that the results are ordered by homolog clustering score.

    c. Limiting the Genomic Regions Displayed in the Graphic

i. Assume that you are interested in comparing the genomic neighborhoods of homologs in the *Shewanella* genomes only. Deselect all the homologs in the table by clicking on the toggle checkbox icon  . Then reselect each homolog in the *Shewanella* genome individually.
ii. Click the **Update Image** button. Note that only the *Shewanella* genomes now appear in the graphic.

d. Finding Genes by Product Description
   i. Select to **Edit Settings** again.
   ii. Remove 'pilQ' from the **Gene name** text box and enter **cytochrome** into the **Description** text box.
   iii. Resubmit the form.
   iv. A list of genes should be displayed with Product including the string 'cytochrome'.
   v. Select the gene **cydB** in the list of results.
   vi. Note that the total number of hits total is greater than the number of genomes (indicating there may be duplicated regions in some genomes).
   vii. Click the 'Genome' column header at the top of the results table to sort by genome. Deselect all checkboxes for all hits by clicking the toggle checkbox icon at the top of the results table.
   viii. Note that the first 2 hits are in the same genome. Select the checkboxes for these 2 hits and scroll down to select the Update Image button.
   ix. View the genomic neighborhoods graphic for the 2 results

e. Viewing Top Hits Only
   i. Select to **Edit Settings** again.
   ii. Under option **4. Select Display Options**, change the setting under **Limit results based on top hits** to select **Show only top hits for each comparison genome**. Resubmit the form.
   iii. The results will now only show the top hit for each comparison genome (the hit with the lowest e-value will be considered the 'top' hit)

f. Filtering Homologs Based on Homolog Cluster Score
   iv. Select to **Edit Settings** again.
   v. Under option **4. Select Display Options**, change the number of hits to display to **100**
   vi. Change the display setting to **Only show hits with a homolog clustering score of at least '1'** and resubmit the form.
   vii. Note that the number of hits has increased with the cluster score constraint removed.
   viii. Browse the table of results, noting the scores for each hit. You will notice that some scores have a value of 1. These results are not filtered based on clustering of genes and so may contain more spurious hits.

Comments, problems, suggestions?
nwrce_tools@u.washington.edu