VIRUS PATHOGEN RESOURCE (ViPR) EXERCISES

I. Getting familiar with the ViPR site

Upon completion of this exercise, you will be able to navigate the ViPR site, find virus families, view basic information about these groups of viruses, and know how to contact the ViPR team with questions, suggestions or problems.

a. Go to http://www.viprbrc.org using any Internet browser (e.g. Firefox, Safari, Internet Explorer).
b. The ViPR site has each virus family separated from the others, so you need to select a virus family before proceeding to search and analysis. Select a virus family that you work with. You will be taken to the virus family home page.
c. On the virus family home page, you will notice a grey navigation bar consisting of the following tabs: Search Data, Analyze & Visualize, Workbench, Virus Families, and Home. These tabs are consistent across the ViPR site and are designed to help you navigate the site.
d. Scroll down the page and click on the “Information about the virus family” below the “Data Summary” bar.
e. In the blue banner, pull down the “Resources” menu and you can view the virus family’s About page, other virus pathogen resources, anti-viral drug information, and immunology resources.
f. Pull down the “Support” menu and view how to contact the ViPR team when you have with questions, suggestions, or problems.
g. Pull down the “Announcements” menu and view “ViPR Newsletters”.
h. Migrate back to the ViPR homepage by clicking the ViPR logo or the “Home” tab. Now select “Dengue” in the “Featured Viruses” section to get to the Dengue Virus page. Browse the Dengue Virus landing page.

What kind of genome does Dengue virus have?

II. Conduct Comparative Genomics Analysis

Upon completion of this exercise, you will be able to: search for virus sequences and view detailed information about these sequences in ViPR, perform a multiple sequence alignment and phylogenetic tree construction on a select set of sequence records to infer their evolutionary relationships, and identify nucleotide or amino acid positions that differ significantly between groups of viruses distinguished by specific host attributes.

In the following exercise, you will use Dengue virus 2 as an example to search and analyze sequences. The Dengue virus is native in tropical areas of the world and is endemic in areas where it co-localizes with the preferred Aedes aegypti mosquito vector. It can logically be assumed that DENV infections reported in clinics located in non-tropical regions of the world, and the United
States, are likely due to recent travel to endemic regions by the patient. These import cases can thus establish viral lineage in new regions as a result of human travel. The CDC has previously determined that by examining the recent travel history of patients having been clinically diagnosed with DENV between 1999 and 2000 validates such an explanation (17). As an example ViPR use case, we will extend the CDC study by performing an in-depth comparative genomics analysis of all DENV 2 isolates taken between the years of 1999 and 2000. This will involve the following bioinformatics workflow: 1) identify sequence records in the ViPR database using the metadata-driven query interface; 2) save selected sequence records as a working set in a personal workbench; 3) construct a phylogenetic tree using an optimal model of evolution and color the tree to highlight metadata differences; 4) visualize multiple sequence alignments to verify lineage relationships inferred from the phylogenetic analysis. Similar tasks can be performed for other viruses to address other biological questions by combining the wealth of relevant data with the suite of bioinformatics tools integrated into the ViPR resource.

1. Search for Dengue virus Type 2 genome sequences isolated from human between 1999 to 2000
   a. From the Dengue home page, mouse-over the “Search Data” tab and click “Genomes” to load the Genome Search page. You will notice you have many options to search: virus attributes, host attributes, clinical attributes, etc. From within the Taxonomy tree, click “Select All” next to “Dengue virus type 2”, enter 1999-2000 in “Collection Year”, select Human in “Host Selection” and click the orange “Search” button.

   ![Image of ViPR Genome Search](Image)

   The search result will be displayed in a table as shown below. Each column is sortable by clicking the header. Now click the “Country” header to sort records by country. Note: You can do advanced sorting by clicking the “Display Settings” button located above the result table.
b. On the Genome search result page, click on \( \textcolor{blue}{\text{ } } \text{ } \) to view the Strain Details.

i. What’s the name of the strain? Where was it isolated?

ii. Is there any clinical metadata associated with this strain?

iii. In the Genome section, click on “View Nucleotide Sequence” to retrieve the sequence. Click the “FASTA Download” button above the sequence to download the FASTA file to your computer for later use.

d. View the genome image map. Click “E” in the image map or \( \textcolor{blue}{\text{ } } \text{ } \) next to E in the protein information table to display the details of the Envelope protein. Look at the Genomic Annotation section. How long is the CDS?
v. Can you find the HMM/Pfam Domains information?

vi. Click “Prediction Details” in the “Predicted Epitopes” section. What is the sequence and supertype of the first epitope listed?

<table>
<thead>
<tr>
<th>HMM/Pfam Domains (999)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accession</td>
</tr>
<tr>
<td>FF0106.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicted Epitopes (10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHC Supertype</td>
</tr>
<tr>
<td>A2</td>
</tr>
<tr>
<td>A3</td>
</tr>
<tr>
<td>A4</td>
</tr>
</tbody>
</table>

c. Click on the “Results” breadcrumb at the top of the page to return to the Genome search result page. You can select records and run analysis on the selected records by mousing-over the “Run Analysis” button and clicking the desired analysis option. In the next exercise, we are going to store the sequences as a working set in the Workbench feature so that we can run various analyses on the working set.

2. Register for a Workbench and store sequences as a working set for further analysis

The ViPR Workbench provides free online storage space for you to:

- Save and organize working sets of sequences, analysis results and search criteria
- Share working sets and analysis results with others
- Upload personal sequences and combine with existing working sets

a. Select all records by checking the checkbox above the table and then click the “Add to Working Set” button. Now you need to register for a Workbench account in order to use this feature.

b. In the pop-up lightbox, you will be prompted to log in or register for a new account. If you don’t have a Workbench account yet, choose the “Register for a new account” button, enter your email, password of your choice, name and institution, and click “Register” to complete the registration process.

Alternatively, you can click either the “Register for a Workbench” link under the “Workbench” tab or the “Workbench Sign In” link in the top-right corner to get to the Workbench Registration page.
c. Enter a name and description for the working set. After the working set is saved to your Workbench, you will be able to view and run analysis on the working set during subsequent visits.

d. Click on the “Workbench” tab from the grey navigation bar to go to your Workbench. You’ll see the saved working set as well as unsaved searches and analyses you just ran during this session.

e. Click next to the working set that you just saved to display items in the working set. Now you will see the analysis tools available for your data type under the “Run Analysis” drop-down menu. Move on to the next exercise.

3. Multiple sequence alignment

a. Select all genomes in the working set by checking the checkbox above the table. Then mouse-over “Run Analysis” and click “Align Sequences (MSA)”.

b. Click “Continue” using default “Nucleic Acid (Genome)” option.

c. On the next page, select FASTA or any other desired output format and click “Run”.

d. Once the analysis finishes, mouse-over the blue “Run Analysis” button and click on “Visualize Aligned Sequences”.
e. On the Visualize Aligned Sequences Customization page, you can customize the sequence label in the alignment by selecting the “Custom” radio button and selecting one or more options from the list, e.g. Strain Name and Country. Click “Run” to load the visualized alignment.

f. Customize sequence labels in your alignment:
On the Visualize Aligned Sequences page, right-click on a strain name in the alignment, mouse-over the sequence name in pop-up menu, click on “Edit Name/Description”, modify the name as desired, and click “Accept”.

g. Highlight a sequence region on your alignment:
Within the JalView alignment visualization window, click and drag a desired region of sequence alignment, right-click on the selected region, mouse-over “Selection” and click on “Create Sequence Feature”.
h. Color alignment based on sequence identity cutoff:
   Examine various options to adjust the alignment display (e.g. File, Format, etc.).
   Click on “Colour” pulldown menu and then the “Above Identity Threshold” option. Using
   sliding bar, adjust color display such that only residues with >80% sequence identity are colored.
   Scroll left and right to view results.

4. Build a phylogenetic tree
   a. Create a phylogenetic tree from the aligned sequences using the “Quick Tree” model.
      On the Visualize Aligned Sequences page, click on the “Generate Phylogenetic Tree” button.
   b. On the next page, select “Quick Tree” and click on “Build Tree”. The analysis will take a
      couple of minutes to run. You can save the analysis to your Workbench by entering a name and
      then clicking “Save to Workbench”. Move to other parts of the exercises, and then come back to
      the Workbench to retrieve the analysis results.
   c. After the analysis is completed, click on “View Tree” to open the tree window.
      In the “Tree Decorations” section of the tree viewer window, click on the drop-down menu
      below “Basic Decoration Options” and choose “country”. Click on “Show Legend” to display
      the color code for different countries.
d. Now, color tree leaves by the “Advanced Decoration”.
   In the “Advanced Decorations” section, click on the drop-down menu and choose “Region”. You will see the tree is now colored by continents: Asia, North America, and South America.

e. Change the decoration colors.
   Click the “Advanced Decoration” again. In the pop-up “Advanced Decoration Options” window, click “Manual Decoration” then click “Go” to change the color. Check “North America” and choose red in the color palette, then click “Apply”.

f. Run tree model compare to determine which model fits your data best.
   Return to the Generate Phylogenetic Tree page using the breadcrumb feature. Choose the “Custom Tree” model and then “Run ModelCompare tool for recommendation of evolutionary model that best fits my data”.
   Which model of evolution fits best?

5. Metadata Genome Compare (meta-CAT)

In this exercise, we are going to compare the genome sequences from the previous phylogenetic tree.

a. Click the “Workbench” tab, find the working set you just saved and click to display the working set. Or click the breadcrumb of “Working Set” at the top of the page.

b. Select all sequences from the working set. Then mouse-over the “Run Analysis” tab and click “Metadata Genome Compare”.

c. We are going to separate these sequences into two groups by the two clades from the phylogenetic tree analysis excluding the Brazil strain and the Viet Nam strain and use “0.05” as our significance cutoff value, so keep the default settings on the page and click “Continue”.

d. Now, assign the sequences into two groups by double-clicking on the strains from the top clade (DENV_2/US/BID_1428/1999, DENV_2/US/BID_1425/1999, strains from other countries except Brazil and Viet Nam) in the top box and clicking the “Add” link above either of the boxes at the bottom of the page. Repeat the process to assign all the other US strains and the Brazil strain to the other group and click the “Run” button.

e. This analysis may take a few minutes to finish. You can save the analysis to your Workbench and retrieve it later. To do so, enter in a name and click “Save to Workbench”.

Virus Pathogen Database and Analysis Resource (ViPR, www.viprbrc.org) is funded by the NIAID under Contract No. HHSN272200900041C and is a collaboration between Northrop Grumman Health IT, University of Texas Southwestern Medical Center and Vecna Technologies. Influenza Research Database (IRD, www.fludb.org) is funded by the NIAID under Contract No. HHSN266200400041C and is a collaboration between Northrop Grumman Health IT, University of Texas Southwestern Medical Center, Vecna Technologies, SAGE Analytica and Los Alamos National Laboratory.

f. When the analysis completes, two tables will be displayed on the page. The first gives the statistical results for individual positions, while the second table reports the statistical results for pairs of groups found to contribute to the significant result at each column.

g. Click the blue “Show P-Values Bar Plot” button to view the p-values for the positions identified as significantly different between your specified groups.

h. Go back to the same working set, generate and view a new multiple sequence alignment. Find the positions of the significant nucleotide differences.

6. Saving and sharing working set, searches and analysis results
   a. Save an unsaved search or analysis to your Workbench:
      Go to your Workbench. Check an unsaved search or analysis, click the orange “More Actions” button and then “Save Unsaved Searches/Analysis” to save the search/analysis permanently to your Workbench.
   b. Share a working set/search/analysis or with others:
      Click the checkbox next to the item that you want to share, then mouse-over the orange “Sharing” button, and click “Share items with collaborator”. You’ll be prompted to enter your collaborator’s email address. Enter eva.rab@utsouthwestern.edu here.
   c. Upload your own file to the Workbench and run analysis of your choice:
i. On the Workbench page, click the orange “More Actions” button, then “Upload File”. Upload the FASTA file you downloaded in the previous exercise, or any FASTA-, Phylip-, Newick-, or PDB-formatted file.

ii. After the upload process is finished, you will be taken to the Workbench page. Find the file you just uploaded and click to display the file.

iii. On the next page, mouse-over the blue “Run Analysis” button at the top of the page to view available analysis options for your file.

III. 3D Protein Structure Visualization

At the end of this exercise you will be able to find and visualize protein structures for specific proteins and to adjust the images to highlight selected protein features.

a. From the grey navigation bar, mouse-over the “Search data” tab and click “3D protein structure”.

b. Search for the 3D structures of Dengue virus type 2. Select all Dengue virus type 2 and click “Search”.

c. On the 3D protein structure search result page, select “1OK8” and click on “View Structure” to display the structure.

d. You’ll be taken to the Jmol protein structure visualization page. Use mouse in display window to change view. In the “Display Options” section, change the Display Type to “Secondary Structure in Cartoon” to view the secondary structures of your protein.

e. In the “Highlight Ligands” section, check “Highlight Ligands in” green.

f. In the “Highlight by Swiss-Prot Position” section, highlight residues 617-624, which have been found to form beta-strand.

g. Rotate the structure as you need. Download the protein structure image with highlighted residues by clicking on “Save View As Image” beneath the image.
IV. Sequence Feature Variant Type Analysis

At the end of this exercise you will be able to identify regions of proteins with known structural, functional and immune epitope properties, and will be able to assess the level of sequence variation in these regions.

Note: The Sequence Feature Variant Type (SFVT) tool is available for Hepatitis C (subtype 1a), Dengue (serotype 2), and Pox (Vaccinia) viruses in ViPR. SFVT for other virus taxa within ViPR are being developed.

Sequence Feature Variant Type (SFVT) can help you identify sequence variations that may correlate with phenotypic characteristics, e.g., drug sensitivity/resistance, virulence, transmissibility, etc.

- Specific regions defined based on functional properties, structural properties, and immune epitope locations
- Obtained from literature and other databases and validated by domain experts

a. Make sure you are on the Dengue site. From the grey navigation bar, mouse-over the “Search Data” tab and click on “Sequence Feature Variant Types”.

b. You will be taken to the Sequence Feature Variant Types page. Review the description of Sequence Feature Variant Types. After you are finished, click on “Go to Sequence Feature List” to view all Sequence Features of Dengue virus 2 proteins.

c. Look at the Sequence Feature list of Dengue 2 proteins. How many Sequence Features are there for the E protein?

d. Click on the number in the “Structural SFs” category for E to view all structural Sequence Features of E.

e. Find Dengue_Virus_2_polyprotein_SF70 and click on to view its details. What structure does this Sequence Feature form?

f. How many variant types exist for this Sequence Feature in ViPR? What is the reference strain that defines the Variant Type-1 (VT-1)?

g. Click on the number in the “Strain Count” column for VT-1 to view all the strains that harbor this Variant Type. Then click on the next to any of the strains to obtain detailed metadata about that particular strain harboring this Variant Type. Click the “Country” column heading to re-order table by country. How many strains were isolated from Australia?

h. Go back to the Sequence Feature Details page by clicking the breadcrumb Sequence Feature Details (Dengue_Virus_2_Beta-strand_617(8)) under the grey navigation bar. Click the “Excel Download” button above the Variant Type table to download the full data set.
i. Click on the blue “Find a VT(s)” button to expand the VT search panel. Now search for Variant Types with mutations at 620 and 624. Change M620 and K624 to “?” to search for any residue at these positions, i.e., 617-624 FEI?DLE?
How many Variant Types do you find that differ from the reference strain at amino acid positions 620 and 624?

References: