Influenza Research Database & Virus Pathogen Resource Exercises

Table of Contents
Exercise I. Getting familiar with the IRD/ViPR site 1
Exercise II. Searching sequences and exploring sequence annotations 4
Exercise III. Conducting comparative genomics analysis 9
Exercise IV. Visualizing 3D protein structures 19
Exercise V. Browsing and analyzing host factor data 21
Exercise VI. Accessing immunology and serology data 24
Exercise VII. Submitting or annotating your own sequences 28
Exercise VIII. Analyzing Sequence Feature Variant Types 34

Exercise I. Getting familiar with the IRD/ViPR site

Upon completion of this exercise, you will be able to navigate the IRD/ViPR site, have a general idea of where to find the data and tools provided by IRD, and know how to contact the IRD/ViPR team with questions, suggestions or problems.

A. Getting familiar with the Influenza Research Database

a. Go to the IRD homepage (http://www.fludb.org) using any Internet browser.

b. On the IRD home page, you will notice: “Search -> Analyze -> Save to Workbench” in a light blue box. This suggests a workflow for using the IRD site and corresponds to the three core components in IRD:

• Data – browse and search primary and derived data
• Tools – analysis, submission and visualization
• Workbench – personal informatics workspaces
c. Above the light blue box is a grey navigation bar consisting of the following tabs: **Search Data, Analyze & Visualize, Workbench, Submit Data, and Home**. These tabs are consistent across the IRD site and are designed to help you navigate the site.

i. Mouse-over or click the “**Search Data**” tab to view available data search options.

ii. Mouse-over or click the “**Analyze & Visualize**” tab to view analysis and visualization tools provided by IRD.

d. In the blue banner, there is a compilation of links to useful resources.

i. Pull down the “**About Us**” menu and view Citing IRD, Our Publications, and Research Using IRD.

ii. Pull down the “**Announcements**” menu and view Meetings and Events and IRD Newsletters.

iii. Pull down the “**Resources**” menu and view the WHO vaccine strains, BEI reagent resources, anti-viral drug information, Reactome flu pathways, and other resources.

iv. Click the “**Support**” menu and view how to contact the IRD team when you have questions, suggestions, or problems.

v. Pull down the “**Support**” menu and click the “**Tutorials and Training Material**” link to view available tutorials and training materials.

---

B. Getting familiar with the Virus Pathogen Resource

a. Go to the ViPR homepage ([http://www.viprbrc.org](http://www.viprbrc.org)) using any Internet browser.
b. The ViPR site has each virus family separated from the others, so you will 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: “Search -> Analyze -> Save to Workbench” in a light blue box. This suggests a workflow for using the ViPR site and corresponds to the three core components in ViPR:

- Data – browse and search primary and derived data
- Tools – analysis, submission and visualization
- Workbench – personal informatics workspaces

d. Above the light blue box is 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.

  i. Mouse-over or click the “Search Data” tab to view available data search options.

  ii. Mouse-over or click the “Analyze & Visualize” tab to view analysis and visualization tools provided by ViPR.

e. Scroll down the page and click the “Information about the virus family” link below the “Data Summary” bar.

f. In the blue banner, there is a compilation of links to useful resources.

  i. Pull down the “About Us” menu and view Citing ViPR, Our Publications, and Research Using ViPR.

  ii. Pull down the “Announcements” menu and view Meetings and Events and ViPR Newsletters.

  iii. Pull down the “Resources” menu and view the virus family’s About page, other virus pathogen resources, anti-viral drug information, and immunology resources.

  iv. Click the “Support” menu and view how to contact the ViPR team when you have questions, suggestions, or problems.

  v. Pull down the “Support” menu and click the “Tutorials and Training Material” link to view available tutorials and training materials.

g. Return to the virus family homepage by clicking the virus family name at the right end of the grey navigation bar.

h. Migrate back to the ViPR homepage by clicking the ViPR logo or the “Home” tab. Now click a different virus family to get to the virus family page.
Exercise II. Searching sequences and exploring sequence annotations

Upon completion of this exercise, you will be able to search for virus genome or gene/protein sequences and view detailed sequence annotations in IRD.

A. Search for influenza virus segment sequences in IRD

For this exercise, you will search for 2009 H1N1 pandemic HA nucleotide sequences.

1. BLAST for similar sequences

The IRD BLAST tool utilizes the NCBI BLAST program set and has a collection of custom influenza sequence databases to search against.

a. Go to the IRD site (http://www.fludb.org/), mouse-over “Analyze & Visualize” in the grey navigation bar and click “Identify Similar Sequences (BLAST)” to load the BLAST page.

b. Select a format of sequences provided and a database to search.

c. Input sequences from one the following ways:

   • Use sequences found in the IRD database as BLAST input sequences
   • Upload a FASTA-formatted sequence file
   • Paste FASTA-formatted sequences
   • Use a working set consisting of a group of sequences saved to your Workbench

d. Adjust BLAST parameters in the Advanced Options. Click the “Run” button.
e. A BLAST ticket number will be generated for the process, you can either wait till the process to finish or copy the ticket number and use it to retrieve the search later on by clicking the “Retrieve an Analysis” option under the “Analyze & Visualize” tab and then entering the ticket number.

f. On the BLAST report page, all the nearest hits are listed in the top table. Click a hit to view its detailed alignment. Click an IRD link to view the hit’s segment/protein details page in IRD. You can select hits by checking the corresponding checkboxes and click the “Add to Working Set” button located above the table to save the hits to a working set for further analysis.

<table>
<thead>
<tr>
<th>Id</th>
<th>Sequence header</th>
<th>Bit Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>39852</td>
<td>h339052 (Country:China) Influenza A virus (A/Yangzhou/94(2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds</td>
<td>3372</td>
<td>0.0</td>
</tr>
<tr>
<td>293070</td>
<td>h325070 (Country:USA) Influenza A virus (A/Peru/1410/2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds</td>
<td>3372</td>
<td>0.0</td>
</tr>
<tr>
<td>344614</td>
<td>h324414 (Country:USA) Influenza A virus (A/California/04/2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds</td>
<td>3372</td>
<td>0.0</td>
</tr>
<tr>
<td>777277</td>
<td>h327727 (Country:USA) Influenza A virus (A/California/04/2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds</td>
<td>3372</td>
<td>0.0</td>
</tr>
<tr>
<td>532778</td>
<td>h325278 (Country:Laboratory) Influenza A virus (A/California/04/2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds</td>
<td>3356</td>
<td>0.0</td>
</tr>
<tr>
<td>251026</td>
<td>h3251026 (Country:USA) Influenza A virus (A/California/08/2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds</td>
<td>3356</td>
<td>0.0</td>
</tr>
<tr>
<td>252216</td>
<td>h3252216 (Country:Japan) Influenza A virus (A/Fukuoka-G102009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds</td>
<td>3356</td>
<td>0.0</td>
</tr>
<tr>
<td>252228</td>
<td>h3252228 (Country:Japan) Influenza A virus (A/Fukuoka-G102009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds</td>
<td>3356</td>
<td>0.0</td>
</tr>
<tr>
<td>252274</td>
<td>h3252274 (Country:Japan) Influenza A virus (A/Fukuoka-G102009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds</td>
<td>3356</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Note that IRD also provides the “Identify Short Peptides in Proteins” tool accessible from the “Analyze & Visualize” tab, which allows you to find epitopes, ligand binding sites, sequence domains, or other short amino acid strings from among a target set of proteins.

2. Search for sequences using structured search interfaces

a. Go to the IRD homepage (http://www.fludb.org/), mouse-over “Search Data” in the grey navigation bar and click “Nucleotide Sequences” to load the sequence search page.

b. You will notice you have many options to search: sequence type (nucleotide or protein), strain, virus type, segment number, host attributes, etc. As you can see, by default, segment/nucleotide and influenza A are pre-selected. Type H1N1 in the subtype box, select “4 HA” in “Select Segments” and “Swine” from “Host”, and enter 1999 in the From box of “Data Range”.

c. Note that IRD shows instant count of search results here to help you search quickly and efficiently. When you select search criteria on the search page, you will instantly know how many records match your search criteria without clicking the “Search” button and actually running the search. If there are too many or not enough search results, you can quickly adjust the search criteria on the search page to better fit your needs.

d. If you want to find strains with high similarity to 2009 pH1N1 sequences, click “Include/exclude records with high similarity to 2009 pH1N1 sequences” below “Select Segments” and select “Include only pH1N1 sequences”. Click “Search” to run the search.
e. The search result will be displayed in a table as shown below. Each column is sortable by clicking the header. Now click the “Flu Season” header to sort records by flu season. Note: You can do advanced sorting by clicking the “Display Settings” button located above the result table.

f. On the search result page, click on 1 to view the details of a segment.

i. What’s the name of the strain?

ii. How long is the segment?

iii. Does the sequence have a perfect match with the commonly used diagnosis PCR primer set? Hint: Look at the PCR Primer & Probe details.
iv. Can you find the HMM/Pfam domain information?

v. Any predicted epitopes for this segment?

<table>
<thead>
<tr>
<th>Predicted Epitopes (SCOP) Prediction Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIC Subtype</td>
</tr>
<tr>
<td>A3</td>
</tr>
<tr>
<td>A2</td>
</tr>
<tr>
<td>A24</td>
</tr>
<tr>
<td>B7</td>
</tr>
<tr>
<td>B44</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

vi. Scroll up to the Segment Information section. Click “View Sequence and design PCR primers” to retrieve the sequence.

vii. On the next page, a “PCR Primer Design” button is located above the sequence box. Click the button to get to the PCR Primer Design page.

viii. Here you can see the sequence you just viewed is pre-populated into the sequence box. To design primers, specify desired regions on which primers should be excluded, included, or targets, or to overlap, adjust the parameters if needed, and click “Pick Primers”.

ix. A primer report will be loaded which shows the PCR product size, primer sequences, and computed parameters.

g. Click the “Results” breadcrumb at the top of the page to return to the 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. You can also store the sequences as a working set in the Workbench so that you can run various analyses on the working set.
B. View consensus sequence across all influenza A virus strains in IRD

- IRD provides pre-computed consensus sequence and polymorphism score at each position for all influenza A strains.
- IRD’s Sequence Variation Analysis tool allows you to calculate polymorphism of IRD sequences or your own sequences.

For this exercise, you will search for the consensus and all sequence variations existing in all GenBank influenza type A sequences based on host of isolation, flu subtype and segment.

a. Go to the IRD site (http://www.fludb.org/), mouse-over “Analyze & Visualize” in the grey navigation bar and click “Analyze Sequence Variation (SNP)” to load the Analyze Sequence Variation page.

b. On the tool landing page, select “Pre-computed analysis using sequences in the IRD database for a specified host, segment, and subtype” in search type.

c. Next, select desired host, subtype, analysis type (genomic or protein sequence), and segment/protein to analyze. Click “Run”.

d. The analysis result page will be loaded, which shows the polymorphism score, consensus, and counts for each different base/amino acid at each position. The conservation score ranges from 0 (no polymorphism) to 200 (highest polymorphism). At each position, the consensus is the allele with frequency greater than 50%. If no allele exceeds 50%, N (for nucleotide) or Xaa (for amino acid) is used to indicate ambiguity. Sequence polymorphism plot, consensus sequence, and raw alignment are available for download.
Exercise III. Conducting 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 IRD, build a phylogenetic tree on a set of sequences to infer their evolutionary relationships, perform a multiple sequence alignment, SNP analysis and Meta-CATS to identify nucleotide or amino acid positions that differ between groups of virus sequences, and examine the geospatial locations of the surveillance records.

For this exercise, you will use IRD to find influenza HA nucleotide sequences associated with North American duck and shorebird surveillance records and then conduct comparative genomics analysis on these sequences. This will include: (a) finding interested sequences and saving them to a working set, (b) constructing a phylogenetic tree to infer evolutionary relationships between the sequences, (c) aligning the nucleotide sequences, (d) determining positions that differ among the sequences, and (e) visualize related surveillance data on a map for possible migration-based influenza virus spreading analysis.

1. Search for HA nucleotide sequences from H4 subtype viruses isolated from Anatidae (duck) and Scolopacidae (shorebird) surveillance samples and save them to a working set

   a. Go to the IRD website (http://www.fludb.org). From the grey navigation bar, mouse over “Search Data” and click “Animal Surveillance Data”.

   b. On the animal surveillance data search page, select the following parameters:

      Surveillance Data Type: Avian
      Sample Selection: Only samples linked to sequence data
      Subtype of flu positive samples: H4
      Geographic Grouping: North America
      Host selection: Select family from list. Then select Anatidae and Scolopacidae from the list.

   c. The search results are presented in a table. How many H4 positive surveillance samples isolated from Anatidae and Scolopacidae are currently found in the IRD?
d. Click the “Subtype” header to sort the results by subtype. If you need advanced sorting options and want to display additional fields, click the “Display Settings” button.

e. Save segment 4 HA nucleotide sequences to a working set

i. Select all records by clicking the checkbox above the results table. Then click the “Add to Working Set” button.

Note: A working set is a container where you can save your regularly used sequences, strains or surveillance records. With your interested data saved in a working set, you can perform various analyses on the working set without having to search and find your interested data every time you do an analysis. To use this feature, you will need to register for an IRD Workbench account for free so that you can save your working sets to your online Workbench account. You will also be able to save and share other types of data and analyses online via the Workbench.

ii. You will be prompted to log in to your Workbench if you haven’t done so. If you don’t have an IRD Workbench account yet, register for an account.

iii. A lightbox of “Add to Working Set” will pop up. Now choose the data type. For this exercise, we will analyze the HA nucleotide sequences associated with the surveillance records, so choose “segments” as the data type and then segment 4. Name the working set to be “North American H4 ducks and shorebirds” and click “Add to Working Set” to save the HA sequences to a working set.

iv. Access your Workbench by clicking the “Workbench” tab in the grey navigation bar. You will see the new working set at the top of the Workbench content list. Click 🕵️‍♂️ to display items in the working set. How many segment 4 records are there in the working set?

f. Add an outlier to the H4 working set to help comparative genomics analysis

i. From the Workbench table, select the H4 ducks and shorebirds working set by ticking the checkbox next to it. Click “More Actions” and then “Copy” in the pop-up lightbox. Name the copied working set to be: North American H4 ducks and shorebirds + H3 outlier.

ii. Now search for HA sequence from A/mallard_duck/ALB/26/1976 and add it as an outlier to the copied working set.

• Mouse over the “Search Data” tab and click “Nucleotide Sequences”.
• On the Nucleotide Sequence Search page, select “4 HA” from the “Select Segments” list, type “A/mallard_duck/ALB/26/1976” in the strain name box, and click “Search” to run the search.
• The search result page will show 1 segment. Select the segment by ticking the checkbox next to it. Then click the “Add to working set” button and add it to the copied working set: “North American H4 ducks and shorebirds + H3 outlier”.

iii. Click the “Workbench” tab in the grey navigation bar and view the “North American H4 ducks and shorebirds + H3 outlier” working set. How many records are there in the working set now?

2. Phylogenetic analysis on North American H4 ducks and shorebirds

a. From your Workbench, click next to “North American H4 ducks and shorebirds + H3 outlier” to display items in the working set.

b. On the working set page, select all records by checking the checkbox above the table, mouse-over the “Run Analysis” button and click “Generate Phylogenetic Tree”.

c. You will get a warning message when you have mixed subtypes in your data input. Click “OK” to proceed.

d. On the “Generate Phylogenetic Tree” page, you have the options of choosing your desired tree model and tree tip labels. For this exercise, choose following parameters then click “Build Tree”.

   Tree Generation: Quick Tree
   Input: Use all 214 segments and run MUSCLE to align them
   Label Tree Tips (Ends) with: Strain Name

e. When you have a large amount of input, the analysis may take some time to finish. While the analysis is running, you can choose to save the analysis result to your Workbench upon completion by typing an analysis name in the “Save Analysis to Workbench” box and clicking the “Save to Workbench” button. Now you can move to other parts of the site. The analysis result can be retrieved from your Workbench later.

f. After the analysis is finished, a “View Phylogenetic Tree” page will be loaded. Here you can save the phylogenetic file in Newick or PhyloXML format to your computer. Click “View Tree” to load the Archaeopteryx Tree Viewer window.
g. A Tree Viewer window will pop up. Many tree customization options exist including: reroot the tree, collapse/expand/display subtree, swap descendants, decorate tree leaves by color, resize the tree, zoom in/out, fit the tree to window, change the font size, etc.

i. In this exercise, we will first reroot the tree by clicking the node of the outgroup: A/mallard_duck/ALB/26/1976. How many major branches does the tree have? Can you find the strains that are most similar to the shorebird strains?

Hint: You will see that the shorebirds form a branch in the tree.

ii. To rotate or swap descendants, choose “Swap Descendants” in the dropdown menu in the “Tree Manipulations” section to the left of the tree and then click a node.

iii. Choose “Country” from the “Basic Decoration Options” drop-down list to color the tree by country. Click “Show Legend”.

iv. Now use the advanced decoration option to color USA strains in blue and Canada strains in red. Click the “Advanced Decoration” button in the “Tree Decorations” section. A dialog box will pop up. Choose decorate by country and then “Manual Decoration” and click “Go”. A manual decoration dialog box will pop up. Here you can select a country, choose a desired color for that country, and click “Apply” to see the decorations on the tree.

v. Next, color strains by isolation year grouped into 5-year intervals. To do so, click the “Advanced Decoration” button. In the pop-up dialogue box, choose “Year” and then “Scale Year” 5 and click “Apply”. To choose your own colors for the isolation years, tick the checkbox next to “Manual Decoration”, click “Go” and follow prompts.

vi. You can export the tree image by using options under the “File” menu.
Save the tree analysis to your Workbench. There are two ways of doing this:

- Return to the “View Phylogenetic Tree” page and save the tree analysis to your Workbench by clicking the “Save Analysis” button and entering a name.

- Or, you can go to your Workbench, select the TREE analysis listed at the top of the workbench table, and click “More Options” and then “Save unsaved analysis” to save the tree analysis to your Workbench. You will be able to retrieve the tree analysis from your Workbench later.

3. Multiple sequence alignment

- From your Workbench, click next to “North American H4 ducks and shorebirds” to display items in the working set.

- On the working set page, select all records by checking the checkbox above the table, mouse-over the “Run Analysis” button and click “Align Sequences (MSA)”.

- In the pop-up “Select Sequence Type” lightbox, choose “Nucleic Acid (Segment)” and click “Continue”.

- The Align Sequences (MSA) page will be loaded. Here you can select desired output format and sequence output order. Keep the default selections and click “Run”.
e. As soon as the alignment is finished, the Alignment Report page will be loaded. To visualize the alignment, mouse-over “Run Analysis” and click “Visualize Aligned Sequences”. Click “Run” on the next page.

f. The visualized sequence alignment window will be loaded. Many customization options are available in the JalView visualized alignment window.

i. Scroll down the alignment to include sequences from shorebird, Alberta duck, and new duck. Scroll right to region 465-520. Can you find the residues that distinguish the shorebird and Alberta duck from the new duck sequences from the United States, i.e. residues shared by ruddy turnstone and old Alberta duck but not new duck from the United States?

ii. Color alignment based on sequence identity cutoff. To do so, click the “Colour” menu and then “Above Identity Threshold” from the list. Using the sliding bar to adjust color display such that only residues with >80% sequence identity are colored. Scroll left and right to view the alignment.

iii. View the consensus sequence and bar graph of conservation score at the bottom of the alignment window.

iv. If you need any help with JalView, click the “Help” menu and then “Documentation” to access the Jalview Documentation site.
4. Sequence variation analysis (SNP)

a. From your Workbench, click next to “North American H4 ducks and shorebirds” to display items in the working set.

b. On the working set page, select all records by checking the checkbox above the table, mouse-over the “Run Analysis” button and click “Analyze Sequence Variation (SNP)”.

c. After the analysis is finished, the SNP analysis result will be displayed. Click the “Score” column to sort the results. Write down 3 positions that are highly variable.

<table>
<thead>
<tr>
<th>Position</th>
<th>Score</th>
<th>Consensus</th>
<th>T</th>
<th>G</th>
<th>A</th>
<th>C</th>
<th>Deletion</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>421</td>
<td>145</td>
<td>G</td>
<td>47</td>
<td>47</td>
<td>0</td>
<td>0</td>
<td></td>
<td>207</td>
</tr>
<tr>
<td>1009</td>
<td>117</td>
<td>G</td>
<td>27</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td></td>
<td>208</td>
</tr>
<tr>
<td>1108</td>
<td>113</td>
<td>G</td>
<td>10</td>
<td>43</td>
<td>0</td>
<td>0</td>
<td></td>
<td>207</td>
</tr>
<tr>
<td>1096</td>
<td>113</td>
<td>G</td>
<td>10</td>
<td>43</td>
<td>0</td>
<td>0</td>
<td></td>
<td>207</td>
</tr>
<tr>
<td>1051</td>
<td>106</td>
<td>T</td>
<td>25</td>
<td>156</td>
<td>0</td>
<td>37</td>
<td></td>
<td>208</td>
</tr>
<tr>
<td>277</td>
<td>104</td>
<td>G</td>
<td>24</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td></td>
<td>207</td>
</tr>
<tr>
<td>467</td>
<td>104</td>
<td>A</td>
<td>106</td>
<td>1</td>
<td>103</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>517</td>
<td>104</td>
<td>A</td>
<td>152</td>
<td>0</td>
<td>150</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1249</td>
<td>100</td>
<td>T</td>
<td>0</td>
<td>151</td>
<td>9</td>
<td>51</td>
<td></td>
<td>206</td>
</tr>
</tbody>
</table>

5. Metadata-driven Comparative Analysis Tool for Sequences (Meta-CATS)

Use Meta-CATS to identify positions that are significantly different between the shorebird, old Alberta duck, and the new duck sequences from the United States. For your convenience, segment sequences in each of these major branches have been saved in a separate working set for Meta-CATS analysis.

- A unique comparative genomics analysis tool in IRD to identify nucleotide / amino acid positions that significantly differ between two or more groups of virus sequences.
- Meta-CATS consists of three parts: a multiple sequence alignment (using MUSCLE), a chi-square goodness of fit test to identify positions (columns) of the multiple sequence alignment that significantly differ from the expected (random) distribution of residues between all metadata groups, and a Pearson's chi-square test to identify the specific pairs of metadata groups that contribute to the observed statistical difference.

a. Go to your Workbench and check the “Public” checkbox in the Access section on the left of the page. Find and select these three working sets: shorebird NJ DE, duck ALB, and duck MN ND TX AK CA. Click the “More Actions” button and then “Copy” in the lightbox to copy the working sets to your own Workbench.

b. Mouse-over “Run Analysis” and click “Metadata-driven Comparative Analysis Tool” to input selected sequences to Meta-CATS.
c. The Meta-CATS landing page will be loaded. You can choose “Manual Grouping” if you want to manually group your sequences. If you want to group sequences by host, country, year, viral species, virus type, host age, host gender, or cohort, you can easily do so by using the “Auto Grouping” option. In this exercise, we will group the sequences by the three major branches from the phylogenetic tree analysis. Choose “Use working sets” and select the three working sets you just copied to your Workbench: shorebird NJ DE, duck ALB, and duck MN ND TX AK CA. Then click “Continue”.

d. On the next page, you will see that your sequences are grouped by working set and pre-populated in three groups. You can manually remove any sequence from the lists by selecting a sequence and then clicking “Remove”. After you are finished with grouping, click “Run”.

---

Created by the ViPR/IRD team and licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License
e. While the analysis is running, you can choose to save the analysis result to your own Workbench upon completion by typing an analysis name in the “Save Analysis to Workbench” box and clicking the “Save to Workbench” button. Now you can move to other parts of the site. The analysis result can be retrieved from your Workbench later.

f. The Meta-CATS analysis result has two reports: a Chi-square Goodness of Fit test result table listing the positions that have a significant non-random distribution between your specified groups, and a Pearson's chi-square test result table listing the specific pairs of groups that contribute to the observed statistical difference.

g. Review the Chi-square test results to see the positions that differ significantly between shorebird, old Alberta duck, and the new duck sequences from the United States. Sort the results by the Goodness of Fit C-value to push the most different positions to the top of the table. What is the position number with the most significant C-value?

### 6. Display surveillance records on a map

a. From your Workbench, click 🗺 next to the surveillance search you did in step 1.c. to load the surveillance search results. Select all records by clicking the checkbox above the result table. To view surveillance records on a map, you can click the orange “View on Map” button. A map will be displayed as shown below.

b. On the surveillance map, click 📍 to view details about the records at that location. Can you find locations where Scolopacidae (shorebird) samples were collected?

c. From step b, in the pop-up window, follow “Click here to view all the records at the location”. This will bring up all surveillance records collected from this location. Click 🆩 next to a flu-positive sample to see the sample collection, the host, etc.
d. Highlight bird flyways and see if there is a correlation between flyways and bird samples. Check “Atlantic Americas”, “Mississippi Americas”, and “Pacific Americas”.

Compare the phylogenetic tree generated from step 2.g.iv-v with the surveillance map with bird flyways highlighted. Can you make any hypothesis on why the Alberta duck sequences are more similar to the shorebirds?

Reference:

Exercise IV. Visualizing 3D Protein Structures

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

Use the Influenza Research Database to examine an influenza virus protein structure

a. Go to the IRD homepage (http://www.fludb.org/). From the grey navigation bar, mouse over “Search Data” and click “3D Protein Structures”.

b. Search for the 3D structures of influenza A (H3N2) HA protein.
   
   Virus Type: \[\text{H3N2}\]
   Subtype: \[\text{H3N2}\]
   Select Proteins to search: \[\text{HA}\]

   c. On the 3D protein structure search result page, choose the “1EO8” protein structure and click “View Structure” to get to the structure page.
   
   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. Highlight the structure with a sequence conservation score by selecting “Sequence conservation computed using all sequences” within the “Highlight Sequence Conservation” section.
   
   f. In the “Highlight Ligands” section, check “Highlight Ligands in” green.
   
   g. In the “Highlight Epitopes” section, highlight B-cell epitopes. Are there any B cell epitopes that are entirely conserved among influenza virus strains in the database?
   
   h. In the “Highlight Sequence Features” section, highlight in pink the following sequence feature on the structure: \text{Influenza A\_H3\_sialic-acid-binding\_98(19)} (Influenza A\_H3\_ SF61). Examine relationship between amino acid residues known to be required for sialic acid binding and position of sialic acid ligands in the protein crystal structure.
   
   i. In the “Highlight by Swiss-Prot Position” section, highlight residues 226 and 228, which have been found to influence virus host range.
j. Rotate the structure to your needs, then click the “Rock” button in the Display Options section and adjust the rocking parameters if needed.

k. Download the protein structure image with highlighted residues by clicking “Save View As Image” beneath the image. Generate and download a spinning protein structure movie by clicking the “Generate Video” button.
Exercise V. Browsing and analyzing host factor data

Browse host factor data that is generated through the NIAID-funded Systems Biology projects, ViPR-funded Driving Biological Projects, and other related studies.

IRD and ViPR now contain differential expression data and structured metadata derived from gene expression studies generated through the NIAID-sponsored Systems Biology for Infectious Diseases Research programs.

Host factor data from siRNA high throughput screens and proteomics experiments as well as tools to visualize protein-protein interaction networks and functional pathways will be added in the near future.

Currently, the host factor data available in IRD (http://www.fludb.org) and ViPR (http://www.viprbrc.org) is derived from multiple host factor experiments involving influenza or SARS-CoV virus infections in human cell line and in experimental animals.

1. View host factor data

To access host factor data, mouse-over the “Search Data” tab and click “Host Factor Data”. The Host Factor Data landing page lists all available host factor studies. Here, you can:

- View host gene differential expression patterns from an experiment. To do so, click next to a study on the Host Factor Study Data landing page to show the experiments associated with the study; click next to an experiment name to load the Host Factor Experiment Details page; next, click the “View Result Summary” button above the displayed table to view the Host Factor Experiment Result Summary page, which contains a summary table of differential expression patterns; clicking the hyperlinked number in any row will link to the list of genes that have the selected pattern.

- View all differentially expressed host genes obtained from various experimental conditions. To do so, follow steps described above to get to the Host Factor Experiment Result Summary page; click the “View All Differentially Expressed Genes” button at the top of the results.
table to show all differentially expressed genes at various experimental conditions obtained from the experiment.

- Search for a specific gene on any list of differentially expressed host genes and find genes with the same expression pattern from an experiment. To do so, on the Experiment Result Summary page, type one or more gene symbol(s) or Probe IDs in the “Gene Symbol” search box and click the “Find” button.

- View combined expression patterns of multiple biosets from the same experiment. Since each bioset is generated by comparing infected samples with mock-infected samples under different conditions, this allows for more customized data-mining analyses. To do so, click next to a study on the Host Factor Study Data landing page to show the experiments associated with the study; check the box next to an experiment name and click the “View Associated Biosets” button to load the Bioset Information page; a list of biosets generated at each experiment measurement point will be loaded on a new page; select two or more biosets from the same experiment and click the “View Bioset Patterns” button above the table to calculate the corresponding patterns; you will see a Bioset Pattern page showing a summary table of differential expression patterns for your selected biosets; clicking the hyperlinked number in the right column of any row will display the list of genes that have the expression pattern shown on that row.

2. Identify shared, unique or combined lists of differentially expressed host genes across experiments

IRD allows you to compare significant host factor lists from multiple biosets/experiments for customized data-mining analyses. This facilitates the identification of subsets of differentially expressed genes that are shared or unique under different experimental conditions or across different experiments.

a. On the host factor data landing page, click next to one or more studies to show the experiments associated with the study, then check the box next to one or more experiment names and click the “View Associated Biosets” button to load the Host Factor Bioset Information page.

b. On the next page, select two desired biosets, mouse-over the “Run Analysis” button located above the bioset list table and click one of the Boolean analysis options: “Find shared factors”, “Find all factors”, or “Find unique factors”. This will generate a list of genes found to be significantly up- or down-regulated across your selected experiment(s).
### Boolean Analysis Result for Derived Bioset

#### Study Name
- IM001:A/Vietnam/1203/2004(H5N1) infection in C57BL6 mice with variable doses and times post infection.

#### IM004:VN1203 HA avirulent mutant virus (A/Vietnam/1203-CIP048_RG1/2004 (H5N1)) infection in C57BL6 mice: A time course

#### Find data for
- PROBE ID OR GENE SYMBOL

Use comma to separate multiple entries. Ex. DDX58

Save search to search multiple times.

Your search returned 993 records.

Data sorted by Gene Symbol, Probe ID descending.

<table>
<thead>
<tr>
<th>Row Number</th>
<th>Bioset Name</th>
<th>Probe ID</th>
<th>Entrez Gene ID</th>
<th>Gene Symbol</th>
<th>Genbank Accession</th>
<th>Gene Name</th>
<th>Immport Link</th>
<th>Log2 FC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IM004_2day_H5N1</td>
<td>A_51_P479352</td>
<td>68874</td>
<td>1190002J23Rik</td>
<td>NM_001033039</td>
<td>RIKEN cDNA 1190002J23 gene</td>
<td>1.7</td>
<td>6.4E-4</td>
</tr>
<tr>
<td>2</td>
<td>IM004_2day_H5N1</td>
<td>A_51_P170725</td>
<td>74152</td>
<td>1300002K09Rik</td>
<td>NM_028788</td>
<td>RIKEN cDNA 1300002K09 gene</td>
<td>2.8</td>
<td>4.1E-4</td>
</tr>
<tr>
<td>3</td>
<td>IM004_2day_H5N1</td>
<td>A_51_P160083</td>
<td>70258</td>
<td>1500035N22Rik</td>
<td>AK005363</td>
<td>RIKEN cDNA 1500035N22 gene</td>
<td>3.3</td>
<td>3.3E-5</td>
</tr>
<tr>
<td>4</td>
<td>IM004_2day_H5N1</td>
<td>A_52_P67678</td>
<td>74152</td>
<td>1300002K09Rik</td>
<td>NM_028788</td>
<td>RIKEN cDNA 1300002K09 gene</td>
<td>2.6</td>
<td>3.6E-5</td>
</tr>
<tr>
<td>5</td>
<td>IM004_2day_H5N1</td>
<td>A_51_P199552</td>
<td>72244</td>
<td>1600014C10Rik</td>
<td>NM_028166</td>
<td>RIKEN cDNA 1600014C10 gene</td>
<td>2.3</td>
<td>2.5E-5</td>
</tr>
<tr>
<td>6</td>
<td>IM004_2day_H5N1</td>
<td>A_52_P148184</td>
<td>72244</td>
<td>1600014C10Rik</td>
<td>NM_028166</td>
<td>RIKEN cDNA 1600014C10 gene</td>
<td>2.6</td>
<td>1.2E-4</td>
</tr>
<tr>
<td>7</td>
<td>IM004_2day_H5N1</td>
<td>A_52_P77093</td>
<td>72244</td>
<td>1600014C10Rik</td>
<td>NM_028166</td>
<td>RIKEN cDNA 1600014C10 gene</td>
<td>1.8</td>
<td>5.0E-4</td>
</tr>
</tbody>
</table>

**SET TYPE KEY:**
- **Intersect (Boolean AND)**
- **Union (Boolean OR)**
- **XOR (Boolean XOR)**
Exercise VI. Accessing immunology and serology data

A. Serology

IRD is proud to host serology data that were collected from avian, non-human mammalian, and human subjects and tested for the presence of antibodies to specific influenza serotypes. This data was captured as part of larger studies and is freely accessible in IRD.

1. To access this component in IRD, mouse-over the “Search Data” item in the top menu bar and click on the “Serology Experiments (beta)” option.
   a. Click a hyperlinked number in either the “Number of Serum Samples” or “Positive Samples” column, representing a specific dataset. For this exercise, we will use the Avian data, collected by National Taiwan University, and submitted to IRD by the IPIRC.
   
   ![Serology Experiments (Beta) table]

   b. Alternatively, click on the “Go to Serology Experiment Search Page” button and enter in the desired search criteria.

2. From the search results page, click the blue “i” next to any record to view the Experiment Details page. This will display information about the collection and host as well as the sample test (test type and result) data. IRD also lets you see where the sample was collected by clicking on the “View on Map” button.

3. Any records that are selected can be saved to a Working Set within the IRD Workbench feature.
B. Laboratory Experiment Data (through ImmPort)

IRD is working with data submitters to provide access to their experimental results through the Immunology Database and Analysis Portal (ImmPort, immport.niaid.nih.gov) system.

1. To begin, select the “Laboratory Experiments (beta)” from the “Search Data” menu. Click the radio button next to the “Laboratory studies/experiments deposited in the ImmPort System” and any other search criteria, then click “Search”.

![Image of Experiment Search]

2. For this workflow, click on the “FCM analysis of IgA, IgG and IgM plasmablasts (FCM)” experiment (under the “Systems Biology of Seasonal Influenza Vaccination in Humans” study title).
   
   a. The first time you access ImmPort, you will be required to signup for a username and password since these data are currently semi-public. (More information is available by following the “Instructions for obtaining access to the ImmPort system” link.)
   
   b. On subsequent visits, click on the “View detailed data for Experiment in ImmPort (ImmPort login required)” link to view the desired data.

3. You are now on the Study Detail page in the ImmPort system.
   
   a. This page contains information about the demographics, subjects, assays and documentation associated with the study.

   ![Image of Study Detail]

   b. In the “Mechanistic Assays” dropdown, select the checkbox next to “EXP10602” and then click the “View Details” button at the top of the panel.
4. You are now on the Experiment Details page that contains information about a flow cytometry experiment examining IgA, IgG and IgM plasmablasts.

   a. Expand the blue menu bars, select the checkbox next to the desired item and click either “Download Results” or “View Details” to get more information about any experimental sample, protocol, subject, biological sample, reagent, result file, etc.

   b. In the “Result Files” dropdown, select the “FC 016 Day7.001.fcs” file and then click the “Download Results” button.

5. To analyze the flow data using FLOw cytometry Clustering without K (FLOCK), mouse-over the “Tools” item in the gray menu bar at the top of the page and navigate to the “Analyze Individual Files” under the FLOCK heading.

   a. Select the checkbox to “Include Semi-Public files in the Search”.

   b. Select the checkbox next to the “FC 016 Day7.001” item, then click “Run FLOCK” from panel menu bar.
c. Set the desired parameters, including FLOCK Version 2.0 (and a job name) and click the “Run FLOCK” button at the bottom of the page.

d. After the analysis finishes, click on the “Results” link in the “Status” column to view the color-coded cell populations.

i. Results can be adjusted, and cross-sample comparisons can also be done using tools available in ImmPort.
Exercise VII. Submitting or annotating your own virus genome sequences

IRD has a unique sequence curation/annotation pipeline developed by Dr. Catherine Macken’s group at Los Alamos National Laboratory, which can determine the input sequence’s influenza type, segment number, subtype for segments 4 and 6 of type A, and translate the input nucleotide sequence using a translator specific for influenza sequences.

A. Annotating an influenza virus segment sequence

After this exercise you should be able to use the annotation pipelines provided by the Influenza Research Database (IRD) to annotate your own segment sequences.

For this exercise, you will use IRD (http://www.fludb.org) to annotate an influenza virus segment sequence.

a. From the grey navigation bar, mouse over “Analyze & Visualize” and click “Annotate Nucleotide Sequences”.

b. If you have your own sequence, prepare the sequence in FASTA format, save it in plain text and use .fasta as the file extension. FASTA file example:

```
>gb:GQ293081|Organism:Influenza A virus A/Perth/16/2009|Segment:4|Subtype:H3N2|
Host:Human
AAAGCAGGGGATAATTCTATTAACCATGAAGACTATCATTGCTTTGAGCTACATTCTATGTCTGGTTTTCGCTCAAAAAC
TTCCTGGAAATGACAACAGCACGGCAACGC
```

Otherwise, use a sample sequence from: http://tinyurl.com/6v9hdks

c. Either paste your sequence in FASTA format to the sequence box or upload your FASTA sequence. Provide your email address so that IRD can contact you if there are problems in the annotation process. Click “Validate Sequence(s)” to start the annotation process.
d. After the annotation process is finished, a Sequence Annotation Result page will be loaded. Here you will see flu type, segment number, subtype (if you provided HA or NA sequences), and translated protein sequence. You can also download the annotation report by clicking the “Annotation Report” button.

B. Submitting influenza virus segment sequences

After this exercise you should be able to use the data submission tool provided by the Influenza Research Database (IRD) to submit sequences to GenBank via IRD.

IRD supports the submission of four data types:

- Influenza virus sequences
- Experimental data on virus phenotypic characteristics, such as pathogenicity, infectivity, transmissibility, and drug resistance
- Surveillance (avian, non-human mammalian, and human) and influenza virus testing data
- Sequence Features

The benefits of submitting sequence data to IRD include:

- IRD facilitates influenza virus sequence submission to GenBank using our sequence analysis pipeline for automatically defining segment identity, coding region location, subtype designation, etc.
- IRD captures experimental virus phenotypic characteristics (such as pathogenicity, infectivity, transmissibility, and drug resistance), surveillance data, Sequence Features, and other sequence metadata (descriptive information about sequences) that are not supported by GenBank. If you submit these types of data and metadata, IRD will integrate them together and provide tools to integrated analysis of phenotypic characteristics with geospatial surveillance data and detailed genomic and proteomic data. If you have publications...
associated with experimental data, Sequence Features, etc., IRD will not only integrate these types of data with sequence data, but also link your data to publications, thus helping you expose your data and publications to a wider scientific community.

Sequence submission options:

- Single sequence submission: Submit sequence data for wild type or genetically manipulated viruses (e.g. reverse genetic, animal adapted, reassortants, etc.) to IRD. All segment sequences for a single strain can be submitted at one time.
- Multiple sequence submission: You can submit up to 2500 sequences for up to 2500 wild-type influenza strains (i.e. one sequence/strain) simultaneously, which will be curated and annotated to facilitate their submission to GenBank using our sequence analysis pipeline for automatically defining segment identity, coding region location, subtype designation, etc.

Sequence submission process:

1. Data preparation
   a. Sequence data
      Prepare your sequences in a single FASTA format file. Please remove primer/vector sequence from 5' and/or 3' end of influenza virus sequence.
      - For single sequence submission, if you have more than one segment sequences for a single strain, include all segment sequences (up to 8) in a FASTA file.
      - For multiple sequence submission, the defline of the FASTA sequence file should be in the format of "->Unique_Sample_Identifier: yourSampleID|Unique_Sequence_Identifier: yourSampleID-segmentNumber".
   b. Descriptive information associated with sequence data
      IRD’s sequence submission utility will collect information about the submitter, citation, and strain/sample required by GenBank along with additional descriptive information captured by IRD. Download the sequence metadata template (http://www.fludb.org/brcDocs/documents/Batch_test_metadata_6_samples-B2.xls) and view the metadata fields needed for sequence data submission (sheet 1 – Sample data) along with descriptions of the metadata fields (sheet 2 – Field definitions). A portion of the template is illustrated below.

<table>
<thead>
<tr>
<th>Sample Identifier</th>
<th>Sample Description</th>
<th>Strain Name</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edtest11</td>
<td>Test description 1</td>
<td>A/chicken/New York/B32/63/1989(H1N1) chicken</td>
<td>H1N1</td>
<td>1989</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edtest12</td>
<td>Test description 2</td>
<td>A/mallard duck/Alberta/155/1990(H1N1) mallard duck</td>
<td>H1N1</td>
<td>1990</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edtest13</td>
<td>Test description 3</td>
<td>A/mallard duck/Alberta/191/1990(H1N1) mallard duck</td>
<td>H1N1</td>
<td>1990</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edtest14</td>
<td>Test description 4</td>
<td>A/mallard duck/Alberta/225/1999(H1N2) mallard duck</td>
<td>H1N2</td>
<td>1990</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edtest15</td>
<td>Test description 5</td>
<td>A/mallard duck/Alberta/276/1985(H1N3) mallard duck</td>
<td>H1N3</td>
<td>1985</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Single sequence submission: You will follow the web submission forms to submit sequence and sequence descriptive information, so there is no need to fill in this template. This template shows you what descriptive information is needed.
- Multiple sequence submission: Fill out the sequence metadata spreadsheet using the following template with one row for each sample/strain in the FASTA file. Sample Identifiers must be listed in the order in which the respective samples appear in the
FASTA file. Required fields are marked in bold font. Please DO NOT CHANGE THE ORDER OF THE COLUMNS IN THE SPREADSHEET!

2. Data submission

a. Go to the IRD site (www.fludb.org) and click the “Submit Data” tab. The Submit Data to IRD landing page will appear.

b. Select the “Influenza virus sequences” radio button. Two sequence submission options will appear. Select the appropriate option and click “Continue Submission” to proceed.

c. On the following page, indicate whether you have submitted data previously. If not, you will need to enter your personal and citation information.

d. On the GenBank Submission – Step 2 page, upload your sequences either by pasting in your sequence in FASTA format, or by uploading a FASTA-formatted file on your machine. Click the button “Validate Sequence(s)”.

e. The input sequences will be processed through a curation/annotation pipeline.
   • Single sequence submission: Pipeline analysis results will be returned to you for review and comment before continuing submission.
• Batch sequence submission: A batch number will be assigned. The pipeline analysis results can be retrieved for up to one week via the “Retrieve Previous Batch Submission” option from the “Submit Data” menu by using either your email address or the ticket number.

f. You can review the results of the pipeline and decide which sequences to continue submitting. The report flags submitted sequences that fail the pipeline and permits deselecting them from further submission. When the report is approved, click “Continue Submission”.

g. On the GenBank Submission – Step 3 page:

• Single sequence submission: Complete all required data fields marked with red asterisks and as many optional fields in the “Additional Information” section. This additional information will not be submitted to GenBank, but will be stored and made viewable and searchable in IRD.

• Batch sequence submission: Submit descriptive information associated with sequence data by browsing to the metadata spreadsheet you prepared. A link on this page can create a spreadsheet for you, with the Sample Identifier field pre-populated from the FASTA file. There is no need to adjust the spreadsheet if one or more sequences failed the pipeline and was deselected.
h. Click “Continue Submission”. If the page refreshes and there are no issues, click “Continue Submission”.

- Single sequence submission: This completes the submission and assigns IRDACCESSION_NUMBERS to the individual sequences.
- Batch sequence submission: The final Batch Submission Result page summarizes the data submission and assigns IRD Accession numbers. Note that the failed, deselected sequence shown in the previous screenshot does not appear on the summary. If at a later date you correct any failed sequences and resubmit them, use the unique sample and sequence identifiers you recorded from the FASTA file and spreadsheet of this first submission.

3. Post data submission

a. Following submission to IRD, all sequences and strain data are reviewed by an IRD scientist. If any issues arise during this process or the data format needs to be altered to satisfy GenBank requirements, submitter will be contacted by email. After all issues are resolved, sequences are forwarded to GenBank. According to GenBank policy, if no significant issues are identified, GenBank accessions will be provided to the submitter within 48 hours. Release of these sequences in the public NCBI database can take as much as four weeks. Large batches may take even longer.

b. When your sequence data and associated metadata become available in GenBank, they will be imported from GenBank to IRD and made accessible on the IRD site. If you have submitted additional metadata that is not captured by GenBank, such as sampling location, gender of host, age of host, etc., the metadata will also be available for query in IRD.
Exercise VIII. Analyzing Sequence Feature Variant Types

A. Analyzing Sequence Feature Variant Types

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

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.

- **Sequence Features (SFs):** specific regions defined based on functional properties, structural properties, and immune epitope locations. Obtained from literature and/or imported from other databases and validated by domain experts.
- **Variant Types (VTs):** polymorphisms in each Sequence Feature are identified as Variant Types of the Sequence Feature (SFVT).
- The SFVT analysis is available for influenza A virus in the Influenza Research Database and Dengue virus serotypes 1-4, Hepatitis C subtype 1a and Vaccinia viruses in the Virus Pathogen Resource.

For this exercise, use the Influenza Research Database to examine a protein region required for nuclear export of the NS1 protein and sequence polymorphism in this region.

a. Go to the IRD homepage (http://www.fludb.org). From the grey navigation bar, mouse over “Search Data” and click “Sequence Feature Variant Types”.

b. The SFVT landing page will be loaded. Here you can search for Sequence Features or click “Go to Sequence Feature List” to browse all Sequence Features of influenza A proteins.

c. Now search for functional Sequence Features of the NS1 protein. Select “8 NS1” from the segments list and “Functional” from the Sequence Feature Type list. Then click “Search”.

d. The Sequence Feature search results will be displayed in a table.
i. How many functional Sequence Features are there for the NS1 protein?

ii. Can you find the Sequence Feature that is involved in the nuclear export of NS1 protein? Write down the name of the Sequence Feature.

e. Click for the nuclear export signal Sequence Feature to view its details.

i. Which is the name of the reference strain for this Sequence Feature (i.e., defines Variant Type-1)?

ii. What are the sources of information that support its functional classification?

iii. How many variant types exist for this Sequence Feature in IRD?

iv. Click the number in the “Strain Count” column for VT-1 to view detailed metadata about the strains harboring this Variant Type. Click the “Host” column heading to sort the list by host. Are the strains from the same host?

v. Return to the previous page by clicking the breadcrumb. Click the “Strain Count” number for VT-8 to view the strains harboring this Variant Type. Sort by host again. What do these strains have in common?

vi. Return to the previous page by clicking the breadcrumb. Look at the variant type table. Note that VT-8 has E at 139, which may be related to host range.
vii. Now search for Variant Types with E at position 139. Click the blue “Find a VT(s)” button to expand the VT search panel. Click “Fill wildcards” to have “?” populated in all positions and then change position 139 to “E” to search for variant types harboring 139E, i.e., amino acids 137-147 ??E?????????. Click “Search”.

viii. How many variant types did you find?

ix. Examine the hosts for VT-71, VT-72, and VT-73. Do they have the same host range as VT-8?

B. Submitting new Sequence Features to IRD

In an effort to keep the Sequence Feature (SF) list up-to-date and to facilitate greater community involvement in annotation efforts, we encourage you to submit new SFs to IRD.

The SF submission page is accessible from the “Submit Data” tab.

- This tool provides a user-friendly interface for online submission of new SFs. Drop-down menus are made available for fields such as virus type, protein, SF category, etc. to maximize the efficiency.
- Newly submitted SFs are first manually inspected and then validated using appropriate Quality Control measures to ensure accuracy and authenticity, and to avoid duplicates.
- Following Quality Control process, the new SFs are either accepted or rejected and the contributors are notified of the decisions. Submitters of accepted SFs are acknowledged on the IRD site.