Section A. Sequence search and sequence annotation exploration

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/ViPR.

I. Search for influenza virus segment sequences in IRD

1. Search for sequences using structured search interfaces

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

a. Go to the IRD homepage (http://www.fludb.org/), mouse-over “Search Data” in the gray navigation bar, then “Search Sequences” 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. Segment/nucleotide and influenza A are pre-selected by default. 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 an instant count of search results above the search criteria to help you search quickly and efficiently. When you select search criteria on search pages, 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 the radio button next to “Include only pH1N1 sequences” below “2009 pH1N1 Sequences”. Click “Search” to run the search.

![Search Page Screenshot]

[Diagram of the search page with highlighted search criteria]

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.

---

Created by the ViPR/IRD team and licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License
f. On the Search Results page, click on “View” 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?

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.

g. Click the “Results” breadcrumb at the top of the page to return to the results page. Here you can:

   i. select records and run an analysis on the selected records by mousing-over the “Run Analysis” button and clicking a desired analysis option.
ii. store selected sequences as a working set in the Workbench so that you can run various analyses on the working set.

iii. download sequences in various formats by selecting sequences and then clicking the “Download” button.

iv. save the search query to your Workbench by clicking the “Save Search” button. A Save Searches lightbox will pop up. You can choose to receive email alerts on newly released data that match your search criteria by checking the “Subscribe” box.

2. 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.
4. 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.

3. Identify short peptides in proteins

IRD provides a Short Peptide Search tool for finding short amino acid strings such as epitopes, ligand binding sites, sequence domains, etc. in target influenza proteins.

a. Go to the IRD site (http://www.fludb.org), mouse-over “Analyze & Visualize” in the gray navigation bar and click “Identify Short Peptides in Proteins” to load the search page.

b. Three search types are provided:

• EXACT Match finds only exact matches of the query string;
• FUZZY Match finds amino acid strings with over 50% identity to the query string;
• PATTERN Match finds amino acid strings with 100% pairwise identity to the input pattern, e.g. [LIT]-x(3)-[LT]-[NQK]-x-{G}.

For this exercise, we will search for proteins containing an epitope sequence ATVAGSL, so select the EXACT Match option and type ATVAGSL in the “String to Find” box.

c. Input sequences from one of the following ways:

• Upload a FASTA-formatted sequence file
• Paste FASTA-formatted sequences
• Use a working set consisting of a group of sequences saved to your Workbench
• Select a pre-compiled IRD database

For this example, select “Influenza HA Hemagglutinin proteins”.

d. After you have selected your criteria, click “Run” to perform the query.

e. On the results page, all HA proteins harboring the input string are listed in the table. Click “View” to view the hit’s segment/protein details page in IRD.

---

Created by the ViPR/IRD team and licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License
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.

4. Search for immune epitopes

IRD imports experimentally determined immune epitope data from the Immune Epitope Database (IEDB, http://www.immuneepitope.org) and integrates such data with existing protein sequence data. IRD also generates predicted MHC Class I epitopes from influenza amino acid sequences using the NetCTL algorithm.

a. To search for immune epitopes in IRD, mouse-over the “Search Data” tab in the grey navigation bar and click on the “Immune Epitopes” option.

b. On the Immune Epitope Search page, you can search for experimentally determined epitopes or predicted NetCTL epitopes. For this exercise, we will search for experimentally determined epitopes, so make sure the “Experimentally Determined Epitopes” option is selected.

c. Next, you have the options of searching by subtype, strain name, protein, epitope sequence, IEDB ID, other public database ID, assay type, assay result, host, and more. Now we will search for
positive B cell epitopes in the HA protein: choose “HA” in the Protein(s) box; select B-cell “Positive” in the Immune Recognition Context section; click “Search” to run the search.

d. The search results page returns a list of epitopes matching your search criteria. Click “View” next to the third record (IEDB ID: 2790) to view the Epitope Details page.

e. This will display information about the epitope, a summary of assay results as well as protein structure(s) containing the epitope. Was this epitope also tested by T cell assay?

5. 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 gray navigation bar and click “Analyze Sequence Variation (SNP)”.

b. On the Analyze Sequence Variation (SNP) 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.

II. Search for Hepatitis C virus sequences in ViPR

For this exercise, you will search for full-length HCV genome sequences using keyword search. Note that ViPR also provides “BLAST” and “Identify Short Peptides in Proteins” for sequence-based sequence search.

a. From the ViPR homepage (http://www.viprbrc.org/), click “Flaviviridae” to get to the family page.

b. Mouse-over the “Search Data” tab and click “Genomes” to load the Genome Search page.
c. You will notice you have many options to search: virus taxonomy, collection year, sample location, host, genome sequence length, etc. Select the following search criteria and click the orange “Search” button to execute the search. As you add search criteria, you will notice the number of matching hits above the light blue search box dynamically change.

From within the taxonomy tree, click the plus sign next to “Hepacivirus”, next “Hepatitis C virus”, then “Select All” next to “Genotype 2”. Please note that once Hepatitis C virus is selected, an HCV Virus Metadata panel will be displayed. Only some HCV sequence records have associated clinical metadata, so queries based on these metadata fields only retrieve sequences for which those fields are defined.

Complete Genome: [ ] Complete Genome Only [ ] Geographic Grouping: [ ] Asia
Advanced Options: [ ] Remove Duplicate Sequences [ ] Country: [ ] China

![ViPR GUI screenshot](image-url)

**Results matching your criteria:**

**SELECT VIRUSES TO INCLUDE IN SEARCH**

<table>
<thead>
<tr>
<th>Complete Genome Only</th>
</tr>
</thead>
</table>

**COLLECTION YEAR**

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
</tr>
</thead>
</table>

**GENETIC GROUPING**

<table>
<thead>
<tr>
<th>Country</th>
</tr>
</thead>
</table>

**SEARCH WITH HCV VIRUS METADATA**

**HOST Attributes**

<table>
<thead>
<tr>
<th>Host Name</th>
<th>Organism</th>
<th>Country</th>
</tr>
</thead>
</table>

**VIRUS Attributes**

<table>
<thead>
<tr>
<th>Virus Type</th>
</tr>
</thead>
</table>

**Displaying 90 records per page, sorted by Species Name, Strain Name, GenBank Accession in ascending order.**

**Search Result**

<table>
<thead>
<tr>
<th>Species Name</th>
<th>GenBank Accession</th>
<th>Sequence Length</th>
<th>Collection Date</th>
<th>Host</th>
<th>GenBank Host</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hepatitis C virus</td>
<td>KF673531</td>
<td>9667</td>
<td>-9A-</td>
<td>Human</td>
<td>Homo sapiens</td>
<td>China</td>
</tr>
<tr>
<td>Hepatitis C virus</td>
<td>KF673532</td>
<td>9665</td>
<td>-9A-</td>
<td>Unknown</td>
<td>-9A-</td>
<td>China</td>
</tr>
<tr>
<td>Hepatitis C virus</td>
<td>Hq838343</td>
<td>1015</td>
<td>-9A-</td>
<td>Human</td>
<td>Homo sapiens</td>
<td>China</td>
</tr>
</tbody>
</table>

**d.** 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.
e. On the Search Result page, click “View” next to a strain of interest to view the Strain Details.

i. What’s the name of the strain? What subtype is it?

ii. In the Genome section, right-click “View Nucleotide Sequence” to open the sequence in a new window. Click the “FASTA Download” button above the sequence to download the sequence file to your computer.

Strain Details for Hepatitis C virus Strain PR63-2a

i. What’s the name of the strain? What subtype is it?

ii. In the Genome section, right-click “View Nucleotide Sequence” to open the sequence in a new window. Click the “FASTA Download” button above the sequence to download the sequence file to your computer.

iii. View the genome image map. Click “E2” in the image map or “View” next to E2 in the Protein Information table to display the details of the protein. Look at the Genomic Annotation section. How long is the CDS?

iv. Click the epitope count in the Protein Sequence Features section. How many epitopes have been found on E2?
III. Search for Human Herpesvirus 1 sequences in ViPR

1. Search for glycoprotein D protein sequences in Human Herpesvirus 1
   a. From the ViPR homepage ([http://www.viprbrc.org/](http://www.viprbrc.org/)), click “*Herpesviridae*” to get to the family homepage.

   b. Mouse-over “Search Data” in the grey navigation bar and click “Genes & Proteins”.

   c. The Gene/Protein Search page allows you to search for sequences based on taxonomy, collection year, sample location, host selection, gene symbol, gene product name, genomic location, ortholog group, etc. A dynamic number of matching search results is displayed at the top of the page to help you search more efficiently. Select the following criteria and click the orange “Search” button to run the query.

   **Taxonomy**: Human herpesvirus 1 (Herpesviridae -> Alphaherpesvirinae -> Simplexvirus -> Human Herpesvirus 1).
   **Complete Genome Only**: uncheck
   **Gene Symbol**: US6

   d. The Search Results page will be displayed. How many sequences did you find?

   e. Now try a search based on ViPR ortholog clusters. Either:
      - Return to the Search page by clicking “Gene/Protein Search” in the breadcrumb, OR
      - Click the “Search Criteria” button located at the top of the search results table.

   f. Select the following criteria and click the orange “Search” button to run the query.
      **Taxonomy**: Human herpesvirus 1 (Herpesviridae -> Alphaherpesvirinae -> Simplexvirus -> Human Herpesvirus 1).
      **Complete Genome Only**: uncheck
      **Ortholog Group**: US6
The Search Results page will be displayed. How many sequences did you get this time?

Note that some records do not have Gene Symbol/Gene Product Name assigned. Searching by ViPR-clustered Ortholog Group retrieves all US6 orthologs regardless of the presence of a standard Gene Symbol/Gene Product Name.

On the Search Results page, you can:

i. Save the search query to your Workbench and rerun the search again later.

ii. Download the sequences (CDS, protein) by clicking “Download”.

iii. Select records and run an analysis on the selected records by mousing-over the “Run Analysis” button and clicking a desired analysis option.

iv. Store selected sequences as a working set in the Workbench so that you can run various analyses on the working set.

v. View the details of a gene/protein. In this exercise, find Strain F, UniProt Accession number D3YPD0; then click “View” for this record.

vi. On the Gene/Protein Details page, you will find strain information, protein information, genomic annotation, isoelectric point/molecular weight, HMM/P/am domains, related protein structures, predicted and experimentally determined immune epitopes, GO annotations, nearest blastp hits, etc.
2. View Human Herpesvirus 1 genome annotations in GBrowse

a. From the previous Gene/Protein Details page, scroll up and find the Strain Information section. Then click “View Strain Details” in the Strain Name row.

b. The Strain Details page will be displayed.

c. Strain F has multiple genome sequences and all of the information has been integrated on one page. In the GU7347711 section, click “View in GBrowse” to display the genome browser.

d. The GBrowse window will be displayed as shown below. The “Overview” panel displays the entire genome. The “Region” panel displays a portion of the genome surrounding the region of interest. The “Details” panel displays several tracks with each track corresponding to a different type of genomic feature.

e. In the “Protein-coding genes” track, hover over a gene to view its brief description. Then click the gene to view its details. The gene/protein details page will be open in a new window.

f. Enter coordinate numbers (GU7347711:136,000..140,000) in the “Landmark or Region” box and click “Search” to view this region. What genes are in this region?

g. In addition to the ViPR annotation tracks, you can display additional annotations on the genome by uploading custom tracks. The file format of a custom track can be: BED, Feature File Format, GFF, GFF3, Wiggle, BAM or SAM. Detailed information about the accepted file formats is available at: [http://www.viprbrc.org/gbrowse2/annotation_help.html](http://www.viprbrc.org/gbrowse2/annotation_help.html).

i. For this exercise, we are going to upload a polymorphism track of gD protein based on the Sequence Variation Analysis result. Download the track from: [http://tinyurl.com/lrd7xz](http://tinyurl.com/lrd7xz)

ii. Click the “Custom Tracks” tab. Then click Add custom tracks “From a file”, find the track you just downloaded and click “Upload”.

iii. Now click the “Browser” tab, the uploaded track is shown in the “Details” panel.
h. Zoom into a 2 kbp region by selecting “Show 2 kbp” from the Scroll/Zoom drop-down menu.

i. Scroll left or right using the < or > buttons.

j. Flip the orientation by selecting the “Flip” checkbox such that the minus strand points to the right. This is sometimes useful for looking at minus strand genes.

k. Click the ruler to show the current base position.

l. Review the SNP:gD track together with the Protein-coding genes track and the IEDB epitope track. Zoom in to region 139150-139220. Can you find B-cell epitopes on the glycoprotein D protein that are conserved across strains?

m. Find and click epitope 19577 located around 138800. This will open the Epitope and Protein Affiliation Details page. Write down its amino acid positions on the gD protein ADD60053.

n. You can export the genome annotations/images in multiple formats: PNG image, SVG image, GFF annotation table, or FASTA sequence file. To do so, mouse-over the “File” drop-down menu, then the “Export as …” item and click a desired format to download the region you are looking at with all the features that are displayed in the GBrowse window.