IRD and ViPR Hands-on Workshop

JCVI-GSCID/NIAID Workshop
Empowering Genomics in Southern Africa
Application to Infectious Disease

IRD and ViPR

Prepared by: Richard Scheuermann
Bioinformatics Research Group
University of Texas Southwestern Medical Center at Dallas

University of Limpopo, Turfloop Campus, South Africa
May 30th – June 1st, 2011

# Table of Contents

## Part I. Influenza Research Database (IRD) Exercises ................................................. 1
1. Search and View Sequences .................................................................................. 1
2. Comparative Genomics Analysis ........................................................................ 2
3. Sequence Feature Variant Type Analysis ........................................................... 5
4. 3D Protein Structure Visualization ..................................................................... 6
5. Search and Visualize Avian Surveillance Data .................................................... 8

## Part II. Virus Pathogen Resource (ViPR) Exercises .................................................. 10
1. View Basic Virus Information ............................................................................. 10
2. Multiple Sequence Alignment and Phylogenetic Trees ...................................... 10
3. Workbench .......................................................................................................... 11
4. Metadata-driven Comparative Genomics ............................................................ 12

## Part III. Scientific Use Cases .................................................................................... 14
A. Identification of candidate adaptive drivers of species jumps in avian H5N1 viruses ........ 14
B. Computational exploration of sequence relationships and variation in Bunyaviridae ....... 16

## Part IV. Answer Key .................................................................................................... 17

## Part V. Bibliography .................................................................................................. 17

## Part VI. Appendixes
PART I. INFLUENZA RESEARCH DATABASE (IRD) EXERCISES

1. Search and View Sequences

Upon completion of this exercise, you will be able to search for influenza genome segment sequences and view detailed information about these sequences in IRD.

a. Go to [http://www.fludb.org/](http://www.fludb.org/) using any Internet browser (e.g. Firefox, Safari, Internet Explorer).

b. From the grey navigation bar, mouse-over “Search Data” and click “Nucleotide Sequences”.

c. Search for influenza A NA nucleotide sequences from human host in Africa. Choose the following parameters on the search page and click the orange “Search” button to find the sequences.

   Virus Type: A
   Host: Human
   Geographic Location: Africa

   Note: You can also search for proteins or strains from this page as well.

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

   i. What’s the name of the strain?
   ii. How long is the segment?
   iii. Click on “View Sequence” to retrieve the sequence.
   iv. Click on the “Segment Details” breadcrumb at the top of the page to navigate back to the Segment Details page. Can you find the HMM/Pfam domain information?
   v. Click on “Prediction Details” in the “Predicted Epitopes” section. What is the sequence and Supertype of the first epitope listed?

   ![Prediction Details](image)

   Predicted Epitopes (SCD) Prediction Details

<table>
<thead>
<tr>
<th>Supertype</th>
<th># Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3</td>
<td>16</td>
</tr>
<tr>
<td>A2</td>
<td>10</td>
</tr>
<tr>
<td>A24</td>
<td>25</td>
</tr>
<tr>
<td>B7</td>
<td>10</td>
</tr>
<tr>
<td>B44</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
</tr>
</tbody>
</table>

   e. Click on the “Results” breadcrumb at the top of the page to return to the search result page. Click on the blue “Search Criteria” button to open the Search Criteria window and revise your search. Now, in the “Country” section, select “South Africa” and click on the “Search” button to search for sequences from South Africa. How many sequences are there from South Africa?
2. Comparative Genomics Analysis

Upon completion of this exercise, you will be able to perform a multiple sequence alignment and phylogenetic tree construction on a select set of sequence records to infer their evolutionary relationships, and determine the level of sequence conservation at each nucleotide and amino acid position.

a. From Exercise 1e, select all NA segments from South Africa, and then perform a multiple sequence alignment (MSA).

   i. Mouse-over “Run Analysis” and click on the orange “Align Sequences (MSA)” button.
   ii. Click “Continue” using default “Nucleic Acid (Segment)” option.
   iii. Click “Run” on the Align Sequences (MSA) screen.
   iv. Once the analysis finishes, mouse-over the blue “Run Analysis” button and click on “Visualize Aligned Sequences”.

b. Customize sequence labels in your alignment.

   Option 1: On the Visualize Aligned Sequences Customization page, you can customize the contents of a strain name in the alignment by selecting “Custom” and selecting the labels of interest (e.g. Strain Name and Subtype).

   Option 2: 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”.

---

2 Influenza Research Database (IRD, www.fludb.org) is funded by the U.S. National Institute of Allergy and Infectious Diseases (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. 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.
c. 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”.

d. Colour 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.

e. Create a phylogenetic tree from the aligned sequences using the “Quick Tree” model.

On the Visualize Aligned Sequences page, click on the blue “Generate Phylogenetic Tree” button.
On the next page, select “Quick Tree” and click on “Build Tree”. The analysis will take a few seconds to minutes. After the analysis is completed, click on “View Tree” to open the tree window.

f. In the pop-up tree window, color tree leaves by NA subtype. 
In the “Tree Decorations” section, click on the drop-down menu and choose “A/NA subtype”.

g. Now, color tree leaves by Year.
In the “Tree Decorations” section, click on the drop-down menu and choose “Year”. Also, click on “Show Legend” option.

h. Run tree model compare to determine which model fits your data best. 
Return to the Generate Phylogenetic Tree page using the “breadcrumbs” feature, and 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?
3. Sequence Feature Variant Type Analysis

At the end of this exercise you will be able to identify regions of influenza 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.

- 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. From the grey navigation bar, mouse over “Search Data” and click on “Sequence Feature Variant Types”.

b. On the next page, review the description of Sequence Feature Variant Type. After you are finished, click on “Go to Sequence Feature List” to view all Sequence Features of influenza A proteins.

c. Look at the Sequence Feature list of influenza A. How many Sequence Features are there for the NS1 protein?

d. Click on the number in the “Functional SF” category for NS1 to view all functional Sequence Features of NS1.

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

f. Click on \( \text{to view details about this Sequence Feature. What are the sources of information that support its functional classification? In which strain was this sequence feature defined?} \)

g. How many variant types exist for this Sequence Feature in IRD? Which reference strain defines Variant Type – 1 (VT-1)?

h. Click on the number in the “Strain Count” column for VT-1 to view detailed metadata about the strains harboring this Variant Type. How do you think the strains are ordered in the table? Select the “Date” column heading to re-order table by date of isolation. In what year was the earliest strain isolated?
i. Click on the number in the “Strain Count” column for VT-8 to view detailed metadata about the strains harboring this Variant Type. What do these strains have in common? 
   *Note: You can use the toolbar to download the full data set.*

![Variant Types Table]

j. Return to the previous page. Click on the blue “Find a VT(s)” button to expand the VT search panel. Now search for Variant Types with a single mutation at I145. Change the I145 to “?” to search for any residue at this position, i.e., amino acids 137-147 IFDRLETL?LL

How many Variant Types do you find that differ from the reference strain at only amino acid position 145 of NS1?

4. 3D Protein Structure Visualization

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

   a. From the grey navigation bar, mouse over “Search Data” and click on “3D Protein Structures”.

   b. Search for the 3D structures of influenza A (H3N2) HA protein.
      
      *Virus Type: H3N2*  
      *Subtype: H3N2*  
      *Select Proteins to search: 4 HA*

   c. On the 3D protein structure search result page, choose the “1EO8” protein structure and click on “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: **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, positions, which have been found to influence virus host range.

j. Rotate the structure as you need. Download the protein structure image with highlighted residues by clicking on “Save View As Image” beneath the image.
5. Search and Visualize Avian Surveillance Data

After this exercise you should be able to identify influenza surveillance samples from selected parts of the world, determine the proportion that are flu positive and view their geospatial relationships with known bird flyways.

Wild birds are the natural reservoirs of influenza viruses. Influenza surveillance in different wild bird populations is critical for understanding the transmission and evolution of these viruses, and assists in the early detection and warning of highly pathogenic avian influenza.

IRD serves as a central repository of influenza surveillance data for the NIH-NIAID funded Centers of Excellence for Influenza Research and Surveillance (CEIRS) program.

a. Search surveillance data from avian host collected in Africa.

From the grey navigation bar, mouse over “Search Data” and click on “Animal Surveillance Data”. On the next page, choose the following parameters:

- Surveillance Data Type: Avian
- Sample Selection: Uncheck all boxes
- Sampling location: Benin, Cote D’Ivoire, Egypt, Sudan, Togo, Uganda

Note: Use the “Search Criteria” button to go back and change search criteria to answer the following questions.

- How many total African avian surveillance samples are currently found in the IRD?
- How many have been tested for the presence of influenza virus?
- How many samples are flu-positive?
- How many have associated sequence records currently in IRD?

b. Display surveillance records for all tested samples on a map.

Adjust Search Criteria to only select all tested samples and select all records; click on the orange “View on Map” button.

c. Color code by the percent of flu-positive samples.

Check the checkbox next to “Show percent flu-positive samples”.

d. Highlight bird flyways and see if there’s a correlation between flyways and bird samples.

Check “Black sea/Mediterranean”. Which countries might be expected to share similar influenza virus strains based on this bird flyway?
e. Zoom in to find at least one sampling location in Egypt that has flu-positive samples. What’s the name of the location (City, Province)?

*On the surveillance map, click on 📍 to view details about the records at that location.*

f. View the surveillance records of flu-positive samples.

i. From Step e., 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.

ii. Sort records by the “Positive for flu” column.

iii. Click on 📊 next to a flu-positive sample to see the sample collection, the host, etc.
PART II. VIRUS PATHOGEN RESOURCE (ViPR) EXERCISES

1. View Basic Virus Information

At the end of this exercise you should be able to navigate to different virus home pages and view basic information about these groups of viruses.


What kind of genome does Dengue virus have?

How many mature proteins are encoded by the Dengue virus genome?

How many distinct virus types have been identified?

b. Migrate back and find the Flaviviridae home page.

How many classified Flaviviridae genera are represented?

How many different Pestivirus species are represented?

How many complete genomes of Pestivirus are available?

c. Pull down “Announcements” menu and view upcoming “Meetings and Events”.

d. Pull down “Resources” menu and view “Anti-Viral Drug Information”, including details about Ribavirin. Which host enzyme is involved in metabolizing this anti-viral drug?

2. Multiple Sequence Alignments and Phylogenetic Trees

Note: ViPR has most of the same analysis and visualization tools as IRD.
3. Workbench

At the end of this exercise you should be able to register to set up a private workbench, and save and share data and analysis results for future use.

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. Click on the “Workbench” tab from the grey navigation bar to go to your Workbench.

Note: To save and share your data and searches, you’ll need to have an account with ViPR.

b. Click on the orange “Register” button in the right-up corner to register an account.

After completing the registration process you will be taken to your own personal workbench page.
You’ll see the searches and analyses you just ran during this session.

c. On the Workbench page, select a search or analysis by clicking on the checkbox next to the name.
Click on the orange “More Actions” button and then “Save Unsaved Searches” or “Save Unsaved Analysis” to save your search/analysis permanently to your Workbench.

d. Select a search or analysis and share it with Dr. Richard Scheuermann.
From your Workbench, select items, mouse-over the orange “Sharing” button, and click on “Share items with collaborator”. You’ll be prompted to enter your collaborator’s email address. Enter Richard.scheuermann@utsouthwestern.edu here.

e. Extra exercise on your own time: Log in to your Workbench, click on the orange “More Actions” button, then “Upload File”, and run analysis of your choice.
4. Metadata-driven Comparative Genomics

At the end of this exercise you should be able to identify nucleotide or amino acid positions that differ significantly between groups of viruses distinguished by specific host attributes.

Metadata Genome Compare - customized comparative genomic analysis tool
- Identify significant polymorphisms that correlate with metadata, e.g., host attributes, sample location, etc.

a. From the Dengue page, mouse-over the grey “Search Data” menu item and select “Genomes”.

b. From within the Taxonomy tree, click on “Select All” next to “Dengue virus type 2”, scroll to the bottom of the page and click on the “Search” button.

c. Select the 4 sequences isolated during 1995 from the USA, and at least 4 sequences during 2007 from Venezuela by clicking on the checkbox next to the sequence name. **Hint: Try resorting list based on “Date” and/or “Country” to help find the records of interest.**

d. Save selected sequences as a new working set. Give descriptive name.

e. Go to your newly-created working set in Workbench. Select all sequences. At the top of the page, mouse-over the “Run Analysis” button and click on “Metadata Genome Compare”.

---

**Influenza Research Database (IRD,** [www.fludb.org](http://www.fludb.org)) is funded by the U.S. National Institute of Allergy and Infectious Diseases (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. **Virus Pathogen Database and Analysis Resource (ViPR, [www.viprbrc.org](http://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.**
f. We are going to separate these sequences into two groups and use “0.05” as our significance cutoff value, so keep the default settings on the page and click “Continue”.

![Sequence Grouping](image)

---

g. Now, assign the sequences into two groups by double-clicking on the strains isolated from the USA 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 the strains from Venezuela to the other group and click the “Run” button.

![MAIN LIST OF SEQUENCES](image)

---

h. 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.

Note the first three positions with significant nucleotide difference between sequence groups (e.g. 197, 245, 272).

i. Go back to the same working set, generate and view a new multiple sequence alignment. Find the positions of the significant nucleotide differences.
PART III. SCIENTIFIC USE CASES

Use IRD or ViPR to answer hypothetical scientific questions. You may consider using the functionalities that you have just used to address the scientific questions outlined below.

A. Identification of candidate adaptive drivers of species jumps in avian H5N1 viruses

Influenza is a zoonotic virus that circulates in a variety of reservoir species, especially wild birds. Over the last century, each of the human influenza pandemics occurred as a result of a “species jump” event in which a reservoir virus established an initial infection in a human host, followed by sustained human-to-human transmission. Nucleotide sequence substitutions that occur in multiple independent species jump viruses suggest convergent evolution: these are candidates of adaptive drivers of human transmission [Lloyd-Smith 2009; Pepin 2010].

1. Divide into 10 groups – each group will be assigned one of ten influenza proteins (excluding PB1-F2)
2. Select avian reservoir virus sequences
   a. Query: Influenza A, H5N1, protein X, avian, Southeast Asian Countries (China, Hong Kong, Indonesia, Thailand, Viet Nam), up to 2003
   b. Select all sequences
   c. Save as working set to workbench with descriptive name (e.g. NS1_avian_H5N1_SEA_up to 2003)
3. Select human stuttering virus sequences
   a. Query: Influenza A, H5N1, protein X, human, Southeast Asian Countries (China, Hong Kong, Indonesia, Thailand, Viet Nam), after 2003
   b. Select all sequences
   c. Save as working set to workbench with descriptive name (e.g. NS1_human_H5N1_SEA_after 2003)
4. Calculate sequence conservation score for avian reservoir viruses
   a. Open saved avian working set in workbench
   b. Select all sequences
   c. In “Run Analysis” pull down menu, select “Analyze Sequence Variation (SNP) option
   d. Select amino acid option
   e. Download result
5. Calculate sequence conservation score for human stuttering viruses
   a. Open saved human working set in workbench
   b. Select all sequences
   c. In “Run Analysis” pull down menu, select “Analyze Sequence Variation (SNP) option
   d. Select amino acid option
   e. Download result
6. Create .xls spreadsheet with combined results
   a. Open avian conservation score download in Excel
   b. Save as an Excel spreadsheet with a descriptive name (e.g. NS1_H5N1_avian_human_comparative analysis)
c. Open human conservation score download in Excel
d. Copy all columns
e. Paste next to avian data in saved Excel spreadsheet
f. Save

7. Validate results
   a. Compare the consensus sequence from human with consensus sequence from avian – they should be very similar throughout the length.
   b. If they match, move to step 8.
   c. If they don’t match, is may be necessary to add or remove a position to get the two lists in the correct register

8. Identify highly conserved amino acids in avian reservoir viruses
   a. In avian AA count column visually find examples in which the second most frequent amino acid is found in only one virus (e.g. Arg=1, Glu=144; Met=189, Val=1)
   b. Find the conservation score value for these examples (e.g. 6; 5)
   c. Use this information to select a conservation score cutoff to capture sequences with at most one variant virus at a given position (e.g. <7)
   d. Select Data=>Filter=>Autofilter
   e. In Avian conservation score column, select custom filter, is less than “cutoff score determined in step 8c” (7)
   f. These are the Avian H5N1 with <=1 variant amino acid

9. Identify variable amino acids in human stuttering viruses and combine results
   a. In human AA count column visually find examples in which the second most frequent amino acid is found in two or more viruses (e.g. Leu=3, Val=154; Gly=2, Ser=156)
   b. Find the conservation score value for these examples (e.g. 14; 10)
   c. Use this information to select a conservation score cutoff to capture sequences with at least two variant virus at a given position (e.g. >9)
   d. Select Data=>Filter=>Autofilter
   e. In Avian conservation score column, select custom filter, is greater than “cutoff score determined in step 8c” (9)
   f. These are the human H5N1 with >= 2 variant amino acid
   g. Manually select those in which the same amino acid is observed >= 2 times
   h. The end result of the combined filters applied in step 8 and 9 will be - Avian H5N1 with <=1 variant amino acid and human H5N1 with >= 2 of the same variant amino acid as evidence of convergent evolution.

10. Identify strains with evidence of adaptive drivers
    a. Copy filtered data and paste into new worksheet
    b. Curate list to remove any spurious results (e.g. multiple deletions)
    c. Sort entire sheet based on human conservation score in descending order
    d. Select amino acid position with highest score
    e. Go back to IRD workbench
    f. Select human data set (e.g. NS1_human_H5N1_SEA_after 2003)
    g. Select all sequence records
    h. In Run Analysis pull down menu, select Identify Point Mutations in Proteins option
1. Type in amino acid position from step 10d
2. Sort list by last column
3. Select all records corresponding to the second most common amino acid
4. Save to a new working set with informative name (e.g. “Point Mutation X”)

11. Outgroup sequence selection
   a. A/Goose/Guangdong/1/96

12. Phylogenetic analysis for evidence of independent lineages
   a. Select “Point Mutation X” and “Outgroup” working sets
   b. Under “More Actions” button, select “Combine” option
   c. Give new working set and informative name (e.g. “Point Mutation X_with Outgroup”)
   d. Select all sequences in this new working set
   e. In the Run Analysis pull down menu, select “Generate Phylogenetic Tree” option
   f. Use Quick tree option and view resulting tree for evidence of independent lineages

13. Report
14. Repeat steps from 10d – 13 (excluding 11)

Report: For which positions is there evidence for convergent evolution? Any positions that you are not sure about?

B. Computational exploration of sequence relationships and variation in Bunyaviridae

Extra exercise on your own time.

Suppose you are interested in analyzing the relationships between Bunyaviridae viruses (either the Nairovirus or Orthobunyavirus genera) with different hosts, isolation years, or locations. Use the “S” segment in Bunyaviridae to do analysis. Make use of multiple analysis methods and compare the results generated from different methods.

Perform a genome search for the “S” (small) segment in Bunyaviridae, choose sequences with different hosts, isolation years, or isolation locations and save them as a Working Set. Next, generate a: multiple sequence alignment, phylogenetic tree, metadata genome comparison, and SNP analysis.

   a. Use the results from the phylogenetic tree analysis to assign the sequences into different groups for input into the metadata-driven comparative genomics analysis tool. How does the topology of the tree reflect the different metadata associated with the groups of sequences?
   b. Compare the results generated from the MGC pipeline with those generated from the simple SNP analysis. What are the differences? Why?
   c. Verify the MGC results by looking at the statistically-significant positions while visualizing the multiple sequence alignment. Do you notice anything about the sequences segregating at these positions?
PART IV. ANSWER KEY

Part I
1e. 20 (as of May 22, 2011)
2h. HKY
3c. 136
3e. Influenza A_NS1_nuclear-export-signal_137(11) (Influenza A_NS1_SF18)
3f. PubMed: 9560194, Uniprot: P03496; A/Puerto Rico/8/34(H1N1)
3g. 189; A/Udorn/72(H3N2)
3h. Alphabetical by Strain Name; 1902
3i. They are largely isolated from horse and dog.
3j. 4 (excluding VT-1)
4g. No
5a. All: 4772; Tested: 2594; Flu-positive: 94; With sequence records: 0 (as of May 22, 2011)
5d. Egypt and Cote d’Ivoire

Part II
1a. 10.7 kilobase positive-sense, single-stranded RNA; 11 mature proteins; 4 types
1b. 3; 7; 127 (as of May 22, 2011)
1d. Adenosine kinase

PART V. BIBLIOGRAPHY
