



INFLUENZA RESEARCH DATABASE & VIRUS PATHOGEN RESOURCE

EXERCISES

Table of Contents

Section A. Get familiar with the IRD/ViPR site	1
Section B. Comparative Genomics Analysis of 2013 H7N9 Influenza A Viruses	4
I. Search for sequences and save sequences into working sets	6
II. Construct a phylogenetic tree	9
III. Metadata-driven Comparative Analysis Tool for Sequences	11
IV. Determine if the significant positions are located in Sequence Features	12
V. View multiple sequence alignment	14
VI. Visualize 3D protein structures	15
Section C. Annotate your own virus genome sequences	17
I. Annotate an influenza virus segment sequence	17
II. Annotate a Hepatitis C Virus genome sequence	19

Section A. Get 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.

I. 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

The screenshot shows the IRD homepage with a blue header containing the IRD logo and navigation links: About Us, Community, Announcements, Links, Resources, Support, and Sign Out. Below the header is a navigation bar with tabs for SEARCH DATA, ANALYZE & VISUALIZE, WORKBENCH, SUBMIT DATA, and HOME. The main content area is divided into three columns: Search, Analyze, and Save to Workbench. Each column lists various data and tool options. Red arrows point to the 'Search', 'Analyze', and 'Save to Workbench' headers. A red oval highlights the top navigation menu.



- 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.
- Mouse-over or click the “**Search Data**” tab to view available data search options.
 - 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.
- Pull down the “**About Us**” menu and view Citing IRD, Our Publications, and Research Using IRD.
 - Pull down the “**Announcements**” menu and view Meetings and Events and IRD Newsletters.
 - Pull down the “**Resources**” menu and view the WHO vaccine strains, BEI reagent resources, anti-viral drug information, Reactome flu pathways, and other resources.
 - Click the “**Support**” menu and view how to contact the IRD team when you have questions, suggestions, or problems.
 - Pull down the “**Support**” menu and click the “**IRD Protocols**” link to view the protocols used to generate derived data.
 - Pull down the “**Support**” menu and click the “**Tutorials and Training Material**” link to view available tutorials and training materials. View the **YouTube video**.

II. Getting familiar with the Virus Pathogen Resource

- a. Go to the ViPR homepage (<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 (e.g. *Flaviviridae*). 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 “**ViPR Protocols**” link to view the protocols used to generate derived data.
 - vi. 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 (e.g. *Togaviridae*) to get to another virus family page.



Section B. Comparative Genomics Analysis of 2013 H7N9 Influenza A Viruses

Objective

Upon completion of this exercise, you will be able to use the Influenza Research Database (IRD; <http://www.fludb.org/>) to:

- Search for virus sequences and view detailed information about these sequences in IRD
- Save selected sequences as a working set in your private Workbench space
- Combine multiple working sets
- Build a phylogenetic tree on a set of sequences to infer their evolutionary relationships
- Use Meta-CATS to identify nucleotide or amino acid positions that significantly differ between groups of virus sequences
- Determine if significant positions are located in viral protein Sequence Features and examine Sequence Feature Variant Type reports
- Perform a multiple sequence alignment to observe sequence conservation and variations
- Search for 3D protein structures and highlight various features and positions on a structure

Background

H7 viruses normally circulate in birds and horses. Before March 2013, a search for H7 influenza strains in IRD returned a total of 1485 strains in IRD, with 1306 from birds, 102 from environmental samples (usually bird droppings), 33 from horses, and only 15 from humans (11 H7N7, H7N1, H7N2, 2 H7N3). No human isolates were H7N9.

In March 2013, several cases of Influenza virus A H7N9 subtype were identified in Shanghai, China and surrounding provinces. As of May 20, 2013, a total of 132 cases have been confirmed, including 37 deaths. Fortunately, no evidence of ongoing human-to-human transmission for the current H7N9 outbreak has yet been found. This suggests that the virus is undergoing “stuttering transmission” in which a virus that normally circulates in an animal reservoir infects a person, but further human-to-human transmission does not occur. In general, viruses capable of stuttering transmission have acquired novel sequence variations that allow them to infect humans (human adaptation), but have yet to acquire sequence variations that allow them to sustain efficient transmission between humans.

We will perform an in-depth statistical analysis using sequence records and analysis tools available in IRD (www.fludb.org) to clarify the origin of the HA segment and to identify candidate sequence variations that might be involved in this type of human adaptation.



Analysis Workflow

Search for sequences and save sequences into working sets:

- search for H7N9 HA nucleotide sequences and save sequences as working set (1)
- BLAST for HA nucleotide sequences similar to H7N9 2013 outbreak sequences and save them as working set (2)
- combine working sets (1)-(2) into one (3)
- convert nucleotide working set (3) into protein working set (4)

Nucleotide phylogenetic tree:

- construct phylogenetic tree using working set (3)
- color tree to reveal host and subtype specific branching patterns

Metadata-driven Comparative Analysis Tool (Meta-CATS):

- input the protein working set (4) to Meta-CATS
- group the sequences based on the tree topology
- identify positions that are significantly different in the older Eurasian lineage H7N9 isolates compared to the human H7N9 2013 outbreak isolates

Determine if the significant positions are located in Sequence Features:

- follow the Sequence Feature linkage on the Meta-CATS report
- examine Sequence Features containing the significant positions

Multiple sequence alignment:

- align HA protein sequences
- observe the variant positions on the alignment

Highlight significant positions on protein structure:

- search for H7 HA 3D protein structures
- highlight significant positions identified by Meta-CATS on a structure



I. Search for sequences and save matching sequences into working sets

1. Search for H7N9 HA sequences using structured search interfaces

- Go to the IRD homepage (<http://www.fludb.org/>), mouse-over “**Search Data**” in the grey navigation bar, then “**Search Sequences**” and click “**Nucleotide Sequences**”.
- The Nucleotide Sequence Search page allows you to search for sequences based on data type, virus type, subtype, strain name, segment, host, geographical region, complete sequences or not, H1N1 pandemic sequences or not, and date range.
- For this exercise, we are going to search for HA segment sequences from H7N9 strains. Select the following criteria and click the orange “**Search**” button to run the query.

Virus Type: A Complete Sequences: Complete Sequences Only
 Select Segments: 4 HA Advanced Options: Remove Duplicate Sequences
 Sub Type: H7N9

- Note that IRD shows instant count of search results here 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.

SEARCH DATA ANALYZE & VISUALIZE WORKBENCH SUBMIT DATA

Home > Nucleotide Sequence Search

Nucleotide Sequence Search

Search for influenza sequences, proteins, and strains using two types of searches. Use the advanced search to allow you to refine your search with the more fine grained search, and you can pick your viewing options.

Results matching your criteria: 59

DATA TO RETURN	SELECT SEGMENTS	HOST	GEOGRAPHIC GROUPING
<input checked="" type="radio"/> Segment / Nucleotide <input type="radio"/> Protein <input type="radio"/> Strain	All 1 PB2 2 PB1/PB1-F2 3 PA/PA-X 4 HA 5 NP 6 NA 7 M1/M2 8 NS1/NS2	All Avian Bat Blow Fly Camel Cheetah Chicken Civet Dog Domestic Cat Donkey Environment Ferret Horse Human Lab	All Africa Asia Europe North America Oceania

VIRUS TYPE

A
 B
 C

SUB TYPE

H7N9
 * Use comma to separate multiple entries.
 Ex: H1N1, H7, H3N2.

STRAIN NAME

* Use comma to separate multiple entries.
 Ex: A/chicken/Israel/1055/2008, A/chicken/Laos/16/2008.

COMPLETE SEQUENCES

Complete Sequences only
 Include near-complete sequences (VFR)

2009 pH1N1 SEQUENCES (SOP)

Include pH1N1 sequences
 Include only pH1N1 sequences
 Exclude all pH1N1 sequences

DATE RANGE

From: YYYY To: YYYY
 To add month to search, see Advance Options: Month Range

Tip: To select multiple or deselect, Ctrl-click (Windows) or Cmd-click (MacOS)

ADVANCED OPTIONS Show All

Select Advanced Option

Remove Duplicate Sequences Remove

* Add Another Advanced Option

Clear Search

- The Search Results page will be displayed. Here you can:
 - Save the search query to your Workbench and rerun the search again later.
 - Download the sequences (gene, CDS, protein) or the displayed table 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 for any item in the results table by clicking on “**View**” next to any row.
- f. Now click the “**Date**” header in the results table to sort records by date. If you need advanced sorting options and want to display additional fields, click the “**Display Settings**” button.

Segment	Protein Name	Sequence Accession	Complete Genome	Segment Length	Subtype	Date	Host Species	Country	State/Province	Flu Season (SOP)	Strain Name
4	HA	CY146908	Yes	1683	H7N9	05/03/2013	Chicken/Avian	China	-NA-	-NA-	A/chicken/Guangdong/SD641/2013(H7N9)
4	HA	CY147100	Yes	1683	H7N9	05/03/2013	Environment	China	-NA-	-NA-	A/environment/Shandong/SD038/2013
4	HA	CY147108	Yes	1683	H7N9	05/03/2013	Environment	China	-NA-	-NA-	A/environment/Shandong/SD039/2013
4	HA	CY146948	Yes	1683	H7N9	05/03/2013	Chicken/Avian	China	-NA-	-NA-	A/chicken/Jiangxi/SD001/2013

How many HA segment sequences did you find from the current outbreak?

- g. To analyze these sequences, we will select records by ticking the checkbox and adding them to a working set by clicking the “**Add to working set**” button. This way, we will be able to retrieve the data from the Workbench later and run various analyses on the same data set.
- h. You’ll be prompted to log in to your Workbench account in order to save data to a working set. If you don’t have an account already, simply register for an account for free by choosing the “**Register for a new account**” option and following the prompts.
- i. A lightbox of “Add to Working Set” will pop up. Now create a new working set and name it “H7N9 HA complete sequences”. Click “**Add to Working Set**” to save the sequences to a working set.

2. BLAST for HA sequences similar to H7N9 2013 outbreak sequences

Now we are going to expand the sequence set by including HA sequences that are highly similar to the outbreak sequences. We will select a representative isolate and perform a BLAST search of nucleotide sequences. The IRD BLAST tool utilizes the NCBI BLAST program set and has a collection of custom influenza sequence databases to search against.

- a. Select A/Shanghai/02/2013 from the results table, mouse over “**Run Analysis**” and click “**BLAST**”.
- b. In the Select Sequence Type lightbox, select “Nucleic Acid (Segment)” and click “**Continue**”.



- c. Now the BLAST setting page is loaded. IRD provides a collection of custom influenza sequence databases to search against using BLAST. Select “Blastn”, then “Nucleotides for segment 4 HA”. Use the default parameter settings. Click “Run”.

- d. On the BLAST Report page, all nearest hits are listed in the table. Click a hit to view its alignment. Click the IRD link (e.g., ird|982104) to view the hit’s segment/ protein details page in IRD. What are the host and subtype of the sequences that are most similar to the H7N9 outbreak sequences?

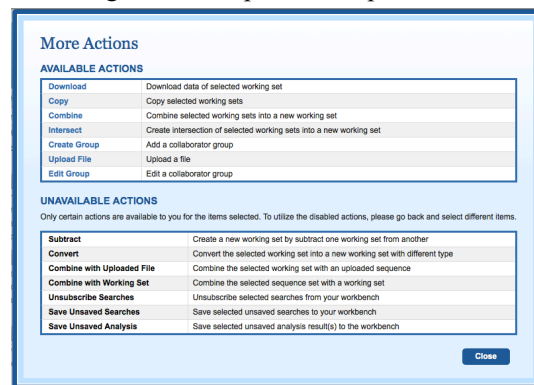
Id	Sequence header	Bit Score	E Value
<input type="checkbox"/> 982104	>ird 982104 Country:China Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 4 hemagglutinin (HA) gene, complete cds. gbj KF021597	3386	0.0
<input type="checkbox"/> 970288	>ird 970288 Country:China Influenza A virus (A/Zhejiang/DTID-ZJU01/2013(H7N9)) segment 4 hemagglutinin (HA) gene, complete cds. gbj KC885956	3378	0.0
<input type="checkbox"/> 973128	>ird 973128 Country:China Influenza A virus (A/Fujian/1/2013(H7N9)) segment 4 hemagglutinin (HA) gene, complete cds. gbj KC994453	3328	0.0
<input type="checkbox"/> 979408	>ird 979408 Country:China Influenza A virus (A/environment/Hangzhou/34/2013(H7N9)) segment 4 hemagglutinin (HA) gene, complete cds. gbj KF001519	3320	0.0

- e. Return to the BLAST Report page by clicking the Results breadcrumb. Now select the top 20 hits that are not directly associated with the current outbreak based on isolation year and subtype (non-H7N9) and click “Add to Working Set” to add these sequences to a new working set named “A/Shanghai/02/2013 BLAST top 20 hits”.



3. Construct segment and protein working sets

- Now we are going to combine the working sets of H7N9 HA complete sequences and A/Shanghai/02/2013 BLAST top 20 hits and use the combined working set for downstream analyses. Click the “**Workbench**” tab from the grey navigation bar to go to your Workbench. You’ll see the saved working set listed at the top of the Workbench table.
- Click the checkboxes for the two working sets we just saved. Click “**More Actions**” then “**Combine**”. Name the combined working set to “H7N9 HA complete seqs + H7N9 blastn top 20”. This working set contains the HA segment sequences from H7N9 isolates and 20 segment sequences that are most similar to the HA segment sequences of the current H7N9 outbreak.
- Next, we are going to convert the combined segment working set into a protein sequence working set. To do so, select the combined working set. Click “**More Actions**” then “**Convert**”.
- In the Convert Working Set lightbox, select Protein as data type and name the working set to: H7N9 HA complete seqs + H7N9 blastn top 20 protein. Click “**Convert**”.
- Access your Workbench by clicking the “**Workbench**” tab. You will see the newly created working sets at the top of the content list.



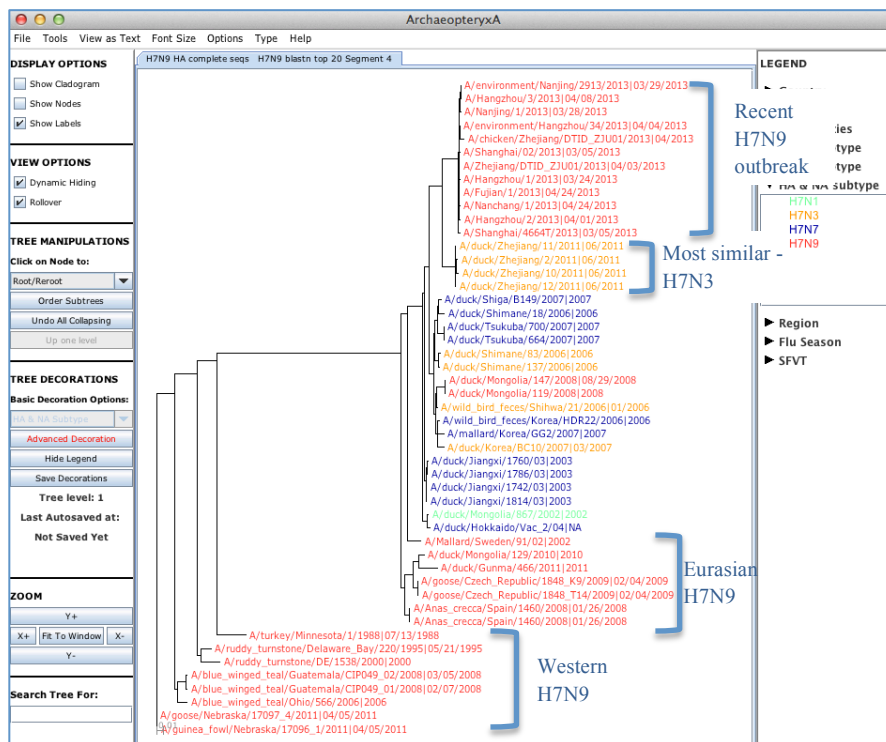
II. Construct an HA segment phylogenetic tree

- Now we will construct a phylogenetic tree using HA segment sequences from H7N9 isolates and segment sequences that are most similar to the current H7N9 outbreak isolates. On the Workbench page, click “**View**” next to H7N9 HA complete seqs + H7N9 blastn top 20.
- The working set details page displays the sequence records saved in the working set. Select all records by clicking the checkbox above the table. Mouse over “**Run Analysis**” and click “**Generate Phylogenetic Tree**”.
- On the Tree setting page, select “**Quick Tree**”, choose strain name and date as tree tip label, and click “**Build Tree**”.
- While the analysis is running, you can save the analysis to your Workbench by entering a name and then clicking “**Save to Workbench**”. Once it is saved, you can come back to the Workbench at any time to retrieve the analysis results.
- 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.
- A Tree Viewer window will pop up. Many tree customization options exist including: reroot the tree, collapse/expand/display subtree, swap descendants, decorate (color) the tree leaves by any



associated metadata (e.g. host or year of isolation, etc.), resize the tree, zoom in/out, fit the tree to window, change the font size, etc.

- i. In the “Tree Decorations” section, select “**HA&NA Subtype**” from the “Basic Decoration Options”. Click “**Show Legend**” to display the color code for different subtypes.
 - ii. The default colors may or may not be ideal for your purpose. You can change the color by using the “**Advanced Decoration**”. In the Advanced Decoration Options dialog box, select “**HA&NA Subtype**”, click the Manual Decoration checkbox and click “**Go**”.
 - iii. Check H7N9 and choose red in the color palette, then click “**Apply**”. Now the H7N9 strains are colored in red.
 - iv. The tree shows that the HA sequences from the H7N9 outbreak are most similar to the HA sequences from a series of duck H7N3 isolates from Zhejiang, suggesting that this new H7N9 outbreak is likely a result of a reassortment event in which the HA segment was derived from an H7N3 ancestor.
 - v. You can save the tree image by clicking the “**File**” menu and then a file format.
- g. Return to the Tree Results page. Save the tree analysis to your Workbench by clicking “**Save Analysis**”. Rename the analysis so that you can recognize it later, for example, “H7N9 HA segment phylogeny”. Then click “**Save**”.
- h. Go to your Workbench. You can see the tree is listed at the top of the Workbench table. Click “**View**” to retrieve the tree analysis result. The parameters used to generate the tree are also saved.





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

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

- 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.
- Picket BE, et al. (2013) "Metadata-driven Comparative Analysis Tool for Sequences (meta-CATS): an Automated Process for Identifying Significant Sequence Variations Dependent on Differences in Viral Metadata." *Virology* (in press).



Now we will use Meta-CATS to identify amino acid positions that are significantly different between the human isolates from the current outbreak and the older Eurasian lineage H7N9 isolates. These two groups are based on the phylogenetic tree topology and will identify positions that differ between these two groups in a statistically significant way.

- We are going to analyze the protein sequences we saved previously in the working set: H7N9 HA complete seqs + H7N9 blastn top 20 protein. So go to your Workbench, find the working set and click “**View**” to display the sequences.
- Sort the list by the “**Date**” column. Now select the following protein sequence records:
 - All human H7N9 from 2013
 - All non-human H7 isolated from all European and Asian Countries before 2012
 - Then mouse over “**Run Analysis**” and click “**Metadata-driven Comparative Analysis Tool**”.
- You are taken to the meta-CATS tool setting page. Here we will separate these sequences into two groups according to the phylogenetic tree analysis: 2013 H7N9 isolates as a group and the rest sequences as another group. We can do so by grouping the sequences by year. Select “Auto Grouping”, and then select “**Year**” from the dropdown list. Now enter year break point “2012” to get groups of: (1) 2012 & before (Eurasian H7 ancestral isolates), and (2) > 2012 (human H7N9 2013 human outbreak isolates). Select “Unaligned FASTA”, use “0.05” as our significance cutoff value and click “**Continue**”.
- On the next page, you will see the sequences are separated into two groups: Group 1 containing <=2012 sequences, and Group 2 containing >2012 sequences. Click the “**Run**” button.
- This analysis may take a few minutes to finish. You can save the analysis to your Workbench and retrieve it later. To do so, enter in a name (Ex., human H7N9 2013 HA vs. older Eurasian) and click “**Save to Workbench**”.



- f. The Meta-CATS analysis result has two reports: a Chi-square Test of Independence 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. Since this analysis only deals with two groups of sequences, we will primarily focus on the first result table.
- g. Review the Chi-square test results to see the positions that differ significantly between the current H7N9 outbreak isolates and other isolates. The residue diversity column lists the counts for each residue within a group. Now sort the results by the Chi-square value to push the most different positions to the top of the table. What is the position number with the highest Chi-square value?
- h. Save the analysis result to your Workbench by clicking the “Save Analysis” button.

Position	Chi-square Value	P-value	Degree Freedom	Residue Diversity	Sequence Feature
235*	37.991	5.627E-9	2	group1(27 Q) group2(1 I, 10 L)	View SF
188*	37.991	5.627E-9	2	group1(1 A, 26 I) group2(11 V)	View SF
541	33.289	7.944E-9	1	group1(27 A) group2(11 V)	View SF
455	33.289	7.944E-9	1	group1(27 N) group2(11 D)	View SF
410	33.289	7.944E-9	1	group1(27 T) group2(11 N)	View SF
195	33.289	7.944E-9	1	group1(27 G) group2(11 V)	View SF
183	33.289	7.944E-9	1	group1(27 D) group2(11 S)	View SF
211*	25.797	3.793E-7	1	group1(25 I, 2 V) group2(11 V)	View SF
307*	20.304	6.606E-6	1	group1(4 D, 23 N) group2(11 D)	View SF
198*	20.304	6.606E-6	1	group1(4 A, 23 T) group2(11 A)	View SF
11*	19.792	1.874E-4	3	group1(1 E, 5 I, 1 M, 18 V) group2(11 I)	View SF
130*	19.119	7.052E-5	2	group1(6 A, 1 I, 20 T) group2(11 A)	View SF
321	18.073	2.126E-5	1	group1(22 E, 5 R) group2(11 R)	View SF

IV. Determine if the significant positions are located in Sequence Features

Sequence Features (SFs) are defined as interesting protein regions with known structural or functional properties. They are obtained from literature and other databases and validated by domain experts. Once a Sequence Feature region has been defined, the number of distinct amino acid sequences observed in the sequence database are determined and each defined as a unique variant type. The reference strain is always Variant Type 1.

The Sequence Feature (SF) column in the meta-CATS table provides a convenient link out to a list of all Sequence Features that contain that amino acid position.

- a. Select the SF link for residue position 235. Is this position located within any predefined Sequence Features?
- b. Click “View” for SF5 to get to the Sequence Feature (SF) Details page.
 - i. This SF is an experimentally determined epitope. It begins at residue 226 and is 14 amino acids long. What is the reference strain used to define the position coordinates of this SF? What is the position range on the reference strain?



- ii. The majority of 2013 outbreak isolates have a Leucine at position 235, which corresponds to Variant Type 7. Click the Strain Count for VT-7. How many strains harbor this substitution? Any strains from the current outbreak?

SEQUENCE FEATURE DEFINITION

Protein Name	HA
Sequence Feature Name	Influenza A_H7_experimentally-determined-epitope_226(14)
Sequence Feature ID	Influenza A_H7_SF5
VT-1 Strain (reference strain)	A/Turkey/Italy/220158/2002(H7N3)
Reference Sequence Accession	AY586409
Reference Position	226(210 HA1)-239

SOURCE STRAIN(S)

Source Strain	VT Number	Source Position	Source Accession	3D Protein Structure	Publication	Epitope Type	Evidence Codes	Epitope Sequence	Comment
A/England/268/1996(H7N7)	VT-1	226-239	AF028020	-N/A-	IEDB:177209 ↗	B Cell	-N/A-	PGARPVVNGQSGRI	-N/A-

VARIANT TYPES

[Excel Download](#)
[MSA Download](#)
[View Phylogenetic Tree](#)
[Find a VT\(s\)](#)

Strain Count	Variant Type	Sequence Variation													Total Variations	
		226	227	228	229	230	231	232	233	234	235	236	237	238		239
772	VT-1	P	G	A	R	P	Q	V	N	G	Q	S	G	R	I	0
356	VT-2	•	•	•	•	-	-	-	-	-	-	-	-	•	•	8
38	VT-3	•	•	E	•	•	•	•	•	•	•	•	•	•	•	1
18	VT-4	•	•	P	•	•	•	I	•	•	•	•	•	•	•	2
16	VT-5	•	•	T	•	-	-	-	-	-	-	-	-	•	•	9
15	VT-6	•	E	•	•	•	•	•	•	•	•	•	•	•	•	1
12	VT-7	•	•	•	•	•	•	•	•	•	L	•	•	•	•	1

- c. Return to the meta-CATS report by clicking the breadcrumb. The recent 2013 isolates possess G195V and T198A substitutions. Click “**View SF**” for 195. Positions 195 and 198 are located within determinants of receptor binding in the 194-198 loop (SF4).
- d. Click “**View**” for SF4.
- How many Variant Types does this SF have?
 - The older H7 proteins mostly have SGSTT, which corresponds to VT-1.

SEQUENCE FEATURE DEFINITION

Protein Name	HA
Sequence Feature Name	Influenza A_H7_determinants-of-receptor-binding_194(5)
Sequence Feature ID	Influenza A_H7_SF4
VT-1 Strain (reference strain)	A/Turkey/Italy/220158/2002(H7N3)
Reference Sequence Accession	AY586409
Reference Position	194(178 HA1)-198

SOURCE STRAIN(S)

Source Strain	VT Number	Source Position	Source Accession	3D Protein Structure	Publication	Evidence Codes	Comment
A/duck/HONG KONG/293/1978(H7N2)	VT-1	185-189	U20461	-N/A-	PubMed:22345462 ↗	EXP	Atypical European viruses show G186V whereas the N. American strains show G186A or G186E.

VARIANT TYPES

[Excel Download](#)
[Find a VT\(s\)](#)

Phylogenetic tree view disabled because there are not enough variant types to generate the tree.

Edit specific positions in this VT-1 sequence with IUPAC symbols or use “?” as a wild-card. If necessary, use the horizontal scroll bar to access the entire SF. Click Search to find VT(s) conforming to the edited sequence. Click Reset to restore this panel to the default VT-1 sequence.

Enter Sequence Variation to Find					
194	195	196	197	198	
?	V	?	?	A	

SEQUENCE VARIATION

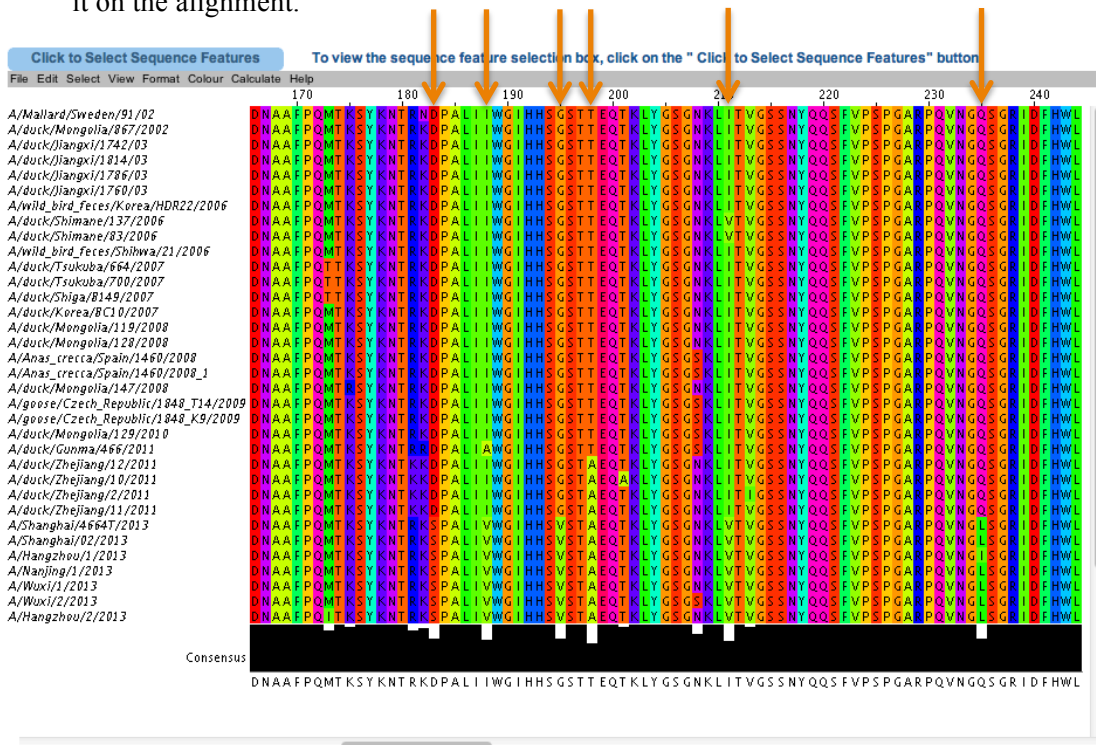
Strain Count	Variant Type	194	195	196	197	198	Total Variations
388	VT-1	S	G	S	T	T	0
14	VT-11	•	V	•	•	A	2

- iii. Now we are going to search for all strains harboring Valine at position 195 and Alanine at position 198. Click “**Find a VT**”, fill all positions with wildcards and fill in V at 195 and A at 198. Then click “**Search**”.
- iv. Did you find a VT harboring SVSTA in the 194-198 loop? Click strain count for the VT. Are they from the current outbreak?

V. View protein sequence alignment

Now we are going to view the protein sequence alignment to confirm the meta-CATS results and to verify clade relationships inferred from the phylogenetic analysis.

- a. Follow breadcrumb back to Working Set with selected sequences. Click the “**Visualize Aligned Sequences**” option from the “**Run Analysis**” pull down menu. Select Sort Sequences By: “**Date**”.
- b. The alignment is presented in the JalView visualization window. The window is interactive.
 - i. The consensus sequence is shown at the bottom of the window. You can choose to show sequence logos by right-clicking on consensus and then selecting “**Show logo**”.
 - ii. You can manually adjust the alignment and display using various gray menu options.
 - iii. Scroll right up to the region of 183-235. Several amino acid substitutions, including D183S, I188V, G195V, T198A, I211V, and Q235L are observed in the vast majority of the recent H7N9 isolates, but are absent from the older H7 proteins.
 - iv. You can highlight Sequence Features on the alignment. “**Click to select Sequence Features**”, you will see a list of Sequence Features curated by IRD. Click SF4 to highlight it on the alignment.



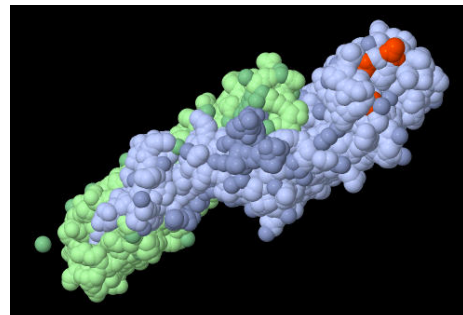


- v. We can change the View Option to “Conserved vs. reference” such that only the first sequence shows full characters, for the remaining sequences, only the nucleotides/residues differing from the reference sequence are shown as full characters.
- vi. You can download the input sequences or alignment in various formats, or save the alignment to your Workbench. Click “**Save Analysis**”, give it a name, and click “**Save**”.

VI. Highlight significant positions on a protein structure

IRD imports experimentally-determined virus protein structures from the Protein Data Bank, integrates data from IEDB and UniProt and provides various visualization options. To investigate the structural implications of the sequence variants identified by the IRD meta-CATS analysis, we are going to highlight the positions on a related H7 protein structure.

- a. From the grey navigation bar, mouse over “**Search Data**” and click “**3D Protein Structures**”.
- b. Search for the 3D structures of influenza A HA subtype H7.
Virus Type: A Subtype: H7 Select Proteins to search: 4 HA
- c. The Search Results page displays a list of matching structures. We are going to examine the HA structure of H7N3 subtype, so click “**View Structure**” for 1TI8 to display the structure.
- d. Now we are on the 3D protein structure viewer page. Click and drag with your mouse in display window to change the focus point.
 - i. In the “Display Options” section, you can change the Display Type to line, stick, space, primary structure, secondary structure, etc. We are going to select “**Space**”.
 - ii. Click “**Spin**” to view the structure spinning. Then click “**Rock**” to rock the structure back and forth.
 - iii. You can overlay the structure with a sequence conservation heat map, highlight ligands, immune epitopes, Sequence Features, or specific residues on the structure.
 - iv. This structure is obtained from A/turkey/Italy/214845/02 and the position numbering in our meta-CATS analysis is the same as this strain. Now type in 195, 198, 235 to highlight these binding determinants on the structure.
 - v. T198A is within the 190-helix and related to mammal adapting (Sorrel, 2009). G195V and Q235L could increase the binding of avian H5 and H7 viruses to human-type receptors (Yamada, 2006; Srinivasan, 2013). Q235L is also located within an experimentally determined epitope.
- e. Rotate the structure as you need. The custom highlighted protein structure can be downloaded as an image by clicking “**Save View As Image**” beneath the image, or a 3D movie of either a spinning structure or a rocking structure by clicking “**Generate Video**”.





References

Noronha JM, et al. Influenza virus sequence feature variant type analysis: evidence of a role for NS1 in influenza virus host range restriction. *J Virol.* 2012 May;86(10):5857-66. doi: 10.1128/JVI.06901-11. PMID:22398283

Pickett BE, et al. Metadata-driven Comparative Analysis Tool for Sequences (meta-CATS): an Automated Process for Identifying Significant Sequence Variations Dependent on Differences in Viral Metadata. *Virology.* 2013 (in press)

Sorrel EM, et al. Minimal molecular constraints for respiratory droplet transmission of an avian-human H9N2 influenza A virus. *Proc Natl Acad Sci U S A.* 2009 May 5;106(18):7565-70. doi: 10.1073/pnas.0900877106. PMID:19380727

Srinivasan K, et al. Quantitative description of glycan-receptor binding of influenza A virus h7 hemagglutinin. *PLoS One.* 2013;8(2):e49597. doi: 10.1371/journal.pone.0049597. PMID:23437033

Yamada S, et al. Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors. *Nature.* 2006 Nov 16;444(7117):378-82. PMID:17108965



Section C. Annotate your own virus genome sequences

After this exercise you should be able to use the annotation pipelines provided by the Influenza Research Database (IRD) and Virus Pathogen Resource (ViPR) to annotate your own virus genome sequences.

I. Annotate an influenza virus segment sequence

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

1. Influenza virus segment sequence annotation

You can annotate your influenza sequences using IRD's unique sequence curation/annotation pipeline, which will determine the influenza type, segment number, subtype (if appropriate), and translated amino acid sequence(s) for each segment submitted.

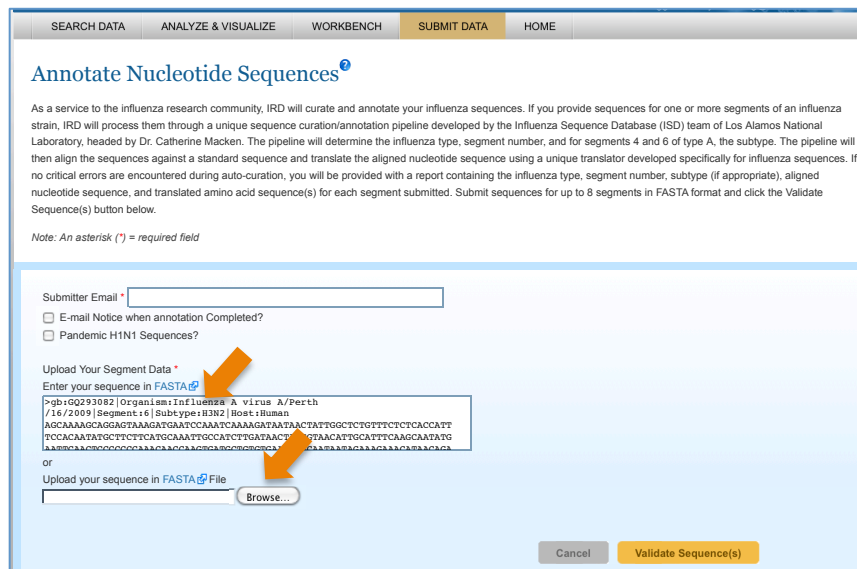
- From the grey navigation bar, mouse over “Analyze & Visualize” and click “Annotate Nucleotide Sequences”.
- If you have your own sequence, prepare the sequence in FASTA format, save it in plain text and use .fasta as the file extension. FASTA file example:

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

AAAGCAGGGGATAATTCTATTAACCATGAAGACTATCATTTGCTTTGAGCTACATTCTATGTCTGGTTTTCGCTCAAAAAC
TTCTTGAAATGACAACAGCACGGCAACGC
```

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

- 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.

- e. If you would like to deposit influenza sequences in GenBank, you can easily submit sequences to GenBank through the IRD site using IRD’s sequence submission utility. To do so, click the “**Submit Data**” tab in the grey navigation bar and follow prompts.

2. H5N1 Clade Classification

IRD has a Highly Pathogenic H5N1 Clade Classification Tool developed by Dr. Catherine Macken’s group at Los Alamos National Laboratory, which can classify the clade of the HA gene of highly pathogenic H5 viruses. The IRD algorithm has been verified as highly accurate (> 99%) for sequences of at least 300 nucleotides of HA1.

- a. From the grey navigation bar, mouse over “**Analyze & Visualize**” and click “**H5N1 Clade Classification**”.
- b. If you have your own H5 sequence, prepare the sequence in FASTA format, save it in plain text and use .fasta as the file extension. FASTA file example:

```
>gb:AM911100|Organism:Influenza A virus A/Anas acuta/Slovenia/470/06|Segment:4|
Subtype:H5N1|Host:Northern Pintail
AGCAAAAGCAGGGTTCAATCTGTCAAATGGAGAAAATAGTGCTTCTTCTTGCAATAGTCAGTCTTGTT
```

Otherwise, use a sample sequence from: <http://tinyurl.com/cer8h3c>

- c. Either paste your sequence in FASTA format to the sequence box or upload your FASTA sequence. Click “**Run**” to proceed.



**Highly Pathogenic H5N1 Clade Classification Tool**

The IRD team has implemented an algorithm for classifying the clade of the hemagglutinin gene of influenza A viruses whose HA belongs to the A/goose/Guangdong/1/96 (H5N1) lineage, that is, the HA lineage of the so-called highly pathogenic H5 viruses. This algorithm was developed by IRD team member Catherine Macken, of Los Alamos National Laboratory. It uses phylogenetic analysis to place HA (H5) sequences within the WHO classification scheme presented [here](#). The IRD algorithm has been verified as highly accurate (> 99%) for sequences of at least 300 nucleotides of HA1. See [SOP](#) for more details.

This tool only handles segment 4 sequences with confirmed H5 serotype and lengths greater than 300 nucleotides. Sequences from other serotypes of HA, or other segments will yield unpredictable and likely incorrect results. If unsure of your sequence's segment or serotype, we suggest you use the IRD Sequence Annotation Tool found on the Analyze and Visualize menu by clicking the Annotate Nucleotide Sequences link.

INPUT SEQUENCES

Upload a file containing my sequences in FASTA [format](#).

Paste sequences in FASTA [format](#).

```
>gb|AM911100|Organism:Influenza A virus A/Anas_acuta/Slovenia/470/06 [Segment:4] [Subtype:H5N1] [Host:Northern Pintail]
AGCAAAAGAGGGGTTCACTCTCCGAAATGGAGAAATAGGCTCTCTCCCAATGTCAGTCTGTT
AAAGATGATGAGATTCGATTTGTTACATGCAACACATCCAGAGAGAGTTGACAGAAATAGAAA
AGAGAGCTGCTGTTTACAGCTCCGACAGCTATGGAGAAAGACAGACAGAGGAAATCTGCGTCTG
TGGAGTGAAGCCCTCAATTTTAMGAGATGTGATGATGAGTGGATGCGCTCCGGAGACCAATGTGKAC
GATTCCTCAATVCCGGAAAGGCTTACATATGGAGAGATCAATCCAGCCAGGCCCTGTTTACC
TAGGAGATTTACGCTATGAGAAATGAGACCTCTGTCAGATGATACCTGTGGAGAAATCTC
```

Clear Run

- d. After the annotation process is finished, a H5N1 Clade Classification Report page will be loaded. Here you will see the clade assigned to your input sequence. You can download the report by clicking the “**Download Raw Result**” button.

H5N1 Clade Classification Report

Save Analysis Download Raw Result

Sequence Identifier	Clade Assignment
gb AM911100 Organism:Influenza_A_virus_A/Anas_acuta/Slovenia/470/06 Segment:4 Subtype:H5N1 Host:Northern_Pintail	2.2.1

II. Annotate a Hepatitis C Virus (HCV) genome sequence

For this exercise, you will use ViPR (www.viprbrc.org) to annotate a Hepatitis C Virus genome sequence, determine its genotype and identify sites of recombination if applicable.

ViPR provides a Genome Annotator (GATU) to help you annotate your own virus genome sequences. To use GATU, you will need to select a previously annotated reference sequence and then use GATU to transfer the annotations to a target genome sequence.

1. Annotating an HCV genome sequence

- Go to www.viprbrc.org and click “**Hepatitis C Virus**” to get to the HCV page.
- Mouse-over the “**Analyze & Visualize**” tab from the grey navigation bar and click “**Genome Annotator (GATU)**”.
- In order to annotate your own sequence, you need to select a previously annotated reference sequence. If you already have an annotated reference sequence in .gb format, click “**Launch GATU**” to proceed directly to launch GATU. If not, you can use ViPR BLAST to search for a closely-related annotated sequence as your reference.

- i. 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:

```
>gi|375127704|gb|JN714194.1| Hepatitis C virus subtype 3a isolate
RASILBS2-SR-PO polyprotein gene, complete cds
AGATACCTGCCTCTTACGAGGCGACACTCCACCATGGATCACTCCCCTGTGAGGAACTTCTGTCTTCACGCGG
AAAGCGCCTAGCCATGGCGTTAGTACGAGTGTCTGTGCAGCCTCCAGGACCCCCCTCCCGGGAGA
```

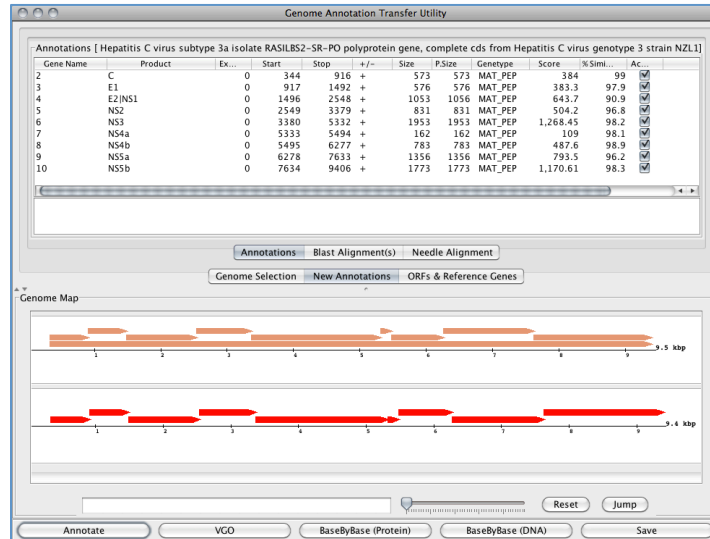
Otherwise, you can use a sample sequence from: <http://tinyurl.com/7mfry6e>

- ii. Click “**Browse**”, find the target sequence file on your computer, and click “**Go**” to run a BLAST search again annotated HCV reference sequences in ViPR.
- iii. After BLAST is finished, a list of recommended reference sequences will be displayed. Choose a closely-related sequence and download its GenBank file to your computer.

Download GenBank File	Sequence header	Bit Score	E Value
EXT445955	>gi 445955 Country Hepatitis C virus genotype 3, genome .gb 157781216	9587	0.0
EXT383780	>gi 383780 Country Hepatitis C virus genotype 1, complete genome .gb 22129792	896	0.0
EXT446165	>gi 446165 Country Hepatitis C virus genotype 2, complete genome .gb 157781212	767	0.0
EXT446144	>gi 446144 Country Hepatitis C virus genotype 4, genome .gb 157781208	755	0.0
EXT445982	>gi 445982 Country Hepatitis C virus genotype 6, complete genome .gb 157781214	737	0.0

- d. Now, click “**Launch GATU**” to run the GATU application. A dialog box will pop up. Click “**Allow**” to allow the GATU applet to be loaded on your computer.

- e. In the GATU window, upload your .gb file as the “Reference Genome” and your target genome FASTA file as the “Genome to Annotate”.
- f. Click “Annotate” to execute annotation process. When done, a table is displayed which summarizes the similarities of transferred annotations and provides users with checkbox control over which to accept.



- g. Click the “Save” button to save the annotated target genome in multiple file formats: Genbank, EMBL, or XML.

2. HCV genome sequence genotype determination and recombination detection

ViPR provides a Genotype Determination and Recombination Detection Tool for Hepatitis C virus, Dengue virus, St. Louis Encephalitis virus, West Nile virus, Japanese Encephalitis virus, Tick-borne Encephalitis virus, Yellow Fever virus, Bovine viral diarrheal virus, and Murray Valley encephalitis virus. This tool estimates the most likely genotype for the input sequences and identifies sites of recombination.

- a. Mouse-over the “Analyze & Visualize” tab from the grey navigation bar and click “Genotype Determination and Recombination Detection”.



- b. On the “Genotype Determination and Recombination Detection” landing page, select “HCV” from the “Select Species” drop-down list.
- c. Download a sample HCV sequence file from: <http://tinyurl.com/7vanutq>
- d. Input your sequence by uploading the FASTA-formatted sequence file or pasting the FASTA-formatted sequence in the box. Then click “Run”. Note that you can also input sequences from a working set saved in your Workbench.
- e. After the analysis is finished, the Report page will be displayed. Here you can:
 - View the predicted genotype and recombination type (if applicable).
 - Download a spreadsheet listing the detailed results of recombination determination.
 - View the genotyping results in graphical format.
 - Download or view the alignment of your sequence with representative sequences from each taxon selected by ViPR.
 - Download or view the phylogenetic tree based on the alignment of your sequence with representative sequences from each taxon selected by ViPR.

SEARCH DATA ANALYZE & VISUALIZE WORKBENCH SUBMIT DATA VIRUS FAMILIES HOME *Flaviviridae*

Home > Genotype determination and Recombination detection > Results

Genotype Report

Save Analysis Run Analysis ▼

Genotype Information

Whole Genome Genotype prediction: 2a
Whole Genome Recombination Type: 2a.2b

Genotype

The genotype results include a tab separated file listing the sequence name, a single consensus genotype result for the entire genome, and the confidence metric.

Download

Recombination

This is an excel spreadsheet listing the results for all windows for the sequence.

Download

Genotyping results in graphical format

Branching index profile for JF343783

Alignment position (nt)

Alignment

This is the multiple sequence alignment of your sequence with a ViPR reference sequence alignment that consists of at least 2 representatives from each taxon

Download Aligned Fasta Visualize Aligned Sequences

Tree

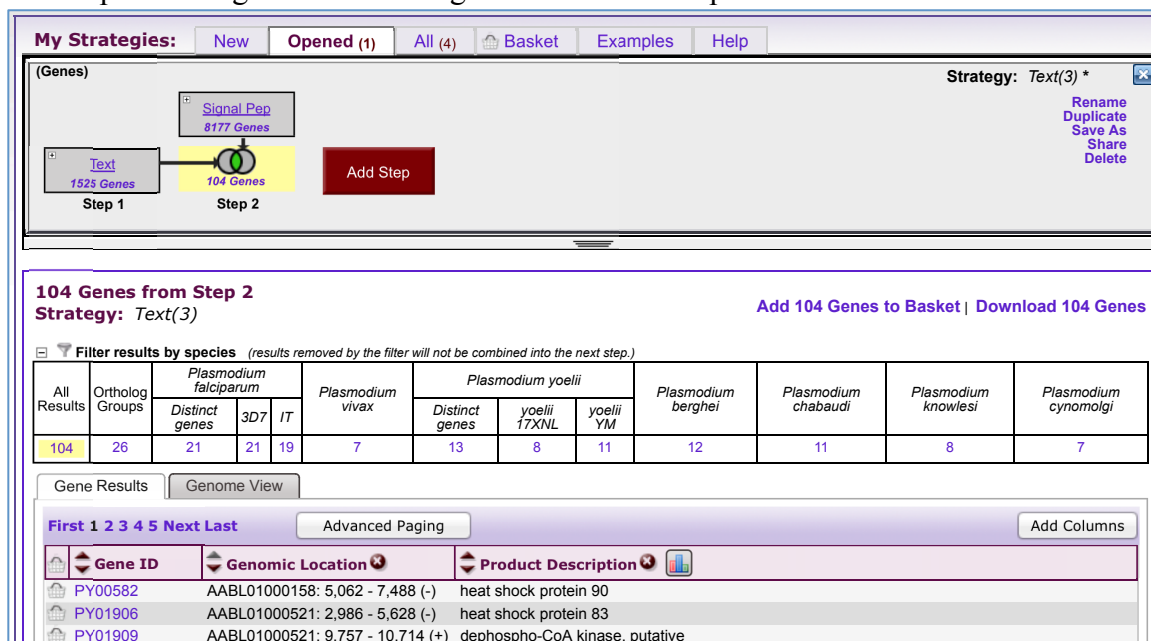
This is the tree generated by PAUP based on the input alignment for the whole genome

Download Newick File View phylogenetic tree

PlasmoDB Exercise: Finding Genes and Exploring the Gene Page

Upon completion of this exercise, you will be able to find genes on the PlasmoDB website and explore gene details.

1. Go to the PlasmoDB homepage (<http://www.plasmodb.org>).
2. Find all possible kinases in Plasmodium.
 - a. In the “Identify Genes by” column, click “Text, IDs, Organism”, then “Text”.
 - b. On the Identify Genes based on text page, select all organism, type “kinase” in the text term box and then add wildcard “*” to both ends of the word (i.e., “*kinase*”) to retrieve genes such as “phosphofructokinase” or “kinases”, select all Fields except “similar proteins”. Click “Get Answer”.
 - c. The next page displays the My Strategies panel in the middle (Text search shown as Step 1) and the Results panel at the bottom.
3. Identify kinases that are likely secreted, i.e., genes with likely secretory signal peptides.
 - a. Click “Add Step”.
 - b. In the pop up box, click “Run a new search for” -> “Genes” -> “Cellular Location” -> “Predicted signal peptide”.
 - c. Next, combine results from Step 1 with Step 2 using “1 Intersect 2”.
 - d. Click “Run Step”.
 - e. The returned page shows the number of possible secreted kinases in the Strategies panel along with the list of genes in the bottom panel.



104 Genes from Step 2
Strategy: Text(3) Add 104 Genes to Basket | Download 104 Genes

Filter results by species (results removed by the filter will not be combined into the next step.)

All Results	Ortholog Groups	Plasmodium falciparum			Plasmodium vivax	Plasmodium yoelii			Plasmodium berghei	Plasmodium chabaudi	Plasmodium knowlesi	Plasmodium cynomolgi
		Distinct genes	3D7	IT		Distinct genes	yoelii 17XNL	yoelii YM				
104	26	21	21	19	7	13	8	11	12	11	8	7

Gene Results | Genome View

First 1 2 3 4 5 Next Last Advanced Paging | Add Columns

Gene ID	Genomic Location	Product Description
PY00582	AABL01000158: 5,062 - 7,488 (-)	heat shock protein 90
PY01906	AABL01000521: 2,986 - 5,628 (-)	heat shock protein 83
PY01909	AABL01000521: 9,757 - 10,714 (+)	dephospho-CoA kinase, putative

4. Visiting a specific gene page.
 - a. From the gene list, pick a gene (“heat shock protein 90” in this case) and click the Gene ID.
 - b. You are directed to the Gene page.
 - i. Write down the location of the gene on the genome.
 - ii. What genes are located upstream of this gene in *P. yoelii*?

PY00582
heat shock protein 90

Previous ID(s): 159.m00048

[Add the first user comment](#)
[Add to Basket](#)
[Add to Favorites](#)

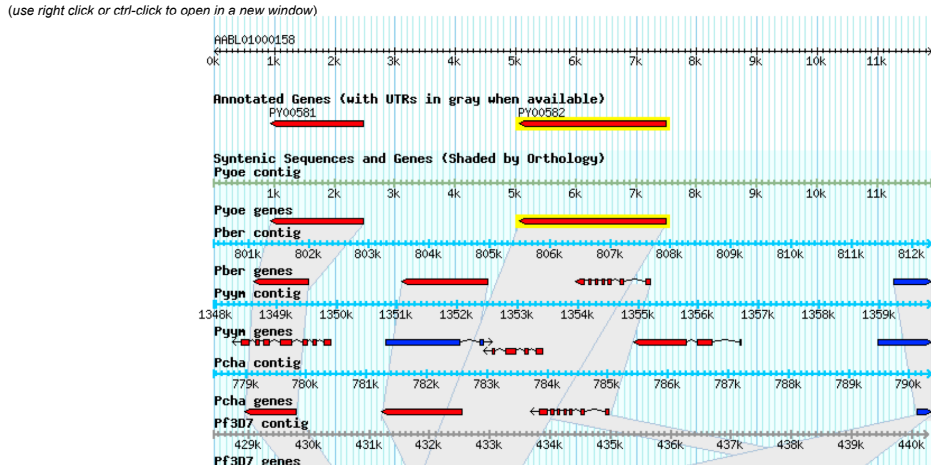
Overview

P. yoelii yoelii 17XNL protein coding gene on AABL01000158 from 5,062 to 7,488 (Chromosome: Not Assigned)

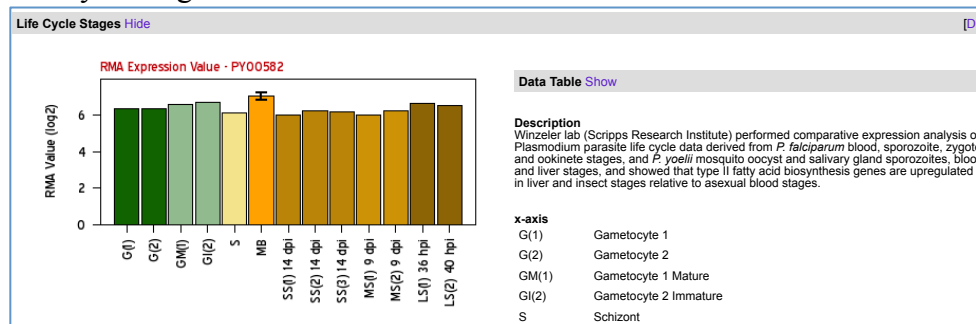
Genomic Context [Hide](#)

[View in Genome Browser](#)

(use right click or ctrl-click to open in a new window)



- iii. Look at the Protein section, what kind of data in PlasmoDB provides evidence for the expression of this gene?
- iv. Now view the Expression section. Is the gene more abundant at certain life cycle stages?



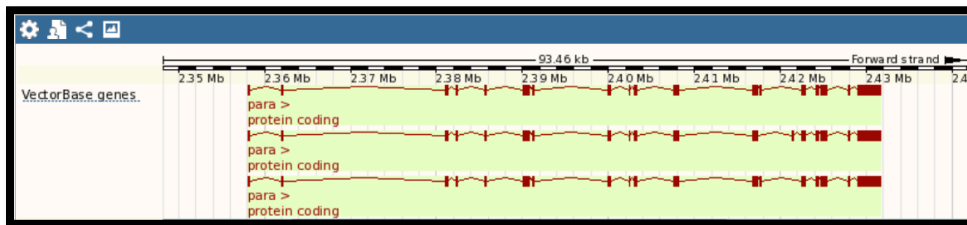
- v. Go back to the Genomic Context section and click “View in Genome Browser”.
- c. The GBrowse page has several tabs at the top. In the Browser window, the Search panel offers several options for you to navigate to a specific region of the genome; the Overview panel gives a bird-eye view of the genome; the Region panel displays the selected region; the Details panel displays the annotation tracks for the selected region.
- i. Click the “Select Tracks” panel, and then select “Synteny” and “Protein expression Evidence” – “P.yoelii”. Click “Back to Browser”.
 - ii. Now you will see the selected tracks are added to the Details panel. What kind of data supports the expression of the gene? Is synteny (chromosome organization) in this region maintained in other species?



VectorBase Exercise: Visualizing Variations in Genomic DNA

Upon completion of this exercise, you will be able to find a gene on the VectorBase website and view its sequence variations.

1. Go to the VectorBase homepage (www.vectorbase.org).
2. In the search box at the top-right corner of the page, type in the gene symbol or gene ID—“para” or “AGAP004707” in this case, and click the “GO” button.
3. On the Search Results page, click on the first result--named “para (AGAP004707)”.
 - a. Notice that the Search Results page lets you apply additional filters using the criteria on the left side of the page (e.g. domain or species).
4. The Gene Browser page displays information about the gene as well as a visual representation of the corresponding introns and exons in the genome.



5. Click on the “Sequence” link in the navigation box on the left side of the page to load the sequence for this gene.
 - a. Note that the exons will be displayed in red text, while the introns will be displayed with black text.
 - b. To display variation within this gene:
 - i. Click on the “Configure this Page” option—in a blue box located at the bottom of the navigation box (on the left side of the screen).
 - (1) Change the “Show Variations” display option from “No” to “Yes and show links”.
 - (2) Change the “Line Numbering” display option from “None” to “Relative to this Sequence”.
 - (3) Click on the checkmark at the top-right side of the pop-up window.
 - ii. Now you will see variations highlighted with various colors representing upstream/downstream, intronic, missense, splice region, or synonymous variations.
 - (1) Links to the right of each sequence variation provide additional information about the source and attributes of the displayed variant.

```

73321 ATCCGACGGTACACAATACGTTTCGATATGATCAGCTTCAGACTTTTGGATGTGCTGG 73380 73358: rs180291684;
73381 AACCGCCTCTACAGATTCATAAACCAATCGTTATAAGATTATTCGATGGATATCCGA 73440
73441 TATGCCGCGGAGATATGATTTCTGTGTCGATATCTAGATGCCTAACGAAAGATTTT 73500 73448: 2L.2431005;
73501 TTGCTAGAAAGGAAATCCTATAGAAGAACAGCCGAATTAGGTGAAGTCAACAACGCC 73560
73561 CAGACGAAGTGGTTACGAACCAATATCATCAACACTTTGGAGGCAGCGTGAAGAGTACT 73620
73621 GTGCTCGAGATACAATCATGCGTGGAAACGCTATAAACAGCGTCACGGAGGCGAACAG 73680 73637: 2L.2431194;
73681 ACGCTTCAGAGATGATCTTGAATAGATGCCGTGTGATAACCGTTGGTGGTGGTAAATG 73740
73741 GCAATGAAATGATGATAGTGGAGATGGTCAACAGGTAGTGGTACAAACGGAAGTCAAC 73800 73799: rs5181107;
73801 ATGGTGGTGGCAGCATAAGTGGCGGAGGAGGAACCTCGTGGTGGTAAAAGTAAAGGAR 73860 73860: WTSI-Ag-GVP-0.1-SNP-2L-2431417
73861 TTATTGGCAGTACTCAGGTAAACATAGGCAATAGTGGATAGTAATATATCACCAAAGGAAT 73920
73921 CACCGGATAGCATCGCGGATCCCAAGGTCGTGACAGCGCCGCTCTTGTGGAGAGCGACG 73980
73981 GATTGTGACGAAAACGGTCCCGTGTGCTCATACTCTCGATCTCCAGCATAACAT 74040
74041 CGCGAACCGCAGATGCTGAGCCAGGTCTCGCCCCCTCTCGGATTCAGATTCGGAAG 74100
74101 CACCACGAAATAATATTTGAAATGACATGCAATGTAAGGTTTAAAGCATCAAAGAACAT 74160 74105: WTSI-Ag-GVP-0.1-SNP-2L-2431662

```

PATRIC Exercise: Finding a gene on the PATRIC website

Upon completion of this exercise, you will be able to select an organism on the PATRIC website and search for a gene.

A. From the home page click on the organism tab (1), then on *Brucella* (2).

B. On the *Brucella* genus landing page, click on Feature Finder (3).

C. On the Feature Finder page, enter the name of a gene, like Propionate CoA-transferase (4) and click Search (5).

F. Clicking on the Pathways tab (*) of the feature page shows all pathways (a) that the gene is involved in.

Pathway ID	Pathway Name	Pathway Class	Annotation	EC Number	Occurrences	Description
0060	Pyruvate metabolism	Carbohydrate Metabolism	PATRIC	2.8.3.1	1	Propanoate CoA transferase
0040	Essential metabolism	Carbohydrate Metabolism	PATRIC	2.8.3.1	2	Propanoate CoA transferase
0043	Gene annotation	Metabolic Biodegradation and Metabolism	PATRIC	2.8.3.1	1	Propanoate CoA transferase

Clicking on a specific Pathway (a, above) takes you to the pathway summary, mapped onto a KEGG pathway, for the genome you are exploring, with the gene you are interested in highlighted in blue (b).

KEGG Map

EC Table

EC Number	Genome G	Feature C	Genome Count	Occurrences
1.1.1.37	1	1	0	1
1.1.1.38	1	1	0	1
1.1.1.40	1	1	0	1
1.1.1.79	1	1	0	1
1.1.2.3	1	1	0	1
1.1.2.4	1	1	0	1
1.1.2.5	1	1	0	1
1.2.1.22	1	3	0	2
1.2.3.3	1	11	0	1
1.2.4.1	1	2	0	2
1.8.1.4	1	1	0	1
2.3.1.12	1	1	0	1
2.3.1.9	1	1	0	1
2.3.3.13	1	1	0	1
2.3.3.9	1	1	0	1
2.7.1.40	1	1	0	1
2.7.2.1	1	1	0	1
2.7.9.1	1	1	0	1
2.8.3.1	1	1	0	1
3.1.2.6	1	3	0	1
4.1.1.	1	3	0	1
4.1.1.49	1	2	0	1
4.1.3.	1	1	0	1
4.2.3.3	1	1	0	1
4.4.3.5	1	1	0	1
6.2.1.1	1	2	0	2
6.4.1.1	1	1	0	1
6.4.1.2	1	2	0	1

G. Clicking on Transcriptomics (*) shows all the experiments, recently collected and curated from GEO, in which this gene is expressed. On this page you can use filters to search for specific keywords (a), apply specific cut-offs to see in which experiments the gene is significantly expressed (b), a log ratio or Z-score distribution graph of the experiments where the gene is significantly expressed (c) a pie chart/bar chart of key metadata attributes (d) and a table that provides a summary of all the comparisons that match the filtering criteria the researcher has chosen (e).

Bacteria • Proteobacteria • Alphaproteobacteria • Rhizobiales • Brucellaceae • Brucella • Brucella melitensis bv. 1 str. 16M • VBIBruMel92729_0021, Propionate CoA-transferase (EC 2.8.3.1)

Genome Browser Compare Regions View Pathways **Transcriptomics** Correlated Genes Literature

a summarizes the transcriptomics data for this gene. The list of comparisons (and respective visual summaries) can be filtered by keyword, log ratio and Z-score. To learn more, see [Gene Page Transcriptomics](#).

b

10 comparisons

keyword: [] | Log Ratio: 0 | Z-score: 0 | Filter | Reset Filter | Show All Comparisons

c

d

Download table in Excel file (.xlsx) | Text file (.txt)

Title	PubMed	Accession	Strain	Gene Modifi	Experimental Condition	Time Point	Avg Intensit	Log Ratio	Z-score
log phase / stationary phase	19419566	GSE11192	16M		growth phase		0	0	0
Brucella melitensis 16M vjBr mutant strain / Brucella...	20387905	GSE8844	16M	vjBr	mutant vs wild type		8.564	0.07	0.167
Brucella melitensis 16M babR mutant strain / Brucell...	20387905	GSE8844	16M	babR	mutant vs wild type		8.526	-0.006	-0.037
Brucella melitensis 16M, logarithmic phase of growt...	20529360	GSE13634	16M		C12-HSL		8.732	0	-0.008
Brucella melitensis 16M, stationary phase of growth...	20529360	GSE13634	16M		C12-HSL		7.709	-0.24	-0.686
Brucella melitensis 16MvBr, logarithmic phase of gr...	20529360	GSE13634	16M	vjBr	C12-HSL		8.475	-0.692	-1.197
Brucella melitensis 16MvBr, stationary phase of gro...	20529360	GSE13634	16M	vjBr	mutant vs wild type		7.708	-0.351	-0.901
Brucella melitensis 16MvBr, logarithmic phase of gr...	20529360	GSE13634	16M	vjBr	mutant vs wild type		8.912	0.148	0.387
Brucella melitensis 16MvBr, stationary phase of gro...	20529360	GSE13634	16M	vjBr	C12-HSL		0	0	0
Brucella melitensis 16Mmuckr, late logarithmic phase...	GSE13647	16M	mucR	mutant vs wild type			8.742	0.273	0.539

e

H. Clicking on Correlated genes tab (*) will show a list of genes that have correlated expression profiles (positively or negatively) across all available data sets along with their functions. The correlation coefficient for each of the correlated genes is provided (a).

Bacteria • Proteobacteria • Alphaproteobacteria • Rhizobiales • Brucellaceae • Brucella • Brucella melitensis bv. 1 str. 16M • VBIBruMel92729_0021, Propionate CoA-transferase (EC 2.8.3.1)

Genome Browser Compare Regions View Pathways **Transcriptomics** **Correlated Genes** Literature

Correlations Graph | Comparisons | (0) | (16) | (16)

460 features found

Gene Name	Accession	Length (bp)	Start	End	Strand	Product Description	Correlation	Comparisons
Brucella melitensis bv. 1 str. 16M	NC_003312	20776	2266	1491	+	Propionate CoA-transferase (EC 2.8.3.1)	1	0
Brucella melitensis bv. 1 str. 16M	NC_003318	107304	107498	1045	+	putative glutathione-regulated potassium-efflux system protein hskB	0.99	0
Brucella melitensis bv. 1 str. 16M	NC_003319	67979	68054	876	+	Transcriptional regulator, ArcA family	0.989	0
Brucella melitensis bv. 1 str. 16M	NC_003317	45182	45179	288	+	FKBP490346 hypothetical protein	0.984	0
Brucella melitensis bv. 1 str. 16M	NC_003317	58711	58705	345	+	FKBP490314 hypothetical protein	0.872	0
Brucella melitensis bv. 1 str. 16M	NC_003319	20551	20520	389	+	COG3_C02812	0.07	0
Brucella melitensis bv. 1 str. 16M	NC_003319	64209	64201	963	-	Proteobacteriaceae ABC transporter, permease protein 1	0.959	0
Brucella melitensis bv. 1 str. 16M	NC_003318	58332	58340	309	+	Endonuclease V (EC 3.1.25.1)	0.959	0
Brucella melitensis bv. 1 str. 16M	NC_003319	1804	1801	488	+	Putative activity regulator of membrane protease hskB	0.968	0
Brucella melitensis bv. 1 str. 16M	NC_003318	42912	43004	433	+	Oxidative hydroperoxide resistance protein	0.966	0
Brucella melitensis bv. 1 str. 16M	NC_003317	94393	94376	305	+	Nucleosuccinate kinase protein	0.965	0
Brucella melitensis bv. 1 str. 16M	NC_003319	44032	44004	1833	+	Putative regulatory protein hskB	0.962	0
Brucella melitensis bv. 1 str. 16M	NC_003317	139123	139145	153	+	Salivary histone H4 nucleosome precursor (EC 3.2.1.1)	0.959	0
Brucella melitensis bv. 1 str. 16M	NC_003318	7497	7505	799	+	COG2041 Salivary nuclease and related enzymes	0.956	0
Brucella melitensis bv. 1 str. 16M	NC_003319	107495	107492	882	-	hskB203 protein hskB	0.956	0
Brucella melitensis bv. 1 str. 16M	NC_003317	123905	123945	441	-	Brucella melitensis inducible transcriptional regulator hskR	0.955	0
Brucella melitensis bv. 1 str. 16M	NC_003317	173292	173256	105	-	hypothetical protein	0.951	0
Brucella melitensis bv. 1 str. 16M	NC_003317	382049	382031	483	-	Transcriptional regulator hskB of prokaryotic cadherins (TMR family)	0.951	0
Brucella melitensis bv. 1 str. 16M	NC_003317	49823	49804	142	-	Blood ABC transport system, ATP-binding protein hskA (EC 3.1.2.1)	0.95	0
Brucella melitensis bv. 1 str. 16M	NC_003317	158804	158860	627	-	ATP synthase F1 chain (EC 3.6.3.14)	0.945	0

a