



INFLUENZA RESEARCH DATABASE & VIRUS PATHOGEN RESOURCE EXERCISES

Table of Contents

Section A	A. Get familiar with the IRD/ViPR site	1
Section E	3. 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	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

Influenza Research Databas	About Us	Community	Announceme You ai	Ints Links Resources Support Sign Out
SEARCH DATA ANALYZE & VISUALIZE	WORKBENCH SUBN	IT DATA	HOME	
Search Search our comprehensive database for: 4 Influenza segment and protein sequences 4 Avian and non-human mammalian surveillance data 4 Virus phenotypic characteristics 4 Influenza strain information 5 Inmune epitope data 5 3D protein structures Browse All Search Types	Analyze Analyze data online: Align sequences Align sequences Analyze Sequence Vari Analyze Sequence Vari Generate a phylogeneti Browse All Tools	es (BLAST) n flu proteins lation (SNP) ic tree	•••••	Save to Workbench Use your workbench to: Store sequences or other data in working sets for future analysis Combine working sets Integrate IRD data with your laboratory data Store analysis results Share results

1





- c. Above the light blue box is a grey navigation bar consisting of the following tabs: Search Data, Analyze & Visualize, Workbench, Submit Data, and Home. These tabs are consistent across the IRD site and are designed to help you navigate the site.
 - i. Mouse-over or click the "Search Data" tab to view available data search options.
 - ii. Mouse-over or click the "Analyze & Visualize" tab to view analysis and visualization tools provided by IRD.
- d. In the blue banner, there is a compilation of links to useful resources.
 - i. Pull down the "About Us" menu and view Citing IRD, Our Publications, and Research Using IRD.
 - ii. Pull down the "Announcements" menu and view Meetings and Events and IRD Newsletters.
 - iii. Pull down the "**Resources**" menu and view the WHO vaccine strains, BEI reagent resources, anti-viral drug information, Reactome flu pathways, and other resources.
 - iv. Click the "**Support**" menu and view how to contact the IRD team when you have questions, suggestions, or problems.
 - v. Pull down the "**Support**" menu and click the "**IRD 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. View the YouTube video.

II. Getting familiar with the Virus Pathogen Resource

a. Go to the ViPR homepage (<u>http://www.viprbrc.org</u>) 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; <u>http://www.fludb.org/</u>) 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 sequenes 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

5





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

1. Search for H7N9 HA sequences using structured search interfaces

- a. Go to the IRD homepage (<u>http://www.fludb.org/</u>), mouse-over "Search Data" in the grey navigation bar, then "Search Sequences" and click "Nucleotide Sequences".
- b. 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.
- c. 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:Image: AComplete Sequences:Image: Complete Sequences OnlySelect Segments:4 HAAdvanced Options:Image: Remove Duplicate SequencesSub Type:H7N9Image: H7N9Image: H7N9

d. 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.

		WORKPENCH		
SEARCH DATA AN	ALTZE & VISUALIZE	WORKBENCH 3	SOBINIT DATA	
tome Mucleotide Sequence Sea	irch	1 0		
Search for influenza sequences and you can pick your viewing	s, proteins, and strains us options.	n –	Jse the advanced search to allow you to refin	e your search with the more fine grained search
DATA TO RETURN	SELEC	T SEGMENTS	HOST	GEOGRAPHIC GROUPING
Segment / Nucleotide Protein Strain VIRUS TYPE A B	All 1 PB2 2 PB1/f 3 PA/P 4 HA 5 NP 6 NA 7 M1/h 8 NS1/	№1-F2 ۹-X 42 NS2	All Avian Bat Camel Cheetah Chicken Civit Dog Domestic Cat	All Aria Aria Europe North America Oceania COUNTRY Afohanistan
C SUB TYPE H7N9 Use comma to separate mentries. Ex: H1N1, H7, H3N2.	Jtiple COMP	LETE SEQUENCES plete Sequences only lude near-complete juences (IVR)	Environment Ferret Horse Human Lab	Algeria Angola Argentina Australia Austria
STRAIN NAME [•] Use comma to separate m entries. Ex: A/chicken/Israel/1055/20 A/chicken/Laos/16/2008.	Jiliple Status (SOP) Ilicius Ilicius Ilicius Inclu	H1N1 SEQUENCES de pH1N1 sequences de all pH1N1 sequences de all pH1N1 sequences RANGE YYY To: YYYY nonth to search, see		
* ADVANCED OPTION	Advance	Options: wonth Range	Tip: To select multiple or desele	ct, Ctrl-click (Windows) or Cmd-click (MacOS)
Select Advanced Option	REN	OVE DUPLICATE SE	QUENCES	× Remove
Remove Duplicate Seque	nces 🗹 R	emove Duplicate Sequence ucleotide sequences are exa	s - Only include one of the results if multiple actly identical.	
 Add Another Advanced O 	ption			Clear Search

- e. The Search Results page will be displayed. Here you can:
 - i. Save the search query to your Workbench and rerun the search again later.
 - ii. 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.

	SEA	RCH DATA	ANA	ALYZE & VISU	ALIZE	WORKB	ENCH	SUBMIT DAT	A				
Hom N	∎ • <u>Nuc</u>	leotide Seque leotide	ence Sea Seq	uence S	Search	Rest	ılts						
	Your S	ieles i Item Id to Workin	s: 59 iten g Set	ns selected Save Sea	Deselect / arch	Run Ana	lysis v	Downloa	ł				
Yo	ur sea Seleo	rch returned i	59 segme ents	nts. Searc	h Criteria		C	Displaying 50 re	cords per page , sor	ted by Date in	descending	Disp	Page: 1 of 2
¢	ø	Segment	Protein <u>Name</u>	Sequence Accession	Complete Genome	Segment Length	Subtype *	Date	Host Species	Country	State/Province	<u>Flu</u> <u>Season</u> (SOP)	Strain Name
	☑ <u>Vie</u>	<u>w</u> 4	HA	CY146908 *	Yes	1683	H7N9	05/03/2013	Chicken/Avian	China	-N/A-	-N/A-	*A/chicken /Guangdong /SD641 /2013(H7N9)
0	Vie	w 4	HA	CY147100	Yes	1683	H7N9	05/03/2013	Environment	China	-N/A-	-N/A-	A/environment /Shandong /SD038/2013
6	✓ Vie	<u>w</u> 4	HA	CY147108	Yes	1683	H7N9	05/03/2013	Environment	China	-N/A-	-N/A-	A/environment /Shandong /SD039/2013
6	⊻ <u>Vie</u>	<u>w</u> 4	HA	CY146948	Yes	1683	H7N9	05/03/2013	Chicken/Avian	China	-N/A-	-N/A-	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**".

SEARCH DATA	ANALYZE & VISUALIZE	WORKBENCH	SUBMIT DATA			
Home Mucleotide Sequence	e Search + Results + Identify S	milar Sequences (BLA	ST)			
Identify Sim	ilar Sequences	(BLAST) [®]				
	1	()				
Compare sequences you	provide or select from the IRD	database consisting of	selected sets of seque	ences from the IRD da	atabase or create your own data	abase for comparison
from a working set on you	ur workbench.					
Note: An asterisk (*) = red	quired field					
FORMAT OF SEQ	UENCES PROVIDED	SELECT DATAB	ASE TO SEARCH	I* INP	UT SEQUENCES	
 Blastn (nucleotide))	Use generated day	itabase	1 rec	cord was previously selected fro	m search
Blastx (search pro	, tein usina nucleotide	Nucleotides for seg	ment 3 PA	resul	lts	
translated in 6 rea	ding frames)	Nucleotides for seg	ment 4 H1			
		Nucleotides for seg	ment 4 H3 ment 4 H5			
		Nucleotides for seg	ment 4 HA			
		Tip: To select multiple	or deselect, Ctrl-click	(Windows)		
		or Cmd-click (MacOS	i)			
		 Use working set to 	run blast			
	_					
OUTPUTFORMA	1					
Remove data from c	renetically manipulated strains	from the result				
Number of Results to D	Display for each Input Sequence	8				
10 25						
50						
▶ ADVANCED OP	TIONS Show All					
<u>La trata La cr</u>						
Select Advanced Op	tion OTH	ER PARAMETER	S			× Remove
Select An Advanced	Option Stran	d	Gap C	osts		
	Both	(Default)	Existe	ence: 0 Extension: 4		
	Bott	om	Existe	ence: 6 Extension: 2		
			Existe	ence: 5 Extension: 2		
	Appl	/ Low Complexity Filt	er			
	True	a				
	Desfe	rm ungapped aligner	ont			
	Feite	ini ungappeu anginn	onc			
+ Add Another Advand	ced Option					
					_	
					Clea	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?

	SEARCH [OATA ANALYZE & VISUA	LIZE WORKBENCH	SUBMIT DATA					
lome	Nucleotide	e • Results • Identify Similar Se	quences (BLAST) Results						
BI	LAST	Report							
BLA Lipn gb:K	STN 2.2.2 nan (1997) F021597 C	2 [Sep-27-2009] Reference: Alt , "Gapped BLAST and PSI-BL/ Drganism:Influenza#A#virus#A/S	schul, Stephen F., Thomas L AST: a new generation of pro hanghai/02/2013 Segme Data	Madden, Alejandro A otein database search ibase: Influenza nucleo	A. Schaffer, Jing programs", Nu tide sequences :	hui Zhang, Zheng Zhang, W cleic Acids Res. 25:3389-340 Seg4 70,570 sequences; 95,79	Vebb Miller, a 02. Query: 90,805 total l	and Dav	id J.
	Cours Amol	A del én Montrine G	2.04						
	Save Anal	ysis Add to Working S	Set						
•	Save Anal	ysis Add to Working S	Set	Sequence header				Bit Score	E Value
	Save Anal Id 982104	ysis Add to Working \$ >ird 982104 Country:C	Set Shina Influenza A virus (A/Sha	Sequence header anghai/02/2013(H7N9)) cds. gb KF021597	segment 4 hema	gglutinin (HA) gene, complete	e	Bit Score 3386	E Value 0.0
	Save Anal Id 982104 970288	sis Add to Working \$ >ird[982104] Country:C >ird[970288] Country:China	set hina Influenza A virus (A/Sha () Influenza A virus (A/Zhejiang)	Sequence header anghai/02/2013(H7N9)) cds. gb KF021597 y/DTID-ZJU01/2013(H7I cds. gb KC885956	segment 4 hema N9)) segment 4	gglutinin (HA) gene, complete nemagglutinin (HA) gene, com	e nplete	Bit Score 3386 3378	E Value 0.0
	Id 982104 970288 973128	sis Add to Working S	iet ihina Influenza A virus (A/She () I Influenza A virus (A/Zhejjang) () Ifluenza A virus (A/Fujjan/1/20)	Sequence header anghai/02/2013(H7N9)) cds.]gb]KF021597 g/DTID-ZJU01/2013(H7I cds.]gb]KC885956 113(H7N9)) segment 4 h	segment 4 hema N9)) segment 4 hemagglutinin (H	gglutinin (HA) gene, complete nemagglutinin (HA) gene, com A) gene, complete cds.]gb KC	e nplete 2994453	Bit Score 3386 3378 3328	E Value 0.0 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

- a. 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.
- b. 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.
- c. 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".
- d. 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".
- e. Access your Workbench by clicking the "**Workbench**" tab. You will see the newly created working sets at the top of the content list.

AILABLE ACTI	ONS	
Download	Downloa	d data of selected working set
Сору	Copy sel	lected working sets
Combine	Combine	e selected working sets into a new working set
Intersect	Create in	tersection of selected working sets into a new working set
Create Group	Add a co	ilaborator group
Upload File	Upload a	fie
Edit Group	Edit a col CTIONS a available to you	llaborator group u for the items selected. To utilize the disabled actions, please go back and select different iten
Edit Group	Edit a col CTIONS a available to you	laborator group u for the items selected. To utilize the disabled actions, please go back and select different item Create a new working set for subtract one working set from another
Edit Group NAVAILABLE Av Ny certain actions an Subtract Convert	Edit a col CTIONS a available to you	laborator group u for the items selected. To utilize the disabiled actions, please go back and select different item Create a new working set by subtract one working set than another Convert the selected working set than working set than different type
Edit Group NAVAILABLE A Ily certain actions are Subtract Convert Combine with Uplo	Edit a col CTIONS a available to you	laborator group I for the items selected. To utilize the disabled actions, please go back and select different item Centes a new working set by subtract one working set than another Convert the selected working set into a new working are the different type Combine the selected working set with opticaded sequence
Edit Group NAVAILABLE A ly certain actions are Subtract Convert Combine with Uplo Combine with Work	Edit a col CTIONS a available to you aded File	laborator group u for the items selected. To utilize the disabled actions, please go back and select different item Create a new working set by subtract one working set from another Convert the selected working set that a new working set with different type Combine the selected sequences with waive set.
Edit Group NAVAILABLE A Ny certain actions are Subtract Convert Combine with Uplo Combine with Work Unsubscribe Searc	Edit a col CTIONS a available to you aded File ting Set hes	Intercentor group for the items selected. To utilize the disabled actions, please go back and select different item Center is new working set by subtract one working set than another Convert the selected working set than an we working set with different type Combine the selected working set with a selected sequence Combine the selected sequences set with a working set
Edit Group NAVAILABLE A ily certain actions are Subtract Convert Combine with Uplo Combine with Work Unsubscribe Searc Save Unsaved Sear	Edit a col CTIONS a available to you added File ting Set thes ches	Internation group International Internationa

II. Construct an HA segment phylogenetic tree

- a. 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.
- b. 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**".
- c. On the Tree setting page, select "Quick Tree", choose strain name and date as tree tip label, and click "Build Tree".
- d. 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.
- e. 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.
- f. 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



10



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.

- a. 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.
- b. Sort the list by the "Date" column. Now select the following protein sequence records:
 - i. All human H7N9 from 2013
 - ii. All non-human H7 isolated from all European and Asian Countries before 2012
 - iii. Then mouse over "Run Analysis" and click "Metadata-driven Comparative Analysis Tool".
- c. 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".
- d. 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.
- e. This analysis may take a few minutes to finish. You can save the analysis to your Workbench and retrieve it later. To do so, enter in a name (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?

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

h. Save the analysis result to your Workbench by clicking the "Save Analysis" button.

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?

s	EQUENCE FE	ATURE DEFI	NITION																
	Protein Name					НΛ												1	
	Sequence Feat	ture Name				Influ	enza A	H7 exp	erime	entally	/-dete	mined-	enitone	226(14	0				
	Sequence Fea	ture ID				Influ	enza A	H7 SE	5	sincury	ucio	in incu	ophopo_	220(14	·/				
	VT-1 Strain (re	ference strair	0			∆/tu	rkov/lta	w/22015	9 8/200	2(H7	N3)								
	Reference Sec	uence Acces	sion			AY5	86409	19/22010	0/200	2(111)	10)								
	Reference Pos	ation	51011			2260	210 HA	1)-239											
		Autori				LEO	21011	1) 200											
s	OURCE STRA	IN(S)																	
	Source Str	rain Nu	VT mber	Source Position	n A	Source ccessic))n	3D Prote Structu	ein re	Pu	ublica	tion	Epitop Type		Evidence Codes	Ep	itope Se	quence	Comment
	A/Englan /268/1996(H	nd V 17N7)	T-1	226-239	A	F02802	0	-N/A-		IED	B:177	209 ଜ୍ମ	B Cel	1	-N/A-	PGA	ARPQVN	IGQSGRI	-N/A-
v	ARIANT TY	PES	MSA D	ownload		View	Phylog	enetic T	ree		Fin	d a VT(s)						
								S	equei	nce V	/ariati	on							
	Strain Count	Variant Type	226	227	228	229	230	231	23	2	233	234	235	236	237	238	239	Total Va	riations
	772	VT-1	Р	G	Α	R	Р	Q	V	'	Ν	G	Q	S	G	R	1	0	
	356	VT-2	•	•	•	•	-	-	-		-	-	-	-	-	•	•	8	
	38	VT-3	•	•	E	•	•	•	•		•	•	•	•	•	•	•	1	
	18	VT-4	•	•	Р	•	•	•	1		•	•	•	•	•	•	•	2	
	16	VT-5	•	•	Т	•	-	-	-		-	-	-	-	-	•	•	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.
 - i. How many Variant Types does this SF have?
 - ii. The older H7 proteins mostly have SGSTT, which corresponds to VT-1.

Protein Name			HA					
Sequence Feature Na	me		Influenz	a A_H7_detern	ninants-of-receptor-bindin	ig_194(5)		1
Sequence Feature ID			Influenz	a A_H7_SF4				
/T-1 Strain (reference	strain)		A/turkey	y/Italy/220158/2	002(H7N3)			
Reference Sequence /	Accession		AY5864	09				
Reference Position			194(178	8 HA1)-198				
OURCE STRAIN(S) Source Strain	VT Number	Source Position	Source	3D Protein Structure	Publication	Evidence	Com	ment
A/duck/HONG KONG/293 /1978(H7N2)	VT-1	185 -189	U20461	-N/A-	PubMed:22345462	EXP	Atypical European v whereas the N. Am G186A o	viruses show G18 verican strains sho r G186E.
ARIANT TYPES Excel Download ylogenetic tree view o	Find a lisabled becau	VT(s) se there are r	not enough vari	ant types to ger	nerate the tree.			
ARIANT TYPES Excel Download hylogenetic tree view of Edit specific positions Click Search to find V	Find a lisabled becau in this VT-1 se T(s) conformir	VT(s) use there are r equence with ing to the edite	not enough vari IUPAC symbols	ant types to ger s or use "?" as a lick Reset to res	nerate the tree. a wild-card. If necessary, store this panel to the defa	use the horizo ault VT-1 sequ	ntal scroll bar to access ence.	s the entire SF.
ARIANT TYPES Excel Download hylogenetic tree view of Edit specific positions Click Search to find V	Find a disabled becau in this VT-1 se T(s) conformir	VT(s) ise there are r equence with ig to the edite	not enough vari IUPAC symbols d sequence. Cl	ant types to ger s or use "?" as a lick Reset to res Enter Sequ	nerate the tree. a wild-card. If necessary, store this panel to the def ence Variation to Find	use the horizo ault VT-1 sequ	ntal scroll bar to access ence.	s the entire SF.
ARIANT TYPES Excel Download vylogenetic tree view of Edit specific positions Click Search to find V Search	Find a lisabled becau in this VT-1 sr T(s) conformir	VT(s) ise there are r equence with 1g to the edite	not enough vari IUPAC symbols ed sequence. Cl	ant types to ger s or use "?" as a lick Reset to res Enter Sequ	nerate the tree. a wild-card. If necessary, tore this panel to the defi ence Variation to Find 196	use the horizo ault VT-1 sequ 197	ntal scroll bar to access ence. 198	s the entire SF.
ARIANT TYPES Excel Download aylogenetic tree view of Edit specific positions Click Search to find V Search Reset	Find a lisabled becau in this VT-1 sr T(s) conformin	VT(s) see there are r equence with ng to the edite	not enough vari IUPAC symbols ed sequence. Cl 195	ant types to ger s or use "?" as a lick Reset to res Enter Sequ	a wild-card. If necessary, tore this panel to the defi ence Variation to Find	use the horizo ault VT-1 sequ	ntal scroll bar to access ence. 198	s the entire SF.
ARIANT TYPES Excel Download Hylogenetic tree view of Edit specific positions Click Search to find V Search Reset	Find a lisabled becau in this VT-1 sr T(s) conformin	se there are r equence with ng to the edite 194 ?	not enough vari IUPAC symbols ed sequence. Cl 195	ant types to ger s or use "?" as a lick Reset to res Enter Sequ 5	a wild-card. If necessary, to develop the definition of the defini	use the horizo ault VT-1 sequ 197 ?	ntal scroll bar to access ence. 198 A	s the entire SF.
ARIANT TYPES Excel Download hylogenetic tree view of Edit specific positions Click Search to find V Search Reset Fill wildcards	Find a disabled becau in this VT-1 sr T(s) conformin	svT(s) set here are r equence with 194 ?	not enough vari IUPAC symbols ed sequence. Cl 195 V	ant types to ger s or use "?" as a lick Reset to res Enter Sequ	a wild-card. If necessary, store this panel to the defi ence Variation to Find 196 2	use the horizo ault VT-1 sequ 197 ?	ntal scroll bar to access ence. 198 A	s the entire SF.
ARIANT TYPES Excel Download lylogenetic tree view or Edit specific positions Click Search to find V Search Reset Fill wildcards	Find a lisabled becau in this VT-1 st T(s) conformin	VT(s) ise there are I equence with 194 ?	not enough vari IUPAC symbols d sequence. Cl 195 V	ant types to ger s or use "?" as a lick Reset to res Entor Sequ 5 	erate the tree. a wild-card. If necessary, tore this panel to the defi ence Variation to Find 196 [2] uence Variation	use the horizo ault VT-1 sequ 197 7	ntal scroll bar to access ence. 198 A	s the entire SF.
ARIANT TYPES Excel Download lydogenetic tree view of Edit specific positions Click Search to find V Search Reset Fill wildcards	Find a lisabled becau in this VT-1 sr T(s) conformin t(s) conformin t(s) conformin	VT(s) use there are i equence with 194 ? 194	IUPAC symbol d sequence. Cl 195 V	ant types to ger s or use "?" as a ick Reset to res Enter Sequ 5 Sequ 5	verate the tree. a wild-card. If necessary, store this panel to the defa enco Variation to Find 196 7 7 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9	use the horizo ault VT-1 sequ 197 2 197	ntal scroll bar to access ence. 198 A 198	s the entire SF.
ARIANT TYPES Excel Download Vylogenetic tree view of Edit specific positions Click Search to find V Search Reset Fill wildcards	Find a lisabled becau in this VT-1 s: T(s) conformin t Typeo	VT(s) see there are 1 equence with 194 ? 194 S	not enough vari IUPAC symbolic did sequence. Cl 198 V V 198 G G	ant types to ger cor use "?" as a lick Reset to res Entor Sequ 5 Seq 5	verate the tree. a wild-card. If necessary, to the panel to the definence Variation to Find 196 7 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	use the horizo uult VT-1 sequ 197 2 2 197 T	ntal scroll bar to access ence. 198 A 198 T	the entire SF.





- 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

14

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.



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





- 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: A 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

Picket 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 avianhuman 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.

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

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

AAAGCAGGGGATAATTCTATTAACCATGAAGACTATCATTGCTTTGAGCTACATTCTATGTCTGGTTTTCGCTCAAAAAC TTCCTGGAAATGACAACAGCACGGCAACGC

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

c. Either paste your sequence in FASTA format to the sequence box or upload your FASTA sequence. Provide your email address so that IRD can contact you if there are problems in the annotation process. Click "Validate Sequence(s)" to start the annotation process.

SEARCH DATA ANALYZE & VISUALIZE	WORKBENCH	SUBMIT DATA	HOME	
Annotate Nucleotide Seque	nces [®]			
As a service to the influenza research community, IRD wi strain, IRD will process them through a unique sequence Laboratory, headed by Dr. Clatherine Macken. The pipelin then align the sequences against a standard sequence a no critical errors are encountered during auto-curation, y nucleotide sequence, and translated amino acid sequence Sequence(s) button below.	Il curate and annotate curation/annotation pi e will determine the ir nd translate the aligne pu will be provided wit e(s) for each segmen	your influenza sequer ipeline developed by the fluenza type, segment ed nucleotide sequence h a report containing the t submitted. Submit ser	ces. If you prov e Influenza Sec number, and fo using a unique le influenza type quences for up	ide sequences for one or more segments of an influenza quence Database (ISD) team of Los Alamos National resements 4 and 6 of type A, the subtype. The pipeline will translator developed specifically for influenza sequences. If a, segment number, subtype (if appropriate), aligned to 8 segments in FASTA format and click the Validate
Note: An asterisk (*) = required field				
Submitter Email *				
Pandemic H1N1 Sequences? Upload Your Segment Data* Enter your sequence in FASTAr2				
>gb:gQ293082 Organism:Influenza A virus A/Pe /16/2009 Segment:6 Subtype:B382 Host:Human AGCANARCGAGGATANAGATGATCAAARCAAARGATAATA TCCACAATATCCTTCTTCATOCAATTGCCATCTTGATAACT Astrocoatancocococabacaacaatacaatacaatagatagata	erth CTATTGGCTCTGTTTCT GTAACATTGCATTTCA CAATAATACAAACAAA	CTCACCATT AGCAATATG		
or Upload your sequence in FASTAR File Browse)			
			Can	cel Validate Sequence(s)





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.

juence Annotati	ion Result [®]			
YOUR TICKET NUMBER	: <u>160142745001</u>	(Select/Deselect All ▼	Annotation Report
1 sequences submitted				
1 sequences pass auto-curati	on without having any warning messag	je(s)		
0 sequences pass auto-curati	on with warning message(s)			
0 sequences fail auto-curation	with error message(s)			
A numerications Destations				
Annotation Details:				
Section Details:	Influenza A virus A/Perth/16/2009 Seg	gment:6 Subtype:H3N2	Host:Human	
Flu Type:	Influenza A virus A/Perth/16/2009 Seg	gment:6 Subtype:H3N2	Host:Human Segment number: 6	
Annotation Details: Subtype:	Influenza A virus A/Perth/16/2009 Seg A N2	gment:6 Subtype:H3N2	Host:Human Segment number: 6	
Protection Details: Subtype: Sequence Length:	Influenza A virus A/Perth/16/2009 Seg A N2 1507	gment:6 Subtype:H3N2	Host:Human Segment number: 6	
Protection Details: Subtype: Su	Influenza A virus A/Perth/16/2009 Seg A N2 1507 AGCAAAAGCAGGAGTAAAGATGA	gment:6 Subtype:H3N2	Host:Human Segment number: 6 AATAACTATTG [full]	

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 "HPAI 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:

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.



The IRD team has implemented an algorithm for classifying the clade of the hemagglutinin gene of Infuerza A viruses whose HA belongs to the Algosea/Guangdong/169 (HSN1) lineage, that is, the HA lineage of the so-called highly pathogenic HS viruses. This algorithm was developed by IRD team member Catherine Macker, of Los Alamos National Laboratory, ruses phylogenetic analysis by pice HA (HS) sequences within the WHO classification scheme presented here G . The IRD algorithm has been verified as highly accurate (> 9%) for sequences of at least 300 nucleotides of HA1. See SOP for more details. This tool only handles segment 4 sequences with confirmed HS serotype and lengths greater than 300 nucleotides. Sequences from other serotypes of HA, or other segments will yield unprecisite and likely incore results. If names of your sequencity 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 Inix.	
INPUT SEQUENCES • Upload a file containing my sequences in FASTA & format. • Pate sequences in FASTA & format.	

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:Nort err_Pintal	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

- a. Go to www.viprbrc.org and click "Hepatitis C Virus" to get to the HCV page.
- b. Mouse-over the "Analyze & Visualize" tab from the grey navigation bar and click "Genome Annotator (GATU)".
- c. 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.

a Research D	atabase		Virus Path	'iPR ogen Resour	rce
SEARCH DATA	ANALYZE & VISUALIZE	WORKBENCH	VIRUS FAMILIES	HOME	Flavivirida
Genome Annta previously annotated refe controls for uploading a re transferred annotations ar be annotated. The annota	tion Transfer Utility (Tcherepa rence to a new, closely-relate ference .gb file of the relevan d provides users with checkt ted target genome can be sa	D) nov, et al., BMC Genomics d target genome. VIPR use t viral family, along with the vox control over which to a ved in multiple file formats.	s 2006, 7:150 PubMed: 1 rrs should ensure that the target genome in .gb o ccept. GATU also detect	6772042) is an initial-stage toc eir system has Java 1.6 or high r Fasta format. When done, a t s ORFs in the target and bioinfi	ol to transfer annotations from a ter. The GATU interface provides able summarizes the similarities of ormatics tools to assess if these sho
Originally developed at th	e University of Victoria, GATU	was adapted for use with	VIPR.		
REFERENCE SEC To use GATU, you will r can use ViPR Blast to s sequence file, and dow respective controls.	QUENCE need to select a reference sec rearch for one. Browse to you nload it to your directory in Ge	quence. If you already have r target sequence in a FAS enBank format. Then clic	e an appropriate GenBar TA@format, then click of Use your to to BLAST	ik file, proceed directly to Laun on Go. Run a Blast search, pick arget sequence for a closely-	tch GATU. If not, you k a reference nces using the
File Path: /Users/yzhang/Downle	oads/sequence.fasta	Browse	elated anno	otated sequence	ce .
Go Launch GATU	Launch GA you have sequence i	TU directly i a reference in .ab format	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:

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.

SEARCH DATA	ANALYZE & VISUALIZE	WORKBENCH	VIRUS FAMILIES	HOME	H	epatitis C vi
me Identify Similar Sequer	ices (BLAST) Results					•
GATU						
GATU, a Genome Anno previously annotated re controls for uploading a transferred annotations be annotated. The anno Originally developed at	tation Transfer Utility (Tcherepan l'erence to a new, dosely-related reference, ghi lico the relevan and provides users with checkb fatled target genome can be sav the University of Victoria, GATU SEQUENCE	nov, et al., BMC Genom 1 target genome. VIPR u t viral family, along with nox control over which to ved in multiple file forma was adapted for use wi	ics 2006, 7:150 PubMed sers should ensure that the target genome in .gb accept. GATU also dete ts. th VIPR.	: 16772042) is an initial their system has Java 1 or Fasta format. When cts ORFs in the target a	stage tool to tran .6 or higher. The done, a table su and bioinformatic	nsfer annotations from a GATU interface providi mmarizes the similaritie s tools to assess if thes
Here are some re Now click Launch	commended Reference Sequen GATU. Under Genome Selectio	ices. Select one, click th on > Reference Genome	e link Download GenBar e, click Upload Genome F	ik File, and save the file file and browse to the s	e to your local ma aved Reference.	E Valuo
EXT445	955 >nil4459	55I Country: I Henatitis C	virus genotype 3 genor	ne labi157781216	9587	
EXT383	780 >qi 383780 C	Country: Hepatitis C viru	is genotype 1, complete	aenome.labl22129792	896	0.0
EXT446	165 >gi 446165 C	ountry: Hepatitis C viru	s genotype 2, complete g	enome. gb 157781212	767	0.0
EXT446	144 >gi 4461/	44 Country: Hepatitis C	virus genotype 4, genor	ne. gb 157781208	755	0.0
EXT445	982 >gi 445982 C	ountry: Hepatitis C viru	s genotype 6, complete g	enome. gb 157781214	737	0.0
Launch GAT						

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 FASTAformatted 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	n and Recombination detection Resu	ults				
Genotype R	Report					
	-					
Save Analysis	Run Analysis V					
Genotype Inform	ation					
Whole Genome Genoty	pe prediction: 2a					
Whole Genome Recom	bination Type: 2a,2b					
Genotype						
The genotype results in	clude a tab separated file listing	the sequence name, a	single consensus gen	otype result for the entire	genome, and th	e confidence metric.
Download						
Recombination						
This is an excel spread	sheet listing the results for all wir	dows for the sequence	e.			
Download						
E Construing roo	ulto in graphical format					
- Genotyping res	Branchir	n⊄ index profile	for JF343783			
1						
H.1111111						
•						
0.8 -						
•.7 - ×						
- 0.6 -						
₩ 						
۵						
0.3 - -						
0.2 -						
0.1 -						
004000000000000000000000000000000000000	74862-055-44074882-055-44074882	000-4004000-000-000000	00000000000000000000000000000000000000	0049940956400000000000000000000000000000	101778999 101778999	
		2 b 2 a	Alignment posit	ion (nt)		
■ Alignment						
This is the multiple sequ	uence alignment of your sequence	e with a ViPR reference	e sequence alignment	that consists of at least 2	representatives	from each taxon
Download Aligned	Fasta Visualize Align	ed Sequences				
E Tree						
This is the tree generate	ed by PAUP based on the input a	alignment for the whole	genome			
Download Newick	File View phylogentic	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.

My	/ Sti	rategie	s: Ne	w	0	pened (1)	All (4)	Basket	Exa	mples Help			
(Ge	enes)											Strategy:	Text(3) * 🛛 🔛
				al Pep Genes									Rename Duplicate Save As
] 1525	Text 5 Genes	104	D		Add Step							Share Delete
	S	tep 1	Ste	ep 2									
<u>ا</u>									_				
r—													
10 St	4 G rate	enes fr gy: Te	om Step xt(3)	2							Add 104 Genes	to Basket Down	load 104 Genes
Ξ	🝸 Filt	ter results	s by species	(res	ults re	moved by the filter	will not be com	bined into the	next step	.)			
$\left \right $	AII	Ortholog	Plasmo falcipa	dium rum		Plasmodium	Plas	modium yoel	lii	Plasmodium	Plasmodium	Plasmodium Plasmodium	Plasmodium
Re	sults	Groups	Distinct genes	3D7	IT	vivax	Distinct genes	yoelii 17XNL	yoelii YM	berghei	chabaudi	knowlesi	cynomolgi
1	04	26	21	21	19	7	13	8	11	12	11	8	7
	Gene	Results	Genom	ne Vie	w								
F	irst 1	12345	Next Las	t	(Advanced Pa	aging					(Add Columns
đ	•	Gene ID) 🇘	ieno	mic	Location 🕹	🗘 Pi	roduct Des	criptio	n 🥝 📊			
1	PY	00582	AAE	3L010	0001	58: 5,062 - 7,488	3 (-) heat	shock prote	ein 90				
1	PY	01906	AAE	3L010	00052	21: 2,986 - 5,628	B (-) heat	shock prote	ein 83				
1	PY	01909	AAE	3L010	00052	21: 9,757 - 10,71	14 (+) deph	ospho-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?



Add the first user comment Add to Basket Add to Basket Add to Favorites Add to Favorites eview you fil you fil 17XNL protein coding gene on AABL01000158 from 5,062 to 7,488 (Chromosome: Not Assigned) momic Context Hide (use right click or ctrl-click to open in a new window)		PY00582 heat shock protein 90	
erview yoelii yoelii 17XNL protein coding gene on AABL01000158 from 5,062 to 7,488 (Chromosome: Not Assigned) inomic Context Hide View in Genome Browser (use right click or ctrl-click to open in a new window) AHBL01000158 0K 1k 2k 3k 4k 5k 6k 7k 8k 9k 10K 11k Annotated Genes (uith UTRs in gray when available) PY00582 Syntenic Sequences and Genes (Shaded by Orthology) Pyoe contig 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k Pyoe genes Pyor contig 1349k 1350k 1351k 1352k 1353k 1354k 1355k 1355k 1355k 1355k 1355k 1359k Pyor contig 1349k 1349k 1350k 1351k 1352k 783 794k 785k 786k 787 788k 789K 799K 799K Phor genes Pyon contig 1349k 1349k 1350k 431k 432k 433k 435k 435k 435k 435k 435k 435k 435		Previous ID(s): 159.m00048 Add the first user comment 🗭 Add to Basket 魪 Add to Favorites 🚖	
yoelii yoelii yoelii 17XNL protein coding gene on AABL01000158 from 5,062 to 7,488 (Chromosome: Not Assigned) momic Context Hide (use right click or ctrl-click to open in a new window) AFEL01000158 ok 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k Renotated Genes (uith UTRs in gray when available) PY00581 Syntenic Sequences and Genes (Shaded by Orthology) Pyoe contig 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k Pyoe genes Pyoe actig 1348k 1349k 1350k 1351k 1352k 1353k 1350k 1350k 1357k 1356k 1357k 1357k 1356k 1357k 1356k 1357k 1357k 1356k 1357k	verview		
Nomic Context Hide View in Genome Browser (use right click or ctrl-click to open in a new window) AFEL 01000158 ok 4k 5k 6k 7k 8k 9k 10k 11k AFEL 01000158 ok 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k Brontated Genes (with UTRs in gray when available) PY00581 PY00582 PY00582 PY00582 Pyoe genes 1k 2k 3k 4k 5k 6k 7k 9k 10k 11k Pyoe genes 1k 2k 3k 4k 5k 6k 7k 9k 811k 812k Poor genes 02k 803k 804k 805k 807k 808k 810k 811k 812k Poor genes 1348k 1349k 1350k 1351k 1352k 1353k 1356k 1357k 1358k 1357k 1358k 1357k 1358k 1357k 1358k 1357k	. yoelii yoelii 17XNL pro	otein coding gene on AABL01000158 from 5,062 to 7,488 (Chromosome: Not Assigned)	
View in Genome Browser (use right click or ctri-click to open in a new window) AmBL 01000158 ok 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k Amnotated Genes (with UTRs in gray when available) PY00581 PY00582 <	enomic Context Hide		
(use right click or ctrl-click to open in a new window) PABL 01000158 0k 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k Renotated Genes (with UTRs in gray when available) PY00581 Syntenic Sequences and Genes (Shaded by Orthology) Pyge contig 1k 2k 3k 4k 5k 6k 7k 9k 9k 10k 11k Pyge genes Pober contig 301k 802k 803k 804k 805k 806k 807k 808k 809k 810k 811k 812k Pber contig 1348k 1349k 1350k 1351k 1352k 1353k 1359k 1356k 1357k 1357k 1356k 1357k 1356k 1357k 1357k			
(use right click or ctrl-click to open in a new window)		view in Genome Browser	
APRE 01000158 ok 1k 2k 3k 4k 5k 6k 7k 9k 10k 11k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k 2k 3k 4k 5k 6k 7k 803k 804k 805k 806k 807k 803k 1378k 1378k <th colspan<="" td=""><td>(use right click or ctrl-clic</td><td>k to open in a new window)</td></th>	<td>(use right click or ctrl-clic</td> <td>k to open in a new window)</td>	(use right click or ctrl-clic	k to open in a new window)
ok 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k Rnnotated Genes (uith UTRs in gray when available) Py00581 Py00582 Py00584 Py00584 Py00584 Py00584 <td></td> <td>AABL01000158</td>		AABL01000158	
Annotated Genes (with UTRs in gray when available) PY00551 PY00552 Syntenic Sequences and Genes (Shaded by Orthology) 1k 2k 1k 801k 801k 803k 801k 803k 801k 803k 801k 803k 1348k 1349k 1348k 1349k 1348k 1349k 1348k 1349k 1348k 1349k 1348k 1349k 1349k 1349k		0k 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k	
Pyroossi Pyroossi Syntenic Sequences and Genes (Shaded by Orthology) Pyroossi Pyroe contig 1k 2k 3k 4k 5k 6k 7k 9k 10k 11k Pyroe contig 1k 2k 3k 4k 5k 6k 7k 9k 9k 10k 11k Pyroe contig 1k 2k 3k 4k 805k 806k 807k 808k 809k 810k 811k 812k Pber contig 1348k 1349k 1350k 1351k 1352k 1353k 1356k 1357k 1358k 1359k 1359k 1358k 1359k 1358k 1359k 1359k 1358k 1359k 1359k 1358k 1359k			
Syntemic Sequences and Genes (Shaded by Orthology) Jk 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k Pyoe genes 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k Pyoe genes 801k 802k 803k 804k 805k 806k 807k 808k 809k 810k 811k 812k B01k 802k 803k 804k 805k 806k 807k 808k 809k 810k 811k 812k B01k 802k 803k 1351k 1352k 1353k 1355k 1356k 1357k 1358k 1359k Pyyn contig 90 90 90 90 90 90 90 90 Pyyn genes 90 783k 783k 783k 784k 785k 786k 787k 788k 790 k P1307 contig 90 900 900 k 430 k		Hnnotated Genes (with UIKs in gray when available) PY00581 PY00582	
Syntenic Sequences and Genes (Shaded by Orthology) 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k Pyoe genes 802k 803k 804k 805k 807k 808k 809k 810k 811k 612k Pber contig 802k 803k 804k 805k 807k 808k 809k 810k 811k 612k Pyon contig 901k 1348k 1349k 1350k 1351k 1352k 1353k 1356k 1357k 1358k 1359k 1358k 1359k 1358k 1359k 1350k 1359k			
Pyoe contig 01k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k Pyoe genes 801k 802k 803k 804k 805k 806k 809k 810k 811k 812k Pber contig 801k 802k 803k 804k 805k 806k 809k 810k 811k 812k Pyon genes 9 9 1350k 1351k 1352k 1353k 1356k 1357k 1356k 1357k 1356k 1359k Pyyn genes 9 9 9 781k 782k 783k 784k 787k 788k 789k 790k Pcha contig 9		Syntenic Sequences and Genes (Shaded by Orthology)	
Pyoe genes Por contig 801k 802k 803k 804k 805k 806k 809k 810k 811k 812k Pber contig 801k 802k 803k 804k 805k 807k 808k 809k 810k 811k 812k Pber genes Pyyn contig 9000 <t< td=""><td></td><td>ryoe contig 1, 2, 3, 4, 5, 6, 7, 8, 0, 10, 11,</td></t<>		ryoe contig 1, 2, 3, 4, 5, 6, 7, 8, 0, 10, 11,	
Pber contig 901k 802k 803k 804k 805k 806k 807k 806k 809k 810k 811k 812k Pber genes Pyy contig		Pyoe genes	
801k 802k 803k 804k 805k 806k 807k 808k 809k 810k 811k 812k 812k <th< td=""><td></td><td>Pber contig</td></th<>		Pber contig	
Pber genes Purpose Pyyn contig 13434k 1350k 1351k 1352k 1353k 1356k 1357k 1356k 1357k 1358k 1359k 1348k 1349k 1350k 1351k 1352k 1353k 1354k 1355k 1356k 1357k 1358k 1359k Poha contig Poha genes Poha genes Poha genes Poha genes Pf307 contig Pf307 contig Pf307 genes Pf307 genes Pf307 genes Pf307 genes Pf307 contig		801k 802k 803k 804k 805k 806k 807k 808k 809k 810k 811k 812k	
Pyyn contig 1349k 1349k 1350k 1351k 1352k 1353k 1354k 1355k 1356k 1357k 1358k 1359k Pyn genes Pcha contig 779k 780k 781k 782k 783k 784k 785k 786k 787k 788k 789k 790k Pcha genes Pf307 contig 429k 430k 431k 432k 433k 439k 435k 436k 437k 438k 439k 440k		Pber genes	
1349k 1349k 1390k 1391k 1352k 1391k 1352k 1393k 1390k		Pyym contig	
Pcha contig 779k 780k 781k 782k 783k 784k 785k 786k 789k 790k Pcha genes 71307 contig 1		1348k 1349k 1350k 1351k 1352k 1353k 1354k 1355k 1356k 1357k 1358k 1359k	
Pcha contig 779k 780k 781k 782k 783k 784k 785k 786k 787k 788k 789k 799k Pcha genes Pf307 contig 429k 430k 431k 432k 433k 434k 435k 436k 437k 438k 439k 440k Pf307 genes			
779k 780k 781k 782k 783k 784k 785k 786k 787k 788k 789k 789k 799k 790k Pchagenes Pf307 contig 429k 430k 431k 432k 433k 434k 435k 436k 437k 438k 439k 440k Pf307 genes		Pcha contig	
Pf307 contig 429k 430k 431k 432k 433k 434k 435k 436k 437k 438k 439k 440k Pf307 genes 1		779k 780k 781k 782k 783k 784k 785k 786k 787k 788k 789k 790k Pcha genes	
429k 430k 431k 432k 433k 434k 435k 436k 437k 438k 439k 440k Pf307 genes		Pf307 contig	
Pf307 genes		429k 430k 431k 432k 433k 434k 435k 436k 437k 438k 439k 440k	
		Pf307 genes	

- 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 Brower 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 (<u>www.vectorbase.org</u>).
- 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.

🌣 🧟 < 🖂								
	L				93.46 kb	 	 – Forward stra	nd 💻 –
VectorBase genes	2.35 Mb	236 Mb para > protein coding para > protein coding para > protein coding	<u>237 МБ</u>	238 Mb	239 Mb	2.41 Mb		2.4+

- 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	ATCCTGACGGTACACAATACGTTCGATATGATCAGCT <mark>R</mark> TCAGACTTTTTGGATGTGCTGG	73380 <u>73358: rs180291684</u> ;
73381	AACCGCCTCTACAGATTCATAAACCAAATCGTTATAAGATTATTTCGATGGATATTCCGA	73440
73441	TATGCCG <mark>Y</mark> GGAGATATGATGTTCTGTGTCGATATTCTAGATGCACTAACGAAAGATTTTT	73500 <u>73448: 2L.2431005;</u>
73501	TTGCTAGAAAAGGAAATCCTATAGAAGAAACAGCCGAATTAGGTGAAGTTCAACAACGCC	73560
73561	CAGACGAAGTTGGTTACGAACCAGTATCATCAACACTTTGGAGGCAGCGTGAAGAGTACT	73620
73621	GTGCTCGACTGATACA <mark>R</mark> CATGCGTGGAAACGCTATAAACAGCGTCACGGAGGCGGAACAG	73680 <u>73637: 2L.2431194;</u>
73681	ACGCTTCAGGAGATGATCTTGAAATAGATGCCTGTGATAACGGTTGTGGTGGTGGTAATG	73740
73741	gcaatgaaaatgatgatagtggagatggtgcaacaggtagtggtgacaacggaagtca <mark>g</mark> c	73800 73799: rs5181107;
73801	ATGGTGGTGGCAGCATAAGTGGCGGAGGAGGAACTCCTGGTGGTGGTAAAAGTAAAGGA <mark>R</mark>	73860 73860: WTSI-Ag-GVP-0.1-SNP-2L-2431417
73861	TTATTGGCAGTACTCAGGCTAACATAGGCATAGTGGATAGTAATATATCACCAAAGGAAT	73920
73921	CACCGGATAGCATCGGCGATCCCCAAGGTCGTCAGACGGCCGTTCTTGTGGAGAGCGACG	73980
73981	GATTTGTGACGAAAAACGGTCACCGTGTCGTCATACACTCTCGATCTCCCAGCATAACAT	74040
74041	CGCGAACGGCAGATGTCTGAGCCAGGTCTCGCCCCCCTCTCTCGGATTCAGATTCCGAAG	74100
74101	CACGRCAGAAATAATATTTGAAATGACATGCAATGTAAGGTTTAAGCATTCAAAGAACAT	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).

PATRIC			Login Not Registered? Sign Lin J
Pathosystems Resource Integration Center	ORGANISMS SEARCH	ES & TOOLS DOWNLOA	DS ABOUT Learn About Registering
Search	Genera Containing Ni Emerging / Re-emerg Pathogens	AID Category A-C / jing Bacteria	MY WORKSPACE: No Items, No Groups •
	 Bacillus 	 Francisella 	0104-94 etc TV-2482
FINDING GENOMIC	 Bartonella 	 Helicobacter 	0101.11 30. 11-2102
ISLANDS	 Borrelia 	 Listeria 	Construction of the second
IN EMERGENT E. COLLSTRA 7	Brucella	 Mycobacterium 	
	 Burkholderia 	 Rickettsia 	or To or O protein protein protein report
genomes were sequenced as a result of a European <i>E. co</i>	 Campylobacter 	 Salmonella 	or5 page for page for page for page for page for page for page page page page page page page page
PATRIC, users can identify areas of these genomes that ar missing proteins, unique proteins) when compared to their	Chlamydophila	 Shigella 	<pre>exercise to the top of top of the top of top of</pre>
phylogenetically-related <i>E, coll</i> genomes. In some cases, w	 Clostridium 	 Staphylococcus 	werver po according and according and according and according and according and according accord
that these proteins were obtained via lateral transfer.	 Coxiella 	 Streptococcus 	\$ \$ 2 2 2 6 6 F F F F F F F F F F F F F F F
	 Ehrlichia 	▶ Vibrio	
LEARN MORE	 Escherichia 	 Yersinia 	
	Complete Lists of Bac	teria:	
	 All Bacteria 		DOMES
GENOMIC ISLANDS PROTEINS	FROM OUTBREAKS	DIVERSE PATHWAY	S ALL PATRIC TOOLS

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

Bacteria • Proteobacteria • Alphaproteobacteria	Rhizobiales + Brucellaceae	Brucella		
Overview Taxonomy Phylogeny Genome List	Feature Protein Table Families	Pathways Transcriptomics Diseases Literature		
Search Tools	Taxonomy Summary			
GF Genome Finder	Taxonomy ID	234		
FFF Feature Finder	Lineage Bacteria > Proteobacteria > Alphaproteobacteria > Rhizobiales > Brucellaceae > Brucella			
CP Comparative Pathway Tool	External Links	Immune Epitope Database and Analysis Resource		
PFS Protein Family Sorter	Summary Terms - Click on n	umber to view genomes associated with term (see PATRIC FAQs)		
GO Search	Genome Status	WGS (<u>59</u>), Complete (<u>16</u>)		
EG EC Search	Isolation Country	China (<u>17</u>), Australia (<u>2</u>), <u>show all 22 genomes</u>		
	Host Name	Homo sapiens (10), Ovis aries (2), show all 27 genomes		
Experiment Summary	Disease	Brucellosis (24), Spontaneous abortion (7), show all 37 genomes		

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

Select organism(s)	Inter keyword	
My Groups Taxonomy Tree A-2 List		_
Jump to:	Feature Type: CDS	
(8)	Propionate CoA-transferase	
	Keyword: Example: DNA polymerase	-
Search within: Brucella	Example: Urlan Example: VBIBruSui107850_0001	
	Sequence Status: ALL 💌	
5	Annotation: PATRIC Search	



D. Choose any locus tag from the Results Table (6).

(Kolo Search crean) Browngreuzh bri vlář Zi Kalansk Kondi To semina be filo vr. ot. narovšela, vrřet v stál se v da veří ni PATRO Bestvr stálen, čeles ve Textus z tále 1905														
Filter By	(K) Workspace View		w Do	Download Tools		sols								
Annotation (72) Annotation (72) Patralic (72) Feature Type (72) CDG (72)	Oser Al	Add Fee	ure(s) 8 FASTA	DNA 🛄 Protein 👂 A	rable • 4STA •	Pathway CER M	S.A MAP MAP IDs to •							
		🗇 Genome P	Cenome Name			Locus Tag		Amotation	Peature Type	Start	End	Length (NT)	Strand	Product Description
		E Bucela abortus A13204		CP000178		VIIIhu4bo220367_2933		PATRIC	CDS	1003747	2004754	1000	+	Undecaprenyl-glycosyltransferase Wbid*
		Bucela caris ATCC 23365		NC 01010		VRIBruCari25663-0527		PATRIC	CDS	509630	511399	1530		Undecaprenyl-glycosyltransferase Wbid ⁺
		E Bucela o	Bucela ceti 61/94		51	VRIBruCet113143-1692	- b	PATRIC	CDS	830854	831861	1008		Undecaprenyl-glycosyltransferase WbirF
		E Brucela in	Brucela inopinata BO1		022	VEEBruSp109945_1928		PATRIC	CDS	82005	83013	1008		Undecaprenyl-glycosyltransferase WbkF
		E Brucela al	Erupela abortus 104M		19	VEEhu8bo245177_0525		PATRIC	CDS	111572	112579	1000		Undecaptertyl-glycosyltransferase WbidF
		E Bucela o	Brucela cett M644/93/1		2	VEBruCet111375 1985		PATRIC	CDS	63687	64694	1008	+	Undecaprenyl-glycosyltransferase Wbid ¹
		E Bucela o	Brucela cett M13/05/1		0	VEIBruCet83544_1699	60	PATRIC	CDS	846195	847162	1008		Undecaprenyl-glycosyltransferase WbirF
		E Brucela o	Brucela ceti MH90/95/1		8	VEEBruCet48569_1839	60	PATRIC	CDS	830176	831183	1008		Undecaprenyl-glycosyltransferase WbkF
		🖾 Brucela m	Erucela meltensis MS-10		45	VEED:uMei250065_2234	60	PATRIC	CDS	72423	73430	1009		Undecapters/lightcosyltransferase WbkF
		🖾 Dszelam	Brucela melterais MS-90			VEEbuMil157313_0529	600	PATRIC	CDS	532053	533060	1000		Undecaprenyl-glycosyltransferase Wbid!
		E Brucelarr	Brucela melitensis M28			VEID1040176544_0524	603	PATRIC	CDS	531806	532813	1008		Undecaprenyl-glycosyltransferase WbirF
		E Brucola m	Brucela meltensis biovar Abor - Brucela meltensis ATCC 22457			VEEBruMe86222_0558	60	PATRIC	CDS	529628	530635	1008		Undecaprenyl-glycosyltransferase WbkF
		E Brucela m				VEEBruMe(14466_0557	10	PATRIC	CDS	531821	532828	1009		Undecaprenyl-glycosyltransferase WbkF
		Bucela minoti CON 4915 Bucela abortus NCTC 9038 Bucela abortus NCTC 9038 Bucela meteros NI Bucela meteros NI		NC 01312		VIIIbuMc92249_1727		PATRIC	CDS	513030	514045	1000		Undecaprenyl-glycosyltransferase Wbid?
				NZ 007037	12	VEIBru4b097623_1897	63	PATRIC	CDS	537652	\$38699	1008	+	Undecaprenyl-glycosyltransferase WbitF
				NZ E09995	12	VEIBru/Neo114381_1695	60	PATRIC	CDS	845995	847002	1008		Undecaprenyl-glycosyltransferase WbirF
				CP002933		VEEBruMe(196544_0525	60	PATRIC	CDS	531867	532874	1008		Undecaprenyl-glycosyltransferase WbkF
		🗄 Brucela o	Brucela ons ATCC 25840 Brucela onspecials M163/99			VEEDruOv126990_1923	100	PATRIC	CDS	533336	534400	1065		Undecaprenyl-glycosyltransferase WbidF
		E Bucela p			2	VEBuPn128057_2473	66	PATRIC	CDS	167300	168205	906	+	Undecaprenyl-glycosyltransferase Wbid*
		E Bucela p	nipodali: 62/94	NZ 099998	8	VEIBruPin17457_0475	60	PATRIC	CDS	846697	847704	1008		Undecaprenyl-glycosyltransferase WbirF
					(have)									

E. The feature page for the chosen locus tag (gene), the gateway to relevant pathways (a), expression data (b), genes that have been found with similar expression values in the correlated genes tab (c), two different browsers that feature this gene (d), the corresponding RefSeq locus tag (e) and other information.

a	b c						
Bacteria • Proteobacteria • Aphaproteobar ria VBIBruMel92729_0021, Propionate C Overview Genome Compare Region Browser Viewer Pathwar	a · Rhizoblata · Brucellacs · Brucella · Br rransferat (EC 2.8.3.1) // Transcriptomics Correlated Literatur	ucella melitensis bv. 1 str. 16M •					
	5 . D						
Recent PubMed Articles	Feature Properties	add Easter	to Markenese				
No pubmed record is available.		Add Feadline	to workspace				
	Genome Browser						
	Compare Region Viewer						
	Accession						
	Annotation Source						
	Eosture ID	17940050					
	Corresponding RefSeq Locus Tag						
	Corresponding RefSeg Protein ID	NP 538939.1					
	Uniprot Mapping	27 IDs are mapped (UniProtKB-Accession: <u>OGYJR1</u>)					
		Submit a request for structure determination to SSGCID					
	Pseudo Gene	No					
	Genomic Properties						
		View	NT Sequence 🕅				
	Feature Type	CDS					
	Product	Propionate CoA-transferase (EC 2.8.3.1)					
	Location	2077622266					
	Strand	forward					
	Functional Properties						
	Protein Properties		15				
		View	AA Sequence 2				
	AA Length	496 AA					
	Gene symbol	· Deminustra Call transformers /CC 2.0.2.1					
	Product 60 Accimponents	Propurate CuA-transferate (cc 2.8.3.1)					
	EC Assignments	EC:2.8.3.1 :Propionate CoAutransferase					
	EL Assignments	Electronic Propulate CoArtrainferase (EC 2.8.3.1)					
	r saran naaigi menta	KEGG:00640 :Pronancate metaholism(Metaholism: Carbohydrate Metaholism)					
	Pathway Assignments	KEGG:00643 :Styrene degradation(Metabolism) Carbohydrate Metabolism) KEGG:00643 :Styrene degradation(Metabolism) Carbohydrate Metabolism)	etabolism)				



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



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).



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).





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).

				;	*					
			Bacharia - Protechacharia - A VBIBruhte92729_0021, P Dvereiew Genome C	phaprotechacteria - Phaobaina - Druce II ropionate CoA-transferase (EC 2.8.3 organ legen Pathways Transcriptomics Deven	+ Brucela + Brucela melterata melatod Genes	by, 1 atr. 2000)			a	
Correlation Ca	toff: 0.6 🗶 Correlation: positive	* Filter								
463 features fo	ind									
Workspace	View Download S FASTA DNA Table pAT FASTA Protein 2 FASTA	Pathway BER MSA	MAP IDs to +						Ļ	Help
8	Genome Name	Accession	Locus Tag	Start	End +	Length(NT)	Strand	Product Description	Correlation -	Comparisons
8	Erucela meltensis bv. 1 str. 16M	NC 003317	VBIB/UM992729_0021	20776	22266	1491		Propionate CoA-transferase (EC 2.8.3.1)	1	8
12	Etucela meltensis by, 1 str. 16M	NC 003338	VEIBruMe/92729_2390	1073094	1074938	1945	+	putative Gutathione-regulated potassium-efflux system protein KelB	0.99	9
10	brucela meltansis by, 1 str. 16M	NC 003338	VEID-UNI02729_2972	679739	680614	876	+	Transcriptional regulator, AraC family	0.988	8
8	Erucela meltensis bv. 1 str. 16M	NC.003317	VBIBruMe92729_0485	451152	451379	228		FIG00450185: hypothetical protein	0.994	8
13	Etucela melteras by, 1 str. 16M	NC 003317	VBIB: uMe/92729_0600	557211	557555	245		FIG00450314: hypothetical protein	0.972	9
12	Enucela meltensis by, 1 str. 16M	NC 003338	VED/UN/02729_2599	281613	282200	583	+	COGs COG2672	0.97	8
8	Enucela melitensis bv. 1 str. 16M	NC.003318	VEIBruMe92729_2925	642239	643201	963		Petrobactin ABC transporter, permease protein 1	0.969	8
10	Enacela melberata by, 1 str. 16M	NC 003338	VEIBruMe/92729_2978	683332	693640	309	+	Endonuclease V (EC 3.1.25.1)	0.969	8
-	Erucela meltensis by, 1 str. 16M	NC 003338	VEEHUN/62729_2294	18524	18001	468	+	Putative activity regulator of membrane protease Ybbit	0.968	8
8	Erucela meltensis bv. 1 str. 16M	NC.003318	VEIBruMe/92729_2716	429912	490394	423	+	Organic hydroperoxide resistance protein	0.966	8
23	Erucela meltersis by, 1 str. 16M	NC 003317	YEERUM#92729_1027	949391	949756	305	+	Ntronecluctase family protein	0.965	8
	brucela meltansis by, 1 str. 16M	NC 003318	YEEBU/M652729_2772	482032	483054	1023	+	Possible regulatory protein BirS	0.962	8
13	Brucela meltensis bv. 1 str. 16M	NC_003317	VEIBruMd92729_1458	1349123	1351045	1923	+	Soluble lytic murein transplycosylase precursor (EC 3-2-1)	0.998	9
12	Etucela meltensa by. 1 str. 16M	NC 003338	YEIBruMe92729_2251	24292	75565	799	+	C0G2041: Sulfite oxidase and related enzymes	0.956	9
	brucela melferos by, 1 str. 16M	NC 003338	VEB 04052729 2091	1074951	1075832	882		UPROCESS protein YCHK	0.996	8
	Description of the second seco	N. 00331/	1000 (West 2729 1332	1233805	1234245	991		Appropriate resource or surgery construction regulator Not	0.995	8
10	Income contracts for 1 of 150	N. 1002337	1000 (1000) 1001	1732392	1/4/300	105		To exception of the first of exercises of the first freek.)	0.951	
	En ande meditancia har 1 str. 16M	NC 002217	VERN 44402720 0427	409933	409344	40.3		Reven APC transport nation at Disinder protein Ref. (TC 3.4.1.2.1)	0.951	0
121	Proved residence by 1 day 14M	NC 007217	VERD-04/07720 1712	1599124	1900///	627		ATE perform (C 16.2.2.2.14)	0.95	0