

PANDEMIC FLU GENOMICS EXERCISE

Dr. Robert Moss
Wofford College

Feel free to modify and distribute this exercise for educational purposes.

The Pandemic Influenza outbreak of 2009 was perhaps the first major outbreak where DNA sequencing was used for clinical diagnosis of viral infection. The CDC and NIH have developed some amazing genomics tools that we can use to learn about the virus and its spread.

Go to <http://www.cdc.gov/h1n1flu/>

Click on "Related Links", go to the Genbank resources. You might want to bookmark this page. Restricting your work to this site will limit all searches to influenza viruses, so will make your work easier.

You have been sent a blood sample from a suspected flu sample. You have sequenced a portion of the genome:

```
1 atgaaggcaa tactagtagt tctgctatat acatttgcaa cgcgaaatgc agacacatta
61 tgtatagggt atcatgcgaa caattcaaca gacactgtag acacagtact agaaaagaat
121 gtaacagtaa cacactctgt taaccttcta gaagacaagc ataacgggaa actatgcaaa
181 ctaagagggg tagccccatt gcatttgggt aaatgtaaca ttgctggctg gatcctggga
241 aatccagagt gtgaatcact ctcccacagca agctcatggt cctacattgt ggaaacatct
301 agttcagaca atggaacgtg ttaccagga gatttcatcg attatgagga gctaagagag
361 caattgagct cagtgtcatc atttgaaggg ttgagatat tcccacagac aagttcatgg
421 cccaatcatg actogaacaa aggtgtaacg gcagcatgtc ctcatgctgg agcaaaaagc
481 ttctacaaaa atttaatatg gctagttaaa aaaggaaatt catacccaaa gctcagcaaa
541 tcctacatta atgataaagg gaaagaagtc ctctgctat ggggcattca ccatccatct
601 actagtgctg accaacaagg tctctatcag aatgcagatg catatgtttt tgtggggta
661 tcaagataca gcaagaagt caagccggaa atagcaataa gaccacaaat gagggatcaa
721 gaagggagaa tgaactatta ctggacacta gttagaccgg gagacaaaat aacattcgaa
781 gcaactggaa atctagtggg accgagatat gcattcgcaa tggaaaagaa tgctggatct
841 ggtattatca tttcagatac accagtccac gattgcaata caacttgtca gacacccaag
901 ggtgctataa acaccagcct cccatttcag aatatacatc cgatcacaat tggaaaatgt
961 ccaaaatattg taaaagcac aaaattgaga ctggccacag gattgaggaa tgtcccgtct
1021 attcaatcta gaggcctatt tggggccatt gccggttca ttgaaggggg gtggacaggg
1081 atggtagatg gatggtacgg ttatcaccat caaaatgagc aggggtcagg atatgcagcc
1141 gacctgaaga gcacacagaa tgccattgac gaaattacta acaaagtaaa ttctgttatt
1201 gaaaagatga atcacagtt cacagcagta ggtaagagat tcaaccacct ggaaaaaaga
1261 atagagaatt taaataaaaa agttgatgat ggtttcctgg acatttggac ttacaatgcc
1321 gaactgttgg ttctatttga aaatgaaaga actttggact accacgatcc aaatgtgaag
1381 aacttatatg aaaaggtaa agccagcta aaaaacaatg ccaaggaat tggaaacggc
1441 tgctttgaat ttaccacaa atgcgataac acgtgcatgg aaagtgtcaa aaatgggact
1501 tatgactacc caaaatactc agaggaagca aaatataaca gagaagaaat agatgggta
1561 aaactggaat caacaaggat ttaccagatt ttggcgatct attcaactgt cgccagttca
1621 ttggtactgg tagtctccct gggggcaatc agtttctgga tgtgctctaa tgggtctcta
1681 cagtgtagaa tatgtattta a
```



Influenza Virus Resource presents data obtained from the NIAID Influenza Genome Sequencing Project as well as from GenBank, combined with tools for flu submission to GenBank. In addition, it provides links to other resources that contain flu sequences, publications and general information about

Read more about: [This resource](#) | [Flu database](#) | [Flu sequence submission to GenBank](#) | [NIAID Influenza Sequencing Project](#) | [Influenza](#)

- NCBI
- Growth of flu sequences
- GenBank sequences from the NIAID Project
- Assembly Archive
- Trace Archive
- NIAID data releasing status
- RefSeq genomes
- RefSeq proteins
- Protein Structures
- Flu resources
- NIAID Project
- ICVI Flu

GenBank sequences from pandemic (H1N1) 2009 viruses

All submitted influenza sequences are available in GenBank as soon as they are processed. The pandemic (H1N1) 2009 virus sequences are listed on this page and are available for BLAST searching [here](#), and are also available in the **NCBI Influenza Virus Sequence Database**, and can be retrieved with sequences from other influenza viruses for further analyses using tools integrated to the database. [Go to a tutorial for instructions downloading these sequences.](#) A complete list of GenBank sequences for these viruses can also be obtained through a [special genome project page](#) or directly from [here](#). The result of RPS-BLAST against PDB database, and a summary of amino acid differences in proteins of these viruses are available at [Riken National Institute of Japan](#).

The following pandemic (H1N1) 2009 virus sequences were submitted to NCBI and are available in GenBank:

July 28, 2009, 5 submitted by Adolfo Lutz Institute, Brazil:

PB2 PB1 PA HA NP NA MP NS

Influenza A virus GQ414764

Click to sort columns

Description	Max score	Total score	Query coverage	E value	Max id
Influenza A virus (A/Oklahoma/01/2009(H1N1)) segment 4 hemaqa	3142	3142	100%	0.0	100%
Influenza A virus (A/Texas/12/2009(H1N1)) segment 4 hemaqlutin	3142	3142	100%	0.0	100%
Influenza A virus (A/Texas/05/2009(H1N1)) segment 4 hemaqlutin	3142	3142	100%	0.0	100%
Influenza A virus (A/Texas/09/2009(H1N1)) segment 4 hemaqlutin	3142	3142	100%	0.0	100%
Influenza A virus (A/Texas/22/2009(H1N1)) segment 4 hemaqlutin	3142	3142	100%	0.0	100%
Influenza A virus (A/Kansas/02/2009(H1N1)) segment 4 hemaqluti	3142	3142	100%	0.0	100%
Influenza A virus (A/Texas/08/2009(H1N1)) segment 4 hemaqlutin	3142	3142	100%	0.0	100%
Influenza A virus (A/Texas/06/2009(H1N1)) segment 4 hemaqlutin	3142	3142	100%	0.0	100%
Influenza A virus (A/Texas/04/2009(H1N1)) segment 4 hemaqlutin	3142	3142	100%	0.0	100%
Influenza A virus (A/Texas/04/2009(H1N1)) segment 4 hemaqlutin	3142	3142	100%	0.0	100%
Influenza A virus (A/Texas/04/2009(H1N1)) segment 4 hemaqlutin	3142	3142	100%	0.0	100%
Influenza A virus (A/Texas/05/2009(H1N1)) segment 4 hemaqlutin	3142	3142	100%	0.0	100%
Influenza A virus (A/Texas/05/2009(H1N1)) segment 4 hemaqlutin	3142	3142	100%	0.0	100%
Influenza A virus (A/Texas/05/2009(H1N1)) segment 4 hemaqlutin	3142	3142	100%	0.0	100%

Swine influenza A (H1N1) virus isolated during human swine flu outbreak of 2009. For more information, see <http://www.cdc.gov/>.

Some of the information does not have GenBank feature identifiers and is being provided in the comment section.

##EpifluData-START##

```

Isolate A/Texas/12/2009
Subtype H1N1
Segment_name HA
Host_gender F
Host_age 12
Passage_history C2
Adamantane_resistance resistant
Zanamivir_resistance sensitive
Oseltamivir_resistance sensitive
Country USA
State/Province Texas state
Collection_day 24
Collection_month 4
    
```

1. Based upon this fragment:
 - A. What viral gene is this from?
 - B. What strain of the virus does this seem to be from?
 - C. Based upon the sequence information available, where would you guess this person was exposed to this virus?
 - D. What antivirals do you believe the virus will or will not respond to, based upon the record associated with this gene?

Download and save the protein sequence.

The screenshot shows the NCBI Influenza Virus Sequence Database Query Builder. A blue thought bubble contains the text: "Can pull up protein sequence of any species of flu, from any nation; time". A red box highlights the date range selector, which includes fields for "year", "month", and "day", and "From:" and "To:" labels. The search criteria table is also highlighted with a red box.

Virus Species	Host	Country/Region	Segment/Protein
any	any	any	any
Influenzavirus A	Avian	Africa	PB2
Influenzavirus B	Blow fly	Asia	PB1
Influenzavirus C	Camel	Europe	PB1-F2

2. Go to the [NCBI Influenza Virus Sequence Database](#). You'll make a tree. Select about 4 sequences each from each:
 - USA H1N1 human flu from 2009
 - USA H1N1 swine flu, and then avian flu, and a flu from another species, all from 2000-2009

Make sure you're comparing the same gene segments for each virus. Click on "full length", and "Remove identical".

Host	Country/Region	Protein	Subtype	Date	Length	Key word	Full-length	Remove identical sequences	NIAID project	Include Lab strains	Pandemic (H1N1) 2009	Number of sequences
Avian	any	HA	H1N1	2000/01/01-2009/07/01	-		yes	yes	no	no	included	23
Human	any	HA	H1N1	2009/01/01-2009/07/01	-		yes	yes	no	no	included	136
Swine	any	HA	H1N1	2000/01/01-2009/07/01	-		yes	yes	no	no	included	60
Camel	any	HA	H1N1	2000/01/01-2009/07/01	-		yes	yes	no	no	included	0
Canine	any	HA	H1N1	2000/01/01-2009/07/01	-		yes	yes	no	no	included	0
Cat	any	HA	H1N1	2000/01/01-2009/07/01	-		yes	yes	no	no	included	0
Civet	any	HA	H1N1	2000/01/01-2009/07/01	-		yes	yes	no	no	included	0
Equine	any	HA	H1N1	2000/01/01-2009/07/01	-		yes	yes	no	no	included	0
Ferret	any	HA	H1N1	2000/01/01-2009/07/01	-		yes	yes	no	no	included	1

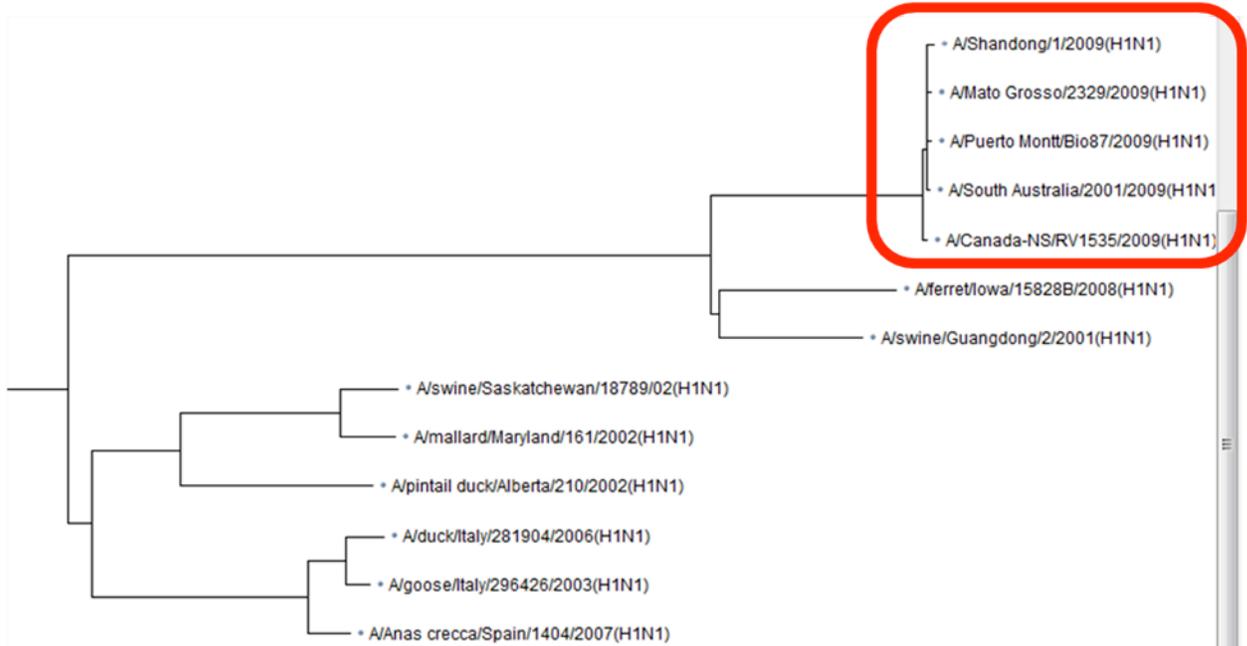
elds

▼ year ▼

- Select FASTA sequences to download - ▼ - Select accession list to download - ▼

host	protein	subtype	country	year	Virus name	219 protein sequences
Avian	HA	H1N1	Canada	2002/08/16	Influenza A virus (A/pintail duck/Alberta/210/2002(H1N1))	
Avian	HA	H1N1	Italy	2003	Influenza A virus (A/goose/Italy/296426/2003(H1N1))	
Avian	HA	H1N1	Italy	2006	Influenza A virus (A/duck/Italy/281904/2006(H1N1))	
Avian	HA	H1N1	Italy	2007	Influenza A virus (A/duck/Italy/69238/2007(H1N1))	
Avian	HA	H1N1	Spain	2007	Influenza A virus (A/Anas crecca/Spain/1404/2007(H1N1))	
Avian	HA	H1N1	Spain	2007	Influenza A virus (A/Anas crecca/Spain/1384/2007(H1N1))	
Avian	HA	H1N1	Spain	2007	Influenza A virus (A/Anas platyrhynchos/Spain/1365/2007(H1N1))	
Avian	HA	H1N1	USA	2002	Influenza A virus (A/mallard/Maryland/161/2002(H1N1))	
Avian	HA	H1N1	USA	2002	Influenza A virus (A/mallard/Maryland/369/2002(H1N1))	
Avian	HA	H1N1	USA	2002	Influenza A virus (A/blue winged teal/TX/43/2002(H1N1))	
Avian	HA	H1N1	USA	2002	Influenza A virus (A/mallard/Maryland/02-181/2002(H1N1))	
Avian	HA	H1N1	USA	2002	Influenza A virus (A/mallard/Maryland/02-184/2002(H1N1))	
Avian	HA	H1N1	USA	2002	Influenza A virus (A/mallard/Maryland/02-359/2002(H1N1))	
Avian	HA	H1N1	USA	2002	Influenza A virus (A/blue winged teal/TX/27/2002(H1N1))	
Avian	HA	H1N1	USA	2002/07/25	Influenza A virus (A/mallard/Maryland/350/2002(H1N1))	
Avian	HA	H1N1	USA	2003	Influenza A virus (A/mallard/Maryland/42/2003(H1N1))	
Avian	HA	H1N1	USA	2003	Influenza A virus (A/mallard/Maryland/13/2003(H1N1))	
Avian	HA	H1N1	USA	2003	Influenza A virus (A/mallard/Maryland/199/2003(H1N1))	
Avian	HA	H1N1	USA	2003	Influenza A virus (A/mallard/Maryland/53/2003(H1N1))	
Avian	HA	H1N1	USA	2005/03/01	Influenza A virus (A/muscovy duck/New York/21211-5/2005(H1N1))	
Avian	HA	H1N1	USA	2007/08/11	Influenza A virus (A/northern pintail/Interior Alaska/1/2007(H1N1))	

Which viruses are these genes from the human H1N1 flu viruses most closely related to? Print screen and paste a copy of the screen from which you're making your conclusions into this document here.



3. What part of the protein is most important for vaccines? Open another window, and we'll examine the protein structure. First, go to the protein database "pdb.org". There are many entries for hemagglutinin, but we'd like to see how it interacts with the immune system. So search for "[HEMAGGLUTININ and ANTIBODY](#)". Take note of the 4 character code next to the check box.

We could examine the protein structure by clicking on the name, but that requires a plug-in that isn't installed on campus computers. So instead, google "firstglance", and search for the code you copied down. "1QFU". Chain 'A' is the viral protein we're looking at; H and L are chains of an ANTIBODY molecule that's bound to the virus.

3. Based upon your alignment, and the "firstglance" model for this protein, if you were to make an 'artificial' vaccine against this new flu, which part of the protein would you want to include [give a rough range of about 50 amino acids], and why?

4. A. Go back to the Genbank page. Find "vaccines" on the menu bar on the left. Look up the contents of the 2008-2009 flu vaccine. What is the name of the H1N1 virus in the vaccine?

Influenza Virus Vaccine for the 2009-2010 Season

FDA's Vaccines and Related Biological Products Advisory Committee (VRBPAC) met in Silver Spring, Maryland, on February 18, 2009, to select the influenza virus strains for the composition of the influenza vaccine for use in the 2009-2010 U.S. influenza season. During this meeting, the advisory panel reviewed and evaluated the surveillance data related to epidemiology and antigenic characteristics, serological responses to 2008/2009 vaccines, and the availability of candidate strains and reagents.

The panel recommended that vaccines to be used in the 2009-2010 influenza season in the U.S. contain the following.

- an A/Brisbane/59/2007 (H1N1) like virus; *
- an A/Brisbane/10/2007 (H3N2) like virus; **
- a B/Brisbane/60/2008-like virus.

Query Builder

Subtype	Date	Length	Key word	Full-length	Remove identical sequences	NIAID project	Include Lab strains	Pandemic (H1N1) 2009	Number of sequences
any	-	-	A/Brisbane/59/2007	yes	yes	no	yes	included	2
any	2009/02/01-2009/06/01	-		yes	yes	no	yes	included	164
H1N1	2009/02/01-2009/06/01	-		yes	yes	no	yes	included	123

B. Using these same tools, add the sequence of the hemagglutinin protein of that H1N1 virus that the 2009-2010 vaccine was based upon. Align a few of these with a few of the 2009 pandemic H1N1.

What % identities do you find? Are they very similar in the area you found to be important in question #4?

D. Based upon this result, without further information, would you guess that last year's vaccine would provide much protection against the current pandemic flu? Explain your reasoning in a few sentences.

Alignment for 5 protein sequences. Alignment length is 566.

[a tree](#)
[Download alignment](#)
[Print-friendly version](#)
 Go to position

19 57 85 114 142 170 199 227 255 284 312 3

```

..60.....70.....80.....90.....100.....110.....120.....130.....140.....150.
LCKLRGVAPLHLGKCNIAAGWILGNPECESLSTASSWSYIVETSSSDNGTCYPGDFIDYEELREQLSSVSSFERFEIFPKTSWPNHDTINKGVTIAC
..L.K.I..Q..N.SV.....L.ISKE.....KPNPE.....H.A.....E.....-..VT..S.S.
..L.K.I..Q..N.SV.....L.ISKE.....KPNPE.....H.A.....E.....-..VT..S.S.
.....S.
.....S.
    
```

340

```

270.....
YAFAMERNAG
...LS.GF.
...LS.GF.
.....
.....
    
```