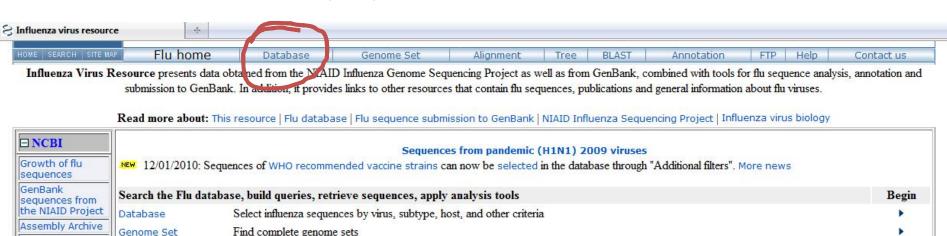
ASSIGNMENT 2

- **Dataset 2** Design of a study scheme from available sequence information of influenza viruses and collecting data.
 - Phylogeny analyses by MEGA-software (neighbor-joining and UPGMA trees and parsimony cladogram). No mrBayes- analyses with this dataset (you can do them, if you want). If enough time, network-phylogies.

■ Time schedule:

- Wed 2. Feb from 12 to 20 there will be extra help for you in C128. This concerns data collection and alignments for datasets 1 and 2. The week after (maybe in Wed 9. Feb, need to be agreed) you can again come to meet extra help and check your alignments.
- Data-collection and MEGA-analyses ready during weeks (3) and 4. Work should be ready in 22. Feb.
- Recommendation is that you don't work alone, instead form groups.

http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html



Find complete genome sets Trace Archive Alianment Align your sequence(s) and others in the database (up to 1000 sequences) NIAID data releasing status Tree Build a clustering or phylogenetic tree RefSeq genomes BLAST BLAST a flu sequence against the database RefSeg proteins Annotate flu genomic sequences Annotation Protein Structures Submission Submit flu sequences to GenBank ∃Flu resources Retrieve database and sequence data through ftp NIAID Project JCVI Flu Sample Searches Begin Influenza Research Full-length HA proteins of the H3 subtype Influenza A virus in USA from 1998 to 2002. Database Complete genome sets of the Pandemic (H1N1) 2009 Influenza A virus from Japan with the "H274Y" Drug-resistance mutaion in the NA protein. CDC Flu Vaccine Selection New records in Entrez WHO Flu · Publications on influenza viruses from the past 2 weeks **■NCBI Viruses**

All influenza sequences updated in GenBank from the past week

Viral Genomes Virus Variation

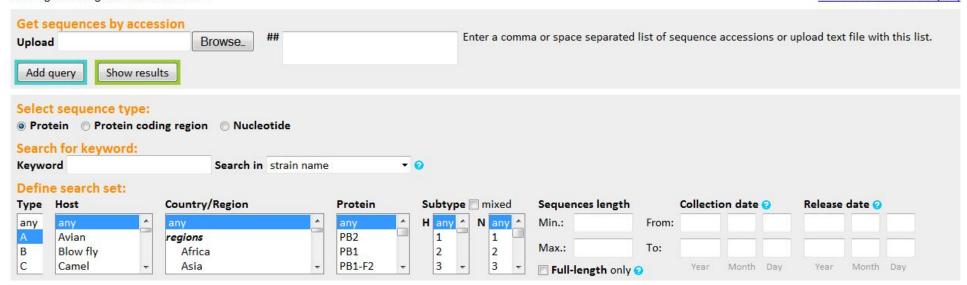


Influenza Virus Sequence Database

Protein or nucleotide sequences can be retrieved from the database using GenBank accession numbers or search terms.

Multiple queries can be built by clicking the "Add Query" button every time a new query is made, and queries in any combination from the Query Builder can be selected to get sequences in the database. Sequences can be downloaded, and it is possible to analyze them using the multiple sequence alignment or tree building tool integrated to the database.

Permanent link for this query



■ The recent pandemic H1N1 is an A-virus which has type 1 of H (haemagglutinin) gene and type 1 of N (neuraminidase) gene. Within type 1 (and other types as well, for example the bird flu H5N1) very much sequence evolution occurs => lots of information for tracing the virus (its movement, geographical, temporal differences etc.)

- Construct a comparative plan, for example H1N1 in different countries or continents, maybe at a certain time, temporal changes within a country / continent, virus sequences from human vs animals, or H1N1 vs H5N1, whatever you like, based on the existing information.
- Take as many sequences as you like. Working with more than 500 is not very practical and alignments are very time-consuming.
- It is not a good idea to take many identical or nearly identical sequences (how to avoid that, is your problem.....).
- Restrict your study to H (taking both H and N is not practical, too much work).
- Take only "full-length" sequencies (if you include one sequence from which, say, 200 nucleotides are missing from the 3´ end, this part is not considered in your phylogeny analyses, i.e. you miss the information from your whole data).
- Probabaly not much editing after alignments, some sequence items have a short "extra" part at the 5´ end. Discard that.

■ In the query result below, from one virus "particle" from Bamako information from many genes: In addition to your study object, H (HA) there are the genes PB2, PA, NP, NA, M1, M2, NS1,NS2. From the samples of Senegal, Djibouti, South Africa and Egypt only HA.

	1200000				2.02.00		
▼ ADM14980	219	Human	NS1	H1N1	Ethiopia	2009/12/01	Influenza A virus (A/Addis Ababa/WR2848T/2009(H1N1)
▼ <u>ADM14981</u>	121	Human	NS2	H1N1	Ethiopia	2009/12/01	Influenza A virus (A/Addis Ababa/WR2848T/2009(H1N1)
▼ ADM14955	759	Human	PB2	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
ADM14956	716	Human	PA	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
ADM14957	566	Human	HA	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
ADM14958	498	Human	NP	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
ADM14959	469	Human	NA	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
ADM14960	252	Human	M1	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
ADM14961	97	Human	M2	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
ADM14962	219	Human	NS1	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
ADM14963	121	Human	NS2	H1N1	Mali	2009/12/18	Influenza A virus (A/Bamako/WR2361N/2009(H1N1))
ACT85993	29	Human	NP	H1N1	Mauritius	2009/06/25	Influenza A virus (A/Candos/1/2009(H1N1))
ACG50709	466	Human	NA	H1N1	South Africa	2008/05/27	Influenza A virus (A/CapeTown/26/2008(H1N1))
ACG50719	325	Human	HA	H1N1	South Africa	2008/05/27	Influenza A virus (A/CapeTown/26/2008(H1N1))
CAD29928	346	Human	HA	H1N1	Senegal	1997	Influenza A virus (A/Dakar/11/97(H1N1))
CAD29924	355	Human	HA	H1N1	Senegal	2000	Influenza A virus (A/Dakar/17/2000(H1N1))
ADG21157	566	Human	HA	H1N1	Djibouti	2009/09	Influenza A virus (A/Djibouti/N11092/2009(H1N1))
ADG21185	566	Human	HA	H1N1	Djibouti	2009/12/08	Influenza A virus (A/Djibouti/N13142/2009(H1N1))
ABQ53688	470	Human	NA	H1N1	South Africa	1997	Influenza A virus (A/Durbin/113/1997(H1N1))
CAD29908	363	Human	HA	H1N1	Egypt	2001	Influenza A virus (A/Egypt/101/2001(H1N1))
ADM26271	566	Human	HA	H1N1	Egypt	2010/01	Influenza A virus (A/Egypt/N00124/2010(H1N1))

- When you click the accession number of a certain sequence, you first get the protein (amino acid) sequence.
- Note the link to nucleotide sequence and links to many other facilities.

```
AUTHORS Lin, Y.
 TITLE
             Direct Submission
                                                                                                                                           All links from this record
  JOURNAL Submitted (19-APR-2002) Lin Y., Virology, National Institute for
             Medical Research, The Ridgeway, Mill Hill, London, NW7 1AA, UNITED
                                                                                                                                           BLink
                                                                                                                                           Related Sequences
FEATURES
                       Location/Qualifiers
                                                                                                                                           CDD Search Results
     source
                       /organism="Influenza A virus (A/Egypt/101/2001(H1N1))"
                                                                                                                                           Conserved Domains (Concise)
                       /strain="(A/Egypt/101/2001(H1N1))"
                                                                                                                                           Conserved Domains (Full)
                       /serotype="A"
                       /db xref="taxon:192544"
                                                                                                                                           Domain Relatives
                       1..363
     Protein
                                                                                                                                           Encoding mRNA
                       /product="haemagglutinin"
                                                                                                                                          ►Nucleotide
                       20..>363
     Region
                                                                                                    Nucleotide records that are the source of
                       /region name="Hemagglutinin"
                                                                                                    protein records in the current set. The
                       /note="Hemagglutinin; pfam00509"
                                                                                                    protein sequences are generated through
                                                                                                                                           Related Structures (List)
                       /db xref="CDD:109560"
                                                                                                    translation of coding region features on
                       1..363
     CDS
                                                                                                                                           Related Structures (Summary)
                                                                                                    the nucleotide records.
                                                                                                                                           Taxonomy
                       /coded by="AJ457871.1:<4..>1092"
                       /db xref="GOA:Q8AZG4"
```

■ FASTA-link is at the left up corner of the page.

■ You get the sequence in FASTA-format like this:

>gi|22859141|emb|AJ457871.1| Influenza A virus (A/Egypt/101/2001(H1N1)) partial HA gene for haemagglutinin, genomic RNA

- The title (what follows the mark >) is too long and in your subsequent analyses (already after Clustal alignment) you see only some first part of it .
- Edit the title into a reasonable form, discard everything that is not necessary!
- This is an example of a reasonable title: >Egypt/101/2001 AJ457871