This document discusses the genomes/INFLUENZA folder found on the ncbi website, found here: https://ftp.ncbi.nlm.nih.gov/genomes/INFLUENZA/ [link 1]

NCBI is the National Center for Biological Information.

NCBI Server

The INFLUENZA folder is a subfolder of the ncbi file transfer protocol (ftp) server. Basically, This means that it has files. This parent folder can be found here: https://ftp.ncbi.nlm.nih.gov/ [link 2]
The REAMDE of this folder states “You have reached the NCBI ftp server. NOTE: ALL DATA HERE IS PUBLIC, NON-SENSITIVE, UNRESTRICTED SCIENTIFIC DATA SHARING AMONG SCIENTIFIC COMMUNITIES. THESE SERVERS ARE INTENTIONALLY PUBLIC.” Which you can also read here: https://ftp.ncbi.nlm.nih.gov/README.ftp [link 3], or find within the folder by clicking on it.

Figure 1: screenshot of the ncbi ftp server
(taken on December 11 2022)
All links and text on the website are included in the screenshot.

The genomes folder is within this, and then the INFLUENZA folder is within the genomes folder. All files and folders not within the genomes/INFLUENZA folder are not discussed here; only a small portion of the server’s files are within the INFLUENZA folder.

genomes/INFLUENZA

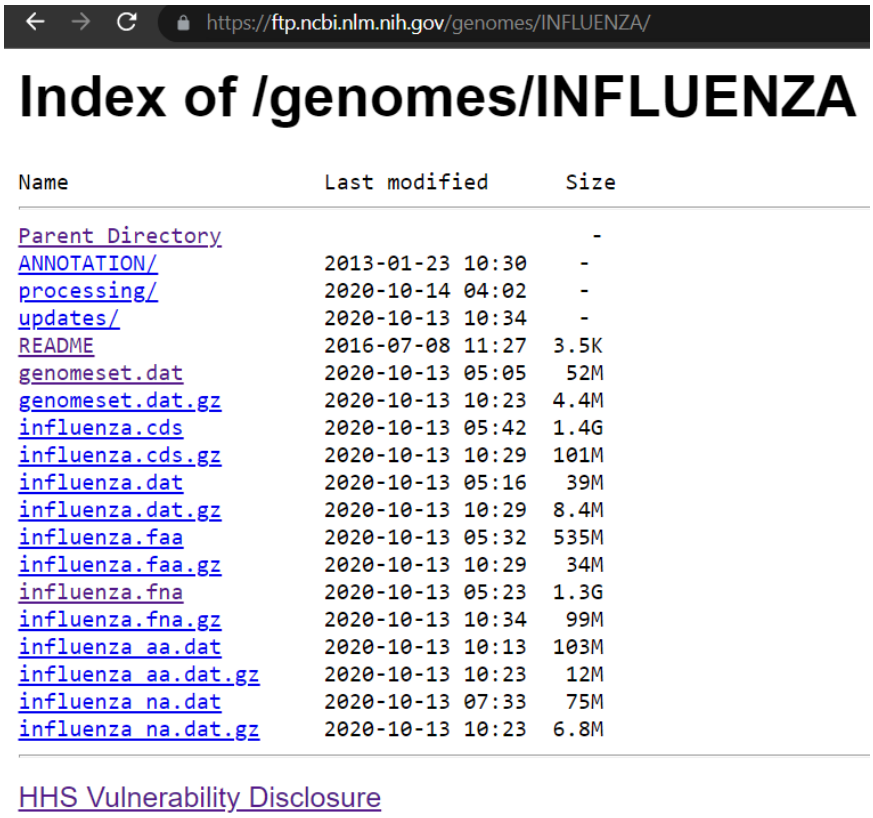
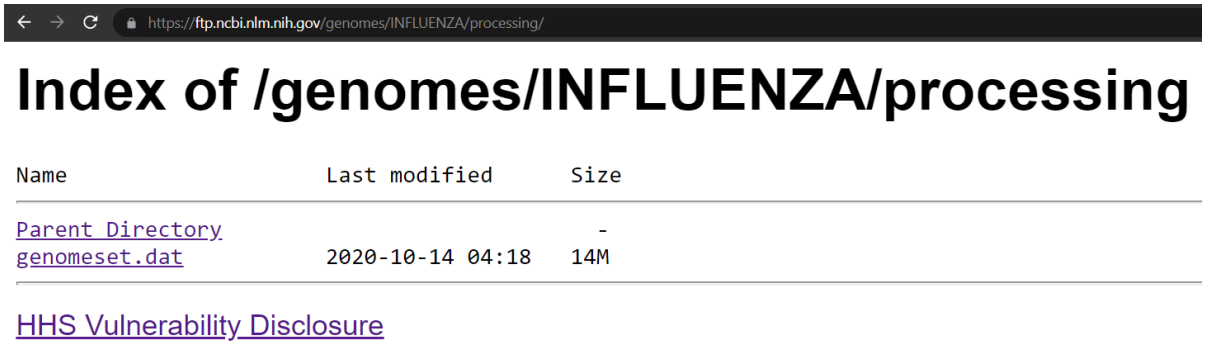


Figure 2: screenshot of the genomes/INFLUENZA folder, again found here: https://ftp.ncbi.nlm.nih.gov/genomes/INFLUENZA/ [link 1] (taken on December 11 2022)
All links and text on the website are included in the screenshot.

The Parent Directory link goes back to the genomes folder.

The ANNOTATION folder contains reference sequences used in the Influenza Virus Sequence Annotation Tool http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/annotation.cgi

The processing folder contains the file genomeset.dat. It was modified more recently than the genomeset.dat in the genomes/INFLUENZA folder, and is a different size.



The updates folder contains folders for each day from 2020-09-13 to 2020-10-13.
README file

genomeset.dat

contains a table with supplementary genomeset data
I have also split up the data in this file into multiple files, and uploaded this data to github: https://github.com/belle172/NCBI_data/tree/main/NCBI_genomes_influenza_genomeset_dat

genomeset.dat.gz - .gz compressed file of genomeset.dat

influenza.cds influenza.cds.gz

influenza.dat influenza.dat.gz - Nucleotide, protein, and coding region IDs

influenza.faa influenza.faa.gz - FASTA protein sequences as amino acids

influenza.fna - influenza.fna.gz
FASTA nucleotide sequences

influenza_aa.dat influenza_aa.dat.gz - Supplementary protein data
influenza_na.dat influenza_na.dat.gz - Supplementary nucleotide data

FUTURE WORK:
Look into the files other than genomeset.dat and influenza.fna.
Why hasn't the INFLUENZA folder been updated since 2020? Does NCBI have genomes from the last few years available somewhere else?

Analysis of genomeset.dat and influenza.fna

genomeset.dat

file of header lines for complete genomes. According to the README from NCBI: "The genomeset.dat file contains information for sequences of viruses with a complete set of segments in full-length (or nearly full-length). Those of the same virus are grouped together (using an internal group ID that is shown in the last column of the file) and separated by an empty line from those of other viruses."
https://ftp.ncbi.nlm.nih.gov/genomes/INFLUENZA/README
This file is important as it only contains information for at least nearly complete genomes, whereas influenza.fna also contains sequences for samples that do not have all proteins sequenced. This file also has some columns of information that the header lines in influenza.fna do not contain.
The file was split up by metadata information using the following script:
https://github.com/belle172/NCBI_data/blob/main/NCBI_genomes_influenza_genomeset_dat/protein_extractor_humans.py
As well as with a few other scripts found in the relevant folders on github, until I got metadata I was interested in.

influenza.fna

This file contains the nucleotide [ATGC] sequences for 1.4 GB of influenza sequences, so I think it is all of the data in one file. Compressed, the file is still too big for me to upload to github, so if you also want the whole file download it from here:
https://ftp.ncbi.nlm.nih.gov/genomes/INFLUENZA/influenza.fna
Alternatively, you could download the compressed file from NCBI at this link
https://ftp.ncbi.nlm.nih.gov/genomes/INFLUENZA/influenza.fna.gz
but I downloaded the whole file directly. When doing that, there was a lot of newline characters ['\n'] in the middle of sequences, so I wrote the following python script to put all the sequences on one line, found on my github here:
https://github.com/belle172/NCBI_data/blob/main/influenza_fna/influenza_file_fixer.py
After removing the whitespace, the original 1,395,602 KB went to 1,379,411 KB.

The next section discusses the code to combine the header lines of the complete genomes from genomeset.dat with their sequences in influenza.fna. Doing so will work with the entire influenza_fixed.fasta file, but for time I also split up influenza_fixed.fasta by metadata into only the sequences from Minnesota, using the script found here:
https://github.com/belle172/NCBI_data/blob/main/influenza_fna/protein_extractor_from_total.py

Which created Minnesota_influenza.txt, found here:
https://github.com/belle172/NCBI_data/blob/main/influenza_fna/Minnesota_influenza.txt

Combining genomeset.dat and influenza.fna

In the folder of the genomeset.dat file split up by metadata, there is the script genome_assembler.py:
https://github.com/belle172/NCBI_data/tree/main/NCBI_genomes_influenza_genomeset_dat/genomeset_dat/human_influenza/human_MN

This script takes in the genomeset.dat file or an equivalent file of influenza headers, such as the file human_MN.txt, which is comprised of the lines from genomeset.dat that are from both humans and Minnesota. Running the genome_assembler.py script produces the file MN_genome_headers.txt. You can change the line 'seqs_file = open('human_MN.txt')' to genomeset.dat or any of its other metadata files found in the repository, but the script does also expect complete genomes, and for those complete genomes to be 8 segments. For running the script on a file of genomes that has a different number of segments per genome, the while loop line would also have to be changed.

MN_genomes_headers.txt, the file created, is a header file like human_MN.txt, just reformatted in the way the next script expects. Crucially, it organizes the genomes by genome, so now each genome has a header line, and then the header lines for its segments, like so:

```
>(A/Minnesota/22/2014(H3N2))  H3N2    2014/09/20      1
KT837179        segment 1       length 2280
KT837213        segment 2       length 2274
KT837318        segment 3       length 2151
KT837098        segment 4       length 1701
KT837183        segment 5       length 1497
KT837194        segment 6       length 1410
KT837230        segment 7       length 982
KT837173        segment 8       length 838
```

With the data for the whole genome stored on the first line, in the format
>(sample) \t (virus subtype) \t (date) \t (genome number)
Then the segment data for that sample's genome in the following lines, in the format
(genBank id) \t (segment number) \t (length of segment)

influenza_fna folder

Now the file MN_genome_headers.txt contains the information for assembled genomes, in the format we want, of the human Minnesota metadata we want. With the sequences file influenza.fna or a file created of just some of its data, such as Minnesota_influenza.txt, again found here:
https://github.com/belle172/NCBI_data/blob/main/influenza_fna/Minnesota_influenza.txt
these two files can finally be joined to create a file for each genome, using the script genome_files_writer.py, also found in the influenza_fna folder on github:
https://github.com/belle172/NCBI_data/tree/main/influenza_fna
For the sake of just downloading and running the code, I have also added the MN_genome_headers.txt file to this folder. The script needs a folder to put the genome files, in this instance influenza_MN_genomes/. After running the script, the influenza_fna/influenza_MN_genomes folder now contains the assembled genome files.

https://github.com/belle172/NCBI_data/tree/main/influenza_fna/influenza_MN_genomes
We have reached the source of assembled genomes, with the metadata of being found in Minnesota.

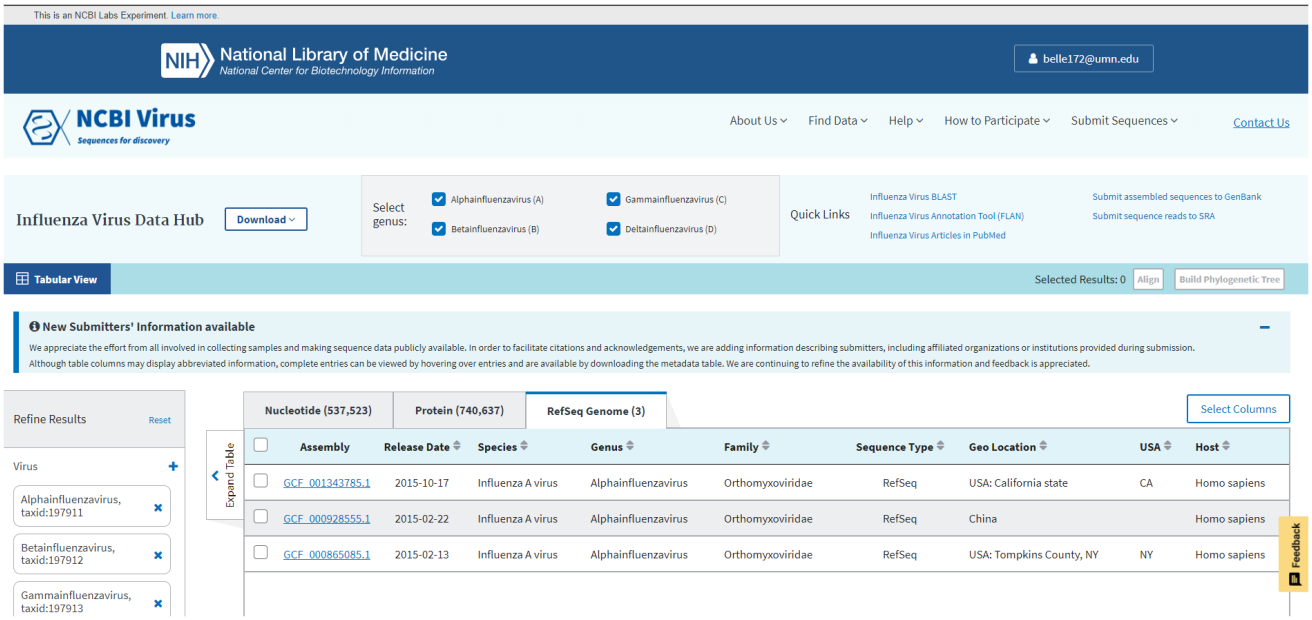
There are 382 genomes spread out over year of collecting, all from samples in Minnesota.

Separate gene reference files

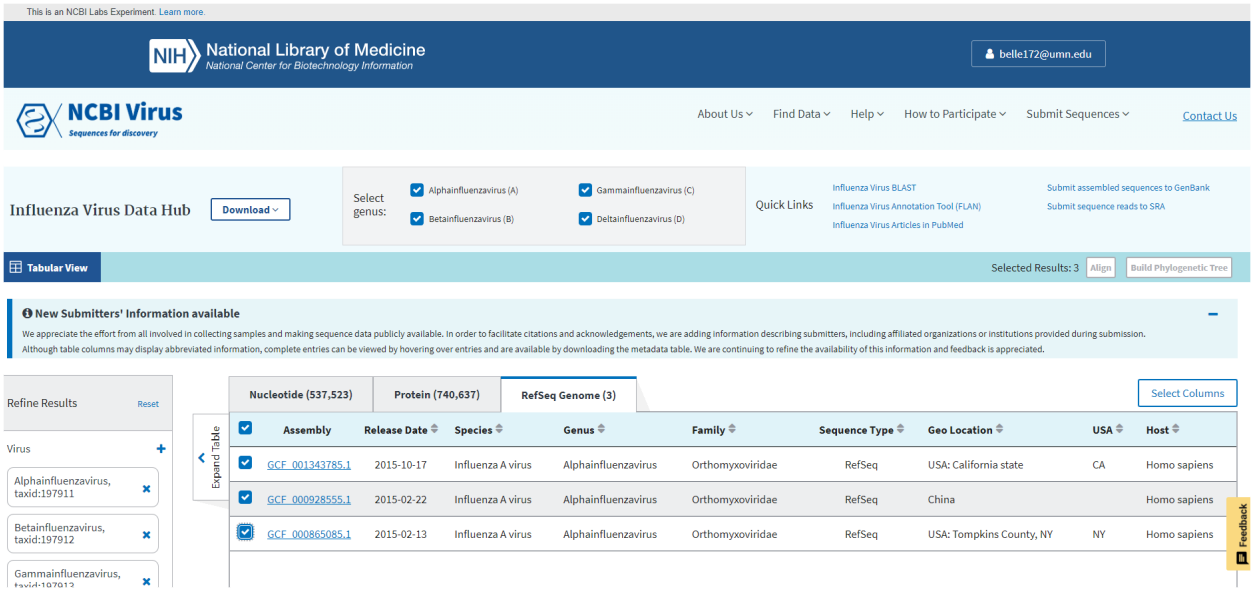
On NCBI, when filtering for Influenza A and B in humans, there are 3 reference genomes:

https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Genome&VirusLineage_ss=taxid:197911&VirusLineage_ss=taxid:197912&VirusLineage_ss=taxid:197913&VirusLineage_ss=taxid:1511083&HostLineage_ss=Homo%20sapiens%20(human),%20taxid:9606

This is what the website looks like:

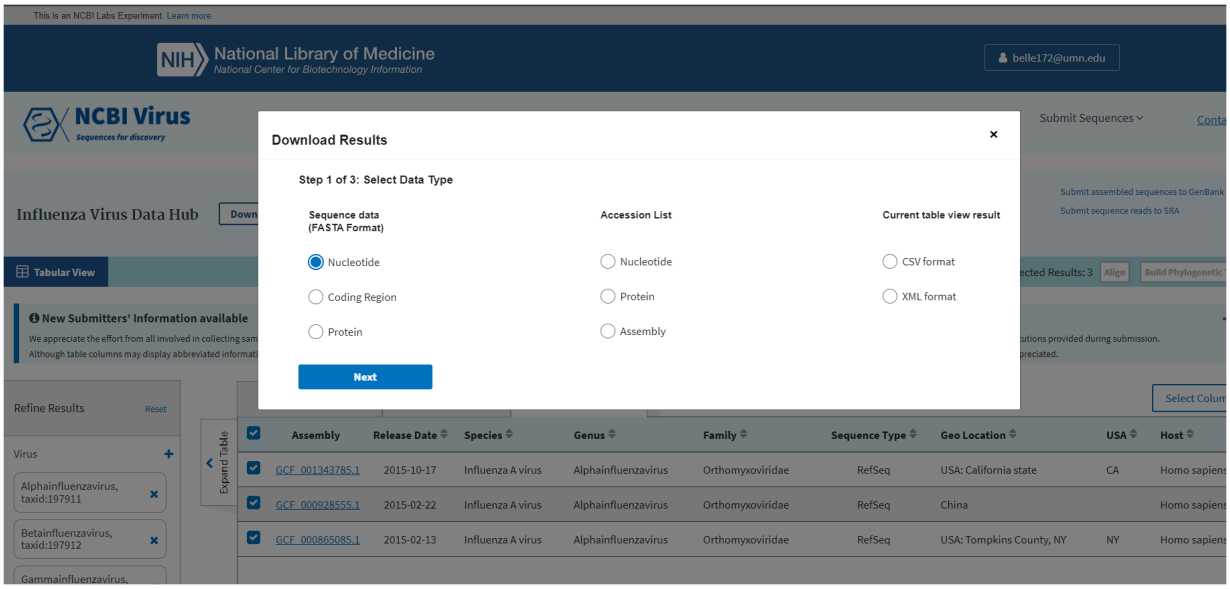


Select the three reference genomes by clicking on the check boxes next to them:

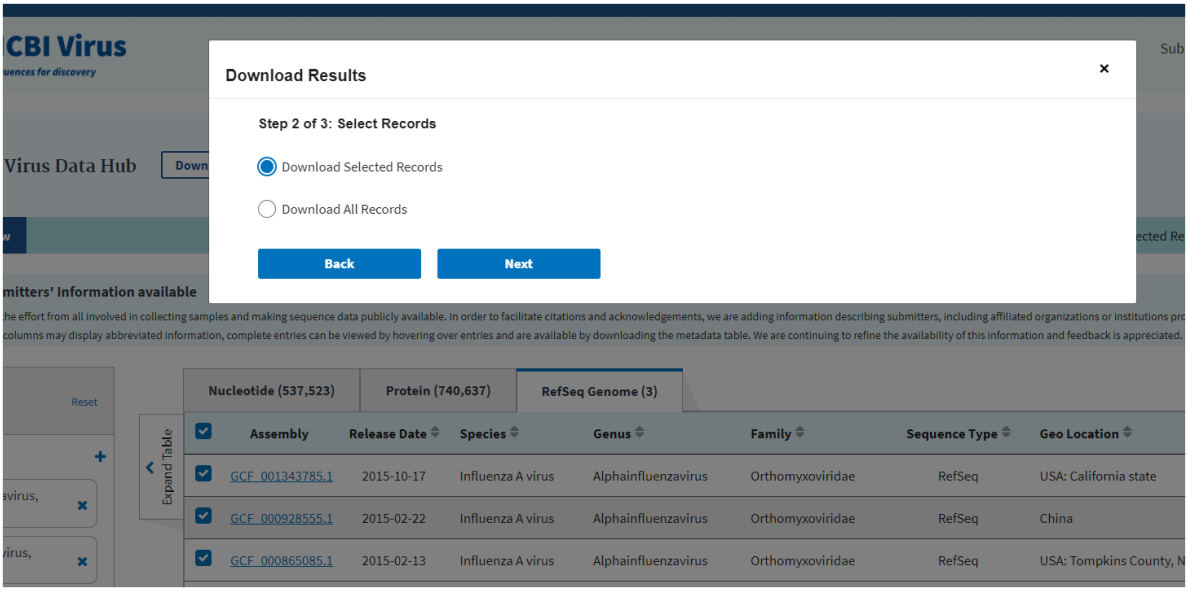


Click on the Download button next to Influenza Virus Data Hub on the left side.

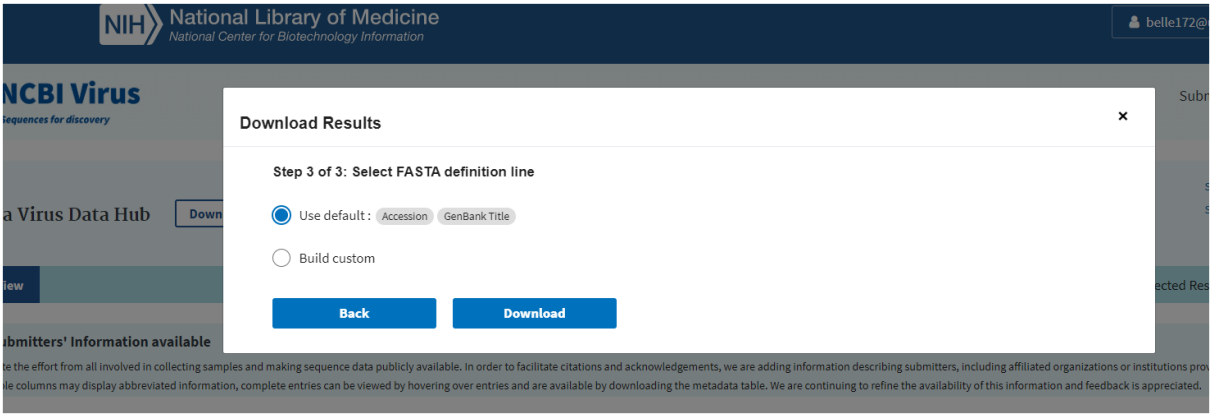
That comes to this view



With Nucleotide selected (blue dot next to it) click Next,



Then with Download Selected Records click Next,



Use default, then Download. The download is named sequences.fasta as its default:

You can also find this file on github here:

https://github.com/belle172/NCBI_data/blob/main/sequences.fasta

Notice that the segments are not in order, the genomes aren't in separate files, and the sequences span multiple lines instead of each sequence being on one line. You could just select one genome at a time and download, but since separating a whole file that contains multiple genomes is already what we did to get the Minnesota genomes, I just did that again to get this folder of the separate gene reference genomes:

https://github.com/belle172/NCBI_data/tree/main/human_influenza_ref_genomes

Extract regions of each genome

Each influenza genome file in the human_influenza_ref_genomes folder and the influenza_MN_genomes folder contain 8 regions.

Running the script genomes_into_regions.py

found here https://github.com/belle172/NCBI_data/blob/main/genomes_into_regions.py

creates the folder proteins

https://github.com/belle172/NCBI_data/tree/main/proteins

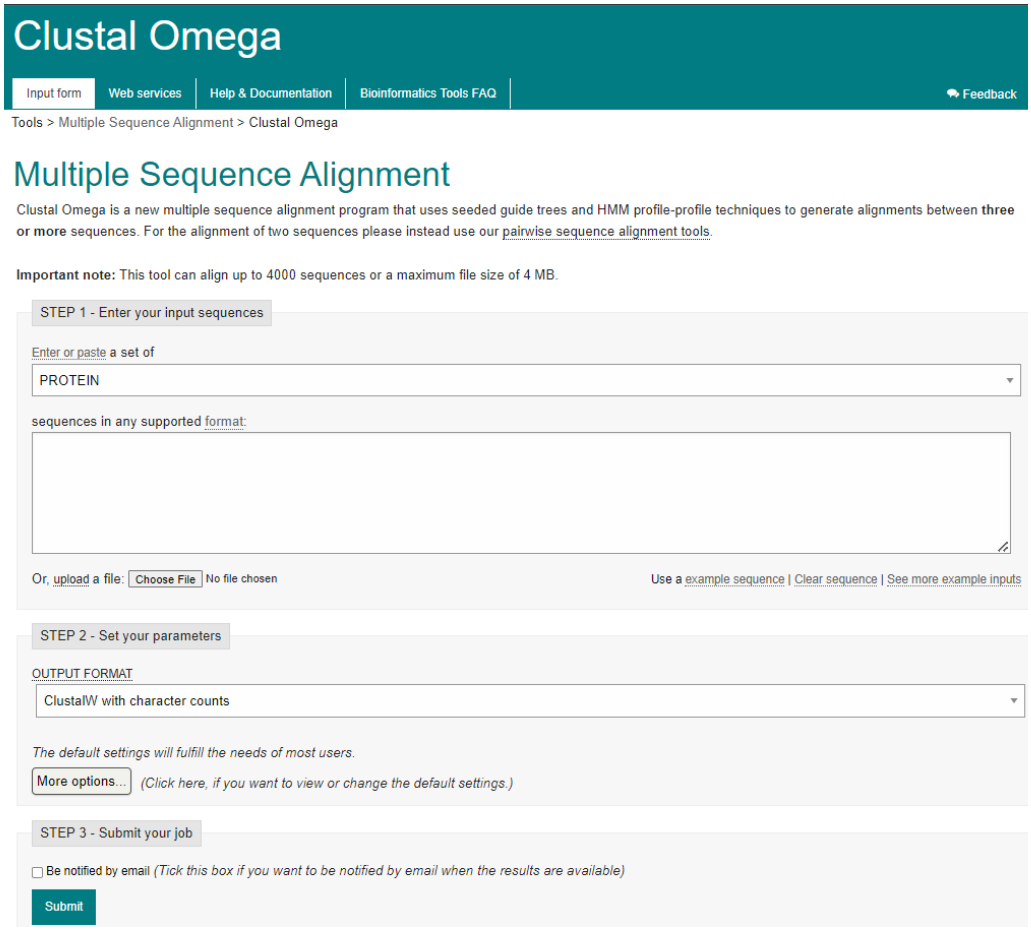
which then contains a folder for each reference genome in the human_influenza_ref_genomes folder. Each of the reference genome folders contains 8 files, one for each region of the influenza genome.

Now we have used alignment again the separate gene reference files to extract the regions of each genome.

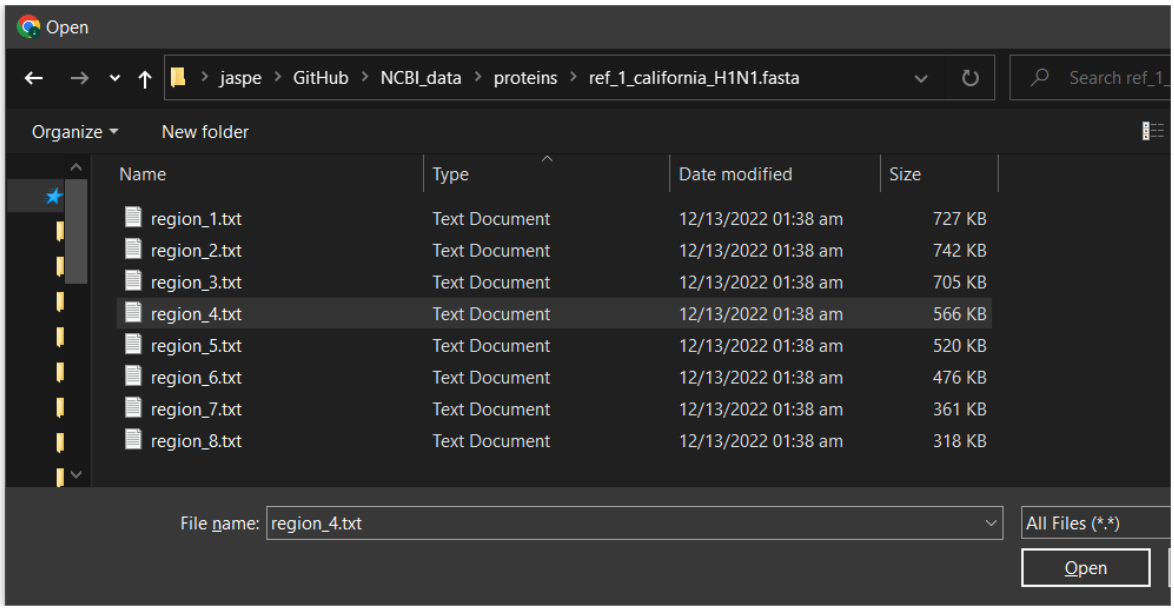
Multiple Sequence Alignment (MSA)

Using the online tool clustal, https://www.ebi.ac.uk/Tools/msa/clustalo/ , I obtained multiple sequence alignment files for each protein region by uploading the region files obtained in the last step to clustal.

https://www.ebi.ac.uk/Tools/msa/clustalo/ webpage:

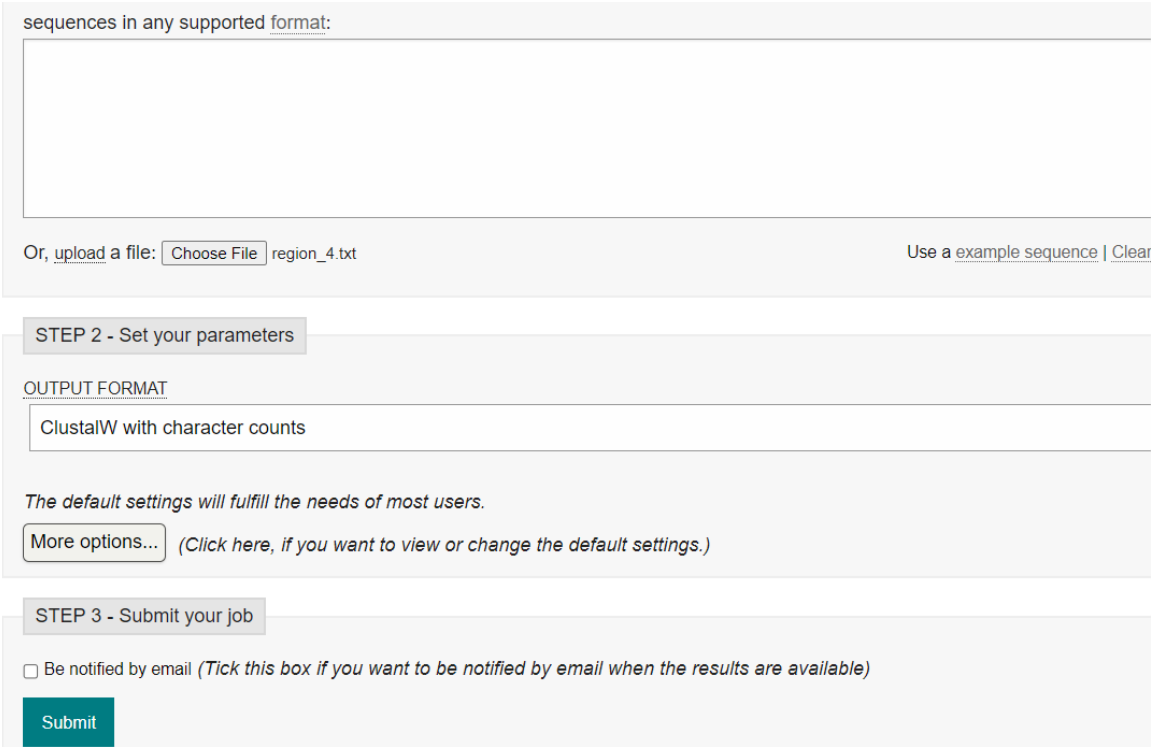


Click the Choose File button, then navigate to the proteins/<reference genome> folder, for our current usage that will be ref_1 regions 4,5,6.



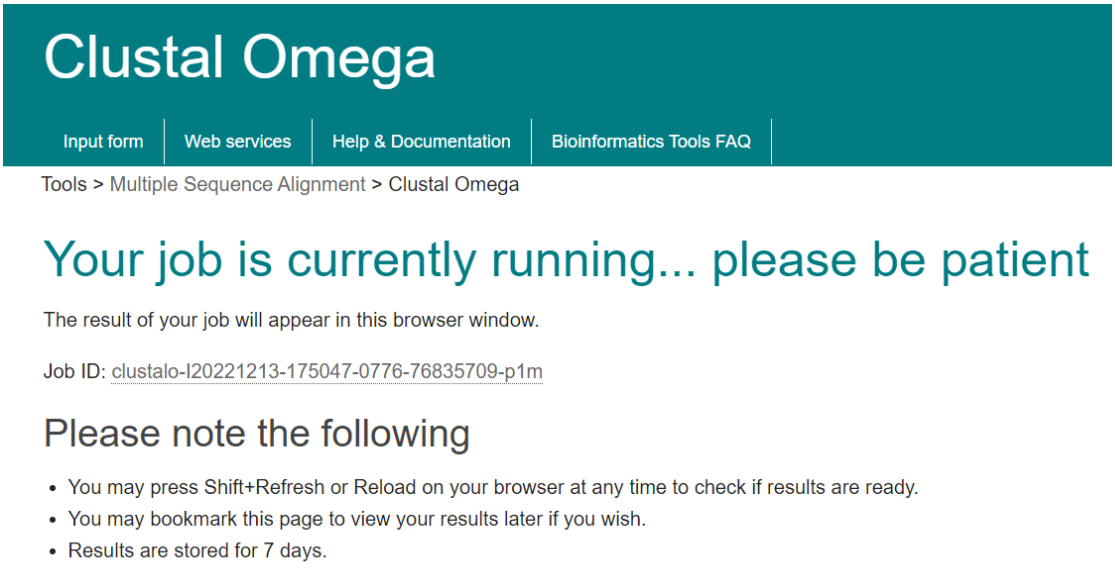
Click Open.

Then you should see the file name next to the Choose File button on the clustal website (region_4.txt)



Click the Submit button, and wait.

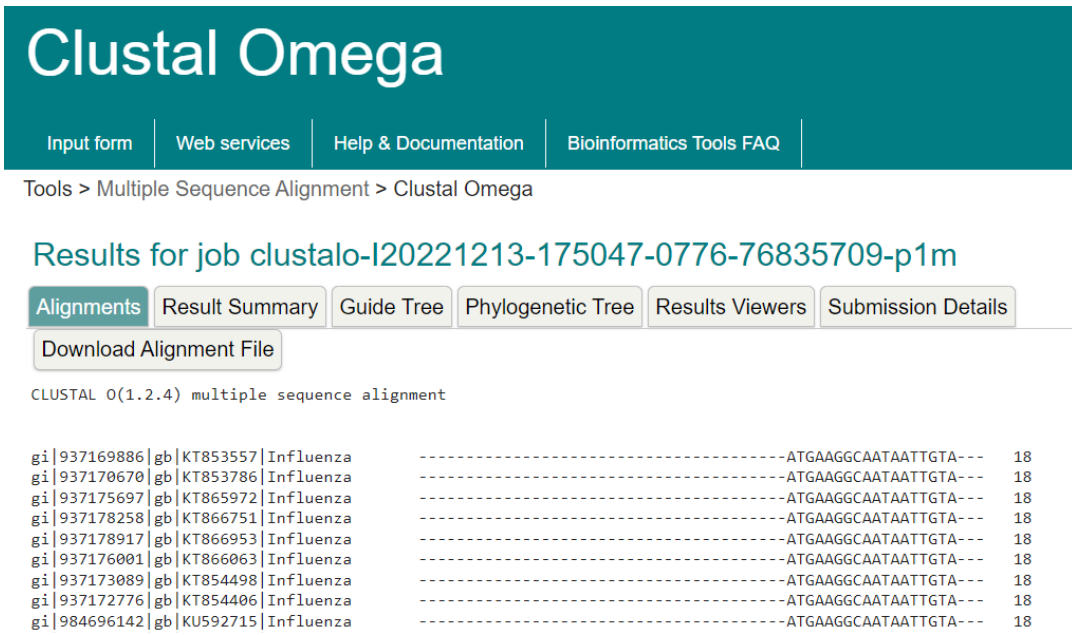
After a few seconds you should get this page:



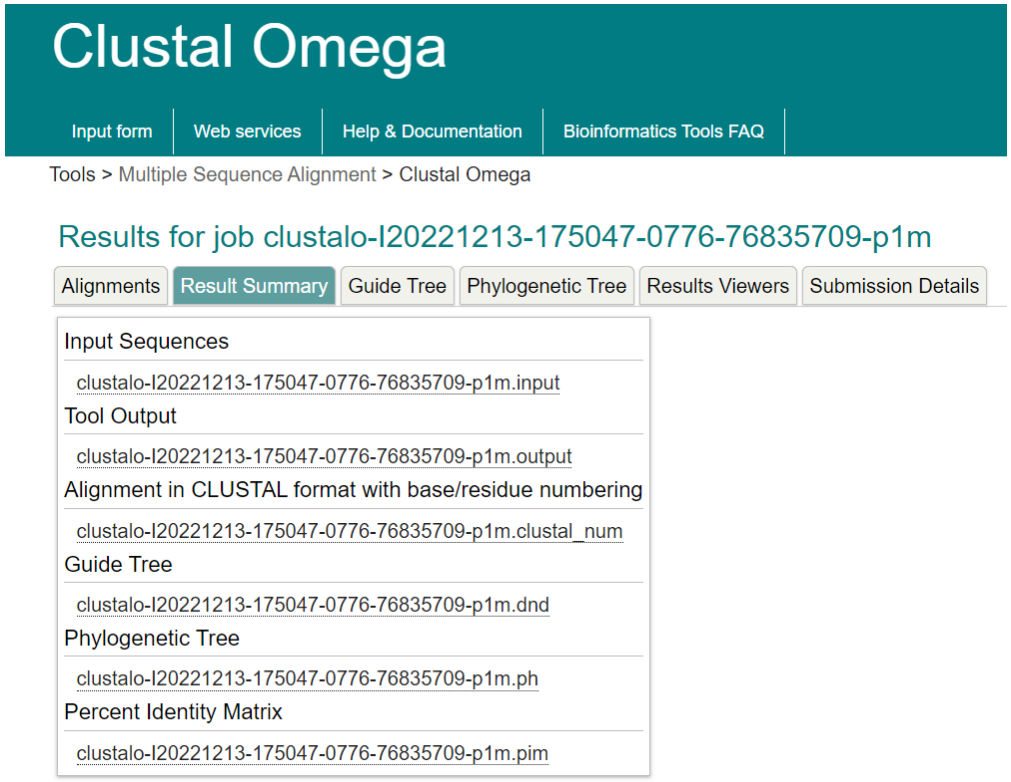
Then after about 5 minutes for each of the files, the results were ready. Note that you do not need to keep the tab open, as long as you save the link you can access the results for 7 days, for instance the results of uploaded region_1.txt can be found here, until December 20 2022:

https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20221213-175047-0776-76835709-p1m

The results page:



Click on the second tab, Results Summary.



For the rest of the analysis, we only used the ‘Phylogenetic Tree’ and ‘Alignment in CLUSTAL format’ files. Download the 6 files on the clustal page, or just the ones you want. You can find the files from running clustal on github here: https://github.com/belle172/NCBI_data/tree/main/clustal_files Where each folder in the clustal_files/ folder has the clustal results from uploading the corresponding file in the proteins folder.

The file named alignment_clustal_format.clustal_num in each of the clustal folders is the multiple sequence alignment of that protein in all the genomes. For instance, the genome reference 1 segment 4 file that was uploaded to clustal can be found here: https://github.com/belle172/NCBI_data/blob/main/proteins/ref_1_california_H1N1.fasta/region_4.txt
The results from that file, downloaded from clustal, are here:
https://github.com/belle172/NCBI_data/tree/main/clustal_files/clustal_ref1_segment4
The clustal format alignment file: https://github.com/belle172/NCBI_data/blob/main/clustal_files/clustal_ref1_segment4/alignment_clustal_format.clustal_num

Clustal citation:
Madeira F, Pearce M, Tivey ARN, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research. 2022 Apr:gkac240. DOI: 10.1093/nar/gkac240. PMID: 35412617; PMCID: PMC9252731.
Publication: https://europepmc.org/article/MED/35412617

The last step of the multiple sequence alignment just gets the MSA files ready for genetic distance calculations. Convert the .clustal_num files to fasta format, which you can do with this webpage:
https://sequenceconversion.bugaco.com/converter/biology/sequences/clustal_to_fasta.php
All of the multiple sequence alignment files, now in fasta format, can be found here:
https://github.com/belle172/NCBI_data/tree/main/RefGen_fasta
Lastly we once again need to put all the sequences on one line each.

This final version of the multiple sequence alignment files can be found in the fixed_fastas folder:
https://github.com/belle172/NCBI_data/tree/main/fixed_fastas

Calculate genetic distances

All the files discussed for genetic distances are in the fixed_fastas folder, with the multiple sequence alignment files.
To calculate the genetic distance/edit distance between all pairs of sequences in each multiple sequence alignment files, use the script eric_hw3.py found here on each multiple sequence alignment file:
https://github.com/belle172/NCBI_data/blob/main/fixed_fastas/eric_hw3.py
The genetic distance file generated by the script is genetic-distances.txt, which were then manually renamed and saved in the format ‘genetic-distances_ref<refnum>_seg<segment num>.txt’. The genetic distance files we used are in the fixed_fastas folder, for instance the genetic distance file from the multiple sequence alignment of reference genome 1’s segment 4 is genetic-distances_ref1_seg4.txt, found here:
https://github.com/belle172/NCBI_data/blob/main/fixed_fastas/genetic-distances_ref1_seg4.txt

These files contain the calculated genetic distance/edit distance between all pairs of sequences in each MSA.