

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.

The screenshot shows a web browser window with the address bar displaying 'https://ftp.ncbi.nlm.nih.gov'. The main content area is titled 'Index of /' and contains a table listing the files and folders available on the server. The table has three columns: 'Name', 'Last modified', and 'Size'. The 'Name' column lists various folders and files, including '1000genomes/', 'ReferenceSamples/', 'SampleData/', 'asn1-converters/', 'bigwig/', 'bioproject/', 'biosample/', 'blast/', 'cgap/', 'cn3d/', 'comparative-genome-viewer/', 'dbgap/', 'diffexpIR-notebook/', 'entrez/', 'epigenomics/', 'eqtl/', 'fa2htgs/', 'genbank/', 'gene/', 'genomes/', 'geo/', 'giab/', 'hapmap/', 'hmm/', 'mmdb/', 'ncbi-asn1/', 'nist-immse/', 'osiris/', 'pathogen/', 'pub/', 'pubchem-scratch/', 'pubchem/', 'pubmed/', 'rapt/', 'refsam/', 'refseq/', 'repository/', 'seqc/', 'sequin/', 'sky-cgh/', 'snp/', 'sra/', 'tech-reports/', 'toolbox/', 'tpa/', 'variation/', '10GB', '1GB', '50GB', '5GB', 'README.ftp', 'favicon.ico', and 'robots.txt'. The 'Last modified' column shows dates and times, and the 'Size' column shows file sizes, with some files having a size of 0 or a specific size like 10GB, 1.0GB, 50GB, 5.0GB, 2.3K, 3.2K, and 26.

Name	Last modified	Size
<a href="#">1000genomes/</a>	2022-12-10 22:48	-
<a href="#">ReferenceSamples/</a>	2022-12-07 08:32	-
<a href="#">SampleData/</a>	2022-02-07 22:48	-
<a href="#">asn1-converters/</a>	2021-09-22 15:29	-
<a href="#">bigwig/</a>	2022-02-07 22:48	-
<a href="#">bioproject/</a>	2022-12-11 11:10	-
<a href="#">biosample/</a>	2022-12-11 07:58	-
<a href="#">blast/</a>	2022-12-10 22:48	-
<a href="#">cgap/</a>	2004-09-13 11:14	-
<a href="#">cn3d/</a>	2014-10-03 10:22	-
<a href="#">comparative-genome-viewer/</a>	2022-12-11 02:12	-
<a href="#">dbgap/</a>	2022-09-13 10:11	-
<a href="#">diffexpIR-notebook/</a>	2017-09-21 12:38	-
<a href="#">entrez/</a>	2013-07-18 18:49	-
<a href="#">epigenomics/</a>	2022-02-07 22:48	-
<a href="#">eqtl/</a>	2017-09-19 15:34	-
<a href="#">fa2htgs/</a>	2006-08-04 17:02	-
<a href="#">genbank/</a>	2022-12-10 22:48	-
<a href="#">gene/</a>	2022-02-07 22:48	-
<a href="#">genomes/</a>	2022-12-10 22:48	-
<a href="#">geo/</a>	2022-12-11 05:17	-
<a href="#">giab/</a>	2022-12-10 22:48	-
<a href="#">hapmap/</a>	2011-09-20 10:18	-
<a href="#">hmm/</a>	2022-11-02 16:52	-
<a href="#">mmdb/</a>	2022-11-29 14:21	-
<a href="#">ncbi-asn1/</a>	2022-12-10 22:48	-
<a href="#">nist-immse/</a>	2019-08-29 11:59	-
<a href="#">osiris/</a>	2021-09-01 10:16	-
<a href="#">pathogen/</a>	2022-12-10 22:48	-
<a href="#">pub/</a>	2022-06-27 22:26	-
<a href="#">pubchem-scratch/</a>	2013-04-19 11:06	-
<a href="#">pubchem/</a>	2022-12-05 15:30	-
<a href="#">pubmed/</a>	2022-12-08 15:13	-
<a href="#">rapt/</a>	2022-02-07 22:48	-
<a href="#">refsam/</a>	2022-12-07 08:32	-
<a href="#">refseq/</a>	2022-11-14 10:57	-
<a href="#">repository/</a>	2022-02-07 22:48	-
<a href="#">seqc/</a>	2022-12-10 22:48	-
<a href="#">sequin/</a>	2021-01-26 17:04	-
<a href="#">sky-cgh/</a>	2016-06-23 15:20	-
<a href="#">snp/</a>	2022-12-10 22:48	-
<a href="#">sra/</a>	2022-12-10 22:48	-
<a href="#">tech-reports/</a>	2004-09-29 08:49	-
<a href="#">toolbox/</a>	2013-07-25 14:03	-
<a href="#">tpa/</a>	2021-09-19 15:26	-
<a href="#">variation/</a>	2016-08-09 10:36	-
<a href="#">10GB</a>	2019-09-18 16:38	10GB
<a href="#">1GB</a>	2019-09-18 16:38	1.0GB
<a href="#">50GB</a>	2019-09-18 16:39	50GB
<a href="#">5GB</a>	2019-09-18 16:38	5.0GB
<a href="#">README.ftp</a>	2022-06-28 15:16	2.3K
<a href="#">favicon.ico</a>	2019-08-29 12:00	3.2K
<a href="#">robots.txt</a>	2019-08-29 11:59	26

## 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.

← → ↻ 🔒 <https://ftp.ncbi.nlm.nih.gov/genomes/INFLUENZA/>

## Index of /genomes/INFLUENZA

Name	Last modified	Size
<a href="#">Parent Directory</a>		-
<a href="#">ANNOTATION/</a>	2013-01-23 10:30	-
<a href="#">processing/</a>	2020-10-14 04:02	-
<a href="#">updates/</a>	2020-10-13 10:34	-
<a href="#">README</a>	2016-07-08 11:27	3.5K
<a href="#">genomeset.dat</a>	2020-10-13 05:05	52M
<a href="#">genomeset.dat.gz</a>	2020-10-13 10:23	4.4M
<a href="#">influenza.cds</a>	2020-10-13 05:42	1.4G
<a href="#">influenza.cds.gz</a>	2020-10-13 10:29	101M
<a href="#">influenza.dat</a>	2020-10-13 05:16	39M
<a href="#">influenza.dat.gz</a>	2020-10-13 10:29	8.4M
<a href="#">influenza.faa</a>	2020-10-13 05:32	535M
<a href="#">influenza.faa.gz</a>	2020-10-13 10:29	34M
<a href="#">influenza.fna</a>	2020-10-13 05:23	1.3G
<a href="#">influenza.fna.gz</a>	2020-10-13 10:34	99M
<a href="#">influenza_aa.dat</a>	2020-10-13 10:13	103M
<a href="#">influenza_aa.dat.gz</a>	2020-10-13 10:23	12M
<a href="#">influenza_na.dat</a>	2020-10-13 07:33	75M
<a href="#">influenza_na.dat.gz</a>	2020-10-13 10:23	6.8M

[HHS Vulnerability Disclosure](#)

← → ↻ 🔒 <https://ftp.ncbi.nlm.nih.gov/genomes/INFLUENZA/processing/>

## Index of /genomes/INFLUENZA/processing

Name	Last modified	Size
<a href="#">Parent Directory</a>		-
<a href="#">genomeset.dat</a>	2020-10-14 04:18	14M

[HHS Vulnerability Disclosure](#)

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](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](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](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](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](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](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](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](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](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](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:

This is an NCBI Labs Experiment. [Learn more.](#)

**NIH** National Library of Medicine  
National Center for Biotechnology Information

**NCBI Virus**  
Sequences for discovery

About Us ▾ Find Data ▾ Help ▾ How to Participate ▾ Submit Sequences ▾ [Contact Us](#)

**Influenza Virus Data Hub** [Download ▾](#)

Select genus: ☒ Alphainfluenzavirus (A) ☒ Gammainfluenzavirus (C) ☒ Betainfluenzavirus (B) ☒ Deltainfluenzavirus (D)

Quick Links  
[Influenza Virus BLAST](#)  
[Influenza Virus Annotation Tool \(FLAN\)](#)  
[Influenza Virus Articles in PubMed](#)  
[Submit assembled sequences to GenBank](#)  
[Submit sequence reads to SRA](#)

**Tabular View** Selected Results: 0 [Align](#) [Build Phylogenetic Tree](#)

**New Submitters' Information available**  
We appreciate the effort from all involved in collecting samples and making sequence data publicly available. In order to facilitate citations and acknowledgements, we are adding information describing submitters, including affiliated organizations or institutions provided during submission. Although table columns may display abbreviated information, complete entries can be viewed by hovering over entries and are available by downloading the metadata table. We are continuing to refine the availability of this information and feedback is appreciated.

**Refine Results** [Reset](#)

Virus [+](#)

☒ Alphainfluenzavirus, taxid:197911 [×](#)

☒ Betainfluenzavirus, taxid:197912 [×](#)

☒ Gammainfluenzavirus, taxid:197913 [×](#)

**Expand Table**

	Nucleotide (537,523)	Protein (740,637)	RefSeq Genome (3)							
<input type="checkbox"/>	Assembly	Release Date	Species	Genus	Family	Sequence Type	Geo Location	USA	Host	
<input type="checkbox"/>	<a href="#">GCF_001343785.1</a>	2015-10-17	Influenza A virus	Alphainfluenzavirus	Orthomyxoviridae	RefSeq	USA: California state	CA	Homo sapiens	
<input type="checkbox"/>	<a href="#">GCF_000928555.1</a>	2015-02-22	Influenza A virus	Alphainfluenzavirus	Orthomyxoviridae	RefSeq	China		Homo sapiens	
<input type="checkbox"/>	<a href="#">GCF_000865085.1</a>	2015-02-13	Influenza A virus	Alphainfluenzavirus	Orthomyxoviridae	RefSeq	USA: Tompkins County, NY	NY	Homo sapiens	

[Select Columns](#)

[Feedback](#)

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

This is an NCBI Labs Experiment. [Learn more.](#)

**NIH** National Library of Medicine  
National Center for Biotechnology Information

**NCBI Virus**  
Sequences for discovery

About Us ▾ Find Data ▾ Help ▾ How to Participate ▾ Submit Sequences ▾ [Contact Us](#)

**Influenza Virus Data Hub** [Download ▾](#)

Select genus: ☒ Alphainfluenzavirus (A) ☒ Gammainfluenzavirus (C) ☒ Betainfluenzavirus (B) ☒ Deltainfluenzavirus (D)

Quick Links  
[Influenza Virus BLAST](#)  
[Influenza Virus Annotation Tool \(FLAN\)](#)  
[Influenza Virus Articles in PubMed](#)  
[Submit assembled sequences to GenBank](#)  
[Submit sequence reads to SRA](#)

**Tabular View** Selected Results: 3 [Align](#) [Build Phylogenetic Tree](#)

**New Submitters' Information available**  
We appreciate the effort from all involved in collecting samples and making sequence data publicly available. In order to facilitate citations and acknowledgements, we are adding information describing submitters, including affiliated organizations or institutions provided during submission. Although table columns may display abbreviated information, complete entries can be viewed by hovering over entries and are available by downloading the metadata table. We are continuing to refine the availability of this information and feedback is appreciated.

**Refine Results** [Reset](#)

Virus [+](#)

☒ Alphainfluenzavirus, taxid:197911 [×](#)

☒ Betainfluenzavirus, taxid:197912 [×](#)

☒ Gammainfluenzavirus, taxid:197913 [×](#)

**Expand Table**

	Nucleotide (537,523)	Protein (740,637)	RefSeq Genome (3)							
<input checked="" type="checkbox"/>	Assembly	Release Date	Species	Genus	Family	Sequence Type	Geo Location	USA	Host	
<input checked="" type="checkbox"/>	<a href="#">GCF_001343785.1</a>	2015-10-17	Influenza A virus	Alphainfluenzavirus	Orthomyxoviridae	RefSeq	USA: California state	CA	Homo sapiens	
<input checked="" type="checkbox"/>	<a href="#">GCF_000928555.1</a>	2015-02-22	Influenza A virus	Alphainfluenzavirus	Orthomyxoviridae	RefSeq	China		Homo sapiens	
<input checked="" type="checkbox"/>	<a href="#">GCF_000865085.1</a>	2015-02-13	Influenza A virus	Alphainfluenzavirus	Orthomyxoviridae	RefSeq	USA: Tompkins County, NY	NY	Homo sapiens	

[Select Columns](#)

[Feedback](#)

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

That comes to this view

This is an NCBI Labs Experiment. Learn more.

NIH National Library of Medicine  
National Center for Biotechnology Information

belle172@umn.edu

NCBI Virus  
Sequences for discovery

Influenza Virus Data Hub

Tabular View

New Submitters' Information available  
We appreciate the effort from all involved in collecting samples and making sequence data publicly available. In order to facilitate citations and acknowledgements, we are adding information describing submitters, including affiliated organizations or institutions provided during submission. Feedback is appreciated.

Refine Results

Virus

Alphainfluenzavirus, taxid:197911

Betafluenzavirus, taxid:197912

Gammairfluenzavirus,

Download Results

Step 1 of 3: Select Data Type

Sequence data (FASTA Format)

☒ Nucleotide

☐ Coding Region

☐ Protein

Accession List

☐ Nucleotide

☐ Protein

☐ Assembly

Current table view result

☐ CSV format

☐ XML format

Next

Assembly	Release Date	Species	Genus	Family	Sequence Type	Geo Location	USA	Host	
<input checked="" type="checkbox"/>	GCF_001343785.1	2015-10-17	Influenza A virus	Alphainfluenzavirus	Orthomyxoviridae	RefSeq	USA: California state	CA	Homo sapiens
<input checked="" type="checkbox"/>	GCF_000928555.1	2015-02-22	Influenza A virus	Alphainfluenzavirus	Orthomyxoviridae	RefSeq	China		Homo sapiens
<input checked="" type="checkbox"/>	GCF_000865085.1	2015-02-13	Influenza A virus	Alphainfluenzavirus	Orthomyxoviridae	RefSeq	USA: Tompkins County, NY	NY	Homo sapiens

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

NCBI Virus  
Sequences for discovery

Virus Data Hub

Download Results

Step 2 of 3: Select Records

☒ Download Selected Records

☐ Download All Records

Back

Next

Nucleotide (537,523)

Protein (740,637)

RefSeq Genome (3)

Assembly	Release Date	Species	Genus	Family	Sequence Type	Geo Location	
<input checked="" type="checkbox"/>	GCF_001343785.1	2015-10-17	Influenza A virus	Alphainfluenzavirus	Orthomyxoviridae	RefSeq	USA: California state
<input checked="" type="checkbox"/>	GCF_000928555.1	2015-02-22	Influenza A virus	Alphainfluenzavirus	Orthomyxoviridae	RefSeq	China
<input checked="" type="checkbox"/>	GCF_000865085.1	2015-02-13	Influenza A virus	Alphainfluenzavirus	Orthomyxoviridae	RefSeq	USA: Tompkins County, NY

Then with Download Selected Records click Next,

NIH National Library of Medicine  
National Center for Biotechnology Information

belle172@

NCBI Virus  
Sequences for discovery

a Virus Data Hub

Download Results

Step 3 of 3: Select FASTA definition line

☒ Use default: Accession GenBank Title

☐ Build custom

Back

Download

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](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](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](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](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:

# Clustal Omega

[Input form](#)[Web services](#)[Help & Documentation](#)[Bioinformatics Tools FAQ](#)[Feedback](#)

Tools > Multiple Sequence Alignment > Clustal Omega

## Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

**Important note:** This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

Or, [upload a file](#): 

Choose File

 No file chosen

[Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

The default settings will fulfill the needs of most users.

More options...

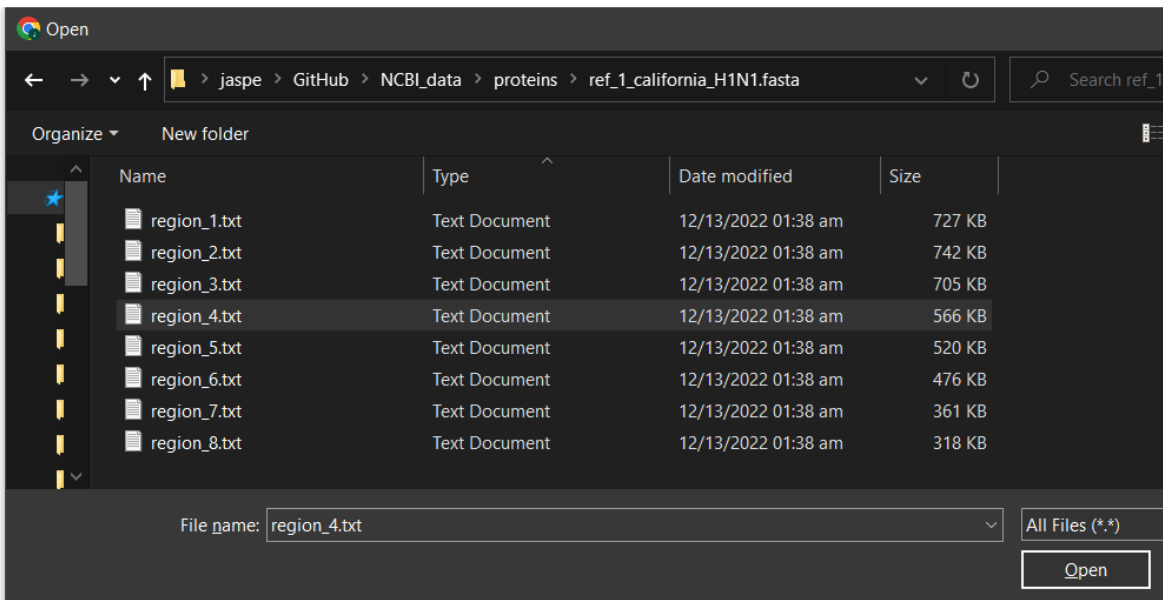
(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

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)

sequences in any supported format:

Or, upload a file:  region\_4.txt [Use a example sequence](#) | [Clear](#)

**STEP 2 - Set your parameters**

**OUTPUT FORMAT**

ClustalW with character counts

*The default settings will fulfill the needs of most users.*

(Click here, if you want to view or change the default settings.)

**STEP 3 - Submit your job**

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

Click the Submit button, and wait.

After a few seconds you should get this page:

# Clustal Omega

[Input form](#) | [Web services](#) | [Help & Documentation](#) | [Bioinformatics Tools FAQ](#)

[Tools](#) > [Multiple Sequence Alignment](#) > [Clustal Omega](#)

## Your job is currently running... please be patient

The result of your job will appear in this browser window.

Job ID: [clustalo-l20221213-175047-0776-76835709-p1m](#)

### Please note the following

- You may press Shift+Refresh or Reload on your browser at any time to check if results are ready.
- You may bookmark this page to view your results later if you wish.
- Results are stored for 7 days.

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-l20221213-175047-0776-76835709-p1m>

The results page:

# Clustal Omega

[Input form](#) | [Web services](#) | [Help & Documentation](#) | [Bioinformatics Tools FAQ](#)

Tools > Multiple Sequence Alignment > Clustal Omega

Results for job clustalo-l20221213-175047-0776-76835709-p1m

[Alignments](#) | [Result Summary](#) | [Guide Tree](#) | [Phylogenetic Tree](#) | [Results Viewers](#) | [Submission Details](#)

[Download Alignment File](#)

CLUSTAL O(1.2.4) multiple sequence alignment

gi 937169886 gb KT853557 Influenza	-----ATGAAGGCAATAATTGTA---	18
gi 937170670 gb KT853786 Influenza	-----ATGAAGGCAATAATTGTA---	18
gi 937175697 gb KT865972 Influenza	-----ATGAAGGCAATAATTGTA---	18
gi 937178258 gb KT866751 Influenza	-----ATGAAGGCAATAATTGTA---	18
gi 937178917 gb KT866953 Influenza	-----ATGAAGGCAATAATTGTA---	18
gi 937176001 gb KT866063 Influenza	-----ATGAAGGCAATAATTGTA---	18
gi 937173089 gb KT854498 Influenza	-----ATGAAGGCAATAATTGTA---	18
gi 937172776 gb KT854406 Influenza	-----ATGAAGGCAATAATTGTA---	18
gi 984696142 gb KU592715 Influenza	-----ATGAAGGCAATAATTGTA---	18

Click on the second tab, Results Summary.

# Clustal Omega

[Input form](#) | [Web services](#) | [Help & Documentation](#) | [Bioinformatics Tools FAQ](#)

Tools > Multiple Sequence Alignment > Clustal Omega

Results for job clustalo-l20221213-175047-0776-76835709-p1m

[Alignments](#) | [Result Summary](#) | [Guide Tree](#) | [Phylogenetic Tree](#) | [Results Viewers](#) | [Submission Details](#)

Input Sequences

[clustalo-l20221213-175047-0776-76835709-p1m.input](#)

Tool Output

[clustalo-l20221213-175047-0776-76835709-p1m.output](#)

Alignment in CLUSTAL format with base/residue numbering

[clustalo-l20221213-175047-0776-76835709-p1m.clustal\\_num](#)

Guide Tree

[clustalo-l20221213-175047-0776-76835709-p1m.dnd](#)

Phylogenetic Tree

[clustalo-l20221213-175047-0776-76835709-p1m.ph](#)

Percent Identity Matrix

[clustalo-l20221213-175047-0776-76835709-p1m.pim](#)

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](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](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](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](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](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](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](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](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](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.