Course Manual - Phylogenetic Analysis

Scenario 3 – Retrieval of sequences and phylogenetic analysis

There has been a recent spike in Influenza cases in your local hospital. The clinicians are worried there may be an outbreak and ward to ward transmission. They have asked you to sequence the virus from clinical samples to determine if they are related.

For this practical we will use the HA sequence of the viruses to identify whether an outbreak has occurred at the hospital. The file used is Hospital_HA.fas

Note We have NA sequences also available if you would like to repeat the practical later on your own. The naming convention used is similar to the HA example. All files can be found in the **Example_Data** sub-folder.

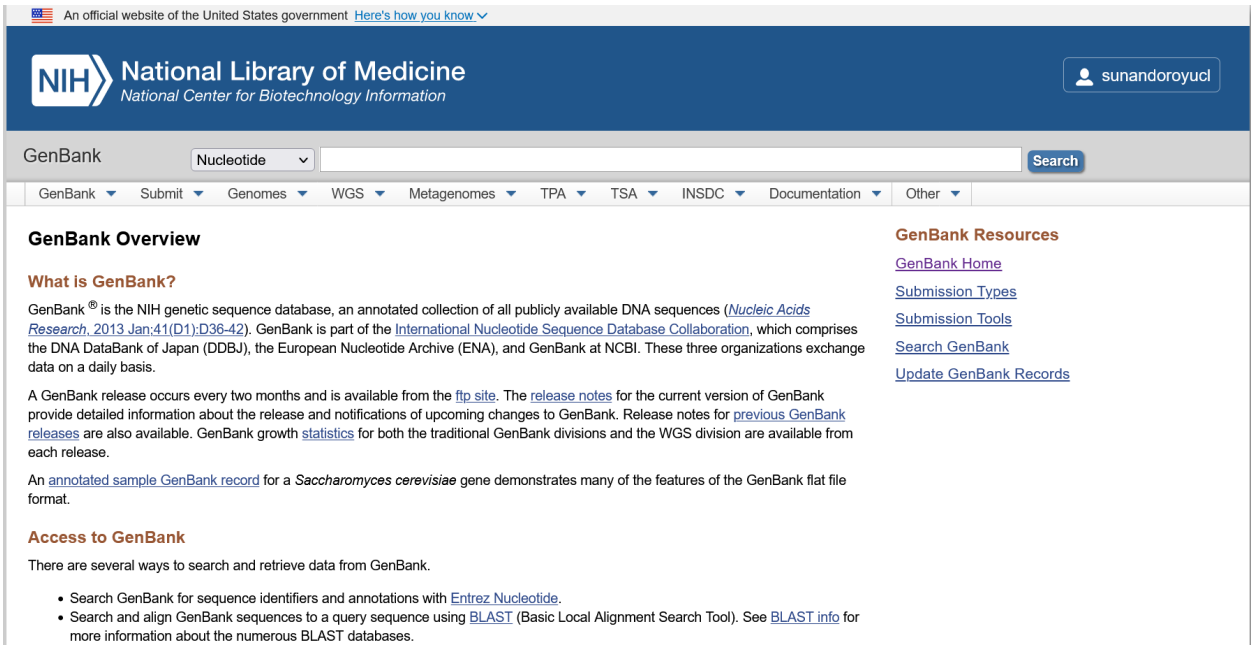
Software used in this session

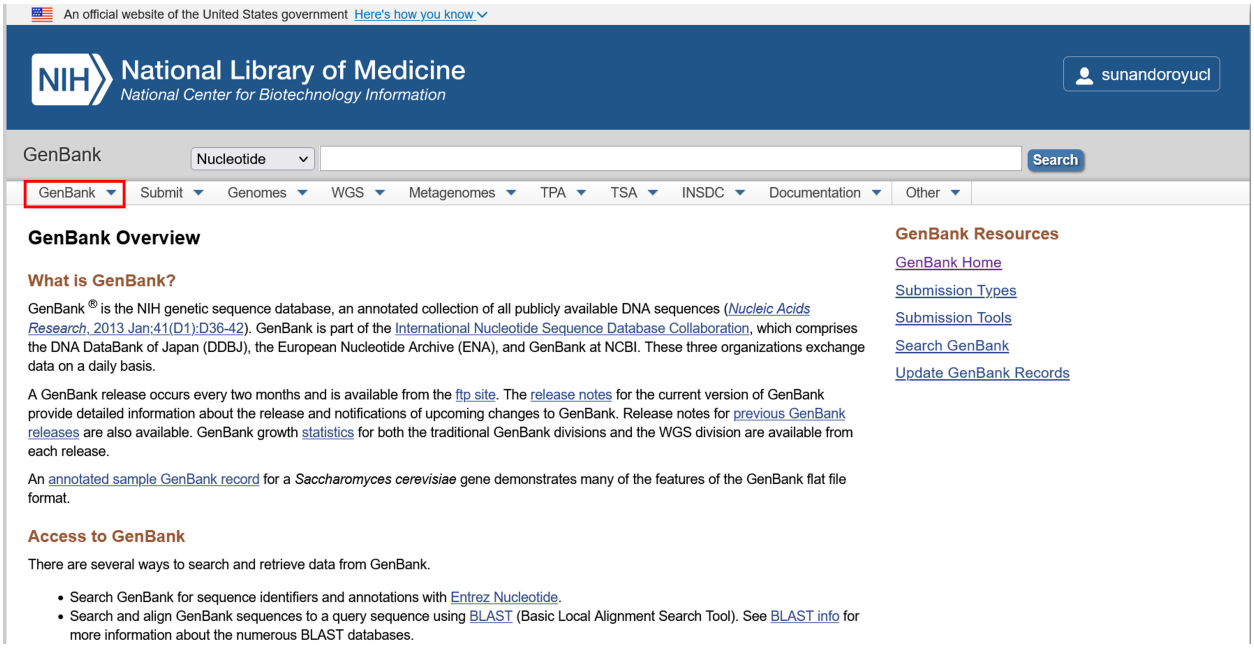
1. **Mafft** - Alignment Tool
2. **MEGA** - Alignment Viewer and Editor
3. **Modeltest-ng** - Model Testing
4. **IQ-TREE** - Tree Building
5. **Figtree** - Tree Viewer and Editor

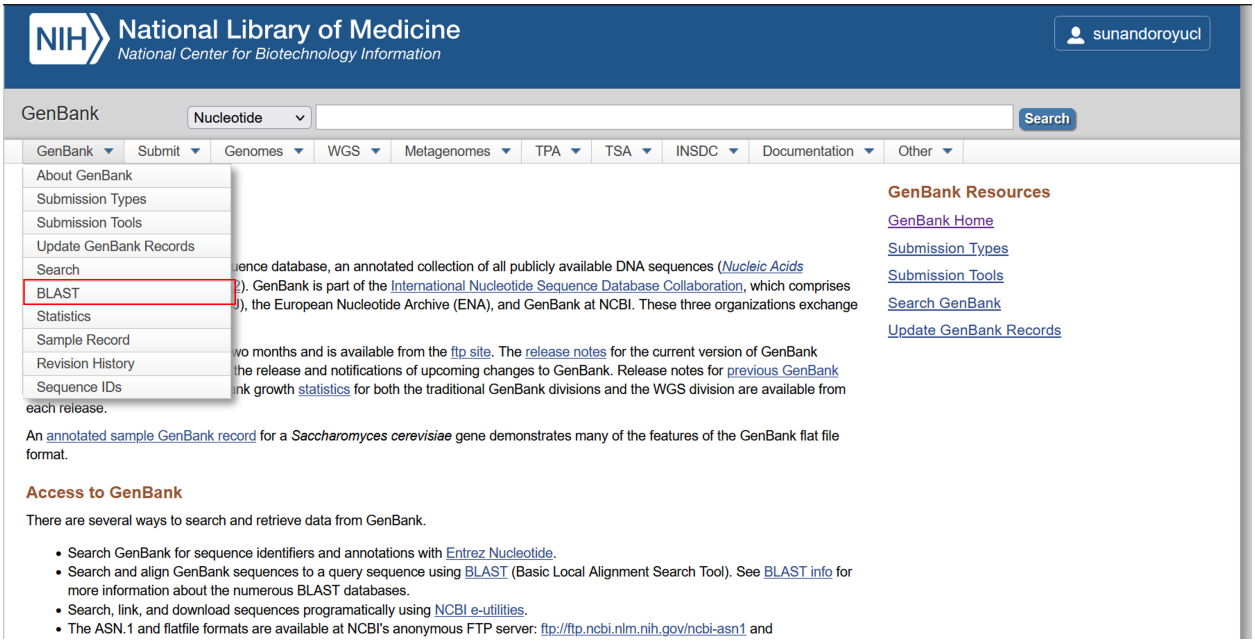
Step 1 - Downloading related sequences.

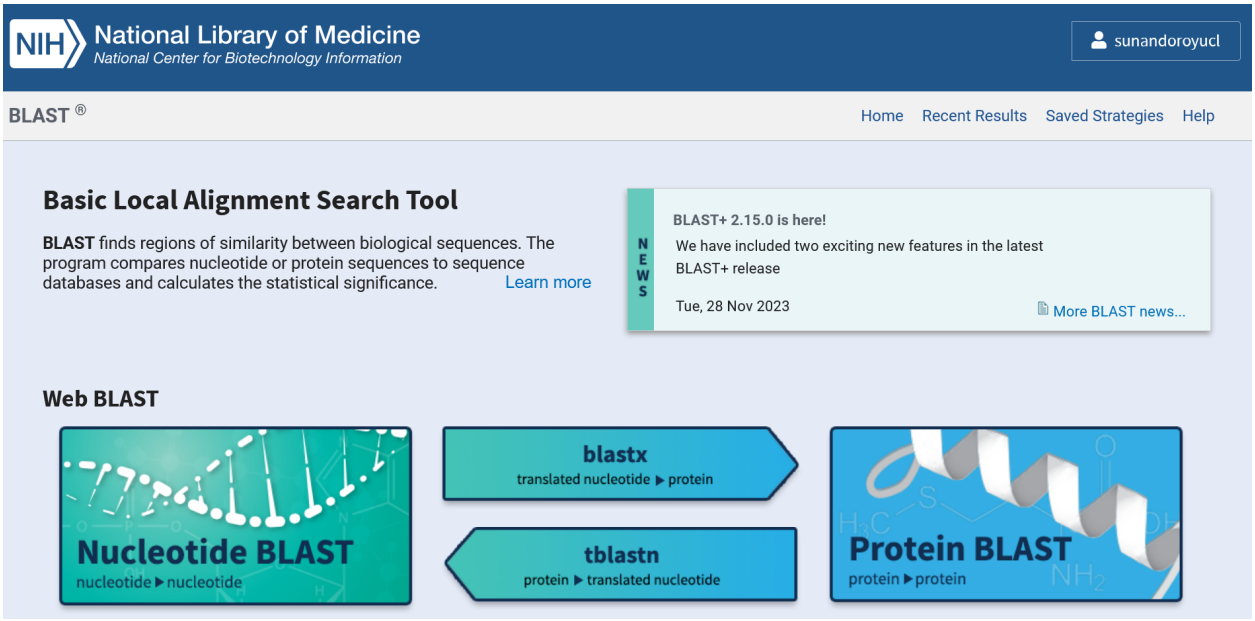
```
cd ~/Phylogenetics
ls
cp Example_Data/Hospital_HA.fas .
ls
```

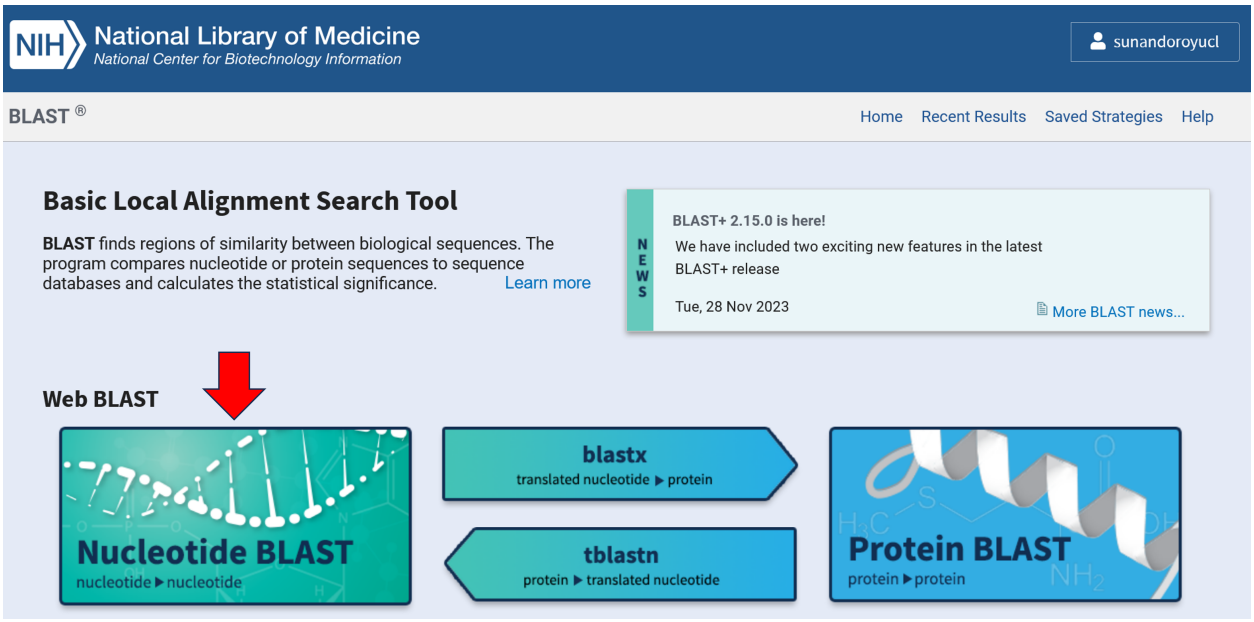
Website - Genbank https://www.ncbi.nlm.nih.gov/genbank/

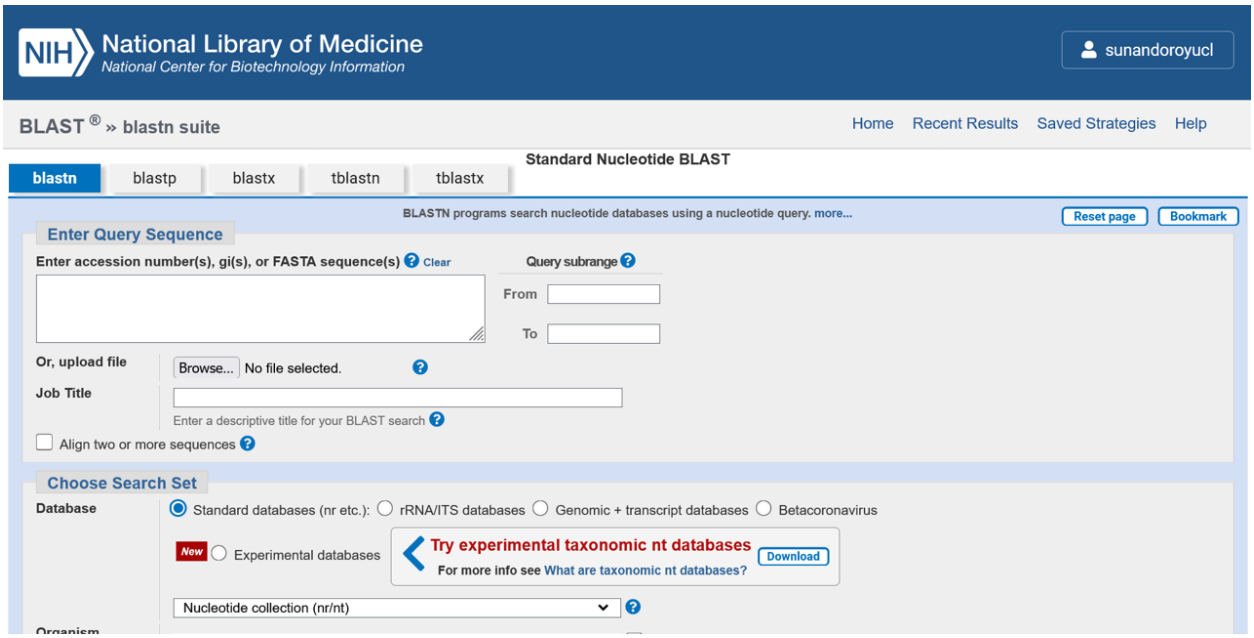

An official website of the United States government Here's how you know
National Library of Medicine
National Center for Biotechnology Information
sunandoroyucl
GenBank
Nucleotide
Search
GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Documentation Other
GenBank Overview
What is GenBank?
GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.
A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.
An annotated sample GenBank record for a Saccharomyces cerevisiae gene demonstrates many of the features of the GenBank flat file format.
Access to GenBank
There are several ways to search and retrieve data from GenBank.
Search GenBank for sequence identifiers and annotations with Entrez Nucleotide.
Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). See BLAST info for more information about the numerous BLAST databases.
GenBank Resources
GenBank Home
Submission Types
Submission Tools
Search GenBank
Update GenBank Records

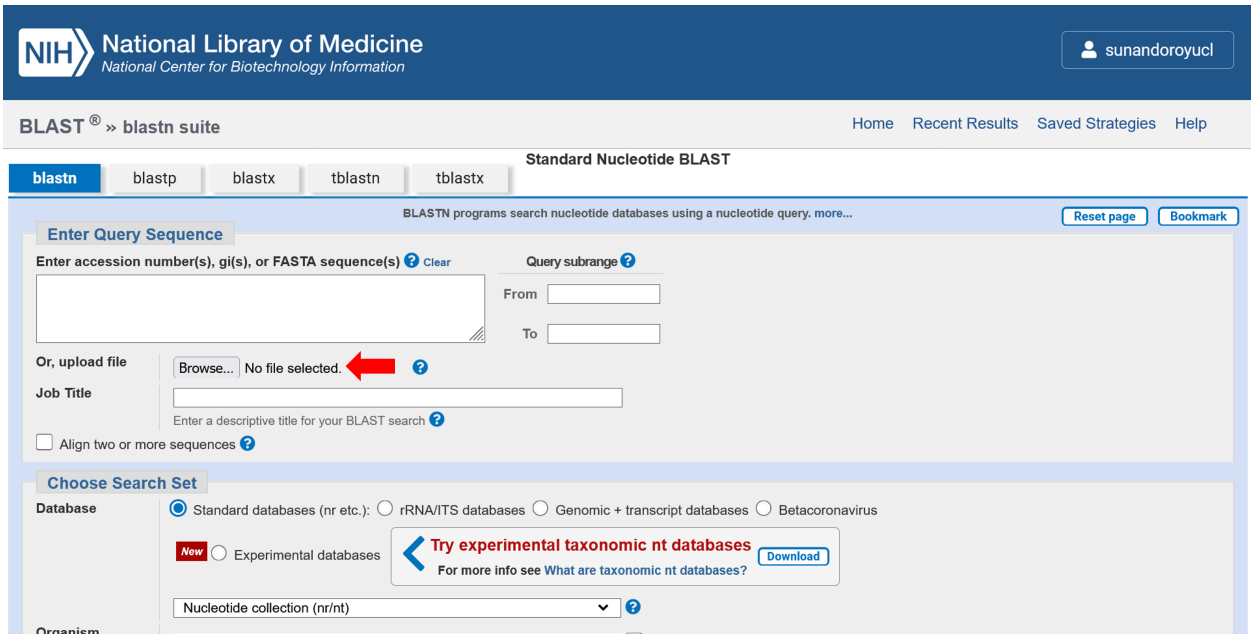

An official website of the United States government Here's how you know
National Library of Medicine
National Center for Biotechnology Information
sunandoroyucl
GenBank
Nucleotide
Search
GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Documentation Other
GenBank Overview
What is GenBank?
GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.
A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.
An annotated sample GenBank record for a Saccharomyces cerevisiae gene demonstrates many of the features of the GenBank flat file format.
Access to GenBank
There are several ways to search and retrieve data from GenBank.
Search GenBank for sequence identifiers and annotations with Entrez Nucleotide.
Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). See BLAST info for more information about the numerous BLAST databases.
GenBank Resources
GenBank Home
Submission Types
Submission Tools
Search GenBank
Update GenBank Records


National Library of Medicine
National Center for Biotechnology Information
sunandoroyucl
GenBank
Nucleotide
Search
GenBank
Submit
Genomes
WGS
Metagenomes
TPA
TSA
INSDC
Documentation
Other
About GenBank
Submission Types
Submission Tools
Update GenBank Records
Search
BLAST
Statistics
Sample Record
Revision History
Sequence IDs
GenBank Resources
GenBank Home
Submission Types
Submission Tools
Search GenBank
Update GenBank Records
Access to GenBank
There are several ways to search and retrieve data from GenBank.
Search GenBank for sequence identifiers and annotations with Entrez Nucleotide.
Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). See BLAST info for more information about the numerous BLAST databases.
Search, link, and download sequences programatically using NCBI e-utilities.
The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1 and

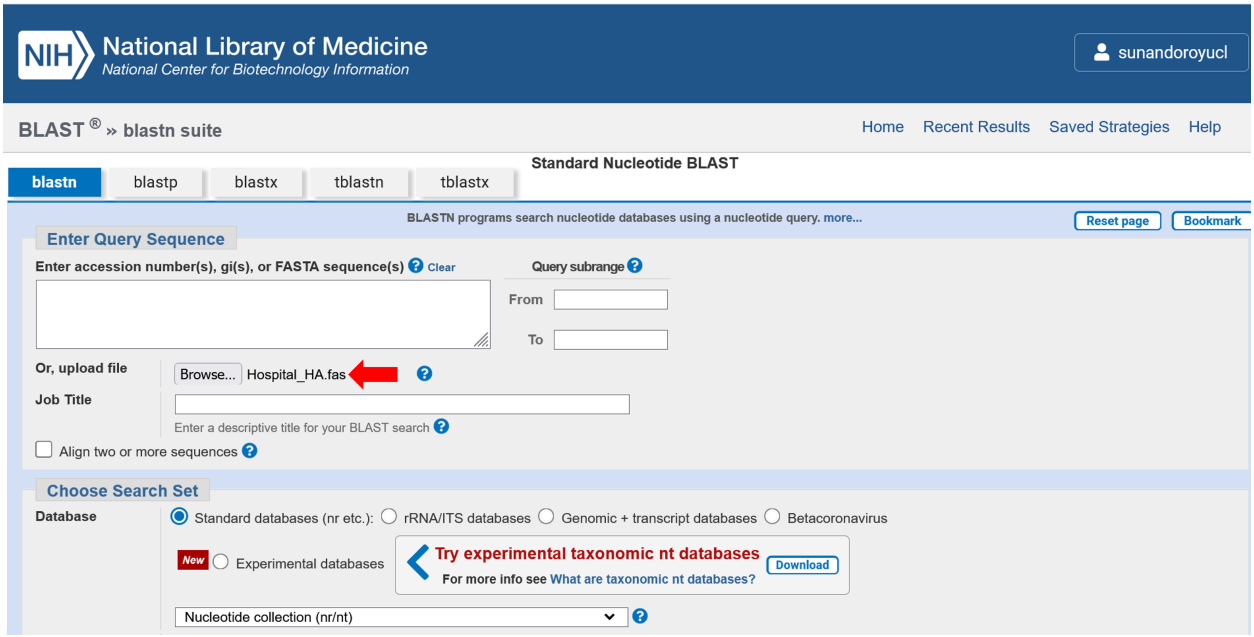

National Library of Medicine
National Center for Biotechnology Information
sunandoroyucl
BLAST ®
Home
Recent Results
Saved Strategies
Help
Basic Local Alignment Search Tool
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.
Learn more
NEWS
BLAST+ 2.15.0 is here!
We have included two exciting new features in the latest BLAST+ release
Tue, 28 Nov 2023
More BLAST news...
Web BLAST
Nucleotide BLAST
nucleotide ▶ nucleotide
blastx
translated nucleotide ▶ protein
tblastn
protein ▶ translated nucleotide
Protein BLAST
protein ▶ protein

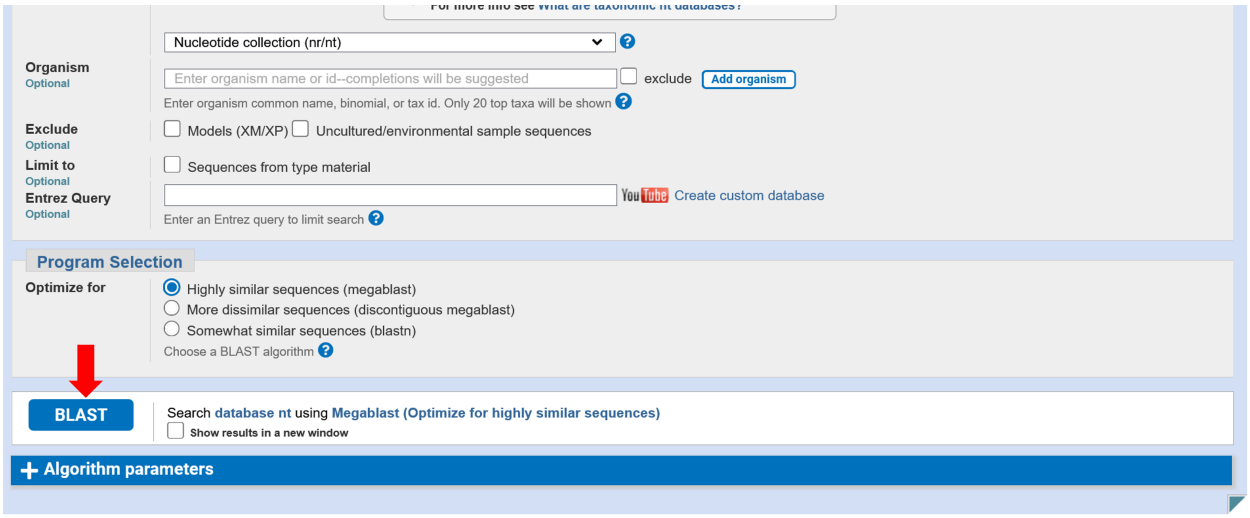

National Library of Medicine
National Center for Biotechnology Information
sunandoroyucl
BLAST ®
Home
Recent Results
Saved Strategies
Help
Basic Local Alignment Search Tool
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.
Learn more
NEWS
BLAST+ 2.15.0 is here!
We have included two exciting new features in the latest BLAST+ release
Tue, 28 Nov 2023
More BLAST news...
Web BLAST
Nucleotide BLAST
nucleotide ▶ nucleotide
blastx
translated nucleotide ▶ protein
tblastn
protein ▶ translated nucleotide
Protein BLAST
protein ▶ protein

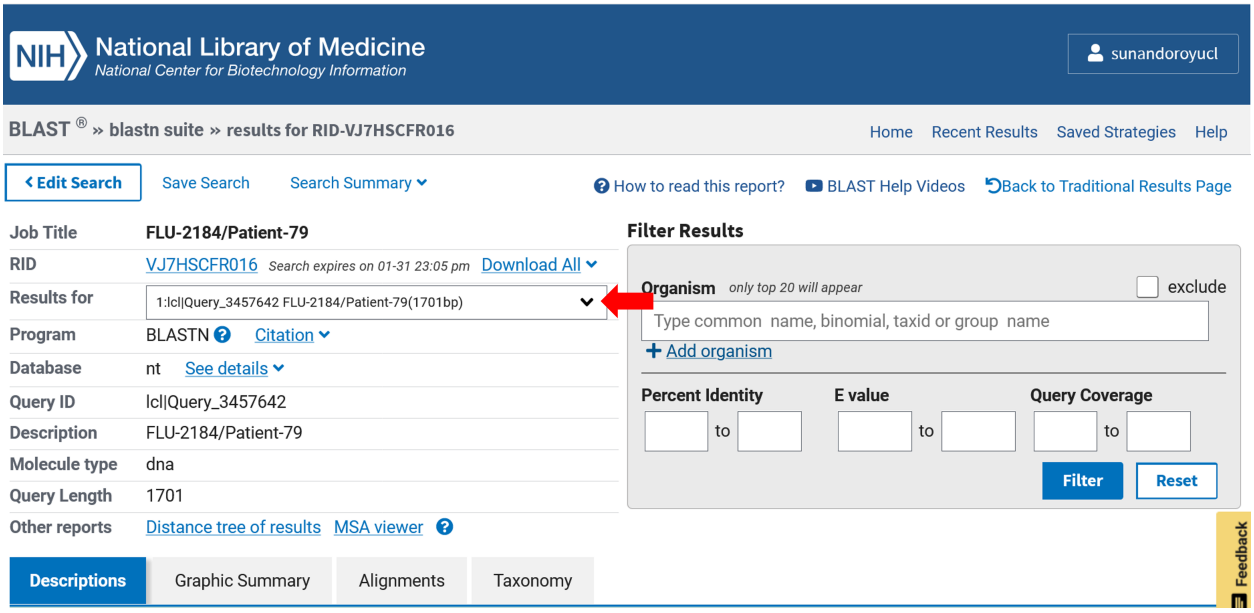

National Library of Medicine
National Center for Biotechnology Information
sunandoroyucl
BLAST ® » blastn suite
Home
Recent Results
Saved Strategies
Help
Standard Nucleotide BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTN programs search nucleotide databases using a nucleotide query. more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Browse...
No file selected.
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Standard databases (nr etc.):
rRNA/ITS databases
Genomic + transcript databases
Betacoronavirus
New
Experimental databases
Try experimental taxonomic nt databases
Download
For more info see What are taxonomic nt databases?
Nucleotide collection (nr/nt)
Organism



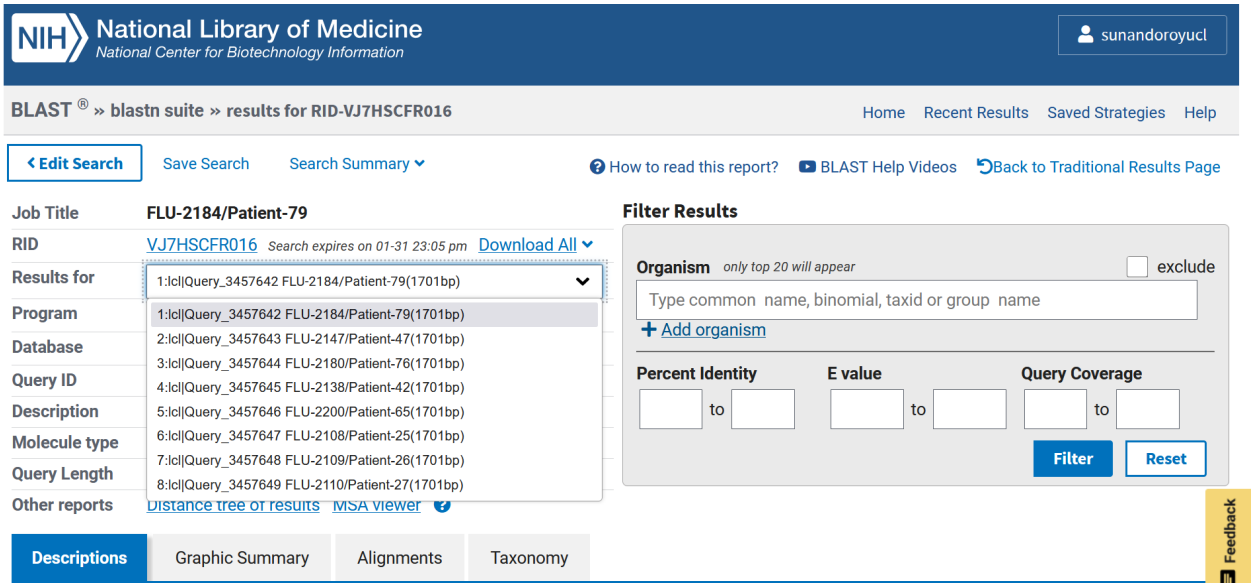
Navigate to the phylogenetics folder and upload the **Hospital_HA.fas** file.



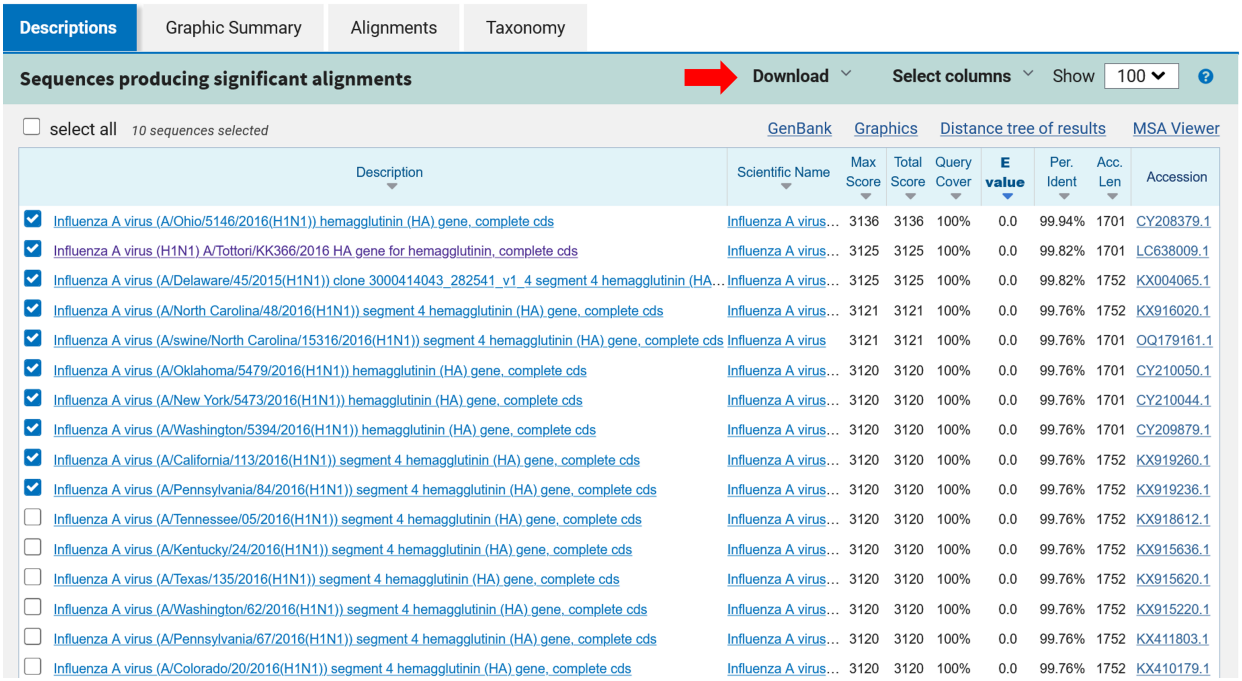


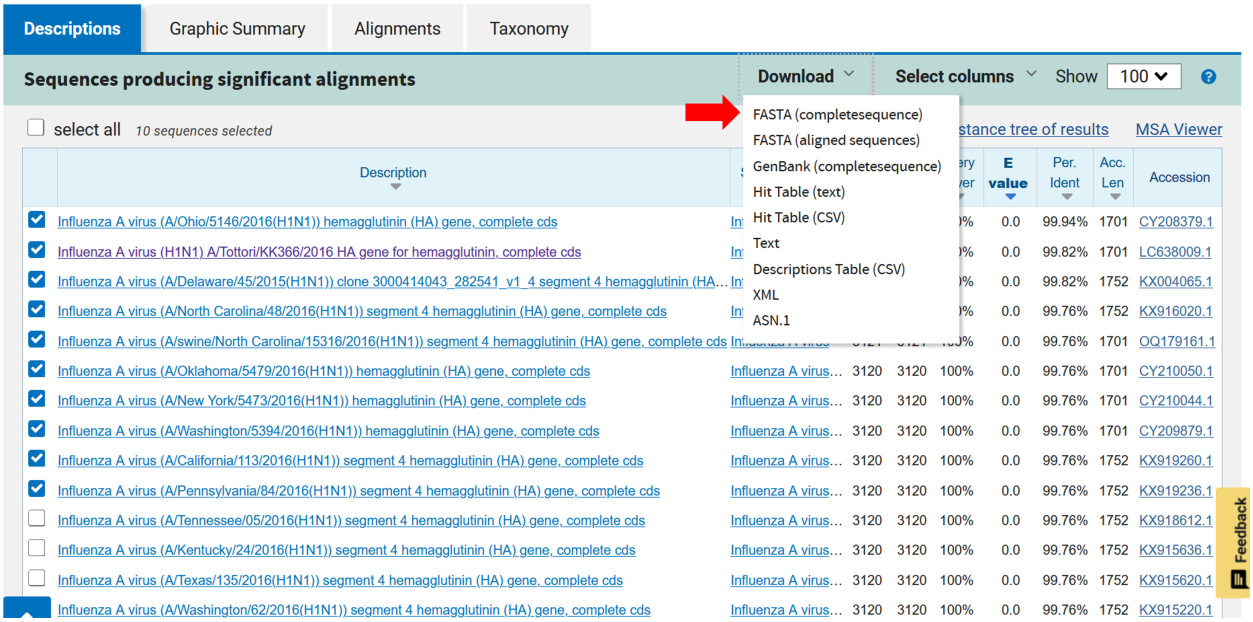


You can browse the hits for each reference sequence by clicking on the bar and selecting the strain of interest in the dropdown menu.



Scroll down to select the top hits for each reference sequence.





For each sequence run using Blast we can download corresponding 10 closest hit. We have already downloaded this for every sequence and saved in the **Example_Data** folder with the sequence number as the file name. You can open these files in any text editor (GUI) or using nano from the terminal.

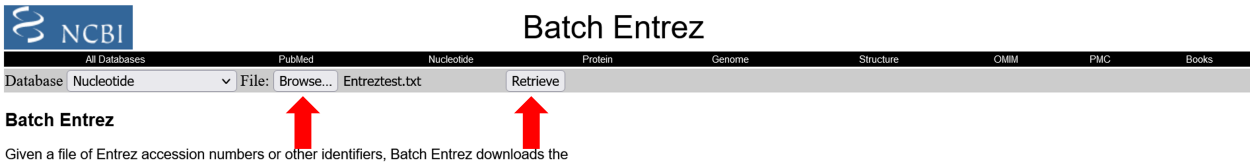
We have the file header cleaned up and available along with some other sequences from GISAID in the file named **Combined_HA_genbank.fas**

Alternate Ways to Download from GenBank

Batch Entrez (https://www.ncbi.nlm.nih.gov/sites/batchentrez)

To test batch downloads we will create a file with the 3 Accession numbers KX004485,MT540610,KU821082

Use your skills learnt in the course so far to create a file in the ~/**Phylogenetics**/ folder called Entreztest.txt with the 3 Accession numbers above and use batchentrez to download the sequences. Each Accession number should be in a new line in the file.



Batch Entrez

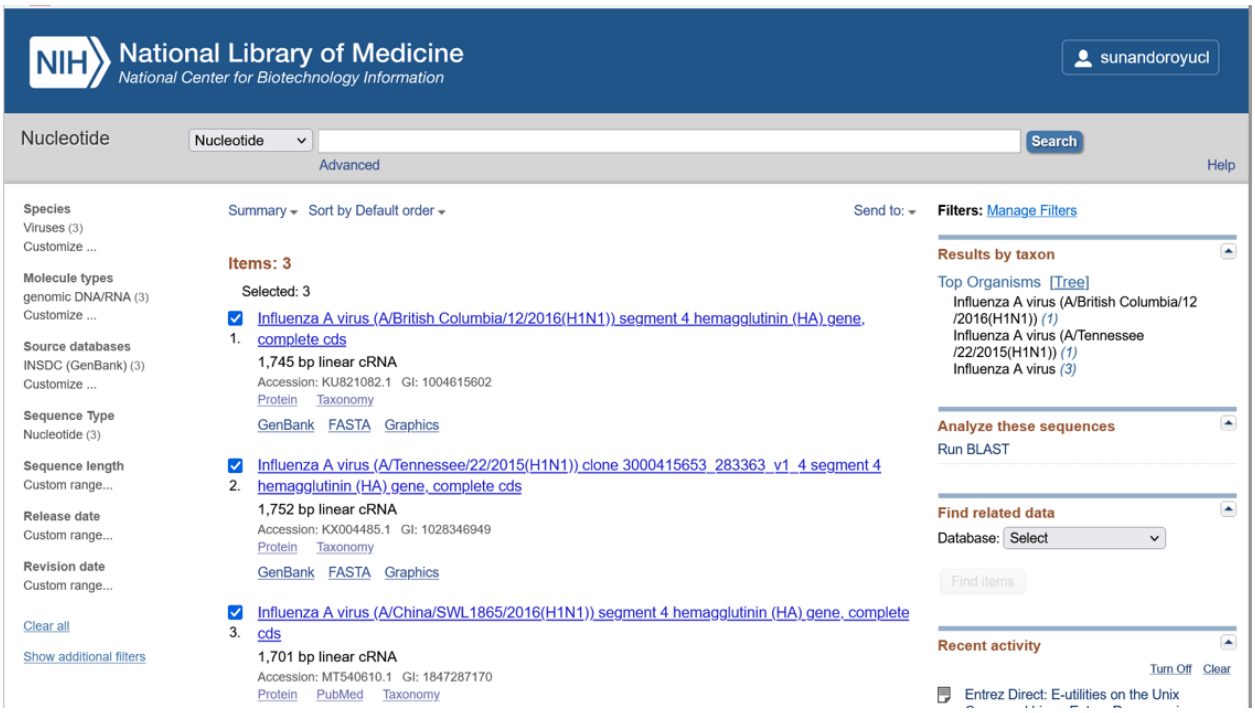
Given a file of Entrez accession numbers or other identifiers, Batch Entrez downloads the corresponding records.

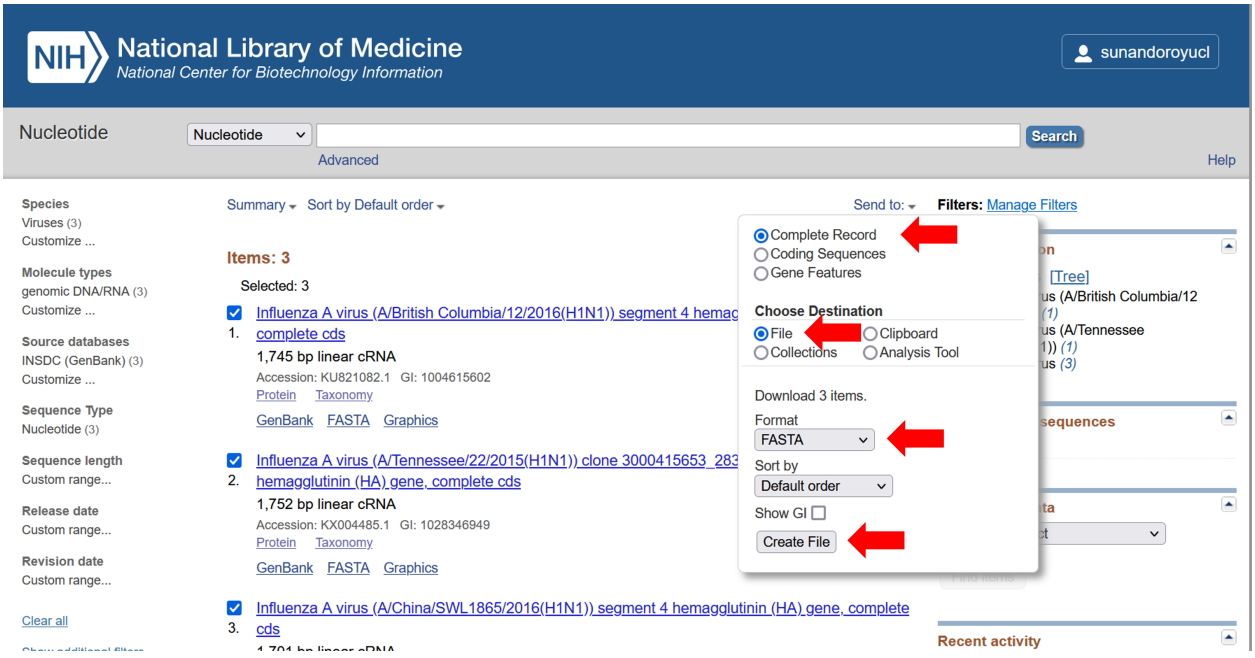
Instructions

1. Start with a local file containing a list of accession numbers or identifiers
2. Select the database corresponding to the type of accession numbers or identifiers in your input file
3. Use the **Browse** or **Choose File...** button to select the input file
4. Press the **Retrieve** button to see a list of document summaries
5. Select a format in which to display the data for viewing, and/or saving
6. Select 'Send to file' to save the file.

Tips

- To download entire genome records, check the NCBI FTP site, instead of using Batch Entrez.
- Some lists of record identifiers can be tens of thousands of lines long, so Batch Entrez may not retrieve all records from one list. Split the list of identifiers into smaller files using a file splitting software or a file split command at the command prompt in UNIX or LINUX systems.
- When loading large numbers of genome records, put several thousand record identifiers per file, one per line, left-adjusted.





Entrez E-utilities ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/

Manual - https://www.ncbi.nlm.nih.gov/books/NBK179288/

QuickStart - http://bioinformatics.cvr.ac.uk/blog/ncbi-entrez-direct-unix-e-utilities/

Browser

```
#To Download 4 SARS-CoV2 genomes as an example
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&id=AY278488, AY304486, MN908947
```

Terminal

```
#To download all complete L protein of taxon rhabdoviridae from Refseq
cd ~/Phylogenetics

esearch -db "protein" -query "txid11270[Organism] AND L Protein Complete AND refseq[filter]" | efetch -

grep '>' esearchoutput.fasta
```

Step 2 – Aligning Sequences.

Software Used

Mafft (https://mafft.cbrc.jp/alignment/software/)

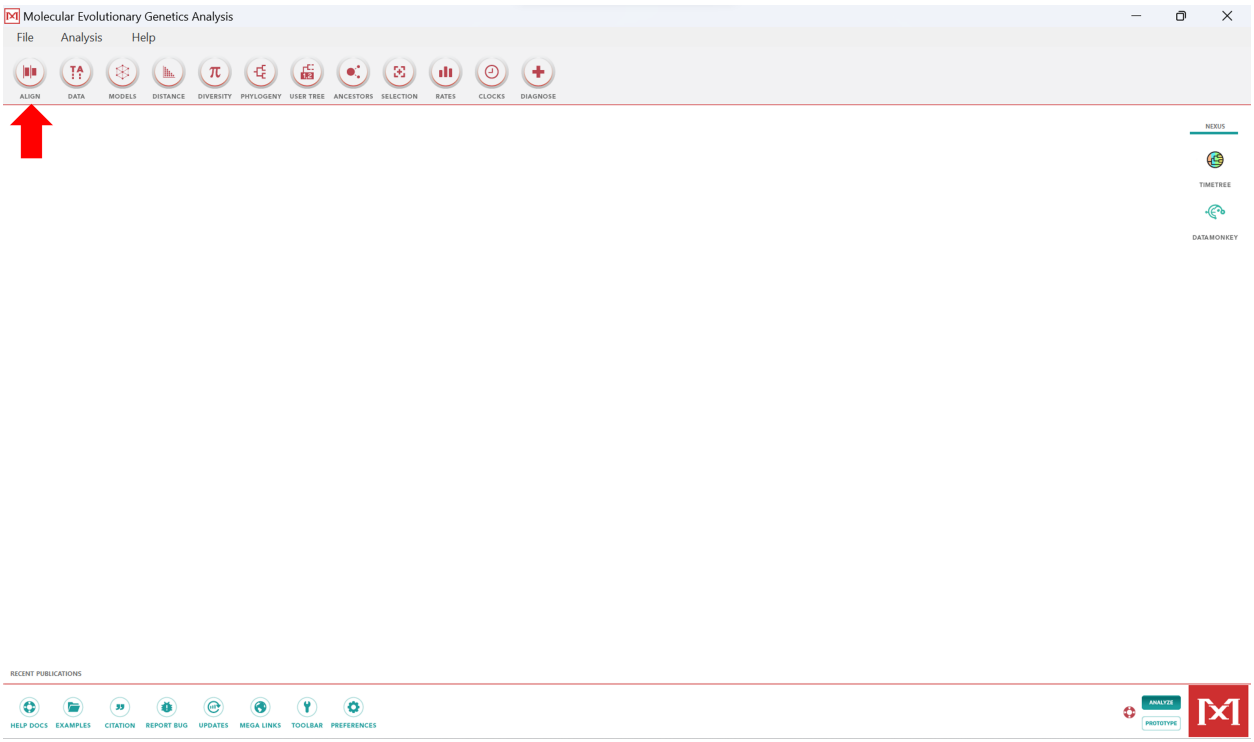
MEGA (https://www.megasoftware.net/)

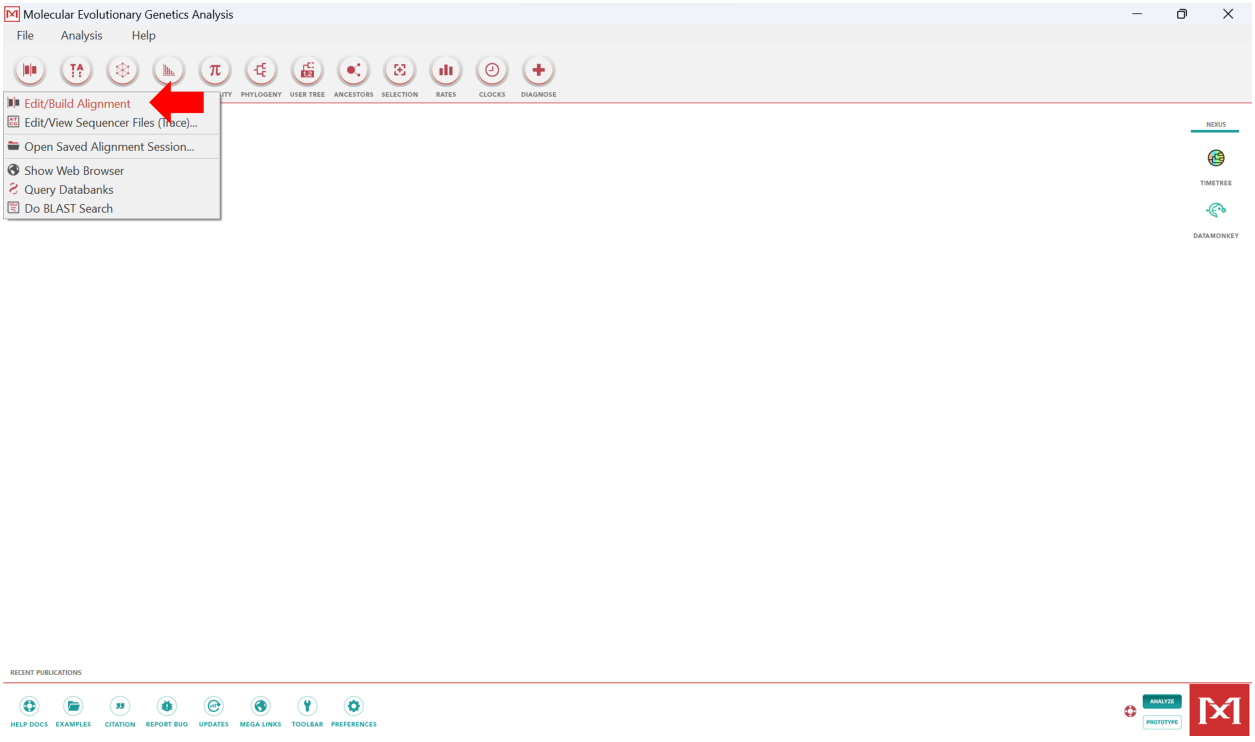
Alternate Software – MUSCLE, CLUSTALW

To view/edit sequence files we will use MEGA. MEGA has a GUI and will launch as a standalone program. Before we start we will first copy over the dataset that includes our hospital sequences and all the reference strains we have downloaded. The file name is **Combined_HA_genbank.fas**

Make sure you are in the Phylogenetics folder for this part of the practical

```
cd ~/Phylogenetics

cp Example_Data/Combined_HA_genbank.fas .

#To Launch MEGA
mega
```


Molecular Evolutionary Genetics Analysis
File
Analysis
Help
ALIGN
DATA
MODELS
DISTANCE
DIVERSITY
PHYLOGENY
USER TREE
ANCESTORS
SELECTION
RATES
CLOCKS
DIAGNOSE
NEXUS
TIMETREE
DATAMONKEY
RECENT PUBLICATIONS
HELP DOCS
EXAMPLES
CITATION
REPORT BUG
UPDATES
MEGA LINKS
TOOLBAR
PREFERENCES
ANALYZE
PROTOTYPE


Molecular Evolutionary Genetics Analysis
File
Analysis
Help
PHYLOGENY
USER TREE
ANCESTORS
SELECTION
RATES
CLOCKS
DIAGNOSE
Edit/Build Alignment
Edit/View Sequencer Files (Trace)...
Open Saved Alignment Session...
Show Web Browser
Query Databanks
Do BLAST Search
NEXUS
TIMETREE
DATAMONKEY
RECENT PUBLICATIONS
HELP DOCS
EXAMPLES
CITATION
REPORT BUG
UPDATES
MEGA LINKS
TOOLBAR
PREFERENCES
ANALYZE
PROTOTYPE

We are now going to use mafft to align the file. You can use any file name you prefer in place of outputfile.fas

below.

```
cd ~/Phylogenetics

mafft Combined_HA_genbank.fas > outputfile.fas
```

This uses the default models to align the sequences. For highly divergent sequences this may produce inaccurate alignments.

Alternatives are if you have a curated alignment mafft –add works to add new sequences to existing alignments which puts more weightage on the existing alignment.

We can also use the L-INS-i algorithm in Mafft that aligns more divergent sequences using pairwise local alignments.

Once the alignment is complete you can open and view the aligned file in MEGA by repeating the steps above. You can continue to build a tree using the file you just created. Just remember to use the file name of your output in the commands below. We will be using the file **Combined_HA_genbank_aln.fas**

```
cd ~/Phylogenetics

cp Example_Data/Combined_HA_genbank_aln.fas .

ls
```

Step 3 – Constructing a Phylogeny.

Software Used

Modeltest-ng (https://github.com/ddarriba/modeltest)

IQ-TREE (http://www.iqtree.org/)

Alternate Software – PhyML,RAxML

For this session we will start with the aligned nucleotide sequences created in the last step. You can use the alignment you have created or the file already prepared for you **Combined_HA_genbank_aln.fas**

Model Testing
To run model testing we will use Modeltest-ng. You can use any filename in place of **modeloutputfile** below.

```
cd ~/Phylogenetics

modeltest-ng -d nt -i Combined_HA_genbank_aln.fas -o modeloutputfile -t ml -p 2
```

Where:
-d : datatype
-i : Input file
-o : Output file
-t : Sets the starting tree topology
-p : Number of threads

We test three criteria to select the best fitting models BIC, AIC and AICc. The modeltest results for this alignment are in ~/Phylogenetics/Example_Data/HA_model.out

```
less ~/Phylogenetics/Example_Data/HA_model.out
```

```
AICc      model             K           lnL          score          delta    weight
------------------------------------------------------------------------------------
      1   HKY+G4            5     -3388.1832      6913.3664         0.0000    0.1377
      2   HKY+I             5     -3388.4961      6913.9922         0.6258    0.1007
      3   TrN+G4            6     -3387.7826      6914.5652         1.1988    0.0756
      4   TPM2uf+G4         6     -3388.0754      6915.1508         1.7844    0.0564
      5   TrN+I             6     -3388.0972      6915.1945         1.8281    0.0552
      6   HKY+I+G4          6     -3388.1096      6915.2191         1.8528    0.0545
      7   TPM1uf+G4         6     -3388.1101      6915.2201         1.8538    0.0545
      8   TPM3uf+G4         6     -3388.1664      6915.3328         1.9664    0.0515
      9   TPM2uf+I          6     -3388.3877      6915.7755         2.4091    0.0413
     10   TPM1uf+I          6     -3388.4237      6915.8474         2.4810    0.0398
------------------------------------------------------------------------------------
Best model according to AICc
---------------------------
Model:              HKY+G4
lnL:                -3388.1832
Frequencies:        0.3515 0.1873 0.2214 0.2398
Subst. Rates:       1.0000 14.9141 1.0000 1.0000 14.9141 1.0000
Inv. sites prop:    -
Gamma shape:        0.3150
Score:              6913.3664
Weight:             0.1377
```

Tree building

To build a tree we are going to use IQ-TREE. You can use any output file name in place of **treefileout**

```
cd ~/Phylogenetics

iqtree -s Combined_HA_genbank_aln.fas -bb 1000 -st DNA -nt 4 -alrt 1000 -pre treefileout
```

where:
-s : Input File
-bb : ultrafast bootstrap
-st : data type
-nt : Number of threads
-alrt : SH-like approximate likelihood ratio test
-pre : Prefix for output file

IQ-TREE outputs multiple files. The final tree file we will use has an extension of **.contree.** The final output file that we will take for tree viewing and editing is **HA_UFbootstrap_alrt_genbank.contree**

Step 4 – Viewing and Modifying a Tree File.

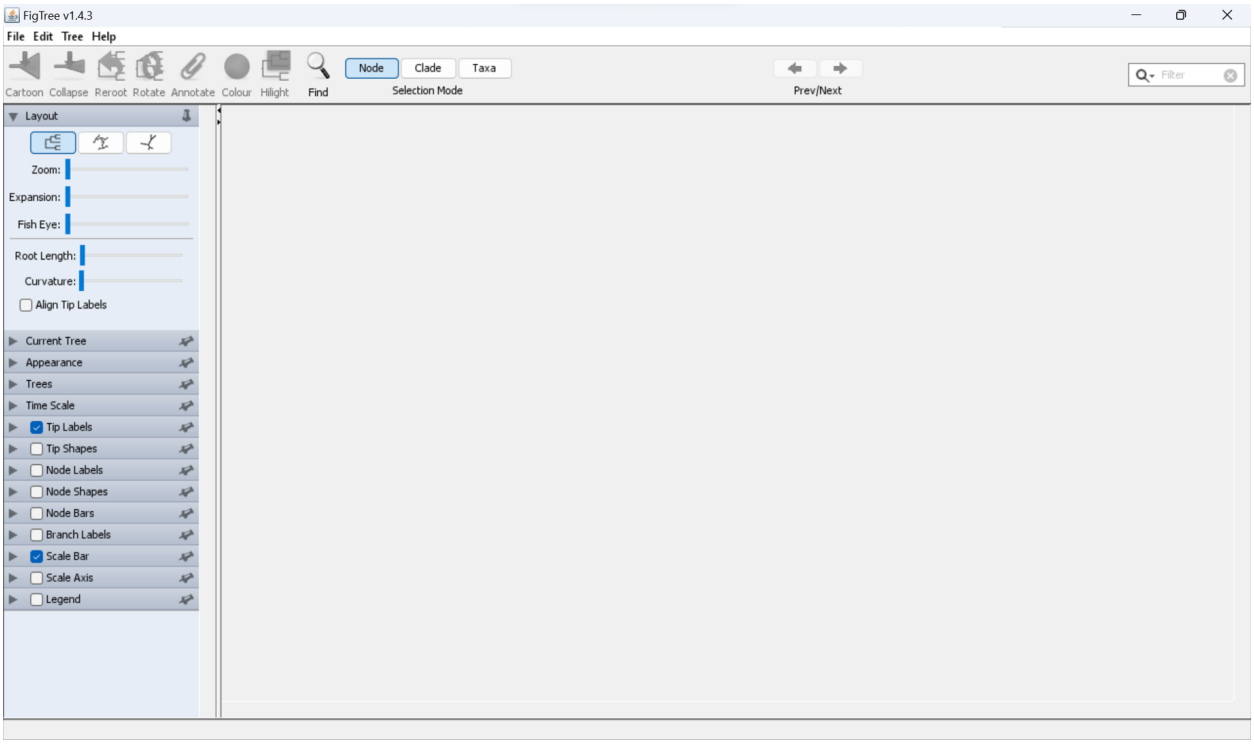
Software Used

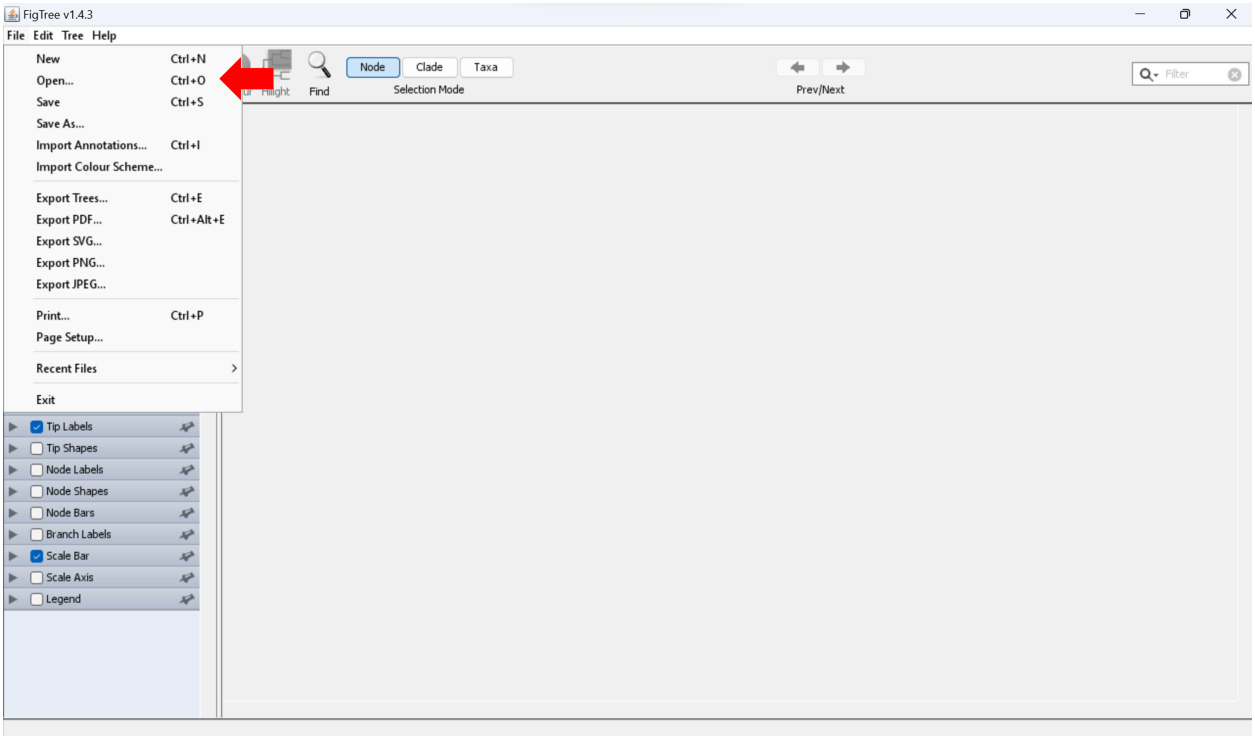
FigTree (http://tree.bio.ed.ac.uk/software/figtree/)

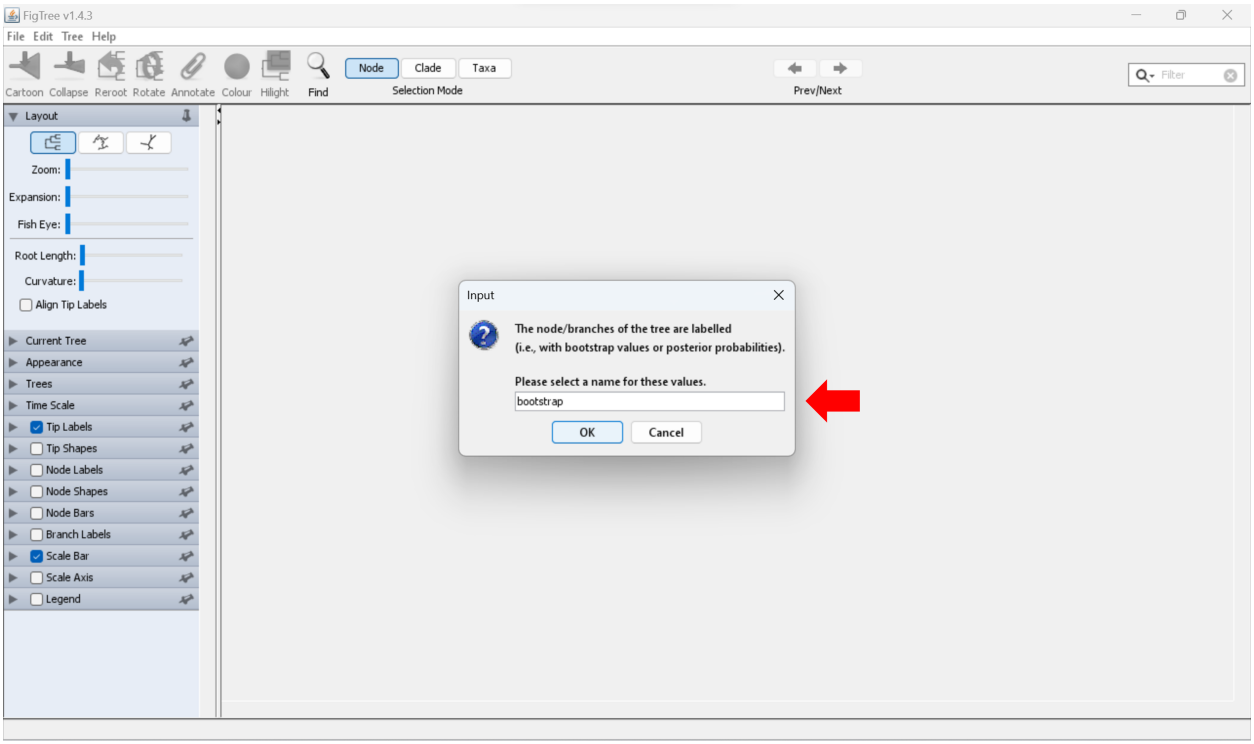
Alternate Software MEGA (https://www.megasoftware.net/)

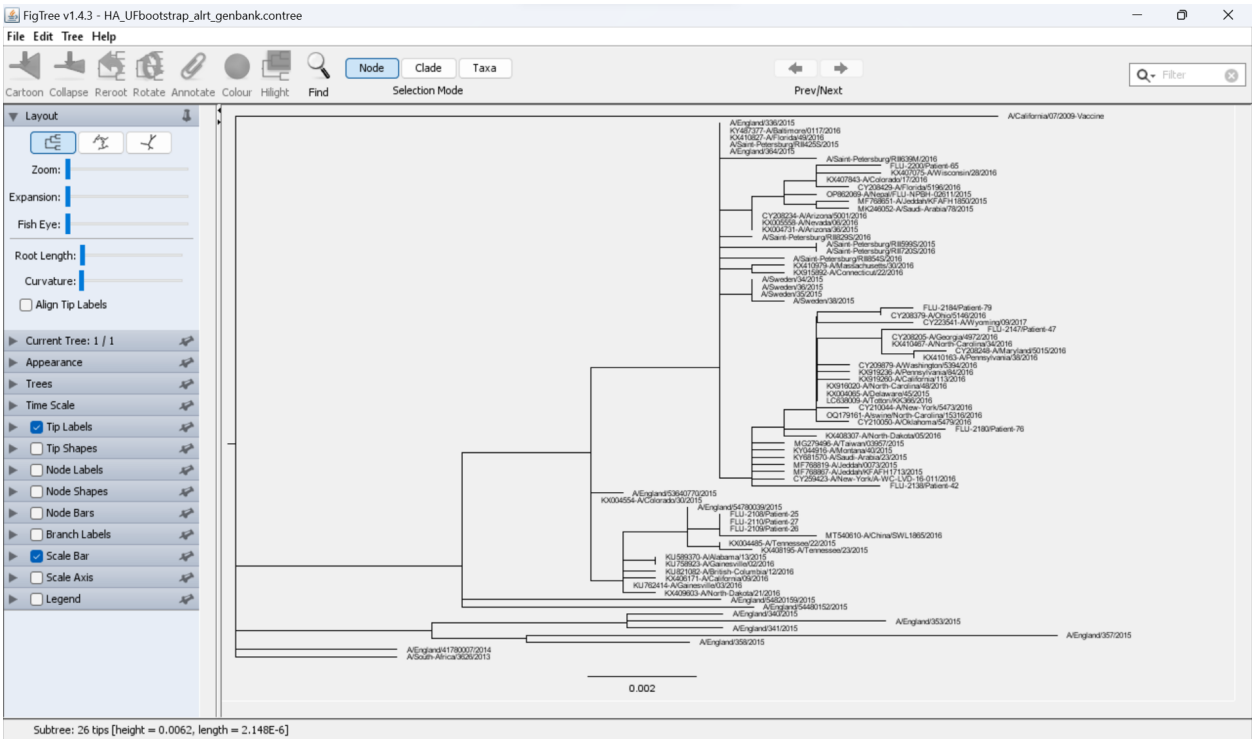
Figtree is Java based and will launch a GUI

```
cd ~/Phylogenetics

cp Example_Data/HA_UFbootstrap_alrt_genbank.contree .

#To launch figtree
figtree
```


FigTree v1.4.3
File Edit Tree Help
Cartoon Collapse Reroot Rotate Annotate Colour Hilight Find
Node Clade Taxa
Selection Mode
Prev/Next
Filter
Layout
Zoom:
Expansion:
Fish Eye:
Root Length:
Curvature:
Align Tip Labels
Current Tree
Appearance
Trees
Time Scale
Tip Labels
Tip Shapes
Node Labels
Node Shapes
Node Bars
Branch Labels
Scale Bar
Scale Axis
Legend

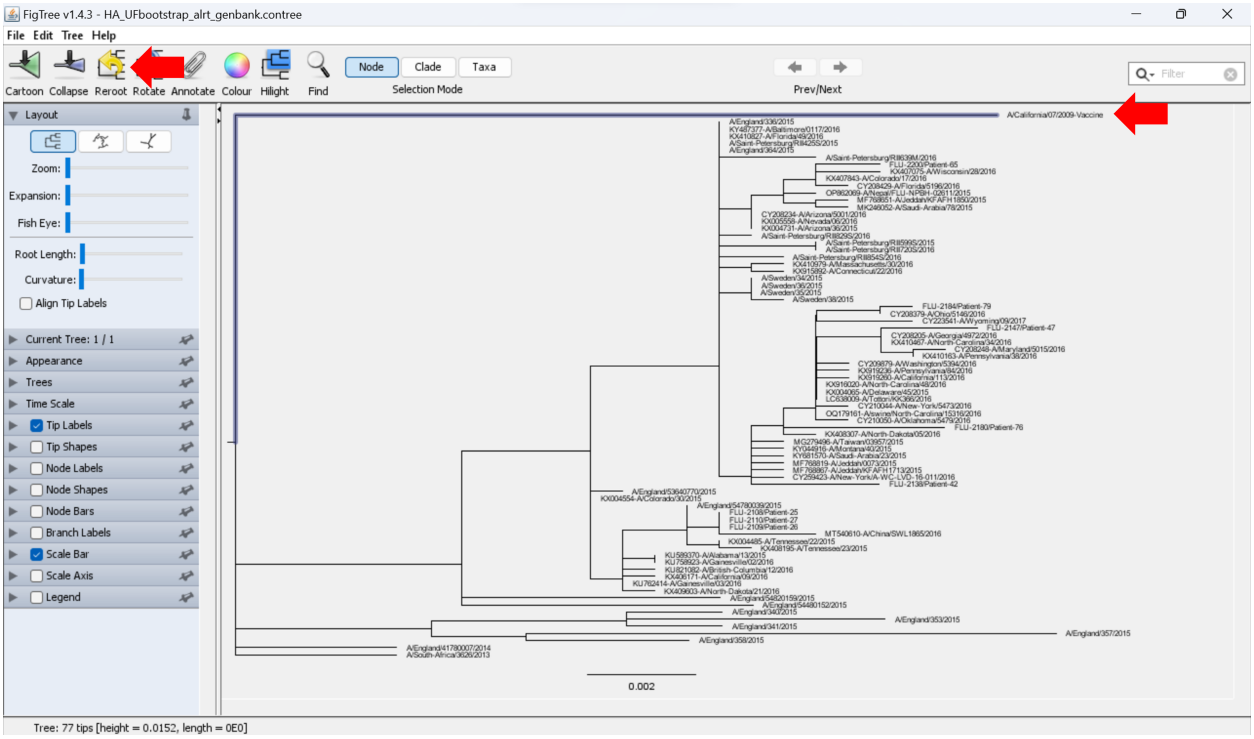

FigTree v1.4.3
File Edit Tree Help
New Ctrl+N
Open... Ctrl+O
Save Ctrl+S
Save As...
Import Annotations... Ctrl+I
Import Colour Scheme...
Export Trees... Ctrl+E
Export PDF... Ctrl+Alt+E
Export SVG...
Export PNG...
Export JPEG...
Print... Ctrl+P
Page Setup...
Recent Files
Exit
Hilight Find
Node Clade Taxa
Selection Mode
Prev/Next
Filter
Tip Labels
Tip Shapes
Node Labels
Node Shapes
Node Bars
Branch Labels
Scale Bar
Scale Axis
Legend

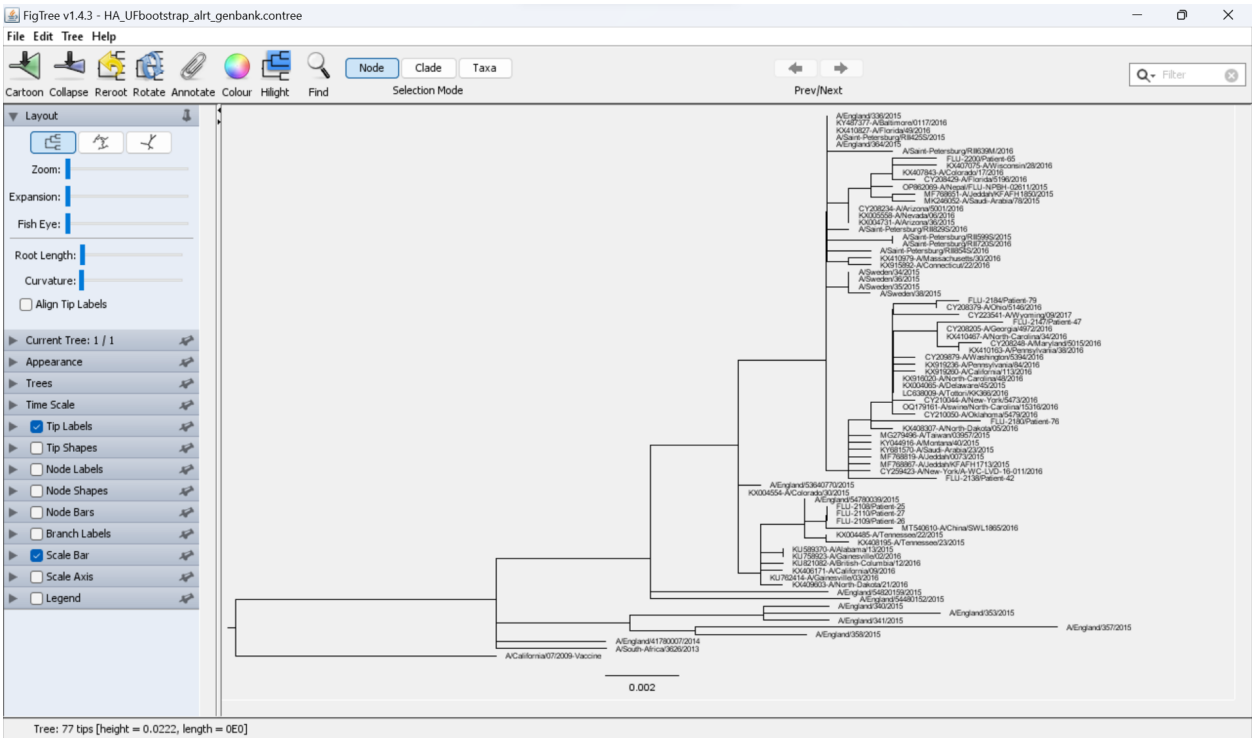




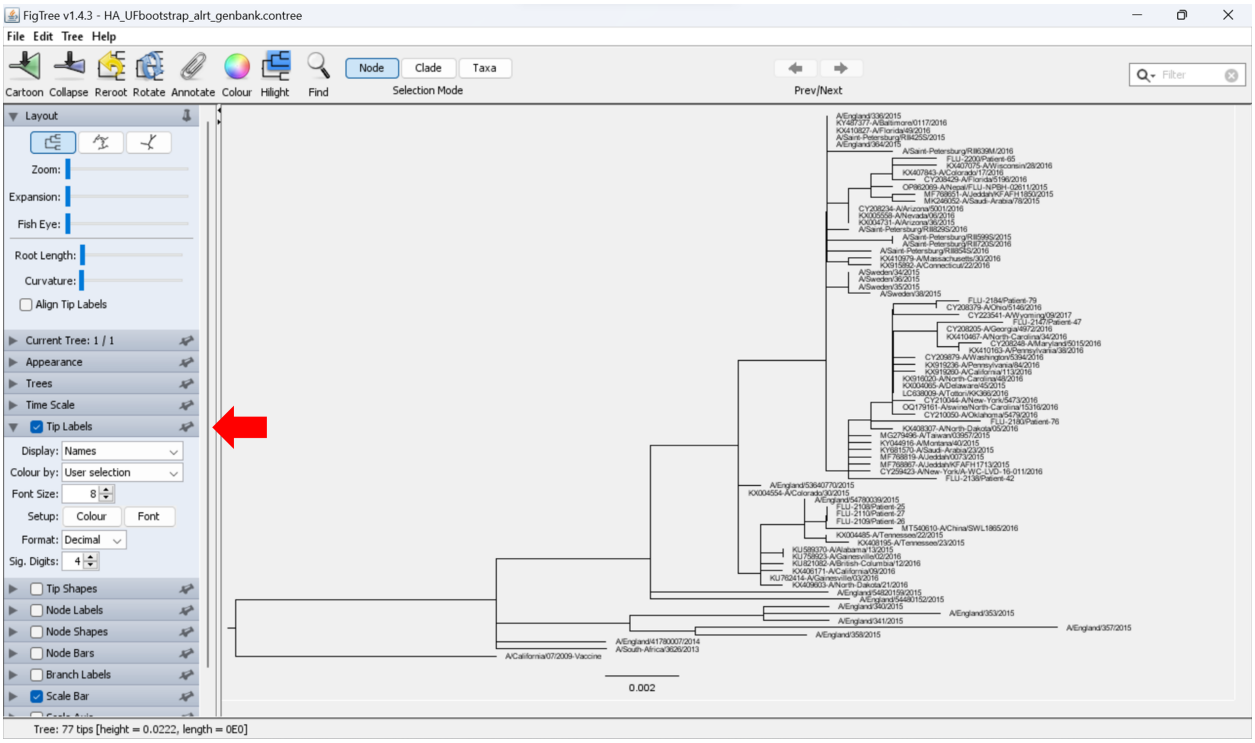
We will root the tree using the **A/California/07/2009-Vaccine Strain**. This is an unrelated strain of

the same genotype.



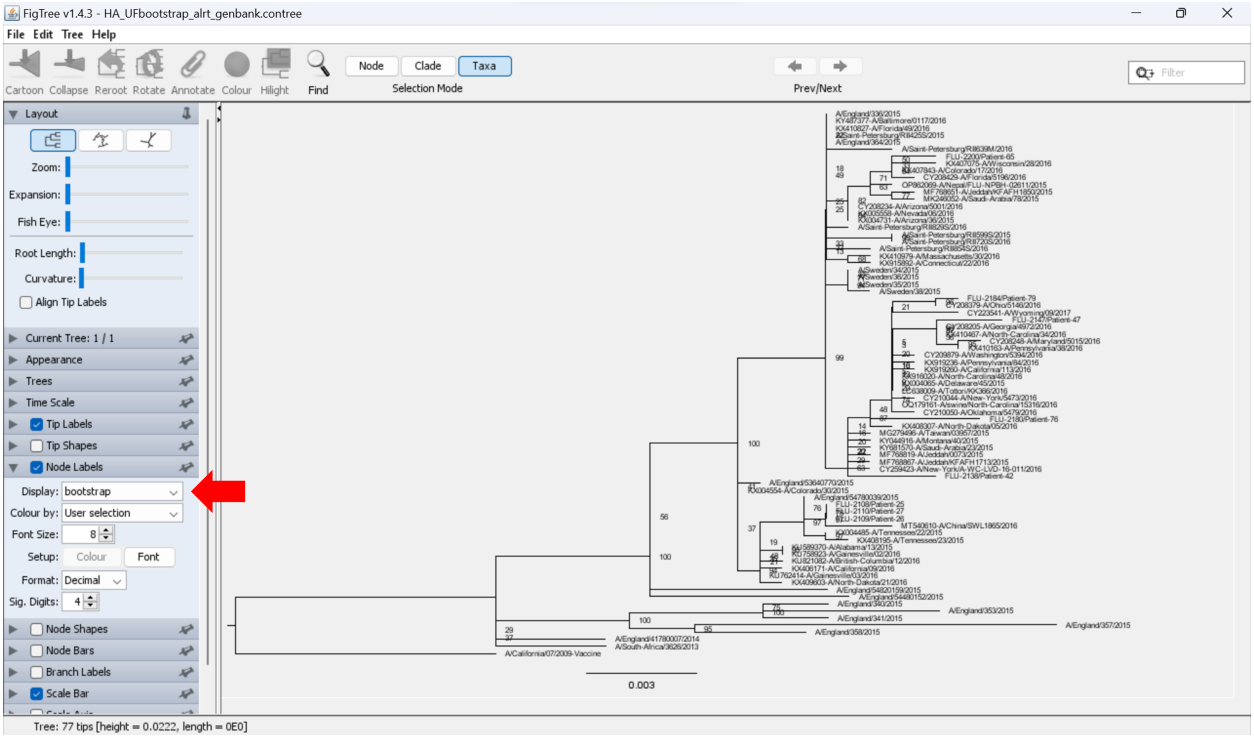


You can change the font type or size under **Tip Labels**

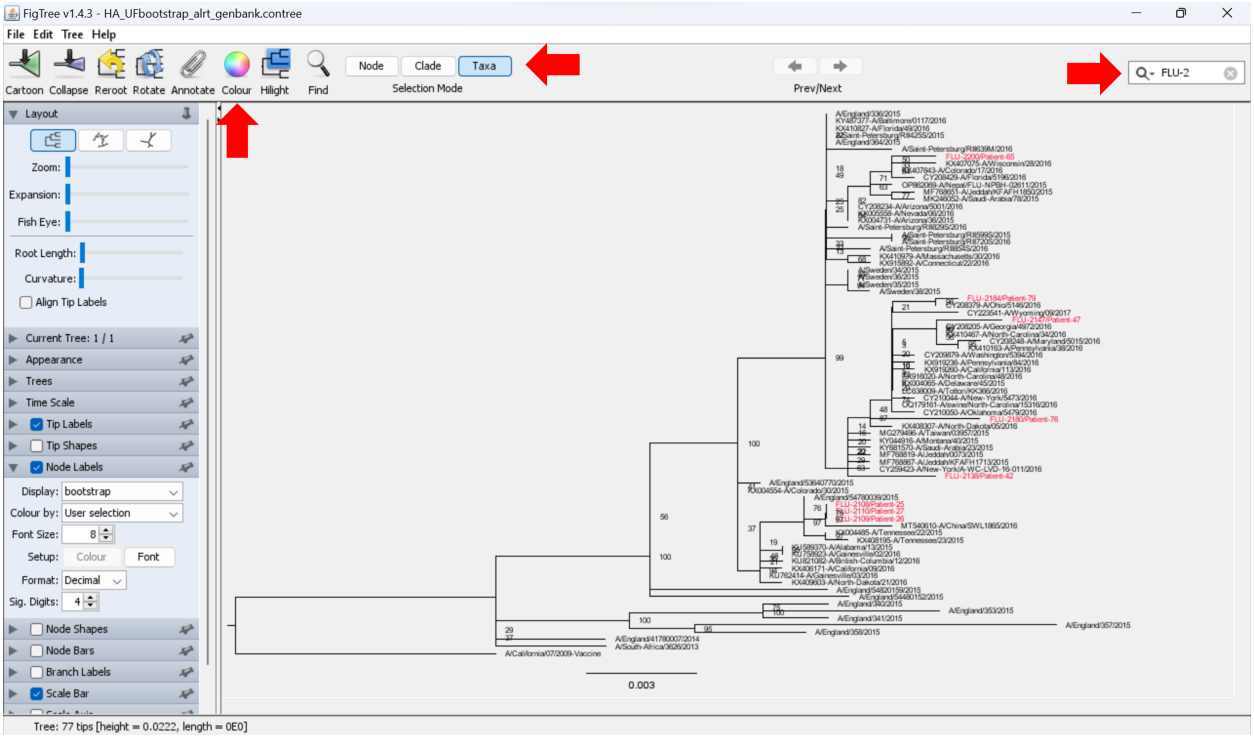


You can add the bootstrap values under **Node Labels**

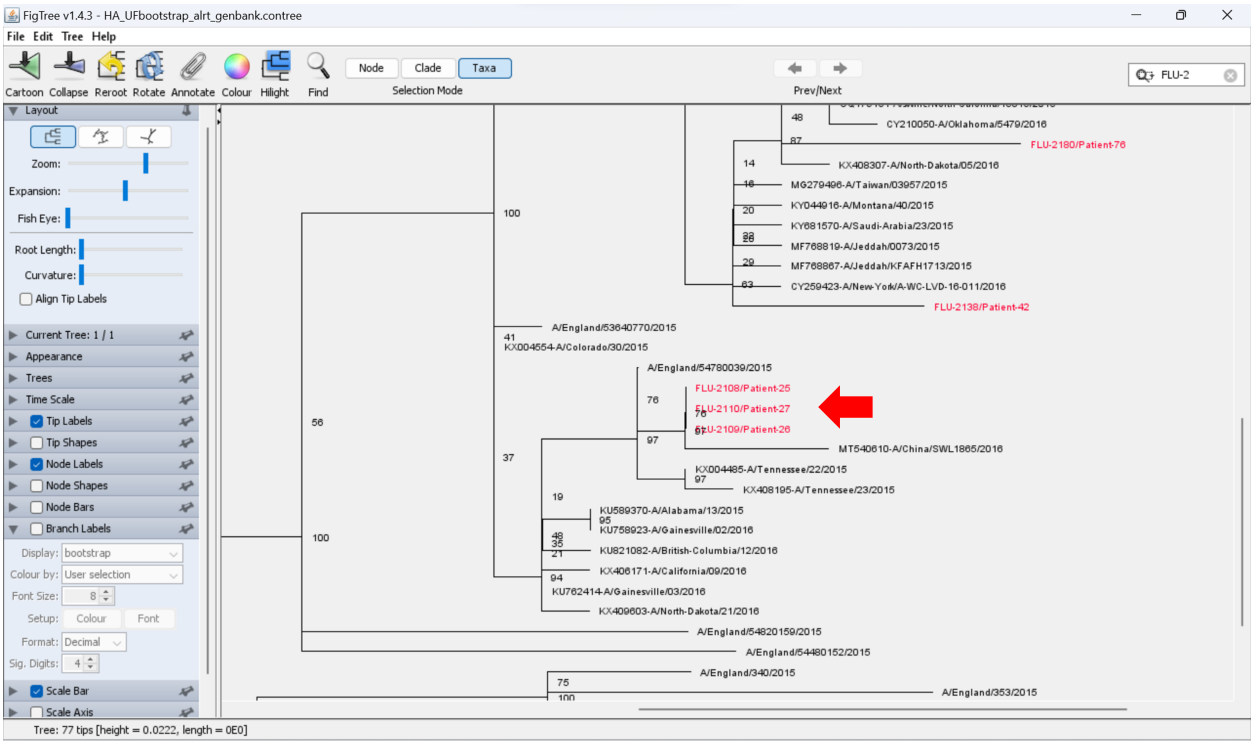
Note - We are using ultrafast bootstrap so values over 90 is considered significant. aLRT values above 80-85 is considered significant and can be found in the .treefile. If you use regular bootstrap with the -b option you can use 70 as a cutoff.



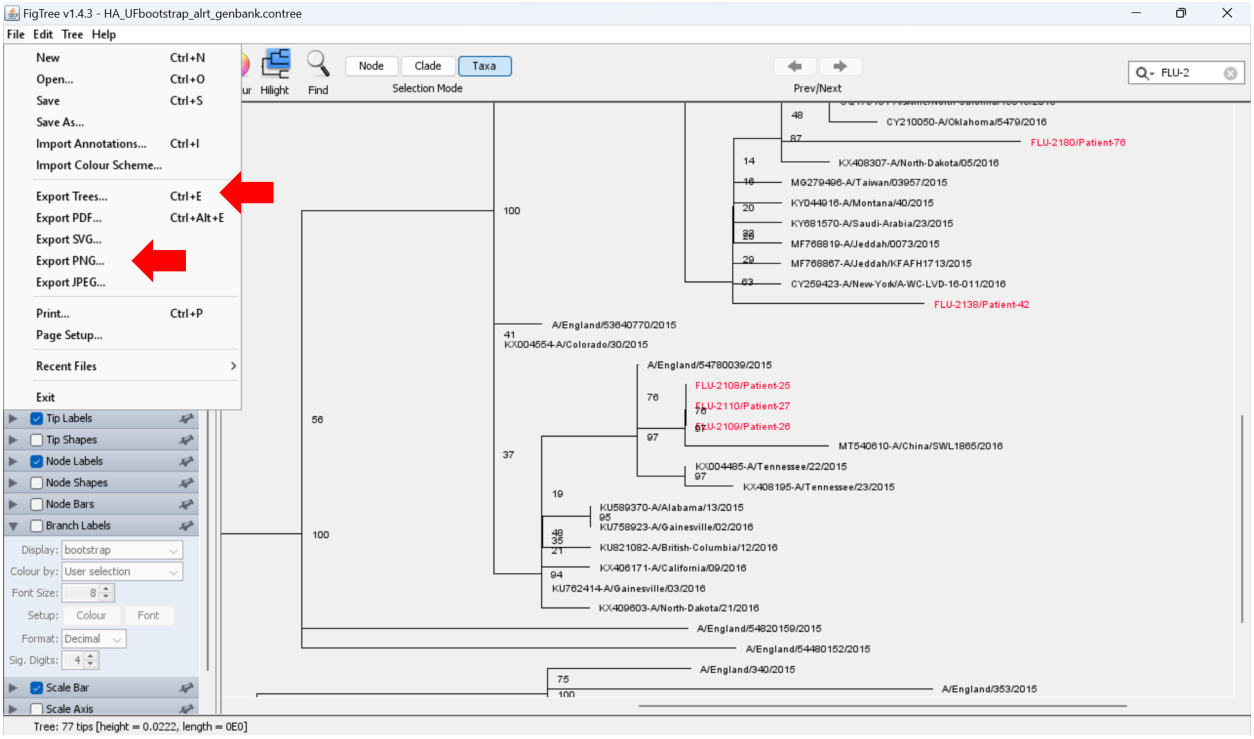
You can select our hospital strains and colour them by searching for **FLU-2** in the search bar above to highlight first and then colour the taxa.



Once highlighted we can find the hospital strains scattered across the tree, except the 3 strains below. They form one cluster and are identical in sequence suggesting a potential outbreak.



We can export the trees either in standard Nexus or Newick format. Trees can also be exported as PDFs or image formats like jpg.



We can further process the trees in **PowerPoint/ggtree/Illustrator** or other image editing tools to add metadata for each sequence in the alignment.