

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

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.

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

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.

National Library of Medicine
National Center for Biotechnology Information

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

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.
- Search, link, and download sequences programmatically 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>

GenBank Resources

[GenBank Home](#)
[Submission Types](#)
[Submission Tools](#)
[Search GenBank](#)
[Update GenBank Records](#)

National Library of Medicine
National Center for Biotechnology Information

sunandoroyucu

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

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

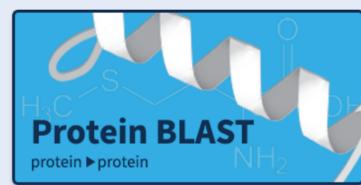
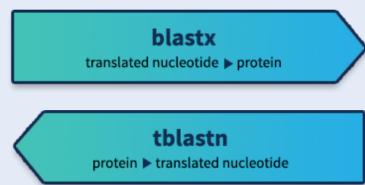
BLAST+ 2.15.0 is here!

We have included two exciting new features in the latest BLAST+ release

Tue, 28 Nov 2023

 [More BLAST news...](#)

Web BLAST



[blastn](#) [blastp](#) [blastx](#) [tblastn](#) [tblastx](#)

Standard Nucleotide BLAST

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 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?](#)

Organism

Nucleotide collection (nr/nt)

National Library of Medicine
National Center for Biotechnology Information

BLAST® > blastn suite

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?](#)

Organism [?](#)

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

National Library of Medicine
National Center for Biotechnology Information

BLAST® > blastn suite

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...](#) Hospital_HA.fas [?](#)

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?](#)

Organism [?](#)

FOR MORE INFO SEE [WHAT ARE TAXONOMIC IN DATABASES?](#)

Organism Optional

Exclude Optional
 Enter organism name or id--completions will be suggested exclude

Limit to Optional
 Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Entrez Query Optional
 Models (XM/XP) Uncultured/environmental sample sequences
 Sequences from type material
 Create custom database
 Enter an Entrez query to limit search

Program Selection

Optimize for
 Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)
 Choose a BLAST algorithm

BLAST Search database nt using Megablast (Optimize for highly similar sequences)
 Show results in a new window

+ Algorithm parameters



National Library of Medicine
National Center for Biotechnology Information 

BLAST® > blastn suite > results for RID-VJ7HSCFR016 [Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

[Edit Search](#) [Save Search](#) [Search Summary](#) [How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

Job Title	FLU-2184/Patient-79
RID	VJ7HSCFR016 <small>Search expires on 01-31 23:05 pm</small> Download All <input type="button" value="▼"/>
Results for	1:clQuery_3457642 FLU-2184/Patient-79(1701bp) <input type="button" value="▼"/>
Program	BLASTN <input type="button" value="?"/> Citation <input type="button" value="▼"/>
Database	nt See details <input type="button" value="▼"/>
Query ID	lclQuery_3457642
Description	FLU-2184/Patient-79
Molecule type	dna
Query Length	1701
Other reports	Distance tree of results MSA viewer <input type="button" value="?"/>

Filter Results

Organism only top 20 will appear exclude
 Type common name, binomial, taxid or group name

Percent Identity to **E value** to **Query Coverage** to

[Descriptions](#) [Graphic Summary](#) [Alignments](#) [Taxonomy](#) 

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

[Edit Search](#)

[Save Search](#)

[Search Summary](#)

[How to read this report?](#)

[BLAST Help Videos](#)

[Back to Traditional Results Page](#)

Job Title	FLU-2184/Patient-79
RID	VJ7HSCFR016 Search expires on 01-31 23:05 pm Download All ▾
Results for	1:iclQuery_3457642 FLU-2184/Patient-79(1701bp) ▾
Program	1:iclQuery_3457642 FLU-2184/Patient-79(1701bp)
Database	2:iclQuery_3457643 FLU-2147/Patient-47(1701bp)
Query ID	3:iclQuery_3457644 FLU-2180/Patient-76(1701bp)
Description	4:iclQuery_3457645 FLU-2138/Patient-42(1701bp)
Molecule type	5:iclQuery_3457646 FLU-2200/Patient-65(1701bp)
Query Length	6:iclQuery_3457647 FLU-2108/Patient-25(1701bp)
Other reports	7:iclQuery_3457648 FLU-2109/Patient-26(1701bp)
	8:iclQuery_3457649 FLU-2110/Patient-27(1701bp)

Filter Results

Organism	only top 20 will appear	<input type="checkbox"/> exclude
Type common name, binomial, taxid or group name		
+ Add organism		
Percent Identity	E value	Query Coverage
<input type="text"/> to <input type="text"/>	<input type="text"/> to <input type="text"/>	<input type="text"/> to <input type="text"/>
Filter Reset		

[Descriptions](#)

[Graphic Summary](#)

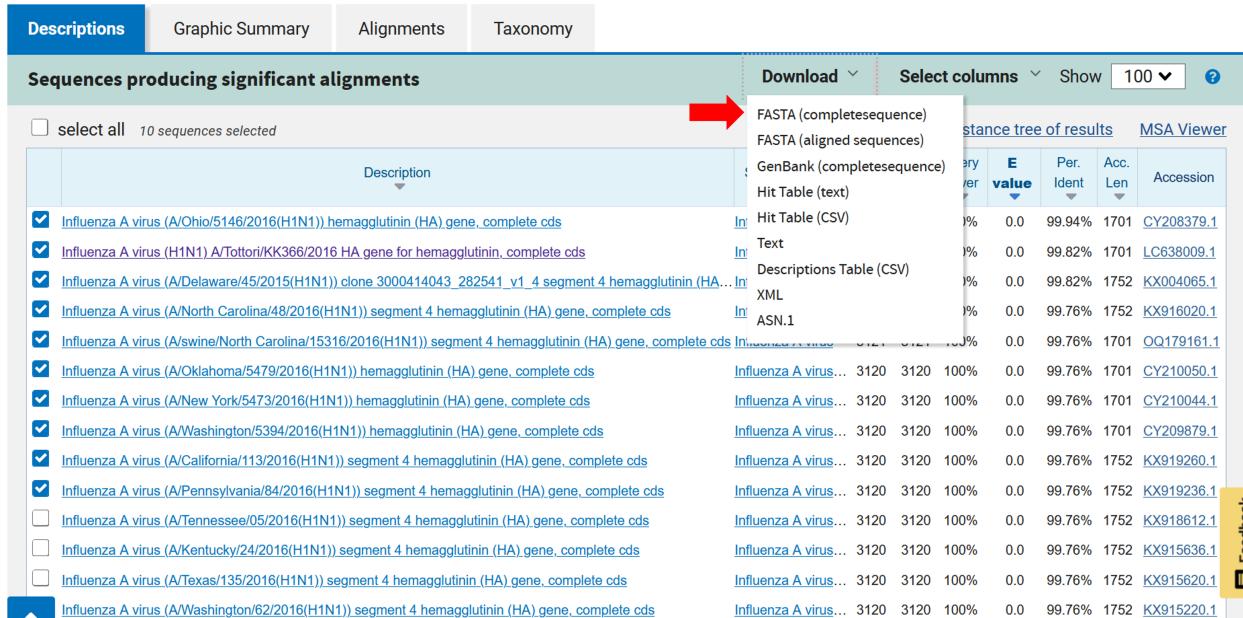
[Alignments](#)

[Taxonomy](#)

[Feedback](#)

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

Descriptions	Graphic Summary	Alignments	Taxonomy	Download ▾	Select columns ▾	Show 100 ▾	?		
Sequences producing significant alignments									
<input type="checkbox"/> select all 10 sequences selected	GenBank	Graphics	Distance tree of results	MSA Viewer					
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Influenza A virus (A/Ohio/5146/2016(H1N1)) hemagglutinin (HA) gene, complete cds	Influenza A virus...	3136	3136	100%	0.0	99.94%	1701	CY208379.1
<input checked="" type="checkbox"/>	Influenza A virus (H1N1) A/Tottori/KK366/2016 HA gene for hemagglutinin, complete cds	Influenza A virus...	3125	3125	100%	0.0	99.82%	1701	LC638009.1
<input checked="" type="checkbox"/>	Influenza A virus (A/Delaware/45/2015(H1N1)) clone 3000414043_282541_v1_4 segment 4 hemagglutinin (HA) gene, complete cds	Influenza A virus...	3125	3125	100%	0.0	99.82%	1752	KX004065.1
<input checked="" type="checkbox"/>	Influenza A virus (A/North Carolina/48/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	Influenza A virus...	3121	3121	100%	0.0	99.76%	1752	KX916020.1
<input checked="" type="checkbox"/>	Influenza A virus (A/swine/North Carolina/15316/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	Influenza A virus...	3121	3121	100%	0.0	99.76%	1701	QJ179181.1
<input checked="" type="checkbox"/>	Influenza A virus (A/Oklahoma/5479/2016(H1N1)) hemagglutinin (HA) gene, complete cds	Influenza A virus...	3120	3120	100%	0.0	99.76%	1701	CY210050.1
<input checked="" type="checkbox"/>	Influenza A virus (A/New York/5473/2016(H1N1)) hemagglutinin (HA) gene, complete cds	Influenza A virus...	3120	3120	100%	0.0	99.76%	1701	CY210044.1
<input checked="" type="checkbox"/>	Influenza A virus (A/Washington/5394/2016(H1N1)) hemagglutinin (HA) gene, complete cds	Influenza A virus...	3120	3120	100%	0.0	99.76%	1701	CY209879.1
<input checked="" type="checkbox"/>	Influenza A virus (A/California/113/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	Influenza A virus...	3120	3120	100%	0.0	99.76%	1752	KX919260.1
<input checked="" type="checkbox"/>	Influenza A virus (A/Pennsylvania/84/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	Influenza A virus...	3120	3120	100%	0.0	99.76%	1752	KX919236.1
<input type="checkbox"/>	Influenza A virus (A/Tennessee/05/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	Influenza A virus...	3120	3120	100%	0.0	99.76%	1752	KX918612.1
<input type="checkbox"/>	Influenza A virus (A/Kentucky/24/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	Influenza A virus...	3120	3120	100%	0.0	99.76%	1752	KX915636.1
<input type="checkbox"/>	Influenza A virus (A/Texas/135/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	Influenza A virus...	3120	3120	100%	0.0	99.76%	1752	KX915620.1
<input type="checkbox"/>	Influenza A virus (A/Washington/62/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	Influenza A virus...	3120	3120	100%	0.0	99.76%	1752	KX915220.1
<input type="checkbox"/>	Influenza A virus (A/Pennsylvania/67/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	Influenza A virus...	3120	3120	100%	0.0	99.76%	1752	KX411803.1
<input type="checkbox"/>	Influenza A virus (A/Colorado/20/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	Influenza A virus...	3120	3120	100%	0.0	99.76%	1752	KX410179.1



The screenshot shows a table of search results for 'Sequences producing significant alignments'. The columns include 'Description', 'E value', 'Per. Ident.', 'Acc. Len.', and 'Accession'. A red arrow points from the 'select all' checkbox at the top left to the 'Download' dropdown menu.

Sequences producing significant alignments		Download	Select columns	Show	100	?
		FASTA (complete sequence)	Sequence tree of results		MSA Viewer	
		FASTA (aligned sequences)				
	Description	E value	Per. Ident.	Acc. Len.	Accession	
<input checked="" type="checkbox"/>	Influenza A virus (A/Ohio/5146/2016(H1N1)) hemagglutinin (HA) gene, complete cds	In	99.94%	1701	CY208379.1	
<input checked="" type="checkbox"/>	Influenza A virus (H1N1) A/Tottori/KK366/2016 HA gene for hemagglutinin, complete cds	In	99.82%	1701	LC638009.1	
<input checked="" type="checkbox"/>	Influenza A virus (A/Delaware/45/2015(H1N1)) clone 3000414043_282541_v1_4 segment 4 hemagglutinin (HA) gene, complete cds	In	99.82%	1752	KX004065.1	
<input checked="" type="checkbox"/>	Influenza A virus (A/North Carolina/48/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	In	99.76%	1752	KX916020.1	
<input checked="" type="checkbox"/>	Influenza A virus (A/swine/North Carolina/15316/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	In	99.76%	1701	OQ179161.1	
<input checked="" type="checkbox"/>	Influenza A virus (A/Oklahoma/5479/2016(H1N1)) hemagglutinin (HA) gene, complete cds	In	99.76%	1701	CY210050.1	
<input checked="" type="checkbox"/>	Influenza A virus (A/New York/5473/2016(H1N1)) hemagglutinin (HA) gene, complete cds	In	99.76%	1701	CY210044.1	
<input checked="" type="checkbox"/>	Influenza A virus (A/Washington/5394/2016(H1N1)) hemagglutinin (HA) gene, complete cds	In	99.76%	1701	CY209879.1	
<input checked="" type="checkbox"/>	Influenza A virus (A/California/113/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	In	99.76%	1752	KX919260.1	
<input checked="" type="checkbox"/>	Influenza A virus (A/Pennsylvania/84/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	In	99.76%	1752	KX919236.1	
<input type="checkbox"/>	Influenza A virus (A/Tennessee/05/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	In	99.76%	1752	KX918612.1	
<input type="checkbox"/>	Influenza A virus (A/Kentucky/24/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	In	99.76%	1752	KX915636.1	
<input type="checkbox"/>	Influenza A virus (A/Texas/135/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	In	99.76%	1752	KX915620.1	
<input type="checkbox"/>	Influenza A virus (A/Washington/62/2016(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds	In	99.76%	1752	KX915220.1	

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.

Received lines: 3
Rejected lines: 0
Removed duplicates: 0
Passed to Entrez: 3
[Retrieve records for 3 UID\(s\)](#)



National Library of Medicine
National Center for Biotechnology Information

Nucleotide Nucleotide Advanced Search Help

Species
Viruses (3)
Customize ...

Molecule types
genomic DNA/RNA (3)
Customize ...

Source databases
INSDC (GenBank) (3)
Customize ...

Sequence Type
Nucleotide (3)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

[Clear all](#)
[Show additional filters](#)

Items: 3
Selected: 3

- [Influenza A virus \(A/British Columbia/12/2016\(H1N1\)\) segment 4 hemagglutinin \(HA\).gene.complete cds](#)
1,745 bp linear cRNA
Accession: KU821082.1 GI: 1004615602
Protein Taxonomy
[GenBank](#) [FASTA](#) [Graphics](#)
- [Influenza A virus \(A/Tennessee/22/2015\(H1N1\)\) clone 3000415653_283363_v1_4 segment 4 hemagglutinin \(HA\).gene.complete cds](#)
1,752 bp linear cRNA
Accession: KX004485.1 GI: 1028346949
Protein Taxonomy
[GenBank](#) [FASTA](#) [Graphics](#)
- [Influenza A virus \(A/China/SWL1865/2016\(H1N1\)\) segment 4 hemagglutinin \(HA\).gene.complete cds](#)
1,701 bp linear cRNA
Accession: MT540610.1 GI: 1847287170
Protein PubMed Taxonomy
[GenBank](#) [FASTA](#) [Graphics](#)

Send to: [Filters: Manage Filters](#)

Results by taxon
[Top Organisms \[Tree\]](#)
Influenza A virus (A/British Columbia/12/2016(H1N1)) (1)
Influenza A virus (A/Tennessee/22/2015(H1N1)) (1)
Influenza A virus (3)

Analyze these sequences
[Run BLAST](#)

Find related data
Database: [Select](#)

[Find items](#)

Recent activity
[Turn Off](#) [Clear](#)

[Entrez Direct: E-utilities on the Unix](#)

National Library of Medicine
National Center for Biotechnology Information

Nucleotide Nucleotide Advanced Search Help

Species
Viruses (3)
Customize ...

Molecule types
genomic DNA/RNA (3)
Customize ...

Source databases
INSDC (GenBank) (3)
Customize ...

Sequence Type
Nucleotide (3)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

[Clear all](#)

Items: 3
Selected: 3

- [Influenza A virus \(A/British Columbia/12/2016\(H1N1\)\) segment 4 hemagglutinin \(HA\).gene.complete cds](#)
1,745 bp linear cRNA
Accession: KU821082.1 GI: 1004615602
Protein Taxonomy
[GenBank](#) [FASTA](#) [Graphics](#)
- [Influenza A virus \(A/Tennessee/22/2015\(H1N1\)\) clone 3000415653_283363_v1_4 segment 4 hemagglutinin \(HA\).gene.complete cds](#)
1,752 bp linear cRNA
Accession: KX004485.1 GI: 1028346949
Protein Taxonomy
[GenBank](#) [FASTA](#) [Graphics](#)
- [Influenza A virus \(A/China/SWL1865/2016\(H1N1\)\) segment 4 hemagglutinin \(HA\).gene.complete cds](#)
1,701 bp linear cRNA
Accession: MT540610.1 GI: 1847287170
Protein PubMed Taxonomy
[GenBank](#) [FASTA](#) [Graphics](#)

Send to: [Filters: Manage Filters](#)

Complete Record
 Coding Sequences
 Gene Features

Choose Destination
 File
 Clipboard
 Collections
 Analysis Tool

Download 3 items.
Format: [FASTA](#)

Sort by: [Default order](#)
Show GI

Recent activity

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/>)

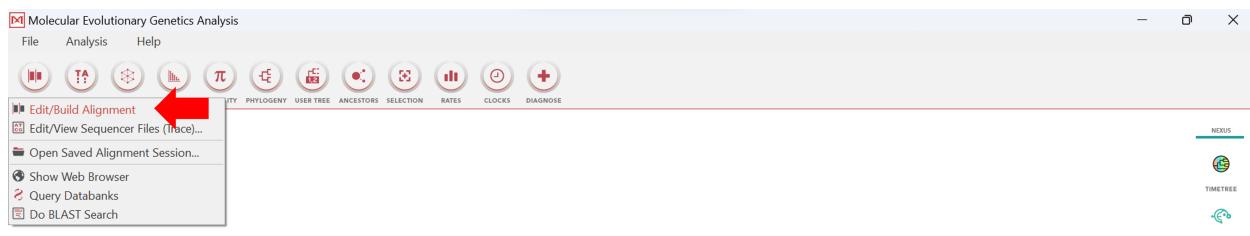
MEGA (<https://www.megasoftware.net/>)

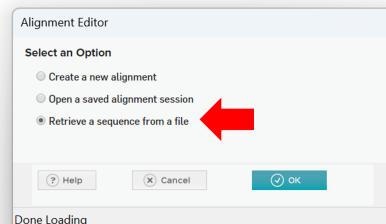
Alternate Software – MUSCLE, CLUSTALW

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
```





MX: Alignment Explorer (Combined_HA_genbank.fas)

Data Edit Search Alignment Web Sequencer Display Help

DNA Sequences Translated Protein Sequences

Species/Abbrv

24 A/Sweden/38/2015	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C I G C A A A T G C A S C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
25 FLU-2184/Patient-79	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
26 FLU-2147/Patient-47	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
27 FLU-2180/Patient-76	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
28 FLU-2138/Patient-42	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
29 FLU-2200/Patient-65	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
30 FLU-2108/Patient-25	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
31 FLU-2109/Patient-26	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
32 FLU-2110/Patient-27	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
33 CY208234/A/Georgia/4972/2016	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
34 CY208234/A/Arizona/5001/2016	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
35 CY208248/A/Maryland/5015/2016	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
36 CY208379/A/Ohio/5146/2016	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
37 CY208429-A/F/Iorida/5196/2016	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
38 CY20879-A/Washington/5394/2016	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
39 CY210044/A/NewYork-A/WC-LVD-16-011/2016	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
40 CY210050/A/Oklahoma/5479/2016	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C T T A T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
41 CY223541-A/Wyoming/09/2017	G G A A A C C A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
42 CY259423-A/Wyoming/09/2017	G G A A A C C A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
43 KU0589370/A/Alabama/3/2015	G G A A A C C A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
44 KU075892-A/Gainesville/02/2016	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
45 KU092414-A/Gainesville/03/2016	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
46 KU0821062-A/British-Columbia/12/2016	T G A A G G C A A T A C T A G T A G T T G C T A T A T A C A T T T A C A C C G C A A A T G C A G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
47 KX004065-A/Delaware/45/2015	G G A A A C C A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
48 KX004485-A/Tennessee/22/2015	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
49 KX004554-A/Colorado/30/2015	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
50 KX004731-A/Arizona/36/2015	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
51 KX005558-A/Nevada/06/2016	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
52 KX0406171-A/California/09/2016	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
53 KX0407075-A/Wisconsin/28/2016	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
54 KX0407843-A/Colorado/17/2016	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
55 KX0408195-A/Tennessee/23/2015	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
56 KX0408307-A/North-Dakota/05/2016	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
57 KX0409603-A/North-Dakota/21/2016	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
58 KX0410163-A/Pennsylvania/38/2016	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
59 KX0410467-A/North-Carolina/34/2016	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
60 KX0410827-A/Florida/49/2016	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
61 KX0410979-A/Massachusetts/30/2016	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
62 KX0915892-A/Connecticut/22/2016	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C
63 KX091620-A/North-Carolina/48/2016	G G A A A C C A A A A G C A A C C A A A T G C A A T A C T A G C A G T G C T A T G C A G T T T C A A C C G C A A A T G C A C A C G C A T T T G T A T A G G T T A T C A T G C G A A C A A T T C A A C

Site # 1 with w/o gaps Selected genetic code: Standard

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
<hr/>						
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
<hr/>						
Best model according to AICc						
<hr/>						
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.

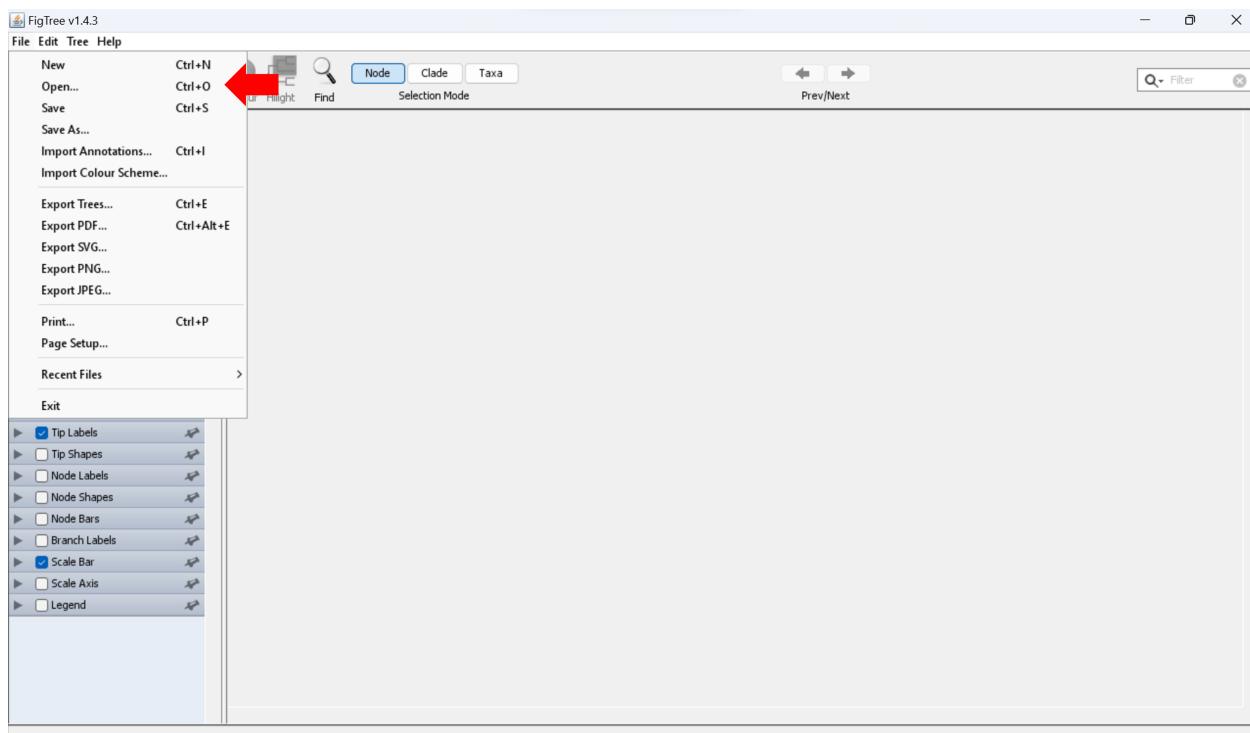
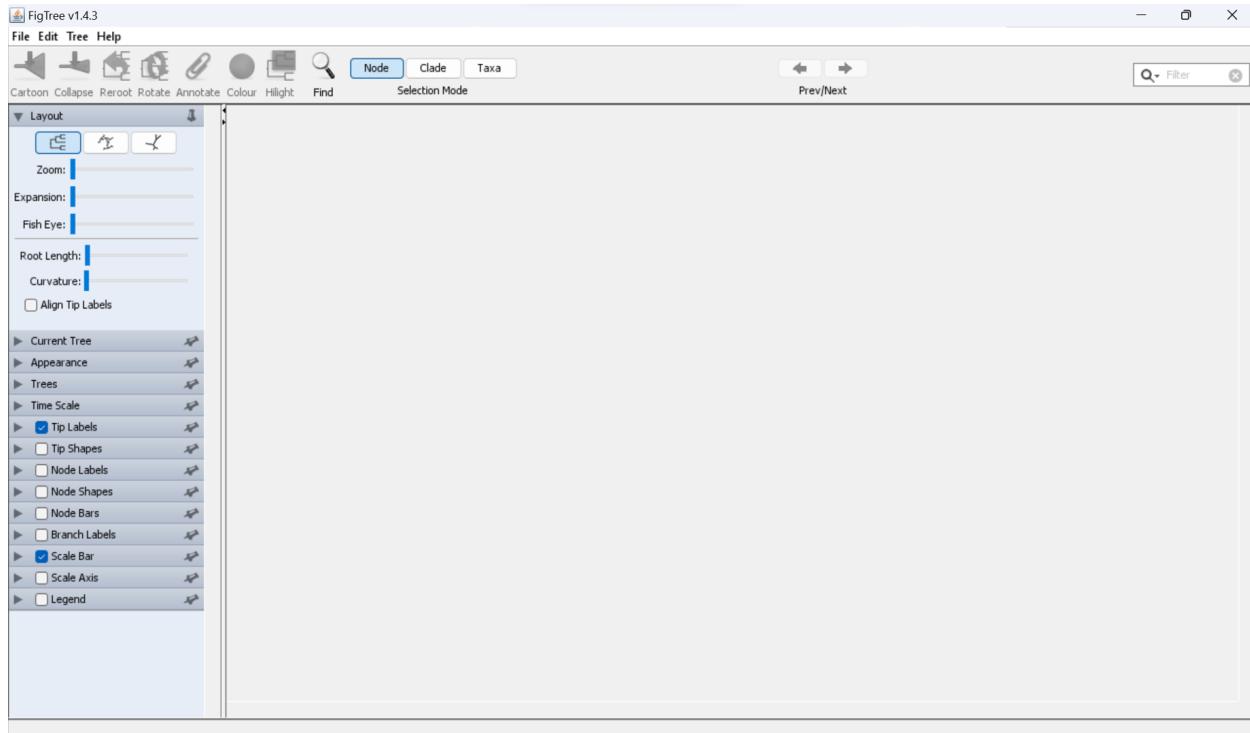
Software Used

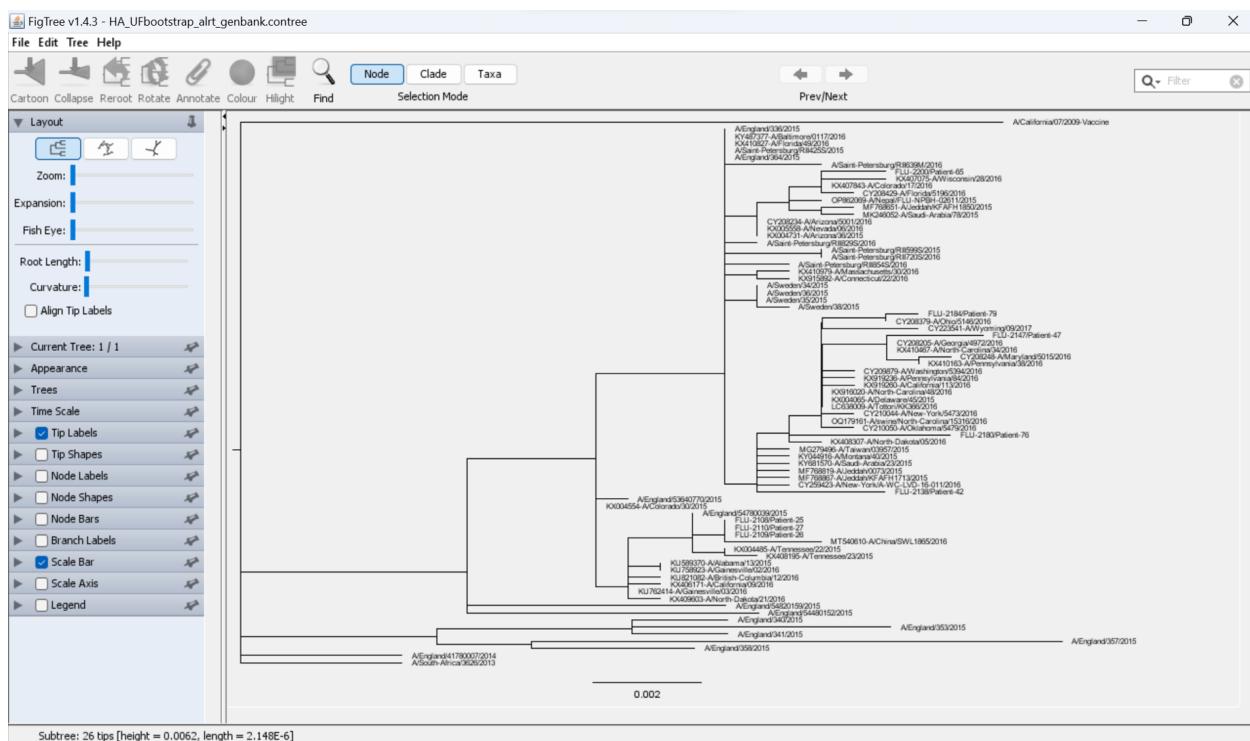
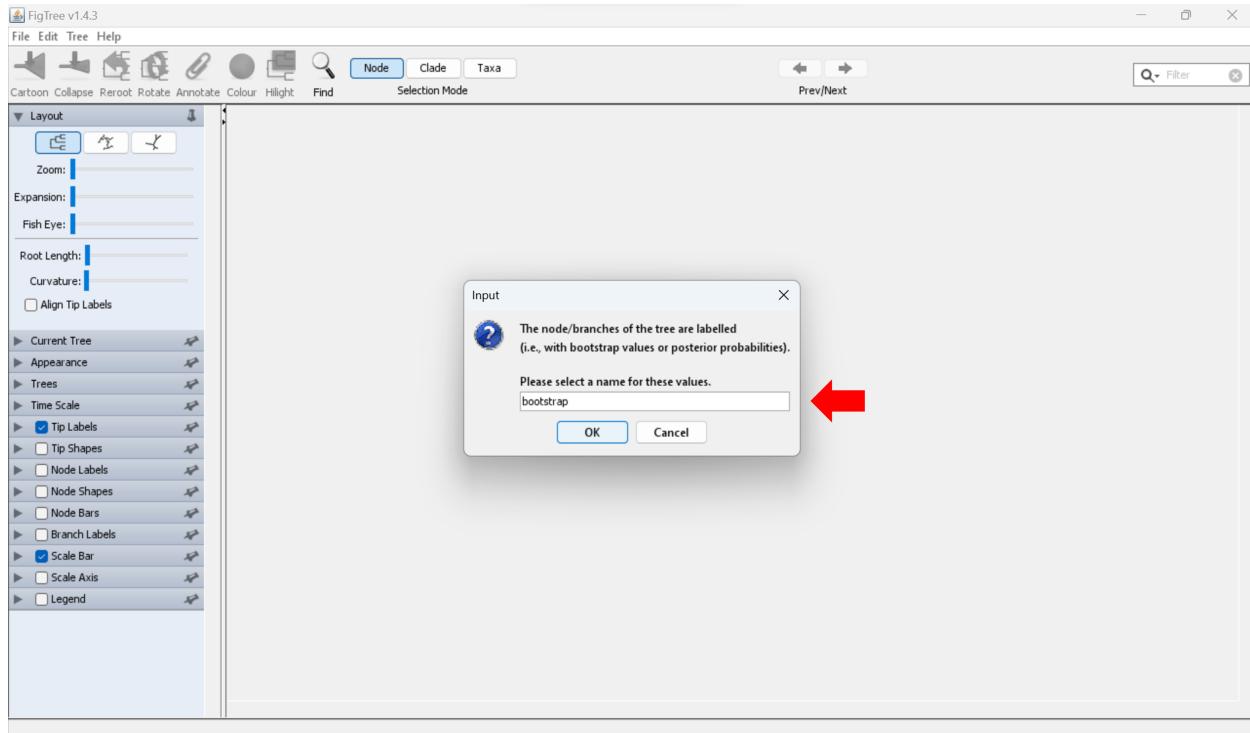
FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>)

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

Figtree is Java based and will launch a GUI

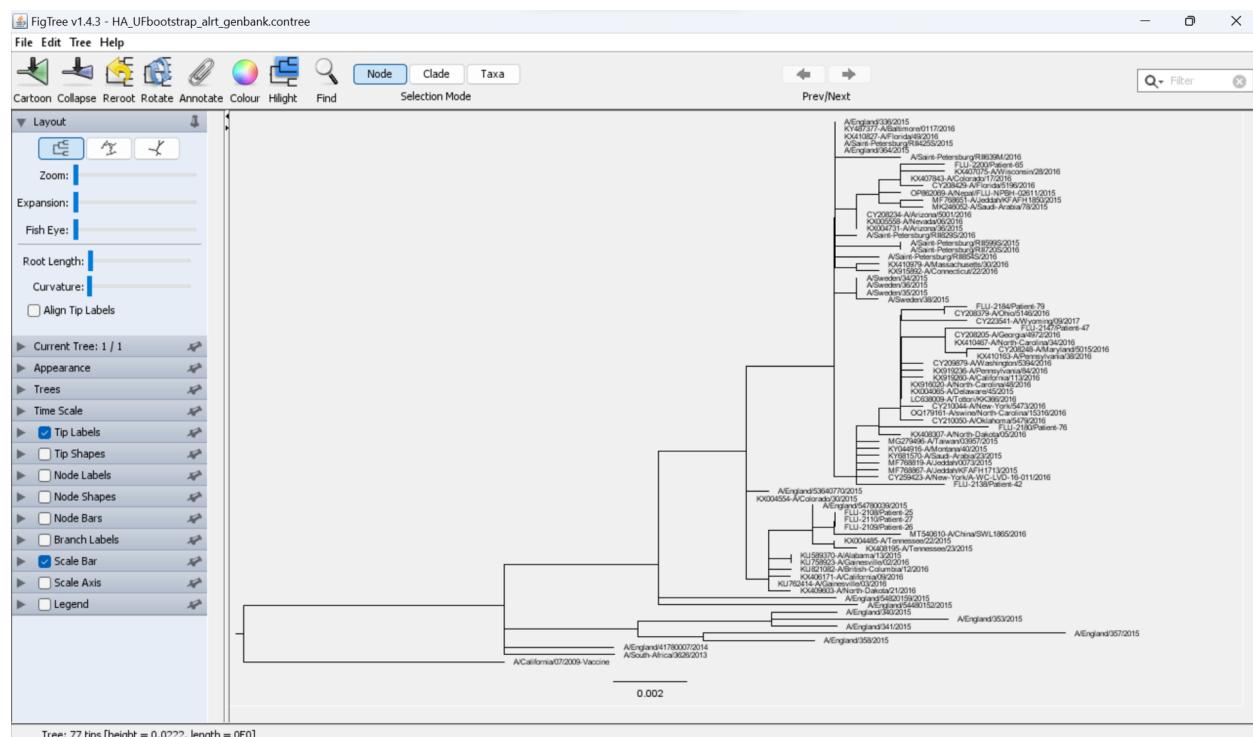
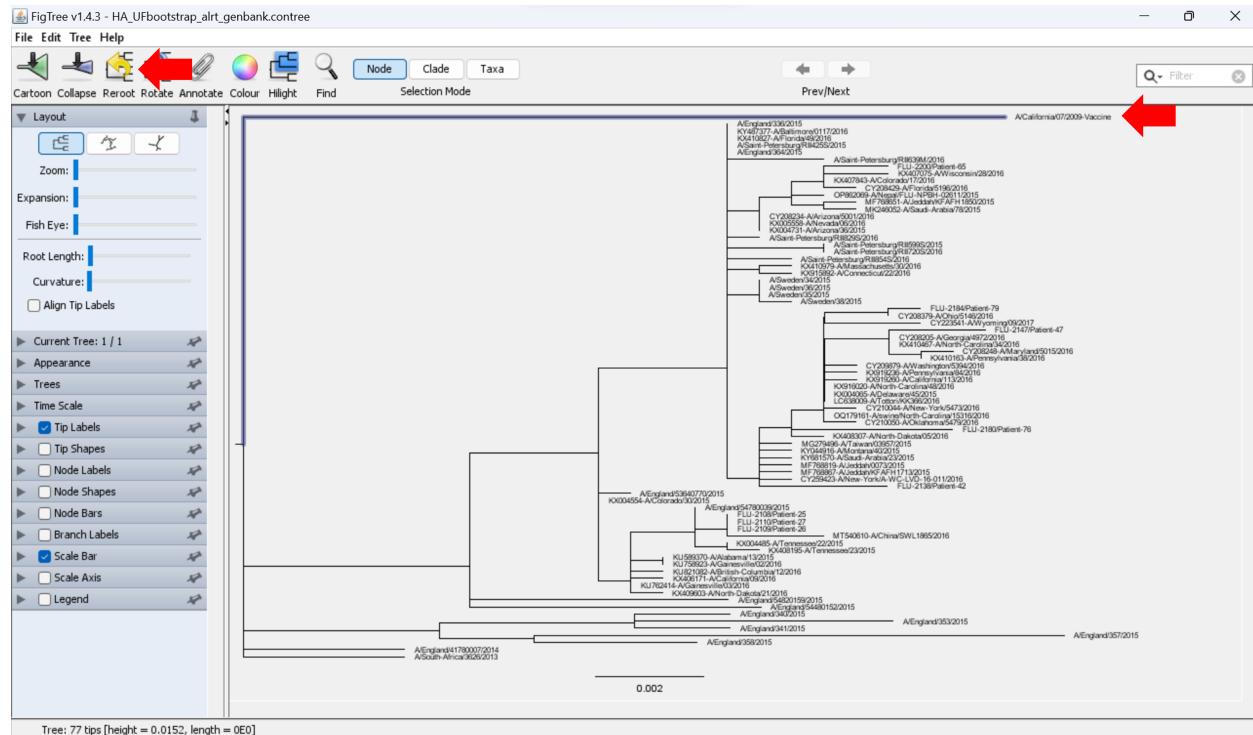
```
cd ~/Phylogenetics
cp Example_Data/HA_UFbootstrap_alrt_genbank.contree .
#To launch figtree
figtree
```



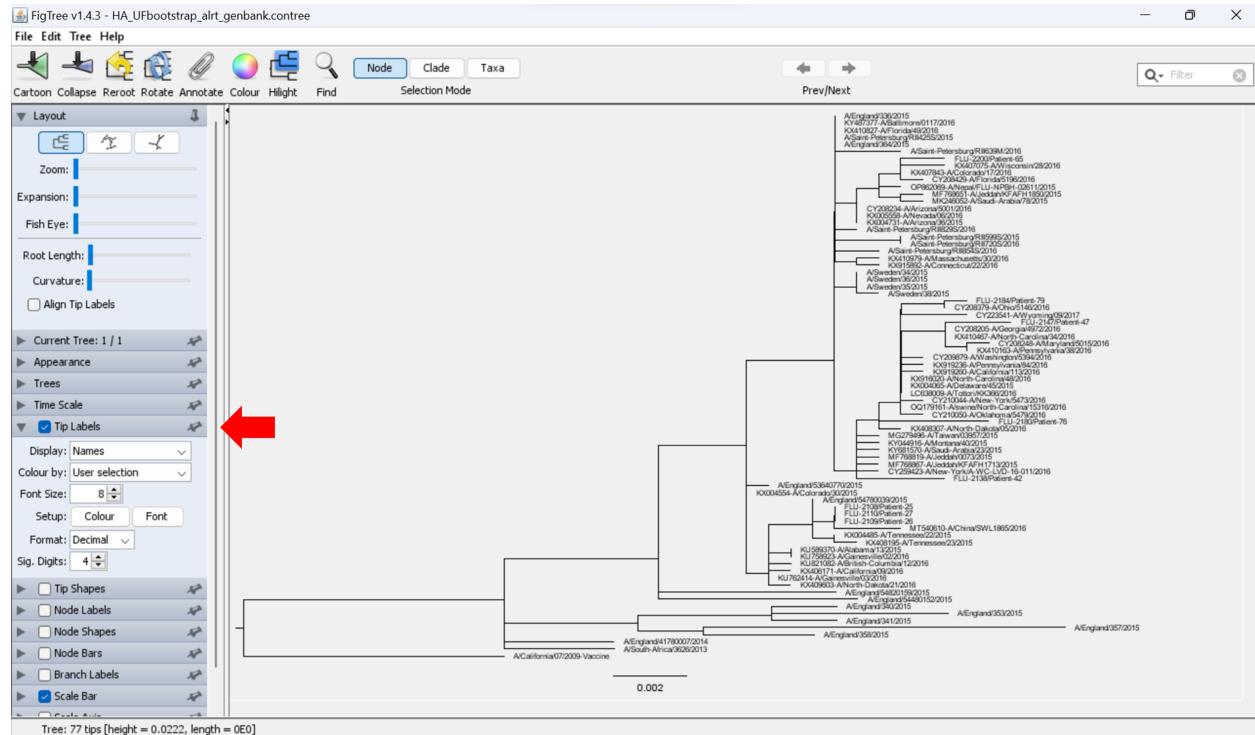


We will root the tree using the **A/California/07/2009-Vaccine Strain**. This is an unrelated strain of

the same genotype.

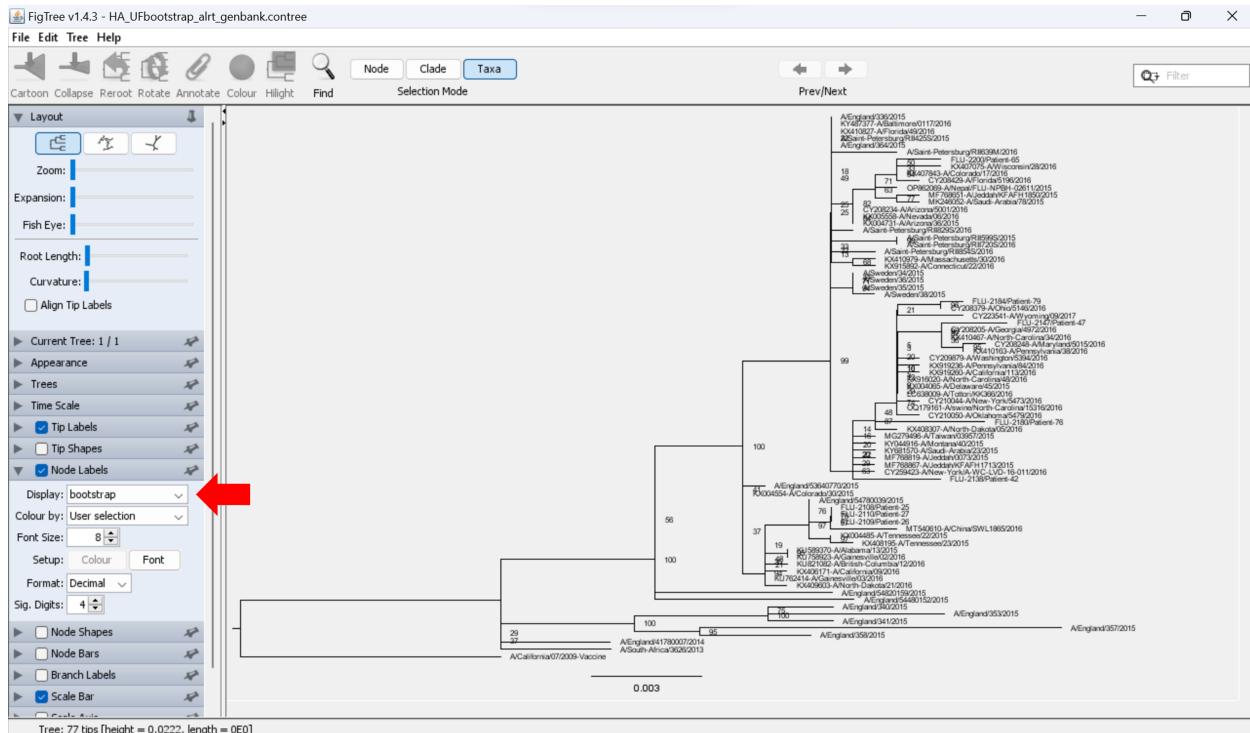


You can change the font type or size under **Tip Labels**

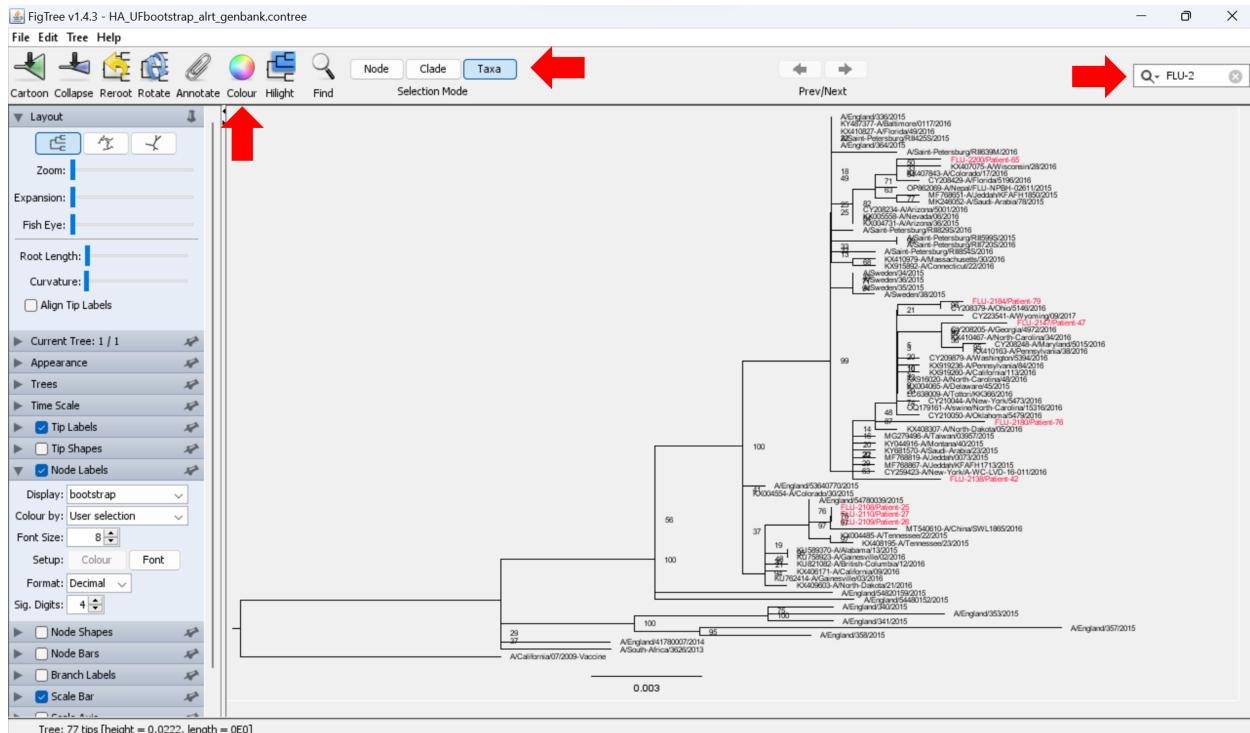


You can add the bootstrap values under **Node Labels**

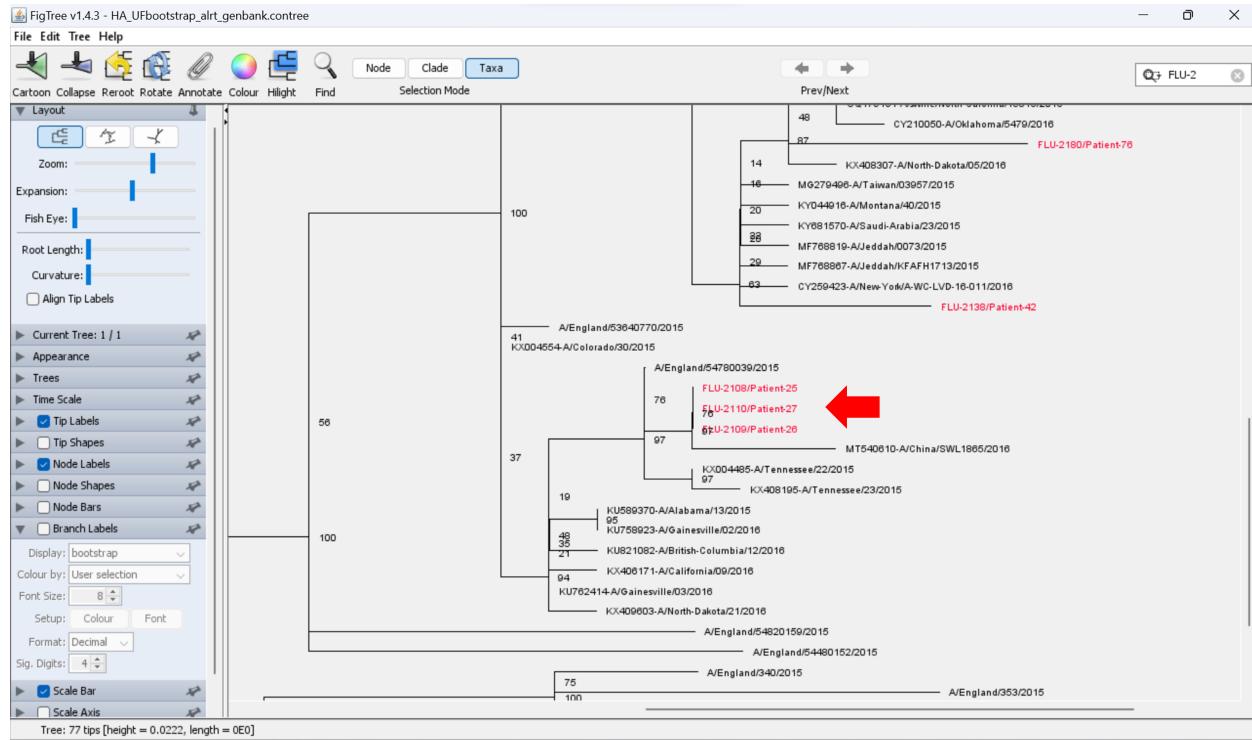
Note - We are using ultrafast bootstrap so values over 90 is considered significant. aLRT values above 80-85 is considered significant and can be found in the .treefile. If you use regular bootstrap with the -b option you can use 70 as a cutoff.



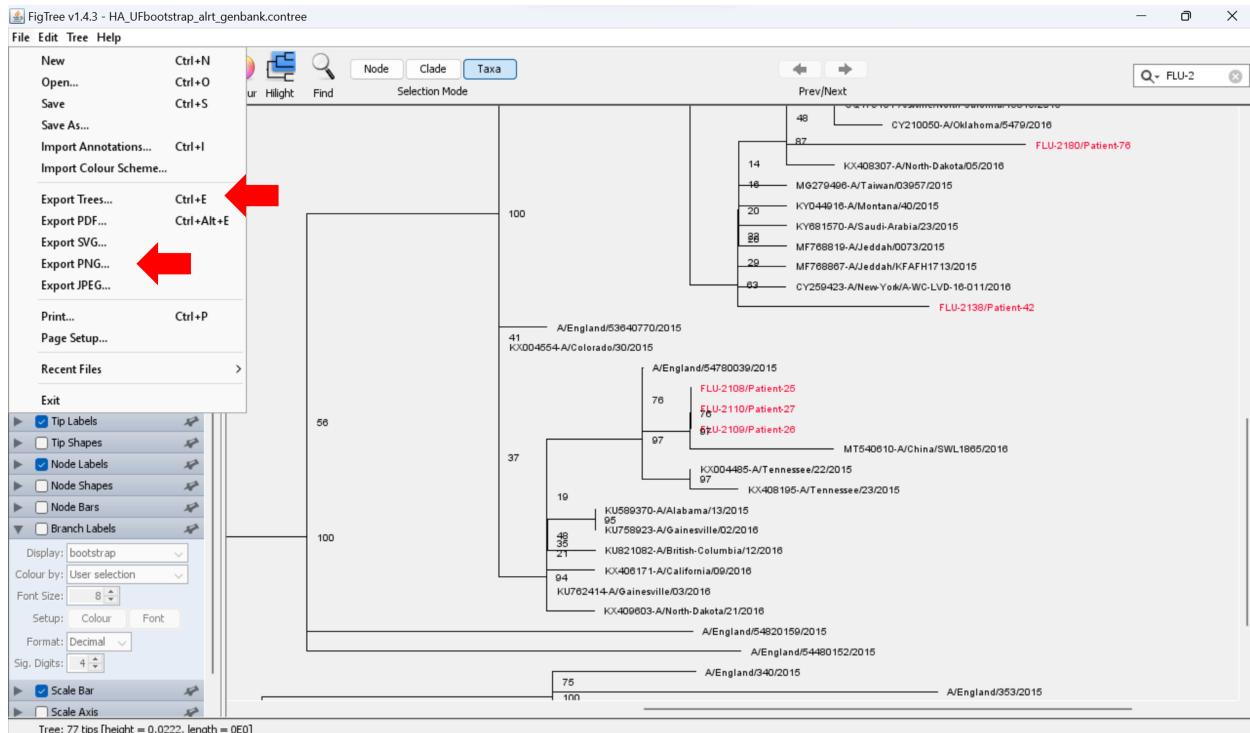
You can select our hospital strains and colour them by searching for **FLU-2** in the search bar above to highlight first and then colour the taxa.



Once highlighted we can find the hospital strains scattered across the tree, except for 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.