wellcome connecting science

Course Manual - Phylogenetic Analysis

Scenario 3 – Retrieval of sequences and phylogenetic analysis

From the metagenomic analysis carried out you have determined the pathogen responsible for the outbreak above and now wish to identify a likely source of infection. Using phylogenetic analysis, what is the likely source of the infection?

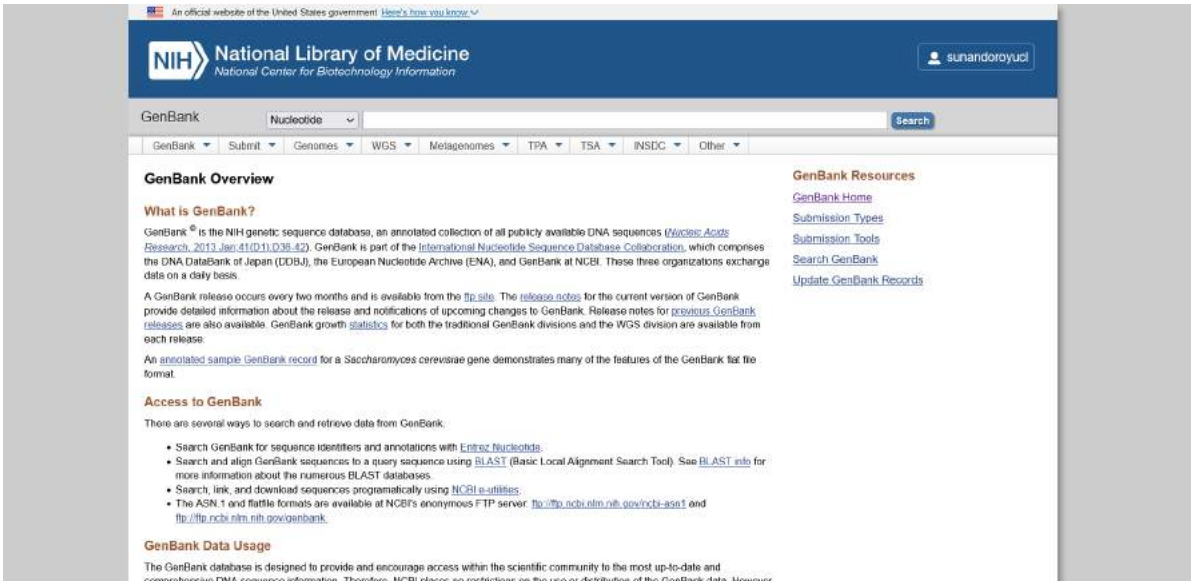
Note: All analysis in this section will be carried out on the viral spike protein sequence – NCBI Protein id - **QHD43416.1** from MN908947.3 (Wuhan-1 strain)

Software for the session:

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

Step 1 – Downloading related sequences

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 | MN908947 | Search

GenBank | Submit | Genomes | WGS | Metagenomes | TPA | TSA | INSDC | 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.
- 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 ftp://ftp.ncbi.nlm.nih.gov/genbank.

GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and

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

Nucleotide | Nucleotide | Search | Advanced | Help

GenBank | Send to

Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

GenBank: MN908947.3

FASTA | Graphics

Go to:

```
LOCUS       MN908947             29903 bp ss-RNA     linear   VRL 18-MAR-2020
DEFINITION  Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,
            complete genome.
ACCESSION   MN908947
VERSION     MN908947.3
KEYWORDS    .
SOURCE      Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
            Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae;
            Betacoronavirus; Sarbecovirus.
REFERENCE   1  (bases 1 to 29903)
  AUTHORS   Wu,F., Zhao,S., Yu,B., Chen,Y.M., Wang,W., Song,Z.G., Hu,Y.,
            Tao,Z.W., Tian,J.H., Pei,Y.Y., Yuan,M.L., Zhang,Y.L., Dai,F.H.,
            Liu,Y., Wang,Q.M., Zheng,J.J., Xu,L., Holmes,E.C. and Zhang,Y.Z.
  TITLE     A new coronavirus associated with human respiratory disease in
            China
  JOURNAL   Nature 579 (7798), 265-269 (2020)
   PUBMED   32015508
REFERENCE   2  (bases 1 to 29903)
```

Change region shown

Customize view

Analyze this sequence

- Run BLAST
- Pick Primers
- Highlight Sequence Features
- Find in this Sequence

NCBI Virus

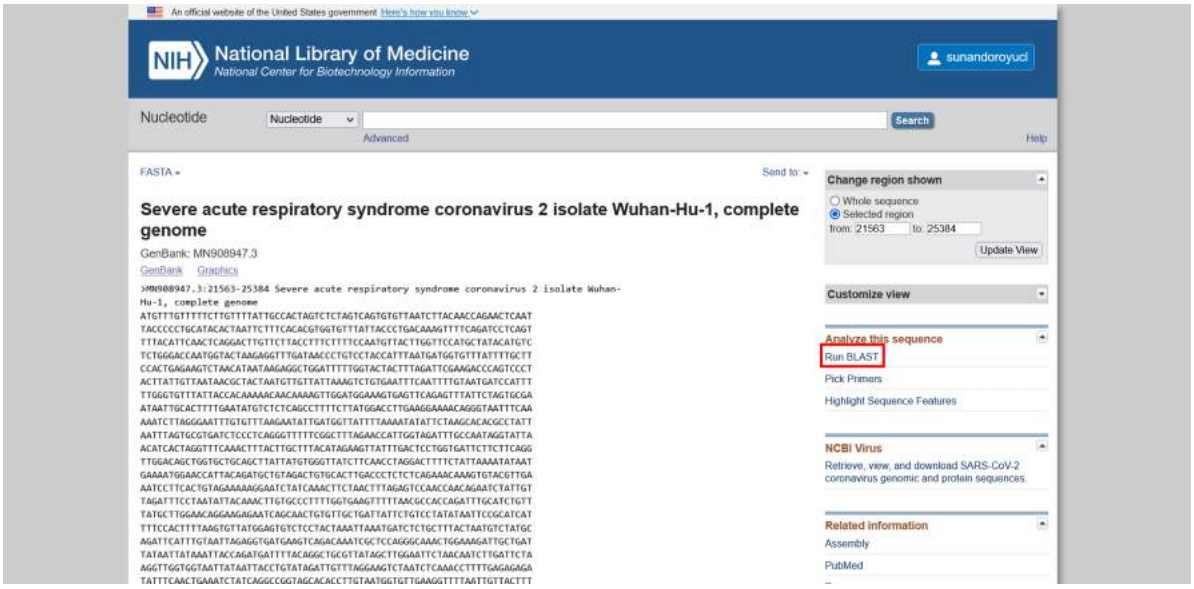
Retrieve, view, and download SARS-CoV-2 coronavirus genomic and protein sequences.

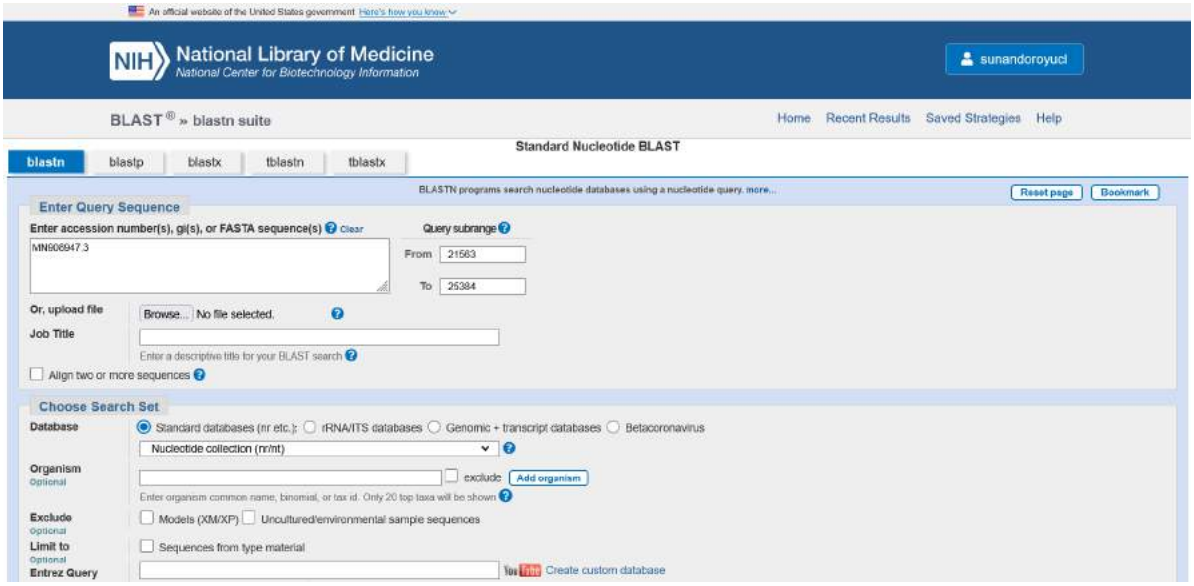
Related information

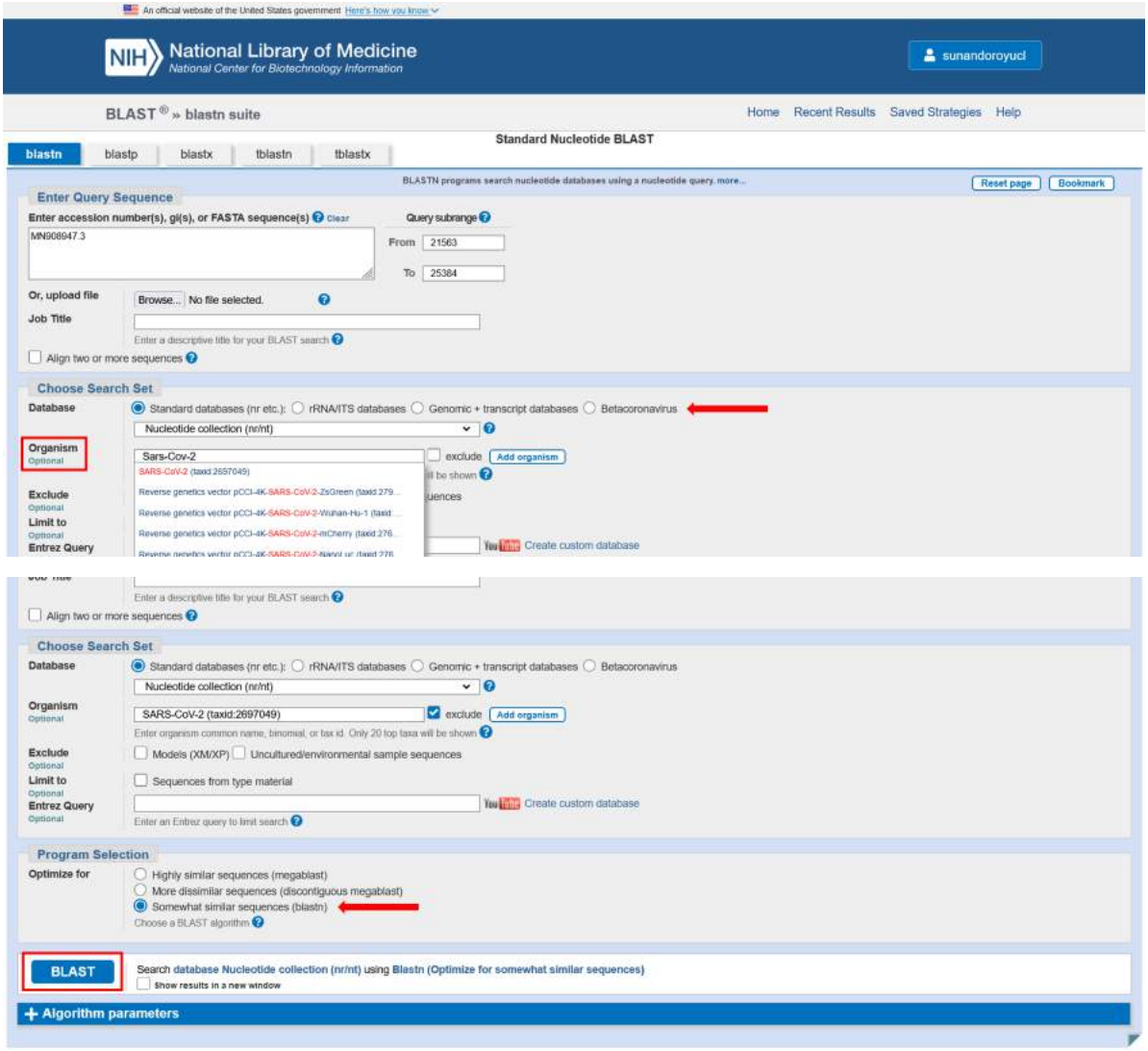
- Assembly
- PubMed
- Taxonomy
- Full text in PMC
- Gene

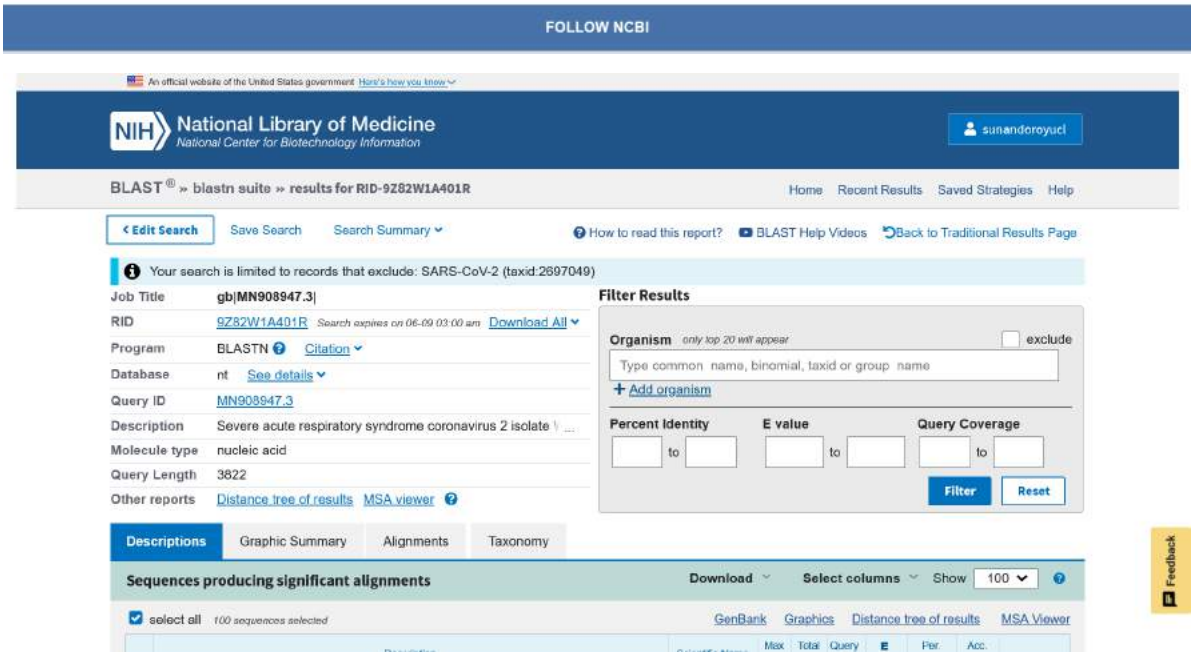
```
                     PRSQMEIDFLELAMDEFIERYKLEGYAFEHIVYGDFSHSQLGGLHLLIGLAKRFKESP
                     FELEDFIPMDSTVKNYFITDAQTGSSKCVCSVIDLLLDDFVEIIKSQDLSVVSKVVKV
                     TIDYTEISFMLWCKDGHVETFYPKLQSSQAWQPGVAMPNLYKNQRMLLEKCDLQNYGD
                     SATLPKGIMMNVAKYTQLCQYLNTLTLAVPYNMRVIHFGAGSDKGVAPGTAVLRQWLP
                     TGTLLVDSDLNDFVSDADSTLIGDCATVHTANKWDLIISDMYDPKTKNVTKENDSKEG
                     FFTYICGFIQQKLALGGSVAIKITEHSWNADLYKLMGHFAWWTAFVTNVNASSSEAFL
                     IGCNYLGKPREQIDGYVMHANYIFWRNTNPIQLSSYSLFDMSKFPLKLRGTAVMSLKE
                     GQINDMILSLLSKGRLIIRENNRVVISSDVLVNN"
gene                 21563..25384
                     /gene="S"
CDS                  21563..25384
                     /gene="S"
                     /note="structural protein"
                     /codon_start=1
                     /product="surface glycoprotein"
                     /protein_id="QHD43416.1"
                     /translation="MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFR
                     SSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIR
                     GWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVY
                     SSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQ
                     GFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFL
                     LKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITN
                     LCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
                     TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
                     YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY
                     RVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG
                     RDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAI
                     HADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPR
                     RARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTM
                     YICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFG
                     GFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFN
                     GLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQN
                     VLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGA
                     ISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMS
                     ECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
                     FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELD
                     SFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELG
                     KYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSE
                     PVLKGVKLHYT"
gene                 25393..26220
                     /gene="ORF3a"
CDS                  25393..26220
                     /gene="ORF3a"
```

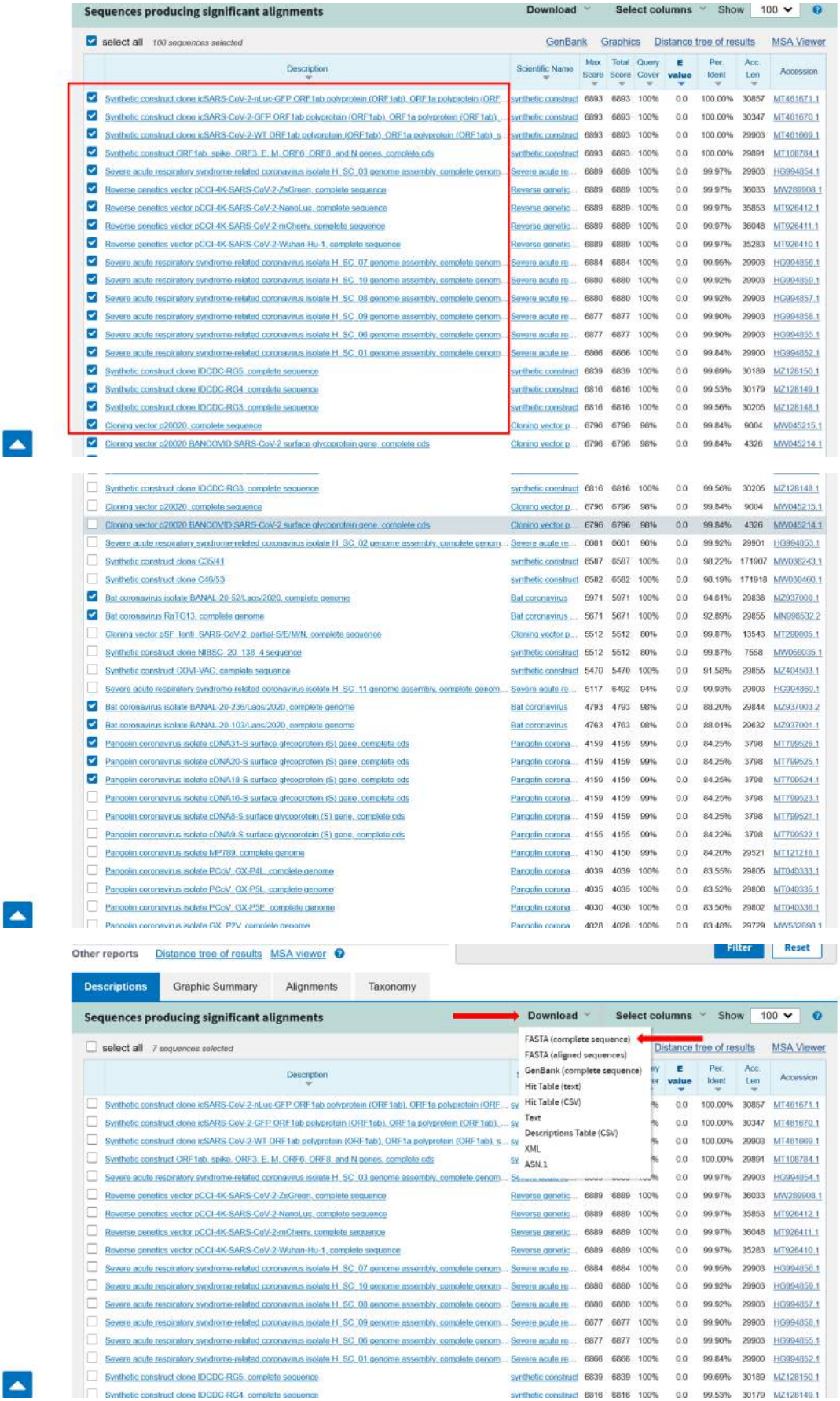

21563..25384
/gene="S"
/note="structural protein"
/codon_start=1
/product="surface glycoprotein"
/protein_id="
/translation="
CDS
Feature
2 of 10
MN908947 : 1 segment
Details
Display
FASTA
GenBank
Help


An official website of the United States government
National Library of Medicine
National Center for Biotechnology Information
sunandoroyucl
Nucleotide
Search
Advanced
Help
FASTA
Send to
Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
GenBank: MN908947.3
GenBank
Graphics
>MN908947.3:21563-25384 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
Change region shown
Whole sequence
Selected region
from: 21563
to: 25384
Update View
Customize view
Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
NCBI Virus
Retrieve, view, and download SARS-CoV-2 coronavirus genomic and protein sequences.
Related information
Assembly
PubMed


An official website of the United States government
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
MN908947.3
From
21563
To
25384
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
Nucleotide collection (nr/nt)
Organism
Optional
exclude
Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Create custom database

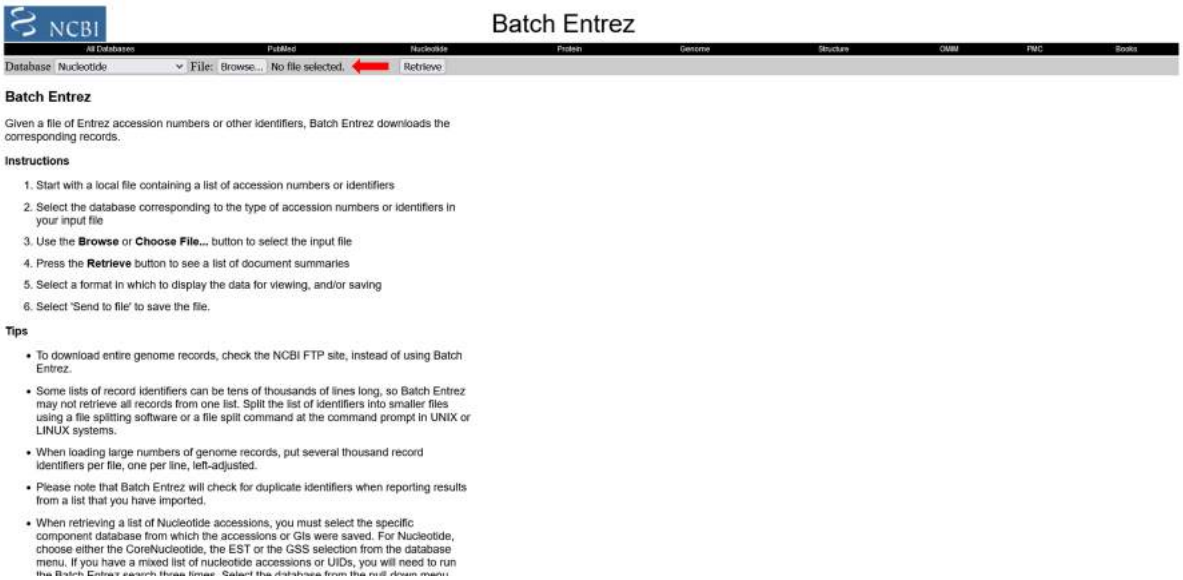

An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
sunandoroyucl
BLAST® » blastn suite
Home Recent Results Saved Strategies Help
Standard Nucleotide BLAST
blastn blastp blastx tblastn tblastx
BLASTN programs 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
MN908947.3
From 21563
To 25384
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
Nucleotide collection (nr/nt)
Organism
Optional
Sars-Cov-2
exclude Add organism
SARS-CoV-2 (taxid:2697049)
Exclude
Optional
Limit to
Optional
Entrez Query
Create custom database
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
Nucleotide collection (nr/nt)
Organism
Optional
SARS-CoV-2 (taxid:2697049)
exclude Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown
Exclude
Optional
Models (XM/XP) Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Optional
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 Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)
Show results in a new window
+ Algorithm parameters


FOLLOW NCBI
An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
sunandoroyucl
BLAST® » blastn suite » results for RID-9Z82W1A401R
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
Your search is limited to records that exclude: SARS-CoV-2 (taxid:2697049)
Job Title gb|MN908947.3|
RID 9Z82W1A401R Search expires on 06-09 03:00 am Download All
Program BLASTN Citation
Database nt See details
Query ID MN908947.3
Description Severe acute respiratory syndrome coronavirus 2 isolate
Molecule type nucleic acid
Query Length 3822
Other reports Distance tree of results MSA viewer
Filter Results
Organism only top 20 will appear
exclude
Type common name, binomial, taxid or group name
+ Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Select columns
Show 100
select all 100 sequences selected
GenBank Graphics Distance tree of results MSA Viewer
Feedback



Alternate Ways to Download from GenBank

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



Note: If you have Accession from different databases you will have to run Batch Entrez multiple times each for a unique sequence database

File to Use ~/Sunando/Betacorona_Accession.txt

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

```
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?
db=nucleotide&id=AY278488, AY304486, MN908947,
MT782115&rettype=fasta&retmode=text
```

Command Line

```
$ esearch -db "protein" -query "txid11270[Organism] AND L Protein Complete AND
refseq[filter]" | efetch -format fasta > outputfile.fasta

$ head outputfile.fasta
```

For the next step we will start with ~/Sunando/Spike.fas

Step 2 – Aligning Sequences.

Software Used

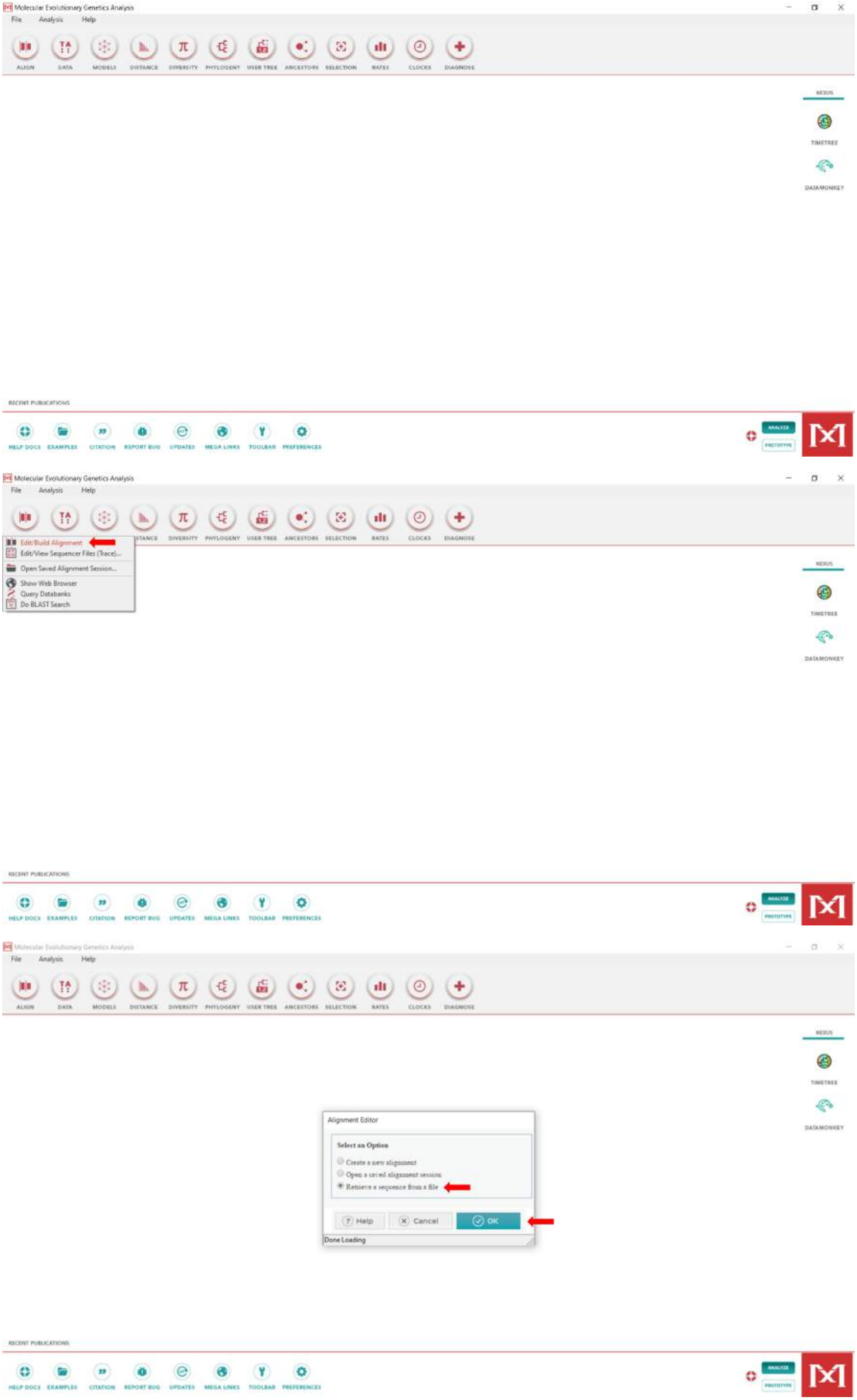
Mafft (https://mafft.cbrc.jp/alignment/software/)

MEGA (http://doua.prabi.fr/software/seaview)

Alternate Software – MUSCLE, CLUSTALW

To view/edit sequence files we will use MEGA. MEGA has a GUI and will launch as a standalone program

```
$ 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
Edit/Build Alignment
Edit/View Sequencer Files (Trace)...
Open Saved Alignment Session...
Show Web Browser
Query Databanks
Do BLAST Search
Alignment Editor
Select an Option
Create a new alignment
Open a saved alignment session
Retrieve a sequence from a file
Help
Cancel
OK
Done Loading

Picture40

```
$ mafft ~/Sunando/Spike.fas > ~/Sunando/outputfile.fas
```

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

```
$ mafft --maxiterate 1000 --localpair ~/Sunando/Spike.fas > outputfilelinsi.fas
```

The final alignment file is ~/Sunando/Spike_aln.fas

You can also use the alignment file you have generated.

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.

Model Testing

To run model testing we will use Modeltest-ng

```
$ modeltest-ng -d aa -i ~/Sunando/Spike_aln.fas -o modeloutputfile -t ml -p 2
```

-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 ~/Sunando/Spike_model.out and ~/Sunando/Spike_model.log

```
$ nano ~/Sunando/Spike_model.out
$ nano ~/Sunando/Spike_model.log
```

```
AICc      model              K            lnL          score          delta    weight
-----------------------------------------------------------------------------------
       1  TVM+I+G4           9    -89441.2343    179094.4687         0.0000    0.5632
       2  GTR+I+G4          10    -89440.4884    179094.9768         0.5081    0.4368
       3  TVM+G4             8    -89467.5351    179145.0703        50.6016    0.0000
       4  GTR+G4             9    -89467.0653    179146.1306        51.6620    0.0000
       5  TPM3uf+I+G4        7    -89514.2112    179236.4224       141.9537    0.0000
       6  TIM3+I+G4          8    -89514.1910    179238.3821       143.9134    0.0000
       7  TPM2uf+I+G4        7    -89530.2483    179268.4965       174.0279    0.0000
       8  TIM2+I+G4          8    -89530.0605    179270.1211       175.6524    0.0000
       9  TPM3uf+G4          6    -89541.5959    179289.1919       194.7232    0.0000
      10  TIM3+G4            7    -89541.6048    179291.2095       196.7409    0.0000
-----------------------------------------------------------------------------------
Best model according to AICc
----------------------------
Model:              TVM+I+G4
lnL:                -89441.2343
Frequencies:        0.2626 0.1931 0.1752 0.3691
Subst. Rates:       2.6387 5.3471 1.7471 1.9174 5.3471 1.0000
Inv. sites prop:    0.0673
Gamma shape:        0.9500
Score:              179094.4687
Weight:             0.5632
```

The model that we will use for tree building is TVM+I+G4

Tree building

To build a tree we are going to use IQ-TREE

```
$ iqtree -s ~/Sunando/Spike_aln.fas -bb 1000 -st DNA -nt 4 -alrt 1000 -pre
treeoutfile
```

-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 we will take forward to the next step while IQ-TREE completes running will be ~/Sunando/Spike_Tree.contree

Step 4 – Viewing and Modifying a Tree File.

Software Used

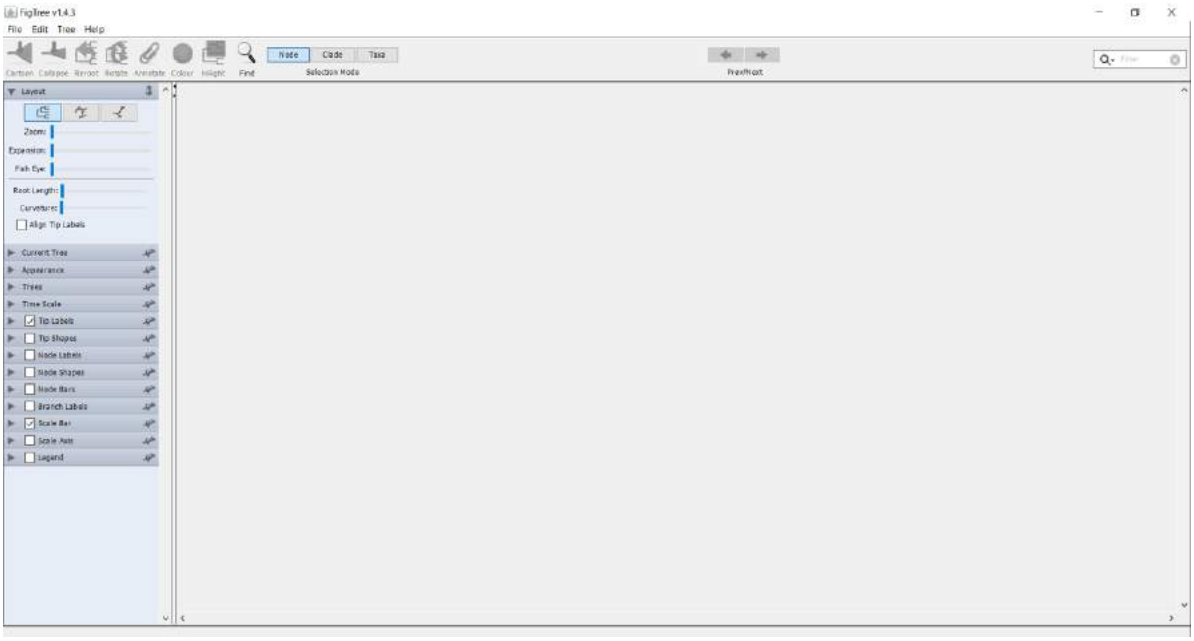
FigTree (http://tree.bio.ed.ac.uk/software/figtree/)

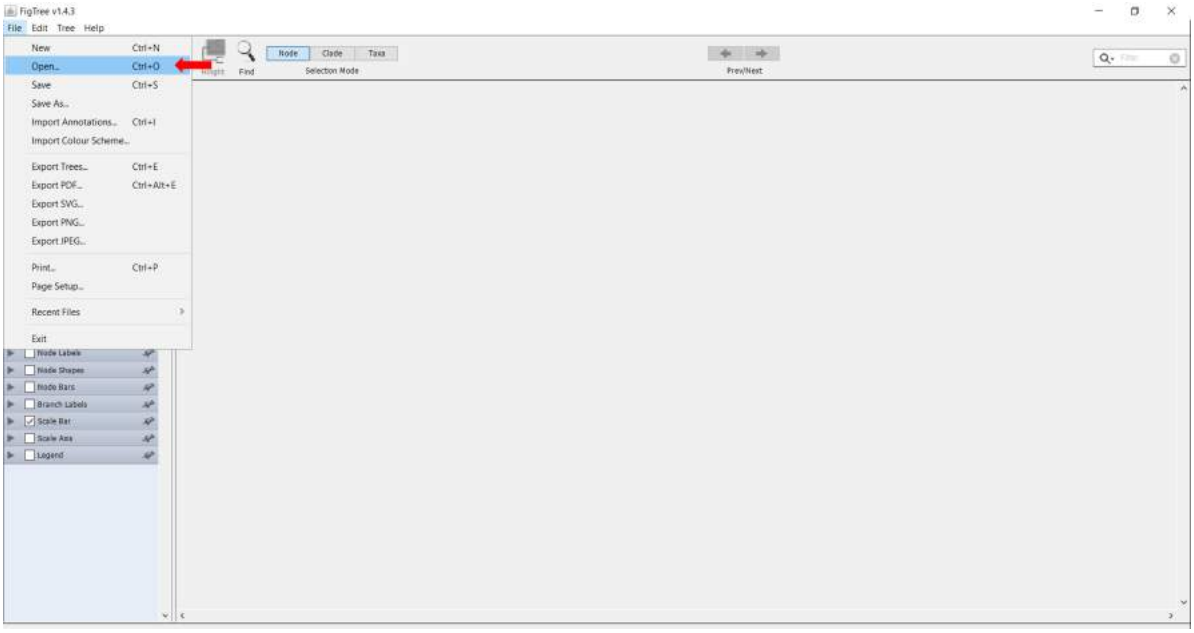
MEGA (https://www.megasoftware.net/)

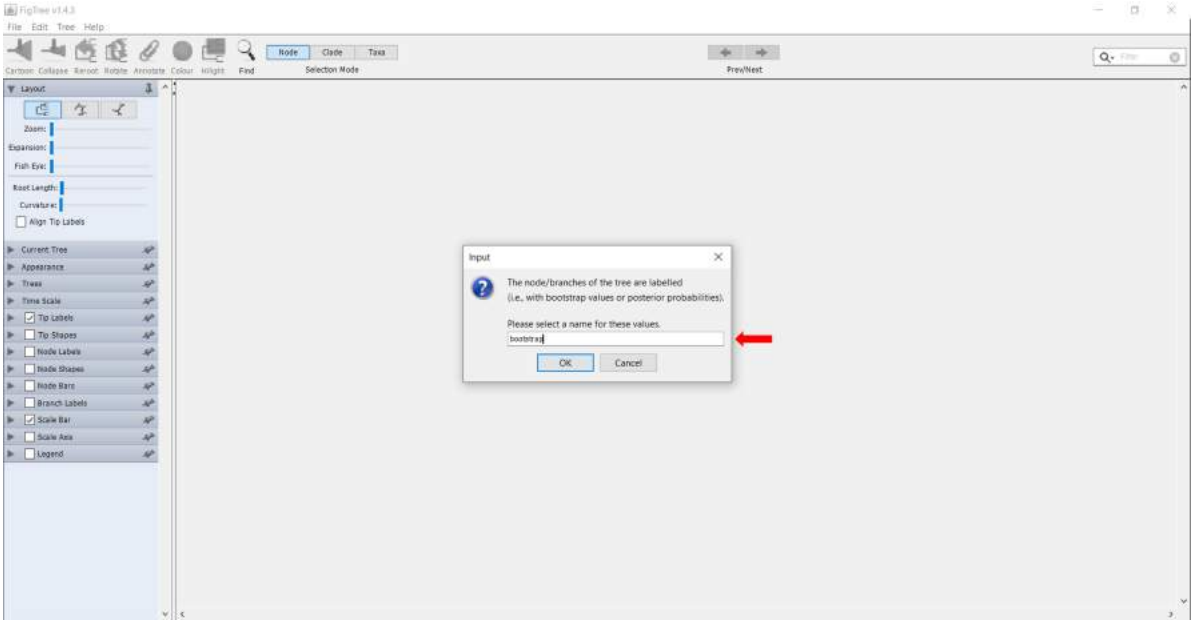
```
$ figtree
```

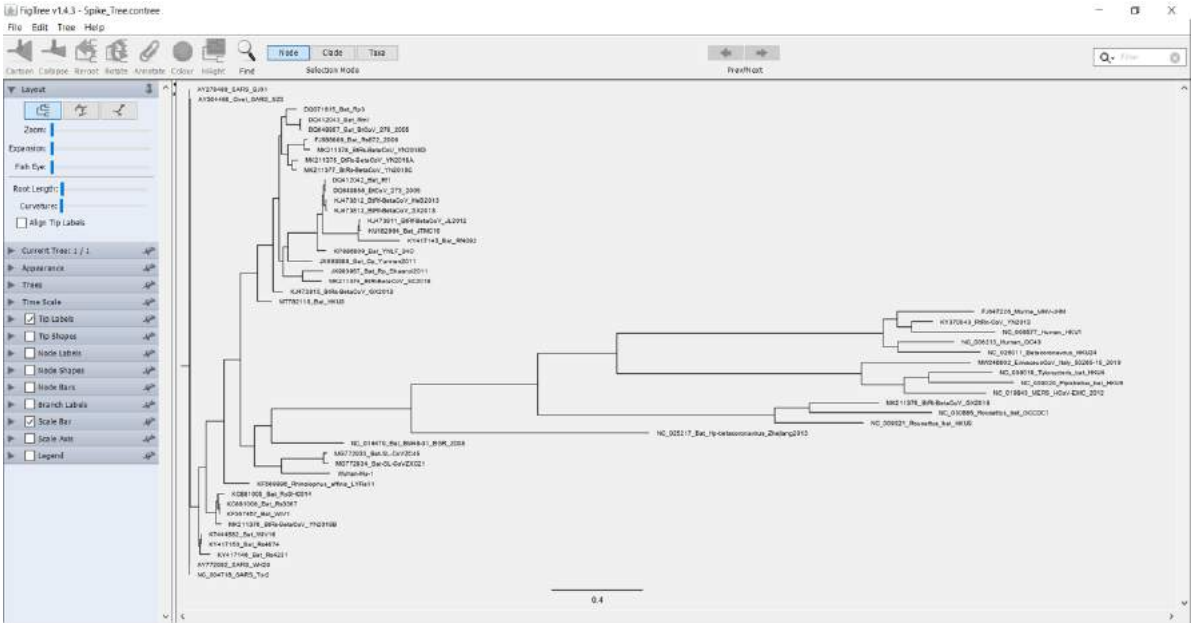
Figtree is Java based and will launch a GUI

Open file - ~/Sunando/Spike_Tree.contree

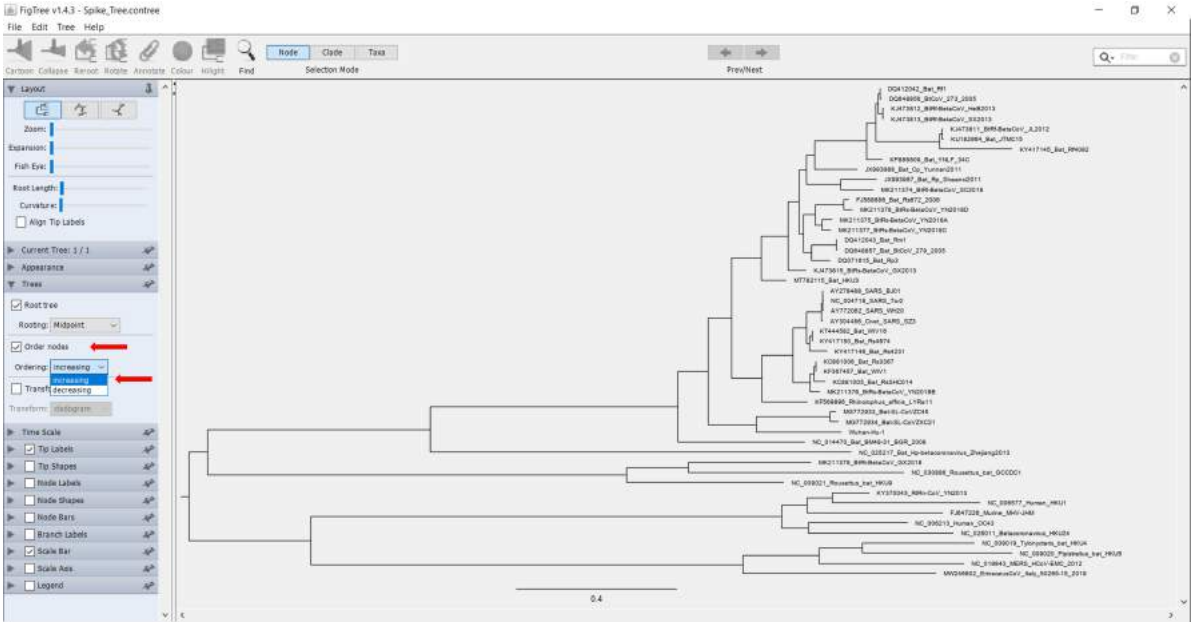




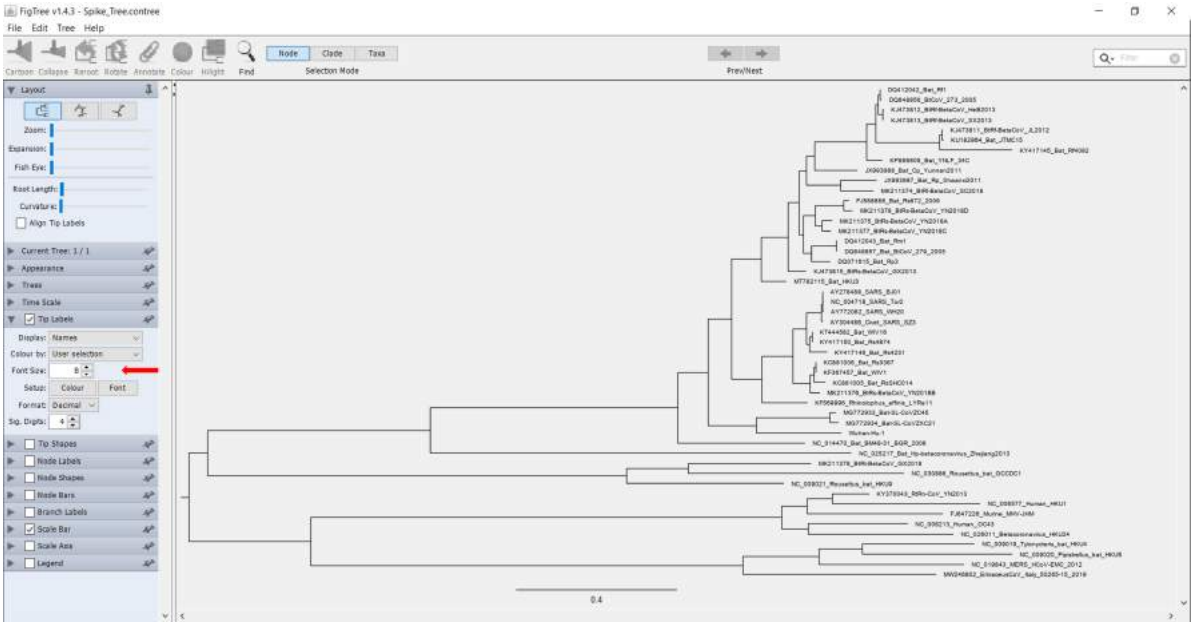




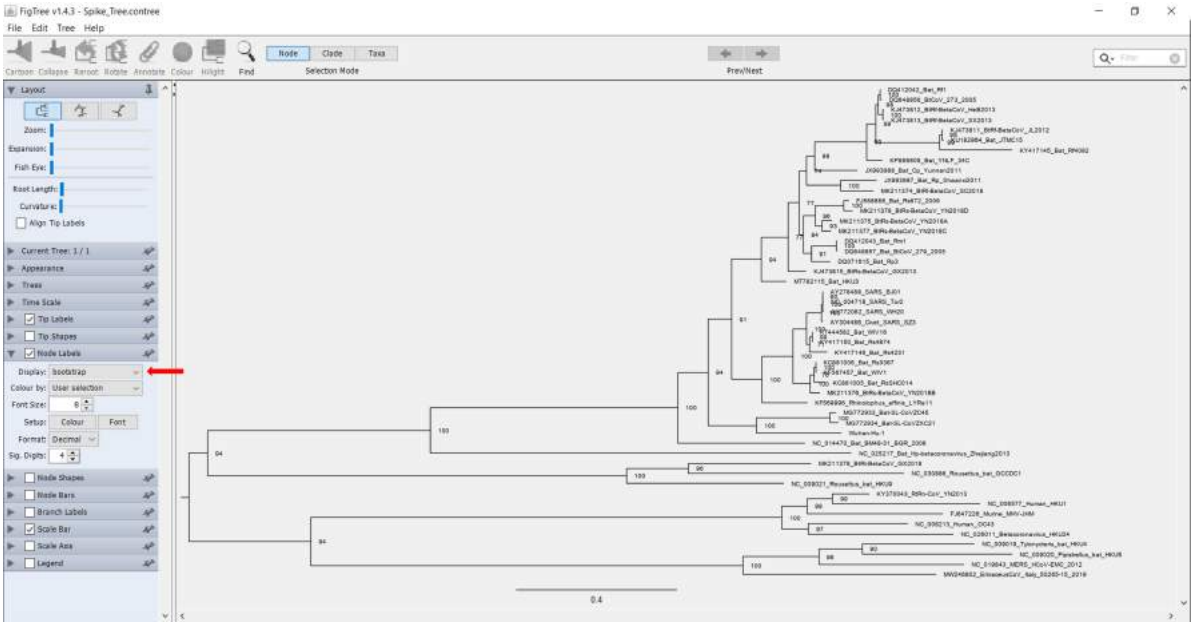
Rooting a Tree



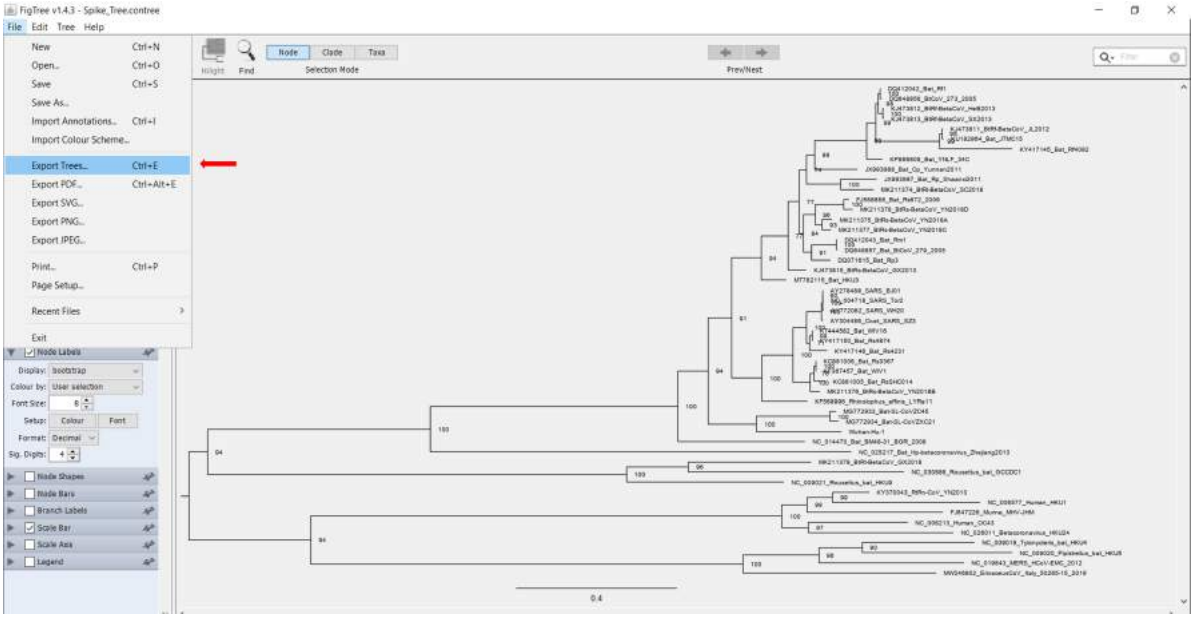
Modify Labels



Add bootstrap support



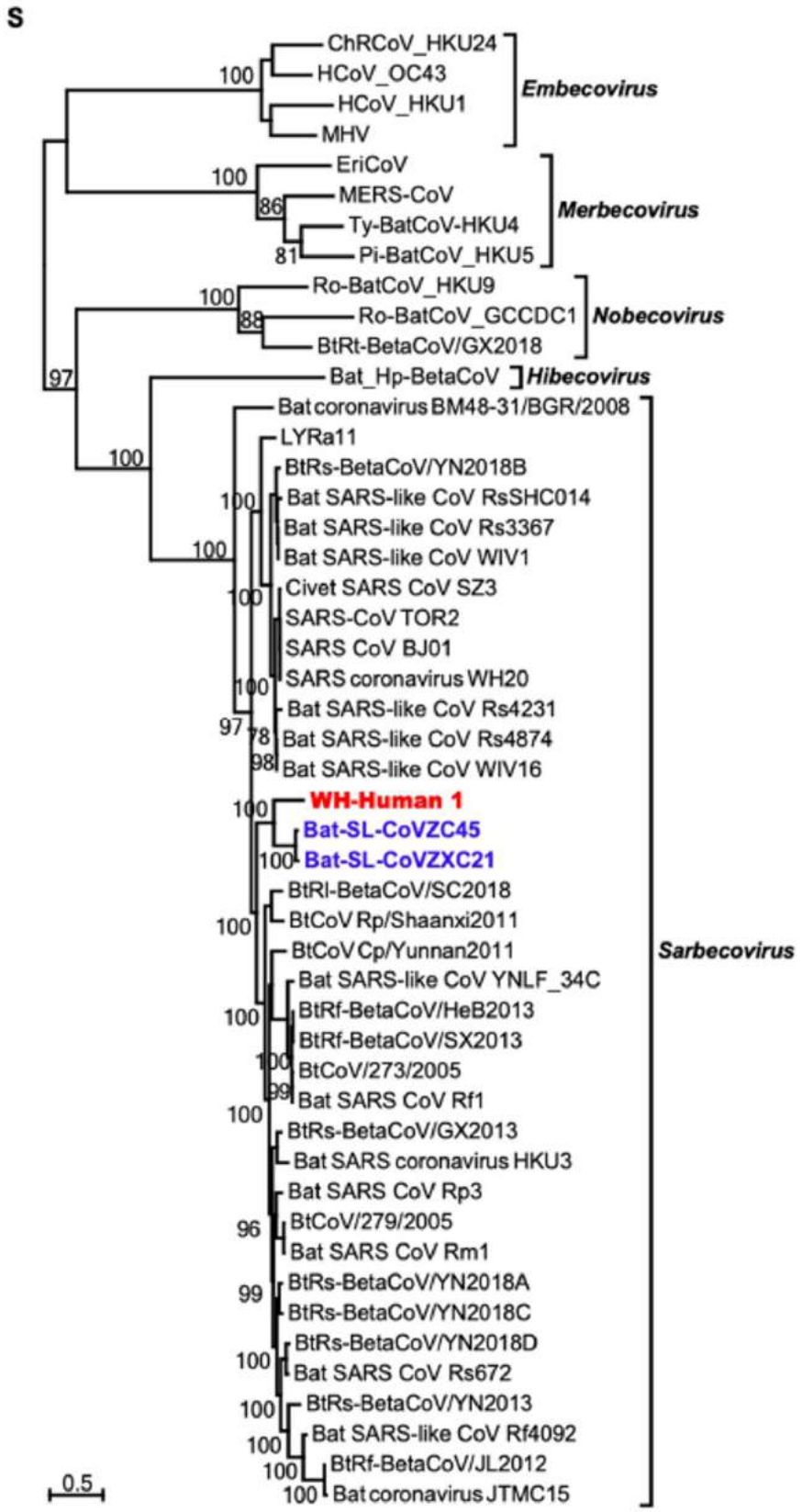
Export Trees



We can go through similar tree viewing in MEGA

Final Results

We further process the trees in **PowerPoint/ggtree** or other image editing tools to add metadata for each sequence in the alignment. In this case we have added virus groups for each virus in the alignment.



Based on this data one would infer that the isolated virus most likely lies within the Sarbecovirus group which includes the original SARS coronavirus but is quite distinct from it. The closest related species both are from bats which could suggest a potential origin.