

Sequence Alignments and Phylogeny

Genomics and Clinical Virology

Sunando Roy

sunando.roy@ucl.ac.uk

30/06/2022

Scenario: You have determined the pathogen responsible for the outbreak above and now wish to identify a likely source of infection. Consensus sequence data from bats, rodents, humans and reference sequences are available for further analysis. Using phylogenetic analysis, what is the likely source of the infection?

Outline

- Sequence retrieval from GenBank
- Multiple Sequence Alignments
- Model Testing and Maximum Likelihood based tree building
- Viewing and modifying a Tree file

Software

1.Mafft - Alignment tool

2.MEGA/SeaVIEW- Alignment Viewer and Editor

3.Modeltest-ng - Model Testing

4.IQ-TREE - Tree Building tool

5.MEGA/Figtree - Tree Viewer and Editor

Sequence Retrieval

- **Where?**

- GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>)
- ENA (<https://www.ebi.ac.uk/ena>)
- DDBJ (<http://www.ddbj.nig.ac.jp/>)

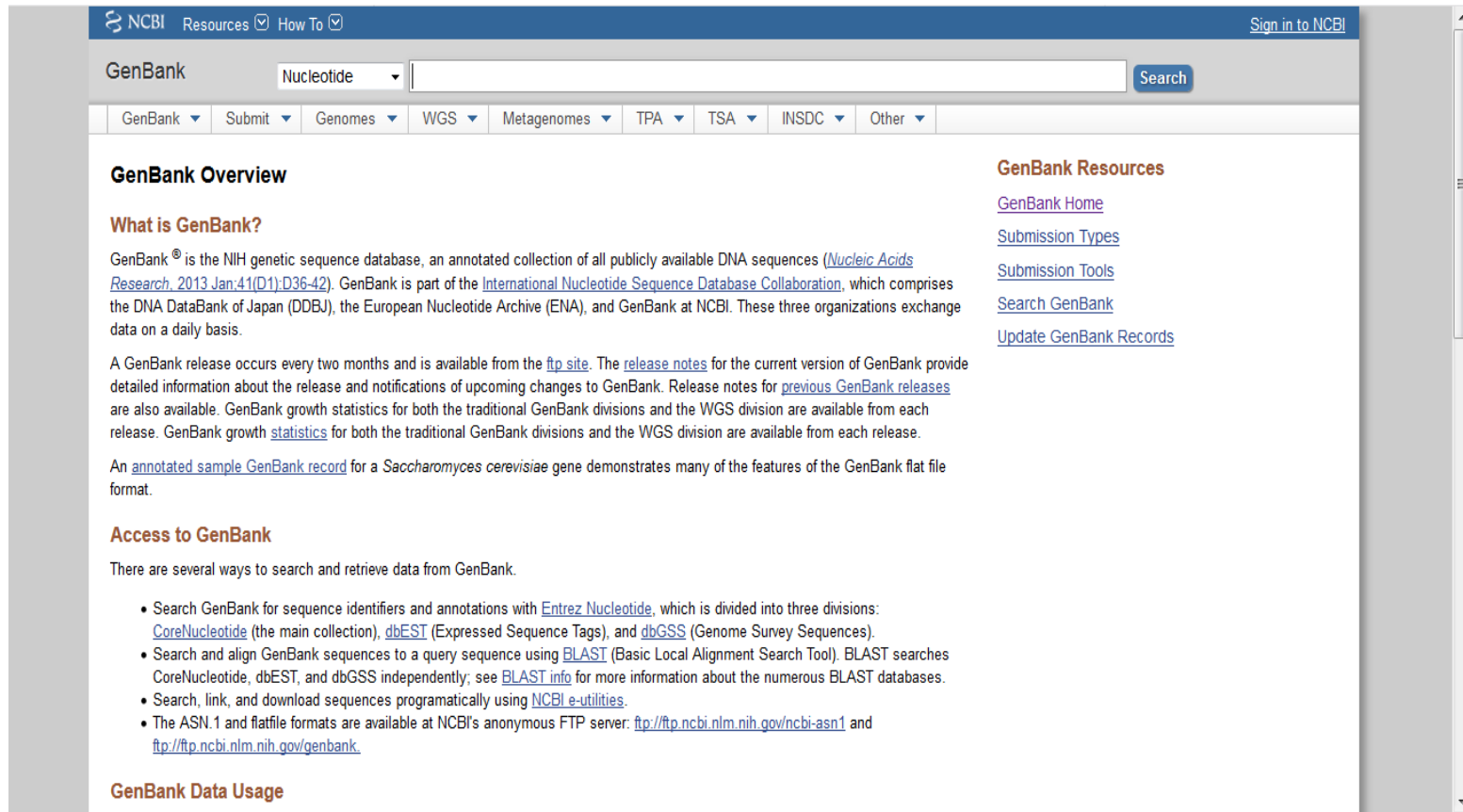
For all other databases

(https://en.wikipedia.org/wiki/List_of_biological_databases)

There are now pathogen specific databases like
GISAID(Influenza/SARS-CoV-2) and NoroNet (Norovirus)

- How?

- GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>)



The screenshot shows the NCBI GenBank homepage. At the top, there is a navigation bar with 'NCBI', 'Resources', and 'How To' links, along with a 'Sign in to NCBI' button. Below this is a search bar with 'GenBank' selected and a 'Nucleotide' dropdown menu. A 'Search' button is to the right of the search bar. Below the search bar is a horizontal menu with various database categories: GenBank, Submit, Genomes, WGS, Metagenomes, TPA, TSA, INSDC, and Other. The main content area is divided into two columns. The left column is titled 'GenBank Overview' and contains a section 'What is GenBank?' which describes the database as a collection of publicly available DNA sequences. It also mentions the 'GenBank release' and provides links to 'release notes' and 'previous GenBank releases'. The right column is titled 'GenBank Resources' and contains links to 'GenBank Home', 'Submission Types', 'Submission Tools', 'Search GenBank', and 'Update GenBank Records'. At the bottom of the left column, there is a section 'Access to GenBank' which lists several ways to search and retrieve data, including using 'Entrez Nucleotide', 'BLAST', and 'NCBI e-utils'. The bottom of the page features a 'GenBank Data Usage' section.

NCBI Resources How To Sign in to NCBI

GenBank Nucleotide 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. 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](#), which is divided into three divisions: [CoreNucleotide](#) (the main collection), [dbEST](#) (Expressed Sequence Tags), and [dbGSS](#) (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using [NCBI e-utils](#).
- 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

GenBank Resources

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

• Alternatives

- Batch Entrez

(<https://www.ncbi.nlm.nih.gov/sites/batchentrez>)

The screenshot shows the NCBI Batch Entrez web interface. At the top is the NCBI logo and the title "Batch Entrez". Below this is a navigation bar with tabs for "All Databases", "PubMed", "Nucleotide", "Protein", "Genome", "Structure", "OMIM", "PMC", and "Books". The "Nucleotide" tab is selected. Below the navigation bar is a form with a "Database" dropdown menu set to "Nucleotide", a "File:" label, a "Browse..." button, and the text "No file selected.". To the right of the "Browse..." button is a "Retrieve" button. Below the form is a section titled "Batch Entrez" with a paragraph explaining its function: "Given a file of Entrez accession numbers or other identifiers, Batch Entrez downloads the corresponding records." Below this is a section titled "Instructions" with a numbered list of six steps: 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. Below the instructions is a section titled "Tips" with a bulleted list of three 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

- **Alternatives**

- Entrez E-utilities

[\(ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/\)](ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/)

- Command – Browser example

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

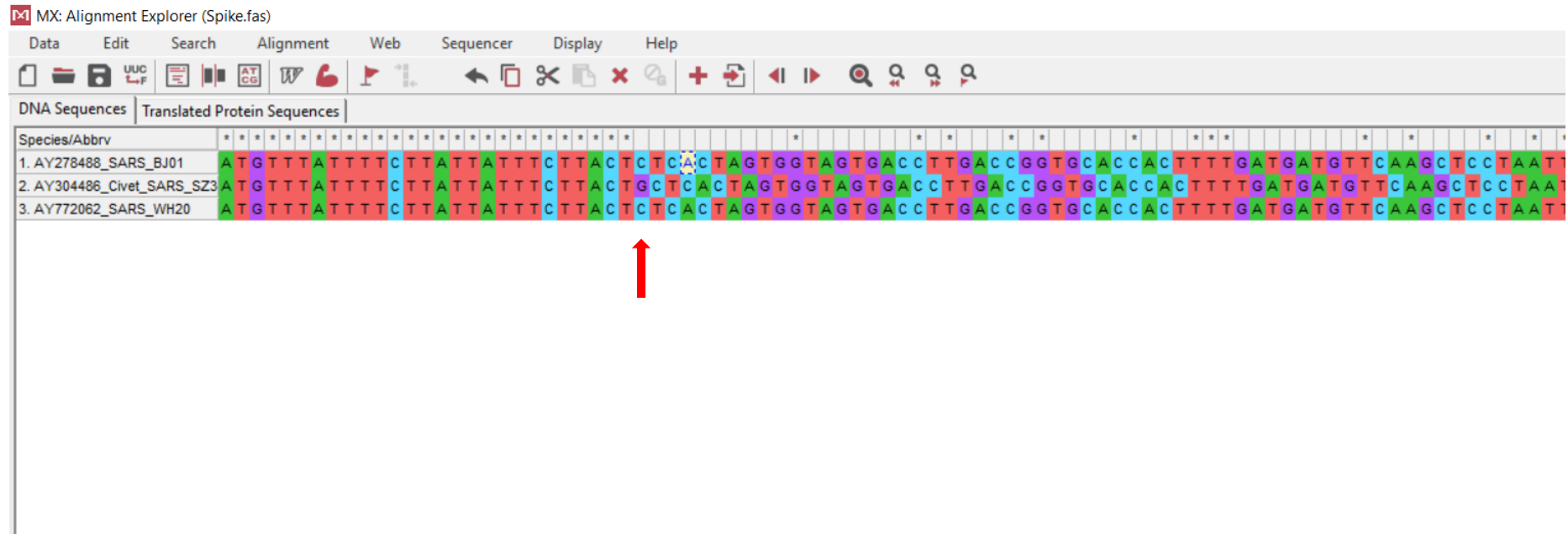
- Command – Terminal example

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

Multiple Sequence Alignments

- **Why?**

- Necessary for every sequence analysis





DNA Sequences

Translated Protein Sequences

Species/Abbrv	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
---------------	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---



MX: Alignment Explorer (Spike.fas)

Data Edit Search Alignment Web Sequencer Display Help

Translated Protein Sequences

Species/Abbrv

1. AY278486_SARS_BJ01
2. AY304486_Civet_SARS_SZ3
3. AY772062_SARS_WH20
4. DQ071615_Bat_Rp3
5. DQ412042_Bat_Rf1
6. DQ412043_Bat_Rm1
7. DQ648856_Bat_BiCoV_273_2005
8. DQ648857_Bat_BiCoV_279_2005
9. FJ588686_Bat_Rs672_2006
10. FJ647226_Murine_MHV-JHM
11. JX993987_Bat_Rp_Shaanxi2011
12. JX993988_Bat_Cp_Yunnan2011
13. KC881005_Bat_RsSHC014
14. KC881006_Bat_Rs3367
15. KF367457_Bat_WIV1
16. KF569996_Rhinolophus_affinis_LYRa11
17. KJ473811_BtRf-BetaCoV_JL2012
18. KJ473812_BtRf-BetaCoV_HeB52013
19. KJ473813_BtRf-BetaCoV_SX2013
20. KJ473815_BtRs-BetaCoV_GX2013
21. KP886809_Bat_YNLF_34C
22. KT444582_Bat_WIV16
23. KU182964_Bat_JTMC15
24. KY370043_RtRn-CoV_YN2013
25. KY417145_Bat_Rf4092
26. KY417146_Bat_Rs4231
27. KY417150_Bat_Rs4874
28. MG772933_Bat-SL-CoVZC45
29. MG772934_Bat-SL-CoVZXC21
30. MK211374_BtRf-BetaCoV_SC2018
31. MK211375_BtRs-BetaCoV_YN2018A
32. MK211376_BtRs-BetaCoV_YN2018B
33. MK211377_BtRs-BetaCoV_YN2018C
34. MK211378_BtRs-BetaCoV_YN2018D
35. MK211379_BtRf-BetaCoV_GX2018
36. Wuhan-Hu-1

Site # 1 with w/o gaps

- **How?**

- Maximize identity between sequences in your alignment.
- Uses scores for matches (+), mismatches (-), gap penalty (-).
- Changing scoring parameters change alignments.
- Visual inspection is always necessary.
- Bad alignment = Bad phylogenetic inferences.

Code: `$ mafft --maxiterate 1000 --localpair ~/Sunando/Spike.fas > outputfile.fas`

MAFFT version 7
Multiple alignment program for amino acid or nucleotide sequences

Download version
[Mac OS X](#)
[Windows](#)
[Linux](#)
[Source](#)

Online version
[Alignment](#)
[mafft --add](#)
[Merge](#)
[Phylogeny](#)
[Rough tree](#)
[Merits / limitations](#)
[Algorithms](#)
[Tips](#)
[Benchmarks](#)
[Feedback](#)

Follow

For a large number of short sequences, try [an experimental service](#) (2017/Jul).
This service will be unavailable for maintenance, Feb.10 6:00PM - Feb.11 (JST).

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:
Paste protein or DNA sequences in fasta format. [Example](#)

or upload a **plain text** file: No file selected.

☐ Use structural alignment(s)
☐ Allow unusual symbols (Selenocysteine "U", Inosine "I", non-alphabetical characters, etc.) [Help](#)

<https://mafft.cbrc.jp/alignment/server/>

Model Testing and Maximum Likelihood based tree building

- **Why?**

- To infer evolutionary relationships and identify novel pathogens.
- Identify geographical location for the source of infection.
- Identify potential host species.
- Infer time of transmission.

Tree Building

Seq 1 – ATTGCAAT

Seq 2 – ATTGCAAT

Seq 3 – TTTGCTAT

Seq 4 – TTTGCTAT

Seq 5 – ATTCCTAC

Tree Building

Seq 1 – ATTGCAAT

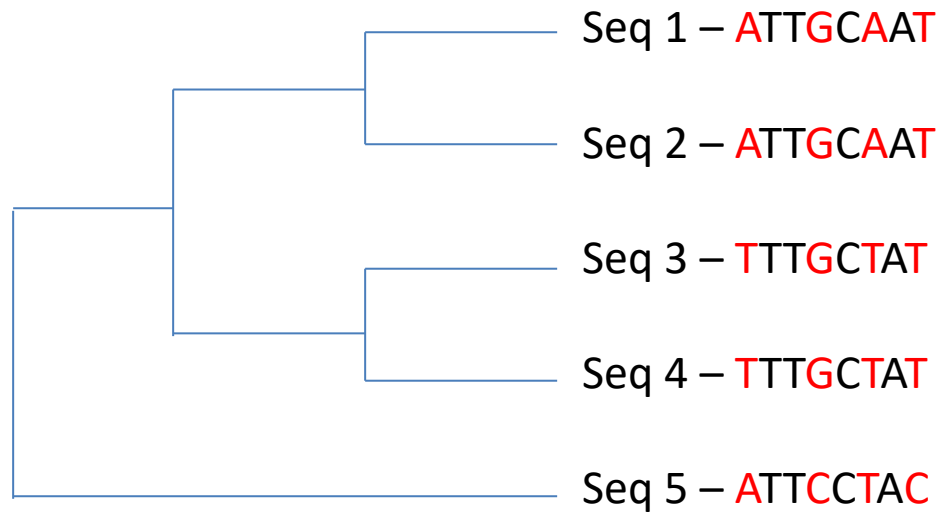
Seq 2 – ATTGCAAT

Seq 3 – TTTGCTAT

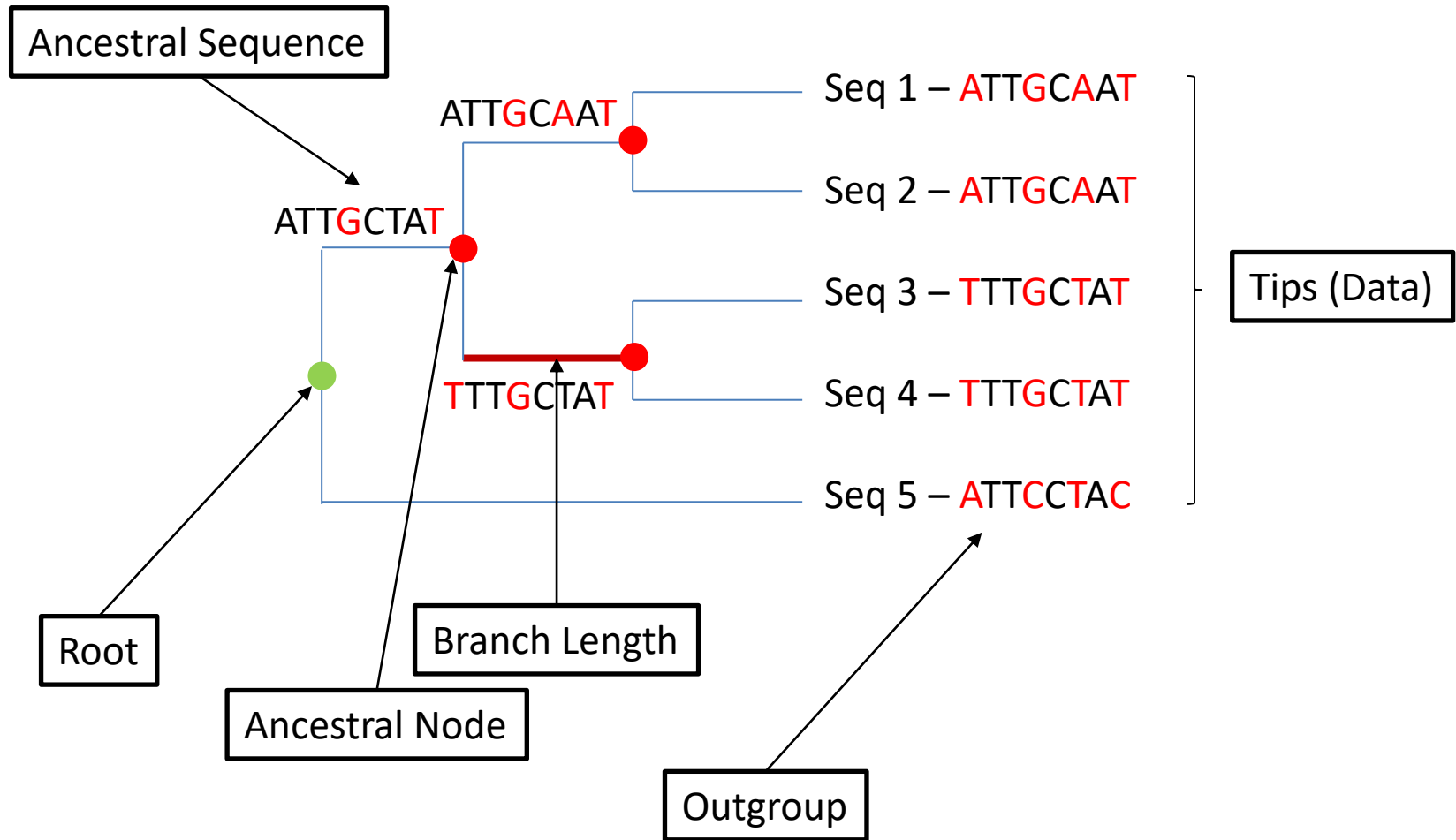
Seq 4 – TTTGCTAT

Seq 5 – ATTCCTAC

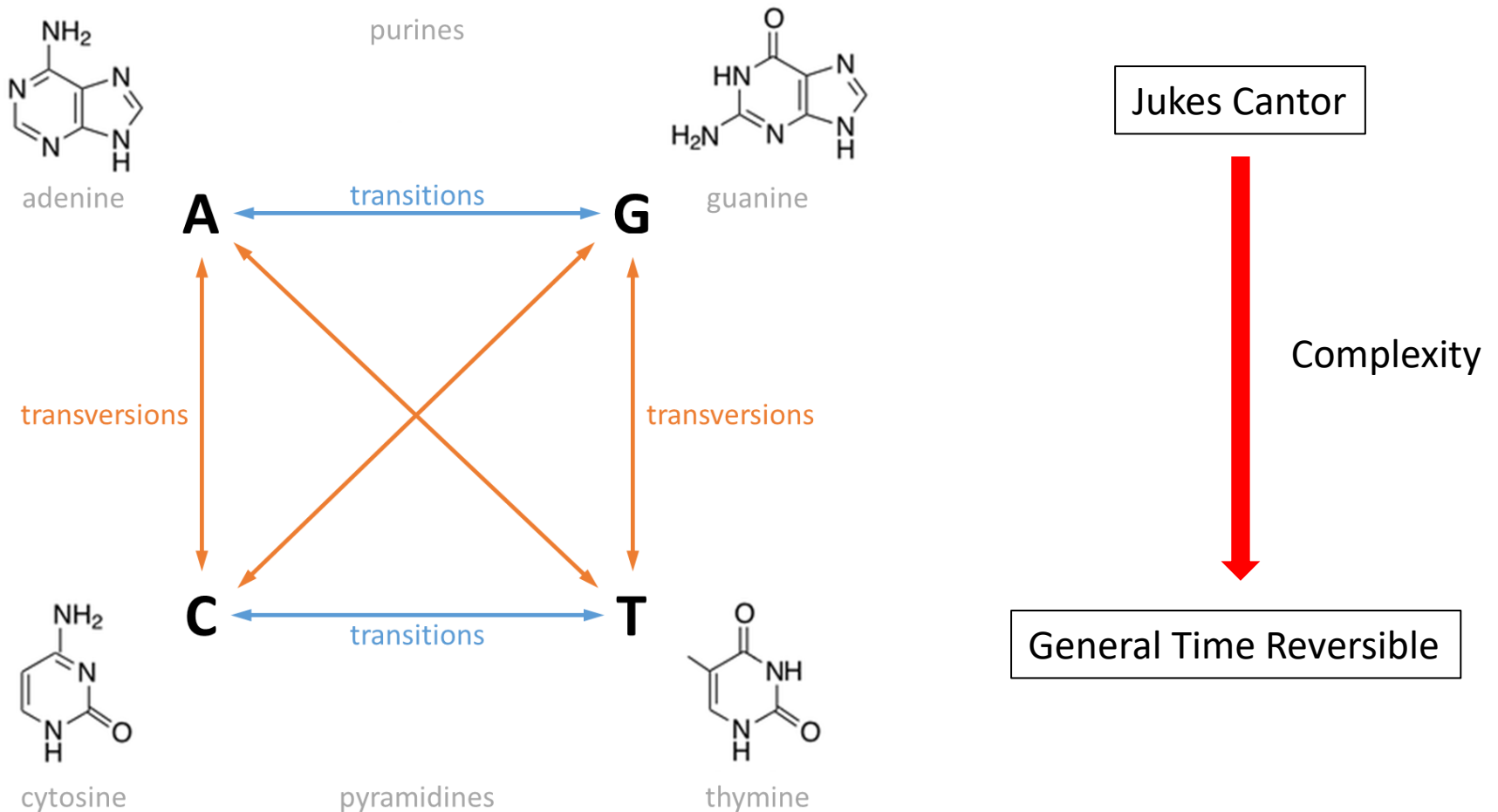
Tree Building



Tree Building



Evolutionary models - DNA



Evolutionary models - Proteins

													Model	Training data				References										
Ala	4												General models:															
Arg	-1	5											JTT	—				Jones <i>et al.</i> (1992)										
Asn	-2	0	6										LG	—				Le and Gascuel (2008)										
Asp	-2	-2	1	6									PAM (Dayhoff)	—				Dayhoff <i>et al.</i> (1978)										
Cys	0	-3	-3	-3	9								PMB	—				Veerassamy <i>et al.</i> (2003)										
Gln	-1	1	0	0	-3	5							VT	—				Müller and Vingron (2000)										
Glu	-1	0	0	2	-4	2	5						WAG	—				Whelan and Goldman (2001)										
Gly	0	-2	0	-1	-3	-2	-2	6					Specialized models:															
His	-2	0	1	-1	-3	0	0	-2	8				HIVb	HIV (eight proteins)				Nickle <i>et al.</i> (2007)										
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4			rtREV	retroelement <i>pol</i>				Dimmic <i>et al.</i> (2002)										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4																	
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5																
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5															
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6														
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7													
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4												
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5											
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11										
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7									
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4								
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val								

- Tree Building Algorithms
 - Neighbor Joining – BioNJ
 - Maximum Parsimony – MEGA
 - Maximum Likelihood – PhyML, RAxML, IQ-TREE
 - Bayesian Inference – Mr Bayes, BEAST

Neighbour Joining

- Estimates relationships based on a pairwise distance matrix.
- Distance matrix calculation does take into consideration evolutionary substitution models.
- Collapses closest distance pair into one taxa and repeats steps until all tips are clustered.
- Samples only one possible tree out of all possible outcomes.
- Fast but struggles in estimating relationships over longer evolutionary times.

Maximum Parsimony

- Character based tree estimation
- Evaluates multiple tree topologies.
- Scores the best tree on minimum number of character changes required to explain the data.
- Does not use evolutionary models of substitution.
- Does not perform well over longer evolutionary time scales.
- Suffers from effects of Long Branch Attraction.

Maximum Likelihood

- Calculates the probability of a tree topology at each individual site across the sequence and a final product across sites is computed.
- Can use independent rate measurement across each site.
- Evaluates multiple tree topologies using branch swaps, nearest neighbour interchange etc. to find trees with the best probability.
- The probability is presented as a log likelihood thus less negative the number the greater the probability

Bayesian Inference

- Trees built by estimating posterior probability from a set of user defined priors.
- Used in Phylodynamics and Phylogeography analysis.
- Computationally intensive.
- Sensitive to priors and evolutionary assumptions.

Bootstrap

- Start from a reference tree
- Alignment sites are sampled with replacement.
- Trees are built for each resampled dataset
- The frequency of each node occurring in the bootstrapped trees is computed
- Gives a statistical confidence value to each node.
- Bootstrap values of 70-75 is used as an indicator for good support.

- **Caveats**

- Model testing – Modeltest-ng
- Recombination Detection – GARD, Simplot

Code:

```
$ 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

Note:

- This will run with model testing included
- Output will have .treefile that has both the UF bootstrap and alrt values and the .contree consensus tree
- For UF bootstrap values >95 and aLRT values >80 considered as strong support
- Traditional bootstrap can be done using **-b** option
- Models can be specified using **-m** option

← → ↺ 🏠

🛡️ iqtree.cibiv.univie.ac.at ☆

📄 ⬇️ 📏 📄 📄 📄

IQ-TREE web server: fast and accurate phylogenetic trees under maximum likelihood

Server load: 33% Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) *Nucl. Acids Res.* 44 (W1): W232-W235. [doi: 10.1093/nar/gkw256](https://doi.org/10.1093/nar/gkw256)

Tree Inference | Model Selection | Analysis Results

For a quick start, take a look at the [tutorial](#) for the IQ-TREE web server.
Please visit the [IQ-TREE homepage](#) for more information or if you want to download the main software.
Data Privacy Statement: All your personal data are strictly confidential and will not be shared with any third parties. Your data will be automatically deleted after 180 days.

Input Data

Alignment file : [Browse...](#) [Show example >](#)

Use example alignment: ☐ Yes ?

Sequence type: ☒ Auto-detect ☐ DNA ☐ Protein ☐ Codon ?
☐ DNA->AA ☐ Binary ☐ Morphology

Partition file: This field is optional. [Browse...](#) [Show example >](#)

Partition type: ☒ Edge-linked ?
☐ Edge-unlinked

Substitution Model Options

Substitution model: Auto ?

FreeRate heterogeneity: ☐ Yes [+R] ?

Rate heterogeneity: ☐ Gamma [+G] ☐ Invar. sites [+I] ?

#rate categories: 4

State frequency: ☒ Empirical (from data) ☐ AA model (from matrix) ☐ ML-optimized
☐ Codon F1x4 ☐ Codon F3x4

Ascertainment bias correction: ☐ Yes [+ASC] ?

Branch Support Analysis

Bootstrap analysis: ☐ None ☒ Ultrafast ☐ Standard ?

Number of bootstrap alignments: 1000

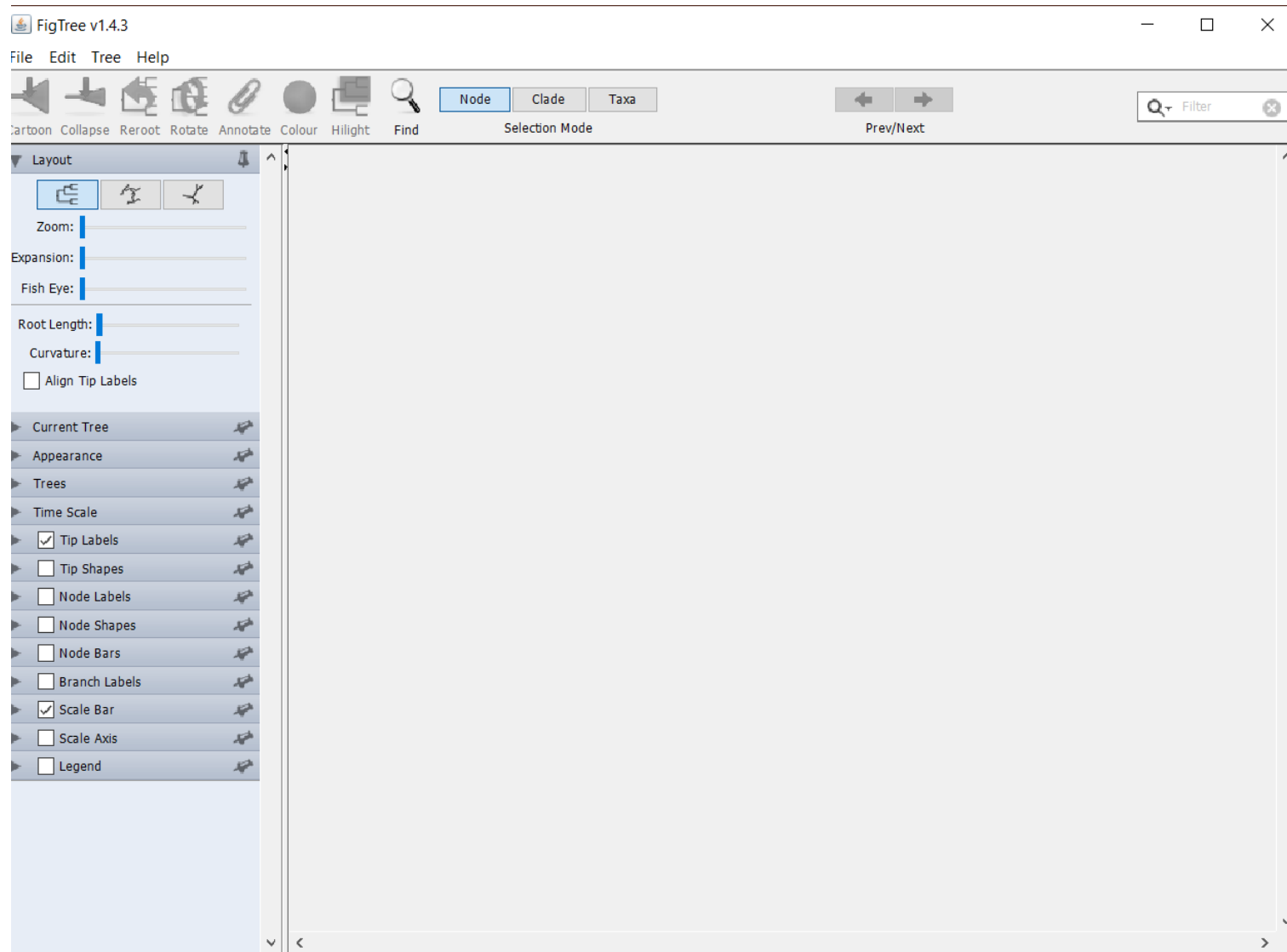
<http://iqtree.cibiv.univie.ac.at/>

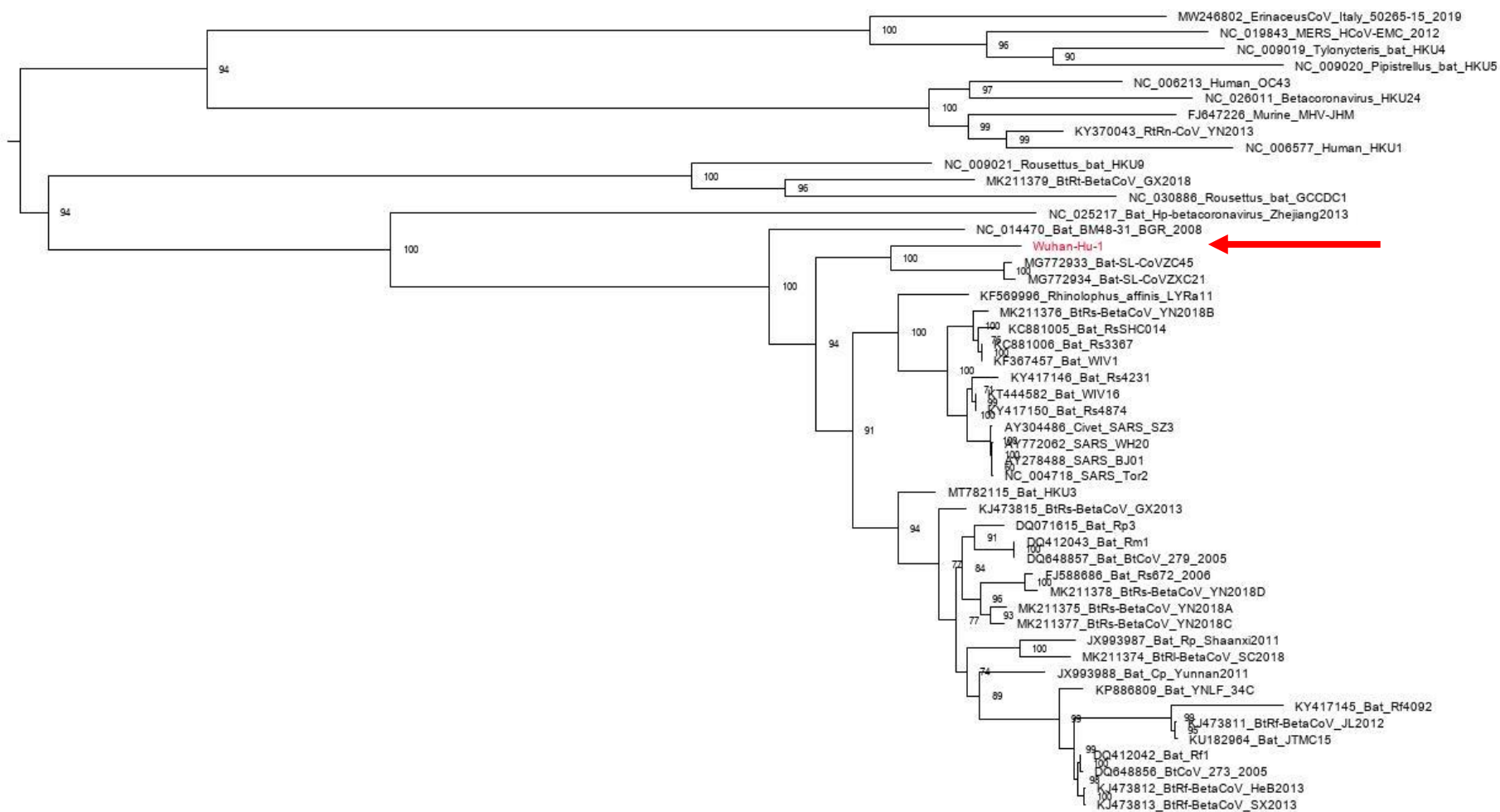
Viewing and modifying a Tree file

- **Why?**

- To visualize final phylogenetic relationships and draw inferences.
- To create final figures for publications.

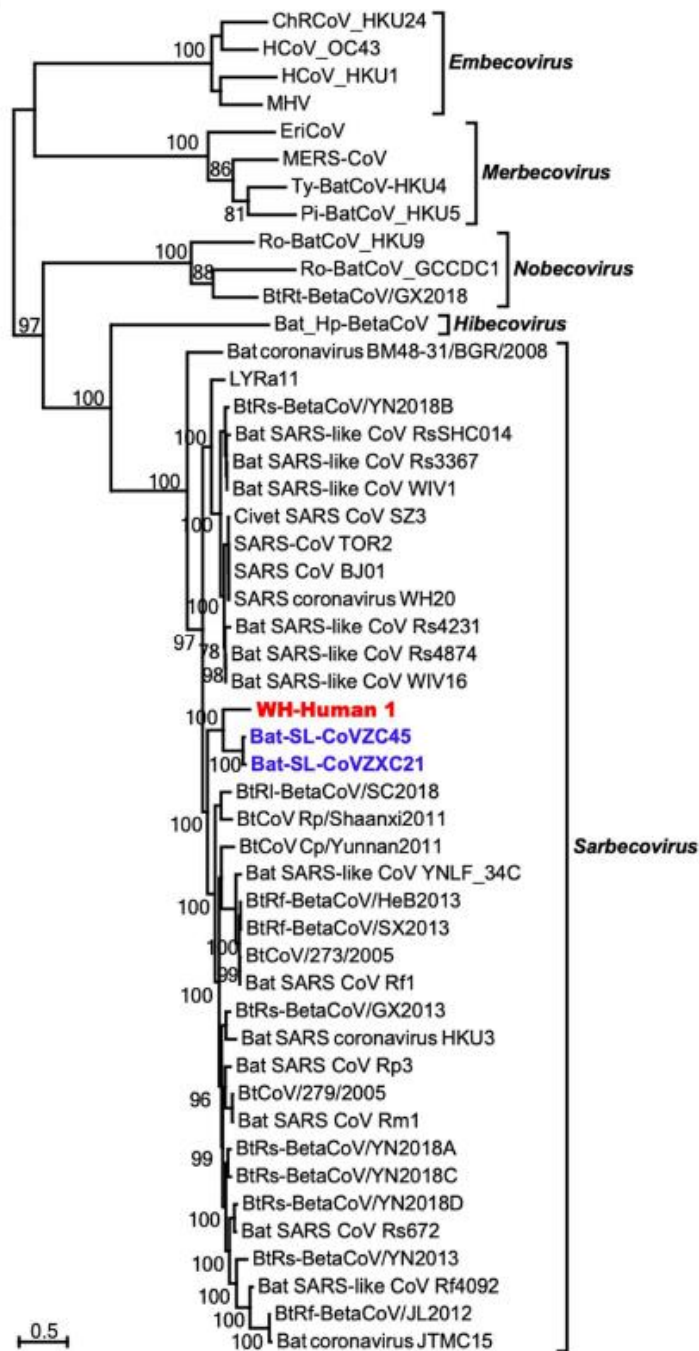
Code: \$ figtree





0.3

S



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

Questions