Technical University of Denmark

A Tale of Two Viruses: Lassa and SARS-CoV-2

Maximum Likelihood and R-based Phylogenetics

Author
Swati Tak – s220868

March 2023

Table of Contents

Contents

Introduction

Emergence of novel viruses poses a constant threat to public health, with potentially catastrophic consequences as we have seen in the very recent past. Lassa fever and COVID-19, caused by the Lassa and SARS-CoV-2 viruses respectively, are two of the most significant viral epidemics in recent history. Understanding the origins of these viruses is essential for developing effective strategies to prevent and control future outbreaks. In this project, Maximum Likelihood and R-based Phylogenetics were applied to trace the evolutionary history of Lassa and SARS-CoV-2. By analysing genetic data from a variety of sources,[1] this project report sheds some light on the origins of these two deadly viruses.

Construction of Dataset

Steps taken to construct dataset for SARS-CoV-2:

1. Search SARS-CoV-2 on NCBI Virus using "Search by Virus" option on the main page
2. Refine results from the left-hand-side menu by choosing:
    a. "complete" under "Nucleotide Completeness"
    b. "Human" under "Host"
    c. "01/01/2003" to "01/01/2004" under "Release Date"
3. From the results table, select the entry with ID "NC_004718" as the 2003 SARS-CoV-2 sequence (for the outgroup)
4. Download the 2003 outgroup sequence by using following options:
    a. Click "Download" next to "Explore Virus Data"
    b. Select "Nucleotide" under "Sequence data (FASTA Format)" and click "Next"
    c. Select "Download Selected Records" and click "Next"
    d. Select "Use default" and click "Download"
5. Reset the Refine Results options to find other SARS sequences this time as follows:
    a. Under Virus, type "SARS" and select the entry with taxid "694009"
    b. Choose "complete" under "Nucleotide Completeness"
    c. Choose "Human" under "Host"
6. From the results table, now select around 30-35 sequences from different locations and time frames from the result table

[1] Dataset for the Lassa virus was obtained from Anders Gorm Pedersen, Professor at DTU. In Appendix I, of this report, you can find a list of all sequences included in the SARS-CoV-2 dataset.

7. Click “Download” and follow the same options as earlier for downloading them in fasta format
8. Also download “CSV” from “Current table view result” and follow the same options for download as above for selected records

Steps for construction of non-human sars virus sequences:

1. From the above dataset, copy any one of the sequences and paste it in the **Enter Query Sequence** field on BLASTN
2. Under **Choose Search Set**, write “Human” in the **Organism** field and check the **exclude** option next to it
3. Click “BLAST” button at the end to run the search
4. From the search results, choose all sequences that have “bat” and “pangolin” words in their names and have a percent identity of more than 90%
5. Avoid selecting sequences that have words like “synthetic” or “constructs” in their names
6. Ensure there are at least 10 sequences selected
7. Click “Download” button on the top of the search table and choose “Fasta” from the drop-down

The two sets of sequences (with human and non-human hosts) were combined manually by copy pasting all in one file.

Alignment of Sequences

After combining the sequences, we aligned them using the Mafft algorithm.[2] Mafft is a popular tool for aligning DNA or protein sequences. It uses a progressive alignment approach that builds the alignment iteratively, starting with the most similar sequences and gradually adding more divergent sequences. This algorithm also incorporates various techniques to improve the accuracy and speed of the alignment process, such as gap opening penalties, iterative refinement, and consistency-based scoring.

[2] “Multiple alignment program for amino acid or nucleotide sequences”, *Mafft version 7*, https://mafft.cbrc.jp/alignment/software/, Accessed: March 2003.

Aligning the sequences using Mafft

The below command was used to run the mafft algorithm on the dataset.

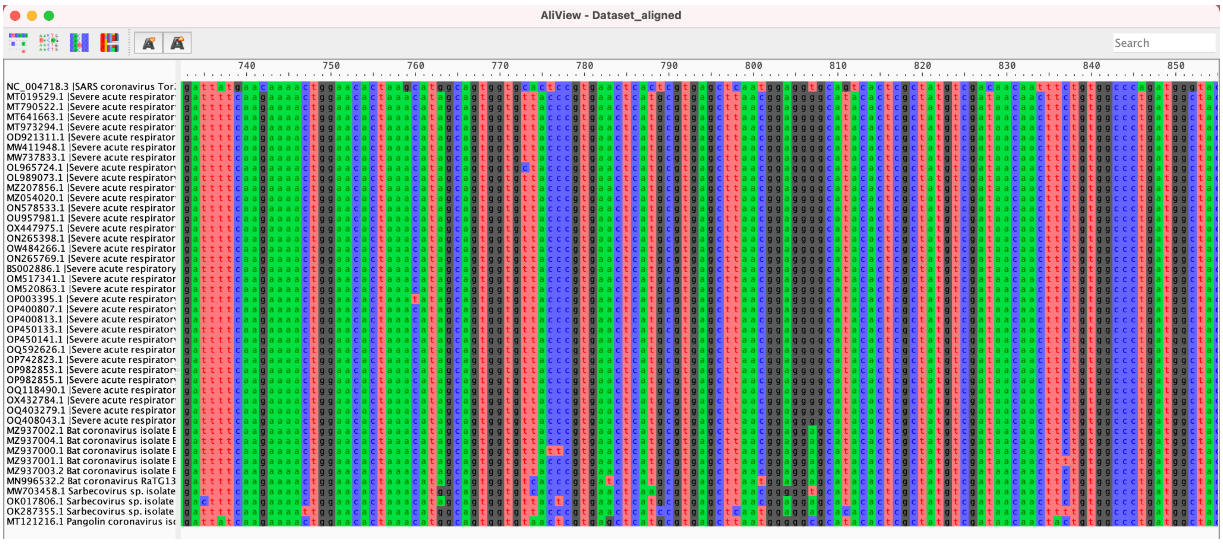
mafft --auto Dataset.fasta > Dataset_aligned

It aligns the sequences in a FASTA file called Dataset.fasta using the MAFFT software and saves the output in a new file called Dataset_aligned.

The --auto option is used to automatically select the appropriate algorithm and parameters for the alignment based on the input data.

The > operator is used to redirect the standard output (i.e. the aligned sequences) of the mafft command to a file called Dataset_aligned instead of printing it to the terminal.

Inspecting the alignment in Aliview



```
NC_004718.3 |SARS coronavirus Tor2
MT019529.1 |Severe acute respiratory
MT790522.1 |Severe acute respiratory
MT641663.1 |Severe acute respiratory
MT973294.1 |Severe acute respiratory
OD921311.1 |Severe acute respirator
MW411948.1 |Severe acute respirator
MW737833.1 |Severe acute respirator
OL965724.1 |Severe acute respiratory
OL989073.1 |Severe acute respiratory
MZ207856.1 |Severe acute respiratory
MZ054020.1 |Severe acute respiratory
ON578533.1 |Severe acute respirator
OU957981.1 |Severe acute respirator
OX447975.1 |Severe acute respiratory
ON265398.1 |Severe acute respirator
OW484266.1 |Severe acute respirator
ON265769.1 |Severe acute respirator
BS002886.1 |Severe acute respiratory
OM517341.1 |Severe acute respirator
OM520863.1 |Severe acute respirator
OP003395.1 |Severe acute respiratory
OP400807.1 |Severe acute respiratory
OP400813.1 |Severe acute respiratory
OP450133.1 |Severe acute respiratory
OP450141.1 |Severe acute respiratory
OQ592626.1 |Severe acute respirator
OP742823.1 |Severe acute respiratory
OP982853.1 |Severe acute respiratory
OP982855.1 |Severe acute respiratory
OQ118490.1 |Severe acute respirator
OX432784.1 |Severe acute respiratory
OQ403279.1 |Severe acute respirator
OQ408043.1 |Severe acute respirator
MZ937002.1 Bat coronavirus isolate E
MZ937004.1 Bat coronavirus isolate E
MZ937000.1 Bat coronavirus isolate E
MZ937001.1 Bat coronavirus isolate E
MZ937003.2 Bat coronavirus isolate E
MN996532.2 Bat coronavirus RaTG13
MW703458.1 Sarbecovirus sp. isolate
OK017806.1 Sarbecovirus sp. isolate
OK287355.1 Sarbecovirus sp. isolate
MT121216.1 Pangolin coronavirus iso
```

Figure 1: Homologous regions visualised through Aliview

As you can see here, the sequences are very much similar at numerous sites and are thus quite homologous in nature.

Phylogenetic Study Using PAUP

Maximum Likelihood and GTR+I+G Model

Maximum likelihood (ML) is a statistical method used to infer evolutionary relationships among DNA or protein sequences by estimating the likelihood of the observed data under different evolutionary models. In the context of phylogenetic reconstruction, the method involves constructing a tree that maximizes the likelihood of observing the sequence data given the model of evolution and the tree topology. The model of evolution accounts for different rates of substitutions, transitions, and transversions among nucleotides, and different frequencies of occurrence of each nucleotide at different positions in the sequence.[3]

The GTR+I+G model is one of the commonly used models in ML phylogenetic reconstruction. GTR stands for the General Time Reversible model, which assumes that the rates of substitutions among nucleotides vary among different positions in the sequence and that the nucleotide frequencies at different positions evolve independently. The I parameter accounts for the unequal base frequencies in the data, while the G parameter accounts for rate variation across sites, following a gamma distribution. Together, the GTR+I+G model accounts for both the heterogeneity of substitution rates among different positions in the sequence and the heterogeneity of substitution rates across sites due to different evolutionary pressures, such as selection or mutation bias.[4]

The above-mentioned method and model can be applied using PAUP, a computational program that stands for Phylogenetic Analysis Using Parsimony and other methods.

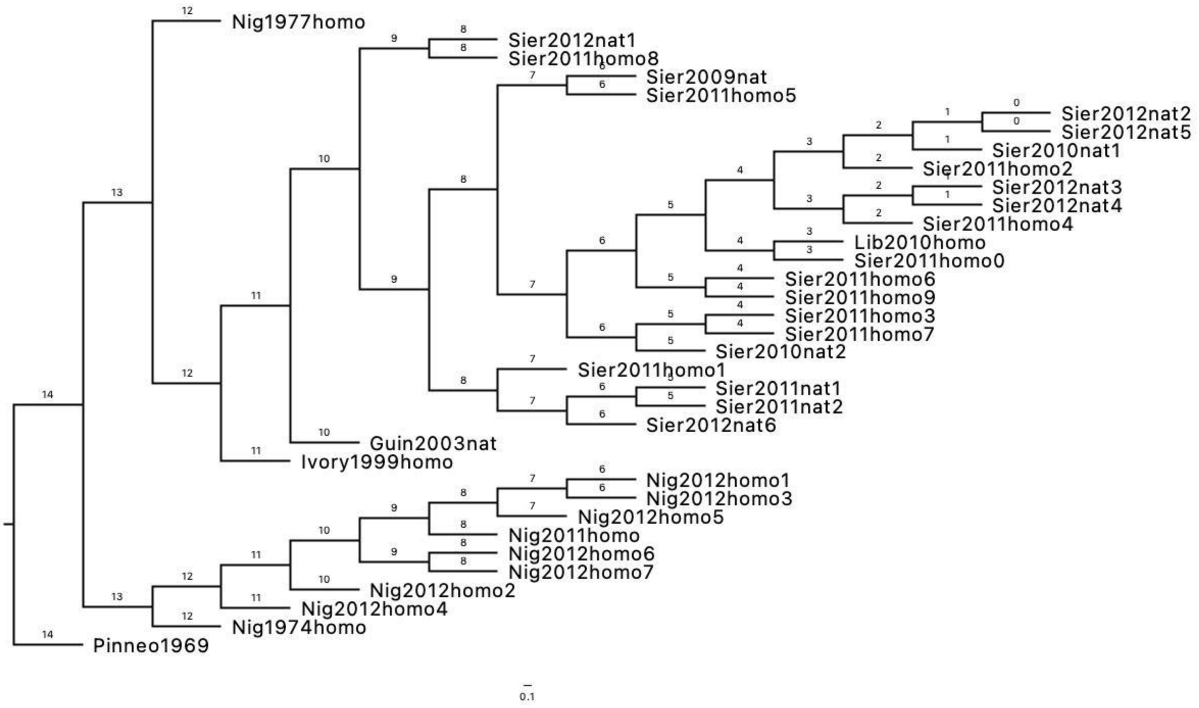
Phylogenetic Tree Reconstruction: Lassa Virus

To construct a phylogenetic tree for Lassa virus sequences, the following steps were applied in PAUP.

[3] "Maximum Likelihood Method", *ScienceDirect*, www.sciencedirect.com/topics/medicine-and-dentistry/maximum-likelihood-method, Accessed: March 2023.

[4] "Trends in substitution models of molecular evolution", *NCBI*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4620419/, Accessed: March 2023.

Step	**Command used**
Start paup and load the dataset	paup lassa.nexus
Set criteria as maximum likelihood	set criterion=likelihood
Set model as GTR with fraction of invariant sites and gamma distributed rates	lset nst=6 rmatrix=estimate basefreq=empirical rates=gamma shape=estimate pinvar=estimate
Set the outgroup to Pinneo1969 sequences	outgroup Pinneo1969
Set to root the tree according to the outgroup and also as monophyletic	set root=outgroup outroot=monophyl
Run hsearch to construct the trees	hsearch swap=nni
Find likelihood score of the best tree	lscore **Likelihood of the 3 best trees:** Tree 1 = 10889.946 Tree 2 = 10889.946 Tree 3 = 10889.947
Check the tree through paup command line	describe
Save the tree in Nexus format and exit paup	saveTrees format=Nexus file=lassa_tree.nexus brLens=yes q
Visualise the tree in figtree and then export it as a jpeg	figtree lassa_tree.nexus



There are multiple instances (6 instances as per the tree) when the lassa virus jumped from rodents to humans, and also evolved while doing so. Mastomys natalensis seems to be acting as a consistent host for the virus throughout this evolution and thus has been key to the spread of the virus. So one of the ways to manage or stop the spread of the lassa virus could be to control the population of Mastomys natalensis. By reducing the number of these rodents and minimizing human contact with them, the risk of transmission can be decreased.

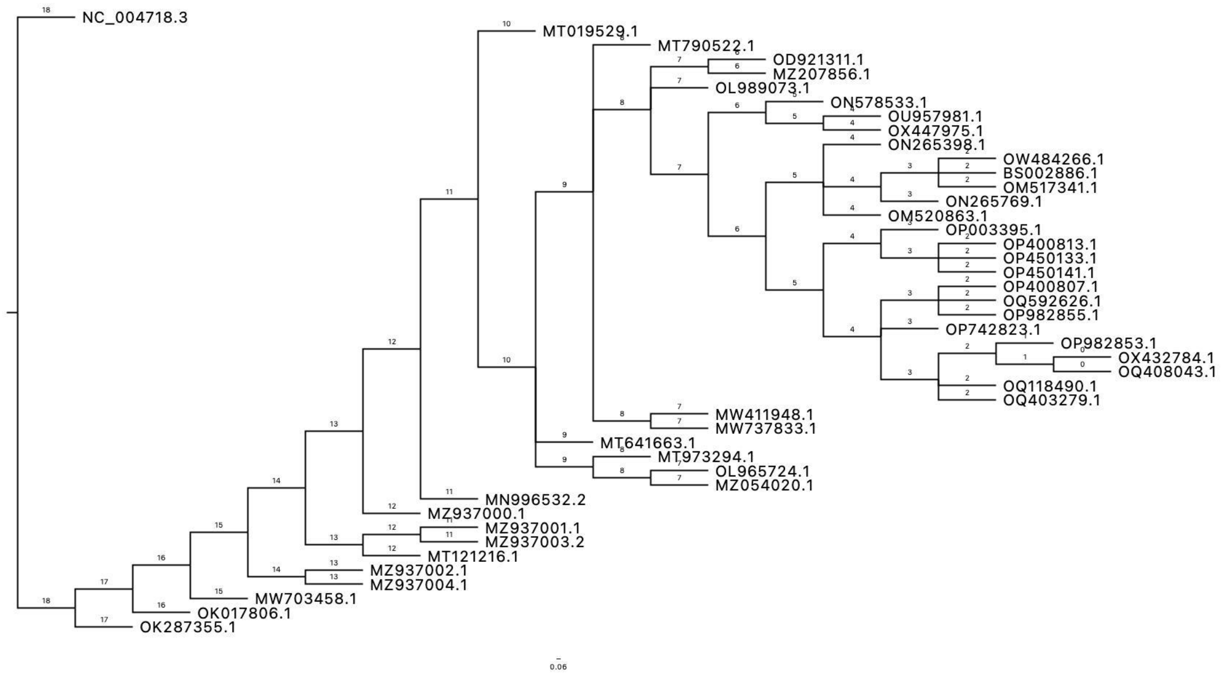
Phylogenetic Tree Reconstruction: SARS-Cov-2

To construct a phylogenetic tree for SARS-CoV-2 sequences, the alignment file was first converted to Nexus format to be able to use it in PAUP. For the conversion, the following command was used.

```
seqconverter -I auto -O nexus Dataset_aligned > Dataset_aligned.nexus
```

The next steps for phylogenetic reconstruction were carried out in PAUP as follows.

Step	**Command used**
Start paup and load the dataset	paup Dataset_aligned.nexus
Set criteria as maximum likelihood	set criterion=likelihood
Set model as GTR with fraction of invariant sites and gamma distributed rates	lset nst=6 rmatrix=estimate basefreq=empirical rates=gamma shape=estimate pinvar=estimate
Set the outgroup to 2003 SARS sequences	outgroup NC_004718.3
Set to root the tree according to the outgroup and also as monophyletic	set root=outgroup outroot=monophyl
Run hsearch to construct the trees	hsearch swap=nni
Find likelihood score of the best tree	lscore **Likelihood score of the one best tree**: 110235.46
Check the tree through paup command line	describe
Save the tree in Nexus format and exit paup	saveTrees format=Nexus file=sars_tree.nexus brLens=yes q
Visualise the tree in figtree and then export it as a jpeg	figtree sars_tree.nexus



While there has been a common ancestor with the human host for all the sequences here, it can be seen that the virus seems to have jumped from bats to humans at some point before the period of the recent pandemic. This means that interaction or exposure to bats can be a contributing factor for the control of this virus. Additionally, the virus seems to have evolved quite fast once it jumped to the human host, probably because it needed to adjust its genome for survival in the human body and evade immune responses.

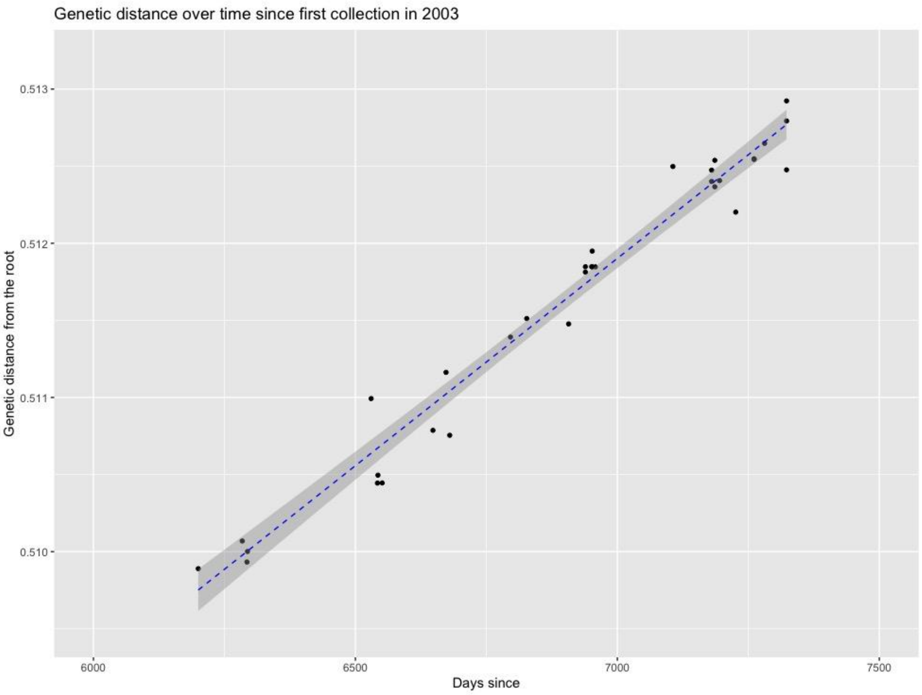
There has been one instance where the virus jumped from bat host (MN996532.2 Bat coronavirus RaTG13, complete genome) to human host (MT019529.1|Severe acute respiratory syndrome coronavirus 2 isolate).

R-based Phylogenetics

Analysis of clock-like behaviour and origin of SARS-CoV-2

Steps taken for the analysis in R

1. Install and load required packages (packages used: ape, apTreeshape, dplyr, ggplot2, tibble, nlme)
2. Read the tree and subset the data frame for human sequences by dropping non-human sequences from the original tree
3. Match and merge, on the basis of Accession, all the human sequences data with the csv data that has collection dates
4. Create a scatter plot using ggplot2 package and using geom_smooth(method=lm)



The points seem to be falling along a straight line as you can see in the plot so we can presume the substitution rate is relatively constant over time.

5. Perform linear regression using **nlme** package

Intercept = 0.03652345
Slope = Substitution Rate = 6.875702e-05

It seems the viral genome has been actually evolving constantly and fast with time and is thus evolving according to a molecular clock, which means that the virus is accumulating mutations in its genetic makeup at a steady and predictable rate.

6. Calculate the number of substitutions per site per year
Number of substitutions happening per site per day = 6.875702e-05
So we can now calculate the estimate clock rate as the number of substitutions happening per site per year as follows: 6.875702e-05 * 365 = 0.02509631 = 25096.31 x 10^{-6}.

The speed of the rate of 25096.31 x 10^{-6} substitutions per site per year can be considered as relatively fast in the context of RNA viruses like SARS-CoV-2. RNA viruses typically have high mutation rates due to their error-prone replication machinery, and this leads to frequent substitutions and rapid evolution over short periods of time. The estimated substitution rate for SARS-CoV-2 is consistent with previous estimates for RNA viruses and suggests that the virus is evolving rapidly, which can have important implications for the development of vaccines, therapeutics, and other control measures.[5]

7. Calculate the number of days when the most common ancestor might have existed before the first sequence collection using the intercept, the substitution rate and the genetic distance between the most recent common ancestor (from 2019) and root (2003), which was calculated through cophenetic function, that is, 0.50989.
The formula for the regression line is: Genetic distance = Intercept + (Substitution rate * Time), where Intercept is the estimated intercept of the regression line, Substitution rate is the estimated slope or substitution rate of the regression line, and Time is the time variable.

Genetic distance = Intercept + (Substitution Rate * Number of days)
0.50989 = 0.03652345 + (6.875702e-05 * Number of days)

[5] Duffy, S. (2018). “Why are RNA virus mutation rates so damn high?”, *PLOS Biology*, 16(8), e3000003, https://doi.org/10.1371/journal.pbio.3000003.

Number of days = (0.50989 – 0.03652345)/6.875702e-05 = 6884.629 days

The date prior to the first collection on 23-12-2019 by 6884.629 days was 15-02-2001.
Thus, a common ancestor for the SARS-Cov-2 for humans existed on **15 February 2001**.

Conclusion

- The SARS-CoV-2 virus is evolving rapidly, with an estimated rate of 25096.31 x 10^{-6} substitutions per site per year.
- The impact of these mutations on the virus's survival and transmission in the human host is difficult to determine without further analysis and experimentation.
- The molecular clock analysis indicates that the SARS-CoV-2 virus has been evolving steadily over time, which could have implications for the development of effective control and prevention strategies.
- Continued research and surveillance are necessary to better understand the evolution and transmission of emerging viral pathogens and to develop effective strategies for their control and prevention, particularly in the context of RNA viruses.

Appendix I: SARS-CoV-2 Dataset Sequence Names

Earliest SARS sequence from 2003:

- NC_004718.3 |SARS coronavirus Tor2, complete genome

33 human SARS-CoV-2 sequences:

- MT019529.1 |Severe acute respiratory syndrome coronavirus 2 isolate BetaCoV/Wuhan/IPBCAMS-WH-01/2019, complete genome
- MT790522.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ZMB/29/2020, complete genome
- MT641663.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/AUS/NT19/2020, complete genome
- MT973294.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/AUS/TAS220/2020, complete genome
- OD921311.1 |Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: monopartite
- MW411948.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/West Bank/AAS27/2020, complete genome

- MW737833.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/West Bank/AAS17/2020, complete genome
- OL965724.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/hCov_19_USA_ID_IVREF_689736_2020/2020, complete genome
- OL989073.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ARG/210316-1/2021, complete genome
- MZ207856.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/West Bank/AAS79/2021, complete genome
- MZ054020.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/NH-CDCBI-CRSP_FCP2TXEQKC44AWE5/2021, complete genome
- ON578533.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/BRA/RJ-NVBS5320GENOV827956611926/2021, complete genome
- OU957981.1 |Severe acute respiratory syndrome coronavirus 2 isolate RNA genome assembly, complete genome: monopartite
- OX447975.1 |Severe acute respiratory syndrome coronavirus 2 isolate 20214901293 genome assembly, complete genome: monopartite
- ON265398.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/FRA/IHUCOVID-060638-Nova1E/2021, complete genome
- OW484266.1 |Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: monopartite
- ON265769.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/FRA/IHUCOVID-061135-Nova1E/2022, complete genome
- BS002886.1 |Severe acute respiratory syndrome coronavirus 2 hCoV-19/Japan/SZ-NIG-Y212769/2022 RNA, complete genome
- OM517341.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/OH-CDC-ASC210584594/2022, complete genome
- OM520863.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/IA-CDC-LC0505804/2022, complete genome
- OP003395.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CO-CDPHE-2103416775/2022, complete genome
- OP400807.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/TX-CDC-STM-7VFMXAUCG/2022, complete genome
- OP400813.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/IL-CDC-STM-33U67VPVP/2022, complete genome
- OP450133.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/GA-CDC-STM-JHAJYQFF2/2022, complete genome
- OP450141.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/FL-CDC-STM-3MG34U9JX/2022, complete genome
- OQ592626.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/VNM/NHTD-OUCRU3749/2022, complete genome
- OP742823.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/IL-CDC-QDX42513300/2022, complete genome

- OP982853.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/NJ-CDC-QDX43796093/2022, complete genome
- OP982855.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/NV-CDC-QDX43796834/2022, complete genome
- OQ118490.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CA-CDC-LC0960078/2022, complete genome
- OX432784.1 |Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: monopartite
- OQ403279.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/WA-CDC-LC0998749/2023, complete genome
- OQ408043.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/UT-UPHL--23020774478/2023, complete genome

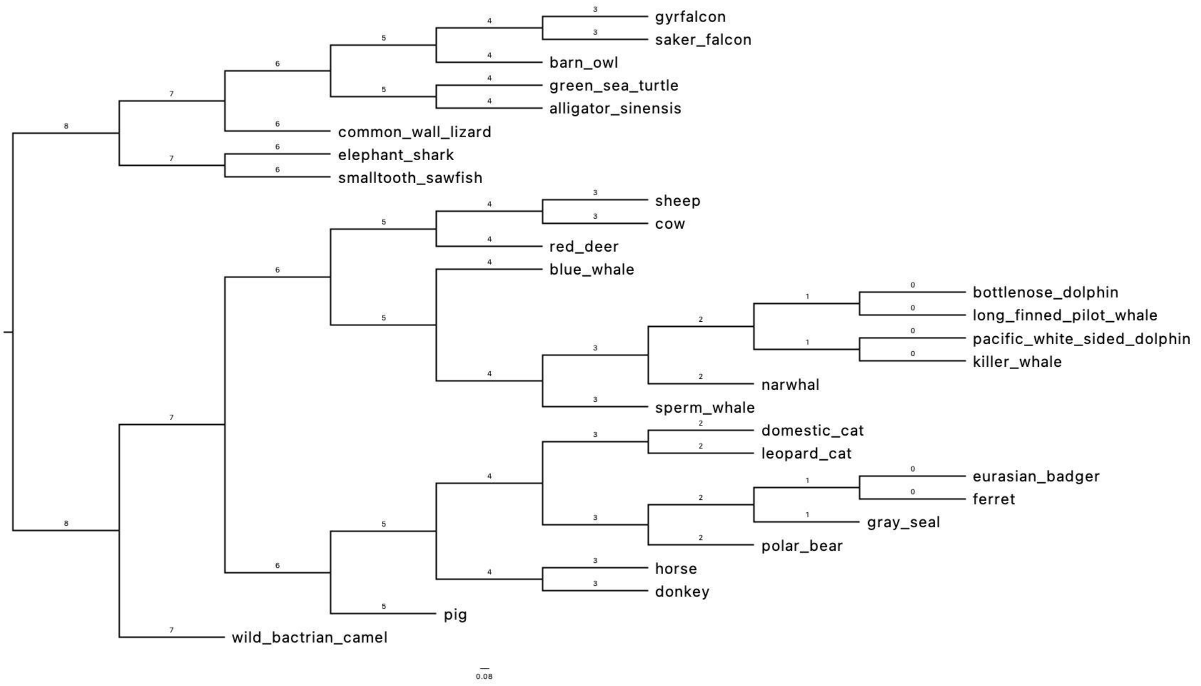
10 non-human SARS virus sequences:

- MZ937002.1 Bat coronavirus isolate BANAL-20-116/Laos/2020, complete genome
- MZ937004.1 Bat coronavirus isolate BANAL-20-247/Laos/2020, complete genome
- MZ937000.1 Bat coronavirus isolate BANAL-20-52/Laos/2020, complete genome
- MZ937001.1 Bat coronavirus isolate BANAL-20-103/Laos/2020, complete genome
- MZ937003.2 Bat coronavirus isolate BANAL-20-236/Laos/2020, complete genome
- MN996532.2 Bat coronavirus RaTG13, complete genome
- MW703458.1 Sarbecovirus sp. isolate PrC31, complete genome
- OK017806.1 Sarbecovirus sp. isolate YN2021, complete genome
- OK287355.1 Sarbecovirus sp. isolate BetaCoV_Yunnan_Rp_JCC9_2020, complete genome
- MT121216.1 Pangolin coronavirus isolate MP789, complete genome

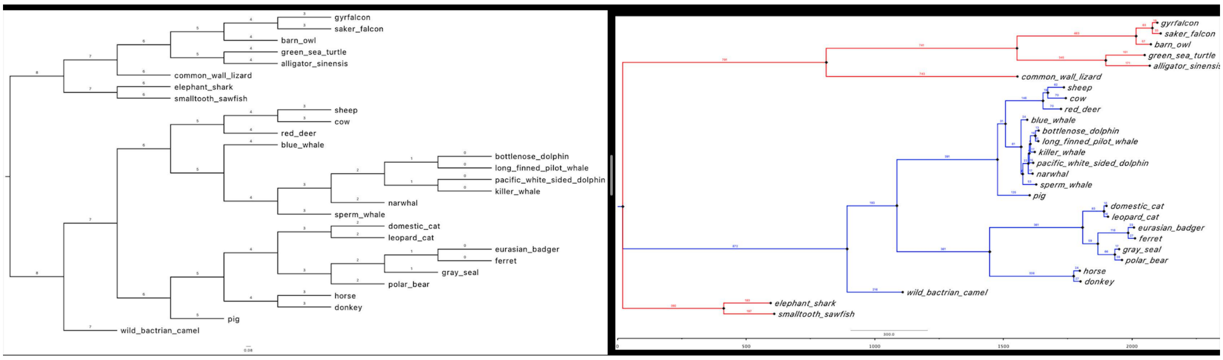
Appendix II: Maximum Likelihood and GTR+I+G Model for Whale Phylogenetic Study from Project 1

Step	Commands used
Start paup and load the dataset	paup mafft.nexus
Rationale for choosing maff alignment: *Mafft alignment was chosen for redoing the maximum likelihood analysis for the whale evolution because mafft returned the most homologous results compared to the other two algorithms, namely clustalw and muscle.*	

Set criteria as maximum likelihood	set criterion=likelihood
Set model as GTR with fraction of invariant sites and gamma distributed rates	lset nst=6 rmatrix=estimate basefreq=empirical rates=gamma shape=estimate pinvar=estimate
Set the outgroup to 2003 SARS sequences	outgroup elephant_shark smalltooth_sawfish green_sea_turtle alligator_sinensis common_wall_lizard gyrfalcon saker_falcon barn_owl
Set to root the tree according to the outgroup and also as monophyletic	set root=outgroup outroot=monophyl
Run hsearch to construct the trees	hsearch swap=nni
Find likelihood score of the best tree	lscore **Likelihood score of the best tree:** 49367.197
Check the tree through paup command line	describe
Save the tree in Nexus format and exit paup	saveTrees format=Nexus file=whale_tree.nexus brLens=yes
Visualise the tree in figtree and then export it as a jpeg	figtree whale_tree.nexus



Comparison between trees made with parsimony and maximum likelihood methods



The two trees pretty much look similar except for a couple of minor differences in the terms of duration when certain whale species evolved. The two trees still depict the same evolutionary relationship with land-dwelling mammals and thus convey similar results.