EVOLUTION AND POPULATION DIVERSITY STUDY OF INFLUENZA D VIRUS (IDV) USING BEAST ALGORITHM

A REPORT SUBMITTED TO
SAVITRIBAI PHULE PUNE UNIVERSITY

TOWARDS PARTIAL FULFILLMENT OF DEGREE
OF
MASTER OF SCIENCE (M. SC.)
IN STATISTICS
IN THE FACULTY OF SCIENCE AND TECHNOLOGY

SUBMITTED BY

Dangat Ashitosh Vilas

UNDER THE GUIDANCE OF

Dr. Mohan Kale

DEPARTMENT OF STATISTICS
AND
CENTRE FOR ADVANCED STUDIES IN STATISTICS,
SAVITRIBAI PHULE PUNE UNIVERSITY,
PUNE-411007, INDIA. MAY, 2023

Certificate of the Guide

This is to certify that, the following students of M.Sc. Statistics,

(1) Dangat Ashitosh Vilas

have successfully completed their project titled **Evolution and Population Diversity Study of Influenza D Virus (IDV) using BEAST Algorithm** under the guidance of **Dr. Mohan Kale** and have submitted this project report on **May,19,2023** as a part of the course ST-402, towards partial fulfillment of requirements for the degree of M.Sc. Statistics in Savitribai Phule Pune University in the academic year 2022-2023.

Dr. Mohan Kale
(Project Guide)

Prof. T. V. Ramanathan
(Head of the Department)

Acknowledgments

We would like to express our deep and sincere gratitude to **Dr. Mohan Kale** for his valuable guidance and motivation for the successful completion and execution of the project. His profound knowledge and expertise came in handy for our project.

It gives us immense pleasure to thank **Mr. Sanket Limaye**, Bioinformatics Centre, Savitribai Phule Pune University, for providing timely inputs and suggestions required for the project.

We would also like to thank **Prof. T.V. Ramanathan, Head, Department of Statistics,** and the teaching and non-teaching staff for their invaluable assistance.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Summary

1.1 Introduction

Influenza D Virus (IDV) was initially isolated from pigs and subsequently from bovine (cattle). Since the seroprevalence in bovine is much higher than in swine, bovine is considered to be the primary natural reservoir for IDV. The pathology and mode of transmission of IDV were investigated by experimental infection of cattle and swine, but also of guinea pigs, ferrets, camels, and horses. The virus was also transmitted to naive animals by direct contact. After experimental infection with IDV in cattle, the virus can be detected both in the upper and lower respiratory tracts and transmitted to contact animals. Previous studies have reported that IDV detection rates are higher in clinically sick cattle than in apparently healthy cattle. The recent continuous outbreaks of IDV in bovines and other hosts imply that IDV can transmit among hosts with high efficiency. The intercontinental transmission and high prevalence of IDV, especially in cattle highlight its potential threat to other agricultural animals and humans.

Bayesian analysis has been conducted to address the origin and evolutionary history of IDV. Bayesian evolutionary analysis by sampling trees (BEAST) combines statistical techniques with molecular sequence data to reconstruct phylogenetic trees and estimate evolutionary parameters, allowing researchers to explore the genetic relationships, divergence times, and population dynamics of pathogens.

Evolutionary rate study revealed that IDV evolves with a high substitution rate. Hence, continuous monitoring of the evolution of IDV needs to be conducted in the future. As with other types of influenza viruses, IDVs will constantly evolve into more diverse lineages. The IDV vaccine research efforts in the future need to accommodate for such antigenic drift phenomena and focus on developing a universal vaccine that can effectively protect animals or humans from multiple strains or lineages of IDV. This study aims to enlighten the intricate evolutionary processes and population diversity of the IDV using the BEAST algorithm.

1.2 Motivation

IDV primarily affects cattle and other ruminants, leading to respiratory illnesses that can have significant implications for animal health, welfare and productivity. The cattle industry is extremely important as it stimulates the self-sufficiency of the rural economy by providing auxiliary income to farmers and in turn, initiates the economic growth of a country and consequently contributes to the Gross Domestic Product.

IDV represents a relatively new and understudied member of the influenza virus family. As an emerging pathogen, it poses a potential threat to animal and potentially to human health. Understanding the virus's biology, transmission patterns, and evolutionary dynamics can aid in the development of targeted interventions, control measures, and prevention strategies to safeguard animal populations. The bovine respiratory disease complex costs the global cattle industry billions of dollars per year. As observed earlier, IDV has emerged an important role in this disease complex, thereby indicating a need for an effective and safe vaccine against IDV infection in the bovines. Currently, there are no specific vaccines or treatments for IDV in animals.

The interspecies transmission and the international appearance of IDV in worldwide animal populations represent a potential risk to global human health which clearly motivates further investigation of this novel influenza virus.

1.3 Objectives

- To estimate the rates of substitution (per site, per year) evolution.
- The study the time of evolution for the most recent common ancestor. IDV continuously evolves into more diverse lineages, its vaccine research efforts in the future need to accommodate for such an antigenic drift phenomenon and focus on developing a universal vaccine that can effectively protect animals or humans from multiple strains or lineages of IDV.
- To model the substitution rates for a nucleotide.
- To determine the maximum parsimony rate.
- To estimate the linkage disequilibrium.
- To conduct molecular phylogeny analysis.
- To perform biodiversity study.

1.4 About virus and disease:

The influenza viruses are members of the family Orthomyxoviridae. The influenza viruses A, B, C and D represent the four antigenic types of influenza viruses. In there four antigenic types, the influenza A virus is the most severe, the influenza B virus is less severe but can still cause outbreaks and the influenza C virus is usually only associated with minor symptoms.

IDV primarily spreads among animals, particularly cattle, through direct contact with infected animals or exposure to respiratory secretions. The virus can also spread indirectly through contaminated surfaces, although the exact duration of survival on surfaces is not well studied for IDV.

The respiratory secretions of infected animals, such as nasal discharge, saliva or respiratory droplets expelled during coughing or sneezing contain the virus. A close contact between infected and susceptible animals, such as being housed together in the same area or through nose-to-nose contact, facilitates the spread of IDV.

Taxonomy of IDV is as follows:

Realm: Riboviria
Kingdom: Orthornavirae
Phylum: Negarnaviricota
Class: Insthoviricetes
Order: Articulavirales
Family: Orthomyxoviridae
Genus: Deltainfluenzavirus
Species: Influenza D virus

1.5 Data Source and Description

We collect secondary data from National Center for Biotechnology Information GenBank (Reference: `https://www.ncbi.nlm.nih.gov/genbank/`). It provides a large suite of online resources for the biological information and data. GenBank sequence database is an open access annotated of all publicly available nucleotide sequences. IDV has an identical genome structure. It consists of 7 gene segments namely PB2, PB1, P3, HEF, NP, P42 and NS. The NS1 and NS2 genes has been separated from the NS segment as the independent segments. The gene sequences are further extracted, curated and aligned by a Ph. D. student of the Bioinformatics Centre, Savitribai Phule Pune University and then used for analysis.
The sequences are as follows:

SR NO.	SEQUENCE NAME	NO. OF SEQUENCES
1	PB2	175
2	PB1	170
3	P3	189
4	HEF	188
5	NP	183
6	P42	179
7	NS	147
8	NS1	147
9	NS2	147

Table 1.1: Number of Sequences

1.6 Terminologies and Definitions

Allele: An alternative form of a gene (one member of a pair).

Biodiversity: Biodiversity is the degree of variation of life forms in a given ecosystem.

Genetic biodiversity: Genetic biodiversity refers to the total number of genetic characteristics in the genetic makeup of a species.

Genome: Complete set of genetic information found within an individual organism.

Genotype: An individual's genotype for a gene is the set of alleles, which the gene contains. Such as the letter Bb (B-dominant genotype and b-recessive genotype).

Serotype: Refers to the distinct variations within a species of bacteria or viruses or among immune cells of different individuals.

Linkage: Two genetic loci are said to be in linkage if the alleles at these loci segregate together more often than would be expected by chance.

Population admixture: It occurs when individuals from two or more previously separated populations begin interbreeding.

Recombination: Recombination is the process by which two Deoxyribonucleic acid (DNA) molecules share genetic information or their ancestry, producing a new combination of alleles.

Parsimony informative sites: These are the Sites on a chromosome containing at least two different allelic states, each represented at least twice.

Phylogenetic tree: It is nothing but a branching diagram or "Tree" showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their genotypic or phenotypic characteristics.

Sequence alignment: Sequence alignment is the procedure of comparing two or more sequences by searching for a series of individual characters that are in the same order in the sequences.

Virus: Small infectious agent that replicates only inside the living cells of other organisms.

Strain: A strain is a genetic variant or subtype of a micro-organism (e.g. virus, bacterium, fungus). For example, a ”flu strain” is a certain biological form of influenza or ”flu” virus.

95% HPD Lower: The lower bound of the highest posterior density (HPD) interval. The HPD is a credible set that contains 95% of the sampled values.

95% HPD Upper: The upper bound of the highest posterior density (HPD) interval. The HPD is a credible set that contains 95% of the sampled values.

Auto-Correlation time (ACT): The number of states in the Markov chain Monte Carlo (MCMC) chain that two samples have to be from each other for them to be uncorrelated. The ACT is estimated from the samples in the trace (excluding the burn-in).

Effective sample size (ESS): The ESS is the number of independent samples that the trace is equivalent to. This is essentially the chain length (excluding the burn-in) divided by the ACT.

Maximum clade credibility (MCC) tree: It finds the tree with the highest product of the posterior probability of all its nodes.

Mean substitution rate per site per year: The mean substitution rate per site per year refers to the average rate at which genetic substitutions, such as nucleotide changes, accumulate in a particular region of a genome over a specific period of time. It is commonly used in molecular clock analyses and provides an estimate of the evolutionary rate of a sequence.

Maximum parsimony: Maximum parsimony is a principle and method used in phylogenetic analysis to infer the most likely evolutionary tree or phylogeny from a given set of data, typically molecular sequence data such as DNA or protein sequences. The principle of maximum parsimony aims to identify the tree that requires the fewest evolutionary changes or genetic mutations to explain the observed data.

Linkage disequilibrium (LD): LD is a term used in genetics to describe the non-random association or correlation between alleles at different loci (positions) on a chromosome. In other words, it refers to the tendency of certain alleles at different genetic loci to occur together more often than would be expected by chance.

1.7 Software Used

1. **IQTREE**

   IQTREE is a fast and effective stochastic algorithm to infer phylogenetic trees by the method of maximum likelihood.

   Reference: `http://iqtree.cibiv.univie.ac.at/`

2. **TEMPEST v1.5.3**

   TEMPEST is a graphical program for looking for temporal signals in trees generated from tip-dated sequences.

   Reference: `http://tree.bio.ed.ac.uk/software/tempest/`

3. **BEAUti v1.10.4**

   BEAUti stands for Bayesian Evolutionary Analysis Utility. This program is used to import data, design the analysis, and generate the BEAST control file.

   Reference: `https://beast.community/beauti`

4. **BEAST v1.10.4**

   BEAST stands for Bayesian Evolutionary Analysis Sampling Trees. This main program takes a control file generated by BEAUti and performs the analysis.

   Reference: `https://beast.community/`

5. **LOGCOMBINER v1.10.4**

   Logcombiner will combine log files from different runs and reduce the sampling frequency. It allows you to combine log and tree files from multiple independent runs of BEAST.

   Reference: `https://beast.community/logcombiner`

6. **TRACER v1.7.2**

   Tracer is a graphical program for exploring the output of BEAST, diagnosing problems, and summarizing the results.

Reference: `http://tree.bio.ed.ac.uk/software/tracer/`

7. **TREE ANNOTATOR v1.10.4**

   Tree Annotator is a post-analysis program will produce a summary tree from the output of BEAST.

   Reference: `https://beast.community/treeannotator`

8. **FIG TREE v1.4.4**

   Fig Tree is a graphical program for viewing trees, displaying summary information from TreeAnnotator.

   Reference: `http://tree.bio.ed.ac.uk/software/figtree/`

1.8 BEAST Software:

BEAST is a fast, flexible software architecture for Bayesian analysis of molecular sequences related by an evolutionary tree. The BEAST software package is an ambitious attempt to provide a general framework for parameter estimation and hypothesis testing of evolutionary models from molecular sequence data.

As the BEAST is a Bayesian statistical framework, it provides a role for prior knowledge in combination with the information provided by the data. The purpose behind the development of BEAST is to bring a large number of complementary evolutionary models (substitution models, insertion-deletion models, demographic models, tree shape priors, relaxed clock models, and node calibration models) into a single coherent framework for evolutionary inference. A Bayesian Markov chain Monte Carlo (MCMC) method for performing relaxed phylogenetics that is able to co-estimate phylogeny and divergence times under a new class of relaxed-clock models.

MCMC is a stochastic algorithm that produces sample-based estimates of a target distribution of choice. A Bayesian MCMC algorithm needs to evaluate the likelihood of each state in the chain and thus performance is dictated by the speed at which these likelihood evaluations can be made. BEAST attempts to minimize the time taken to evaluate a state by only recalculating the likelihood for parts of the model that have changed from the previous state. BEAST uses the Java version of the functions, thereby retaining its platform independence.

Chapter 2

Statistical Methods

2.1 Markov Chain Monte Carlo (MCMC)

Notations: X: genotypes of sampled individuals

Z: population of origin of individuals (unknown)

P: allele frequencies in all populations (unknown)

Q: admixture proportion for each individual.

P, Q-Dirichlet distribution (used as a prior distribution in Bayesian statistics)

MCMC is useful for obtaining the distribution $\pi(\theta)$.

In this case $\theta = (Z, P, Q) \quad and \quad \pi(0) = P(Z, P, Q|X)$

The idea is to construct Markov Chain $\theta(0), \theta(1), \theta(2)$..... with stationary distribution $\pi(\theta)$ the $\theta(m)$ is approximately distributed as $\pi(\theta)$ where m is sufficiently large.

This can be formalized and shown to be true provided the Markov chain satisfies certain technical conditions such as ergodicity that hold here.

Furthermore, for sufficiently large $C, 0(m), 0(m+c), 0(m+2c)$ will reasonably independent samples from (0). The values of M and C are referred to as M: Burn-in period of the Markov chain. C: Thinning interval the value of m and C heavily depends on the amount of correlation between two successive states of the Markov Chain.

2.2 Chi-square test:

At the beginning of each run, IQTREE performs a composition chi-square test for every sequence in the alignment. The purpose is to test for homogeneity of character composition (e.g., nucleotide for DNA, amino-acid for protein sequences). A sequence is denoted as failed if its character composition significantly deviates from the average composition of the alignment. More specifically, for each sequence, compute:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i – E_i)^2}{E_i},$$

where k is the size of the alphabet
(e.g. 4 for DNA, 20 for amino acids) and the values 1 to k correspond uniquely to one of the characters. O_i is the character frequency in the sequence tested. E_i is the overall character frequency from the entire alignment. Whether the character composition deviates significantly from the overall composition is done by testing the χ^2 value using the χ^2 distribution with $k-1$ degrees of freedom (df=3 for DNA or df=19 for amino acids).

This test is regarded as an explorative tool that might help to nail down problems in a dataset. One would typically not remove failing sequences by default. But if the tree shows unexpected topology the test might point in the direction of the origin of the problem.

2.3 Maximum likelihood calculations based on AIC, BIC, AICc

IQTREE is a software package that implements maximum likelihood and Bayesian inference methods for phylogenetic analysis. It uses these methods to estimate the evolutionary relationships and parameters of genetic data. IQ-TREE also provides model selection tools, such as Akaike Information Criterion (AIC), Corrected Akaike Information Criterion (AICc) and Bayesian Information Criteria (BIC) to compare different models and select the most appropriate one for the data at hand. It scores each model on the basis of AIC, AICc and BIC. In BIC, a stronger penalty is applied to model complexity compared to AIC and AICc, making it more effective in selecting parsimonious models.

1) AIC: The AIC is a measure of the relative quality of a statistical model. Akaike proposed an information criterion, AIC, based on maximizing the expected entropy of the model. Entropy is a measure of the expected information, in this case, the Kullback - Leibler information measure. Essentially, the AIC is a penalized log-likelihood measure. Let L be the likelihood function for a specific model. The AIC is,

$$AIC = -2lnL + 2k,$$

where, k includes the parameters from the substitution model, such as base frequencies, substitution rates, the proportion of invariable sites, or rate variation among sites.

2) BIC: BIC aims to select the model that maximizes the likelihood of the data while also considering model complexity. It penalizes complex models more heavily than the AIC does. The BIC is,

$$BIC = -2lnL + k * ln(n),$$

where k is the number of estimable parameters and n is the sample size (for now we assume that n can be approximated by the total number of characters in the alignment.)

3) AICc: The corrected AIC (AICc) is a modified version of the AIC that addresses the issue of small sample sizes in statistical model selection. It adjusts the AIC value to provide a more reliable estimate of model fit when the number of observations is relatively small compared to the number of parameters in the model. The AIC is given as,

$$AICc = AIC + \frac{2k(k+1)}{(n-k-1)},$$

where sample size is approximated by the total number of characters in the alignment.

4)Weighted AIC (W-AIC): W-AIC in the context of fitting hierarchical models using Markov Chain Monte Carlo (MCMC) methods. While traditional AIC is based on the maximized likelihood, W-AIC takes into account the full posterior distribution of model parameters.

5)Weighted BIC (W-BIC): The W-BIC is a variation of the BIC that incorporates additional weights to account for specific considerations in model selection. These weights are assigned to different components of the BIC formula to emphasize certain aspects of model evaluation. The standard BIC formula is given by:

$$BIC = -2(log - likelihood) + klog(n),$$

where log-likelihood is the logarithm of the likelihood function of the model, k is the number of free parameters in the model and n is the sample size.

6)Weighted AICc (W-AICc): W-AICc is a further extension of the AICc that incorporates weights to account for the possibility of model selection uncertainty. Instead of selecting a single best model based on the AICc, weighted AICc provides a weighted average of multiple models, taking into account their relative fit to the data and the complexity of each model. The weights reflect the likelihood that each model is the best one among the set of candidate models.

2.4 Marginal likelihood calculations using path sampling and stepping stone sampling

Path sampling and stepping stone sampling are both methods used in Bayesian statistical analysis to estimate marginal likelihoods and to compute Bayes factors. These techniques are particularly relevant when comparing different models or hypothesis testing in the context of Bayesian model selection. Accurate estimates of marginal likelihoods can be obtained using stepping-stone sampling and path-sampling techniques.
The marginal likelihood is the probability of the observed data X under a given model M_i that is averaged over all possible values of the parameters of the model θ_i with respect to the prior density on θ_i

$$P(X|M_i) = \int P(X|\theta_i)P(\theta_i)dt$$

These methods require Markov chain Monte Carlo samples from a range of power posterior distributions, which represent the path between the prior and the posterior distributions.

- **Path sampling** Path sampling involves estimating the marginal likelihood for every power posterior sample. Marginal likelihood represents the probability of the observed data under a specific model, integrating over all possible values of the model's parameters. It serves as a measure of the goodness of fit of a model to the data and can be used for model comparison.

- **Stepping-stone sampling** Stepping-stone sampling estimates the likelihood ratio for a series of power posterior samples. It is a variation of path sampling that employs multiple intermediate distributions between the prior and posterior. Each intermediate distribution is obtained by progressively weighting the likelihood component of the posterior distribution. The accuracy of these two methods is very similar, but stepping-stone sampling is more computationally efficient because it requires fewer samples from the power posterior.

Although both of these methods are more accurate than the harmonic-mean estimator, they come at the cost of a substantially increased computational burden.

Chapter 3

Data Analysis

3.1 Nucleotide Substitution Models

- Jukes-Cantor (JC) Model:

  This is a simple and widely used model that assumes all nucleotide substitutions occur at equal rates. It assumes a single parameter for the overall substitution rate and does not account for the differences in mutation rates between different types of nucleotide substitutions.

- Hasegawa-Kishino-Yano (HKY) Model:

  The HKY model is a variant of the General Time Reversible (GTR) model that assumes a stationary state for nucleotide frequencies but allows for different substitution rates for transitions and transversions. It is commonly used for phylogenetic analyses and incorporates more parameters than the JC or Kimura models.

- Tamura-Nei (TN93) Model:

  This model of substitution assumes that nucleotide frequencies may vary among different sequences or lineages. In addition to nucleotide frequencies, the TN93 model considers different rates for transitions (purine-to-purine or pyrimidine-to-pyrimidine substitutions) and transversions (purine-to-pyrimidine or vice versa substitutions). It assumes that the transition rate is different from the transversion rate, acknowledging the known bias in mutation rates observed in nucleotide substitutions.

- General Time Reversible (GTR) Model:
  This model considers six different substitution rates, one for each possible pair of nucleotides.

  $$(A \longleftrightarrow C, A \longleftrightarrow G, A \longleftrightarrow T, C \longleftrightarrow G, C \longleftrightarrow T, G \longleftrightarrow T)$$

  allowing for a more detailed representation of the evolutionary processes. It also takes into account rate heterogeneity among different sites in a sequence.

- Transitional Interchangeability Matrix (TIM) Model:
  The TIM model assumes that substitutions occur at different rates depending on the type of nucleotide or amino acid being replaced and the surrounding context. The TIM model calculates the substitution probabilities between different nucleotides or amino acids based on observed patterns of substitution in a given dataset.

3.2 Molecular Clocks

1. Strict Clock:

   Constant rate among branches.

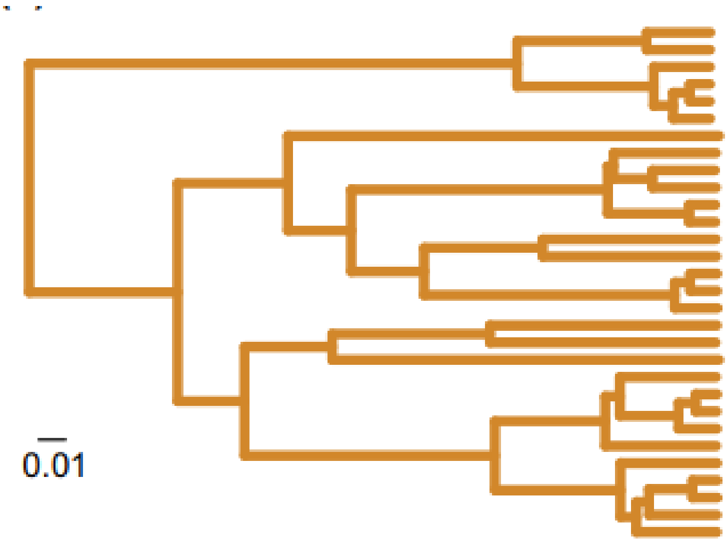


Figure 3.1: Strict Clock

2. Uncorrelated Relax Clock:

   Distinct rate along each branch drawn from a chosen probability distribution.

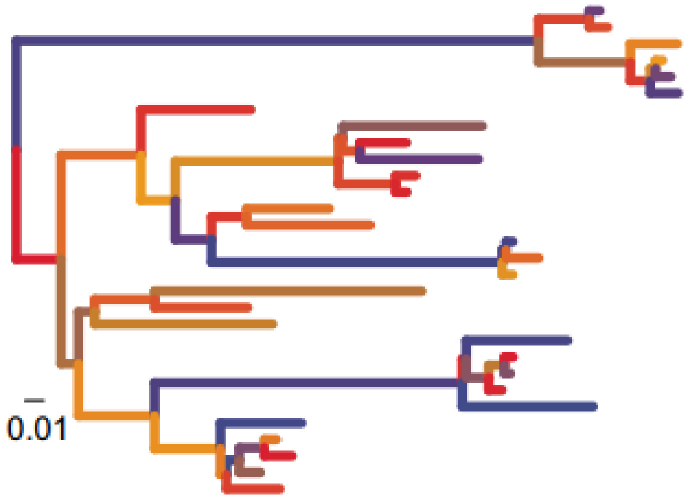


Figure 3.2: Uncorrelated Relax Clock

3. Fixed Local Clock:

   Branches having the same rate are clustered together.

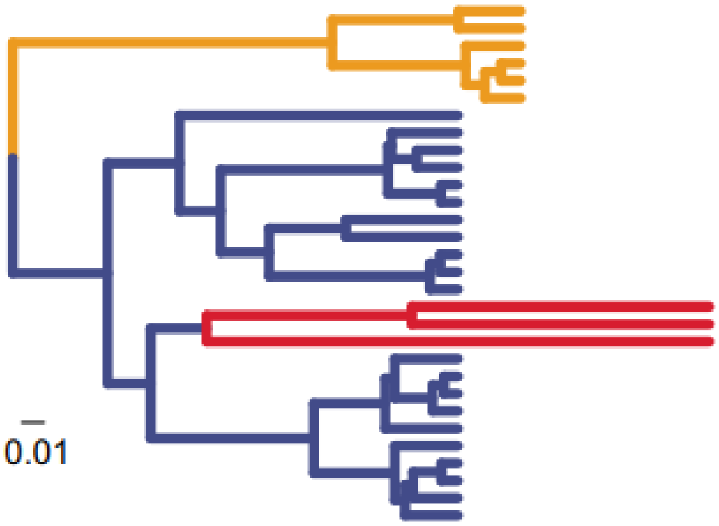


Figure 3.3: Fixed Local Clock

4. Random Local Clock:

   Branches having the same rates are distributed throughout the tree rather than being locally clustered.

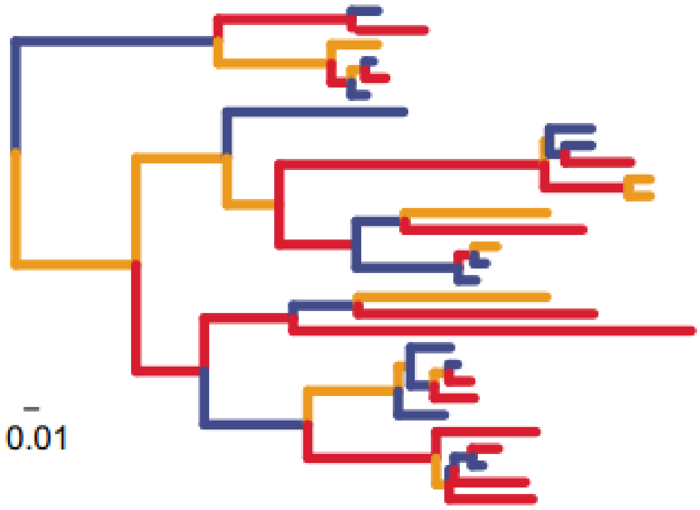


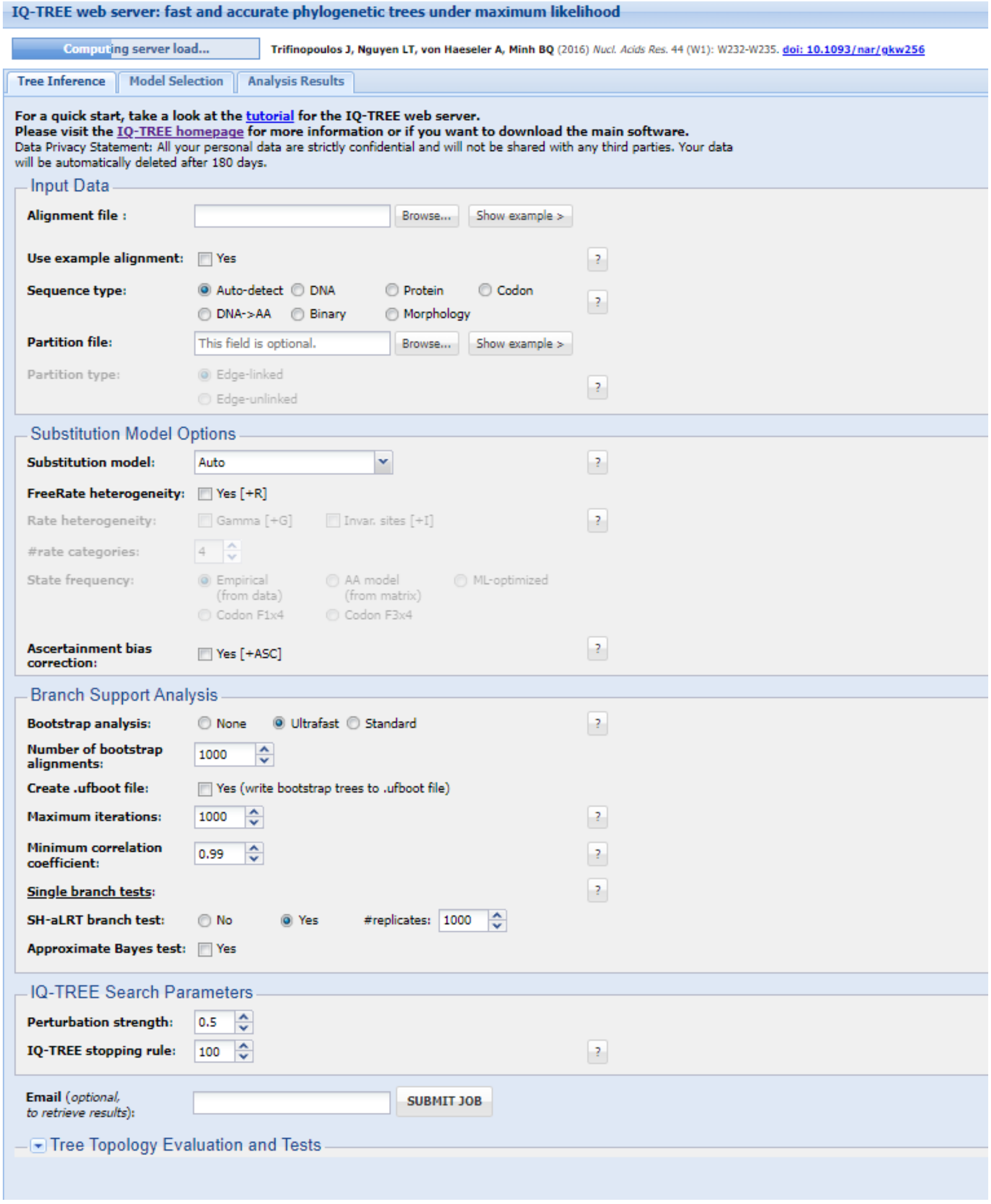
Figure 3.4: Random Local Clock

3.3 Work flow for analysis

3.3.1 IQTREE

IQTREE is a fast and effective stochastic algorithm to infer phylogenetic trees by maximum likelihood.

1) Choose the alignment file and then click on submit the job.



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

Computing server load...

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) *Nucl. Acids Res.* 44 (W1): W232-W235. doi: 10.1093/nar/gkw256

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
Create .ufboot file:	□ Yes (write bootstrap trees to .ufboot file)
Maximum iterations:	1000
Minimum correlation coefficient:	0.99
Single branch tests:	
SH-aLRT branch test:	○ No ● Yes #replicates: 1000
Approximate Bayes test:	□ Yes

IQ-TREE Search Parameters

Perturbation strength:	0.5
IQ-TREE stopping rule:	100

Email (*optional, to retrieve results*): SUBMIT JOB

Tree Topology Evaluation and Tests

Figure 3.5: IQTREE

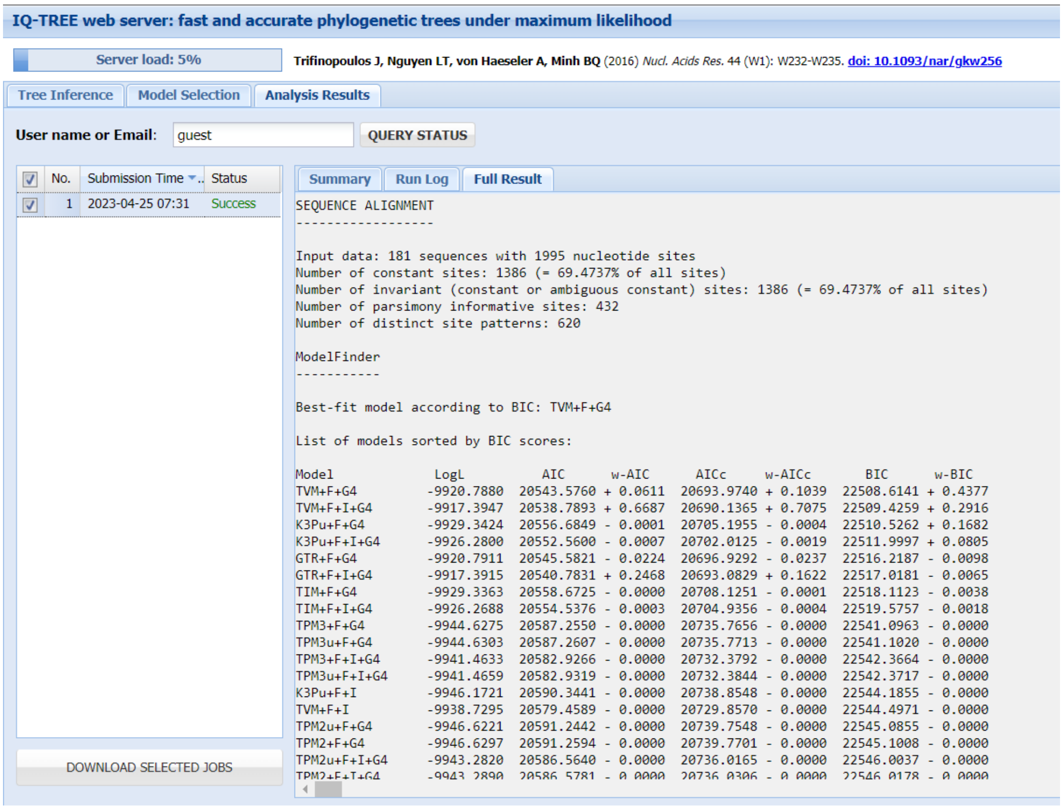
2) IQTREE scores each model on the basis of AIC, w-AIC, AICc, w-AICc, BIC and w-BIC and the value corresponding to the least BIC value will be considered as the best model.

AIC, w-AIC: Akaike information criterion scores and weights.

AICc, w-AICc: Corrected AIC scores and weights.

BIC, w-BIC: Bayesian information criterion scores and weights.

3) Select "DOWNLOAD SELECTED JOBS" at the lower left-hand corner. All the files required for further analysis will be hence downloaded.



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

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) *Nucl. Acids Res.* 44 (W1): W232-W235. doi: 10.1093/nar/gkw256

User name or Email: guest

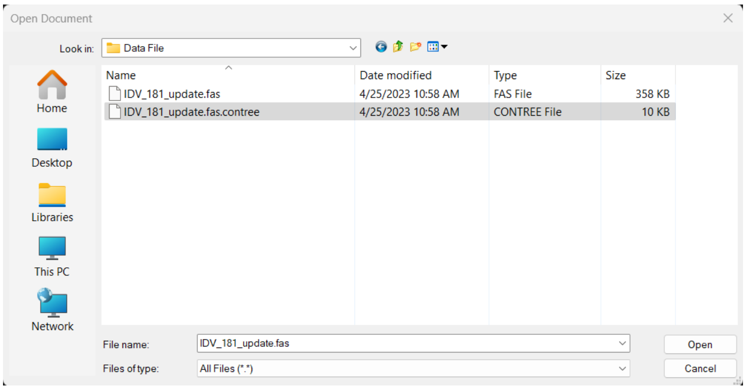
☑	No.	Submission Time	Status
☑	1	2023-04-25 07:31	Success

```
SEQUENCE ALIGNMENT
------------------

Input data: 181 sequences with 1995 nucleotide sites
Number of constant sites: 1386 (= 69.4737% of all sites)
Number of invariant (constant or ambiguous constant) sites: 1386 (= 69.4737% of all sites)
Number of parsimony informative sites: 432
Number of distinct site patterns: 620

ModelFinder
-----------

Best-fit model according to BIC: TVM+F+G4

List of models sorted by BIC scores:

Model                  LogL          AIC      w-AIC        AICc     w-AICc         BIC      w-BIC
TVM+F+G4          -9920.7880  20543.5760 + 0.0611  20693.9740 + 0.1039  22508.6141 + 0.4377
TVM+F+I+G4        -9917.3947  20538.7893 + 0.6687  20690.1365 + 0.7075  22509.4259 + 0.2916
K3Pu+F+G4         -9929.3424  20556.6849 - 0.0001  20705.1955 - 0.0004  22510.5262 + 0.1682
K3Pu+F+I+G4       -9926.2800  20552.5600 - 0.0007  20702.0125 - 0.0019  22511.9997 + 0.0805
GTR+F+G4          -9920.7911  20545.5821 - 0.0224  20696.9292 - 0.0237  22516.2187 - 0.0098
GTR+F+I+G4        -9917.3915  20540.7831 + 0.2468  20693.0829 + 0.1622  22517.0181 - 0.0065
TIM+F+G4          -9929.3363  20558.6725 - 0.0000  20708.1251 - 0.0001  22518.1123 - 0.0038
TIM+F+I+G4        -9926.2688  20554.5376 - 0.0003  20704.9356 - 0.0004  22519.5757 - 0.0018
TPM3+F+G4         -9944.6275  20587.2550 - 0.0000  20735.7656 - 0.0000  22541.0963 - 0.0000
TPM3u+F+G4        -9944.6303  20587.2607 - 0.0000  20735.7713 - 0.0000  22541.1020 - 0.0000
TPM3+F+I+G4       -9941.4633  20582.9266 - 0.0000  20732.3792 - 0.0000  22542.3664 - 0.0000
TPM3u+F+I+G4      -9941.4659  20582.9319 - 0.0000  20732.3844 - 0.0000  22542.3717 - 0.0000
K3Pu+F+I          -9946.1721  20590.3441 - 0.0000  20738.8548 - 0.0000  22544.1855 - 0.0000
TVM+F+I           -9938.7295  20579.4589 - 0.0000  20729.8570 - 0.0000  22544.4971 - 0.0000
TPM2u+F+G4        -9946.6221  20591.2442 - 0.0000  20739.7548 - 0.0000  22545.0855 - 0.0000
TPM2+F+G4         -9946.6297  20591.2594 - 0.0000  20739.7701 - 0.0000  22545.1008 - 0.0000
TPM2u+F+I+G4      -9943.2820  20586.5640 - 0.0000  20736.0165 - 0.0000  22546.0037 - 0.0000
TPM2+F+I+G4       -9943.2890  20586.5781 - 0.0000  20736.0306 - 0.0000  22546.0178 - 0.0000
```

Figure 3.6: IQTREE

3.3.2 TEMPEST

1. When started, TempEst displays a file selection dialog box in which you can select the contree file that you downloaded in the previous section.



Open Document

Look in: Data File

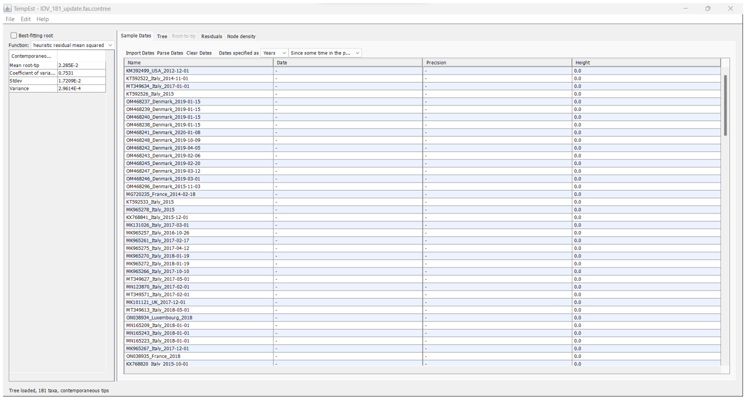
Name	Date modified	Type	Size
IDV_181_update.fas	4/25/2023 10:58 AM	FAS File	358 KB
IDV_181_update.fas.contree	4/25/2023 10:58 AM	CONTREE File	10 KB

File name: IDV_181_update.fas

Files of type: All Files (*.*)

Figure 3.7: TEMPEST 1

2. Once the tree is loaded the main window appears:



TempEst - IDV_181_update.fas.contree

☐ Best-fitting root

Function: heuristic residual mean squared

Contemporaneo...	
Mean root-tip	2.285E-2
Coefficient of varia...	0.7531
Stdev	1.7209E-2
Variance	2.9614E-4

Dates specified as Years Since some time in the p...

Name	Date	Precision	Height
KM392499_USA_2012-12-01	-	-	0.0
KT592522_Italy_2014-11-01	-	-	0.0
MT349634_Italy_2017-01-01	-	-	0.0
KT592526_Italy_2015	-	-	0.0
OM468237_Denmark_2019-01-15	-	-	0.0
OM468239_Denmark_2019-01-15	-	-	0.0
OM468240_Denmark_2019-01-15	-	-	0.0
OM468238_Denmark_2019-01-15	-	-	0.0
OM468241_Denmark_2020-01-08	-	-	0.0
OM468248_Denmark_2019-10-09	-	-	0.0
OM468242_Denmark_2019-04-05	-	-	0.0
OM468243_Denmark_2019-02-06	-	-	0.0
OM468245_Denmark_2019-02-20	-	-	0.0
OM468247_Denmark_2019-03-12	-	-	0.0
OM468246_Denmark_2019-03-01	-	-	0.0
OM468296_Denmark_2015-11-03	-	-	0.0
MG720235_France_2014-02-18	-	-	0.0
KT592533_Italy_2015	-	-	0.0
MK965278_Italy_2015	-	-	0.0
KX768841_Italy_2015-12-01	-	-	0.0
MK131026_Italy_2017-03-01	-	-	0.0
MK965257_Italy_2016-10-26	-	-	0.0
MK965261_Italy_2017-02-17	-	-	0.0
MK965275_Italy_2017-04-12	-	-	0.0
MK965270_Italy_2018-01-19	-	-	0.0
MK965272_Italy_2018-01-19	-	-	0.0
MK965266_Italy_2017-10-10	-	-	0.0
MT349627_Italy_2017-05-01	-	-	0.0
MN123870_Italy_2017-02-01	-	-	0.0
MT349571_Italy_2017-02-01	-	-	0.0
MK101121_UK_2017-12-01	-	-	0.0
MT349613_Italy_2018-05-01	-	-	0.0
ON038934_Luxembourg_2018	-	-	0.0
MN165209_Italy_2018-01-01	-	-	0.0
MN165243_Italy_2018-01-01	-	-	0.0
MN165223_Italy_2018-01-01	-	-	0.0
MK965267_Italy_2017-12-01	-	-	0.0
ON038935_France_2018	-	-	0.0
KX768820_Italy_2015-10-01	-	-	0.0

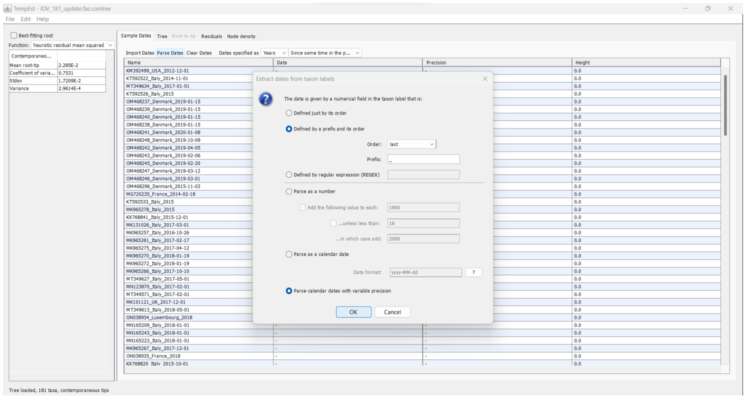
Tree loaded, 181 taxa, contemporaneous tips

Figure 3.8: TEMPEST 2

3. Parsing dates of sampling:

   This dialog box provides a wide range of other options for extracting dates from taxon labels.

   - Defined by a prefix and its order: This option finds fields that are prefixed by a particular character or string and then uses the order to specify which one. Here, the year that each virus was isolated is given at the end of the labels; hence, the order is last and separated.



☐ Best-fitting root

Function: heuristic residual mean squared

Dates specified as: Years; Since some time in the p...

Contemporaneo...	
Mean root-tip	2.285E-2
Coefficient of varia...	0.7531
Stdev	1.7209E-2
Variance	2.9614E-4

Extract dates from taxon labels

The date is given by a numerical field in the taxon label that is:

○ Defined just by its order

● Defined by a prefix and its order

Order: last

Prefix: _

○ Defined by regular expression (REGEX)

○ Parse as a number

☐ Add the following value to each: 1900

☐ ...unless less than: 16

...in which case add: 2000

○ Parse as a calendar date

Date format: yyyy-MM-dd ?

● Parse calendar dates with variable precision

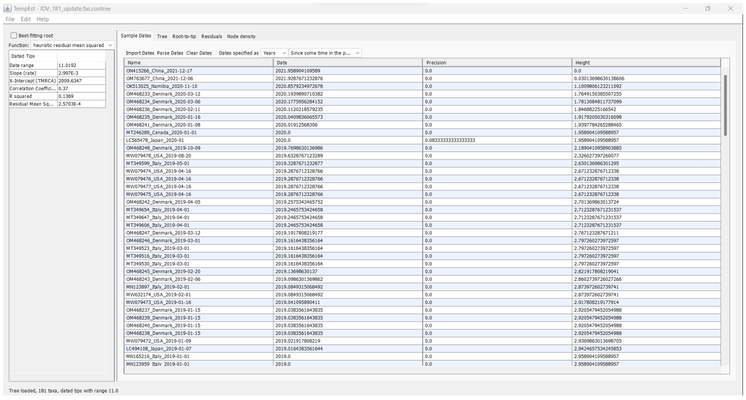
Name	Date	Precision	Height
KM392499_USA_2012-12-01	-	-	0.0
KT592522_Italy_2014-11-01			0.0
MT349634_Italy_2017-01-01			0.0
KT592526_Italy_2015			0.0
OM468237_Denmark_2019-01-15			0.0
OM468239_Denmark_2019-01-15			0.0
OM468240_Denmark_2019-01-15			0.0
OM468238_Denmark_2019-01-15			0.0
OM468241_Denmark_2020-01-08			0.0
OM468248_Denmark_2019-10-09			0.0
OM468242_Denmark_2019-04-05			0.0
OM468243_Denmark_2019-02-06			0.0
OM468245_Denmark_2019-02-20			0.0
OM468247_Denmark_2019-03-12			0.0
OM468246_Denmark_2019-03-01			0.0
OM468296_Denmark_2015-11-03			0.0
MG720235_France_2014-02-18			0.0
KT592533_Italy_2015			0.0
MK965278_Italy_2015			0.0
KX768841_Italy_2015-12-01			0.0
MK131026_Italy_2017-03-01			0.0
MK965257_Italy_2016-10-26			0.0
MK965261_Italy_2017-02-17			0.0
MK965275_Italy_2017-04-12			0.0
MK965270_Italy_2018-01-19			0.0
MK965272_Italy_2018-01-19			0.0
MK965266_Italy_2017-10-10			0.0
MT349627_Italy_2017-05-01			0.0
MN123870_Italy_2017-02-01			0.0
MT349571_Italy_2017-02-01			0.0
MK101121_UK_2017-12-01			0.0
MT349613_Italy_2018-05-01			0.0
ON038934_Luxembourg_2018			0.0
MN165209_Italy_2018-01-01	-	-	0.0
MN165243_Italy_2018-01-01	-	-	0.0
MN165223_Italy_2018-01-01	-	-	0.0
MK965267_Italy_2017-12-01	-	-	0.0
ON038935_France_2018	-	-	0.0
KX768820_Italy_2015-10-01	-	-	0.0

Tree loaded, 181 taxa, contemporaneous tips

Figure 3.9: TEMPEST 3

   - The parsed dates are as follows:

     The Height column lists the ages of the tips relative to time 0.



☐ Best-fitting root

Function: heuristic residual mean squared

Dates specified as: Years; Since some time in the p...

Dated Tips	
Date range	11.0192
Slope (rate)	2.997E-3
X-Intercept (TMRCA)	2009.6347
Correlation Coeffici...	0.37
R squared	0.1369
Residual Mean Sq...	2.5703E-4

Name	Date	Precision	Height
ON415266_China_2021-12-17	2021.958904109589	0.0	0.0
OM763677_China_2021-12-06	2021.9287671232876	0.0	0.03013698630138606
OK513025_Namibia_2020-11-10	2020.8579234972678	0.0	1.1009806123211092
OM468233_Denmark_2020-03-12	2020.1939890710382	0.0	1.7649150385507255
OM468234_Denmark_2020-03-06	2020.1775956284152	0.0	1.7813084811737099
OM468236_Denmark_2020-02-11	2020.1120218579235	0.0	1.84688225166542
OM468235_Denmark_2020-01-16	2020.0409836065573	0.0	1.9179205030316098
OM468241_Denmark_2020-01-08	2020.01912568306	0.0	1.9397784265288465
MT246289_Canada_2020-01-01	2020.0	0.0	1.958904109588957
LC565478_Japan_2020-01	2020.0	0.08333333333333333	1.958904109588957
OM468248_Denmark_2019-10-09	2019.7698630136986	0.0	2.1890410958903885
MW079478_USA_2019-08-20	2019.6328767123289	0.0	2.326027397260077
MT349599_Italy_2019-05-01	2019.3287671232877	0.0	2.630136986301295
MW079474_USA_2019-04-16	2019.2876712328766	0.0	2.671232876712338
MW079476_USA_2019-04-16	2019.2876712328766	0.0	2.671232876712338
MW079477_USA_2019-04-16	2019.2876712328766	0.0	2.671232876712338
MW079475_USA_2019-04-16	2019.2876712328766	0.0	2.671232876712338
OM468242_Denmark_2019-04-05	2019.2575342465752	0.0	2.701369863013724
MT349654_Italy_2019-04-01	2019.2465753424658	0.0	2.7123287671231537
MT349647_Italy_2019-04-01	2019.2465753424658	0.0	2.7123287671231537
MT349606_Italy_2019-04-01	2019.2465753424658	0.0	2.7123287671231537
OM468247_Denmark_2019-03-12	2019.1917808219177	0.0	2.767123287671211
OM468246_Denmark_2019-03-01	2019.1616438356164	0.0	2.797260273972597
MT349523_Italy_2019-03-01	2019.1616438356164	0.0	2.797260273972597
MT349516_Italy_2019-03-01	2019.1616438356164	0.0	2.797260273972597
MT349530_Italy_2019-03-01	2019.1616438356164	0.0	2.797260273972597
OM468245_Denmark_2019-02-20	2019.13698630137	0.0	2.821917808219041
OM468243_Denmark_2019-02-06	2019.0986301369862	0.0	2.8602739726027266
MN123897_Italy_2019-02-01	2019.0849315068492	0.0	2.873972602739741
MW632174_USA_2019-02-01	2019.0849315068492	0.0	2.873972602739741
MW079473_USA_2019-01-16	2019.041095890411	0.0	2.917808219177914
OM468237_Denmark_2019-01-15	2019.0383561643835	0.0	2.9205479452054988
OM468239_Denmark_2019-01-15	2019.0383561643835	0.0	2.9205479452054988
OM468240_Denmark_2019-01-15	2019.0383561643835	0.0	2.9205479452054988
OM468238_Denmark_2019-01-15	2019.0383561643835	0.0	2.9205479452054988
MW079472_USA_2019-01-09	2019.021917808219	0.0	2.9369863013698705
LC494108_Japan_2019-01-07	2019.0164383561644	0.0	2.9424657534245853
MN165216_Italy_2019-01-01	2019.0	0.0	2.958904109588957
MN123959_Italy_2019-01-01	2019.0	0.0	2.958904109588957

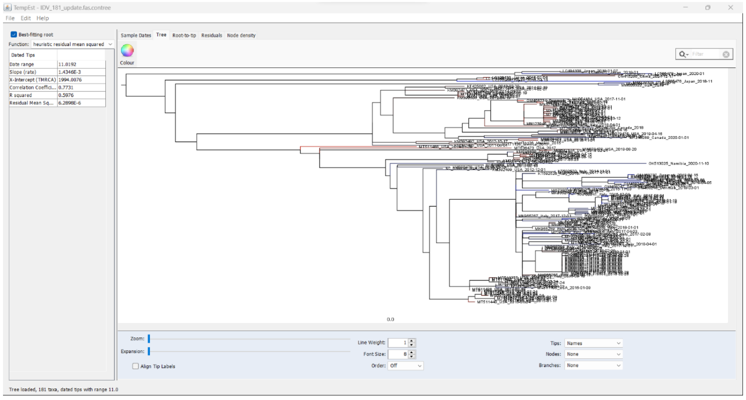
Tree loaded, 181 taxa, dated tips with range 11.0

Figure 3.10: TEMPEST 4

4. The temporal signal and rooting

   We can explore data using the tabs at the top of the window-Tree, Root-to-tip, and Residuals.

   - Tree - Since the tree is constructed using a non-molecular clock model, it will be arbitrarily rooted.



☑ Best-fitting root

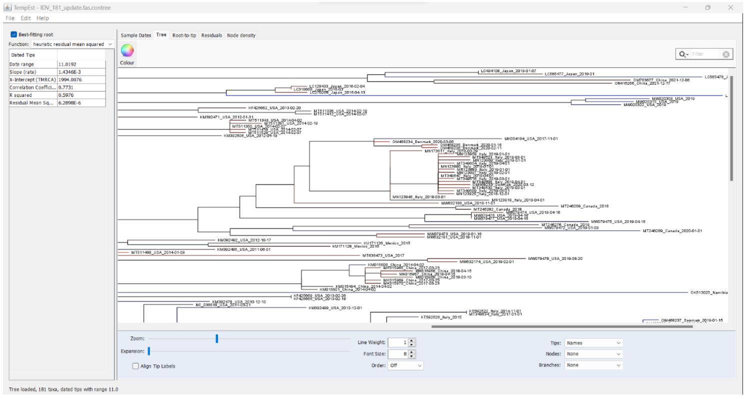
Function: heuristic residual mean squared

Dated Tips	
Date range	11.0192
Slope (rate)	1.4346E-3
X-Intercept (TMRCA)	1994.0076
Correlation Coeffici...	0.7731
R squared	0.5976
Residual Mean Sq...	6.2898E-6

Zoom:
Expansion:
☐ Align Tip Labels
Line Weight: 1
Font Size: 8
Order: Off
Tips: Names
Nodes: None
Branches: None

Tree loaded, 181 taxa, dated tips with range 11.0

Figure 3.11: TEMPEST 5



☑ Best-fitting root

Function: heuristic residual mean squared

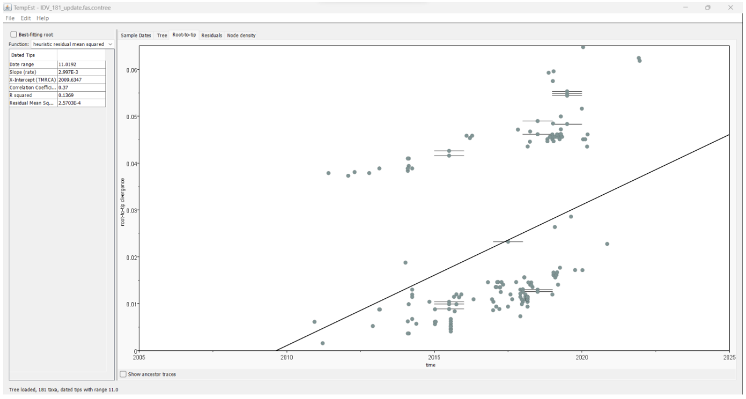
Dated Tips	
Date range	11.0192
Slope (rate)	1.4346E-3
X-Intercept (TMRCA)	1994.0076
Correlation Coeffici...	0.7731
R squared	0.5976
Residual Mean Sq...	6.2898E-6

Zoom:
Expansion:
☐ Align Tip Labels
Line Weight: 1
Font Size: 8
Order: Off
Tips: Names
Nodes: None
Branches: None

Tree loaded, 181 taxa, dated tips with range 11.0

Figure 3.12: TEMPEST 6

   - Root-to-tip panel shows a plot of the divergence from the root of the tree against the time of sampling.



□ Best-fitting root

Function: heuristic residual mean squared

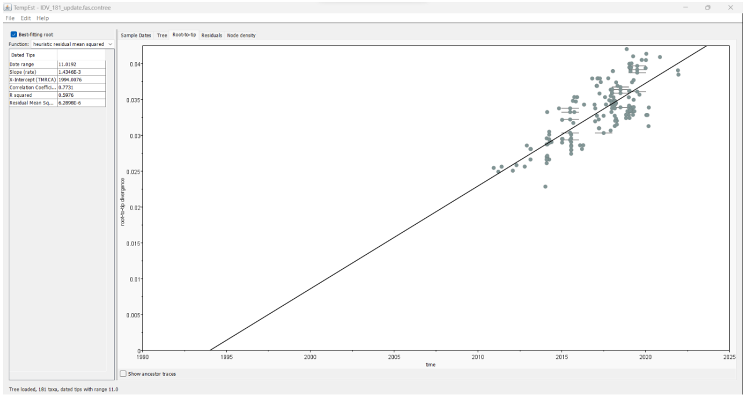
Dated Tips	
Date range	11.0192
Slope (rate)	2.997E-3
X-Intercept (TMRCA)	2009.6347
Correlation Coeffici...	0.37
R squared	0.1369
Residual Mean Sq...	2.5703E-4

□ Show ancestor traces

Tree loaded, 181 taxa, dated tips with range 11.0

Figure 3.13: TEMPEST 7

- By clicking on the Best-fitting root, it minimizes the mean of the squares of the residuals.



☑ Best-fitting root

Function: heuristic residual mean squared

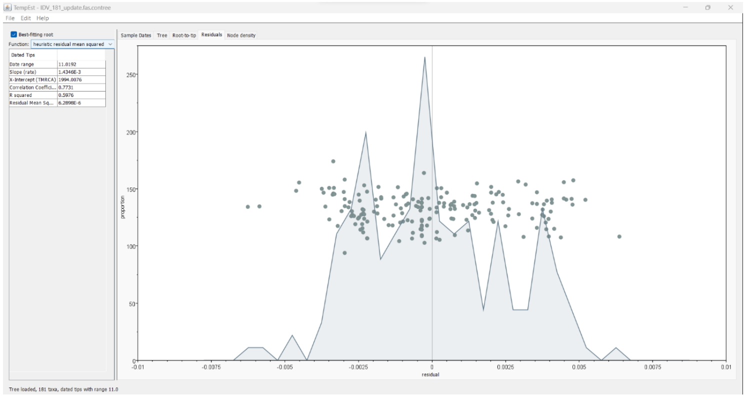
Dated Tips	
Date range	11.0192
Slope (rate)	1.4346E-3
X-Intercept (TMRCA)	1994.0076
Correlation Coeffici...	0.7731
R squared	0.5976
Residual Mean Sq...	6.2898E-6

□ Show ancestor traces

Tree loaded, 181 taxa, dated tips with range 11.0

Figure 3.14: TEMPEST 8

- If one switches to the Residual panel you will see a plot of all the residuals.



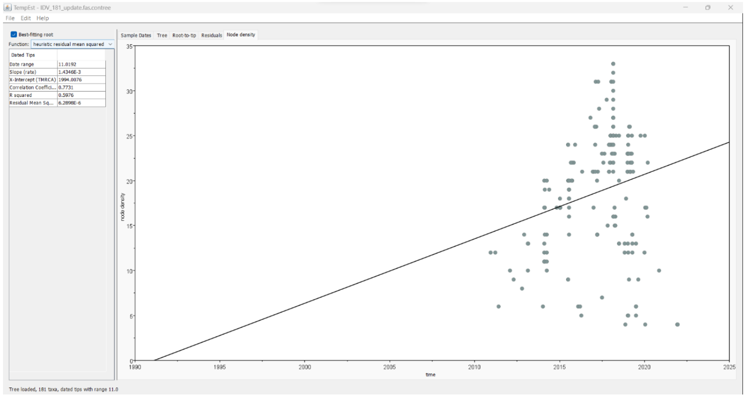
☑ Best-fitting root

Function: heuristic residual mean squared

Dated Tips	
Date range	11.0192
Slope (rate)	1.4346E-3
X-Intercept (TMRCA)	1994.0076
Correlation Coeffici...	0.7731
R squared	0.5976
Residual Mean Sq...	6.2898E-6

Tree loaded, 181 taxa, dated tips with range 11.0

Figure 3.15: TEMPEST 9



☑ Best-fitting root

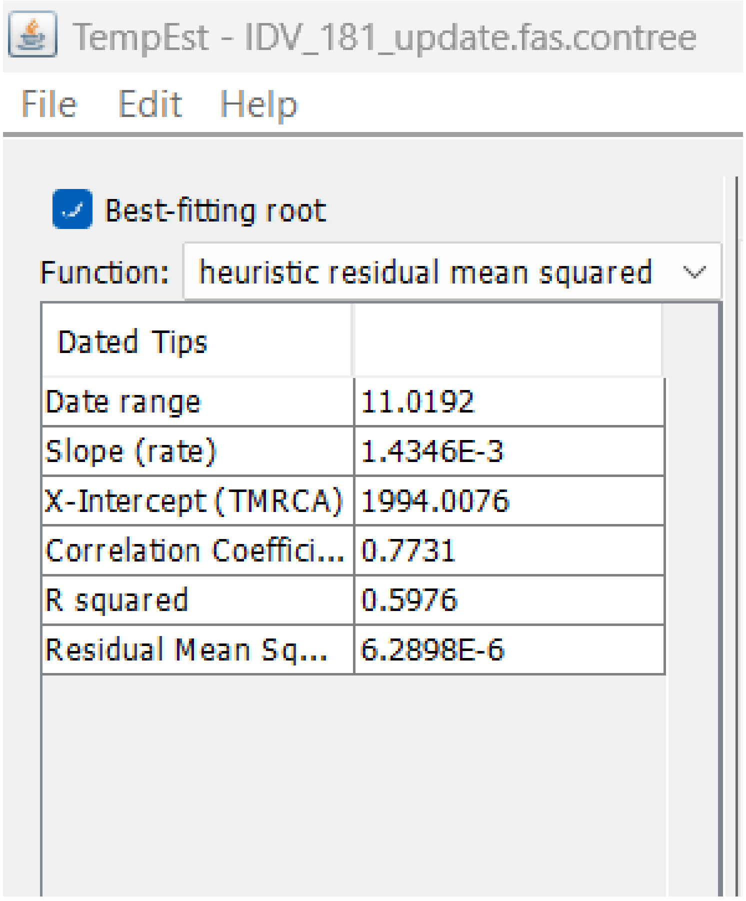
Function: heuristic residual mean squared

Dated Tips	
Date range	11.0192
Slope (rate)	1.4346E-3
X-Intercept (TMRCA)	1994.0076
Correlation Coeffici...	0.7731
R squared	0.5976
Residual Mean Sq...	6.2898E-6

Tree loaded, 181 taxa, dated tips with range 11.0

Figure 3.16: TEMPEST 10

- On the left-hand side of the window there is a table of statistics



☑ Best-fitting root

Function: heuristic residual mean squared

Dated Tips	
Date range	11.0192
Slope (rate)	1.4346E-3
X-Intercept (TMRCA)	1994.0076
Correlation Coeffici...	0.7731
R squared	0.5976
Residual Mean Sq...	6.2898E-6

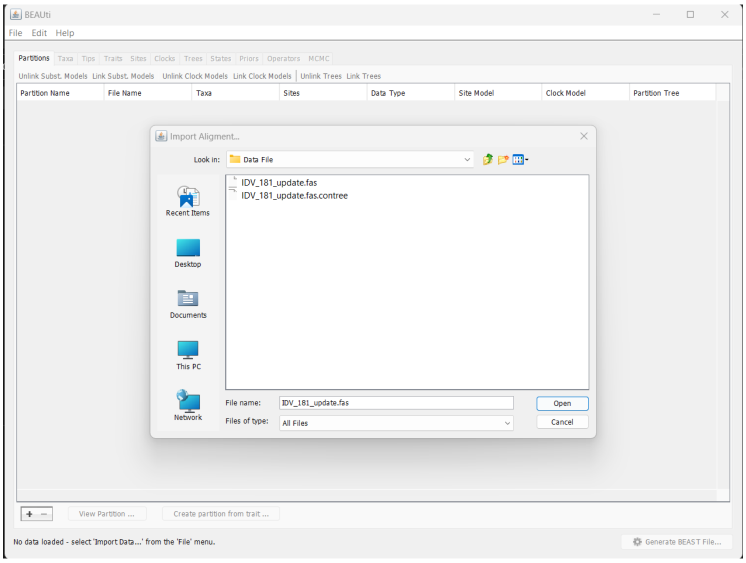
Figure 3.17: TEMPEST 11

- Slope(rate) is an estimate of the rate of evolution in substitutions per site per year.
- X-Intercept is an estimate of the date of the root of the tree.

3.3.3 BEAUti

Following are the steps for generating a BEAST XML input file:

1. Loading the NEXUS file: To load a NEXUS format alignment, simply select the Import Alignment option from the File menu.


BEAUti
File Edit Help
Partitions Taxa Tips Traits Sites Clocks Trees States Priors Operators MCMC
Unlink Subst. Models Link Subst. Models Unlink Clock Models Link Clock Models Unlink Trees Link Trees
Partition Name File Name Taxa Sites Data Type Site Model Clock Model Partition Tree
Import Aligment...
Recent Items
Desktop
Documents
This PC
Network
Open
Cancel
View Partition ...
Create partition from trait ...
Generate BEAST File...

Look in: Data File

IDV_181_update.fas
IDV_181_update.fas.contree

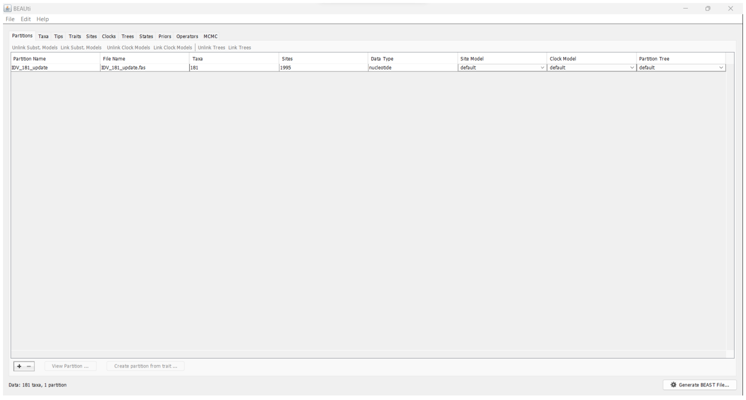
File name: IDV_181_update.fas

Files of type: All Files

No data loaded - select 'Import Data...' from the 'File' menu.

Figure 3.18: BEAUti 1

Once loaded, the alignment will be displayed in the main window in a table:


BEAUti
File Edit Help
Partitions Taxa Tips Traits Sites Clocks Trees States Priors Operators MCMC
Unlink Subst. Models Link Subst. Models Unlink Clock Models Link Clock Models Unlink Trees Link Trees
View Partition ...
Create partition from trait ...
Generate BEAST File...

Partition Name	File Name	Taxa	Sites	Data Type	Site Model	Clock Model	Partition Tree
IDV_181_update	IDV_181_update.fas	181	1995	nucleotide	default	default	default

Data: 181 taxa, 1 partition

Figure 3.19: BEAUti 2

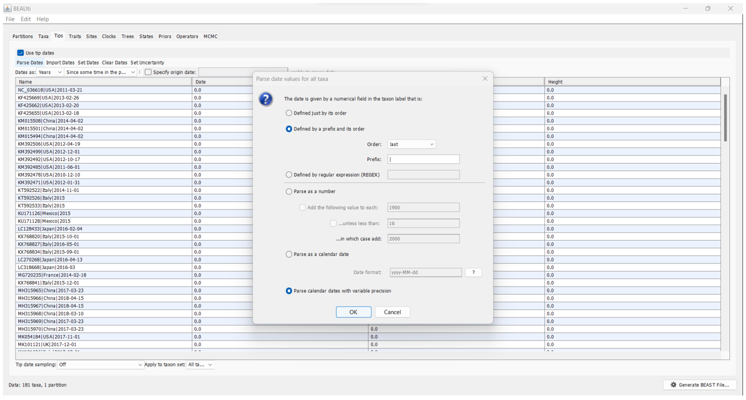
2. Setting the dates of the taxa:

   In the Tips options, you will see a table with all of the taxa that were in alignment. This panel allows you to give the taxa dates.

   - Here, the year that each virus was isolated is given at the end of the labels and hence the order is last and is separated by.

- Parse calendar date with variable precision: This parses a very specific way of specifying dates which are by far the most reliable. This format is the equivalent of using yyyy-MM-dd in the box above. This format has an additional advantage in that the day can be omitted if it is not known. Both the day and the month can be omitted if only the year is known.

  When using this option cases where the day or day and month are omitted are noted and the uncertainty is recorded in the precision column.



BEAUti
File Edit Help
Partitions Taxa Tips Traits Sites Clocks Trees States Priors Operators MCMC
Parse Dates Import Dates Set Dates Clear Dates Set Uncertainty
OK
Cancel
Generate BEAST File...

☑ Use tip dates

Dates as: Years · Since some time in the p... · □ Specify origin date:

Name	Date	Height
NC_036618|USA|2011-03-21	0.0	0.0
KF425669|USA|2013-02-26	0.0	0.0
KF425662|USA|2013-02-20	0.0	0.0
KF425655|USA|2013-02-18	0.0	0.0
KM015508|China|2014-04-02	0.0	0.0
KM015501|China|2014-04-02	0.0	0.0
KM015494|China|2014-04-02	0.0	0.0
KM392506|USA|2012-04-19	0.0	0.0
KM392499|USA|2012-12-01	0.0	0.0
KM392492|USA|2012-10-17	0.0	0.0
KM392485|USA|2011-06-01	0.0	0.0
KM392478|USA|2010-12-10	0.0	0.0
KM392471|USA|2012-01-31	0.0	0.0
KT592522|Italy|2014-11-01	0.0	0.0
KT592526|Italy|2015	0.0	0.0
KT592533|Italy|2015	0.0	0.0
KU171126|Mexico|2015	0.0	0.0
KU171128|Mexico|2015	0.0	0.0
LC128433|Japan|2016-02-04	0.0	0.0
KX768820|Italy|2015-10-01	0.0	0.0
KX768827|Italy|2016-05-01	0.0	0.0
KX768834|Italy|2015-09-01	0.0	0.0
LC270268|Japan|2016-04-13	0.0	0.0
LC318668|Japan|2016-03	0.0	0.0
MG720235|France|2014-02-18	0.0	0.0
KX768841|Italy|2015-12-01	0.0	0.0
MH315965|China|2017-03-23	0.0	0.0
MH315966|China|2018-04-15	0.0	0.0
MH315967|China|2018-04-15	0.0	0.0
MH315968|China|2018-03-10	0.0	0.0
MH315969|China|2017-03-23	0.0	0.0
MH315970|China|2017-03-23	0.0	0.0
MK054184|USA|2017-11-01	0.0	0.0
MK101121|UK|2017-12-01	0.0	0.0

Tip date sampling: Off · Apply to taxon set: All ta...

Data: 181 taxa, 1 partition

Parse date values for all taxa

The date is given by a numerical field in the taxon label that is:

○ Defined just by its order

● Defined by a prefix and its order

Order: last

Prefix: |

○ Defined by regular expression (REGEX)

○ Parse as a number

□ Add the following value to each: 1900

□ ...unless less than: 16

...in which case add: 2000

○ Parse as a calendar date

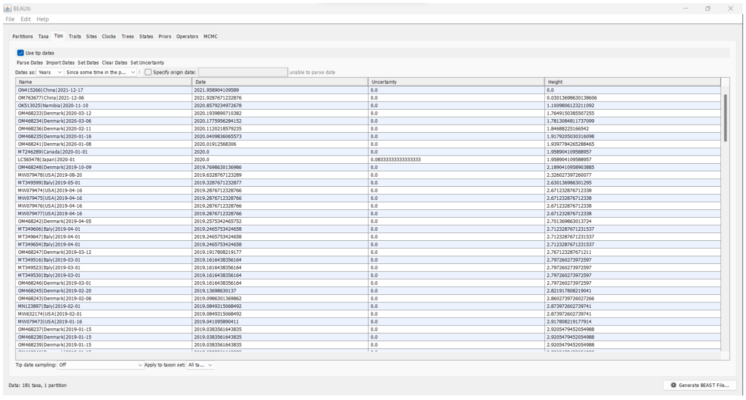
Date format: yyyy-MM-dd ?

● Parse calendar dates with variable precision

Figure 3.20: BEAUti 3

- The parsed dates are as follows:

  The Height column lists the ages of the tips relative to time 0.


BEAUti
File Edit Help
Partitions Taxa Tips Traits Sites Clocks Trees States Priors Operators MCMC
Parse Dates Import Dates Set Dates Clear Dates Set Uncertainty
Generate BEAST File...

☑ Use tip dates

Dates as: Years | Since some time in the p... | □ Specify origin date: | unable to parse date

Name	Date	Uncertainty	Height
ON415266|China|2021-12-17	2021.958904109589	0.0	0.0
OM763677|China|2021-12-06	2021.9287671232876	0.0	0.03013698630138606
OK513025|Namibia|2020-11-10	2020.8579234972678	0.0	1.1009806123211092
OM468233|Denmark|2020-03-12	2020.1939890710382	0.0	1.7649150385507255
OM468234|Denmark|2020-03-06	2020.1775956284152	0.0	1.7813084811737099
OM468236|Denmark|2020-02-11	2020.1120218579235	0.0	1.84688225166542
OM468235|Denmark|2020-01-16	2020.0409836065573	0.0	1.9179205030316098
OM468241|Denmark|2020-01-08	2020.01912568306	0.0	1.9397784265288465
MT246289|Canada|2020-01-01	2020.0	0.0	1.958904109588957
LC565478|Japan|2020-01	2020.0	0.08333333333333333	1.958904109588957
OM468248|Denmark|2019-10-09	2019.7698630136986	0.0	2.1890410958903885
MW079478|USA|2019-08-20	2019.6328767123289	0.0	2.326027397260077
MT349599|Italy|2019-05-01	2019.3287671232877	0.0	2.630136986301295
MW079474|USA|2019-04-16	2019.2876712328766	0.0	2.671232876712338
MW079475|USA|2019-04-16	2019.2876712328766	0.0	2.671232876712338
MW079476|USA|2019-04-16	2019.2876712328766	0.0	2.671232876712338
MW079477|USA|2019-04-16	2019.2876712328766	0.0	2.671232876712338
OM468242|Denmark|2019-04-05	2019.2575342465752	0.0	2.701369863013724
MT349606|Italy|2019-04-01	2019.2465753424658	0.0	2.7123287671231537
MT349647|Italy|2019-04-01	2019.2465753424658	0.0	2.7123287671231537
MT349654|Italy|2019-04-01	2019.2465753424658	0.0	2.7123287671231537
OM468247|Denmark|2019-03-12	2019.1917808219177	0.0	2.767123287671211
MT349516|Italy|2019-03-01	2019.1616438356164	0.0	2.797260273972597
MT349523|Italy|2019-03-01	2019.1616438356164	0.0	2.797260273972597
MT349530|Italy|2019-03-01	2019.1616438356164	0.0	2.797260273972597
OM468246|Denmark|2019-03-01	2019.1616438356164	0.0	2.797260273972597
OM468245|Denmark|2019-02-20	2019.13698630137	0.0	2.821917808219041
OM468243|Denmark|2019-02-06	2019.0986301369862	0.0	2.8602739726027266
MN123897|Italy|2019-02-01	2019.0849315068492	0.0	2.873972602739741
MW632174|USA|2019-02-01	2019.0849315068492	0.0	2.873972602739741
MW079473|USA|2019-01-16	2019.041095890411	0.0	2.917808219177914
OM468237|Denmark|2019-01-15	2019.0383561643835	0.0	2.9205479452054988
OM468238|Denmark|2019-01-15	2019.0383561643835	0.0	2.9205479452054988
OM468239|Denmark|2019-01-15	2019.0383561643835	0.0	2.9205479452054988

Tip date sampling: Off | Apply to taxon set: All ta...

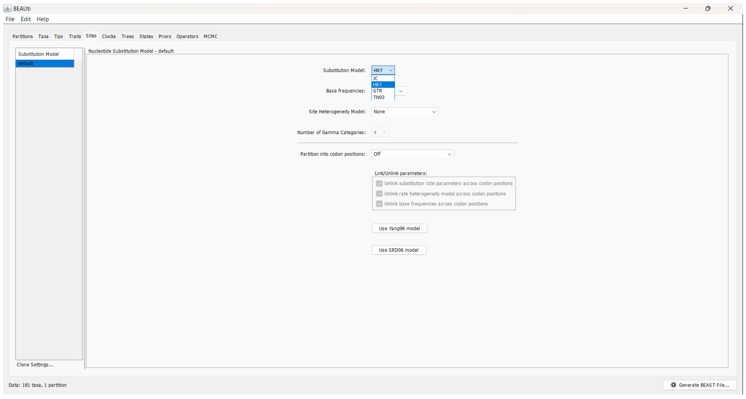
Data: 181 taxa, 1 partition

Figure 3.21: BEAUti 4

3. Setting the evolutionary model:

   In the Sites tab, you can set the model of molecular evolution for the sequence data you have loaded.

   - Substitution models specify the evolutionary process that describes how the data was generated on the tree.



Nucleotide Substitution Model - default

Substitution Model: HKY

Base frequencies:

Site Heterogeneity Model: None

Number of Gamma Categories: 4

Partition into codon positions: Off

Link/Unlink parameters:

☑ Unlink substitution rate parameters across codon positions

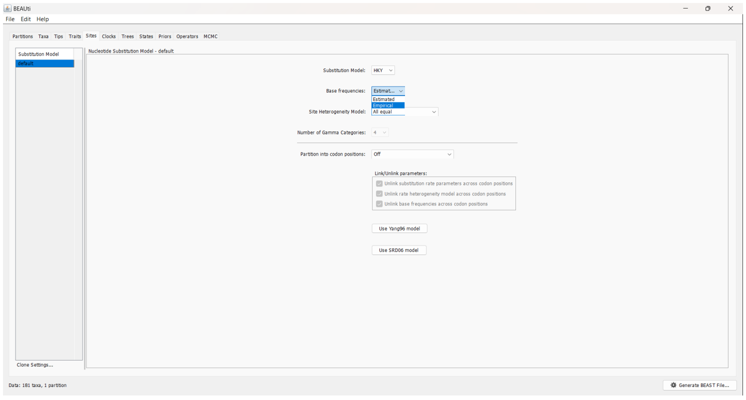
☑ Unlink rate heterogeneity model across codon positions

☑ Unlink base frequencies across codon positions

Data: 181 taxa, 1 partition

Figure 3.22: BEAUti 5

- Base frequencies:



Nucleotide Substitution Model - default

Substitution Model: HKY

Base frequencies: Estimat...

Site Heterogeneity Model: All equal

Number of Gamma Categories: 4

Partition into codon positions: Off

☑ Unlink substitution rate parameters across codon positions

☑ Unlink rate heterogeneity model across codon positions

☑ Unlink base frequencies across codon positions

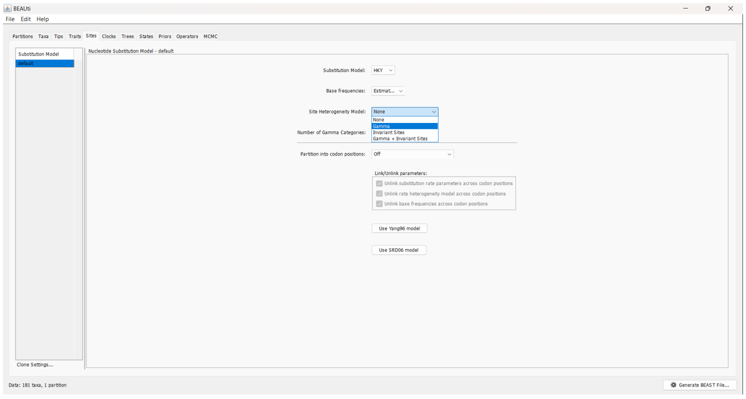
Data: 181 taxa, 1 partition

Figure 3.23: BEAUti 6

- Site Heterogeneity Model:

  Gamma distribution (G): gamma-distributed rate variation among sites

  The proportion of invariable sites (I): invariant sites across all taxa



Nucleotide Substitution Model - default

Substitution Model: HKY

Base frequencies: Estimat...

Site Heterogeneity Model: None

Number of Gamma Categories:

Partition into codon positions: Off

☑ Unlink substitution rate parameters across codon positions

☑ Unlink rate heterogeneity model across codon positions

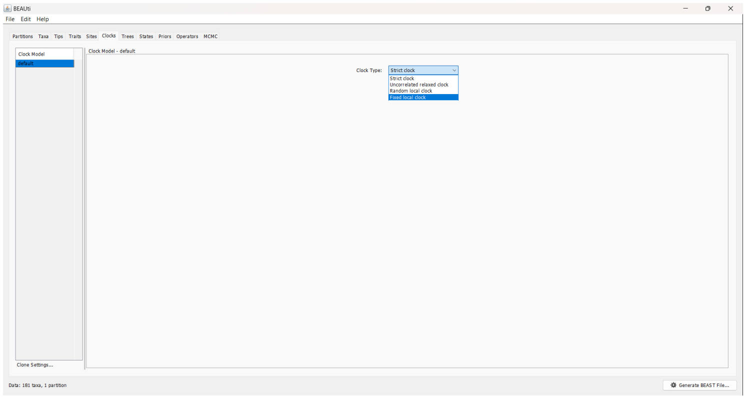
☑ Unlink base frequencies across codon positions

Data: 181 taxa, 1 partition

Figure 3.24: BEAUti 7

4. Setting the molecular clock model:

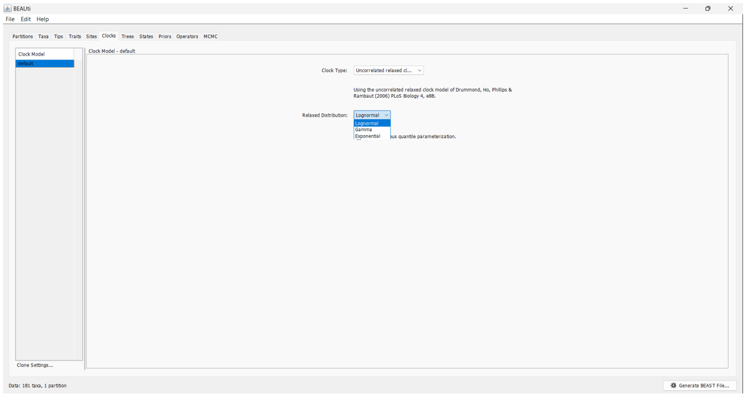
   In the next tab Clock we set the model of a molecular clock we will use. BEAST exclusively uses molecular clock models so that trees have a timescale.



Clock Type: Strict clock

Data: 181 taxa, 1 partition

Figure 3.25: BEAUti 8



Clock Type: Uncorrelated relaxed cl...

Using the uncorrelated relaxed clock model of Drummond, Ho, Phillips & Rambaut (2006) PLoS Biology 4, e88.

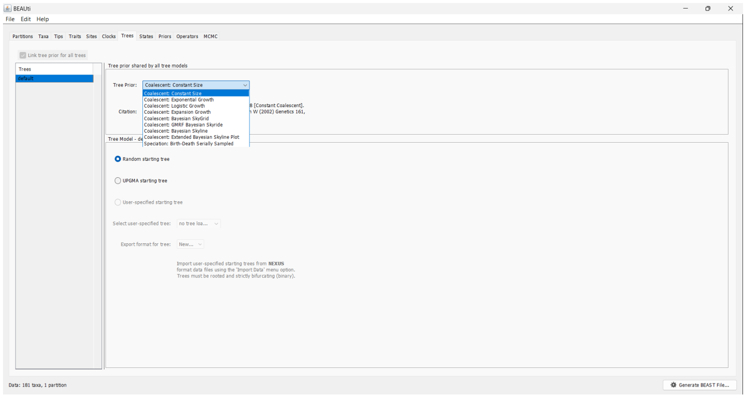
Relaxed Distribution: Lognormal

ous quantile parameterization.

Data: 181 taxa, 1 partition

Figure 3.26: BEAUti 9

5. Setting the tree prior: In this panel, one can set the model that provides a prior on the tree and some choices about the starting tree in the MCMC run.


BEAUti
File Edit Help
Partitions Taxa Tips Traits Sites Clocks Trees States Priors Operators MCMC
Trees
default
Generate BEAST File...

☑ Link tree prior for all trees

Tree prior shared by all tree models

Tree Prior: Coalescent: Constant Size

- Coalescent: Constant Size
- Coalescent: Exponential Growth
- Coalescent: Logistic Growth
- Coalescent: Expansion Growth
- Coalescent: Bayesian SkyGrid
- Coalescent: GMRF Bayesian Skyride
- Coalescent: Bayesian Skyline
- Coalescent: Extended Bayesian Skyline Plot
- Speciation: Birth-Death Serially Sampled

Citation: 8 [Constant Coalescent]. W (2002) Genetics 161,

Tree Model - de

● Random starting tree

○ UPGMA starting tree

○ User-specified starting tree

Select user-specified tree: no tree loa...

Export format for tree: New...

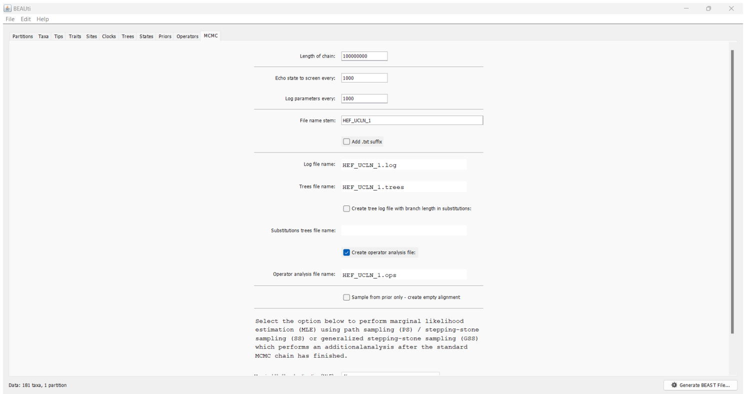
Import user-specified starting trees from **NEXUS** format data files using the 'Import Data' menu option. Trees must be rooted and strictly bifurcating (binary).

Data: 181 taxa, 1 partition

Figure 3.27: BEAUti 10

6. Setting MCMC options:

   - The next tab, MCMC, provides more general settings to control the length of the MCMC run and the file names. The Chain Length is the number of steps the MCMC will make in the chain before finishing and it depends on the size of the data set, the complexity of the model, and the quality of the answer required. The next options specify how often the parameter values in the Markov chain should be displayed on the screen and recorded in the log file


BEAUti
File Edit Help
Partitions Taxa Tips Traits Sites Clocks Trees States Priors Operators MCMC
Generate BEAST File...

Length of chain: 100000000

Echo state to screen every: 1000

Log parameters every: 1000

File name stem: HEF_UCLN_1

□ Add .txt suffix

Log file name: HEF_UCLN_1.log

Trees file name: HEF_UCLN_1.trees

□ Create tree log file with branch length in substitutions:

Substitutions trees file name:

☑ Create operator analysis file:

Operator analysis file name: HEF_UCLN_1.ops

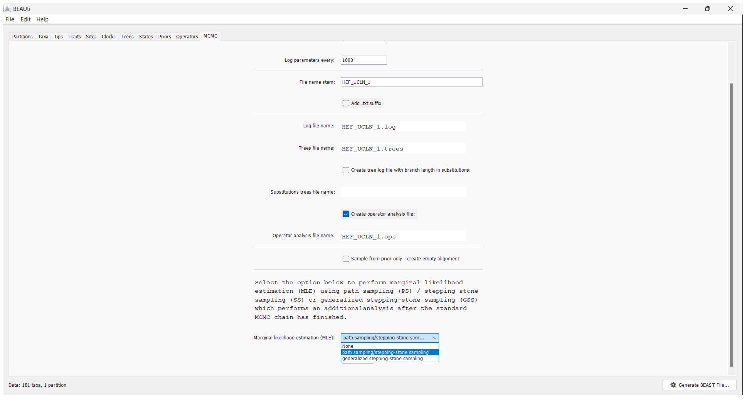
□ Sample from prior only - create empty alignment

```
Select the option below to perform marginal likelihood
estimation (MLE) using path sampling (PS) / stepping-stone
sampling (SS) or generalized stepping-stone sampling (GSS)
which performs an additionalanalysis after the standard
MCMC chain has finished.
```

Data: 181 taxa, 1 partition

Figure 3.28: BEAUti 11

- This option allows to performance of marginal likelihood estimation using path sampling (PS) / stepping-stone (SS) or generalized stepping-stone (GSS) which performs an additional analysis after the standard MCMC has finished.


BEAUti
File Edit Help
Partitions Taxa Tips Traits Sites Clocks Trees States Priors Operators MCMC
Generate BEAST File...

Log parameters every: 1000

File name stem: HEF_UCLN_1

□ Add .txt suffix

Log file name: HEF_UCLN_1.log

Trees file name: HEF_UCLN_1.trees

□ Create tree log file with branch length in substitutions:

Substitutions trees file name:

☑ Create operator analysis file:

Operator analysis file name: HEF_UCLN_1.ops

□ Sample from prior only - create empty alignment

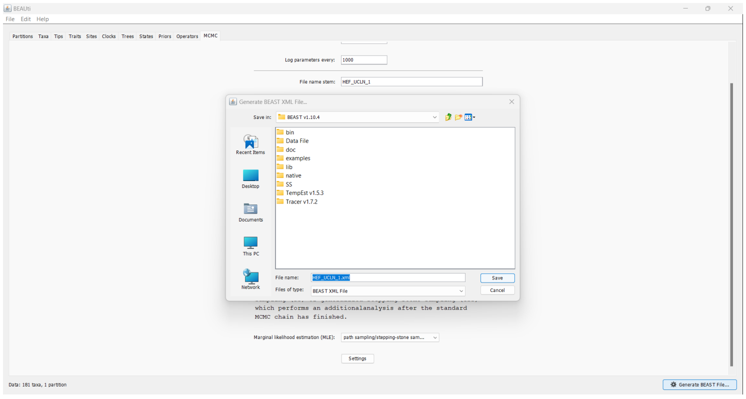
```
Select the option below to perform marginal likelihood
estimation (MLE) using path sampling (PS) / stepping-stone
sampling (SS) or generalized stepping-stone sampling (GSS)
which performs an additionalanalysis after the standard
MCMC chain has finished.
```

Marginal likelihood estimation (MLE): path sampling/stepping-stone sam...
None
path sampling/stepping-stone sampling
generalized stepping-stone sampling

Data: 181 taxa, 1 partition

Figure 3.29: BEAUti 12

7. Generating the BEAST XML file: We are now ready to create the BEAST XML file. Select Generate XML from the File menu. It ends with '.xml'.



Log parameters every: 1000

File name stem: HEF_UCLN_1

Save in: BEAST v1.10.4

bin
Data File
doc
examples
lib
native
SS
TempEst v1.5.3
Tracer v1.7.2

File name: HEF_UCLN_1.xml

Files of type: BEAST XML File

```
which performs an additionalanalysis after the standard
MCMC chain has finished.
```

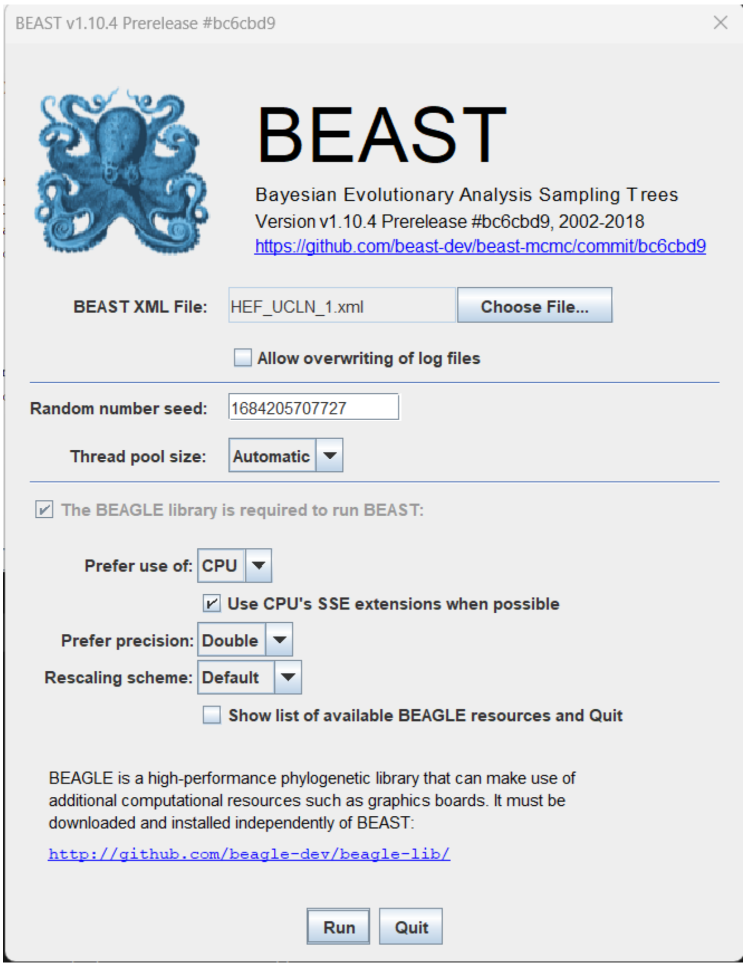
Marginal likelihood estimation (MLE): path sampling/stepping-stone sam...

Data: 181 taxa, 1 partition

Figure 3.30: BEAUti 13

3.3.4 BEAST

- Select the XML file you created in BEAUti and press Run.



BEAST v1.10.4 Prerelease #bc6cbd9

BEAST

Bayesian Evolutionary Analysis Sampling Trees
Version v1.10.4 Prerelease #bc6cbd9, 2002-2018
https://github.com/beast-dev/beast-mcmc/commit/bc6cbd9

BEAST XML File: HEF_UCLN_1.xml

□ Allow overwriting of log files

Random number seed: 1684205707727

Thread pool size: Automatic

☑ The BEAGLE library is required to run BEAST:

Prefer use of: CPU

☑ Use CPU's SSE extensions when possible

Prefer precision: Double

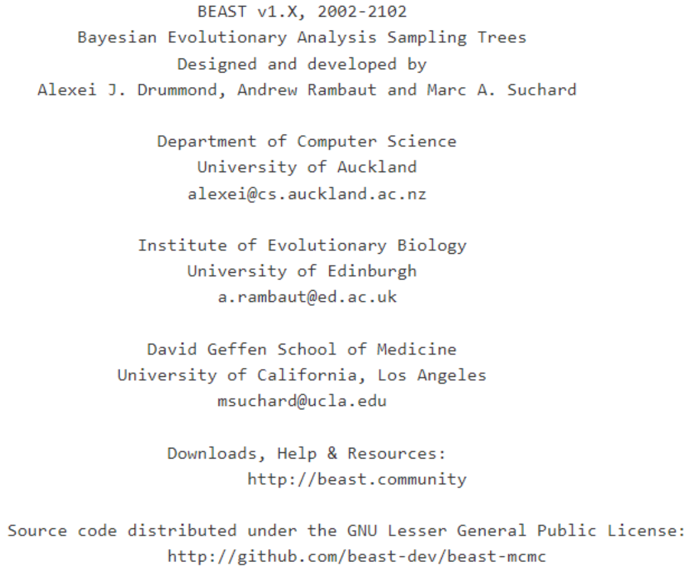
Rescaling scheme: Default

□ Show list of available BEAGLE resources and Quit

BEAGLE is a high-performance phylogenetic library that can make use of additional computational resources such as graphics boards. It must be downloaded and installed independently of BEAST:
http://github.com/beagle-dev/beagle-lib/

Figure 3.31: BEAST 1

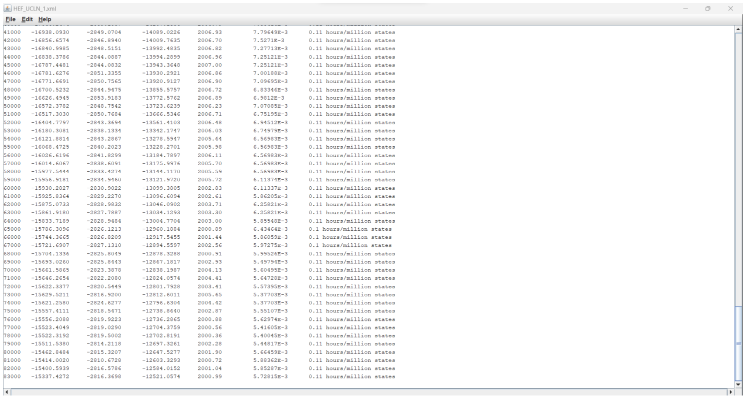
- All citations relevant for the analysis are mentioned at the start of the run.



```
                    BEAST v1.X, 2002-2102
        Bayesian Evolutionary Analysis Sampling Trees
                  Designed and developed by
  Alexei J. Drummond, Andrew Rambaut and Marc A. Suchard

               Department of Computer Science
                   University of Auckland
                  alexei@cs.auckland.ac.nz

              Institute of Evolutionary Biology
                  University of Edinburgh
                     a.rambaut@ed.ac.uk

               David Geffen School of Medicine
            University of California, Los Angeles
                     msuchard@ucla.edu

                Downloads, Help & Resources:
                     http://beast.community

 Source code distributed under the GNU Lesser General Public License:
                http://github.com/beast-dev/beast-mcmc
```

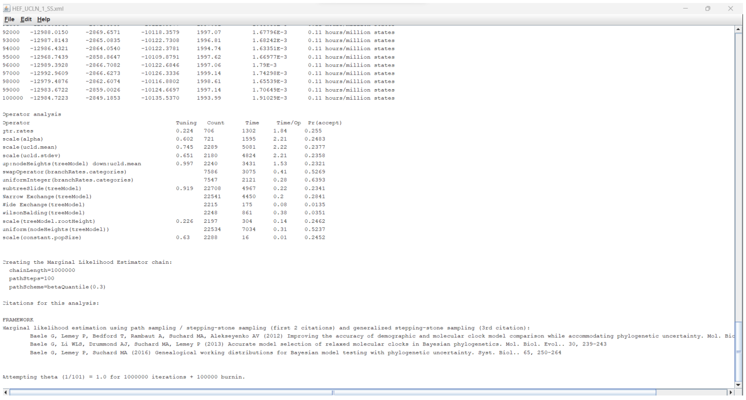
Figure 3.32: BEAST 2

- The first column is the ‘state’ number — in this case it is incrementing by 1000 so between each of these lines it has made 1000 operations. The screen log shows only a few of the metrics and parameters but it is also recording a log file to disk with all of the results in it (along with a ‘.trees’ file containing the sampled trees for these states). After a few thousand states it will start to report the number of hours per million states.



```
41000   -16938.0930   -2849.0704   -14089.0226   2006.93   7.79649E-3   0.11 hours/million states
42000   -16856.6574   -2846.8940   -14009.7635   2006.70   7.5271E-3    0.11 hours/million states
43000   -16840.9985   -2848.5151   -13992.4835   2006.82   7.27713E-3   0.11 hours/million states
44000   -16838.3786   -2844.0887   -13994.2899   2006.96   7.25121E-3   0.11 hours/million states
45000   -16787.4481   -2844.0832   -13943.3648   2007.00   7.25121E-3   0.11 hours/million states
46000   -16781.6276   -2851.3355   -13930.2921   2006.86   7.00188E-3   0.11 hours/million states
47000   -16771.6691   -2850.7565   -13920.9127   2006.90   7.09695E-3   0.11 hours/million states
48000   -16700.5232   -2844.9475   -13855.5757   2006.72   6.83346E-3   0.11 hours/million states
49000   -16626.4945   -2853.9183   -13772.5762   2006.89   6.9812E-3    0.11 hours/million states
50000   -16572.3782   -2848.7542   -13723.6239   2006.23   7.07085E-3   0.11 hours/million states
51000   -16517.3030   -2850.7684   -13666.5346   2006.71   6.75195E-3   0.11 hours/million states
52000   -16404.7797   -2843.3694   -13561.4103   2006.48   6.94512E-3   0.11 hours/million states
53000   -16180.3081   -2838.1334   -13342.1747   2006.03   6.74979E-3   0.11 hours/million states
54000   -16121.8814   -2843.2867   -13278.5947   2005.64   6.56983E-3   0.11 hours/million states
55000   -16068.4725   -2840.2023   -13228.2701   2005.98   6.56983E-3   0.11 hours/million states
56000   -16026.6196   -2841.8299   -13184.7897   2006.11   6.56983E-3   0.11 hours/million states
57000   -16014.6067   -2838.6091   -13175.9976   2005.70   6.56983E-3   0.11 hours/million states
58000   -15977.5444   -2833.4274   -13144.1170   2005.59   6.56983E-3   0.11 hours/million states
59000   -15956.9181   -2834.9460   -13121.9720   2005.72   6.11374E-3   0.11 hours/million states
60000   -15930.2827   -2830.9022   -13099.3805   2002.83   6.11337E-3   0.11 hours/million states
61000   -15925.8364   -2829.2270   -13096.6094   2002.61   5.86205E-3   0.11 hours/million states
62000   -15875.0733   -2828.9832   -13046.0902   2003.71   6.25821E-3   0.11 hours/million states
63000   -15861.9180   -2827.7887   -13034.1293   2003.30   6.25821E-3   0.11 hours/million states
64000   -15833.7189   -2828.9484   -13004.7704   2003.00   5.85548E-3   0.11 hours/million states
65000   -15786.3096   -2826.1213   -12960.1884   2000.89   6.43464E-3   0.1 hours/million states
66000   -15744.3665   -2826.8209   -12917.5455   2001.44   5.86059E-3   0.1 hours/million states
67000   -15721.6907   -2827.1310   -12894.5597   2002.56   5.97275E-3   0.1 hours/million states
68000   -15704.1336   -2825.8049   -12878.3288   2000.91   5.99526E-3   0.11 hours/million states
69000   -15693.0260   -2825.8443   -12867.1817   2002.93   5.49794E-3   0.11 hours/million states
70000   -15661.5865   -2823.3878   -12838.1987   2004.13   5.60495E-3   0.11 hours/million states
71000   -15646.2654   -2822.2080   -12824.0574   2004.41   5.64728E-3   0.11 hours/million states
72000   -15622.3377   -2820.5449   -12801.7928   2003.41   5.57395E-3   0.11 hours/million states
73000   -15629.5211   -2816.9200   -12812.6011   2005.65   5.37703E-3   0.11 hours/million states
74000   -15621.2580   -2824.6277   -12796.6304   2004.42   5.37703E-3   0.11 hours/million states
75000   -15557.4111   -2818.5471   -12738.8640   2002.87   5.55107E-3   0.11 hours/million states
76000   -15556.2088   -2819.9223   -12736.2865   2000.88   5.62974E-3   0.11 hours/million states
77000   -15523.4049   -2819.0290   -12704.3759   2000.56   5.41605E-3   0.11 hours/million states
78000   -15522.3192   -2819.5002   -12702.8191   2000.36   5.40045E-3   0.11 hours/million states
79000   -15511.5380   -2814.2118   -12697.3261   2002.28   5.44817E-3   0.11 hours/million states
80000   -15462.8484   -2815.3207   -12647.5277   2001.90   5.66459E-3   0.11 hours/million states
81000   -15414.0020   -2810.6728   -12603.3293   2000.72   5.88362E-3   0.11 hours/million states
82000   -15400.5939   -2816.5786   -12584.0152   2001.04   5.85287E-3   0.11 hours/million states
83000   -15337.4272   -2816.3698   -12521.0574   2000.99   5.72815E-3   0.11 hours/million states
```

Figure 3.33: BEAST 3

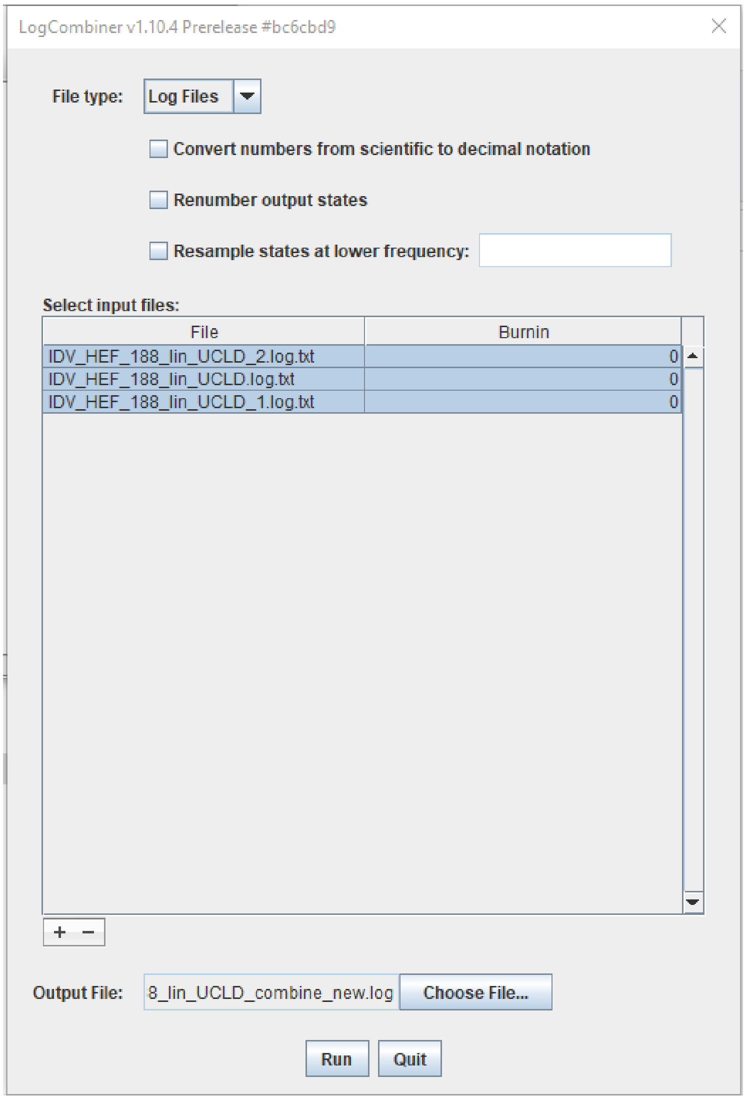


```
92000   -12988.0150    -2869.6571    -10118.3579    1997.07     1.67796E-3    0.11 hours/million states
93000   -12987.8143    -2865.0835    -10122.7308    1996.81     1.68242E-3    0.11 hours/million states
94000   -12986.4321    -2864.0540    -10122.3781    1994.74     1.63351E-3    0.11 hours/million states
95000   -12968.7439    -2858.8647    -10109.8791    1997.62     1.66977E-3    0.11 hours/million states
96000   -12989.3928    -2866.7082    -10122.6846    1997.06     1.79E-3       0.11 hours/million states
97000   -12992.9609    -2866.6273    -10126.3336    1999.14     1.74298E-3    0.11 hours/million states
98000   -12979.4876    -2862.6074    -10116.8802    1998.61     1.65539E-3    0.11 hours/million states
99000   -12983.6722    -2859.0026    -10124.6697    1997.14     1.70649E-3    0.11 hours/million states
100000  -12984.7223    -2849.1853    -10135.5370    1993.99     1.91029E-3    0.11 hours/million states

Operator analysis
Operator                                      Tuning  Count    Time    Time/Op  Pr(accept)
gtr.rates                                     0.224   706      1302    1.84     0.255
scale(alpha)                                  0.602   721      1595    2.21     0.2483
scale(ucld.mean)                              0.745   2289     5081    2.22     0.2377
scale(ucld.stdev)                             0.651   2180     4824    2.21     0.2358
up:nodeHeights(treeModel) down:ucld.mean      0.997   2240     3431    1.53     0.2321
swapOperator(branchRates.categories)                  7586     3075    0.41     0.5269
uniformInteger(branchRates.categories)                7547     2121    0.28     0.6393
subtreeSlide(treeModel)                       0.919   22708    4967    0.22     0.2341
Narrow Exchange(treeModel)                            22541    4450    0.2      0.2841
Wide Exchange(treeModel)                              2215     175     0.08     0.0135
wilsonBalding(treeModel)                              2248     861     0.38     0.0351
scale(treeModel.rootHeight)                   0.226   2197     304     0.14     0.2462
uniform(nodeHeights(treeModel))                       22534    7034    0.31     0.5237
scale(constant.popSize)                       0.63    2288     16      0.01     0.2452


Creating the Marginal Likelihood Estimator chain:
  chainLength=1000000
  pathSteps=100
  pathScheme=betaQuantile(0.3)

Citations for this analysis:

FRAMEWORK
Marginal likelihood estimation using path sampling / stepping-stone sampling (first 2 citations) and generalized stepping-stone sampling (3rd citation):
        Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Bic
        Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P (2013) Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Mol. Biol. Evol.. 30, 239-243
        Baele G, Lemey P, Suchard MA (2016) Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty. Syst. Biol.. 65, 250-264


Attempting theta (1/101) = 1.0 for 1000000 iterations + 100000 burnin.
```

Figure 3.34: BEAST 2

3.3.5 LOGCOMBINER

- Selecting log files from “file type” menu. The input files are those log files generated from the above step. Choosing an output file will save the combined log files into it. Then select “Run”.



LogCombiner v1.10.4 Prerelease #bc6cbd9

File type: Log Files

□ Convert numbers from scientific to decimal notation

□ Renumber output states

□ Resample states at lower frequency:

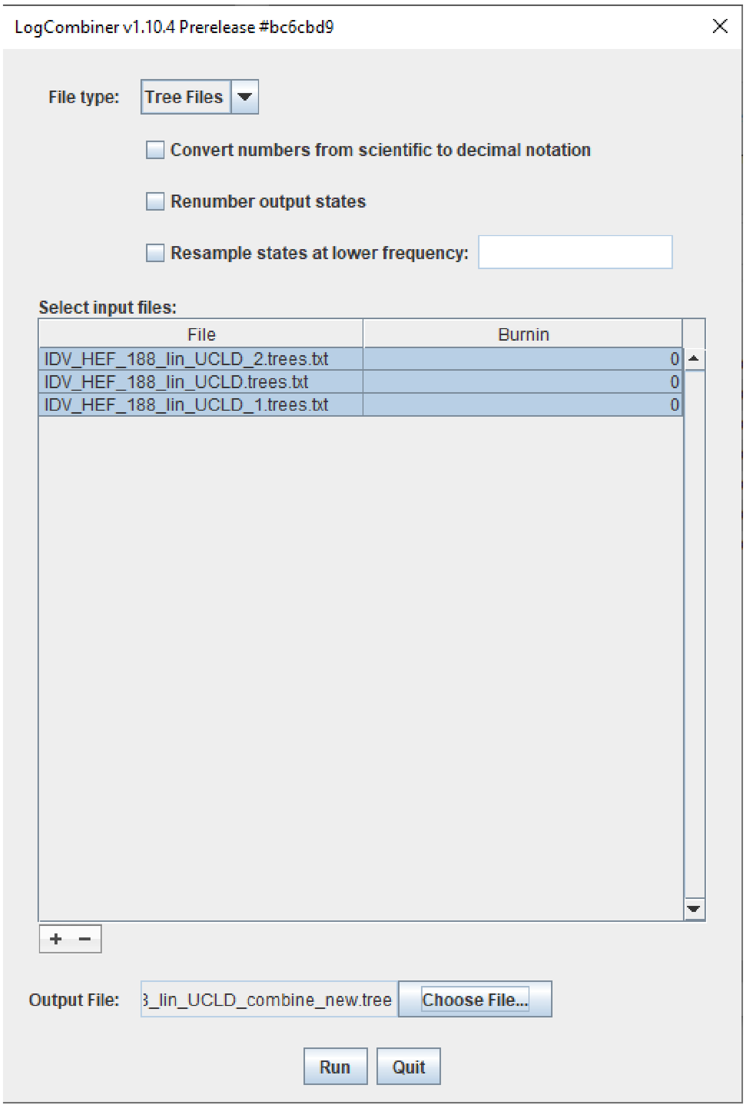
Select input files:

File	Burnin
IDV_HEF_188_lin_UCLD_2.log.txt	0
IDV_HEF_188_lin_UCLD.log.txt	0
IDV_HEF_188_lin_UCLD_1.log.txt	0

Output File: 8_lin_UCLD_combine_new.log

Figure 3.35: LOGCOMBINER 1

- Similarly, for tree files.



LogCombiner v1.10.4 Prerelease #bc6cbd9

File type: Tree Files

☐ Convert numbers from scientific to decimal notation

☐ Renumber output states

☐ Resample states at lower frequency:

Select input files:

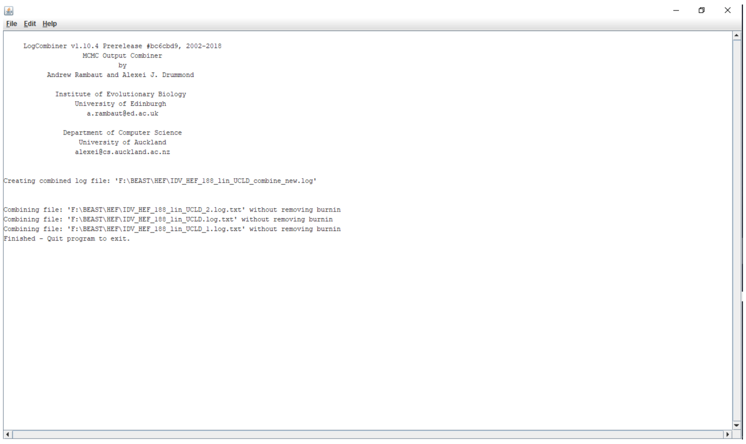
File	Burnin
IDV_HEF_188_lin_UCLD_2.trees.txt	0
IDV_HEF_188_lin_UCLD.trees.txt	0
IDV_HEF_188_lin_UCLD_1.trees.txt	0

Output File: 3_lin_UCLD_combine_new.tree

Figure 3.36: LOGCOMBINER 2

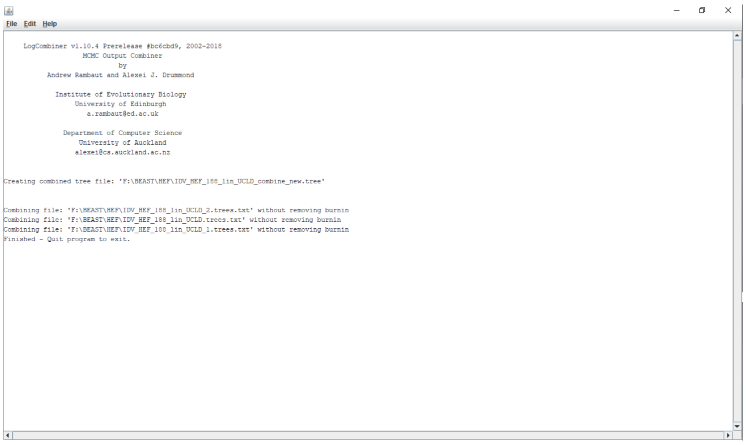
- When the program says ‘Finished’, the combined log file you selected, above, will be ready. Close the program by selecting the Quit option. Once LogCombiner has finished you can analyze the combined log file in Tracer.

- Similarly, for tree files.



```
    LogCombiner v1.10.4 Prerelease #bc6cbd9, 2002-2018
                  MCMC Output Combiner
                           by
           Andrew Rambaut and Alexei J. Drummond

              Institute of Evolutionary Biology
                   University of Edinburgh
                      a.rambaut@ed.ac.uk

                Department of Computer Science
                    University of Auckland
                   alexei@cs.auckland.ac.nz


Creating combined log file: 'F:\BEAST\HEF\IDV_HEF_188_lin_UCLD_combine_new.log'


Combining file: 'F:\BEAST\HEF\IDV_HEF_188_lin_UCLD_2.log.txt' without removing burnin
Combining file: 'F:\BEAST\HEF\IDV_HEF_188_lin_UCLD.log.txt' without removing burnin
Combining file: 'F:\BEAST\HEF\IDV_HEF_188_lin_UCLD_1.log.txt' without removing burnin
Finished - Quit program to exit.
```

Figure 3.37: LOGCOMBINER 3



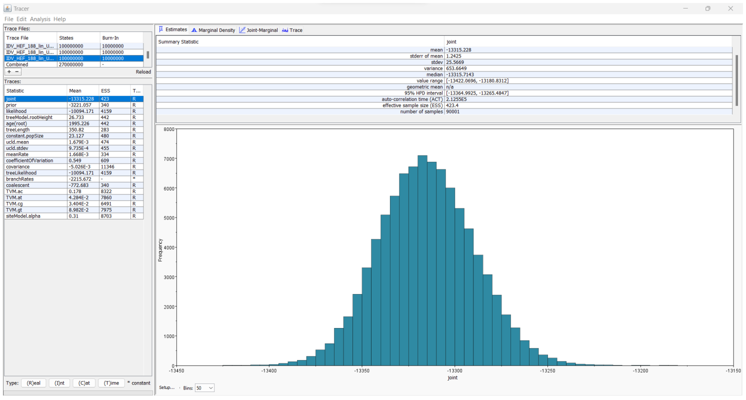
```
    LogCombiner v1.10.4 Prerelease #bc6cbd9, 2002-2018
                  MCMC Output Combiner
                           by
           Andrew Rambaut and Alexei J. Drummond

              Institute of Evolutionary Biology
                   University of Edinburgh
                      a.rambaut@ed.ac.uk

                Department of Computer Science
                    University of Auckland
                   alexei@cs.auckland.ac.nz


Creating combined tree file: 'F:\BEAST\HEF\IDV_HEF_188_lin_UCLD_combine_new.tree'


Combining file: 'F:\BEAST\HEF\IDV_HEF_188_lin_UCLD_2.trees.txt' without removing burnin
Combining file: 'F:\BEAST\HEF\IDV_HEF_188_lin_UCLD.trees.txt' without removing burnin
Combining file: 'F:\BEAST\HEF\IDV_HEF_188_lin_UCLD_1.trees.txt' without removing burnin
Finished - Quit program to exit.
```

Figure 3.38: LOGCOMBINER 4

3.3.6 TRACER

- We start by loading the output .log files of both replicates of the same BEAST XML into Tracer 1.7. To load the log file(s), select the Open option from the File menu or drag and drop the log file into the Tracer window. The files will load and you will be presented with a window similar to the one below.

- As with loading a single log file, the name of the log file loaded and the traces that it contains can be seen on the left hand side. When the different files loaded contain the same set of logged parameters, then a



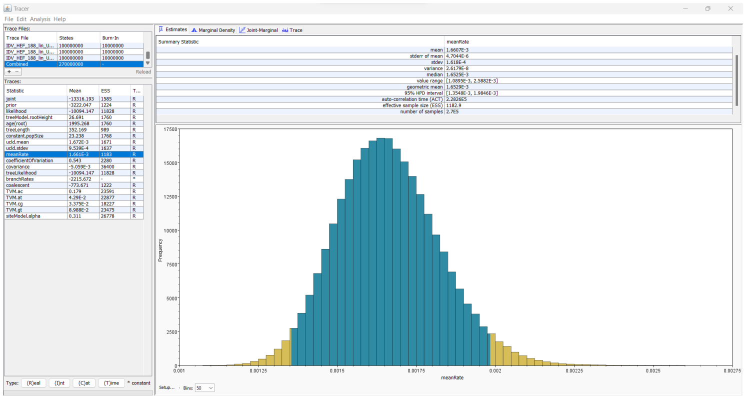
Trace File	States	Burn-In
IDV_HEF_188_lin_U...	100000000	10000000
IDV_HEF_188_lin_U...	100000000	10000000
IDV_HEF_188_lin_U...	100000000	10000000
Combined	270000000	-

Statistic	Mean	ESS	T...
joint	-13315.228	423	R
prior	-3221.057	340	R
likelihood	-10094.171	4159	R
treeModel.rootHeight	26.733	442	R
age(root)	1995.226	442	R
treeLength	350.82	283	R
constant.popSize	23.127	480	R
ucld.mean	1.679E-3	474	R
ucld.stdev	9.735E-4	455	R
meanRate	1.668E-3	334	R
coefficientOfVariation	0.549	609	R
covariance	-5.026E-3	11346	R
treeLikelihood	-10094.171	4159	R
branchRates	-2215.672	-	*
coalescent	-772.683	340	R
TVM.ac	0.178	8322	R
TVM.at	4.284E-2	7860	R
TVM.cg	3.404E-2	6491	R
TVM.gt	8.982E-2	7975	R
siteModel.alpha	0.31	8703	R

Summary Statistic	joint
mean	-13315.228
stderr of mean	1.2425
stdev	25.5669
variance	653.6649
median	-13315.7143
value range	[-13422.0696, -13180.8312]
geometric mean	n/a
95% HPD interval	[-13364.9925, -13265.4847]
auto-correlation time (ACT)	2.1255E5
effective sample size (ESS)	423.4
number of samples	90001

Figure 3.39: TRACER 1

Combined trace will automatically appear (with a number of iterations equal to the sum of the two traces minus their burn-in length).



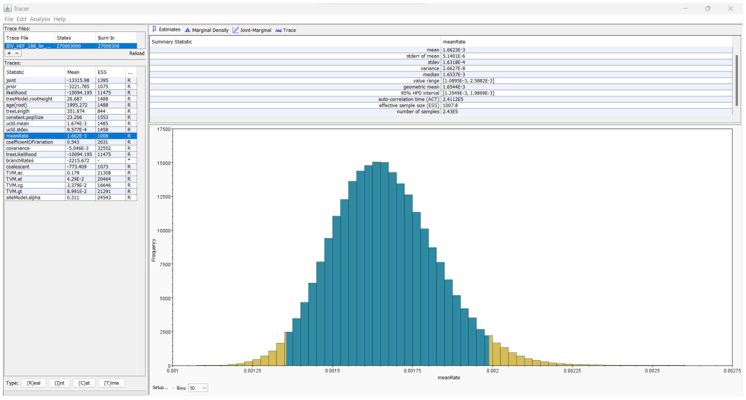
Trace File	States	Burn-In
IDV_HEF_188_lin_U...	100000000	10000000
IDV_HEF_188_lin_U...	100000000	10000000
IDV_HEF_188_lin_U...	100000000	10000000
Combined	270000000	-

Statistic	Mean	ESS	T...
joint	-13316.193	1585	R
prior	-3222.047	1224	R
likelihood	-10094.147	11828	R
treeModel.rootHeight	26.691	1760	R
age(root)	1995.268	1760	R
treeLength	352.169	989	R
constant.popSize	23.238	1768	R
ucld.mean	1.672E-3	1671	R
ucld.stdev	9.539E-4	1637	R
meanRate	1.661E-3	1183	R
coefficientOfVariation	0.543	2280	R
covariance	-5.059E-3	36400	R
treeLikelihood	-10094.147	11828	R
branchRates	-2215.672	-	*
coalescent	-773.671	1222	R
TVM.ac	0.179	23591	R
TVM.at	4.29E-2	22877	R
TVM.cg	3.375E-2	18227	R
TVM.gt	8.988E-2	23475	R
siteModel.alpha	0.311	26778	R

Summary Statistic	meanRate
mean	1.6607E-3
stderr of mean	4.7044E-6
stdev	1.618E-4
variance	2.6179E-8
median	1.6525E-3
value range	[1.0895E-3, 2.5882E-3]
geometric mean	1.6529E-3
95% HPD interval	[1.3548E-3, 1.9846E-3]
auto-correlation time (ACT)	2.2826E5
effective sample size (ESS)	1182.9
number of samples	2.7E5

Figure 3.40: TRACER 2

- Selecting the Combined trace allows to explore a concatenation of the log files. Tracer will plot a (marginal posterior) histogram for the selected statistic and also give you summary statistics such as the mean and median. The 95% HPD interval stands for highest posterior density interval and represents the most compact interval on the selected parameter that contains 95% of the posterior probability. In the top right of the window is a table of calculated statistics for the selected trace



Trace Files:

Trace File	States	Burn-In
IDV_HEF_188_lin_...	270003000	27000300

Traces:

Statistic	Mean	ESS	...
joint	-13315.98	1395	R
prior	-3221.785	1075	R
likelihood	-10094.195	11475	R
treeModel.rootHeight	26.687	1488	R
age(root)	1995.272	1488	R
treeLength	351.874	844	R
constant.popSize	23.206	1553	R
ucld.mean	1.674E-3	1485	R
ucld.stdev	9.577E-4	1458	R
meanRate	1.662E-3	1008	R
coefficientOfVariation	0.543	2031	R
covariance	-5.046E-3	32552	R
treeLikelihood	-10094.195	11475	R
branchRates	-2215.672	-	*
coalescent	-773.409	1073	R
TVM.ac	0.179	21308	R
TVM.at	4.29E-2	20464	R
TVM.cg	3.379E-2	16646	R
TVM.gt	8.991E-2	21291	R
siteModel.alpha	0.311	24543	R

Summary Statistic	meanRate
mean	1.6623E-3
stderr of mean	5.1401E-6
stdev	1.6318E-4
variance	2.6627E-8
median	1.6537E-3
value range	[1.0895E-3, 2.5882E-3]
geometric mean	1.6544E-3
95% HPD interval	[1.3549E-3, 1.9899E-3]
auto-correlation time (ACT)	2.4112E5
effective sample size (ESS)	1007.8
number of samples	2.43E5

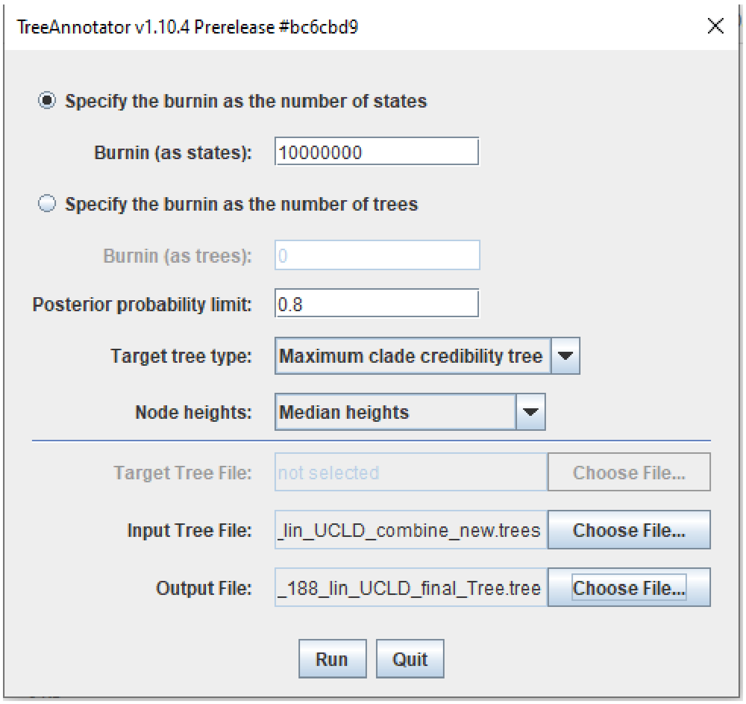
Bins: 50

Figure 3.41: TRACER 3

3.3.7 TREE ANNOTATOR

- This is a post-analysis program that will produce a summary tree from the output of BEAST. It's simply not feasible to inspect every tree that was visited during the BEAST analysis, hence we will create a consensus tree summarizing the posterior tree distribution. Within the BEAST package, this is done by constructing a maximum clade credibilty (MCC) tree using the program TreeAnnotator.

  Upon running TreeAnnotator, you will be presented with the following window: Typically, 10% of the total number of iterations is used as the burn-in for analysis, provided that this is sufficient to have made it past the actual burn-in phase (which can be inspected/checked in Tracer). For the Input Tree File, select the file which was generated during the BEAST run needs to be selected. For the Output File, no such file can be selected but rather its file name needs to be entered manually in the following window:



TreeAnnotator v1.10.4 Prerelease #bc6cbd9

● Specify the burnin as the number of states

Burnin (as states): 10000000

○ Specify the burnin as the number of trees

Burnin (as trees): 0

Posterior probability limit: 0.8

Target tree type: Maximum clade credibility tree

Node heights: Median heights

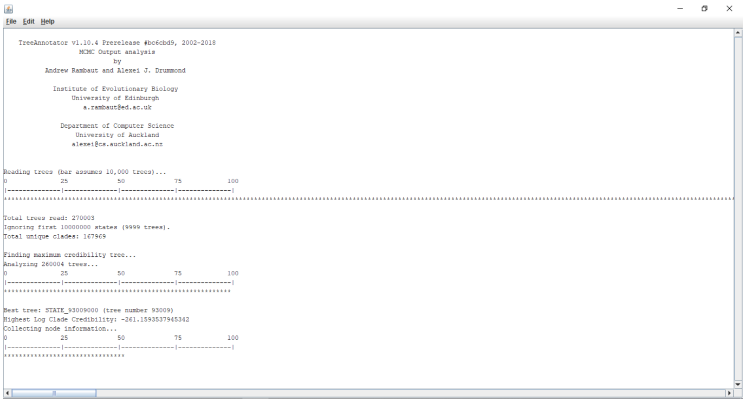
Target Tree File: not selected

Input Tree File: _lin_UCLD_combine_new.trees

Output File: _188_lin_UCLD_final_Tree.tree

Figure 3.42: TREE ANNOTATOR 1

- After all the required settings have been entered, you can start constructing the MCC tree by clicking ‘Run. Progress can monitored in the following window, at the end of which you will be asked to quit the program and the MCC tree will have been written to the selected output file:



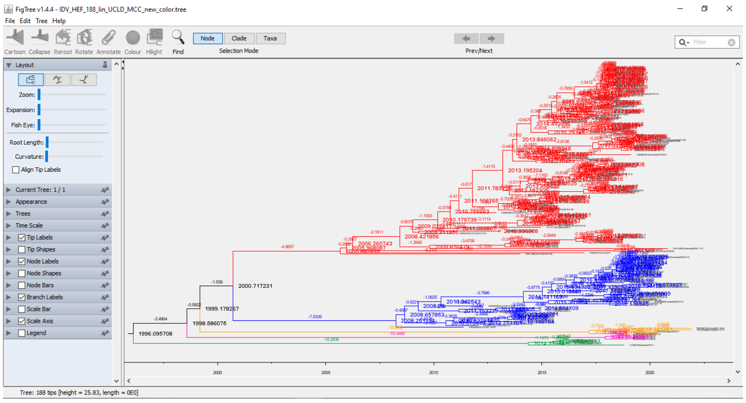
```
   TreeAnnotator v1.10.4 Prerelease #bc6cbd9, 2002-2018
                  MCMC Output analysis
                           by
          Andrew Rambaut and Alexei J. Drummond

             Institute of Evolutionary Biology
                  University of Edinburgh
                     a.rambaut@ed.ac.uk

               Department of Computer Science
                   University of Auckland
                  alexei@cs.auckland.ac.nz


Reading trees (bar assumes 10,000 trees)...
0              25             50             75            100
|--------------|--------------|--------------|--------------|
**************************************************************************************************************************************************************************************************

Total trees read: 270003
Ignoring first 10000000 states (9999 trees).
Total unique clades: 167969

Finding maximum credibility tree...
Analyzing 260004 trees...
0              25             50             75            100
|--------------|--------------|--------------|--------------|
************************************************************

Best tree: STATE_93009000 (tree number 93009)
Highest Log Clade Credibility: -261.1593537945342
Collecting node information...
0              25             50             75            100
|--------------|--------------|--------------|--------------|
********************************
```

Figure 3.43: TREE ANNOTATOR 2

3.3.8 FIG TREE

The final step is to load the constructed MCC tree into FigTree, which allows visualizing of the tree and accompanying summary information produced by TreeAnnotator. After starting FigTree, simply go to File and Open

- The following tree is the tree with node ages.



FigTree v1.4.4 - IDV_HEF_188_lin_UCLD_MCC_new_color.tree

Selection Mode: Node

☐ Align Tip Labels

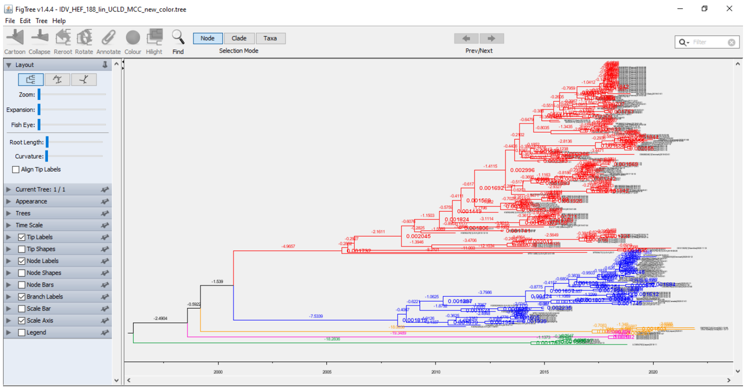
Current Tree: 1 / 1

☑ Tip Labels
☐ Tip Shapes
☑ Node Labels
☐ Node Shapes
☐ Node Bars
☑ Branch Labels
☐ Scale Bar
☑ Scale Axis
☐ Legend

Tree: 188 tips [height = 25.83, length = 0E0]

Figure 3.44: FIGTREE 1

- The following tree mentions the mean substitution rate.



FigTree v1.4.4 - IDV_HEF_188_lin_UCLD_MCC_new_color.tree

Selection Mode: Node

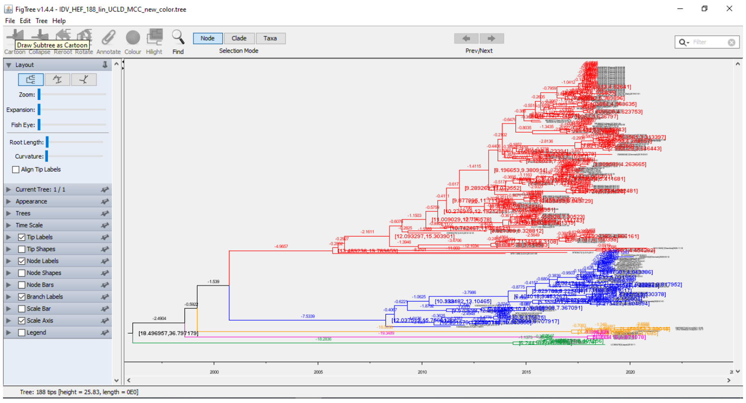
☐ Align Tip Labels

Current Tree: 1 / 1

☑ Tip Labels
☐ Tip Shapes
☑ Node Labels
☐ Node Shapes
☐ Node Bars
☑ Branch Labels
☐ Scale Bar
☑ Scale Axis
☐ Legend

Figure 3.45: FIGTREE 2

- The following tree mentions the 95% HPD intervals



□ Align Tip Labels
☑ Tip Labels
□ Tip Shapes
☑ Node Labels
□ Node Shapes
□ Node Bars
☑ Branch Labels
□ Scale Bar
☑ Scale Axis
□ Legend

Tree: 188 tips [height = 25.83, length = 0E0]

Figure 3.46: FIGTREE 3

Chapter 4

Results and Conclusions

4.1 Results

- The following table summarizes our analysis which includes the best models for each gene along with the substitution rate and the time for a most recent ancestor.

Dataset	Gene name	Substitution Model	Clock	R^2	Correlation coefficient
1	PB2	TIM+F+G4	UCLN	0.7447	0.863
2	PB1	TIM+F+G4	STRICT	0.0019	0.045
3	P3	TIM+F+I+G4	STRICT	0.4904	0.7003
4	HEF	TVM+F+G4	UCLN	0.5927	0.7699
5	NP	TN+F+G4	STRICT	0.6533	0.8082
6	P42	GTR+F+G4	STRICT	0.0640	0.2531
7	NS	HKY+F+G4	UCED	0.2622	0.512
7.1	NS1	HKY+F+G4	UCLN	0.2939	0.5421
7.2	NS2	TN93+F+G4*	UCLN	0.1826	0.4273

Dataset	Gene name	MEAN NSR (sub/site/year)	NSR (95% HPD Interval)	tMRCA
1	PB2	1.35×10^{-3}	$(1.122 \times 10^{-3}, 1.5935 \times 10^{-3})$	1997.47
2	PB1	1.16×10^{-3}	$(9.9688 \times 10^{-4}, 1.3341 \times 10^{-3})$	1997.04
3	P3	1.31×10^{-3}	$(1.1203 \times 10^{-3}, 1.4832 \times 10^{-3})$	1997.53
4	HEF	1.66×10^{-3}	$(1.4047 \times 10^{-3}, 2.0118 \times 10^{-3})$	1996.09
5	NP	1.47×10^{-3}	$(1.2389 \times 10^{-3}, 1.6881 \times 10^{-3})$	1998.97
6	P42	1.35×10^{-3}	$(1.106 \times 10^{-3}, 1.6039 \times 10^{-3})$	1997.30
7	NS	1.31×10^{-3}	$(9.0492 \times 10^{-4}, 1.7197 \times 10^{-3})$	1997.36
7.1	NS1	1.32×10^{-3}	$(9.572 \times 10^{-4}, 1.7028 \times 10^{-3})$	1998.62
7.2	NS2	1.63×10^{-3}	$(1.0584 \times 10^{-3}, 2.2291 \times 10^{-3})$	2003.16

indicate that In the NS2 gene, TPM2+F+G4 is the actual best model but this model is not available in BUAUTi software hence we use the second-best model which is TN93+F+G4.

- **Model Explanation :**
- F: Empirical Base Frequency.
- $G4$: Site heterogeneity model in Gamma and number of Gamma category is 4
- UCLN: Uncorrelated Clock Lognormal Distribution
- UCED: Uncorrelated Clock Exponential Distribution

Substitution model:

- TIM: Transition model
- GTR: General time reversible
- TN: Tamura and Nei plot
- HKY: Hasegawa-Kishino-Yano
- TN93: Tamura and Nei plot 93
- TVM: Transversion-Varying Model

4.2 Conclusion on the basis of Molecular Phylogenetic Tree

The following tree best summarizes our Bayesian phylogenetic analysis.

- The following tree is the tree with node ages.

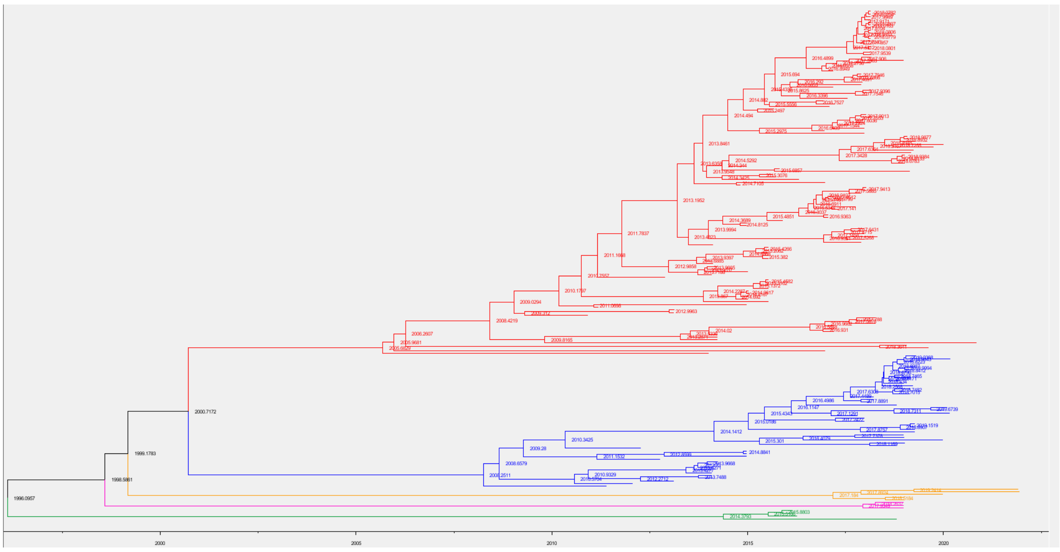


Figure 4.1: Best tree with node ages

- The following tree mentions the mean substitution rate.

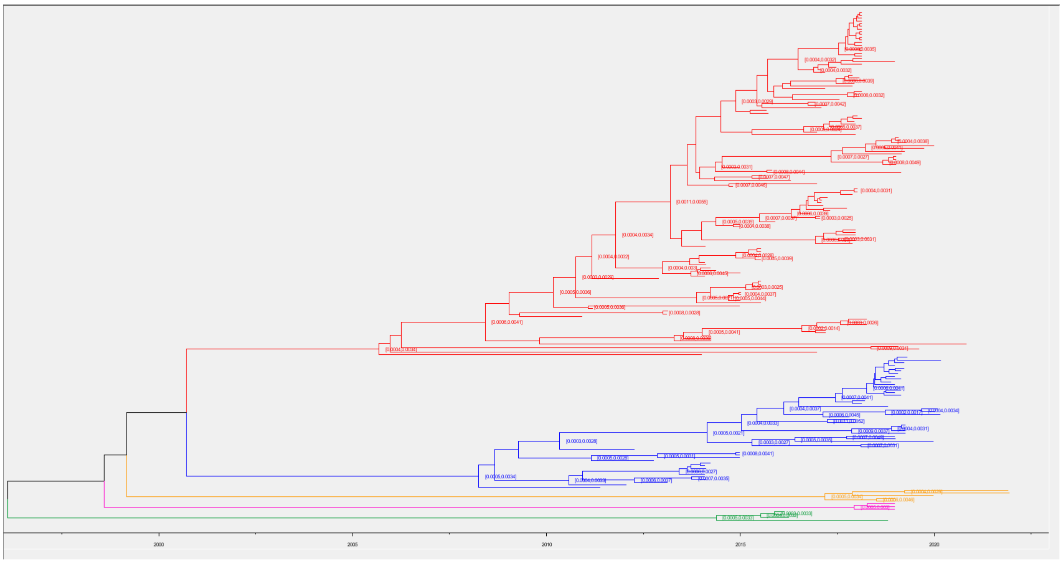


Figure 4.2: Best tree with substitution rates per site per year

- The following tree mentions the 95% HPD intervals

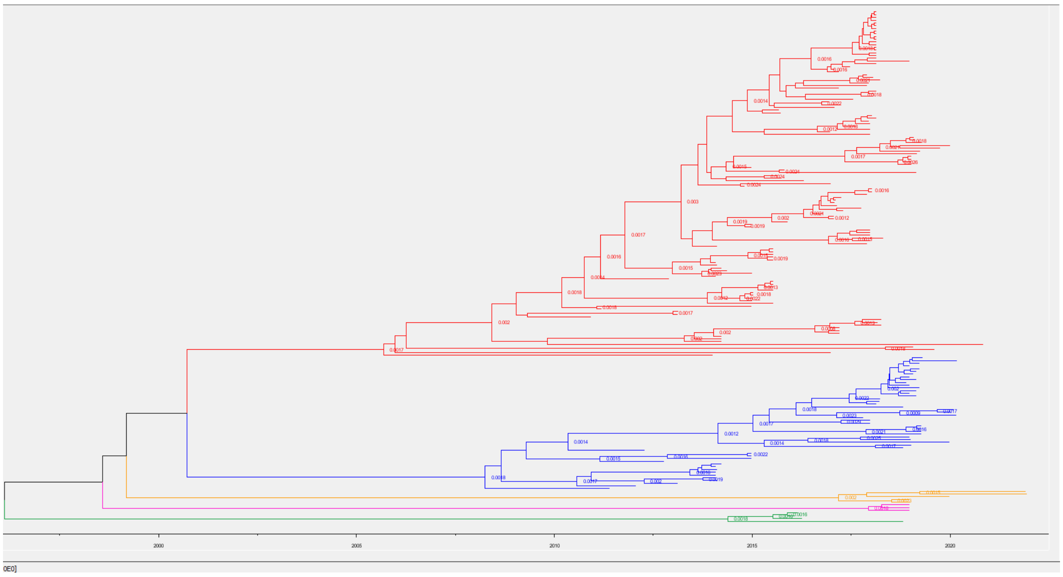


Figure 4.3: Best tree with 95% HPD intervals.

4.3 Statistical Conclusions:

- Particularly, for the 4^{th} gene (HEF) of IDV, we study its molecular clock behavior by analyzing the rate at which mutations accumulate in the HEF gene sequence and also by comparing the differences in the HEF gene sequences from different strains or isolates of IDV.

- In the case of IDV-HEF, the TVM+F+G4 model was identified as the best fit. The Transversion-Varying Model model (TVM) is a widely used substitution model that allows for different rates of nucleotide substitutions. The F parameter accounts for rate heterogeneity among sites, and the G4 parameter models the rate variation among different evolutionary categories.

- By applying uncorrelated lognormal distribution to the clock model for IDV-HEF, researchers can estimate the rates of molecular evolution and divergence times more accurately. This model accounts for rate variation and provides a more realistic representation of the evolutionary dynamics of the gene sequences.

- For the analysis of HEF, using the BEAST software, the convergence occurred at 10^8 (MCMC chain length) for all the parameters ($\geq$200) of Effective Sample Size (ESS).

- The mean rate for HEF, is estimated to be 1.66×10^{-3} substitutions per site per year. This rate represents the average rate of evolutionary change in the gene sequence over time.

- Additionally the 95% HPD interval of mean substitution rates, which is a credible interval, is given as (1.4047×10^{-3}, 2.0118×10^{-3}). This interval provides a range of plausible values for the true rate of evolution. It suggests that with 95% credible confidence the actual rate of evolutionary change in gene 4 IDV-HEF falls within this credible interval.

- The analysis of HEF resulted in the identification of five distinct clusters based on the marginal likelihood. This suggests that the data supports the presence of five separate groups or lineages within the gene sequence.

Chapter 5

Limitations and Scope

5.1 Limitations

- Bayesian statistical analysis relies on the availability of high-quality and comprehensive data and hence, problems like underreporting, surveillance gaps or difficulties in identifying and characterizing IDV strains can restrict the accuracy and reliability of the Bayesian analysis.

- Bayesian analysis incorporates prior information or assumptions about the data and model parameters. However, for emerging or newly discovered viruses like IDV, there may be a lack of prior knowledge making it challenging to establish informative prior distributions.

- Bayesian analysis often involves complex mathematical models which require numerous assumptions and they introduce uncertainties and potential biases. The accuracy of the analysis depends on the appropriateness of the chosen model and the quality of parameter estimates.

- Bayesian analysis is usually computationally intensive, particularly when dealing with large datasets or complex models. The analysis may require significant computational resources, time and expertise.

- The accuracy of uncertainty estimates depends on several factors, including the quality of data, model assumptions and parameter estimates and in situations where data or prior knowledge is limited, the uncertainty estimates may be less precise or less reliable.
- The results obtained from Bayesian analysis of IDV evolution may not be readily generalizable to other viruses or populations. Factors such as geographic location, host species and ecological factors can influence the evolutionary dynamics of IDV.

5.2 Scope

- Bayesian computational methods can be used to estimate the rate of molecular evolution in IDV which helps in understanding mode of its evolution and helps predict future evolutionary trends.
- Understanding the population dynamics of IDV is crucial for developing effective control and prevention strategies.
- Bayesian methods can be used to study the evolution of IDV and help in the development of improved vaccines and vaccination strategies against IDV.
- Bayesian statistical analysis is essential for understanding the virus's behaviour, improving surveillance and control measures and guiding public health interventions.

5.3 Bibliography

- Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS biology. 2006 May;4(5):e88

- Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology. 2007 Dec;7(1):1-8.

- Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Systematic biology. 2018 Sep 1;67(5):901-4.

- Ho SY, Duchêne S. Molecular-clock methods for estimating evolutionary rates and timescales. Molecular ecology. 2014 Dec;23(24):5947-65.

- Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus evolution. 2018 Jan;4(1):vey016.

- Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic acids research. 2016 Apr 15;44(W1):W232-5.