UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

# Faculty of Sciences
## *School for Information Technology*
Master of Statistics and Data Science

***Master's thesis***

***Evaluation of Different Molecular Clock Models in a Mycobacterium Tuberculosis dataset***

**Furaha Maine Chaula**

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

**SUPERVISOR :**

dr. Jade Vincent MEMBREBE

**2023**
**2024**

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

**www.uhasselt.be**
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

# Faculty of Sciences
## *School for Information Technology*
## Master of Statistics and Data Science

***Master's thesis***

***Evaluation of Different Molecular Clock Models in a Mycobacterium Tuberculosis dataset***

**Furaha Maine Chaula**

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

**SUPERVISOR :**

dr. Jade Vincent MEMBREBE

**Abstract**

**Background:** Mycobacterium tuberculosis (MTB), as the causative agent of tuberculosis, continues to pose a significant global health threat. The rate of evolution within MTB remains a contentious issue, marked by divergent estimates from studies utilizing different calibration methods. Molecular clock models are statistical models used to estimate evolutionary rates and divergence times in various genome sequence data. These methods have been applied to various genomic data sets, including those from pathogens like MTB, viruses, and other organisms. However, there is a lack of consensus on the rate of evolution of MTB among different studies. These discrepancies in these estimates arise when different metrics are employed, indicating potential variation in the evolutionary tree's evolution rates.

**Objectives:** To evaluate and compare the performance of the different molecular clock models in reconstructing the phylogeny of Mycobacterium tuberculosis to uncover the clock-like nature of the MTB dataset, hence resolving the discrepancies in evolutionary rate estimates and shedding light on the underlying factors influencing the observed variability.

**Methodology:** Three clock models, such as Strict clock, Random local clock, and Uncorrelated lognormal clock models, were explored, evaluated, and compared to uncover the clocklike nature of the MTB dataset. The software BEAST was employed to reconstruct the phylogeny and analyze the performance of different clock models. Bayes Factor with marginal likelihood estimation was considered for model selection.

**Results:** For the clade L2.3.1, L2.3.2, L2.3.4, and L2.3.5, A Strict clock model gives an estimated overall evolutionary rate that ranges approximately between $(2.588, 6.556) \times 10^{-8}$ substitutions per site per year for the clade L2.3.1, L2.3.2, L2.3.4, and L2.3.5. For the clades L2.3.3 and L2.3.6, the Random clock model was found to be an optimal clock model to explain the evolutionary rate with an initial rate that ranges between $(2.644, 8.659) \times 10^{-8}$ substitutions per site per year. Subsequently, in terms of tMRCA for the clade L2.3.1 to L2.3.3 were found to have quite high ages ranging from approximately [950,000 to 1,900,000] years. Incontrast, clades L2.3.4 to L.2.3.6 were found to have quite similar small ages of the most recent common ancestor range between approximately (400,000 to 660,000) years. This indicates that the clades L2.3.1 to L2.3.3 were early diverged as compared to the clades L2.3.4 to L.2.3.6.

**Conclusion:** In conclusion, this study shows that both Strict clock and Random local clock models fit well for the MTB dataset, leads to relative uniform evolutionary rates, across the trees. This might be as a result of using a strict prior on substitutional rates or other evolutionary factors such as due to a lack of calibration points as a results of an uninformative dataset. While the tMRCA difference might be due to possible potential differences that align with the data set. Hence there is a need for further investigation concerning other factors which might be influencing these evolutionary dynamics of the underlying dataset, possibly including the calibration information . This information can be utilized by other reserchers to assess and investigate an underlying divergence process and adaptation factors within the studied groups. Also incase of different dataset, a preliminary check should be done on specific dataset making sure if there is relatively uniformity in the evolutionary machanism, then as strict clock can be used for the simplification inthe analysis and less computational cost.

**Keywords:** Mycobacterium tuberculosis, Molecular clock models, Phylogenetics, Substitutional models, BEAST, tMRCA, ESS.

**Acknowledgement**

I am highly grateful to Almighty God for his grace and abundant blessings in my life.

I extend my deepest appreciation to my supervisor, Dr.Jade Vincent Membrebe(PhD), for his invaluable mentorship and constructive feedback throughout the research process. His expertise and insightful suggestions greatly enriched the quality of this work.

I am very thankful to Hasselt University for providing me with the necessary resources and facilities and creating an environment conducive to research and learning.

Very Special thanks to VLIR-OUS for believing in me and granting me a scholarship opportunity. Without their support, it would be difficult for me to be here.

My heartfelt appreciation to my mentor, Prof.Emmanuel Mpolya, I am sincerely grateful for the positive impact he has had on my academic journey from the beginning.

I am so grateful for my family. Your love, understanding, and patience have been my constant motivation. Thank you for standing by me through the challenges and celebrations and providing a nurturing environment that allowed me to focus on my studies.

I would like to express my sincere gratitude to my seniors, friends, and all those who played a role, directly or indirectly, in the completion of this research report. Your support has been invaluable, and I am truly thankful for the collaborative spirit that made this journey complete.

# Contents

# 1 Introduction

## 1.1 Background

Tuberculosis (TB) is an ancient infectious disease that is widely spread across the world. It is caused by Mycobacterium tuberculosis (MTB). TB has become a leading re-emerging infectious disease with a serious impact on global health challenges, especially in developing countries, but also as a dominant cause of death worldwide (Obi et al., 2010). WHO (2023) reported that in 2020 approximately 10 million people developed TB, and 1.5 million died from the disease globally.

In the realm of infectious diseases, MTB, as the causative agent of TB, continues to pose a significant global health threat (Grange, 2009). Many studies have been done on characterizing MTB strains in endemic countries, understanding the MTB evolution process, disease transmission, mechanisms of resistance, and adaptation to anti-tubercular therapies (Nguyen et al., 2004); (Mbugi et al., 2016); (Al-Mutairi et al., 2019). Understanding the evolutionary dynamics of MTB has become pivotal in explaining the transmission patterns, tracing the origins of outbreaks, and informing effective control strategies (Loiseau, 2020); (Sobkowiak et al., 2023). MTB is assumed to have diverged from its common ancestor as recently as 15,000 years ago. Subsequently, various genetic elements have evolved at different rates and can be used to understand the patterns of TB infection (Arnold, 2007). However, there is a lack of consensus on the rate of evolution of MTB. Some studies, particularly those relying on calibration based on sampling time, suggest a rate in the order of magnitude around $10^{-7}$ nucleotide substitutions/year, (Menardo et al., 2019). The discrepancies in these estimates arise when different metrics and approaches are employed, indicating potential variation in the rates of evolution among different species such as MTB.

Molecular clocks are among the approaches used for estimating the divergence times and the rate of evolution in genomic sequence datasets. These clocks have been widely used in molecular phylogenetics to infer the pattern and timing of evolutionary divergences (Lee, 2020). Different types of clock models have been developed to understand the dynamics of the evolutionary rate for the different species. Previously, a molecular clock model was developed based on the ideal concept of constant evolutionary rate (Zuckerkandl, 1962); (Zuckerkandl and Pauling, 1965) as Strict clock model, later was extended to account for the rate heterogeneity (Kishino and Hasegawa, 1989); (Kishino and Hasegawa, 1990) as Relaxed clock models (Drummond et al., 2006); (Drummond and Suchard, 2010).

The choice of a suitable clock model can be challenging, but some of the model selection techniques have been suggested(Ho and Duchêne, 2014). However, it is important to assess the adequacy of clock models to ensure the reliability of estimates. Methods such as Bayesian phylogenetic approaches and posterior predictive simulations have been proposed to evaluate the statistical evidence and model performance of clock models (Duchêne et al., 2015); (Didelot et al., 2021). These methods have been applied to various genomic data sets, including those from pathogens like MTB (Zhu et al., 2023) and foamy viruses (Membrebe et al., 2019). Particularly, this study evaluated these clock models to compare their performance while understanding the underlying MTB dataset.

### 1.2 Problem Statement and Research Questions

The rate of evolution within MTB remains a contentious issue, marked by divergent estimates from studies utilizing different calibration methods. Some of the current studies, such as Menardo et al. (2019), suggested the existence of varying clock-like behavior across the evolutionary tree. This lack of consensus hinders our comprehensive understanding of the evolutionary dynamics of MTB. To address this complexity, this study evaluated the performance of the different molecular clocks models using an advanced computational software called BEAST (Bayesian Evolutionary Analysis Sampling Tree) (Drummond and Rambaut, 2007) on previously published genome TB sequences dataset (Zhu et al., 2023). To uncover the clock-like nature of the MTB dataset, hence resolving the discrepancies in evolutionary rate estimates and shedding light on the underlying factors influencing the observed variability by addressing the following questions:

- How do different molecular clock models perform in reconstructing the phylogeny of MTB?
- What is the optimal molecular clock model for the MTB dataset, and how does it vary?

This study report consists of 9 consecutive sections: Section 1: an Introduction, which consists of Background information, a Problem statement, and research questions. Section 2: Detailed with Data Description and Data Exploration. Section 3 comprises the outlined methodologies that were used in the study. Section 4: Presents the results of all the analyses done in the study. Section 5: Discussion of the study findings. Section 6: Highlights the Drawbacks of the methods used in the study. Section 7: Gives Ethical thinking, Societal relevance, and Stakeholder awareness concerning the conducted study. Section 8: Gives the Conclusion of the study. Section 9: Outlined the Ideas for future research in the field.

## 2 Data

### 2.1 Data description

A dataset used in this study comprises 349 genome sequences of L2.3 strains of MTB for the 6 clades, which were obtained from 51 different countries, as published by Zhu et al. (2023). The dataset was available as a raw dataset in different formats, including (i) XML file format for the Bayesian skyline plot analysis, (ii)the fasta file containing the aligned and concatenated SNP sequences, and (iii) the original tree files generated by IQ-Tree, with bootstrap values. All these data were found uploaded in an online open-access repository website Figshare [Figshare]. Particularly, in this study data (i) XML file format dataset with the direct link: [xml.file] and (ii)the fasta file containing the aligned and concatenated SNP sequence with the direct link: [fasta.file], were used for all the analyses.

## 2.2 Data Exploration

Exploratory data analysis is an important part of the phylogenetic analysis. It involves evaluating the characteristics of the data before proceeding to the definitive analysis as discussed by Morrison (2010). In phylogenetic analysis, exploring a clocklike structure of the dataset is the key concept for the estimation of the evolutionary rate and timescales (Menardo et al., 2019). Some of the popular studies have assessed a clocklike structure using a root-to-tip regression and the date randomization test(DRT) (Menardo et al., 2019). Particularly in this study, a root-to-tip regression was employed to explore the data clocklike structure.

A root-to-tip regression(RTT) was employed to explore a clock-like behavior structure of the data. As root-to-tip can be used only for exploratory data analyses and not hypothesis testing (Rambaut et al., 2016), hence the method was suitable for this case. A root-to-tip regression is a regression of the root-to-tip distances as a function of sampling times of phylogenetic trees with branch lengths in units of nucleotide changes per site, where the slope corresponds to the rate. Under perfect clock-like behavior, the distance between the root of the phylogenetic tree and the tips is a linear function of the tip's sampling year (Menardo et al., 2019).

# 3 Methodology

## 3.1 Molecular Evolutionary Models

A mathematical model describing the evolution in time of the sequences can be built empirically, using properties calculated through comparisons of observed sequences or parametrically, using chemical or biological properties of DNA and amino acids (Liò and Goldman, 1998). These models help to estimate the genetic distance between sequences, measured by expected nucleotide substitutions per site along evolutionary lineages from a common ancestor. These distances are often depicted as branch lengths in phylogenetic trees, with extant/actual sequences as tips and ancestral sequences as internal nodes, typically unknown/unobserved (Liò and Goldman, 1998). To capture those unobserved events, substitutional models can be employed to compute the substitution rates to describe how sites within a sequence alignment change (Membrebe et al., 2022).

Substitutional models, which are sometimes called evolutionary sequence models, are defined as Markov models that describe changes in macro-molecules over evolutionary time (Arenas, 2015). These macromolecules can be in different forms, such as nucleotide sequence-based, amino acids sequence-based, and codon sequence-based macromolecules. Particularly, this study focused on DNA nucleotide sequence-based, which basically comprises thymine(T), cytosine(C), adenine(A), and guanine(G). T and C are pyrimidines, while A and G are purines. Two types of substitutions occur based on these biochemical categories: transitions within a category and transversions between categories, with their respective rates, denoted as $\alpha$ and $\beta$ as shown in figure 1,(Anisimova, 2019).

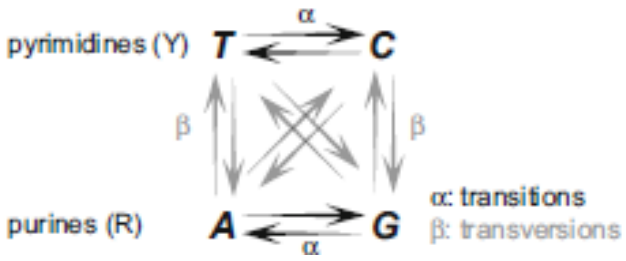



Figure 1: A Markov-chain model of nucleotide Substitution

### 3.1.1 Markov model

When modeling evolving biological systems, it is not strictly necessary to know the details of the underlying mechanisms. Instead, the changes can be modeled as a sequence of state transitions. Each transition is assigned a probability that defines the chance of the system changing from one state to another. Together with the states, these transition probabilities define a stochastic model with the Markov property: such a model is known as the Markov model(Grewal et al., 2019). A 'Markov property' implies a memoryless stochastic process which means that the probability of the current event depends only on the attained state of the previous event (Blumenthal, 1957)

### 3.1.2 Continuous-time Markov chains

Substitution models utilize a continuous-time Markov chain (CTMC) process to elucidate the transitions of nucleotides across the evolutionary tree towards the observed sequences (Bielejec et al. (2014)). A CTMC is a Markovian stochastic process represented by a substitution rate matrix, denoted as Q, which specifies the instantaneous rates of changes between pairs of characters ($q_{ij} \geq 0$ for $i \neq j$). Combined with the time parameter t, the rate matrix is employed to compute a matrix of finite-time transition probabilities of character changes through matrix exponentiation as given below:

$$P(t) = e^{Qt} \tag{1}$$

Where Q is assumed to be independent of $t$, the transition probabilities do not depend on $t$. A Markov process is assumed to exhibit time-reversibility, expressed as:

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t) \tag{2}$$

$$\textit{for all } i \neq j \textit{ and } t.$$

where $\pi_i$ and $\pi_j$ represent frequencies for corresponding nucleotide bases. This assumption presents another convenient mathematical simplification, suggesting that the substitution process is indistinguishable whether observed forwards or backward (Cannarozzi and Schneider, 2012); (Membrebe et al., 2022). However, this assumption may not align with biological reasoning.

### 3.1.3 Nucleotide substitution model

A nucleotide substitutional model describes the process of one nucleotide being substituted for another. When examining evolution at the DNA level, 12 potential rates of changes exist among the nucleotide bases (King and Jukes, 1969). Various nucleotide substitution models have been suggested, each employing distinct parameterizations for these rates. Such models include; Jukes-Cantor(JC) model, Hasegawa-Kishino-Yano (HKY) model (Hasegawa et al., 1985), general time-reversible(GRT) model (Tavare, 1986); (Yang, 1994); (Zharkikh, 1994), Kimura two-parameter (K80 or K2P) model (Kimura, 1980).

Specifically, this study employed an HKY model due to its capacity to consider different parameters for transitions(changes within purines or pyrimidines) and transversions(changes between purines and pyrimidines) while at the same time accounting for the different nucleotide equilibrium base compositions (Membrebe et al., 2022). The HKY model was formerly introduced for better modeling the substitution process in primate mtDNA (Hasegawa et al., 1985). A $4 \times 4$ substitution rate matrix of the HKY model ($Q_{HKY}$) has the following structure:

$$Q_{HKY} = \begin{pmatrix} . & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & . & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & . & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & . \end{pmatrix} \tag{3}$$

## 3.2 Phylogenetic reconstruction

Phylogenetic reconstruction is defined as the process of inferring and depicting the evolutionary relationships among a group of organisms or species based on shared genetic, morphological, or other biological characteristics (De Bruyn et al., 2014). The main role of this process is to construct an evolutionary tree known as a phylogeny or phylogenetic tree that represents the branching patterns and their hypothetical ancestors of the studied taxa or sequences (Nei and Kumar, 2000); (Felsenstein, 2004). The reconstruction of the phylogenetic trees using data collected from the extant species is a fundamental problem in evolutionary biology (Steel, 2016). Due to the uncertainty of the true evolutionary histories, this historical information can only be estimated. Hence, evolutionary trees are hypotheses about what happened in the past (Gregory, 2008). In a statistical framework, these computations depend on assumptions, methods used for analysis, and metrics that may affect the estimation accuracy of the evolutionary trees as a result of more complexity in the analysis.

Over the past decades, there have been many approaches for the reconstruction of evolutionary histories using sequence data, where Maximum likelihood was among the principle methodologies (Li, 1996). Due to technological advancement, recently, new statistical and computational methods such as the IQ-Tree (Nguyen et al., 2015), Bayesian-based (Rannala and Yang, 1996), Bootstrapping methods (Sanderson, 1995), have emerged in the field of molecular and evolutionary biology. Particularly, the analysis of this study based on the application of the Bayesian approach.

### 3.2.1 Bayesian evolutionary analysis

A Bayesian method for inferring evolutionary trees was once proposed by Rannala and Yang (Rannala and Yang, 1996) using nucleotide data as an alternative to maximum likelihood estimation (Felsenstein, 1981). Unlike parsimony and distance-based approaches, maximum likelihood and Bayesian estimation approaches rely on the likelihood function. The likelihood function is defined as the probability of the observed data conditioned on the model's parameters.

In particular, Bayesian analysis aims to compute the joint posterior probability distribution of the parameters. This is done using Bayes' theorem (Koch and Koch, 1990) which is given by:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}, \tag{4}$$

where; $p(.)$ is denoted as either a probability or a probability density, depending on whether $\theta$ and $D$ are discrete or continuous. In the phylogenetic framework, the data D is generally a discrete multiple sequence alignment, and the parameters ($\theta$) include continuous components such as branch lengths and substitution rate parameters, hence a probability distribution is defined as:

$$f(\theta|D) = \frac{Pr(D|\theta)f(\theta)}{Pr(D)}, \tag{5}$$

where; $Pr(D|\theta)$ is a likelihood, $f(\theta)$ is the prior distribution for parameters($\theta$) and $Pr(D)$ is the marginal likelihood.

In the Bayesian phylogenetics framework, evolutionary trees, and their associated parameters are estimated as probability distributions, and the statistical inference can be performed by using the Markov Chain Monte Carlo (MCMC) algorithm, using software such as BEAST (Drummond and Rambaut, 2007), BEAST 2 (Bouckaert et al., 2014), MrBayes (Huelsenbeck and Ronquist, 2001) and RevBayes (Höhna et al., 2016). These Bayesian approaches allow for the incorporation of prior knowledge, uncertainty, and complex models, making them well-suited for molecular clock models. Particularly in this study, BEAST (Bayesian Evolutionary Analysis Sampling Trees) software was chosen for the analysis to answer the research questions of interest due to its flexibility.

## 3.3 Molecular clock models

A molecular clock model presents a means of estimating evolutionary rates and timescales using genetic data (Ho and Duchêne, 2014). These models originated from the molecular clock hypothesis, which stated that the evolutionary rates of biological sequences are approximately constant through time (Zuckerkandl, 1962). In this study, three molecular clock models were considered, a Strict clock, Random local relaxed clock, and Uncorrelated relaxed clock models, to answer the questions of interest. These models were fitted in BEAST based on their different assumptions and parameter specifications. In constructing the molecular clock models, it is important to consider the rates of character change through time and lineages (Tay et al., 2023). Clock models can be applied to both molecular and morphological characters, which eventually improves estimates of the tree of life (Warnock et al., 2017). Statistical methods have been developed to cope with uncertainties in molecular evolution, such as variation in rates and difficulty in calibrating clocks (Kumar, 2005). Finally, the substitution process in molecular evolution is often overdispersed, and models that take temporal variation into account have been developed as relaxed clock models. Generally, as explained by Drummond and Bouckaert (2015), the Bayesian methods require specifying priors, which require extra effort and care but also allow containing the analysis when information about a certain parameter is available from independent sources or literature or from fossil records. For example, in this case, the substitution rate was retrieved from the previously done studies on the same dataset for the MTB (Bos et al., 2014); (Zhu et al., 2023).

### 3.3.1 Strict clock Model

A Strict clock model is defined as the simplest molecular clock model that assumes a constant evolutionary rate for the sequences of interest, hence it's a one-parameter model (Zuckerkandl, 1962); (Zuckerkandl and Pauling, 1965). The parameter of which represents the conversion rate between branch lengths and evolutionary time. This rate is commonly expressed as substitutions per site per year or substitutions per site per million years (Brown and Yang, 2011).

A Strict clock (STR) was fitted for the six clades separately in the BEAUti package in BEAST. A nucleotide substitution model HKY was chosen without specifying the heterogeneity rate of the site-specific due to its simplicity in fitting clock models. The prior for the clock rate was specified based on the previous studies (Bos et al., 2014); (Zhu et al., 2023) as a normal distribution with a mean of $4.6 \times 10^{-8}$ and a standard deviation of $1 \times 10^{-8}$. The remaining parameters were left with set by default values. Thereafter, the MCMC chain of

size 10,000,000 was run with 10% burn-in excluded, for the different clades separately. A demonstrational Strict clock is given below as figure (2), which shows the same evolutionary rate presented by the same color across the entire evolutionary tree.

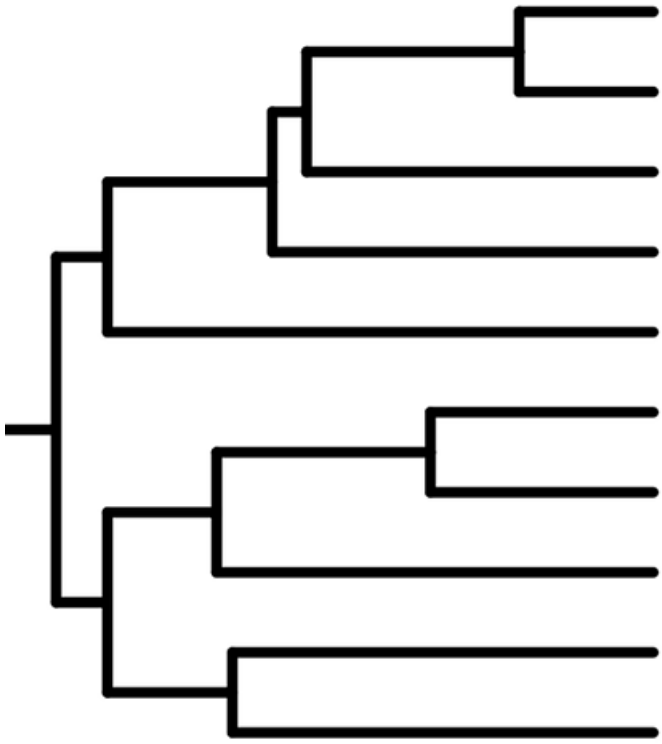


Figure 2: A demonstrational Strict Clock

### 3.3.2 A random local clock model

A random local clock model was developed to relax the restrictive assumption of the 'strict' clock by allowing different rates in different regions of the evolutionary tree, but within each region, the rate must be the same (Drummond and Suchard, 2010). Theoretically, when the number of rate changes becomes zero, the resulting clock model is a Strict clock model, while if the number of rate changes is equal to the number of branches, the resulting model is a relaxed clock model that estimates an evolutionary rate for each branch (Membrebe et al., 2022).

A Random local clock (RL) was fitted for the six clades separately in the BEAUti package in BEAST. Prior distributions were specified quite similarly to those in a strict clock, whereby a nucleotide substitution model HKY was considered without specifying the heterogeneity rate of the site-specific due to its simplicity in fitting clock models. The prior distribution for clock rate was specified based on the findings of the previous studies (Bos et al., 2014); (Zhu et al., 2023) as a normal distribution with a mean of $4.6 \times 10^{-8}$ and a standard deviation of $1 \times 10^{-8}$. The remaining parameters were left with set by default values. Thereafter, the MCMC chain of size 200,000,000 and 500,000,000 was run with 10% excluded as burn-in; this was done for the different clades separately. A demonstrational Random local clock is given below as figure

(3), which shows different evolutionary rates between different regions across the evolutionary tree, presented in different colors between the two regions(lineage A and lineage B).

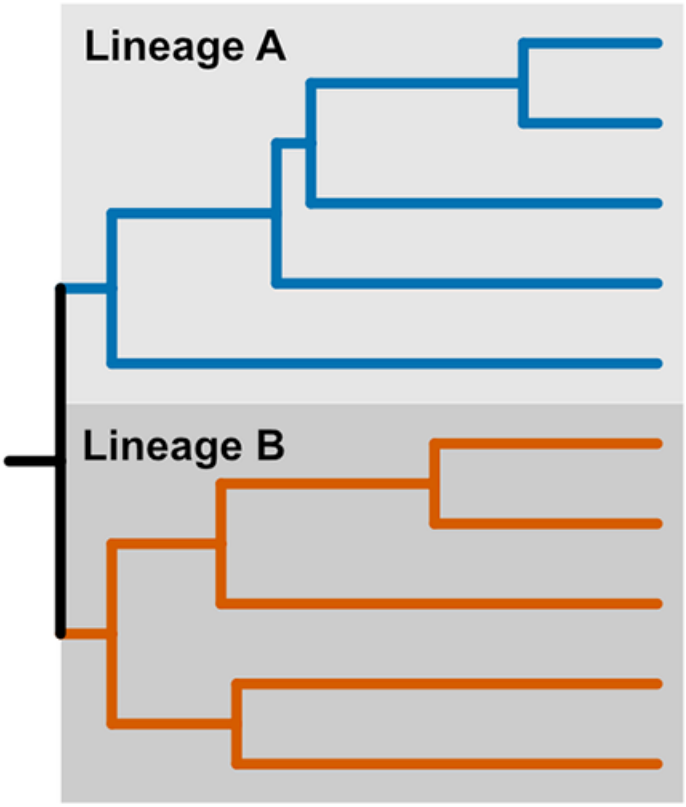



Figure 3: A demonstrational Random local Clock

### 3.3.3 Uncorrelated relaxed clock Model

The uncorrelated clock model is defined as a relaxed model that assumes a priori no dependency on the evolutionary rates of the adjacent branches of the evolutionary tree. Instead, the rate on each branch of the evolutionary tree is drawn independently and identically from an underlying rate distribution (Drummond et al., 2006); (Lepage et al., 2007); (Rannala and Yang, 2007). An uncorrelated lognormal relaxed clock (UCL) was fitted for the six clades separately in the BEAUti package in BEAST. Six different analyses with different clade datasets were run BEAST but with the settings as specified below; The Uncorrelated relaxed clock model (Drummond et al., 2006) with lognormal distribution was chosen, and a nucleotide substitution model HKY was considered without specifying the heterogeneity rate of the site-specific as it was specified in the above analyses of Strict and Random local clocks. The parameter of the uncorrelated relaxed clock, such as the mean of the branch rate(ucld.mean), was assumed to follow a normal distribution with a mean of $4.6 \times 10^{-8}$ and a standard deviation of $1 \times 10^{-8}$ and standard deviation (ucld.stdev or $\sigma$) was assumed to follow lognormal with a mean of $\frac{1}{3}$ and standard deviation of 1 respectively as their prior distributions. The remaining parameters were left with set by default values. Thereafter, the MCMC chain of size 200,000,000 and 500,000,000 was run with 10% burn-in; this was done for the different clades separately. A demonstrational Uncorrelated relaxed clock is given below as figure (4), which shows different

evolutionary rates between and within different branches across the evolutionary tree, presented in different colors.

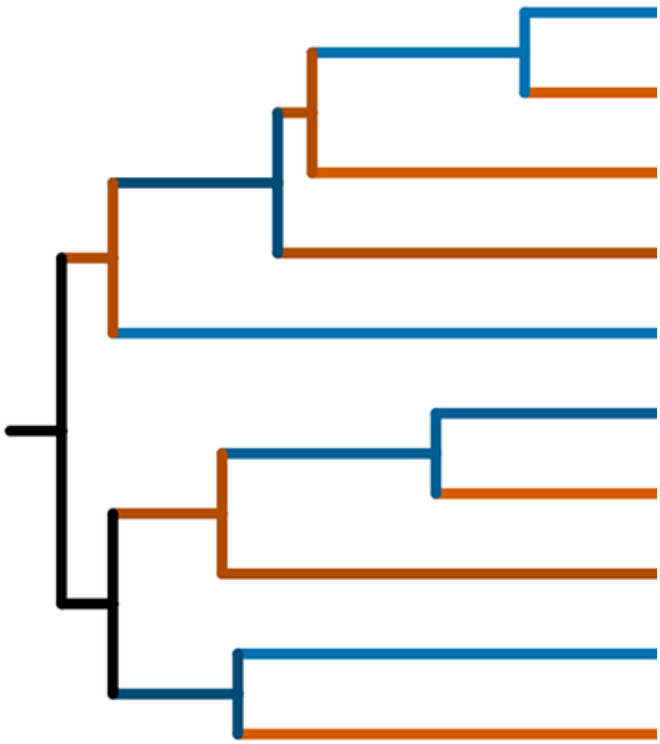


Figure 4: A demonstrational Uncorrelated Relaxed Clock

### 3.4 Model performance assessment

Model performance assessment plays a crucial role in the comparative analysis of the different models. The fitted models were then assessed using a Tracer package(version 1.7.1) in BEAST. This tool analyzes and visualizes output files generated during Bayesian phylogenetic analyses (Rambaut et al., 2018). The primary purpose of using a tracer is to examine and interpret the results of Markov chain Monte Carlo (MCMC) simulations conducted in BEAST. In this study, the fitted clock models were assessed based on MCMC chain convergence shown by trace plots, ESS (Effective Sample Size), and information about parameter estimates with their respective uncertainties.

#### 3.4.1 MCMC Convergence Diagnostics

Markov chain Monte Carlo (MCMC) is a powerful means for generating random samples used in statistical computations (Metropolis et al., 1953); (Hastings, 1970). MCMC plays a crucial role in Bayesian analysis in tackling complex problems where the direct computation of posteriors(analytical solution) is infeasible. This method relies on dependent sequences (Markov chains) that eventually converge to a limiting distribution, which corresponds to the distribution of interest. In other words, MCMC allows us to explore complex probability distributions

by simulating sequences of correlated samples (Spall, 2003). When exploring the output from the phylogenetic Markov chain, Monte Carlo (MCMC) simulation visualization and diagnostic analysis provide intuitive and often crucial insights into the success and reliability of the analysis. Different diagnostic tools such as trace plots, Effective sample size(ESS), and posterior distributions can be used for assessing convergence, identifying potential issues, and validating their results (Nylander et al., 2008). Particularly, this study utilized trace plots and ESS to assess convergence.

**Trace plots**

Trace plots are essential diagnostic tools in Bayesian phylogenetic inference using the Markov chain Monte Carlo(MCMC) approach in assessing the convergence(Ali et al., 2017). These plots are utilized in MCMC analyses within software like BEAST to visually assess the convergence, monitor parameter behavior, and ensure reliable results. Hence, the chain was considered converged when the trace plots for each variable were mixed well.

**Effective sample size**

The effective sample size is a measure of the number of independent samples from the marginal posterior distribution that the trace is equivalent to (Thiébaux and Zwiers, 1984). ESS needs to be above 100 for valid estimates of the posterior distribution of all parameters in the model(Li, 2010). However, it is recommended that ESS should be at least 200 to ensure more accurate inference of the posterior distributions (Drummond et al., 2006). Therefore, the higher the ESS, the better the accuracy and reliability of the parameter estimation of the tree topology. For the low ESS, there are different ways to increase ESSs of the parameters, whereby in this study, this was done by increasing the chain length of the MCMC. Hence, the chain was considered converged when the ESS for each variable was found to be above 100.

## 3.5 Model comparison and Model selection

A comparative analysis in the Bayesian context is done based on the Bayes Factor(BF), which is defined as the ratio of the two marginal likelihoods from the two models of interest (Jeffreys, 1935). In this study, marginal likelihood estimation(MLE) was done in BEAST using a generalized stepping-stone sampling (GSS) algorithm as a working distribution, which performs an additional analysis after the standard MCMC chain has finished (Baele et al., 2016). The use of Bayesian approaches in phylogenetics has increased in recent years due in part to the availability of software, including BEAST (Drummond et al., 2012) and MrBayes (Ronquist et al., 2012).

### 3.5.1 Marginal Likelihood estimation

In a Bayesian framework, the marginal likelihood is how the data update our prior beliefs about the models, which is commonly used to compare the model fit that is grounded in probability theory (Oaks et al., 2018). In phylogenetic analysis, when learning about the evolutionary patterns and processes by using Bayesian-based methods, comparing the models obtained requires the calculation of marginal likelihoods. A marginal likelihood of a given phylogenetic model $M$ can be derived using Bayes' theorem in expression (4) and given as:

$$P(D|M) = \int_{\theta} P(D|\theta, M)P(\theta|M)d\theta, \tag{6}$$

where $D$ represents the observed molecular data, $\theta$ denotes the model's parameters, and M represents the mode. Hence, the ratio of the two marginal likelihoods from the two models of interest, let's say $M_1$ and $M_2$ which is the Bayes Factor(BF), is given as:

$$BF = \frac{P(D|M_1)}{P(D|M_2)}, \tag{7}$$

However, the computation of marginal likelihood estimate has its complexities, especially in phylogenetics, where models are composed of a large number of parameters (Membrebe et al., 2022). As a result, several methods have been proposed for the marginal likelihood estimation(MLE), such as path sampling (PS)(Lartillot and Philippe, 2006), stepping-stone sampling (Xie et al., 2011) and generalized stepping-stone sampling(GSS) (Fan et al., 2011). A GSS approach has been proposed as a powerful tool for MLE computation under specified working distributions, which accommodates phylogenetic uncertainty while avoiding numerical issues as compared to PS and SS, which tend to overestimate the log(marginal likelihood) when vague priors are used (Baele et al., 2016). Particularly, this study employed a GSS with a coalescent working distribution since a constant population size of a Coalescent was chosen prior to the tree for the fitted clock models.

However, the MLE in BEAST is given in terms of ln(MLE); hence, the BF was calculated as $ln(BF) = ln(MLE)_{M_1}$ - $ln(MLE)_{M_2}$ and their interpretation in this study, were according to Kass and Raftery (1995) proposed threshold as follows;

Table 1: Interpretations of the Bayes factors(BFs)

| **BF range** | **ln(BF) range** | **Interpretation** |
|---|---|---|
| 1-3 | 0-1.1 | Weak evidence |
| 3-20 | 1.1-3 | Positive evidence |
| 20-150 | 3-5 | Strong evidence |
| 150+ | 5+ | Very strong evidence |

### 3.5.2 tMRCA estimation

The time to the most recent common ancestor (tMRCA) is the height of the genealogical/evolutionary tree that unites all the sampled sequences from a given locus that originated from a common ancestor, in the absence of intralocus recombination (King and Wakeley, 2016). In this study, estimates for the age of the tree (tMRCA) serve as a crucial parameter of interest, representing the age/time of the most recent common ancestor (tMRCA) of all taxa under study. To estimate tMRCA, the molecular clock models have been employed, which can be strict by assuming a constant evolutionary rate or relaxed by allowing variation in evolutionary rate across lineages. Analysis of this study utilized the Bayesian inference approach implemented in BEAST (Bayesian Evolutionary Analysis Sampling Trees) software (Drummond and Rambaut, 2007), incorporating prior information from previous studies to compare the performance

of the different clock models applied. Therefore, this study considered the tMRCA estimates of the evolutionary trees generated by the optimal selected clock models after a comparative analysis conducted to answer the questions of interest.

# 4 Results

## 4.1 Data Description

Table (2.) below, is the summary of the dataset that was used in all the analysis of the study. A total of 349 sampled sequences with a Phred base quality above 20 and read length longer than 30 (Zhu et al., 2023). These sequences with different numbers of sites, were categorized by six different clades of the modern Beijing strains of MTB.

Table 2: Summary of the data description

| Clade | Taxa(Sequences) | Sites |
|---|---|---|
| **L2.3.1** | 34 | 2138 |
| **L2.3.2** | 20 | 1371 |
| **L2.3.3** | 37 | 2541 |
| **L2.3.4** | 77 | 4711 |
| **L2.3.5** | 103 | 5971 |
| **L2.3.6** | 78 | 4696 |

## 4.2 Data Exploration

A root-to-tip(RTT) regression, by using mid-rooted trees for the six clades differently, was used to explore a clock-likeness behavior on the data. Below are the resulting regression plots, which give an insight into the clock-like behavior of the dataset for the different clades. Figure (5, 7, 9 and 10), give weak negative slopes for the relation between the root to tip distance and the tips for clades L2.3.1, L2.3.3, and L2.3.5 which might mean that there is not enough evidence to detect clock-like structure in the dataset. While Figure (6 and 8) for the clade L2.3.2 and L2.3.4, respectively, give a weak positive slope for the relation between the root-to-tip distance and the tips which give insight that there's weak clock-likeness behavior for the given clades. These results are not significantly proven; further analysis needs to be done to determine the significance of the results. Hence, these will be proven by model fitting and comparative analysis in the later sections.

File: tree.midL1.txt
Slope: -0.08925941838930994
Intercept: 5.5192317135150075
R-squared: 0.09020200318925205
P-value: 0.08437605001135726
<Figure size 640x480 with 0 Axes>

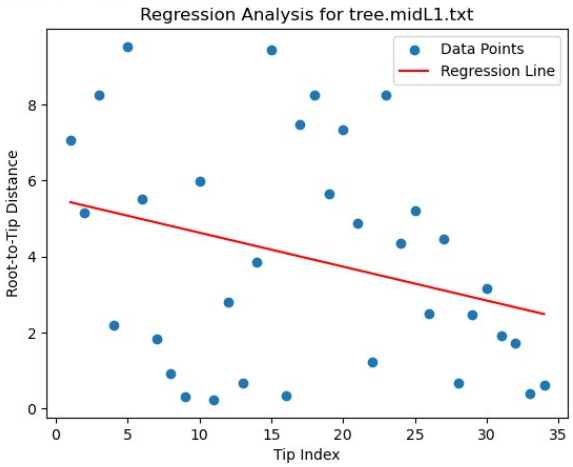



Figure 5: Root-to-tip regression for clade L2.3.1

File: tree.midL2.txt
Slope: 0.07413656851970689
Intercept: 3.7802472478838416
R-squared: 0.030045992106090755
P-value: 0.46487896738926904

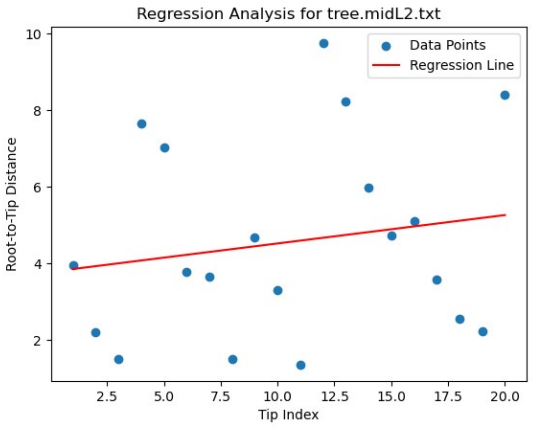



Figure 6: Root-to-tip regression for clade L2.3.2

File: tree.midL3.txt
Slope: -0.08302560756276481
Intercept: 7.587465270869642
R-squared: 0.09474006459404233
P-value: 0.06384146020372257

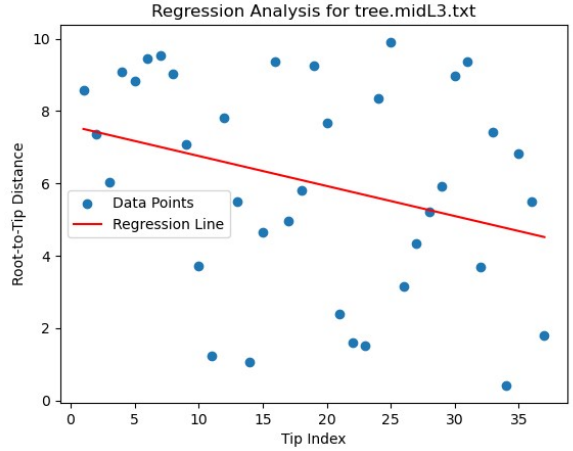



Figure 7: Root-to-tip regression for clade L2.3.3

File: tree.midL4.txt
Slope: 0.00838905578665922
Intercept: 4.57804359078921
R-squared: 0.00431896832457632 44
P-value: 0.5701282096404325

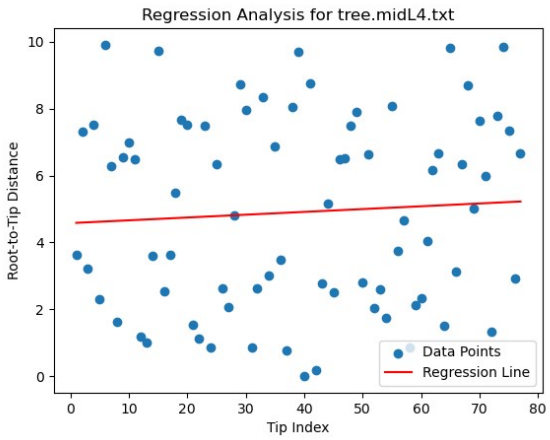



Figure 8: Root-to-tip regression for clade L2.3.4

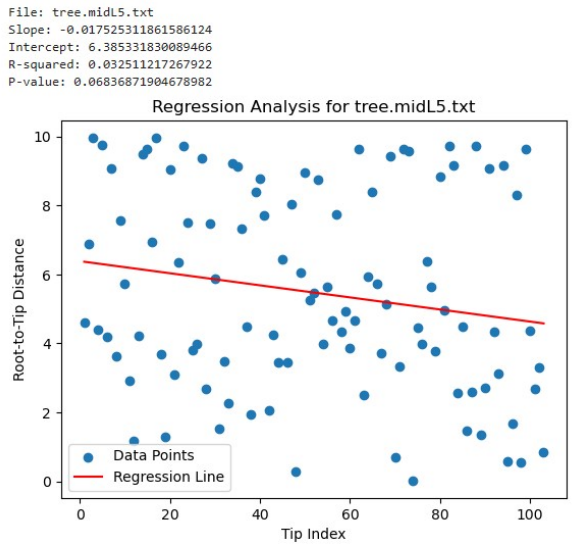



Figure 9: Root-to-tip regression for clade L2.3.5

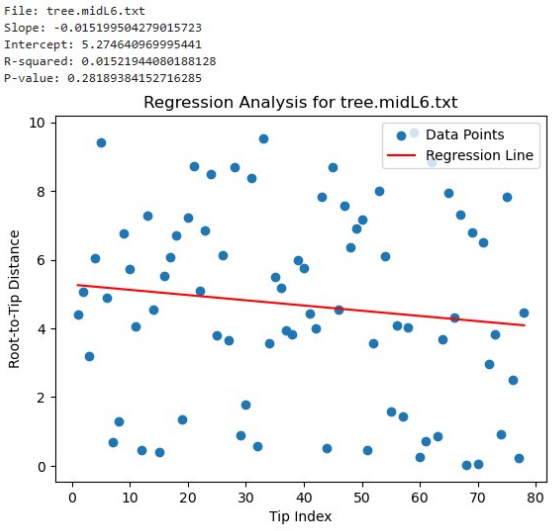



Figure 10: Root-to-tip regression for clade L2.3.6

## 4.3 Model perfomance assessment

### 4.3.1 Trace plot and ESS diagnostics

After fitting all clock models for each clade separately, by running the MCMC simulation, trace plots and ESS diagnostics were used to assess the convergence of the chain as given below;

**Strict clock Model**

The trace plots for all the parameters of the fitted strict clock in all of the six-clades were well converged and well mixed. Specifically, the trace plots for the estimated tMRCA( age of the most recent common ancenstor) were presented on this study given as figure (17, 20, 23, 26, 29, and 32)(in appendix). Furthermore, interms of ESS, all the parameters included in a fitted Strict clock model were above 200 as shown in the tables in figure (35,36, 37, 38, 39, and 40)(in appendix). These results were obtained from an MCMC run of 10,000,000 iterations without adjustment on the chain iteration number. Therefore, the results suggest that a strict clock model could be a suitable fit for the data of all the clades. However, a comprehensive comparative analysis of the different clock models is necessary to ensure meaningful and statistically significant results.

**Random Local Clock Model**

The results of the fitted random local clocks show that the trace plots for all the parameters of the fitted in all of the six-clades were well converged and well mixed. The trace plots of the parameter of interest are presented in figure (18, 21, 24, 27, 30 and 33)(in appendix). Furthermore, all parameters of the four clades(L2.3.1, L2.3.2 ,L2.3.3 and L2.3.6), have $ESS > 200$. The clades(L2.3.4 and L2.3.5) have respectively six parameters and one parameter which have $100 < ESS < 200$, while the remaining parameters including the ones of interest have ESS above 200. Tables for the ESS for the parameters of the six clades are shown in figures(41, 42, 43, 44, 45 and 46)(in appendix). Therefore, the results obtained on ESS may suggest that a

Random local clock model could be a suitable fit for the data of all the clades L2.3.1, L2.3.2, L2.3.3, and L2.3.6. However application of random local clock might not be suitable for the clade L2.3.4 and clade L2.3.5, which may lead into unreliable conclusion. Hence, a comprehensive comparative analysis of the different clock models is necessary to ensure meaningful and statistically significant results.

**Uncorrelated Relaxed Clock Model**

The results of the fitted Uncorrelated Relaxed clock model show that all parameters in clade L2.3.1, L2.3.2, L2.3.4, L2.3.5, L2.3.6 had $ESS > 200$ except for some parameters(four parameters, exclusive of the interested ones) in clade L2.3.3 which had $100 < ESS < 200$ as shown in the tables (47, 48, 50, 51, 52 and 49)(in appendix), respectively. Also, the trace plots were well converged and well mixed in the chain for all the parameters of the five clades, except for clade L2.3.3 even after adjusting the number of iterations. The trace plots for the parameter of interest tMRCA were presented in the figures (19, 22, 28, 31, 34 and 23)(in appendix), respectively. These results suggest that an Uncorrelated relaxed clock model could be a suitable fit for the data of all the clades except clade L2.3.3. However, a comprehensive comparative analysis of the different clock models is necessary to ensure meaningful and statistically significant results.

## 4.4 Model comparison and model selection

### 4.4.1 Marginal Likelihood estimation

A comparative analysis of the different molecular clock models considered in this study was done based on Bayes factor (7) calculated from the log of the marginal likelihood estimates of the models to be compared, ie. log(MLE) as presented on the table of results table (3). Subsequently, the interpretation of the results was done on the selected optimal clock model in each clade separately with their corresponding phylogenetic trees.

Table 3: Estimates of the log(MLE) per each clade

| Clade | Clock | ln(MLE) | Clade | Clock | ln(MLE) |
|---|---|---|---|---|---|
| L2.3.1 | Str | -14981.998 | L2.3.4 | Str | -37496.966 |
| | Rl | -14982.571 | | Rl | -37496.226 |
| | Ucl | -14998.742 | | Ucl | -37502.853 |
| L2.3.2 | Str | -8962.809 | L2.3.5 | Str | -49698.391 |
| | Rl | -8963.363 | | Rl | -49699.369 |
| | Ucl | -8980.160 | | Ucl | -49701.176 |
| L2.3.3 | Str | -18101.161 | L2.3.6 | Str | -37120.490 |
| | Rl | -18052.215 | | Rl | -37114.284 |
| | Ucl | -18085.682 | | Ucl | -37118.986 |

**Note:** **Str**=Strict Clock, **Ucl**=Uncorrelated Lognormal Clock, **Rl**=Random Local Clock, **MLE** = Marginal likelihood estimate

**Clade L2.3.1:** The results in table (3); show that the STR and RL were found to have quite

the same ln(MLE) lead to a small Bayes factor(lnBF) of approximately less than 1, but higher than UCL by approximately equal or greater than 15 Bayes factor, when compared to both STR and RL respectively. This suggests that there is weak evidence of the differences between STR and RL in explaining well the clade L2.3.1 dataset, while there is very strong evidence against UCL in favor of either STR or RL.

**Clade L2.3.2:** The results in table (3); show that the STR and RL were found to have quite the same ln(MLE) led to a small Bayes factor(lnBF) approximately less than 1 Bayes factor, but higher than UCL ln(MLE) by approximately equal to 17 Bayes factor, when compared to both STR and RL respectively. This suggests that there is weak evidence of the difference between STR and RL in explaining well the Clade L2.3.2 dataset, while there is very strong evidence against UCL in favor of either STR or RL.

**For the Clade L2.3.3:** The results in table (3); show that the RL has the highest ln(MLE) as compared to UCL by approximately Bayes factor(lnBF) of 17.89 and to STR by approximately Bayes factor of 33.47. This suggest that, there is very strong evidence in favor of RL against either STR or UCL. This makes an obvious selection of RL as an optimal clock model to explain the Clade L2.3.3 dataset.

**For Clade L2.3.4:** The results in table (3); show that the STR and RL were found to have quite the same ln(MLE) led to a small Bayes factor(lnBF) approximately equal to 0.18 Bayes factor when compared to both STR and RL respectively. This suggests that there is an insignificant difference between STR and RL in explaining well the Clade L2.3.4 dataset, while there is positive evidence against UCL in favor of either STR or RL.

**For Clade L2.3.5:** The results in table (3); show that the STR and RL were found to have quite the same ln(MLE) with a difference of approximately equal to 1 as Bayes factor(lnBF), but approximately a bit higher than the ln(MLE) of UCL by approximately to 3 as Bayes factor(lnBF) when compared to both sc and RL respectively. This suggests that there is an weak evidence of the difference between STR and RL in explaining the Clade L2.3.5 dataset.

**Clade L2.3.6:** The results in table (3); show that the RL has the highest ln(MLE) as compared to STR by approximately Bayes factor(lnBF) of 6.21 and to UCL by approximately Bayes factor of 4.7. This suggest that, there is strong evidence in favor of RL against either STR or UCL. This makes an obvious selection of RL as an optimal clock model to explain the Clade L2.3.6 dataset.

Therefore, based on the BF mode of comparison, for the clades(L2.3.1, L2.3.2, L2.3.4, and L2.3.5) where both STR and RL were favored, then STR was preferred over RL based on the parsimony principle for statistical models (Coelho et al., 2019). A Strict clock model (STR) is considered a simple model that contains few parameters as compared to the Random local clock model (RL). While for the clade(L2.3.3, and L2.3.6), RL was selected as it was strongly favoured against both STR and UCL. Hence, the reported tMRCA estimates and phylogenetic trees in the next subsection were based on the selected optimal clocks only.

### 4.4.2 tMRCA Estimation

Here, the final results on tMRCA estimates as shown in the table (4) and phylogenetic trees of the optimal clock models were presented based on the overall results in Appendix table (5).

Table 4: Estimates for the Optimal clock models per clade

| Clade | Clock | tmrca mean(years) | 95% HPD | Par | Mean | 95% HPD |
|---|---|---|---|---|---|---|
| L2.3.1 | STR | 1.192E6 | [7.205E5,1.982E6] | O.R | 4.627E-8 | [2.732E-8, 6.614E-8] |
| L2.3.2 | STR | 1.878E6 | [1.126E6,2.854E6] | O.R | 4.615E-8 | [2.647E-8, 6.567E-8] |
| L2.3.3 | RL | 9.776E5 | [5.745E5,1.497E6] | I.R | 4.612E-8 | [2.644E-8, 6.550E-8] |
| | | | | R.C | 1.194 | [1, 2] |
| L2.3.4 | STR | 4.867E5 | [2.868E5, 7.220E5] | O.R | 4.629E-8 | [2.748E-8, 6.694E-8] |
| L2.3.5 | STR | 4.113E5 | [2.469E5, 6.023E5] | O.R | 4.616E-8 | [2.682E-8, 6.556E-8] |
| L2.3.6 | RL | 6.575E5 | [3.403E5, 1.132E6] | O.R | 4.589E-8 | [2.685E-8, 6.597E-8] |
| | | | | | 3.572 | [1, 6] |

**Clade L2.3.1:** A simpler model STR was selected over RL, as the optimal clock model to explain the Clade L2.3.1 dataset, with a constant overall rate of $4.627 \times 10^{-8}$ substitutions per site per year, with the estimated tMRCA of 1,191,600 (720,460 to 1,982,000) years, as shown in table (4). Hence, a reconstructed phylogenetic tree of the strict clock for the clade L2.3.1 is presented in figure (11), showing a constant evolutionary rate(all branches with the same color) across the phylogeny.

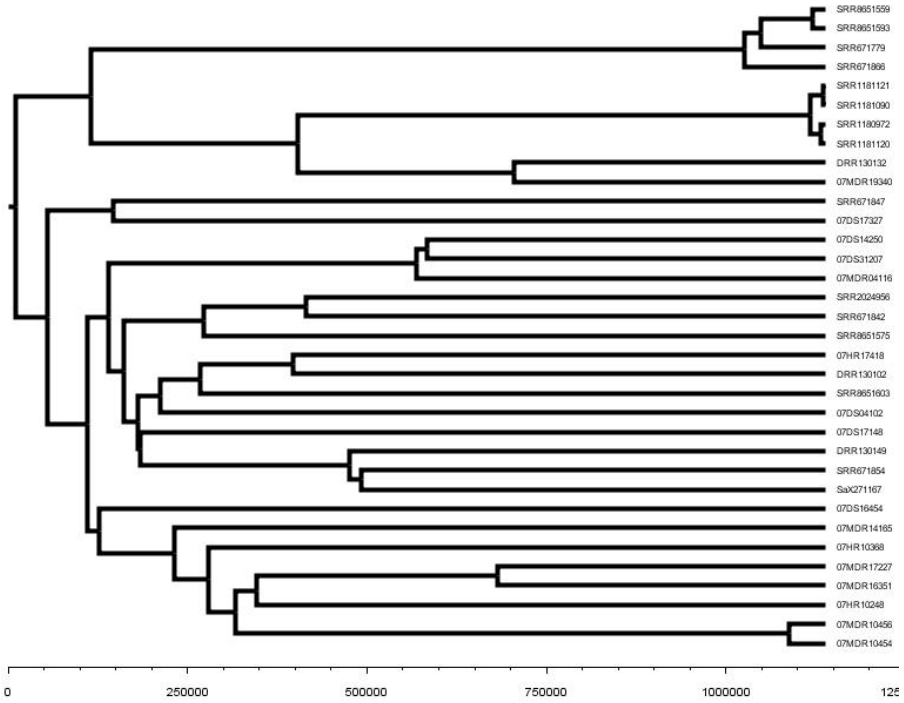



Figure 11: Phylogenetic tree for clade L2.3.1

**Clade L2.3.2:** A simpler model STR was selected over RL, as an optimal clock model to explain the Clade L2.3.2 dataset, with a constant overall rate of $4.615(2.647, 6.567) \times 10^{-8}$

substitutions per site per year, with the estimated tMRCA of 1,878,000 (1,126,000 to 2,854,000) years, as shown in table (4). Hence, a reconstructed phylogenetic tree of the strict clock for the clade L2.3.2 is presented in figure (12), showing a constant evolutionary rate(all branches with the same color) across the phylogeny.

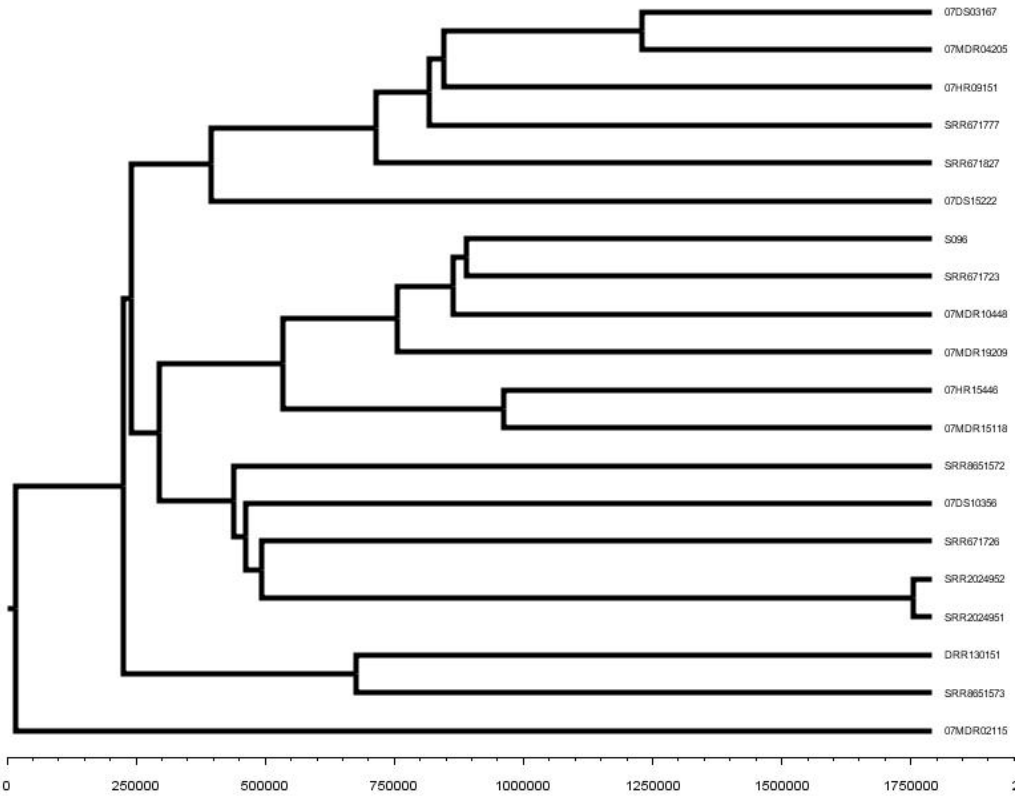



Figure 12: Phylogenetic tree for clade L2.3.2

**Clade L2.3.3:** A RL was selected as the optimal clock model to explain the Clade L2.3.3 dataset with the initial rate of $4.612(2.644, 6.550) \times 10^{-8}$ substitutions per site per year and the corresponding estimated tMRCA of 977,600 (574,510 to 1,497,200) years as shown in table (4). Hence, a reconstructed phylogenetic tree for the clade L2.3.3 is presented in figure (13) shows different evolutionary rates for at least three regions(rate by color) across the phylogeny.

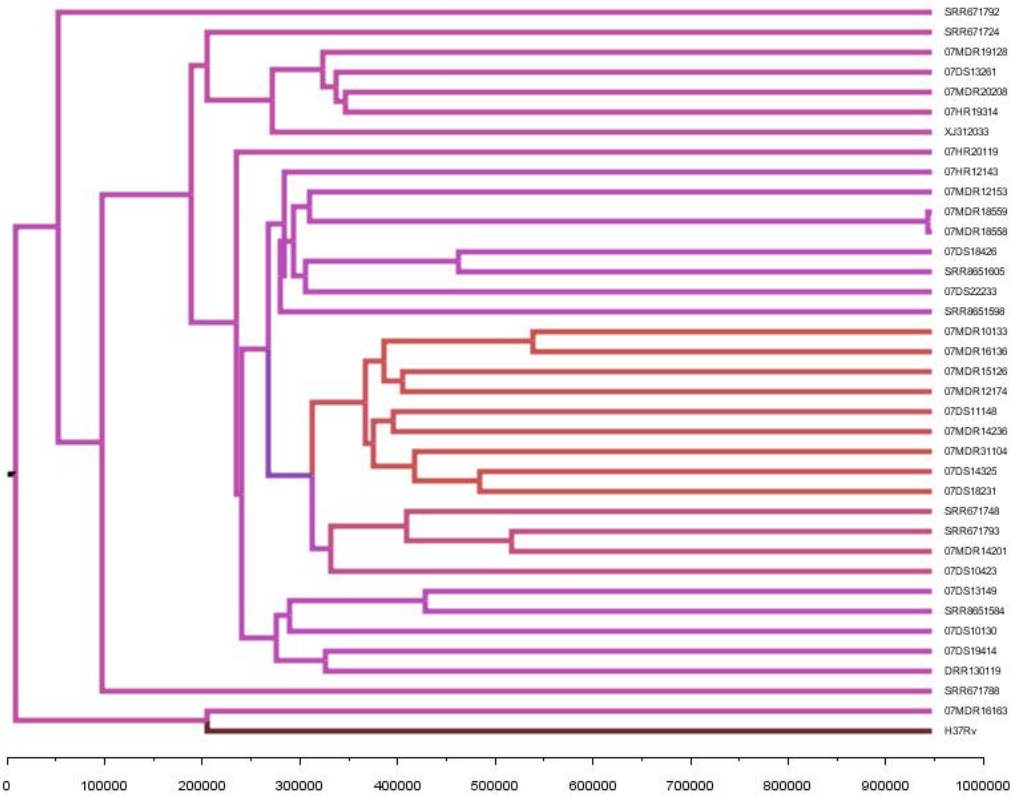



Figure 13: Phylogenetic tree for clade L2.3.3

**Clade L2.3.4:** A simpler model STR was selected over RL, as an optimal clock model to explain the Clade L2.3.4 dataset, with a constant overall rate of $4.629(2.748, 6.694) \times 10^{-8}$ substitutions per site per year, with the estimated tMRCA of 486,700 (286,780 to 722,000) years, as shown in table (4). Hence, a reconstructed phylogenetic tree of the strict clock for the clade L2.3.4 is presented in figure (14), showing a constant evolutionary rate(all branches with the same color) across the phylogeny.

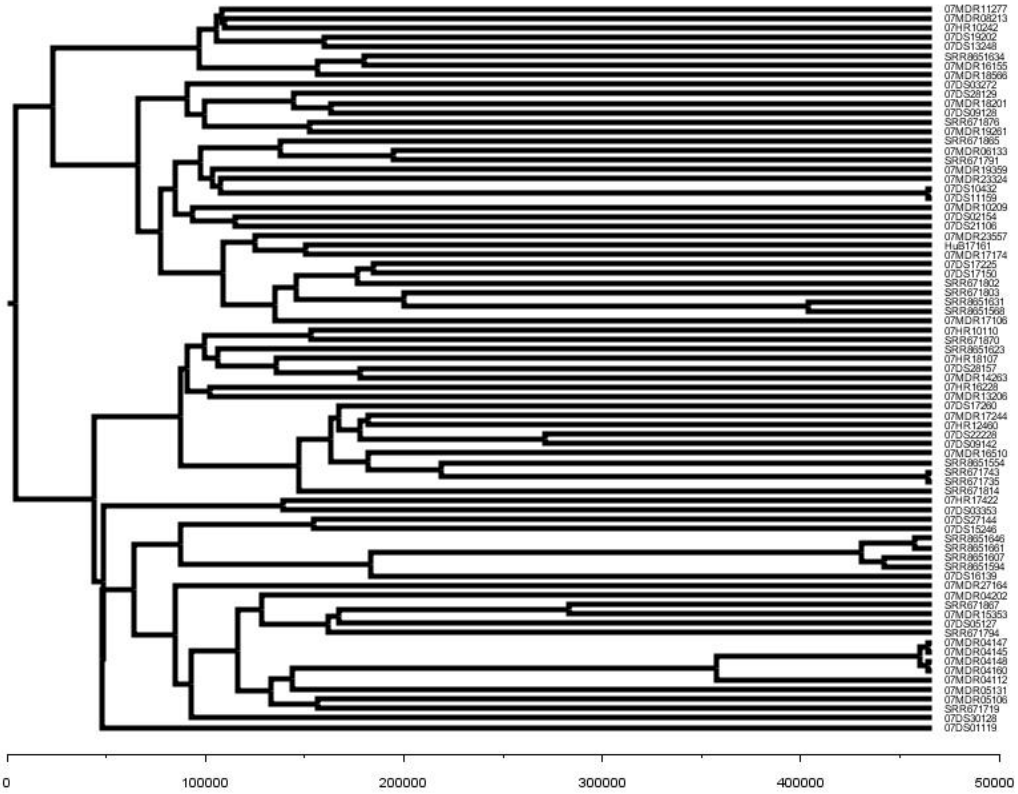



Figure 14: Phylogenetic tree for clade L2.3.4

**Clade L2.3.5:** A simpler model STR was selected over RL, as an optimal clock model to explain the Clade L2.3.5 dataset, with a constant overall rate of $4.616(2.682, 6.556) \times 10^{-8}$ substitutions per site per year, with the estimated tMRCA of 411,300 (246,900 to 602,300) years, as shown in table (4). A reconstructed phylogenetic tree of the strict clock for the clade L2.3.5 is presented in figure (15), showing a constant evolutionary rate(all branches with the same color) across the phylogeny.

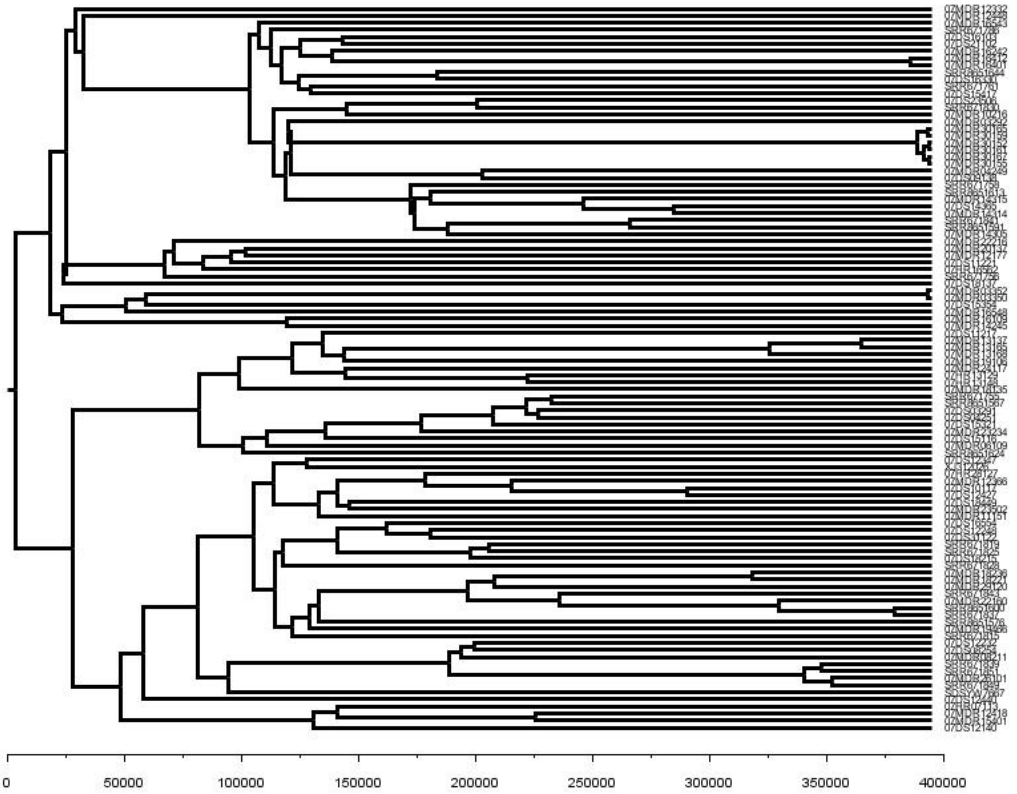



Figure 15: Phylogenetic tree for clade L2.3.5

**Clade L2.3.6:** A RL was selected as the optimal clock model to explain the Clade L2.3.6 dataset with the initial rate of $4.612(2.644, 6.550) \times 10^{-8}$ substitutions per site per year and the corresponding estimated tMRCA of 977,600 (574,510 to 1,497,200) years as shown in table (4). Hence, a reconstructed phylogenetic tree for the clade L2.3.6 is presented in figure (16) shows different evolutionary rates for at least five regions(rate by color) across the phylogeny.

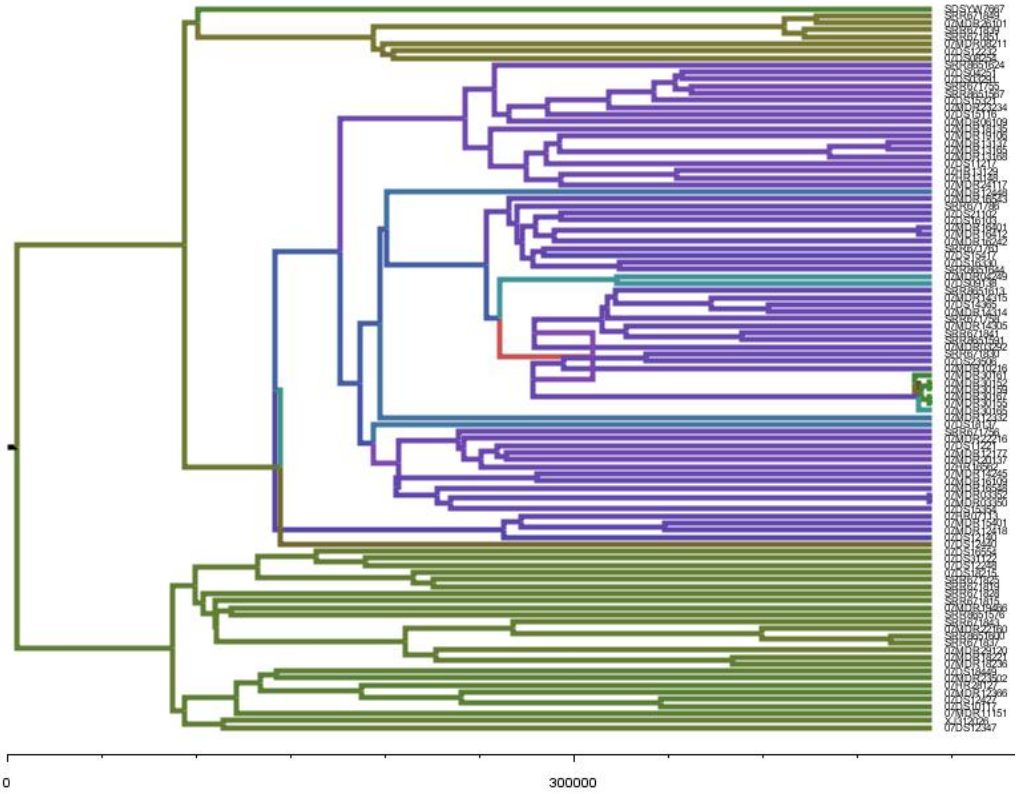



Figure 16: Phylogenetic tree for clade L2.3.6

# 5 Discussion

The objective of the study was to evaluate the performance of the different clock models applied in the MTB dataset. A Bayesian approach has been employed to fit the particular models of interest, such as strict clock, random local clock, and uncorrelated relaxed clock models.

Firstly, before model fitting, an investigation was done to explore the clock-likeness behavior on the MTB dataset by using a root-to-tip regression analysis, whereby it was found that there was weak clock-like behavior in the dataset; further analyses were needed to be done to attain significant results. Hence, the three clock models were fitted, evaluated, and compared based on their performances in the particular dataset.

Secondly, to answer a primary research question, the three clock models of interest were fitted based on their different assumptions, and they were then evaluated separately per each clade dataset with a Bayesian approach in BEAST software. The results indicate that a Strict clock model might be suitable to fit all clades datasets, while the random local clock Model might be suitable to fit clades L2.3.1, L2.3.2, L2.3.3, and L2.3.6, but not good fit for clades L2.3.4 and L2.3.5, where some parameters had ESS between 100 and 200. Lastly, Uncorrelated might be suitable for all other clades except for clade L2.3.3, where some parameters had ESS between 100 and 200. However, the trace plots of all the parameters for all the clocks in each clade were very well converged and mixed even after multiple adjustments on the MCMC chain iterations.

Thirdly, to answer a secondary research question, It was required to find optimal clock models and to explain how they vary. This question was answered by comparing different models fitted in every clade separately based on the Bayes factor, which was computed based on the ln(MLE) of the clock models using a generalized stepping stone sampling approach in BEAST. The attained results revealed that for the clade L2.3.1, L2.3.2, L2.3.4, and L2.3.5; a Strict Clock and Random Local Clock models equally fit the data well as compared to the Uncorrelated Relaxed Clock Model. However, a Strict clock was selected over a Random local clock for these clades, as a result of model simplicity based on the parsimony principle for statistical models (Coelho et al., 2019). Hence, a simpler Strict clock model with one parameter was considered to explain the evolutionary rate in the mentioned clades. This clock model suggests an estimated overall evolutionary rate that range approximately between $(2.588, 6.556) \times 10^{-8}$ substitutions per site per year for the clade L2.3.1, L2.3.2, L2.3.4, and L2.3.5. A Relaxed local clock model that fit the data similarly well as a strict clock was not considered due to their complexities due to having many parameters to be estimated. Consequences of fitting relaxed models may sometimes lead to overfitting and model misspecification, slow analysis, as it was faced in fitting random local for clade L2.3.4 and L2.3.5 in this study, despite of its advantage of taking into account the heterogeneity evolutionary rate. Furthermore, for the clade L2.3.3 and L2.3.6, the Random clock model was found to be an optimal clock model to explain the evolutionary rate of the dataset with an initial rate that ranges between $(2.644, 8.659) \times 10^{-8}$ substitutions per site per year.
However, the findings concerning evolutionary rates estimation, show that both strict clock

and random local give, on average, the same evolutionary rate but different in how they vary across the tree. This might suggest relative uniformity in this dataset's evolutionary rate, which could be a result of using a strict prior on substitutional rate or other evolutionary factors across various branches of the evolutionary tree or due to a lack of calibration points as a results of an uninformative dataset.

Additionally, In assessing the age of the most recent common ancestor (tMRCA), based on the fact that all the clades were assumed to have emerged at a relatively similar time. The finding of the optimal models revealed that clade L2.3.1 to L2.3.3 were found to have quite a high age of the most recent common ancestor in a range of approximately [950,000 to 1,900,000] years. In contrast, clades L2.3.4 to L.2.3.6 were found to have quite similar small ages of the most recent common ancestor range between approximately (400,000 to 660,000) years. This indicates that the clades L2.3.1 to L2.3.3 were early diverged as compared to the clades L2.3.4 to L.2.3.6. These discoveries give an insight that these two groups with some differences in divergence timeline, might be due to environmental, biological differences, or other factors which might influence the evolutionary process. As for the clade L2.3.1 to L2.3.3 found to undergo evolutionary process earlier as compared to clade L2.3.4 to L.2.3.6. Consequently, these age findings interms of divergence mechanism were found to coincide relatively with that of the previous study done by Zhu et al. (2023), but differ in age units due to metrics used differences.

# 6 Possible Drawbacks of the used Methods

**Computational Intensity**

- Bayesian approaches that employ molecular clock models are computationally intensive and might require substantial resources when dealing with large datasets to run the analysis.

**Analysis complexity**

- For the Relaxed clock models; as a result of additional parameters due to variation in evolutionary rate, increasing model complexity.
- Lack of calibration points, as a results of uninformative dataset.

# 7 Ethics, Relevance, and Stakeholders

## 7.1 Ethical thinking

**Data Transparency**

- In this study, ethical thinking guided our approach to data handling. We prioritize transparency by clearly showing how this publicly available data was obtained and used in this study.

## 7.2 Societal relevance

- Understanding the application of molecular clock in MTB is crucial for tracking the evolution of strains and predicting disease spreading. Hence, contributing to the development of more effective public health strategies.

## 7.3 Stakeholder awareness

**Researchers and academia**

- Awareness of the study's methodologies, ethical considerations, and identified limitations will contribute to informed decision-making and guide future research directions in the field of TB research.

**Healthcare practitioners**

- The study's findings could significantly impact our comprehension of the pathogen's evolution, particularly MTB, which may influence strategies for disease control, treatment, and prevention.
- The relevance of the dataset in public health and its potential implications for advancing molecular epidemiology.

# 8 Conclusion

In conclusion, this study shows that both Strict and Random local clock models fit well for the MTB dataset, leading to relatively uniform evolutionary rates across the trees. This might result from using a strict prior on substitutional rates or other evolutionary factors, such as a lack of calibration points due to an uninformative dataset. The tMRCA difference might be due to potential uninformed factor variations that align with the data set. Hence, there is a need for further investigation concerning other factors that might influence the underlying dataset's evolutionary dynamics, possibly including the calibration information. Other researchers can utilize this information to assess and investigate an underlying divergence process and adaptation factors within the studied groups. Also, in the case of different datasets, a preliminary check should be done on specific datasets, ensuring relative uniformity in the evolutionary mechanism. In that case, a strict clock can be recommended to simplify the analysis and reduce computational costs.

# 9 Ideas for Future Research

- Future studies may consider using another approach to test for the clock-like behavior structure based on hypothesis testing for the significant results for the clock model to be applied since this study employed a Root-to-Tip regression analysis, which is only suitable for data exploration and not for significance testing about the hypothesis of constant evolutionary rate.
- Future studies can also extend this study to transmission analysis. By examining the transmissions across the phylogeny using the R package TransPhylo.

## References


Al-Mutairi, N. M., Ahmad, S. and Mokaddas, E. M. (2019), 'Molecular characterization of multidrug-resistant mycobacterium tuberculosis (mdr-tb) isolates identifies local transmission of infection in kuwait, a country with a low incidence of tb and mdr-tb', *European Journal of Medical Research* **24**, 1–13.

Ali, R. H., Bark, M., Miró, J., Muhammad, S. A., Sjöstrand, J., Zubair, S. M., Abbas, R. M. and Arvestad, L. (2017), 'Vmcmc: a graphical and statistical analysis tool for markov chain monte carlo traces', *BMC bioinformatics* **18**, 1–8.

Anisimova, M. (2019), *Evolutionary genomics: statistical and computational methods*, Humana.

Arenas, M. (2015), 'Trends in substitution models of molecular evolution', *Frontiers in genetics* **6**, 163122.

Arnold, C. (2007), 'Molecular evolution of mycobacterium tuberculosis', *Clinical microbiology and infection* **13**(2), 120–128.

Baele, G., Lemey, P. and Suchard, M. A. (2016), 'Genealogical working distributions for bayesian model testing with phylogenetic uncertainty', *Systematic Biology* **65**(2), 250–264.

Bielejec, F., Lemey, P., Baele, G., Rambaut, A. and Suchard, M. A. (2014), 'Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography', *Systematic biology* **63**(4), 493–504.

Blumenthal, R. M. (1957), 'An extended markov property', *Transactions of the American Mathematical Society* **85**(1), 52–72.

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S. A., Bryant, J. M., Harris, S. R., Schuenemann, V. J. et al. (2014), 'Pre-columbian mycobacterial genomes reveal seals as a source of new world human tuberculosis', *Nature* **514**(7523), 494–497.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A. and Drummond, A. J. (2014), 'Beast 2: a software platform for bayesian evolutionary analysis', *PLoS computational biology* **10**(4), e1003537.

Brown, R. P. and Yang, Z. (2011), 'Rate variation and estimation of divergence times using strict and relaxed clocks', *BMC Evolutionary Biology* **11**, 1–12.

Cannarozzi, G. M. and Schneider, A. (2012), *Codon evolution: mechanisms and models*, Oxford University Press.

Coelho, M. T. P., Diniz-Filho, J. A. and Rangel, T. F. (2019), 'A parsimonious view of the parsimony principle in ecology and evolution', *Ecography* **42**(5), 968–976.

De Bruyn, A., Martin, D. P. and Lefeuvre, P. (2014), 'Phylogenetic reconstruction methods: an overview', *Molecular Plant Taxonomy: Methods and Protocols* pp. 257–277.

Didelot, X., Siveroni, I. and Volz, E. M. (2021), 'Additive uncorrelated relaxed clock models for the dating of genomic epidemiology phylogenies', *Molecular Biology and Evolution* **38**(1), 307–317.

Drummond, A. J. and Bouckaert, R. R. (2015), *Bayesian evolutionary analysis with BEAST*, Cambridge University Press.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J. and Rambaut, A. (2006), 'Relaxed phylogenetics and dating with confidence', *PLoS biology* **4**(5), e88.

Drummond, A. J. and Rambaut, A. (2007), 'Beast: Bayesian evolutionary analysis by sampling trees', *BMC evolutionary biology* **7**(1), 1–8.

Drummond, A. J. and Suchard, M. A. (2010), 'Bayesian random local clocks, or one rate to rule them all', *BMC biology* **8**(1), 1–12.

Drummond, A. J., Suchard, M. A., Xie, D. and Rambaut, A. (2012), 'Bayesian phylogenetics with beauti and the beast 1.7', *Molecular biology and evolution* **29**(8), 1969–1973.

Duchêne, D. A., Duchêne, S., Holmes, E. C. and Ho, S. Y. (2015), 'Evaluating the adequacy of molecular clock models using posterior predictive simulations', *Molecular Biology and Evolution* **32**(11), 2986–2995.

Fan, Y., Wu, R., Chen, M.-H., Kuo, L. and Lewis, P. O. (2011), 'Choosing among partition models in bayesian phylogenetics', *Molecular biology and evolution* **28**(1), 523–532.

Felsenstein, J. (1981), 'Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates', *Evolution* pp. 1229–1242.

Felsenstein, J. (2004), Inferring phylogenies, *in* 'Inferring phylogenies', pp. 664–664.

Grange, J. M. (2009), 'The genus mycobacterium and the mycobacterium tuberculosis complex', *Tuberculosis: a comprehensive clinical reference* **5**, 44–59.

Gregory, T. R. (2008), 'Understanding evolutionary trees', *Evolution: Education and Outreach* **1**(2), 121–137.

Grewal, J. K., Krzywinski, M. and Altman, N. (2019), 'Markov models—markov chains', *Nat. Methods* **16**(8), 663–664.

Hasegawa, M., Kishino, H. and Yano, T.-a. (1985), 'Dating of the human-ape splitting by a molecular clock of mitochondrial dna', *Journal of molecular evolution* **22**, 160–174.

Hastings, W. K. (1970), 'Monte carlo sampling methods using markov chains and their applications'.

Ho, S. Y. and Duchêne, S. (2014), 'Molecular-clock methods for estimating evolutionary rates and timescales', *Molecular ecology* **23**(24), 5947–5965.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P. and Ronquist, F. (2016), 'Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language', *Systematic biology* **65**(4), 726–736.

Huelsenbeck, J. P. and Ronquist, F. (2001), 'Mrbayes: Bayesian inference of phylogenetic trees', *Bioinformatics* **17**(8), 754–755.

Jeffreys, H. (1935), Some tests of significance, treated by the theory of probability, *in* 'Mathematical proceedings of the Cambridge philosophical society', Vol. 31, Cambridge University Press, pp. 203–222.

Kass, R. E. and Raftery, A. E. (1995), 'Bayes factors', *Journal of the american statistical association* **90**(430), 773–795.

Kimura, M. (1980), 'A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences', *Journal of molecular evolution* **16**, 111–120.

King, J. L. and Jukes, T. H. (1969), 'Non-darwinian evolution: Most evolutionary change in proteins may be due to neutral mutations and genetic drift.', *Science* **164**(3881), 788–798.

King, L. and Wakeley, J. (2016), 'Empirical bayes estimation of coalescence times from nucleotide sequence data', *Genetics* **204**(1), 249–257.

Kishino, H. and Hasegawa, M. (1989), 'Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea', *Journal of molecular evolution* **29**, 170–179.

Kishino, H. and Hasegawa, M. (1990), '[34] converting distance to time: Application to human evolution'.

Koch, K.-R. and Koch, K.-R. (1990), 'Bayes' theorem', *Bayesian Inference with Geodetic Applications* pp. 4–8.

Kumar, S. (2005), 'Molecular clocks: four decades of evolution', *Nature Reviews Genetics* **6**(8), 654–662.

Lartillot, N. and Philippe, H. (2006), 'Computing bayes factors using thermodynamic integration', *Systematic biology* **55**(2), 195–207.

Lee, M. S. (2020), 'Clock models for evolution of discrete phenotypic characters', *The molecular evolutionary clock: theory and practice* pp. 101–113.

Lepage, T., Bryant, D., Philippe, H. and Lartillot, N. (2007), 'A general comparison of relaxed molecular clock models', *Molecular biology and evolution* **24**(12), 2669–2680.

Li, S. S. (1996), *Phylogenetic tree construction using Markov chain Monte Carlo*, The Ohio State University.

Li, W. L. (2010), Rates of Molecular Evolution and Phylogenomic Inference, PhD thesis, ResearchSpace@ Auckland.

Liò, P. and Goldman, N. (1998), 'Models of molecular evolution and phylogeny', *Genome research* **8**(12), 1233–1244.

Loiseau, C. (2020), Evolutionary Epidemiology of the Mycobacterium tuberculosis Complex, PhD thesis, University_of_Basel_Associated_Institution.

Mbugi, E. V., Katale, B. Z., Streicher, E. M., Keyyu, J. D., Kendall, S. L., Dockrell, H. M., Michel, A. L., Rweyemamu, M. M., Warren, R. M., Matee, M. I. et al. (2016), 'Mapping of mycobacterium tuberculosis complex genetic diversity profiles in tanzania and other african countries', *PloS one* **11**(5), e0154571.

Membrebe, J. V., Lemey, P. and Baele, G. (2022), 'Bayesian phylogenetic model development for viral evolutionary reconstruction'.

Membrebe, J. V., Suchard, M. A., Rambaut, A., Baele, G. and Lemey, P. (2019), 'Bayesian inference of evolutionary histories under time-dependent substitution rates', *Molecular biology and evolution* **36**(8), 1793–1803.

Menardo, F., Duchêne, S., Brites, D. and Gagneux, S. (2019), 'The molecular clock of mycobacterium tuberculosis', *PLoS pathogens* **15**(9), e1008067.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The journal of chemical physics* **21**(6), 1087–1092.

Morrison, D. A. (2010), 'Using data-display networks for exploratory data analysis in phylogenetic studies', *Molecular Biology and Evolution* **27**(5), 1044–1057.

Nei, M. and Kumar, S. (2000), *Molecular evolution and phylogenetics*, Oxford University Press, USA.

Nguyen, D., Brassard, P., Menzies, D., Thibert, L., Warren, R., Mostowy, S. and Behr, M. (2004), 'Genomic characterization of an endemic mycobacterium tuberculosis strain: evolutionary and epidemiologic implications', *Journal of clinical microbiology* **42**(6), 2573–2580.

Nguyen, L., Schmidt, H. A., Von Haeseler, A. and Minh, B. Q. (2015), 'Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Molecular biology and evolution* **32**(1), 268–274.

Nylander, J. A., Wilgenbusch, J. C., Warren, D. L. and Swofford, D. L. (2008), 'Awty (are we there yet?): a system for graphical exploration of mcmc convergence in bayesian phylogenetics', *Bioinformatics* **24**(4), 581–583.

Oaks, J., Cobb, K. and Minin, V. (2018), 'Marginal likelihoods in phylogenetics: a review of methods and'.

Obi, R., Orji, N., Nwanebu, F., Okangba, C. and Ndubuisi, U. (2010), 'Emerging and re-emerging infectious diseases: the perpetual menace', *Asian J qf Eperimental Biol Sci* **2**(1), 271–282.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G. and Suchard, M. A. (2018), 'Posterior summarization in bayesian phylogenetics using tracer 1.7', *Systematic biology* **67**(5), 901–904.

Rambaut, A., Lam, T. T., Max Carvalho, L. and Pybus, O. G. (2016), 'Exploring the temporal structure of heterochronous sequences using tempest (formerly path-o-gen)', *Virus evolution* **2**(1), vew007.

Rannala, B. and Yang, Z. (1996), 'Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference', *Journal of molecular evolution* **43**, 304–311.

Rannala, B. and Yang, Z. (2007), 'Inferring speciation times under an episodic molecular clock', *Systematic biology* **56**(3), 453–466.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. and Huelsenbeck, J. P. (2012), 'Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space', *Systematic biology* **61**(3), 539–542.

Sanderson, M. J. (1995), 'Objections to bootstrapping phylogenies: a critique', *Systematic biology* **44**(3), 299–320.

Sobkowiak, B., Romanowski, K., Sekirov, I., Gardy, J. L. and Johnston, J. C. (2023), 'Comparing mycobacterium tuberculosis transmission reconstruction models from whole genome sequence data', *Epidemiology & Infection* **151**, e105.

Spall, J. C. (2003), 'Estimation via markov chain monte carlo', *IEEE Control Systems Magazine* **23**(2), 34–45.

Steel, M. (2016), *Phylogeny: discrete and random processes in evolution*, SIAM.

Tavare, N. (1986), 'Mixing in continuous crystallizers', *AIChE journal* **32**(5), 705–732.

Tay, J. H., Baele, G. and Duchene, S. (2023), 'Detecting episodic evolution through bayesian inference of molecular clock models', *bioRxiv* pp. 2023–06.

Thiébaux, H. J. and Zwiers, F. W. (1984), 'The interpretation and estimation of effective sample size', *Journal of Applied Meteorology and Climatology* **23**(5), 800–811.

Warnock, R. C., Yang, Z. and Donoghue, P. C. (2017), 'Testing the molecular clock using mechanistic models of fossil preservation and molecular evolution', *Proceedings of the Royal Society B: Biological Sciences* **284**(1857), 20170227.

WHO (2023), 'Who standard: universal access to rapid tuberculosis diagnostics'.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L. and Chen, M.-H. (2011), 'Improving marginal likelihood estimation for bayesian phylogenetic model selection', *Systematic biology* **60**(2), 150–160.

Yang, Z. (1994), 'Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods', *Journal of Molecular evolution* **39**, 306–314.

Zharkikh, A. (1994), 'Estimation of evolutionary distances between nucleotide sequences', *Journal of molecular evolution* **39**, 315–329.

Zhu, C., Yang, T., Yin, J., Jiang, H., Takiff, H. E., Gao, Q., Liu, Q. and Li, W. (2023), 'The global success of mycobacterium tuberculosis modern beijing family is driven by a few recently emerged strains', *Microbiology Spectrum* pp. e03339–22.

Zuckerkandl, E. (1962), 'Molecular disease, evolution, and genic heterogeneity', *Horizons in biochemistry* pp. 189–225.

Zuckerkandl, E. and Pauling, L. (1965), Evolutionary divergence and convergence in proteins, *in* 'Evolving genes and proteins', Elsevier, pp. 97–166.

# 10 Appendix

## 10.1 Overall estimates for the Molecular clock models

Table 5: Molecular Clock Models Estimates for the MTB dataset.

| Clade | Clock | log(MLE) | tMRCA | | Pars | mean | 95%HPD |
|---|---|---|---|---|---|---|---|
| | | | Mean | 95%HPD | | | |
| | **STR** | -14981.998 | 1.1916E6 | [7.2046E5,1.982E6] | **O.R** | 4.627E-8 | [2.732E-8, 6.614E-8] |
| | **RL** | -14982.571 | 1.266E6 | [7.205E5,1.982E6] | **I.R** | 4.612E-8 | [2.658E-8, 6.560E-8] |
| L2.3.1 | | | | | **R.C** | 0.807 | [0, 2] |
| | **UCL** | -14998.742 | 1.154E6 | [7.176E5,1.708E6] | **Mean** | 4.81E-8 | [2.875E-8, 6.650E-8] |
| | | | | | $\sigma$ | 4.982E-9 | [1.014E-10, 9.691E-9] |
| | **STR** | -8962.809 | 1.878E6 | [1.126E6,2.854E6] | **O.R** | 4.615E-8 | [2.6465E-8, 6.567E-8] |
| | **RL** | -8963.363 | 1.904E6 | [1.097E6, 2.955E6] | **I.R** | 4.616E-8 | [2.697E-8, 6.603E-8] |
| L2.3.2 | | | | | **R.C** | 0.235 | [0, 1] |
| | **UCL** | -8980.160 | 1.773E6 | [1.091E6, 2.638E6] | **Mean** | 4.812E-8 | [2.963E-8, 6.765E-8] |
| | | | | | $\sigma$ | 4.445E-9 | [1.789E-12, 9.974E-9] |
| | **STR** | -18101.161 | 8.794E5 | [5.404E5, 1.352E6] | **O.R** | 4.622E-8 | [2.695E-8, 6.623E-8] |
| | **RL** | -18052.215 | 9.776E5 | [5.7451E5,1.4972E6] | **I.R** | 4.612E-8 | [2.644E-8, 6.550E-8] |
| L2.3.3 | | | | | **R.C** | 1.194 | [1,2] |
| | **UCL** | -18085.681 | 1.6222E8 | [5.1201E5, 6.7463E8] | **Mean** | 4.829E-8 | [2.947E-8, 6.761E-8] |
| | | | | | $\sigma$ | 3.3389E-6 | [8.2763E-9, 1.4113E-5] |
| | **STR** | -37496.966 | 4.867E5 | [2.8678E5, 7.220E5] | **O.R** | 4.629E-8 | [2.748E-8, 6.694E-8] |
| | **RL** | -37496.226 | 9.386E5 | [4.9309E5,1.5093E6] | **I.R** | 4.615E-8 | [2.647E-8, 6.541E-8] |
| L2.3.4 | | | | | **R.C** | 6.893 | [4, 9] |
| | **UCL** | -37502.853 | 4.706E5 | [3.023E5, 6.913E5] | **Mean** | 4.812E-8 | [2.972E-8, 6.755E-8] |
| | | | | | $\sigma$ | 7.626E-9 | [3.838E-9, 1.184E-8] |
| | **STR** | -49698.391 | 4.113E5 | [2.469E5, 6.023E5] | **O.R** | 4.616E-8 | [2.682E-8, 6.556E-8] |
| | **RL** | -49699.369 | 4.743E5 | [2.772E5, 7.192E5] | **I.R** | 4.627E-8 | [2.672E-8, 6.560E-8] |
| L2.3.5 | | | | | **R.C** | 1.981 | [1, 4] |
| | **UCL** | -49701.176 | 4.099E5 | [2.569E5, 6.065E5] | **Mean** | 4.789E-8 | [2.881E-8, 6.690E-8] |
| | | | | | $\sigma$ | 6.85E-9 | [3.497E-9, 1.043E-8] |
| | **STR** | -37120.490 | 5.9977E5 | [3.479E5, 9.032E5] | **O.R** | 4.589E-8 | [2.588E-8, 6.594E-8] |
| | **RL** | -37114.284 | 6.575E5 | [3.4033E5, 1.1318E6] | **I.R** | 4.616E-8 | [2.685E-8, 6.597E-8] |
| L2.3.6 | | | | | **R.C** | 3.57 | [1, 6] |
| | **UCL** | -37118.986 | 5.828E5 | [3.569E5, 8.575E5] | **Mean** | 4.818E-8 | [2.922E-8, 6.716E-8] |
| | | | | | $\sigma$ | 9.147E-9 | [4.67E-9, 1.392E-8] |

**Note: Pars**=Parameters, **STR**=Strict Clock, **RL**=Random Local Clock, **UCL**=Uncorrelated Lognormal Clock, **OR**=Overall Rate, **I.R**=Initial Rate, **R.C**=Rate change, $\sigma$ = Dispersion, **95% HPD**= Highest posterior density intervals

## 10.2 Trace plots for each clock per clade

Below are the trace plots for the clock models per each clade separately.

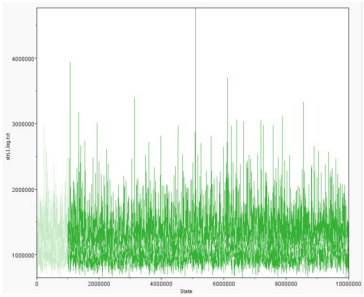


Figure 17: STRL1 traceplot

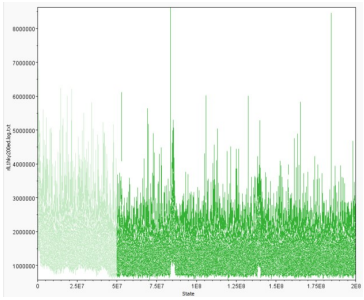


Figure 18: RLL1 traceplot

.

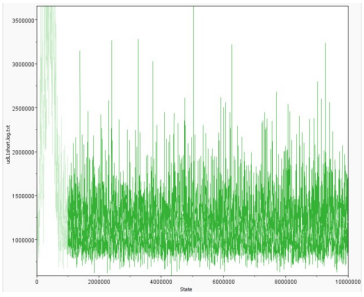


Figure 19: UCLL1 traceplot

.

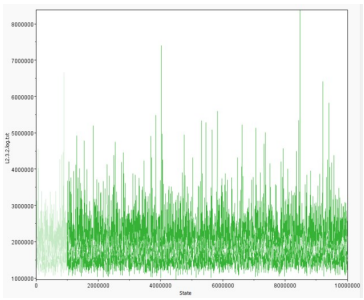


Figure 20: STRL2 traceplot

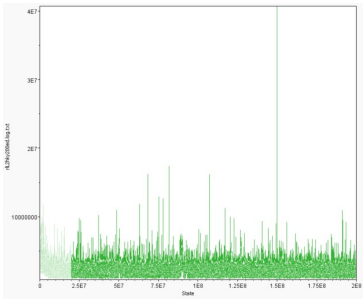


Figure 21: RLL2 traceplot

.

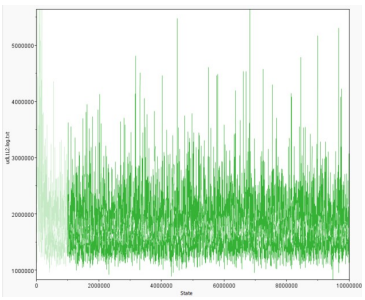


Figure 22: UCLL2 traceplot

.

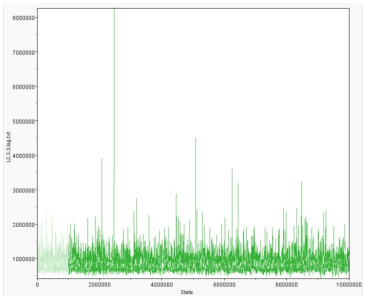


Figure 23: STRL3 traceplot

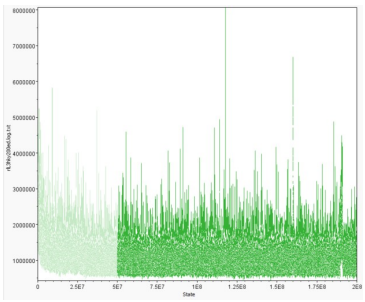


Figure 24: RLL3 traceplot

.

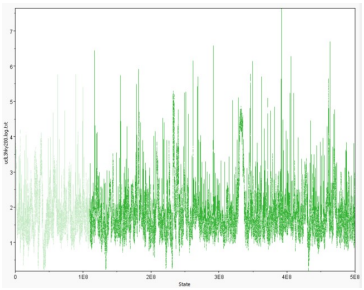


Figure 25: UCLL3 traceplot

.

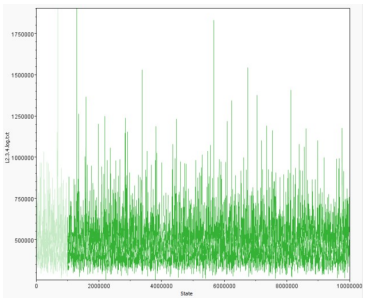


Figure 26: STRL4 traceplot

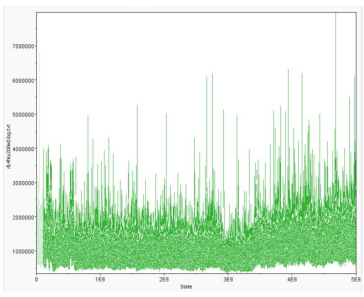


Figure 27: RLL4 traceplot

.

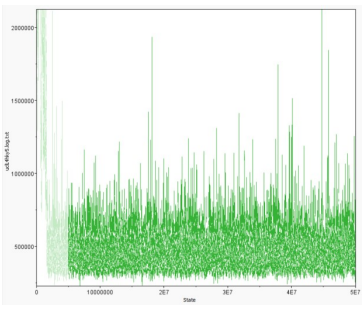


Figure 28: UCLL4 traceplot

.

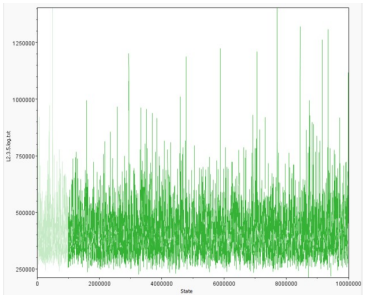


Figure 29: STRL5 traceplot

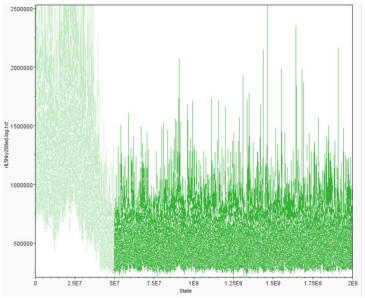


Figure 30: RLL5 traceplot

.

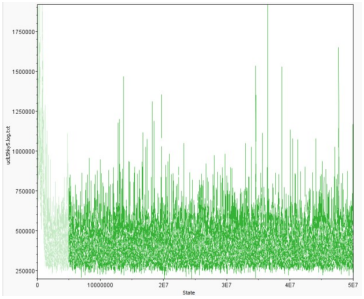


Figure 31: UCLL5 traceplot

.

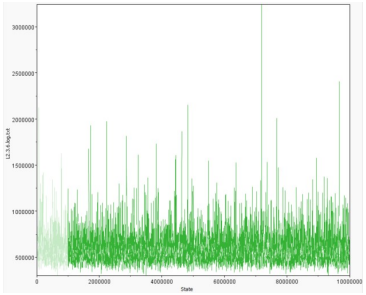


Figure 32: STRL6 traceplot

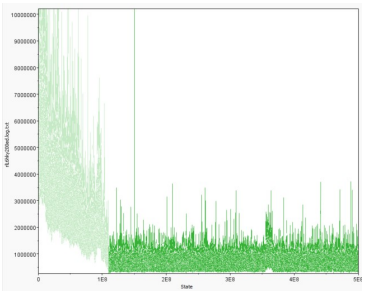


Figure 33: RLL6 traceplot

.

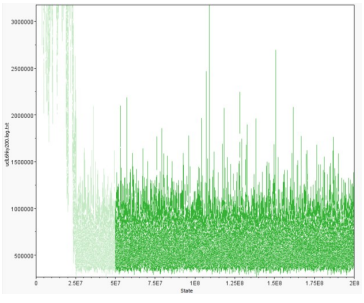
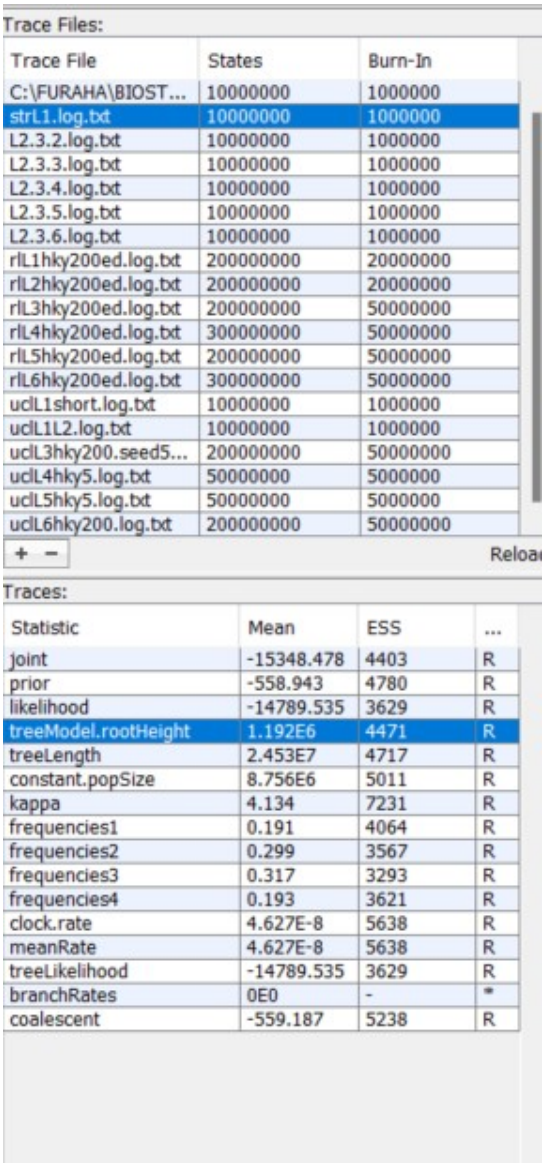


Figure 34: UCLL6 traceplot

.

## 10.3 Effective sample size

Below are the tables for effective sample size output in Tracer.



Trace Files:

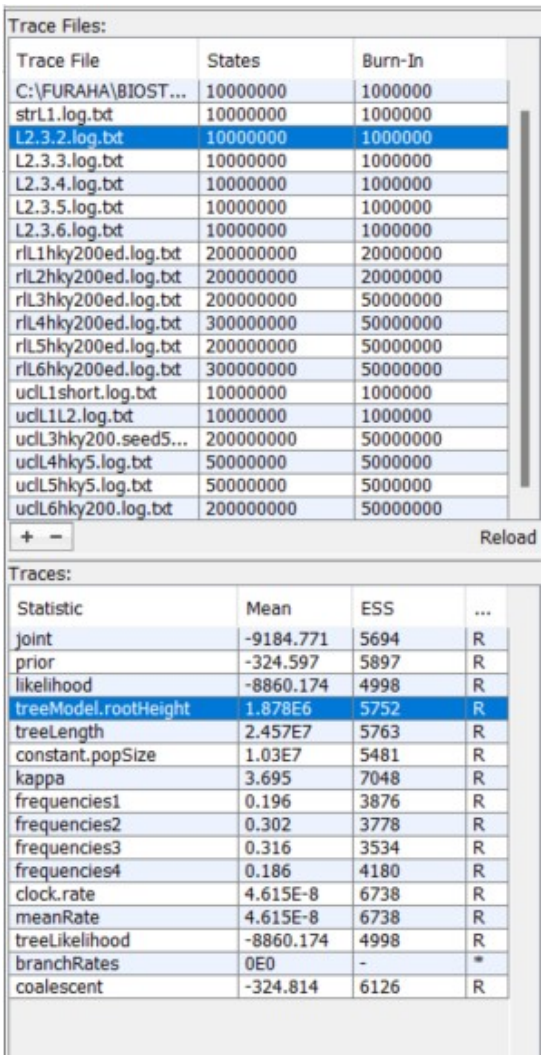
| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

+ −

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -15348.478 | 4403 | R |
| prior | -558.943 | 4780 | R |
| likelihood | -14789.535 | 3629 | R |
| treeModel.rootHeight | 1.192E6 | 4471 | R |
| treeLength | 2.453E7 | 4717 | R |
| constant.popSize | 8.756E6 | 5011 | R |
| kappa | 4.134 | 7231 | R |
| frequencies1 | 0.191 | 4064 | R |
| frequencies2 | 0.299 | 3567 | R |
| frequencies3 | 0.317 | 3293 | R |
| frequencies4 | 0.193 | 3621 | R |
| clock.rate | 4.627E-8 | 5638 | R |
| meanRate | 4.627E-8 | 5638 | R |
| treeLikelihood | -14789.535 | 3629 | R |
| branchRates | 0E0 | - | * |
| coalescent | -559.187 | 5238 | R |

Figure 35: Strict L2.3.1 ESS Table



Trace Files:

| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

+ −

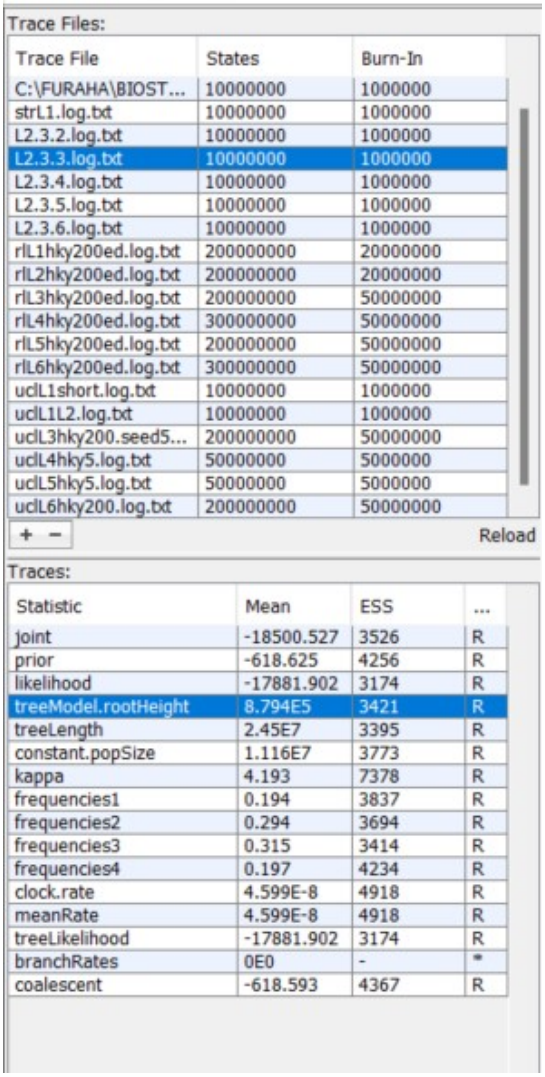
Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -9184.771 | 5694 | R |
| prior | -324.597 | 5897 | R |
| likelihood | -8860.174 | 4998 | R |
| treeModel.rootHeight | 1.878E6 | 5752 | R |
| treeLength | 2.457E7 | 5763 | R |
| constant.popSize | 1.03E7 | 5481 | R |
| kappa | 3.695 | 7048 | R |
| frequencies1 | 0.196 | 3876 | R |
| frequencies2 | 0.302 | 3778 | R |
| frequencies3 | 0.316 | 3534 | R |
| frequencies4 | 0.186 | 4180 | R |
| clock.rate | 4.615E-8 | 6738 | R |
| meanRate | 4.615E-8 | 6738 | R |
| treeLikelihood | -8860.174 | 4998 | R |
| branchRates | 0E0 | - | * |
| coalescent | -324.814 | 6126 | R |

Figure 36: Strict Clock L2.3.2 ESS Table

.



Trace Files:

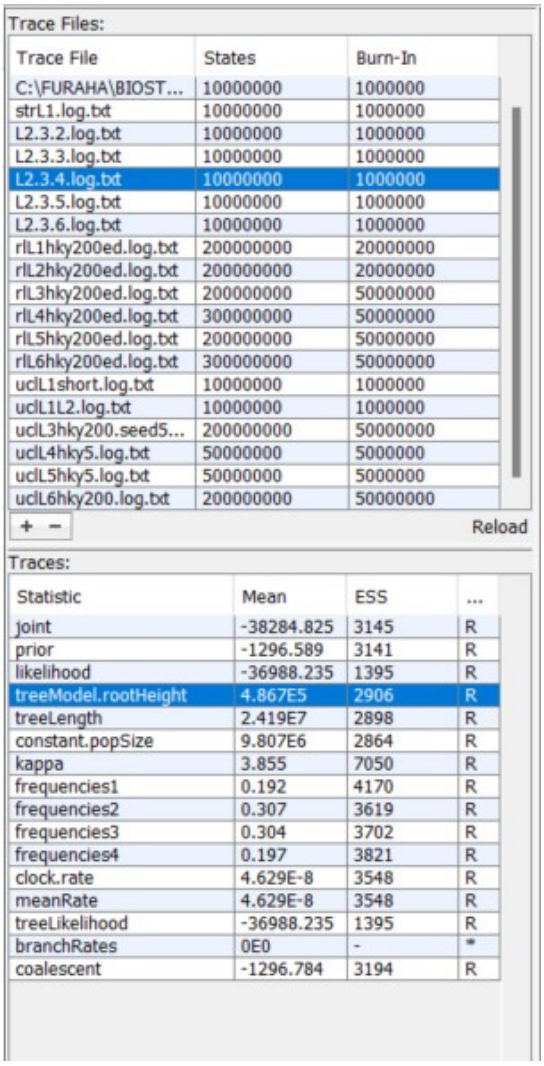
| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

+ −  Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -18500.527 | 3526 | R |
| prior | -618.625 | 4256 | R |
| likelihood | -17881.902 | 3174 | R |
| treeModel.rootHeight | 8.794E5 | 3421 | R |
| treeLength | 2.45E7 | 3395 | R |
| constant.popSize | 1.116E7 | 3773 | R |
| kappa | 4.193 | 7378 | R |
| frequencies1 | 0.194 | 3837 | R |
| frequencies2 | 0.294 | 3694 | R |
| frequencies3 | 0.315 | 3414 | R |
| frequencies4 | 0.197 | 4234 | R |
| clock.rate | 4.599E-8 | 4918 | R |
| meanRate | 4.599E-8 | 4918 | R |
| treeLikelihood | -17881.902 | 3174 | R |
| branchRates | 0E0 | - | * |
| coalescent | -618.593 | 4367 | R |

Figure 37: Strict L2.3.3 ESS Table



Trace Files:

| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

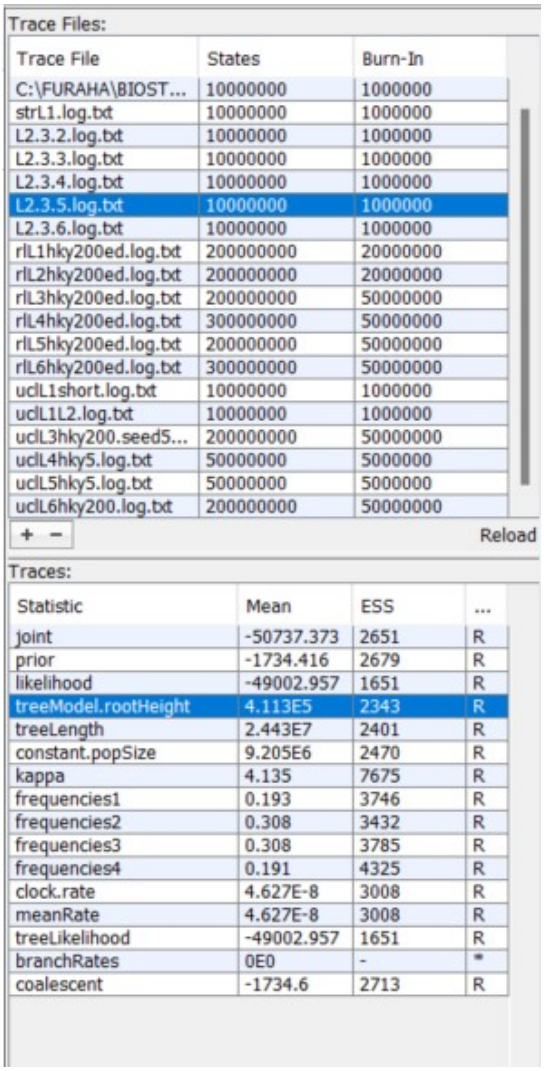
+ −  Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -38284.825 | 3145 | R |
| prior | -1296.589 | 3141 | R |
| likelihood | -36988.235 | 1395 | R |
| treeModel.rootHeight | 4.867E5 | 2906 | R |
| treeLength | 2.419E7 | 2898 | R |
| constant.popSize | 9.807E6 | 2864 | R |
| kappa | 3.855 | 7050 | R |
| frequencies1 | 0.192 | 4170 | R |
| frequencies2 | 0.307 | 3619 | R |
| frequencies3 | 0.304 | 3702 | R |
| frequencies4 | 0.197 | 3821 | R |
| clock.rate | 4.629E-8 | 3548 | R |
| meanRate | 4.629E-8 | 3548 | R |
| treeLikelihood | -36988.235 | 1395 | R |
| branchRates | 0E0 | - | * |
| coalescent | -1296.784 | 3194 | R |

Figure 38: Strict Clock L2.3.4 ESS Table

.



Trace Files:

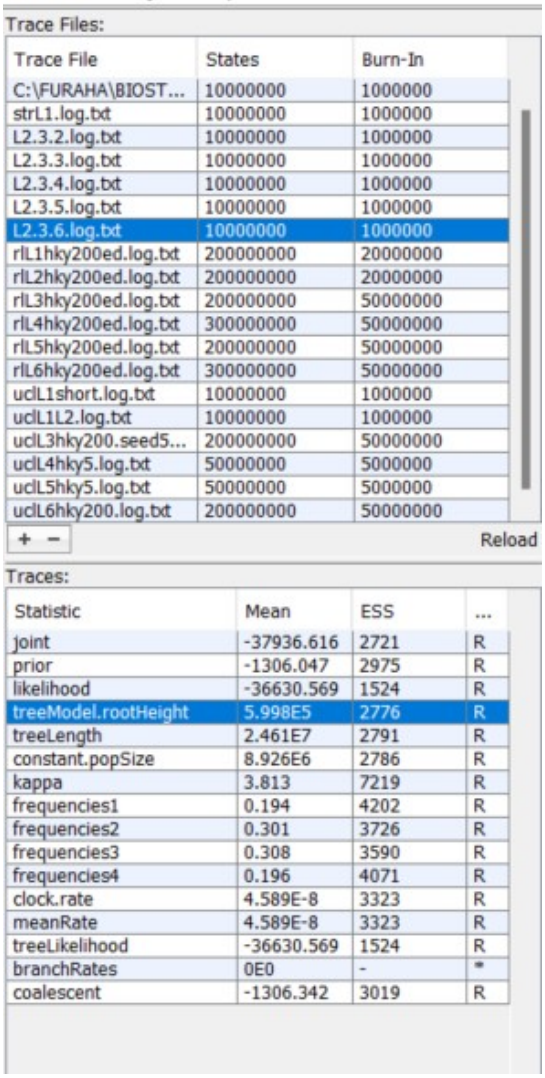
| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

+ −  Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -50737.373 | 2651 | R |
| prior | -1734.416 | 2679 | R |
| likelihood | -49002.957 | 1651 | R |
| treeModel.rootHeight | 4.113E5 | 2343 | R |
| treeLength | 2.443E7 | 2401 | R |
| constant.popSize | 9.205E6 | 2470 | R |
| kappa | 4.135 | 7675 | R |
| frequencies1 | 0.193 | 3746 | R |
| frequencies2 | 0.308 | 3432 | R |
| frequencies3 | 0.308 | 3785 | R |
| frequencies4 | 0.191 | 4325 | R |
| clock.rate | 4.627E-8 | 3008 | R |
| meanRate | 4.627E-8 | 3008 | R |
| treeLikelihood | -49002.957 | 1651 | R |
| branchRates | 0E0 | - | * |
| coalescent | -1734.6 | 2713 | R |

Figure 39: Strict L2.3.5 ESS Table



Trace Files:

| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

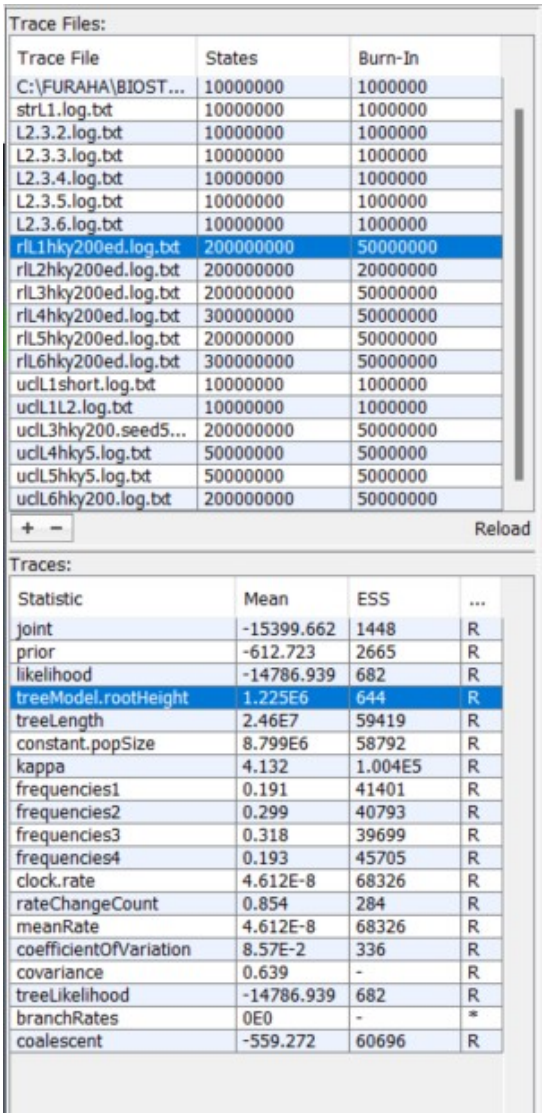
+ −  Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -37936.616 | 2721 | R |
| prior | -1306.047 | 2975 | R |
| likelihood | -36630.569 | 1524 | R |
| treeModel.rootHeight | 5.998E5 | 2776 | R |
| treeLength | 2.461E7 | 2791 | R |
| constant.popSize | 8.926E6 | 2786 | R |
| kappa | 3.813 | 7219 | R |
| frequencies1 | 0.194 | 4202 | R |
| frequencies2 | 0.301 | 3726 | R |
| frequencies3 | 0.308 | 3590 | R |
| frequencies4 | 0.196 | 4071 | R |
| clock.rate | 4.589E-8 | 3323 | R |
| meanRate | 4.589E-8 | 3323 | R |
| treeLikelihood | -36630.569 | 1524 | R |
| branchRates | 0E0 | - | * |
| coalescent | -1306.342 | 3019 | R |

Figure 40: Strict Clock L2.3.6 ESS Table

.



Trace Files:

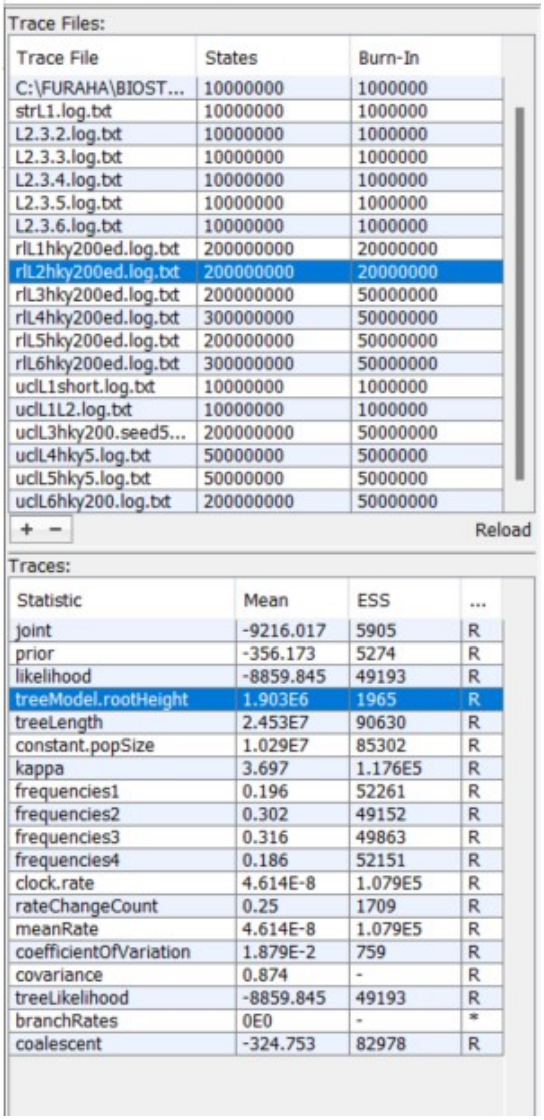
| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 50000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

+ −  Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -15399.662 | 1448 | R |
| prior | -612.723 | 2665 | R |
| likelihood | -14786.939 | 682 | R |
| treeModel.rootHeight | 1.225E6 | 644 | R |
| treeLength | 2.46E7 | 59419 | R |
| constant.popSize | 8.799E6 | 58792 | R |
| kappa | 4.132 | 1.004E5 | R |
| frequencies1 | 0.191 | 41401 | R |
| frequencies2 | 0.299 | 40793 | R |
| frequencies3 | 0.318 | 39699 | R |
| frequencies4 | 0.193 | 45705 | R |
| clock.rate | 4.612E-8 | 68326 | R |
| rateChangeCount | 0.854 | 284 | R |
| meanRate | 4.612E-8 | 68326 | R |
| coefficientOfVariation | 8.57E-2 | 336 | R |
| covariance | 0.639 | - | R |
| treeLikelihood | -14786.939 | 682 | R |
| branchRates | 0E0 | - | * |
| coalescent | -559.272 | 60696 | R |

Figure 41: RL L2.3.1 ESS Table



Trace Files:

| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

+ −  Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -9216.017 | 5905 | R |
| prior | -356.173 | 5274 | R |
| likelihood | -8859.845 | 49193 | R |
| treeModel.rootHeight | 1.903E6 | 1965 | R |
| treeLength | 2.453E7 | 90630 | R |
| constant.popSize | 1.029E7 | 85302 | R |
| kappa | 3.697 | 1.176E5 | R |
| frequencies1 | 0.196 | 52261 | R |
| frequencies2 | 0.302 | 49152 | R |
| frequencies3 | 0.316 | 49863 | R |
| frequencies4 | 0.186 | 52151 | R |
| clock.rate | 4.614E-8 | 1.079E5 | R |
| rateChangeCount | 0.25 | 1709 | R |
| meanRate | 4.614E-8 | 1.079E5 | R |
| coefficientOfVariation | 1.879E-2 | 759 | R |
| covariance | 0.874 | - | R |
| treeLikelihood | -8859.845 | 49193 | R |
| branchRates | 0E0 | - | * |
| coalescent | -324.753 | 82978 | R |

Figure 42: RL L2.3.2 ESS Table

.

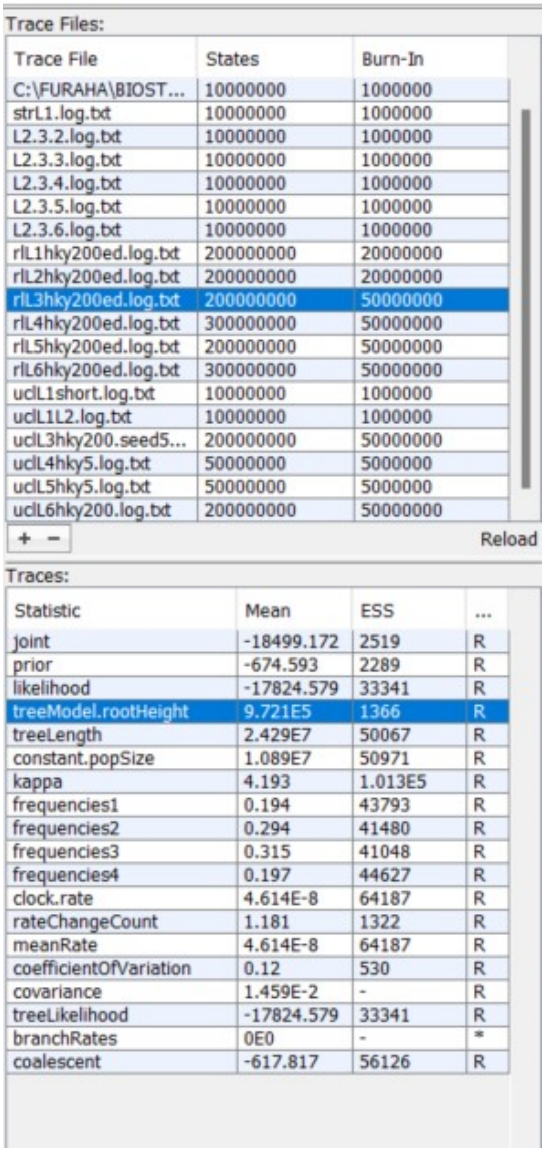



Figure 43: RL L2.3.3 ESS Table

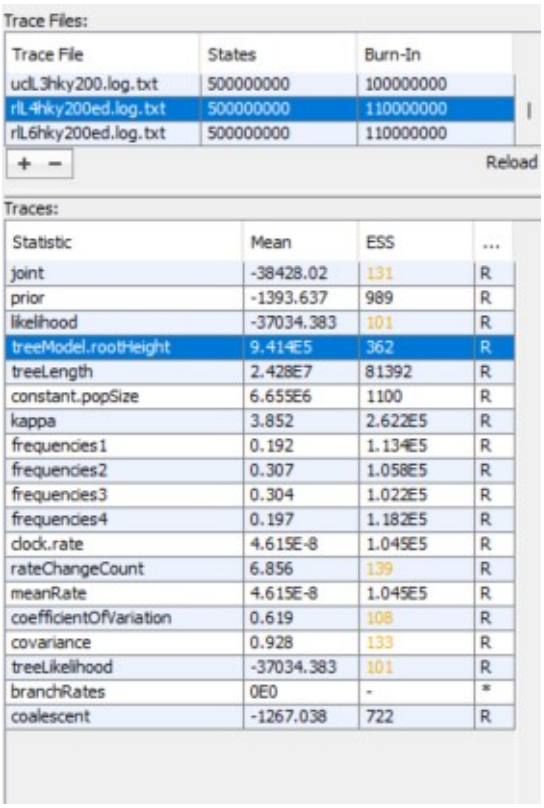
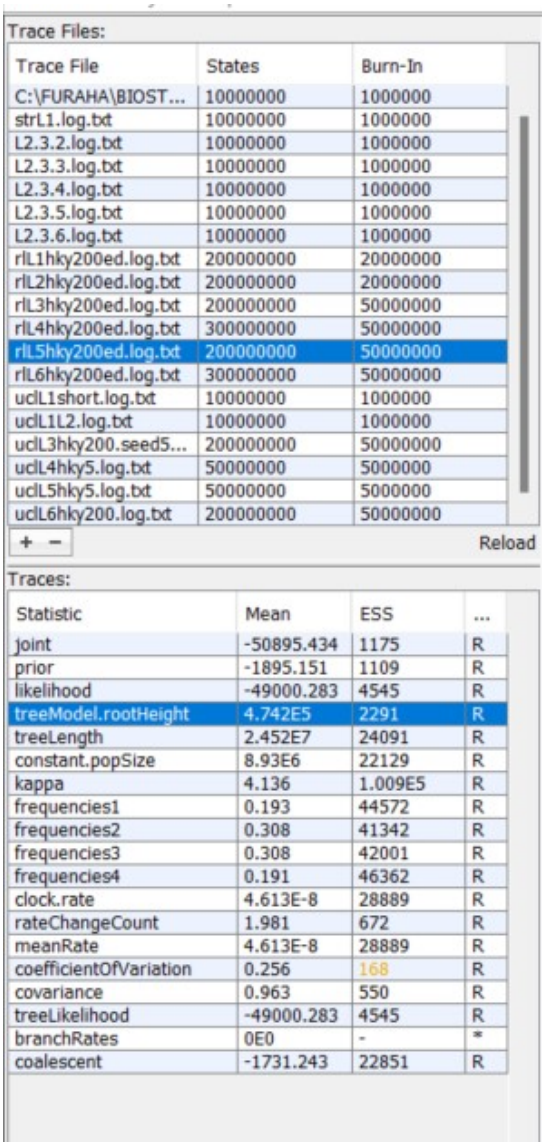



Figure 44: RL L2.3.4 ESS Table

.



Trace Files:

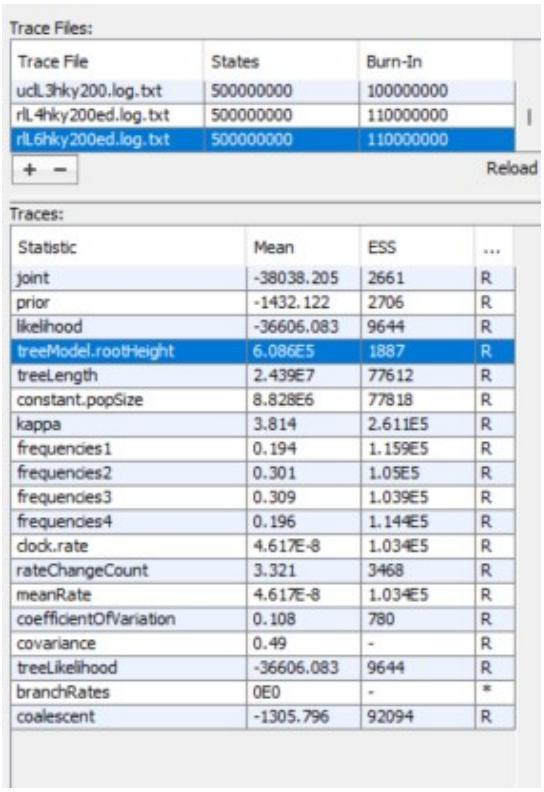
| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

+ −  Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -50895.434 | 1175 | R |
| prior | -1895.151 | 1109 | R |
| likelihood | -49000.283 | 4545 | R |
| treeModel.rootHeight | 4.742E5 | 2291 | R |
| treeLength | 2.452E7 | 24091 | R |
| constant.popSize | 8.93E6 | 22129 | R |
| kappa | 4.136 | 1.009E5 | R |
| frequencies1 | 0.193 | 44572 | R |
| frequencies2 | 0.308 | 41342 | R |
| frequencies3 | 0.308 | 42001 | R |
| frequencies4 | 0.191 | 46362 | R |
| clock.rate | 4.613E-8 | 28889 | R |
| rateChangeCount | 1.981 | 672 | R |
| meanRate | 4.613E-8 | 28889 | R |
| coefficientOfVariation | 0.256 | 168 | R |
| covariance | 0.963 | 550 | R |
| treeLikelihood | -49000.283 | 4545 | R |
| branchRates | 0E0 | - | * |
| coalescent | -1731.243 | 22851 | R |

Figure 45: RL L2.3.5 ESS Table



Trace Files:

| Trace File | States | Burn-In |
|---|---|---|
| uclL3hky200.log.txt | 500000000 | 100000000 |
| rlL4hky200ed.log.txt | 500000000 | 110000000 |
| rlL6hky200ed.log.txt | 500000000 | 110000000 |

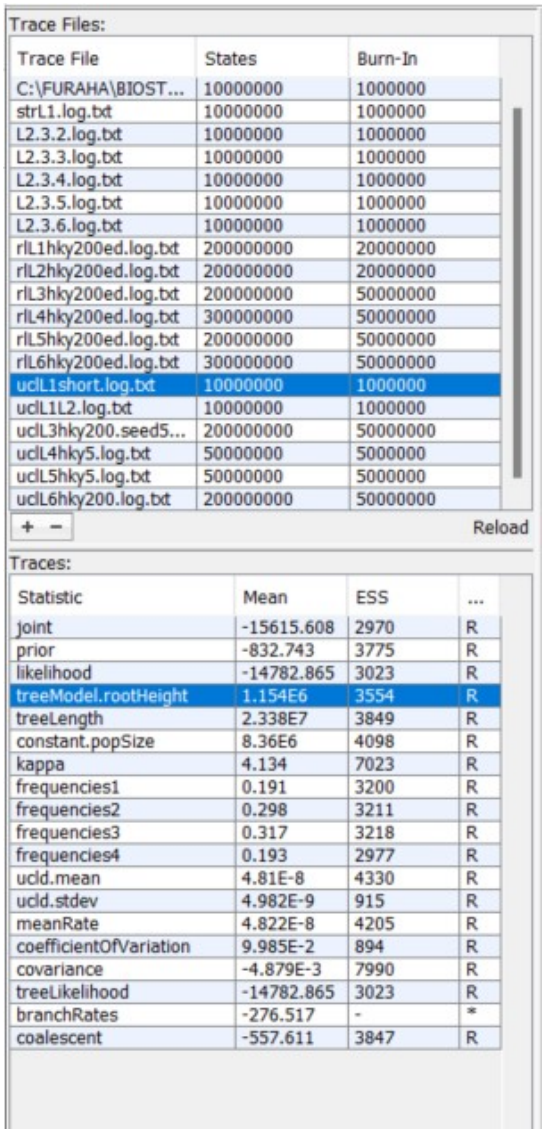
+ −  Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -38038.205 | 2661 | R |
| prior | -1432.122 | 2706 | R |
| likelihood | -36606.083 | 9644 | R |
| treeModel.rootHeight | 6.086E5 | 1887 | R |
| treeLength | 2.439E7 | 77612 | R |
| constant.popSize | 8.828E6 | 77818 | R |
| kappa | 3.814 | 2.611E5 | R |
| frequencies1 | 0.194 | 1.159E5 | R |
| frequencies2 | 0.301 | 1.05E5 | R |
| frequencies3 | 0.309 | 1.039E5 | R |
| frequencies4 | 0.196 | 1.144E5 | R |
| clock.rate | 4.617E-8 | 1.034E5 | R |
| rateChangeCount | 3.321 | 3468 | R |
| meanRate | 4.617E-8 | 1.034E5 | R |
| coefficientOfVariation | 0.108 | 780 | R |
| covariance | 0.49 | - | R |
| treeLikelihood | -36606.083 | 9644 | R |
| branchRates | 0E0 | - | * |
| coalescent | -1305.796 | 92094 | R |

Figure 46: RL L2.3.6 ESS Table

.



Trace Files:

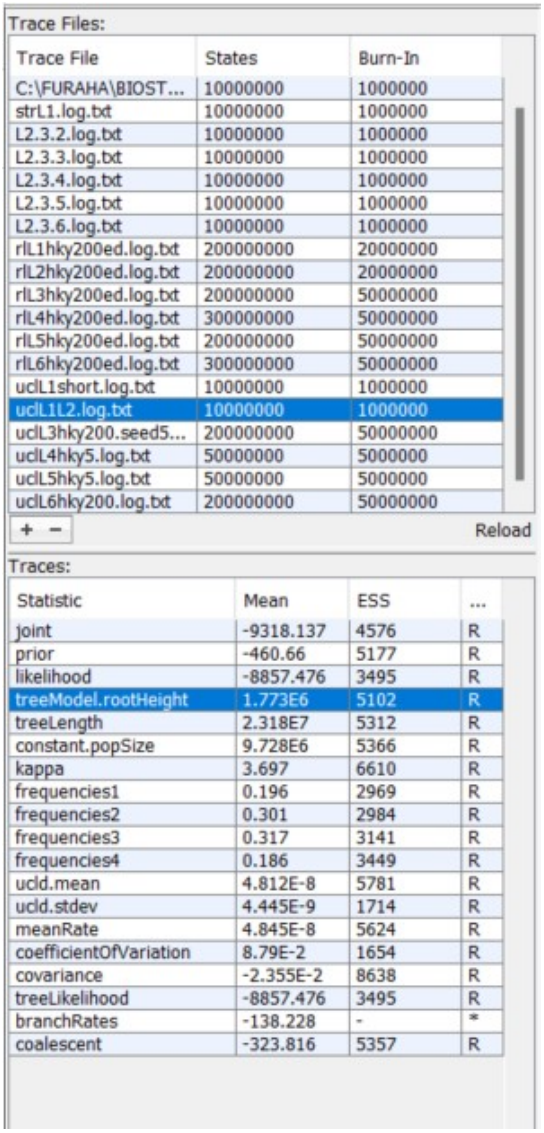
| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

+ −  Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -15615.608 | 2970 | R |
| prior | -832.743 | 3775 | R |
| likelihood | -14782.865 | 3023 | R |
| treeModel.rootHeight | 1.154E6 | 3554 | R |
| treeLength | 2.338E7 | 3849 | R |
| constant.popSize | 8.36E6 | 4098 | R |
| kappa | 4.134 | 7023 | R |
| frequencies1 | 0.191 | 3200 | R |
| frequencies2 | 0.298 | 3211 | R |
| frequencies3 | 0.317 | 3218 | R |
| frequencies4 | 0.193 | 2977 | R |
| ucld.mean | 4.81E-8 | 4330 | R |
| ucld.stdev | 4.982E-9 | 915 | R |
| meanRate | 4.822E-8 | 4205 | R |
| coefficientOfVariation | 9.985E-2 | 894 | R |
| covariance | -4.879E-3 | 7990 | R |
| treeLikelihood | -14782.865 | 3023 | R |
| branchRates | -276.517 | - | * |
| coalescent | -557.611 | 3847 | R |

Figure 47: UCL L2.3.1 ESS Table



Trace Files:

| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

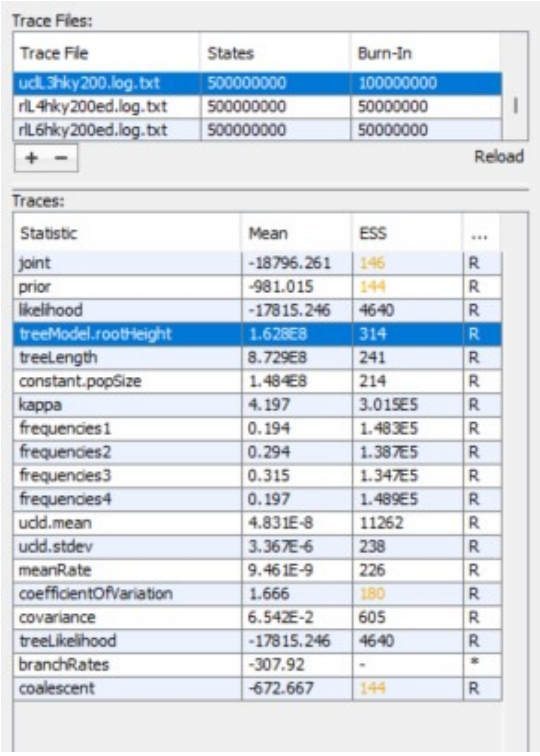
+ −  Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -9318.137 | 4576 | R |
| prior | -460.66 | 5177 | R |
| likelihood | -8857.476 | 3495 | R |
| treeModel.rootHeight | 1.773E6 | 5102 | R |
| treeLength | 2.318E7 | 5312 | R |
| constant.popSize | 9.728E6 | 5366 | R |
| kappa | 3.697 | 6610 | R |
| frequencies1 | 0.196 | 2969 | R |
| frequencies2 | 0.301 | 2984 | R |
| frequencies3 | 0.317 | 3141 | R |
| frequencies4 | 0.186 | 3449 | R |
| ucld.mean | 4.812E-8 | 5781 | R |
| ucld.stdev | 4.445E-9 | 1714 | R |
| meanRate | 4.845E-8 | 5624 | R |
| coefficientOfVariation | 8.79E-2 | 1654 | R |
| covariance | -2.355E-2 | 8638 | R |
| treeLikelihood | -8857.476 | 3495 | R |
| branchRates | -138.228 | - | * |
| coalescent | -323.816 | 5357 | R |

Figure 48: UCL L2.3.2 ESS Table

.



Trace Files:

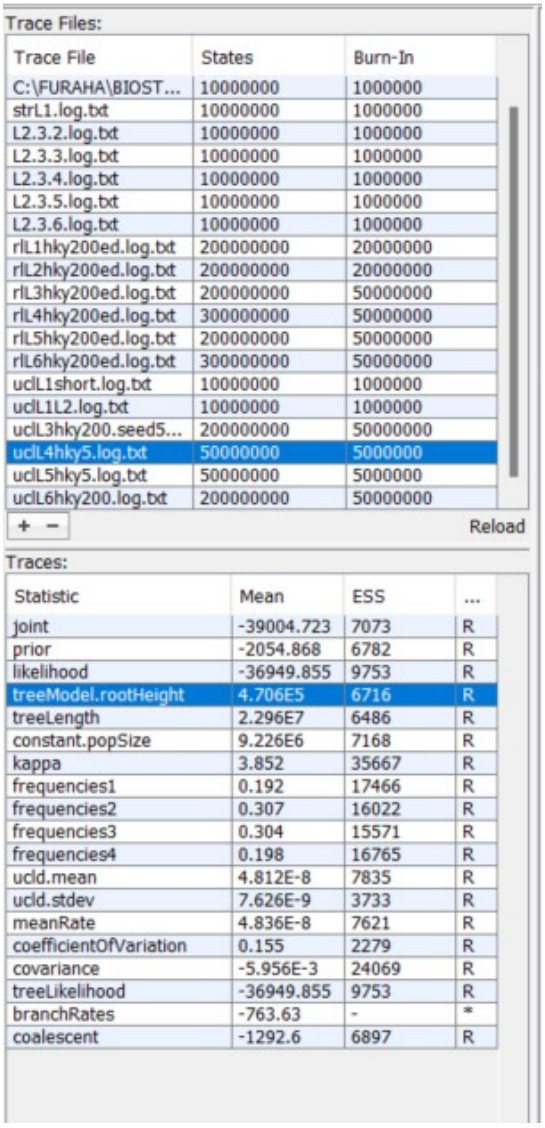
| Trace File | States | Burn-In |
|---|---|---|
| uclL3hky200.log.txt | 500000000 | 100000000 |
| rlL4hky200ed.log.txt | 500000000 | 50000000 |
| rlL6hky200ed.log.txt | 500000000 | 50000000 |

+ − Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -18796.261 | 146 | R |
| prior | -981.015 | 144 | R |
| likelihood | -17815.246 | 4640 | R |
| treeModel.rootHeight | 1.628E8 | 314 | R |
| treeLength | 8.729E8 | 241 | R |
| constant.popSize | 1.484E8 | 214 | R |
| kappa | 4.197 | 3.015E5 | R |
| frequencies1 | 0.194 | 1.483E5 | R |
| frequencies2 | 0.294 | 1.387E5 | R |
| frequencies3 | 0.315 | 1.347E5 | R |
| frequencies4 | 0.197 | 1.489E5 | R |
| ucld.mean | 4.831E-8 | 11262 | R |
| ucld.stdev | 3.367E-6 | 238 | R |
| meanRate | 9.461E-9 | 226 | R |
| coefficientOfVariation | 1.666 | 180 | R |
| covariance | 6.542E-2 | 605 | R |
| treeLikelihood | -17815.246 | 4640 | R |
| branchRates | -307.92 | - | * |
| coalescent | -672.667 | 144 | R |

Figure 49: UCL L2.3.3 ESS Table



Trace Files:

| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

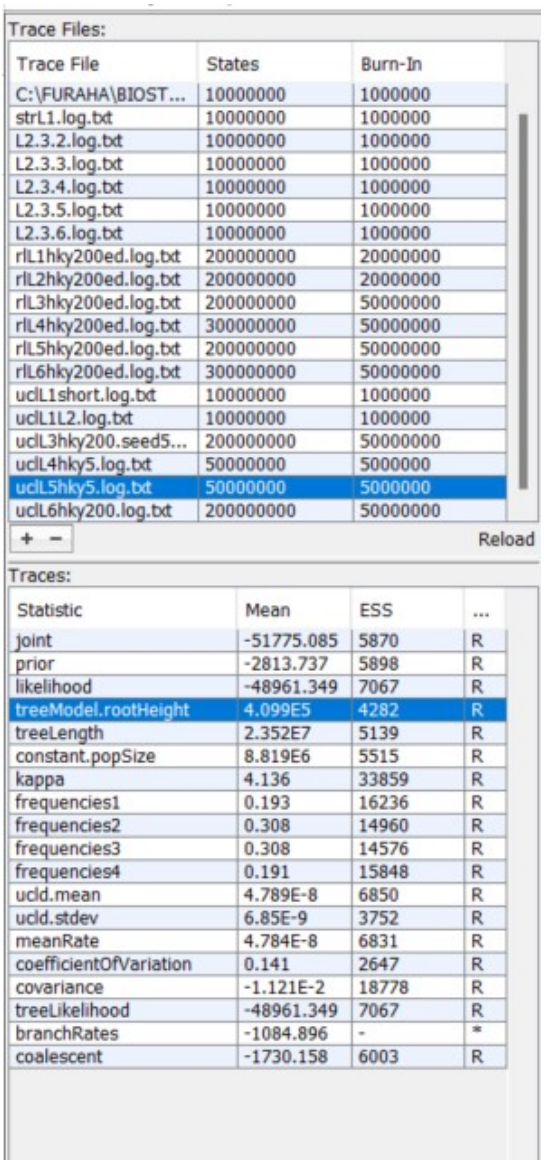
+ − Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -39004.723 | 7073 | R |
| prior | -2054.868 | 6782 | R |
| likelihood | -36949.855 | 9753 | R |
| treeModel.rootHeight | 4.706E5 | 6716 | R |
| treeLength | 2.296E7 | 6486 | R |
| constant.popSize | 9.226E6 | 7168 | R |
| kappa | 3.852 | 35667 | R |
| frequencies1 | 0.192 | 17466 | R |
| frequencies2 | 0.307 | 16022 | R |
| frequencies3 | 0.304 | 15571 | R |
| frequencies4 | 0.198 | 16765 | R |
| ucld.mean | 4.812E-8 | 7835 | R |
| ucld.stdev | 7.626E-9 | 3733 | R |
| meanRate | 4.836E-8 | 7621 | R |
| coefficientOfVariation | 0.155 | 2279 | R |
| covariance | -5.956E-3 | 24069 | R |
| treeLikelihood | -36949.855 | 9753 | R |
| branchRates | -763.63 | - | * |
| coalescent | -1292.6 | 6897 | R |

Figure 50: UCL L2.3.4 ESS Table

.



Trace Files:

| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

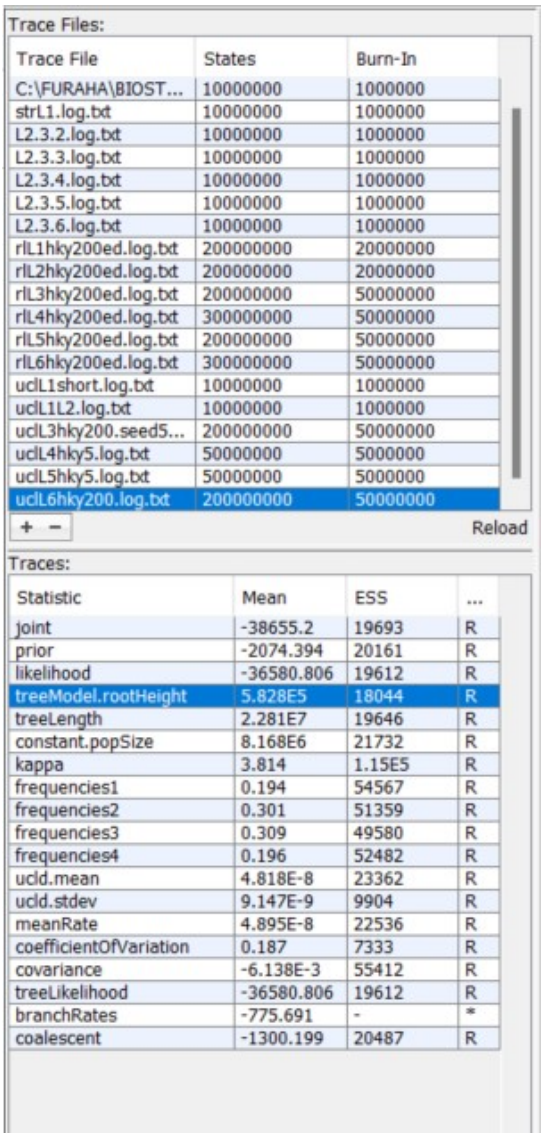
+ −

Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -51775.085 | 5870 | R |
| prior | -2813.737 | 5898 | R |
| likelihood | -48961.349 | 7067 | R |
| treeModel.rootHeight | 4.099E5 | 4282 | R |
| treeLength | 2.352E7 | 5139 | R |
| constant.popSize | 8.819E6 | 5515 | R |
| kappa | 4.136 | 33859 | R |
| frequencies1 | 0.193 | 16236 | R |
| frequencies2 | 0.308 | 14960 | R |
| frequencies3 | 0.308 | 14576 | R |
| frequencies4 | 0.191 | 15848 | R |
| ucld.mean | 4.789E-8 | 6850 | R |
| ucld.stdev | 6.85E-9 | 3752 | R |
| meanRate | 4.784E-8 | 6831 | R |
| coefficientOfVariation | 0.141 | 2647 | R |
| covariance | -1.121E-2 | 18778 | R |
| treeLikelihood | -48961.349 | 7067 | R |
| branchRates | -1084.896 | - | * |
| coalescent | -1730.158 | 6003 | R |

Figure 51: UCL L2.3.5 ESS Table



Trace Files:

| Trace File | States | Burn-In |
|---|---|---|
| C:\FURAHA\BIOST... | 10000000 | 1000000 |
| strL1.log.txt | 10000000 | 1000000 |
| L2.3.2.log.txt | 10000000 | 1000000 |
| L2.3.3.log.txt | 10000000 | 1000000 |
| L2.3.4.log.txt | 10000000 | 1000000 |
| L2.3.5.log.txt | 10000000 | 1000000 |
| L2.3.6.log.txt | 10000000 | 1000000 |
| rlL1hky200ed.log.txt | 200000000 | 20000000 |
| rlL2hky200ed.log.txt | 200000000 | 20000000 |
| rlL3hky200ed.log.txt | 200000000 | 50000000 |
| rlL4hky200ed.log.txt | 300000000 | 50000000 |
| rlL5hky200ed.log.txt | 200000000 | 50000000 |
| rlL6hky200ed.log.txt | 300000000 | 50000000 |
| uclL1short.log.txt | 10000000 | 1000000 |
| uclL1L2.log.txt | 10000000 | 1000000 |
| uclL3hky200.seed5... | 200000000 | 50000000 |
| uclL4hky5.log.txt | 50000000 | 5000000 |
| uclL5hky5.log.txt | 50000000 | 5000000 |
| uclL6hky200.log.txt | 200000000 | 50000000 |

+ −

Reload

Traces:

| Statistic | Mean | ESS | ... |
|---|---|---|---|
| joint | -38655.2 | 19693 | R |
| prior | -2074.394 | 20161 | R |
| likelihood | -36580.806 | 19612 | R |
| treeModel.rootHeight | 5.828E5 | 18044 | R |
| treeLength | 2.281E7 | 19646 | R |
| constant.popSize | 8.168E6 | 21732 | R |
| kappa | 3.814 | 1.15E5 | R |
| frequencies1 | 0.194 | 54567 | R |
| frequencies2 | 0.301 | 51359 | R |
| frequencies3 | 0.309 | 49580 | R |
| frequencies4 | 0.196 | 52482 | R |
| ucld.mean | 4.818E-8 | 23362 | R |
| ucld.stdev | 9.147E-9 | 9904 | R |
| meanRate | 4.895E-8 | 22536 | R |
| coefficientOfVariation | 0.187 | 7333 | R |
| covariance | -6.138E-3 | 55412 | R |
| treeLikelihood | -36580.806 | 19612 | R |
| branchRates | -775.691 | - | * |
| coalescent | -1300.199 | 20487 | R |

Figure 52: UCL L2.3.6 ESS Table

.