UNIVERSITY OF CALGARY

Parallelization of Bayesian Phylogenetics to Greatly Improve Run Times

by

David Yang

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

MARCH, 2024

Abstract

Phylogenetic analyses are invaluable to understanding the transmission of viruses, especially during disease outbreaks. In particular, Bayesian phylogenetics has great potential in modeling viral transmission due to the numerous phylogenetic models that can be incorporated. Currently, the availability of user-friendly software and accessibility to sequence data makes phylogenetic analyses easy to perform. However, to date, Bayesian phylogenetic analyses are still limited by long computational run-times which are especially unfavorable during ongoing and evolving disease outbreaks that demand real-time phylogeny results. Current optimization methods of Bayesian phylogenetic analysis mainly focus on iteration-level parallelization and mostly overlook the potential of larger-scale parallelization approaches. In this thesis, we provide an in-depth overview of topics including phylogenetic analysis, relevant biological information, and phylogenetic analysis optimization methods. We also proposed a novel parallelized Markov Chain Monte Carlo method that greatly improved Bayesian phylogenetic run times and integrated the approach into a data pipeline to allow for the direct analysis of viral samples. We demonstrated the validity of our methods by performing phylogenetic analyses on two sets of HIV simulation data and one set of real-world SARS-CoV-2 data. Our results suggested that the parallelization of MCMC in Bayesian phylogenetic analyses drastically reduces run times by 29-fold without causing significant deviations in parameter estimates and predicted phylogenetic trees.

Preface

This thesis is an original, unpublished work by the author, David Yang. This work was the result of graduate research performed under the supervision of Dr. Paul Gordon and Dr. Qingrun Zhang.

Acknowledgements

I would like to express my deepest appreciation and gratitude to the following individuals who have played a significant role in the completion of my thesis. First and foremost, I am immensely grateful to my supervisor, Dr. Qingrun Zhang and Co-supervisor, Dr. Paul Gordon, for their invaluable guidance, unwavering support, and expertise throughout the entire research process. Their mentorship has been instrumental in shaping the direction of this work and my growth as a researcher.

I am also thankful to my committee members, Dr. Wenyuan Liao and Dr. Frank Van der Meer, for their constructive feedback and expertise. I extend my appreciation to our research group for their collaboration and encouragement during this academic journey. In particular, their explanations of the TransCOVID pipeline has been exceptionally helpful for the completion of my thesis. Special thanks to my family and friends for their unwavering support and understanding. I am also grateful to Alberta Innovates for the financial support that made this research possible.

Finally, I want to acknowledge the countless individuals who contributed directly or indirectly to the completion of this thesis. Your assistance has been instrumental, and I am thankful for the collective effort that has shaped this work.

Contents

List of Figures

List of Tables

Acronyms

AIDS Acquired Immunodeficiency Syndrome. 6

BDSKY Birth-Death Skyline plot. vi, 13, 14, 17, 21

BEAGLE broad-platform evolutionary analysis general likelihood evaluator. 18, 21

BEAST2 Bayesian Evolutionary Analysis Sampling Trees 2. 18–21, 25

COVID-19 Coronavirus disease 2019. 6, 22, 26

DNA Deoxyribonucleic Acid. 4, 5

ESS effective sample size. 12, 25

FAVITES The software Framework for Viral Transmission and Evolution Simulation. vii, 26–30

FFT Fast Fourier Transform. 8

GeoUnrooted Geodesic Unrooted distance. 15, 30

GISAID The Global Initiative on Sharing All Influenza Data. 1, 2, 48

GTR General Time-Reversible. 9, 10, 17, 22, 27

HIV Human Immunodeficiency Virus. 5, 6, 25, 27, 32

HKY Hasegawa-Kishino-Yano. 9

JC69 Jukes and Cantor 1969. 9

MAFFT Multiple Alignment using Fast Fourier Transform. 8, 22

MatchingSplit Matching split distance. 15, 30

MCC maximum clade credibility. 13, 29, 32, 34, 43, 49

MCMC Markov Chain Monte Carlo. i, 12, 19–21, 25, 29, 30, 32, 34, 38, 43, 47, 50

MP maximum parsimony. 10, 11

mRNA messenger Ribonucleic Acid. 4

MSA Multiple Sequence Alignment. 8

NNI Nearest-Neighbour-Interchange. 18

RF Robinson-Fould's distance. 15, 30

RFWeighted Weighted Robinson-Fould's distance. 15, 30

RNA Ribonucleic Acid. 5, 6

SARS-CoV-2 Severe Acute Respiratory Syndrome Coronavirus 2. 5, 6, 25

SPR Subtree-Pruning-and-Regrafting. 18

TN93 Tamura-Nei 93. 9

UMAST Unrooted Maximum Agreement Subtree distance. 15, 30

Chapter 1

Introduction

In this chapter, we provide an introduction for this thesis. In Section 1.1, we discuss the context of this project and the motivation behind our work. In Section 1.2 we introduce fundamental biological background information in order to understand this project. In Section 1.3 we provide an overview of phylogenetic analysis. In particular, we explain Bayesian phylogenetic analyses and any complications that require attention when implementing it in practice. Finally, in Section 1.4, we outline the various limitations of phylogenetic research and summarize current attempts to solve these issues.

1.1 Motivation

The emergence of viral outbreaks poses an immediate and salient threat to global health, demanding swift and effective responses from the scientific community. The COVID-19 pandemic has emphasized the need for advanced tools and methodologies for unraveling the complexities of viral evolution. Phylogenetic analysis is a key component in elucidating genetic relationships among transmittable pathogens and plays a pivotal role in informing public health interventions. However, the unprecedented scale and speed of the COVID-19 pandemic have exposed the limitations of current phylogenetic analysis methods. This has driven the exploration for innovative approaches to accelerate phylogenetic research and allow for timely implementation of preventative measures. The considerations that have prompted this thesis are discussed in greater detail.

Availability and access to exorbitant amounts of viral genomic data: Unprecedented amounts of genomic data were generated during the COVID-19 pandemic. Genomic repositories such as The Global Initiative on Sharing All Influenza Data (GISAID) have amassed an abundance of viral sequences which are invaluable to phylogenetic research. To

date, more than 2 million SARS-CoV-2 sequences are publicly available on GISAID. However, current phylogenetic analysis methods are incapable of processing this sheer volume of data. Simpler analysis can be performed, however they often fail to capture the complexity of the viral outbreak. The need to process and analyse these large datasets quickly and accurately have motivated the search for optimization techniques.

Demand for real-time phylogeny results: The demand for immediate and urgent public health interventions during viral outbreaks have underscored the need to expedite phylogenetic research. Traditional phylogenetic methods often falls short in terms of providing timely insights into viral transmission and evolution. For example, a naive analysis involving a couple hundred sequences can take up to days and even weeks to complete [1]. Not only are these long run times especially impractical in the sense of informing public health decisions, but they also fail to keep up with the ongoing and evolving COVID-19 pandemic. Through optimization techniques, researchers can significantly reduce the computational time required for phylogenetic analysis.

Reduction of computational demand: The scarcity of computational resources, combined with the growing demand for high-performance computing in genomics, have highlighted the need for optimization methods. Parallelization which leverages concurrent processing, emerges as a solution to handle large-scale genomic data quick and efficiently. This imperative has inspired the search for parallelization methods tailored specifically for phylogenetic analysis. By doing so, we would not only addresses the current challenges posed by the COVID-19 pandemic but we would also lay the groundworks for handling future viral outbreaks with even greater computational demands.

Hence, in this paper, we present a novel parallelization method designed for Bayesian phylogenetic analysis with the software BEAST2. To demonstrate the validity of our method, we performed phylogenetic analyses on two simulated HIV datasets and a SARS-CoV-2 dataset. To enhance our understanding, we utilized various distance metrics and statistical tests to assess the output phylogenetic trees and MCMC predicted parameters. Our results suggested that parallelization of MCMC in Bayesian phylogenetic analyses drastically reduced run-times by 28-fold, all while maintaining the integrity of parameter estimates and predicted phylogenetic trees.

1.2 Biological Background

Theory of Evolution:

The theory of evolution was first proposed by Charles Darwin in the 19th century. It is a fundamental biological concept that governs the process of change in all forms of "life" over successive generations. Though at the time, it's initial applications were for organisms, it is also relevant for other biological entities such as viruses and prions that by some are not considered "living". In Darwin's work "On the Origin of Species", Darwin proposed that all forms of life are related and share a common ancestor at some point in time [2]. However, over time, populations of organisms undergo changes in their traits and characteristics. In his work, he introduced the mechanism *natural selection* as the driving force behind evolutionary change. The four main components of natural selection are the following:

1. Variation: Individuals within populations exhibit variation. Some individuals possess traits that are beneficial to their fitness and survival.
2. Inheritance: Traits that are genetically determined are inherited by the individual's offspring.
3. Competition: Resources within a population are scarce which creates competition.
4. Survival of the fittest: Individuals with favorable traits survive and reproduce, thereby passing on their advantageous traits to the next generation.

Over successive generations, the accumulation of these advantageous traits leads to adaptations of a population to their environment. Through time, these populations sufficiently diverge and result in speciation. Although these principles were first proposed in the 19th century, they are still integrated within modern evolution theories today such as Modern Synthesis.

Contemporary theories of evolution considers and describes advanced complexities unaccounted for in Darwin's theory of evolution [3]. Modern evolution theories builds upon Darwin's fundamental principles but also incorporates new insights from genetics, molecular biology, population genetics, developmental biology, and other fields. Through our understanding of genetics and molecular biology, we can attribute the inheritance of individual traits to factors such as genes, epigenetics, and developmental conditions. Moreover, population genetics suggests that in addition to natural selection, other mechanisms such as

genetic drift and gene flow also drive evolutionary changes. Ultimately, the theory of evolution implies that there is a tree of life which can be studied through phylogenetic analysis and represented as phylogenetic trees.

Molecular and Population Genetics:

Genetics play a fundamental role in the process of evolution, providing the hereditary information that is passed through successive generations [4]. The basic building blocks of genetics are molecules called Deoxyribonucleic Acid (DNA), which are constructed by smaller units called nucleotides. Nucleotides contain three main components: a phosphate group, a deoxyribose sugar molecule, and a nitrogenous base (A = adenine, T = thymine, C = cytosine, G = guanine). The distinctive characteristics of DNA, including its complementary base pairing of A-T and G-C, along with its double helix structure, enable the preservation of integrity and the storage of genetic instructions. Within the genome, specific sequences of DNA, genes, encode for the synthesis of proteins and regulate cellular processes. Genes undergo an intermediary stage, converting first into messenger Ribonucleic Acid (mRNA) sequences. In mRNA, groups of three consecutive nucleotides (codon), correspond to unique amino acids depending on their combination of nucleotide bases. Strands of amino acids then fuse together to form intricate protein structures observable by the naked eye.

Within this complex process, genes are the source of phenotypic variation [4]. The accumulation of mutations within genes leads to the expression of different variants of a trait. Types of gene mutations include:

- Substitution mutations: one nucleotide within a codon is substituted with a different nucleotide resulting in the instruction for a different amino acid. These mutations can be classified as transitions or transversions as shown in Figure 1.1.
- Insertion and deletion mutations: entire sequences of nucleotides are removed or added within a gene.
- Frameshift mutations: the removal or addition of some amount of nucleotides resulting in a shift of the reading frame for the codons.

Other factors such as gene duplications and deletions, and chromosomal changes can also dictate the phenotype of an organism [4]. These phenotypic differences and their corresponding genetic information are used to make inferences on evolutionary relatedness. Hence, genetic sequencing is a prerequisite to phylogenetic analysis. In most organisms this involves

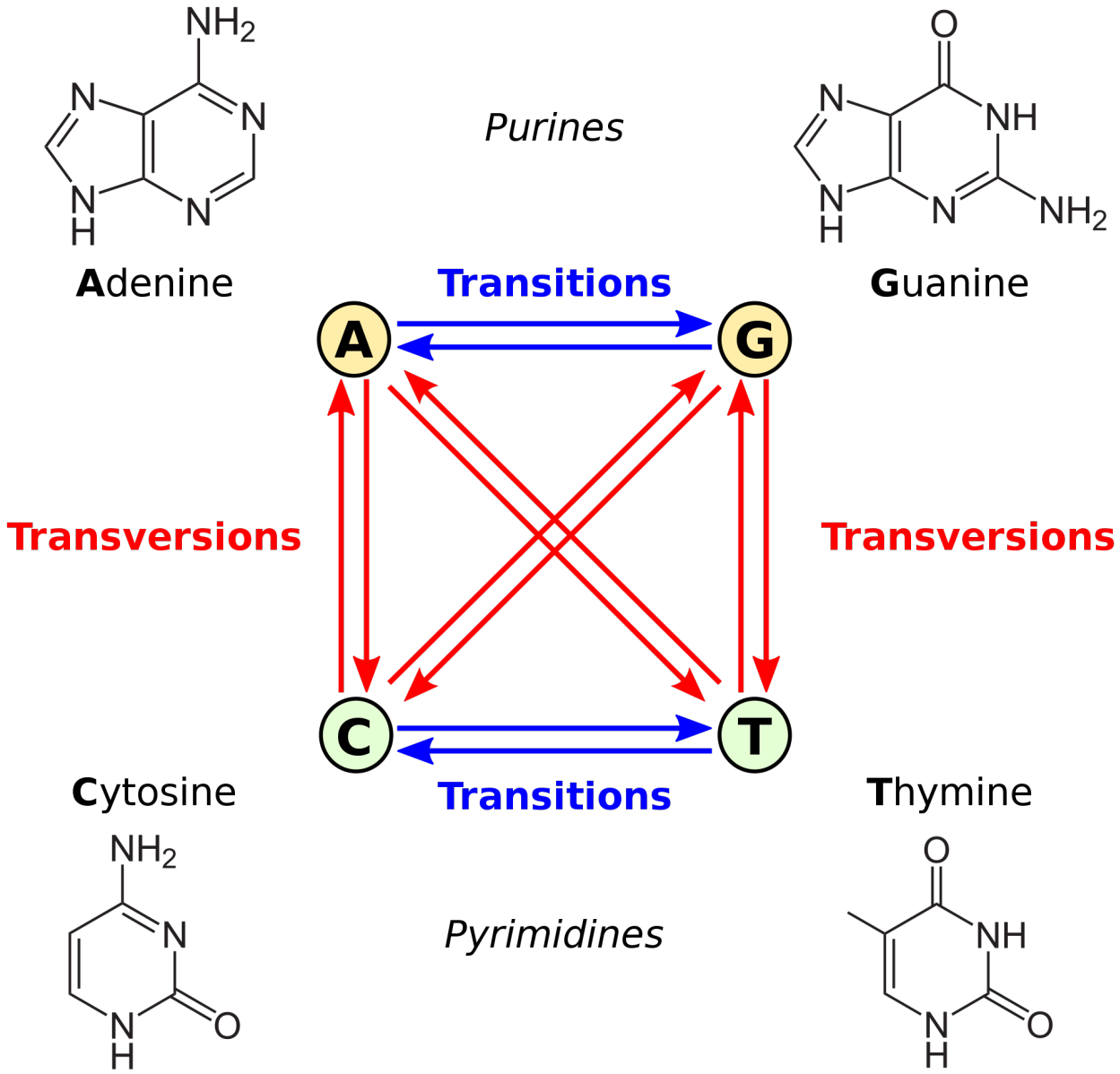


Figure 1.1: Diagram illustrating transversion and transition mutations in Deoxyribonucleic Acid molecules [5].

sequencing DNA, however in many Ribonucleic Acid (RNA) viruses such as Human Immunodeficiency Virus (HIV) and Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), this involves RNA instead. The sampled results are usually then stored as FASTA files which can be utilized in phylogenetic analysis.

Relating to viral outbreaks:

For the purpose of this thesis, it is important to discuss the concepts of evolution and genetics in the context of viral outbreaks. Viruses are infectious agents that rely on infecting host cells to propagate and replicate. Since viruses are not able to carry out basic metabolic functions themselves, it is disputed whether they are considered "living". Although viruses exhibit extreme diversity in terms of morphology, most viruses contain some sort of genetic information (RNA or DNA) and a protein shell called a capsid [6]. Viruses such as HIV and SARS-CoV-2 are heavily studied since they are the cause of the human diseases Acquired Immunodeficiency Syndrome (AIDS) and Coronavirus disease 2019 (COVID-19) respectively.

The discussion of evolution and genetics are important for understanding the dynamics of a viral outbreak. While viruses are not classified as living organisms due to their inability to carry out metabolic processes independently, the evolution of viruses is governed by the same evolutionary principles [6]. Viruses engage in an ongoing battle with their hosts, propelling swift evolutionary changes. Notably, certain viruses like HIV and SARS-CoV-2 store genetic information in the more mutable RNA form, leading to a higher incidence of genetic mutations. These elevated mutation rates contribute significantly to the genetic diversity within viral populations. This diversity, in turn, facilitates rapid adaptation to dynamic environments, selective pressures, and the challenges posed by host immune responses. Certain mutations can result in key advantages in terms of immune evasion, host invasion, and drug resistance which are selected for by natural selection. Often, these mutations are used to classify viruses into strains. As mutations accumulate over time in the genomes of viruses, the analysis of genetic differences between viral sequences can reveal evolutionary relationships. In the context of viral outbreaks, viruses can replicate through disease spread and thus the evolutionary relationship indicate viral transmission.

1.3 Phylogenetic Analysis

Phylogenetic analysis is the elucidation of evolutionary relationships amongst biological entities and has always been a central theme in scientific research. Work in these areas has not only furthered our understanding of historical processes that shaped the diversity of life, but they continue to provide meaningful insights into current age phenomena. Namely, Phylogenetic analyses serves as a critical tool for extracting significant information crucial to understanding the complex dynamics of viral evolution and transmission. Namely, phylogenetic trees can be utilized to infer transmission networks that elucidate disease spread within a population. Works involving applications of phylogenetic analysis to the SARS epidemic,

MERS outbreak, and other transmittable diseases such as HIV can be readily found [7, 8].

In consideration of the recent COVID-19 pandemic, phylogenetic analysis was showcased as an invaluable tool in understanding viral outbreaks and mapping out the transmission of viruses. Through phylogenetic studies of SARS-CoV-2, the probable zoonotic origin of COVID-19 was identified [9]. Furthermore, similar studies also explained the cause for multiple SARS-CoV-2 strains during the early pandemic as the outcome of multiple episodes of the founder effect from Wuhan China (Gomez-Carballa et al., 2020). Concerns for the extensive spread of COVID-19 also led many phylogenetic studies to work on containing the transmission of the virus. Some studies, for example, clarified infection sources for infected individuals [10] while other studies detected "super-spreaders" which contributed prominently to SARS-CoV-2 spread [11]. Remarkably, other studies also utilized phylogenetic analyses to identify elusive asymptomatic individuals that drove COVID-19 infections during the early stages of the pandemic [12]. Ultimately, such work was essential to infection tracking which was demonstrated to be crucial during the ongoing and rapidly evolving COVID-19 pandemic.

Phylogeny refers to the evolutionary relationships between organisms and entities. A central goal of phylogenetic analysis is to construct phylogenetic trees that depict the branching patterns of evolution. An example of a phylogenetic tree is shown in Figure 1.2. In phylogenetic trees, nodes represent common ancestors and divergence events. Meanwhile, branches are used to depict evolutionary lineages with branch lengths indicating the degree of change or time elapsed. Finally, terminal taxa or tips in phylogenetic trees usually represent sequences that are being studied or compared. Phylogenetic trees can be categorized as rooted or unrooted, which differs in the fact that rooted phylogenetic trees possess a common ancestor for all samples. In the context of viral outbreaks, transmission networks can be inferred from phylogenetic trees. Transmission networks are powerful tools for understanding the spread of a pathogen within a community. In transmission networks, sequences are grouped in clusters instead, which represent individuals that were infected from the same source. This knowledge is invaluable for informing public health decisions.

In phylogenetic analysis, phylogenies are primarily inferred through molecular data such as nucleotide and amino acid sequences. In viral outbreaks, samples are collected from infected individuals through blood and respiratory samples. The quality and representativeness of sampled sequences significantly affect the outcome of phylogenetic analysis. Many studies implement robust sampling strategies in order to capture the viral diversity [14, 15]. It is important to note that adding more sequences to the analysis in hopes of obtaining better results is simply not enough [16].

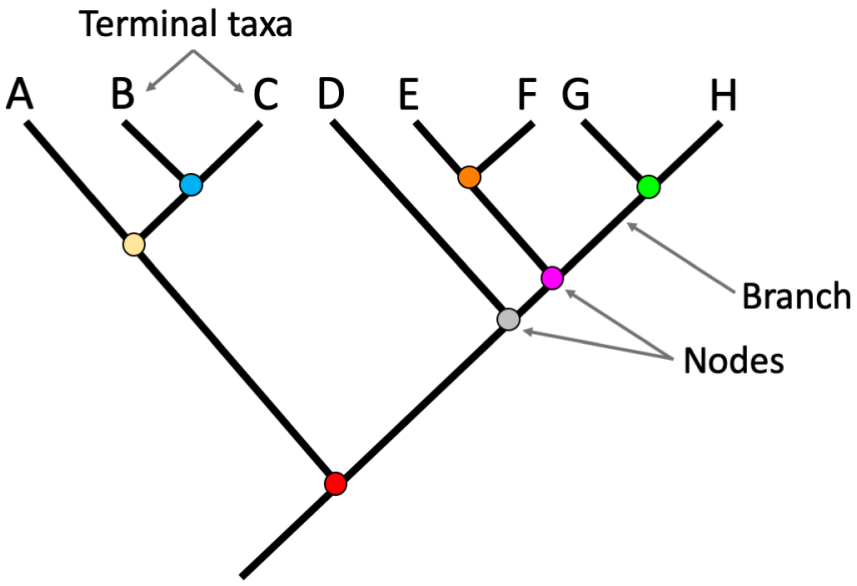


Figure 1.2: Diagram outlining various components in a rooted phylogenetic tree [13].

Sampled molecular sequences are aligned through Multiple Sequence Alignment (MSA) in order to identify homologous sequences that are conserved across different samples. Multiple Alignment using Fast Fourier Transform (MAFFT) is a widely used tool for performing MSA.

Multiple Alignment using Fast Fourier Transform:

MAFFT involves the application of the Fast Fourier Transform (FFT) to efficiently align multiple molecular sequences [17, 18]. The key algorithm in MAFFT is based on a progressive alignment strategy, where sequences are aligned in a stepwise manner ordered by similarity. The general algorithm in MAFFT follows these five key steps:

1. **Pairwise alignment**: Identification of conserved regions in all sequences.
2. **Distance matrix:** Calculation of distance matrix using the pairwise alignment to evaluate dissimilarity between alignments. For each alignment, an alignment score is calculated.
3. **Guide tree**: Construction of a guide tree using the distance matrix where nodes represent the sequences and branch lengths represent the distances [18].
4. **Progressive Alignment**: Progressive alignment is performed using the guide tree, starting from the leaves and working towards the root.

5. **Iterative alignment**: Repetition of the process with adjustments to refine alignment accuracy.

Basic Phylogenetic Models:

Aligned sequences are then used to construct phylogenetic trees. Several methods exist for constructing phylogenetic trees. The main approaches are based on maximum likelihood, maximum parsimony, and Bayesian statistics. Understanding these tree construction processes first requires understanding the process of sequence evolution. These understandings are incorporated into phylogenetic studies through substitution models which describe the probabilities of different nucleotide or amino acid changes (Table 1.1). These models allow researchers to estimate relatedness between samples based on the differences in the sequences.

Table 1.1: Substitution models

Model	Transitions	Transversions	Base Frequencies
JC69	No distinction		All are equal and fixed (0.25)
HKY	Equal	Equal	Unequal, free values
TN93	Independent	Equal	Unequal, free values
GTR	Independent	Independent	Unequal, free values

The simplest substitution model is Jukes and Cantor 1969 (JC69) [19]. This model assumes that the probability of all bases is equal, $\pi_A = \pi_G = \pi_C = \pi_T = 0.25$ and that they all share the same substitution rate, μ. Meanwhile, other substitution models such Hasegawa-Kishino-Yano (HKY) [20], Tamura-Nei 93 (TN93) [21], and General Time-Reversible (GTR) [22] assume that there are unequal base frequences, $\pi_A \neq \pi_G \neq \pi_C \neq \pi_T$. These models also make distinctions between the substitution rates of transitions and transversions. In the HKY model, the transition and transversion rates can vary, however, all transition rates are equal and all transversion rates are equal. In the TN93 model, the rates of transition substitutions are allowed to vary, while in the GTR model, rates of both transitions and transversions are independent. In the GTR model, the substitution rate matrix is

$$Q = \begin{bmatrix} - & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & - & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & - & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & - \end{bmatrix}$$

where, α = substitution rate between A and G, β = substitution between A and C, γ = substitution rate between A and T, δ = substitution rate between G and C, ϵ = substitution rate between G and T, and η = substitution rate between C and T.

The GTR substitution model is most commonly used in genetic studies. Its flexibility and independence of substitution rates allow it to provide realistic representations of substitution rates compared to simpler models. These assumptions of the GTR model are important for representing viral outbreaks. Viral genomes have unequal compositions of nucleotides each with varying substitution rates. Moreover, evolutionary pressures cause differential rates of mutation within different regions of the genome [23]. For instance, natural selection favors mutations in regions encoding proteins involved with host infection, immune evasion, and drug resistance. Hence, the GTR model shows great promise for modeling the sequence evolution of viral outbreaks.

While substitution models are important for inferring phylogenetic tree topology, molecular clock models are crucial in understanding the temporal aspects of phylogenetic trees. Molecular clock models assume a rate of evolution over time which is essential for inferring divergence times between species or strains, and estimating tree branch lengths [24]. Two main categories of molecular clock models are strict and relaxed molecular clocks. Strict molecular clocks assume a constant rate of mutation across all lineages. Correspondingly, in phylogenetic studies involving strict molecular clocks, genetic differences directly indicate time since divergence. While strict clocks are simple to implement in practice, they overlook the fact that evolutionary rates can vary across lineages. Hence, relaxed molecular clocks emerge as the favorable alternative. Relaxed molecular clock models are more realistic in that they capture the heterogeneity of mutation rates across different lineages by allowing variation in mutation rates [24]. In uncorrelated log-normal relaxed molecular clock models, the mutation rates in each lineage are assumed to be independent and drawn from a log-normal distribution [25]. This aligns with the understanding of viral outbreaks such as SARS-CoV-2, which evolved into strains with dissimilar mutation rates [26]. Together, these phylogenetic models provide the assumptions necessary to construct phylogenetic trees for ongoing and evolving viral outbreaks such as COVID-19.

Maximum parsimony phylogenetic methods:

The maximum parsimony (MP) phylogenetic approach seeks to find the tree topology that minimizes the number of evolutionary changes required to explain the sequence data [27, 28]. These methods involve the calculation of a parsimony criterion. Some parsimony criterion involves the Hamming distance, $H(x, y)$ which is used to compare two equal length

sequences x and y [27]. The Hamming distance is equal to the number of positions in both sequences that are not equal.

$$\sum_{i=1}^{n} I(x_i \neq x_j)$$

where n = the length of the sequences and $I()$ = the indicator function. In phylogenetic trees, the Hamming distance, $H(e) = H(x, y)$ is calculated for each edge in the tree, where x and y are the two endpoints of the edge. Letting e denote the edge, and $E(t)$ denoting all edges in the phylogenetic tree, the parsimony criterion can be calculated as $\sum_{e \in E(t)} H(e)$. Although in this example, the Hamming distance is utilized, other maximum parsimony-based methods may employ different distance metrics. Ultimately, maximum parsimony-based phylogenetic techniques are intuitive in principle but suffer from several limitations.

MP methods require an exhaustive search for the most parsimonious phylogenetic tree within the "tree space" which consists of all possible phylogenetic trees for the data. While heuristic methods are available, these do not guarantee that the most parsimonious phylogenetic tree is selected [29]. Moreover, MP methods assume constant substitution rates within all lineages which is usually not representative of the real world, such as during viral outbreaks[30]. MP methods also assume that the simplest explanation is the correct one even though this is not always the truth. For example, this assumption implies that homoplasic mutations (convergent evolution) are uncommon. Therefore, in cases with significant amounts of homoplasy, such as in viral genomes, MP methods will incorrectly infer evolutionary relationships [31, 32]. Finally, MP cannot incorporate phylogenetic models for sequence evolution, and thus fail to account for complex evolutionary processes [33]. Ultimately the limitations illustrated above steer researchers away from MP based methods, especially for complex datasets.

Maximum likelihood phylogenetic methods:

The essence of maximum likelihood-based phylogenetic methods is to maximize the likelihood of the observed sequence data given a particular tree topology, tree branch lengths, and model parameters. Bases within molecular sequences are considered independent and their corresponding log-likelihoods are calculated separately. The log-likelihood for all sites in a sequence are then summed and for a given set of evolutionary parameters. The likelihood function used in this process is the following:

$$L(Data|\theta, \tau) = \prod_{i=1}^{n} P(x_i|\theta, \tau)$$

where θ = the set of model parameters,τ = the phylogenetic tree, n = number of sites,

x_i = i-th site in the alignment. In the likelihood equation, $P(x_i|\theta,\tau)$ is determined by the substitution model used. In the naive case using the Jukes-Cantor model, the probability of observing a nucleotide at a site is given by $P(x_i|\theta,\tau) = \frac{1}{4} \times \frac{3}{4}e^{-\frac{4}{3}\mu}$ where μ = the overall substitution rate. The process of calculating the likelihood is repeated for all possible tree topologies, branch lengths, and model parameters and the set of parameters yielding the highest likelihood is selected. Typically this task of inference is carried out by optimization algorithms such as the Newton-Raphson method or heuristic search methods. A main advantage of the maximum likelihood approach is that it is very consistent. However, the search for the highest likelihood is also an exhaustive process and is therefore very computationally demanding [34].

Bayesian phylogenetic methods:

Bayesian-based phylogenetic methods use Baye's theorem to infer phylogenetic trees from sequence data [35]. The key characteristic of Baye's theorem is the incorporation of prior distributions, $f(\theta)$, which represent our prior knowledge of various parameters of interest. The posterior distribution given the data is:

$$f(\theta,\tau|D) = \frac{f(\theta,\tau)f(D|\theta,\tau)}{\int f(\theta,\tau)f(D|\theta,\tau)}$$

where θ are all phylogenetic model parameters, τ is the phylogenetic tree, $f(D|\theta,\tau)$ is the likelihood, and $\int f(\theta,\tau)f(D|\theta,\tau) = f(D)$ is the marginal likelihood which acts as the normalizing constant. Since the marginal likelihood is constant given a dataset and difficult to derive, the posterior distribution can be estimated with $f(\theta,\tau|D) \propto f(\theta,\tau)f(D|\theta,\tau)$ instead.

In Bayesian-based approaches, algorithms such as Markov Chain Monte Carlo (MCMC) are essential for approximating the posterior distribution. MCMC methods sample model parameters such as tree topologies, substitution rates, and clocks rates, from the posterior distribution. This process generates Markov chains that only depend on the preceding value. The Markov chain explores the parameter space and eventually reaches a stationary distribution that is representative of the posterior distribution. Metrics such as the effective sample size (ESS) are used to assess whether the parameter space was sufficiently explored. Burn-in periods are also implemented to discard initial Markov chain values as they are not representative of the posterior distribution. An example of a MCMC method is the Metropolis-Hastings algorithm [36, 37]:

1. **Initialization**: Specify initial values, θ_0, for the parameters that are being sampled.

2. **Proposal distribution**: Choose a proposal distribution $q(\theta^*|\theta)$ where θ is the current state, and θ^* is the proposed state.

3. Sample θ^* from the proposal distribution $q(\theta'|\theta)$.

4. **Acceptance ratio**: calculate the acceptance ratio which is defined as $\alpha(\theta^*, \theta) = min(1, \frac{\pi(\theta^*)q(\theta|\theta^*)}{\pi(\theta)q(\theta^*|\theta)})$

5. Accept or reject the proposed value by generating a uniform random number $u \in [0, 1]$. Accept the proposed θ^* if $u \leq \alpha(\theta^*, \theta)$, else reject the proposed θ^*

6. Repeat steps 3 to 5, until a n-length Markov chain of $\theta_1, .., \theta_n$ is sampled.

These estimated parameters are then used to build the maximum clade credibility (MCC) tree - the most reasonable phylogenetic tree given the estimated Bayesian parameters.

Bayesian phylogenetic analysis has been widely used for viral outbreaks. This is because Bayesian phylogenetic methods are flexible frameworks that can capture very complex evolutionary dynamics. Bayesian phylogenetic methods also facilitate the integration of both genomic and epidemiological data such as contact tracing [38]. However, most notably, there is a variety of Bayesian prior models that can be easily incorporated depending on the focus of the study. In particular, the Birth-Death Skyline plot (BDSKY) model is especially effective for constructing phylogenetic trees for evolving populations such as RNA viruses which can be problematic for traditional coalescent-based models [39]. This is because one of the main assumptions of coalescent models is that the population remains constant [40].

On the other hand, the BDSKY model is based on a birth-death process Figure 1.3 [39]. In the model, infected individuals transmit the disease at a rate of λ and become noninfectious at a rate of δ. Moreover, individuals are sampled in the population at a rate of ψ and die at a rate of μ. In the model, death can imply several outcomes including death and recovery. The model assumes that once an individual is sampled, they are removed from the population since they are no longer at risk for transmission. An extension of BDSKY plot model is the BDSKY serial model which assumes that these rates of transmission, death, sampling, and noninfectiousness can vary within different lineages. These parameters in the BDSKY are very informative as they can be used to calculate epidemiological quantities such as effective reproductive number ($R = \frac{\lambda}{\mu+\psi}$) the duration of infection ($T = \frac{1}{\mu+\psi}$) and the probability of being sampled ($\pi = \frac{\psi}{\psi+\mu}$).

While the flexibility of Bayesian-based phylogenetic frameworks is an advantage, this freedom may also be problematic if incorrect assumptions are enforced. Hence, assessing the inferences from phylogenetic methods is very important. Parameters involving point

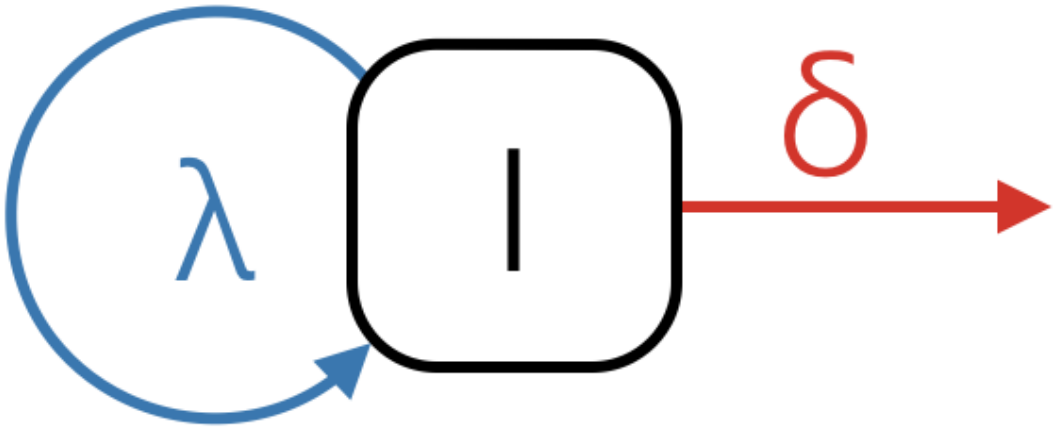


Figure 1.3: Diagram demonstrating the interactions between model parameters in the Birth-Death Skyline plot (BDSKY) model [41].

estimates can be evaluated using parametric and non-parametric tests such as the Student's t-test and Mann-Whitney U test respectively. The Student's t-test is a parametric statistical test used to compare two different groups [42]. In a one-sample t-test, the mean of a population is compared to the sample mean. The corresponding test statistic is

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

where $\bar{X}$ is the sample mean, μ is the population mean, $\hat{\sigma}$ is the population standard deviation.

The Mann-Whitney U test on the other hand is a non-parametric statistical test used to compare two samples that are not necessarily normally distributed [43]. In the U test, values from both samples are ranked based on order from smallest to largest. The test then involves the calculation of the test statistic:

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{i=n_1+1}^{n_2} R_i$$

where n_1, n_2 represent the sample sizes and R_i is the rank of the i^{th} element. These test statistics can then be compared to critical values for the desired α.

On the other hand, the comparison of proposed phylogenetic trees is typically performed using distance metrics that take into account the tree topology and branch lengths. Some common distance metrics used are described below.

Robinson-Fould's distance (RF): RF distance is calculated by $RF = \frac{(A+B)}{2}$ where A is the number of unique partitions in the first tree and B is the number of unique partitions in the second tree [44–46]. Note that a partition is defined as subsets of each tree where a branch is removed.

Weighted Robinson-Fould's distance (RFWeighted): RFWeighted distance is an extension of the RF distance [46, 47]. This metric assigns weights to branches based on branch length and sums the differences in weights between the partitions. This allows this distance metric to not only account for tree topology differences, but also branch length differences.

Unrooted Maximum Agreement Subtree distance (UMAST): UMAST distance is based on finding the largest common subtree, T_c, shared between two unrooted phylogenetic trees with the same taxa [46, 48]. The distance between the two trees, T_1, T_2 is equal to the difference in cardinality (number of unique elements) between T_c and T_1.

Geodesic Unrooted distance (GeoUnrooted): The GeoUnrooted distance is calculated based on this concept of orthants, which are Euclidean regions corresponding to tree topologies within the continuous tree space [49, 50]. Within each orthant, are coordinates of points corresponding to branch lengths within a tree. The geodesic distance of two phylogenetic trees, T_1, T_2, is the sum of the shortest possible distances between all coordinates in the orthants of T_1 and T_2. Similar to the RFWeighted distance, the GeoUnrooted distance also takes into account tree branch lengths [46].

Matching split distance (MatchingSplit): The Matching split distance (MatchingSplit)

distance is calculated by comparing all splits in the two phylogenetic trees [45, 46]. The distance computes the minimum weight perfect matching ($h_s(A_1|B_1, A_2|B_2)$) for all splits in both trees. The sum of the minimum weight perfect matching for all splits gives the MatchingSplit distance. The calculation for the minimum weight perfect matching is

$$h_s(A_1|B_1, A_2|B_2) = min\{|A_1| + |A_2| - 2|A_1 \cap A_2|, |L| - (|A_1| + |A_2| - 2|A_1 \cap A_2|)\}$$

In the equation, $A_1|B_1$ corresponds to the split between the subset A_1 and B_1 in the first phylogenetic tree, and $A_2|B_2$ corresponds to the split in the second phylogenetic tree. Moreover, $|L|$ denotes the cardinality of L, which consists of all taxa in the trees.

Path difference distance: The Path difference distance is calculated by obtaining the square of the difference in distance between a set of two leaves in the two phylogenetic trees being compared [46, 51]. This difference is summed in all pairs of leaves in both trees and square rooted to give the Path difference distance.

$$Path\ difference = \sqrt{\sum_{i=1}^{p} D_i^2}$$

where p is the number of pairs of leaves in the trees, and D_i is the difference in distance between a pair of leaves in the compared trees.

1.4 Limitations of Phylogenetic Analyses

In general, despite the power of phylogenetic analysis in unraveling evolutionary relationships, phylogenetic studies are limited by intense computational costs and long run times. A simple analysis involving 10 sequences can yield more than 2 million different unrooted and 34 million different rooted phylogenetic trees (Table 1.2). In addition, the number of possible trees grows factorially as more sequences are included. The relationship between the number of possible phylogenetic trees and the number of sequences (n) are shown by these equations:

$$rooted\ trees = \frac{(2n-3)!}{2^{n-2}(n-2)!} \quad , \ unrooted\ trees = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

Hence, the tree space that must be explored in phylogenetic analyses consists of overwhelming amounts of tree topologies, thereby requiring high computational demand and long runtimes [52]. This tree-search problem is accentuated by the large amounts of sequencing data available that can be included in phylogenetic studies.

Table 1.2: Phylogenetic tree complexity

Number of sequences	Number of rooted trees	Number of unrooted trees
2	1	1
5	105	15
10	3.44×10^7	2.03×10^6
15	2.13×10^14	7.91×10^12
20	8.20×10^21	2.21×10^20

On the other hand, the incorporation of complex phylogenetic models such as the GTR substitution and BDSKY model introduces many parameters into phylogenetic studies that further complicate inferences. Although each phylogenetic method discussed previously have their own specific limitations, these discussed limitations are ubiquitous for all phylogenetic methods. In the next part of this subsection, we will discuss current advancements made in order to overcome these limitations.

Maximum parsimony based methods are generally faster than maximum likelihood and Bayesian-based phylogenetic methods [53]. This is owing to the simplicity of its optimality criterion, parsimony. In addition, software such as MPBoot has further sped up maximum parsimony-based methods by decreasing the runtime for bootstrap calculations [54]. Another software, matOptimize, which was inspired by the overwhelming number of SARS-CoV-2 sequences available, also optimized maximum parsimony-based phylogenetic

analyses through parallelization and memory-efficient data structures [55]. While maximum parsimony methods are computationally efficient, they are only suitable for simple analysis. This is because maximum parsimony methods are based on naive assumptions previously discussed, which do not hold true. Hence, maximum parsimony methods are typically unfavored when handling complex evolutionary models and accommodating site-specific rate variations such as during viral outbreaks.

Maximum likelihood based phylogenetic analysis is widely considered to be more reliable than parsimony-based methods [53, 56, 57]. As a result, several fast maximum likelihood phylogenetic programs have been developed including PhyML [58], RAxML [34], FastTree [59], IQ-TREE [60], and MAPLE. PhyML utilizes the Subtree-Pruning-and-Regrafting (SPR) and Nearest-Neighbour-Interchange (NNI) algorithms to speed up maximum likelihood searches. RAxML also uses the SPR algorithm but provides features for parallelizing maximum likelihood calculations to speed up computations [34, 61]. Meanwhile, FastTree achieves improved speeds compared to PhyML and RAxML by making use of heuristics to limit tree searches and maximum likelihood calculations in the SPR and NNI algorithms. The use of pure hill climbing algorithms such as SPR and NNI suffers from the limitation of being trapped within local maxima. Hence, IQ-tree utilizes stochastic processes by selecting random candidate trees at each iteration to perform the NNI algorithm which greatly improves computational efficiency [60]. Finally, MAPLE achieves reasonable accuracy with less memory usage and computational runtime by incorporating parsimony-based heuristics in the Felsenstein pruning algorithm [62].

Bayesian based methods are also suitable for modeling complex evolutionary dynamics and substitution models. The main advantage of Bayesian-based methods is the incorporation of prior information which cannot be achieved in parsimony-based and maximum likelihood-based methods. Several programs and approaches have been developed to speed up Bayesian phylogenetic analysis, addressing the computational challenges associated with this method. For example, some researchers have utilized variational inference to accelerate the tree-search and parameter-search process [63]. Meanwhile, other programs use computational parallelization to speed up parameter space exploration. ExaBayes utilizes nonblocking parallelization of Metropolis-coupled chains to accelerate MCMC sampling [64]. Moreover, other programs such as broad-platform evolutionary analysis general likelihood evaluator (BEAGLE) and DNAml utilize GPU parallelization to speed up posterior distribution computations [65, 66]. BEAGLE can be directly integrated into phylogenetic analysis with Bayesian Evolutionary Analysis Sampling Trees 2 (BEAST2). BEAST2 is an open-source Markov chain Monte Carlo (MCMC) based Bayesian phylogenetic software that focuses on being “useable”, “open”, and “extensible” [67, 68]. Owing to its useability,

BEAST2 is also incorporated in many dedicated pipelines such as TransCOVID which can infer COVID-19 transmission networks and predict asymptomatic individuals in an infected population [12]. However, we found that optimizations involving BEAST2 mostly utilize GPU and CPU to speed up computations, but do not parallelize MCMC sampling chains. Hence, the inspiration for our novel approach and thesis.

Chapter 2

Methodology

In this Chapter, we introduce introduce our methodology for performing phylogenetic analysis. In Section 2.1, we discuss our novel approach to parallelize MCMC sampling for Bayesian phylogenetic analysis using the software BEAST2. In Section 2.2, we discuss how the novel approach is implemented in our study and outline the overall computational procedure that was used to perform the analysis in the upcoming studies involving simulation and SARS-CoV-2 sequence data.

2.1 Novel Approach

Bayesian-based phylogenetic analysis using the MCMC algorithm is computationally expensive and requires long run times. Currently, many programs have been developed to speed up the MCMC algorithm, but most focus on the utilization of GPUs and multiple processors to improve computational times within each iteration of the MCMC algorithm [65, 66]. While this effectively accelerates each individual MCMC iteration, overall, it does not provide substantial improvements as phylogenetic analysis often requires millions of MCMC iterations to effectively explore parameter spaces [69, 70]. This is because phylogenetic analysis can involve complex evolutionary dynamics that entail multimodal and multidimensional parameter spaces [69]. Accordingly, very long MCMC chains, in the magnitudes of millions or even billions, must be sampled in order to achieve convergence and approximate the posterior distribution. Furthermore, in most cases, researchers are unsure whether the "true" posterior distribution is approximated and thus rely on greater chain lengths to achieve certainty [70]. Hence, this demand for long MCMC chains in phylogenetic studies negates the improvements brought by GPU and CPU parallelizations.

Having said that, researchers have developed methods that parallelize the inherently

serial MCMC sampling algorithm [71–73]. This process involves partitioning the original parameter space into smaller subspaces, then independently sampling from each subspace [71]. By doing so, multiple independent MCMC chains corresponding to each subspace are generated. Accordingly, combining each of these independent MCMC chains would provide an approximation of the entire parameter space [71, 72]. Thus, such parallelization divides the serial task of MCMC approximation of the posterior distribution into multiple smaller parallel problems that can be solved with ease. Though this strategy is trivial for simple parameter spaces, it is especially effective for large and complex parameter spaces. Hence, we believe this strategy would be particularly useful for phylogenetic analysis. Currently, the program ExaBayes implements both iteration-level and chain-level parallelization for MCMC algorithms in Bayesian-based phylogenetic analysis which drastically reduces run-times and computational demand [64]. However, such advancements are not available in the software BEAST2, which is more popular due to its ease of use, well-documentation, framework flexibility, and tools for model comparison [64, 67]. Most importantly, BEAST2 features complex evolutionary models such as the BDSKY model which is particularly useful for viral outbreaks [39].

Therefore, in this thesis, we implemented a novel methodology to perform chain-level parallelization for the MCMC algorithm in Bayesian-based phylogenetic studies using BEAST2. In our approach, we sample multiple asynchronous and independent MCMC chains from our parameter space. This differs from other parallelizations in BEAST2 which specify starting points along a chain of stored MCMC states [74]. Similar to the algorithms proposed previously, this partitions the phylogenetic parameter space into smaller and more manageable subspaces [71]. Since the MCMC chains sampled are independent and asynchronous, in other words the parallel chains do not exchange information, this method can also leverage distributed computing. After parallel MCMC chains are acquired, these independent MCMC chains are directly combined with the Logcombiner program that is native to the BEAST2 software. This ensures usability as additional steps are not required from the user and users remain in the BEAST2 environment. In our method, we also utilize the built-in feature of BEAGLE in the software BEAST2 to perform iteration-level parallelization with GPU [66, 67]. Ultimately, the integration of both chain-level and iteration-level parallelization of the MCMC algorithm greatly reduces computational demand and run times. The improvements in run-time from chain-level parallelization directly corresponded with the number of chains parallelized.

Table 2.1: Parameters specified in BEAUTi when performing Bayesian phylogenetic analysis using the BEAST2 software.

Model Specified	Configuration
Generalized Time Reversible site model	Gamma category count = 4; Default parameters
Relaxed log-normal clock model	Default parameters
Birth Death Skyline Serial tree prior	Default parameters

2.2 Phylogenetic Procedure

The overall computational procedure used in this thesis was inspired by the TransCOVID pipeline which incorporates the software BEAST2 COVID-19 [12]. The TransCOVID pipeline can infer COVID-19 transmission networks and predict asymptomatic individuals in an infected population [12]. Hence, the parameterizations specified in the dedicated pipeline serve as an excellent starting point for the purposes of analyzing viral outbreaks using BEAST2. The computational procedure employed in this study can be seen in Figure 2.1 and will be discussed in greater detail. In Figure 2.1, the computational procedure involving both parallelized MCMC chains and traditional serial MCMC chains are illustrated along with the run times associated with each step in the procedure.

Upon retrieving samples of the molecular sequence data, **multiple sequence alignment** was performed on all sequence subsamples using MAFFT [17]. Next, BEAUTi from the BEAST2 software package was used to set up the parameters used in the phylogenetic analyses [67]. In the analysis, tip dates were activated which specify the collection date of each sequence or tip in the tree. In terms of parameterization, the GTR site model with a gamma category count of 4, the relaxed log-normal clock model, and the Birth Death Skyline Serial models were employed as priors in the parallelized tool (Table 2.1). Default settings in BEAST 2.7.4 were used for each prior unless otherwise specified. These phylogenetic models were employed as they align with the evolutionary dynamics of viral outbreaks such as SARS-CoV-2 and HIV. Note that in order to duplicate XML configuration files from BEAUTi to reduce set up time for parallelized runs, file names for the tracelog and treelog in the BEAUTi configuration file were set as "$(filebase).log" and "$(filebase).tree" respectively.

In the computational procedure, Bayesian phylogenetic analysis using the MCMC algorithm was performed using BEAST2. For each sample, parallelized phylogenetic analyses were ran with MCMC chain lengths of 100 million for 29 separate runs, while sequential runs were run for MCMC chains of at least 2.9 billion. These chain lengths were selected to achieve ESS of at least 200 for parameters sampled from MCMC which is recommended by

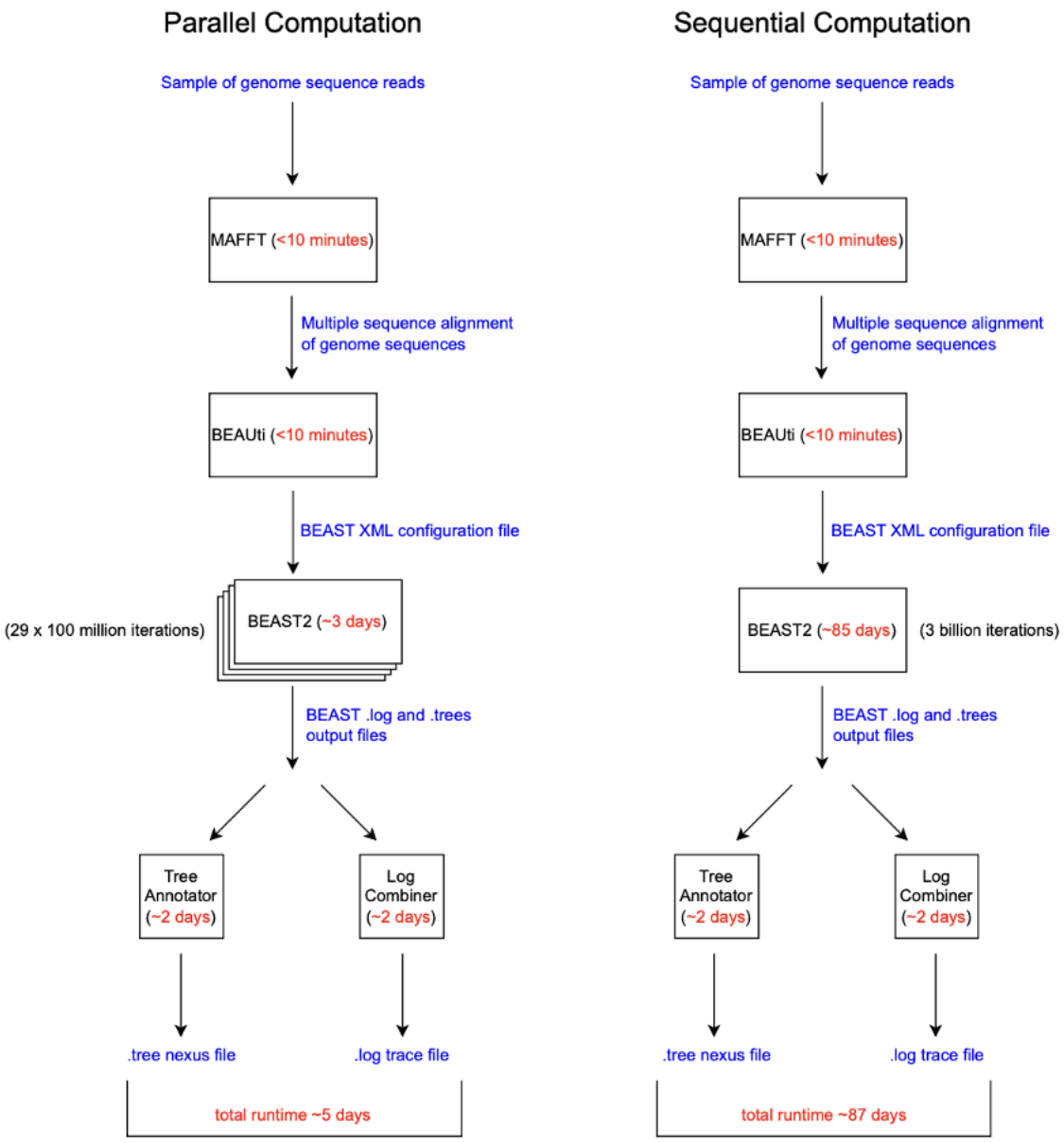


Figure 2.1: Overview of the Bayesian phylogenetic analyses performed in the study. Differences in methodology between MCMC chains ran in parallel and sequentially are shown along with the corresponding run times for each step in the procedure.

BEAST2 [67]. Meanwhile, both sequential and parallel MCMC chains were executed with the exact same hardware specifications (i.e. memory, central processing unit). In the proce-

dure, independent MCMC chains obtained from the parallelized MCMC runs were combined afterward using the built-in program LogCombiner. All corresponding tree and log output files were then resampled with LogCombiner from BEAST2 with a resample frequency of 5000. Finally, maximum clade likelihood phylogenetic trees were extracted with TreeAnnotator from BEAST2 using common ancestor node heights and a 10% burn-in. Finally, trace file outputs from MCMC phylogenetic analyses were summarized using LogAnalyser from BEAST2 and viewed in Tracer [67, 75]. NEXUS trees obtained from TreeAnnotator were also visualized and converted into the Newick tree file format using the software FigTree with "save all trees" selected [76]. More details regarding the parallelized tool can be found at this github. Overall, we found that the implementation of the parallelized MCMC chains reduced experiment times by more than 10 times. This drastic reduction (>28 times) in run-time was observed because of rapid improvements caused by the parallelization of the MCMC algorithm; MCMC computational times were reduced from 85 days to 3 days using independent parallelized chains.

Chapter 3

Simulations and Real Data Analysis of HIV and SARS-CoV-2

In this Chapter, we present the application of the computational procedure introduced in Chapter 3 on two simulation HIV datasets and one SARS-CoV-2 dataset. Particularly, we compare and contrast the results obtained from parallel and serial MCMC chains (Figure 2.1). In Section 3.1, we provide an introduction to new tools and concepts used specifically in these studies. In Section 3.2, we discuss specific data collection and phylogenetic model comparison procedures. In Section 3.3, parameter predictions and tree topology results from each study were reported and analyzed.

3.1 Introduction

Bayesian phylogenetic analyses using the MCMC algorithm were performed in these study with the software BEAST2 and modified TransCOVID pipeline as discussed in Section 2.2 [12, 67]. In these studies, MCMC computations from parallel computations were compared to MCMC sampling that was performed sequentially. The purpose of these studies was to evaluate the validity of the novel MCMC parallelization approach introduced in Section 2.1 on both simulated data and real-world examples. MCMC computations were parallelized by asynchronously running multiple smaller independent MCMC chains rather than one continuous long MCMC chain. In the parallel computations, 29 separate computations of 100 million MCMC chain lengths were performed whereas in the sequential computation, runs were performed with MCMC chain lengths of at least 2.9 billion. In the studies, these chain lengths were selected to achieve an ESS of at least 200 for parameters sampled from MCMC. Three different datasets were applied with our experimental design; this included SARS-CoV-2 sequence data from the early pandemic, HIV simulation data with perfect sampling

(100% sampling rate of population), and HIV simulation data with imperfect sampling (10% sampling rate of population).

In this thesis, we conducted two simulation studies involving HIV data to evaluate our implemented method. This is because simulation studies have become a standard for assessing the reliability of newly proposed methods. In simulated phylogenetic studies, the ground truth is known and parameters within the phylogenetic models are easily controlled, allowing us to gain additional insight on our implemented methods. Specific programs are also available for simulating viral outbreaks such as The software Framework for Viral Transmission and Evolution Simulation (FAVITES). FAVITES allows for the concurrent simulation of both a viral phylogenetic tree as well as viral evolution [77]. FAVITES is superior to other simulation programs as it simulates the full end-to-end epidemic dataset including incomplete sampling, viral phylogeny, and transmission histories [77]. Furthermore, at each step within the simulation, several different phylogenetic models are available to choose from in FAVITES. Relating to the design of our thesis, this allows us to simulate epidemic data resembling our knowledge of real-world viral outbreaks. Since our novel parallelization method was inspired by the COVID-19 pandemic, we simulated one dataset with a 10% sampling rate. This sampling rate was selected as it aligned with studies from the United Kingdom (UK) reporting similar sequencing rates for COVID-19 in the UK, which hosts one of the best sequence sampling programs worldwide [78, 79]. Moreover, as a result of the design of the simulation experiments, tree topologies and parameter estimates proposed from MCMC methods were also easily compared with the ground truth defined in the simulation studies. In particular, tree topologies were compared using various rooted and unrooted distance metrics previously discussed in Section 1.3 with the software TreeCMP [45].

3.2 Details of data collection and analysis

In the phylogenetic studies, the computational procedure discussed in Section 2.2 was employed with no modifications unless otherwise specified. As a result, only processes that were distinct from these phylogenetic studies will be explained. The Data collection subsection describes the process that was performed to obtain the three different datasets that were analyzed using the parallelized tool. The Data analysis subsection describes the statistical tests that were applied to outputs from the computational procedure to assess the validity of our parallelized MCMC chain approach.

3.2.1 Data collection

Simulation data:

HIV1 B-DNA-POL-LITTLE simulation data was generated via the software FAVITES using the configuration file `HIV_FAVITES_ART_0.125x` found on FAVITES with minor modifications. The configuration file of the simulation data can be found in the Appendix. Notably, the phylogenetic configuration utilizes a GTR gamma sequence evolution model and a "VirusTreeSimulator" node evolution model. For the GTR model, the parameterization is shown in Table 3.1.

Table 3.1: Parameters defined in the configuration for FAVITES when simulating HIV datasets with perfect (100%) and imperfect sampling (10%)

Dataset	rateAC	rateAG	rateAT	rateCG	rateGT	gammaShape
HIV with perfect sampling	0.182382	1	0.072316	0.097792	0.100663	2
HIV with 10% sampling rate	0.182382	1	0.072316	0.097792	0.100663	2

In the first set of simulation data, 10 samples of 1000 FASTA sequences along with their corresponding phylogenetic trees were generated with perfect sequence sampling. In the second set of simulation data, 10 samples of 1000 FASTA sequences along with their corresponding phylogenetic trees were generated with 10% sequence sampling.

SARS-CoV-2 data:

A total of six 1000 FASTA sequence subsamples were obtained from GISAID's SARS-CoV-2 database using random and weighted sampling on SARS-CoV-2 sequences between the dates of February 1st, 2020, to October 31st, 2020 that were available on GISAID as of March 1st, 2021. Random sampling was performed using the Software Augur (augur filter) from NextStrain on the FASTA files and associated metadata obtained from GISAID on March 1st, 2021 [80]. The following parameters were used:

```
--min-date 2020-02-01 --max-date 2020-10-31 -subsample-max-sequences 1000
```

Weighted sampling was conducted in order to minimize sampling bias caused by differential sequencing rates and SARS-CoV-2 prevalence in different regions around the world in different time frames [81]. The following steps were performed to determine the number of desired sequences to be sampled from each region and month in the weighted sampling strategy. First, the number of sequences to be obtained from each month globally (monthly

sequencing allowance) was determined based on the proportion of SARS-CoV-2 infections that occurred in each month out of all months. Next, the number of sequences to be obtained from each country in each month was determined based on the proportion of SARS-CoV-2 infections that occurred in each country out of all countries during each respective month. The proportion of infections in each country was then multiplied by the monthly sequencing allowance to ensure that the number of sequences obtained in each month and country was proportional to their respective infection proportions. In total, 25,000 sequences were obtained. The number of SARS-CoV-2 infections during each month and in each country was estimated using the IHME epidemiological model [82]. Using the following strategy, a CSV file containing the number of sequences to be obtained from each region and month was generated and fed into the software Nybbler which allows this weighted sampling strategy to be performed on sequencing data and its associated metadata [83]. Ultimately, 3 datasets of 1000 SARS-CoV-2 sequences were obtained from weighted sampling, and 3 datasets of 1000 SARS-CoV-2 sequences were obtained from random sampling.

3.2.2 Data analysis

Tree analysis:

Phylogenetic trees generated from the computational procedure were investigated using distance metrics for unrooted phylogenetic trees. Distance metrics for the phylogenetic trees were calculated by using the software TreeCMP with the following arguments [45]:

```
-d qt pd rf ms um rfw gdu
```

In TreeCMP, the unrooted distance metrics between two trees were calculated. Accordingly, distance metrics were computed for the following comparisons:

1. Simulated HIV samples: comparison between parallel or sequential phylogenetic trees with FAVITES simulation tree as the ground-truth reference trees,
2. Simulated HIV samples: comparison between parallel phylogenetic trees with sequential phylogenetic trees as the reference trees, and
3. SARS-CoV-2 samples: comparison between the parallel phylogenetic trees with sequential phylogenetic trees as the reference trees.

Parameter estimate analysis:

Parameter estimates provided by the phylogenetic computational procedure and computed distance metrics were analyzed. To test for statistical significance between different treatments (i.e. parallel, sequential), the Mann-Whitney U-test was performed on the distance metrics and MCMC parameter estimates. The U-test was used to compare the following:

- Parameter estimates obtained from MCMC ran in parallel vs MCMC ran sequentially,
- Distance metrics obtained from the comparison of sequential vs "true" MCC trees, and parallel vs "true" MCC trees in the analyses involving simulated data, and
- Distance metrics obtained from the comparison of sequential vs parallel MCC trees in all analyses.

The two-tailed student's t-test was also performed to compare the parameters from MCMC with "true" values in the analyses involving simulated data.

All boxplots and violin plots shown in the figures were created in R with the package "ggplot2" and modified using Inkscape [84, 85]. Phylogenetic trees were visualized using FigTree and MCMC trace plots were viewed using Tracer [75, 76].

3.3 Results

To validate the results from our novel parallelization method, we examined parameters sampled from MCMC as well as maximum clade credibility phylogenetic trees constructed through MCMC. In particular, the following parameters were inspected: substitution rates from A to C (rateAC), G to T (rateGT), C to G (rateCG), A to T (rateAT), A to G (rateAG), and gamma shape (gammaShape). These parameters were selected as their 'true' values were known and defined in the configuration files of each simulation sample (Table 3.1). Thus, providing us with a convenient method to assess the performance of our MCMC sampling algorithms, and allowing us to gain additional insight regarding our implemented method. These parameters were also compared in the study involving SARS-CoV-2 data for consistency. However, the ground truth for SARS-CoV-2 remains unknown and therefore parameter estimates were compared with other scientific literature.

Maximum clade credibility (MCC) trees obtained from MCMC chains performed in parallel and sequentially were compared with the predefined phylogenetic trees specified in the FAVITES simulations. MCC trees obtained from MCMC chains performed in parallel and

sequentially were also compared with each other. Differences in phylogenetic trees when visualized were difficult to contrast. Hence, differences in phylogenetic trees were quantified with various unrooted similarity metrics including RF distance, RFWeighted distance, UMAST distance, MatchingSplit distance, GeoUnrooted distance, and Path difference distance.

3.3.1 HIV dataset with perfect sampling

In the first HIV dataset, perfect sampling was specified in the FAVITES software. The purpose of this was to mimic situations where all infected samples of a population are reported and sequenced. Although this does not occur in the real world, it's important for understanding how well the parallelized MCMC algorithm performs in ideal circumstances. A total of 10 different phylogenetic analyses were performed on the 10 different samples. In each analysis, two treatments were specified. The first involved MCMC chains ran serially, and the latter involved MCMC chains ran in parallel.

Table 3.2: Parameter estimates obtained from MCMC phylogenetic analyses ran in parallel and sequentially on simulated HIV data with perfect sampling rate. Ground-truth parameter values specified in the configuration of the simulation are also shown for reference.

Group	rateAC	rateAG	rateAT	rateCG	rateGT	gammaShape
Parallel #1	0.190592	1.134454	0.077401	0.100687	0.118022	1.578291
Sequential #1	0.191285	1.14102	0.077605	0.100965	0.118341	1.580005
Parallel #2	0.201034	1.190655	0.078193	0.110933	0.108724	2.513343
Sequential #2	0.20091	1.190402	0.078139	0.110801	0.108773	2.518466
Parallel #3	0.195524	1.175345	0.080087	0.095341	0.11052	1.488687
Sequential #3	0.195478	1.180296	0.080277	0.095812	0.111228	1.492906
Parallel #4	0.19303	1.116982	0.080121	0.103838	0.10181	1.363009
Sequential #4	0.193384	1.120493	0.080211	0.104038	0.102009	1.363877
Parallel #5	0.197616	0.961815	0.072424	0.094886	0.088884	1.754129
Sequential #5	0.197285	0.961559	0.072436	0.094982	0.088946	1.755635
Parallel #6	0.200705	1.181526	0.076423	0.112712	0.109088	1.955163
Sequential #6	0.200735	1.182132	0.076427	0.11275	0.109121	1.959853
Parallel #7	0.212087	1.35096	0.083232	0.10869	0.111258	1.45266
Sequential #7	0.212125	1.348053	0.082975	0.108458	0.111034	1.454552
Parallel #8	0.189194	1.049234	0.075652	0.099794	0.103445	2.548589
Sequential #8	0.189392	1.050123	0.075552	0.100004	0.10351	2.557024

Parallel #9	0.20731	1.412333	0.092544	0.108753	0.123644	1.332105
Sequential #9	0.207681	1.41019	0.092237	0.1089	0.123455	1.335203
Parallel #10	0.20175	1.180115	0.076844	0.109024	0.102437	1.824798
Sequential #10	0.201268	1.179504	0.076672	0.109126	0.102475	1.830874
Reference	**0.182382**	**1**	**0.072316**	**0.097792**	**0.100663**	**2**

Parameter estimates from phylogenetic analyses in all ten samples from both treatments were summarized in Table 3.2. In Figure 3.1, parameter estimates from MCMC performed sequentially and in parallel from all ten trials were compared with boxplots to illustrate differences in parameter estimates. To test for statistical significance, the Mann-Whitney U-test was used to compare parameter estimates from the two distinct groups of MCMC runs; the simulation datasets ($n = 10$) sampled with parallel MCMC chains, and the simulation datasets ($n = 10$) sampled with sequential MCMC chains. The student's t-test was also utilized to compare the samples of MCMC parameter estimates from the parallel and sequential MCMC runs, respectively, with the parameter values predefined in the simulated data.

From the U-test, it was concluded that all parameters obtained from the sequential and parallel MCMC runs were not significantly different. When comparing the parameters obtained from MCMC with the "true" values via the t-test however, we found that the rateAC values obtained from the parallel (p-value = 0.010) and sequential (p-value = 0.011) MCMC runs differed significantly Figure 3.1. The "true" value for rateAC was 0.182382, which is less than the minimum rateAC values obtained from all trials in both the sequential and parallel treatments.

Within each of the ten different parallel MCMC trials, 29 different MCMC chains of 100 million iterations were produced. Parameter estimates were directly obtained from each of these 100 million iteration MCMC chains. These estimates were compiled for the first MCMC trial involving HIV data with perfect sampling and compared with the parameters obtained from the combined parallelized MCMC chains and sequential MCMC chain for the first trial (Figure 3.2). In all violin plots, estimates from combined parallel MCMC chains fell near the center/median of the violin plot distributions. Estimates from sequential MCMC chains were also encompassed within the distribution of the parallelized chains but did not align with the median of the distribution. In all cases, the estimates from the sequential MCMC runs were greater than those of the estimates from the combined parallel runs. Finally, in

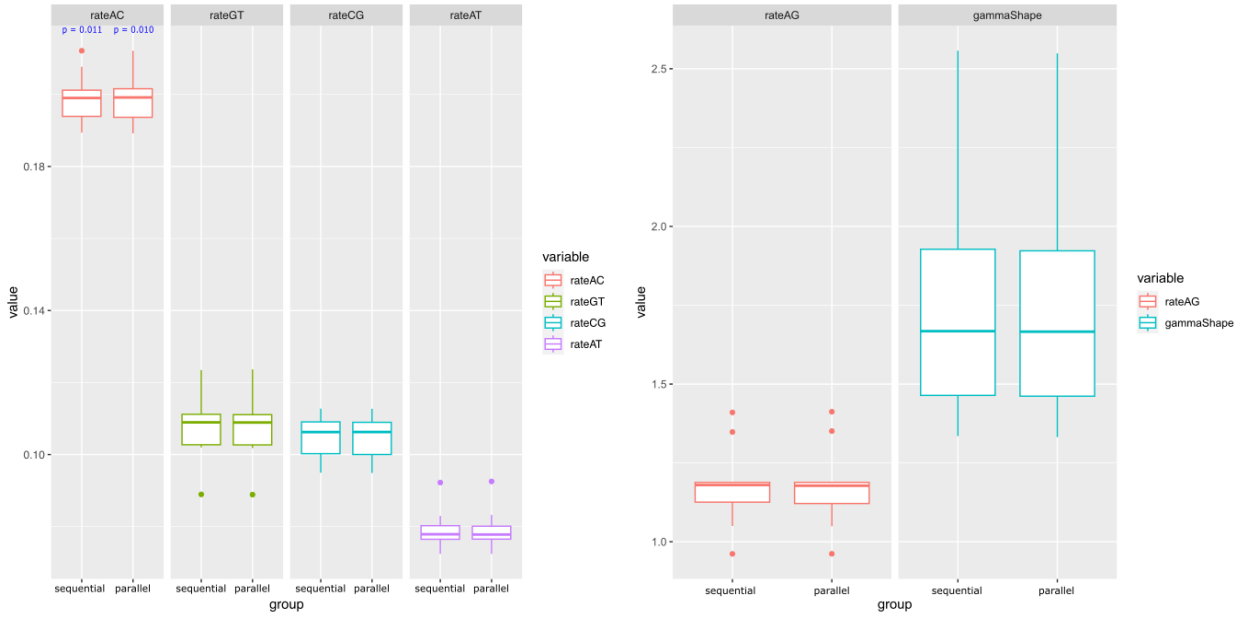


Figure 3.1: Boxplots from ten parameter estimates from MCMC phylogenetic analyses ran in parallel and sequentially on simulated HIV data with perfect sampling rate. Significant p-values from the U-test (as a line) and t-test (directly above) are labeled. Figures with different vertical axis scaling were used due to differences in the ranges of values.

all violin plots, the parallelized parameter distributions did not agree with the ground-truth values specified for each parameter estimate.

MCC trees obtained from the parallel and sequential MCMC runs were compared visually with the ground-truth phylogenetic tree in Figure 3.3. The differences in the topology of the trees were observed but were hard to quantify. Hence, MCC trees obtained from the parallel and sequential MCMC runs were compared with the "true" phylogenetic trees simulated in FAVITES with various distance metrics (Figure 3.4). The Mann-Whitney U-test was performed to compare the distance metrics calculated from the MCC trees obtained from the parallel computations versus the predefined tree, and the distance metrics calculated from the MCC trees obtained from the sequential computations versus the predefined tree. The results were presented in Figure 3.4. A near significant difference (p-value = 0.052) was observed in the comparison of the RF-weighted distances via the U-test, but not in the RF distances.

Finally, MCMC trace plots for rateAC of the analyses involving parallel and sequential MCMC chains for the first sample of the HIV dataset with perfect sampling were plotted

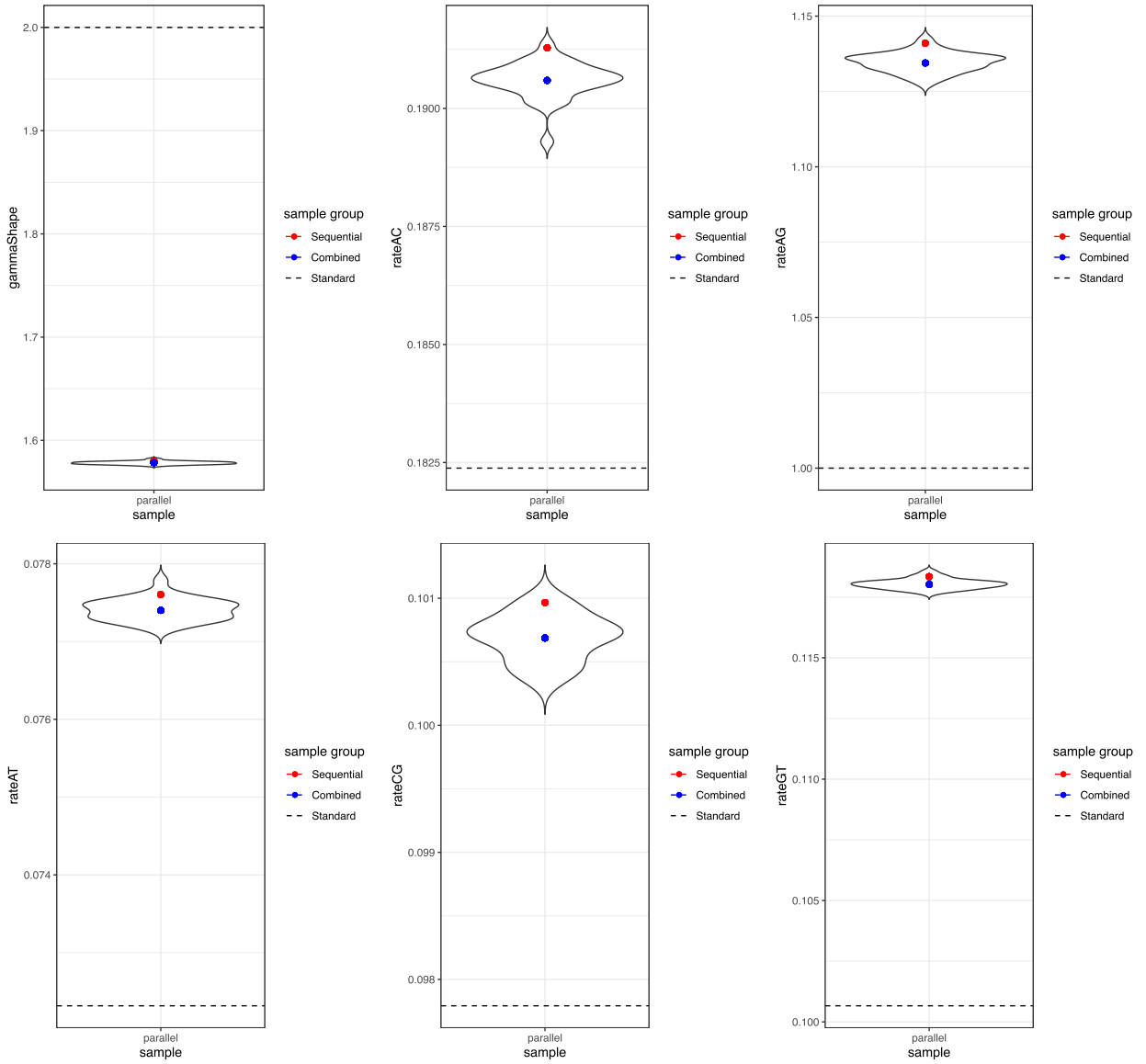


Figure 3.2: Violin plots from parameter estimates from twenty-nine independent MCMC chains (100-million iterations) from the phylogenetic analysis performed the first replicate of sequences in the HIV dataset perfect sampling rate. The parameter estimates for all 29 MCMC chains combined ("combined") and parameter estimates from the MCMC run sequentially ("sequential") on the same dataset are also shown. Horizontal dashed lines represent the true value (ground truth) for each parameter. Figures with different vertical axis scaling were used due to differences in the ranges of values.

in Figure 3.5. Trace plots for both treatments were fairly similar. The dissimilarities in the lengths of the plots for the parallel and sequential chains were attributed to the difference in the number of iterations that were performed for each treatment.

(a) Parallel (b) Sequential (c) Ground-truth

Figure 3.3: Phylogenetic trees obtained from MCMC phylogenetic analyses ran in parallel (a) and sequentially (b) on sample #1 of simulated HIV data with perfect sampling rate. Ground-truth trees are shown as reference (c).

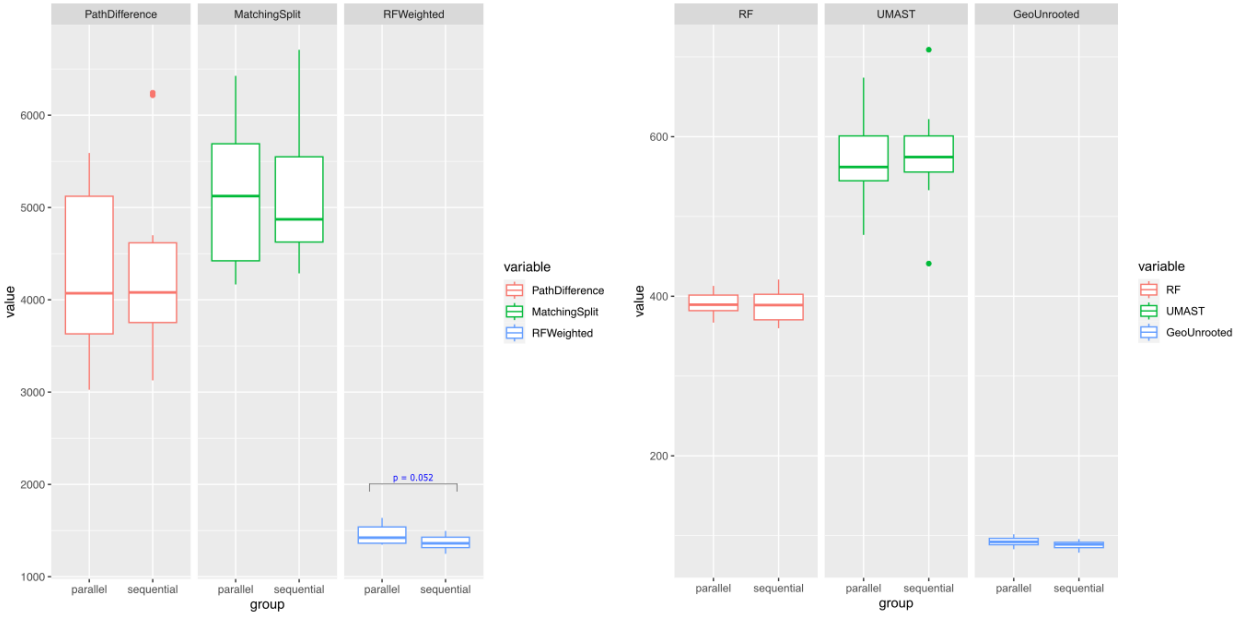


Figure 3.4: Boxplots from ten distance metrics calculated by comparing MCC phylogenetic trees obtained from MCMC with "true" trees defined in the simulated HIV data with perfect sampling rate. Significant p-values from the U-test comparing the distances from the sequential and parallel samples are labeled. Figures with different vertical axis scaling were used due to differences in the ranges of values.

3.3.2 HIV dataset with imperfect sampling

In the second HIV dataset, a 10% sampling rate was specified in the FAVITES software. The purpose of this was to mimic situations where only a proportion of infected individuals

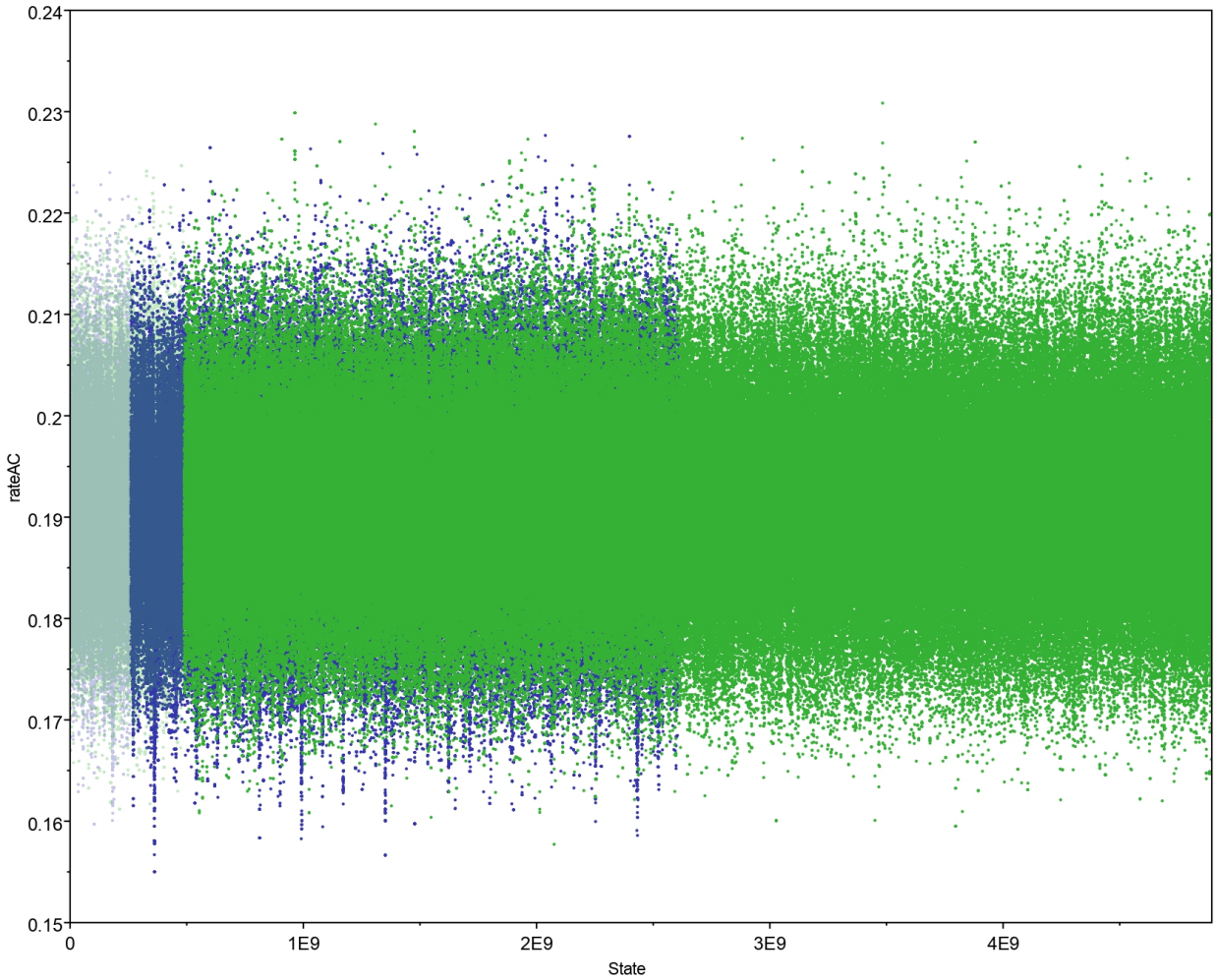


Figure 3.5: rateAC MCMC traces obtained from MCMC phylogenetic analyses ran in parallel (blue) and sequentially (green) on sample #1 in simulated HIV data with perfect sampling rate.

are reported and sequenced in a population such as during the COVID-19 pandemic. A total of 10 different phylogenetic analyses were performed. In each analysis, two treatments were specified. The first involved MCMC ran sequentially, and the latter involved MCMC runs in parallel.

Parameter estimates from phylogenetic analyses in all ten samples from both treatments were summarized in Table 3.2. Meanwhile in Figure 3.6, parameter estimates from MCMC performed sequentially and in parallel from all ten trials were compared with boxplots. Similar to Subsection 3.3.1, the U-test and t-test were employed to test for statistical significance of the parameter estimates. All parameters obtained from the sequential and parallel MCMC runs were not significantly different according to the U-test. When comparing the param-

eters obtained from MCMC with the "truth" values via the t-test, we found that only the rateAT values obtained from the parallel (p-value = 0.029) and sequential (p-value = 0.032) MCMC runs differed significantly (Figure 3.6). The "true" value for rateAC was 0.072316.

Table 3.3: Parameter estimates obtained from MCMC phylogenetic analyses ran in parallel and sequentially on simulated HIV data with a 10% sampling rate. Ground-truth parameter values specified in the configuration of the simulation are also shown for reference.

Group	**rateAC**	**rateAG**	**rateAT**	**rateCG**	**rateGT**	**gammaShape**
Parallel #1	0.175669	1.005885	0.081875	0.087289	0.104952	2.038026
Sequential #1	0.17604	1.009665	0.081671	0.087774	0.105268	2.044189
Parallel #2	0.186381	1.144597	0.085848	0.0936	0.114723	1.468562
Sequential #2	0.185692	1.13994	0.085598	0.093791	0.114896	1.476175
Parallel #3	0.183636	1.034714	0.074356	0.098208	0.101928	1.668151
Sequential #3	0.182851	1.03363	0.073985	0.098408	0.102122	1.673062
Parallel #4	0.183326	1.132486	0.088278	0.096867	0.113298	2.045434
Sequential #4	0.183301	1.132246	0.088118	0.097036	0.113384	2.053384
Parallel #5	0.207762	1.297526	0.089648	0.101182	0.117328	1.298873
Sequential #5	0.208471	1.300504	0.089141	0.101505	0.11711	1.302002
Parallel #6	0.194458	1.299333	0.085355	0.107477	0.114207	1.421041
Sequential #6	0.195227	1.300556	0.085155	0.10788	0.113943	1.426607
Parallel #7	0.209818	1.19147	0.082522	0.09799	0.101082	1.31933
Sequential #7	0.209575	1.188851	0.081834	0.098154	0.100785	1.365843
Parallel #8	0.2204	1.449389	0.086725	0.114677	0.118455	1.420865
Sequential #8	0.220662	1.44755	0.086753	0.114494	0.117876	1.422999
Parallel #9	0.19009	1.075906	0.080872	0.093892	0.111917	2.577868
Sequential #9	0.189655	1.07679	0.080718	0.093927	0.112128	2.590755
Parallel #10	0.181555	0.884891	0.073055	0.084083	0.089966	1.455722
Sequential #10	0.189655	1.07679	0.080718	0.093927	0.112128	2.590755
Reference	**0.182382**	**1**	**0.072316**	**0.097792**	**0.100663**	**2**

Like in Subsection 3.3.1, parameter estimates from individual parallelized MCMC chains prior to combination in sample 1 were directly obtained. These estimates were compiled and compared with the parameter estimates obtained from the combined parallelized MCMC chain and sequential MCMC chain (Figure 3.5). In all violin plots, estimates from the

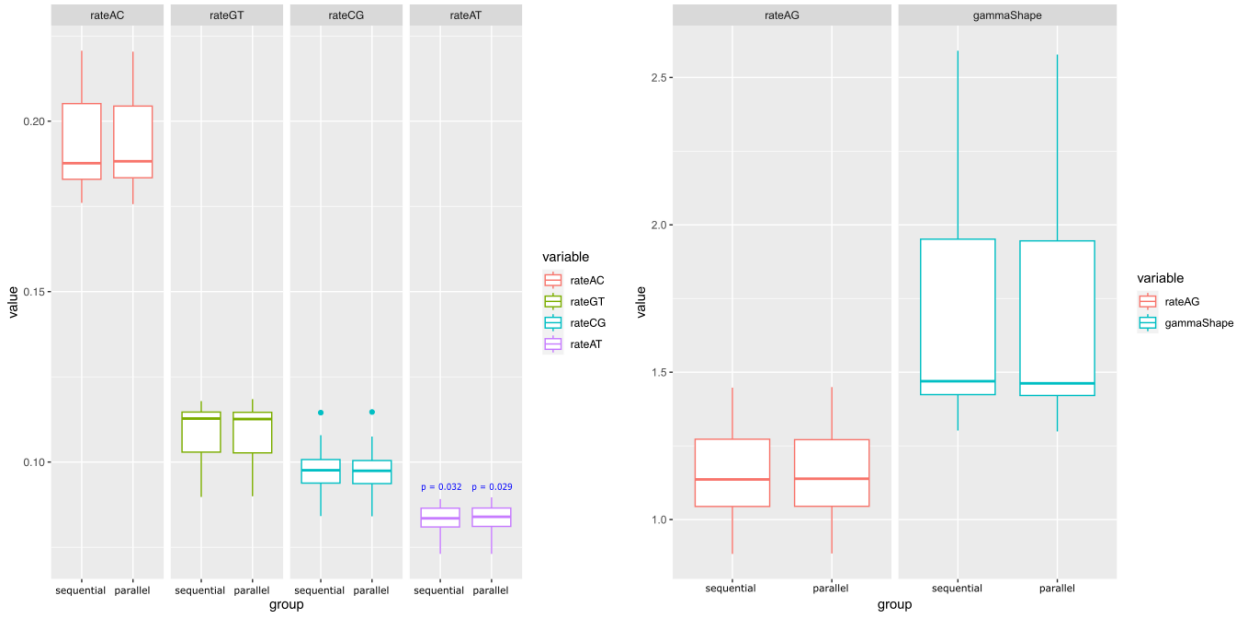


Figure 3.6: Boxplots from parameter estimates from MCMC phylogenetic analyses ran in parallel and sequentially on simulated HIV data with a 10% sampling rate. Significant p-values from the U-test (as a line) and t-test (directly above) are labeled. Figures with different vertical axis scaling were used due to differences in the ranges of values.

combined parallel MCMC chain were located near the center/median of the violin plot distributions. Most estimates from sequential MCMC chains were also encompassed within the distribution but were not situated near the median of the distribution. The gamma shape and rateCG estimates from the sequential MCMC chains, however, fell outside distributions produced by the independent parallelized chains. In all cases except rateAT, the estimates from the sequential MCMC runs were greater than that of the estimates from the combined parallel runs. Finally, in all violin plots, the parallelized parameter distributions did not agree with the ground-truth values specified for each parameter estimate.

Again, similar to Subsection 3.3.1, MCC trees obtained from the parallel and sequential MCMC runs were compared visually with the ground-truth phylogenetic tree in Figure 3.8. Differences in the topology of the trees were observed but hard to quantify. Hence, MCC trees obtained from the parallel and sequential MCMC runs were compared with the "ground-truth" phylogenetic trees simulated in FAVITES using various distance metrics. The Mann-Whitney U-test was also performed to compare the distance metrics calculated from the MCC trees obtained from the parallel computations versus the predefined tree, and MCC trees obtained from the sequential computations versus the predefined tree. The results were presented in Figure 3.9. A significant difference was observed in the comparison of the

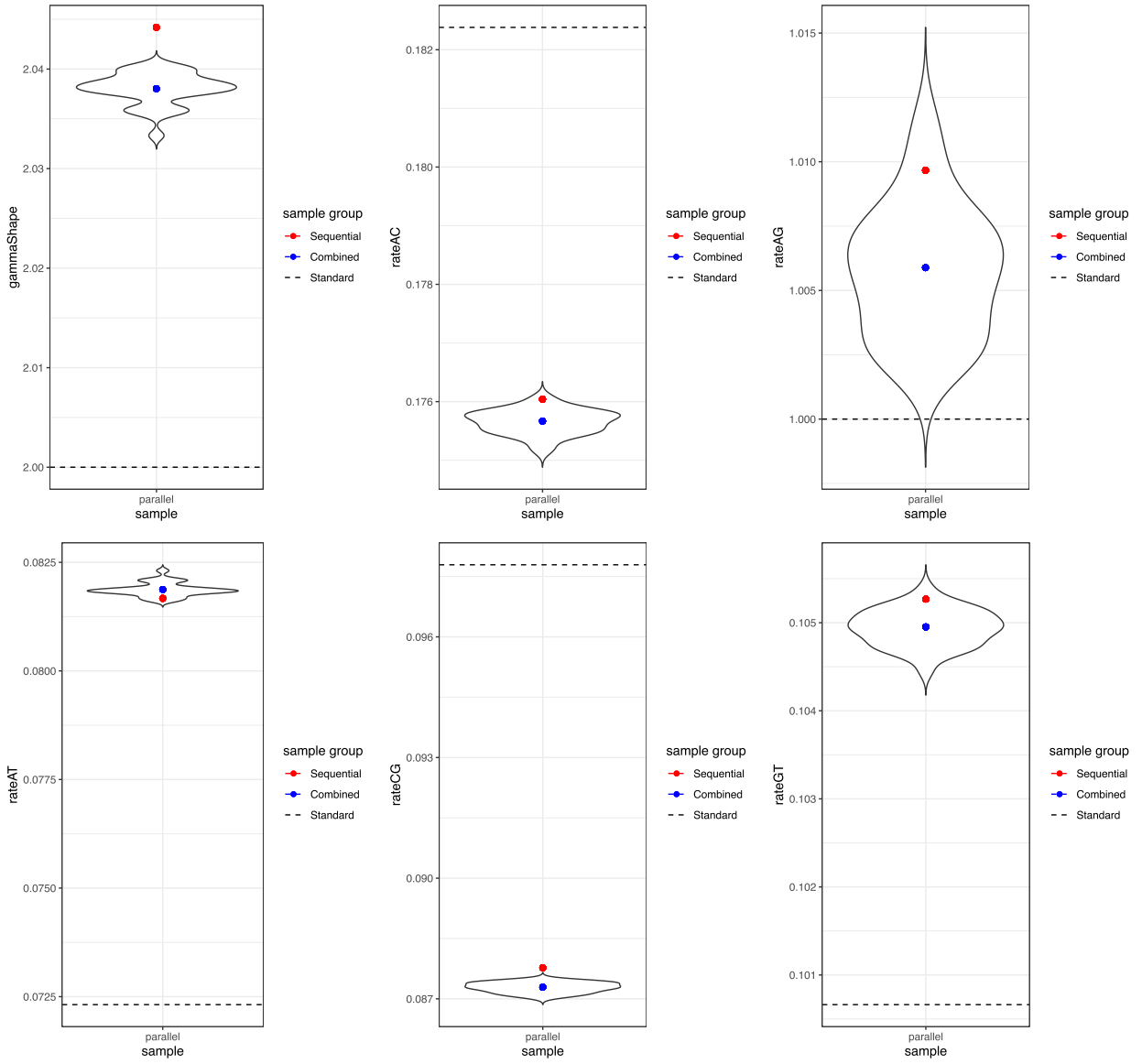
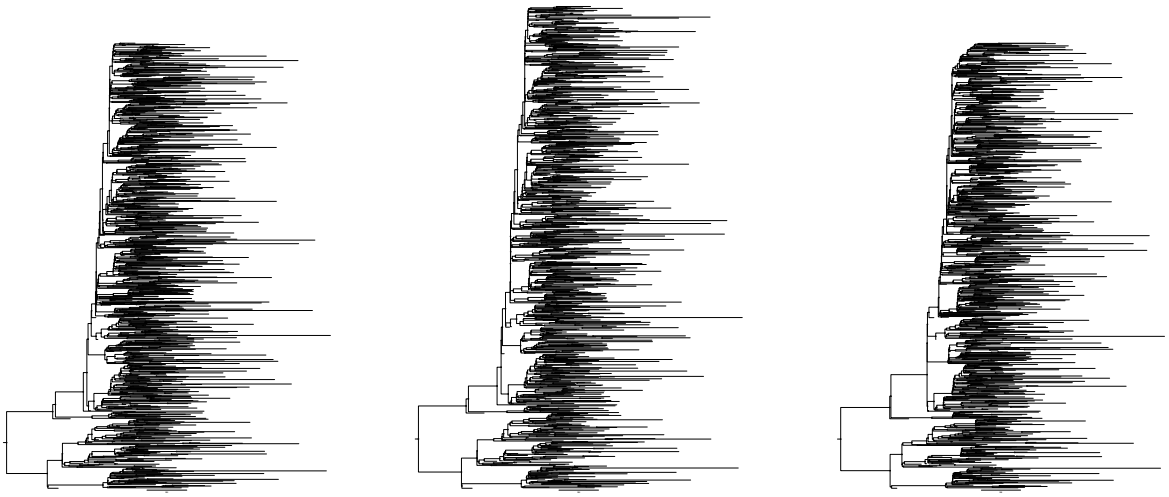


Figure 3.7: Violin plots of parameter estimates from twenty-nine independent MCMC chains (100-million iterations) from the phylogenetic analysis performed the first replicate of sequences in the HIV dataset 10% sampling rate. The parameter estimates for all 29 MCMC chains combined ("combined") and parameter estimates from the MCMC ran sequentially ("sequential") on the same data are also shown. Horizontal dashed lines represent the true value (ground truth) for each parameter. Figures with different vertical axis scaling were used due to differences in the ranges of values.

RF-weighted (p-value = 1.08E-5) and GeoUnrooted (p-value = 2.17E-5) distances, but not in the RF distances.

Finally, MCMC trace plots for rateAC of the analyses involving parallel and sequential



(a) Parallel (b) Sequential (c) Ground-truth

Figure 3.8: Phylogenetic trees obtained from MCMC phylogenetic analyses ran in parallel (a) and sequentially (b) on sample #1 of simulated HIV data with 10% sampling rate. Ground-truth trees are shown as reference (c).

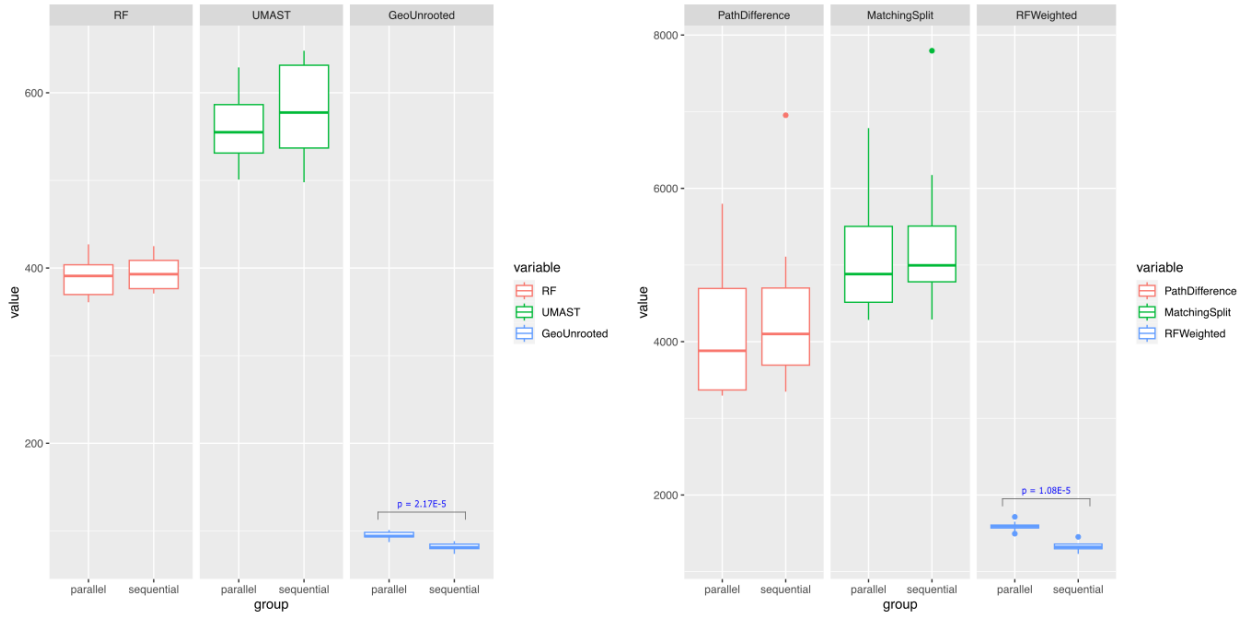


Figure 3.9: Boxplots from ten distance metrics calculated by comparing MCC phylogenetic trees obtained from MCMC with "true" trees defined in the simulated HIV data with a 10% sampling rate. Significant p-values from the U-test comparing the distances from the sequential and parallel samples are labeled. Figures with different vertical axis scaling were used due to differences in the ranges of values.

MCMC chains for the first sample of the HIV dataset with imperfect sampling were plotted in Figure 3.10. Trace plots for both treatments were fairly similar. The dissimilarities in the

lengths of the plots for the parallel and sequential chains were attributed to the difference in the number of iterations that were performed for each treatment.

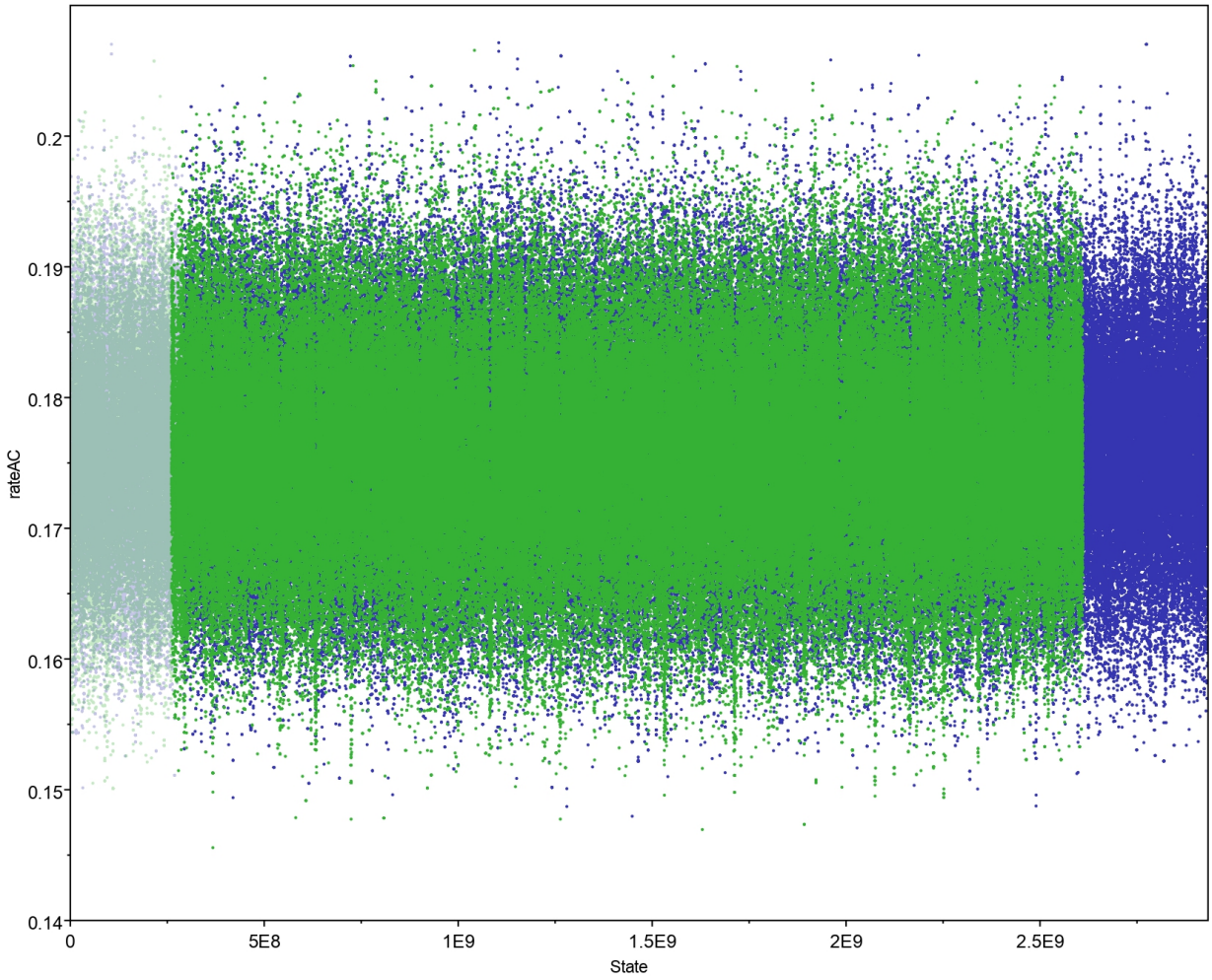


Figure 3.10: rateAC MCMC traces obtained from MCMC phylogenetic analyses ran in parallel (green) and sequentially (blue) on sample #1 in simulated HIV data with 10% sampling rate.

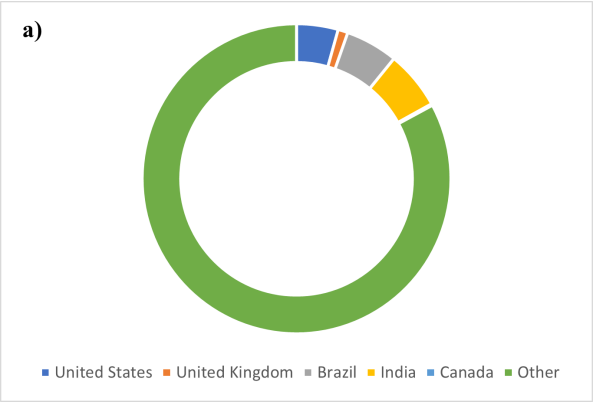
3.3.3 SARS-CoV-2 dataset

In the final dataset with SARS-CoV-2 data, one thousand sequences were sampled from all sequences available on GISAID. The purpose of this was to apply our implemented methodology to real-world data that contain complexities that aren't found in simulated data. A total of six different phylogenetic analyses were performed. In each analysis, two treatments were specified. The first involved MCMC ran sequentially, and the latter involved parallelized MCMC runs. In each treatment, three samples were obtained from weighted sampling, and three samples were obtained from random sampling. Differences in population representation from the weighted and random sampling strategies were illustrated in Figure 3.11.

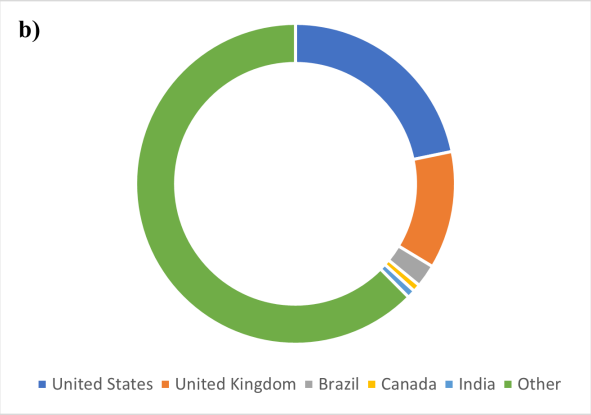
Parameter estimates from phylogenetic analyses in all six samples from both treatments were summarized in Table 3.4. In Figure 3.12, the parameter estimates from MCMC performed sequentially and in parallel from all six trials were compared with boxplots. Similar to the simulation studies, the U-test was employed to test for statistical significance. However, as "true" values for the parameters could not be ascertained, the t-test could not be applied. Accordingly, all parameters obtained from the sequential and parallel MCMC runs were not significantly different. To summarize all parameter estimates, in Table 3.5, the statistical test results for the parameter estimates of all three datasets are shown.

Table 3.4: Parameter estimates obtained from MCMC phylogenetic analyses ran in parallel and sequentially on SARS-CoV-2 data.

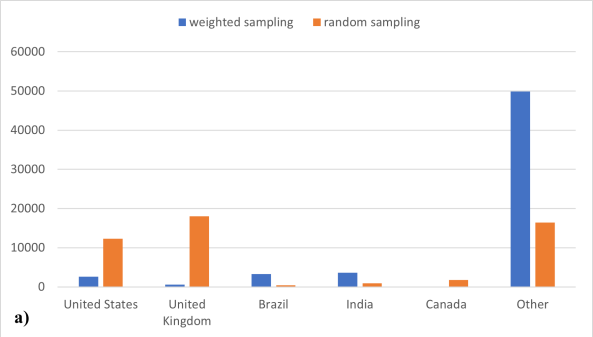
Group	rateAC	rateAG	rateAT	rateCG	rateGT	gammaShape
Parallel #1	0.092015	0.312976	0.056073	0.061551	0.386661	0.113105
Sequential #1	0.091023	0.311887	0.055249	0.0615	0.386104	0.114918
Parallel #2	0.08455	0.314067	0.056178	0.056135	0.417575	0.067258
Sequential #2	0.08455	0.313014	0.055337	0.055535	0.417353	0.069693
Parallel #3	0.088383	0.312436	0.058792	0.057802	0.393203	0.066863
Sequential #3	0.08852	0.312758	0.058895	0.057779	0.393688	0.068286
Parallel #4	0.107849	0.331881	0.075939	0.072337	0.413398	0.144432
Sequential #4	0.107845	0.331345	0.075838	0.072371	0.412555	0.145243
Parallel #5	0.103514	0.34382	0.07733	0.078817	0.406101	0.118399
Sequential #5	0.10304	0.344147	0.076792	0.079149	0.406347	0.120389
Parallel #6	0.094476	0.321044	0.071323	0.071273	0.415934	0.121527
Sequential #6	0.094676	0.322732	0.071535	0.07212	0.417309	0.122755



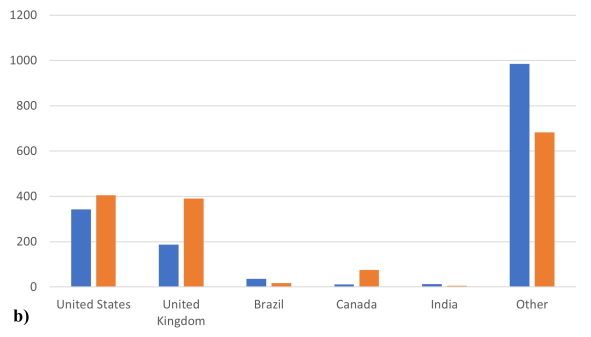
(a) Network 1



(b) Network 2



(c) Network 3



(d) Network 4

Figure 3.11: Doughnut chart of the number of sequences to be obtained from each country using weighted sampling to obtain a) 25,000 sequences from February 1st, 2020 to October 31st, or b) 1576 sequences from March 2020. These proportions represent the ideal number of sequences to be obtained from each country when using the weighting sampling strategy. The number of sequences to be obtained from each country when using the random versus weighted sampling strategy. a) Depicts the number of sequences obtained from each country if 25,000 sequences were obtained from February 1st, 2020 to October 31st, 2020. b) Shows the number of sequences to be obtained from each country if 1576 sequences from March 2020 were obtained.

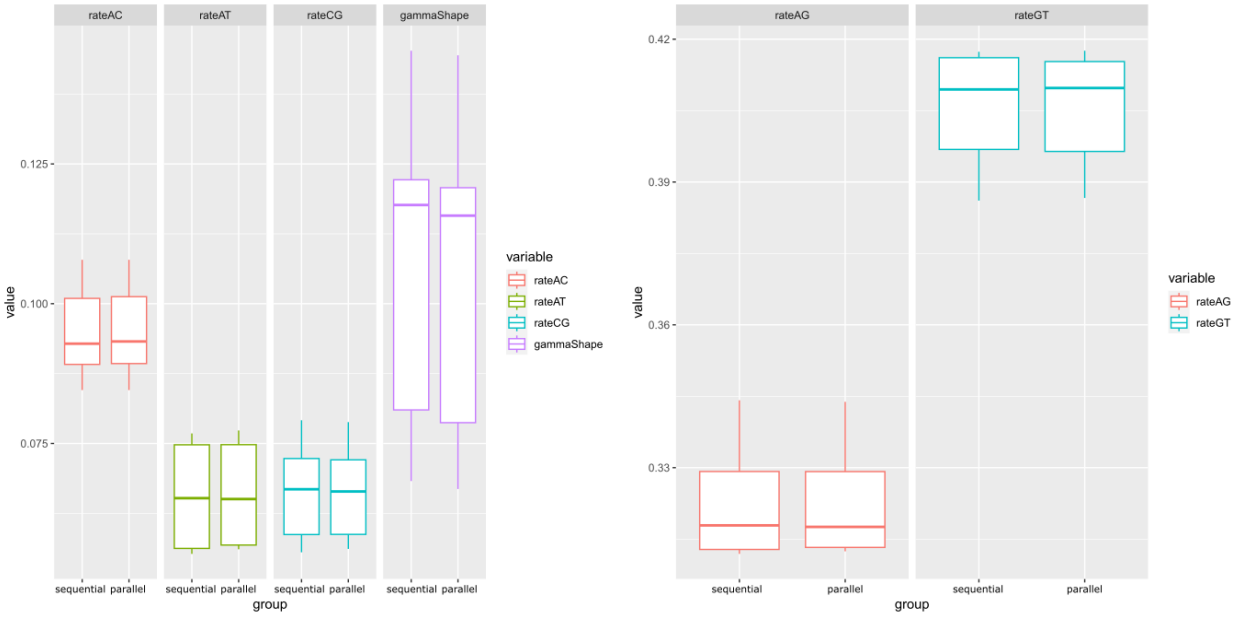
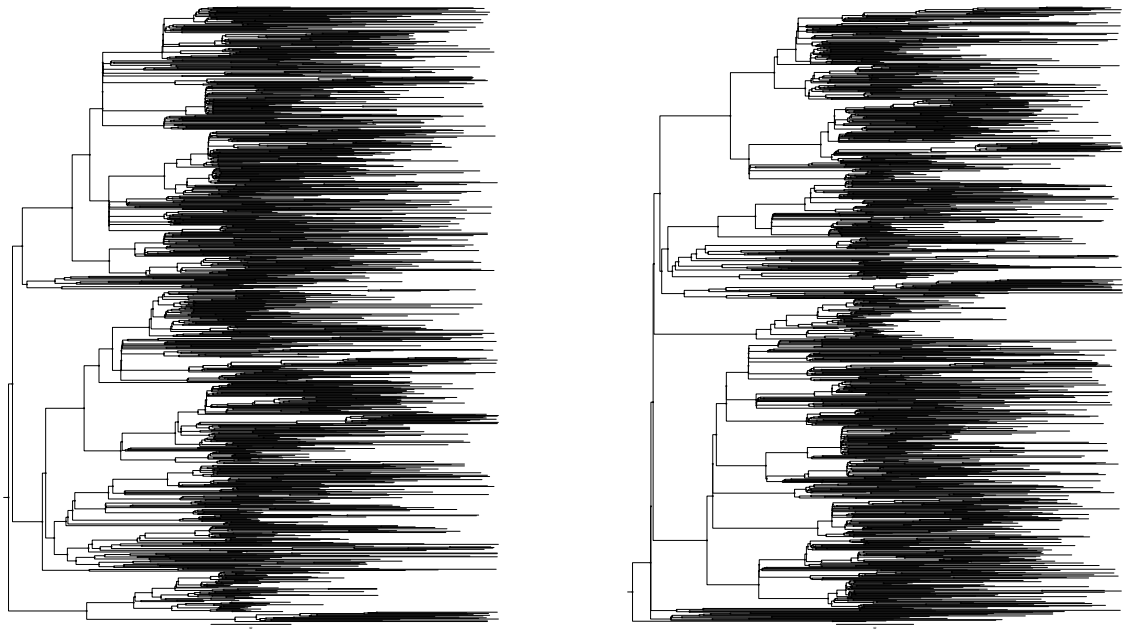


Figure 3.12: Boxplots from six parameter estimates from MCMC phylogenetic analyses ran in parallel and sequentially on SARS-CoV-2 data. Significant p-values from the U-test (as a line) and t-test (directly above) are labeled. Figures with different vertical axis scaling were used due to differences in the ranges of values.

Next, MCC trees obtained from the parallel and sequential MCMC runs were visualized and compared in Figure 3.13. Since the ground-truth phylogenetic trees were unknown, MCC trees obtained from the parallel and sequential MCMC runs in all three datasets were directly compared with each other with distance metrics in Figure 3.14. The U-test was employed to identify any significant differences between calculated distances. In all metrics except for the GeoUnrooted distances, only significant differences in distance metrics were observed between the comparisons of either SARS-CoV-2 vs HIV (100% sampling) or SARS-CoV-2 vs HIV (10% sampling). For the GeoUnrooted metric, distances calculated were statistically significant for all three datasets.

Finally, MCMC trace plots for rateAC of the analyses involving parallel and sequential MCMC chains for the second sample of the SARS-CoV-2 dataset retrieved with weighted sampling was plotted in Figure 3.15. Trace plots for both treatments were fairly similar. The



(a) parallel (b) sequential

Figure 3.13: Phylogenetic trees obtained from MCMC phylogenetic analyses ran in parallel (a) and sequentially (b) on SARS-CoV-2 data.

Table 3.5: Mann-whitney U-test and Student t-test p-values for parameter estimates obtained from MCMC Phylogenetic analyses ran in parallel and sequentially for SARS-CoV-2 data, and simulated data with perfect sampling rate (Simulation 1), and 10% sampling rate (Simulation 2)

Group	**rateAC**	**rateAG**	**rateAT**	**rateCG**	**rateGT**	**gammaShape**
Simulation 1 Parallel t-test	0.01129	0.089964	0.102195	0.151682	0.227696	0.689272
Simulation 1 Sequential t-test	0.010094	0.086547	0.099789	0.142497	0.224073	0.685559
Simulation 1 U-test	0.970512	0.970512	0.970512	0.853428	0.853428	0.739364
Simulation 2 Sequential t-test	0.232056	0.179726	0.03169	0.503558	0.185432	0.777523
Simulation 2 Parallel t-test	0.227036	0.179334	0.029217	0.511842	0.184631	0.784161
Simulation 2 U-test	0.911797	1	0.795936	0.795936	1	0.684211
SARS-CoV-2 U-test	1	0.937229437	0.818181818	1	1	0.588744589

dissimilarities in the lengths of the plots for the parallel and sequential chains were attributed to the difference in the number of iterations that were performed for each treatment.

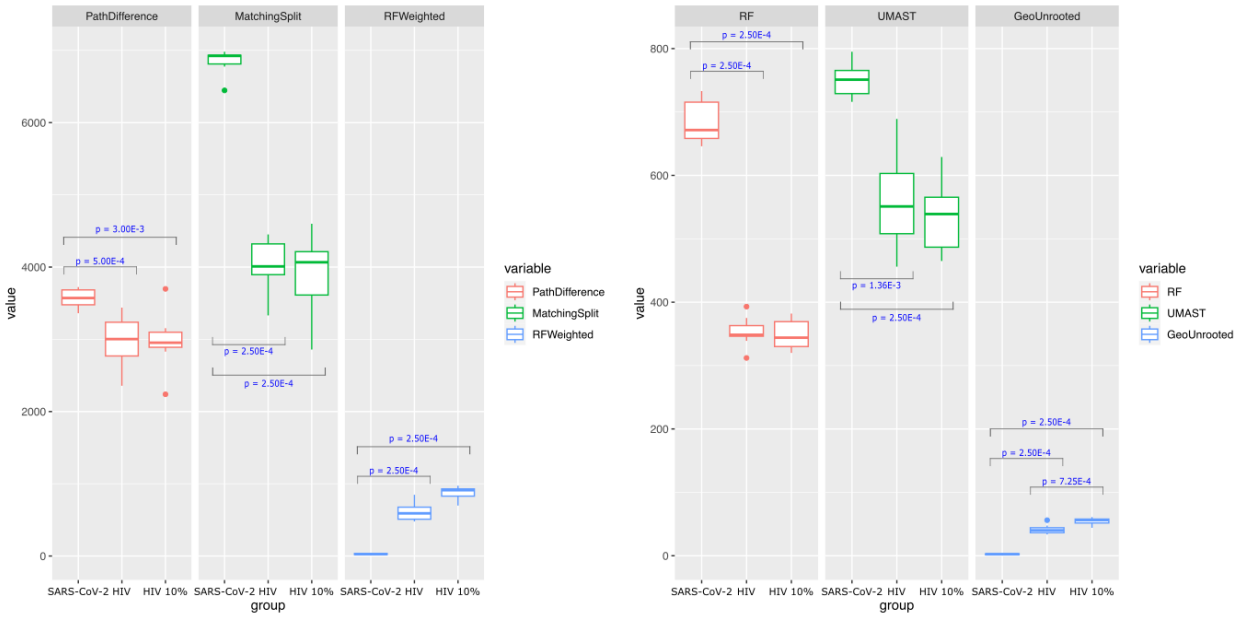


Figure 3.14: Boxplots from distance metrics calculated by comparing MCC phylogenetic trees obtained from sequential and parallel MCMC runs on simulated HIV data with perfect sampling, simulated HIV data with 10% sampling rate, and SARS-CoV-2 data. Significant p-values from the U-test comparing the distances between the metrics from each dataset are labeled. Figures with different vertical axis scaling were used due to differences in the ranges of values.

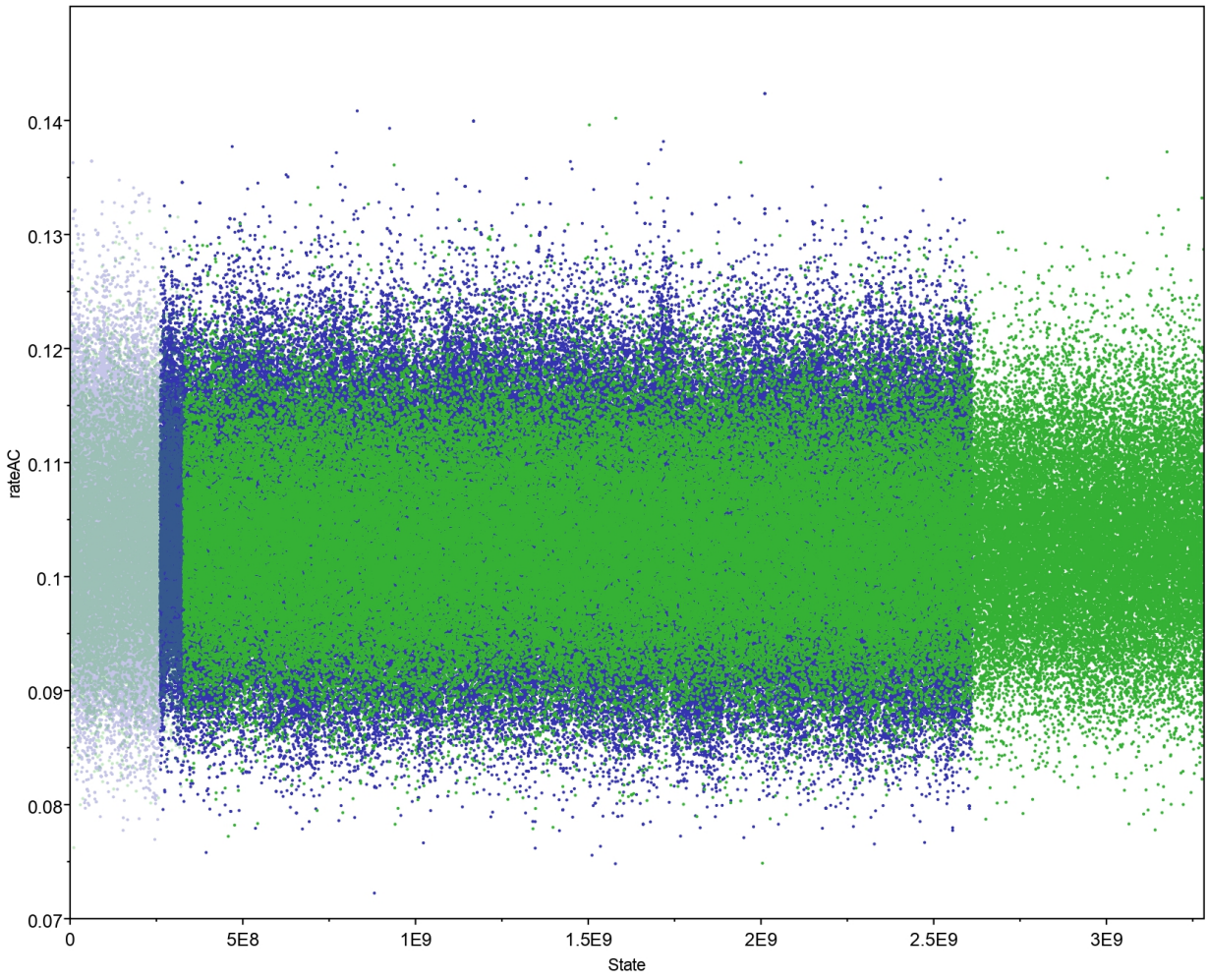


Figure 3.15: rateAC MCMC traces obtained from MCMC phylogenetic analyses ran in parallel (blue) and sequentially (green) on sample #1 in SARS-CoV-2 data from the early pandemic.

Chapter 4

Concluding remarks

Owing to the novel parallelization approach of the MCMC algorithm in Bayesian phylogenetic analysis created in Section 2.1, overall experimental run-times were reduced by almost 84 days in the simulation and SARS-CoV-2 studies. This is because the run time in BEAST2 is linear with the number of MCMC iterations, and therefore by running 29 parallel chains, we effectively improved our run time by 29-fold. This was equivalent to an average run time of 0.025 hours per million states as opposed to 0.9 hours per million states reported by other Bayesian phylogenetic studies involving SARS-CoV-2 sequences and the software BEAST2 [86]. In terms of modeling disease outbreaks, this run time difference is a tremendous improvement for inferring real-time changes in disease-causing entities such as viruses which evolve continuously and rapidly. Accordingly, these computational run time advancements can accelerate public health decisions and interventions. Even in analyses of biological entities where evolution is gradual and demand for quicker run-times is lower, this parallelization method provides an option for researchers to gain a quick understanding of the phylogenies they are working with. While our method mainly focuses on the time complexity of phylogenetic analysis, the implementation of our method does not sacrifice space complexity. In short, our method requires no additional memory or processing overhead. Moreover, our results suggest that the parameter estimates from MCMC parallelized were comparable in terms of accuracy with the estimates obtained from MCMC performed sequentially. Furthermore, comparison between the trace plots of parallel and sequential MCMC chains yielded no differences, reinforcing the theoretical viability of partitioning the parameter space into smaller subspaces (Figure 3.5, Figure 3.10,Figure 3.15). Ultimately, this outlined that minimal drawbacks ensue from the use of our MCMC parallelization method.

4.1 Parameter estimates

In the analyses involving all datasets, parameter predictions were generally similar even in the SARS-CoV-2 dataset which involved much lower sampling rates and other complications such as uneven sampling. In both treatments involving parallel and sequential MCMC chains of SARS-CoV-2 data, samples retrieved from a weighted surveying strategy were also included ($n = 3$). This strategy appeared to alleviate uneven sampling of the COVID-19 population both spatially and temporally (Figure 3.11). Such sampling strategies should be explored in future studies as data sharing initiatives such as GISAID will inspire future sampling programs involving exorbitant amounts of sampling data. This process is made easier with the novel parallelization approach which expedites phylogenetic research, making comparisons between sampling strategies convenient. Meanwhile, in the simulation studies, because of our understanding of the "true" parameter values, we were able to assess the performance of both the treatments with MCMC performed in parallel and sequentially. We found that when estimates deviated from the "true" values in the parallel sample, the same estimates also deviated in the sequential sample (Figure 3.1, Figure 3.6). This was seen for both the parameters rateAC in the HIV dataset with perfect sampling and rateAT in the HIV dataset with imperfect sampling. These results suggest that the sequential MCMC runs provided no significant advantages in terms of predicting phylogenetic parameters over the parallel MCMC runs when involving simulated data. Even in the parameter estimates from SARS-CoV-2 data which may have contained confounding factors, no significant differences in parameter estimates were observed, further reinforcing the applicability of this method for real-world datasets (Table 3.4).

The violin plots in our results showed that the independent MCMC chains of each parallel computation provided distinct individual parameter estimates, possibly due to an exploration of different regions in the parameter space (Figure 3.1, Figure 3.6). This aligned with our understanding of the novel parallelization approach which partitions the parameter space into various subspaces. When these individual MCMC chains were combined, we found that the parameter estimates of the combined MCMC chains were generally near the median of the individual estimates. In the HIV dataset with perfect sampling, we found that the parameter estimates from the sequential MCMC runs fell within the distributions of the estimates from the independent MCMC trials. However, in the HIV dataset with a 10% sampling rate, there were two parameters (rateCG, gammaShape) for which the estimates from the sequential MCMC runs were greater than the maximum estimates of the distribution of the individual MCMC runs. This suggested that with lowered sampling rates, the accuracy of parameter estimation may decrease in the novel parallelization method. However, nonetheless, when

multiple replicates were combined ($n = 10$) such as that in the experiment, the differences in the estimates from the parallel and sequential MCMC runs became mostly negligible. It is important to note that the differences in parameter estimates appeared deceptively large in the violin plots due to the scaling of the y-axis.

4.2 Phylogenetic tree prediction

Distance metrics calculated in the simulation studies also suggested that MCC phylogenetic trees obtained from parallel and sequential MCMC were mostly similar (Figure 3.4, Figure 3.9). When comparing the differences between the parallel and ground-truth trees with the differences between the sequential and ground-truth trees, we found that the trees obtained from the parallel MCMC differed more significantly in terms of the RF-weighted and GeoUnrooted metrics in the HIV dataset with imperfect sampling. In the HIV dataset involving perfect sampling, however, only the RF-weighted metric was near significant (p-value = 0.052). This suggests that longer MCMC chains are required to make inferences on branch lengths and sufficiently explore parameter spaces when sampling rates are imperfect. Meanwhile. all other distance metrics which mostly reflect tree topologies were similar, implying that the sampling rate of sequences had a greater effect on the predicted branch lengths than the tree topologies constructed. Furthermore, in both datasets, the RF metric was not significant while the RF-weighted metric was significant. This suggested that deviations in the predicted phylogenetic trees mostly lay in the differences in branch length which was accounted for in the RF-weighted metric but not the RF metric. Hence, overall, most of the differences lay in the scaling of the phylogenetic trees and not in the topology of the phylogenetic trees generated. Moreover, the fact that deviations were observed in the phylogenetic trees generated and not the parameters estimated also connects back to the idea that the phylogenetic tree spaces are more complex and difficult to predict.

We also calculated the distance metrics directly between phylogenetic trees obtained from the sequential and parallel MCMC samples. We found that in general, our parallelization methodology performed worse for the real-world SARS-CoV-2 dataset (Figure 9) than the simulated datasets. This was expected as there were likely more confounding factors and a lower sampling rate in the SARS-CoV-2 dataset. Moreover, in the simulated HIV datasets, we did not notice any significant effects from the differences in sampling rate on the phylogenetic trees produced from the parallel and sequential MCMC runs. Only in the GeoUnrooted metric was there a significant difference observed in phylogenetic trees produced with perfect and imperfect sampling, which again is related to the tree edge lengths.

4.3 Conclusion

Ultimately, our results suggested that parallelizing the MCMC runs provided a much-improved run time while losing little to no accuracy in terms of parameter estimates and generating phylogenetic trees. The sequential MCMC runs provided no significant advantage in predicting phylogenetic parameters over the parallel MCMC runs in our analyses involving simulated and SARS-CoV-2 data. Meanwhile, distance metrics calculated in the simulation study also suggested that MCC phylogenetic trees obtained from parallel and sequential MCMC were mostly similar except for branch lengths. In general, our parallelization methodology was less consistent for the real-world SARS-CoV-2 data. Most of the differences lay in the scaling of the phylogenetic trees and not in the topology of the phylogenetic trees for the simulation dataset. On the other hand, in the SARS-CoV-2 dataset, differences were found in the topology as well. In the context of viral outbreaks, which often entail the inference of transmission networks, this is an acceptable level of difference as transmission networks are much more sensitive and dependent on the topological structure of phylogenetic trees rather than the branch lengths [87]. Through the experiments performed, we demonstrated that our novel parallelization method is consistent for samples with imperfect sampling and even real-world datasets involving enormous population sizes like the COVID-19 pandemic. Hence, our methodology provides an acceptable trade-off between accuracy and run time, and offers a practical alternative to traditional long sequential MCMC runs for modeling future viral outbreaks. Future studies should investigate how this parallelization method can be integrated with other current optimization methods to attain even better performance. An intriguing avenue worth exploring involves optimizing the space complexity of our parallelization method to reduce memory requirements that often hinder phylogenetic studies as well.

An important consideration in our parallelization method involves addressing the potential limitation of insufficient convergence in each of the shorter parallelized MCMC chains. This problem plagued previous applications of analogous MCMC parallelization approaches when they were first implemented. However, through computational advancements, MCMC chain lengths that used to be considered "long" for modeling viral outbreaks like SARS-CoV (e.g. 10 million, n=40, 2009) [88], MERS-COV (e.g. 30 million, n=131, 2020) [89], and HIV (e.g. 50 million, n=2907, 2012) [90] can now be easily implemented as shorter chains in our parallelization method. For example, in the studies performed in this thesis, our "short chains" consisted of 100 million iterations each. Although this alone does not guarantee sufficient convergence in our studies, we demonstrated sufficient convergence through the consistency of our parameter estimates with traditional methods, convergence to the station-

ary distribution in our trace plots, and fulfillment of an ESS of at least 200 in all samples. In future analyses involving our parallelization methodology, it is possible that select shorter parallel chains do not converge to the stationary distribution. However, as long as most parallel chains converge, the correct MCC tree and corresponding parameter estimates will be obtained as they will have a greater posterior probability compared to the outputs from the chains with insufficient convergence. In future studies, if needed, sufficient convergence can be further explored by employing other metrics such as the Gelman-Rubin convergence diagnostic to assess whether the parallelized chains converge sufficiently to the posterior distribution [91]. In short, although our parallelization method may suffer in accuracy due to insufficient convergence in select parallel chains, in the context of viral outbreaks, its much-improved run-times greatly outweigh the insignificant loss of accuracy in phylogenetic analyses as demonstrated by the simulation and SARS-CoV-2 studies in this thesis.

In conclusion, the parallelization of phylogenetic analyses for viral outbreaks is a pivotal stride in the realm of phylogenetic and infectious disease research. As we grapple with the constant threat of emerging pathogens, the ability to swiftly and accurately analyze their evolutionary dynamics is cardinal. Streamlining the computational process involved in phylogenetic research facilitates this process and enables real-time decision-making for public health responses. This is because timely identification of transmission networks and infection tracking through viral evolution can significantly enhance our understanding of the epidemiological landscape, aiding in the development of targeted interventions and mitigation strategies. Therefore, the optimization efforts discussed in this paper not only contribute to the advancement of scientific methodologies but also carry profound implications for the public health of populations worldwide.

Bibliography

1. Tamura, K. *et al.* Estimating divergence times in large molecular phylogenies. *Proceedings of the National Academy of Sciences* **109.** Publisher: Proceedings of the National Academy of Sciences, 19333–19338. https://www.pnas.org/doi/10.1073/pnas.1213199109 (2023) (Nov. 20, 2012).

2. Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (John Murray, London, 1859).

3. Hendry, A. P., Schoen, D. J., Wolak, M. E. & Reid, J. M. The Contemporary Evolution of Fitness. *Annual Review of Ecology, Evolution, and Systematics* **49,** 457–476. ISSN: 1545-2069. http://dx.doi.org/10.1146/annurev-ecolsys-110617-062358 (Nov. 2018).

4. Alberts, B. *et al. Molecular Biology of the Cell* 4th. ISBN: 978-0815332183 (Garland Science, New York, 2002).

5. Wikipedia contributors. *Transversion* Accessed on December 28, 2023. https://en.wikipedia.org/wiki/Transversion.

6. Holmes, E. C. The Evolutionary Genetics of Emerging Viruses. *Annual Review of Ecology, Evolution, and Systematics* **40,** 353–372 (2009).

7. Li, S.-W. & Lin, C.-W. Human coronaviruses: Clinical features and phylogenetic analysis. *BioMedicine* **3,** 43–50. ISSN: 2211-8020. http://dx.doi.org/10.1016/j.biomed.2012.12.007 (Mar. 2013).

8. Raj, V. S., Osterhaus, A. D., Fouchier, R. A. & Haagmans, B. L. MERS: emergence of a novel human coronavirus. *Current Opinion in Virology* **5,** 58–62. ISSN: 1879-6257. http://dx.doi.org/10.1016/j.coviro.2014.01.010 (Apr. 2014).

9. Helmy, Y. A. *et al.* The COVID-19 Pandemic: A Comprehensive Review of Taxonomy, Genetics, Epidemiology, Diagnosis, Treatment, and Control. *Journal of Clinical Medicine* **9,** 1225. ISSN: 2077-0383. http://dx.doi.org/10.3390/jcm9041225 (Apr. 2020).

10. Wang, J.-T. *et al.* The role of phylogenetic analysis in clarifying the infection source of a COVID-19 patient. *Journal of Infection* **81,** 147–178. ISSN: 0163-4453. http://dx.doi.org/10.1016/j.jinf.2020.03.031 (July 2020).

11. Jungck, J. R. & Ko, H. Phylogenetic Analysis to Detect COVID Superspreaders. *Microbiology Research Journal International* **33,** 36–43. ISSN: 2456-7043. http://dx.doi.org/10.9734/mrji/2023/v33i81400 (Oct. 2023).

12. Perera, D. *et al.* Reconstructing SARS-CoV-2 infection dynamics through the phylogenetic inference of unsampled sources of infection. *PLOS ONE* **16** (ed Kalendar, R.) e0261422. ISSN: 1932-6203. http://dx.doi.org/10.1371/journal.pone.0261422 (Dec. 2021).

13. Digital Atlas of Ancient Life. *Reading Trees: Phylogenetics* Accessed on December 28, 2023. https://www.digitalatlasofancientlife.org/learn/systematics/phylogenetics/reading-trees/.

14. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian Phylogeography Finds Its Roots. *PLoS Computational Biology* **5** (ed Fraser, C.) e1000520. ISSN: 1553-7358. http://dx.doi.org/10.1371/journal.pcbi.1000520 (Sept. 2009).

15. Young, A. D. & Gillung, J. P. Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology* **45,** 225–247. ISSN: 1365-3113. http://dx.doi.org/10.1111/syen.12406 (Dec. 2019).

16. Philippe, H. *et al.* Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biology* **9** (ed Penny, D.) e1000602. ISSN: 1545-7885. http://dx.doi.org/10.1371/journal.pbio.1000602 (Mar. 2011).

17. Katoh, K., Misawa, K., Kuma, K.-i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30,** 3059–3066. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkf436 (2023) (July 15, 2002).

18. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* **9,** 286–298. ISSN: 1477-4054. http://dx.doi.org/10.1093/bib/bbn013 (Mar. 2008).

19. Jukes, T. H. & Cantor, C. R. Evolution of Protein Molecules. *Mammalian Protein Metabolism* **3,** 21–132 (1969).

20. Hasegawa, M., Kishino, H. & Yano, T.-a. Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA. *Journal of Molecular Evolution* **22,** 160–174 (1985).

21. Tamura, K. & Nei, M. Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humans and Chimpanzees. *Molecular Biology and Evolution* **10,** 512–526 (1993).

22. Tavare, S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences* **17,** 57–86 (1986).

23. Rahimi, A., Mirzazadeh, A. & Tavakolpour, S. Genetics and genomics of SARS-CoV-2: A review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection. *Genomics* **113,** 1221–1232. ISSN: 0888-7543. http://dx.doi.org/10.1016/j.ygeno.2020.09.059 (Jan. 2021).

24. Dos Reis, M., Donoghue, P. C. J. & Yang, Z. Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics* **17,** 71–80. ISSN: 1471-0064. http://dx.doi.org/10.1038/nrg.2015.8 (Dec. 2015).

25. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed Phylogenetics and Dating with Confidence. *PLOS Biology* **4.** Publisher: Public Library of Science, e88. ISSN: 1545-7885. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040088 (2023) (Mar. 14, 2006).

26. Tay, J. H., Porter, A. F., Wirth, W. & Duchene, S. The Emergence of SARS-CoV-2 Variants of Concern Is Driven by Acceleration of the Substitution Rate. *Molecular Biology and Evolution* **39,** msac013. ISSN: 1537-1719. https://doi.org/10.1093/molbev/msac013 (2023) (Feb. 1, 2022).

27. Jin, G., Nakhleh, L., Snir, S. & Tuller, T. Inferring Phylogenetic Networks by the Maximum Parsimony Criterion: A Case Study. *Molecular Biology and Evolution* **24,** 324–337. ISSN: 1537-1719. http://dx.doi.org/10.1093/molbev/msl163 (Oct. 2006).

28. Kannan, L. & Wheeler, W. C. Maximum Parsimony on Phylogenetic networks. *Algorithms for Molecular Biology* **7,** 9. ISSN: 1748-7188. https://doi.org/10.1186/1748-7188-7-9 (2023) (May 2, 2012).

29. Hendy, M. & Penny, D. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences* **59,** 277–290. ISSN: 0025-5564. http://dx.doi.org/10.1016/0025-5564(82)90027-X (June 1982).

30. Sanderson, M. J. & Donoghue, M. J. Reconstructing shifts in diversification rates on phylogenetic trees. *Trends in Ecology amp; Evolution* **11,** 15–20. ISSN: 0169-5347. http://dx.doi.org/10.1016/0169-5347(96)81059-7 (Jan. 1996).

31. Van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution* **83,** 104351. ISSN: 1567-1348. http://dx.doi.org/10.1016/j.meegid.2020.104351 (Sept. 2020).

32. Felsenstein, J. Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Systematic Zoology* **27,** 401. ISSN: 0039-7989. http://dx.doi.org/10.2307/2412923 (Dec. 1978).

33. Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology amp; Evolution* **11,** 367–372. ISSN: 0169-5347. http://dx.doi.org/10.1016/0169-5347(96)10041-0 (Sept. 1996).

34. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313. ISSN: 1367-4803. http://dx.doi.org/10.1093/bioinformatics/btu033 (Jan. 2014).

35. Nascimento, F. F., Reis, M. d. & Yang, Z. A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology amp; Evolution* **1,** 1446–1454. ISSN: 2397-334X. http://dx.doi.org/10.1038/s41559-017-0280-x (Sept. 2017).

36. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21,** 1087–1092. ISSN: 1089-7690. http://dx.doi.org/10.1063/1.1699114 (June 1953).

37. Larget, B. & Simon, D. L. Markov Chasin Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. *Molecular Biology and Evolution* **16,** 750–759. ISSN: 1537-1719. http://dx.doi.org/10.1093/oxfordjournals.molbev.a026160 (June 1999).

38. Grenfell, B. T. *et al.* Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science* **303,** 327–332. ISSN: 1095-9203. http://dx.doi.org/10.1126/science.1090727 (Jan. 2004).

39. Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A. J. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences* **110,** 228–233. ISSN: 1091-6490. http://dx.doi.org/10.1073/pnas.1207965110 (Dec. 2012).

40. Kingman, J. The coalescent. *Stochastic Processes and their Applications* **13,** 235–248. ISSN: 0304-4149. http://dx.doi.org/10.1016/0304-4149(82)90011-4 (Sept. 1982).

41. Drummond, A. J. & Rambaut, A. *Skyline Plots* Accessed on December 28, 2023. https://taming-the-beast.org/tutorials/Skyline-plots/.

42. Mishra, P., Singh, U., Pandey, C., Mishra, P. & Pandey, G. Application of student's t-test, analysis of variance, and covariance. *Annals of Cardiac Anaesthesia* **22,** 407. ISSN: 0971-9784. http://dx.doi.org/10.4103/aca.ACA_94_19 (2019).

43. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18,** 50–60. ISSN: 0003-4851. http://dx.doi.org/10.1214/aoms/1177730491 (Mar. 1947).

44. Robinson, D. & Foulds, L. Comparison of phylogenetic trees. *Mathematical Biosciences* **53,** 131–147. ISSN: 0025-5564. http://dx.doi.org/10.1016/0025-5564(81)90043-2 (Feb. 1981).

45. Bogdanowicz, D., Giaro, K. & Wróbel, B. TreeCmp: Comparison of Trees in Polynomial Time. *Evolutionary Bioinformatics* **8,** EBO.S9657. ISSN: 1176-9343. http://dx.doi.org/10.4137/EBO.S9657 (Jan. 2012).

46. Goluch, T., Bogdanowicz, D. & Giaro, K. ¡scp¿Visual TreeCmp¡/scp¿: Comprehensive Comparison of Phylogenetic Trees on the Web. *Methods in Ecology and Evolution* **11** (ed Price, S.) 494–499. ISSN: 2041-210X. http://dx.doi.org/10.1111/2041-210X.13358 (Feb. 2020).

47. Robinson, D. F. & Foulds, L. R. in *Combinatorial Mathematics VI* (eds Horadam, A. F. & Wallis, W. D.) 119–126 (Springer, Berlin, Heidelberg, 1979).

48. Dong, S. & Kraemer, E. *Calculation, visualization, and manipulation of MASTs (maximum agreement subtrees)* in *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.* (IEEE). http://dx.doi.org/10.1109/csb.2004.1332453.

49. Khodaei, M., Owen, M. & Beerli, P. Geodesics to characterize the phylogenetic landscape. *PLOS ONE* **18** (ed Yoshida, R.) e0287350. ISSN: 1932-6203. http://dx.doi.org/10.1371/journal.pone.0287350 (June 2023).

50. Owen, M. & Provan, J. S. *A Fast Algorithm for Computing Geodesic Distances in Tree Space* 2009. https://arxiv.org/abs/0907.3942.

51. Steel, M. A. & Penny, D. Distributions of Tree Comparison Metrics–Some New Results. *Systematic Biology* **42,** 126–141. ISSN: 1076-836X. http://dx.doi.org/10.1093/sysbio/42.2.126 (June 1993).

52. Money, D. & Whelan, S. Characterizing the Phylogenetic Tree-Search Problem. *Systematic Biology* **61,** 228. ISSN: 1063-5157. http://dx.doi.org/10.1093/sysbio/syr097 (Mar. 2012).

53. Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* **13,** 303–314. ISSN: 1471-0064. http://dx.doi.org/10.1038/nrg3186 (Mar. 2012).

54. Hoang, D. T. *et al.* MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evolutionary Biology* **18.** ISSN: 1471-2148. http://dx.doi.org/10.1186/s12862-018-1131-3 (Feb. 2018).

55. Ye, C. *et al.* matOptimize: a parallel tree optimization method enables online phylogenetics for SARS-CoV-2. *Bioinformatics* **38** (ed Schwartz, R.) 3734–3740. ISSN: 1367-4811. http://dx.doi.org/10.1093/bioinformatics/btac401 (June 2022).

56. Zhou, X., Shen, X.-X., Hittinger, C. T. & Rokas, A. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Molecular Biology and Evolution* **35,** 486–503. ISSN: 1537-1719. http://dx.doi.org/10.1093/molbev/msx302 (Nov. 2017).

57. Whelan, S. & Morrison, D. A. in *Bioinformatics* 349–377 (Springer New York, Nov. 2016). ISBN: 9781493966226. http://dx.doi.org/10.1007/978-1-4939-6622-6_14.

58. Guindon, S. & Gascuel, O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology* **52** (ed Rannala, B.) 696–704. ISSN: 1063-5157. http://dx.doi.org/10.1080/10635150390235520 (Oct. 2003).

59. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5** (ed Poon, A. F. Y.) e9490. ISSN: 1932-6203. http://dx.doi.org/10.1371/journal.pone.0009490 (Mar. 2010).

60. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32,** 268–274. ISSN: 0737-4038. http://dx.doi.org/10.1093/molbev/msu300 (Nov. 2014).

61. Aberer, A. J., Pattengale, N. D. & Stamatakis, A. Parallel computation of phylogenetic consensus trees. *Procedia Computer Science* **1,** 1065–1073. ISSN: 1877-0509. http://dx.doi.org/10.1016/j.procs.2010.04.118 (May 2010).

62. De Maio, N. *et al.* Maximum likelihood pandemic-scale phylogenetics. *Nature Genetics* **55,** 746–752. ISSN: 1546-1718. http://dx.doi.org/10.1038/s41588-023-01368-0 (Apr. 2023).

63. Zhang, C. & IV, F. A. M. *Variational Bayesian Phylogenetic Inference* in *International Conference on Learning Representations* (2019). https://openreview.net/forum?id=SJVmjjR9FX.

64. Aberer, A. J., Kobert, K. & Stamatakis, A. ExaBayes: Massively Parallel Bayesian Tree Inference for the Whole-Genome Era. *Molecular Biology and Evolution* **31,** 2553–2556. ISSN: 0737-4038. http://dx.doi.org/10.1093/molbev/msu236 (Aug. 2014).

65. Rajapaksa, S., Rasanjana, W., Perera, I. & Meedeniya, D. *GPU Accelerated Maximum Likelihood Analysis for Phylogenetic Inference* in *Proceedings of the 2019 8th International Conference on Software and Computer Applications* (ACM, Feb. 2019). http://dx.doi.org/10.1145/3316615.3316630.

66. Baele, G., Ayres, D. L., Rambaut, A., Suchard, M. A. & Lemey, P. in *Evolutionary Genomics* 691–722 (Springer New York, 2019). ISBN: 9781493990740. http://dx.doi.org/10.1007/978-1-4939-9074-0_23.

67. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology* **10.** Publisher: Public Library of Science, e1003537. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003537 (2023) (Apr. 10, 2014).

68. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology* **10** (ed Prlic, A.) e1003537. ISSN: 1553-7358. http://dx.doi.org/10.1371/journal.pcbi.1003537 (Apr. 2014).

69. Whidden, C. & Matsen, F. A. Quantifying MCMC Exploration of Phylogenetic Tree Space. *Systematic Biology* **64,** 472–491. ISSN: 1063-5157. http://dx.doi.org/10.1093/sysbio/syv006 (Jan. 2015).

70. Harrington, S. M., Wishingrad, V. & Thomson, R. C. Properties of Markov Chain Monte Carlo Performance across Many Empirical Alignments. *Molecular Biology and Evolution* **38** (ed Pupko, T.) 1627–1640. ISSN: 1537-1719. http://dx.doi.org/10.1093/molbev/msaa295 (Nov. 2020).

71. Hafych, V., Eller, P., Schulz, O. & Caldwel, A. Parallelizing MCMC sampling via space partitioning. *Statistics and Computing* **32.** ISSN: 1573-1375. http://dx.doi.org/10.1007/s11222-022-10116-z (June 2022).

72. Solonen, A. *et al.* Efficient MCMC for Climate Model Parameter Estimation: Parallel Adaptive Chains and Early Rejection. *Bayesian Analysis* **7.** ISSN: 1936-0975. http://dx.doi.org/10.1214/12-BA724 (Sept. 2012).

73. VanDerwerken, D. N. & Schmidler, S. C. *Parallel Markov Chain Monte Carlo* 2013. https://arxiv.org/abs/1312.7479.

74. Bouckaert, R., Collienne, L. & Gavryushkin, A. Online Bayesian Analysis with BEAST 2. http://dx.doi.org/10.1101/2022.05.03.490538 (May 2022).

75. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology* **67** (ed Susko, E.) 901–904. ISSN: 1076-836X. http://dx.doi.org/10.1093/sysbio/syy032 (Apr. 2018).

76. Rambaut, A. *Figtree* 2010.

77. Moshiri, N., Ragonnet-Cronin, M., Wertheim, J. O. & Mirarab, S. FAVITES: simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics* **35** (ed Schwartz, R.) 1852–1861. ISSN: 1367-4811. http://dx.doi.org/10.1093/bioinformatics/bty921 (Nov. 2018).

78. Brito, A. F. *et al.* Global disparities in SARS-CoV-2 genomic surveillance. http://dx.doi.org/10.1101/2021.08.21.21262393 (Aug. 2021).

79. Furuse, Y. Genomic sequencing effort for SARS-CoV-2 by country during the pandemic. *International Journal of Infectious Diseases* **103,** 305–307. ISSN: 1201-9712. http://dx.doi.org/10.1016/j.ijid.2020.12.034 (Feb. 2021).

80. Huddleston, J. *et al.* Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. *Journal of Open Source Software* **6,** 2906. ISSN: 2475-9066. http://dx.doi.org/10.21105/joss.02906 (Jan. 2021).

81. David Yang Michael Hynes, Q. L. *Investigating the effect of sampling bias on SARS-CoV-2 phylogneetic inference* Honours project.

82. Schroeder, S. A. How to Interpret Covid-19 Predictions: Reassessing the IHME's Model. *Philosophy of Medicine* **2.** ISSN: 2692-3963. http://dx.doi.org/10.5195/pom.2021.43 (Mar. 2021).

83. Paul Gordon, D. Y. *Nybbler: Targeted subsampling of SARS-CoV-2 genome sequence ensembles used in genomic epidemiology*

84. Valero-Mora, P. M. ggplot2:Elegant Graphics for Data Analysis. *Journal of Statistical Software* **35.** ISSN: 1548-7660. http://dx.doi.org/10.18637/jss.v035.b01 (2010).

85. *Inkscape project* 2010. https://inkscape.org.

86. Dellicour, S. *et al.* A Phylodynamic Workflow to Rapidly Gain Insights into the Dispersal History and Dynamics of SARS-CoV-2 Lineages. *Molecular Biology and Evolution* **38** (ed Stuart, N.) 1608–1613. ISSN: 1537-1719. http://dx.doi.org/10.1093/molbev/msaa284 (Nov. 2020).

87. Colijn, C. & Gardy, J. Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, Medicine, and Public Health* **2014,** 96–108. ISSN: 2050-6201. http://dx.doi.org/10.1093/emph/eou018 (Jan. 2014).

88. Yip, C. W. *et al.* Phylogenetic perspectives on the epidemiology and origins of SARS and SARS-like coronaviruses. *Infection, Genetics and Evolution* **9,** 1185–1196. ISSN: 1567-1348. http://dx.doi.org/10.1016/j.meegid.2009.09.015 (Dec. 2009).

89. Chen, X. *et al.* The Phylogeography of MERS-CoV in Hospital Outbreak-Associated Cases Compared to Sporadic Cases in Saudi Arabia. *Viruses* **12,** 540. ISSN: 1999-4915. http://dx.doi.org/10.3390/v12050540 (May 2020).

90. Volz, E. M., Koopman, J. S., Ward, M. J., Brown, A. L. & Frost, S. D. W. Simple Epidemiological Dynamics Explain Phylogenetic Clustering of HIV from Patients with Recent Infection. *PLoS Computational Biology* **8** (ed Fraser, C.) e1002552. ISSN: 1553-7358. http://dx.doi.org/10.1371/journal.pcbi.1002552 (June 2012).

91. Roy, V. Convergence Diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application* **7,** 387–412. ISSN: 2326-831X. http://dx.doi.org/10.1146/annurev-statistics-031219-041300 (Mar. 2020).

Appendix

An example of the configuration file used in the pipeline is shown below. In the example configuration file, a 10% sampling rate is specified.

```
{
    # Module Implementations
    "Driver": "Default",
    "Logging": "File",
    "TreeNode": "Simple",
    "ContactNetwork": "NetworkX",
    "ContactNetworkGenerator": "BarabasiAlbert",
    "SeedSelection": "Random",
    "SeedSequence": "VirusNonHomYuleHeightGTRGamma",
    "EndCriteria": "GEMF",
    "TransmissionTimeSample": "HIVARTGranichGEMF",
    "TransmissionNodeSample": "GEMF",
    "NodeEvolution": "VirusTreeSimulator",
    "SequenceEvolution": "GTRGammaSeqGen",
    "SourceSample": "Random",
    "NumTimeSample": "Once",
    "TimeSample": "Gamma",
    "NumBranchSample": "Single",
    "TreeUnit": "TruncatedNormal",
    "NodeAvailability": "TransmissionWeighted",
    "Sequencing": "Perfect",

    # Parameter Choices
    "end_time": 10000,
    "gemf_path": "GEMF",
    "hiv_a1_to_a2": 4.333333,
```

```
"hiv_a1_to_d": 0,
"hiv_a1_to_i1": 0.48,
"hiv_a2_to_a3": 0,
"hiv_a2_to_d": 0,
"hiv_a2_to_i2": 0.48,
"hiv_a3_to_a4": 0,
"hiv_a3_to_d": 0,
"hiv_a3_to_i3": 0,
"hiv_a4_to_d": 0,
"hiv_a4_to_i4": 0,
"hiv_freq_a1": 0,
"hiv_freq_a2": 0,
"hiv_freq_a3": 0,
"hiv_freq_a4": 0,
"hiv_freq_d": 0,
"hiv_freq_i1": 2,
"hiv_freq_i2": 0,
"hiv_freq_i3": 0,
"hiv_freq_i4": 0,
"hiv_freq_ns": 0,
"hiv_freq_s": 99000,
"hiv_i1_to_a1": 0.125,
"hiv_i1_to_d": 0,
"hiv_i1_to_i2": 8.66666666666,
"hiv_i2_to_a2": 0.125,
"hiv_i2_to_d": 0,
"hiv_i2_to_i3": 0,
"hiv_i3_to_a3": 0,
"hiv_i3_to_d": 0,
"hiv_i3_to_i4": 0,
"hiv_i4_to_a4": 0,
"hiv_i4_to_d": 0,
"hiv_ns_to_d": 0,
"hiv_ns_to_s": 999999,
"hiv_s_to_d": 0,
"hiv_s_to_i1_by_a1": 0.005625,
```

```
"hiv_s_to_i1_by_a2": 0,
"hiv_s_to_i1_by_a3": 0,
"hiv_s_to_i1_by_a4": 0,
"hiv_s_to_i1_by_i1": 0.1125,
"hiv_s_to_i1_by_i2": 0.0225,
"hiv_s_to_i1_by_i3": 0,
"hiv_s_to_i1_by_i4": 0,
"hiv_s_to_i1_seed": 0,
"hmmemit_path": "hmmemit",
"java_path": "java",
"node_sample_fraction": 0.1,
"num_cn_nodes": 1000,
"num_edges_from_new": 2,
"num_seeds": 2,
"out_dir": "/OUTPUT_DIR",
"seed_height": 25,
"seed_speciation_rate_func": "exp(-t**2)+1",
"seqgen_a_to_c": 1.8118,
"seqgen_a_to_g": 9.93411,
"seqgen_a_to_t": 0.7184,
"seqgen_c_to_g": 0.97148,
"seqgen_c_to_t": 9.93411,
"seqgen_freq_a": 0.392,
"seqgen_freq_c": 0.164,
"seqgen_freq_g": 0.212,
"seqgen_freq_t": 0.232,
"seqgen_g_to_t": 1.0,
"seqgen_gamma_shape": 2,
"seqgen_num_gamma_rate_categories": "",
"seqgen_path": "seq-gen",
"tree_rate_loc": 0.0008,
"tree_rate_max": float('inf'),
"tree_rate_min": 0,
"tree_rate_scale": 0.0005,
"ts_gamma_scale": 10,
"ts_gamma_shape": 1,
```

```
    "viral_sequence_type": "HIV1-B-DNA-POL-LITTLE",
    "vts_growthRate": 2.851904,
    "vts_max_attempts": 100,
    "vts_model": "logistic",
    "vts_n0": 1,
    "vts_t50": -2
}
```