The Impact of the Tree Prior and Purifying Selection on Estimating Clock Rates During Viral Epidemics

Master Thesis

Author(s):
Möller, Simon

Publication date:
2016-08-29

Permanent link:
https://doi.org/10.3929/ethz-b-000184546

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

The impact of the tree prior and purifying selection on estimating clock rates during viral epidemics

Master Thesis

Simon Möller

Monday 29th August, 2016

Advisors: Prof. Dr. Tanja Stadler, Dr. David Alan Rasmussen
Department of Biosystems, Science and Engineering, ETH Zürich

Abstract

This thesis is concerned with Bayesian phylogenetic inference of clock rates for viral epidemics. In this and other areas of application of phylogenetics it has been observed that the inferred rate decreases with an increase in the sampling period. Purifying selection is a likely biological factor that contributes to this phenomenon since it purges slightly deleterious mutations from a population over time thus decreasing the overall genetic diversity that is observed per unit time. However, other factors such as methodological biases also play a role and make a biological interpretation of results difficult. We aim to contribute towards disentangling these different influences. With a simulation study we demonstrate that a misspecified tree prior can upwardly bias the inferred clock rate and that the interplay of the different models involved in the inference can be complex and non-intuitive. We also show that the choice of tree prior can influence the inference of clock rate on a real world Ebola dataset from Sierra Leone, but fail to see such influence for a larger dataset from Guinea. Furthermore, we detect signs of purifying selection in the Guinea dataset by comparing rate estimates on internal and pendant branches.

Acknowledgements

Working on my thesis in the past six months has been an interesting and rewarding experience. I would like to thank the entire Computational Evolution group, especially Tanja Stadler and Louis du Plessis, for stimulating discussions and great advice. The time in the group has been great from a personal as well as a scientific perspective. I also want to thank Rudolphus Goclenius, who gave the working title – *Rudolphus* – to my project since his name sounded pleasant and he was born exactly 468 years before I officially started my thesis.

RODOLPHVS GOCLENIVS PHILOSOPHVS MARPVRGENSIS
QVISQVIS Aristotelem noſti, magnumq; Platona;
GOCLENI Haſſiaci talis imago fuit.
Dd d3

Figure 1: A picture of Rudolphus Goclenius (obtained from Wikipedia, original source: http://www.uni-mannheim.de/mateo/desbillons/aport/seite206.html).

Contents

Chapter 1

Introduction

Determining the depth to cover in the introduction to a Master's thesis is probably always a difficult task. The supervisor should not require any introduction to understand the content, whereas including enough information so that the "general audience" can follow would mean writing a couple of text books. As my target audience I chose an ETH Master's student from a field other than Computational Biology and Bioinformatics with some background in mathematics and statistics. I do not cover much of biology and hope that any reader is roughly familiar with DNA, Darwin and diseases in general. While the results in this thesis are in some sense quite technical and constrained to a narrow field, I hope that the introduction also conveys some of the beauty of the foundations that this work is built upon. Since everything in this thesis has to do with Bayesian phylogenetic inference I start by describing phylogenetics (section 1.1) and Bayesian inference (section 1.2) separately before discussing their conjunction (section 1.3).

The second purpose of this chapter is to outline the problems that are addressed in the remainder of the thesis. To this end, section 1.4 introduces the clock rate which is the main parameter we are interested in and section 1.5 briefly describes our contributions.

1.1 Phylogenetics

Yang and Rannala [1] recently published an excellent introduction to phylogenetic methods including the ones that are based on sequence distances or parsimony alone which I do not cover here. For even more details, Yang's book [2] is a valuable resource.

History Phylogenetic trees as we know and use them today originated with Darwin's concept of random variation and natural selection (see Fig. 1.1 (a) for a draft in his original notebook). These trees represent the evolutionary relationships between species (or, more generally, organisms) and are based on their difference in heritable traits. For a long time, scientists used expert knowledge and morphological characters to construct them. Usually, by using a parsimony argument stating that the best tree is the one that requires the least number of changes throughout the tree.

With the advances in sequencing technology it has now for some decades been possible to construct phylogenetic trees based on DNA and RNA sequence data. These are potentially much more accurate since evolution is actually happening on the nucleotide level. The data also suit themselves for a stringent mathematical analysis because, instead of having to weigh differences between having four and six legs or being herbivore or carnivore, algorithms can operate on a sequence of letters. In this section I will describe the framework used for such analyses.

Models of evolution and the Felsenstein likelihood We assume that as data we are given a multiple sequence alignment, i.e. a set of sequences in which the positions of the nucleotides are thought to correspond to one another such that differences can be interpreted as mutations in the past. Our goal is to compute the likelihood, $P(\mathcal{D}|\mathcal{T},\mathcal{M})$, of the data, $\mathcal{D}$, given an unrooted phylogenetic tree, $\mathcal{T}$, that describes the evolutionary relationships between the species whose sequences we are comparing and a substitution model, $\mathcal{M}$, which describes the dynamics of nucleotide changes. The tree consists of a topology and a set of branch lengths given in evolutionary distance, i.e. units of expected number of substitutions per site. An example is shown in Fig. 1.1 (b). The concept of a substitution will be explained further below. Any reader, who is not familiar with them, can for now think of mutations instead. Multiple substitution models with varying degrees of complexity and number of parameters exist, but they all essentially provide us with a transition probability matrix, $P(d)$. Its entries, $p_{i,j}(d)$, describe the probability that nucleotide i is substituted by nucleotide j over a distance d.

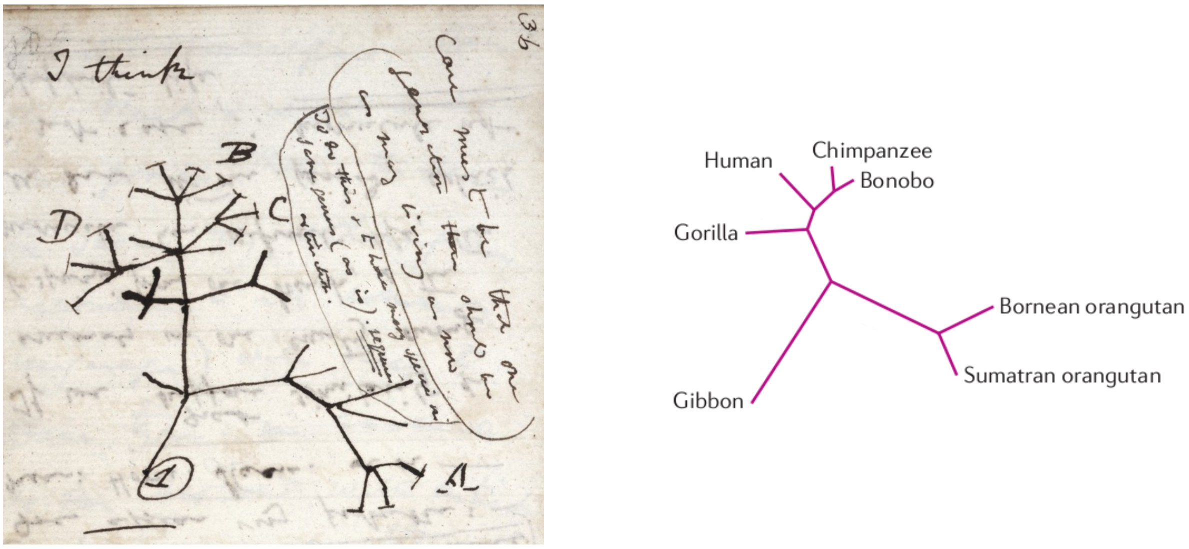
Let us start by stating explicitly some of the assumptions that we make:

1. Independence: Sites (i.e. different positions) in the alignment evolve independently of each other.
2. Memorylessness: Nucleotide substitutions at one site happen independently from the history at that site.
3. Neutrality: Nucleotide substitutions do not cause a change in fitness.
4. Reversibility: After correcting for varying nucleotide frequencies, substitutions happen at the same rate in both directions between any two nucleotides.

The first assumption allows us to compute the likelihood for each site separately and then multiply the outcomes to obtain the final result. The second assumption allows us to treat model evolution at each site as a Markov chain. The third assumption allows us to treat all substitutions from one nucleotide to another the same. This makes the Markov chain time homogeneous. The final assumptions makes the Markov chain reversible.

$\mathcal{D}$ provides us with the nucleotides at the *tips* of the tree. If we, for a moment, assume that we also know all nucleotides on internal nodes, we can compute the likelihood (for a single site in the alignment) by simply multiplying $p_{i_k,j_k}(d_k)$, where k enumerates all branches in the tree. Note that it does not matter which end of the branch is i and which one is j because of the reversibility assumption. The challenge is to enumerate all possible assignments of internal nodes and then multiply the likelihoods for each one. This can be done efficiently using Felsenstein's pruning algorithm [3].

Clock models So far we were concerned with a tree in which branch lengths were given in expected number of substitutions. Usually it is more interesting to have branch lengths in units of real time. For this we introduce the clock rate, r, a model parameter which scales the (temporal) branch length to give us expected number of substitutions again. Mathematically we have $d_k = t_k \cdot r$, where t_k is the branch length given in units of time and d_k is the distance used to compute the likelihood across branch k.


36
I think
B
C
D
1
A
Case must be that one generation then should be as many living as now
To do this & to have many species in same genus (as is) requires extinction.
Human
Chimpanzee
Bonobo
Gorilla
Bornean orangutan
Sumatran orangutan
Gibbon

(a) Draft of a phylogenetic tree on page 36 in Darwin's notebook (obtained from http://darwin-online.org.uk/).

(b) A tree representing the evolutionary relationship between human and apes (taken from [1], Box 3, bootstrap values removed).

Figure 1.1: Two examples of unrooted phylogenetic trees.

Theoretically, each branch could have its own substitution rate but this would result in too many parameters. The *molecular clock* assumptions states that the rate is constant across the tree. Still, the parameter r is unidentifiable as illustrated by a simple example: A distance of 5 expected substitutions can be explained by a temporal branch length of 5 years and a substitution rate of 1 substitution/year or by 1 year and a substitution rate of 5 substitutions/year. To fix this, we have to add some form of absolute temporal calibration. For macro-evolution, researchers often use fossils to determine the absolute time of internal nodes. For fast evolving pathogens, it can be sufficient to get samples over a time-span of months or years so that enough genetic change is captured to calibrate the clock.

A strict molecular clock is the most restrictive clock model and often the assumption of having an equal clock rate across all branches is violated by the data. Therefore, *relaxed clock models* have been developed, which allow the rate to vary across branches either in a correlated or in an uncorrelated way [4].

1.2 Bayesian inference

History If we want to learn about the world in a scientific way, we need a method to draw general conclusions from available data. Since deductive reasoning, i.e. drawing conclusions that are *certain*, is often impossible, science has largely adopted an inductive process, as e.g. described by Popper. In this framework, a hypothesis can never be proven true. It can only become *more plausible* as it defeats repeated attempts to falsify it.

Although he lived long before Popper, Reverend Thomas Bayes was concerned with assigning probabilities to unobserved events (or hypotheses) given some observed events (or data). In his famous "Essay towards Solving a Problem in the Doctrine of Chances" [5] that was published after his death, he considered the following thought experiment: Suppose you are blindfolded and sit in front of a table. The table is such that if a ball is thrown on it, it has equal chance to land anywhere on it. Your assistant throws a ball and it is your task to guess where on the table it landed. To achieve this you can ask your assistant to throw another ball and tell you whether it landed to the right or left of the original ball. If it landed to the right, it is more likely that the original ball was on the left half of the table and vice versa. By repeating this procedure, you can become increasingly certain about the position of the original ball.

Bayes' formula can be used to formalize this process of accumulating knowledge. Today, it is usually stated in the following form:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \qquad (1.1)$$

where A and B are some events and $P(B) \neq 0$. Though he never stated it explicitly, Bayes did prove a special case of it and "the way to the formula is apparent in several instances of the Essay" [6]. The theorem was also discovered by Laplace who put it to use in the fields of astronomy and made substantial contribution to the mathematical framework surrounding it.

Today, Bayes' formula can be derived in one line using the definition of conditional probabilities following Kolmogorov axioms, though it is much older than those axioms and Bayes himself did not have a notion of random variables. Much more excitingly, though, it can also be derived thoroughly as an extension of boolean logic and a set of axioms that try to capture our human "common sense". This was formally done by Cox [7] and together with the works by Pòlya [8] it leads to a deep foundation of Bayesian inference as stated in the preface of Jaynes book "Probability Theory: The Logic of Science" [9]:

> [W]hen one added Pòlya's qualitative conditions to [the consistency theorems of R. T. Cox] the result was a proof that, if degrees of plausibility are represented by real numbers, then there is a uniquely determined set of quantitative rules for conducting inference. That is, any other rules whose results conflict with them will necessarily violate an elementary – and nearly inescapable – desideratum of rationality or consistency.

It is fascinating that these rules agree with the general theorems of probability derived by Bernoulli and Laplace. However, in their new interpretation they are a result of logic and do not require inherent randomness in nature. As such, they ultimately follow a different definition of probability as a degree of believe for an event rather than the frequency at which that event happens in an infinite number of (random) experiments.

Bayes' formula in science We can rewrite Bayes' formula with different symbols to highlight it's application in science:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}, \qquad (1.2)$$

where

- $\mathcal{H}$ is some hypothesis,
- $\mathcal{D}$ are some data,
- $P(\mathcal{H}|\mathcal{D})$, the *posterior*, is the probability that our hypothesis is true, given the observed data (and our prior knowledge and some model),

- $P(\mathcal{D}|\mathcal{H})$, the *likelihood*, is the probability that the observed data is generated under the hypothesis,
- $P(\mathcal{H})$, the *prior*, is our prior believe that the hypothesis is true,
- $P(\mathcal{D})$, the *marginal likelihood*, is the overall probability that we observe the data under any hypothesis. For a discrete set of hypotheses, $\mathbb{H}$, we can rewrite it as $P(\mathcal{D}) = \sum_{\mathcal{H}^* \in \mathbb{H}} P(\mathcal{D}|\mathcal{H}^*)$. For a continuous hypothesis space – e.g. a continuous range of parameters – we use an integral instead of a sum.

The posterior is often the quantity that is of direct interest in science as it gives a measure of support for a given hypothesis. The prior, however, was and continues to be a cause for debates. Especially during the 20th century, there was a discussion on an almost religious level, whether Bayesian inference should be used in science and industry. Most famously, this debate was held between Jeffreys on the side of the Bayesians and Fisher on the side of the frequentists. While Bayesians believe that the prior allows to incorporate previous knowledge in a meaningful and robust way, the other side insists on *letting the data speak* for itself. Books have been written about the underlying philosophical differences and their technical implications (e.g. [10]), but by now many examples of successful use of Bayesian inference in the real world have been documented [11]. Jaynes therefore concludes:

> We are now in possession of proven theorems and masses of worked-out numerical examples. As a result, the superiority of Bayesian methods is now a thoroughly demonstrated fact in a hundred different areas.

Markoc Chain Monte Carlo methods On top of the philosophic difficulties, Bayesian inference could for a long time not be used on a large scale, because the computational burden in computing the posterior was too high. This is mainly caused by the difficulty in computing the marginal likelihood which is potentially a large sum or integral. This changed with the introduction of modern computers and a class of algorithms named Markov Chain Monte Carlo (MCMC) algorithms. The original algorithm was introduced in 1953 [12] and has had a huge impact on many scientific fields over the years.

This algorithm can be used to draw samples from any given distribution, $P(x)$. In fact, it suffices to provide a function, $f(x)$, that is proportional to the desired target distribution. The algorithm starts at some initial point, x_0, and then proposes a new point, x_1, nearby. Then the ratio $r = f(x_1)/f(x_0)$ is computed (and possibly corrected with a Hastings ratio [13]) and the new point is accepted with probability $\max(1, r)$. This can be repeated until a sufficient number of samples are drawn. By viewing this process as a Markov chain, one can show that – under some reasonable assumptions – it produces samples from the desired distribution in the limit of infinitely many steps. Since we only require f to be proportional to P, we can disregard the marginal likelihood when sampling from the posterior. Another advantage of MCMC algorithms is that they also provide us with estimated marginal posterior densities of the model parameters. In fact, all we have to do is take the samples for that parameter while disregarding the others to get a marginal estimate of the parameter's distribution that incorporates all remaining uncertainty.

1.3 Bayesian phylogenetic inference

Using Bayesian methods for phylogenetic inference brings along advantages such as seamless integration of complicated models and prior knowledge. As discussed above, the latter can also be seen as a weakness depending on a philosophical standpoint or on the amount of available information. This introduction to Bayesian phylogenetic inference is very brief. While I leave out many mathematical details, I hope that the reader will understand *in principle* how we can address various biological questions – phrased in the form of inference of model parameters – using nothing but sequence data and a bit of calibration as our data.

The posterior We can write the posterior very closely related to the way we wrote Bayes' theorem in eq. 1.2:

$$P(\mathcal{T},\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{T},\theta)P(\mathcal{T},\theta)}{P(\mathcal{D})}, \qquad (1.3)$$

where we expanded the hypothesis, $\mathcal{H}$, to the tree, $\mathcal{T}$, and a parameter vector, θ. What exactly is contained in θ depends on the chosen models, but the computation of the priors is straightforward since they are usually specified using one of the standard probability distributions. How to design a prior for $\mathcal{T}$ is described below. The likelihood $P(\mathcal{D}|\mathcal{T},\theta)$ can be computed by the pruning algorithm as explained above and the marginal likelihood, $P(\mathcal{D})$, does not have to be computed when we are using an MCMC algorithm. Different implementations of such a phylogenetic MCMC algorithm exist in a form that makes them easy to use by individual researchers (e.g. [14–16]).

Tree priors We still need to specify a prior over tree space, i.e. we need a way to assign a probability to a topology and branch lengths. Generally, this is done by designing a model that generates a tree and then determining how likely a given a tree is under that model. Two different approaches exist for this. The first one is the *coalescent* which is a theoretical result for a Wright-Fisher process with a large population [17]. A Wright-Fisher process describes an idealized population with synchronous change of generations in which each individual "picks" one parent at random (see e.g. [18]). Given the population size, the coalescent allows us to compute the probability of a tree. The second approach is the so-called *birth death process*. While the coalescent is a "backward-in-time" model that starts at the tips of the tree, the birth death process is a "forward-in-time" model that starts out with a single ancestor and then models branching events according to a set of population parameters. Various extensions exist for both models, including non-contemporary tip dates or changes in the parameter values over time.

1.4 Clock rate, substitution rate and mutation rate

In this thesis, I will talk a lot about inferring rates of evolution which are measured in evolutionary change/site/time. It is useful to distinguish between three different such rates:

- Clock rate: a model parameter that is needed to compute the likelihood for sequence evolution along a branch in a temporal phylogenetic tree
- Mutation rate: the rate at which errors happen in the DNA or RNA duplication process of a cell / virus; mainly a biochemical "parameter" of the polymerase
- Substitution rate: the rate at which substitutions (i.e. fixed mutations) occur in a population

It is worth to state some details about the definition of the substitution rate. *Fixation* means that within one population a mutation completely replaces the original nucleotide – either because it is advantageous or by random genetic drift. We therefore require a concept of a *population*. In the beginning of phylogenetics, when macro-evolution was the main subject of interest, populations were simply the different species. Generally, however, what is considered a population depends on the scale of the analysis. Overall, the substitution rate is a complex measure that is influenced by the mutation rate but also by generation time, population size and fitness [19].

When the clock rate is estimated from genetic data, it usually falls somewhere between the mutation and the substitution rate. For short time frames it is closer to the former, for long time frames closer to the latter. This is because on short time scales, when the most common ancestor of the samples is not too far in the past, we still observe many *slightly deleterious mutations*. Those mutations produce a viable organism but result in a slight decrease in fitness. Over longer time frames, these mutations will be purged from the population again by a process called *purifying selection*. At that point, most of the observed genetic diversity stems from actual substitutions. A more detailed example is presented in [20].

For viruses, the primary organisms of interest in this thesis, defining a substitution rate is difficult. Firstly, there are many options when defining populations. A population could e.g. contain all the viruses in one patient, one city or one country. On top of that, it is "intrinsically contradictory" [21] to talk about fixation within any viral population given their quasi-species nature. They mutate so rapidly that they are in a constant transient state with many mutations occurring concurrently at any time.

We can, of course, infer the *clock* rate from viral datasets, e.g. by sampling the viral sequences in different patients during an outbreak of a viral disease. But interpretation of that parameter should be done with care – especially when comparing it across time-scales or populations. Still one of the first studies for the 2013–16 outbreak of Ebola in West Africa stated the following about their inferred clock rate [22]:

> The observed substitution rate is roughly twice as high within the 2014 outbreak as between outbreaks. Mutations are also more frequently nonsynonymous during the outbreak. Similar findings have been seen previously and are consistent with expectations from incomplete purifying selection.

With the last sentence, the authors tried to put their findings into context. Yet, their paper triggered the headline "Ebola virus mutating rapidly as it spreads" [23] in the Nature news section and caused concerns among scientists and public health officials. Further studies [24, 25] later on revealed that the longer term "substitution" rate of the virus within that epidemic proved to be the same as the estimate between outbreaks. This shows the need for careful interpretation of these parameters which can only happen when we properly understand the underlying models and their interaction.

1.5 Motivation

The decline in inferred clock rate with increasing sampling period caused some debate in the recent years [20, 26–28]. While it is widely acknowledged that purifying selection contributes to this phenomenon, the degree to which other factors like sampling biases and model misspecifications play a role is less clear. The thesis aims to contribute to this understanding.

In chapter 2 we use a simulation study to investigate the influence of the tree prior on clock rate inference and show that it can lead to inflated estimates. With sufficient data, this bias can be overcome and therefore a misspecified tree prior may contribute to the observed decline in rate. We also demonstrate that the choice of tree prior can influence clock rate inference on an empirical dataset. In chapter 3 we find some evidence for purifying selection in an Ebola dataset and introduce a modified relaxed clock model that may be helpful for researchers in the future.

Chapter 2

The interplay of tree prior and clock rate estimation

2.1 Introduction

Bayesian inference is a powerful tool for the study of phylodynamics. It allows seamless integration of complicated models with various parameters along with varying degrees of uncertainty. Using nothing but nucleotide sequences and temporal calibration, one can answer questions about past speciation times, past migration events of species or pathogens infecting host populations and temporal changes in population parameters such as speciation or transmission rates. Rather than point estimates, we can compute marginal posteriors of our parameters of interest which provide full distributions, incorporating the overall uncertainty in the model provided the model fits the data. In this chapter we will focus on the inference of the clock rate parameter that determines how quickly nucleotide changes happen along branches in the tree.

While the Bayesian phylogenetic framework as a whole is conceptually straightforward, carrying out an analysis can be very complex and thus dedicated software tools have been developed [14–16]. Sequence data alone allows us to infer genealogies in which branch lengths correspond to the expected number of substitutions along that branch. External calibration is required to compute the length in temporal units. For studies of macro-evolution, fossils [29] or ancient DNA (aDNA) [30] are frequently used, whereas for fast-evolving pathogens, like RNA viruses, serially sampled data can be sufficient when evolution happens on the time scale that was sampled [31, 32].

Applying phylodynamic tools to pathogens allows insights into different aspects of the pathogen's evolutionary and transmission dynamics. We can infer the effective reproductive number (R_e), the time of origin or the substitution rate. All of them can be of interest to public health officials who need to decide where and how to react to an epidemic. While R_e and the time of origin are straightforward to define, the concept of a substitution rate for viruses is less clear. Usually, the substitution rate is defined as the rate at which mutations become fixed in a population [19]. This however requires to define populations and fixations, both of which are problematic terms when talking about an ongoing epidemic [21, 33].

As a model parameter, the clock rate is well defined and together with the branch length determines the expected amount of nucleotide change along that branch. However, many models and analyses acknowledge the fact that there are multiple different rates varying between branches and sites. Comparison of clock rates should therefore be done very carefully and only where applicable. Nevertheless it was shown for many different real datasets stemming from viral outbreaks that the clock rate decreases as the sampling period is increased [34]. This was also observed in real time during the 2013–16 Ebola epidemic in West Africa. One early study [22] caused great concern among scientist and public health officials who were afraid of an Ebola virus that was accumulating substitutions much more rapidly than previously observed and would thus evade screenings and render vaccines and medication useless [23]. Further data collection, however, made it evident that substitutions were occurring at the same speed as long term observations suggested [24, 25, 35].

The fact that the clock rate estimate depends on the time scale that is used for calibration does not only hold for viruses and has first been observed and publicized more than ten years ago [36]. In that paper, Ho et al. suggest that the most likely cause is incomplete purifying selection: On shorter time scales slightly deleterious mutations are still observed in the data and artificially inflate the clock rate. Over a longer timeframe these mutations are purged due to purifying selection (see [20] for an illustrative example). It was quickly shown, though, that purifying selection alone cannot explain the observed decline [37]. Multiple other factors such as calibration errors, model misspecification and sequencing errors can all contribute to inflated clock rate estimates as well (see [20] for a review). The debate about which of these factors contribute to which degree is still very much ongoing, in particular with regard to the question how big of a role purifying selection plays [26–28]. To understand the complex interplay, simulation studies and analyses of real datasets are both important.

For a phylogenetic analysis in a Bayesian framework we need to specify at least a clock model, a substitution model and a tree prior. All these components interact in a way that is sometimes non-intuitive. There have been efforts to make it easy for researchers to select the best clock and substitution model [38–43] but less so with respect to the right tree prior. However, even if we are only interested in the clock rate and integrate out the uncertainty in tree space, the tree prior can have an impact on the posterior distribution. Even though the models for the clock and the tree are independent components of the analysis, the tree length (i.e. the sum of all branch lengths) and clock rate are highly negatively correlated as their product needs to explain the overall diversity that is observed in the data. While we put an explicit prior directly on the clock rate this is not true for the tree length. Rather, the tree length obtains a prior indirectly from the specified tree prior. This influence has not been studied in detail, except for some analytical results for a coalescent model with contemporary tips [44]. Results for serially sampled tips or for birth death processes are to our knowledge not available.

With this chapter we wish to contribute to the current debate about clock rate inference in Bayesian phylogenetics. To this end we point out some non-trivial conceptual issues using a simulation study. New models for tree priors are regularly investigated using simulation studies in which the model itself or more simple models are used to generate the data [45–47]. While this is a valuable contribution to show that the model can recover true values under ideal circumstances, it offers no information about the robustness of the inference to violations of the underlying model assumptions. In our simulation study, we use an empirical rather than a simulated tree and simulate sequence evolution on that tree. This allows us to assess the robustness of the inference of the (known) mutation rate from the simulated sequences when the tree prior potentially poorly models the underlying tree. We obtained the tree from an analysis of sequences from Guinea during the latest Ebola outbreak. We then analyse this dataset as well one of the earliest datasets of the outbreak [22] using different tree priors and show that for the latter a much lower estimate of the clock rate is obtained when using a model accounting for structure within the population.

2.2 Methods

2.2.1 Simulation study

We simulated sequence evolution along a fixed tree using very simple clock and substitution models and subsequently analysed the resulting alignment using the same models and a constant population size coalescent tree prior. To obtain the tree, we used the coding part of 236 whole genome sequences of the Ebola virus from patients in Guinea sampled during a period of 10 months. These data have previously been described in [48,49]. We removed 3 Guinea sequences with unknown district from [49]. For the analysis in Beast2 we used a serially sampled birth death skyline model with 3 intervals for R0 and the sampling proportion along with an HKY substitution model and a lognormal relaxed clock. We ran the chain for 100 million steps and discarded 10% of burn-in. We used `treeannotator` which is included in Beast2 to create the maximum clade credibility tree which is shown in Fig. 2.1.

We then used the sequence simulator included in Beast2 and a Jukes-Cantor model (JC69) [50] with a fixed clock rate of 0.1 substitutions/site/year (s/s/y) to simulate sequences of length 100, 500, 1000 and 15000 base pairs. We used the alignment obtained from the simulations as input to a Beast2 analysis in which we also employed a strict clock and JC69 model. We chose a normal distribution with standard deviation of 0.02 around the true value of 0.1 as a prior for the clock rate and a lognormal distribution with $M = 0$ and $S = 0.5$ for the population size of a constant size coalescent. We also used different coalescent and birth-death priors to check the robustness of our findings. The specific settings are highlighted in the respective figure captions. For each sequence length we did 10 independent simulations. The parameters in the setup of this simulation are not meant to be directly biologically relevant but rather an illustrative example. Chains were run for 10^8 steps, except for the alignment of length 15000 for which they were run for 6×10^7 steps. We used R [51] and the `coda` package [52] to analyse posterior samples of the clock rate, tree height, tree length and total divergence (product of clock rate and tree length) after discarding 10% of burn-in and verifying that the effective sample size was above 200 for all parameters. Plots were also created in R using the package `ggplot2` [53].

2.2.2 Inference under different tree priors

We investigated the dependence of clock rate inference on the chosen tree prior on two Ebola datasets. The first one is the same as the one used to generate the tree for the simulation study and we will refer to it as the Guinea dataset. For the analysis under a structured coalescent model, we assigned the tips to four clades based on their location in the tree that was used for the simulation. The second dataset comes from an early publication of the Ebola epidemic in West Africa in 2014 which concluded that the substitution rate for the virus was about twice as high during that outbreak as between outbreaks [22]. It contains

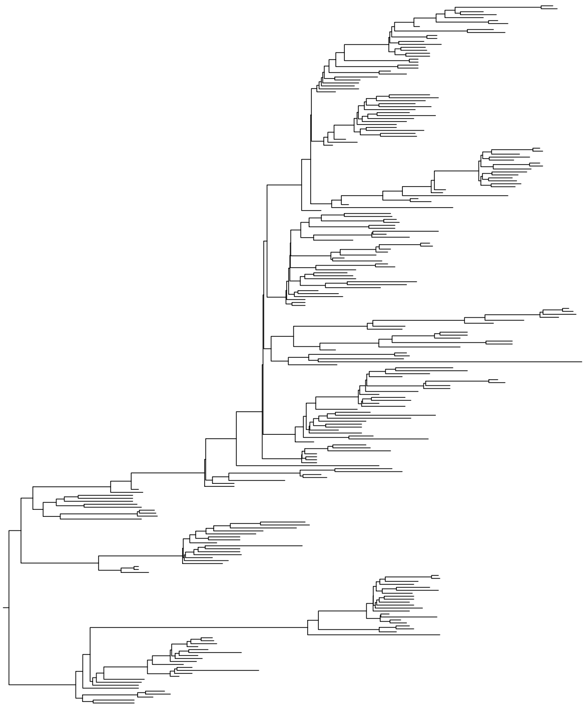


Figure 2.1: The tree that was used for the simulation study.

whole genome data (i.e. almost 20 k sites) of 81 sequences sampled during three months. 78 of the sequences were sampled in Sierra Leone by [22] and 3 had been sampled previously in Guinea [54]. For the structured coalescent analysis, the sequences were assigned to three clades based on the location (one for Guinea and two for Sierra Leone). We will refer to this dataset as the Sierra Leone dataset.

For the analysis in Beast2 we used a strict clock and an HKY substitution model [55] without site heterogeneity. For the Sierra Leone dataset this follows the analysis that was carried out in the original publication. We used six different tree priors: Constant rate birth death, birth death skyline [45], constant population size coalescent, exponential growth coalescent, coalescent skyline [47] and structured coalescent [56]. The distributions and parameters for the priors are listed in the supplementary material (Table A.1). We ran all chains for 10^8 steps, except for the structured coalescent analysis of the Guinea dataset which was run for 5×10^8 steps. We always discarded 10% burn-in and verified that the effective sample size for all parameters was above 190.

We subsequently used path sampling [57] to assess the relative goodness of fit of the different models. We used the implementation in Beast2 with 16 steps, $\alpha = 0.3$, a chain length of 5×10^7 steps and a burn-in of 20%.

2.3 Results

2.3.1 Simulation study

The median estimate and 95% HPD intervals for clock rate, tree height, tree length and total divergence for each replicate of the simulation study are shown in Fig. 2.2. The dashed lines in each panel indicate the true values. The HPD intervals for tree height and total divergence almost always include the true value and become smaller with increasing sequence length. The estimates for clock rate and tree length, however, are biased upwards and downwards, respectively. The bias and the variance decreases as the sequence length increases, but the true value is only covered by the HPD intervals of some runs for sequence length 15000. Without any sequence data, the inferred values are close to the truth (see Fig. A.1).

For the first out of ten replicates, Fig. 2.3 shows the posterior distribution of topologies using the `treescape` tool [58] after down-sampling to 101 trees per sequence length and discarding 10% of burn-in. `treescape` turns topologies of a set of labelled trees into a real valued vector representation, computes the pairwise Euclidean distance between all these vectors and then does a principal component analysis to visualize the distances in two dimensions. For this plot we also included data from a sim-

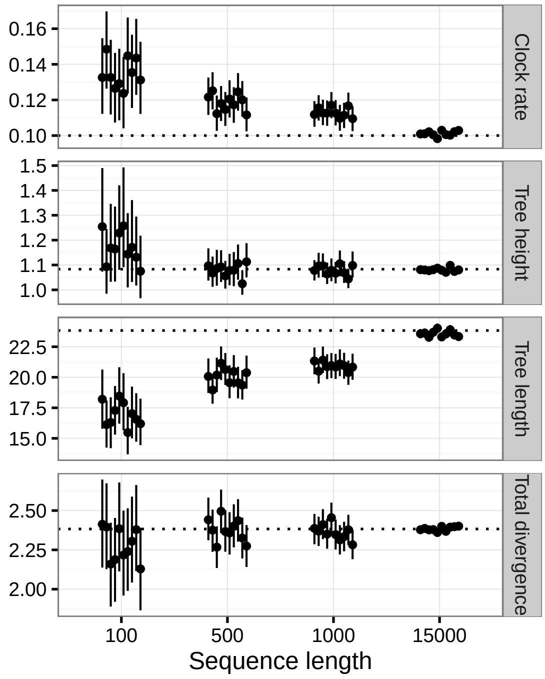


Figure 2.2: Median values and 95% HPD intervals for some parameters in the simulation study. The dashed lines indicate the true values. Clock rate is given in substitutions/site/year, tree height and tree length in years and total divergence (product of clock rate and tree length) in substitutions/site.

ulation without sequence data (i.e. sequence length is 0). The points representing topologies obtained with sequence data form a cluster around the true topology (shown with a red cross) while the topologies coming from the analysis without sequence data are clearly separated.

The results from the simulation study are robust towards changing the tree prior. As displayed in Fig. A.3, changing the coalescent population size prior to either very fast or very slow exponential growth does not alter the clock rate estimates. Further, using a birth-death prior with either a very high or very low basic reproductive number (i.e. fast or slow growth) does not change the observed pattern either (Fig. A.4).

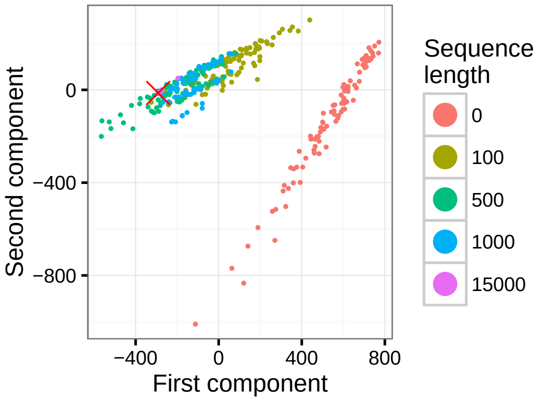


Figure 2.3: Distribution of topologies of the posterior tree samples for analyses of simulated datasets of different sequence lengths using the `treescape` plotting tool [59]. The red cross marks the true tree.

Fig. A.5 shows an analysis of the simulated alignments in a maximum likelihood framework using the tools RAxML [60] and least-squares dating [61]. For sequences of length 500 and more, the true value is within one standard deviation of the inferred mean.

2.3.2 Inference under different tree priors

Fig. 2.4 shows the results for the two Ebola datasets. For the Guinea dataset the birth death model leads to the highest clock rate with a median of roughly 1.3×10^{-3} s/s/y. Under all the other models the inferred rate is slightly below 1.2×10^{-3} s/s/y. The HPD intervals, however, are largely overlapping. The inferred tree length shows the opposite trend with the birth death model leading to a median estimate of 17 years whereas the other models result in an estimate between 18 and 19 years. For the tree height the large HPD intervals for the constant and structured coalescent stand out. These two also have the highest median of around 1.1 years, while the lowest estimate of slightly below 1 year is obtained under the birth death model. For total divergence there is no noticeable difference between any of the models.

For Sierra Leone all of the unstructured models lead to a median estimate of the clock rate of about 2×10^{-3} s/s/y which is a lot higher than the long term rate during that epidemic (approximately 1.25×10^{-3} s/s/y [24]). The structured coalescent model results in a median rate of 1.3×10^{-3} s/s/y. The opposite trend is observed for the tree length. While the unstructured models result in estimates of around 1.5 years, the structured coalescent leads to a median value of 2.3 years. For the tree height the medians for the unstructured models is around 0.3 years but the birth death models result in much narrower HPD intervals than the coalescent models. Using the structured coalescent we obtain even more variance around a median of 0.4 years. Again we find no noticeable difference for total divergence.

Fig. 2.5 shows the results of the model comparison. For Guinea the structured coalescent is clearly the best model, with a Bayes factor between 60 and 70 compared to all the other models. We checked the robustness of this finding by running path sampling with a varying number of steps (Fig A.6). While the structured coalescent always presents the best fit, the ranking of the other models varies. For the Sierra Leone dataset we find that the birth death and the skyline coalescent models present the worst and best fit, respectively. However, the log Bayes factor of the birth death model is only around -1.5 and for a different number of steps, the structured coalescent model was sometimes estimated to be the worst (Fig. A.6). The skyline coalescent and the birth death skyline, on the other hand, yield consistently high performances.

We also analysed an influenza dataset containing 273 sequences sampled during only one month [62] and find little difference in the inferred clock rate but some variation in the estimated tree height. Details are presented in the supplementary material (Section A.1 and Fig. A.7).

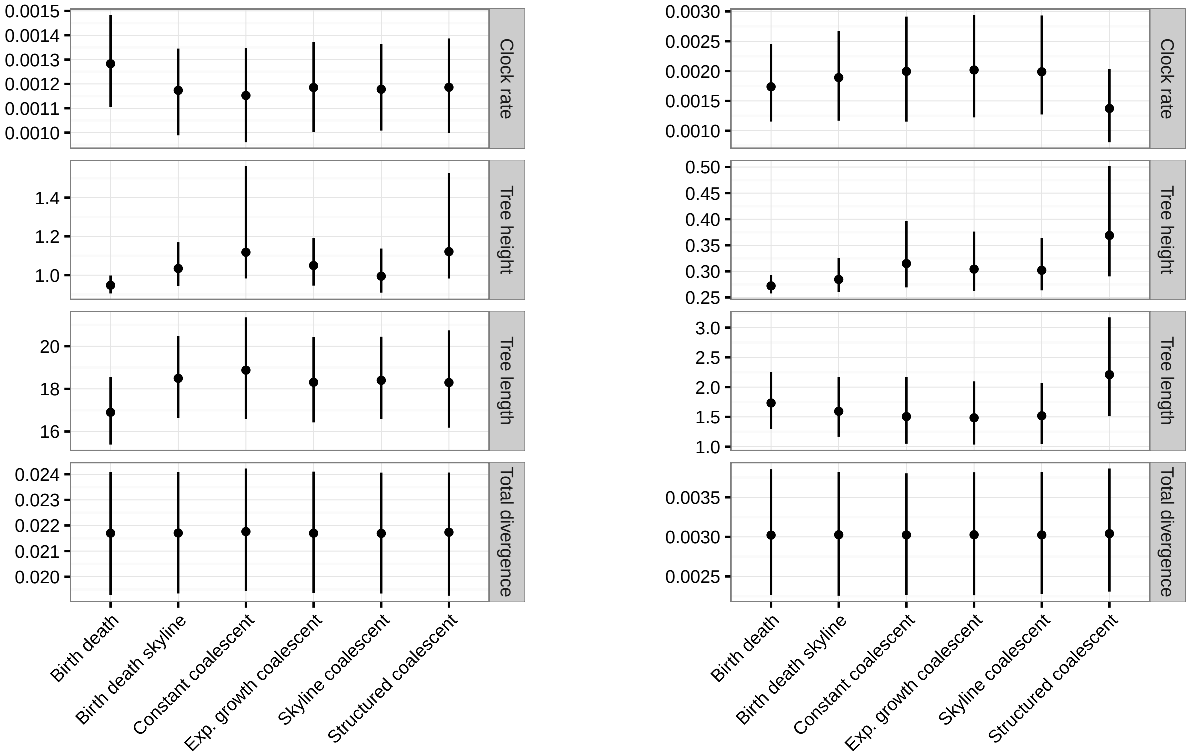
2.4 Discussion and conclusion

The simulation study shows that when simulating along a tree that is based on real data it can be surprisingly difficult to recover the true clock rate, even when very simple clock and substitution models are used. This, despite the fact that our prior on the clock rate is correct. The problem arises from a misspecification of the tree prior (see Fig. 2.3 and A.1) which will be difficult to detect in real datasets where the truth is unknown. The tree prior implies an indirect prior for the tree length which then leads to a bias in the clock rate estimate. Matters are complicated further by the fact that, in this particular case, the tree prior seems appropriate when looking at the tree length as an aggregate statistic, since the bias disappears when sampling without sequence data (Fig. A.1).

These results may initially seem counter-intuitive. To explain them, we will briefly review the Bayesian phylogenetic framework. Let the clock rate be denoted by μ, the tree by $\mathcal{T}$, the remaining parameters by θ, the tip dates by τ and the data (a multiple sequence alignment) by $\mathcal{D}$. In a simple form, the posterior is given by

$$P(\mu, \mathcal{T}, \theta | \mathcal{D}, \tau) \propto P(\mathcal{D} | \mu, \mathcal{T}, \theta) f(\mathcal{T} | \tau) f(\mu) f(\theta). \tag{2.1}$$

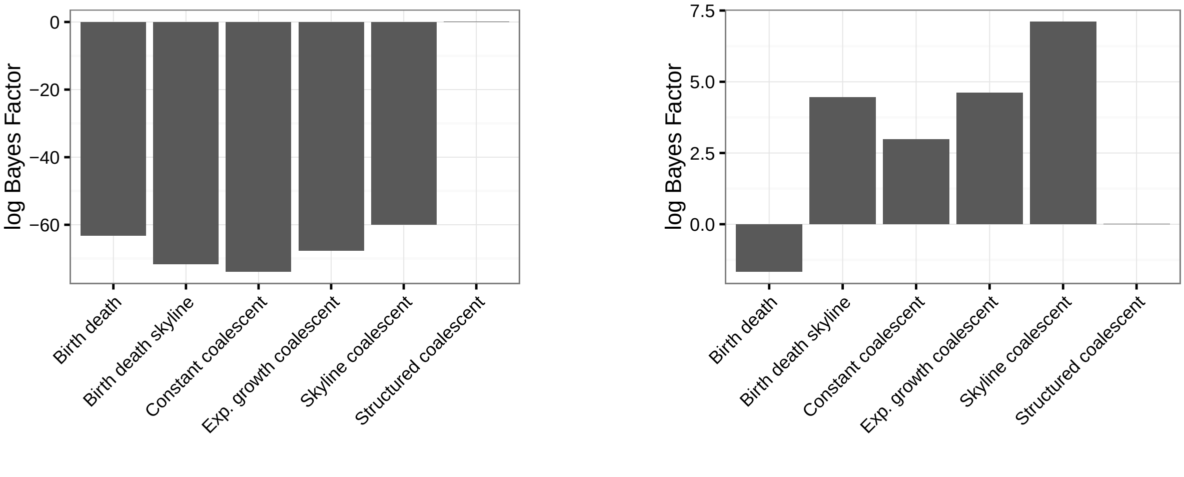
Here, we chose to condition the tree prior on the tip dates rather than to include them in the likelihood. Without any sequence data, we have $P(\mathcal{D}|\mu, \mathcal{T}, \theta) = 1$ and therefore, when sampling from the posterior, $\mathcal{T}$ and μ can change independently. Adding only a little bit



(a) Guinea

(b) Sierra Leone

Figure 2.4: Median and 95% HPD intervals for clock rate, tree height, tree length and total divergence inferred from the two datasets under different tree priors. For units refer to the caption of Fig. 2.2.



(a) Guinea

(b) Sierra Leone

Figure 2.5: Log Bayes factors of all models compared to the structured coalescent model. Higher values indicate a better fit.

of sequence data can make a large difference in the inference as it links the tree and the clock rate via $P(\mathcal{D}|\mu, \mathcal{T}, \theta)$. If the tree prior causes the tree length to be underestimated, the model will compensate for this by increasing the clock rate to explain the overall diversity in the data.

The prior distribution on the tree space, $f(\mathcal{T}, \tau)$, is a distribution over topologies and branch lengths. This indirectly gives rise to the prior on the tree length which is depicted in Fig. A.1. `treescape` allows computation of a median tree from a set of trees which can be used as a representative tree similar to the maximum clade credibility tree. The median tree without sequence data is depicted in Fig. A.2 and looks very different from the original tree (Fig. 2.1), despite the tree length being almost the same. Upon adding sequence data the topologies that before had a high prior support become very unlikely (Fig. 2.3) and under this constrained topological space, the indirect prior on the tree length is altered as well. It is the change in the likelihood of topologies that causes the downward bias in tree length for sequence lengths 100, 500 and 1000 which in turn causes the clock rate to be overestimated. This bias can be overcome if sufficient data are added.

We see that the tree height can be estimated much more reliably than the tree length (Figs. 2.2 and A.3). This is because information coming from all sequences can be used for its estimation. Like the total diversity, it is in that sense a global parameter and little data is already informative about it. There are, however, many trees of the same height but with different lengths (see e.g. Fig. 2.6) and inferring the length correctly is a much harder problem and thus more susceptible to biases from the tree prior. The same holds true for the topology. Fig. 2.3 shows that the tree prior alone produces topologies that are very different from the true topology. It also shows, however, that for short sequences, there is still a lot of uncertainty in the topology and while we do move away from the prior towards the truth, the tree prior still causes a bias. This is important to keep in mind in studies with little data where the tree topology is treated as a nuisance parameter since bias in topology can result in biased inference of other parameters as well. Our simulation indicates that none of the non-structured tree priors can overcome the problem (Fig 2.2, A.3, A.4) – even with very different priors for the dynamics in the underlying non-structured population, too much weight is given to certain tree topologies.

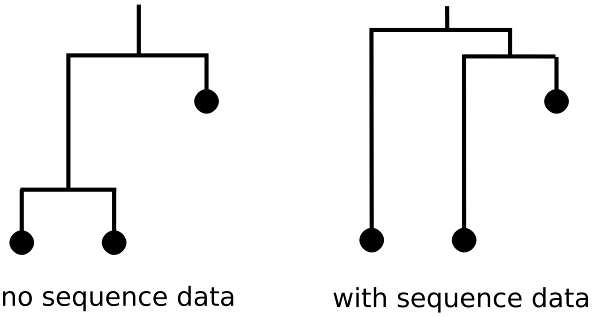


Figure 2.6: A toy example of how the sequence data can influence the branch length via changing the topology.

The influence of sequence data on the topological space and on the tree length can be illustrated with a toy example (Fig. 2.6): Consider a tree with two contemporary and one past sample. For a small population size a coalescent tree prior would give high probability to the tree topology in which the two contemporary tips form a cherry. When sequence data are added, it may become obvious though, that the cherry should be be formed between one of the contemporary tips and the sample from the past. This effectively puts a lower bound on the tree length.

The practical problem for researchers is that they obviously do not know the true value for clock rate and tree length. The usual approach is to sample from the posterior without sequence data and compare the resulting distribution for the parameters of interest with the one obtained including sequence data. This approach would fail for the clock rate and the tree length on the simulated datasets for sequence lengths smaller than 15000 which would result in false confidence in wrong estimates. To make things worse, comparing the prior against the posterior for short sequences shows an increase in the clock rate. Thus, bearing in mind that we only have little data, one would typically conclude that more data would lead to an even higher clock rate.

These observations are similar to recent publications that point out the complexity of defining a correctly calibrated prior for divergence time estimation [63–65]. In that case the convolution happens between the calibration of internal nodes and the tree prior which can lead to an unexpected realization of the prior and result in biased estimates and too narrow confidence intervals. In our example the convolution only happens when sequence data are added that contain a strong signal about the topological space which in turn leads to an unexpected constrain on the tree length. We also showed that a maximum likelihood analysis which does not employ an (explicit) tree prior did not suffer from the bias reinforcing our interpretation that it is the tree prior that causes the problem.

The analysis of the two empirical datasets also confirms that the tree prior can influence the inferred clock rate: For Sierra Leone we see that the choice of tree prior can heavily influence the estimated clock rate. If a structured model had been used in the original analysis, then the difference between the short and long term estimates would have disappeared. Simulation studies in the context of aDNA have shown that complex population structure in the past can lead to biased estimates of clock rate if the data is analysed under a model that is too simple [66]. We do not claim that the estimates using the structured model are correct in this case. Even though the sequence data came from two different countries and a structured model therefore represents a reasonable choice, the model comparison suggests that it has a rather poor fit compared to the other models. Instead, the data seem to demand a model which allows a change of the parameter values over time. However, correct estimation of marginal likelihood is a difficult and computationally demanding task and the results should therefore be taken with a grain of salt. Also, marginal likelihoods and Bayes factors say nothing about the *absolute* goodness of fit [67]. This can only be done with computationally even more demanding methods like posterior predictive simulations [68]. Regardless, our results show the importance of carefully choosing a tree prior and that this choice can strongly influence the clock rate estimates.

For the Guinea dataset the median under each model is contained in the HPD intervals of all others, but the birth death model still leads to a markedly higher estimate. This, despite the fact that the data cover a span of 10 months and should therefore suffer less from short term biases. Also for the inference of tree height the choice of tree prior would have played a big role. While this is a more obvious conclusion, it underlines the need to check the influence of the tree prior on the inference and to check overall goodness of fit of the model. We should further note that the assignment to clades was not so clear a-priori. Even though the data suggests that the structured model already fits the data the best, it may therefore be fruitful, to additionally apply a structured model that does not rely on classified tips (e.g. [69] or [Barido-Sottani & Stadler, in prep.]).

In this chapter we pointed out some conceptual problems in the inference of the clock rate when using Bayesian phylogenetic tools. The interaction between tree prior and clock rate estimation can be complex and non-intuitive. We used a simulation study to demonstrate that deviation of the posterior clock rate distribution from the prior does not necessarily imply a signal in the data and can be a mere artefact of the chosen tree prior. The reanalysis of an Ebola data from Sierra Leone showed that the high mutation rate that was reported originally could be wrong and that the inferred rate under a different tree prior comes very close to the long term estimate. Overall this stresses the need to choose the tree prior carefully even if the parameter of interest is the clock rate and demands further investigations on how overall model fit in Bayesian phylogenetic analyses can be assessed.

Chapter 3

Purifying selection during the Ebola epidemic in Guinea

3.1 Introduction

One of the key biological factors that is often called upon to explain a decrease in inferred clock rates with increasing sampling period is purifying selection (see [20] for an excellent review of this and other possible factors). The reasoning is that most (non-lethal) mutations that occur in a population are slightly deleterious (i.e. they lead to a decrease in fitness) and will therefore be removed again by purifying selection. On a short time-scale, we still observe more of those mutations and thus infer a higher rate than over a longer period of time. By the same reasoning, we expect to find more slightly deleterious mutations and thus a higher clock rate on pendant branches of a phylogenetic tree.

We wanted to test if we can detect purifying selection in a dataset from the 2013–16 Ebola epidemic in West Africa. To this end, we used a relaxed clock model and analysed the estimated clock rate on internal and pendant branches. Since purifying selection will only act on non-synonymous mutations, we estimated the clock rate parameters for codon position 1, 2 and position 3 separately. This is not perfect, as some mutations on position 1, 2 can be synonymous and others on position 3 can be non-synonymous, but it is methodologically convenient and a good approximation. We also employed a constrained relaxed clock model, in which only two different clock rates are used throughout the tree, hoping to amplify the difference between internal and pendant branches.

3.2 Data

We used the coding part of the viral genome sampled from 236 patients in Guinea over the span of ten months starting from March 27, 2014. This is the same data that was used for the simulation study in chapter 2. The sequences were divided into six intervals which are temporally equally spaced, each containing between 19 and 94 sequences (see table 3.1 for details). We then analysed the cumulative dataset, mimicking how one would analyse an on-going epidemic as more and more data become available. For this dataset, an overall decrease in clock rate estimation with increasing sampling period has been observed [70].

Table 3.1: The table shows the latest date for each interval as well as the number of new sequences added during that interval and the total number of sequences. The first sequence was sampled on 2014-03-27.

Interval	Last date	No. of new sequences	Total no. of sequences
1	2014-05-14	31	31
2	2014-06-25	19	50
3	2014-08-28	33	83
4	2014-10-19	94	177
5	2014-12-09	33	210
6	2015-01-31	26	236

3.3 Analysis using a relaxed clock

The analysis was done in Beast2 using a birth death skyline model [45] with $i+1$ intervals for R0 and the sampling proportion when analysing data up to interval i. We used an HKY substitution model and a relaxed lognormal clock and estimated the parameters for those models separately for codon position 1, 2 and codon position 3. Details about the prior distributions of the parameters are listed in table 3.2. We ran the chains for 50 million steps and discarded 10% of burn-in. Except for the full dataset (i.e. interval 6) the ESS for all parameters were larger than 200.

Instead of reporting either the relaxed clock mean or the mean rate across all branches, we analysed the rates for internal and pendant branches separately. To this end, we extracted the weighted average of the clock rate on internal and pendant branches based on their length for each posterior tree sample. Formally, for each set of branches, B, we computed an aggregated clock rate with

$$r^{(B)} = \sum_{b \in B} r_b t_b / \sum_{b \in B} t_b, \tag{3.1}$$

where r_b and t_b are the rate and length of branch b, respectively (see eq. (8) in [4]). This captures how much genetic change happens per time across all branches of the respective type.

Median clock rate estimates in substitutions/site/year (s/s/y) and 95% HPD intervals are shown in Fig. 3.1. We can see for the internal branches on position 1, 2 and both branch types on position 3 that the clock rate decreases with increasing sampling period. In those three cases, the final median is barely contained in the first HPD interval. The results for internal and pendant branches are nearly indistinguishable for position 3 but show some difference on position 1, 2 with pendant branches having a higher median clock rate than internal branches for intervals 2 through 6. Since changes on codon position 1, 2 are much more likely to cause a non-synonymous change in the protein, this agrees with our expectations and indicates that, indeed, purifying selection acts mainly on the internal branches.

It should also be noted that the absolute rates are quite different on position 1, 2 and position 3. For the first interval, they are around 1.1×10^{-3} and 3.6×10^{-3} s/s/y, respectively, for both branch types. On position 1, 2 the rate on the internal and pendant branches decrease to 0.6×10^{-3} and 0.9×10^{-3} s/s/y, respectively. On position 3, the final rate for both branches is around 1.9×10^{-3} s/s/y.

Table 3.2: Priors used for the serial birth death skyline analysis.

Parameter	Prior distribution
Clock rate mean	LogNormal(-6.6, 2)
Clock rate standard deviation	LogNormal(0.05, 1.2)
Kappa	LogNormal(1, 1.25)
R0	LogNormal(0, 1.25)
Become uninfectious rate	Gamma(19.734, 1.36)
Sampling proportion	Beta(2, 10)
Origin	Uniform(0, ∞)

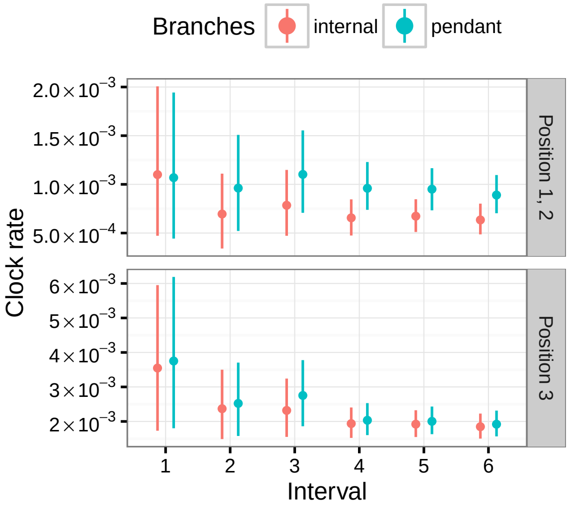


Figure 3.1: Median clock rate estimates in substitutions/site/year and 95% HPD intervals for increasing number of sequences. Separate relaxed clock models were used for codon position 1, 2 and codon position 3. The rate estimates were then extracted from the branches of the posterior tree sample (see eq. 3.1) and are presented separately for internal (red) and pendant (turquoise) branches.

3.4 Analysis using a constrained relaxed clock

We applied a variation of the relaxed clock model which only allowed two different rates across the tree to the same data. We were hoping to amplify the signal for the difference between internal and pendant branches or be able to generally detect certain types of branches that would be fast or slow. A clock model that explicitly models two different rates is not available. However, we show that when using a relaxed lognormal clock restricted to two rate categories, BEAST2 effectively allows independent choice of the two rates (see Appendix D).

For this analysis we used an exponential growth coalescent, an HKY substitution model and the constrained relaxed lognormal clock. We again estimated the parameters for codon position 1, 2 and position 3 separately. The prior distributions of the parameters are listed in table 3.3. We ran the chains for 50 million steps and discarded 10% of burn-in. Almost all parameters had an ESS of above 200.

Fig. 3.2 shows the clock rate estimates for the fast and slow rate on position 1, 2 and position 3. The decrease is clearly visible for the slow rate at position 1, 2 and the fast rate at position 3, whereas the other two only show a slight decline. As expected, HPD intervals become smaller with increasing data. We can see that the third interval appears to be an outlier from the general trend. This will be investigated further below.

We can now also directly compare how often pendant and internal branches are assigned the two different rates. The relative frequencies are shown in Fig. 3.3. Overall the distribution is quite balanced with around half the internal and half the pendant branches using the slow

Table 3.3: Priors used for the exponential growth coalescent.

Parameter	Prior distribution
Clock rate mean	LogNormal(-6.6, 2)
Clock rate standard deviation	LogNormal(0.05, 1.2)
Population size	One on x
Growth rate	Laplace(0.001, 2.246)

and fast rate. However, it is noticeable that there is a tendency for the internal branches on position 1, 2 to be slow which is in agreement with the findings in the previous section and again indicates purifying selection.

To understand why the estimates from the third interval do not follow the general trend, we looked at the maximum clade credibility trees for intervals 2, 3 and 4. These were plotted with ape [71] in R and are shown in Fig. 3.4. The tips are coloured according to the interval they belong to. We can see that most of the sequences that are added in the third interval form a new clade. This changes the overall shape of the tree which now appears to have more structure. The clade at the bottom of the tree represents a burst-like event which is probably not captured well under the exponential growth coalescent (cf. chapter 2). While the fourth interval adds another burst, most of the sequences are added to the clade that was created in the third interval. This is interesting, because we can think of the tree in interval 3 to represent two separate outbreaks for which we do not have much data. In that case we find a strong signal for a fast clock rate. Interval 4 then adds more data and the rate decreases again.

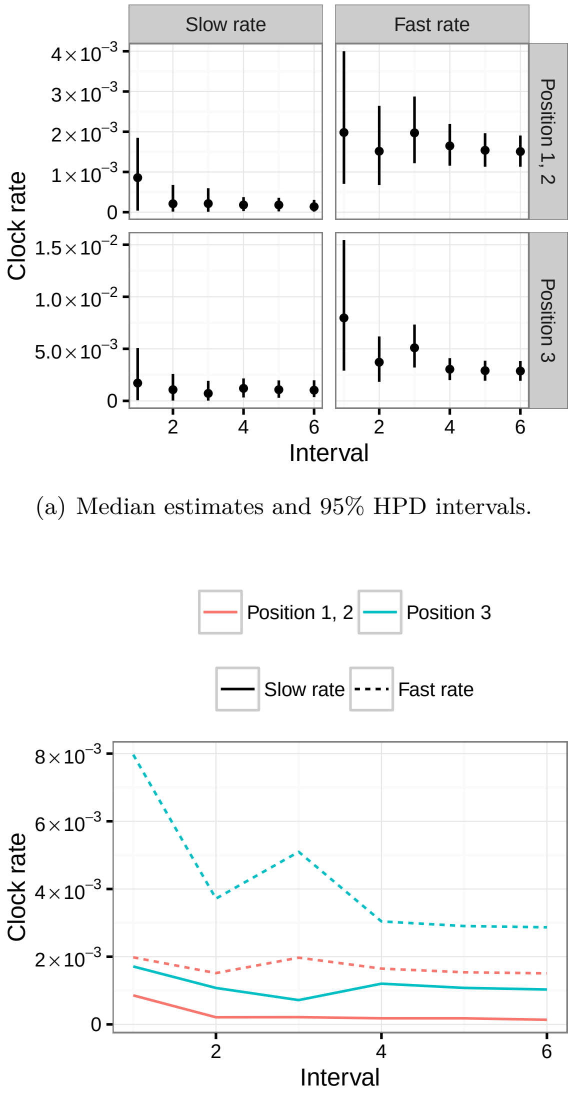
We have also analysed an Ebola dataset that was prepared in a slightly different way and contained non-coding sequences. For details and results refer to Appendix B.

3.5 Conclusion and further work

Using a relaxed clock model, we found some small differences between clock rate estimates on internal and pendant branches on codon position 1, 2. This difference indicates that purifying selection is taking place during the epidemic which has also been confirmed by other studies using more data and a different method [24]. We found similar differences when using a constrained relaxed clock model that only allowed two different rates across the tree.

It may be interesting to reanalyse some data in light of the results described in chapter 2. Overall, how well does the tree prior we use fit the data in each interval? In particular, can we see for the third interval, which adds a burst-like event, that the prior fits worse than for the second interval? Using a codon model that accurately models non-synonymous and synonymous changes could also provide a stronger signal than our approximation that only splits the alignment based on codon position. Maybe after removing methodological biases, the slight signal we find for internal branches having a slower rate than external branches (especially in coding regions on codon position 1, 2), will become more pronounced. In that case, a model that explicitly estimates two different clock rates on internal and pendant branches may prove useful.

Overall, for further work on this project a careful definition of the goal will be vital. A decrease in the inferred rate over time is not a



(a) Median estimates and 95% HPD intervals.

(b) Median estimates only to make comparison between results for the two positions easier.

Figure 3.2: Clock rate estimates (s/s/y) using an exponential growth coalescent and a constrained lognormal relaxed clock with two rate categories.

problem in itself. Further model development is only part of any solution and needs to be accompanied by a solid interpretation of the estimated parameters. Thus it may be helpful to make some of our assumptions about the underlying biological processes (in particular about purifying selection) more explicit. Either with a simulation study or a more stringent mathematical framework which allows us to formalise hypotheses and which may guide the thinking process. It seems intuitive to say that "purifying selection should lead to an increased rate on pendant compared to internal branches", but what are the underlying assumptions that make this true? How do we expect the rate of a pendant branch to behave as its length changes? What happens to the rate of internal branches, when the tree becomes deeper (increase of tree height) or longer (increase of tree length)? These and many more questions will be difficult to answer with intuition alone.

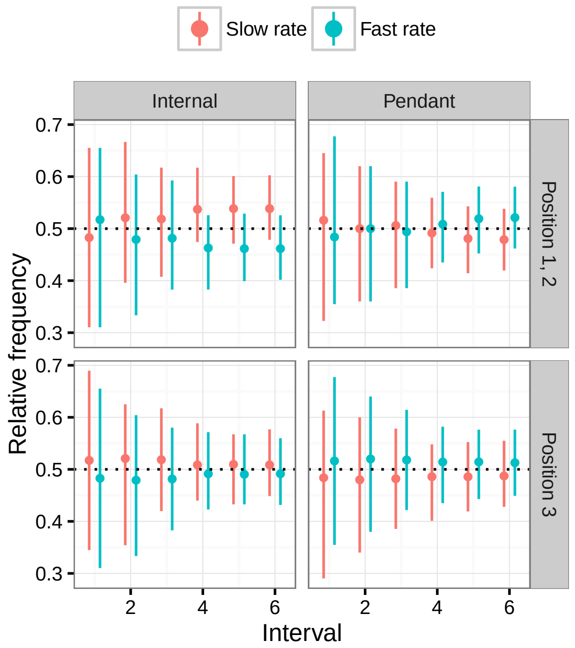
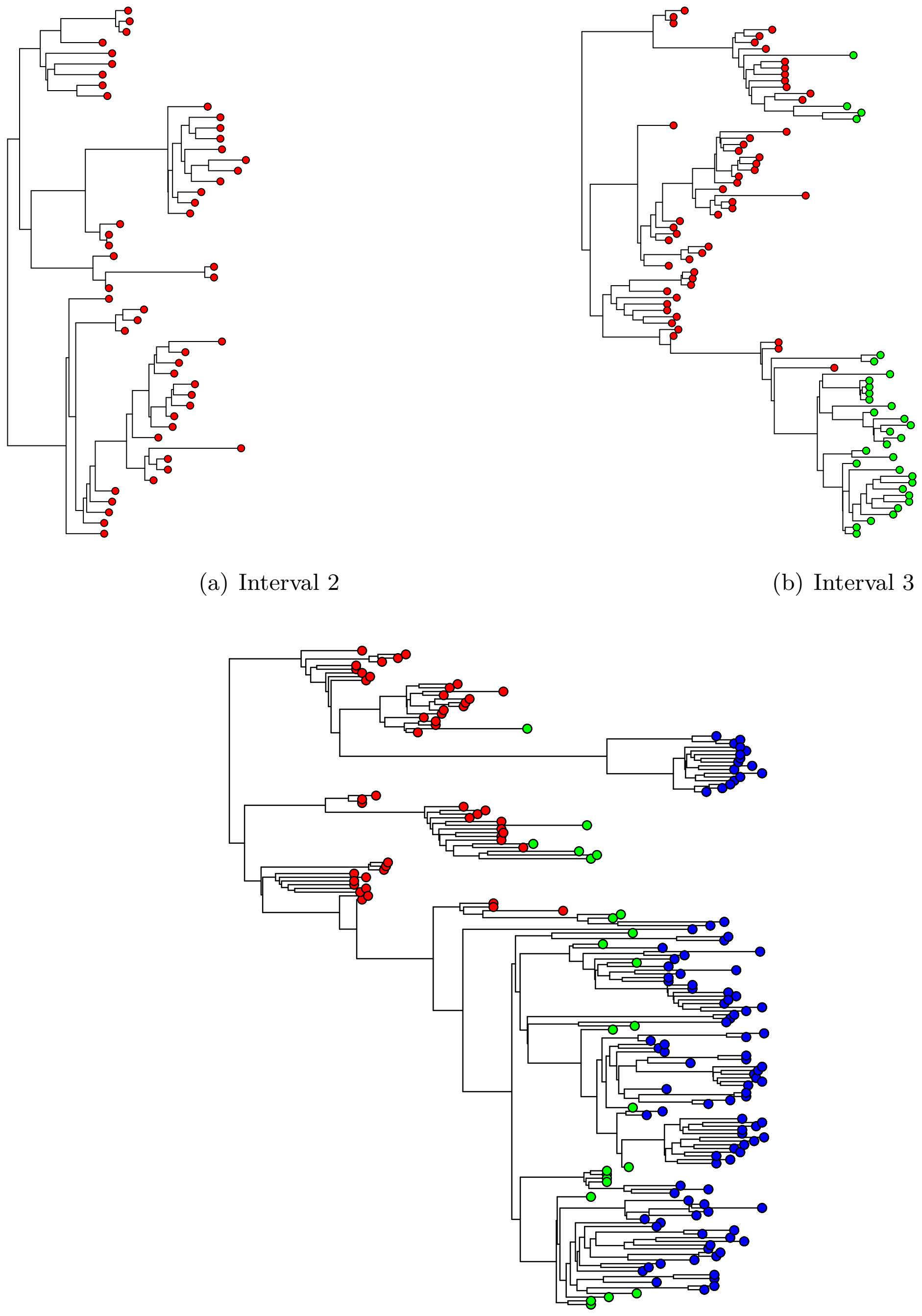


Figure 3.3: Relative frequency with which internal and pendant branches "choose" the slow (red) and fast (turquoise) rate. The dotted line marks 0.5.



(a) Interval 2

(b) Interval 3

(c) Interval 4

Figure 3.4: Maximum clade credibility (MCC) trees for interval 2, 3 and 4. Tips are coloured according to the interval they belong to: 1 and 2 in red, 3 in green and 4 in blue.

Chapter 4

Discussion

In the discussion about the decline of estimated clock rate with increasing sampling period it is important to distinguish two things: methodological errors and biological causes. For the former we should strive to find solutions, for the latter we should offer explanations. From a biological point of view, we expect to see a decrease in the estimated rate as we increase the sampling time since the estimate shifts from something close to the spontaneous mutation rate to the long term evolutionary substitution rate. The current debate [20, 26–28] is more about understanding how much of the observed decline is caused by the biology and how much of it stems from problems in the methods. Methodological errors in the form of biases may be more pronounced for small time scales – either because of their nature or because on small time scales we usually have less data – and can therefore contribute to the observed decline. We need to disentangle these different factors to be able to draw conclusions from real world data.

In chapter 2 we explored potential biases in the clock rate estimate stemming from the tree prior. We talk about biases when we systematically over- or underestimate the value of a parameter. Thus, it is difficult to detect biases in real world datasets where the truth is unknown. Using a simulation study, we highlighted the non-intuitive interactions between the data, the tree prior and the clock model. We demonstrated that misspecifications of the tree prior can lead to inflated clock rate estimates if not enough data are available and showed that this may have been a problem in an early analysis of the Ebola outbreak. To what extent this finding contributes to the various cases in which a decline of rate have been observed is unclear. This should rather be investigated for each case individually, for example by performing a goodness-of-fit test of the applied model.

In chapter 3 we tried to detect and quantify purifying selection in an Ebola dataset. Purifying selection is the most prominent biological phenomenon that is offered as an explanation for decrease in rate estimation over time. Under purifying selection, the clock rate on internal branches should be higher than on pendant branches, because the latter are more likely to still contain slightly deleterious mutations. We found small evidence for this and also showed that the decrease of rate is more pronounced for the internal branches which may reflect an increasing contribution of purifying selection over time.

Generally, interpreting the clock rate should be done with great care. First and foremost, the clock rate is a model parameter, needed to compute the Felsenstein likelihood. Biologically it can correspond to very different things, depending on what type of data and what models are used. For the same set of samples, one can estimate clock rates for coding, non-coding or for whole genome sequences – those estimates will in almost all cases be very different. Going from clock rate to substitution rate is conceptually difficult. Already defining the term substitution rate is not straightforward and is depending on the dataset at hand. It is thus almost meaningless to compare such rates across datasets or timescales.

Appendix A

Supplementary material to chapter 2

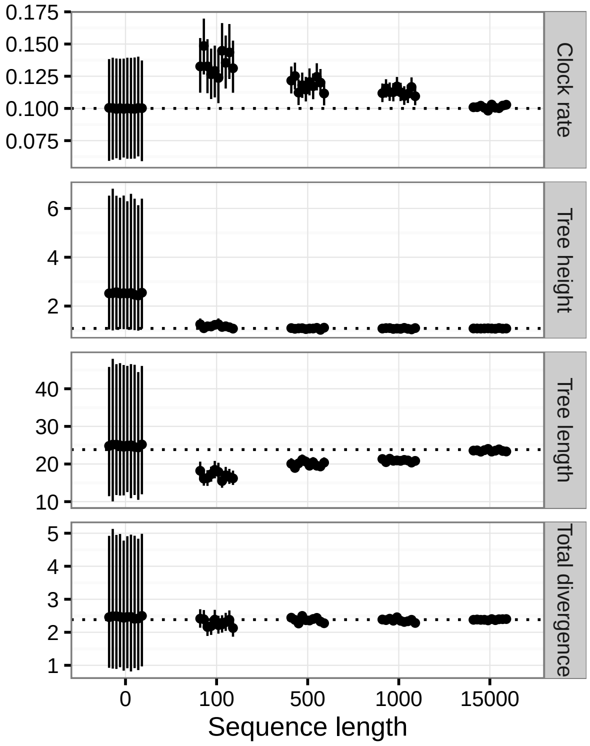


Figure A.1: Similar to Fig. 2.2 in the main text. This also includes results for sequences of length 0 (i.e. analyses that did not use any sequence data but only tip dates and the prior). For those analyses, the chain length was increased to 6×10^8 steps.

Table A.1: Priors used for the analyses.

	Ebola: Guinea	Ebola: Sierra Leone	Influenza
Clock rate	LogNormal(-3, 1.25)		LogNormal(-6, 1.25)
Population size	LogNormal(1, 2)	LogNormal(1, 1.25)	LogNormal(-4, 4)
Growth rate	Laplace(0, 2)		
Migration rate	LogNormal(0, 1.25)		
Kappa	LogNormal(1, 1.25)		
R0	LogNormal(0, 1)		
Become uninfectious rate	LogNormal(3, 1)		
Sampling proportion	Beta(2, 10)		Beta(1, 50)
Origin	LogNormal(0, 1)	LogNormal(-1, 1)	
Skyline intervals	5	3	5

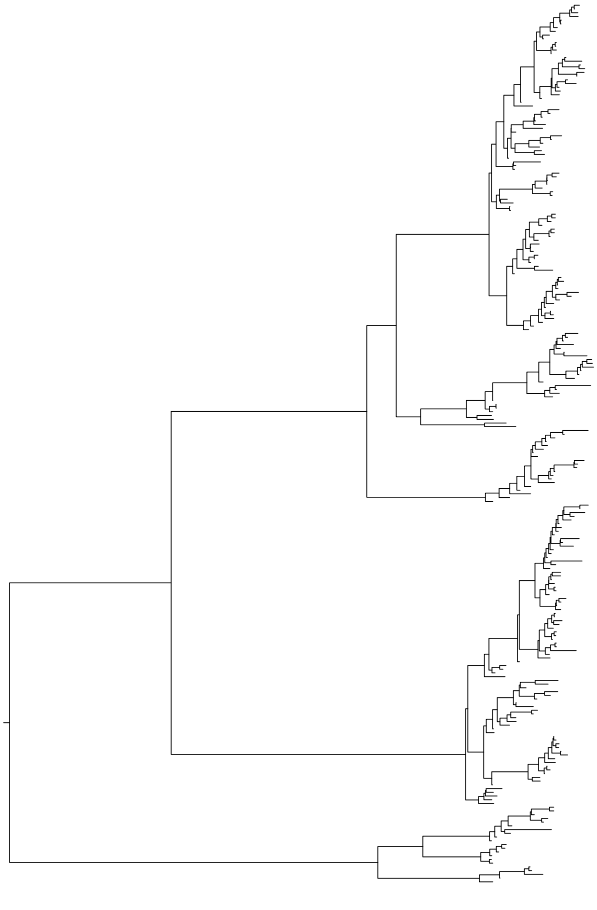


Figure A.2: Median tree (computed with `treescape`) for the first run of the simulation study without any sequence data.

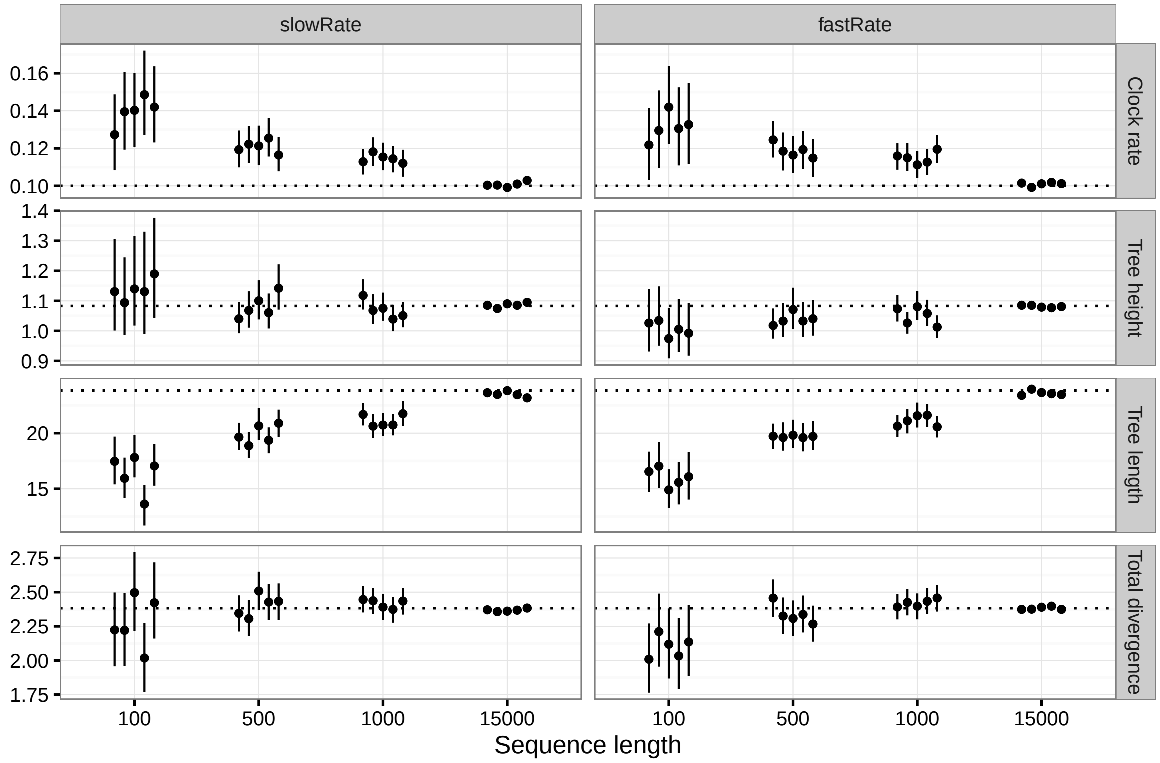


Figure A.3: This is equivalent to the results shown in Fig. 2.2 but for analyses using an exponential growth coalescent model with the growth rate fixed to two different values: 0.001 and 4.0 for the slow and fast growth rate, respectively.

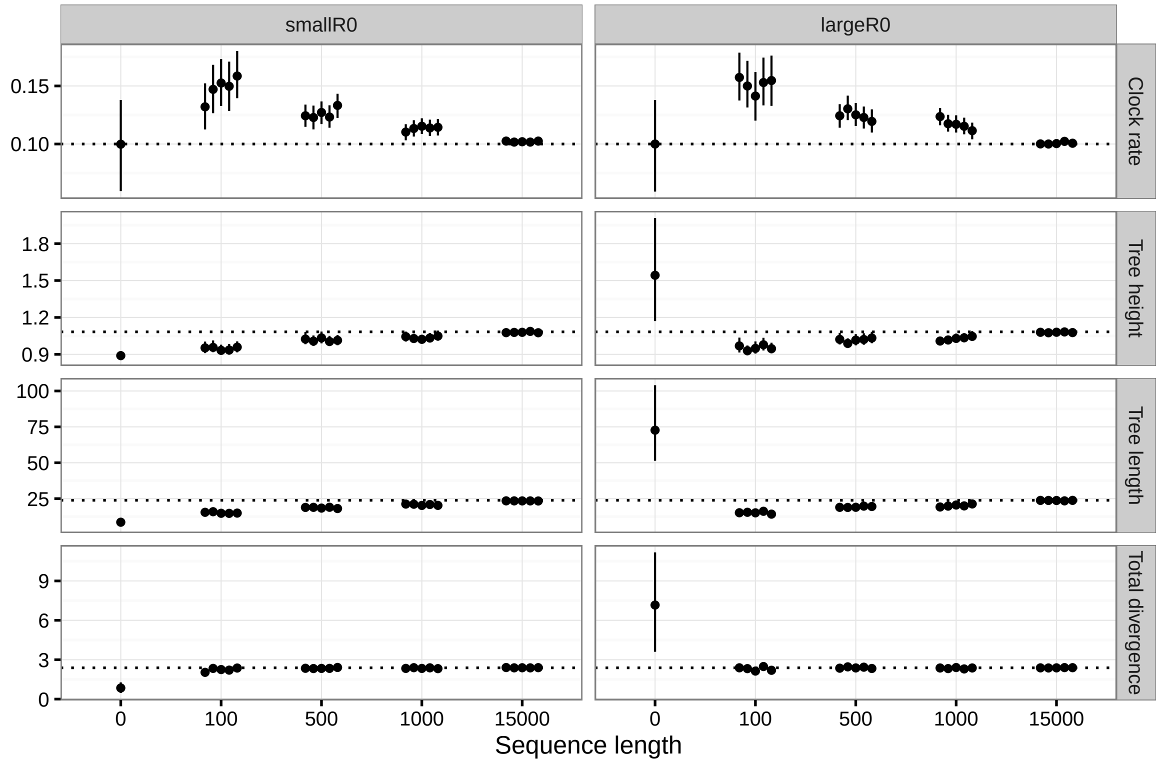


Figure A.4: This is equivalent to the results shown in Fig. 2.2 but for analyses using a serially sampled birth death model with R0 fixed to two different values: 1.000001 and 3.0 for the small and large value, respectively. The results also include an analysis without any sequence data and show that in that case the estimate for the tree length are very different under the two models. However, little sequence data is enough to overcome this prior and lead to an underestimation of tree length and therefore an overestimation of the clock rate.

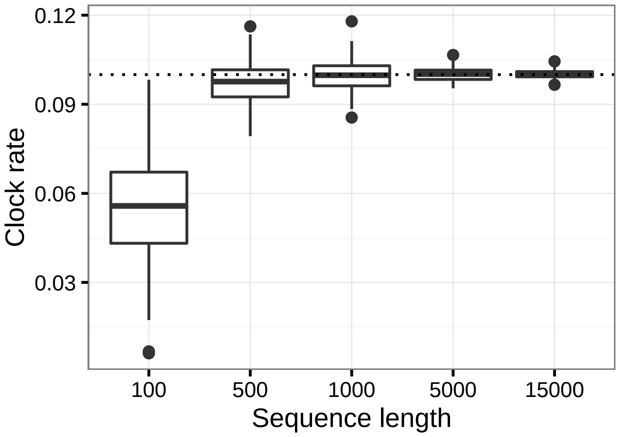
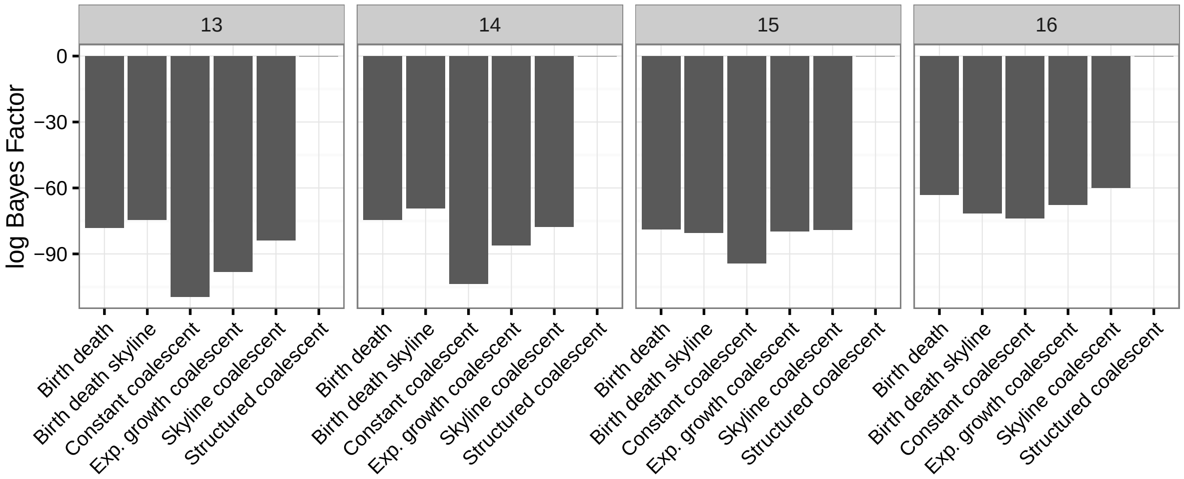
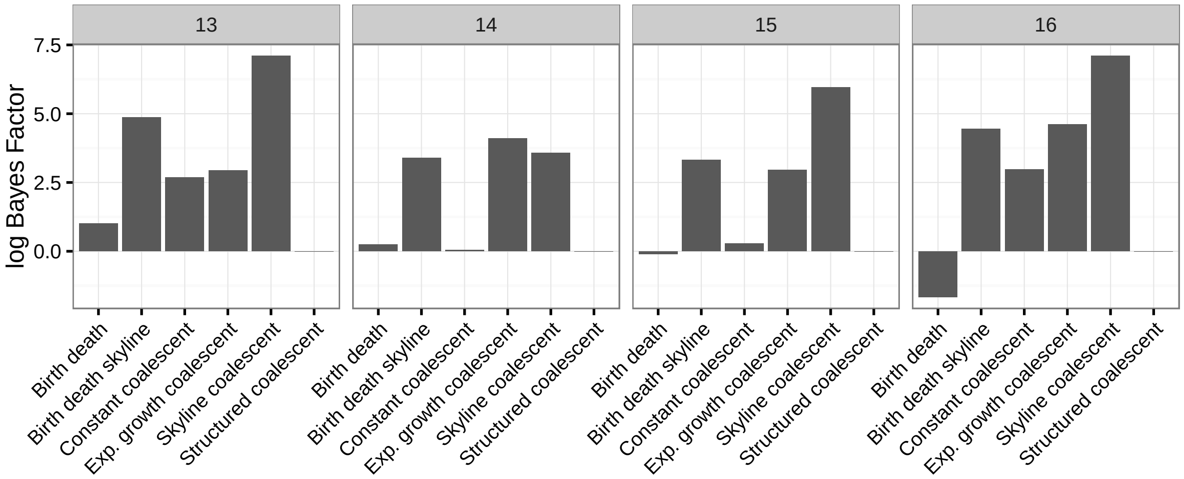


Figure A.5: Results of estimating the clock rate using RAxML and LSD. For this analysis we created 100 alignments per sequence length and inferred the maximum likelihood tree with RAxML which we then fed into LSD to infer the clock rate.



(a) Guinea



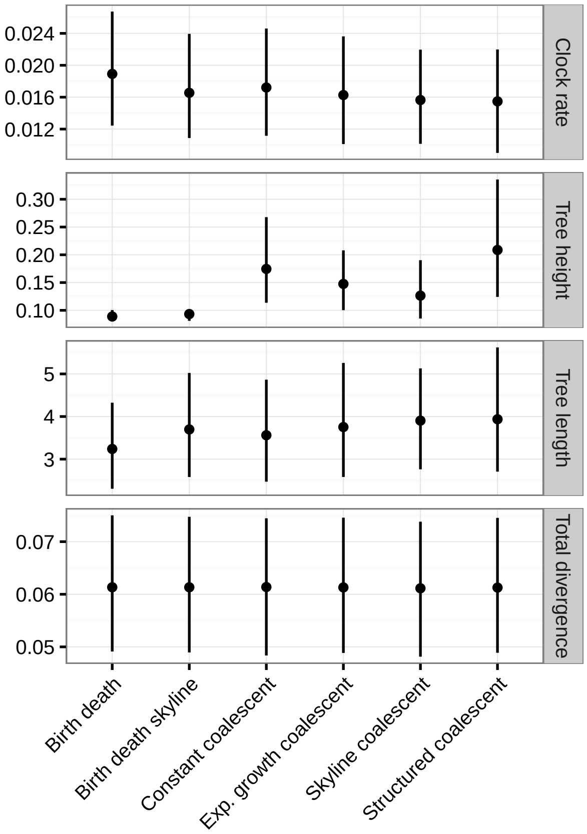
(b) Sierra Leone

Figure A.6: Extension of Fig. 2.5 in the main text showing the results of path sampling analyses for a varying number of steps.

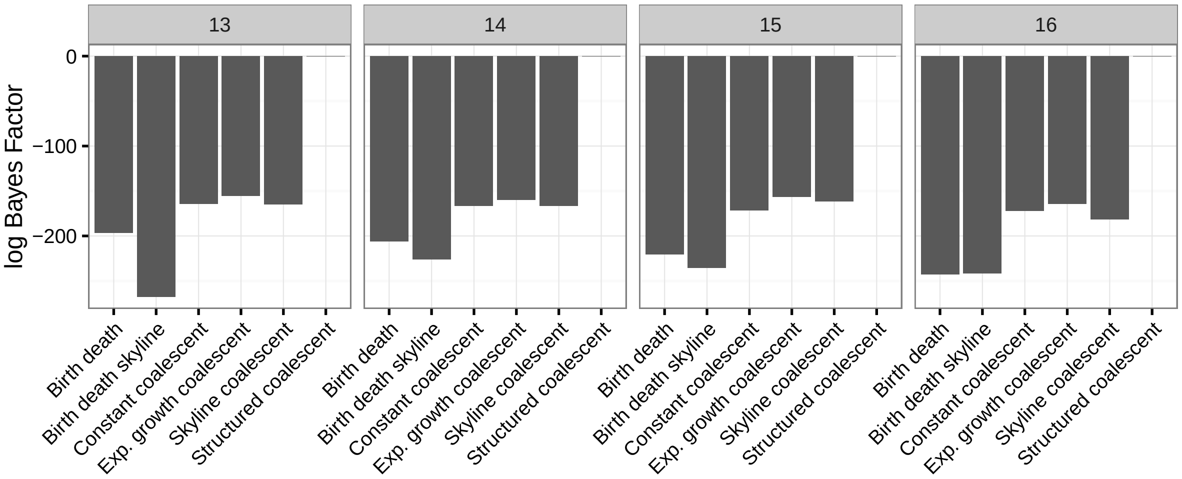
A.1 Analysis of influenza dataset

We also analysed an influenza dataset from 2009/10 that was used in [62] to show that the inferred substitution rate decreases with an increasing sampling period under the same six tree priors. For this dataset, we used sequences for the neuraminidase gene (1407 sites) from the first month only (273 sequences). We assigned these sequences to two clades using a simple k-means clustering algorithm ($k = 2$) on a distance matrix based on the Hamming distance of the sequence data. We used 4×10^8 for all but the structured coalescent model for which we used 3×10^8 steps. For the path sampling analysis we increased the chain length to 10^8 steps.

Results are shown in Fig. A.1. All six models lead to a similar estimate of the clock rate with a median between 16 and 19×10^{-3} s/s/y and largely overlapping HPD intervals. The long term rate for this dataset is around 5×10^{-3} s/s/y [62], but none of the HPD intervals include this value. The estimates for the tree length are very similar as well under all models with medians ranging from 3.2 years for the birth death model to 4 years for the structured coalescent and largely overlapping HPD intervals. There are, however, clear differences in the estimated tree height, ranging from a median of 0.1 years for the birth death models to a median of 0.2 years for the structured coalescent model. Also, the HPD intervals for the structured coalescent model are much wider than for the unstructured coalescent models which in turn are much wider than those for the birth death models. Again, we find no noticeable difference for the total divergence between any of the models.



(a) Results of parameter inference. Analogous to Fig. 2.4 in the main text.



(b) Model comparison. Analogous to Fig. A.6.

Figure A.7: Results for an influenza dataset analysed under different tree priors.

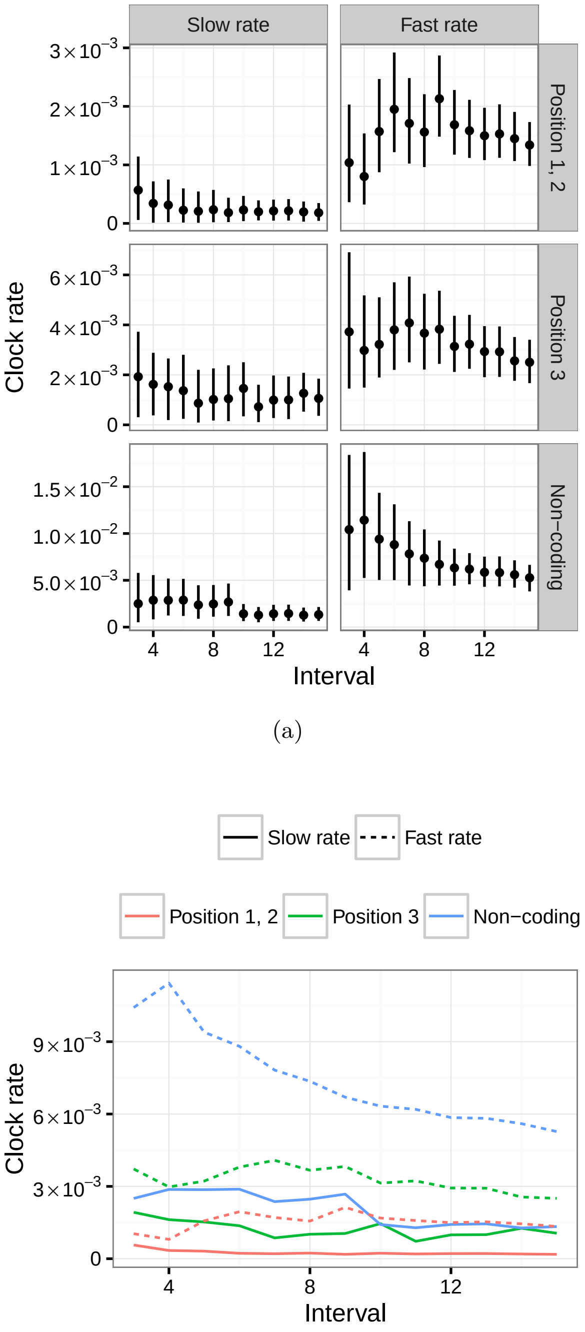
Appendix B

Supplementary material to chapter 3

This analysis follows the one presented in section 3.4 for a slightly different dataset. We used the sequences from the first four intervals of the original analysis and split them into 15 intervals, each containing 11 or 12 sequences (see table B.1 for details). Additionally, we also included the non-coding parts of the genome for those sequences. In the results, we exclude the first two intervals because they only contained 11 and 22 sequences, respectively, which was too little for proper inference. Figs. B.1 and B.2 are analogous to Figs. 3.2 and 3.3, respectively. While not many new conclusions can be drawn from these results, they now also present data for non-coding regions of the genome which may be interesting at a future point.

Table B.1: This table shows for each interval the date of the last sequence as well as the total number of sequences at that point and how many days they spanned. The first sequence was sampled on 2014-03-27.

Interval	Last date	Total no. of days spanned	Total no. of sequences
1	2014-03-31	4	11
2	2014-04-18	22	22
3	2014-05-21	55	33
4	2014-06-09	74	45
5	2014-07-24	119	57
6	2014-08-14	140	69
7	2014-08-27	153	81
8	2014-09-04	161	93
9	2014-09-19	176	105
10	2014-09-25	182	117
11	2014-10-01	188	129
12	2014-10-05	192	141
13	2014-10-09	196	153
14	2014-10-14	201	165
15	2014-10-19	206	177



(a)

(b)

Figure B.1: Analogous to Fig. 3.2, but for the analysis that included non-coding sequences and divided the data differently.

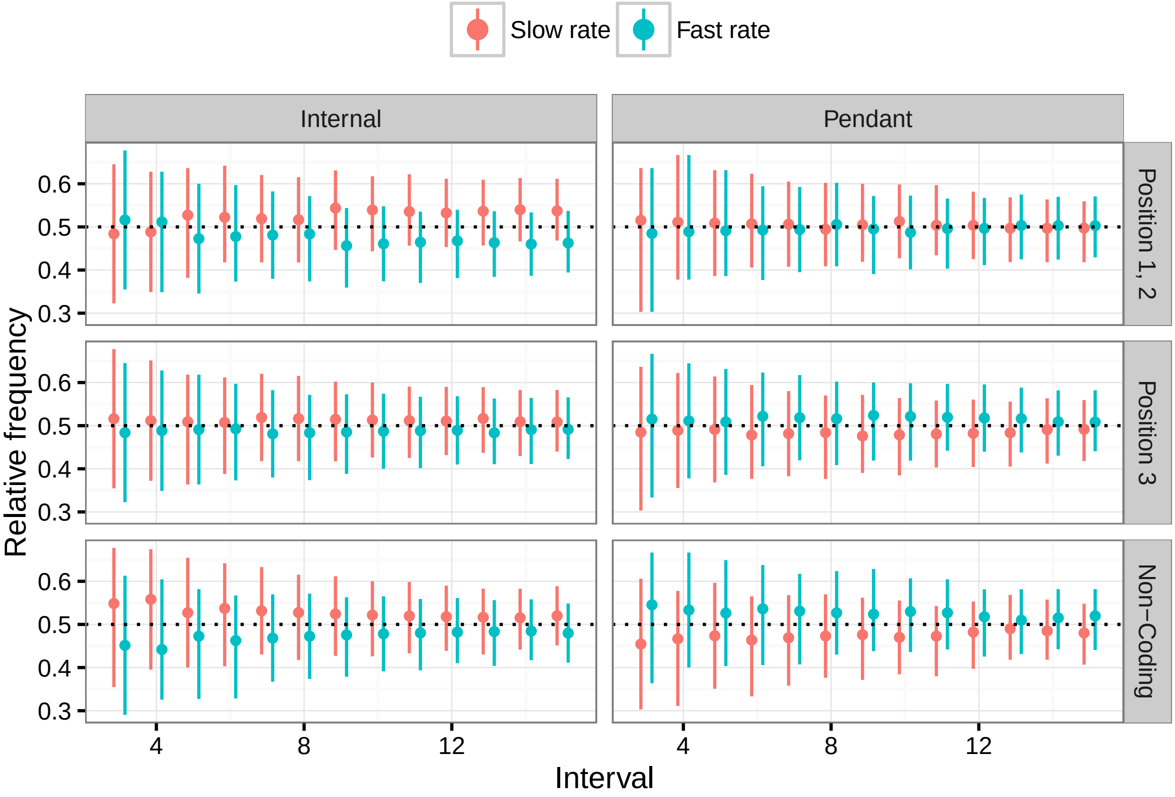


Figure B.2: Analogous to Fig. 3.3 for the second analysis.

Appendix C

Analytic results for the tree length under a serially sampled constant size coalescent model

I made some progress in deriving the indirect prior distribution of the tree length that is created by a serially sampled constant population size coalescent model. Let's first establish some notation:

- n: number of tips
- $\mathbf{z} = \{z_1, ..., z_n\}$: times of tips (0 is present), sorted in increasing order
- $\mathbf{x} = \{x_1, ..., x_{n-1}\}$: times of internal nodes, sorted in increasing order
- $\mathbf{t} = \{t_1, ..., t_{2n-1}\}$: union of all z and x, sorted in increasing order
- p: (effective and scaled) population size (corresponding to the `popSize` parameter in Beast2 and θ in [72])

For given p and $\mathbf{z}$ we would like to compute the expected tree length by integrating over all possible timings of internal nodes $\mathbf{x}$. Given $\mathbf{x}$ it is straightforward to calculate the tree length, T:

$$T(\mathbf{x}, \mathbf{z}) = x_{n-1} + \sum_{i=1}^{n-1} x_i - \sum_{i=1}^{n} z_i. \tag{C.1}$$

Also, Drummond et. al derived a formula for the probability density of a tree under the constant size coalescent process [72]:

$$f(\mathbf{t}, \mathbf{k}, p, n) = \frac{1}{p^{n-1}} \prod_{i=2}^{2n-1} \exp\left[(-k_i(k_i - 1)/2p)(t_i - t_{i-1})\right], \tag{C.2}$$

where $\mathbf{k} = \{k_2, ..., k_{2n-1}\}$ and k_i denotes the number of lineages in time interval t_{i-1} and t_i.

This is the probability density for one specific tree. To account for all trees that have the same timings of internal nodes, we need to apply a factor $m(\mathbf{k}) = \prod_i \binom{k_i}{2}$ for all i that correspond to coalescent events.

Thus the expected tree length is given by

$$E[T|\mathbf{z}, p, n] = \int_x m(\mathbf{k}) f(\mathbf{t}, \mathbf{k}, p, n) T(\mathbf{x}, \mathbf{z}) d\mathbf{x} \tag{C.3}$$

Note that we can clearly compute $\mathbf{t}$ from $\mathbf{z}$ and $\mathbf{x}$ so here I use each term when it is more convenient.

What remains to be done is to find the integration boundaries for all the x_i. Let's start from the root. The upper bound for the root is clearly infinity and the lower bound is given by the oldest tip. Similarly, for the next node, the lower bound is given by the next tip while the upper bound is given by the root. And so on. Formally, the integration domain for x_i is $[z_{i+1}, x_{i+1}]$ for $i = 1, ..., n-2$ and $[z_{i+1}, \infty]$ for $i = n-1$ (i.e. the root).

Putting all this into Mathematica allows us to find the expected value for a tree of a fixed n while leaving all other parameters ($\mathbf{z}$ and p) unspecified:

- $E[T|\mathbf{z}, p, 3] = 3p - z_1 + z_3$
- $E[T|\mathbf{z}, p, 4] = 4p - \frac{1}{3} p e^{\frac{z_2 - z_3}{p}} - z_1 + z_4$
- $E[T|\mathbf{z}, p, 5] = 5p - \frac{1}{3} p e^{\frac{z_2 - z_3}{p}} - \frac{1}{6} p e^{\frac{z_2 - z_4}{p}} - \frac{1}{3} p e^{\frac{z_3 - z_4}{p}} - z_1 + z_5$

For $n = 6$ the equation becomes quite long and looses the structure it had before:

$$\begin{aligned} E[T|\mathbf{z}, p, 6] = &-\frac{1}{60} p e^{-\frac{z_3+2z_4+z_5}{p}} \left(10 e^{\frac{z_2+z_3+z_4+z_5}{p}} + 5 e^{\frac{z_2+z_3+2z_4}{p}} + e^{\frac{z_2+3z_3}{p}} \right. \\ &\left. + 20 e^{\frac{z_2+2z_4+z_5}{p}} + 20 e^{\frac{2z_3+z_4+z_5}{p}} - 360 e^{\frac{z_3+2z_4+z_5}{p}} + 10 e^{\frac{2(z_3+z_4)}{p}} + 20 e^{\frac{z_3+3z_4}{p}} \right) - z_1 + z_6 \end{aligned}$$

I also computed the result for $n = 7$ but it does little good to display it here. It took roughly 4 hours using 24 kernels on the Euler cluster.

We can also compute the variance analytically, but this is even more computationally expensive so I have only done it for $n = 3$:

$$V[T|\mathbf{z}, p, 3] = 5p^2$$

In a next step it would be useful to compute the distribution or just the expected value of T for a given topology. To this end, one would have to change eq. C.3 to integrate only over those $\mathbf{x}$ that are consistent with that topology. For the case of $n = 3$ and two contemporary tips, I have such results for the two different topologies. Let tip A be the sample from time t_A in the past. For the topology ((A, B), C), we get an expected tree length of $3p + 2t_A$ with a variance of $5p^2$. For (A, (B, C)) we get $3p + t_A + \frac{2t_A}{2-3e^{t_A/p}}$ and $5p^2 - \frac{6t_A^2 e^{t_A/p}}{\left(2-3e^{t_A/p}\right)^2}$. I have checked all the above results against simulation in Beast2 and found no deviations.

Appendix D

Implementation of relaxed clock in Beast2

Usually, an uncorrelated relaxed clock model is thought of as allowing each branch to have its own rate that is drawn independently from some prior distribution (e.g. a lognormal or exponential). The probability for the rate vector can then be included in the computation of the posterior. Internally, however, Beast2 discretizes the prior distribution into a number of categories and then assigns branches to those categories. This means that (i) branches can only be assigned a finite number of discrete values and (ii) those values can generally not be changed independently. However, we show that in the case of two rate categories and a lognormal prior distribution, the distribution can be adjusted by changing the two parameters μ and σ such that the two rates can in effect be chosen arbitrarily.

The p quantile of a lognormal distribution with mean μ and standard deviation σ in real space is given by

$$F^{-1}(p) = \exp\left[\mu + \sigma\varphi^{-1}(p)\right], \tag{D.1}$$

where $\varphi^{-1}(p)$ is the quantile function of the standard normal distribution.

For two rate categories, the two quantiles that determine the rates in Beast2 are at $p_1 = 0.25$ and $p_2 = 0.75$. Let the corresponding rates be denoted by $r_1 = F^{-1}(p_1)$ and $r_2 = F^{-1}(p_2)$. We can then solve for μ and σ explicitly and get

$$\sigma = \log\frac{r_1}{r_2}/\left(\varphi_1^{-1} - \varphi_2^{-1}\right) = \log\frac{r_1}{r_2}/\left(2\varphi_1^{-1}\right) \tag{D.2}$$

and

$$\mu = \log r_1 - \frac{1}{2}\log\left(\frac{r_1}{r_2}\right) = \frac{1}{2}\log(r_1 \cdot r_2), \tag{D.3}$$

where $\varphi_1^{-1} = \varphi^{-1}(p_1) \approx -0.67 \approx -\varphi^{-1}(p_2) = \varphi_2^{-1}$. Thus we can choose μ and σ to accommodate any choice of rates (provided they are supported by their respective prior).

Beast2 reports the parameters `ucld.mean` $= \exp(\mu + \sigma^2/2)$ and `ucld.stdev` $= \sigma$ and I have verified that these computations are correct by comparing them to some posterior samples.

Appendix E

Reproducibility

There has been substantial debate in recent years about the “replication crisis” in the sciences (see e.g. [73–75]). While replication means that an entire experiment should be repeated from scratch, *reproducibility* is a weaker goal, simply stating that the data analysis should be reproducible by other researchers. For computational sciences, all findings should in theory be reproducible by making code and data available. I was for example able to download all necessary material to reproduce the relevant results of the Ebola [22] and influenza [62] study. The latter, however, took considerable effort since there was no clear separation of data and scripts and no description of how to use the code.

In this chapter I want to describe the system that I put in place when I began working on my thesis. Hopefully this will not only allow people who are interested in my work to understand what I did and how I did it, but also enable them to reproduce and build upon my results. I was certainly inspired by Sandve et al.'s “Ten Simple Rules for Reproducible Computational Research” [76] and Noble's article on how to organize Computational Biology projects [77].

Wiki For wet lab scientists it is a standard requirement to use a lab journal to record protocols and results. Some computational scientists also use this method, but I find it more natural to keep track of things in an electronic form. Thus I chose `TiddlyWiki` to document my work in a personal wiki. This allows linking different entries, including figures and full text search. I used it to document meetings, thoughts about the project, the literature review, analyses setup, bugs and results with their interpretation.

Generally, I tried to split my work into different analyses that are identified (i.e. named) by the date I started them. Each analysis should correspond roughly to one project or idea. It happens that the findings of chapter 2 correspond to analyses `2014-04-20` (simulation study) and `2014-08-18` (data analysis), whereas those of chapter 3 correspond to `2014-03-13`. Each such analysis has its own directory as well as page in the wiki. For small analyses the wiki entry contains some description of what I was trying to achieve. For larger analyses, i.e. those that took more time, it is just a collection of further wiki pages that are tagged with the title (i.e. date) of that analysis. This is because, over time, the analyses usually diverged from the initial plan so it was difficult to write a general statement. For each result, I create a new wiki

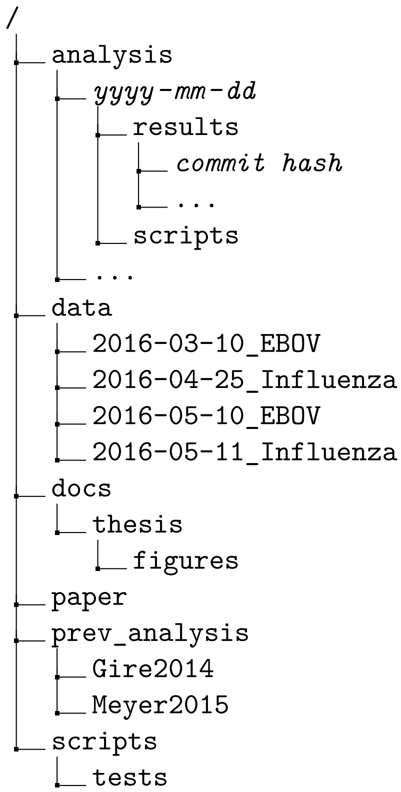


Figure E.1: The directory structure of my thesis directory. Italics indicate that the term should be replaced by a value.

page that is tagged with the name of the original analysis and titled as: “date: commit hash - real title”. I use the hash key of the git commits to structure my results as this should enable me or others to go back to exactly the code that was used to produce them. Since all results are tagged with the original analysis, a list of all those pages is included automatically at the bottom of the main analysis page in chronological order. I include plots in the wiki so that they are version controlled and backed up, even though the raw results are not.

Directory structure Fig. E.1 shows the directory structure I used during the work on my thesis. In reality, there are more folders than displayed, but the figure explains the general structure and should suffice to find the things that ended up working. I hope that many things are self explanatory but it may be helpful to state some additional information: Each subdirectory of `data` contains a `README` explaining where the data came from. `docs` contains any form of documentation like the original project description, the wiki and the thesis. `paper` contains copies of all data and scripts necessary to produce the results in chapter 2 along with a manuscript for submission to a journal. `prev_analysis` contains the data and scripts from the two studies that were reanalysed (i.e. for the most part just the downloaded supplementary material). The global `scripts` folder contains scripts that are used by multiple analyses. Most of them do quite high level, generic things like filling an analysis xml template with sequence data from a fasta file. Raw results and the corresponding figures are usually contained in the same results folder.

I used `git` as version control for everything but the results. The entire content (including the repository history) is contained in a zip archive.

Thoughts on this system I generally think that my organization of the project worked quite well and could also scale to larger projects. Using version control, a wiki and the commit hashs to identify what version produced which results is in my opinion a great framework. Two things come to my mind that could be improved:

Firstly, the separation into different analyses felt at times somewhat artificial. Sometimes, what started out as a seemingly new idea ended up converging with something I had already done and other times analyses moved away from their original intention. It may be better to simply separate `data`, `scripts` and `results`. Additionally, the `scripts` directory could then be divided into `plotting` and `analysis`. Still one would eventually end up with a messy collection of scripts and sometimes scripts require only slight modifications to deal with data from different analyses leading to either manual interventions or duplicated code. I noticed, though, that the scripts that I put into the global `scripts` folder tended to become more robust over time as I used them for various things.

The second point is rather small: I used a single file as my `TiddlyWiki`. This is convenient but does not scale well. Luckily, `TiddlyWiki` also provides ways to split a wiki into multiple files so this should not be a problem.

Bibliography

[1] Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet. 2012;13(5):303–314. Available from: http://dx.doi.org/10.1038/nrg3186.

[2] Yang Z. Molecular Evolution: A Statistical Approach. Oxford University Press; 2014.

[3] Felsenstein J. Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. Syst Biol. 1973;22(3):240–249. Available from: http://dx.doi.org/10.1093/sysbio/22.3.240.

[4] Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed Phylogenetics and Dating with Confidence. PLoS Biology. 2006;4(5):e88. Available from: http://dx.doi.org/10.1371/journal.pbio.0040088.

[5] Bayes T, Price R. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. Philos Trans Roy Soc London. 1763;53(0):370–418. Available from: http://dx.doi.org/10.1098/rstl.1763.0053.

[6] Bruss FT. 250 years of "An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S. communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S.". Jahresber Dtsch Math Ver. 2013;115(3-4):129–133. Available from: http://dx.doi.org/10.1365/s13291-013-0069-z.

[7] Cox RT. Probability, Frequency and Reasonable Expectation. American Journal of Phyics. 1946;14(1):1. Available from: http://dx.doi.org/10.1119/1.1990764.

[8] Polya G. Mathematics and Plausible Reasoning, Volume 2: Patterns of plausible inference. Princeton University Press; 1954. Available from: http://www.ebook.de/de/product/3239255/gyorgy_polya_george_polya_g_polya_mathematics_and_plausible_reasoning_volume_2_logic_symbolic_and_mathematical.html.

[9] Jaynes ET. Probability Theory: The Logic of Science. Cambridge University Pr.; 2003. Available from: http://www.ebook.de/de/product/3260119/e_t_jaynes_probability_theory.html.

[10] Tschirk W. Statistik: Klassisch oder Bayes. Springer Berlin Heidelberg; 2014. Available from: http://www.ebook.de/de/product/23083052/wolfgang_tschirk_statistik_klassisch_oder_bayes.html.

[11] McGrayne SB. The Theory That Would Not Die. Yale University Press; 2012. Available from: http://www.ebook.de/de/product/18982663/sharon_bertsch_mcgrayne_the_theory_that_would_not_die.html.

[12] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics. 1953;21(6):1087. Available from: http://dx.doi.org/10.1063/1.1699114.

[13] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57(1):97–109. Available from: http://dx.doi.org/10.1093/biomet/57.1.97.

[14] Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29(8):1969–1973. Available from: http://dx.doi.org/10.1093/molbev/mss075.

[15] Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology. 2014;10(4):e1003537. Available from: http://dx.doi.org/10.1371/journal.pcbi.1003537.

[16] Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19(12):1572–1574. Available from: http://dx.doi.org/10.1093/bioinformatics/btg180.

[17] Kingman JFC. On the Genealogy of Large Populations. Journal of Applied Probability. 1982;19:27. Available from: http://dx.doi.org/10.2307/3213548.

[18] Rosenberg NA, Nordborg M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet. 2002;3(5):380–390. Available from: http://dx.doi.org/10.1038/nrg795.

[19] Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9(4):267–276. Available from: http://dx.doi.org/10.1038/nrg2323.

[20] Ho SYW, Lanfear R, anAlan Cooper Phillips LB, Soubrier J, Rodrigo AG, Cooper A. Time-dependent rates of molecular evolution. Mol Ecol. 2011;20(15):3087–3101. Available from: http://dx.doi.org/10.1111/j.1365-294X.2011.05178.x.

[21] Domingo E, Sheldon J, Perales C. Viral Quasispecies Evolution. Microbiol Mol Biol Rev. 2012;76(2):159–216. Available from: http://dx.doi.org/10.1128/MMBR.05023-11.

[22] Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345(6202):1369–1372. Available from: http://dx.doi.org/10.1126/science.1259657.

[23] Hayden EC. Ebola virus mutating rapidly as it spreads. Nature. 2014;Available from: http://dx.doi.org/10.1038/nature.2014.15777.

[24] Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG, et al. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. Cell. 2015;161(7):1516–1526. Available from: http://dx.doi.org/10.1016/j.cell.2015.06.007.

[25] Tong YG, Shi WF, Liu D, Qian J, Liang L, Bo XC, et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. Nature. 2015;524(7563):93–96. Available from: http://dx.doi.org/10.1038/nature14490.

[26] Emerson BC, Hickerson MJ. Lack of support for the time-dependent molecular evolution hypothesis. Mol Ecol. 2015;24(4):702–709. Available from: http://dx.doi.org/10.1111/mec.13070.

[27] Emerson BC, Alvarado-Serrano DF, Hickerson MJ. Model misspecification confounds the estimation of rates and exaggerates their time dependency. Mol Ecol. 2015;24(24):6013–6020. Available from: http://dx.doi.org/10.1111/mec.13451.

[28] Ho SYW, Duchêne S, Molak M, Shapiro B. Time-dependent estimates of molecular evolutionary rates: evidence and causes. Mol Ecol. 2015;24(24):6007–6012. Available from: http://dx.doi.org/10.1111/mec.13450.

[29] Donoghue PCJ, Yang Z. The evolution of methods for establishing evolutionary timescales. Phil Trans R Soc B. 2016;371(1699):20160020. Available from: http://dx.doi.org/10.1098/rstb.2016.0020.

[30] Lambert DM, Ritchie PA, Millar CD, Holland B, Drummond AJ, Baroni C. Rates of Evolution in Ancient DNA from Adelie Penguins. Science. 2002;295(5563):2270–2273. Available from: http://dx.doi.org/10.1126/science.1068105.

[31] Drummond A, Pybus GO, Rambaut A. Inference of Viral Evolutionary Rates from Molecular Sequences. In: Advances in Parasitology Volume 54. Elsevier BV; 2003. p. 331–358. Available from: http://dx.doi.org/10.1016/S0065-308X(03)54008-8.

[32] Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. Trends in Ecology & Evolution. 2003;18(9):481–488. Available from: http://dx.doi.org/10.1016/S0169-5347(03)00216-7.

[33] Belshaw R, Sanjuán R, Pybus OG. Viral mutation and substitution: units and levels. Current Opinion in Virology. 2011;1(5):430–435. Available from: http://dx.doi.org/10.1016/j.coviro.2011.08.004.

[34] Duchene S, Holmes EC, Ho SYW. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proceedings of the Royal Society B: Biological Sciences. 2014;281(1786):20140732–20140732. Available from: http://dx.doi.org/10.1098/rspb.2014.0732.

[35] Hoenen T, Safronetz D, Groseth A, Wollenberg KR, Koita OA, Diarra B, et al. Mutation rate and genotype variation of Ebola virus from Mali case sequences. Science. 2015;348(6230):117–119. Available from: http://dx.doi.org/10.1126/science.aaa5646.

[36] Ho SYW, Phillips MJ, Cooper A, Drummond AJ. Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times. Mol Biol Evol. 2005;22(7):1561–1568. Available from: http://dx.doi.org/10.1093/molbev/msi145.

[37] Woodhams M. Can deleterious mutations explain the time dependency of molecular rate estimates? Mol Biol Evol. 2006;23(12):2271–2273.

[38] Ho SYW, Duchêne S. Molecular-clock methods for estimating evolutionary rates and timescales. Mol Ecol. 2014;23(24):5947–5965. Available from: http://dx.doi.org/10.1111/mec.12953.

[39] Duchêne DA, Duchêne S, Holmes EC, Ho SYW. Evaluating the Adequacy of Molecular Clock Models Using Posterior Predictive Simulations. Mol Biol Evol. 2015;32(11):2986–2995. Available from: http://dx.doi.org/10.1093/molbev/msv154.

[40] Duchêne S, Giallonardo FD, Holmes EC. Substitution Model Adequacy and Assessing the Reliability of Estimates of Virus Evolutionary Rates and Time Scales. Mol Biol Evol. 2015;33(1):255–267. Available from: http://dx.doi.org/10.1093/molbev/msv207.

[41] Bouckaert R, Drummond A. bModelTest: Bayesian phylogenetic site model averaging and model comparison; 2015. Available from: http://dx.doi.org/10.1101/020792.

[42] Kelchner SA, Thomas MA. Model use in phylogenetics: nine key questions. Trends in Ecology & Evolution. 2007;22(2):87–94. Available from: http://dx.doi.org/10.1016/j.tree.2006.10.004.

[43] Ripplinger J, Sullivan J. Assessment of Substitution Model Adequacy Using Frequentist and Bayesian Methods. Mol Biol Evol. 2010;27(12):2790–2803. Available from: http://dx.doi.org/10.1093/molbev/msq168.

[44] Eriksson A, Mehlig B, Rafajlovic M, Sagitov S. The Total Branch Length of Sample Genealogies in Populations of Variable Size. Genetics. 2010;186(2):601–611. Available from: http://dx.doi.org/10.1534/genetics.110.117135.

[45] Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences. 2013;110(1):228–233. Available from: http://dx.doi.org/10.1073/pnas.1207965110.

[46] Minin VN, Bloomquist EW, Suchard MA. Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. Mol Biol Evol. 2008;25(7):1459–1471. Available from: http://dx.doi.org/10.1093/molbev/msn090.

[47] Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. Mol Biol Evol. 2005;22(5):1185–1192. Available from: http://dx.doi.org/10.1093/molbev/msi103.

[48] Simon-Loriere E, Faye O, Faye O, Koivogui L, Magassouba N, Keita S, et al. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. Nature. 2015;524(7563):102–104. Available from: http://dx.doi.org/10.1038/nature14612.

[49] Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A, et al. Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. Nature. 2015;524(7563):97–101. Available from: http://dx.doi.org/10.1038/nature14594.

[50] Jukes TH, Cantor CR. Evolution of protein molecules. Mammalian protein metabolism. 1969;3(21):132.

[51] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2016. Available from: https://www.R-project.org/.

[52] Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News. 2006;6(1):7–11. Available from: http://CRAN.R-project.org/doc/Rnews/.

[53] Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2009. Available from: http://ggplot2.org.

[54] Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, et al. Emergence of Zaire Ebola Virus Disease in Guinea. N Engl J Med. 2014;371(15):1418–1425. Available from: http://dx.doi.org/10.1056/NEJMoa1404505.

[55] Hasegawa M, Kishino H, Yano Ta. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985;22(2):160–174.

[56] Vaughan TG, Kühnert D, Popinga A, Welch D, Drummond AJ. Efficient Bayesian inference under the structured coalescent. Bioinformatics. 2014;30(16):2272–2279. Available from: http://bioinformatics.oxfordjournals.org/content/30/16/2272.abstract.

[57] Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty. Mol Biol Evol. 2012;29(9):2157–2167. Available from: http://dx.doi.org/10.1093/molbev/mss084.

[58] Jombart T, Kendall M, Almagro-Garcia J, Colijn C. treescape: Statistical Exploration of Landscapes of Phylogenetic Trees; 2016. R package version 1.9.17. Available from: https://CRAN.R-project.org/package=treescape.

[59] Kendall M, Colijn C. Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution. Mol Biol Evol. 2016;Available from: http://mbe.oxfordjournals.org/content/early/2016/07/17/molbev.msw124.abstract.

[60] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–1313. Available from: http://dx.doi.org/10.1093/bioinformatics/btu033.

[61] To TH, Jung M, Lycett S, Gascuel O. Fast Dating Using Least-Squares Criteria and Algorithms. Syst Biol. 2015;65(1):82–97. Available from: http://dx.doi.org/10.1093/sysbio/syv068.

[62] Meyer AG, Spielman SJ, Bedford T, Wilke CO. Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak. Virus Evol. 2015;1(1):vev006. Available from: http://dx.doi.org/10.1093/ve/vev006.

[63] dos Reis M, Donoghue PCJ, Yang Z. Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet. 2015;17(2):71–80. Available from: http://dx.doi.org/10.1038/nrg.2015.8.

[64] Rannala B. Conceptual issues in Bayesian divergence time estimation. Phil Trans R Soc B. 2016;371(1699):20150134. Available from: http://dx.doi.org/10.1098/rstb.2015.0134.

[65] dos Reis M. Notes on the birth–death prior with fossil calibrations for Bayesian estimation of species divergence times. Phil Trans R Soc B. 2016;371(1699):20150128. Available from: http://dx.doi.org/10.1098/rstb.2015.0128.

[66] Navascués M, Emerson BC. Elevated substitution rate estimates from ancient DNA: model violation and bias of Bayesian methods. Mol Ecol. 2009;18(21):4390–4397. Available from: http://dx.doi.org/10.1111/j.1365-294X.2009.04333.x.

[67] Gatesy J. A tenth crucial question regarding model use in phylogenetics. Trends in Ecology & Evolution. 2007;22(10):509–510. Available from: http://dx.doi.org/10.1016/j.tree.2007.08.002.

[68] Brown JM. Detection of Implausible Phylogenetic Inferences Using Posterior Predictive Assessment of Model Fit. Syst Biol. 2014;63(3):334–348. Available from: http://dx.doi.org/10.1093/sysbio/syu002.

[69] Stadler T, Bonhoeffer S. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. Philosophical Transactions of the Royal Society B: Biological Sciences. 2013;368(1614):20120198–20120198. Available from: http://dx.doi.org/10.1098/rstb.2012.0198.

[70] du Plessis L. Understanding the spread and adaptation of infectious diseases using genomic sequencing data [PhD Thesis]. ETH Zurich; 2016.

[71] Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–290.

[72] Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. Genetics. 2002;161(3):1307–1320. Available from: http://www.genetics.org/content/161/3/1307.

[73] Allison DB, Brown AW, George BJ, Kaiser KA. Reproducibility: A tragedy of errors. Nature. 2016;530(7588):27–29. Available from: http://dx.doi.org/10.1038/530027a.

[74] Nuzzo R. How scientists fool themselves – and how they can stop. Nature. 2015;526(7572):182–185. Available from: http://dx.doi.org/10.1038/526182a.

[75] Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533(7604):452–454. Available from: http://dx.doi.org/10.1038/533452a.

[76] Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten Simple Rules for Reproducible Computational Research. PLoS Comput Biol. 2013;9(10):e1003285. Available from: http://dx.doi.org/10.1371/journal.pcbi.1003285.

[77] Noble WS. A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol. 2009;5(7):e1000424. Available from: http://dx.doi.org/10.1371/journal.pcbi.1000424.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

The impact of the tree prior and purifying selection on estimating clock rates during viral epidemics

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):

Möller

First name(s):

Simon

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date

Zürich, 26.08.2016

Signature(s)

Simon Möller

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.