Master Thesis

Estimation of epidemiological parameters based on simulated phylogenetic trees using birth-death process and coalescent models

Author(s):
Boskova, Veronika

Publication Date:
2013

Permanent Link:
https://doi.org/10.3929/ethz-a-010208418

Rights / License:
In Copyright - Non-Commercial Use Permitted

This page was generated automatically upon download from the ETH Zurich Research Collection. For more information please consult the Terms of use.
Estimation of epidemiological parameters based on simulated phylogenetic trees using birth-death process and coalescent models

Veronika Bošková

Master Thesis
ETH Zürich
D-INFK: Computational Biology and Bioinformatics
March 2013 - August 2013

Supervised by: Tanja Stadler
D-USYS: Institute of Integrative Biology
Acknowledgements

I am very much thankful to my direct supervisor, Tanja Stadler, for wonderful 6 months, great project and excellent supervision. I enjoyed the project very much for not only it being novel an interesting, giving me the opportunity to learn new bio-mathematical models and practice skills in programming and mathematics, but also for Tanja always finding time for my questions, troubleshooting and critical discussion of my work, which improved my thesis immensely.

Next, I would like to thank Denise Kühnert for assistance, troubleshooting and useful discussions of BEAST code and Gabriel Leventhal and Helen Alexander for discussions and guidance in the math world.

I also thank Prof. Sebastian Bonhoeffer for giving me the opportunity to join the group for my master thesis. Thanks to the entire group, especially all of the members mentioned above and Victor Garcia, Mathieu Legros and Sébastien Wielgoss for pleasant and friendly working and social atmosphere created. It set a high standard for me of how excellent research is done. Special thanks goes to Phine Fingerhuth for being the greatest pal.

Finally, I would like to thank my family and my boyfriend for infinite support and trust in my work and decisions.
## Contents

(1) Introduction .......................... 1
  1.1 From epidemiological model to tree dynamics ....................... 1
  1.2 From phylogenetic tree to dynamics of epidemics .................... 2
  1.3 The coalescent .................................. 3
  1.4 Birth-death process ................................ 8
  1.5 Coalescent vs birth-death process for tree analysis ............ 9

(2) Aims of the project ..................... 11

(3) Methods ................................. 12
  3.1 Birth-death tree simulations .................................. 12
  3.2 Coalescent tree simulations .................................. 13
  3.3 Parameter inference based on simulated datasets .................. 14
     3.3.1 Coalescent model ....................................... 14
     3.3.2 Birth-death model ....................................... 14
  3.4 MCMC procedure ............................................ 17
     3.4.1 Prior settings in birth-death process ..................... 17
     3.4.2 Prior settings in coalescent with exponential growth rate ... 18
     3.4.3 Effective sampling size ................................. 18

(4) Results ................................. 19
  4.1 Analysis of I, SI, SIR and coalescent trees ...................... 19
     4.1.1 Stochastic I trees and coalescent trees .................. 19
     4.1.2 Stochastic SI and SIR trees ............................ 49

(5) Discussion .............................. 53

(6) Conclusions ............................. 56

(7) Outlook ................................ 57

(8) References .............................. 58

(9) Appendix ............................... 62
**List of Figures**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I, SI and SIR model scheme</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>Comparison of birth-death model and coalescent with exponential growth rate</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>model performance in <em>growth rate</em> parameter recovery when simulating trees</td>
<td></td>
</tr>
<tr>
<td></td>
<td>with birth-death (I) model and with coalescent</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Tracking of cumulative infecteds, infecteds and sampled individuals over</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>time in I model tree, $\lambda = 0.55, \delta = 0.5$ and $p = 0.5$</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Example tree topology for tree # 1 in the I model simulation under $\lambda$</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>= 0.55, $\delta = 0.5$ and $p = 0.5$ and coalescent tree with same tip times</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Example tree topology for tree # 75 in the I model simulation under $\lambda$</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>= 0.55, $\delta = 0.5$ and $p = 0.5$ and coalescent tree with same tip times</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Role of sampling proportion in <em>growth rate</em> parameter recovery</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>Comparison of birth-death model and coalescent with exponential growth</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>rate 95% HPD intervals size for <em>growth rate</em> parameter</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Interchangeability of sampling proportion in <em>growth rate</em> parameter</td>
<td>41</td>
</tr>
<tr>
<td>9</td>
<td>Uniform unfixed sampling proportion in <em>growth rate</em> parameter</td>
<td>49</td>
</tr>
<tr>
<td>10</td>
<td>Comparison of birth-death model and coalescent with exponential growth</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>rate performance in <em>growth rate</em> recovery when decreasing population of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>susceptible individuals in SI and SIR model-based trees</td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>Comparison of birth-death model and coalescent with exponential growth</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>rate model performance in <em>growth rate</em> parameter recovery when simulating</td>
<td></td>
</tr>
<tr>
<td></td>
<td>trees with birth-death (I) model and with coalescent at scaled branch</td>
<td></td>
</tr>
<tr>
<td></td>
<td>lengths</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>Tracking of cumulative infecteds, infecteds and sampled individuals over</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>time in I model trees</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>Role of sampling proportion in <em>growth rate</em> parameter recovery at medium $R_0$</td>
<td>79</td>
</tr>
<tr>
<td>S4</td>
<td>Comparison of birth-death model and coalescent with exponential growth</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>rate 95% HPD intervals size for <em>growth rate</em> parameter at fixed $R_0$ and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>varying $p$</td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>Interchangeability of sampling proportion in $\lambda \delta p$ parameter</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>recovery</td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>Interchangeability of sampling proportion in $\lambda$ parameter recovery</td>
<td>92</td>
</tr>
<tr>
<td>S7</td>
<td>Interchangeability of sampling proportion in $\delta$ parameter recovery</td>
<td>96</td>
</tr>
<tr>
<td>S8</td>
<td>Interchangeability of sampling proportion in $R_0$ parameter recovery</td>
<td>100</td>
</tr>
<tr>
<td>S9</td>
<td>Uniform unfixed sampling proportion in $\lambda \delta p$ parameter</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>recovery</td>
<td></td>
</tr>
<tr>
<td>S10</td>
<td>Uniform unfixed sampling proportion - exploration of prior on $p$</td>
<td>116</td>
</tr>
<tr>
<td>S11</td>
<td>Uniform unfixed sampling proportion in $\lambda$ parameter recovery</td>
<td>124</td>
</tr>
<tr>
<td>S12</td>
<td>Uniform unfixed sampling proportion in $\delta$ parameter recovery</td>
<td>132</td>
</tr>
<tr>
<td>S13</td>
<td>Uniform unfixed sampling proportion in $R_0$ parameter recovery</td>
<td>140</td>
</tr>
<tr>
<td>S14</td>
<td>Tracking of cumulative infecteds, infecteds and sampled individuals over</td>
<td>146</td>
</tr>
<tr>
<td></td>
<td>time in SI and SIR model-based trees</td>
<td></td>
</tr>
<tr>
<td>S15</td>
<td>Comparison of birth-death model and coalescent with exponential growth</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>rate 95% HPD intervals size for <em>growth rate</em> parameter in SI and SIR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>model-based trees</td>
<td></td>
</tr>
</tbody>
</table>
(1) Introduction

Relationships of diverse species or of individuals within one (fast) evolving species can be related in a phylogenetic tree evaluating pairwise sequence similarity. In recent years, several methods using samples from extant or extinct species have been developed to estimate parameters of interest from phylogenies, taking into account uncertainties of inference of the tree topology [Yang and Rannala, 1997, Drummond et al., 2002, Drummond and Rambaut, 2007, Stadler et al., 2012].

For many applications, determining the past dynamics of populations is of interest. Especially in epidemiological context, the interest lies in knowing the basic reproductive ratio $R_0$ and the rate of spread of the epidemics. $R_0$ is the predictor of the probability and extent of spread of the disease in the population, that is to say, the number of secondary infections a single pathogen can cause when being introduced in a population of susceptible individuals. If this quantity is larger than one, the disease or the pathogen is expected to spread within the population [Diekmann et al., 1990]. The growth rate determines the speed of spread of the pathogen. These quantities are informative on the extent of the preventive measures (e.g. vaccinations or isolation of infected individuals) needed to be taken in order to impede the spread of the pathogen in the population. These parameters were originally only deduced using prevalence and incidence epidemiological data; however, joining the studies of phylogenetics with those of mathematical epidemiology is the recently nascent field of phylodynamics (for review see [Kühnert et al., 2011]), allowing estimation of population dynamics directly from pathogen sequencing data which are becoming widely available.

1.1 From epidemiological model to tree dynamics

To describe dynamics of an epidemics, popular SIR models and adaptations have been previously presented [Kermack and McKendrick, 1927, Anderson and May, 1991]. When using phylogenetic tools to understand epidemic spread, immediately arising questions are: How does the SIR model dynamics give rise to a phylogenetic tree? How we can estimate the parameters of the tree that would be informative about the dynamics of the epidemics?

The SIR-type of models are based on separating the population into different classes of individuals, typically susceptibles ($S$), infecteds ($I$) and recovered ($R$). Individuals can change their status (switch from one class to another). This dynamics is determined by two parameters, birth rate $\lambda$ and death rate $\delta$. Birth rate or transmission rate is a rate at which one infected individual will infect another uninfected individual or rate at which a new pathogen proliferates. In the reconstructed tree
birth event is displayed as bifurcation or split of one branch into two branches, thus the birth rate is sometimes also termed bifurcation rate. This bifurcation happens at the branch which belongs to the individual that infects a new individual. After bifurcation, one of the branches corresponds to the "old" individual, that was already infected earlier, and the second branch corresponds to the newly infected individual. The death rate, or recovery rate, is the rate at which a pathogen/infected individual dies (i.e. recovers from the disease). In the phylogenetic tree, death is marked as a branch in the tree that stops growing and does not continue nor give a progeny at the present time. The transmission chain of epidemics can thus easily be defined in phylogenetic terms. For epidemics following the SIR-type of dynamics, the growth of the population size at the initial stage of the spread follows exponential trend, though, depending on the exact model, it can slow down at later stages. Such early exponential growth has also been observed for within-host dynamics of viruses, like HIV [Rinaldo et al., 1998], confirming the fitness of approximation of the real-life within-host epidemics with the SIR-type of models.

How do we now know which individual infected whom to properly mark the transmission events in the tree? Often, in study of epidemics, the only data that are available are many sequences from different individuals. To relate them in a tree, numerous distance-based or sequence similarity-based methods have been developed for use of sequencing data to reconstruct the most likely phylogeny for all sequences under study (please refer to books and references therein [Felsenstein, 2004, Page and Holmes, 2009]). Methods using Bayesian approach producing, instead of a single most likely tree, an entire distribution of trees have also been popularized [Drummond and Rambaut, 2007]. Unlike most DNA-based organisms, HIV and other RNA viruses display high mutation rate, which predisposes them to be an ideal candidate for phylogenetic reconstruction and application of phylodynamic methods, even if the samples taken from the patients are spaced apart on the scale of months [Drummond et al., 2003].

1.2 From phylogenetic tree to dynamics of epidemics

Once the phylogenetic tree is reconstructed, it can provide us with insights into the dynamics of the epidemics by enabling us to estimate parameters such as transmission rate ($\lambda$), recovery rate ($\delta$), growth rate ($r = \lambda - \delta$), or basic reproductive ratio ($R_0 = \frac{\lambda}{\delta}$). Popular methods for parameter estimation are based on Bayesian inference coupled with Markov chain Monte Carlo (MCMC) procedure [Drummond and Rambaut, 2007], where the likelihood of the parameters is coupled with prior information (if available). Posterior distribution of parameters in phylogenetic context is defined as:
where parameters are $\mathcal{T}$, representing the phylogeny, $\theta$, being a unified term for all parameters of the sequence evolution model, and $\eta$, covering all the parameters of the epidemiological (disease dynamics) model. $f[\mathcal{T}, \eta, \theta|\text{data}]$ in the equation (1) is the posterior probability density distribution of the parameters, the $f[\text{data}|\mathcal{T}, \theta]$ is the likelihood of the parameters ($\mathcal{T}$ and $\theta$), that is usually computed by Felsenstein algorithm [Felsenstein, 1981], $f[\mathcal{T}|\eta]$ is the probability density of the tree given the epidemiological parameters giving rise to the topology via the SIR-type of dynamics, $f[\theta]$ and $f[\eta]$ are priors for evolutionary model and epidemiological parameters, respectively and $f[\text{data}]$ is the normalizing constant representing the integral of the numerator over all tree topologies and parameters. This quantity is a constant (when data are fixed) and is irrelevant for the estimation of the posterior probability density of the parameters in the MCMC procedure.

Two models are currently widely used in epidemiological context to estimate phylodynamic parameters (i.e. calculate $f[\mathcal{T}|\eta]$) from phylogenies: coalescent [Kingman, 1982a, Kingman, 1982b, Griffiths and Tavare, 1994, Drummond et al., 2002] (uses and review of the model are described in [Rosenberg and Nordborg, 2002]) and birth-death process [Kendall, 1948a, Kendall, 1948b, Nee et al., 1994] (reviewed in [Novozhilov et al., 2006]). Estimation of the epidemiological parameters in this context is exemplified by attempts to estimate $R_0$ and/or growth rate parameter of HCV virus [Pybus et al., 2001, Drummond et al., 2005] or HIV epidemic [Rodrigo and Felsenstein, 1999, Stadler et al., 2012, Stadler et al., 2013].

### 1.3 The coalescent

According to the coalescent model, the nodes are progressively merged going back in time until there is only single one left. Coalescent thus provides a prior distribution of trees given the epidemiological parameter estimates. This probability density function allows for estimation of the probability of the tree given parameter constellation [Balding et al., 2007, Griffiths and Tavare, 1994, Rodrigo and Felsenstein, 1999, Drummond et al., 2002]. The coalescent seems to be a good approximation to many processes arising in real life [Hudson et al., 1994, Griffiths and Tavare, 1994, Drummond et al., 2005], however, violations of the model assumption can lead to consequences whose nature and size are still not fully explored [Rodrigo and Felsenstein, 1999].

In the coalescent with constant population size $N$, $n$ genes/individuals are sampled. (Note, that for the purpose of this thesis we will assume that if the population departs from Wright-Fisher model [Fisher, 1930, Wright, 1931] the effective population...
size \((N_e)\) is the population size of interest, and we set \(N = N_e\).) The coalescent reconstructs the evolution of the sampled individuals towards the most recent common ancestor (MRCA), by reconstructing probabilistic structure of the underlying true tree [Donnelly and Tavare, 1995]. Assuming we are dealing with a Fisher-Wright population, where generations are discrete [Fisher, 1930, Wright, 1931], the waiting time \(\tau\), measured in units of \(N\) generations, till the next event (\(j\) genes/individuals merge to produce \(j-1\) genes/individuals) is distributed as:

\[
P\{T(j) > \tau\} = \left( \prod_{i=1}^{j-1} \left(1 - \frac{i}{N}\right) \right)^{N\tau} \xrightarrow{N \to \infty} \exp \left[-\left(\frac{j}{2}\right)\tau\right]
\]

[Balding et al., 2007, Chapter 22.6].

Using time \(t\) in calendar units, we can rewrite the above approximation in terms of \(\phi = N\rho\), were \(\rho\) is length of a generation in calendar units, and \(N\) is the population size as defined above, transforming \(\tau = \frac{\tau}{\phi}\), such that:

\[
P\{T(j) > t\} = \left( \prod_{i=1}^{j-1} \left(1 - \frac{i}{N}\right) \right)^{Nt/\phi} \xrightarrow{N \to \infty} \exp \left[-\left(\frac{j}{2}\right)t\phi\right].
\]

The coalescent with constant population size \(N\), and with all leaves sampled at the same time point (e.g. in present), is characterized by producing trees of average age 2, where 1 time unit is \(N\) generations, which would be \(2\phi\) in time counted in calendar units, and where the time to coalescence of last two branches takes about half of the total coalescence time of sample of size \(n\). This follows from analytic expression of \(E[T(2)]\), which, assuming \(E[T(j)] = \frac{1}{\binom{j}{2}}\) and counting time in units of \(N\) generations, can be expressed from general form of \(E[T_{MRCA}(n)]\) as:

\[
E[T_{MRCA}(n)] = \sum_{j=2}^{n} E[T(j)] = \sum_{j=2}^{n} \frac{2}{j(j-1)} = 2 \sum_{j=2}^{n} \left(\frac{1}{j-1} - \frac{1}{j}\right) = 2 \left(1 - \frac{1}{n}\right)
\]

[Balding et al., 2007, Chapter 22.6].

Counting time in calendar units, we assume \(E[T(j)] = \frac{\phi}{\binom{j}{2}}\) and obtain:

\[
E[T_{MRCA}(n)] = \sum_{j=2}^{n} E[T(j)] = \sum_{j=2}^{n} \frac{2\phi}{j(j-1)} = 2\phi \sum_{j=2}^{n} \left(\frac{1}{j-1} - \frac{1}{j}\right) = 2\phi \left(1 - \frac{1}{n}\right).
\]
Same as above, let \( \phi = N\rho \), where \( \rho \) is number of calendar units per generation. Let \( \frac{1}{\phi} \) be the instantaneous rate of any pair of lineages merging. Further, let \( i \in Z \) be the leaf nodes and \( i \in Y \) be the internal nodes of the tree (if we have \( n \) leaves in the tree, the number of internal nodes is \( n - 1 \)), which correspond, when going in back in time, to coalescent events. Let \( k_i \) be the number of lineages present in the interval of time between nodes \( i \) and \( i - 1 \) (note that the node labels are ordered according to the increasing age, such that \( i > i - 1 \), implying \( t_i > t_{i-1} \)), where time is measured in calendar units. The probability of a tree with fixed times of events (coalescence and/or sampling) \( t_i \), time being measured in calendar units, can be calculated according to [Drummond et al., 2002] as follows:

\[
f\left[T | \phi \right] = \prod_{i \in Y} \frac{1}{\phi} \exp\left(-\frac{(k_i)}{2\phi}(t_i - t_{i-1})\right) \prod_{i \in Z} \exp\left(-\frac{(k_i)}{2\phi}(t_i - t_{i-1})\right).
\]

The formula above is a generalization of the coalescent process with serially sampled leaves introduced by [Rodrigo and Felsenstein, 1999].

Given we have \( k \) ancestors in the sample at some time point \( t \), the coalescence rate for the model with constant population size is \( \frac{k}{N\rho} \). Often, the population size is not constant but is changing through time. If we assume that the population size is growing deterministically following an exponential function with rate rate \( \beta > 0 \) such that \( N(t) = Ne^{-\beta t} \), where \( N = N(0) \) is the population size at present time \( (t = 0) \) and \( N(t) \) is the population size at time \( t \) in the past, the rate of coalescence accelerates to become \( \frac{k}{N\rho N(t)} = \frac{k}{N(t)\rho} \).

We can thus define the coalescent intensity function \( \lambda(t) \), such that \( \lambda(t) = \frac{N}{N(t)} = e^{\beta t} \) [Donnelly and Tavare, 1995]. Furthermore, let \( N(j) \) be the population size at the time \( j \) generations ago, and let time \( \tau \) be again defined in units of \( N \) generations. Set \( x = \frac{j}{N} \), \( f_N(x) = \frac{N(j)}{N} \), \( \lim_{N \to \infty} N(j) = \lim_{N \to \infty} \frac{1}{f_N(x)} = \lambda(x) \), then:

\[
\Lambda(\tau) = \lim_{N \to \infty} \sum_{j=1}^{[N\tau]} \frac{N}{N(j)} \frac{1}{N} = \lim_{N \to \infty} \sum_{x=1/N}^{\tau} \frac{1}{f_N(x)} \frac{1}{N} = \int_0^\tau \lambda(x)dx
\]

[Balding et al., 2007, Chapter 22.6] (note: the \( \frac{1}{N} \) in the summation represents the size of the interval, so when we transform the summation into an integral, we use this as our \( dx \), small interval size at which we integrate).

Approximating the exact result for population growth governed by \( \Lambda(\tau) \), the waiting time, measured in units of \( N \) generations, till the next event (\( j \) genes/individuals merge to produce \( j-1 \) genes/individuals) is distributed as:
\[ P\{T(j) > \tau | T(n) + ... + T(j + 1) = \varsigma \} = \exp \left( -\frac{j}{2} \left( \Lambda(\tau + \varsigma) - \Lambda(\varsigma) \right) \right) \]

[Griffiths and Tavare, 1994, Donnelly and Tavare, 1995].

Transforming the time to calendar units such that \( \tau = \frac{t}{\phi} = \frac{j}{N\rho} \), \( \rho \) being the length of one generation in calendar time units and at the same time small interval size in units of which we define our sum, and assuming \( \phi \) is a constant, we obtain:

\[ \Lambda(t) = \lim_{N \to \infty} \sum_{j=1}^{\left\lfloor \frac{Nt}{\phi} \right\rfloor} \frac{N}{N(j)} \rho = \lim_{N \to \infty} \sum_{x=\rho=\phi/N}^{t} \frac{N}{N(x)} \phi = \int_{0}^{t} \lambda(x)dx, \]

and

\[ P\{T(j) > t | T(n) + ... + T(j + 1) = s \} = \exp \left( -\frac{(j)}{2} \left( \Lambda(t + s) - \Lambda(s) \right) \right). \]

The last equation, the formulation for waiting time (in calendar units) for next coalescent event has been published and used in [Drummond et al., 2002].

It follows that when \( \lambda(t) = 1 \) for \( t \geq 0 \), so \( N(j) = N(0) = N \), for all \( j > 0 \), then we reduce back to coalescent with constant population size [Griffiths and Tavare, 1994, Donnelly and Tavare, 1995]. In contrast to the coalescent with constant population size, the coalescent with exponential population size growth generates star-like trees [Donnelly and Tavare, 1995].

Further extensions, such as skyline plot have been previously introduced as well [Pybus et al., 2000]. Due to the fact that the parameters of the model may change during the course of the infection or that the apparent parameters of the model may change due to model specifications (e.g. SIR model - see section 3.1), an adaptive method, capable of capturing these temporal changes was developed. In general, the skyline plot can capture the variation in the birth (transmission) and death (or recovery) rates of the process governed by the SIR-type of models. Such models all have initial exponential burst of the population size, in most models followed by a slow down, and possible extinction of the infected population, leading to apparent variation in the growth rate. The skyline plot splits the total time interval (from the origin of the disease till the last sampled individual) into the pre-specified number of intervals, and estimates the epidemiological parameters (i.e. population size of infecteds) for each interval separately. In the coalescent framework coupled with Bayesian probabilistic inference, the method is termed Bayesian skyline
plot [Drummond et al., 2005]. The Bayesian skyline plot implementation in BEAST [Drummond and Rambaut, 2007] allows the length of the intervals to be adaptive to the amount of information provided by the total interval, i.e. the method splits the total internal into smaller intervals, each of different length but containing same amount of information (events such as bifurcation or time of sampling).

Currently, new developments in the field of coalescent theory make the estimations of the coalescent rates, especially in complex populations (evolving under SIR model, or having further sub- or compartmental structure) more precise by avoiding the skyline plots, which have been shown to be biased when faster than exponential growth or slower than exponential growth occur. This is due to the fact that in the classic estimation methods for population size changes, the growth rate and thus also the population size were only linearly correlated with the total population size and the effect of the varying birth rates were neglected [Volz, 2012].

As stated in [Rodrigo and Felsenstein, 1999], the requirements for the studied population to be well approximated by the coalescent theory are:
1) that individuals from one generation give rise to the individuals in next generation (birth-death type of process governing the population growth),
2) existence of sufficient diversity within the population to allow reconstruction of the phylogenetic relationships,
3) population size \( N \) to be large enough (compared to the sample taken), and
4) small enough population size to be able to trace back the MRCA.

From the definition of the coalescent and the theory developed thereafter, yet another requirement can be formulated:
5) population size is well captured in deterministic way.

These assumptions with exception of 1) (no discrete generations are assumed) and 3) (no relationship of \( n \) with \( N \) is defined) are also valid for the populations in continuous time following Moran model, rather than Wright-Fisher population model [Stadler, 2013b].

All of the points above, except for point 3) and 5) are usually met in the microevolutionary studies. In the epidemiological context, the sampling rate of recent epidemics can reach high proportions. In the HIV in UK and in Switzerland was estimated to have sampling proportion of \(~ 0.5 - ~ 0.8\) [Stadler et al., 2013, Stadler et al., 2012]. This high sampling proportion is a misspecification for the coalescent model and its consequences on the model performance are so far unknown. In addition, deviations from the deterministic population size are so far not well explored either.
1.4 Birth-death process

Coalescent methods have been quite widespread due to fast explosion of available inference methods and implementations (see citations above in section 1.3). However, application of this mathematical framework to fit fast-evolving populations have been challenged and birth-death process has been further advanced to be equally available for phylogenetic and phylodynamic inferences [Kendall, 1948a, Stadler, 2009, Stadler, 2010a, Stadler et al., 2012].

A birth-death process is a stochastic process, based on Markov chain jumps from one state to one of two neighboring states on the chain. The chain represents the total population count and at any given next time step the population grows by birth of one individual or shrinks by losing one individual. Thus, the probability of any state is determined by the jump probabilities and the state the system was in in the previous step. The theory has a long history and formulations of probability distributions for phylogenetic trees were developed successfully over time [Kendall, 1948a, Kendall, 1948b, Kendall, 1949, Nee et al., 1994, Rannala and Yang, 1996, Yang and Rannala, 1997, Stadler, 2009, Stadler, 2010a].

In a birth-death process, the tree grows forward in time, and finishes after a fixed amount of time or when it reaches a certain number of leaves. The waiting times between events in the birth-death process are exponentially distributed with parameter $\lambda + \delta$ and the population grows in expectation as:

$$E[X(t)] = X(0)e^{(\lambda - \delta)t},$$

where $X(t)$ is the population size at time point $t$ [Kendall, 1948a, Kendall, 1948b].

If the process is repeated many times, a distribution is induced and one can attribute a probability to a tree under given parameter settings. Alternatively, the probability of a tree under certain parameter settings can be inferred using a master equation approach [Stadler, 2009, Stadler, 2010a, Stadler and Bokma, 2013]. The probability density of the tree must be conditioned on certain quantity, such as the time of origin of the tree ($x_0$), the survival of the tree till present, or the total number of individuals in the tree (for review of different expressions for such probability density see [Stadler, 2013a]).

Define $p$ as the sampling proportion, the probability that once the individual is chosen to die with rate $\delta$, it will appear in the final tree, and set $\psi = \delta p$ and $\mu = \delta (1 - p)$ (for pictorial description of the parameters refer to section 3.1). In addition, let $\rho$ be the sampling probability of the extant individuals. Further, let $p_0(t)$ be the probability that an individual at time $t$ in the past will have no sampled descendants, and $p_1(t)$ be the probability that an individual alive at time $t$ will have exactly 1 sampled
descendant in present and no sampled extinct descendants. We obtain:

\[
\begin{align*}
p_0(t) &= \frac{\lambda + \mu + \psi + c_1 \left( e^{-c_1(1-c_2)} - (1+c_2) \right)}{2\lambda}, \\
q(t) &= 2(1-c_2^2) + e^{-c_1(1-c_2)} + e^{c_1(1+c_2)^2}, \\
p_1(t) &= \frac{4\rho}{q(t)}, \\
c_1 &= \sqrt{(\lambda - \mu - \psi)^2 + 4\lambda\psi}, \\
c_2 &= -\lambda - \mu - 2\lambda\rho - \psi,
\end{align*}
\]

The probability density of a tree topology can be formulated using \( p_0 \) and \( p_1 \). First, the tree probability density distribution under incomplete sampling have been derived by [Stadler, 2009]. Extending the model to heterochronous sampling, with \( m \) serially sampled leaves, and \( n \) leaves sampled at present \( (t = 0) \), conditioned on the time of origin being \( x_0 \), setting the number of individuals which are sampled in past and have one or more their descendant in sampled in present to \( k \), and \( u \) to be the probability that once an individual is sampled, it immediately becomes noninfectious, the probability density formulation of tree becomes:

\[
f[T|\lambda, \mu, \psi, \rho, u, x_0] = \frac{(\psi(1-u))^k}{\lambda} \prod_{i=0}^{n-1} \frac{\lambda^{4\rho} \lambda \psi(u + (1-u)p_0(y_i))q(y_i)}{q(x_i)q(x_{i+n-1})}. \quad (7)
\]

The \( x_i \) are the bifurcation times, with exception of \( x_0 \), being the starting point in the tree, and \( y_i \) are the sampling times of extinct individuals without sampled descendants in present [Stadler, 2010a].

Further extensions of the simple birth-death model include procedure similar to the Bayesian skyline plot, namely the birth-death skyline plot [Stadler et al., 2013]. Disadvantage of current implementation of the birth-death skyline plot in BEAST [Drummond and Rambaut, 2007] is that it only splits the total interval into sub-intervals of equal length but unequal information content, and thus parameter estimates in intervals further in the past can have quite large confidence intervals [Stadler, 2013b, Stadler et al., 2013].

### 1.5 Coalescent vs birth-death process for tree analysis

The major advantage of the birth-death model over the coalescent lies in three factors: 1) the birth-death process produces trees and is able to analyze phylogenies with
stochastic population sizes, whereas coalescent is based on deterministic population growth;
2) the fact that birth-death process models do not expect the proportion of the sampled individuals in the tree to be small but rather estimate this quantity along with the other model parameters; and
3) birth-death process is able to recover birth and death rates separately, and thus provide an estimate for the $R_0$, i.e for the phenotype incidence, whereas coalescent is only able to recover population size variation through time ($N(t)$), i.e prevalence of the phenotype [Stadler et al., 2012, Stadler et al., 2013].

In the study of HIV epidemics in Switzerland, comparison of the birth-death model to the coalescent using trees simulated with SIR-type models revealed the advantage of the birth-death model in the ability to separate death rates from birth rates and thus infer the basic reproductive ratio. In comparison, the coalescent requires to independently estimate the generation length, along with the growth rate of the epidemics. In addition, the 95% HPD intervals for growth rate parameters recovered from 100 simulated trees were larger for, and the accuracy of the method was significantly higher for the birth-death model as compared to the coalescent (growth rate parameter for 97% of trees fell into the 95% HPD interval, while only 55% of the cases were correctly recovered under the coalescent) [Stadler et al., 2012].

This was one example where the comparative analysis of coalescent vs. birth-death models was performed on simulated birth-death trees under very specific parameter settings. We were interested in whether we could reproduce these findings and shed more light on the comparative performance of the two methods in phylodynamic parameter recovery.
(2) Aims of the project

In general, population dynamic parameters, such as growth rate of the population, can be inferred from the genealogical relationships of the species/population under study [Balding et al., 2007].

To properly compare performance of the currently widely but seemingly interchangeably used coalescent and birth-death process models in estimating parameters of an epidemics, we performed a simulation study. We used three SIR-types of models to simulate trees that would resemble real-life epidemics: pure I model, SI and SIR models (described in detail in next section). We also simulated coalescent trees, keeping the sampling times of leaves identical to those observed in I/SI/SIR trees.

We estimated the parameters that gave rise to various phylogenies by both coalescent and birth-death process and compared the two methods on the basis of the precision with which they recovered the growth rate parameter from simulated trees.
(3) Methods

3.1 Birth-death tree simulations

We use three modifications of stochastic SIR models to simulate trees growing forward in time (see Figure 1). All of the models are implemented as Gillespie algorithm [Gillespie, 1977]. The first modification is a pure I model, where we only have a group of infected individuals. The population size of the individuals ($N$) is infinite. An individual gives birth with rate $\lambda$, which can be understood in epidemiological terms as transmission rate, and the leaves/branches are dying with rate $\delta$, which can be understood as recovery rate. From the leaves that die, we sample with a certain probability $p$ those leaves that would appear in the final (observed) phylogeny. Such a process can be interpreted as individuals being sampled and dying with rate $\psi = \delta p$ and the rest of the nodes, not visible in the reconstructed phylogeny dying with rate $\mu = \delta (1 - p)$. This model has been implemented in R package 'TreeSim' [Stadler, 2010b].

In the second model, SI, we have two populations: susceptibles ($S$) and infecteds ($I$). An infected individual infects the individuals from the susceptible population with rate $\lambda \times (# of susceptibles)$ (read $\lambda$ times number of susceptibles) and die with rate $\delta$. We again sample from those individuals that die with a probability $p$ and we can thus divide the total death rate in pure death rate $\mu$ and sampling followed by death rate $\psi$ as in the model I above. Once the individuals die, they return to the pool of susceptible individuals (alternatively, for the purpose of use as epidemiological model, we can say that when an infected individual dies, it is replaced immediately with a new susceptible individual, so that the population size $N$ always remains constant). Thus, we have a maximum limit on the total population size $N = S + I$. This total population size restriction produces logistic curve of number of infecteds over time. The initial exponential phase is followed by the slow down (saturation) of the apparent growth rate of the tree (epidemics) until the equilibrium $\lambda \times (# of susceptibles) = \delta \times (# of infecteds)$ is reached and the growth ceases.

The third model represents three classes of individuals - susceptibles ($S$), infecteds ($I$) and recovereds ($R$). Again, an infected individual infects the susceptibles with rate $\lambda \times (# of susceptibles)$ and recovers with rate $\delta$. Upon death, an individual is sampled and shows up in the final phylogenetic tree with probability $p$. There is only single one-way flow of individuals and those that recover cannot become susceptible again, nor are they replaced by new susceptible individuals. The total population size is again limited, $N = S + I + R$. Given the definition of the model, the number of infecteds over time first increases exponentially, then slows down (when there are still susceptible individuals available) reaching the saturation phase and then declines, while the number of susceptibles constantly decreases until it reaches 0 and
the number of recovereds increases over time until there are $N$ recovereds.

(a) I model

(b) SI model

(c) SIR model

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{model_diagrams}
\caption{I, SI and SIR models}
See text for detailed description of the models.
\end{figure}

The initial population size in trees simulated by these models is increasing at the rate $r = \lambda - \delta$. Thus, the total population size (and similarly the population size of the sampled individuals) is increasing as $N = N(0) = N(t)e^{rt}$, if $t = 0$ is the time at present and we look at the growth from past time $t$.

For each chosen parameter settings, we simulate 100 trees. We stop each tree simulation once we reach 100 serially-sampled leaf nodes.

### 3.2 Coalescent tree simulations

Given that the birth-death process and the coalescent are not two equivalent models (see sections 1.3 and 1.4, in the introduction) we wanted to produce trees generated both by birth-death-like models (see stochastic I, SI and SIR models above) and by coalescent models for optimal comparison of performance of the two methods in recovery of phylodynamic parameters.

We used BEAST package [Drummond and Rambaut, 2007] to simulate coalescent trees with temporarily spaced leaves by sampling from the prior distribution of the trees. We used coalescent models with specified rate of exponential population growth (as under I model the population size also grows exponentially in expectation). The times of sampling of the leaves, the population size parameters and the duration of infectiousness (generation length = $\frac{1}{\delta}$) were specified to match exactly those values seen in the corresponding trees simulated by stochastic birth-death (I) model (section 3.1). We sampled 2,000,000 trees from the prior and chose the 1,700,000th to be analyzed by the birth-death model and coalescent model in order to recover the growth rate parameter.
3.3 Parameter inference based on simulated datasets

The birth-death model and the coalescent model were applied to the simulated trees to find the posterior distribution of parameters. For this purpose, we used equation (1), and only $f[\mathcal{T}|\eta]$ was changed when using birth-death or coalescent. The term $\eta$ summarizes all the epidemiological parameters of the tree considered by a model. We used uniform prior distribution for all the parameters (with exception of root-origin distance in the birth-death model, which appeared not to influence the posterior in a significant way). In the context of our tree analysis, without sequence data, $f[\text{data}|\mathcal{T}, \theta]$, where $\theta$ is the term summarizing the parameters of the sequence evolution model, is invariant to $\eta$, $\mathcal{T}$ and $\theta$.

We assumed in this study that the epidemiological parameters did not change at any time during the time span encompassed by the phylogenetic tree; meaning we assumed simple exponential growth of the epidemic. We explored performance of the MCMC implementation of the birth-death skyline serial model with 1 interval and coalescent model with exponential growth rate as implemented in BEAST.

3.3.1 Coalescent model

Since we had populations (initially) growing at exponential rate, we chose to fit coalescent with exponential growth rate of population, i.e. $\lambda(t) = e^{rt}$, and with serially sampled leaves to our data. The formulation for the probability of a tree given its age $t$, given $\phi = N\rho$, $N(t) = \phi_{\rho}e^{-rt}$, $\frac{1}{\phi}e^{rt}$ being the instantaneous rate of any pair of lineages merging, $i \in Y$ being the internal nodes of the tree (if we have $n$ leaves in the tree, the number of internal nodes is $n - 1$), corresponding to, going back in time, a coalescent event, $i \in Z$ being the leaf nodes (total of $n$) and $k_i$ being the number of lineages present in the interval of time between times $t_i$ and $t_{i-1}$, is as follows:

$$f[\mathcal{T}|\phi, r] = \prod_{i \in Y} \frac{1}{\phi} e^{rt_i} \exp \left( -\frac{(k_i)}{\phi r} (e^{rt_i} - e^{rt_{i-1}}) \right) \prod_{i \in Z} \exp \left( -\frac{(k_i)}{\phi r} (e^{rt_i} - e^{rt_{i-1}}) \right)$$

(Drummond et al., 2002), corrected by [Leventhal et al., 2013].

3.3.2 Birth-death model

When we define $\mathcal{T}$ to be the sampled tree which is obtained after time $x_0$ (origin), using definitions of $\lambda$, $\psi$ and $\mu$ as in section 3.1, and defining $p_0(t)$ as in equation (2), $q(t)$ as in equation (3), and setting $\rho = 0$ so that $c_1$ and $c_2$ becomes:
\[ c_1 = \sqrt{(\lambda - \mu - \psi)^2 + 4\lambda \psi}, \]
\[ c_2 = \frac{\lambda - \mu - \psi}{c_1}, \]

we can formulate:

\[ 1 - p_0(t) = s(t) = \frac{\lambda - \mu - \psi - c_1 e^{c_1 t (1-c_2) - (1+c_2)}}{2\lambda}. \]  

(9)

The probability density of such tree, with \( m \) serially sampled leaves can be formulated using equation (7). Assuming all the sampled individuals become immediately noninfectious \( (u = 1, k = 0) \), conditioning on \( S \), at least one individual being sampled at present (dividing the \( f[T|\lambda, \mu, \psi, x_0] \) by \( \frac{1}{s(x_0)} \)), and age of tree defining as \( x_0 \), the exact expression becomes:

\[ f[T|\lambda, \mu, \psi, x_0, S] = \frac{(\psi \lambda)^m}{\lambda s(x_0)} \prod_{i=0}^{m-1} \frac{1}{q(x_i)} \prod_{i=1}^{m} q(y_i), \]  

(10)

where \( x_i \) are the bifurcation times, with exception of \( x_0 \), which represents the time of origin, in the tree and \( y_i \) are the sampling times. This expression was derived by [Stadler, 2013b] using [Stadler, 2010a, Stadler et al., 2012].

3.3.2.1 Interchangeability of the sampling proportion parameter in the recovery of growth rate \( \lambda - \delta \) and \( \lambda \delta p \)

Let us denote \( r = \lambda - \mu - \psi = \lambda - \delta \), and \( d = \lambda \psi = \lambda \delta p \). The likelihood surface defining the probability of the given tree topology \( f[T|\lambda, \mu, \psi, x_0, S] \), conditioned on the tree height \( (x_0) \), does not change when changing sampling proportion \( p \), while keeping \( r \) and \( d \) constant, as it is dependent only on combined parameters \( d \) and \( r \) [Stadler, 2013b].

In particular, the probability density of a sampled tree \( T \) with parameters \( \lambda, \delta, p \), such that \( \lambda, \delta \geq 0 \) and \( 0 < p \leq 1 \), is equal to the probability density of a sampled tree \( T \) with parameters \( \hat{p} \) and,
\[ \hat{\lambda} = \frac{r + \sqrt{r^2 + \frac{4d}{p}}}{2} \]
\[ \hat{\delta} = \frac{-r + \sqrt{r^2 + \frac{4d}{p}}}{2}. \]

We can thus analyze trees using \( \hat{\lambda}, \hat{\delta}, \hat{p} \) instead of \( \lambda, \delta, p \). Given one of the three parameters \( \hat{\lambda}, \hat{\delta}, \hat{p} \), the other two parameters are obtained for value of the tree topology probability density function defined above. Notice also, that increase in \( p \) leads to decrease in \( \lambda \) and \( \delta \), if \( d \) and \( r \) are kept constant.

Though setting the sampling proportion \( p \) to a false value will disable us from recovering true \( \lambda \) and \( \delta \) parameters as such individually, we will always be able to recover true (respectively, maximum likelihood) growth rate parameter \( r = \lambda - \delta \), and \( \lambda \delta p \).

There is a direct relation between \( d, r \) and \( p \) with \( R_0 \), such as:

\[ R_0 = \frac{\lambda}{\delta} = \frac{\lambda^2}{\delta} = \frac{p \left( 2r^2 + \frac{4d}{p} + 2r \sqrt{r^2 + \frac{4d}{p}} \right)}{4d}. \]

If growth rate \( r \geq 0 \), we fix \( d \), and \( r \) and let \( p \) vary, we can have \( R_0(p) \) bounded by:

\[
\begin{align*}
R_0(p) & = \frac{2(pr^2) + 4d + 2\sqrt{(pr^2)^2 + 4d(pr^2)}}{4d} \\
R_0(0) & = 1 \\
R_0(1) & = 1 + \frac{r^2}{2d} + \frac{r}{2d} \sqrt{r^2 + 4d} \\
R_0(p) & > R_0(pt) \quad \text{(for } p > pt) \\
R_0(p) & \geq \frac{r^2}{d} p + 1 \\
R_0(p) & < \frac{2(pr^2) + 4d + 2(pr^2 + 2d)}{4d} = \frac{r^2}{d} p + 2.
\end{align*}
\]

If growth rate \( r < 0 \), we fix \( d \), and \( r \) and let \( p \) vary, we have:
\[ R_0(0) = 1 \]
\[ R_0(1) = 1 + \frac{r^2}{2d} + \frac{r}{2d} \sqrt{r^2 + 4d} = ... < 1 \]
\[ R_0(p) > 0 \]
\[ \frac{dR_0}{dp} = \frac{2r^2}{4d} - \frac{2}{4d} \frac{1}{\sqrt{(pr^2)^2 + 4d(pr^2)}} = \frac{2r^2}{4d} \left( 1 - \frac{p + 2d}{\sqrt{(pr^2)^2 + 4d(pr^2)}} \right) \]
\[ \leq \frac{2r^2}{4d} - \frac{2}{4d} \left( \frac{p + 2d}{pr^2 + 2d} \right). \]

From the above formulae it is apparent that for \( r \geq 0 \), the larger the \( p \) is the larger the \( R_0 \) becomes.

### 3.4 MCMC procedure

To sample from the posterior distribution of the parameters of interest we apply the Markov chain Monte Carlo (MCMC) computational procedure, which explores and approximates the posterior parameter surface by taking samples from their combined posterior distribution.

For our purpose, in the analysis where we explore parameters within one time interval we run the MCMC chain for 2,000,000 steps and sample every 1,000th step to get effective sampling size (ESS, explained in section 3.4.3) for each parameter above \( \sim 800 \). We remove the first 200,000 steps (10%) as burn-in. Summary plots for each of the parameters are done on 95% HPD intervals, meaning the shortest interval, containing 95% of the data is analyzed.

#### 3.4.1 Prior settings in birth-death process

Priors for four parameters in total were set in the birth-death process:
1) sampling proportion: we set this quantity to the true value to enable recovery of \( \lambda \) and \( \delta \)
   (alternatively, we fix it to false \( p \) - see section 4.1.1.4, or we set a uniform prior over \([0,1]\) - see section 4.1.1.5)
2) \( R_0 \): prior was set to have uniform distribution over interval \([0,100000]\), with starting value of 2.0
3) \( \delta \): prior was set to have uniform distribution over interval \([0,200000]\), with starting value of 1.0
4) root-origin distance: prior was set to have log normal distribution with
parameters $M=1$ and $S=1.25$, with starting value of 1.0

3.4.2 Prior settings in coalescent with exponential growth rate

Priors for two parameters in total were predefined in the coalescent model with exponential growth rate as follows:

1) ePopSize ($\phi = N\rho$, where $\rho$ is the duration of one generation in calendar units):
   prior was set to have uniform distribution over interval $[0,10000000]$, with starting value of 0.3
2) growth rate: prior was set to have uniform distribution over interval $[-150000,150000]$, with starting value of $3 \times 10^{-4}$

3.4.3 Effective sampling size

Due to the fact that we used Bayesian MCMC method to estimate parameters of our trees, we checked for whether the effective sampling size (ESS) of our procedure was satisfactory. Good ESS was empirically estimated to be between 500-1000, meaning that in the sample of MCMC steps there are 500-1000 independent steps (samples from the posterior) or samples that well cover the entire distribution [Stadler, 2013b].

Autocorrelation plots can be examined to determine ESS. The idea is to check whether the MCMC chain has a good mixing [Kass et al., 1998]. Formula for ESS calculation is defined as:

$$ESS = \frac{n}{\tau} = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k(v)}$$

where $n$ stands for total sample size and $\tau$ is the autocorrelation time. $\rho_k(v)$ is the autocorrelation of lag $k$ for $v$, and $v$ represents the total of the MCMC samples. The idea is to then find cutoff point $k$ for which $\tau$ would be minimal, thus for which ESS would me maximized [SAS Institute Inc., 2011, Chapter 7].

With our simulations, we aimed for ESS of around 1000. For this purpose we examined the log output file from BEAST in Tracer [Rambaut and Drummond, 2009] which calculates the ESS. According to this analysis, we specified length of the chain and the sampling points as described at the beginning of this section.
(4) Results

4.1 Analysis of I, SI, SIR and coalescent trees

4.1.1 Stochastic I trees and coalescent trees

We simulated trees with stochastic I model and with coalescent model with exponential growth rate and reanalyzed the trees with both birth-death model (fixed $p$, uniform prior for $R_0$ and $\delta$ and log normal prior for root-origin distance) and coalescent model with exponential growth rate (uniform prior for both ePopSize $\phi$ ($\phi = N\rho$, where $\rho$ is the duration of one generation in calendar units) and growth rate). We were interested in the recovery of the growth rate parameter by each of the two models. We simulated trees under wide range of $R_0$ settings, using sampling proportion $p = 0.5$ and death rate $\delta = 0.5$. For each parameter setting we simulated and analyzed 100 trees.
(a) I model trees: $R_0=64, \lambda=32, \delta=0.5$

(b) Coalescent trees: $r=31.5$, tip times as in (a)

(c) I model trees: $R_0=8, \lambda=4, \delta=0.5$

(d) Coalescent trees: $r=3.5$, tip times as in (c)
(e) I model trees: $R_0=4, \lambda=2, \delta=0.5$

(f) Coalescent trees: $r=1.5$, tip times as in (e)

(g) I model trees: $R_0=2, \lambda=1, \delta=0.5$

(h) Coalescent trees: $r=0.5$, tip times as in (g)
(i) I model trees: $R_0=1.1, \lambda=0.55, \delta=0.5$

(j) Coalescent trees: $r=0.05$, tip times as in (i)

(k) I model trees: $R_0=1, \lambda=0.5, \delta=0.5$

(l) Coalescent trees: $r=0$, tip times as in (k)
Figure 2: Comparison of birth-death model and coalescent with exponential growth rate model performance in growth rate parameter recovery when simulating trees with birth-death (I) model and with coalescent.

For each plot, 100 trees simulated with stochastic I model (left column) or coalescent with exponential growth rate (right column) under given parameter settings were analyzed with birth-death model (blue bars) or with coalescent with exponential growth rate (red bars). 95% highest posterior density (HPD) intervals of the growth rate parameter are shown (y-axis). Within each figure the trees are ordered (x-axis) by the median value of the growth rate parameter recovered by the coalescent (orange dot within the red bar). Median of the posterior interval for the growth rate parameter recovered by birth-death model is indicated as light blue dot within each blue interval. The true value of the growth rate parameter (the value under which the trees were simulated), is displayed as black horizontal bar.

From Figure 2 it can be observed that birth-death model recovers well the growth rate parameter for trees simulated with stochastic I model, whereas coalescent model recovers well the growth rate parameter for those trees that were simulated under the coalescent. This observation is not surprising but confirms the basic expectations that the model used for simulation should be reasonably good when it is also applied for inference.

On the other hand, comparing performances of the birth-death model and the coalescent in terms of growth rate parameter recovery reveals that coalescent seems to miss the true growth rate value on the I-trees more often than the birth-death model on the coalescent trees (see Figure 2). Exploration of the phenomenon was performed and the results summarized in section 4.1.1.1.

In addition, coalescent does not perform well in terms of growth rate parameter recovery on the I model trees in most parameter settings, getting apparently worse with lowering the $R_0$ - compare the plots in the left column of Figure 2, which are ordered according to decreasing $R_0$. With lowering $R_0$ there are less trees, for which the posterior interval covers the true growth rate, with most of these being over-estimates of the growth rate. The same can be observed for the birth-death process, when fitting the growth rate parameter to the coalescent trees for $R_0 < 2$. Further analyses are presented in section 4.1.1.2.

As expected, the above observations (sections 4.1.1.1 and 4.1.1.2) do not change when the branch lengths are being scaled, i.e. rates are multiplied by constant factor, as is documented in Supplementary Figure S1, for some chosen parameter settings.

In section 4.1.1.3 our interest was drawn to the 95% HPD interval sizes of the growth rate estimates produced by birth-death and coalescent models.

Furthermore, we were also interested to confirm by simulations the interchangeability of the sampling proportion during the tree analysis, when the tree is fixed. This hypothesis is further elaborated upon in section 4.1.1.4.

Finally, if the sampling proportion is not fixed, the MCMC procedure explores the full space of the parameters, including all the possible values for $p$. We wanted to confirm that the true growth rate is still recovered, when we set a uniform prior over the interval $[0,1]$ for the sampling proportion parameter - see section 4.1.1.5.
4.1.1.1 Exploration of the reasons of malperformance of the coalescent method

From the Figure 2 above, it is visible (making row-wise comparisons of the plots) that when coalescent model is applied to trees produced by I model, the true growth rate value is missed by the 95% HPD intervals more often than in cases where birth-death model is applied to the coalescent trees. The summary of counts where the true growth rate is properly recovered, i.e. lies in the 95% HPD interval, by the method is summarized in Table 1 and Supplementary Table ST1.

<table>
<thead>
<tr>
<th>True growth rate parameter recovered correctly</th>
<th>I model trees</th>
<th>coalescent trees</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>birth-death</td>
<td>coalescent</td>
</tr>
<tr>
<td>$R_0 = 64, \lambda = 32, \delta = 0.5, p = 0.5$</td>
<td>98</td>
<td>71</td>
</tr>
<tr>
<td>$R_0 = 8, \lambda = 4, \delta = 0.5, p = 0.5$</td>
<td>94</td>
<td>59</td>
</tr>
<tr>
<td>$R_0 = 4, \lambda = 2, \delta = 0.5, p = 0.5$</td>
<td>92</td>
<td>65</td>
</tr>
<tr>
<td>$R_0 = 2, \lambda = 2, \delta = 0.5, p = 0.5$</td>
<td>92</td>
<td>55</td>
</tr>
<tr>
<td>$R_0 = 1, 1, \lambda = 0.55, \delta = 0.5, p = 0.5$</td>
<td>94</td>
<td>28</td>
</tr>
<tr>
<td>$R_0 = 1, \lambda = 0.5, \delta = 0.5, p = 0.5$</td>
<td>78</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 1: Growth rate recovery counts
Out of the 100 trees simulated with the respective method (I model or coalescent) we counted the number of times the method applied (birth-death or coalescent) recovered the true growth rate value in the 95% HPD interval. The summary for each $R_0$ setting with $\delta = 0.5$ and $p = 0.5$ is displayed.

Seeing that coalescent performs much worse on the trees specified by the I model than on those specified by the coalescent, we hypothesized that the expected population growth is the reason for differential performance of the two models. By definition, in the coalescent model the population size is varying deterministically according to a given growth rate and in the birth-death process this population size is rather fluctuating stochastically (the population size can grow and suddenly decrease to a number less than in the previous steps). When we take a lot of trajectories produced by birth-death process, on average, the same population size change is observed as predicted by coalescent. The deterministic population size could be a reason why coalescent over- or under-estimates quite often the growth rate of the tree obtained by the forward-in-time (birth-death) simulation model. This expectation of the coalescent model in terms of population growth would translate into it trying to fit at the steepest initial slope on the infecteds-over-time curve sometimes leading to wrong (most often over-) estimation of the growth rate parameter. The birth-death process fits all the points equally, and so is unlikely to miss the true growth rate in the 95% HPD interval, even though some stochastic fluctuations of the population growth occur. To confirm our intuition, we produced lineages through time (LTT) plots, and plots of cumulative infecteds, infecteds and sampled infecteds over time.
I model trees: $R_0=1.1, \lambda=0.55, \delta=0.5, p=0.5$

Figure 3: Tracking of cumulative infecteds, infecteds and sampled individuals over time in I model tree, $\lambda = 0.55, \delta = 0.5$ and $p = 0.5$

For each of 100 trees simulated with stochastic I model we counted the number of cumulative infected, infected and sampled individuals at each time step. The x-axis represents the time (going backwards from present $t=0$) and y-axis the counts. Each line represents history of one tree over time. The left column shows the trees colored on the basis of whether coalescent was able to correctly recover (grey lines) or not (red lined for over- and orange lines for under-estimated) true growth rate. The right column shows the same for birth-death process applied to the same trees. Here, we use example figure for settings of $\lambda = 0.55, \delta = 0.5$ and $p = 0.5$. See Supplementary Figure S2 for the plots for other parameter settings. The abbreviations in the legend stand for the following; BDSS - birth-death skyline serial model (with 1 interval), CE - coalescent with exponential growth rate, LTT - lineages-through-time.
Indeed, it is apparent from Figures 3 and S2 that coalescent underestimates the growth rate for trees whose growth is severely delayed, i.e. which did not reach the exponential growth phase yet, once the simulations were stopped. Similarly, the coalescent overestimated the growth rate quantity in cases where the tree of sampled infecteds grows faster than deterministically expected, i.e. once the individual in the past is sampled, it is expected to produce a tree with growth: \( N(t) = Ne^{-rt} \) \((r = \lambda - \delta)\).

### 4.1.1.2 Role of sampling proportion at low \(R_0\) values on growth rate recovery

Trees simulated with I model under the settings above seem to be problematic for coalescent model to fit and recover true growth rate parameter. Also, (especially at \(R_0 < 2\)) the coalescent seems to produce trees which are fitted by birth-death model with higher growth rate than true parameter was. For detail, explore Figure 2 and Table 1.

Since sampling proportion is constant \((p=0.5)\) for all the trees simulated above, a lower \(R_0\), translates to more death per birth event, and consequently means higher sampling from the population. We thus further investigated if this increased sampling causes the methods to fail for low \(R_0\).

Upon visual inspection of the trees produced by the birth-death process and by the coalescent for \(R_0 = 1.1\) \((\lambda=0.55, \delta=0.5, p=0.5)\), we noticed that coalescent mostly produced longer trees with longer inner branches, especially close to the root, than the birth-death model (Figure 4). Our suspicion was that the sampling proportion of \(p=0.5\) violated the assumptions of the coalescent model (i.e. low sampling proportion is assumed), caused the model to break down during the analysis stage of I trees and mis-specify the growth rate parameter to be too big for such short trees.

The same argument would hold for coalescent trees (simulated by taking the tip (leaf node) dates, population sizes and number of calendar units per generation from the respective trees produced by the I model) analyzed by birth-death model. High sampling proportion would cause too many lineages to be sampled early on in the tree formed by stochastic I model, causing the coalescent method to reconstruct trees with more coalescent events than expected per given time step, at the period where a lot of leaf nodes are sampled, then attaching to it long inner branches towards the root, resulting in an apparent higher growth rate towards the leaf nodes of the tree. The birth-death model would fit such a tree (due to its capacity to take into account stochastic population size fluctuations), as an instance of the fast growing tree, that by chance has few long tree branches towards to root (i.e. fewer than expected number of lineages) but then picks up the exponential growth towards the leaf nodes. The growth rate would thus be estimated by the birth-death model to
be higher than true.

On the other hand, there were rare cases where the birth-death model under-estimated the growth rate on the coalescent trees. This phenomenon was caused by the fact that the population size at the end of the simulation with I model, was low, compared to the sampling time of the oldest tip (Figure 5). I.e. the expected time to the MRCA, calculated from \( N(t) = Ne^{-rt} \), where \( N(t_{MRCA}) = 1 \) and \( N \) is the population size obtained at the end of the tree simulation with I model, is smaller than time of the oldest sampled tip in the tree. Thus, the time to MRCA has to be delayed in such trees to be larger than the time of sampling of the oldest tip, causing over-stretched coalescent trees to appear.
Figure 4: **Example tree topology for tree # 1 in the I model simulation under** $\lambda=0.55$, $\delta=0.5$ and $p=0.5$ **and coalescent tree with same tip times**

The population size of both trees is $N = 63$. The expected exponential growth follows $N(t) = Ne^{-rt} = 63 * e^{-0.05t}$. It is thus expected in the coalescent model that the last common ancestor for all leaves of the tree will be found at the latest at $t = 82.9$. However, due to the high sampling rate, a lot of coalescent events appear towards the leaf nodes of the coalescent tree, and this number is significantly reduced towards the root. (The estimated median growth rate of the coalescent tree by the birth-death model is: 0.121, and the 95% HPD is [0.073 , 0.169].)
Figure 5: Example tree topology for tree # 75 in the I model simulation under $\lambda = 0.55$, $\delta = 0.5$ and $p = 0.5$ and coalescent tree with same tip times

The population size of both trees is $N = 6$. The expected exponential growth is $N(t) = Ne^{-rt} = 6e^{-0.05t}$. It is thus expected in the coalescent model that the last common ancestor for all leaves of the tree will be found at the latest at $t = 35.8$. However, due to the last tip being sampled at time $t > 35.8$ (i.e. at $t = 37.2$), the time to the MRCA is shifted further in the past, resulting in coalescent tree with apparent slower growth rate. (The estimated median growth rate of the coalescent tree by the birth-death model is: 0.011, and the 95% HPD is [-0.115, 0.163].)
We further picked the parameter constellation of: $\lambda = 10$ and $\delta = 9$ (low $R_0$; $R_0=1.1$ - upscaling the branch length of $\lambda = 0.55, \delta = 0.5$) and $\lambda = 4$ and $\delta = 0.5$ (medium $R_0$; $R_0=8$) to simulate and reanalyze trees with different sampling proportions: $p = [0.001, 0.01, 0.5, 1]$. The results are displayed in Table 2, Figure 6 and Supplementary Figure S3.

<table>
<thead>
<tr>
<th>True growth rate parameter recovered correctly</th>
<th>I model trees</th>
<th>coalescent trees</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>birth-death</td>
<td>coalescent</td>
</tr>
<tr>
<td>$R_0 = 8, \lambda = 4, \delta = 0.5, p = 1$</td>
<td>93</td>
<td>68</td>
</tr>
<tr>
<td>$R_0 = 8, \lambda = 4, \delta = 0.5, p = 0.5$</td>
<td>94</td>
<td>59</td>
</tr>
<tr>
<td>$R_0 = 8, \lambda = 4, \delta = 0.5, p = 0.01$</td>
<td>98</td>
<td>69</td>
</tr>
<tr>
<td>$R_0 = 1.1, \lambda = 10, \delta = 9, p = 1$</td>
<td>79</td>
<td>28</td>
</tr>
<tr>
<td>$R_0 = 1.1, \lambda = 10, \delta = 9, p = 0.5$</td>
<td>93</td>
<td>20</td>
</tr>
<tr>
<td>$R_0 = 1.1, \lambda = 10, \delta = 9, p = 0.01$</td>
<td>95</td>
<td>50</td>
</tr>
<tr>
<td>$R_0 = 1.1, \lambda = 10, \delta = 9, p = 0.001$</td>
<td>94</td>
<td>64</td>
</tr>
</tbody>
</table>

Table 2: Growth rate recovery counts at varying $p$
Out of the 100 trees simulated at fixed $\lambda$ and $\delta$ with the respective method (I model or coalescent) we counted the number of times the method applied (birth-death or coalescent) recovered the true growth rate value in the 95% HPD interval. The summary for two $R_0$ setting with varying $p$ is displayed.
(a) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=1$

(b) Coalescent trees: $r=1$, tip times as in (a)

(c) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.5$

(d) Coalescent trees: $r=1$, tip times as in (c)
(e) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.01$

(f) Coalescent trees: $r=1$, tip times as in (e)

(g) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.001$

(h) Coalescent trees: $r=1$, tip times as in (g)
It appears that the closer we approach in our simulations to the assumptions of the coalescent model in terms of low proportion of sample size in relation to the total population size, the more often true growth rate parameter is recovered by the coalescent when applied to I model trees (notice the decreasing number of misclassification with decreasing $p$ in Figure 6). The same is observed in the case of birth-death fit of the growth rate parameter to the coalescent trees, i.e. with decreasing the sampling proportion in the original simulation, the tip times, the population size and the length of generation in calendar units, plugged in the coalescent model produce trees which can be properly interpreted by the birth-death model. That is to say, coalescent can properly analyze only trees produced by a model that formally complies with all the assumptions of the coalescent, and this property becomes more obvious the lower $R_0$ gets. In contrast, performance of birth-death model is not strictly dependent on the sampling proportion, however, it overestimates the growth rate for trees produced by the coalescent model at low $R_0$ values, when too high sampling proportion is used to simulate tip times with I model.

4.1.1.3 Lengths of the 95% HPD intervals

The lengths of the 95% intervals were also explored, as it was pointed before that birth-death process may be more likely to recover correctly the parameters of interest due to the larger HPD intervals as compared to coalescent [Stadler et al., 2012]. Consistent with previous observations, in our simulations (Figures 7 and S4), for all parameter settings, in a large majority of the trees, we notice longer 95% HPD intervals when trees are analyzed with the birth-death model as compared to the coalescent.

Also, we observed that with lower sampling proportion used for tree simulations, the size of the 95% HPD interval got smaller for both birth-death and coalescent estimates of the growth rate (data shown in Supplementary Figure S4), meaning both methods become more confident in the growth rate estimates the longer the tree gets.
(a) I model trees: $R_0=64, \lambda=32, \delta=0.5$

(b) Coalescent trees: $r=31.5$, tip times as in (a)

(c) I model trees: $R_0=8, \lambda=4, \delta=0.5$

(d) Coalescent trees: $r=3.5$, tip times as in (c)
(e) I model trees: $R_0=4, \lambda=2, \delta=0.5$

(f) Coalescent trees: $r=1.5$, tip times as in (e)

(g) I model trees: $R_0=2, \lambda=1, \delta=0.5$

(h) Coalescent trees: $r=0.5$, tip times as in (g)
(i) I model trees: $R_0=1.1, \lambda=0.55, \delta=0.5$

(j) Coalescent trees: $r=0.05$, tip times as in (i)

(k) I model trees: $R_0=1, \lambda=0.5, \delta=0.5$

(l) Coalescent trees: $r=0$, tip times as in (k)
4.1.1.4 Interchangeability of sampling proportion setting in tree analysis by birth-death model

As outlined in the methods part 3.3.2, we attempted to show that given a tree structure, setting of the sampling proportion $p$ in the analysis by birth-death process would not influence recovery of the growth rate or $\lambda \delta p$ parameters. We used following parameter constellation to prove our hypothesis: $\lambda = 4, \delta = 0.5, p = [0.01, 0.5, 1]$. We analyzed set of 100 trees simulated for each $p$ under each $p = [0.01, 0.5, 1]$, thus, the same 100 trees were analyzed under 3 different settings of $p$, e.g. trees produced under $\lambda = 4, \delta = 0.5, p = 0.01$ were analyzed under $p = 0.01$ (true $p$), $p = 0.5$, and $p = 1$.

We could confirm that no matter what $p$ is used for the analysis, true growth rate is equally well recovered by the birth-death process for both stochastic I model trees or coalescent trees (see Figure 8). The same holds for recovery of $\lambda \delta p$ (Supplementary Figure S5).

During this analysis, we also noticed that for trees generated with I model, when we set $p$ to its true value (i.e. value used during the tree simulation), we are able to recover true $\lambda$ and $\delta$ parameters (Supplementary Figures S6 and S7, respectively), and by consequence also true $R_0$ (Supplementary Figure S8). However, this was not the case for the coalescent trees, as these are not anymore produced under the assumptions for $\lambda$ and $\delta$ used in I-type of model. Additionally, as predicted by the analytic results in section 3.3.2.1, if we fixed $p$ during the analysis to be bigger than the true $p$ used during simulations, then we overestimated $R_0$ and underestimated $\lambda$ and $\delta$ for I model trees and observed upward shift for $R_0$ and downward shift for $\lambda$ and $\delta$ estimates for coalescent trees. The opposite was true when we fixed $p$ to be smaller than true $p$ ($\lambda$, $\delta$ were overestimated and $R_0$ was underestimated for I model trees and $\lambda$, $\delta$ shifted downwards and $R_0$ shifted upwards for coalescent trees).
(a) I model trees: true $p=0.01$, analyzed under $p=0.01$ and $p=0.5$

(b) Coalescent trees: $r=3.5$, tip times as in (a), analyzed under $p=0.01$ and $p=0.5$

(c) I model trees: true $p=0.01$, analyzed under $p=0.01$ and $p=1$

(d) Coalescent trees: $r=3.5$, tip times as in (c), analyzed under $p=0.01$ and $p=1
(e) I model trees: true $p=0.5$, analyzed under $p=0.5$ and $p=0.01$

(f) Coalescent trees: $r=3.5$, tip times as in (e), analyzed under $p=0.5$ and $p=0.01$

(g) I model trees: true $p=0.5$, analyzed under $p=0.5$ and $p=1$

(h) Coalescent trees: $r=3.5$, tip times as in (g), analyzed under $p=0.5$ and $p=1
(i) I model trees: true $p=1$, analyzed under $p=1$ and $p=0.01$

(j) Coalescent trees: $r=3.5$, tip times as in (i), analyzed under $p=1$ and $p=0.01$

(k) I model trees: true $p=1$, analyzed under $p=1$ and $p=0.5$

(l) Coalescent trees: $r=3.5$, tip times as in (k), analyzed under $p=1$ and $p=0.5
Figure 8: Interchangeability of sampling proportion in growth rate parameter recovery
100 trees were simulated with I model with \((R_0 = 8)\) \(\lambda = 4, \delta = 0.5,\) and \(p = [0.01, 0.5, 1]\) - left column or with coalescent when using same sampling times, population size and setting the duration of one generation in calendar units as in the corresponding I model trees - right-hand side columns, and analyzed with birth-death model assuming either the true \(p\) or untrue \(p\). Within each figure the trees are ordered (x-axis) by the median value of growth rate parameter recovered from trees analyzed under true \(p\). The graphs display the 95% HPD of the growth rate parameter and are ordered in increasing order by sampling proportion values used during the tree simulation. The value of the growth rate parameter under which the trees were simulated, is displayed as black horizontal bar.

4.1.1.5 Unfixing the sampling proportion in tree analysis by birth-death model

To mimic the situation where the sampling proportion is unknown and there is no prior knowledge or clues about its value, we analyzed our simulated trees with birth-death process, but instead of specifying \(p\) to a fixed value, we specified a uniform prior over the interval \([0,1]\). This caused no effect on the recovery of the growth rate parameter, as seen from Figure 9, or on recovery of \(\lambda \delta p\) (see Supplementary Figure S9). However, for the parameters \(\lambda, \delta\) and \(R_0\) there were different effects observed, depending on true sampling proportion \(p\) of the tree used for simulation.

First, in cases where true (used during simulation) \(p = 0.5\), unfixing \(p\) resulted in wider 95% HPD intervals that either fully, or mostly, contained the 95% HPD interval produced when \(p\) was fixed to true value. This is because the value \(p = 0.5\) is in the middle of the prior, i.e. prior on \(p\) is uniform around 0.5 (see Supplementary Figure S10).

Second, especially in cases where \(p = [0.001, 0.01, 1]\), the 95% HPD interval produced at unfix ed \(p\) was shifted away from the 95% HPD interval that resulted from analysis where \(p\) was fixed to true value (refer to Supplementary Figures S11, S12, S13). As predicted by derivations in section 3.3.2.1, when the true \(p\) used for tree simulations was below 0.5, the estimated interval for \(\lambda\) and \(\delta\) produced when prior on \(p\) had a uniform distribution around 0.5, was shifted downward from the interval estimated when the \(p\) was fixed to true value during analysis, and the estimated interval for \(R_0\) was shifted upward. When the true \(p\) used for simulations was higher than 0.5, the posterior intervals for \(\lambda\) and \(\delta\) shifted upwards, whereas the posterior interval for \(R_0\) shifted downwards.
(a) I model trees: $R_0=64, \lambda=32, \delta=0.5, p=0.5$

(b) Coalescent trees: $r=31.5$, tip times as in (a)

(c) I model trees: $R_0=64, \lambda=3.2, \delta=0.05, p=0.5$

(d) Coalescent trees: $r=3.15$, tip times as in (c)
(e) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=1$

(f) Coalescent trees: $r=3.5$, tip times as in (e)

(g) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.5$

(h) Coalescent trees: $r=3.5$, tip times as in (g)
(i) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.01$

(j) Coalescent trees: $r=3.5$, tip times as in (i)

(k) I model trees: $R_0=4, \lambda=2, \delta=0.5, p=0.5$

(l) Coalescent trees: $r=1.5$, tip times as in (k)
(m) I model trees: $R_0=2, \lambda=1, \delta=0.5, p=0.5$

(n) Coalescent trees: $r=0.5$, tip times as in (m)

(o) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=1$

(p) Coalescent trees: $r=1$, tip times as in (o)
(q) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.5$

(r) Coalescent trees: $r=1$, tip times as in (q)

(s) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.01$

(t) Coalescent trees: $r=1$, tip times as in (s)
(u) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.001$

(v) Coalescent trees: $r=1$, tip times as in (u)

(w) I model trees: $R_0=1.1, \lambda=1, \delta=0.9, p=0.5$

(x) Coalescent trees: $r=0.1$, tip times as in (w)
(y) I model trees: $R_0=1.1, \lambda=0.55, \delta=0.5, p=0.5$

(z) Coalescent trees: $r=0.05$, tip times as in (y)

(aa) I model trees: $R_0=1, \lambda=0.5, \delta=0.5, p=0.5$

(bb) Coalescent trees: $r=0.0$, tip times as in (aa)
4.1.2 Stochastic SI and SIR trees

In trees produced by SI model, with small total population size ($N$) the curve of infected individuals over time follows a logistic trend. In case of the SIR model, assuming small total population size, the curve of infecteds over time initially follows exponential growth, after which the growth slows down, to then start decreasing (see Supplementary Figure S14). For neither the birth-death model nor the coalescent model with exponential population size growth, it is trivial to reconstruct proper growth rate of the SI/SIR population, assuming simple exponential growth (Figure 10). Birth-death process underestimates the growth rate more severely than the coalescent model, which may be due to the coalescent model fitting to the steepest initial slope of the infecteds-over-time curve, whereas the birth-death process fits the entire trend in the infected-over-time curve. As mentioned in section 4.1.1.1, since coalescent is a deterministic process, the number of lineages present in a tree at any certain moment may not be larger than the population size $N(t)$, which results in the coalescent estimating the growth rate from highest early slope, and later dynamics not influencing growth rate estimate. In contrast, the birth-death process infers an average growth rate from entire dynamics of the tree and thus deals with the fluctuations in the number of lineages present in the tree close to the root equally as with those present next to the leaf nodes of the tree.

Dissimilar to the sizes of the 95% HPD intervals of the growth rate parameter recovered from I model trees, in SI and SIR trees, the coalescent produced larger posterior intervals as compared to birth-death model (see Supplementary Figure S15).
(a) I model trees: $N=\infty$

(b) SI model trees: $N=10,000$

(c) SIR model trees: $N=10,000$
(d) SI model trees: $N=500$

(e) SIR model trees: $N=500$

(f) SI model trees: $N=300$

(g) SIR model trees: $N=300$
Figure 10: Comparison of birth-death model and coalescent with exponential growth rate performance in growth rate recovery when decreasing population of susceptible individuals in SI and SIR model-based trees

For each plot, 100 trees simulated with stochastic SI model (left column) or SIR model (right column) under $\lambda = 4$, $\delta = 0.5$, $p = 0.5$, and population size $N$, indicated above each plot. The trees were analyzed with birth-death model (blue bars) or with coalescent with exponential growth rate (red bars). 95% HPD intervals of the growth rate parameter are shown (y-axis). Within each figure the trees are ordered (x-axis) by the median value of the growth rate parameter recovered by the coalescent (orange dot within the red bar). Median of the posterior interval for the growth rate parameter recovered by birth-death model is indicated as light blue dot within each blue interval. The true value of the growth rate parameter (the value under which the trees were simulated), is displayed as black horizontal bar. Plot a) is identical to Figure 2 c).
(5) Discussion

Under birth-death and coalescent models, the times of coalescence, or bifurcation, are stochastically selected from the distribution of coalescence (bifurcation) event times, thus the individual trees produced by both coalescent and birth-death process are stochastic realization of the respective processes. However, the coalescent model with exponential growth rate of the population assumes deterministic changes in the population size. Thus the trajectory of population size follows pure exponential growth (i.e. decay). It has been pointed out before [Donnelly and Tavare, 1995] that the coalescent model can appropriately approximate evolution happening according to models where the sampled genealogy is conditioned on the total population size varying deterministically. Not surprisingly, application of the coalescent model on the trees generated by the coalescent itself, confirms the recovery of the parameters generating the tree in the first instance (see results in this thesis or for example [Drummond et al., 2002]). This should be a natural property of any model when the model generating and analyzing the trees are identical. As has been shown in this thesis, the stochasticity in the birth-death-like tree topology however, does violate the expectation of the coalescent in terms of deterministic population growth, and the dynamics of trees produced by the I model is generally poorly estimated by the coalescent model (approximately in 70% of the trees true parameter was recovered).

The ability of the coalescent to properly reconstruct the parameters of dynamics of the I-type of trees decreases rapidly with increasing the sampling proportion, and this effect is most apparent at decreasing the basic reproductive ratio \((R_0)\). It has been postulated earlier that coalescent approximations are good for samples where sample size \(n\) is sufficiently smaller than the population size \(N : n(n - 1) \ll N\) [Rodrigo and Felsenstein, 1999] (note, we have always approximated \(N = N_e\)). This approximation is valid for sampling of nodes at present only (i.e. when samples are all extant species), and similar approximation to the serially sampled nodes has not been postulated. We could, however, by intuition extend the claim, and postulate that such inequality should hold for sample from the population at any time point in the process. Nevertheless, from our simulations, we can with certainty conclude that even sampling proportion of 0.01, at low \(R_0\) in I model trees seem to be too high for the coalescent to reconstruct the true dynamics of the process.

The assumption of small sampling proportion \((n \ll N)\) is not included in the coalescent model based on population evolving in continuous time according to the Moran model. Still, as the formulation of \(f(T|\eta)\) is the same for both the Wright-Fisher and Moran population models [Stadler, 2013b], and our findings clearly show importance of the sampling proportion, used during the simulations of I trees, for the recovery of the epidemiological parameters from the tree topology by the coalescent, it seems
that stochastic variation of the population size, captured better with increasing \( p \), is the main culprit for why coalescent does not properly estimate the growth rate parameter from the I model trees.

In contrast to the coalescent, stochasticity of population size fluctuations incorporated in the definition of the birth-death process allows the model to well approximate the growth rate of any tree produced by I-type dynamics, disregarding the sampling proportion used to sample the nodes. The birth-death model is also capable of capturing the dynamics of trees produced by the coalescent when the tip times, population size and the setting of the duration of one generation in calendar units used during the simulation of coalescent trees conform with assumptions of coalescent on low sampling proportion of nodes as compared to total population size. Overestimates of the coalescent tree growth rate are reported by the birth-death model in cases where this assumption of the coalescent is not met and the sampling proportion is apparently too high, again becoming more apparent at \( R_0 < 2 \), where the death rate is too high compared to the birth rate.

We have also shown in this thesis that the birth-death process produces parameter estimates with larger 95% HPD intervals and is a more accurate method for the growth rate parameter recovery on simulated datasets, which is consistent with simulations carried out in [Stadler et al., 2012]. Additionally, we have also shown that the size of the 95% HPD intervals is getting smaller for both coalescent and birth-death model with decreasing the sampling proportion \( p \) in simulations, indicating that longer growing trees provide more precise information for both models.

Furthermore, previously reported correlation in parameter estimates when analyzing \( \delta \), \( \lambda \), and \( p \) simultaneously [Stadler, 2009, Stadler et al., 2013] is confirmed in this research thesis by seeing the shift of and/or increase in the 95% HPD interval size of \( \lambda \), \( \delta \) and \( R_0 \) when setting \( p \) to a wrong value or to uniform prior, respectively. Correlation of the three parameters however does not influence the estimates of growth rate nor of \( \lambda \delta p \) parameters when \( p \) is unknown or set to a wrong value. Therefore, if analyzing a dataset with unknown sampling proportion, using uniform prior on \( p \) or setting it to a wrong random value still enables for growth rate and \( \lambda \delta p \) parameters recovery. Caution must be however taken when drawing any conclusions on \( \lambda \), \( \delta \) and \( R_0 \). These can be well estimated only when setting \( p \) to the true value.

In trees generated by more complex models such as SI and SIR, both coalescent and birth-death process underestimated the growth rate. However, the birth-death model performs much worse than the coalescent. We again suspect that due to the fact that the coalescent tries to fit the exponential growth to the part of the tree with highest increase in number of lineages (and ignores the parts of the \#infecteds-over-time curve, where there are less lineages than predicted by the steepest growing part of the curve), coalescent estimates slightly higher growth rates on SI/SIR trees.
than birth-death model.

Due to the complex nature the growth of population of infected individuals follows when total population size ($N$) is limited (such as in SI and SIR model), more sophisticated models of analysis than simple birth-death process or coalescent are required. Ideally, applying the skyline plot [Drummond et al., 2005, Stadler et al., 2013] analysis to such trees would provide us with more accurate estimates of changes of growth rate over time. The initial exponential growth of the trees could be separately analyzed and the growth rate that was initially used for simulation of the trees recovered. As pointed out by [Stadler et al., 2013], although the 95% HPD intervals for the parameters of the birth-death skyline model may increase, going back in past, the method may be more sensitive to parameter changes (such as $R_0$ in the HIV-1 epidemics in UK) than coalescent model (no change in the population size coinciding with introduction of antiretroviral therapy).

In [Stadler, 2009], higher estimate of the growth rate parameter of the HCV epidemics in Egypt was reported when data were analyzed with the coalescent model as compared to birth-death model. Furthermore, total interval of coalescent posterior for growth rate displayed larger spread (standard deviation) as compared to birth-death model. Whereas we cannot with certainty redefine this observation in terms of the impact on the 95% HPD lengths, it is an indication that coalescent produced larger HPD intervals than birth-death process. Similarly, in [Stadler et al., 2012], higher growth rate estimate by the coalescent with larger 95% HPD interval size as compared to the birth-death model was observed when HIV dataset from Switzerland was analyzed (note the inverse relationship in the 95% HPD sizes for the data simulated under I model in the same paper). Results from both of these studies are consistent with the epidemic following either the SI or the SIR model-like behavior, already reaching the post-exponential growth phase - see Figure 10 and Supplementary Figure S15. Confirming our intuition, application of the SI model to estimating the dynamics of the HIV epidemics in Switzerland, fitting more realistic model to the dataset than any of the previously applied models, reveals higher $R_0$ [Leventhal et al., 2013] than previously estimated [Stadler et al., 2012].
(6) Conclusions

We conclude by summarizing that for the true phylogenetic history of the population of pathogen (birth-death type of tree produced by our simulations) coalescent model is not an appropriate analytic method (for any sampling proportion). In contrast, if the process is governed by the coalescent type of evolution, the sampling proportion of the leaf nodes matters for the parameter recovery by the birth-death model, especially at low basic reproductive ratios. In general, when one knows what kind of process produced the tree, the appropriate method should be applied for analysis of such trees. In case of doubt of the underlying process generating the genealogy and/or in case of doubt of the sampling proportion of the leaf nodes, birth-death process is the method of choice for reliable phylodynamic parameter estimation. In addition, advantage of the birth-death model lies in the fact that it can also recover birth and death parameters individually, assuming the underlying process leading to the observed topology relies on the birth-death type of model behavior. Several studies deduced birth and death rate from the phylogenetic tree, assuming they handled birth-death type of trees (see for instance [Stadler and Bokma, 2013, Stadler et al., 2012]). The type of underlying tree (tree produced by the birth-death model, or by the coalescent process) in combination with the analytic method applied strictly determines whether and how well we can deduce the growth rate \( r = \lambda - \delta \) and whether we can disentangle these two parameters (\( \lambda \) and \( \delta \)).

Moreover, diagnostics of the parameter estimation based on comparison of birth-death tree estimates and coalescent estimates can help to pinpoint the process underlying the observed topology. If the two estimates are off and do not agree - further investigation of the process is required. It could be that none of the models sufficiently captures the dynamics of the studied populations, or alternatively, we suggest three most likely scenarios explaining the disagreement between the birth-death and the coalescent estimates. If the 95% HPD interval size is larger for growth rate parameter when estimated by coalescent as compared to birth-death estimates, then:

1) we are very likely dealing with SI/SIR type of population dynamics and in addition to observing larger HPD interval sizes we would expect to observe higher median estimates of the growth rate by the coalescent as compared to the birth-death model.

2) If the 95% HPD interval size are smaller for growth rate parameter when estimated by coalescent as compared to birth-death estimates, then, in case we observe overall lower parameter estimates by birth-death process than those by coalescent, we are probably dealing with an instance of the birth-death tree with high sampling proportion (especially at low \( R_0 \)).

3) Alternatively, it can be that the sampling proportion (at low \( R_0 \)) is high and the true underlying tree is an instance of realization of the coalescent, in which case the birth-death model also reports higher median value for the growth rate parameter than the coalescent.
(7) **Outlook**

It would be of great interest to further develop the concept of comparison of birth-death process and coalescent on growth rate parameter recovery started in this thesis by extending our simulations to use of sequences instead of fixed topology in the first place. The sequences can easily be simulated by for instance a publicly available Seq-Gen program [Rambaut and Grass, 1997], applying the simplest nucleotide substitution model: Jukes-Cantor (JC69). The tree topology can then be treated as nuisance parameter [Rodrigo and Felsenstein, 1999], and integrated out during the MCMC procedure.

Furthermore, we would be excited to see how well the coalescent and birth-death skyline models preform in estimating the growth rate of the SI/SIR epidemic, in particular when the post-exponential growth phases are included.
(8) References


A list of references in the given text:


(9) Appendix
(a) I model trees: $R_0=64, \lambda=3.2, \delta=0.05$

(b) Coalescent trees: $r=3.15$, tip times as in (a)

(c) I model trees: $R_0=1.1, \lambda=10, \delta=9$

(d) Coalescent trees: $r=1$, tip times as in (c)
(e) I model trees: $R_0=1.1, \lambda=1, \delta=0.9$

(f) Coalescent trees: $r=0.1$, tip times as in (e)

Figure S1: Comparison of birth-death model and coalescent with exponential growth rate model performance in growth rate parameter recovery when simulating trees with birth-death (I) model and with coalescent at scaled branch lengths.

For each plot, 100 trees simulated with stochastic I model (left column) or with coalescent with exponential growth rate (right column) under given parameter settings were analyzed with birth-death model (blue bars) or with coalescent with exponential growth rate (red bars). 95% HPD intervals of the growth rate parameter are shown (y-axis). Within each figure the trees are ordered (x-axis) by the median value of the growth rate parameter recovered by the coalescent (orange dot within the red bar). Median of the posterior interval for the growth rate parameter recovered by birth-death model is indicated as light blue dot within each blue interval. The true value of the growth rate parameter (the value under which the trees were simulated), is displayed as black horizontal bar. The plots a) and b) correspond to trees in Figure 2 a) and b), but the branches were scaled by a factor 0.1, the plots c) and d), and e) and f) correspond to the plots i) and j) in the same figure scaled by factor 18 and 1.8, respectively.
a) I model trees: \( R_0 = 64, \lambda = 32, \delta = 0.5, p = 0.5 \)
b) I model trees: $R_0=64, \lambda=3.2, \delta=0.05, p=0.5$
c) I model trees: \( R_0=8, \lambda=4, \delta=0.5, p=1 \)
d) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.5$
e) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.01$
f) I model trees: $R_0=4, \lambda=2, \delta=0.5, p=0.5$
g) I model trees: $R_0=2, \lambda=1, \delta=0.5, p=0.5$
h) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=1$
i) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.5$
j) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.01$
k) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.001$
1) I model trees: $R_0=1.1, \lambda=1, \delta=0.9, p=0.5$
m) I model trees: \( R_0=1, \lambda=0.5, \delta=0.5, p=0.5 \)

Figure S2: Tracking of cumulative infecteds, infecteds and sampled individuals over time in I model trees
For each of 100 trees simulated with stochastic I model we counted the number of cumulative infected, infected and sampled individuals at each time step. The x-axis represents the time (going backwards from present \( t = 0 \)) and y-axis the counts. Each line represents history of one tree over time. The left column shows the trees colored on the basis of whether coalescent was able to correctly recover (grey lines) or not (red lined for over- and orange lines for under-estimated) true growth rate. The right column shows the same for birth-death process applied to the same trees. Plots a) - k) are for different settings for \( \lambda, \delta \) and \( p \) in simulations - see the header above each sub-plot.
(a) I model trees: \( R_0=8, \lambda=4, \delta=0.5, p=1 \)

(b) Coalescent trees: \( r=3.5 \), tip times as in (a)

(c) I model trees: \( R_0=8, \lambda=4, \delta=0.5, p=0.5 \)

(d) Coalescent trees: \( r=3.5 \), tip times as in (c)
(e) I model trees: \( R_0 = 8, \lambda = 4, \delta = 0.5, p = 0.01 \)

(f) Coalescent trees: \( r = 3.5 \), tip times as in (e)

Figure S3: **Role of sampling proportion in growth rate parameter recovery at medium \( R_0 \)**

100 trees were simulated with I model with \( \lambda = 4, \delta = 0.5 \), and \( p = [0.01, 0.5, 1] \) - left column or with coalescent when using same sampling times, population size and setting the duration of one generation in calendar units as in the corresponding I model trees - right-hand side columns. The trees were analyzed with birth-death model (blue bars) or with coalescent with exponential growth rate (red bars). 95% HPD intervals of the growth rate parameter are shown (y-axis). Within each figure the trees are ordered (x-axis) by the median value of the growth rate parameter recovered by the coalescent (orange dot within the red bar). Median of the posterior interval for the growth rate parameter recovered by birth-death model is indicated as light blue dot within each blue interval. The graphs are ordered in decreasing order by sampling proportion values \( p \). Subfigures c) and d) are identical to subfigures c) and d), respectively, in Figure 2. The value of the growth rate parameter under which the trees were simulated, is displayed as black horizontal bar.
(a) I model trees: $R_0 = 64, \lambda = 3.2, \delta = 0.05, p = 0.5$

(b) Coalescent trees: $r = 3.15$, tip times as in (a)

(c) I model trees: $R_0 = 8, \lambda = 4, \delta = 0.5, p = 1$

(d) Coalescent trees: $r = 3.5$, tip times as in (c)
(e) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.01$

(f) Coalescent trees: $r=3.5$, tip times as in (e)

(g) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=1$

(h) Coalescent trees: $r=1$, tip times as in (g)
(i) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.5$

(j) Coalescent trees: $r=1$, tip times as in (i)

(k) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.01$

(l) Coalescent trees: $r=1$, tip times as in (k)
(m) I model trees: \( R_0 = 1.1, \lambda = 10, \delta = 9, p = 0.001 \)

(n) Coalescent trees: \( r = 1 \), tip times as in (m)

(o) I model trees: \( R_0 = 1.1, \lambda = 1, \delta = 0.9, p = 0.5 \)

(p) Coalescent trees: \( r = 0.1 \), tip times as in (o)
Figure S4: Comparison of birth-death model and coalescent with exponential growth rate 95% HPD intervals size for growth rate parameter at fixed $R_0$ and varying $p$

Sizes of 95% HPD interval of growth rate parameter for each of 100 trees are displayed. For each tree, the estimates of the growth rate by birth-death model (blue rectangles) are joined with the corresponding estimates by coalescent (red rectangles) by a vertical line. If the true growth rate value was not captured in the 95% HPD interval, the rectangle is larger and of more intense color. The trees produced by stochastic I model are displayed in left-hand column and corresponding coalescent trees with same tip sampling times are displayed in right-hand column. Within each figure the trees are ordered (x-axis) by the median value of the growth rate parameter recovered by the coalescent. The abbreviations in the legend stand for the following: BD - birth-death model, CE - coalescent with exponential growth rate.
(a) I model trees: true $p=0.01$, analyzed under $p=0.01$ and $p=0.5$

(b) Coalescent trees: $r=3.5$, tip times as in (a), analyzed under $p=0.01$ and $p=0.5$

(c) I model trees: true $p=0.01$, analyzed under $p=0.01$ and $p=1$

(d) Coalescent trees: $r=3.5$, tip times as in (c), analyzed under $p=0.01$ and $p=1
(e) I model trees: true $p=0.5$, analyzed under $p=0.5$ and $p=0.01$

95% HPD interval of $\lambda d \Delta x p$

(f) Coalescent trees: $r=3.5$, tip times as in (e), analyzed under $p=0.5$ and $p=0.01$

95% HPD interval of $\lambda d \Delta x p$

(g) I model trees: true $p=0.5$, analyzed under $p=0.5$ and $p=1$

95% HPD interval of $\lambda d \Delta x p$

(h) Coalescent trees: $r=3.5$, tip times as in (g), analyzed under $p=0.5$ and $p=1$

95% HPD interval of $\lambda d \Delta x p$
(i) I model trees: true $p=1$, analyzed under $p=1$ and $p=0.01$

(j) Coalescent trees: $r=3.5$, tip times as in (i), analyzed under $p=1$ and $p=0.01$

(k) I model trees: true $p=1$, analyzed under $p=1$ and $p=0.5$

(l) Coalescent trees: $r=3.5$, tip times as in (k), analyzed under $p=1$ and $p=0.5
Figure S5: **Interchangeability of sampling proportion in $\lambda \delta p$ parameter recovery**

100 trees were simulated with I model with ($R_0 = 8$) $\lambda = 4$, $\delta = 0.5$, and $p = [0.01, 0.5, 1]$ - left column or with coalescent when using same sampling times, population size and setting the duration of one generation in calendar units as in the corresponding I model trees - right-hand side columns, and analyzed with birth-death model assuming either the true $p$ or untrue $p$. Within each figure the trees are ordered (x-axis) by the median value of growth rate parameter recovered from trees analyzed under true $p$. The graphs display the 95% HPD of the $\lambda \delta p$ parameter and are ordered in increasing order by sampling proportion values used during the tree simulation. The value of $\lambda \delta p$ under which the trees were simulated, is displayed as black horizontal bar.
(a) I model trees: true $p=0.01$, analyzed under $p=0.01$ and $p=0.5$

95% HPD interval of lambda

(b) Coalescent trees: $r=3.5$, tip times as in (a), analyzed under $p=0.01$ and $p=0.5$

95% HPD interval of lambda

(c) I model trees: true $p = 0.01$, analyzed under $p=0.01$ and $p=1$

95% HPD interval of lambda

(d) Coalescent trees: $r=3.5$, tip times as in (c), analyzed under $p=0.01$ and $p=1$

95% HPD interval of lambda
(e) I model trees: true $p=0.5$, analyzed under $p=0.5$ and $p=0.01$

(f) Coalescent trees: $r=3.5$, tip times as in (e), analyzed under $p=0.5$ and $p=0.01$

(g) I model trees: true $p=0.5$, analyzed under $p=0.5$ and $p=1$

(h) Coalescent trees: $r=3.5$, tip times as in (g), analyzed under $p=0.5$ and $p=1
(i) I model trees: true $p=1$, analyzed under $p=1$ and $p=0.01$

<table>
<thead>
<tr>
<th>95% HPD interval of lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation number</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>95% HPD interval of lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation number</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>100</td>
</tr>
</tbody>
</table>

(j) Coalescent trees: $r=3.5$, tip times as in (i), analyzed under $p=1$ and $p=0.01$

<table>
<thead>
<tr>
<th>95% HPD interval of lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation number</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>95% HPD interval of lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation number</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>100</td>
</tr>
</tbody>
</table>

(k) I model trees: true $p=1$, analyzed under $p=1$ and $p=0.5$

<table>
<thead>
<tr>
<th>95% HPD interval of lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation number</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>95% HPD interval of lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation number</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>100</td>
</tr>
</tbody>
</table>

(l) Coalescent trees: $r=3.5$, tip times as in (k), analyzed under $p=1$ and $p=0.5$

<table>
<thead>
<tr>
<th>95% HPD interval of lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation number</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>95% HPD interval of lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation number</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>100</td>
</tr>
</tbody>
</table>

91
Figure S6: **Interchangeability of sampling proportion in $\lambda$ parameter recovery**

100 trees were simulated with I model with ($R_0 = 8$)$\lambda = 4, \delta = 0.5$, and $p = [0.01, 0.5, 1]$ - left column or with coalescent when using same sampling times, population size and setting the duration of one generation in calendar units as in the corresponding I model trees - right-hand side columns, and analyzed with birth-death model assuming either the true $p$ or untrue $p$. Within each figure the trees are ordered (x-axis) by the median value of growth rate parameter recovered from trees analyzed under true $p$. The graphs display the 95% HPD of the $\lambda$ parameter and are ordered in increasing order by sampling proportion values used during the tree simulation. The value of $\lambda$ under which the trees were simulated, is displayed as black horizontal bar.
(a) I model trees: true $p=0.01$, analyzed under $p=0.01$ and $p=0.5$

(b) Coalescent trees: $r=3.5$, tip times as in (a), analyzed under $p=0.01$ and $p=0.5$

(c) I model trees: true $p=0.01$, analyzed under $p=0.01$ and $p=1$

(d) Coalescent trees: $r=3.5$, tip times as in (c), analyzed under $p=0.01$ and $p=1
(e) I model trees: true $p=0.5$, analyzed under $p=0.5$ and $p=0.01$

(f) Coalescent trees: $r=3.5$, tip times as in (e), analyzed under $p=0.5$ and $p=0.01$

(g) I model trees: true $p=0.5$, analyzed under $p=0.5$ and $p=1$

(h) Coalescent trees: $r=3.5$, tip times as in (g), analyzed under $p=0.5$ and $p=1
(i) I model trees: true $p=1$, analyzed under $p=1$ and $p=0.01$

(j) Coalescent trees: $r=3.5$, tip times as in (i), analyzed under $p=1$ and $p=0.01$

(k) I model trees: true $p=1$, analyzed under $p=1$ and $p=0.5$

(l) Coalescent trees: $r=3.5$, tip times as in (k), analyzed under $p=1$ and $p=0.5
Figure S7: **Interchangeability of sampling proportion in $\delta$ parameter recovery**

100 trees were simulated with I model with ($R_0 = 8$) $\lambda = 4, \delta = 0.5$, and $p = [0.01, 0.5, 1]$ - left column or with coalescent when using same sampling times, population size and setting the duration of one generation in calendar units as in the corresponding I model trees - right-hand side columns, and analyzed with birth-death model assuming either the true $p$ or untrue $p$. Within each figure the trees are ordered (x-axis) by the median value of *growth rate* parameter recovered from trees analyzed under true $p$. The graphs display the 95% HPD of the $\delta$ parameter and are ordered in increasing order by sampling proportion values used during the tree simulation. The value of $\delta$ under which the trees were simulated, is displayed as black horizontal bar.
(a) I model trees: true \( p=0.01 \), analyzed under \( p=0.01 \) and \( p=0.5 \)

(b) Coalescent trees: \( r=3.5 \), tip times as in (a), analyzed under \( p=0.01 \) and \( p=0.5 \)

(c) I model trees: true \( p=0.01 \), analyzed under \( p=0.01 \) and \( p=1 \)

(d) Coalescent trees: \( r=3.5 \), tip times as in (c), analyzed under \( p=0.01 \) and \( p=1 \)
(e) I model trees: true $p=0.5$, analyzed under $p=0.5$ and $p=0.01$

(f) Coalescent trees: $r=3.5$, tip times as in (e), analyzed under $p=0.5$ and $p=0.01$

(g) I model trees: true $p=0.5$, analyzed under $p=0.5$ and $p=1$

(h) Coalescent trees: $r=3.5$, tip times as in (g), analyzed under $p=0.5$ and $p=1
(i) I model trees: true $p=1$, analyzed under $p=1$ and $p=0.01$

(j) Coalescent trees: $r=3.5$, tip times as in (i), analyzed under $p=1$ and $p=0.01$

(k) I model trees: true $p=1$, analyzed under $p=1$ and $p=0.5$

(l) Coalescent trees: $r=3.5$, tip times as in (k), analyzed under $p=1$ and $p=0.5
Figure S8: **Interchangeability of sampling proportion in $R_0$ parameter recovery**

100 trees were simulated with I model with ($R_0 = 8$) $\lambda = 4$, $\delta = 0.5$, and $p = [0.01, 0.5, 1]$ - left column or with coalescent when using same sampling times, population size and setting the duration of one generation in calendar units as in the corresponding I model trees - right-hand side columns, and analyzed with birth-death model assuming either the true $p$ or untrue $p$. Within each figure the trees are ordered (x-axis) by the median value of growth rate parameter recovered from trees analyzed under true $p$. The graphs display the 95% HPD of the $R_0$ parameter and are ordered in increasing order by sampling proportion values used during the tree simulation. The value of $R_0$ under which the trees were simulated, is displayed as black horizontal bar.
(a) I model trees: $R_0=64, \lambda=32, \delta=0.5, p=0.5$

(b) Coalescent trees: $r=31.5$, tip times as in (a)

(c) I model trees: $R_0=64, \lambda=3.2, \delta=0.05, p=0.5$

(d) Coalescent trees: $r=3.15$, tip times as in (c)
(e) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=1$

(f) Coalescent trees: $r=3.5$, tip times as in (e)

(g) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.5$

(h) Coalescent trees: $r=3.5$, tip times as in (g)
(i) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.01$

(j) Coalescent trees: $r=3.5$, tip times as in (i)

(k) I model trees: $R_0=4, \lambda=2, \delta=0.5, p=0.5$

(l) Coalescent trees: $r=1.5$, tip times as in (k)
(m) I model trees: \( R_0 = 2, \lambda = 1, \delta = 0.5, p = 0.5 \)

\[
\begin{align*}
0 & \quad 20 & \quad 40 & \quad 60 & \quad 80 & \quad 100 \\
0.00 & \quad 0.05 & \quad 0.10 & \quad 0.15 & \quad 0.20 & \quad 0.25 & \quad 0.30 \\
95\% \text{ HPD interval of } \lambda \delta p
\end{align*}
\]

(n) Coalescent trees: \( r = 0.5 \), tip times as in (m)

(o) I model trees: \( R_0 = 1.1, \lambda = 10, \delta = 9, p = 1 \)

\[
\begin{align*}
0 & \quad 20 & \quad 40 & \quad 60 & \quad 80 & \quad 100 \\
0 & \quad 50 & \quad 100 & \quad 150 & \quad 200 \\
95\% \text{ HPD interval of } \lambda \delta p
\end{align*}
\]

(p) Coalescent trees: \( r = 1 \), tip times as in (o)
(q) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.5$

(r) Coalescent trees: $r=1$, tip times as in (q)

(s) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.01$

(t) Coalescent trees: $r=1$, tip times as in (s)
(u) I model trees: \( R_0=1.1, \lambda=10, \delta=9, p=0.001 \)

(v) Coalescent trees: \( r=1 \), tip times as in (u)

(w) I model trees: \( R_0=1.1, \lambda=1, \delta=0.9, p=0.5 \)

(x) Coalescent trees: \( r=0.1 \), tip times as in (w)
(y) I model trees: \( R_0=1.1, \lambda=0.55, \delta=0.5, p=0.5 \)

(z) Coalescent trees: \( r=0.05 \), tip times as in (y)

(aa) I model trees: \( R_0=1, \lambda=0.5, \delta=0.5, p=0.5 \)

(bb) Coalescent trees: \( r=0.0 \), tip times as in (aa)
Figure S9: **Uniform unfixed sampling proportion in $\lambda\delta p$ parameter recovery**

100 trees were simulated with I model under various combinations of $\lambda$, $\delta$ and $p$ - left column or with coalescent when using same sampling times, population size and setting the duration of one generation in calendar units as in the corresponding I model trees - right-hand side columns, and analyzed with birth-death model assuming either the true $p$ or $p$ with uniform prior over [0,1]. Within each figure the trees are ordered (x-axis) by the median value of growth rate parameter recovered from trees analyzed under true $p$. The graphs display the 95% HPD of the $\lambda\delta p$ parameter and are ordered in decreasing order of $R_0$. The value of $\lambda\delta p$ under which the trees were simulated, is displayed as black horizontal bar.
(a) I model trees: \( R_0 = 64, \lambda = 32, \delta = 0.5, p = 0.5 \)

(b) Coalescent trees: \( r = 31.5 \), tip times as in (a)

(c) I model trees: \( R_0 = 64, \lambda = 3.2, \delta = 0.05, p = 0.5 \)

(d) Coalescent trees: \( r = 3.15 \), tip times as in (c)
(e) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=1$

(f) Coalescent trees: $r=3.5$, tip times as in (e)

(g) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.5$

(h) Coalescent trees: $r=3.5$, tip times as in (g)
(i) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.01$

(j) Coalescent trees: $r=3.5$, tip times as in (i)

(k) I model trees: $R_0=4, \lambda=2, \delta=0.5, p=0.5$

(l) Coalescent trees: $r=1.5$, tip times as in (k)
(m) I model trees: $R_0=2, \lambda=1, \delta=0.5, p=0.5$

(n) Coalescent trees: $r=0.5$, tip times as in (m)

(o) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=1$

(p) Coalescent trees: $r=1$, tip times as in (o)
(q) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.5$

(r) Coalescent trees: $r=1$, tip times as in (q)

(s) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.01$

(t) Coalescent trees: $r=1$, tip times as in (s)

113
(u) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.001$

(v) Coalescent trees: $r=1$, tip times as in (u)

(w) I model trees: $R_0=1.1, \lambda=1, \delta=0.9, p=0.5$

(x) Coalescent trees: $r=0.1$, tip times as in (w)
(y) I model trees: $R_0=1.1, \lambda=0.55, \delta=0.5, p=0.5$

(z) Coalescent trees: $r=0.05$, tip times as in (y)

(aa) I model trees: $R_0=1, \lambda=0.5, \delta=0.5, p=0.5$

(bb) Coalescent trees: $r=0.0$, tip times as in (aa)
Figure S10: **Uniform unfixed sampling proportion - exploration of prior on p**

100 trees were simulated with I model under various combinations of $\lambda$, $\delta$ and $p$ - left column or with coalescent when using same sampling times, population size and setting the duration of one generation in calendar units as in the corresponding I model trees - right-hand side columns, and analyzed with birth-death model assuming either the true $p$ or $p$ with uniform prior over $[0,1]$. Within each figure the trees are ordered (x-axis) by the median value of growth rate parameter recovered from trees analyzed under true $p$. The graphs display the 95% HPD of the $p$ parameter and are ordered in decreasing order of $R_0$. The value of $p$ under which the trees were simulated, is displayed as black horizontal bar.
(a) I model trees: $R_0=64, \lambda=32, \delta=0.5, p=0.5$

(b) Coalescent trees: $r=31.5$, tip times as in (a)

(c) I model trees: $R_0=64, \lambda=3.2, \delta=0.05, p=0.5$

(d) Coalescent trees: $r=3.15$, tip times as in (c)
(e) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=1$

(f) Coalescent trees: $r=3.5$, tip times as in (e)

(g) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.5$

(h) Coalescent trees: $r=3.5$, tip times as in (g)
(i) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.01$

(j) Coalescent trees: $r=3.5$, tip times as in (i)

(k) I model trees: $R_0=4, \lambda=2, \delta=0.5, p=0.5$

(l) Coalescent trees: $r=1.5$, tip times as in (k)
(m) I model trees: $R_0=2, \lambda=1, \delta=0.5, p=0.5$

(o) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=1$

(n) Coalescent trees: $r=0.5$, tip times as in (m)

(p) Coalescent trees: $r=1$, tip times as in (o)
(q) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.5$

(r) Coalescent trees: $r=1$, tip times as in (q)

(s) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.01$

(t) Coalescent trees: $r=1$, tip times as in (s)
(u) I model trees: \( R_0=1.1, \lambda=10, \delta=9, p=0.001 \)

(v) Coalescent trees: \( r=1 \), tip times as in (u)

(w) I model trees: \( R_0=1.1, \lambda=1, \delta=0.9, p=0.5 \)

(x) Coalescent trees: \( r=0.1 \), tip times as in (w)
(y) I model trees: $R_0=1.1, \lambda=0.55, \delta=0.5, p=0.5$

(z) Coalescent trees: $r=0.05$, tip times as in (y)

(aa) I model trees: $R_0=1, \lambda=0.5, \delta=0.5, p=0.5$

(bb) Coalescent trees: $r=0.0$, tip times as in (aa)
Figure S11: **Uniform unfixed sampling proportion in λ parameter recovery**

100 trees were simulated with I model under various combinations of λ, δ and p - left column or with coalescent when using same sampling times, population size and setting the duration of one generation in calendar units as in the corresponding I model trees - right-hand side columns, and analyzed with birth-death model assuming either the true p or p with uniform prior over [0,1]. Within each figure the trees are ordered (x-axis) by the median value of growth rate parameter recovered from trees analyzed under true p. The graphs display the 95% HPD of the λ parameter and are ordered in decreasing order of R₀. The value of λ under which the trees were simulated, is displayed as black horizontal bar.
(a) I model trees: $R_0=64, \lambda=32, \delta=0.5, p=0.5$

(b) Coalescent trees: $r=31.5$, tip times as in (a)

(c) I model trees: $R_0=64, \lambda=3.2, \delta=0.05, p=0.5$

(d) Coalescent trees: $r=3.15$, tip times as in (c)
(e) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=1$

(f) Coalescent trees: $r=3.5$, tip times as in (e)

(g) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.5$

(h) Coalescent trees: $r=3.5$, tip times as in (g)
(i) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.01$

(j) Coalescent trees: $r=3.5$, tip times as in (i)

(k) I model trees: $R_0=4, \lambda=2, \delta=0.5, p=0.5$

(l) Coalescent trees: $r=1.5$, tip times as in (k)
(m) I model trees: \( R_0 = 2, \lambda = 1, \delta = 0.5, p = 0.5 \)

(n) Coalescent trees: \( r = 0.5 \), tip times as in (m)

(o) I model trees: \( R_0 = 1.1, \lambda = 10, \delta = 9, p = 1 \)

(p) Coalescent trees: \( r = 1 \), tip times as in (o)
(q) I model trees: \( R_0=1.1, \lambda=10, \delta=9, p=0.5 \)

(r) Coalescent trees: \( r=1 \), tip times as in (q)

(s) I model trees: \( R_0=1.1, \lambda=10, \delta=9, p=0.01 \)

(t) Coalescent trees: \( r=1 \), tip times as in (s)
(u) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.001$

(v) Coalescent trees: $r=1$, tip times as in (u)

(w) I model trees: $R_0=1.1, \lambda=1, \delta=0.9, p=0.5$

(x) Coalescent trees: $r=0.1$, tip times as in (w)
(y) I model trees: \( R_0=1.1, \lambda=0.55, \delta=0.5, p=0.5 \)

(z) Coalescent trees: \( r=0.05 \), tip times as in (y)

(aa) I model trees: \( R_0=1, \lambda=0.5, \delta=0.5, p=0.5 \)

(bb) Coalescent trees: \( r=0.0 \), tip times as in (aa)
Figure S12: Uniform unfixed sampling proportion in δ parameter recovery

100 trees were simulated with I model under various combinations of λ, δ and p - left column or with coalescent when using same sampling times, population size and setting the duration of one generation in calendar units as in the corresponding I model trees - right-hand side columns, and analyzed with birth-death model assuming either the true p or p with uniform prior over [0,1]. Within each figure the trees are ordered (x-axis) by the median value of growth rate parameter recovered from trees analyzed under true p. The graphs display the 95% HPD of the δ parameter and are ordered in decreasing order of $R_0$. The value of δ under which the trees were simulated, is displayed as black horizontal bar.
(a) I model trees: $R_0=64, \lambda=32, \delta=0.5, p=0.5$

(b) Coalescent trees: $r=31.5$, tip times as in (a)

(c) I model trees: $R_0=64, \lambda=3.2, \delta=0.05, p=0.5$

(d) Coalescent trees: $r=3.15$, tip times as in (c)
(e) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=1$

(f) Coalescent trees: $r=3.5$, tip times as in (e)

(g) I model trees: $R_0=8, \lambda=4, \delta=0.5, p=0.5$

(h) Coalescent trees: $r=3.5$, tip times as in (g)
(i) I model trees: \( R_0=8, \lambda=4, \delta=0.5, p=0.01 \)

(j) Coalescent trees: \( r=3.5 \), tip times as in (i)

(k) I model trees: \( R_0=4, \lambda=2, \delta=0.5, p=0.5 \)

(l) Coalescent trees: \( r=1.5 \), tip times as in (k)
(m) I model trees: $R_0=2, \lambda=1, \delta=0.5, p=0.5$

(n) Coalescent trees: $r=0.5$, tip times as in (m)

(o) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=1$

(p) Coalescent trees: $r=1$, tip times as in (o)
(q) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.5$

(r) Coalescent trees: $r=1$, tip times as in (q)

(s) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.01$

(t) Coalescent trees: $r=1$, tip times as in (s)
(u) I model trees: $R_0=1.1, \lambda=10, \delta=9, p=0.001$

(v) Coalescent trees: $r=1$, tip times as in (u)

(w) I model trees: $R_0=1.1, \lambda=1, \delta=0.9, p=0.5$

(x) Coalescent trees: $r=0.1$, tip times as in (w)
(y) I model trees: $R_0=1.1, \lambda=0.55, \delta=0.5, p=0.5$

(z) Coalescent trees: $r=0.05$, tip times as in (y)

(aa) I model trees: $R_0=1, \lambda=0.5, \delta=0.5, p=0.5$

(bb) Coalescent trees: $r=0.0$, tip times as in (aa)
Figure S13: **Uniform unfixed sampling proportion in $R_0$ parameter recovery**

100 trees were simulated with I model under various combinations of $\lambda$, $\delta$, and $p$ - left column or with coalescent when using same sampling times, population size and setting the duration of one generation in calendar units as in the corresponding I model trees - right-hand side columns, and analyzed with birth-death model assuming either the true $p$ or $p$ with uniform prior over $[0,1]$. Within each figure the trees are ordered (x-axis) by the median value of growth rate parameter recovered from trees analyzed under true $p$. The graphs display the 95% HPD of the $R_0$ parameter and are ordered in decreasing order of $R_0$. The value of $R_0$ under which the trees were simulated, is displayed as black horizontal bar.
a) SI model trees: $N=10,000$
b) SI model trees: $N=500$
c) SI model trees: $N=300$
d) SIR model trees: $N=10,000$
e) SIR model trees: $N=500$
f) SIR model trees: $N=300$

Figure S14: Tracking of cumulative infecteds, infecteds and sampled individuals over time in SI and SIR model-based trees

For each of 100 trees simulated with stochastic SI or SIR model we counted the number of cumulative infected, infected and sampled individuals at each time step. The x-axis represents the time (going backwards from present $t=0$) and y-axis the counts. One line represents history of one tree over time. The left column shows the trees colored on the basis of whether coalescent was able to correctly recover (grey lines) or not (red lined for over- and orange lines for under-estimated) true growth rate. The right column shows the same for birth-death process applied to the same trees. Plots a) - f) are for different settings of $N$ (population size) in simulations - see the header above each sub-plot. All trees were simulated under $\lambda = 4$, $\delta = 0.5$, $p = 0.5$. 

146
(a) I model trees: $N=\infty$

(b) SI model trees: $N=10,000$

(c) SIR model trees: $N=10,000$
(d) SI model trees: \( N=500 \)  

![Graph of Lengths of 95% HPD intervals for SI model trees with \( N=500 \)]  

(e) SIR model trees: \( N=500 \)  

![Graph of Lengths of 95% HPD intervals for SIR model trees with \( N=500 \)]  

(f) SI model trees: \( N=300 \)  

![Graph of Lengths of 95% HPD intervals for SI model trees with \( N=300 \)]  

(g) SIR model trees: \( N=300 \)  

![Graph of Lengths of 95% HPD intervals for SIR model trees with \( N=300 \)]
Figure S15: **Comparison of birth-death model and coalescent with exponential growth rate 95% HPD intervals size for growth rate parameter in SI and SIR model-based trees**

For each plot, 100 trees simulated with stochastic SI model (left column) or SIR model (right column) under $\lambda = 4$, $\delta = 0.5$, $p = 0.5$ and population size $N$, indicated above each plot. Sizes of 95% HPD interval of growth rate parameter for each of 100 trees are displayed. For each tree, the estimates of the growth rate by birth-death model (blue rectangles) are joined with the corresponding estimates by coalescent (red rectangles) by a vertical line. If the true growth rate value was not captured in the 95% HPD interval, the rectangle is larger and of more intense color. Within each figure the trees are ordered (x-axis) by the median value of the growth rate parameter recovered by the coalescent. Note, plot a) is identical to Figure 7 c).
Table ST1: Growth rate recovery counts at scaled branch lengths
For each of the 100 trees simulated with the respective method (I model or coalescent) we counted the number of times the method applied (birth-death or coalescent) recovered or missed the true growth rate value in the 95% HPD interval. The summary for each $R_0$ setting with $p = 0.5$ across all settings is displayed.