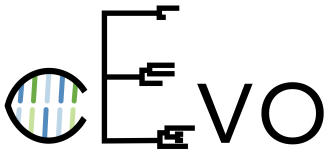
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



Quantitative comparison of phylodynamic models for structured populations

Master Thesis

Sophie Seidel

seidels@ethz.ch

Department of Biosystems Science and Engineering
Computational Evolution
ETH Zürich

Supervisors:
Supervisor 1, Supervisor 2
Prof. Dr. Tanja Stadler, Dr. Timothy Vaughan

February 15, 2019

Acknowledgements

I would like to take these last moments before my submission deadline to thank the whole cEvo group for providing such a great social and intellectual setting during the past six months. Special thanks goes to my supervisors Timothy Vaughan and Tanja Stadler for their advice and help on this way.

Abstract

Quantifying population dynamic parameters from genetic sequence data is one of the central aims of phylodynamic inference. The two most widely used models, namely the birth-death process and the coalescent, have been extended to include population structure in the recent years. Even though both of them are used extensively, there is a clear lack of understanding on how they compare. We performed a simulation study to investigate their performance under a representative set of scenarios including varying growth, sampling and migration. We find that the structured coalescent cannot correctly estimate neither the migration rates nor the effective population sizes for exponentially growing populations. For almost constant population sizes, it provides more accurate estimates of the migration rates than the BD model. We detect signs for improved accuracy in the migration rate estimates when the average branch length is shorter.

Contents

Chapter 1

Introduction

Phylodynamic models are the mathematical glue that tie evolutionary relationships between organisms to the large-scale dynamics of the population to which they belong. They have been used to shed light on the global transmission dynamics of influenza [1, 2, 3, 4], the origin of HIV [5] and informed disease prevention efforts [6]. Further, their usage provided deeper insights into the 2014 Ebola epidemic [7] and into outbreaks of slowly evolving pathogens like tuberculosis [8]. Apart from epidemic applications, they illuminated the demographic history of species like humans [9] and great apes [10], or the origin of polar bears [11].

Although the term "Phylodynamics" was coined by Grenfell et al. [12] as the "melding of immunodynamics, epidemiology, and evolutionary biology" in 2004, some of its conceptual insights date back until 1876. It was already then, that Ernst Häckel drew a hypothetical tree (Fig. 1.1) to depict the distribution of humans across the globe, trying to merge their current location with possible migration routes in the past and hence possible ancestry.

At least three major innovations were needed to move from a hypothetical picture to the statistical models of population dynamics we have today. With the advent of DNA sequencing technology, we became able to make statements about the relationships between individuals based on the similarity of their DNA sequences. The major advantages of DNA are i) it is shared by all living organisms and ii) comparing its 4 bases is easier than sizes of phenomenological traits. Availability of the sequencing data alone does not help if one cannot formulate a model to link the input, the DNA sequences, to the desired output, variables describing the population dynamics. We shall introduce the models providing such a description in section 1.1. Finally, we need to be able to compute our model's output variables in a reasonable amount of time. We will briefly touch on the framework of Bayesian Inference that allows us to do so in section 1.2.

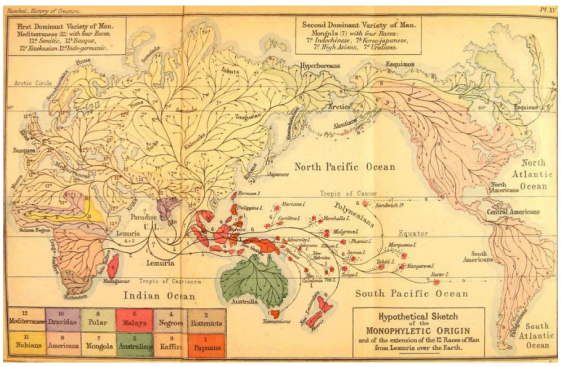


Figure 1.1: Hypothetical tree by Ernst Häckel depicting the spatial dispersal of humans over Earth.

1.1 Structured Phylodynamics

Phylodynamics can be used to elucidate population dynamics from genetic sequence information. For biological intuition we will use the infectious disease scenario in the following section. However, these results hold more generally. We define a population of infected hosts in different geographical locations or demes. Mathematically precise, a population of d subpopulations, consists of $N_1, ..., N_d$ individuals. For the purpose of this thesis, we will restrict ourselves to $d = 2$.

1.1.1 The Multitype Birth-Death Model

In the multitype birth-death (BD) model [13], building on [14, 15, 16, 17, 18], we always start with one individual in location $i \in \{1, 2\}$, at time $t = 0$. After an exponentially distributed waiting time Δt, it can:

- give birth to (infect) another individual at rate λ_i:

$$N_i(t + \Delta t) = N_i(t) + 1$$

- recover at rate μ_i

$$N_i(t + \Delta t) = N_i(t) - 1$$

- migrate to location j at rate m_{ij}:

$$N_i(t + \Delta t) = N_i(t) - 1$$

$$N_j(t + \Delta t) = N_j(t) + 1$$

- be sampled at rate ψ_i, meaning we observe this infected individual, e.g. upon hospitalization. We assume that sampling leads to removal from the population:

$$N_i(t + \Delta t) = N_i(t) - 1$$

This process generates a full tree including all sampled and non-sampled individuals. What we observe in reality is the sampled tree, consisting only of the sampled tips and the branches connecting them.

That the BD model explicitly models the sampling process, can be an advantage since it allows practitioners to include information potentially available to them, e.g. if one location is oversampled compared to another. The situation becomes problematic, if no such information is present. Even worse, if the sampling process in reality deviates from the model's assumptions, estimates of the population dynamic parameters can be biased [19].

Even though apparent from the equations above, we would like to highlight that population size varies stochastically in the BD. For the growth rate $r_i = \lambda_i - \mu_i$ the expected population size at time t can be obtained by [14]:

$$E\left[N_i(t)\right] = N_i(0) \cdot e^{r_i \cdot t} \tag{1.1}$$

We can calculate the likelihood of the rate parameters and a given sampled tree, by evaluating the probability that an individual evolved as observed in the tree. For more details, we refer the reader to [20].

1.1.2 The Structured Coalescent

The structured coalescent (SC) [21, 22, 23] extends Kingman's coalescent [24] in allowing individuals to belong to different subpopulations and to migrate between them.

It assumes that a structured Wright-Fisher model [25] can describe the underlying population dynamics. In that model, N_d individuals in location $d \in \{1, ..., k\}$, evolve through discrete generations of length g. Although the results apply for general k, we shall continue for the case

relevant to us: $k = 2$. At the end of each generation, individuals migrate from d to d' with probability $m_{dd'} \cdot g$ or stay with probability $1 - m_{dd'} \cdot g$. Subsequently, N_d individuals are sampled with replacement from subpopulation d and a new generation starts. Note, that the subpopulation sizes N_d hence always remain constant. For the purpose of this thesis we will set N_d to N_d^e, the effective population size, if the population depart from the Wright-Fisher model.

Finally, we need to be able to calculate the likelihood of a phylogenetic tree under the structured coalescent model. The derivation here should provide an intuition on the parameters that enter the likelihood calculation. For a more rigorous treatment see [26]. Starting with n sampled lineages at present time $t = 0$, after an exponentially distributed waiting time Δt towards the past, one individual i in location d can:

- coalesce with another individual at rate:

$$c_d = \frac{1}{N_d \cdot g}$$

- migrate to d' at rate $m_{dd'}$ as before, corresponding to the immigration rate [23]:

$$q_{d'd} = m_{dd'} \cdot \frac{N_d}{N_{d'}}$$

To approximate the time to coalescence by an exponential distribution, one needs to assume that the infected population is considerably larger than the number of sampled individuals [24, 27]. The time spanned by the tree is divided into adjacent intervals $\boldsymbol{\tau} = (\tau_0, \tau_1, ..., \tau_f)$. In Fig. 1.2, τ_0 would correspond to the interval from $t0$ to $t1$.

Summarising our parameters $\mathbf{m}$, $\boldsymbol{\tau}$ and $\mathbf{N}$ in η, we can calculate the likelihood as the probability, that an individual evolved as observed in the tree

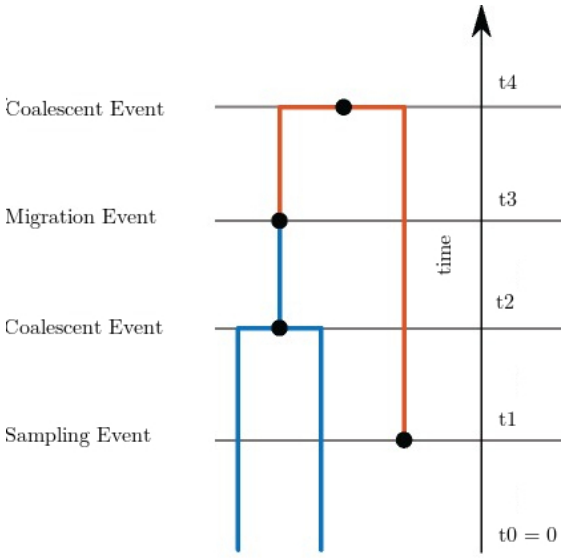


Figure 1.2: Events on an exemplary SC tree. Borrowed with permission from N.Müller.

$\mathcal{T}$ [26]:

$$P(\mathcal{T}|\eta) = (m_{dd'})^{v^m_{dd'}} (c_d)^{v^c_d} \cdot \exp\left[-\sum_{i=0}^{f} \tau_i \sum_{\substack{d=1 \\ d \neq d'}}^{2} \left(\binom{k_{i,d}}{2} c_d + k_{i,d} m_{dd'} \right) \right] \tag{1.2}$$

where $k_{i,d}$ is the number of lineages in location d during interval τ_i, $v^m_{dd'}$ is the total number of migration events from d to d' and v^c_d the total number of coalescent events in deme d.

1.1.3 A first look at differing model assumptions

From their brief introductions, we can already appreciate that the SC and the BD model differ in the way they estimate epidemiological parameters. In this last section, we want to summarise the different assumptions explained earlier for the reader's easier reference:

Table 1.1: Contrasting premises of the SC and the BD model

	SC	BD
N_i	constant	varies
$n\_samples_i$	$\ll N_i$	no assumption
sampling	no assumption	needs a model

where N_i is the total number of infected in location i.

1.2 Bayesian Inference

Bayesian inference is based on what is called "Bayes' Theorem" after Reverend Thomas Bayes (1763) and basically is not much more than today's product rule of probability theory [28] for two events A, B:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{1.3}$$

The interesting point is not so much the formula itself, but its conjunction with an Bayesian interpretation of probabilities. In contrast to the "Frequentists" understanding, the probability of an event or a hypothesis expresses our certainty about it, not the frequency with which it occurs. Taken that way, Bayes' formula can be used to update our beliefs, as new data becomes available. To make that statement clearer, we change the symbols in Eq. 1.2 to D for (new) data, and σ, the parameters of our model:

$$P(\sigma|D) = \frac{P(D|\sigma)P(\sigma)}{P(D)} \tag{1.4}$$

$$= \frac{P(D|\sigma)P(\sigma)}{\sum_\sigma P(D|\sigma) \cdot P(\sigma)} \tag{1.5}$$

where

- $P(\sigma|D)$ is *the posterior*, the probability of our parameters, after including the new evidence
- $P(D|\sigma)$, *the likelihood*, is the probability of observing our data, given that our model and its parameters were true
- $P(\sigma)$ is *the prior*, our belief about the parameters, before we look at the data
- $P(D)$, *the marginal*, is the probability of our data under any parameterisation for discrete σ. In the continuous case, the sum is replaced by an integral.

The usage of Bayes' rule for complex models used to be considerably hindered, mainly because calculating the multidimensional sum or integral in the denominator of equation 1.5 was infeasible. This only changed with the introduction of Markov Chain Monte Carlo (MCMC) sampling methods [29]. Algorithms of this class are able to draw samples from any probability distribution $P(x)$ provided that the return value of a function $f(x)$ is proportional to it. Starting from a starting state x_0 the next state x_1 is proposed and based on its probability either accepted or rejected. If the number of samples drawn is large enough, its distribution is proven to converge to our desired target distribution.

1.3 Bayesian Inference of Phylodynamic models

The variation of the genetic background in a population is not only shaped by the amount of substitutions. If demographic fluctuations occur on the same time scale, then genetic sequences as well contain a footprint of the past population dynamics [30].

Let us illustrate this point with one example. Branch lengths in a timed tree will be longer for a larger number of substitutions separating two nodes. On the other hand, population size

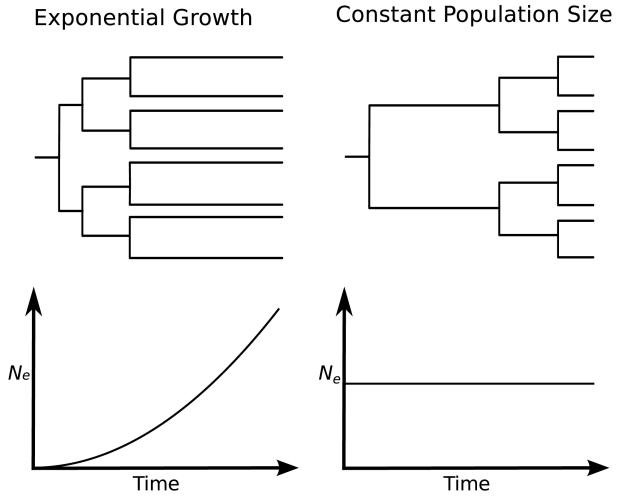


Figure 1.3: Effects of an exponentially growing and a constant population size on the tree topology. Plots show the effective population size N_e through time. Taken from [31].

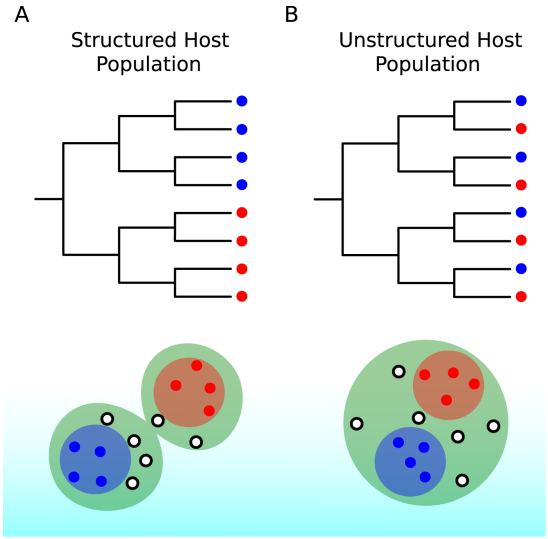


Figure 1.4: Effects of population structure on the tree's topology. Colors indicate different locations of the samples taken at present. Taken from [31].

changes influence the ratio of the external branch lengths (close to the tips) to the internal ones (close to the root) as depicted in Fig. 1.3. For the case of an exponentially growing population, the coalescence rate is $c = 1/N(t)/g$. That means for small population sizes we would expect a high c, hence a small waiting time until coalescence and vice versa. That's why branch lengths are increasing from the root towards the tips.

To disentangle how sequence evolution and population dynamics interact, we need to analyse them jointly. Bayesian phylogenetic inference [32, 16, 33] allows us to simultaneously estimate the DNA substitution parameters θ, the phylogenetic tree τ, of the sampled sequences D, together with the population dynamics parameters η:

$$P(\tau, \theta, \eta|D) = \frac{P(D|\tau, \theta)P(\tau|\eta)P(\eta)P(\theta)}{P(D)} \quad (1.6)$$

The first term, $P(D|\tau, \theta)$, called the phylogenetic likelihood, is not of concern for this thesis. It is sufficient to know, that we can calculate it using the pruning algorithm [34]. Of interest instead is the second term, the phylodynamic likelihood. For $P(\tau|\eta)$ we insert the likelihood of either the multitype birth-death model (compare 1.1.1) or the structured coalescent (Eq. 1.1.2). That means η contains the migration rates for both models. It further includes the birth, death and sampling rates for the BD and the population sizes in the SC model, respectively.

1.3.1 Influence of population structure on the tree topology

Population size variation as just discussed, is just one population dynamic factor that leaves its mark on the tree. Population structure, instead, shapes the topology in clustering leaves from the same location together as illustrated in Fig. 1.4. Early phylodynamic models assumed a homogeneous population mostly for computational reasons. But early theoretical work by [35] showed qualitatively different results for models that ignore structure. These finding were later

confirmed in a number of studies on sexual contact networks [36, 37, 38], leading to the development of the structured version of the coalescent and the birth-death models, as introduced in sections 1.1.2, 1.1.1. Although both models have found wide application in the community, a thorough comparison of them as performed in the unstructured case [39], is lacking. With this thesis we aim to contribute to a better understanding of their behaviour in different scenarios.

To investigate the performance of the structured coalescent and the multitype birth-death model, we designed a simulation study. In such cases, one simulates trees under each model. For each tree, the population dynamic parameters are then inferred by both the SC and the BD model. Knowing the "truth" we can then compare, how often each model has recovered the true parameters.

Chapter 2

Methods

2.1 Tree simulations

When comparing two methodologies, one needs to make sure that one compares indeed the methods and not the conditions under which they are tested. To do so, we set up the tree simulation scenarios with the following two questions in mind:

1. How can we best bring the conditions for the multitype birth-death model (BD) and the structured coalescent (SC) to match?

2. How can we fairly compare the BD and the SC for those cases, where the settings and assumptions cannot be brought to agreement?

To adress the first question, we mapped the parameters of the birth-death model, e.g. migration rates, to the coalescent's counterparts directly, whenever possible. Apart from the model parameterisations, this includes the tree's leaf times and types. That means, we always simulated a birth-death tree under a particular parameter set. Then, from its tip dates and locations, we simulated a coalescent tree using a corresponding set of parameters.

To answer the second question, we point out that the two models differ fundamentally in the way they treat sampling and their assumptions about population size growth. Let us consider the sampling problem first. As explained in the introduction (1), the birth-death model uses the sampling times as a source of information about the underlying population dynamic process, which the coalescent does not. To fairly compare the two models, we designed one simulation scenario, that would provide that additional information to the advantage of the birth-death model, here referred to as "BD with psi sampling", and another, "BD with deterministic sampling", that does not. In the former, individuals in deme i are sampled at random with rate ψ_i. Sampling events, therefore, contain information about ψ_i and the number of individuals in location i alive at the time. In the latter, individuals are collected at pre-defined time points. Hence, those sampling events provide no more information than a sufficient number of individuals having survived to be sampled.

To fairly address the assumptions on population size, we decided to simulate trees for a low and a high growth case. That means, for the first scenario, we use the constant population size SC and a BD model with an almost zero growth rate to generate the trees. For high growth, we use the same growth rates r_i for the SC and the BD model.

These considerations lead us us to six ways to simulate the trees, as summarized in Table (2.1). Notice, that the simulation schemes listed there contain links to their respective method descrip-

Table 2.1: Tree simulation scenarios

BD with psi sampling 2.1.1
BD with deterministic sampling 2.1.1
SC for psi sampling 2.1.2
SC for deterministic sampling 2.1.2
SCexp for psi sampling 2.1.2
SCexp for deterministic sampling 2.1.2

tions.

We aimed to thoroughly evaluate the two models under a representative set of parameter combinations that can be found in Tab. (2.2). The table's columns indicate the parameters that we vary. The case number is used to refer to the particular parameter combination. Let us look at the parameters we have chosen. As discussed in (1 and 2.1, the population size is set to either constant or exponentially growing. Similarly, we model sampling to be either low or high. The former corresponds to a sampling proportion of approximately 1% of the population. The latter depends on the case and will be discussed in the following sections. Lastly, we test for migration being differently strong. In the high migration case, an individual migrates on average once in its life time. In contrast, in the medium and low migration scenarios, we expect an individual to migrate only with a probability of 10% or 1%, respectively. The interested reader can find the exact parameters in Tab. (A.1).

For each scenario and case, we simulate 100 trees each on 100 leaves.

2.1.1 Birth-death tree simulations

Birth-death trees are simulated starting with one individual in location 0. Then, events are sampled according to the rates specified in section (1.1.1) and both the full tree and the sampled tree are kept.

Table 2.2: Parameters for the tree simulations. The columns indicate the parameters varied. The rows name the parameter combination.

Cases	growth	sampling	migration
1	low	low	high
2	low	low	medium
3	low	low	low
4	high	low	high
5	high	low	medium
6	high	low	low
7	low	high	high
8	low	high	medium
9	low	high	low
10	high	high	high
11	high	high	medium
12	high	high	low

Apart from the migration rates, the growth rate in deme i, $r_i = \lambda_i - \delta_i$ is the only parameter we can compare. Therefore we fix the death rate in each location to 1.0 and only vary the birth rates in the BD model.

BD with psi sampling

For each case, we used MASTER [40] to simulate 100 trees for the respective birth and migration rates. It creates a trajectory of events represented in the tree and is stopped, when the total population reaches a maximum value N_f. Then, extant lineages are trimmed to leaves at present time. Afterwards, 50 samples are drawn from all leaves of each type in the tree. These can be either leaves at present or leaves that result from a death event some time in the past.

The parameter N_i^f, the number of leaves at present in deme i, influences the sampling pro-

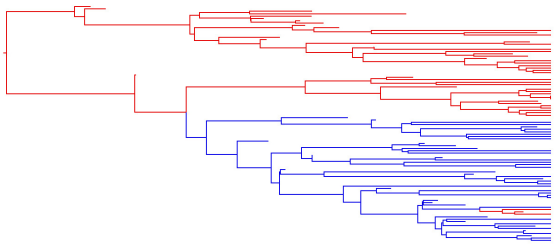


Figure 2.1: A phylogenetic tree generated by the BD with psi sampling. Notice that samples are taken through time. Lineages of type 0 and 1 are coloured in red and blue.

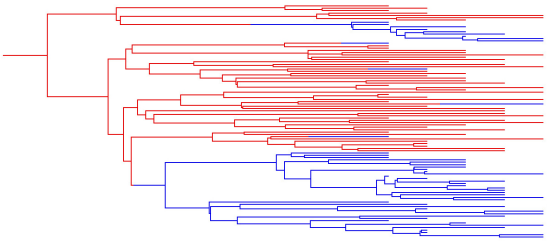


Figure 2.2: A phylogenetic tree as simulated by the BD with deterministic sampling. Individuals are collected at five different time points. Lineages belonging to type 0 or 1 are marked in red and blue.

portion ψ_i, which we can retrieve from the tree as follows:

$$\psi_i = \frac{s_i}{l_i} \tag{2.1}$$

$$= \frac{s_i}{l_i^e + N_i^f} \tag{2.2}$$

where s_i denotes the number of samples and l_i the total number of leaves in deme i. The total tip count can be divided into the number of extinct leaves l_i^e and present leaves N_i^f.

When we simulate until a larger N_f is reached, then we generate, trivially, a larger number of extant lineages. But we also end up with more extinct leaves, because as more time passes, more death events occur. Hence, we reach a smaller sampling proportion for larger N_f.

For the low sampling cases, we choose N_f such, that the sampling proportion is approximately 0.01. This leads to different final population sizes for the high and low growth cases. That's because we have, on average, twice as many birth events per death event in the high compared to the low growth scenario. That means, for an exponentially growing population, we reach the same final population size with fewer deaths. Therefore, we need to set N_f higher to achieve about the same number of leaves and hence a similar sampling proportion.

The high sampling case should be recovered if we simulate up until a smaller final population size (compare Eq. 2.1). On the other hand, N_f should not be too small, as we can end up with too few leaves in our tree to draw 50 samples from. That's why we adjust the simulation end condition, s.t. we have control over the final population sizes N_i^f in each deme i . Both these we set to a certain value F. We choose F to be as small as possible to still retrieve 50 samples and record the minimum, medium and maximum sampling proportion, that result for this setup in Tab A.2. To examine the final population sizes used for each case, consult Tab. A.3.

BD with deterministic sampling

Trees were simulated with our newly implemented simulator that will be introduced in section 2.2 . As before, the simulation starts with one individual in location 0. We draw 10 samples from each deme at 5 predefined time points and stop simulating after the final sampling event. This leaves us with a tree on 100 sampled tips. The exact sampling times can be found in Tab A.4.

2.1.2 Coalescent tree simulations

As we continue to the simulation of the SC trees, we will start discussing how to match them to their BD counterparts. First, we set the leave times and types in the structured coalescent simulation to the same values as in the corresponding birth-death tree. That means, for the same sampling scenario, parameter case and simulation seed, the BD and SC trees should have the same leave types and dates. Secondly, we have to match the population size growth assumptions of the two models. As mentioned in section (2.1), we compare trees simulated under the constant population size SC to the BD trees from the low growth simulation, and those generated under the exponential growth scenario with the high growth case. Further, we map the population size in the coalescent to the present day population (before sampling) $I_i(0)$ and birth rates λ_i attained in the birth death simulation. To do so, we follow [41] and calculate the scaled population size θ_i:

$$\theta_i = \frac{I_i(0)}{2\lambda_i} \tag{2.3}$$

Lastly, we have to bring the backward in time migration rates and their forward in time analogues to agreement. As that mapping depends on the population size assumption, we shall treat it separately in the following subsections.

Generally, a SC simulation starts with the samples taken closest to the present and merges them together, going backwards in time. Events are sampled according to the rates as defined in the model (1.1.2). The simulation stops, once only one individual is left.

For each scenario, we independently simulate 100 trees each on 100 tips.

SC conditioned on psi or deterministic sampling

We simulate trees for the SC model with constant population size using MASTER [40]. The backward in time migration rates q_{ij} can be calculated as derived by Notohara [23]:

$$q_{ij} = \frac{m_{ji} \cdot N_j}{N_i} \tag{2.4}$$

From the point of view of the coalescent, the different sampling schemes do not make any difference to its parameterisation and hence we simulate them just in the same way.

SCexp conditioned on psi or deterministic sampling

To simulate trees from the structured coalescent for an exponentially growing population size, we implemented a new Gillespie algorithm building on the MultiTypeTree package [42]. We calculate the scaled population sizes and the backward migration rates as done for the constant population size case. Then we simulate 100 trees.

2.2 Tree simulators

For most simulations, we could use MASTER [40] to generate the trees we wanted to. Nevertheless, for the BD with deterministic sampling and the structured coalescent with exponential growth, we had the pleasure to implement new simulators.

2.2.1 Simulator for the BD with deterministic sampling

To create trees, that follow the dynamics of a birth-death model with sampling occuring at deterministic time points, we implemented a simulator based on Gillespie's direct method [43]. We

validated it by simulating trees up until the first sampling event and compared the distribution of some tree summary statistics to trees generated by MASTER.

2.2.2 Simulator for the SC with exponential growth

We used Gillespie's first reaction method [43] to implement a simulator for structured coalescent trees under an exponentially growing population.

For equal growth rates $r_i = r_j$ in locations i, j we can map the backward migration rates q_{ij} to the forward in time migration rates m_{ji} as follows:

$$q_{ij} = \frac{m_{ji} \cdot N_j(t)}{N_i(t)} \tag{2.5}$$

$$= \frac{m_{ji} \cdot \theta_j}{\theta_i} \frac{e^{-r_j \cdot t}}{e^{-r_i \cdot t}} \tag{2.6}$$

$$= \frac{m_{ji} \cdot \theta_j}{\theta_i} \tag{2.7}$$

To do so, we assume 1) an exponentially shrinking population from the scaled present day population size θ_i of type i. The second arrangement follows directly from the equality of the growth rates.

Results for $r_i \neq r_j$ have been derived as well and are not much more complex. They just demand a time transformation to sample an exponential waiting time for the time-varying migration rates. Since the results will not be relevant for this thesis, we will not include the derivation here.

2.3 Parameter Inference

We used BEAST2 [44] to sample from the posterior distribution of the parameters of interest from equation 1.6. We provide it with the true tree because we are only interested in the differences arising from our model choice (BD or SC), not yet in additional complications that might arise from reconstructing the tree as well. All chains were run for 10^8 steps and 10% of burn-in was discarded. Additionally, we verified that the effective sample size was higher than 200.

2.3.1 Inference under the birth-death model

We used Beast's BDMM package [20] to infer tip-typed trees. That means we integrate over the type-change processes along the branches and do not estimate the internal node locations.

BD inference of psiSamp trees

For trees simulated under psi sampling conditions, we fix BDMM's sampling through time proportion ψ and the sampling proportion at present ρ to the true value which we calculate from the tree using Eq. 2.1.1. The priors on all parameters can be found in Tab. A.6.

BD inference of determSamp trees

Sampling at deterministic time points violates the BD model's assumption of a sampling rate through time as discussed earlier. Therefore we allow changes in the sampling rate to accommodate the presence of samples at very specific time points. More specifically, we divide the time spanned by the tree into 10 time intervals, 5 of which contain the sampling events. Only for these regions, we allow the BD to estimate the sampling rate. The sampling rate in the remaining 5 intervals without samples, we fix to 0. Nevertheless, we still have to inform 10 more parameters this way, that is 5 sampling rates for each deme. The priors we set on the parameters can be found in Tab. A.5.

2.3.2 Inference under the structured coalescent

We used the MultiTypeTree package [42] to infer the effective population sizes and backward in time migration rates for the SC model. The priors are included in A.7.

Chapter 3

Results

We set out to compare the structured coalescent (SC) and the multitype birth-death model (BD) under a representative set of scenarios. First, we will investigate how well the migration rates are estimated, since they are most directly comparable. We shall start with a parameterisation, that does not violate the assumptions of any of the models. Continuing, we include factors, that the models do not account for and evaluate their influence.

3.1 Comparing the BD and the SC on psi sampled trees

Low growth and low sampling Hence, let us first investigate the low growth and sampling case for varying migration with its summary statistics depicted in Fig. 3.1. For simulated BD (BDpsi) and SC (SCpsi) trees we evaluate the BD and SC inference with respect to the recovery frequency and the root median squared error (RMSE). A model recovers the true parameter, if it falls into the highest posterior density (HPD) interval the model estimates. The RMSE measures the accuracy, by taking the square root of the median of the residuals between the true value and the inferred median.

As expected, the BD performs very well, with recoveries always higher than 95%, on the trees simulated under the BD. Similarly, the SC model recovers the true migration rates better for the SCpsi than for the BDpsi trees. Even so, it clearly struggles to recover the true rates in the low migration scenario, for which it is only correct 50 or 60% of the time. Further, the BD model's recovery frequency for SC trees is higher than 90% for 4 out the 6 migration rates. In contrast, the SC only surpasses 90% for 2 rate on the BD trees. When inferring from the SCpsi trees both models perform worse for decreasing migration. When we compare the models based on their RMSE, we find that the SC's estimated median migration rate is mostly closer to the true value. The BD model can only catch up in accuracy for the low migration cases. In the high migration scenario, the model's challenge probably is to determine how high the rate is. In our simulation, the high migration rate is equal to the birth and death rate; an individual migrates once in its lifetime in expectation. Hence, there is almost no structure present and higher migration rates could also be supported by the data. Before going into greater detail, let us first examine how the picture changes for higher sampling.

Low growth and high sampling As we increase the sampling proportion, we notice some interesting trends for the RMSE. The birth-death model improves upon it for high and medium migration both on the BDpsi and SCpsi trees compared to low sampling. Therefore it only deviates to the performance of the SC in the high mi-

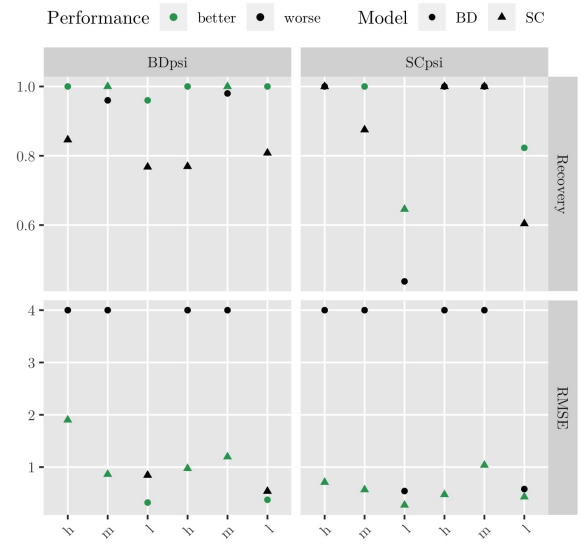


Figure 3.1: **Summary statistics for low growth and low sampling** when inferring the migration rates. We compare trees simulated under the BD with psi sampling (BDpsi) and the SC for psi sampling (SCpsi). Migration rates were set to high, "h", medium, "m" or low, "l". Results are evaluated on the recovery frequency and the RMSE. Values for BD and SC inference are labeled as points or triangles. The model performing better on the particular metric is highlighted in green.

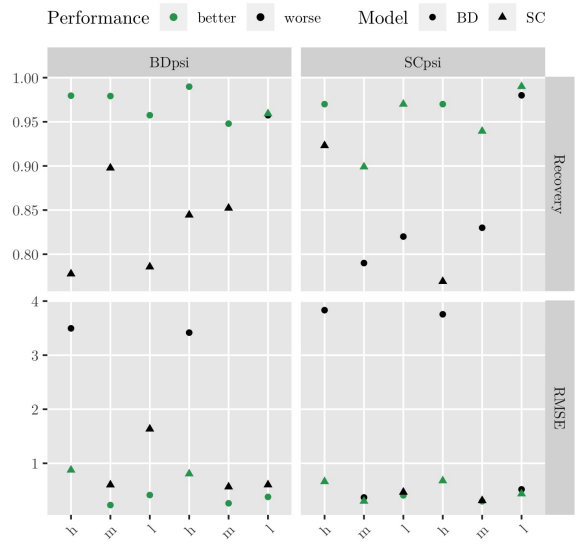


Figure 3.2: **Summary statistics for low growth and high sampling** when inferring the migration rates. See caption of Fig. 3.1. for detailed description.

gration scenario. This development comes with only minor loss in recovery for the BDpsi trees. In contrast, it misses the true value more often in the SCpsi trees for medium migration, while improving on them for low migration. The SC can increase the recovery for low migration in all trees, while it worsens or remains the same for other migration cases. Further, it keeps its good performance with regard to the RMSE for both trees.

To explain these developments, remember that higher sampling corresponds to shorter trees. That mainly results from our simulation setup, where we reach a higher sampling proportion upon stopping at a lower number of individuals (see Methods). It follows, that our samples are not that widely spread in time, which could increase the resolution we have on our migration rates.

To further examine these initial findings, we visualised all 100 inferences by laying the highest posterior density (HPD) intervals inferred by the BD and SC on top of each other in Fig. 3.5 for migration rate mr_{01} and in Fig. A.2.1 for m10. Since we cannot treat all parameter cases, we chose two interesting examples for the low migration and one for the medium migration scenario.

Let us start with the low sampling and low migration setup in the first row of Fig. 3.5. We can observe that the structured coalescent overestimates the migration rate mr_{01} from location 0 to 1 for the BDpsi tree, which leads to its bad recov-

ery in Fig. 3.1. Interestingly, it can recover the truth better for the migration rate in the opposite direction (first row in Fig. A.2.1). The birth-death model, on the other hand, recovers both migration rates from the BDpsi trees equally well but estimates wider posterior intervals for mr_{01}. Nevertheless, for the SCpsi trees it fails to include the true mr_{01} for its narrow HPDs. When we increase sampling and keep migration low (bottom row), we see a similar pattern for the BDpsi trees. On the SCpsi trees, on the contrary, both methods show better recovery. For high sampling and medium migration, both methods perform well on the SCpsi trees for migration in each direction. The SC still seems to overestimate mr_{01} from the BDpsi trees, but the pattern is less pronounced than for low migration.

The are several factors, that could explain the SC consistently overestimating the migration rate mr_{01} in the low migration regimes. First, since we basically condition on at least one migration event from 0 to 1 in our tree simulations, this might already bias the set of trees we get. This bias would be stronger, if there are generally few migration events, s.t. an additional one has a larger impact. This line of reasoning would be consistent with the bias increasing for lower migration and for higher sampling. The BD model estimates a median that is higher than the truth for migration rate mr_{01} in these cases. But it can often still recover the true value because of its wide HPD intervals, that might result from the BD taking into account the stochasticity of the process (compare A.2.1).

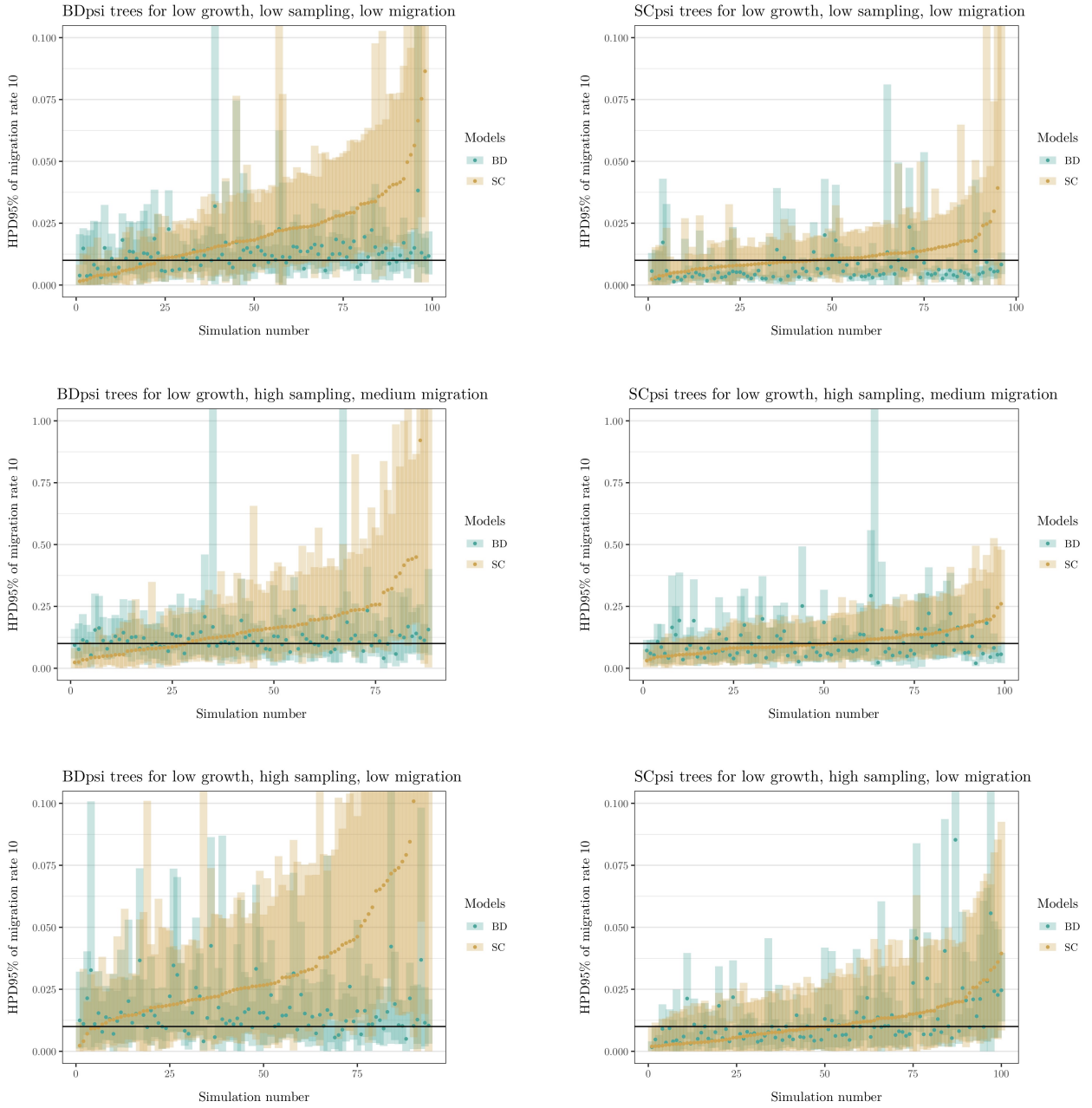


Figure 3.5: For every plot we simulated 100 trees under the BD with psi sampling (BDpsi) or the SC conditioned on psi sampled leaves (SCpsi) and analysed them using the BD (turquoise) and the SC (ochre) model. We show the 95% highest posterior density (HPD) intervals of the estimated migration rate mr_{01} as shaded segments and their medians as dots. The HPD regions are ordered with increasing median of the SC model's estimate. The true migration rate is displayed as a black, horizontal line. The figures' individual captions indicate the growth, sampling and migration parameterisation.

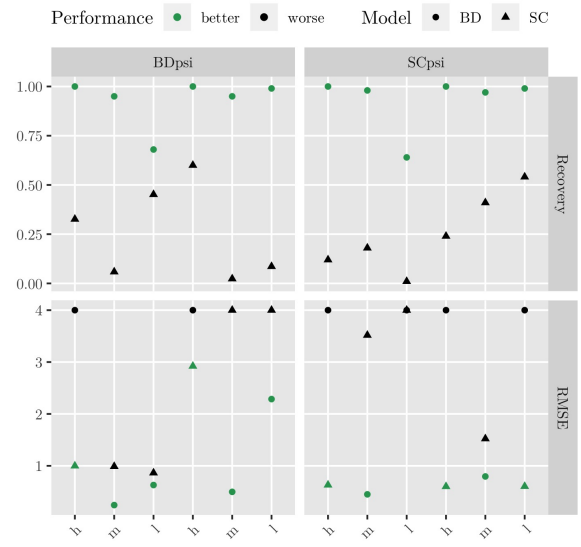


Figure 3.6: **Summary statistics for high growth and low sampling** when inferring the migration rates. See caption of Fig. 3.1. for detailed description.

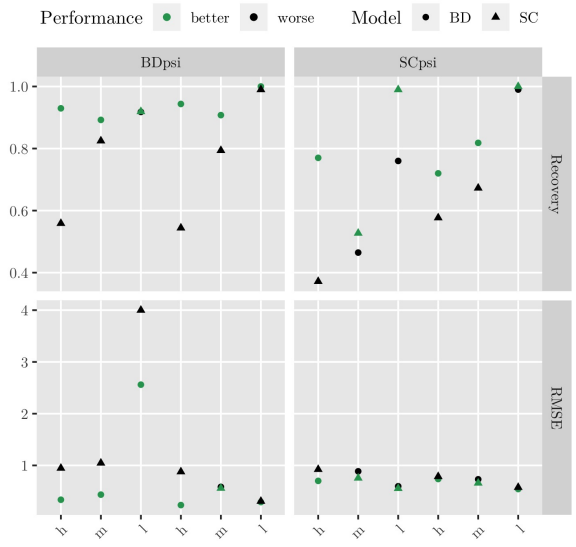


Figure 3.7: **Summary statistics for high growth and high sampling** when inferring the migration rates. See caption of Fig. 3.1. for detailed description.

High growth and low sampling We shall now depart from trees simulated under constant population sizes and move on to the exponential growth case for low sampling as summarised in Fig. 3.6. The BD model show recoveries above 95% for all trees and all cases but the migration rate mr_{01} in the low migration regime. The SC looses significantly in recovery frequency, recuperating the true migration rates at a maximum of 52%. Interestingly, it's median estimates of migration rate mr_{01} in the BDpsi trees and $mr10$ in the SCpsi trees are comparably close to the truth with an RMSE of approx. 1. The BD model achieves RMSEs smaller than 1 for medium migration and does better for low than for high migration. The drop in recovery for the SC is expected, as the distribution over the branching times for a constant population deviates substantially from an exponentially growing one. It is curious, why for each tree simulation variant one inferred migration rate is relative close to the truth. When comparing the BD model's RMSE values to Fig. 3.2, we notice a similar pattern: lower RMSE for low and medium migration. Since both high sampling and high growth lead to shorter trees in our setup, this as well means, that the time between samples should on average be shorter. Shorter times between leaves make longer sequences of events more unlikely and could help to make more precise estimates of the migration rate.

High growth and high sampling Upon increasing the sampling proportion, the SC improves significantly in the recovery of migration rates from the BDpsi trees in the medium (from almost 0% to approx. 80%) and low (from $< 50\%$ to $\geq 90\%$) migration scenario. For both models recovering the true parameter from the SCpsi trees is rather mixed. For the inference under the SC there is a general trend upwards for decreasing migration. The BD model drops a bit

in recovery from the BDpsi trees, but remains at approx. 90%. The RMSE values attained by both models, apart from one outlier, reach their minimum across the different parameter cases we tested so far. At the same time, these trees are the shortest compared to the trees in other simulation schemes. This observation is further evidence for a link between the tree height and/or sampling times and better resolution for the migration rates.

We again plot the inference results for some example cases, to better understand the performance of each model in Fig 3.10. The BD model behaves as we would expect; it reliably recovers the migration rates from the BDpsi trees. For the SCpsi trees it misses the true value more often, but can still capture the general trend. In contrast, the SC completely fails to recover the true migration rates for low sampling, which is very apparent for high migration in Fig. A.2.1 and medium migration as shown in the first row of Fig. 3.10. It always infers one median migration rate to be approx. 0 and the other to be much higher than the truth. The reason for its improved recovery and RMSE for decreasing migration (see medium row in Fig. 3.10 and Tab 3.6) is that the true migration rate approaches zero as well. We will further investigate the reasons for this pattern in the following section. On the other hand, for high sampling we observe this "diverging pattern", one migration rate close to 0 the other much larger, only for high migration (Fig. A.2.1). For decreasing migration the SC is again able to recover the true migration rate, sometimes even outperforming the BD model (compare bottom row in Fig. 3.10).

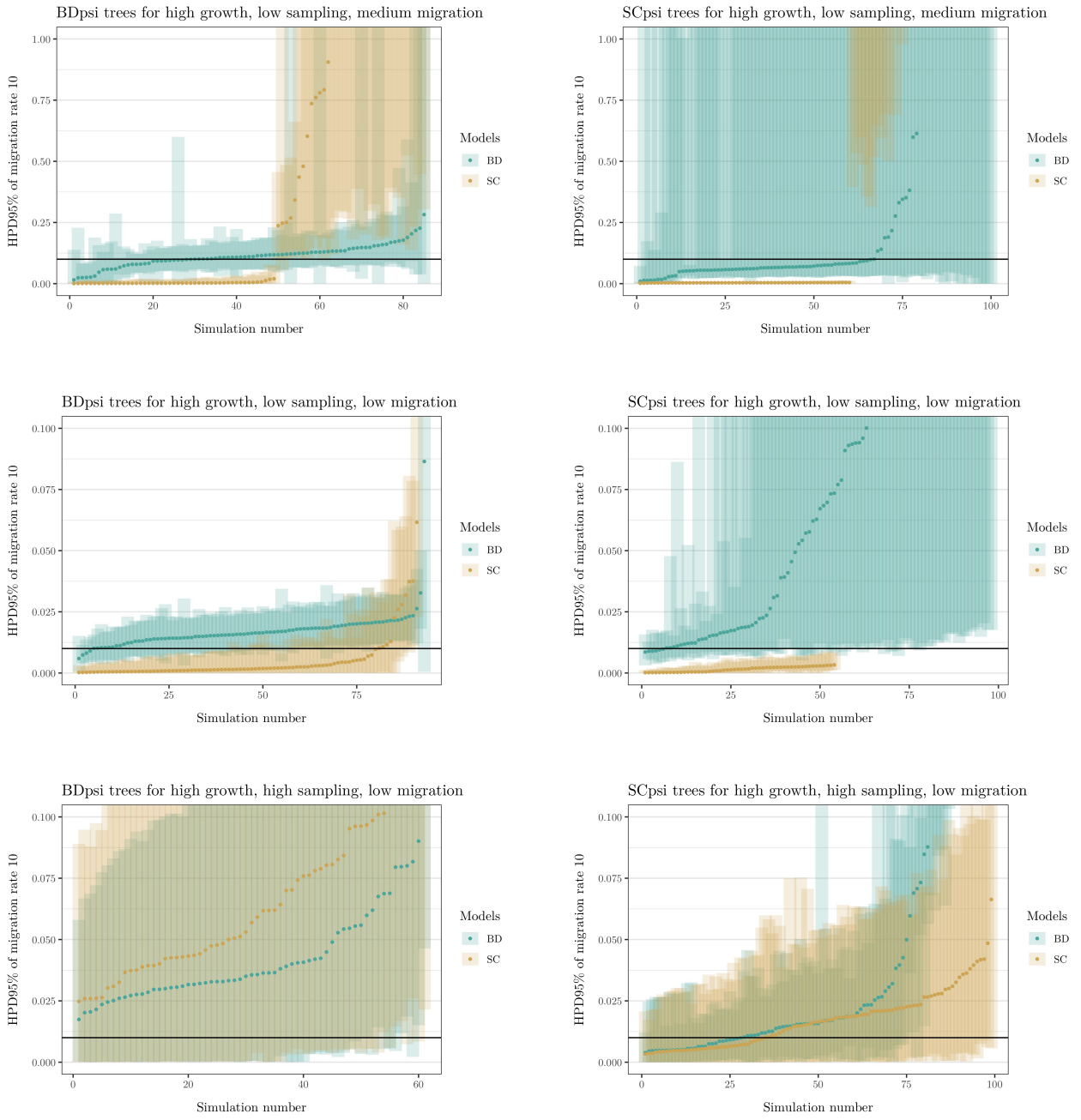


Figure 3.10: For every plot we simulated 100 trees under the BD with psi sampling (BDpsi) or the SC conditioned on psi sampled leaves (SCpsi) and analysed them using the BD (turquoise) and the SC (ochre) model. We show the 95% highest posterior density (HPD) intervals of the estimated migration rate mr_{01} as shaded segments and their medians as dots. The HPD regions are ordered with increasing median. The true migration rate is displayed as a black, horizontal line. The figures' individual captions indicate the growth, sampling and migration parameterisation.

3.2 How the SC interprets population growth

To gain more insights into how the SC recovers the migration rates under high growth conditions, we will examine the combination of parameters it infers.

We plot the estimated population sizes and migration rates next to each other in Fig. 3.13, all ordered by increasing median of population in deme 0. This way we can see what quadruples of parameters are inferred for each tree. For all migration cases the SC estimates one large population of approx 5000, and a small population with 40 individuals. Interestingly the forward migration rate from the small into the large population is always inferred to be almost zero. Migration in the opposite direction is heavily overestimated. Further, the bias increases in order of magnitude for decreasing migration (see scaling of migration rates in 3.13.

How do these paradoxical patterns arise? The main problem the structured coalescent has to face in this situations is how to estimate the population sizes if the branching times close to the root and close to the tips provide conflicting information. That means, close to the tips the comparably slower coalescent events suggest a large population size meanwhile further backwards in time shorter branches show evidence of a small population. The structured coalescent seems to solve the problem by matching the small population i with a large migration rate q_{ij} backward in time (hence small m_{ji} forward). Therefore these lineages coalesce quickly close to the present and can account for coalescent events there. The large migration rate then moves almost all individuals into deme j that has a large enough size, to account for the rest of the tree. For higher sampling the SC can better estimate the true migration rates and the population sizes (see Fig.A.2.1). Since we stop the simulation earlier to reach a higher sampling proportion, it follows that the resulting populations have grown less than their low sampling counterparts. In fact, for low sampling the populations increased by a factor of $500 - 1000$ during the course of the tree simulation, whereas for high sampling they multiplied only by factor 10. Hence, the SC's assumption of a constant population size is not that severely violated in the high sampling case, which could explain the improved performance. We generally observe similar patterns when inferring from the BD trees (compare A.11).

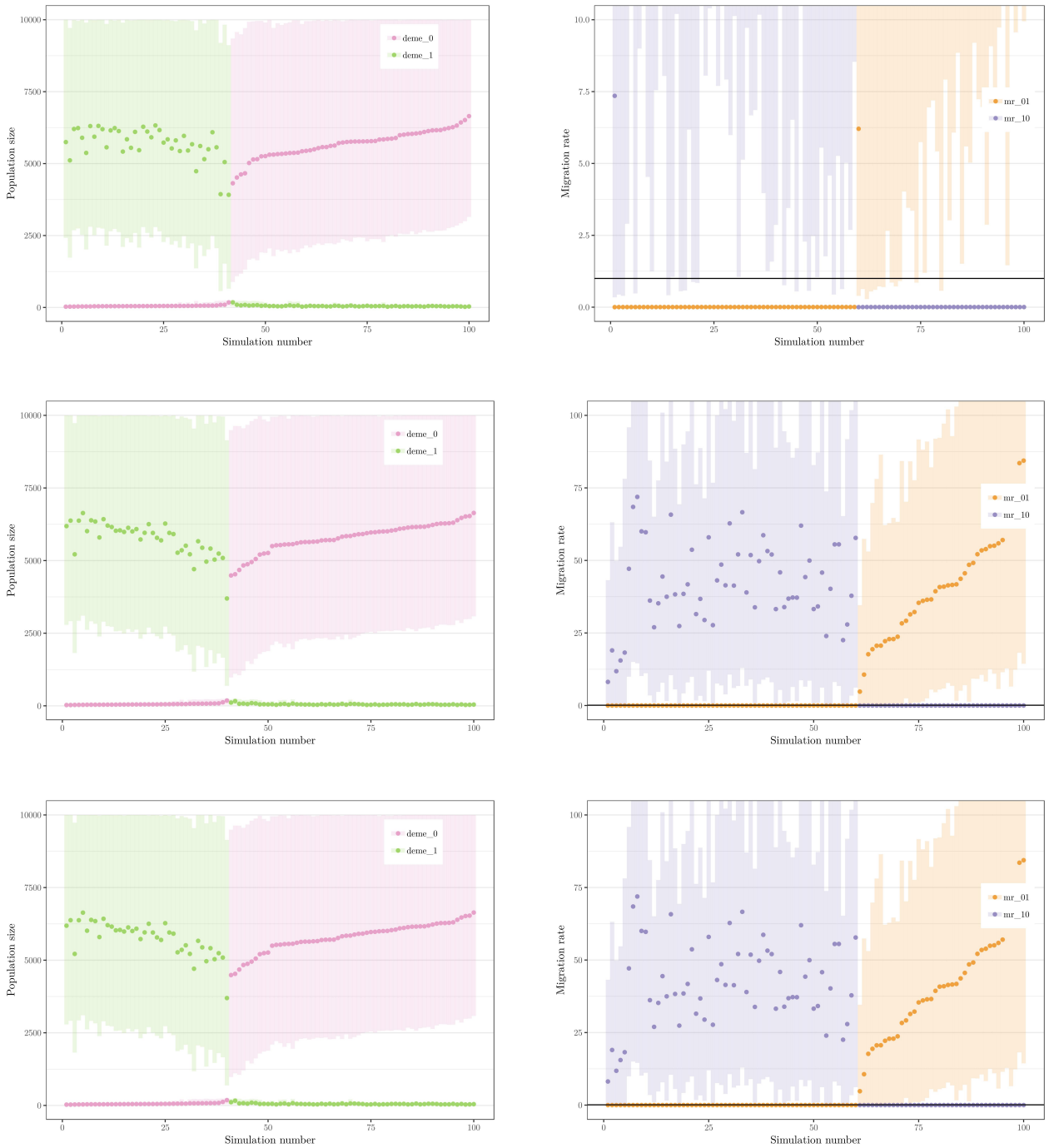


Figure 3.13: **SC interpretation of high growth and low sampling from SCpsi trees.** For each tree the inferred median population sizes in the left and migration rates in the right plot are shown in increasing order of population size in deme 0. The black line indicates the true migration rate parameter. Migration decreases the from "high' over "medium' to "low" from row one to three.

3.3 Comparing the BD and the SC on trees with deterministic sampling times

One advantage the BD model has compared to the structured coalescent is that it exploits the sampling times as a source of information about the underlying process (1). We shall now evaluate its performance if the sampling times do not provide such evidence as they are set a priori.

Low growth and low sampling We shall start examining the summary statistics for the inference of the migration rates, as depicted in Fig. 3.14, for the low growth and low sampling scenario. The BD model's recovery is 0 in all high migration cases, and improves as migration decreases. The structured coalescent recovers the migration rates with frequencies $\geq 80\%$ in most cases. Interestingly, it seems that the BD recovers the migration rates better from the SCdet trees and vice versa. Both methods perform similar on the RMSE for BDdet trees apart from the BD failing at the high migration rate 01. For the SCdet trees, the SC always estimates median values closer to the true value than the BD model.

That the BD model performs worse compared to the psi sampling case is expected, since the sampling times conflict with its assumptions. Further, the inference has to inform 10 more parameters than the SC, since we estimate 5 different sampling proportions for each deme. Interestingly, for low migration the BD performs better on the BDdet and SCdet trees compared to the respective BDpsi and SCpsi variants (compare Fig 3.1). As the structured coalescent only conditions on the sampling times, them being fixed should not influence its performance. Comparing to the inference on the psi trees, the recoveries are generally similar.

If one did not know which of the four simulation schemes for low sampling and low growth generated the tree and wanted to achieve the lowest possible RMSE, one should choose the SC. For highest possible recovery, the SC and the BD model seem to outperform each other almost the same number of times, 10 and 9 respectively. Nevertheless, as the SC exceeds the BD's recovery by larger margins, based on our data, the SC should as well be preferred.

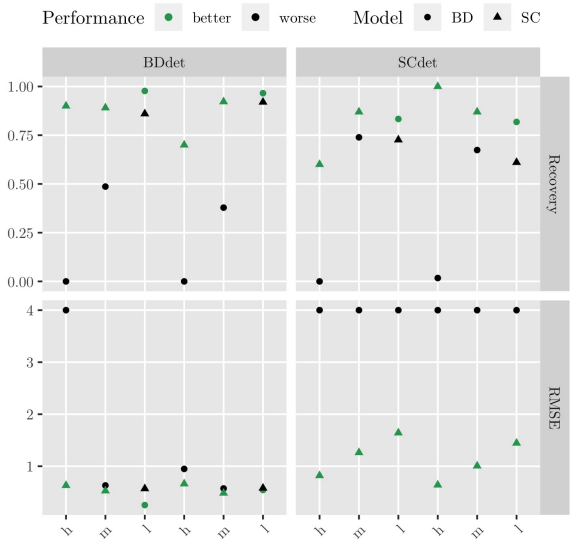


Figure 3.14: **Summary statistics for low growth and low sampling** when inferring the migration rates. We compare trees simulated under the BD with deterministic sampling (BDdet) and the SC for deterministic sampling (SCdet). Migration rates were set to high, "h", medium, "m" or low, "l". Results are evaluated on the recovery frequency and the RMSE. Values for BD and SC inference are labeled as points or triangles. The model performing better on the particular metric is highlighted in green.

Low growth and high sampling For the low growth and high sampling case, we plot the summary statistics for inferring the migration rates using either the SC or the BD model for varying migration intensity in Fig. 3.15. The BD model recovers the migrations rates best ($\geq 90\%$) for

medium and low migration from the BDdet trees and comparably well ($\geq$ 73%) from the SCdet trees. For high migration, it generally performs bad with less than 50% recovery. The SC works best ($\geq$ 89%) in the medium and low migration case for BDdet trees and for high and medium migration on SCdet trees with $\geq$ 78%. For most cases, both models perform similar ($\leq$ 1) with regard to the RMSE. Only the BD model reaches higher values for migration rate 10 for the high migration case on both tree types.

With respect to both recovery and RMSE the SC performed better on the BDdet than the BDpsi trees and the BD better on the SCdet than the SCpsi trees. This is unexpected, since the psi sampling scheme allows the BD to use information from the sampling times. It should be noted, that the tree height for the BDdet trees in this case is 25 (in arbitrary units of time) in contrast to 68 in expectation for the BDpsi trees. Therefore, we rather suspect that the better performance of BD model on the "det" trees results from their shorter heights.

The model that reached lower RMSE values across the different tree simulation schemes and migration cases most often is the structured coalescent. Nevertheless, the BD model recovers the migration rates more frequently.

We visualise some exemplary results as before by overlaying the inferred HPD intervals and medians of migration rate 01 for both models in Fig. 3.18. Interestingly, for medium migration and low sampling, the BD model overestimates the migration rate for about 50% of the BDdet trees and 25% of the SCdet trees by a factor greater than 10. The bias vanishes for the BDdet trees in the low migration scenario, while for the SCdet scheme the migration rate is still strongly overestimated for approx. 10% of trees. In contrast, the SC model's estimates are in reasonable proximity to the true value. For the remaining remaining scenarios, the SC and the BD model perform similarly well.

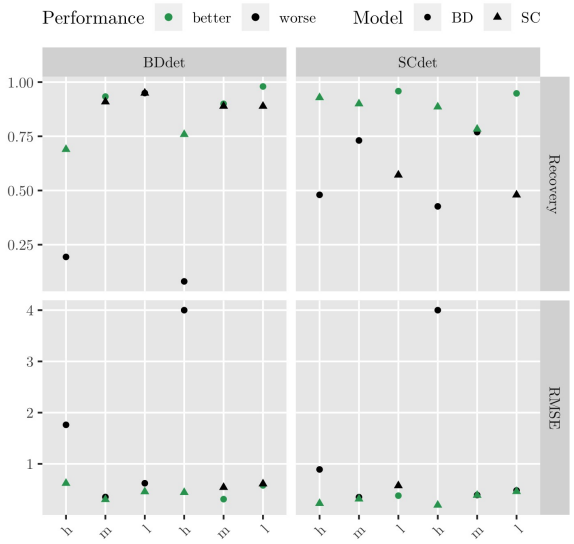


Figure 3.15: **Summary statistics for low growth and higher sampling.** For detailed description refer to caption of Fig. 3.14

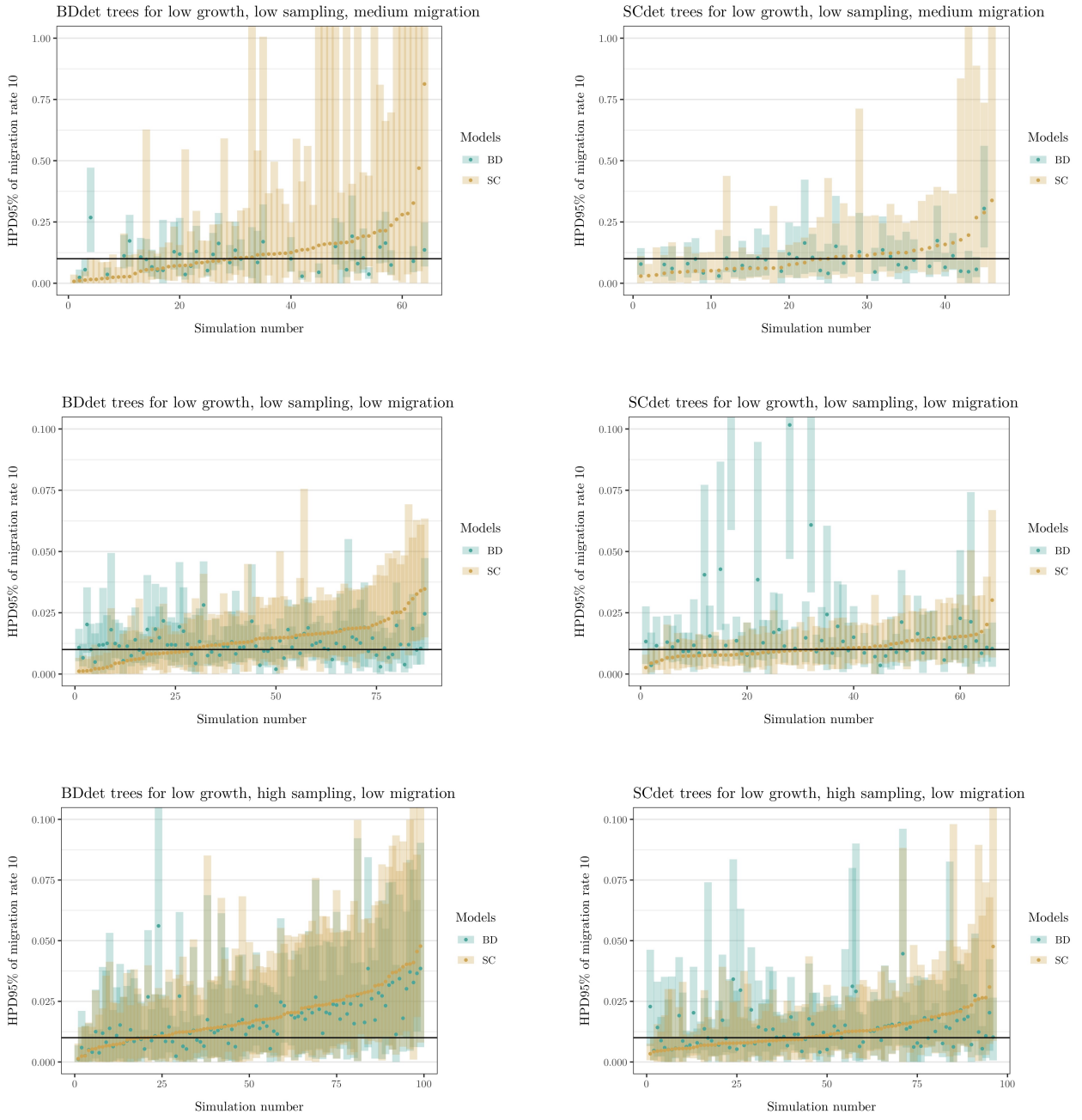


Figure 3.18: For every plot we simulated 100 trees under the BD with deterministic sampling (BDdet) or the SC conditioned on deterministically sampled leaves (SCdet) and analysed them using the BD (turquoise) and the SC (ochre) model. We show the 95% highest posterior density (HPD) intervals of the estimated migration rate mr_{01} as shaded segments and their medians as dots. The HPD regions are ordered with increasing median. The true migration rate is displayed as a black, horizontal line. The figures' individual captions indicate the growth, sampling and migration parameterisation.

High growth and low sampling For high growth and BDdet trees, the BD model estimates the parameters better than in the low growth case, with recoveries $\geq 75\%$ for medium and low compared to only 25% in the high migration case. On the same trees, the SC only exceeds 50% recovery for low migration. For the SCdet trees the picture is rather mixed. The SC never surpasses 50% and generally performs better for migration rate 10. The BD exceeds 50% twice and remains the rest of the time below 25%. With regard to the RMSE both methods perform comparably well on the BDdet trees. For the SCdet trees, both models seem to struggle with precise estimates of migration rate 01 - the BD for high and medium, the SC for medium and low migration. From these findings, we do not obtain clear support for our hypothesis that shorter branch lengths will lead to more precise estimates of the migration rates. The comparably worse performance of the SC can be further explained, as the high growth trees do not agree with its assumption of a constant population size.

Compared to the results for the psi samples trees (compare Tab. 3.6), the BD model now indeed performs better on the psi sampled trees than on the deterministic trees. The tree heights differ very little, as the BDdet trees are 7 and the BDpsi trees in expectation 8.5 time units long.

Across the different tree simulations both models perform similarly on the RMSE, while the BD model recovers the parameters better in 20 of the 24 cases (6 different migration rates for each of the 4 tree simulation schemes).

High growth and high sampling When we increase sampling as shown in Fig. 3.3, the BD model recovers the migrations rates better in the high migration case for the BDdet trees and in all migration scenarios for the SCdet trees compared to low sampling. For low and medium migration its recovery exceeds 80% with one exception. The SC shows a recovery frequency $\geq 76\%$

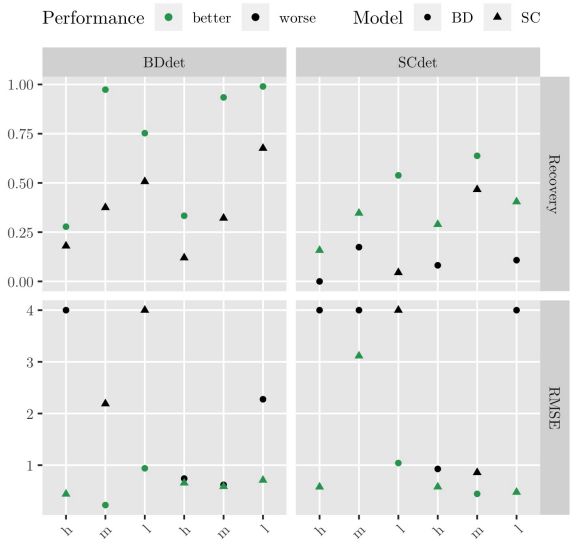


Figure 3.19: **Summary statistics for high growth and low sampling** For detailed description refer to caption of Fig. 3.14.

for all migration scenarios on the BDdet trees. Interestingly it performs worse on the SCdet trees, where its recovery is $\leq 50\%$ for all migration cases but the low one. Both models can improve on the RMSE, leading to values ≤ 1 in most cases. The improved performance of the SC is in line with our expectation, that the population size has not grown as much as in the low sampling case. Therefore, the violation against the constant population size assumption might be less severe.

Comparing to our previous investigation, the BD model has better recovery frequencies and RMSE on the BDpsi trees than on the BDdet trees as expected. The heights of trees are very similar with 3.7 for the BDdet and 3.7 in expectation for the BDpsi trees. As the BD model does not have an advantage on either of the tree sets with respect to height, these findings agree with our intuition, that the BD model should perform better on trees that do not violate its sampling

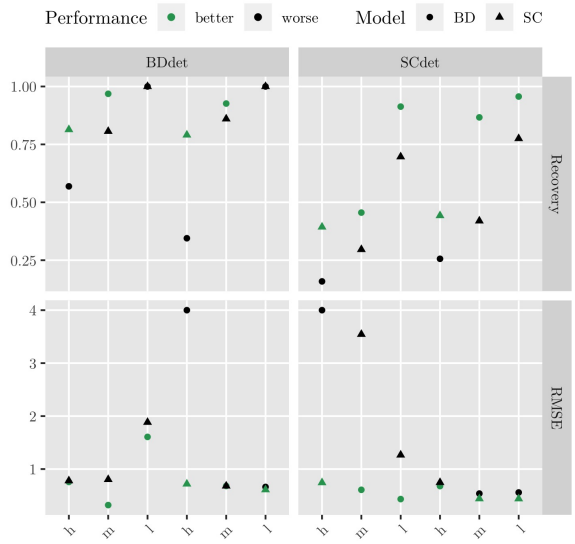


Figure 3.20: **Summary statistics for high growth and high sampling** For detailed description refer to caption of Fig. 3.14

.

assumptions. For the SC the pattern is not as obvious as it recovers e.g. the low migration rates better from the SCpsi trees, but the medium migration rates better from the SCdet trees. These variations might caused by the time between the samples.

Summarising, across all high growth and high sampling trees, the BD model recovers the migration rate better by margin in 12, the SC only in 6 out of 24 cases. As well with respect to the RMSE the BD model outperforms the SC by margin more often (9 times) than vice versa (twice).

Chapter 4
Discussion

In this simulation study we examined the performance of the structured coalescent (SC) and the birth-death model (BD) as they were tested on different growth, migration and sampling regimes.

A very clear finding of our study marks the failure of the SC to accurately estimate the migration rates (and the effective population sizes) for trees generated under exponentially growing populations. As the structured coalescent assumes the effective population sizes to remain constant, exponential growth is a strong violation of its assumption which explains the model's performance. Apart from the specific case of exponential growth, we would not suggest to use the SC, if any large change in the effective population sizes occur. For interesting data sets in practice, that carry traces of population sizes changes, e.g. epidemic outbreak data, we recommend to use the BD or current developments in the field: Mueller et al. [45] propose a method to infer time-varying migration rates and population sizes that are, additional to the sequence data, informed by predictors such as geographic distance. Furthermore, complex epidemiological models that provide more flexibility than just time-varying migration rates and population sizes have been developed [46].

Compared to the simulation study by Boskova et al. [39] for the models' unstructured equivalents, we do not find that the BD model recovers either migration rates or birth rates worse for the critical low growth case. Nevertheless, we simulated the trees until a final population size of 1000 was reached, in contrast to 100 in Boskova's simulations. Therefore our trees should on average be longer and could provide more birth events to inform the inference. One of Boskova et al.'s main findings was that the BD model infers HPD intervals for the growth rate that are generally wider than their SC counterparts. We cannot find such a general pattern for the migration rates based on our data. As the SC infers the effective population size and the BD model the birth rates, we cannot compare the HPD widths of these population dynamic parameter as directly as it was done for the growth rate.

We find, that the BD model can better account for early migration events in the tree. While the structured coalescent tends to overestimate the corresponding migration rate in these cases, the BD stays closer to the true value. This difference increases with decreasing migration. We hypothesise that, for the SC this final (backward time!) migration event occurs relatively fast. Under SC simulations, one would rather see long branches close to the root, where lineages of two different types need to wait until they are in the same deme to coalesce.

Furthermore, the BD model improves significantly in recovery frequency if the sampling times in the tree provide information about the under-

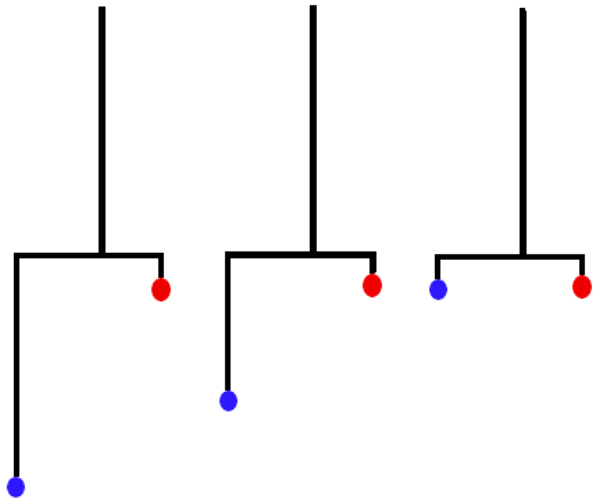


Figure 4.1: Toy example trees

lying process. Nevertheless, even if the sampling times are meaningful to the BD model, its estimates are very often less accurate than those of the SC. From the current data it is not obvious, why the SC is superior to the BD model in this respect.

Our analysis indicates that trees with on average shorter branch lengths allow estimating the migration rates more precisely. This is best illustrated by a little toy example. Imagine we wanted to estimate the migration rates from blue to red and vice versa from the trees in Fig. 4.1. The tree on the right would strongly support a high migration rate from at least on deme to the other. The tree on the left however, has one long enough branch, s.t. multiple migration events could have happened on it. Therefore it could equally well support very fast migration back and forth between the demes or a very small migration rate.

Trees with shorter average branch length can be generated by exponentially growing populations or more dense sampling through time. In practice, this means that collecting more samples could lead to better migration rate estimates.

Chapter 5
Outlook

The parameter space that has to be tested for a thorough comparison of the two models is large. Only for the work shown in this thesis 100 trees for each of the 48 parameter specifications (migration [3] x growth [2] x sampling [2] x sampling process [2] x simulation model [2]; to be read as migration with 3 parameters, growth with 2 ...) had to be simulated and analysed (x models [2]). Nevertheless, a number of interesting questions remains open. Firstly, a major difference of the SC and the BD model compared to the mugration model [47] is that they do not assume the migration process to be independent of the tree generating process. Notwithstanding, in our simulations, the birth, growth and the migration rates between demes never differed. It would be interesting to see, whether asymmetric rates can be better discovered by either model. Further, it would be exciting to see, whether e.g. different birth rates could even lead to more precise estimates of the migration rates, as they would provide more evidence for a lineage being in a certain location in the past.

Furthermore, first attempts were made to evaluate the model's performance on the recovery of the internal node types for the SC (compare A.2.2). Nevertheless, inferring the internal node states with the BD model took too long to complete. At the time of writing this thesis, a software package is being implemented into BEAST which will allow faster inference of the internal node types for the BD model. We would be excited to see how the models compare in that regard.

Bibliography

[1] T. Bedford, S. Cobey, P. Beerli, and M. Pascual, "Global migration dynamics underlie evolution and persistence of human influenza A (H3N2)." *PLoS pathogens*, vol. 6, no. 5, p. e1000918, may 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20523898http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2877742

[2] J. Bahl, M. I. Nelson, K. H. Chan, R. Chen, D. Vijaykrishna, R. A. Halpin, T. B. Stockwell, X. Lin, D. E. Wentworth, E. Ghedin, Y. Guan, J. S. M. Peiris, S. Riley, A. Rambaut, E. C. Holmes, and G. J. D. Smith, "Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans," *Proceedings of the National Academy of Sciences*, vol. 108, no. 48, pp. 19 359–19 364, nov 2011. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1109314108

[3] P. Lemey, A. Rambaut, T. Bedford, N. Faria, F. Bielejec, G. Baele, C. A. Russell, D. J. Smith, O. G. Pybus, D. Brockmann, and M. A. Suchard, "Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2." *PLoS pathogens*, vol. 10, no. 2, p. e1003932, feb 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24586153http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3930559

[4] T. Bedford, S. Riley, I. G. Barr, S. Broor, M. Chadha, N. J. Cox, R. S. Daniels, C. P. Gunasekaran, A. C. Hurt, A. Kelso, A. Klimov, N. S. Lewis, X. Li, J. W. McCauley, T. Odagiri, V. Potdar, A. Rambaut, Y. Shu, E. Skepner, D. J. Smith, M. A. Suchard, M. Tashiro, D. Wang, X. Xu, P. Lemey, and C. A. Russell, "Global circulation patterns of seasonal influenza viruses vary with antigenic drift," *Nature*, vol. 523, no. 7559, pp. 217–220, jul 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26053121http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4499780http://www.nature.com/articles/nature14460

[5] N. R. Faria, A. Rambaut, M. A. Suchard, G. Baele, T. Bedford, M. J. Ward, A. J. Tatem, J. D. Sousa, N. Arinaminpathy, J. Pépin, D. Posada, M. Peeters, O. G. Pybus, and P. Lemey, "HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations." *Science (New York, N.Y.)*, vol. 346, no. 6205, pp. 56–61, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25278604http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4254776

[6] E. M. Volz, N. Ndembi, R. Nowak, G. H. Kijak, J. Idoko, P. Dakum, W. Royal, S. Baral, M. Dybul, W. A. Blattner, and M. Charurat, "Phylodynamic analysis to inform prevention efforts in mixed HIV epidemics," *Virus Evolution*, vol. 3, no. 2, jul 2017. [Online]. Available: https://academic.oup.com/ve/article-lookup/doi/10.1093/ve/vex014

[7] G. Dudas, L. M. Carvalho, T. Bedford, A. J. Tatem, G. Baele, N. R. Faria, D. J. Park, J. T. Ladner, A. Arias, D. Asogun, F. Bielejec, S. L. Caddy, M. Cotten, J. D'Ambrozio, S. Dellicour, A. Di Caro, J. W. Diclaro, S. Duraffour, M. J. Elmore, L. S. Fakoli, O. Faye, M. L. Gilbert, S. M. Gevao, S. Gire,

A. Gladden-Young, A. Gnirke, A. Goba, D. S. Grant, B. L. Haagmans, J. A. Hiscox, U. Jah, J. R. Kugelman, D. Liu, J. Lu, C. M. Malboeuf, S. Mate, D. A. Matthews, C. B. Matranga, L. W. Meredith, J. Qu, J. Quick, S. D. Pas, M. V. T. Phan, G. Pollakis, C. B. Reusken, M. Sanchez-Lockhart, S. F. Schaffner, J. S. Schieffelin, R. S. Sealfon, E. Simon-Loriere, S. L. Smits, K. Stoecker, L. Thorne, E. A. Tobin, M. A. Vandi, S. J. Watson, K. West, S. Whitmer, M. R. Wiley, S. M. Winnicki, S. Wohl, R. Wölfel, N. L. Yozwiak, K. G. Andersen, S. O. Blyden, F. Bolay, M. W. Carroll, B. Dahn, B. Diallo, P. Formenty, C. Fraser, G. F. Gao, R. F. Garry, I. Goodfellow, S. Günther, C. T. Happi, E. C. Holmes, B. Kargbo, S. Keïta, P. Kellam, M. P. G. Koopmans, J. H. Kuhn, N. J. Loman, N. Magassouba, D. Naidoo, S. T. Nichol, T. Nyenswah, G. Palacios, O. G. Pybus, P. C. Sabeti, A. Sall, U. Ströher, I. Wurie, M. A. Suchard, P. Lemey, and A. Rambaut, "Virus genomes reveal factors that spread and sustained the Ebola epidemic," *Nature*, vol. 544, no. 7650, pp. 309–315, apr 2017. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature22040

[8] D. Kühnert, M. Coscolla, D. Brites, D. Stucki, J. Metcalfe, L. Fenner, S. Gagneux, and T. Stadler, "Tuberculosis outbreak investigation using phylodynamic analysis," *Epidemics*, vol. 25, pp. 47–53, dec 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1755436517301068

[9] I. Gronau, M. J. Hubisz, B. Gulko, C. G. Danko, and A. Siepel, "Bayesian inference of ancient human demography from individual genome sequences." *Nature genetics*, vol. 43, no. 10, pp. 1031–4, sep 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21926973http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3245873

[10] T. Mailund, A. E. Halager, M. Westergaard, J. Y. Dutheil, K. Munch, L. N. Andersen, G. Lunter, K. Prüfer, A. Scally, A. Hobolth, and M. H. Schierup, "A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species." *PLoS genetics*, vol. 8, no. 12, p. e1003125, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23284294http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3527290

[11] C. J. Edwards, M. A. Suchard, P. Lemey, J. J. Welch, I. Barnes, T. L. Fulton, R. Barnett, T. C. O'Connell, P. Coxon, N. Monaghan, C. E. Valdiosera, E. D. Lorenzen, E. Willerslev, G. F. Baryshnikov, A. Rambaut, M. G. Thomas, D. G. Bradley, and B. Shapiro, "Ancient hybridization and an Irish origin for the modern polar bear matriline." *Current biology : CB*, vol. 21, no. 15, pp. 1251–8, aug 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21737280http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4677796

[12] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes, "Unifying the Epidemiological and Evolutionary Dynamics of Pathogens," *Science*, vol. 303, no. 5656, pp. 327–332, jan 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14726583http://www.sciencemag.org/cgi/doi/10.1126/science.1090727

[13] T. Stadler and S. Bonhoeffer, "Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic

methods," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1614, pp. 20 120 198–20 120 198, feb 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23382421http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3678323http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2012.0198

[14] D. G. KENDALL, "ON SOME MODES OF POPULATION GROWTH LEADING TO R. A. FISHER'S LOGARITHMIC SERIES DISTRIBUTION," *Biometrika*, vol. 35, no. 1-2, pp. 6–15, may 1948. [Online]. Available: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/35.1-2.6

[15] S. Nee, R. M. May, and P. H. Harvey, "The reconstructed evolutionary process," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 344, no. 1309, pp. 305–311, may 1994. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7938201http://www.royalsocietypublishing.org/doi/10.1098/rstb.1994.0068

[16] Z. Yang and B. Rannala, "Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method," *Molecular Biology and Evolution*, vol. 14, no. 7, pp. 717–724, jul 1997. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9214744https://academic.oup.com/mbe/article-lookup/doi/10.1093/oxfordjournals.molbev.a025811

[17] T. Stadler, "On incomplete sampling under birth–death models and connections to the sampling-based coalescent," *Journal of Theoretical Biology*, vol. 261, no. 1, pp. 58–66, nov 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19631666https://linkinghub.elsevier.com/retrieve/pii/S0022519309003300

[18] T. Stadler, "Sampling-through-time in birth–death trees," *Journal of Theoretical Biology*, vol. 267, no. 3, pp. 396–404, dec 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20851708https://linkinghub.elsevier.com/retrieve/pii/S0022519310004765

[19] E. M. Volz and S. D. W. Frost, "Sampling through time and phylodynamic inference with coalescent and birth-death models," *Journal of The Royal Society Interface*, vol. 11, no. 101, pp. 20 140 945–20 140 945, oct 2014. [Online]. Available: http://rsif.royalsocietypublishing.org/cgi/doi/10.1098/rsif.2014.0945

[20] D. Kühnert, T. Stadler, T. G. Vaughan, and A. J. Drummond, "Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data," *Molecular Biology and Evolution*, vol. 33, no. 8, pp. 2102–2116, 2016.

[21] R. R. Hudson, "Gene genealogies and the coalescent process." *Oxford Surveys in Evolutionary Biology*, vol. 7, pp. 1–44, 1990. [Online]. Available: https://www.cabdirect.org/cabdirect/abstract/19910191040

[22] N. Takahata, "The coalescent in two partially isolated diffusion populations." *Genetical research*, vol. 52, no. 3, pp. 213–22, dec 1988. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3243426

[23] M. Notohara, "The coalescent and the genealogical process in geographically structured population," *Journal of Mathematical Biology*, vol. 29, no. 1, pp. 59–75, oct 1990. [Online]. Available: http://link.springer.com/10.1007/BF00173909

[24] J. F. C. Kingman, "On the genealogy of large populations," *Journal of Applied Probability*, vol. 19, no. A, pp. 27–43, jul 1982. [Online]. Available: https://www.cambridge.org/core/product/identifier/S0021900200034446/type/journal{_}article

[25] M. Notohara, "The coalescent and the genealogical process in geographically structured population," *Journal of Mathematical Biology*, vol. 29, no. 1, pp. 59–75, oct 1990. [Online]. Available: http://link.springer.com/10.1007/BF00173909

[26] P. Beerli and J. Felsenstein, "Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach," *Proceedings of the National Academy of Sciences*, vol. 98, no. 8, pp. 4563–4568, apr 2001. [Online]. Available: https://www.pnas.org/content/98/8/4563

[27] A. G. Rodrigo, E. G. Shpaer, E. L. Delwart, A. K. Iversen, M. V. Gallo, J. Brojatsch, M. S. Hirsch, B. D. Walker, and J. I. Mullins, "Coalescent estimates of HIV-1 generation time in vivo." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 5, pp. 2187–91, mar 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10051616http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC26758

[28] E. T. E. T. Jaynes and G. L. Bretthorst, *Probability theory : the logic of science*. Cambridge University Press, 2003.

[29] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, jun 1953. [Online]. Available: http://aip.scitation.org/doi/10.1063/1.1699114

[30] D. Kühnert, C.-H. Wu, and A. J. Drummond, "Phylogenetic and epidemic modeling of rapidly evolving infectious diseases," *Infection, Genetics and Evolution*, vol. 11, no. 8, pp. 1825–1841, dec 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156713481100284X?via{%}3Dihub{#}f0015

[31] E. M. Volz, K. Koelle, and T. Bedford, "Viral Phylodynamics," *PLoS Computational Biology*, vol. 9, no. 3, p. e1002947, mar 2013. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1002947

[32] B. Mau, M. A. Newton, and B. Larget, "Bayesian phylogenetic inference via Markov chain Monte Carlo methods," *Biometrics*, vol. 55, no. 1, pp. 1–12, mar 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11318142

[33] A. J. Drummond, G. K. Nicholls, A. G. Rodrigo, and W. Solomon, "Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data." *Genetics*, vol. 161, no. 3, pp. 1307–20, jul 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12136032http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1462188

[34] J. Felsenstein, "Evolutionary trees from DNA sequences: A maximum likelihood approach," *Journal of Molecular Evolution*, vol. 17, no. 6, pp. 368–376, nov 1981. [Online]. Available: http://link.springer.com/10.1007/BF01734359

[35] R. Durrett and S. Levin, "The Importance of Being Discrete (and Spatial)," *Theoretical Population Biology*, vol. 46, no. 3, pp. 363–394, dec 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S040058098471032X

[36] G. E. Leventhal, R. Kouyos, T. Stadler, V. von Wyl, S. Yerly, J. Böni, C. Cellerai, T. Klimkait, H. F. Günthard, and S. Bonhoeffer, "Inferring Epidemic Contact Structure from Phylogenetic Trees," *PLoS Computational Biology*, vol. 8, no. 3, p. e1002413, mar 2012. [Online]. Available: http://dx.plos.org/10.1371/journal.pcbi.1002413

[37] D. A. Rasmussen, R. Kouyos, H. F. Günthard, and T. Stadler, "Phylodynamics on local sexual contact networks," *PLOS Computational Biology*, vol. 13, no. 3, p. e1005448, mar 2017. [Online]. Available: http://dx.plos.org/10.1371/journal.pcbi.1005448

[38] K. Robinson, N. Fyson, T. Cohen, C. Fraser, and C. Colijn, "How the Dynamics and Structure of Sexual Contact Networks Shape Pathogen Phylogenies," *PLoS Computational Biology*, vol. 9, no. 6, p. e1003105, jun 2013. [Online]. Available: http://dx.plos.org/10.1371/journal.pcbi.1003105

[39] V. Boskova, S. Bonhoeffer, and T. Stadler, "Inference of Epidemiological Dynamics Based on Simulated Phylogenies Using Birth-Death and Coalescent Models," *PLoS Computational Biology*, vol. 10, no. 11, 2014.

[40] T. G. Vaughan and A. J. Drummond, "A Stochastic Simulator of Birth-Death Master Equations with Application to Phylodynamics," *Molecular Biology and Evolution*, vol. 30, no. 6, pp. 1480–1493, jun 2013. [Online]. Available: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mst057

[41] E. M. Volz, "Complex Population Dynamics and the Coalescent Under Neutrality," *Genetics*, vol. 190, no. 1, pp. 187–201, jan 2012. [Online]. Available: http://www.genetics.org/lookup/doi/10.1534/genetics.111.134627

[42] T. G. Vaughan, D. Kühnert, A. Popinga, D. Welch, and A. J. Drummond, "Efficient Bayesian inference under the structured coalescent," *Bioinformatics*, vol. 30, no. 16, pp. 2272–2279, 2014.

[43] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, dec 1977. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/j100540a008

[44] R. Bouckaert, J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond, "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis," *PLoS Computational Biology*, vol. 10, no. 4, p. e1003537, apr 2014. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1003537

[45] N. F. Mueller, G. Dudas, and T. Stadler, "Inferring time-dependent migration and coalescence patterns from genetic sequence and predictor data in structured populations," *bioRxiv*, p. 342329, jun 2018. [Online]. Available: https://www.biorxiv.org/content/10.1101/342329v1

[46] E. M. Volz and I. Siveroni, "Bayesian phylodynamic inference with complex models," *PLOS Computational Biology*, vol. 14, no. 11, p. e1006546, nov 2018. [Online]. Available: http://dx.plos.org/10.1371/journal.pcbi.1006546

[47] P. Lemey, A. Rambaut, A. J. Drummond, and M. A. Suchard, "Bayesian Phylogeography Finds Its Roots," *PLoS Computational Biology*, vol. 5, no. 9, p. e1000520, sep 2009. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1000520

Appendix A

Appendix

A.1 Methods

A.1.1 Tree simulations

Table A.1: Parameters that will be used in the tree simulations; the columns indicate the parameters varied; the rows growth can be

	birth rate		migration rate		sampling proportion	
Cases	deme 0	deme 1	deme 0	deme 1	deme 0	deme 1
1	1.05	1.05	1.0	1.0	0.01	0.01
2	1.05	1.05	0.1	0.1	0.01	0.01
3	1.05	1.05	0.01	0.01	0.01	0.01
4	2.0	2.0	1.0	1.0	0.01	0.01
5	2.0	2.0	0.1	0.1	0.01	0.01
6	2.0	2.0	0.01	0.01	0.01	0.01
7	1.05	1.05	1.0	5.0	0.01	0.01
8	1.05	1.05	0.1	0.5	0.01	0.01
9	1.05	1.05	0.01	0.05	0.01	0.01
10	2.0	2.0	1.0	5.0	0.01	0.01
11	2.0	2.0	0.1	0.5	0.01	0.01
12	2.0	2.0	0.01	0.05	0.01	0.01

A.1.2 BD simulations

Table A.2: Sampling proportions for BDpsi tree simulations. For each case, the minimum, medium and maximum sampling proportion for each deme are reported.

		deme 0			deme 1		
Case	F	min	median	max	min	median	max
7	27	0,03	0,16	0,51	0,03	0,16	0,51
8	27	0,03	0,13	0,41	0,03	0,14	0,63
9	28	0,008	0,05	0,36	0,09	0,19	0,66
10	40	0,34	0,57	0,78	0,38	0,57	0,83
11	41	0,39	0,51	0,74	0,40	0,63	0,81
12	40	0,32	0,51	0,68	0,36	0,63	0,82

Table A.3: Final population sizes N_{final} for BDpsi simulations and expected simulation end time t_{final}.

Cases	N_{final}	r	t_{final}
1 - 3, 7 - 9	1000	0,05	138
4 - 6, 10 - 12	5000	1,0	8,5
13 - 15	30	0,05	138
16 - 18	40	1,0	3,7

Table A.4: Sampling times for BD with deterministic sampling.

Cases	t1	t2	t3	t4	t5
1 - 3	41	48	54	61	68
4 - 6	5	5.5	6	6.5	7
7 - 9	5	10	15	20	25
10 - 12	2.2	2.6	3.0	3.3	3.7

A.1.3 MCMC settings

Table A.5: Priors on parameters in birth-death inference for trees simulated with deterministic sampling time points.

	Lognormal Prior		HPD 95%	
Parameters	μ	σ	min	max
birth rate	0	3	0	150
migration rate	0	3	0	150
sampling proportions	0	1	0	5

Table A.6: Priors on parameters in birth-death inference for trees simulated with a psi sampling scheme.

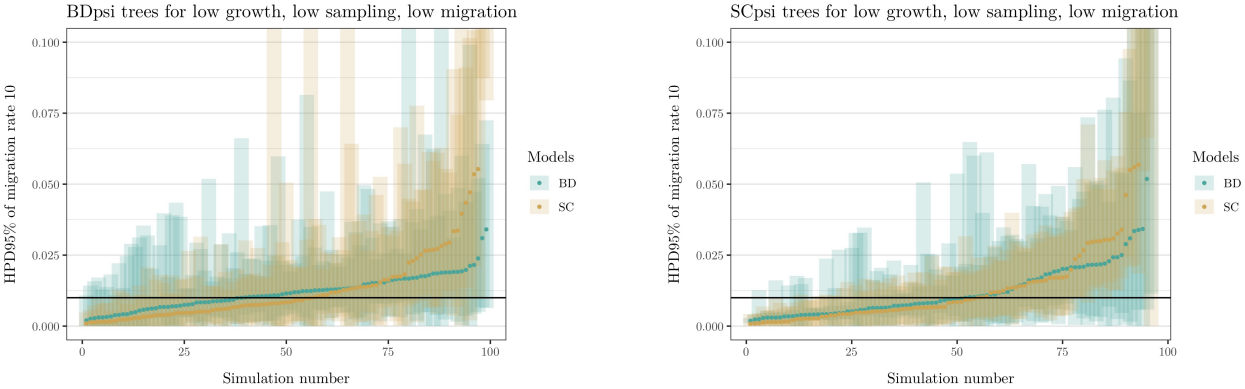
	Lognormal Prior		Prior's HPD95	
Parameters	μ	σ	min	max
birth rate	0	3	0	150
migration rate	0	3	0	150
sampling proportion	m	1		

Table A.7: Priors on parameters in structured coalescence inference.

	Lognormal Prior		Prior's HPD95	
Parameters	μ	σ	min	max
population size	2	3	0	1400
migration rate	0	3	0	150

A.2 Results

A.2.1 Supplement to the Comparison of the BD and SC with psi sampling



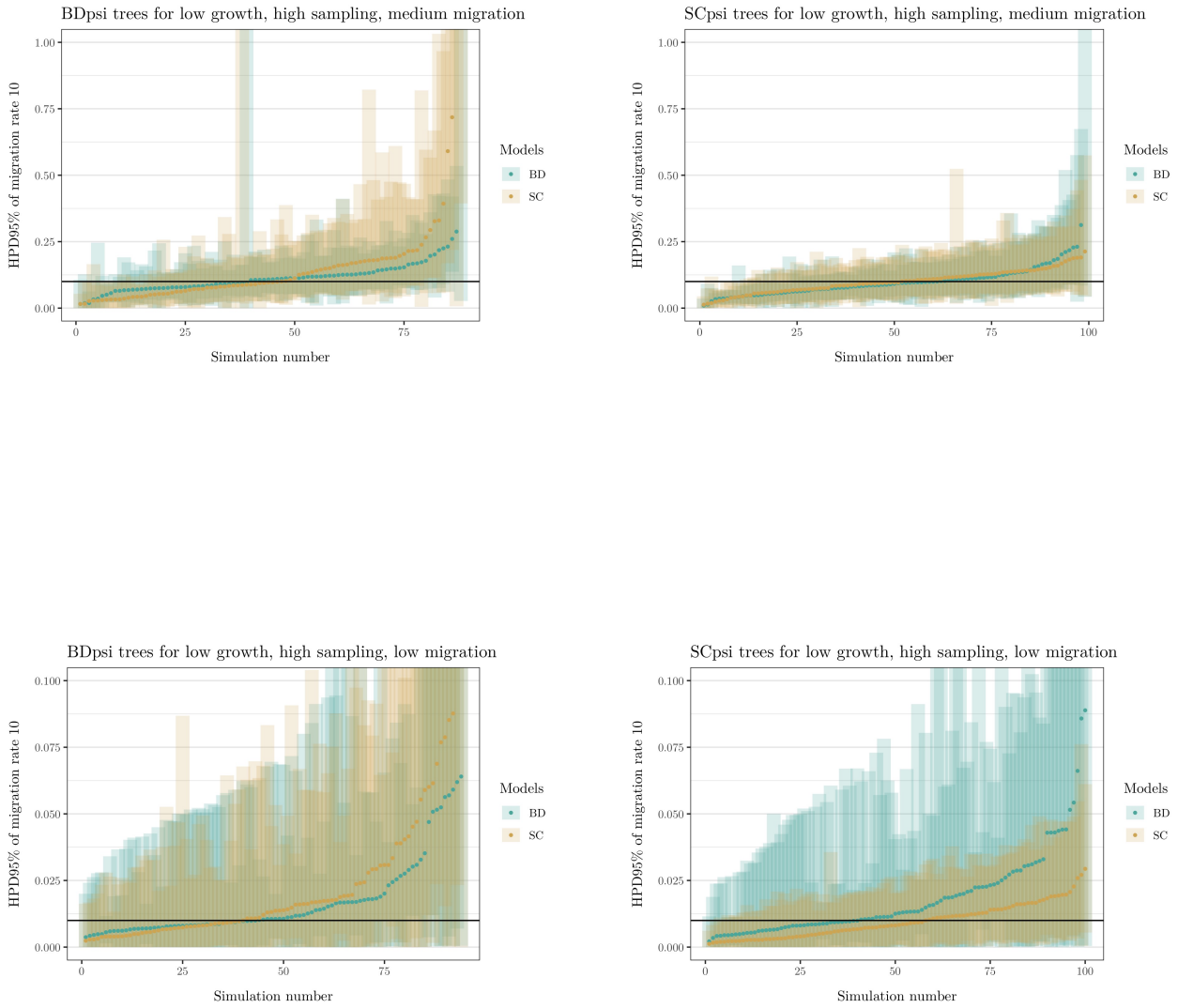


Figure A.3: For every plot we simulated 100 trees under the BD with psi sampling (BDpsi) or the SC conditioned on psi sampled leaves (SCpsi) and analysed them using the BD (turquoise) and the SC (ochre) model. We show the 95% highest posterior density (HPD) intervals of the estimated migration rate m10 as shaded segments and their medians as dots. The HPD regions are ordered with increasing median. The true migration rate is displayed as a black, horizontal line. The figures' individual captions indicate the growth, sampling and migration parameterisation.

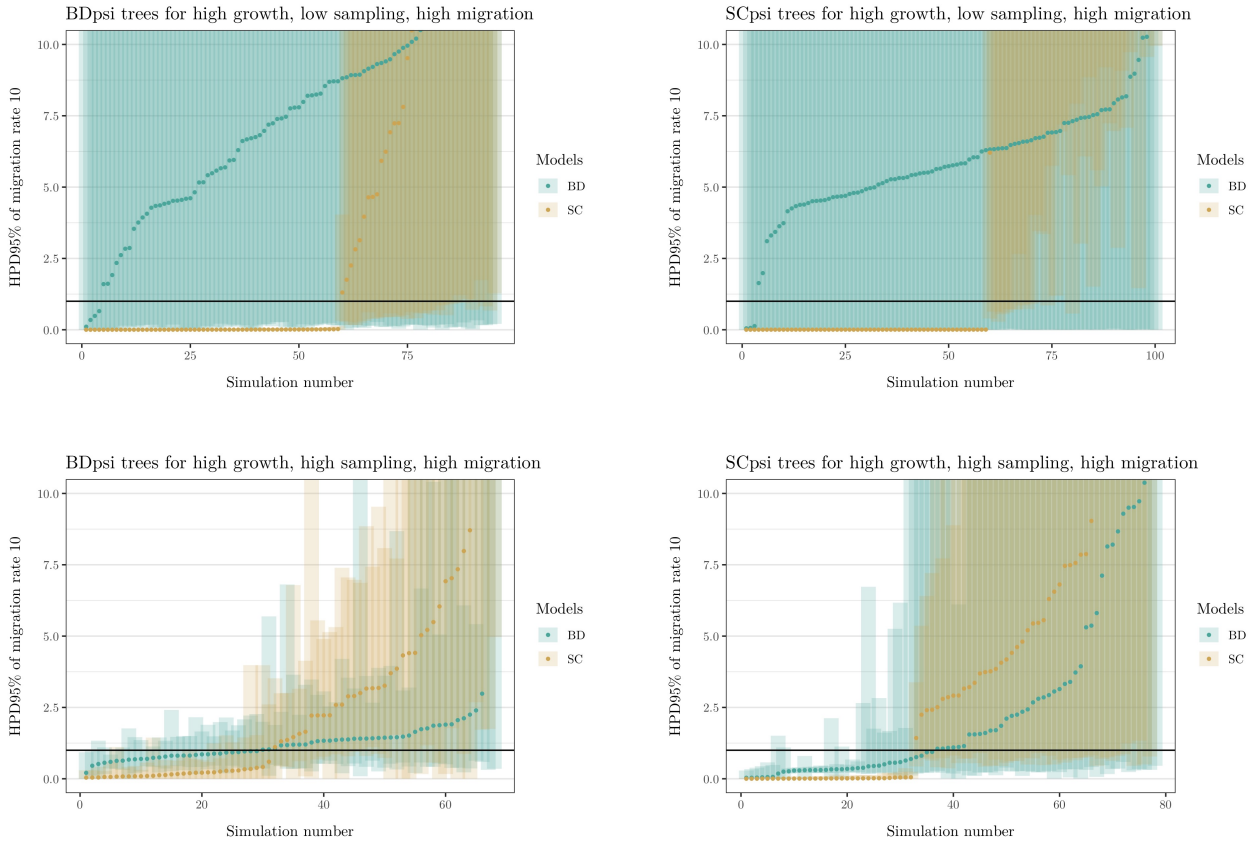
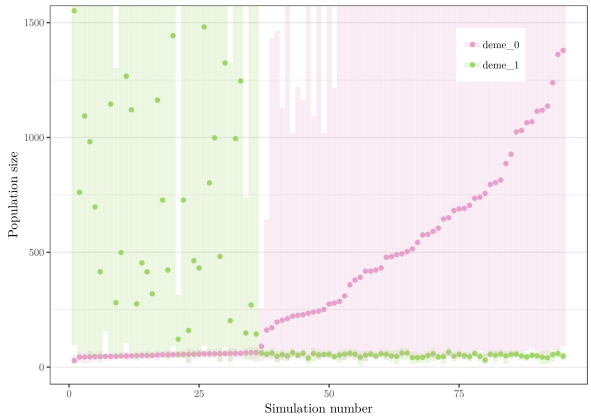
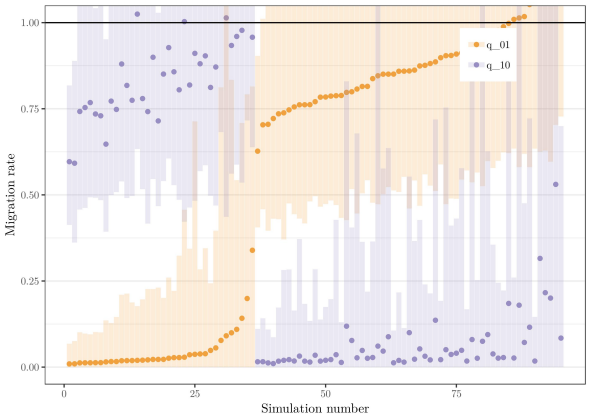
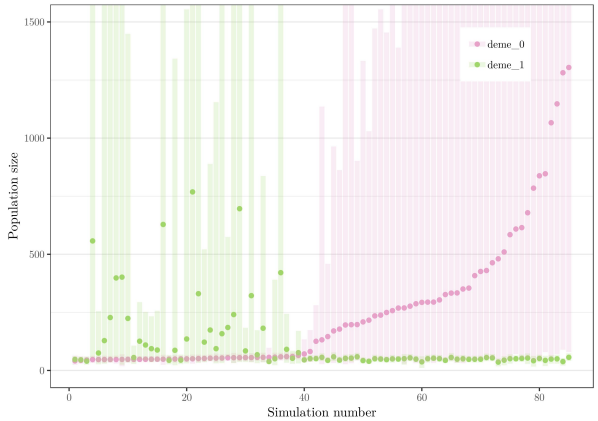


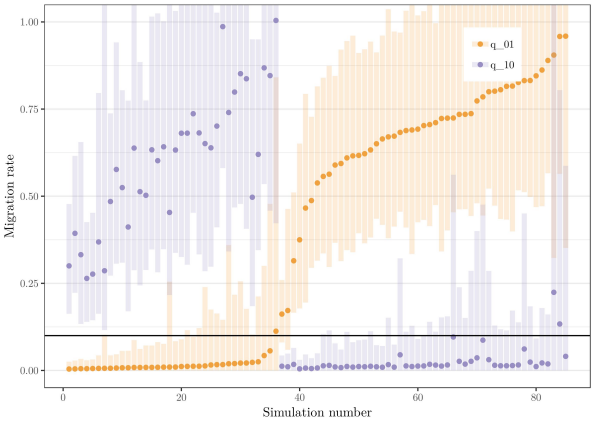
Figure A.5: For every plot we simulated 100 trees under the BD with psi sampling (BDpsi) or the SC for a growing population conditioned on psi sampled leaves (SCpsi) and analysed them using the BD (turquoise) and the SC (ochre) model. We show the 95% highest posterior density (HPD) intervals of the estimated migration rate m01 as shaded segments and their medians as dots. The HPD regions are ordered with increasing median. The true migration rate is displayed as a black, horizontal line. The figures' individual captions indicate the growth, sampling and migration parameterisation.

Figure A.8: **SC interpretation of high growth and high sampling migration from SCpsi trees.** For each tree the inferred median population sizes in the left and migration rates in the right plot are shown in increasing order of population size in deme 0. The black line indicates the true migration rate parameter. Migration decreases the from "high' over "medium' to "low" from row one to three.


deme_0
deme_1
Population size
Simulation number


q_01
q_10
Migration rate
Simulation number


deme_0
deme_1
Population size
Simulation number


q_01
q_10
Migration rate
Simulation number

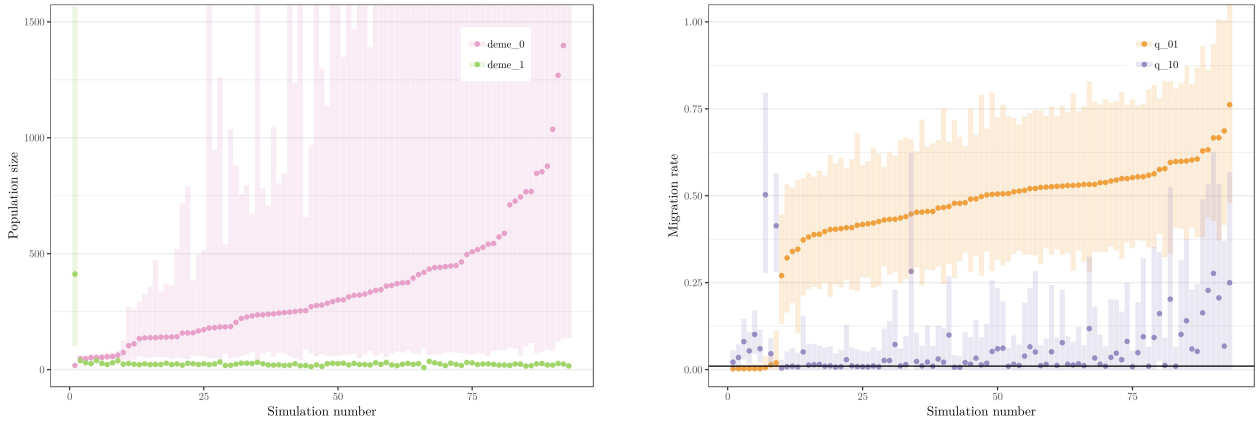
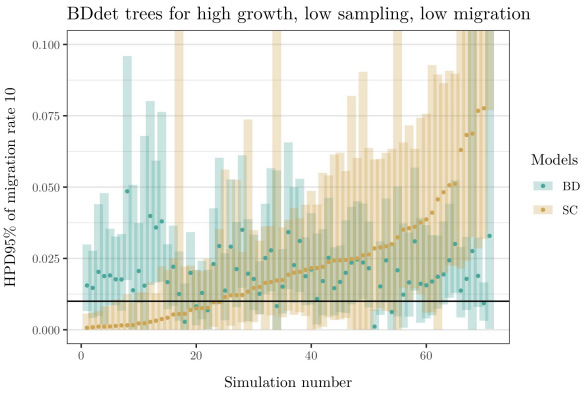
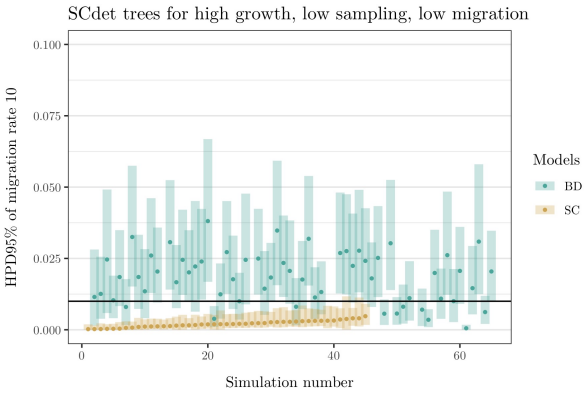
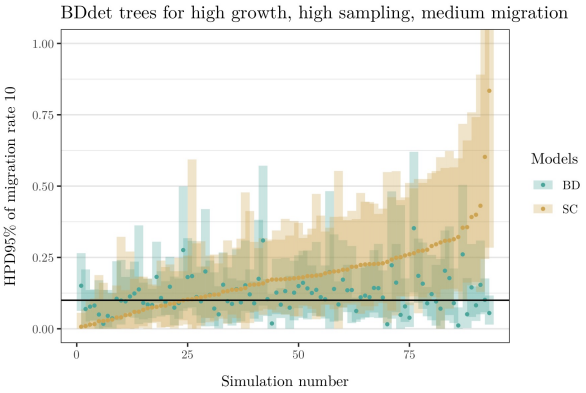
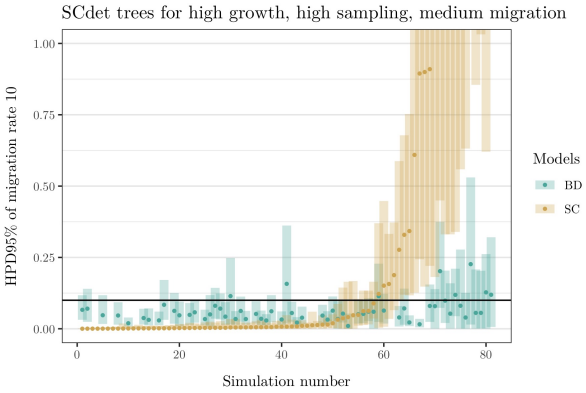


Figure A.11: **SC interpretation of high growth and low sampling from BDpsi trees.** For each tree the inferred median population sizes in the left and migration rates in the right plot are shown in increasing order of population size in deme 0. The black line indicates the true forward migration rate. Migration is decreasing from "high" over "medium" to "low" from row one to three.


BDdet trees for high growth, low sampling, low migration
HPD95% of migration rate 10
0.100
0.075
0.050
0.025
0.000
0
20
40
60
Simulation number
Models
BD
SC


SCdet trees for high growth, low sampling, low migration
HPD95% of migration rate 10
0.100
0.075
0.050
0.025
0.000
0
20
40
60
Simulation number
Models
BD
SC


BDdet trees for high growth, high sampling, medium migration
HPD95% of migration rate 10
1.00
0.75
0.50
0.25
0.00
0
25
50
75
Simulation number
Models
BD
SC


SCdet trees for high growth, high sampling, medium migration
HPD95% of migration rate 10
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Simulation number
Models
BD
SC

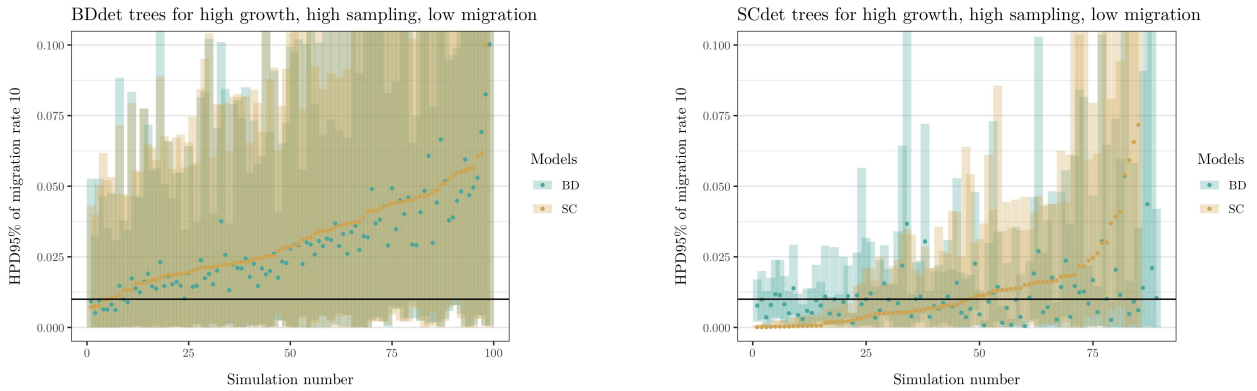


Figure A.14: For every plot we simulated 100 trees under the BD with deterministic sampling (BDdet) or the SC conditioned on deterministically sampled leaves (SCdet) and analysed them using the BD (turquoise) and the SC (ochre) model. We show the 95% highest posterior density (HPD) intervals of the estimated migration rate mr_{01} as shaded segments and their medians as dots. The HPD regions are ordered with increasing median. The true migration rate is displayed as a black, horizontal line. The figures' individual captions indicate the growth, sampling and migration parameterisation.

A.2.2 First attempts at internal node color recovery

As the structured coalescent model infers the locations of the internal node types, the obvious question arose: How good are the estimated internal node locations? In the absence of a way to compare them to the BD model, we here show two exemplary results for the inference from SCdet trees under low growth and high sampling.

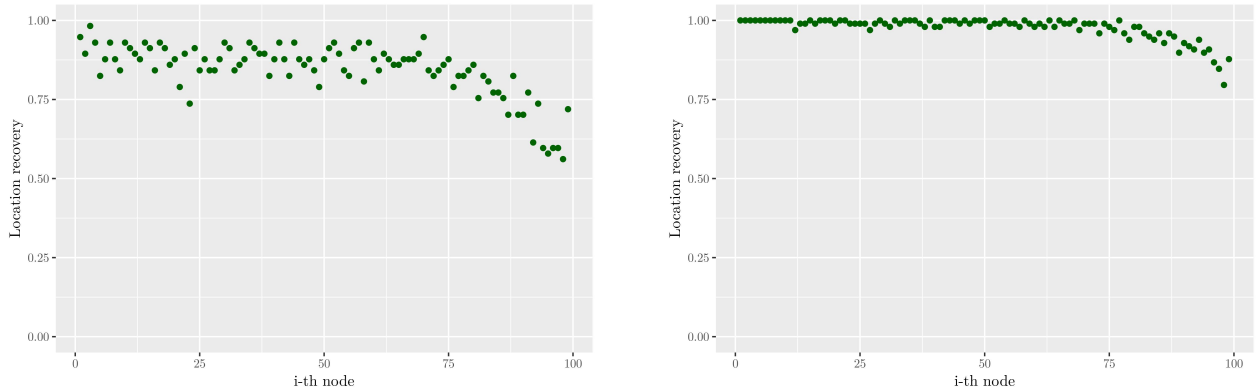


Figure A.15: **Internal node type recovery** For each inferred internal node from 1 to 99, we compute the frequency of correct internal node state estimation from the inferred typed trees of the SC model. Internal nodes are ordered with increasing distance from the tips. The analysis was performed for the SCdet trees for medium migration (left) and low migration (right).