Phylodynamics of Somatic Evolution: A Likelihood-Based Approach for Cellular Reproduction

Tobias Dieselhorst[1] and Johannes Berg*[1]

[1]Institute for Biological Physics, University of Cologne, Zülpicher Str. 77, 50937 Cologne, Germany

The way mutations accumulate over the course of evolution fundamentally differs between macroevolutionary settings and cellular reproduction. In macroevolution, mutations generally accumulate continuously over evolutionary timescales comprising many generations. Under cellular reproduction, like somatic evolution within a tumor or evolution within a bacterial population, mutations can occur in discrete steps at individual reproductive events. In this paper, we address the problem of phylodynamic inference in the context of cellular reproduction. We find that mutations occurring at individual cell divisions generate a compound Poisson statistics for the number of mutations along a phylogenetic branch, and this affects the inference from phylogenetic data. These statistics differ from those of molecular clock models even when only a small fraction of the population is sampled. We set up an inference scheme based on the likelihood of a birth-death model to infer the relative death-to-birth rate and the number of mutations per generation. We apply this method to single-cell data from haematopoietic stem cells. Our approach is computationally efficient thus allowing a broad range of applications, in particular on lineage trees from single-cell data.

*correspondence to bergj@uni-koeln.de

1 Introduction

Evolutionary analysis seeks to infer the past dynamics of a population, or several populations, on the basis of extant data. A wide range of concepts and tools have been developed to determine the evolutionary events that shaped a population, ranging from the problem of inferring rates of birth and death of species to the timing of bifurcations in trees [1, 2, 3, 4, 5, 6].

These inference schemes are based on specific models of (i) the underlying population dynamics generating the population and hence the reconstructed phylogenetic tree and (ii) models specifying the mutations along a phylogenetic tree. The first model typically involves a birth-death process with birth and death rates that may be homogeneous or heterogeneous across the tree [3, 4]. For the second model, the standard choice is the molecular clock hypothesis, which assumes that mutations accumulate at some rate, constant or variable [7], per unit time. This approach is well-suited to a macro-evolutionary setting, where birth and death describe the speciation and extinction of species over long time-scales and many generations, and mutations arise through the procreation of individuals and the fixation of these mutations in a population.

Here, we develop an approach to phylodynamic modelling and inference where mutations are tied to the divisions of individual cells, rather than the passage of time. This link between cell divisions (birth events) and mutations is specific to cellular reproduction, where bifurcations in the phylogenetic tree correspond to divisions of individual cells, rather than speciation events of entire populations. In this case, the mutation statistics is affected by individual birth events along lineages. Some of these birth events correspond to bifurcations of the phylogenetic tree and are easily detected. Other birth events produce a lineage that eventually dies out, and these events are not directly seen in the phylogenetic tree, see Figure 1. Yet these 'hidden' birth events contribute to the accumulation of mutations. Combining the number of hidden birth events along a branch with the number of mutations at each birth event leads to a compound statistics for the number of mutations along each branch. Specifically, we find that the standard Poisson statistics arising in constant-rate molecular clock models is replaced with the so-called compound Poisson statistics, with an increased variance arising from discrete reproductive events along branches compared to a molecular clock. Previous macroevolutionary models introduced compound Poisson statistics in an ad hoc manner to relax the molecular clock assumption [8, 9]. In contrast, we show that this compound structure emerges naturally from the dynamics of cellular reproduction.

Our inference method is designed to handle large-scale single-cell datasets, which are derived from populations of cells undergoing division and death. Examples include the spread of bacteria, tissue development, or the uncontrolled growth of tumor cells. In order to keep the computations simple, we restrict the discussion to the infinite sites model. Under this model, the topology of the phylogenetic tree and the ancestral states are straightforward to infer. To facilitate the efficient computation of the likelihood of a tree (specified by its topology and the mutations along each of

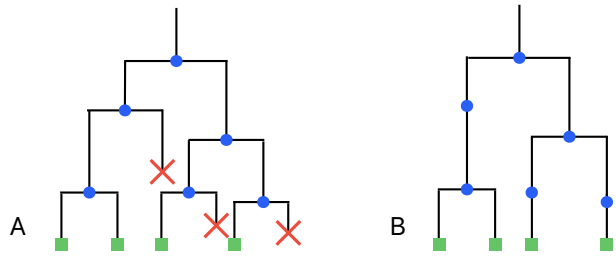

Figure 1: Comparison of a birth-death tree and the corresponding reconstructed phylogenetic tree. (A) A birth-death process of individuals with time running downwards consists of birth events (blue circles) and death events (red crosses) and produces some individuals (green squares) that are alive at the time of observation. Each birth event produces a bifurcation in the birth-death tree. (B) The corresponding phylogenetic tree reconstructed from extant individuals only. Birth events are again indicated by blue circles and can now occur *within* phylogenetic branches ('hidden' birth events), as well as at all bifurcations of the phylogenetic tree.

its branches), we use a dynamic programming approach. This involves iteratively integrating over all branch lengths (in calendar time) and summing over the number of cell divisions per branch. The algorithm achieves running times of just seconds for hundreds of samples and scales linearly with size of the tree.

In Section 2.1, we recapitulate the dynamics of a birth-death process and establish the statistics of mutations along tree branches. We introduce our algorithm to compute the likelihood of the corresponding model parameters and test its capabilities on simulated data. In Section 2.2, we generalize these results to trees reconstructed from a finite fraction of sampled individuals. We apply our method to single-cell data of haematopoietic stem cells [10] in Section 2.3, analyzing prenatal haematopoiesis as well as detecting distinct dynamics in different clades found in an adult donor.

2 Results

2.1 Parameter inference from mutations along branches

We consider the growth of a population under a constant-rate birth-death process. Without loss of generality, we chose the unit of time such that the birth rate is one and denote the rate of cell death (relative to the birth rate) by $q < 1$. By convention, time runs backwards into the evolutionary past from the time of observation. From the reconstructed phylogenetic (topological) tree and the mutations along its branches, we aim to infer the underlying phylodynamic model. Such a model consists of two components: The birth-death process (described by the rates of cell division and cell death), and the statistics of mutations arising at each cell division.

The first component, the birth-death process, is characterized by the joint statistics of the length τ of a branch in calendar time and the number of generations i along it. The following distribution

combines these two measures for a clade starting with a single individual at time τ_s in the past: the joint probability that the individual has undergone i birth events by time $\tau_e = \tau_s - \tau$ and is alive at time τ_e, but all i lineages generated between τ_s and τ_e have died out by the time of sampling is given by

$$p_1(i|\tau_s, \tau_e) = \mathrm{e}^{-(1+q)(\tau_s-\tau_e)} \frac{1}{i!} \left[2q\tau + 2\log\left(\frac{1 - q\,\mathrm{e}^{-(1-q)\tau_e}}{1 - q\,\mathrm{e}^{-(1-q)\tau_s}} \right) \right]^i . \tag{1}$$

The fate of the particular individual alive at time τ_e after that time is not specified. A derivation can be found in Supplementary Information B. The number of generations i along a branch starting at τ_s and ending at τ_e is Poisson distributed with a mean that depends non-linearly on the age of the branch [11].

The second component concerns the statistics of the mutations occurring at each birth event (cell division). Here, we consider the number of mutations per cell division to be Poisson distributed with mean μ, so the number of mutations along a branch with i generations along it plus one additional generation founding the branch is

$$P(m|i) = \exp(-(i+1)\mu)((i+1)\mu)^m/m! \ . \tag{2}$$

Combining this Poisson distribution with the (also Poissonian) distribution of the number of generations (1) along a given branch, the number of mutations along the branch follows a compound Poisson distribution. Compound Poisson statistics is generated by a sum over independent, identically distributed random variables (in our case the number of mutations per cell division) where the number of random variables is itself a Poisson distributed random variable (in our case the number of generations i). The compound Poisson distribution has a larger variance than a simple Poisson distribution; Figure 2 shows two examples. For a sufficiently large number of mutations per generation, one can see the effects of individual cell divisions along a branch in the form of oscillations. For details on the compound Poisson distribution, see Supplementary Information C.

We now combine these two components of a phylodynamic model to construct a likelihood function based on the probability that a tree with the empirically observed numbers of mutations along its branches arises under a birth-death process and the mutation statistics (2). Branch lengths in calendar time and the number of birth events along each branch on the other hand are unobserved variables. In contrast to macroevolutionary data, in single-cell data there are often only few mutations along a branch, so branch lengths cannot be determined precisely. They act as latent variables in this inference problem. To sum over these latent variables, we use a simple iterative procedure, dynamic programming, in the spirit of Felsenstein's recursive algorithm [12]: we iteratively sum over both the times of bifurcations and the number of generations along each branch. The same iterative approach applied to bifurcation times is used in the *TreeTime* algorithm by Sagulenko, Puller, and Neher [6].

$P(\nu|\tau_s)$ denotes the probability that the part of the phylogenetic tree below a given branch ν

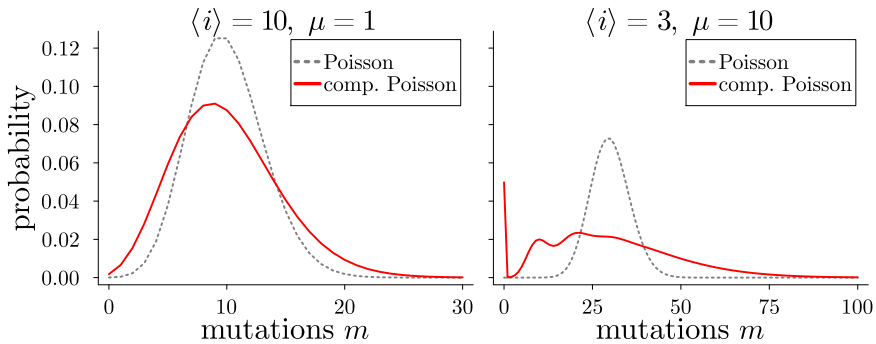

Figure 2: Compound Poisson distributed number of mutations. The distribution of the number of mutations follows a compound Poisson statistics arising from the discrete number of generations along a branch. The red solid lines show the distributions of the number of mutations along branches with mean number of generations $\langle i \rangle$ and mean number of mutations per generation μ. They are compared to Poisson distributions (grey dashed lines) with the same mean $\langle i \rangle \mu$ as they would arise from a molecular clock. Continuous lines are drawn for visibility. See Supplementary Information C for details.

(starting at the top of the branch which emerged at time τ_s) arose under a particular phylodynamic model. In this definition, the phylogenetic tree is specified by its topology and the number of mutations along the branches. The phylodynamic model is specified by the relative death rate q and the mean number μ of mutations per cell division. For a pendant branch (a branch starting at τ_s and ending in a leaf) with m mutations along it, $P(\nu|\tau_s)$ follows from the probability that a clade starting at time τ_s runs until time zero with i generations along it, but has no further extant offsprings (equation (1) with $\tau_e = 0$). From the distribution of the number of mutations along a branch with i generations (2) we directly obtain

$$P(\nu|\tau_s) = \sum_{i=0}^{\infty} P(m|i) p_1(i|\tau_s, \tau_e = 0) \ . \tag{3}$$

On the other hand, for an internal branch ν with m mutations along it and the two branches ν' and ν'' emanating from it we have

$$P(\nu|\tau_s) = \sum_{i=0}^{\infty} \int_0^{\tau_s} \mathrm{d}\tau_e P(m|i) p_1(i|\tau_s, \tau_e) 2 P(\nu'|\tau_e) P(\nu''|\tau_e) \ , \tag{4}$$

where we integrate over the time τ_e of the end of branch ν (and therefore the beginning of branches ν' and ν''). The factor of two arises as either of the two clades arising at τ_e can be ν' or ν''.

The likelihood for the model parameters given mutations along the tree can thus be evaluated iteratively, integrating over the calendar times of internal nodes and the number of generations along each branch in a computationally efficient manner. We start with (3) evaluated for all pendant branches, and then evaluate (4) for all branches ν whose $P(\nu'|\tau)$ and $P(\nu''|\tau)$ have already been evaluated for their descendant branches ν' and ν''. From the last two remaining branches ν' and ν'' emanating from the most recent common ancestor, we finally obtain the likelihood of the model

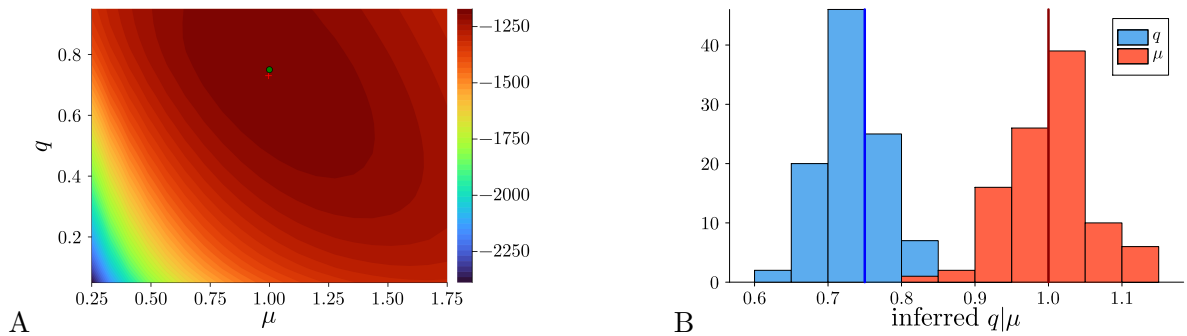

Figure 3: Inference using the likelihood of the birth-death process. (A) A population was grown from a single individual to $N = 300$ individuals with $q = 3/4$ and $\mu = 1$ and the likelihood (5) for the resulting tree (defined by topology and mutations along branches) was computed iteratively as described in the text. The maximum of the log-likelihood near $q = 0.73$ and $\mu = 0.996$ is marked with a red cross, the underlying parameters are marked with a green circle. (B) We grew 100 populations in the same way as in (A) and show the resulting histogram of the inferred values of q (left, in blue), and μ (right, in red) based on the likelihood (5). The underlying values $q = 3/4$ and $\mu = 1$ are marked with blue and red vertical lines, respectively.

parameters given the observed tree (topology and number of mutations along branches)

$$\mathcal{L}(q, \mu) = \int_0^\infty \mathrm{d}\tau_s 2P(\nu'|\tau_s)P(\nu''|\tau_s) \ , \tag{5}$$

where a uniform prior has been used for the time of origin of the tree (the time at which ν' and ν'' arise).

To test this inference scheme, we generate a phylogenetic tree from a population grown from a single individual to size $N = 300$ (stopping when reaching this size) at $q = 3/4$ and $\mu = 1$. Figure 3B shows the likelihood landscape defined by (5). The integrals in (5) were performed numerically using the trapezoidal rule on a grid dividing the interval $[0, 2\log(N/\rho)/(1-q)]$ into 1000 subintervals. The upper limit of the sums over generations i was set to 50. The computation of a single likelihood took about 10 to 15 seconds for a tree of $N = 300$ leaves (4GHz clock speed) and can be sped up by using fewer subintervals for the numerical integration. The running time scales as the second power of the number of subintervals and linearly with the upper limit on the number of generations.

Figure 3C shows the histogram of parameters inferred from 100 different populations grown in the same way. As expected, the inferred values fluctuate from run to run around mean values close to the underlying values. Supplementary Information F shows that a wide range of parameters can be recovered in this way and explores how the inference improves with increasing tree size. We note that to exactly simulate the statistics conditioned on a given population size, one would have to include in the simulation also instances where this size is reached for the second, third, ... time. This is subtly different from the choice here, where the population reaches the given size for the first time, see [13] for a discussion. For supercritical growth and large population sizes, we expect any differences between the two statistical ensembles to be minor, as fluctuations around the mean size are small.

2.2 Finite sampling

In most applications, a phylogenetic tree is not based on the entire extant population at the time of observation but on a subset of sampled individuals or cells. Reconstructing the phylogenetic tree from a subset of individuals changes both the statistics of branch lengths [14] and the number of generations along branches. Hence, sampling must be accounted for in the likelihood function.

In this section, we follow a standard model of sampling [15, 14, 16], where extant individuals are selected independently at the time of observation with equal probability ρ. The key property of a phylogenetic branch then becomes that all extra offspring arising in birth events along that branch either have their clade die out by the time of observation (as before), or, alternatively, no extant element of that clade is sampled.

The probabilities that a birth-death process starting with a single individual at time zero has no *sampled* offspring or a single sampled offspring, respectively, have been calculated by Yang and Rannala [15] as

$$\begin{aligned} p_0^{(\rho)}(\tau) &= 1 - \frac{\rho(1-q)}{\rho + (1-\rho-q)\,\mathrm{e}^{-(1-q)\tau}} \\ p_1^{(\rho)}(\tau) &= \frac{\rho(1-q)^2\,\mathrm{e}^{-(1-q)\tau}}{(\rho + (1-\rho-q)\,\mathrm{e}^{-(1-q)\tau})^2}\ . \end{aligned} \tag{6}$$

These results (6) for the probability that a clade has no sampled offspring can be used to generalize the distribution $p_1(i|\tau_s, \tau_e)$ to the case of finite sampling, following the steps in Appendix B. The result is

$$p_1^{(\rho)}(i|\tau_s, \tau_e) = \mathrm{e}^{-(1+q)(\tau_s-\tau_e)}\,\frac{1}{i!}\left[2q(\tau_s-\tau_e) + 2\log\frac{(1-\rho-q)\,\mathrm{e}^{-(1-q)\tau_e} + \rho}{(1-\rho-q)\,\mathrm{e}^{-(1-q)\tau_s} + \rho}\right]^i\ . \tag{7}$$

Using this equation for the likelihood computation in Section 2.1 incorporates the effect of finite sampling into the inference scheme. In the equation for pendant branches (3) an additional factor of ρ describes the probability that the extant individual is sampled.

In applications of single-cell sequencing such as somatic evolution or evolution of a pathogen within a host, the population size and hence the sampling probability ρ is generally easier to estimate than for populations of individuals in the wild. However, in principle, ρ can also be inferred along with the other parameters q and μ, see Supplementary Information E.

A point that is not related to sampling but has a similar effect on the tree is sequencing noise. In particular in single-cell sequencing, some of the mutations found in the samples will be artifacts produced by sequencing errors. These also lengthen the pendant branches. By incorporating sequencing noise into the likelihood (5), the level of noise can be inferred from the tree. Details are in Supplementary Information G.

2.3 Application to single-cell data of haematopoietic stem cells

2.3.1 Prenatal haematopoiesis

As a first application, we use previously published whole-genome sequencing data of single haematopoietic stem cells (HSC) taken from the cord blood of neonates [10]. A tree based on sequencing data from 216 HSC from donor CB001 is depicted in Figure 4A with branch lengths corresponding to the number of mutations. We estimate the sampling probability to be $\rho = 0.01$, based on an order-of-magnitude estimate of 2×10^4 to 2×10^5 HSC in the donor blood of adults [10]. The maximum number of mutations along a branch is 62 found on a pendant branch. In general, pendant branches carry much larger numbers of mutations than the tree's interior, where most branches only carry a few mutations. This agrees qualitatively with the results of [17] for branch lengths under finite sampling.

HSC differentiate into different blood cells. We assume that differentiation occurs stochastically at some rate per unit time. The inferred relative death rate will therefore correspond to the sum of the rates of cell-death *and* differentiation relative to the rate with which cells replicate. The HSC population expands rapidly prior to birth, so we expect that the majority of mutations arises during cell replication [18].

We compute the likelihood (5) by integrating over the time interval $[0, 2\log(N/\rho)/(1-q)]$ discretized in 1000 steps and setting an upper limit of 200 for the sum over i. The results are shown for different values of the parameters in Figure 4B. The maximum of the likelihood is found to be near $q = 0.56$ and $\mu = 1.68$. Given that on average interior branches carry about three mutations each, this corresponds to about two birth events per branch, (one of them hidden).

We also inferred the parameters for the dataset of a second neonate (donor CB002 in [10]), which contained 390 HSC as well as haematopoietic progenitor cells (HPC). Under the assumption that HPC behave like HSC and that the sampling probability was again $\rho = 0.01$, the maximum lies at approximately $q = 0.64$ and $\mu = 1.45$. For comparison, we also inferred the model parameters under the assumption of a constant-rate molecular clock. For both data sets this leads to a maximum of the likelihood at $q = 0$. This is unexpected and unrealistic since differentiation leads to a non-zero value of q.

The parameter estimates only depend weakly on the sampling probability ρ. For instance setting $\rho = 0.001$ leads to maximum likelihood parameters of $q = 0.52$ and $\mu = 1.24$ for donor CB001 and of $q = 0.61$ and $\mu = 1.03$ for donor CB002.

These results are derived under the assumption of a constant-rate birth-death process (with death events corresponding to either cell death or differentiation). In the following section, we infer clade-specific evolutionary parameters in order to identify signals of selection.

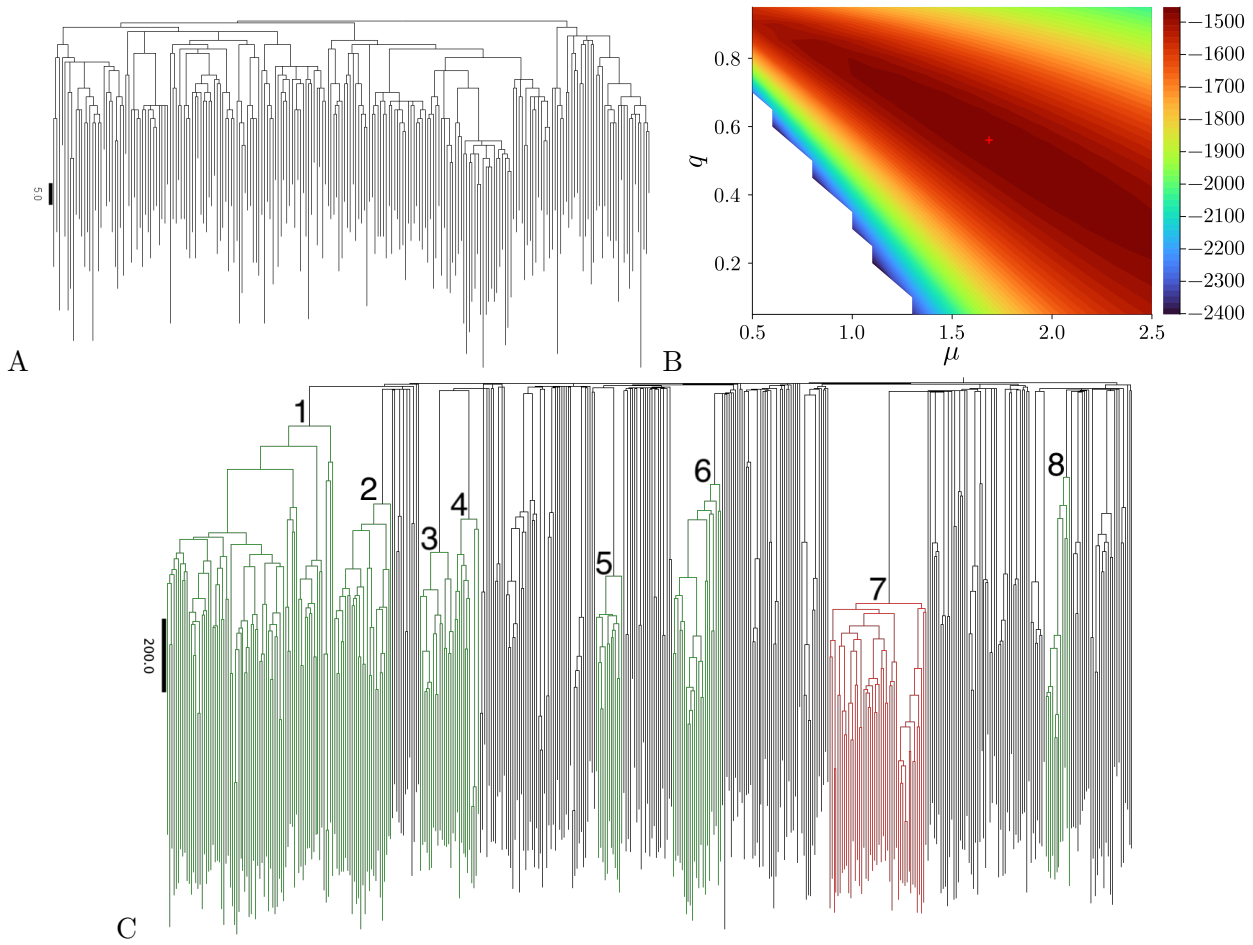

Figure 4: Likelihood of different trees from data. (A) The reconstructed tree from single-cell data of HSC in donor CB001 from [10] with branch lengths proportional to the number of mutations along them. The phylogeny is taken from [10]. The length of the black bar on the left corresponds to five mutations. (B) The resulting log-likelihood landscape of this tree when setting the sampling probability $\rho = 0.01$. Values of the log-likelihood below -2400 are left blank. The maximum (marked by a red cross) is found at around $q = 0.52$ and $\mu = 1.24$. (C) The phylogenetic tree of donor KX004. We find strong evidence that the clade shown in red has evolved under different evolutionary parameters than the other (green) clades, see text.

2.3.2 Heterogeneouty between different clades in adult haematopoiesis

Another phylogenetic tree constructed in [10] was based based on 451 HSC and HPC sampled from bone marrow of 77 year old donor KX004. We aim to detect diversity in the dynamics of different clades in this phylogenetic tree. As we expect a drastic difference between cell dynamics in an embryo and an adult, we will reduce our analysis to clades which have arisen after birth. Based on the neonate trees analyzed in the previous section, we assume that nodes with a distance of more than 100 mutations from the root have arisen after birth of the donor. We further restrict our analysis to clades with at least 23 leaves (i.e. 45 nodes in total). The tree of donor KX004 contains eight clades fulfilling these two criteria, see Figure 4C. We analyze each of these clades independently (thus removing any correlations between the times at which the eight clades arise), as explained above. In SI H we show how to identify, in a tree growing otherwise at uniform rates, a clade under selection by allowing the evolutionary parameters to change at a point within the tree.

To compute the likelihood of a single set of global parameters (q, μ) (i.e. the same parameters in all eight clades), we multiply the probabilities (5) for each clade, resulting in the joint probability of the observed clades. Assuming a sampling probability of $\rho = 0.01$ and using the same discretizations and intervals as in the previous Section 2.3.1, we find the likelihood maximum around $q = 0.91$ and $\mu = 10.81$.

The inference from multiple clades allows for an easy detection of differences between the evolutionary dynamics in the different clades. In principle, each clade can be assigned a separate set of variables. As the optimization of such a high-dimensional problem is not feasible, we consider distinct parameters q_{dist} and μ_{dist} in just one of the eight clades. All other clades follow the model with parameters q_{main} and μ_{main}. We assume a homogeneous sampling probability ρ throughout.

We have computed maximum likelihood estimates for eight different heterogeneous models - each of them proposing distinct parameters to a different clade. The results are shown in Table 1. We also show the likelihood ratio

$$\lambda_{LR} = -2\left(\max\left[\log \mathcal{L}(q, \mu)\right] - \max\left[\log \mathcal{L}(q_{\text{main}}, \mu_{\text{main}}, q_{\text{dist}}, \mu_{\text{dist}})\right]\right) \tag{8}$$

defined as the logarithm of the ratio of the maxima of the homogeneous and heterogeneous model, multiplied by a factor of two. In the limit of infinite data, the likelihood ratio follows a χ-squared distribution (in our case with two degrees of freedom, as this is the number of additional parameters in the heterogeneous model)[19]. Thus, the χ-squared distribution provides an estimate of the p-value of each heterogeneous model, i.e. the probability of falsely rejecting the homogeneous model in favour of the heterogeneous model. We set a significance threshold of $\alpha = 0.05$ to reject the homogeneous model in favour of a heterogeneous model. As we are testing eight different heterogeneous models, only the ones which have a p-value below the Bonferroni corrected significance value $\alpha/8 = 0.00625$ are accepted. Only the model with distinct parameters in clade 7 fulfils this criterion. We conclude that clade 7, shown in red in Figure 4C, has a different dynamics compared

to the rest of the tree, with a lower relative death rate and more average mutations per birth event than in the other subtrees.

Table 1: Maximum likelihood estimates of the parameters in eight of the clades in the phylogenetic tree from donor KX004. In each column, we consider a parameter change in one of the clades.

distinct clade:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
λ_{LR}	6.17	1.20	2.27	1.45	1.74	0.84	16.23	2.17
p-value	0.046	0.549	0.321	0.484	0.419	0.657	0.0003	0.338
q_{main}	0.90	0.91	0.91	0.91	0.90	0.91	0.91	0.91
μ_{main}	10.76	11.1	10.95	10.73	10.98	10.70	10.76	10.60
q_{sub}	0.91	0.92	0.90	0.87	0.89	0.91	0.80	0.85
μ_{sub}	11.18	10.02	9.42	14.39	10.16	11.60	14.45	15.88

3 Discussion

Under cellular reproduction, mutations can arise at individual birth events rather than accumulating continuously over time. We have built a population dynamic framework for the phylodynamic analysis under cellular reproduction. Our approach is geared towards large-scale single-cell data: we use dynamic programming to iteratively sum over the unknown bifurcation times and the number of generations along a branch. This leads to a likelihood function that can be computed in seconds for several hundred samples and scales linearly with the number of samples.

Cellular reproduction generates phylogenetic trees which differ in two respects from their counterparts created in macroevolution. First, when mutations arise at discrete reproductive events, the number of mutations accumulated along a branch follows a compound Poisson distribution. This leads to a larger variance compared to the Poisson distribution observed when mutations arise continuously over time. Second, the mean number of mutations along a branch is not proportional to the branch length in calendar time, even when the parameters of the birth-death process are constant over time. Instead, the mean number of mutations depends non-linearly on the age of a particular branch and the birth and death rates of cells. This is because additional lineages established by birth events along a branch must die out by the time the population is sampled; otherwise, the branch terminates and two new branches begin. This condition is harder to fulfil in recent branches, reducing the number of generations per time in recent branches [11]. As a result, generations accumulate in branches at different rates in different parts of the *reconstructed* tree, even when all underlying parameters are constant through time. This can lead to errors in phylodynamic analysis whenever neutral evolution under cellular reproduction is modelled by a constant-rate molecular clock (see Supplementary Information D).

Throughout, we use the number of mutations along the branches of a tree as a basis for the likelihood. This approach can easily be extended to sequence similarity scores to quantify the rates of different types of mutations. We have also used the infinite sites assumption (no back-mutations), so the tree topology and the characters of internal nodes (ancestral sequences) can be determined

unambiguously. If this assumption breaks down, at large evolutionary times or at large mutation rates, one needs to sum both over the sequences at internal nodes [12] as well as the branching times. This leads to a computationally harder problem, requiring either Markov chain sampling [5, 20], or heuristics like the *TreeTime* algorithm that alternately sums over ancestral states and branch lengths [6]. However, despite the additional computational difficulty, the statistical framework leading to the likelihood (5) generalizes directly to the regime of long evolutionary times.

Our approach differs from coalescent-based phylodynamic inference [21], which tracks bifurcations in the tree backwards in time. The coalescent statistics depends only on the net growth rate of the population (difference between birth and death rates). However, the hidden birth events are closely related to cell death, so the coalescent cannot capture the statistics of hidden birth events. Our approach also differs from the recent analysis of punctuated evolution [22, 20], which uses Monte Carlo Markov sampling, whereas we iteratively sum over the generations along branches/integrate over branch lengths. Punctuated evolution is a macroevolutionary model characterized by bursts of mutations at speciation events in addition to a regular molecular clock. Formally, mutations occurring at birth events can be viewed as an extreme case of punctuated evolution. Our approach also differs from inference based on the distribution of pairwise distances between the tree leaves [23] in that we compute the likelihood of the tree and use the exact statistics of hidden birth events arising from the birth-death process, see also [24].

In this paper, we have focussed on the statistics of mutations that arise at cell divisions. Even in somatic evolution, mutations can arise per unit time: DNA lesions which form spontaneously might not be repaired, and thus accumulate over time. These processes have been analyzed in bulk data on the basis of the different mutational signatures they carry [18, 25]. Under a low rate of cell division, with few replications in a given time interval, more mutations arise due to lesions which accumulate over time, whereas in fast-dividing populations, mutations arise predominantly as part of the reproductive process [18]. In Supplementary Information G we have extended our inference scheme to mixtures of mutations arising at cell divisions and mutations arising per unit time. Inferring the relative weight of these two mechanisms from data is an exiting prospect and will become feasible when larger whole-genome single-cell phylogenies become available.

Code availability: Code is available under MIT license at
`https://github.com/t-dslhrst/SingleCellPhylodynamicInference`.

Acknowledgments This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) grant SFB1310/2 - 325931972. Many thanks to Tibor Antal, Samuel Johnston, Joachim Krug, Martin Peifer, and Shamil Sunya'ev for discussions, and to Tanja Stadler for comments throughout and especially for helpful advice on Section 2.2. We thank Nico Beerenwinkel, Jamie Blundell, Nico Borgsmüller, Donate Weghorn, and Xu Xun for advice on single-cell data. We thank the computing centre of the University of Cologne RRZK for computing time on the DFG-funded clusters CHEOPS (grant number INST 216/512/1FUGG) and RAMSES (grant

number INST 216/512-1 FUGG).

References

[1] Nee, S., May, R. M. & Harvey, P. H. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **344**, 305–311 (1994).

[2] Maddison, W. P., Midford, P. E. & Otto, S. P. Estimating a binary character's effect on speciation and extinction. *Systematic Biology* **56**, 701–710 (2007).

[3] Stadler, T. Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology* **26**, 1203–1219 (2013).

[4] MacPherson, A., Louca, S., McLaughlin, A., Joy, J. B. & Pennell, M. W. Unifying phylogenetic birth–death models in epidemiology and macroevolution. *Systematic Biology* **71**, 172–189 (2022).

[5] Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* **15**, e1006650 (2019).

[6] Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution* **4**, vex042 (2018).

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

[8] Takahata, N. On the overdispersed molecular clock. *Genetics* **116**, 169–179 (1987).

[9] Huelsenbeck, J. P., Larget, B. & Swofford, D. A compound poisson process for relaxing the molecular clock. *Genetics* **154**, 1879–1892 (2000).

[10] Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022). URL `https://www.nature.com/articles/s41586-022-04786-y`. Publisher: Nature Publishing Group.

[11] Bokma, F., van den Brink, V. & Stadler, T. Unexpectedly many extinct hominins. *Evolution* **66**, 2969–2974 (2012).

[12] Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376 (1981).

[13] Hartmann, K., Wong, D. & Stadler, T. Sampling trees from evolutionary models. *Systematic Biology* **59**, 465–476 (2010).

[14] Stadler, T. Sampling-through-time in birth–death trees. *Journal of Theoretical Biology* **267**, 396–404 (2010).

[15] Yang, Z. & Rannala, B. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Molecular Biology and Evolution* **14**, 717–724 (1997).

[16] Stadler, T. & Steel, M. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. *Journal of Theoretical Biology* **297**, 33–40 (2012).

[17] Dieselhorst, T. & Berg, J. Branch length statistics in phylogenetic trees under constant-rate birth-death dynamics (2024). URL `http://arxiv.org/abs/2407.13403`. ArXiv:2407.13403 [q-bio].

[18] Spisak, N., de Manuel, M., Milligan, W., Sella, G. & Przeworski, M. Disentangling sources of clock-like mutations in germline and soma. *bioRxiv* (2023).

[19] Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses **9**, 60–62. URL `https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-9/issue-1/The-Large-Sample-Distribution-of-the-Likelihood-Ratio-for-Testing/10.1214/aoms/1177732360.full`. Publisher: Institute of Mathematical Statistics.

[20] Douglas, J., Bouckaert, R., Harris, S. C., Carter Jr., C. W. & Wills, P. R. Evolution is coupled with branching across many granularities of life. *Proceedings of the Royal Society B: Biological Sciences* **292**, 20250182 (2025). URL `https://royalsocietypublishing.org/doi/full/10.1098/rspb.2025.0182`. Publisher: Royal Society.

[21] Johnson, B., Shuai, Y., Schweinsberg, J. & Curtius, K. clonerate: fast estimation of single-cell clonal dynamics using coalescent theory. *Bioinformatics* **39**, btad561 (2023).

[22] Manceau, M., Marin, J., Morlon, H. & Lambert, A. Model-based inference of punctuated molecular evolution. *Molecular Biology and Evolution* **37**, 3308–3323 (2020).

[23] Werner, B. *et al.* Measuring single cell divisions in human tissues from multi-region sequencing data. *Nature Communications* **11**, 1–9 (2020).

[24] Angaji, A. *et al.* High-density sampling reveals volume growth in human tumours. *eLife* **13** (2024). URL `https://elifesciences.org/reviewed-preprints/95338`.

[25] Satake, A. *et al.* Somatic mutation rates scale with time not growth rate in long-lived tropical trees. *bioRxiv* 2023–01 (2023).

[26] Kendall, D. G. On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics* **19**, 1–15 (1948).

[27] Ross, S. M. *Introduction to probability models* (Elsevier, Amsterdam, 2014).

[28] Blitzstein, J. K. & Hwang, J. *Introduction to probability.* Texts in statistical science (CRC Press Taylor & Francis group, Boca Raton, 2019), 2nd ed edn.

[29] Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (John Wiley & Sons, 2006).

A Stochastic dynamics of a birth-death process

We model the dynamics of a population as a birth-death process with birth rate β and death rate δ per individual and per unit time. A particular clade (one of many) begins to grow from a single individual born at time t_0 until time T, when the population is observed. We limit ourselves to the case of a growing population with $\beta > \delta$.

From a population sampled at one given time, no quantity with units of time can be inferred. A goal will be to determine the relative rate of death $q = \delta/\beta$, and without loss of generality, we set the unit of time such that the birth rate is $\beta = 1$.

The standard approach to such a birth-death process is a master equation for the size of a clade n at time t described by the probability $p_n(t)$ evolving under

$$\partial_t p_n(t) = \beta(n-1)p_{n-1}(t) - \beta n p_n(t) + \delta(n+1)p_{n+1}(t) - \delta n p_n(t) \ . \tag{9}$$

As initial conditions, we will assume the clade to start with a single individual at time $t = 0$, i.e. $p_1(t = 0) = 1$ and $p_n(t = 0) = 0$ for all $n > 1$. This equation can easily be solved using a generating function. A classic result by Kendall [26] gives the probability that the clade has died out by time T

$$p_0(T) = q\frac{1 - \mathrm{e}^{-(1-q)T}}{1 - q\,\mathrm{e}^{-(1-q)T}} \tag{10}$$

and the probability that there is exactly one surviving offspring by time T

$$p_1(T) = \frac{(1-q)^2\,\mathrm{e}^{-(1-q)T}}{(1 - q\,\mathrm{e}^{-(1-q)T})^2} \ , \tag{11}$$

These results can be used directly to determine the statistics of branch lengths and the number of generations (birth events which have not led to an extant clade) along a branch of the phylogenetic tree. Of particular interest for the inference scheme presented in Section 2 is the joint probability of a branch and the number of generations along it, which we derive in the following.

B Joint probability of size-one clades and the number of generations

We measure time relative to the time of observation T and define $\tau = T - t$. It is zero when the population is observed and increases as we look further into the evolutionary past.

We now derive the joint probability that a clade starting with a single individual at time τ_s in the past undergoes i birth events by the time $\tau_e < \tau_s$, all of which do not lead to additional extant clades at the time of observation T. This extends the approach in [17], where we derived the same distribution but without considering the number of birth events. The birth events along a branch

are the generations of a lineage along that branch. Denoting the joint probability as $p_1(i|\tau_s,\tau_s)$ and using the probabilities (10)-(11), its dynamics are described by the master equation

$$-\partial_{\tau_e} p_1(i|\tau_s,\tau_e) = -q p_1(i|\tau_s,\tau_e) - p_1(i|\tau_s,\tau_e) + 2p_0(\tau_e) p_1(i-1|\tau_s,\tau_e) \ . \tag{12}$$

The first term accounts for the death of the individual, the second term for the probability of a birth event (either leading to an additional observed lineage or to a $i+1$ hidden birth event). The last term adds the probability that the i-th birth event occurs and becomes 'hidden' in the phylogenetic tree, as one of the two descending clades dies out before observation. Defining the generating function $G(\mu,\tau_s,\tau_e) = \sum_{i=0}^{\infty} \mu^i p_1(i|\tau_s,\tau_e)$ yields

$$-\partial_{\tau_e} G(\mu,\tau_s,\tau_e) = -\left(q + 1 - 2\mu p_0(\tau_e)\right) G(\mu,t) \ , \tag{13}$$

which is easily solved by integration.

In the case of pendant branches, we integrate from τ_s to $\tau_e = 0$ and get

$$G(\mu,\tau_s,\tau_e=0) = \exp\left\{-(1+q)\tau_s + \mu\left[2q\tau_s + 2\log\left(\frac{1-q}{1-q\,\mathrm{e}^{-(1-q)\tau_s}}\right)\right]\right\} \ . \tag{14}$$

The resulting probability of a clade of given age τ_s having a single member at the time of observation after i generations is

$$\begin{aligned} p_1(i|\tau_s,\tau_e=0) &= \frac{1}{i!}\partial_\mu^i|_{\mu=0} G(\mu,\tau_s,\tau_e=0) \\ &= \mathrm{e}^{-(1+q)\tau_s}\,\frac{1}{i!}\left[2q\tau_s + 2\log\left(\frac{1-q}{1-q\,\mathrm{e}^{-(1-q)\tau_s}}\right)\right]^i \ . \end{aligned} \tag{15}$$

Summing over the number of generations results in equation (11).

Integrating the equation of motion of the generating function (13) only from τ_s to some intermediate $\tau_e > 0$, we find the joint probability of an individual having i birth events in this time interval and all descendants of these births dying out by the time of observation

$$p_1(i|\tau_s,\tau_e) = \mathrm{e}^{-(1+q)(\tau_s-\tau_e)}\,\frac{1}{i!}\left[2q\tau + 2\log\left(\frac{1-q\,\mathrm{e}^{-(1-q)\tau_e}}{1-q\,\mathrm{e}^{-(1-q)\tau_s}}\right)\right]^i \ . \tag{16}$$

For a given branch starting at τ_s and ending at τ_e, this result implies that the number of generations is Poisson distributed with mean

$$\langle i\rangle(\tau_s,\tau_e) = 2q(\tau_s-\tau_e) + 2\log\left(\frac{1-q\,\mathrm{e}^{-(1-q)\tau_e}}{1-q\,\mathrm{e}^{-(1-q)\tau_s}}\right) \ . \tag{17}$$

This agrees with the result derived by Bokma, van den Brink, and Stadler [11] for the mean number of generations. The first term in (17) is proportional to the elapsed time $\tau = \tau_s - \tau_e$; the constant

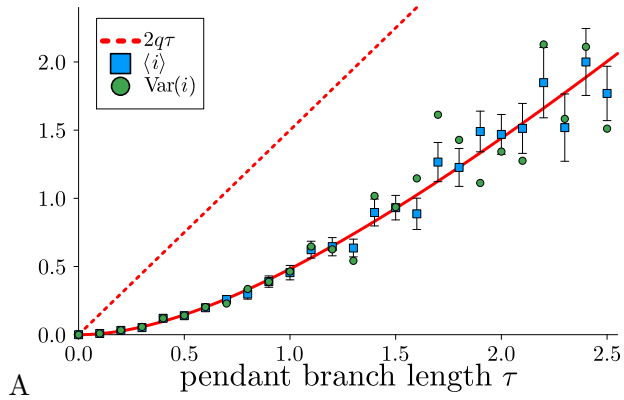

Figure 5: Distribution of the number of generations along pendant branches. Numerical results come from a single population grown to size 7500 at relative death rate $q = 3/4$. We pick all pendant branches of a given length τ (± 0.05) and plot the mean (large blue symbols, with the standard error given by the error bars) and variance (small green symbols) of the number of generations along these branches against τ. The full red line gives the theoretical result (17) (with $\tau_e = 0$) for both the mean and the variance. The dotted red line gives the linear term $2q\tau$ in (17), corresponding to generations appearing along the branch at a constant rate.

of proportionality is twice the relative rate of cell death. Without death events, every birth event would produce two extant lineages, thus producing an interior node in the tree. As a result, without cell death ($q = 0$), the number of generations along each branch is zero. The second, nonlinear term is zero at $\tau = 0$ and grows with increasing branch length. This means that the mean number of generations along a branch is not proportional to the length of the branch calendar time. This counterintuitive result has a simple explanation: all extra lineages arising by birth processes along a branch must die out by the time the population is sampled, and this happens with a smaller probability at short τ, reducing the number of generations for short branches. The effect decreases with increasing τ_e.

Figure 5 compares these results to numerical simulations. A population with relative death rate $q = 3/4$ is grown from a single individual to 7500 individuals (stopping the simulation when this population size is reached) and the phylogenetic tree is reconstructed. Figure 5 shows the mean and variance of the number of generations along pendant branches of a given length τ. Crucially, the mean number of generations is clearly not proportional to the branch length τ, but follows (17) for $\tau_e = 0$. (For a Poisson distribution, the variance equals the mean.)

Summing over the number of generations i in distribution (16) gives the distribution of branch lengths $p_1^{(\tau_e)}(\tau_s - \tau_e)$ [17]. Equation 16 can also be derived by multiplying $p_1^{(\tau_e)}(\tau_s - \tau_e)$ with the result from [11] that the number of generations is Poisson distributed with mean (17).

These results can be easily adapted to finite sampling as discussed in Section 2.2. $p_0(\tau_e)$ in the equation of motion (13) for the generating function needs to be replaced by $p_0^{(\rho)}(\tau_e)$ given by (6).

The result is (16) but with the term in square brackets replaced by

$$\langle i \rangle^{(\rho)}(\tau_s, \tau_e) = 2q(\tau_s - \tau_e) + 2\log \frac{(1-\rho-q)\,\mathrm{e}^{-(1-q)\tau_e} + \rho}{(1-\rho-q)\,\mathrm{e}^{-(1-q)\tau_s} + \rho} \,, \tag{18}$$

which is the mean number of generations on a branch running from τ_s to τ_e generalized to finite sampling. In the iterative algorithm of Section 2.1, each leaf of the tree is assigned an additional factor of ρ accounting for sampling at that probability.

C Distribution of the number of mutations along a branch

Under a constant molecular clock hypothesis, the number of mutations along a branch of a given length is Poisson distributed with a mean given by the branch length (in calendar time) multiplied by a mutation rate per time.

However if, under cellular reproduction, mutations occur at cell divisions, different statistics of the number of mutations along a branch arise. As we have seen in the previous section, the number of generations along a given branch follows a Poisson distribution. If the number of mutations per cell division is itself a random variable this makes the total number of mutations along a branch follow a compound Poisson distribution: The number of accumulated mutations m is the sum over i independent, identically distributed random variables $\hat{m}$, where the number of generations i itself is a Poisson distributed random variable.

Denoting with i the random variable of the number of generations along a branch from τ_s to τ_e and with $\hat{m}$ the (also random) number of mutations at a birth event, we can use Wald's equation [27] to compute the mean number of mutations between bifurcations as

$$\langle m \rangle = \langle i \rangle \langle \hat{m} \rangle \,. \tag{19}$$

For simplicity, we here suppress the dependencies of i and m on τ_s and τ_e. Using the law of total variance (also known as Eve's law, [28]), we find the variance to be

$$\begin{aligned} \mathrm{Var}\,[m] &= \langle \mathrm{Var}\,[m|i] \rangle + \mathrm{Var}\,[\langle m|i \rangle] \\ &= \mathrm{Var}\,[\hat{m}]\,\langle i \rangle + \langle \hat{m} \rangle^2 \mathrm{Var}\,[i] \\ &= \left(\mathrm{Var}\,[\hat{m}] + \langle \hat{m} \rangle^2\right) \langle i \rangle \,. \end{aligned} \tag{20}$$

When we take the number of mutations per birth event to be Poisson distributed with $\langle \hat{m} \rangle = \mathrm{Var}\,[\hat{m}] = \mu$, we obtain

$$\mathrm{Var}\,[m] = \langle i \rangle \mu (1+\mu) \tag{21}$$

which differs by a factor of $(1+\mu)$ from the corresponding result for Poisson-distributed accumulated

mutations. The probability mass function for the mutations along a given branch is

$$P(m|\tau_s,\tau_e) = \sum_{i=0}^{\infty} P(i|\tau_s,\tau_e)P(m|i) \tag{22}$$

$$= \sum_{i=0}^{\infty} \mathrm{e}^{-\langle i\rangle(\tau_s,\tau_e)} \frac{\langle i\rangle(\tau_s,\tau_e)^i}{i!} \mathrm{e}^{-i\mu} \frac{(i\mu)^m}{m!} \ . \tag{23}$$

$P(i|\tau_s,\tau_e)$ denotes the Poisson distribution for the number of generations along a given branch from τ_s to τ_e.

Note that here we focus on the mutations accumulating *between* two tree bifurcations and have not counted the mutations arising at the cell division establishing the branch. For this reason, $P(m|i)$ is a regular Poisson distribution, whereas (2) features an additional generation at the top of the branch. (For this reason, in the computation of the likelihood as described in Section 2, we used the compound Poisson distributed mutations along the branch plus a Poisson distributed number of mutations from the birth event starting the branch.)

Two examples of the probability mass function (22) are shown in Figure 2. On the left, we show the distribution for $\langle i\rangle = 10$ and $\mu = 1$ and on the right for $\langle i\rangle = 3$ and $\mu = 10$ (red lines). The Poisson distributions with the same mean $\langle i\rangle\mu$ are depicted by grey dashed lines.

D Comparison of models with different mutation statistics

By using the statistics of the number of generations along branches, our approach incorporates two differences to strict molecular clock models: The mean number of generations along branches is inhomogeneous throughout the tree (as discussed in Supplementary Information B), which affects the statistics of mutations if they are tied to birth events. Furthermore, mutations from birth events along branches follow compound Poisson statistics rather than a Poisson distribution (Supplementary Information C).

In order to compare approaches based on different assumptions on the mutational process, we have implemented the same likelihood-based inference scheme introduced above but with the assumption that mutations occur at some rate per time. This yields a Poisson distribution for the number of mutations along a given branch. We do this in two different ways. In a first approach, we take the number of mutations along a branch to be Poisson distributed, with a mean depending on the number of generations. Specifically, the mean number of mutations along a branch running from τ_s and τ_e is taken to be $(1 + \langle i\rangle(\tau_s,\tau_e)) \times \mu$, i.e. the expectation value of the number birth events contributing mutations to the branch (one birth event forming the branch plus the mean number of generations along the branch $\langle i\rangle(\tau_s,\tau_e)$) times the mean number of mutations per birth event μ. This replaces the compound Poisson statistics (see Supplementary Information C) by a Poisson distribution with the same mean. In a second approach, we assume a constant molecular clock with rate μ_τ throughout the entire tree. In both approaches, the bifurcation times and number of

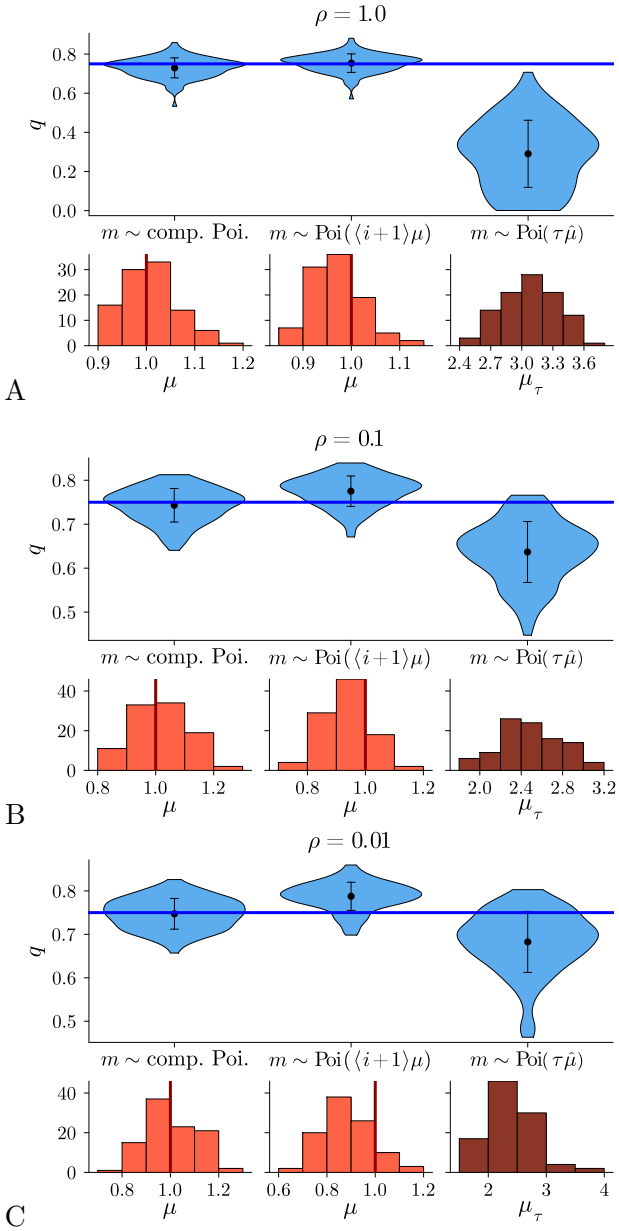

Figure 6: Comparison of different models of mutations. We compare results of our method based on the compound Poisson statistics of mutations along branches ("$m \sim$ comp. Poi.") and the same algorithm but with Poisson statistics (inhomogeneous "$m \sim \mathrm{Poi}(\langle i+1 \rangle \mu)$" and a constant-rate molecular clock "$m \sim \mathrm{Poi}(\tau \mu_\tau)$", see text). (A) Distributions of the inferred parameters from 100 populations grown to size 300. Violin plots show the results for q, with the dots indicating the mean and the error bars the standard deviation. Histograms show the inferred mean number of mutations per cell division μ or mutation rate μ_τ. Solid lines show the underlying parameter values $q = 3/4$ and $\mu = 1$. (B) The same, but for sampling probability $\rho = 0.1$. (C) The same for $\rho = 0.01$.

generations along branches are integrated and summed over as discussed in Section 2.1.

To compare the inference results of these different models, we simulated trees with relative death rate $q = 3/4$ and mean mutations per birth event $\mu = 1$. In Figure 6A, we show the inferred parameters from 100 populations, where we stopped the simulations when a population size of $N = 300$ was reached. For Figure 6B (and C), we stopped when the population reached size 3×10^3 (3×10^4) and reconstructed the tree from 300 randomly chosen individuals (corresponding to sampling probability $\rho = 0.1$ or $\rho = 0.01$, respectively). Violin plots show the distributions of the inferred values of q for the different models of the number of mutations along each branch. On the left of each subfigure, we show the inference results for the original likelihood (5) based on the compound Poisson statistics, in the centre we use the Poisson statistics with the same mean, and on the right the constant-rate molecular clock. The three histograms below depict the corresponding distributions of the inferred mutation parameters μ and μ_τ.

In the case of complete sampling, we find that the first approach (Poisson statistics) gives results comparable to the compound Poisson statistics. However, in the cases of $\rho = 0.1$ and $\rho = 0.01$, a noticeable bias is seen in the results based on Poisson statistics. This is because in the fully sampled case, many (in particular pendant) branches have no hidden generations along the branch. The mutations solely come from the birth event establishing the branch and the number of mutations is thus Poisson distributed. When only a fraction of the extant distribution is sampled, the number of hidden birth events (in particular along pendant) branches increases, and thus, the mutations follow the compound Poisson rather than Poisson statistics.

While the inferred values of q using the molecular clock model are far off in all three cases, a significant improvement is seen when comparing the results under finite sampling to the ones from fully observed populations (note the different y-axes). This is because the inhomogeneity in the mean number of generations across the tree becomes weaker as we look at earlier times reaching an asymptotic regime, see Section A. With decreasing sampling probability, the effects of this inhomogeneity decrease as more branches lie further away from the time of observation. Therefore, a small improvement is also visible in the results for $\rho = 0.01$ compared to $\rho = 0.1$. However, even under low sampling, the results under strict molecular clock assumption are highly inaccurate and therefore unreliable.

E Inference of the sampling probability

In Section 2.2, we discussed the inference of the sampling probability ρ together with the other parameters q and μ. We illustrate this inference by simulating populations with $q = 3/4$ and $\mu = 1$, stopping when the population size reaches 3000. We then reconstruct the phylogenetic trees based on 300 randomly chosen extant individuals, which corresponds to a sampling probability of $\rho = 0.1$. We created 100 trees in this manner and infer the corresponding parameters. Figure 7A shows the inference results for q and μ at given $\rho = 0.1$, while Figure 7B depicts the results of the simultaneous inference of all three parameters. As expected, q and μ are more accurately inferred

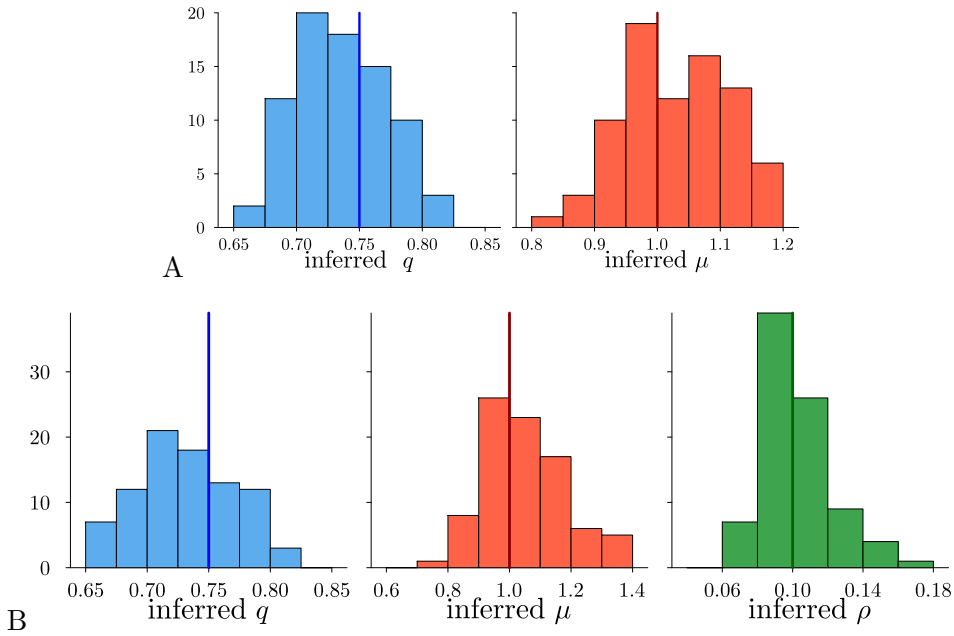

Figure 7: Parameter inference from samples of a population. Depicted are the inference results from 100 populations simulated with $q = 3/4$ and $\mu = 1.0$ up to size 3000. For each population, a tree was reconstructed from 300 samples, corresponding to a sampling probability of $\rho = 0.1$ (see text). (A) Inference results when the value of $\rho = 0.1$ is known. (B) Inference results when ρ is inferred along with the other parameters.

when ρ is known, however, it is still possible to estimate all three parameters.

F Inference bias

In Fig. 3 of the main text we showed how model parameters can be inferred from simulated data for a specific choice of model parameters. Here we repeat the inference from artificial data trees grown to size $N = 300$ across a wide range of model parameters q and μ. The resulting inference error of the model parameters remains small and nearly flat over a wide range of model parameters, see Figure 8.

Maximum likelihood estimates generally have a bias in the parameter estimates [29], which vanishes in the limit of a large data. Figure 9 shows how the bias in the parameter estimates from the likelihood (5) decreases with the number of sampled individuals (equal to the population size at full sampling, $\rho = 1$). This bias can be reduced by conditioning the population e.g. on not dying out. Such conditions are applied by replacing the uniform prior in the integration over the time of the root node (5). We refer to [4] for a discussion on different conditioning scenarios.

G Combining different mutational processes

So far we have focussed on mutations that are caused by the reproduction of cells and arise at each cell division with a certain probability. Also in somatic evolution mutations can arise per unit time,

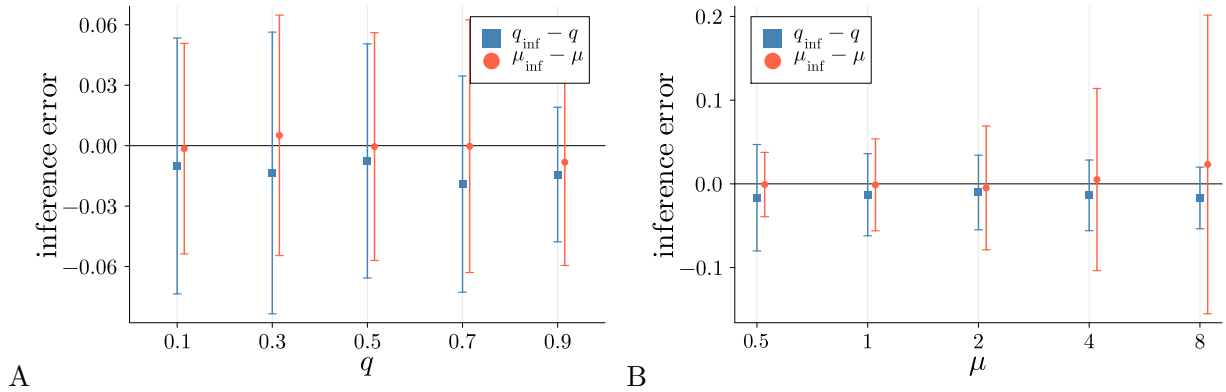

Figure 8: Inference error across different values of the model parameters. The difference between inferred and underlying values of the relative rate of death q and the number of mutations per birth μ is shown for different values of the underlying q and μ. The symbols (blue squares for q and orange circles for μ) show the differences between inferred and underlying values averaged over 100 runs each producing a tree of size $N = 300$ at full sampling ($\rho = 1$). Error bars show the corresponding standard deviation. The inference is based on maximizing the likelihood (5). For visibility, the symbols indicating the results for the mutation parameter μ are shifted a little to the right. (A) The mean number of mutations is kept fixed at $\mu = 1$, while the relative rate of death q varies as shown on the x-axis. (B) The relative rate of death is kept fixed at $q = 0.75$, while μ varies as shown on the x-axis.

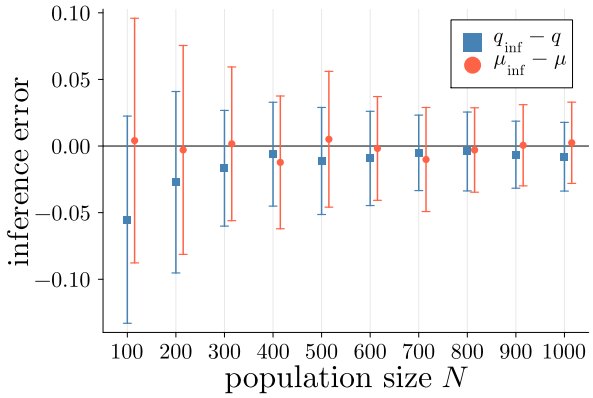

Figure 9: The dependence of the inference error on population size. The difference between inferred and underlying values of the relative rate of death q and the number of mutations per birth μ is shown for populations of different sizes. The symbols (blue squares for q and orange circles for μ) show the differences between inferred and underlying values averaged over 100 runs; error bars show the corresponding standard deviation. The underlying parameters were $q = 3/4$, $\mu = 1$ at full sampling ($\rho = 1$). As expected, the bias of the estimates decreases with the population size N, until reaching a plateau for populations of size $N = 400$ and larger.

or more precisely, DNA lesions are generated at a certain rate per unit time, and lead to mutations unless they have been repaired by the time the cell divides [18]. In this section, we extend our model of mutations to the case where mutations occur both at a constant rate μ_τ per unit time (i.e. following a molecular clock) and, in addition, each division contributes a number of mutations drawn from a Poisson distribution with mean μ as described in Section 2.1.

Mutations that occur at a fixed rate μ_τ per unit time lead to a Poisson distributed number of mutations with mean $\mu\tau$ along a branch of length τ. The total number of mutations arrise both over time and from i birth events is thus

$$P(m|i,\tau) = \exp(-\mu_\tau\tau - (i+1)\mu)(\mu_\tau\tau + (i+1)\mu)^m/m! \ . \tag{24}$$

Using this distribution in equations (3)-(5), we can compute the likelihood of μ_τ along with q and μ.

However, teasing apart the relative contributions of these two processes from a phylogenetic tree will require larger trees with more mutations along the branches than currently available and must be left for future work. Relying on the mutational signatures [18] to tell apart m_τ (the number of mutations generated a process running per unit time) and $m - m_\tau$ (the number of mutations generated by a process arising at cell divisions) would simplify the inference. We note that if the number of mutations in a birth event is not Poisson distributed, an additional sum arises when computing the distribution of mutations (24)

Finally, we look at sequencing noise, modelled by adding a random number m_e of sequencing errors (erroneous mutations) to the tree leaves. These numbers are drawn independently and identically from some distribution $P_e(m_e)$. Incorporating this model into the iterative construction of the likelihood leads in place of (3) to

$$P(\nu|\tau_s) = \sum_{m_e=0}^{m} P_e(m_e) \sum_{i=0}^{\infty} P(m - m_e|i) p_1(i|\tau_s, \tau_e = 0) \ . \tag{25}$$

A natural choice of the distribution $P_e(m_e)$ is, again, a Poisson distribution with mean μ_e to be inferred. In this case, similar to the case with mutations accumulating over time, the sum over m_e can be avoided. The statistics of interior branches is not affected by sequencing noise. Hence, the iterative procedure (4) for the interior branches remains unchanged, and the model parameters including μ_e can be inferred by maximizing the resulting likelihood.

H Inference of selection

The algorithm presented in Section 2.1 can easily be adapted to account for distinct parameters in different clades of the tree. This enables detecting mutations which change the evolutionary dynamics in the corresponding clade. In principle, each such parameter change leads to three additional parameters to infer: The birth and death rates of the distinct subtree, β_{sub} and δ_{sub}

respectively, in units of the main tree's birth rate and the mean number of mutations per generation μ_{sub} in the subtree. In the following, we will consider the case where there is only a single such parameter change, leading to a single distinct subtree.

Taking the parameter change to occur at a particular node in the tree (with both descending lineages following the new parameters), we can easily compute the probability of the distinct subtree using the iterative algorithm from Section 2.1, replacing μ in equation (2) with μ_{sub} and using

$$p_1(i|\tau_s,\tau_e) = \mathrm{e}^{-(\beta_{\text{sub}}-\delta_{\text{sub}})(\tau_s-\tau_e)}\,\frac{1}{i!}\left[2\delta_{\text{sub}}(\tau_s-\tau_e) - \log\frac{\beta_{\text{sub}}-\delta_{\text{sub}}\,\mathrm{e}^{-(\beta_{\text{sub}}-\delta_{\text{sub}})\tau_e}}{\beta_{\text{sub}}-\delta_{\text{sub}}\,\mathrm{e}^{-(\beta_{\text{sub}}-\delta_{\text{sub}})\tau_s}}\right]^i \tag{26}$$

in equations (3) and (4). All iterations over clades ν starting with branches which are not part of the distinct subtree are performed as before with parameters q and μ. The computation of $P(\nu|\tau_s)$ (equation (4)) for the branch above the parameter change also uses these main parameters but uses the results of the two descending clades $P(\nu'|\tau_s)$ and $P(\nu''|\tau_s)$ which were computed with the alternative parameters. Proceeding until the final integration over the root node (5), we obtain the probability of the entire tree with a distinct subtree, and hence the likelihood of parameters q, μ, β_{sub}, δ_{sub} and μ_{sub}.

Proposing parameter changes at different nodes of the tree and comparing the resulting maximum likelihoods allows finding the most likely location of a parameter change. To test this approach, we have simulated 100 trees in the following way: We simulate a population starting with a single individual and with parameters $q = 0.75$ and $\mu = 1$. The two lineages descending from the 200th birth event evolve under a higher birth rate $\beta_{\text{sub}} = 2$, but with otherwise identical parameters $\delta_{\text{sub}} = 0.75$ and $\mu_{\text{sub}} = 1$. The simulation is stopped when the population first reaches a size of $N = 300$ and the phylogenetic tree is reconstructed from all extant individuals. If the population has died out or either of the two distinct parts of the tree have less than 50 leaves, the simulation is repeated.

We then inferred parameters under the assumption of a single parameter change. This parameter change was proposed at every node where the descending subtree includes at least 25 leaves. For each of these proposed parameter changes, the likelihood was determined. We fixed the discretization of time to 1000 equidistant subintervals between 0 and $2\log(N/\rho)/(1-q_{\text{homo}})$, where q_{homo} is the maximum likelihood estimate of the relative death rate under a homogeneous parameter model. The upper limit of the sum over the number of generations was set to 50. Comparing the results of each proposed parameter change, we accepted the one with the highest likelihood maximum.

In order to assess the accuracy with which we have found the correct location of the parameter change, we define the dichotomy error as follows: If the inferred distinct subtree of size n_{inf} overlaps with the true distinct subtree of size n_{true}, the dichotomy error is given by $n_{\text{inf}} - n_{\text{true}}$. If the two do not overlap, the dichotomy error is given by the number of nodes which is in neither of the two subtrees. In our simulations of trees with 300 leaves (and therefore 599 nodes), this dichotomy error is given by $599 - n_{\text{inf}} - n_{\text{true}}$.

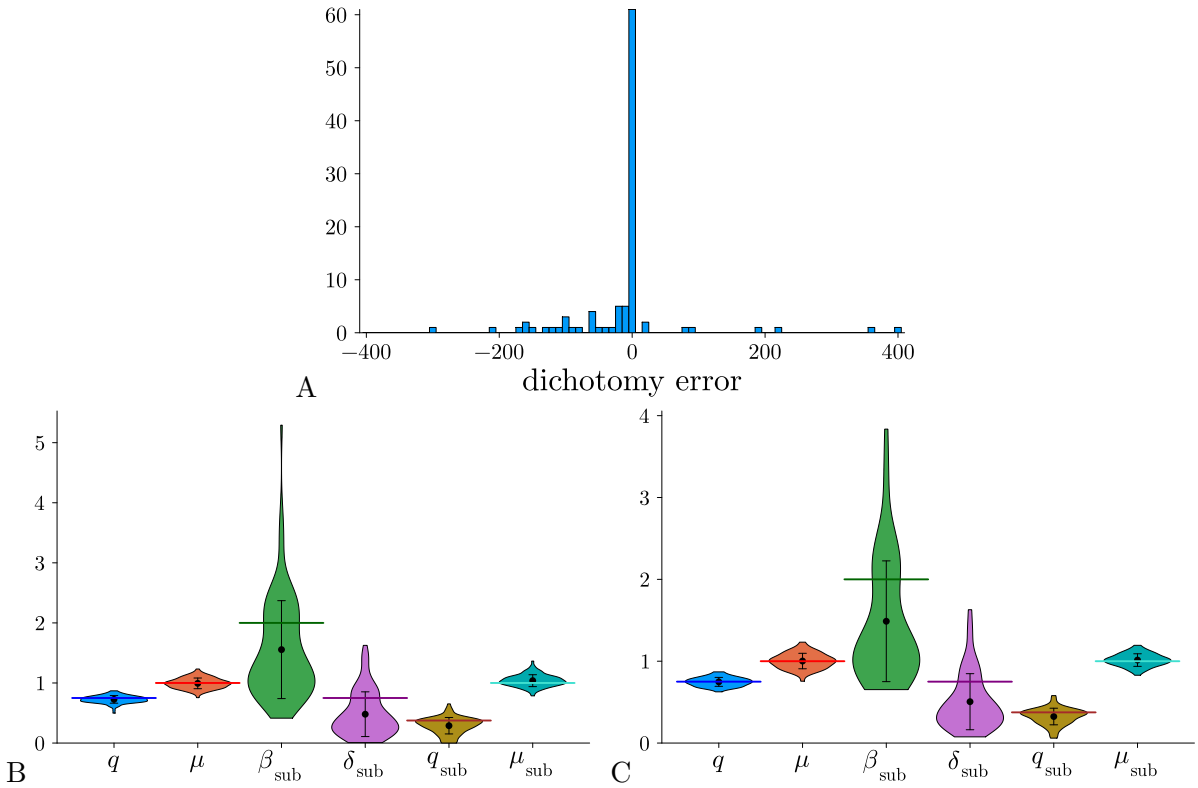

Figure 10: Results of the inference of parameter shifts We show inference results from 100 simulated populations as described in the text. (A) Dichotomy errors of the inferred distinct subtrees, as defined in the text. The correct subtree (dichotomy error zero) was found in 58 of the 100 simulations. (B) Violin plots show the inferred parameters, with black dots marking the mean and error bars the standard deviations of the inference results. Horizontal lines mark the true values used in the simulation. We further depict the relative death rate in the subtree $q_{\text{sub}} = \delta_{\text{sub}}/\beta_{\text{sub}}$. (C) The same as (B), but only depicting the results of the 58 simulations where the distinct subtree was located correctly.

Figure 10 shows the results from 100 simulated populations. In 58 cases, the parameter change was detected at the correct position. Only in four runs did the inferred subtree not overlap with the true distinct subtree. The dichotomy errors are depicted in Figure 10A, the inferred parameters from all 100 simulations are plotted as violin plots in Figure 10B. We also show the inferred relative death rate $q_{\text{sub}} = \delta_{\text{sub}}/\beta_{\text{sub}}$. Figure 10C shows that the results of only the simulations where the subtree was found correctly are only slightly better than when all runs are included. The selective advantage of the distinct subtree, i.e. a smaller relative death rate $q_{\text{sub}} < q$, was inferred correctly in all 96 cases where the inferred subtree overlapped with the true distinct subtree.

So far, we have simply used the likelihood difference as a rule to accept or reject different proposed parameter changes. However, we find that propositions of small distinct subtrees can be accepted due to overfitting stochastic fluctuations. This is why, here, we only consider subtrees with 25 leaves or more. Future research could lead to a principled rule for accepting/rejecting parameter changes that takes into account such fluctuations.