Phylodynamic inference on single-cell data

Tobias Dieselhorst and Johannes Berg*
Institute for Biological Physics, University of Cologne, Zülpicher Straße 77a, 50937 Cologne, Germany
(Dated: December 5, 2024)

The way mutations accumulate over the course of evolution fundamentally differs between macroevolutionary settings and cellular reproduction. In macroevolution, mutations generally accumulate continuously over time. Under cellular reproduction, like somatic evolution within a tumor or evolution within a bacterial population, mutations occur largely in discrete steps at individual reproductive events. In this paper, we address the problem of phylodynamic inference under cellular reproduction. We find that when mutations occur at cell divisions, the number of mutations along a phylogenetic branch follows a compound Poisson statistics rather than a Poisson statistics, and this affects the inference from phylogenetic data. Based on the likelihood of a birth-death model, we set up an inference scheme for the relative death-to-birth rate and the number of mutations per generation and apply this method to haematopoietic stem cells. Our approach is computationally efficient thus allowing a broad range of applications, in particular on lineage trees from single-cell data.

CONTENTS

I. Introduction 1
1

II. Results 3
A. Parameter inference from mutations along branches 3
B. Finite sampling 5
C. Application to single-cell data of haematopoietic stem cells 5

III. Discussion 6

Acknowledgments 7

References 7

A. Stochastic dynamics of a birth-death process 9

B. Joint probability of size-one clades and the number of generations 9

C. Distribution of the number of mutations along a branch 11

D. Comparison of models with different mutation statistics 11

E. Inference of the sampling probability 13

F. Inference bias 13

I. 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

* bergj@uni-koeln.de

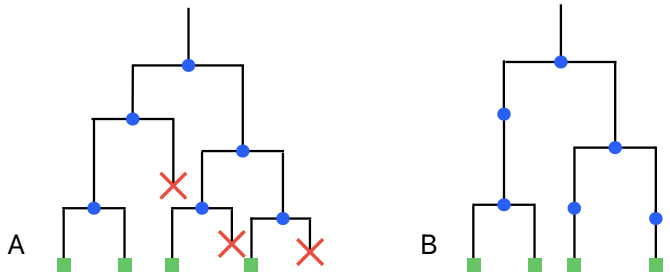

FIG. 1. **Comparison of a birth-death tree and the corresponding reconstructed phylogenetic tree.** (A) A birth-death process of individuals with time running downwards consists of birth events (blue circles) and death events (red crosses) and produces some individuals (green squares) that are alive 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.

population, ranging from the problem to infer rates of birth and death of species to the timing of bifurcations in trees [3, 12, 13, 16, 18, 23].

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 is usually a birth-death model with rates of birth and death that can be homogeneous or inhomogeneous across the tree [12, 23]. For the second model, the simplest choice is the molecular clock hypothesis, which assumes that mutations accumulate at some rate, constant or variable [7], per unit time.

In this paper, we develop an approach where mutations are tied to birth events of the underlying birth-death process. This link between mutations and birth events is specific to cellular reproduction and differs from a macroevolutionary scenario. In macroevolution, birth and death describe the speciation and extinction of species, however mutations generally arise through the procreation of individuals and the fixation of these mutations in a population. As a result, in macroevolution, mutations can be modelled as arising in populations continuously at some rate per unit of time, although that rate may vary over time.

When mutations are linked to birth processes of individuals, however, the mutation statistics is shaped by individual birth events along the branches of a tree. Some of these birth events correspond to bifurcations of the reconstructed phylogenetic tree and are easily detected. Other birth events produce a lineage that eventually dies out, and these events are not seen in the phylogenetic tree. Yet these 'hidden' birth events contribute to the accumulation of mutations and affect the statistics of mutations along branches, see Figure 1. 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 by the so-called compound Poisson statistics, with an increased variance arising from the discrete reproductive events along branches.

Based on the mutations along the branches of a phylogenetic tree, we aim to infer the rates of the birth-death model that produced the tree, as well as the statistics of mutations. Our inference scheme is geared towards large-scale single-cell datasets generated from a populations of individual cells that divide and die. 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. Also, we assume constant rates of cell birth and death through time and across the tree and will discuss extensions in Section III. To facilitate the efficient computation of the likelihood of a tree (specified by its topology and the mutations along each of its branches), we will 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 resulting algorithm has running times of seconds for hundreds of samples and scales linearly in the number of samples.

Our approach differs from the recent analysis of punctuated evolution [6, 14] using a Monte Carlo Markov chain dynamics. Punctuated evolution is a macroevolutionary model characterized by bursts of mutations at speciation events in addition to a regular molecular clock. Formally, our model of cellular reproduction can be viewed as an extreme case of punctuated evolution where mutations occur solely at speciations. However, Monte Carlo dynamics leads to prohibitively long running times for large trees. Our approach also differs from inference based on the distribution of pairwise distances between the tree leaves [26] in that we compute the likelihood of the entire tree and

use the exact statistics of hidden birth events arising from the birth-death process.

In Section II A, we recapitulate the dynamics of a birth-death process and establish the statistics of mutations along tree branches. The corresponding likelihood of the underlying birth-death model can be computed efficiently using a dynamic programming approach. In Section II B, 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 [15] in Section II C. All derivations are collected in the Appendix, along with results from extensive simulations.

II. RESULTS

A. Parameter inference from mutations along branches

We consider a growing population described by 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 Appendix 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 [2].

The second component concerns the statistics of the mutations occurring at each birth event (cell division). For simplicity, we only look at the number of mutations that take place at a cell division; possible extensions will be discussed in Section III. 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, see Appendix 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. To sum over these latent variables, we use a simple iterative procedure, dynamic programming, in the spirit of Felsenstein's recursive algorithm [8]: we iteratively sum over both the times of bifurcations and the number of generations along each branch. The same iterative approach is applied to bifurcation times in the *TreeTime* algorithm by Sagulenko, Puller, and Neher [18].

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

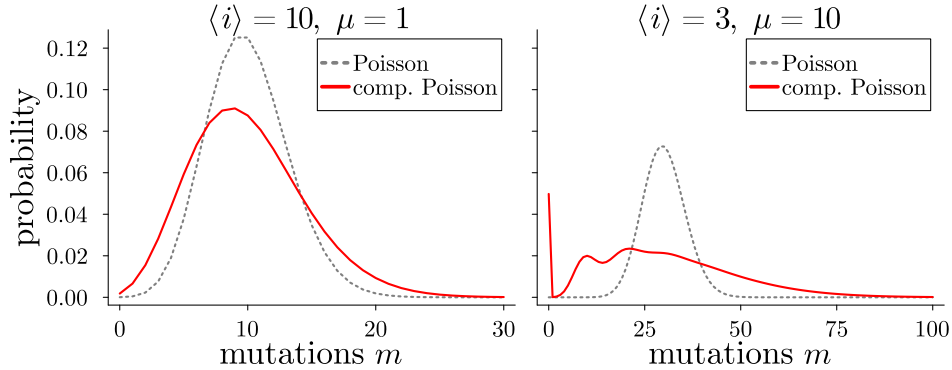

FIG. 2. **Compound Poisson distributed number of mutations.** The distribution of the number of mutations follows a compound Poisson statistics arising from the discrete number of generations along a branch. The red solid lines show the distributions of the number of mutations along branches with mean 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 Appendix C for details.

(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

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

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 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 our approach, 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. Appendix F shows how the inference improves with increasing the number of samples. 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 [10] 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.

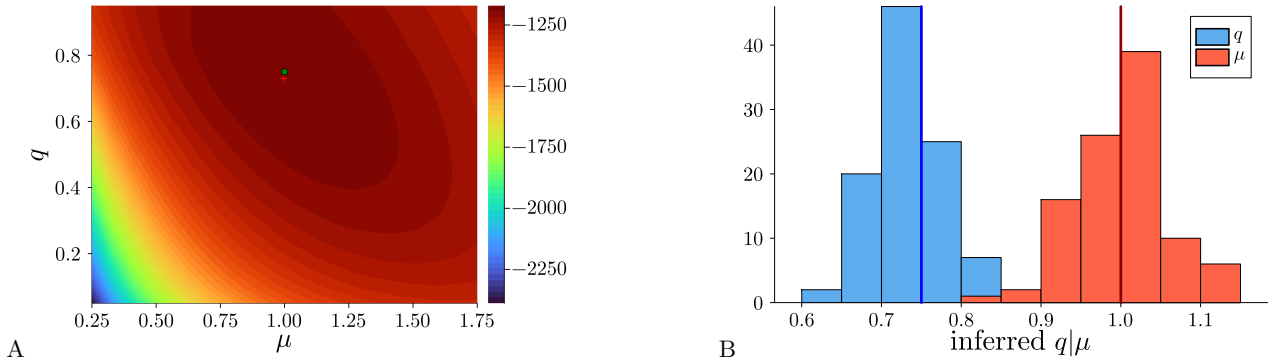

FIG. 3. **Inference using the likelihood of the birth-death process.** (A) A population was grown from a single individual to $N = 300$ individuals with $q = 3/4$ and $\mu = 1$ and the likelihood (5) for the resulting tree (defined by topology and mutations along branches) was computed iteratively as 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.

B. 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 [22] 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 [22, 24, 27] and pick extant individuals at time of observation independently and 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 [27] 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 II A 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 Appendix E. This is different from settings with a molecular clock, where based on the bifurcation times it turns out that there are many values of the sampling probability ρ that are compatible with the data [21].

C. Application to single-cell data of haematopoietic stem cells

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 [15]. A tree based on sequencing data from 216 HSC from donor

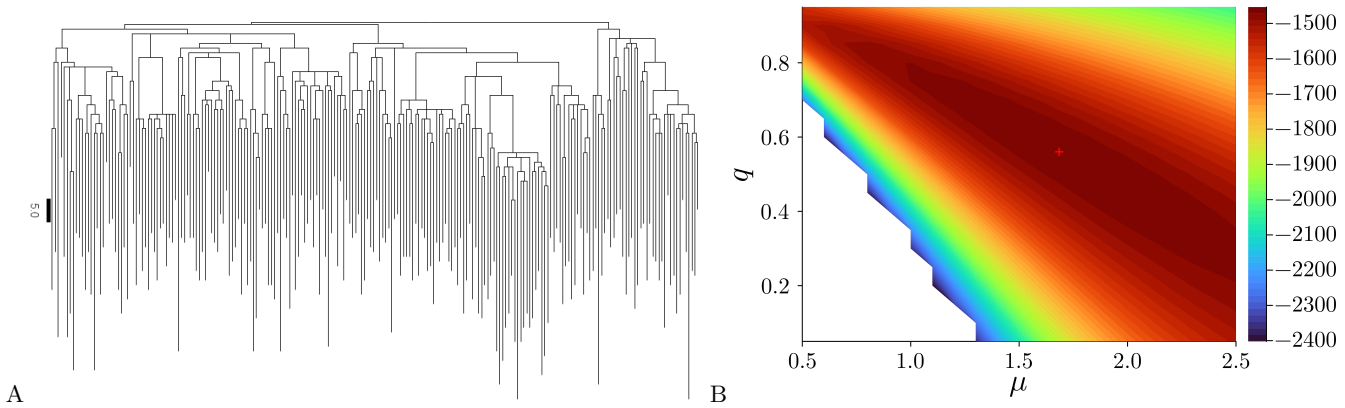

FIG. 4. **Likelihood of different trees from data.** (A) The reconstructed tree from single-cell data of HSC in donor CB001 from [15] with branch lengths proportional to the number of mutations along them. The phylogeny is taken from [15]. The length of the black bar on the left corresponds to five mutations. (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 $\mu = 1.24$.

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 [15]. 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 our results in [5] for branch lengths under finite sampling.

HSC differentiate into different blood cells. We assume that this happens stochastically with some rate per unit time. Thus, the inferred relative death rate will correspond to the sum of the rates of cell-death and differentiation, divided by the rate with which cells replicate.

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 $\mu = 1.68$. These are the parameters maximizing the likelihood under the assumption of a constant-rate birth-death process (with death events corresponding to either cell death or differentiation). Allowing for changes in the parameters across the tree is a generalization left for future work.

We also inferred the parameters for the dataset of a second neonate (donor CB002 in [15]), which contained 390 HSC and 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 $\mu = 1.45$.

We find that the results are not sensitive to changes in the sampling probability ρ. Setting $\rho = 0.001$, the likeliest parameters for the tree of donor CB001 are $q = 0.52$ and $\mu = 1.24$ and for donor CB002 $q = 0.61$ and $\mu = 1.03$.

III. DISCUSSION

Under cellular reproduction, mutations arise generally 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 produces phylogenetic trees which differ in two respects from their counterparts arising in macroevolution. First, we find that 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 fulfill in recent branches, reducing the number of generations per time in recent

branches [2]. As a result, generations accumulate in branches at different rates in different parts of the *reconstructed* tree, even when all underlying parameters are constant throughout evolution. This can lead to errors in phylodynamic analysis whenever neutral evolution under cellular reproduction is modelled by a constant-rate molecular clock (see Appendix 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 (as in Felsenstein's original algorithm [8]) as well as the branching times, leading to a computationally harder problem. This requires either Markov chain sampling approaches as, for instance, implemented in the *BEAST* package [3, 6], or heuristics like the *TreeTime* algorithm that alternately sums over ancestral states and branch lengths [18].

Even in somatic evolution, mutations can arise per unit time [20]. This effect has been studied recently by tracing cells during the growth of long-lived trees [19]. It could be that under a low rate of cell division, most mutations arise between cell divisions and are coupled to the passage of time, whereas in fast-dividing populations, mutations arise predominantly as part of the reproductive process. A mixed model incorporating both mutational processes along the lines of [6, 14] can easily be incorporated into our model of mutations and the iterative scheme to compute the likelihood.

Most importantly, we have so far kept the parameters of our phylodynamic model constant throughout time and throughout the tree. However, the rates of cell birth and death can change over time and across the tree. In the human haematopoietic stem cells analyzed in Section II C, there is a phase of exponential growth that establishes the stem cell population before birth, but after development this is followed by a phase where the stem cell population size is kept constant and most cells are dormant [15]. Later in life, this is followed by a phase characterized by the expansion of individual clades. Similarly, during tumour growth, growth rates and the mutation rate can vary drastically across the different phases of the disease [9]. Also, particular clades can acquire a relative growth advantage relative to others, which needs to be reflected in the phylodynamic parameters. These changes across time and across the tree can be incorporated into the likelihood along the lines of corresponding models in epidemic modelling [25]. This would allow disentangling exogenous temporal heterogeneity (such as distinct growth phases or selection) and intrinsic temporal heterogeneity stemming from the tree reconstruction.

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 and Martin Peifer for discussions, and to Tanja Stadler for comments throughout and especially for helpful advice on section II B. 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 CHEOPS cluster (grant number INST 216/512/1FUGG).

[1] J. K. Blitzstein and J. Hwang. *Introduction to probability*. Texts in statistical science. CRC press Taylor & Francis group, Boca Raton, 2nd ed edition, 2019. ISBN 978-1-138-36991-7.

[2] F. Bokma, V. van den Brink, and T. Stadler. Unexpectedly many extinct hominins. *Evolution*, 66(9):2969–2974, 2012.

[3] R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 15(4):e1006650, 2019.

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

[5] T. Dieselhorst and J. Berg. Branch length statistics in phylogenetic trees under constant-rate birth-death dynamics, July 2024. URL `http://arxiv.org/abs/2407.13403`. arXiv:2407.13403 [q-bio].

[6] J. Douglas, R. Bouckaert, S. C. Harris, C. W. Carter Jr, and P. R. Wills. Evolution is coupled with branching across many granularities of life. *bioRxiv*, pages 2024–09, 2024.

[7] A. J. Drummond, S. Y. W. Ho, M. J. Phillips, and A. Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS biology*, 4(5):e88, 2006.

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

[9] M. Gerstung, C. Jolly, I. Leshchiner, S. C. Dentro, S. Gonzalez, D. Rosebrock, T. J. Mitchell, Y. Rubanova, P. Anur, K. Yu, et al. The evolutionary history of 2,658 cancers. *Nature*, 578(7793):122–128, 2020.

[10] K. Hartmann, D. Wong, and T. Stadler. Sampling trees from evolutionary models. *Systematic Biology*, 59(4):465–476, 2010.

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

[12] A. MacPherson, S. Louca, A. McLaughlin, J. B. Joy, and M. W. Pennell. Unifying phylogenetic birth–death models in epidemiology and macroevolution. *Systematic Biology*, 71(1):172–189, 2022.

[13] W. P. Maddison, P. E. Midford, and S. P. Otto. Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, 56(5):701–710, 2007.

[14] M. Manceau, J. Marin, H. Morlon, and A. Lambert. Model-based inference of punctuated molecular evolution. *Molecular Biology and Evolution*, 37(11):3308–3323, 2020.

[15] E. Mitchell, M. Spencer Chapman, N. Williams, K. J. Dawson, N. Mende, E. F. Calderbank, H. Jung, T. Mitchell, T. H. H. Coorens, D. H. Spencer, H. Machado, H. Lee-Six, M. Davies, D. Hayler, M. A. Fabre, K. Mahbubani, F. Abascal, A. Cagan, G. S. Vassiliou, J. Baxter, I. Martincorena, M. R. Stratton, D. G. Kent, K. Chatterjee, K. S. Parsy, A. R. Green, J. Nangalia, E. Laurenti, and P. J. Campbell. Clonal dynamics of haematopoiesis across the human lifespan. *Nature*, 606 (7913):343–350, June 2022. ISSN 1476-4687. doi:10.1038/s41586-022-04786-y. URL `https://www.nature.com/articles/s41586-022-04786-y`. Publisher: Nature Publishing Group.

[16] 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*, 344(1309):305–311, 1994.

[17] S. M. Ross. *Introduction to probability models*. Elsevier, Amsterdam, 2014.

[18] P. Sagulenko, V. Puller, and R. A. Neher. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, 4(1): vex042, 2018.

[19] A. Satake, R. Imai, T. Fujino, S. Tomimoto, K. Ohta, M. Na'iem, S. Indrioko, Widiyatno, S. Purnomo, A. Mollá-Morales, et al. Somatic mutation rates scale with time not growth rate in long-lived tropical trees. *bioRxiv*, pages 2023–01, 2023.

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

[21] T. Stadler. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, 261(1):58–66, 2009.

[22] T. Stadler. Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, 267(3):396–404, 2010.

[23] T. Stadler. Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology*, 26(6): 1203–1219, 2013.

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

[25] T. Stadler, D. Kühnert, S. Bonhoeffer, and A. J. Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences*, 110(1):228–233, 2013.

[26] B. Werner, J. Case, M. J. Williams, K. Chkhaidze, D. Temko, J. Fernández-Mateos, G. D. Cresswell, D. Nichol, W. Cross, I. Spiteri, W. Huang, I. P. M. Tomlinson, C. P. Barnes, T. A. Graham, and A. Sottoriva. Measuring single cell divisions in human tissues from multi-region sequencing data. *Nature Communications*, 11(1):1–9, 2020.

[27] Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Molecular Biology and Evolution*, 14(7):717–724, 1997.

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

$$\partial_t p_n(t) = \beta(n-1)p_{n-1}(t) - \beta n p_n(t) + \delta(n+1)p_{n+1}(t) - \delta n p_n(t) \ . \tag{8}$$

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 [11] gives the probability that the clade has died out by time T

$$p_0(T) = q\frac{1 - \mathrm{e}^{-(1-q)T}}{1 - q\,\mathrm{e}^{-(1-q)T}} \tag{9}$$

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

$$p_1(T) = \frac{(1-q)^2\,\mathrm{e}^{-(1-q)T}}{(1 - q\,\mathrm{e}^{-(1-q)T})^2} \ , \tag{10}$$

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 II 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 our approach in [5], where we derived the same distribution but without considering the number of birth events. We will also call these birth events (which are hidden in the phylogenetic tree) generations. Denoting the joint probability as $p_1(i|\tau_s, \tau_s)$ and using the probabilities (9)-(10), its dynamics are described by the master equation

$$-\partial_{\tau_e} p_1(i|\tau_s, \tau_e) = -q p_1(i|\tau_s, \tau_e) - p_1(i|\tau_s, \tau_e) + 2p_0(\tau_e)p_1(i-1|\tau_s, \tau_e) \ . \tag{11}$$

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

$$-\partial_{\tau_e} G(\mu, \tau_s, \tau_e) = -\left(q + 1 - 2\mu p_0(\tau_e)\right) G(\mu, t) \ , \tag{12}$$

which is easily solved by integration.

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

$$G(\mu, \tau_s, \tau_e = 0) = \exp\left\{-(1+q)\tau_s + \mu\left[2q\tau_s + 2\log\left(\frac{1-q}{1 - q\,\mathrm{e}^{-(1-q)\tau_s}}\right)\right]\right\} \ . \tag{13}$$

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

$$\begin{aligned} p_1(i|\tau_s, \tau_e = 0) &= \frac{1}{i!}\partial_\mu^i|_{\mu=0} G(\mu, \tau_s, \tau_e = 0) \\ &= \mathrm{e}^{-(1+q)\tau_s}\,\frac{1}{i!}\left[2q\tau_s + 2\log\left(\frac{1-q}{1 - q\,\mathrm{e}^{-(1-q)\tau_s}}\right)\right]^i \ . \end{aligned} \tag{14}$$

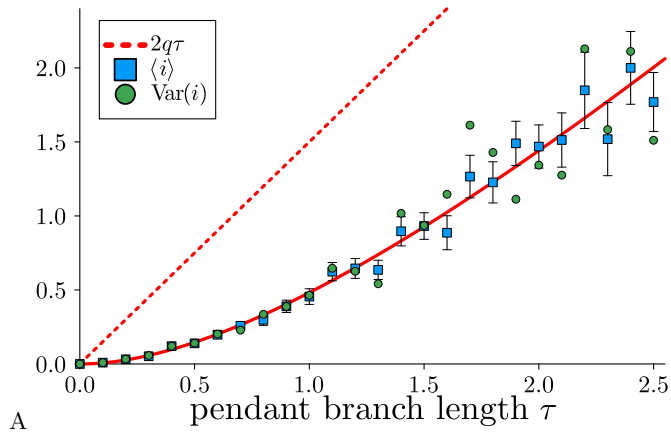

FIG. 5. **Distribution of the number of generations along pendant branches.** Numerical results come from a single population grown to size 7500 at relative death rate $q = 3/4$. We pick all pendant branches of a given length τ (± 0.05) and plot the mean (large blue symbols, with the standard error given 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 (16) (with $\tau_e = 0$) for both the mean and the variance. The dotted red line gives the linear term $2q\tau$ in (16), corresponding to generations appearing along the branch at a constant rate.

Summing over the number of generations results in equation (10).

Integrating the equation of motion of the generating function (12) 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

$$p_1(i|\tau_s, \tau_e) = \mathrm{e}^{-(1+q)(\tau_s-\tau_e)} \frac{1}{i!} \left[2q\tau + 2\log\left(\frac{1 - q\,\mathrm{e}^{-(1-q)\tau_e}}{1 - q\,\mathrm{e}^{-(1-q)\tau_s}} \right) \right]^i . \tag{15}$$

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

$$\langle i \rangle(\tau_s, \tau_e) = 2q(\tau_s - \tau_e) + 2\log\left(\frac{1 - q\,\mathrm{e}^{-(1-q)\tau_e}}{1 - q\,\mathrm{e}^{-(1-q)\tau_s}} \right) . \tag{16}$$

This agrees with the result derived by Bokma, van den Brink, and Stadler [2] for the mean number of generations. The first term in (16) is proportional to the elapsed time $\tau = \tau_s - \tau_e$; the constant 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 (16) for $\tau_e = 0$. (For a Poisson distribution, the variance equals the mean.)

Summing over the number of generations i in distribution (15) gives the distribution of branch lengths $p_1^{(\tau_e)}(\tau_s - \tau_e)$ [5]. Equation 15 can also be derived by multiplying $p_1^{(\tau_e)}(\tau_s - \tau_e)$ with the result from [2] that the number of generations is Poisson distributed with mean (16).

These results can be easily adapted to finite sampling as discussed in Section II B. $p_0(\tau_e)$ in the equation of motion (12) for the generating function needs to be replaced by $p_0^{(\rho)}(\tau_e)$ given by (6). The result is (15) but with the term in square brackets replaced by

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

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 II A, 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 [17] to compute the mean number of mutations between bifurcations as

$$\langle m \rangle = \langle i \rangle \langle \hat{m} \rangle \ . \tag{18}$$

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, [1]), we find the variance to be

$$\begin{aligned} \mathrm{Var}\,[m] &= \langle \mathrm{Var}\,[m|i] \rangle + \mathrm{Var}\,[\langle m|i \rangle] \\ &= \mathrm{Var}\,[\hat{m}]\,\langle i \rangle + \langle \hat{m} \rangle^2 \mathrm{Var}\,[i] \\ &= \left(\mathrm{Var}\,[\hat{m}] + \langle \hat{m} \rangle^2 \right) \langle i \rangle \ . \end{aligned} \tag{19}$$

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

$$\mathrm{Var}\,[m] = \langle i \rangle \mu (1 + \mu) \tag{20}$$

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

$$P(m|\tau_s, \tau_e) = \sum_{i=0}^{\infty} P(i|\tau_s, \tau_e) P(m|i) \tag{21}$$

$$= \sum_{i=0}^{\infty} \mathrm{e}^{-\langle i \rangle (\tau_s, \tau_e)} \, \frac{\langle i \rangle (\tau_s, \tau_e)^i}{i!} \, \mathrm{e}^{-i\mu} \, \frac{(i\mu)^m}{m!} \ . \tag{22}$$

$P(i|\tau_s, \tau_e)$ 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 II, 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 (21) 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 Appendix B), which affects the statistics of mutations if they are tied to birth events. Furthermore,

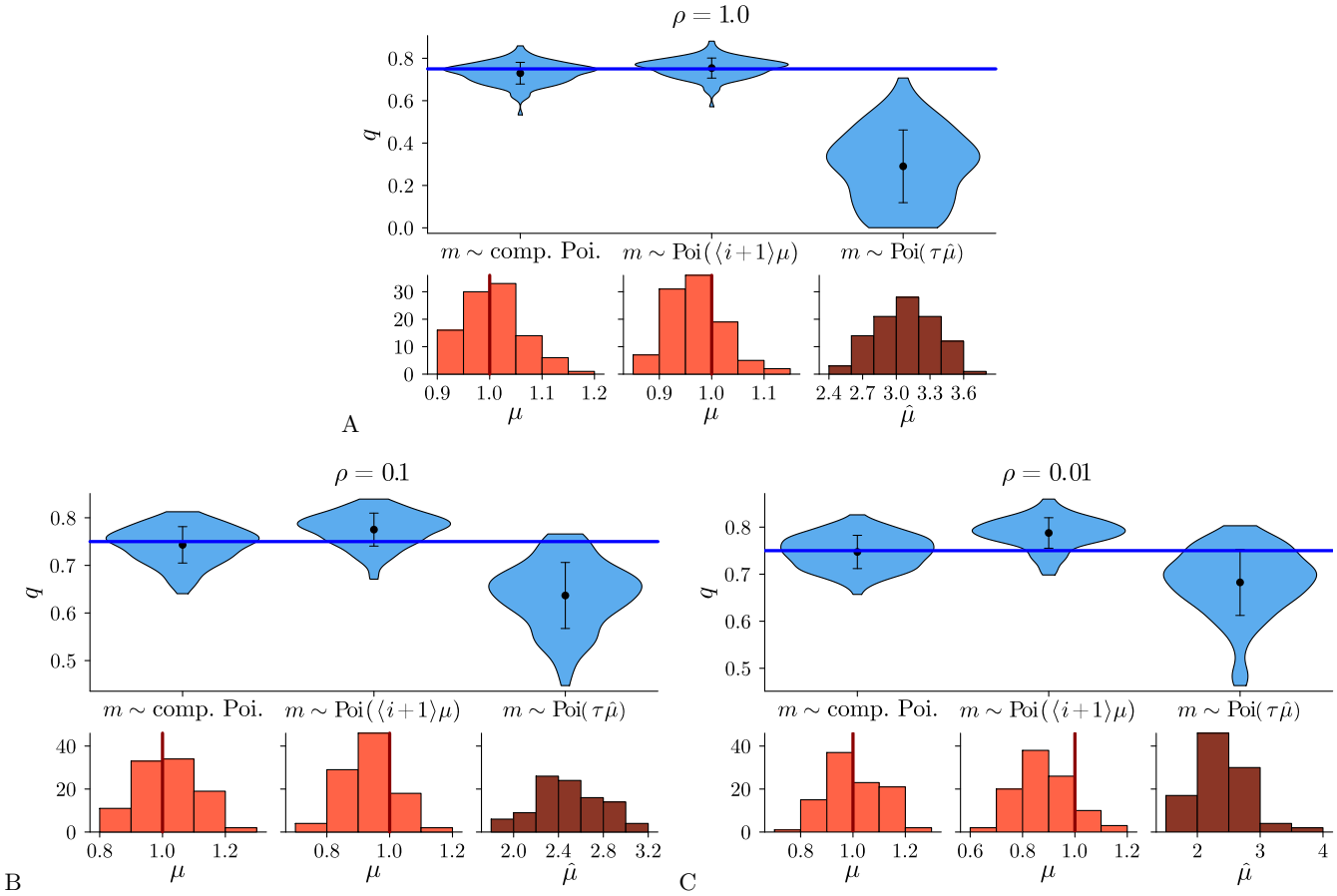

FIG. 6. **Comparison of different models of mutations.** We compare results of our method based on the compound Poisson statistics of mutations along branches ("$m \sim$ comp. Poi.") and the same algorithm but with Poisson statistics (inhomogeneous "$m \sim \mathrm{Poi}(\langle i+1 \rangle \mu)$" and a constant-rate molecular clock "$m \sim \mathrm{Poi}(\tau \hat{\mu})$", 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 $\hat{\mu}$. 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 $\rho = 0.01$.

mutations from birth events along branches follow compound Poisson statistics rather than a Poisson distribution (Appendix 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 Appendix C) by a Poisson distribution with the same mean. In a second approach, we assume a constant molecular clock with rate $\hat{\mu}$ throughout the entire tree. In both approaches, the bifurcation times and number of generations along branches are integrated and summed over as discussed in Section II A.

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 $\hat{\mu}$.

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

The possibility to infer ρ is an important difference to inference schemes based on bifurcation times or branch lengths in calendar time: the distributions of bifurcation times or branch lengths are not unique to a single set of birth rate, death rate and ρ [5, 21]. Instead, an entire family of parameter sets is compatible with each time-scaled tree. However, the joint probability of a lineage and the number of generations along a branch 7 is unique to every set of parameters and this is why our inference scheme can infer all three parameters q, μ and ρ.

F. INFERENCE BIAS

Maximum likelihood estimates generally show a bias in the parameter estimates [4], which vanishes in the limit of a large number of sampled individuals. Figure 8 shows how the bias in the estimates from the likelihood (5) decreases with the number of sampled individuals (equal to the population size at full sampling, $\rho = 1$).

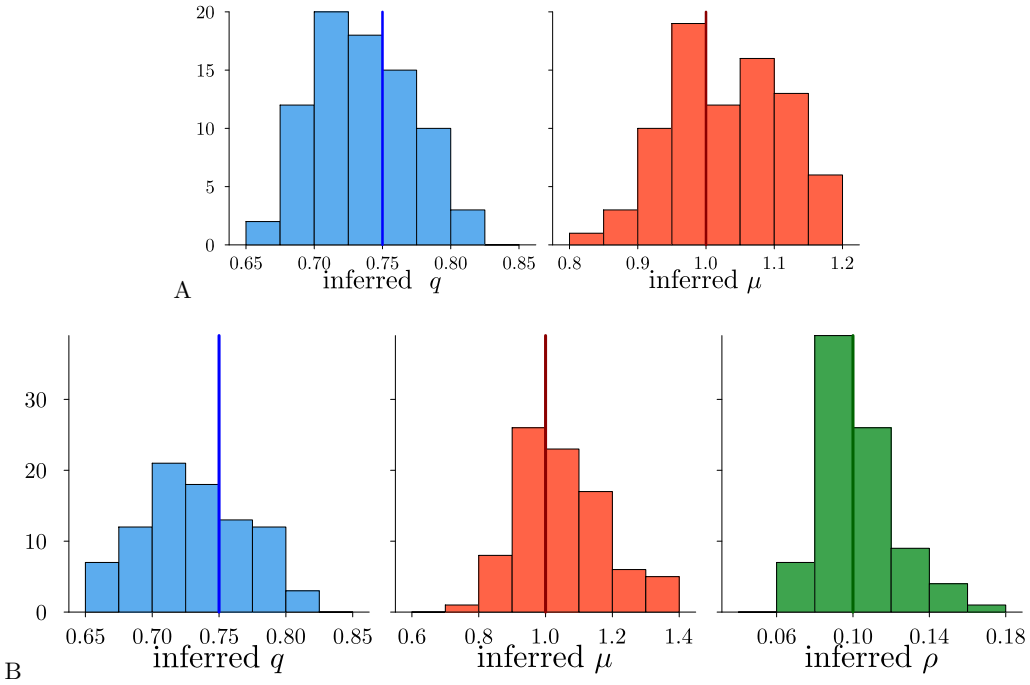

FIG. 7. **Parameter inference from samples of a population.** Depicted are the inference results from 100 populations simulated with $q = 3/4$ and $\mu = 1.0$ up to size 3000. For each population, a tree was reconstructed from 300 samples, corresponding to a sampling probability of $\rho = 0.1$ (see 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.

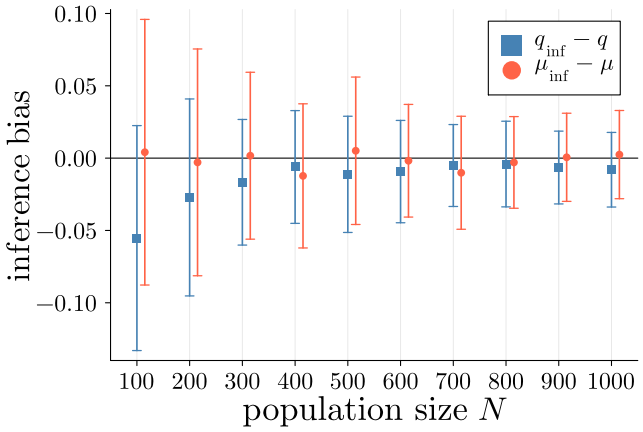

FIG. 8. **The dependence of the inference bias on population size.** The difference between inferred and underlying values of the relative rate of death q and the number of mutations per birth μ is shown for populations of different sizes. The symbols (blue squares for q and orange balls for μ) show the differences between inferred and underlying values averaged over 100 runs; error bars show the corresponding standard deviation. The inference is based on maximizing the likelihood (5), the underlying parameters were $q = 3/4$, $\mu = 1$ at full sampling ($\rho = 1$). For visibility, the symbols indicating the results for the mutation parameter μ are shifted a little to the right. 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.