OXFORD GENETICS

GENETICS, 2026, 232(1), iyaf112
https://doi.org/10.1093/genetics/iyaf112
Advance Access Publication Date: 12 June 2025
Investigation

# Learning evolutionary parameters from genealogies using allelic trees

Antoine Aragon [1], Amaury Lambert [2,*,†], Thierry Mora [1,†], Aleksandra M. Walczak [1,†]

[1]Laboratoire de physique de l'École normale supérieure, CNRS, PSL University, Sorbonne Université, and Université de Paris, Paris 75005, France
[2]Institut de Biologie de l'ENS (IBENS), École normale supérieure, CNRS, INSERM, PSL University, Paris 75005, France

*Corresponding author: Institut de Biologie de l'ENS (IBENS), École normale supérieure, CNRS, INSERM, PSL University, 46 rue d'Ulm, Paris 75005, France. Email: amaury.lambert@ens.psl.eu
†These authors contributed equally to this work.

Cellular diversification in processes from development to cancer progression and affinity maturation is often linked to the appearance of new mutations, generating genetic heterogeneity. Describing the underlying coupled genetic and growth processes that result in the observed diversity in cell populations is informative about the timing, drivers and outcomes of cell fates. Current approaches based on phylogenetic methods do not cover the entire range of evolutionary rates, often making artificial assumptions about the timing of events. We introduce CBA, a probabilistic method that infers the division, degradation and mutation rates from the observed genetic diversity in a population of cells. It uses a summarized backbone tree, intermediary between the true cell tree and the allelic tree representing the ancestral relationships between types, called a monogram, which allows for efficient sampling of possible phylogenies consistent with the observed mutational signatures. We demonstrate the accuracy of our method on simulated data and compare its performance to standard phylogenetic approaches.

Keywords: phylogenetic methods; statistical inference; mathematical models; population genetics; cell lineages; allelic trees; somatic evolution

## Introduction

The growth and diversification of somatic cell populations within a organism offer a perfect example of asexual evolution. Particular instances include cancer progression (Simeonov *et al.* 2021; Yang *et al.* 2022; Jones *et al.* 2023), affinity maturation of antibodies (Uduman *et al.* 2011; Yaari *et al.* 2015; DeWitt III *et al.* 2018; Horns *et al.* 2019; Nourmohammad *et al.* 2019; Hoehn *et al.* 2021; Turner *et al.* 2021), and hematopoiesis (Perié *et al.* 2015; Lee-Six *et al.* 2018; Fabre *et al.* 2022; Mitchell *et al.* 2022). Because these processes may be modeled by traditional evolutionary genetics (Stadler *et al.* 2021), inference methods developed in the fields of phylogenetics and population genetics could be adapted to estimate the evolutionary parameters of somatic processes, such as division, death, and mutation rates, as well as their dependence on time, tissue, or cell type. Efforts in adapting such methods to somatic evolution hold the potential to gain new insight into a variety of biological processes, such as the timing and drivers of cell fate decision during early development (Coorens *et al.* 2021 ; Park *et al.* 2021; Spencer Chapman *et al.* 2021), the mechanisms of hypermutation and selection during antibody affinity maturation (Uduman *et al.* 2011; Yaari *et al.* 2015; Horns *et al.* 2019; Nourmohammad *et al.* 2019; Hoehn *et al.* 2021), or the drivers of cancer onset and progression (Martincorena *et al.* 2017).

Several methods exist to infer the phylodynamics of growing populations. Bayesian inference is widely used to infer phylogenetic trees from molecular sequences and to test evolutionary hypotheses. In particular, the "Swiss knife" BEAST 2.0 platform (Bouckaert *et al.* 2014), based on Monte–Carlo Markov chains, offers a large diversity of evolutionary models allowing for various inference tasks. More recently, Delphy, a reformulation of Bayesian phylogenetics, was designed to improve speed and scalability in the analysis of large epidemiology datasets (Varilly *et al.* 2025). Other methods have been proposed to address specific aims more efficiently. Starting from time-resolved genealogies, phylodyn allows for inference of time-dependent population size (Karcher *et al.* 2017). A distinct method, cloneRate, is optimized to efficiently infer the net growth rate (Johnson *et al.* 2023), with applications to hematopoiesis. Starting from densely sampled sequence data of rapidly evolving populations such as viruses, TreeTime (Sagulenko *et al.* 2018) infers time-resolved genealogies as well as evolutionary models, while another method infers their individual fitnesses (Neher *et al.* 2014). Recently, a method was proposed to infer the mutation rate and the death-to-division ratio from phylogenetic trees with no time information, assuming that mutations occur at cell division (Dieselhorst and Berg 2024).

A major challenge faced by inference methods is that we never have access to the full process that produced the observed data. Typical datasets will only include a sample of existing types, which are defined by their somatic DNA mutations, barcoding, or gene expression profiles measured by RNAseq (Fig. 1a). Estimating evolutionary parameters can be done directly from the whole cell lineage tree (Fig. 1b, left). However, this is not possible since cell deaths leave no signatures in the data. Phylogenetic approaches often start from the "reconstructed"

Received on 28 February 2025; accepted on 02 June 2025

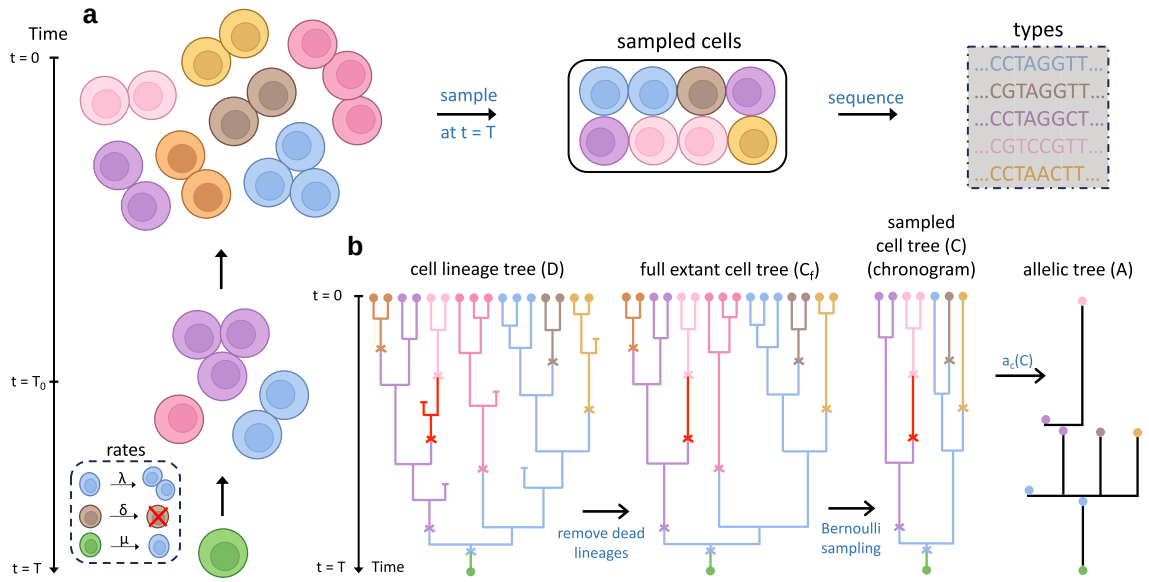


**Fig. 1.** Evolutionary model of growing populations of cells. a) Growing population of cells along with schematic of how data are obtained from typical experiments. The population is sampled at time $t = 0$ and then sequenced, yielding a list of (geno)types. b) Left and middle-left panels: cell lineage tree and extant cell lineage tree generated by the population the population generates. Cells divide at rate $\lambda$, die at rate $\delta$ and acquire mutations at rate $\mu$. In the cell tree $D$, cell deaths are depicted by a horizontal line marking the end of a lineage, and mutations are indicated by crosses. Each mutation corresponds to a change of type, which is represented by a change in shade (infinite-site model). The full extant cell lineage tree ($C_f$) is the cell lineage tree where dead lineages are removed. By contrast, the cell tree $D$ also represents dead cells. Middle-right panel: the sampled cell tree $C$ (chronogram) is obtained through a Bernoulli sampling with parameter $\rho$ on the leaves of $C_f$. Right panel: The allelic tree $A = a_c(C)$ (i.e. the phylogram with abundances of nodes and leaves removed), obtained by keeping only the list and order of mutations across types. The length of branches in $A$ indicates the number of mutations separating the types.

tree $C_f$, which traces the time-resolved genealogy of extant cells (Fig. 1b, middle left), or from the reconstructed tree of sampled cells (Fig. 1b, middle right) called the chronogram, or the sampled cell tree (C). However, what the data provide are a mixture of variants, or "types," that arose at different, unknown times, along with their abundances. Its phylogeny may be inferred from data to get a genealogical tree of extant cells whose branch lengths represent genetic distance, called a phylogram. When the abundances of types cannot be measured accurately and are discarded, as we will assume in this paper, the phylogram reduces to a "tree of mutations," called the allelic tree $A$ (Fig. 1b, right), where each tip represents a type present in the sample, and each branch represents a series of mutations necessary to transition from one type to another.

Many methods approximate the chronogram by the phylogram, by using the genetic distance between two cells as a rough measure of calendar time. That approximation is valid when the genetic distance is neither too small (to avoid small number fluctuations) nor too large, and under the double assumption that mutations are neutral and of constant rate—the molecular clock hypothesis. The genetic divergence between two individuals is on average $2\mu\ell t$, where $t$ is the time since their most recent common ancestor, $\mu$ is the per nucleotide mutation rate and $\ell$ is the number of nucleotides in the stretch of DNA considered. In phylogenetics, individuals are usually sampled in different species so that $t$ is large. In viral evolution (Grenfell *et al.* 2004), $t$ may be small but $\mu\ell$ is large.

In contrast, in somatic evolution $\mu\ell t$ may be small, so that the phylogram is a poor proxy of the chronogram. In specific cases, mutations can be artificially induced through genome editing-based lineage tracing methods to get a better resolution of time and apply classical methods with some adaptations (Feng *et al.* 2021; Seidel and Stadler 2022; Prillo *et al.* 2023; Gao and Feder 2024; Zwaans *et al.* 2024). In many cases however, inferring the chronogram from data are not possible. On the other hand, since mutations are rare, it is relatively straightforward to build the phylogram by maximum parsimony (DeWitt III *et al.* 2018).

Our goal is to propose a probabilistic method that infers the division, death and mutation rates from the allelic tree, which works even in the low mutation limit. Our strategy is based on an intermediate object between the sampled cell tree (chronogram) and the allelic tree (phylogram with no abundances): the *monogram*, which we will refer to from now on as the backbone tree. The backbone tree can be viewed as the allelic tree for which times of division and mutation events have been resolved and calibrated in time. The method we introduce, called CBA, interpolates between these three levels of description: (C) cell tree or chronogram, (B) backbone tree or monogram, and (A) allelic tree.

## Results

### Outline of the approach

We consider a population of cells that proliferates and mutates following a birth–death process with division rate $\lambda$, death rate $\delta$ and a neutral mutation rate $\mu$ (Fig. 1a). We assume that mutations accumulate with absolute time and not upon division. We further assume an infinite sites model of mutation with no recurrence. Time is defined in a backwards manner, such that $t = T$ marks the beginning of the process while $t = 0$ marks the present. As the population grows, it generates a cell lineage tree $D$ shown in Fig. 1b, which tracks divisions, deaths and mutations. The cell lineage tree is what an ideal observer would see given the evolution of the population between $t = T$ and $t = 0$, as it contains all the past relevant information about growth and diversification. Finding estimates for the rates from $D$ is easily accomplished through maximum likelihood using the branching property (Lambert and Stadler 2013). Unfortunately, in most datasets we do not have

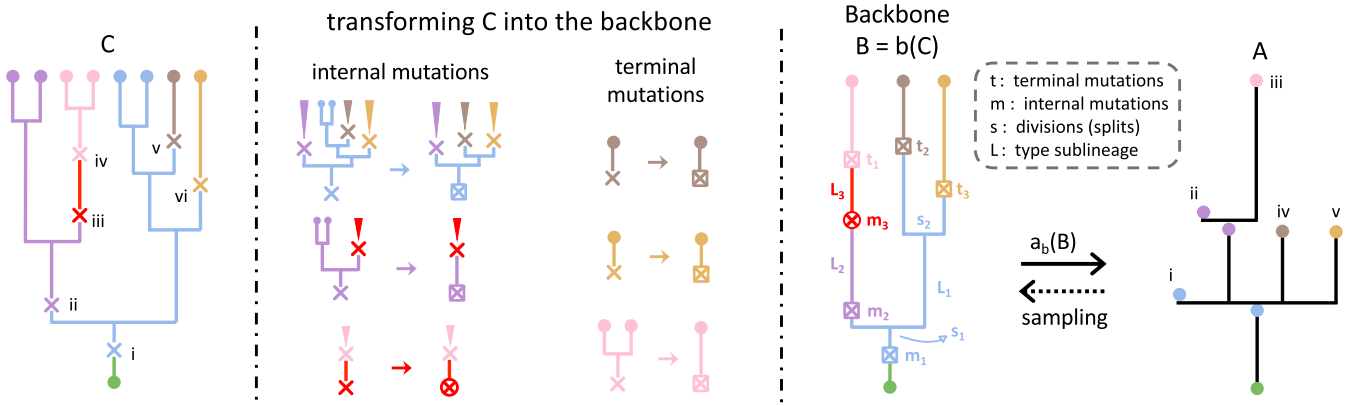


**Fig. 2.** The backbone tree as an intermediate between cell and allelic trees. Left panel: the starting point is the tree of sampled extant cells $C$. Middle panel: set of rules transforming a cell tree into a backbone tree. Subtrees descending from terminal mutations (iv, v, vi) are collapsed into a single edge and the mutation is boxed to indicate that the type emerging from that mutation is observed. Internal mutations leading to subtrees carrying no further mutations are boxed, and the corresponding unmutated subtrees removed (i, ii). Internal mutations from which no such subtree stems (iii) are circled. Right panel: The backbone resulting from $B$ is transformed into an allelic tree through $f(B)$, so that $f(b(C)) = a(C)$.

access to the whole tree but only have access to a sample of cells at one timepoint, or a few timepoints.

Removing death events that give rise to dead end branches from the cell tree $D$ in Fig. 1b gives the lineage tree $C_f$ of extant cells at $t = 0$, where the subscript "$f$" stands for "full sample" and refers to the fact that all the extant cells in $D$ are present as the leaves of $C_f$. This object is often studied in macroevolution (Nee *et al.* 1994; Morlon *et al.* 2011). A quasi-ideal observer, seeing a population of living cells at time $t = 0$ with all their genetic distances, could only reconstruct $C_f$ and not $D$ since the present population contains no information about past death events. The likelihood of the reconstructed cell lineage tree was derived in Nee *et al.* (1994) and could be used to infer the rates using maximum likelihood (see Supplementary Section 1).

Typical experiments give access to neither $D$ nor $C_f$. The population is sampled at time $t = 0$ and then sequenced, yielding a list of (geno)types, in which abundances (number of cells representing each type) are often discarded since experimental biases (sequence amplification error, expression noise) make them unreliable. We call $C$ the subtree of $C_f$ that contains only the phylogeny of sampled cells, obtained by Bernoulli sampling on the leaves of $C_f$ (Fig. 1b, second left). But experiments do not give direct information about the times of bifurcation and mutation events in $C$, nor about tree bifurcations that are not followed by mutations, and so in practice even the sampled tree $C$ is unknown. Instead, experiments typically give the allelic tree $A$, which describes the history of mutations leading to the sequences. The allelic tree may be inferred from the list of type sequences using well-established maximum-likelihood methods such as IQ-Tree (Minh *et al.* 2020) and RAxML (Stamatakis 2014). $A = a(C)$ can be deduced from $C$ (Fig. 1b), where the function $a$ consists in removing branches in $C$ with no mutations and making the length of all remaining branches proportional to the number of mutations. No explicit expression for the likelihood $P(A|\lambda, \delta, \mu)$ exists for the allelic tree, making inference challenging.

Our method fills this gap and infers the rates $\lambda, \delta, \mu$ governing the dynamics of the process through the allelic tree. The lack of an explicit expression for the likelihood of $A$ makes this a difficult task. To overcome this hurdle, we connect $A$ to $C$ through the backbone tree $B$ (Fig. 2). The backbone tree is an intermediate object between $A$ and $C$. It is a function $B = b(C)$ of the sampled cell lineage tree $C$. The rules of $b$ for transforming $C$ into $B$ are illustrated in Fig. 2. Subtrees downstream of terminal mutations, i.e. subtrees that have no further mutations ( iv, v, and vi), are replaced by a single terminal branch. Subtrees downstream of internal (nonterminal) mutations (i, ii, and iii) are collapsed to only keep the part of the tree leading to further mutations. Mutations are boxed if in $C$ there exists a path between the mutation and the present with no further mutation, in other words, if the type that emerged following that mutation is represented in the data. Mutations are circled if there is no such path and all sampled descendants carry additional mutations (e.g. mutation iii). The allelic tree $A = a_b(B)$ can be obtained from $B$ through a function $a_b$ which collapses to length 0 any mutation-free branch in $B$ and instead generates upstream of every mutation an edge of unit length. If in addition this mutation is internal and boxed in the backbone, $f$ adds downstream of it an edge of zero length (e.g. leaves i and ii in the allelic tree in Fig. 2). It follows that $a_c(C) = a_b(b(C))$. The functions $a_b$, $a_c$, and $b$ are independent of both the choice of models for $C$ and the experimental conditions (e.g. sampling, sequencing depth) and algorithmic methods (e.g. maximum parsimony, maximum likelihood) influencing $A$. They are recipes for transforming one tree type into another and depend solely on the input tree $C$ or $B$.

The backbone tree $B$ is the simplest reduction of the cell tree $C$ towards $A$ that we found and whose likelihood $P(B|\lambda, \delta, \nu)$ can still be computed analytically. It is close enough to the allelic tree so that fast sampling schemes exist to generate random backbone trees consistent with a given allelic tree, $B \in a_b^{-1}(A)$. We exploit this last point to develop an efficient inference procedure.

## Likelihood of the backbone tree

To implement our approach we must first derive the likelihood of the backbone tree. A full derivation with explicit calculations is detailed in Supplementary Section 2. This likelihood includes terms corresponding to divisions (splits) at times $(s_1, \ldots, s_{n-1})$, internal boxed mutations at times $(m_1, \ldots, m_{\ell'})$, internal circled mutations at times $(m_{\ell'+1}, \ldots, m_\ell)$, and terminal mutations at times $(t_1, \ldots, t_n)$, where $\ell$ is the total number of mutations, and $\ell'$ the number of those which are boxed. All those times take values between 0 and $T$. The concatenation of intervals downstream of an internal mutation $i$ until the next mutations is denoted by $L_i$ and

represented by the same color as the mutation in the left panel of Fig. 2. The concatenation of the internal intervals of $B$, i.e. all of $B$ except for edges stemming from terminal mutations, is denoted by $\tilde{B}$. In the backbone tree shown on the right of Fig. 2, $\tilde{B}$ would be the concatenation of the green, blue, purple, and red intervals. Putting all the division and mutation terms together, the likelihood of the backbone tree can be written as

$$P(B|\lambda,\delta,\mu) = P_{\text{free}}(\tilde{B})\prod_{i=1}^{n-1}P_{\text{div}}(s_i)\prod_{i=1}^{n}P_{\text{term}}(t_j) \times\prod_{i=1}^{\ell'}P_{\text{box}}(L_i)\prod_{i=\ell'+1}^{\ell}P_{\text{circ}}(L_i). \quad (1)$$

For example, for the tree in Fig. 2 the likelihood is:

$$P(B|\lambda,\delta,\mu) = P_{\text{free}}(\tilde{B})P_{\text{div}}(s_1)P_{\text{div}}(s_2) \times P_{\text{box}}(L_1)P_{\text{box}}(L_2)P_{\text{circ}}(L_3) \times P_{\text{term}}(t_1)P_{\text{term}}(t_2)P_{\text{term}}(t_3). \quad (2)$$

Each term is detailed below. While all mutations occur with rate $\mu$ and divisions with rate $\lambda$, in the backbone tree these rates must be corrected to account for the fate of the downstream subtrees, in particular that they must be represented in the data, and whether they undergo further mutations or not. To calculate the functional form of the terms in the likelihood, it is useful to denote by $p(t)$ the probability that a sublineage downstream of a point on the tree at time $t$ survives until the present, and by $q(t)$ the probability that sublineage survives and also mutates at least once. These probabilities can be computed analytically, as shown in Supplementary Section 2.

By construction, the backbone tree only carries a visible division event if both postdivision branches mutate after the division. The rate of a division occurring at time $t$ in $B$ (Supplementary Fig. 2) is then proportional to the division rate $\lambda$ times the probability that each sublineage survives and mutates:

$$P_{\text{div}}(t) = \lambda q^2(t)/q(t) = \lambda q(t), \quad (3)$$

where the $q(t)$ in the denominator comes from conditioning on the predivision branch having at least one mutation and surviving.

The rate of a terminal mutation occurring at time $t$ is given by the mutation rate $\mu$ times the probability $p(t) - q(t)$ that the branch survives until the present but without any further mutation (Supplementary Fig. 2), conditional again on the predivision branch surviving with at least one further mutation:

$$P_{\text{term}}(t) = \mu\frac{p(t)-q(t)}{q(t)}. \quad (4)$$

We see from both the $P_{\text{div}}(t)$ and $P_{\text{term}}(m_j)$ terms that conditioning on the predivision branch surviving introduces a nonobvious factor to the rates.

An internal circled mutation, such as the red mutation at time $m_3$ in Fig. 2, means that there was no division event between this mutation and the next downstream mutations (in this case during the interval $L_3$) that would lead to a sublineage surviving until the present with no further mutations—the existence of such a sublineage would imply the representation of the type emerging upon that mutation, making it a boxed mutation. The rate of such division events is given by $2\lambda(p(t) - q(t))$, where the factor of 2 stems from the fact that there are 2 potential subtrees at each division that could cause the type to be represented (Supplementary Fig. 2). More rigorously, a division event can lead to 3 alternatives: two surviving subtrees each with at least one mutation with probability $q(t)^2$, both without a mutation in the surviving branches with probability $(p(t) - q(t))^2$, and one of each type with probability $2q(t)(p(t) - q(t))$ (Supplementary Fig. 2). We must condition on the predivision branch surviving with at least one mutation, since we are considering points on the tree between internal and terminal mutations. This means keeping only the case where there is a subtree of each type, since we want the type to be represented, and dividing its probability by $q(t)$: $2q(t)(p(t) - q(t))/q(t) = 2(p(t) - q(t))$.

As a result, the rate of mutations *not* followed by such division events until the next mutations is:

$$P_{\text{circ}}(L_i) = \mu\exp\left(-2\lambda\int_{L_i}(p(t)-q(t))\,dt\right), \quad (5)$$

Internal boxed mutations occur with complementary rate, so that:

$$P_{\text{box}}(L_i) = \mu - P_{\text{circ}}(L_i). \quad (6)$$

Finally, $P_{\text{free}}(\tilde{B})$ corresponds to the probability that no other event occurs on the part of the backbone tree upstream of the terminal mutations (denoted by $\tilde{B}$):

$$P_{\text{free}}(\tilde{B}) = \exp\left(-\int_{\tilde{B}}\left[\mu\frac{p(t)}{q(t)} + \lambda q(t)\right]dt\right), \quad (7)$$

where the integral runs over all branches of $\tilde{B}$. The first term describes the rate of mutations leading to surviving sublineages, conditioned on the predivision branch surviving and having at least one further mutation. The second term corresponds to the rate of divisions (3).

## Inference procedure

Starting from one or multiple allelic trees $A$ generated from a biological process, we seek to infer the true rates $\boldsymbol{\theta}^* = (\lambda^*, \delta^*, \mu^*)$ governing the dynamics of division, death and mutation. Our goal is to maximize $P(A|\boldsymbol{\theta})$ with respect to the rates, thereby obtaining their maximum-likelihood estimate (MLE) $\hat{\boldsymbol{\theta}} = (\hat{\lambda}, \hat{\delta}, \hat{\mu})$. Since we lack an analytic expression for the likelihood of $A$, we use an expectation maximization (EM) algorithm, in which the backbone tree $B$, encoded by the node depths $s$ and mutation times $m$, $t$, constitute our hidden variables, to iteratively find $\hat{\lambda}$, $\hat{\delta}$, and $\hat{\mu}$ (Fig. 3).

To motivate our choice to use EM recall that for an allelic tree $A$, there is an infinite number of backbone trees $B \in a_b^{-1}(A)$ satisfying the equation $A = a_b(B)$. The probability of $A$ given the rates can be formally written as

$$P(A|\boldsymbol{\theta}) = \int_{a_b^{-1}(A)}P(B, A|\boldsymbol{\theta})\,dB. \quad (8)$$

Note that since $A$ is entirely determined by $B$ through $A = a_b(B)$, we have that $P(B, A|\boldsymbol{\theta}) = P(B|\boldsymbol{\theta})$. The backbone $B$ in (8) contains the times at which mutations and divisions occur, i.e. hidden variables which are in theory integrated over to obtain the left-hand side of the equation. In practice, however, (8) cannot be computed analytically, and involves too many dimensions in the integral to be realistically computed numerically. We therefore use an expectation maximization (EM), a two-step iterative algorithm, to approximate the MLE.

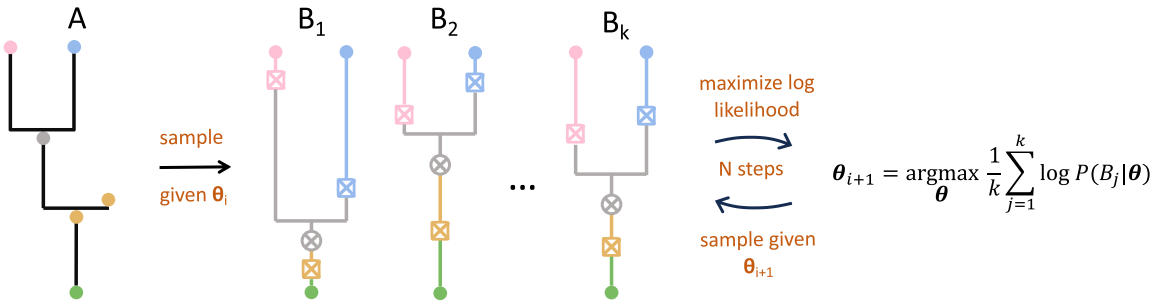


**Fig. 3.** Expectation–maximization. Schematic of the inference scheme based on expectation maximization for one allelic tree $N = 1$. Given the current parameter estimates $\boldsymbol{\theta}_i$, backbone trees $\{B_1, B_2, \ldots, B_k\}$ are sampled from $P(B|A, \boldsymbol{\theta}_i)$ using a Monte–Carlo Markov chain. Their mean log likelihood (expectation) is maximized (maximization) with respect to $\boldsymbol{\theta}$, yielding the next parameter estimates $\boldsymbol{\theta}_{i+1}$. These expectation and maximization steps are repeated a total of $N_{MC}$ times.

In the first step, referred to as the "expectation" step, we calculate the expected value of the log likelihood of $\boldsymbol{\theta}$ taken with respect to the current parameter estimates $\boldsymbol{\theta}_i$ (Fig. 3):

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_i) = \mathbb{E}_{B\sim P(B|A,\boldsymbol{\theta}_i)}[\log P(B|\boldsymbol{\theta})]. \tag{9}$$

In the maximization step, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_i)$ is maximized with respect to $\boldsymbol{\theta}$, yielding the parameter estimates for the next iteration

$$\boldsymbol{\theta}_{i+1} = \text{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_i). \tag{10}$$

The log-likelihood term inside the expectation can be computed for each backbone tree using (1). However, the expectation value requires integrating over all backbone trees $B$, which is not practical. Instead, we evaluate that expectation value using Markov chain Monte Carlo (MCMC), which only requires to be able to compute $P(B|A, \boldsymbol{\theta}_i) \propto P(B|\boldsymbol{\theta}_i)$ up to a constant of $B \in a_b^{-1}(A)$. Specifically, we generate a sequence of samples using the Metropolis Hastings algorithm. Given a backbone $B$, a new backbone $B' \in a_b^{-1}(A)$ is proposed by sampling the depth or time of a random node or mutation from a uniform distribution with the condition that $a_b(B') = A$. $B'$ is then accepted with probability

$$\alpha(B', B) = \min\left(1, \frac{P(B'|\boldsymbol{\theta}_i)}{P(B|\boldsymbol{\theta}_i)}\right), \tag{11}$$

where we have dropped the dependence on $A$ since we sample backbone trees which are strictly consistent with the allelic tree. Starting from a random initializing backbone $B_0$, we run the Metropolis Hastings algorithm for a burn-in period of $50 \times (2n + \ell - 1)$, where $2n + \ell - 1$ is the total number of variables $s, t, m$ in $B$. After this period, we sample $k$ backbone trees $(B_1, \ldots, B_k)$, and we approximate (9) by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_i) \approx \frac{1}{k}\sum_{j=1}^{k} \log P(B_j|\boldsymbol{\theta}). \tag{12}$$

A total of $N_{MC}$ expectation maximization steps are performed, yielding a series of estimates $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{MC}}$. Since at each expectation step we only sample a finite number $k$ of backbones, we do not expect the algorithm to converge exactly to the MLE $\boldsymbol{\theta}_{MLE}$. Instead the estimates $\boldsymbol{\theta}_i$ are expected to fluctuate about $\boldsymbol{\theta}_{MLE}$ after a relaxation period $N_f$, which empirical evidence suggests is short. We report our estimate of the parameters as an average over $\boldsymbol{\theta}_i$ following that relaxation:

$$\hat{\boldsymbol{\theta}} = \frac{1}{N_{MC} - N_f}\sum_{i=N_f+1}^{N_{MC}} \boldsymbol{\theta}_i. \tag{13}$$

We can generalize the inference procedure to take more than one allelic tree as input. This corresponds to a situation where one observes several replicates of the same evolutionary process but with different realizations. While this is unusual in traditional genetics, it is relevant to many processes of somatic evolution. For instance, for B cells, this would correspond to distinct lineages of cells coming from distinct progenitors, sampled from different germinal centers, but subject to the same antigen. Consider a set of allelic trees $(A_1, A_2, \ldots, A_N)$ all generated by processes with the same true rates $\boldsymbol{\theta}^*$. The likelihood of the allelic trees then becomes

$$P(A_1, A_2, \ldots, A_N|\boldsymbol{\theta}) = \int \prod_{u=1}^{N} P(B_u, A_u|\boldsymbol{\theta}) dB_u, \tag{14}$$

and the expected value of the log likelihood of $\boldsymbol{\theta}$ must be taken with respect to all backbones

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_i) = \sum_{u=1}^{N} \mathbb{E}_{B\sim P(B|A_u,\boldsymbol{\theta}_i)}[\log P(B|\boldsymbol{\theta})]. \tag{15}$$

We create different independent realizations of our sampling scheme for each allelic tree, obtaining for each $A_u$ a set $\{B_{u1}, B_{u2}, \ldots, B_{uk}\}$ sampled from $P(B|A_u, \boldsymbol{\theta}_i)$. The approximation to equation (15) is given by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_i) \approx \frac{1}{k}\sum_{u=1}^{N}\sum_{i=1}^{k} \log P(B_{ui}|\boldsymbol{\theta}). \tag{16}$$

The maximization step does not change.

## Validation of inference procedure

We test our inference method on synthetic allelic trees generated with true parameters $(\lambda^*, \delta^*, \mu^*)$, and with comprehensive sampling $(C = C_f)$. Practically, we generate extant cell lineage trees $C$ using the coalescent point process (CPP) (Supplementary Fig. 1) (Lambert and Stadler 2013), and apply the function $a : C \to A$ to each of them to create a synthetic dataset of allelic trees $A$. The generation process along with a brief introduction to the CPP is detailed in Supplementary Section 1.

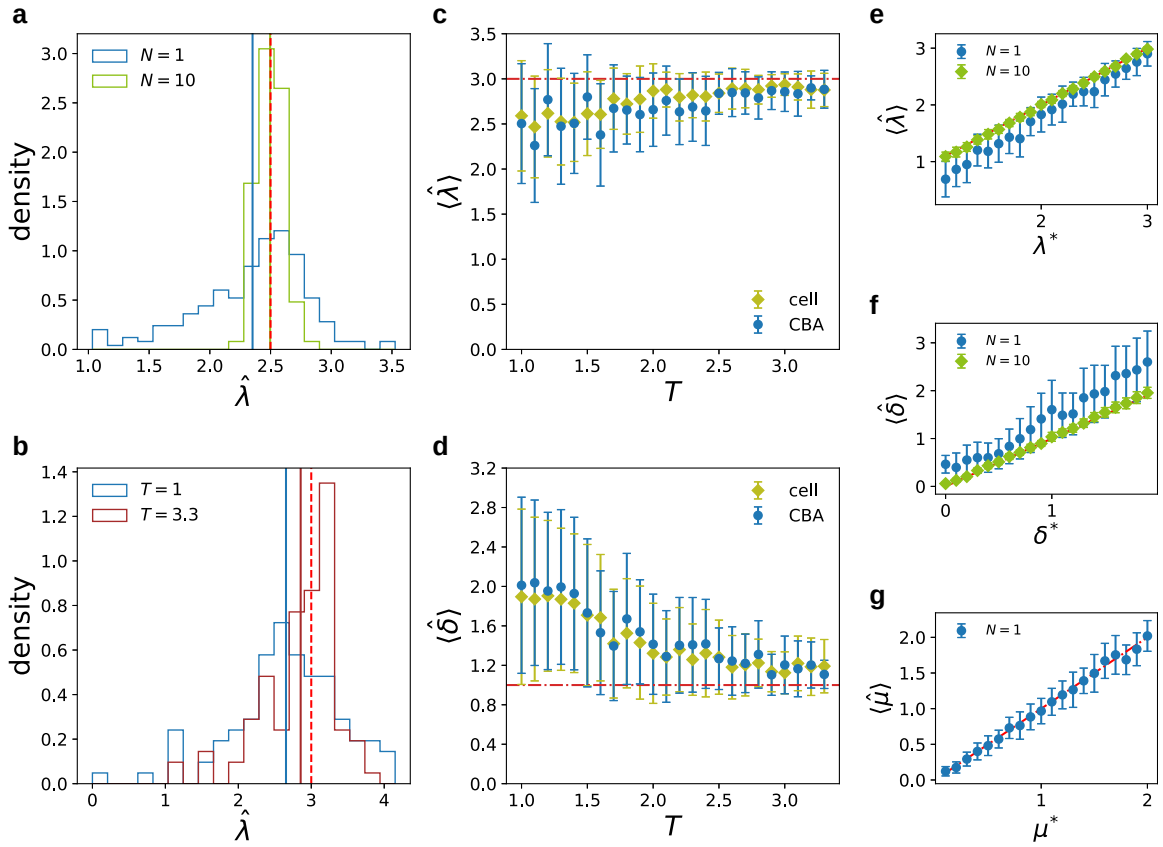


**Fig. 4.** Validation of the inference method on synthetic data. Trees are generated with true parameters $\lambda^*$, $\delta^*$, $\mu^*$. The algorithm is used to infer one of the parameters, given the true values of the other two. Each numerical experiment is repeated 100 times to collect statistics. a) Distribution of $\hat{\lambda}$ inferred from $N = 1$ or 10 trees for $\lambda^* = 2.5$ (dashed line), $\delta^* = 1$, $\mu^* = 1$, and $T = 3$. Solid vertical lines show $\langle\hat{\lambda}\rangle$. Estimates are less biased and more precise for larger $N$. b) Same as (a), but with $\lambda^* = 3$, $\delta^* = 1$, $\mu^* = 1$, $N = 1$ and two values of $T$: $T = 1$, and $T = 3.3$. Longer times lead to larger trees, which allow for better inference. c, d) Average inferred (c) $\langle\hat{\lambda}\rangle$ and (d) $\langle\hat{\delta}\rangle$ and their standard deviations (represented by error bars) vs. $T$, using either inference on the allelic tree $A$ (dots), or the extant cell tree $C$ (diamonds). In (c) $\delta^* = 1$ and $\mu^* = 1$ while in (d) $\lambda^* = 3$ and $\mu^* = 1$. The bias goes down as the tree size increases with $T$. This bias is still present when using the full information of the extant cell tree, although it is smaller. e) Average inferred birth rate $\langle\hat{\lambda}\rangle$ vs its true value $\lambda^*$ while keeping $\delta^* = 1$, $\mu^* = 1$, and $T = 3$ constant, for $N = 1$ (dots) or 10 (diamonds). The bias goes down as $\lambda$ and thus the size of the tree increases. It disappears for $N = 10$. f–g) Same for the death and mutation rates, keeping $\lambda^* = 2$, $T = 3$ and $\delta^* = 1$ while varying the mutation rate, and $T = 3$, $\mu^* = 1$ while varying the death rate. The bias is lowest when $\delta$ is small, and the trees the largest. There is no apparent bias in $\langle\hat{\mu}\rangle$ even at $N = 1$. Error bars are given by the standard deviation of multiple inferences.

We first test the inference for a single parameter given the true values of the other two and the total time $T$, beginning with the birth rate $\lambda$. In Fig. 4a, we show the distribution of birth rates inferred from a set of $N$ allelic trees, where $N$ is varied from 1 to 10, with $\lambda^* = 2.5$, $\delta^* = 1$, $\mu^* = 1.8$, $T = 3$, and $k = 1$. Although the method theoretically is valid for large $k$, when the learning rate (the rate with which the gradient is followed in gradient descent methods) is slow, the ensemble average over $k$ is emulated by an average over many learning steps around convergence. We checked that increasing $k$ up to 10 did not affect the results of the inference (Supplementary Fig. 4).

On average, when $N = 1$, the inferred birth rate is very noisy and underestimated, i.e. $\langle\hat{\lambda}\rangle < \lambda^*$. The large variance is a result of the large variability in the number of leaves of sampled extant cell trees, combined with fluctuations in both the number of observed branches and their lengths due to the random spread of mutations on $C$. Both bias and variance are due to finite sized effects, since maximum-likelihood estimates are only guaranteed to converge asymptotically to the true parameter values in the limit of infinite data $N \to \infty$. Indeed, $\langle\hat{\lambda}\rangle$ gets much closer to the true value $\lambda^*$ when $N$ is increased to 10.

In our context, the large data limit may be directly achieved in two ways: by pooling more trees together (large $N$), or effectively by increasing the size of each tree, and with it the number of informative birth and mutation events that the MLE can exploit. The number of leaves $n_C$ in the tree $C$ of extant cells is on average:

$$\mathbb{E}(n_C) \approx e^{(\lambda-\delta)T}. \tag{17}$$

For a given mutation rate, the number of types in an allelic tree is heavily influenced by the number of leaves $n_C$ in the extant cell tree it came from. In turn, the number of types is linked to how much information the allelic tree $A$ carries about the parameters. Trees with more leaves contain more information and thus lead to better estimates for the rates. It is reasonable to consider the number of leaves in $C$ as a proxy for the signal in $A$.

According to (17), larger trees may be obtained by either increasing the difference $\lambda - \delta$, or by increasing $T$. Figure 4b shows the distribution of inferred $\hat{\lambda}$ for $\lambda^* = 3$, $\delta^* = \mu^* = 1$, and $T = 1$ or $T = 3.3$ with $N = 1$. As expected, much better and less biased inference is achieved on larger trees ($T = 3.3$) than on smaller ones ($T = 1$). We then plotted the average inferred $\langle\hat{\lambda}\rangle$, as well as its standard deviation, as a function of $T$ (Fig. 4c). While $\lambda^*$ is always underestimated, the bias and standard deviation decrease as a function

of $T$. Conversely, $\hat{\delta}$ gets overestimated, and that bias as well as the uncertainty also shrink with $T$ and the tree size (Fig. 4d).

To investigate the origin of these biases in $\hat{\lambda}$ and $\hat{\delta}$, we asked whether inference with full information from the extant cell tree would also suffer from a similar bias. To do so, we used the original cell trees $C$ from which the synthetic allelic trees were generated (and which are known since we are dealing with synthetic data), and inferred the rates $\lambda$ and $\delta$ with a maximum-likelihood estimator, using analytic formulas for the probability $P(C)$ of the realized cell tree from the CPP (Lambert and Stadler 2013). Details of the used formulas are given in Supplementary Section 1. Results for the average inferred parameters and their standard deviations are given by the diamond points in Fig. 4c and d. They show that the bias still exists, albeit slightly less on average than CBA inference on allelic trees. It also decreases as the tree size grows by increasing $T$. The smaller bias is consistent with the fact that cell trees contain more information than allelic trees, including mutation and division times as well as abundances of cells of the same type. But the small difference between the two estimates suggests that the allelic tree encode significant information about $C$. However, the fact that the bias is still present points to a more fundamental origin, for which we propose a plausible explanation below. We are not aware of any previous work in macroevolution or other fields discussing biases in the MLEs of rates obtained from sampled cell trees.

The overestimation of $\delta$, and related underestimation of $\lambda$, may be understood intuitively by considering a simpler case, where the observation is simply the number of leaves $n_C$ in $C$. In that case, the MLE estimator of $\lambda$ would be $\hat{\lambda} = \ln(n_C)/T + \delta^*$, and the MLE estimator of $\delta$ would be $\hat{\delta} = \lambda^* - \ln(n_C)/T$ (see Supplementary Section 3 for a derivation). Since $\langle \ln(n_C) \rangle \leq \ln\langle n_C \rangle = (\lambda^* - \delta^*)T$ by Jensen's inequality, this implies $\langle \hat{\lambda} \rangle \leq \lambda^*$ and $\langle \hat{\delta} \rangle \geq \delta^*$, consistent with our observations. While the inference on the full $C$ or on the backbone tree $B$ contain different sorts of information than just $n_C$, such an effect may affect our MLE estimator. Inspired by the previous argument, we have shown in Supplementary Fig. 5 that when estimating $\hat{\lambda}$ or $\hat{\delta}$ with CBA, calculating $\langle \exp(\hat{\lambda} - \delta^*)T \rangle$ or $\langle \exp(\lambda^* - \hat{\delta})T \rangle$ provides an unbiased estimate of the average number of leaves $\langle n_c \rangle$.

We then study how the inference performs as a function of the parameters. We vary $\lambda^*$ across a given range, keeping $\delta^* = 1$ and $\mu^* = 1$ constant and known, and estimate the average MLE inferred $\langle \hat{\lambda} \rangle$ for a large number of groups of $N$ trees as a function of $\lambda^*$, for $N = 1$ and $N = 10$ and $T = 3$ (Fig. 4e). For $N = 1$, $\lambda^*$ is consistently underestimated, with smaller $\lambda^*$ leading to both larger biases in $\langle \hat{\lambda} \rangle$ and higher uncertainty. This is expected, since for constant $\delta^*$ and $\mu^*$ a higher $\lambda^*$ leads to larger and more informative trees. Increasing the number of trees to $N = 10$ for each inference is sufficient to remove the bias in $\langle \hat{\lambda} \rangle$ (Fig. 4e).

To better understand how the amount of data impacts the accuracy of the estimates, it is interesting to compare the cases of a single large tree ($\lambda^* = 3.0$, $N = 1$) to many small trees ($\lambda^* = 1.1$, $N = 10$), all with $\mu = \delta = 1$ and $T = 3$. We have $n_C(\lambda = 1.1) \approx e^{0.3} \approx 1.35$ and $n_C(\lambda = 3) \approx e^6 \approx 403$. We would expect the signal in a single tree with $\lambda = 3$ to be much larger than 10 times the signal in a tree with $\lambda = 1.1$. However, Fig. 4b shows that we obtain a better estimate of the birth rate when $N = 10$ and $\lambda^* = 1.1$. This indicates that the number of allelic trees used for the inference is much more important than their sizes.

We also varied the other parameters, by applying maximum likelihood to infer a range of $\delta$ and $\mu$. We vary $\delta^* \in [0, 1.9]$, keeping $\lambda^* = 2$ and $\mu^* = 1$ fixed, and $T = 3$. Figure 4f shows the relationship between $\langle \hat{\delta} \rangle$ and $\delta^*$. Inferring the death rate of reconstructed phylogenetic trees is generally considered a harder task than inferring the birth rate, since dead cells are not represented in the data. Comparing the results for $N = 1$ in Fig. 4b and 4c, it is clear that $\langle \hat{\delta} \rangle$ is more biased and more noisy than $\langle \hat{\lambda} \rangle$. Consistent with previous observations, $\langle \hat{\delta} \rangle$ is consistently overestimated, although that bias disappears for $N = 10$. Finally, in Fig. 4g we show the average inferred mutation rate $\langle \hat{\mu} \rangle$ at fixed $\lambda^* = 2$ and $\delta^* = 1$ for $N = 1$. In that case, a single allelic tree is sufficient to obtain an unbiased estimator for the mutation rate. We also verified that inference of other parameters, such as $\lambda$, converges to the result of the inference performed on the cell tree, since we expect $A$ to provide a good approximation of $C$ when $\mu^*$ is large (Fig. 5).

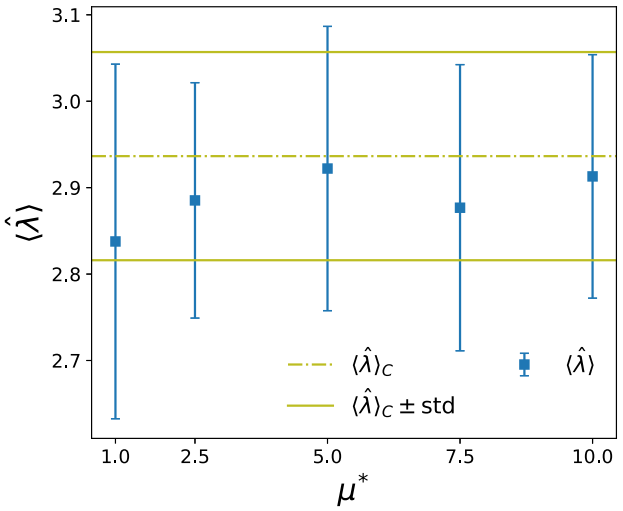


**Fig. 5.** Dependence of birth rate inference on the mutation rate. Average $\hat{\lambda}$ inferred as a function of the mutation rate for $\lambda^* = 3$, $\delta^* = 1$, $T = 3$ and $N = 1$. Larger mutation rates imply larger allelic trees $A$ that more closely resemble $C$. As $\mu^*$ increases we note an increase in $\hat{\lambda}$, asymptotically approaching the inferred average birth rate $\langle \hat{\lambda} \rangle_C$ using $C$. Numerical experiments are repeated 200 times to gather statistics. Error bars are given by the standard deviation of multiple inferences.

In most situations, only a small fraction $\rho$ of cells are sampled and used to collect data. In that case, our inference formulas (1)–(7) are still valid, but with an additional dependency of the $p(t)$ and $q(t)$ function on $\rho$ (see Supplementary Sections 1 and 2). Performing the inference on subsampled trees with $\rho = 10\%$ still works, although accuracy is slightly degraded, as expected from the loss of information caused by subsampling (Fig. 6).

Next, we consider the task of simultaneously inferring $\lambda$ and $\delta$, given $\mu^*$ and $N = 1$ and $T = 3$. To get a sense of the difficulty of the task, we first study inference on synthetic backbone trees. Our method is ultimately designed to make inference on allelic trees, but since backbone trees contain more information, we reason that issues observed in the inference on backbone trees will be inherited by allelic trees. In Fig. 7a, we show a heat maps of the likelihood $P(B|\lambda, \delta, \mu^*)$ of a backbone tree $B$ as a function of $\lambda$ and $\delta$, where $B$ was sampled from $P(B|\lambda^*, \delta^*, \mu^*)$ with $\lambda^* = 3$, $\delta^* = 1$, $\mu^* = 1$, and $T = 3$. The likelihood has a well-defined maximum, but is flat in the difference $r = \lambda - \delta$. This implies that the effective growth rate $r$ is easier to infer than the individual rates $\lambda$ and $\delta$. Louca and Pennell (2021) report that many studies inferring evolutionary rates from reconstructed trees, which they call "extant timetrees," report extinction rates of zero, in direct contradiction with data obtained from fossil records. They attribute this

discrepancy to the difficulty in inferring the exact rates $\lambda$ and $\delta$ rather than $r$.

We then applied the EM algorithm to infer both $\lambda$ and $\delta$ simultaneously from allelic trees. As $T$ is increased, estimates for $\lambda$ and $\delta$ improve both in bias and in variance. This is evidenced by Fig. 7b, where the distribution of $\hat{\lambda}$ is shown for $T = 2$ and $T = 4$, as well as by Fig. 7c, in which we plot the dependence of $\langle\hat{\lambda}\rangle$ and $\langle\hat{r}\rangle = \langle\hat{\lambda} - \hat{\delta}\rangle$ on $T$ for $N = 1$ and $N = 10$. Curiously, when $N = 1$ the birth rate $\lambda$ is consistently overestimated, in contrast to the single-parameter inference case. However, the effective growth rate $r = \lambda - \delta$ is still underestimated, consistent with the convexity argument above. Increasing the number of trees used per simulation to $N = 10$ is enough to provide unbiased estimates of the net growth rate for all values of $T$ considered. Except for $T = 2$, birth rate estimates are also unbiased.

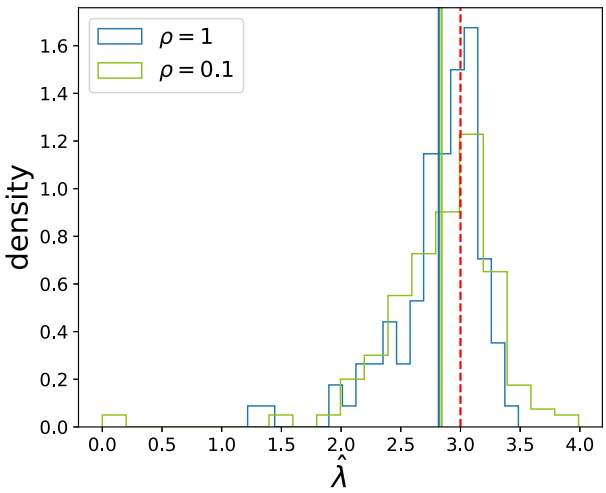


**Fig. 6.** Inference with sampling. Distribution of $\hat{\lambda}$ inferred from $\rho = 1$ and $\rho = 0.1$ for trees generated with $\lambda^* = 3$ (dotted line) and $N = 1$, given $\delta^* = 1$, $\mu^* = 1$ and $T = 3$. No significant change in the average birth rate $\langle\hat{\lambda}\rangle$ (dotted lines) is observed when under sampling. We note that the distribution is wider when $\rho = 0.1$, indicating noisier inference. The distributions are the results of 200 repetitions of the numerical experiments.

## Comparison to other methods

We are not aware of other methods designed to infer rates from allelic trees. However, existing software with broad applicability can be utilized to perform that task. We compare our results to those obtained using the Bayesian evolutionary analysis by sampling trees (BEAST2) (Bouckaert *et al.* 2014) software for the task of simultaneously inferring the birth and death rates $\lambda$ and $\delta$, given the mutation rate $\mu^*$. To do so, we first need to reformulate our problem so it can be handled by BEAST2. In contrast to our inference scheme which begins from the allelic tree A, BEAST2 takes as input a multiple sequence alignment (MSA) $\mathcal{M}$. It then samples from the posterior $P(\boldsymbol{\theta}, A'|\mathcal{M}) \propto P(\mathcal{M}|\boldsymbol{\theta}, A')P(\boldsymbol{\theta}, A')$, i.e. the joint distribution of parameters $\boldsymbol{\theta}$ and trees $A'$ given $\mathcal{M}$, using a Metropolis–Hastings algorithm. The sampled trees $A'$ differ from our allelic trees in two ways, as illustrated in Supplementary Fig. 3. $A'$ is time-resolved, meaning that multifurcations are timed and broken up into bifurcations. Also, the root of $A$ (green node in Supplementary Fig. 3) is not contained in $A'$, and $A'$ starts with the most recent common ancestor.

To generate the input data $\mathcal{M}$ usable by BEAST2, we first generate an allelic tree A as before, and create a set of synthetic sequences corresponding to the leaves of A, and consistent with its topology (see Supplementary Section 4). In addition, we specify priors on all the different subsets of sequences which have a common ancestor in A to ensure that all the trees sampled during the inference performed by BEAST2 are constrained to have the same topology as A. By feeding $\mathcal{M}$ and subsets of leaves with common ancestors to BEAST2, we infer the estimates $\hat{\lambda}_b$ and $\hat{\delta}_b$ as the mean values from the posterior $\hat{\boldsymbol{\theta}}_b = \int d\boldsymbol{\theta}\, \boldsymbol{\theta} P(\boldsymbol{\theta}, A'|\mathcal{M})$. These estimates are then compared with the $\hat{\lambda}$ and $\hat{\delta}$ inferred using CBA on A.

Figure 8 shows a comparison of our MLE estimator with the estimates of BEAST2. We observe that for small mutation rates, our approach is consistently unbiased, while BEAST2 systematically underestimates the birth rate even at large mutation rates, and only converges towards the true value of the parameters as $\mu$ becomes large. The underperformance of BEAST2 is expected because we applied it to an MSA from which duplicates were removed, since abundance information was discarded for the purpose of comparing with CBA. Since BEAST2 is not designed to work with data with no abundance information, we applied it as if the

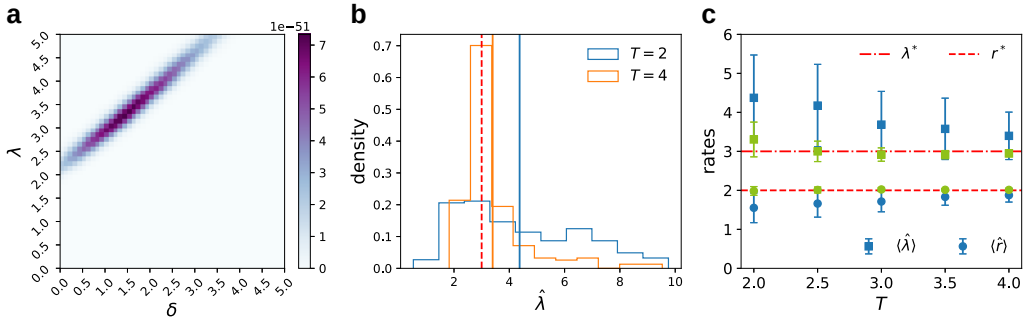


**Fig. 7.** Two-dimensional inference. a) Heatmap representing the normalized likelihood $P(C|\boldsymbol{\theta})/\int d\boldsymbol{\theta} P(B|\boldsymbol{\theta})$ as a function of two unknown parameters $\lambda$ and $\delta$ for $N = 1$ with known $\mu^* = 1$ and $T = 3$, for a backbone tree $B$ drawn with $\lambda^* = 3$, $\delta^* = 1$. The likelihood landscape shows that while the difference $r = \lambda - \delta$ is well represented by the tree, the value of each $\lambda$ and $\delta$ are very uncertain. Note that this inference on $B$ is different from the inference on $A$ we are interested in, but it allows for an explicit computation of the likelihood which our method on $A$ does not permit, giving us insight by proxy. b) Distribution of inferred $\hat{\lambda}$ in the two-dimensional EM search for $\hat{\lambda}$ and $\hat{\delta}$, with the same parameters as above, but for $T = 2$ and $T = 4$. Longer times allow for more accurate inference. c) Mean $\langle\hat{\lambda}\rangle$ (squares) and $\langle\hat{r}\rangle = \langle\hat{\lambda} - \hat{\delta}\rangle$ (dots) as a function of age $T$, for the same parameters as above, but for $N = 1$ and $N = 10$ (lighter shade, closer to true values). Note that the effective growth rate $\hat{r}$ is underestimated for $N = 1$, consistent with previous observations on the single-parameter inference. In contrast, $\hat{\lambda}$ is now overestimated. For each $T$ and for $N = 1$ ($N = 10$), we perform 200 (100) independent simulations to gather statistics. Error bars are given by the standard deviation of multiple inferences.

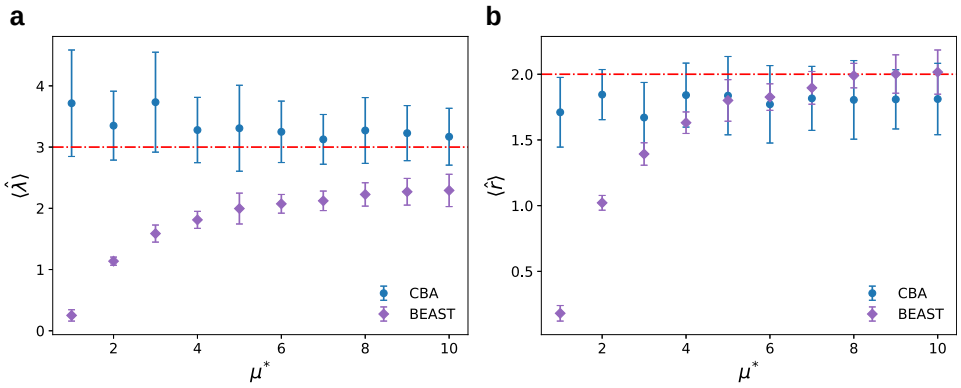


**Fig. 8.** Comparison with BEAST (Bouckaert *et al.* 2014). Average inferred a) birth rate $\langle\hat{\lambda}\rangle$ and b) effective growth rate $\langle\hat{r}\rangle$ as a function of the mutation rate $\mu^*$, for both our method and the posterior average in BEAST, in the two-dimensional inference of $\lambda^* = 2.8$ and $\delta^* = 1$ with known $\mu^*$, from a single allelic tree $A$ ($N = 1$). In order to keep the size of the allelic trees comparable across mutational regimes (number of types $\approx 100$), the stem age of the tree $T$ is decreased for increasing $\mu^*$. Our method performs best compared with BEAST at low mutation rates, where branch lengths are a poor approximation of time. At large mutation rates, both methods perform well. In both (a) and (b), error bars are given by the standard deviation of 100 independent inferences.

abundance of each type was 1, and as if the allelic tree $A$ was equivalent to the cell tree $C$. Consequently, it greatly underestimates the birth rate. As $\mu^*$ is increased, $A$ becomes closer to $C$, which improves the BEAST2 birth and death rate estimates.

That comparison, however, is not ideal, because BEAST2 is not designed to deal with discarded or unreliable abundances. If the abundances of each observed type in $A$ were known, or could at least be reliably estimated from data, then the MSA would better reflect the original tree $C$ and BEAST2 would return better estimates $\hat{\lambda}$ and $\hat{\delta}$ than CBA, by using that additional information. In some experimental settings, however, abundances of types cannot be measured accurately and are subject to a lot of noise. We asked how that noise would affect the performance of BEAST2. Supplementary Fig. 6 shows the average inferred estimates $\hat{\lambda}$ (Supplementary Fig. 6a) and $\hat{r}$ (Supplementary Fig. 6b) for varying degrees of noise in type abundances, implemented by increasing the variance $\sigma^2$ of a negative binomial distribution (see Supplementary Section 4). The birth rate estimate $\hat{\lambda}$ has a positive bias that increases with $\sigma^2$ (Supplementary Fig. 6a) due to the strong variability in the size of subtrees of $C$ near the sampling time, which can be accounted for by allowing both $\lambda$ and $\delta$ to increase, such that recent sublineages have had little time to go extinct. The error in the inference of both $\lambda$ and $r$ (shown by the error bars) steadily increases with $\sigma^2$.

Delphy (Varilly *et al.* 2025) is a recently developed software based on Bayesian inference which, in contrast to BEAST2, works by sampling "explicit mutation-annotated trees" (EMATs). EMATs are timed bifurcating trees with branches containing time-stamped mutations and "missation" events, i.e. points on the tree where a site is missing data in all descending observed sequences. This reformulation of classical Bayesian phylogenetics allows for remarkable increases in speed when mutations in the data are sparse, i.e. when the sequences considered are close in terms of genetic distance, while maintaining state-of-the-art results on various epidemiology datasets. Backbone trees are essentially EMATs with no information on abundances and no "missation" events. While CBA samples backbone trees consistent with an allelic tree assumed to be accurately inferred from data, Delphy samples EMATs consistent with a multiple sequence alignment. Although we have not explicitly tested Delphy on our synthetic data, we believe that it would suffer from similar problems as BEAST2, whereby missing abundances cause growth rates to be underestimated.

## Discussion

We introduced a novel method designed to infer evolutionary rates of a growing population of cells sampled at a single time point from an allelic tree $A$ describing mutational histories. We derived the likelihood of the backbone tree $B$, which enabled us to come up with an efficient inference scheme based on expectation maximization to find the MLEs of the rates. We limited our analysis to rates $\boldsymbol{\theta} = (\lambda, \delta, \mu)$ that were homogenous in time, which allowed for the separate inference of $\lambda$ and $\delta$. All the formalism and derivations introduced are easily generalizable to rates with an explicit time dependence $\boldsymbol{\theta}(t) = (\lambda(t), \delta(t), \mu(t))$ (see (Lambert and Stadler 2013) for a generalization of Supplementary equations 1–2). However, doing so would require parameterizing these dependences with additional parameters, requiring more data to support the inference.

We derived the method in a general framework, since applying it to concrete systems often requires checking the validity of the assumptions. Firstly, we assumed that mutations accumulate continuously over time, instead of upon birth. Both situations occur in cells. Conversely to what we assumed somatic mutations can occur during DNA replication, however they can also occur continuously throughout the cell cycle due to DNA repair errors which lead to mismatches that are incorrectly resolved (Spisak *et al.* 2024). On the discrete side, some single-based substitution (SBS) signatures are known to increase with cellular turnover (e.g. SBS1) (Abascal *et al.* 2021), indicating that errors in DNA replication occurring at discrete times are responsible. Yet other signatures are thought to be mainly caused by errors in DNA repair, owing to their quasi-continuous clock-like accumulation in post mitotic cells (Spisak *et al.* 2024). In the example of affinity maturation, B cells somatically hypermutate when the Activation-Induced cytidine Deaminase (AID) deaminates a cytidine to a uracil, creating mismatches which can be incorrectly repaired by the error prone DNA repair mechanism (Victora and Nussenzweig 2022). These processes usually are assumed to accumulate continuously, although recent work shows that it can depend on the cell cycle (Pae *et al.* 2021). Others have made both continuous and discrete-time assumptions for hematopoietic stem cells. The cloneRate inference method (Johnson *et al.* 2023) infers the net growth rates of cell colonies derived from single hematopoietic stem cells (Van Egeren *et al.* 2021; Fabre *et al.* 2022; Mitchell *et al.* 2022; Williams *et al.* 2022 ) under the same assumption that mutations accumulate through time at a constant

rate, while another recent method (Dieselhorst and Berg 2024) analyzes two hematopoeisis phylogenies from neonate donors (Mitchell *et al.* 2022) under the assumption that mutations occur upon division with a Poisson distribution.

In contrast to BEAST (Bouckaert *et al.* 2014), we do not discuss the task of inferring $A$ from a multiple sequence alignment (MSA)—an active field of research (McKenna *et al.* 2016; Schwartz and Schäffer 2017; DeWitt III *et al.* 2018; Feng *et al.* 2021)—as our inference scheme starts from the allelic tree $A$. We thus assume that data are in a regime where the inference of $A$ is accurate with no degeneracy. We expect this assumption to be true when the total number of mutations is low compared to the length of sequences in the MSA (mutation frequency). For example, the mean nucleotide mutation frequency in B-cell receptor V genes is known to be approximately 5% both in healthy and chronically HIV-1 or HCV-infected individuals (Lupo *et al.* 2022; Kreer *et al.* 2023), which should guarantee reliable allelic tree inference. In addition, a likelihood-based method incorporating genotype abundances, obtained from single-cell data, was shown to accurately lift the degeneracy in B-cell phylogenies inferred using maximum parsimony (DeWitt III *et al.* 2018). In hematopoetic stem cells, whole-genome sequencing shows that they accumulate on average 17 somatic mutations per year (Williams *et al.* 2022). An 80 year old individual would thus have an approximate mean nucleotide frequency of 0.00005%, assuming $\sim 3$ billion base pairs in the genome, making inference of $A$ unambiguous.

Most evolutionary processes are subject to selection and a lot of work has gone into identifying signatures of selection. Cancers are caused by the accumulation of driver mutations which confer selective advantages to cells in which they appear, while viral-immune coevolution places selective pressure on both viruses and the immune system (Cobey *et al.* 2015; Marchi *et al.* 2021). Methods based on the ratio of synonymous to nonsynonymous mutations (dN/dS) can efficiently identify the average number of driver mutations in different tumors (Martincorena *et al.* 2017). For the seasonal A/H3N2 viruses, assuming evolution proceeds through the continual accumulation of weakly deleterious or beneficial mutations, the branching patterns of time resolved phylogenies can be used to infer the fitness of the sampled strains (Neher *et al.* 2014). Many viral and macroevolutionary phylogenies benefit from branches containing a large number of mutations, allowing for fitness dynamics to be approximated by Brownian motion (Neher *et al.* 2014). In the case of B-cell immunoglobulin (Ig) receptors, due to the low average number of mutations per sequence (Briney *et al.* 2019), this approximation cannot be made. Recent efforts in quantifying selection have focused on comparing the frequency of nonsynonymous mutations to their expected frequency assuming neutrality (Uduman *et al.* 2011; Yaari *et al.* 2015), with the latter being difficult to calculate. Inferring signatures of selection from allelic trees has been done for B cells sampled from individuals subjected to the trivalent influenza vaccine (Horns *et al.* 2019) or with chronic HIV-1 infections (Nourmohammad *et al.* 2019) by focusing on statistics such as tree branching asymmetry, terminal branch length distributions and the site frequency spectrum. Multitype birth–death processes (MTBD) model selection by allowing for birth and death rates to change discretely along phylogenies (Barido-Sottani *et al.* 2020). They have been used for Bayesian inference in macroevolution (Barido-Sottani *et al.* 2020), whereby time-resolved phylogenies are sampled along with rates using molecular clock models. We are not aware of any existing methods which infer evolutionary rates using MTBD models for allelic trees when the molecular clock hypothesis breaks down, e.g. for B-cell phylogenies. The generalization of our model past neutrality suffers from our use of coalescent point processes to derive the likelihood of the backbone, specifically through the terms $p(t)$ and $q(t)$ (Supplementary equations 11 and 13). The chief condition for reconstructed trees of birth–death models with rates $\lambda(t)$ and $\delta(t)$ to be CPPs is that while the rates may depend on absolute time $t$, they cannot be subtree dependent. Clearly, this assumption does not hold when selection is present, since the descendants of a cell with a higher birth rate than its ancestor will carry this change in fitness along the subtree. A possible extension for future work would be to use dynamic programing (Neher *et al.* 2014) or to use approximations to compute the likelihood of $B$ with good accuracy.

Besides the issue of selection, a direct application to B cells would require choosing appropriate time dependence for the rates $\boldsymbol{\theta}(t) = (\lambda(t), \delta(t), \mu(t))$. B-cell affinity maturation takes place in germinal centers and is spatially heterogeneous (Victora and Nussenzweig 2022). Germinal centers are divided into dark zones (DZ), where B cells somatically hypermutate and proliferate, and light zones (LZ), where B cells are selected for their affinity to an antigen (Victora and Nussenzweig 2022). Cells migrate independently between the two zones in a process known as cyclic reentry, with migration rates in mice depending on whether cells are transitioning from the LZ to the DZ or vice versa (Victora *et al.* 2010). Additionally, the low percentage of cells which reenter the dark zone indicates the presence of strong bottlenecks where most cells die or go away (Victora *et al.* 2010). To model cyclic reentry, $\lambda(t)$ and $\mu(t)$ could be defined as piecewise functions alternating intervals in which birth is high and where death is high. Increasingly complex choices for $\boldsymbol{\theta}(t)$ could lead to nonidentifiable statistical models (Louca and Pennell 2020), highlighting the importance of simple parametrizations.

In our numerical experiments, we found a consistent bias in the estimation of $\lambda$ and $\delta$, which disappears asymptotically in the large-data limit. In the single parameter inference from a few small trees, $\lambda$ was overestimated and $\delta$ was underestimated. When inferring $\lambda$ and $\delta$ simultaneously, we found that both parameters were overestimated, leading to an underestimation of their difference $r$. To circumvent this bias, we could consider extending CBA to Bayesian inference. Through clever choices for the priors over the rates $P(\lambda)$, $P(\delta)$, and $P(\mu)$, sampling from the posterior $P(\boldsymbol{\theta}, B|A)$ could lead to unbiased estimators as well as get rid of the need for the expectation maximization algorithm.

We showed in both the single and double parameter inference simulations that increasing the number of allelic trees used per inference greatly improves results. The usage of multiple trees requires the trees to come from independent but identically distributed realizations of the same process. In particular, the stem age $T$ must be the same, which in some contexts can be a challenging constraint to satisfy. For example, the stem age of trees of hematopoietic stem cells is the age of the patient at the time of sampling, which can differ significantly, from people in their 20s to 80s, in currently generated datasets. As a result, in hematopoiesis, it is not possible to infer rates using CBA on different trees. However, inference of affinity maturation rates of B cells can be performed on multiple trees at once. In B-cell affinity maturation, trees of clonal families inferred from bulk sequencing experiments are of unknown stem age and do not necessarily originate from the same affinity maturation process. When genetically identical mice are inoculated with the same antigen at the same time and their B cells are sequenced at a later but also identical time, then each germinal center is an independent and identically distributed process (i.e. stem age and true rates are the same)

(Merkenschlager *et al.* 2025; Pae *et al.* 2025 ), allowing for rate inference on multiple trees at once. Finally, in many experiments, the time $T$ at which cellular expansion began is unknown. However, prior knowledge on $T$ can easily be integrated into the inference.

## Data availability

CBA code, as well as code used to simulate the data can be found at https://github.com/ernestamandine/tree-mcmc. Supplemental material available at GENETICS online.

## Funding

This work was supported by CNRS 80 Prime Grant (A.M.W., T.M., and A.L.), the Agence Nationale de la Recherche grant no ANR-19-CE45-0018 "RESP-REP" (A.M.W., T.M., and A.A.) and the CZI Theory Initiative (A.W., T.M., and A.A.).

## Conflicts of interest

The authors declare no conflicts of interest.

## Literature cited

Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing SV, Ellis P, Russell AJC, Alcantara RE, Baez-Ortega A, Wang Y, *et al.* 2021. Somatic mutation landscapes at single-molecule resolution. Nature. 593(7859):405–410. doi:10.1038/s41586-021-03477-4.

Barido-Sottani J, Vaughan TG, Stadler T. 2020. A multitype birth-death model for Bayesian inference of lineage-specific birth and death rates. Syst Biol. 69(5):973–986. doi:10.1093/sysbio/syaa016.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. Beast 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 10(4):1–6. doi:10.1371/journal.pcbi.1003537.

Briney B, Inderbitzin A, Joyce C, Burton DR. 2019. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature. 566(7744):393–397. doi:10.1038/s41586-019-0879-y.

Cobey S, Wilson P, Matsen FA, IV. 2015. The evolution within us. Philos Trans R Soc Lond B Biol Sci. 370(1676):20140235. doi:10.1098/rstb.2014.0235.

Coorens THH, Moore L, Robinson PS, Sanghvi R, Christopher J, Hewinson J, Przybilla MJ, Lawson ARJ, Chapman MS, Cagan A, *et al.* 2021. Extensive phylogenies of human development inferred from somatic mutations. Nature. 597(7876):387–392. doi:10.1038/s41586-021-03790-y.

DeWitt WS, III, Mesin L, Victora GD, Minin VN, Matsen FA, IV. 2018. Using genotype abundance to improve phylogenetic inference. Mol Biol Evol. 35(5):1253–1265. doi:10.1093/molbev/msy020.

Dieselhorst T, Berg J. 2024. Phylodynamic inference on single-cell data. bioRxiv. 627005. doi:10.1101/2024.12.05.627005. Preprint: not peer reviewed.

Fabre MA, de Almeida JG, Fiorillo E, Mitchell E, Damaskou A, Rak J, Orrù V, Marongiu M, Chapman MS, Vijayabaskar MS, *et al.* 2022. The longitudinal dynamics and natural history of clonal haematopoiesis. Nature. 606(7913):335–342. doi:10.1038/s41586-022-04785-z.

Feng J, Dewitt WS, III, McKenna A, Simon N, Willis AD, Matsen FA, IV. 2021. Estimation of cell lineage trees by maximum-likelihood phylogenetics. Ann Appl Stat. 15(1):343–362. doi:10.1214/20-AOAS1400.

Gao Y, Feder AF. 2024. Detecting branching rate heterogeneity in multifurcating trees with applications in lineage tracing data. bioRxiv 601073. https://doi.org/10.1101/2024.06.27.601073. Preprint: not peer reviewed.

Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Holmes EC. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science. 303(5656):327–332. doi:10.1126/science.1090727.

Hoehn KB, Turner JS, Miller FI, Jiang R, Pybus OG, Ellebedy AH, Kleinstein SH. 2021. Human B cell lineages associated with germinal centers following influenza vaccination are measurably evolving. Elife. 10. doi:10.7554/eLife.70873.

Horns F, Vollmers C, Dekker CL, Quake SR. 2019. Signatures of selection in the human antibody repertoire: selective sweeps, competing subclones, and neutral drift. Proc Natl Acad Sci U S A. 116(4):1261–1266. doi:10.1073/pnas.1814213116.

Johnson B, Shuai Y, Schweinsberg J, Curtius K. 2023. clonerate: fast estimation of single-cell clonal dynamics using coalescent theory. Bioinformatics. 39(9). doi:10.1093/bioinformatics/btad561.

Jones MG, Yang D, Weissman JS. 2023. New tools for lineage tracing in cancer in vivo. Annu Rev Cancer Biol. 7(1):111–129. doi:10.1146/cancerbio.2023.7.issue-1.

Karcher MD, Palacios JA, Lan S, Minin VN. 2017. phylodyn: an R package for phylodynamic simulation and inference. Mol Ecol Resour. 17(1):96–100. doi:10.1111/men.2017.17.issue-1.

Kreer C, Lupo C, Ercanoglu MS, Gieselmann L, Spisak N, Grossbach J, Schlotz M, Schommers P, Gruell H, Dold L, *et al.* 2023. Probabilities of developing HIV-1 bNAb sequence features in uninfected and chronically infected individuals. Nat Commun. 14(1):7137. doi:10.1038/s41467-023-42906-y.

Lambert A, Stadler T. 2013. Birth–death models and coalescent point processes: the shape and probability of reconstructed phylogenies. Theor Popul Biol. 90(1):113–128. doi:10.1016/j.tpb.2013.10.002.

Lee-Six H, Øbro NF, Shepherd MS, Grossmann S, Dawson K, Belmonte M, Osborne RJ, Huntly BJP, Martincorena I, Anderson E, *et al.* 2018. Population dynamics of normal human blood inferred from somatic mutations. Nature. 561(7724):473–478. doi:10.1038/s41586-018-0497-0.

Louca S, Pennell MW. 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature. 580(7804):502–505. doi:10.1038/s41586-020-2176-1.

Louca S, Pennell MW. 2021. Why extinction estimates from extant phylogenies are so often zero. Curr Biol. 31(14):3168–3173.e4. doi:10.1016/j.cub.2021.04.066.

Lupo C, Spisak N, Walczak AM, Mora T. 2022. Learning the statistics and landscape of somatic mutation-induced insertions and deletions in antibodies. PLoS Comput Biol. 18(6):e1010167. doi:10.1371/journal.pcbi.1010167.

Marchi J, Lässig M, Walczak AM, Mora T. 2021. Antigenic waves of virus-immune coevolution. Proc Natl Acad Sci U S A. 118(27):e2103398118. doi:10.1073/pnas.2103398118.

Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ. 2017. Universal patterns of selection in cancer and somatic tissues. Cell. 171(5):1029–1041.e21. doi:10.1016/j.cell.2017.09.042.

McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, Shendure J. 2016. Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science. 353(6298):aaf7907. doi:10.1126/science.aaf7907.

Merkenschlager J, Pyo AGT, Silva Santos GS, Schaefer-Babajew D, Cipolla M, Hartweger H, Gitlin AD, Wingreen NS, Nussenzweig MC. 2025. Regulated somatic hypermutation enhances antibody

affinity maturation. Nature. 641(8062):495–502. doi:10.1038/s41586-025-08728-2.
Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 37(5):1530–1534. doi:10.1093/molbev/msaa015.
Mitchell E, Chapman MS, Williams N, Dawson KJ, Mende N, Calderbank EF, Jung H, Mitchell T, Coorens THH, Spencer DH, *et al.* 2022. Clonal dynamics of haematopoiesis across the human lifespan. Nature. 606(7913):343–350. doi:10.1038/s41586-022-04786-y.
Morlon H, Parsons TL, Plotkin JB. 2011. Reconciling molecular phylogenies with the fossil record. Proc Natl Acad Sci U S A. 108(39):16327–16332. doi:10.1073/pnas.1102543108.
Nee S, May RM, Harvey PH. 1994. The reconstructed evolutionary process. Philos Trans R Soc Lond B Biol Sci. 344(1309):305–311. doi:10.1098/rstb.1994.0068.
Neher RA, Russell CA, Shraiman BI. 2014. Predicting evolution from the shape of genealogical trees. Elife. 3:e03568. doi:10.7554/eLife.03568.
Nourmohammad A, Otwinowski J, Łuksza M, Mora T, Walczak AM. 2019. Fierce selection and interference in B-cell repertoire response to chronic HIV-1. Mol Biol Evol. 36(10):2184–2194. doi:10.1093/molbev/msz143.
Pae J, Ersching J, Castro TBR, Schips M, Mesin L, Allon SJ, Ordovas-Montanes J, Mlynarczyk C, Melnick A, Efeyan A, *et al.* 2021. Cyclin D3 drives inertial cell cycling in dark zone germinal center B cells. J Exp Med. 218(4):e20201699. doi:10.1084/jem.20201699.
Pae J, Schwan N, Ottino-Loffler B, DeWitt WS, Garg A, Bortolatto J, Vora AA, Shen JJ, Hobbs A, Castro TBR, *et al.* 2025. Transient silencing of hypermutation preserves B cell affinity during clonal bursting. Nature. 641(8062):486–494. doi:10.1038/s41586-025-08687-8.
Park S, Mali NM, Kim R, Choi JW, Lee J, Lim J, Park JM, Park JW, Kim D, Kim T, *et al.* 2021. Clonal dynamics in early human embryogenesis inferred from somatic mutation. Nature. 597(7876):393397. doi:10.1038/s41586-021-03786-8.
Perié L, Duffy KR, Kok L, de Boer RJ, Schumacher TN. 2015. The branching point in erythro-myeloid differentiation. Cell. 163(7):1655–1662. doi:10.1016/j.cell.2015.11.059.
Prillo S, Ravoor A, Yosef N, Song YS. 2023. ConvexML: scalable and accurate inference of single-cell chronograms from CRISPR/Cas9 lineage tracing data. bioRxiv. 569785. doi:10.1101/2023.12.03.569785. Preprint: not peer-reviewed.
Sagulenko P, Puller V, Neher RA. 2018. TreeTime: maximum-likelihood phylodynamic analysis. Virus Evol. 4(1). doi:10.1093/ve/vex042.
Schwartz R, Schäffer AA. 2017. The evolution of tumour phylogenetics: principles and practice. Nat Rev Genet. 18(4):213–229. doi:10.1038/nrg.2016.170.
Seidel S, Stadler T. 2022. TiDeTree: a Bayesian phylogenetic framework to estimate single-cell trees and population dynamic parameters from genetic lineage tracing data. Proc Biol Sci. 289:20221844. doi:10.1098/rspb.2022.1844.
Simeonov KP, Byrns CN, Clark ML, Norgard RJ, Martin B, Stanger BZ, Shendure J, McKenna A, Lengner CJ. 2021. Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. Cancer Cell. 39(8):1150–1162.e9. doi:10.1016/j.ccell.2021.05.005.
Spencer Chapman M, Ranzoni AM, Myers B, Williams N, Coorens THH, Mitchell E, Butler T, Dawson KJ, Hooks Y, Moore L, *et al.* 2021. Lineage tracing of human development through somatic mutations. Nature. 595(7865):85–90. doi:10.1038/s41586-021-03548-6.
Spisak N, de Manuel M, Milligan W, Sella G, Przeworski M. 2024. The clock-like accumulation of germline and somatic mutations can arise from the interplay of DNA damage and repair. PLoS Biol. 22(6):e3002678. doi:10.1371/journal.pbio.3002678.
Stadler T, Pybus OG, Stumpf MPH. 2021. Phylodynamics for cell biologists. Science. 371(6526):eaah6266. doi:10.1126/science.aah6266.
Stamatakis A. 2014. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30(9):1312–1313. doi:10.1093/bioinformatics/btu033.
Turner JS, O'Halloran JA, Kalaidina E, Kim W, Schmitz AJ, Zhou JQ, Lei T, Thapa M, Chen RE, Case JB, *et al.* 2021. SARS-CoV-2 mRNA vaccines induce persistent human germinal centre responses. Nature. 596(7870):109–113. doi:10.1038/s41586-021-03738-2.
Uduman M, Yaari G, Hershberg U, Stern JA, Shlomchik MJ, Kleinstein SH. 2011. Detecting selection in immunoglobulin sequences. Nucleic Acids Res. 39(suppl):W499–504. doi:10.1093/nar/gkr413.
Van Egeren D, Escabi J, Nguyen M, Liu S, Reilly CR, Patel S, Kamaz B, Kalyva M, DeAngelo DJ, Galinsky I, *et al.* 2021. Reconstructing the lineage histories and differentiation trajectories of individual cancer cells in myeloproliferative neoplasms. Cell Stem Cell. 28(3):514–523.e9. doi:10.1016/j.stem.2021.02.001.
Varilly P, Schifferli M, Yang K, Burcham T, Cronan P, Glennon O, Jacks O, Laning E, Marrs L, Oba K, *et al.* 2025. Delphy: scalable, near-real--time Bayesian phylogenetics for outbreaks. bioRxiv. 645253. doi:10.1101/2025.03.25.645253. Preprint: not peer-reviewed.
Victora GD, Nussenzweig MC. 2022. Germinal centers. Annu Rev Immunol. 40(1):413–442. doi:10.1146/immunol.2022.40.issue-1.
Victora GD, Schwickert TA, Fooksman DR, Kamphorst AO, Meyer-Hermann M, Dustin ML, Nussenzweig MC. 2010. Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. Cell. 143(4):592–605. doi:10.1016/j.cell.2010.10.032.
Williams N, Lee J, Mitchell E, Moore L, Baxter EJ, Hewinson J, Dawson KJ, Menzies A, Godfrey AL, Green AR, *et al.* 2022. Life histories of myeloproliferative neoplasms inferred from phylogenies. Nature. 602(7895):162–168. doi:10.1038/s41586-021-04312-6.
Yaari G, Benichou JIC, Vander Heiden JA, Kleinstein SH, Louzoun Y. 2015. The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. Philos Trans R Soc Lond B Biol Sci. 370(1676):20140242. doi:10.1098/rstb.2014.0242.
Yang D, Jones MG, Naranjo S, Rideout WM, Min KHJ, Ho R, Wu W, Replogle JM, Page JL, Quinn JJ, *et al.* 2022. Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. Cell. 185(11):1905–1923.e25. doi:10.1016/j.cell.2022.04.015.
Zwaans A, Seidel S, Manceau M, Stadler T. 2024. Bayesian phylodynamics of early vertebrate development in BEAST 2. bioRxiv. 601658. doi:10.1101/2024.07.04.601658. Preprint: not reviewed.

*Editor: S. Ott*