Estimating single cell clonal dynamics in human blood using coalescent theory

Brian Johnson [†1], Yubo Shuai[†2], Jason Schweinsberg [*2], and Kit Curtius [*1,3]

[1]Division of Biomedical Informatics, Department of Medicine, University of California San Diego, 9500 Gilman Dr., La Jolla, 92093, CA, USA

[2]Department of Mathematics, University of California San Diego, 9500 Gilman Dr., La Jolla, 92093, CA, USA

[3]Moores Cancer Center, University of California San Diego, 9500 Gilman Dr., La Jolla, 92093, CA, USA

[*]Corresponding Authors: kcurtius@health.ucsd.edu and jschweinsberg@ucsd.edu

[†]These authors contributed equally to this work.

Abstract

While evolutionary approaches to medicine show promise, measuring evolution itself is difficult due to experimental constraints and the dynamic nature of body systems. In cancer evolution, continuous observation of clonal architecture is impossible, and longitudinal samples from multiple timepoints are rare. Increasingly available DNA sequencing datasets at single cell resolution enable the reconstruction of past evolution using mutational history, allowing for a better understanding of dynamics prior to detectable disease. We derive methods based on coalescent theory for estimating the net growth rate of clones from either reconstructed phylogenies or the number of shared mutations. Using single-cell datasets from blood, we apply and validate our analytical methods for estimating the net growth rate of hematopoietic clones, eliminating the need for complex simulations. We show that our estimates may have broad applications to improve mechanistic understanding and prognostic ability. Compared to clones with a single or unknown driver mutation, clones with multiple drivers have significantly increased growth rates (median 0.94 vs. 0.25 per year; $p = 1.6 \times 10^{-6}$). Further, stratifying patients with a myeloproliferative neoplasm (MPN) by the growth rate of their fittest clone shows that higher growth rates are associated with shorter time from clone initiation to MPN diagnosis (median 13.9 vs. 26.4 months; $p = 0.0026$).

Keywords: Coalescent theory, hematopoiesis, somatic evolution, phylodynamics, aging, cancer

Introduction

Clonal expansions of cells that acquire certain mutations post-conception are a direct result of somatic evolution and are prevalent across the human body[1–7]. By estimating the timing of clone initiation and subsequent growth rates of clones, we can characterize evolutionary mechanisms that underlie aging[8] and malignant progression[9–12]. In blood, for example, this evolutionary process is known as *clonal*

hematopoiesis and has been associated with many aging-related disorders, such as anemia[13], impaired immunity[14,15], and cardiovascular disease[16,17], as well as progression to hematopoietic cancers[5,9,18]. Previous analyses found that somatic mutations conferring higher fitness, measured by clonal growth rate, lead to a higher risk of malignant transformation[19–21]. However, validated methods for measuring these important evolutionary parameters, which can vary from patient to patient, remain limited. Fast, accurate estimates of the underlying clonal dynamics using genomic data could serve to improve prognostic ability and ultimately lead to better patient outcomes.

Recent whole genome single-cell sequencing experiments in blood[19,22–24] allow for phylogenetic reconstruction of the ancestral relationships between cells. Information on the growth dynamics of individual clones is contained in the phylogeny of sampled cells from a population. Previous phylodynamics approaches to estimate population size trajectories depend on non-parametric estimation of Kingman's coalescent and its subsequent generalization to variable population size[25–27]. The estimated population size trajectory, with the assumption of an underlying growth model such as Wright-Fisher with selection[23], provides the basis for clone growth rate estimation using the *phylodyn* R package[19,28], as well as a method called Phylofit[22,24]. Due to the lack of an analytical solution, these approaches approximate the population size trajectory with Approximate Bayesian computation (ABC), Markov chain Monte Carlo (MCMC) or Integrated nested Laplace approximations (INLA). We introduce new coalescent methods for estimating the net growth rate of a continuous time, supercritical birth-death process. The birth-death process is consistent with a cellular model of symmetric division (birth) and death or differentiation, and our methods remain valid for other models of clonal expansion that begin with an exponential growth phase (e.g., logistic, Gompertzian, etc.). Importantly, our methods require few assumptions and do not depend on computationally expensive simulation, allowing for near instantaneous estimation of the growth rate and its confidence intervals.

Our methods build on the work of Harris et al.[29] and Lambert[30], who recently discovered a relatively simple way to describe the *exact* genealogy of a sample of size n at time T from a birth-death process. Using Lambert's construction, we derive an approximation to the genealogy when T and n are large and use this approximation to obtain a maximum likelihood estimate of the net growth rate of a clone. We prove a limit theorem which gives the asymptotic distribution for the total lengths of the internal and external branches in the phylogenetic tree. The asymptotic distribution of the total internal branch length leads to a second method for estimating the net growth rate. This also allows us to estimate the net growth rate directly from the number of internal or shared mutations, those which are inherited by more than one of the sampled cells. Additionally, we provide an estimate for clone age which is applicable when the growth rate is known and the mutation rate is unknown.

Recent single-cell sequencing datasets[19,22–24] have generated novel insights about the nature of clones in the blood, identifying high risk mutations and revealing that clonal expansions with known drivers are present decades before symptoms appear[22]. Applying our methods to these datasets generates additional insights on the early growth rates of clones, which appear to be clinically relevant. We validate our estimates with longitudinal data and show that our methods contribute to a better understanding of the overall trajectory of the population size of a clone, refining previous estimates and

further advancing our understanding of hematopoiesis, aging and cancer initiation.

Results

We derived new mathematical estimates for evolutionary parameters (e.g., growth rate of a clone) when analyzing single cell-derived DNA sequencing data from a *sample* of the clone. A sample is a random subset of the total cells in the clone, as is commonly available in a realistic single-cell dataset. In the blood datasets analyzed below, samples of unique clones range from 4-109 cells[19,22–24], while total clone size can hypothetically be as large as the total number of hematopoietic stem cells (HSCs) in the human body (estimates range from 25,000-300,000[20,31,32]). Analysis of coalescence times then requires explicit consideration of the size of the sample and thus new theoretical results were needed. In this section, we provide our estimates for clonal growth parameters under a wide range of applicable modeling assumptions, then apply them to simulated and real data. We also compare our results to those produced using Phylofit, a recent Markov chain Monte Carlo (MCMC)-based approach[22].

Mathematical models for estimating clonal growth

First, we describe the biological rationale for inferring the growth rate from a genealogical tree. All cells sampled from the same clone progeny will have a common ancestor dating back to the clone's origin, i.e., when the first cell acquired the identifying mutation leading to clonal expansion. Any two sampled cells may have a more recent common ancestor, and the most recent time at which the two cells have a common ancestor is called the *coalescence time* for these cells. In a sample of n cells, there will be $n-1$ distinct coalescence times. For larger populations, it is less likely that any two sampled cells will have a recent common ancestor. Therefore, a faster growing (larger) population should have older common ancestors and a slower growing (smaller) population should have more recent common ancestors. Because the probability of a shared ancestor is dependent on the total clone size, the distribution of coalescence times provides information on the clone size trajectory, and here we use it to infer the early growth rate of the clone, r.

To connect growth rates to a genealogical tree, we consider the following birth-death process. We assume each cell divides symmetrically at rate λ and dies or differentiates at rate μ, acquiring mutations through time at rate ν. We wish to estimate $r = \lambda - \mu$, the net growth rate (see Figure 1A). The data consists of a sample of n cells from the clone (of total size = N cells) at a clone age T. We assume that r is positive and constant during the expansion phase of the clone. We also assume the sample size n is much smaller than the total clone size N, which is usually valid for single cell-derived datasets described herein that typically have most coalescence events occurring shortly after clone initiation (i.e., star-shaped genealogies)[19,22–24]. Although our mathematical results are proved for this simple birth-death process, they can be applied to a larger class of models that describe the expansion of clones in blood with an early exponential growth phase (e.g., logistic growth[19,22], purely exponential growth[20], etc.). Our results will be valid when observed coalescence times are not impacted by changing growth rates that may occur after the initial expansion phase (see Figure 1B), and when the

time at sampling, T, does not bias the observed coalescence times.

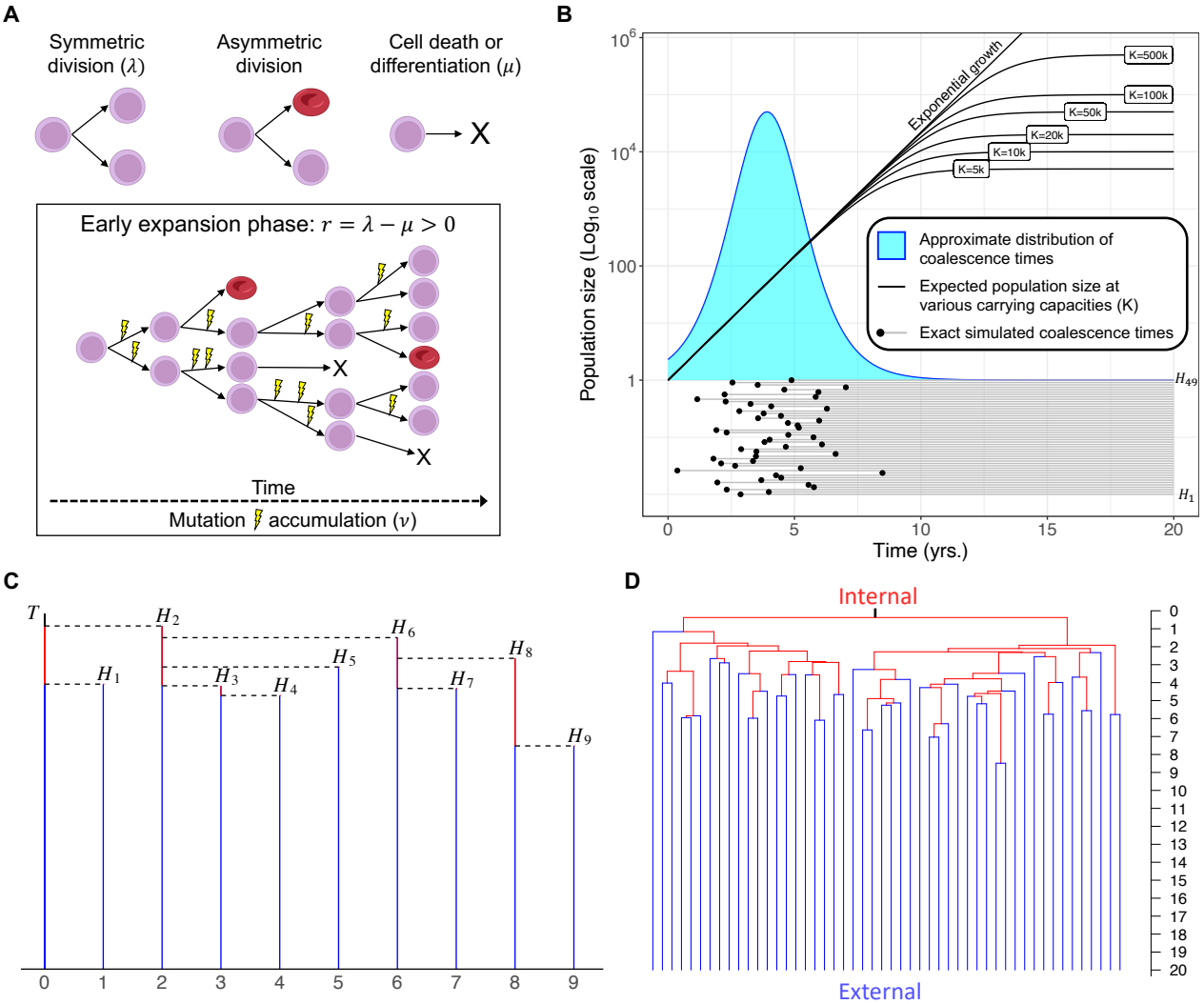


Figure 1: **Model schematic and coalescent results. A:** Stem cells undergo symmetric division at rate λ, increasing the population of stem cells by 1. Asymmetric division does not affect the population size or phylogeny except to introduce mutations. Cell death (or differentiation) occurs at rate μ, which removes the cell's inherited history from the phylogeny and decreases the population size by 1. Our methods seek only to estimate the growth rate during the expansion phase of a clone, when the rate of symmetric division is greater than the rate of cell death ($r = \lambda - \mu > 0$) and both rates are assumed to be constant during this phase. Mutations, which can occur at or between divisions, are assumed to accumulate linearly with time at rate ν. **B:** The approximate distribution of coalescence times for $n = 50$ cells is plotted above one example of $n-1 = 49$ coalescence times drawn from the exact distribution of coalescence times for a birth-death process. The expected population size assuming logistic growth with different carrying capacities shows that most coalescence events occur at smaller population sizes, when the growth trajectory is still approximately exponential. Other parameters: $r = 1$, ($\lambda = 1.5$, $\mu = 0.5$), $T = 20$. Note that a sampling time $T < 10$ years would artificially affect the distribution of coalescence times, introducing bias. **C:** Method overview: Reconstruction of a genealogical tree using the coalescent point process (CPP) can be done by first adding a vertical line of length T, and then adding successive vertical lines representing the coalescence times (H_i). The coalescence times are drawn i.i.d. from the distribution defined in **Simulating the exact genealogy** and are then connected via horizontal lines to form the ultrametric tree. **D:** Tree reconstructed using the coalescent point process with coalescence times from (B). Red edges are internal, representing shared mutational history, and blue edges are external.

Approximating genealogy using a coalescent point process

A recent elegant method by Lambert for computing the exact genealogy of a sample of size n at time T from a birth-death process is described in Simulating the exact genealogy[30]. Here, we derive a useful approximation to the exact genealogy. We model the coalescence times as independent and identically distributed (i.i.d.) random variables having a logistic distribution, plus a random shift which accounts for the randomness in the initial growth of the branching process. The following approximate distribution of coalescence times provides the foundation for our estimates of net growth rate. We note that Ignatieva, Hein, and Jenkins[33] showed using a different method involving a random time change that the coalescence times can be approximated well by i.i.d. logistic random variables. The logistic distribution also appeared in a similar context in work of Delaporte, Achaz, and Lambert[34].

We are mostly interested in the case when the clone age at sampling T and the sample size n are large. We can therefore obtain a useful approximation by letting $T \to \infty$ and then $n \to \infty$ in the construction from section Simulating the exact genealogy in Methods. This leads to the following simpler method for approximating the coalescence times $H_1, \ldots, H_{n-1}$:

1. Let W have an exponential distribution with mean 1.
2. Let $U_1, U_2, \ldots, U_{n-1}$ be i.i.d. random variables having the logistic distribution, which is a symmetric distribution on the real line with density given by
$$f_{U_i}(u) = \frac{e^u}{(1+e^u)^2}.$$
3. Let
$$H_i = T - \frac{1}{r}\Big(\log(1/W) + \log n + U_i\Big).$$

For mathematical details on derivations, see Supplementary section 1.1. Once the H_i have been determined, we can construct the genealogical tree by randomly merging two lineages at each coalescence time, or by using the coalescent point process as shown in Figure 1C. To understand the formula for H_i, note that in a supercritical branching process ($r > 0$), the individuals sampled at time T are likely to have been descended from different ancestors fairly close to time zero, when the clone began expanding. That is, the genealogical tree will be nearly star-shaped, with most coalescence occurring near time 0. Note that the expected population size of the clone at time t is e^{rt}. Considering the case when $\lambda = r$ and $\mu = 0$, the size of the population after a large time t can therefore be approximated by We^{rt}, where W has an exponential distribution with mean 1. This expression equals n when $t = \frac{1}{r}(\log(1/W) + \log n)$. We expect most lineages to coalesce when the size of the population is comparable to n, which is why the coalescence times H_i are close to $T - \frac{1}{r}(\log(1/W) + \log n)$.

Estimating growth rate of a clone

If we can reconstruct the full genealogical tree from data, then we have estimates for the $n-1$ coalescence times $H_1, \ldots, H_{n-1}$. From the discussion above, we can write

$$H_i = a + bU_i, \qquad a = T - \frac{1}{r}\Big(\log(1/W) + \log n\Big), \qquad b = \frac{1}{r}. \tag{1}$$

where the random variables $U_1, \ldots, U_n$ are i.i.d. and have a logistic distribution. Note that we can write $b = 1/r$ instead of $b = -1/r$ because the logistic distribution is symmetric. We can therefore estimate the growth rate r by estimating the parameter b. We introduce here three methods. Alongside the results below, we also created an R package *cloneRate* for implementing growth rate estimation on novel user input data.

Growth rate estimation using maximum likelihood

Maximum likelihood can be used to estimate b from $H_1, \ldots, H_{n-1}$. Because the maximum likelihood estimate does not have a closed form expression, it must be found using numerical methods. We have used the *maxLik* package in R to compute the maximum likelihood estimate $\hat{b}$[35]. From Equation 1, we can estimate r by

$$\hat{r} = 1/\hat{b}. \tag{2}$$

Let $0 < \alpha < 1$. A $100(1-\alpha)\%$ confidence interval (CI) for r is

$$\left[\hat{r}\left(1 - \frac{cz_{\alpha/2}}{\sqrt{n}}\right), \hat{r}\left(1 + \frac{cz_{\alpha/2}}{\sqrt{n}}\right)\right], \qquad c = \frac{3}{\sqrt{3+\pi^2}}, \tag{3}$$

where $z_{\alpha/2}$ is the number such that if Z has a standard normal distribution, then $P(Z > z_{\alpha/2}) = \alpha/2$. Note that $c \approx .836$, which we use to compare to the confidence intervals of the following estimate based on internal branch lengths. See Supplementary section 1.2 for CI derivations.

Growth rate estimation using internal branch lengths

If we are able to reconstruct the full tree, then we know the internal branch lengths L_n^{in} (e.g., sum of the lengths of red branches in Figure 1D). By Theorem 1 in Internal and external branch lengths in Methods, the distribution of L_n^{in} is approximately normal with mean n/r and variance n/r^2. Therefore, we can estimate the growth rate by

$$\hat{r} = \frac{n}{L_n^{in}}, \tag{4}$$

and we obtain an asymptotically valid $100(1-\alpha)\%$ confidence interval for r by

$$\left[\hat{r}\left(1 - \frac{z_{\alpha/2}}{\sqrt{n}}\right), \hat{r}\left(1 + \frac{z_{\alpha/2}}{\sqrt{n}}\right)\right]. \tag{5}$$

This estimate based on the internal branch lengths can be compared directly to the maximum likelihood estimate, as both methods take a time-based ultrametric tree as input. If the coalescence times

are accurate, then considering only the internal branch lengths discards relevant information and one would expect the maximum likelihood estimate to perform better. The confidence bounds of the internal lengths estimate reflect this, as the confidence bounds of the internal lengths method in Equation 5 are identical in form to the confidence bounds for maximum likelihood in Equation 3, except that the internal lengths effectively has $c = 1$. Because $c = 1 > 0.836$, the internal lengths method has wider confidence intervals.

When reconstructing the tree from mutations, there will be some randomness inherent in the number of observed mutations. Because the previous methods use the time-based tree as input, neither accounts for this uncertainty. The following section uses the ideas presented here to estimate the net growth rate directly from the observed mutations, providing confidence bounds which account for the randomness of mutation accumulation.

Growth rate estimation using shared mutations rather than full tree

If we can estimate the mutation rate ν during the expansion phase of the clone, then we can also estimate the growth rate directly from the number of shared mutations, defined as those mutations present in more than 1 but not all of the n sampled cells. The key idea is that there will be more shared mutations when the growth rate is smaller and fewer when the growth rate is larger. As shown in Shared and private mutations in Methods, the distribution of the number of shared mutations M^{in} is approximately normal with mean $n\nu/r$ and variance $\sigma^2 = n(\nu/r + \nu^2/r^2)$. Therefore, if the mutation rate ν is known, we can estimate the growth rate by

$$\hat{r} = \frac{n\nu}{M^{in}}. \tag{6}$$

An asymptotically valid $100(1-\alpha)\%$ confidence interval for r is given by

$$\left[\hat{r}\left(1 - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{1 + \frac{n}{M_n^{in}}}\right),\ \hat{r}\left(1 + \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{1 + \frac{n}{M_n^{in}}}\right)\right]. \tag{7}$$

Accounting for Poissonian fluctuations in the observed number of shared mutations leads to confidence bounds slightly wider than those from the internal lengths method (Equation 5).

Estimation performance on simulated data

To verify the performance of our methods, we generated trees using the exact genealogy reconstruction discussed in full detail in Simulating the exact genealogy in Methods. Recent work by Lambert[30] allows for instantaneous generation of the exact genealogy of a sample from a supercritical birth-death process, removing the need for expensive simulation for a wide range of population genetics and coalescent work. We briefly describe this process here. For a sample of size n at time T from a clone expanding with birth rate λ and death rate μ, $n-1$ exact coalescence times are drawn using the process described in Simulating the exact genealogy. An example of a set of 49 coalescence times for a sample of 50 cells is shown in Figure 1B. Given the coalescence times $H_i, \ldots, H_{n-1}$, the

coalescent point process can be used to quickly reconstruct the genealogy. To reconstruct the tree from the coalescence times, we begin by drawing a vertical line of height T. We then draw vertical lines of heights $H_1, \ldots, H_{n-1}$ and, at the top of each vertical line, draw a horizontal line to the left, stopping when it hits a vertical branch. The resulting tree is ultrametric, meaning that the root to tip distance is the same for all tips. Figure 1C shows a schematic example of generated a tree with 10 tips from $n - 1 = 9$ coalescence times, and Figure 1D shows the tree constructed from the coalescence times in Figure 1B using the coalescent point process.

Then, applying our methods to these reconstructed trees gives a distribution of estimates which allows for benchmarking since we know the 'true' growth rate in the simulated data. We compared the performance of our methods to the Markov chain Monte Carlo (MCMC)-based approach, Phylofit, introduced by Williams et al.[22]. We did not compare to the performance of their Approximate Bayesian computation (ABC)-based estimates, but note that the authors show a strong correlation between estimates from Phylofit and the ABC-based method (correlation coefficient r = 0.96)[22].

Performance across varying sample size n and growth rate r

We found similar performance using our methods compared to using Phylofit. As shown in Figure 2A-B, Phylofit appears to perform slightly better for small sample size n, while both of our methods outperform Phylofit for $n \geq 100$. Of our two methods, maximum likelihood has the lowest root-mean-square error for larger values of n (see Figure 2B). When the sample size n is too low, our approximation of the distribution of coalescence times, which is valid as $n \to \infty$, no longer accurately describes the population. Intuitively, a smaller sample provides less information available to make an accurate estimate of the growth rate. As such, performance deteriorates and the confidence intervals of our estimates expand with decreasing n (see Equations 3, 5, and 7). We use a cutoff sample size of $n = 10$ in a clone when applying to real data below, but note that this cutoff depends on the desired accuracy of the estimate. The confidence intervals for our methods are approximately accurate, as shown in Figure 2C, though we note that the maximum likelihood confidence intervals may be slightly too narrow for small n because the variance estimate is based on the asymptotic Cramer-Rao bound[36]. We also show that Phylofit runtime scales with the number of samples, while our methods are essentially instantaneous regardless of the size of the tree (Figure 2D).

Similarly, we show the performance of our methods across r values. As shown in Figure 3A-B, our maximum likelihood method performs best for most growth rates. Again, Figure 3C shows that the confidence intervals are approximately accurate. In Figure 3, the smallest growth rate $r = 0.25$ shows concerning performance, which motivated further investigation into the impact of small growth rates.

Small growth rate diagnostic for method utilization

We investigated the performance failure at small growth rates and determined why this happens. We then derived a diagnostic to determine when the growth rate is large enough for our methods to be applicable. As shown in Figure 4A, all three methods perform poorly at small growth rates; our

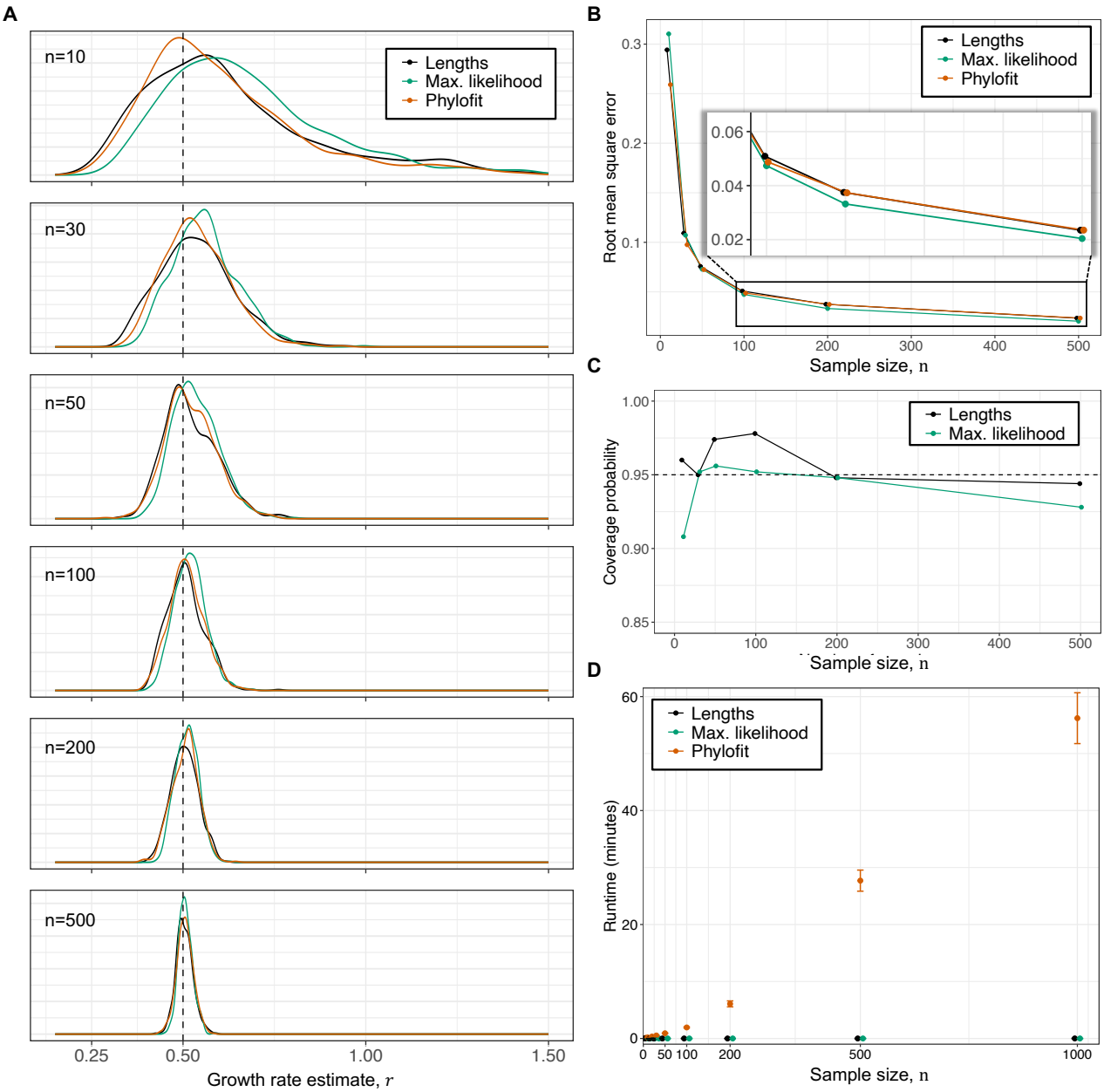


Figure 2: **Performance and number of samples, n. A:** Distribution of estimates from our methods using Equation 2 (green) and Equation 4 (black), and Phylofit (orange) on 500 simulated ultrametric trees from each n value, where n is the number of sampled cells. Simulated trees were generated assuming a continuous time birth-death branching process with $r = 0.5$ and $T = 40$. Birth rate λ is sampled from a uniform distribution on $[0.5, 1.5]$ and death rate $\mu = \lambda - r$. **B:** Root mean square error for each method from the simulations shown in (A) demonstrates improved performance with number of samples. Phylofit is most accurate for small n while Maximum likelihood performs best for large n. **C:** Accuracy of 95% confidence intervals developed for our Internal lengths and Maximum likelihood methods based on simulations in (A). **D:** Runtime (mean +/- st. dev.) of various methods of estimating net growth rate shows that while the MCMC-based Phylofit scales with the number of samples, n, our methods run effectively instantaneously for any tree size.

methods perform worse than Phylofit in this small r regime. To understand why our methods can fail for small r, we note that when $r > 0$, so that the branching process is supercritical, the n sampled cells should all have distinct ancestors that were alive a short time after the initiation of the clone. Consequently, the genealogical tree will be nearly star-shaped, with long external branches and short internal branches near the root (see example in Figure 1D). On the other hand, when $r = 0$, so that the branching process is critical, the population size is nearly stable over time. Then the genealogy of the

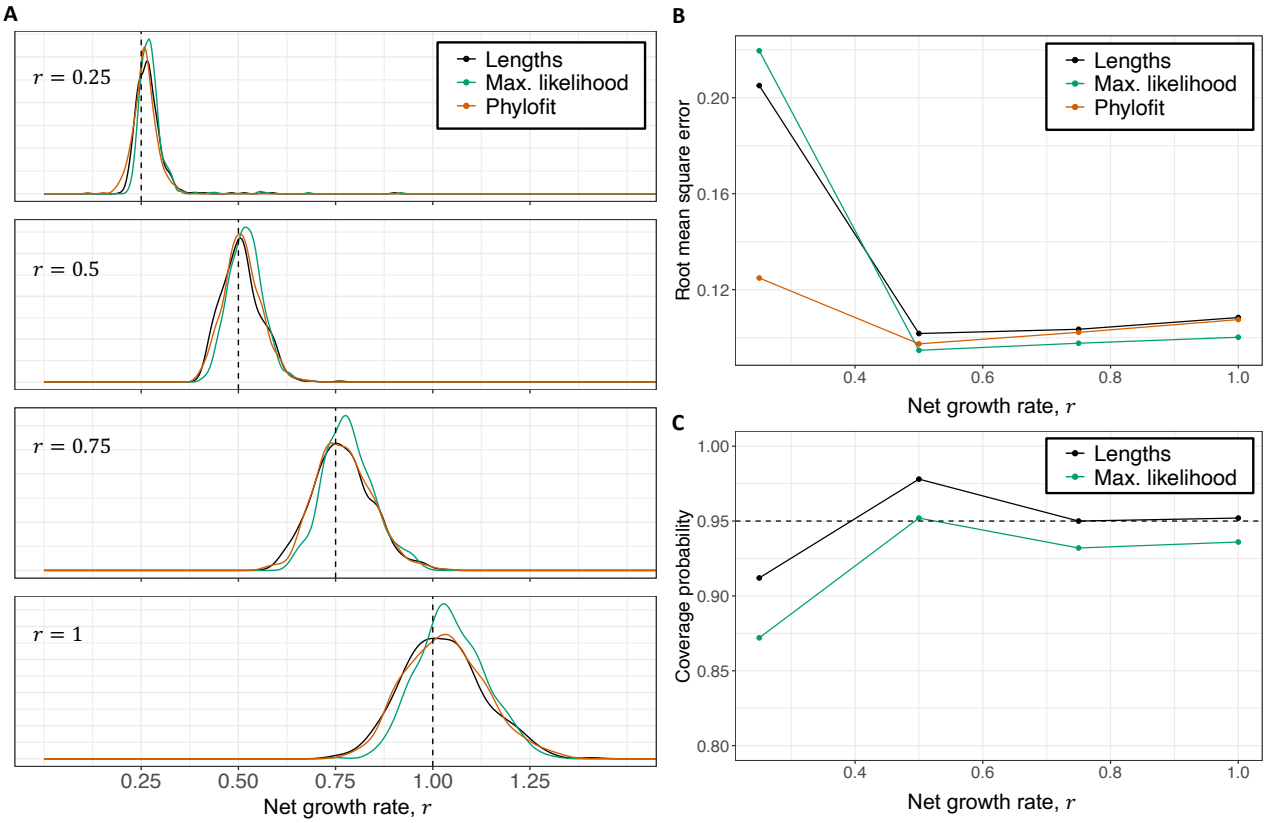
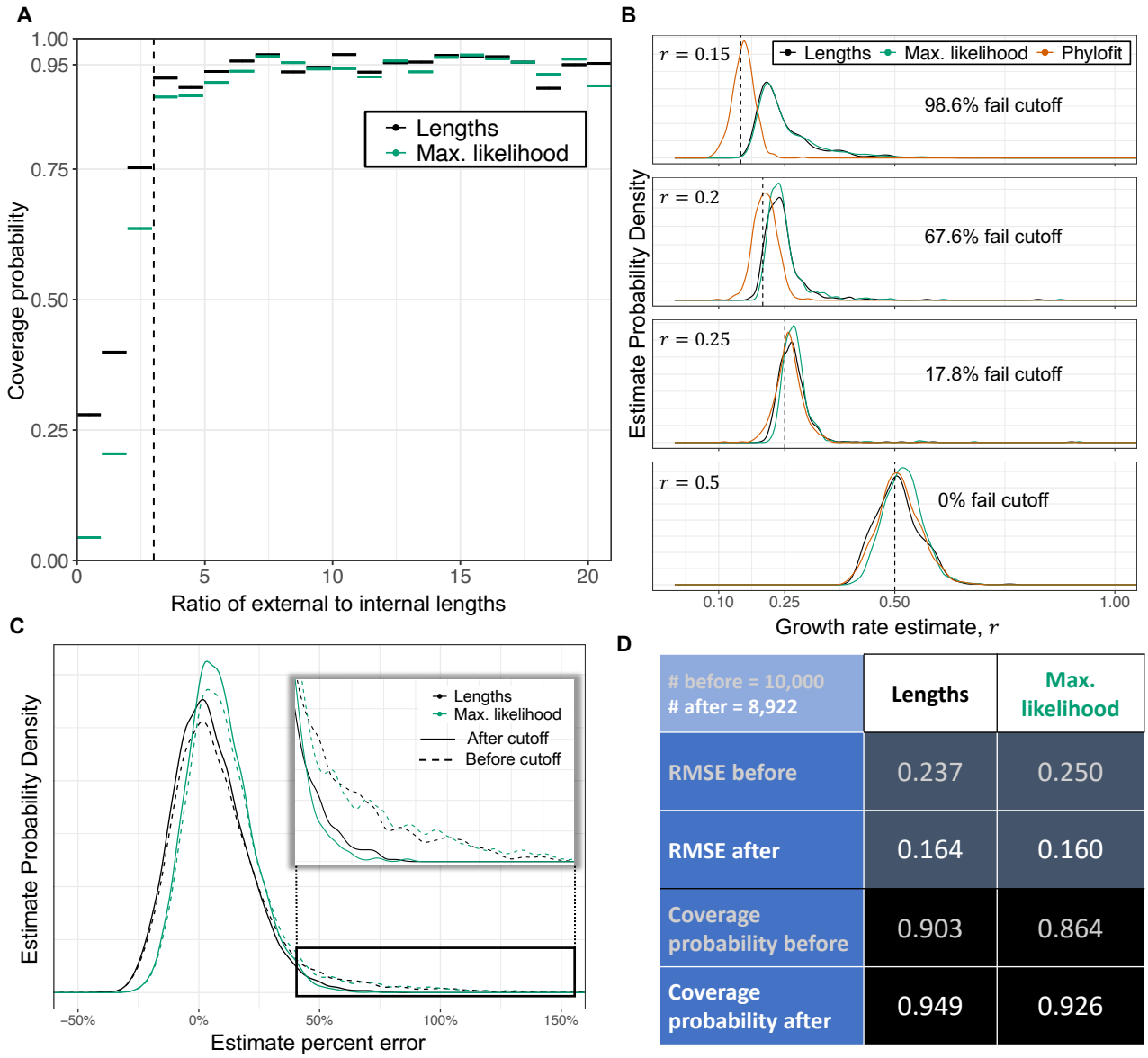


Figure 3: **Performance across different growth rates, r. A:** Distribution of estimates from our methods and Phylofit on 500 simulated ultrametric trees from each r value, where r is the growth rate. Simulated trees generated assuming a continuous time birth-death branching process with $n = 100$ and $T = 40$. **B:** Root mean square error, normalized to account for the different true growth rates, for each method from the simulations shown in (A) demonstrates a decrease in accuracy for small growth rates. Phylofit is most accurate for small r while Maximum likelihood performs best for large r. **C:** Accuracy of 95% confidence intervals developed for our Internal lengths and Maximum likelihood methods based on simulations in (A).

n sampled cells would more closely resemble the classical Kingman's coalescent, in which lineages merge at a constant rate and most coalescence events occur near the time of sampling, leading to long internal branches and short external branches. When r is small but positive, the genealogical tree will be star-shaped if the sampling time T is sufficiently large, but the internal branch lengths will still be longer than when r is larger. Consequently, if T is not sufficiently large, then the constraint that the coalescence times must be less than T will affect the distribution of the internal lengths, and the approximation that we derived will not be accurate.

We found that as long as the total length of the external branches is much larger than the total length of the internal branches, this problem does not arise, and our methods give accurate results. This leads naturally to a diagnostic which determines when our methods are applicable. As shown in Figure 4A, when the ratio of external to internal lengths is greater than or equal to 3, our methods and confidence intervals are accurate. Figure 4B shows that most of the simulated trees with problematic small growth rates fail this diagnostic cutoff. Applying this cutoff to a simulated dataset with growth rates between 0.1 and 1 reduces overestimates and greatly improves performance (Figure 4C-D). Notably, small growth rate clones are unlikely to be observed in enough sampled cells in real data to make an accurate estimate. In fact, none of the 42 clones which we analyze from the blood datasets



# before = 10,000 # after = 8,922	Lengths	Max. likelihood
RMSE before	0.237	0.250
RMSE after	0.164	0.160
Coverage probability before	0.903	0.864
Coverage probability after	0.949	0.926

Figure 4: **Small growth rate r diagnostic. A:** Trees with $n = 50$ tips and $T = 40$ were created from 10,000 randomly sampled r values from a uniform distribution on (0.1, 1) and then binned by the ratio of external to internal lengths in increments of one. For corresponding estimates of r, accuracy of the 95% confidence intervals (CI) shows that the ratio of external to internal lengths can be used as a diagnostic to determine whether our method can accurately estimate growth rate. We use a ratio of 3 as a cutoff value as simulations with a ratio greater than 3 are captured by the 95% confidence intervals approximately 95% of the time. Minimum number of simulations captured in a single bin is 61. **B:** Distribution of estimates from our methods and Phylofit on 500 simulated ultrametric trees from each r value, where r is the growth rate. Simulated trees generated assuming a continuous time birth-death branching process with $n = 100$ and $T = 40$. For each tree, we calculate the ratio of external lengths to internal lengths. For each growth rate, we show the percentage of trees which have a ratio of external to internal lengths below the diagnostic value of 3. **C:** Relative fractional error distribution for four methods from the same simulations shown in (A) before (dashed lines, iterations = 10,000) and after (solid lines, iterations = 8,922) the cutoff of 3 was applied. Inset shows significant reduction in overestimates due to the diagnostic cutoff. **D:** Normalized root mean square error (RMSE) and fraction of samples outside of 95% confidence intervals (coverage probability) for four methods using same simulations shown in (A) and (C) before and after the diagnostic cutoff was applied. The diagnostic provides a significant reduction in error and improvement in accuracy of confidence intervals.

below has an external to internal length ratio less than 4, and only two have a ratio less than 5.

Application to human blood datasets

Number of Individuals	Number of Clones	Data source	Diagnosis	Ref.
11	18 (15 unique)	Adult peripheral blood (PB) and/or bone marrow (BM)	Myeloproliferative Neoplasm (MPN)	22
2	2	Adult BM	MPN	23
3	15	Adult PB and/or BM	Normal	24
3	7	Adult PB	Normal	19
19	**42 (39 unique)**	**Total**		

Table 1: Whole genome sequencing datasets of single-cell derived colonies. Number of clones indicates the number of clonal expansions with $n \geq 10$ cells sampled. As some clones profiled by Williams et al.[22] had $n \geq 10$ cells sampled at multiple timepoints from the same clone, we also specify the number of unique clones. See Supplementary section 3.2 for details.

We applied our methods to single-cell derived sequencing data from human blood (Table 1). The methods for generating the data are fairly similar across the studies: single hematopoietic progenitor cells were clonally expanded and each single-cell derived colony was sequenced to a mean depth of roughly 15x, with slight differences depending on the study[19,22–24]. Time-based ultrametric trees generated in these studies are used as input for our methods and Phylofit. Manual annotation is required to identify clonal expansions, associate clones with specific drivers, and to remove nested subclonal expansions from the clone of interest. We generally designated the clones as annotated in the studies which produced the data[19,22–24] and provide details in Supplementary section 3.2.

First, we check our assumption of neutrality within expanding clones (i.e., all cells within the clone grow at approximately the same rate). Previous authors have studied the expected site frequency spectrum for a sample from a birth-death process. Letting M_n^k denote the number of mutations inherited by k of the n sampled individuals, Durrett[37] showed that as $T \to \infty$, for $k \geq 2$ we have

$$\mathbb{E}[M_n^k] \sim \frac{n\nu}{r} \cdot \frac{1}{k(k-1)}. \tag{8}$$

Gunnarsson, Leder, and Foo[38] calculated the exact expectation in the case when the entire clone is sampled (see also[39,40] for similar calculations). Therefore, we expect the site frequency spectrum to follow the curve $1/k(k-1)$, where k equals the number of cells. In Figure 5A, we show the averaged site frequency spectrum across all clones, with any nested subclones removed, along with the 95% confidence interval of the mean. The agreement between the observed mean and the expectation indicates neutrality within clones, consistent with previous conclusions in blood[19,22–24]. For more detailed data on the site frequency spectrum for each clone, see Supp. Table 5.

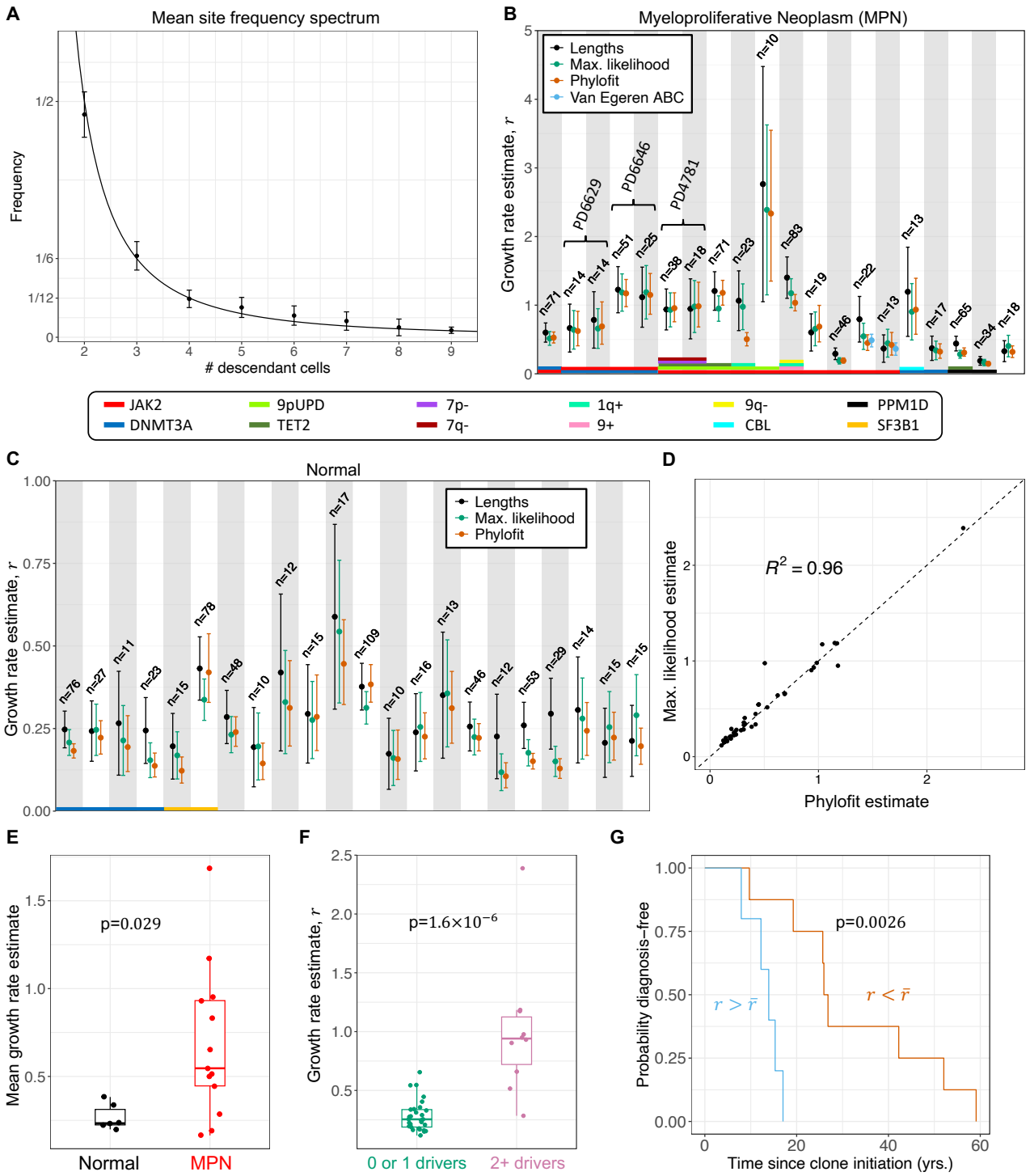


Figure 5: **Applying estimates to blood data. A:** Averaged site frequency spectrum across 42 clones shows agreement with the $\frac{1}{k(k-1)}$ neutral expectation (solid line). Error bars show 95% confidence interval of the mean. **B-C:** Our estimates and Phylofit for clones with $n \geq 10$ tips from individuals with (B) and without (C) Myeloproliferative neoplasms (MPN) shows good agreement across methods. Brackets in (B) group estimates from the same clones in the same patient estimated from two distinct samples taken years apart, showing consistency of estimates. Note that we also include estimates from Van Egeren et al.[23] in light blue for the two clones from their dataset. **D:** Correlation between our maximum likelihood estimate and estimates from Phylofit for all clones from (B) and (C). **E:** Mean max. likelihood net growth rate estimate for clones from patients with (red) and without (black) a diagnosis of MPN shows that more aggressive expansions are associated with MPN. **F:** Max. likelihood net growth rate estimate for clones with single or unknown drivers (green) and multiple drivers (magenta) show that fitness predicted by our methods is consistent with effects of known drivers. Non-parametric Mann-Whitney test used for p-value calculation in (E, F). **G:** In the single most aggressive clone from each patient diagnosed with MPN, stratification by mean net growth rate $\bar{r}$ shows significant differences in Kaplan-Meier survival curves from clone initiation to MPN diagnosis (log-rank test p=0.0026) though sample set was small (13 patients). At time of sampling, mean age of high growth rate group was 60.3 years, median was 50.4 years. Mean age of low growth rate group was 60.9 years, median was 63 years.

In applying our methods to real data, we found agreement across our two methods and agreement with the estimates from Phylofit (see Figure 5B-D). As discussed in Comparing analytical estimates to those using Phylofit in Methods, we only include the estimates from Phylofit without including the sampled clonal fraction as a target because clones have been shown to behave unpredictably at high clonal fractions, decelerating more than would be expected by a logistic growth trajectory[19,24]. Also, sampled clonal fraction and/or Variant Allele Frequency (VAF) may be a poor estimate of mutant allele burden in progenitors and HSCs (see Supp. Figure 3), possibly due to lineage bias in mutated cells, such as the erythroid lineage bias observed in *JAK2* mutants[23].

The most fit clones (those with fastest growth rates) were observed in patients with myeloproliferative neoplasms (MPN). As shown in Figure 5E, we found significantly increased estimates of mean detected clone fitness in individuals diagnosed with MPN as opposed to healthy adults (p=0.029). Additionally, Figure 5F shows that multiple-driver clones have significantly increased rates of expansion as compared to clones with just one or zero known driver mutations (p=1.6×10^{-6}). This suggests increasing fitness effects from the accumulation of additional mutations. Higher growth rates may also be associated with shorter time from clone initiation to cancer diagnosis (log-rank p=0.0026), as shown in Kaplan-Meier curves in Figure 5G. Here the clone initiation time is estimated to occur $1/r$ years before the first coalescence (i.e., first surviving symmetric division). Together, these findings indicate that mechanistic rates for clonal dynamics such as the early growth rate may provide clinically important information in our understanding of hematopoietic cell evolution and transformation to malignancy.

Longitudinal validation of clone growth estimates

We leveraged available longitudinal data to validate our growth rate estimates. Again, for single cell-derived colonies, we used a lower bound of $n = 10$ cells per clone to include in our analysis. For longitudinal bulk data, we restrict to a minimum of 4 timepoint samples of the same bulk cell type from expanding clones. Further, because coalescent estimates are relevant for the early growth rate, we require that longitudinal data have at least two samples with a variant allele frequency in between 0 and 0.25. The longitudinal data consists of peripheral (whole) blood samples[19] and peripheral blood granulocyte samples[22]. It has been suggested that clonal fraction may differ across different blood cell types (granulocyte vs. whole vs. mononuclear)[23]. Data from Williams et al.[22] is consistent with this finding, as there is significantly different clonal fraction across sampled cell type in 3 out of 4 patients where multiple cell types were sampled within a month of each other (see Supp. Figure 3). Therefore, we require that a consistent type be used within each longitudinal growth rate estimate. For more details on the criteria for inclusion of longitudinal data, see Methods section Longitudinal clone inclusion criteria.

We analyzed 4 clones from Williams et al.[22] and 56 clones from Fabre et al.[19] that had appropriate longitudinal data. Of these 60 clones, 3 have sufficient matched data from single cell-derived WGS samples (1 clone from Williams et al.[22] and 2 clones from Fabre et al.[19]) to allow for orthogonal estimates from the same clone. Results for these 3 clones are shown in Figure 6A-C, along with a

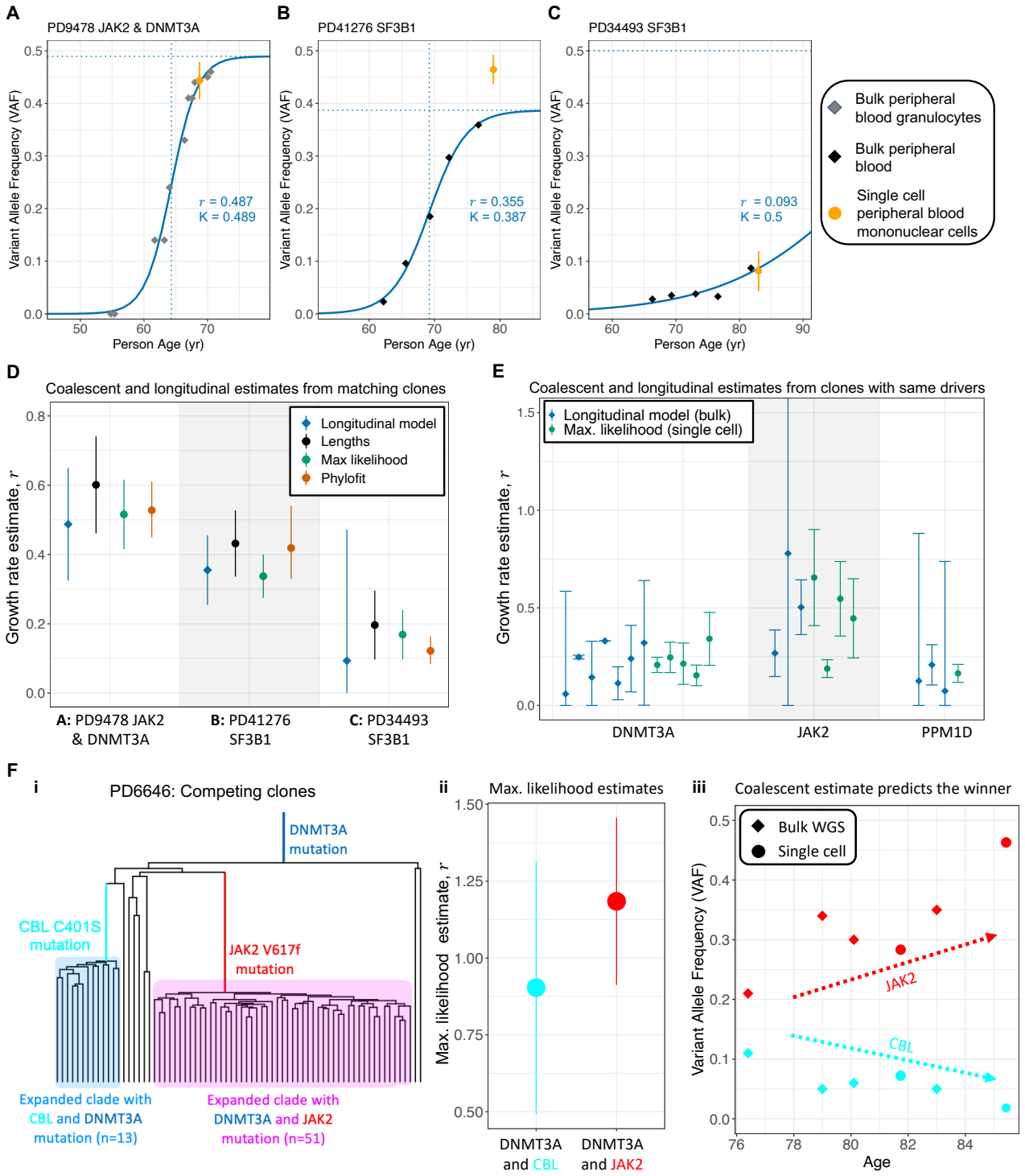


Figure 6: **Longitudinal validation. A-C:** Logistic fit to longitudinal data for three clones which have both single cell and longitudinal data. Only longitudinal bulk WGS data was used for fitting. Single cell colony clonal fraction (divided by 2) and 95% confidence intervals are shown in orange. Source for (A) is Williams et al.[22], and source for (B) and (C) is Fabre et al.[19]. **D:** Longitudinal and Coalescent estimates for each of the clones in (A-C) show agreement across data types. **E:** Longitudinal and coalescent estimates for different clones sharing the same driver. **F:** Clonal competition between a *DNMT3A+CBL* clone and a *DNMT3A+JAK2* clone shown in the reconstructed phylogeny (i). Maximum likelihood coalescent estimate from each clone (ii) shows that the *DNMT3A + JAK2* clone likely has higher fitness. Longitudinal data shows that the *DNMT3A+JAK2* clone increase in VAF over time while the *DNMT3A+CBL* clone decreases, confirming that the *DNMT3A+JAK2* clone has higher fitness, as predicted by our phylogenetic estimate. All error bars represent 95% confidence intervals.

logistic growth model fit. While we show the corresponding single cell colony clonal fraction (orange, divided by 2 to scale to the VAF of a diploid mutant), we do not use this data point in the fitting as the different cell type may affect the clonal fraction, as noted above. The logistic growth model is used primarily to identify the growth rate, r, and is chosen because it has been shown to face fewer parameter identifiability issues than other sigmoid growth models, such as Gompertz or Richards', when applied to similar data[41]. For details on the longitudinal modeling, see Methods section Logistic modeling of longitudinal data.

In comparing the growth rates from the longitudinal fits to our methods and Phylofit (Figure 6D), we found general agreement in the estimates, though we note the wide confidence intervals especially from the logistic fit. Additional longitudinal data from Fabre et al.[19], although not from clones with matched single cell data, was used to compare to our coalescent estimates. First, we identify longitudinal clones with drivers also present in the single cell data. Then, we fit the logistic growth model to these longitudinal clones. After filtering (see Longitudinal clone inclusion criteria) and excluding the 3 clones shown in Figure 6A-C, there were 13 clones with longitudinal data and a driver gene matching the clones profiled by single cell sampling. This data comes from clones with a mutation in one of the following genes: *DNMT3A*, *JAK2*, or *PPM1D*. The estimated growth rates are shown in Figure 6E. Similar growth rates in the same driver genes shows general consistency across all methods, though the small amount of data and wide confidence intervals limit the conclusions that can be drawn.

Finally, we considered competing clones within the same patient. If our estimates are relevant for clonal fitness, we would expect that clones with higher estimated growth rates should out-compete clones with lower estimated growth rates. The only example of competing clones with sufficient single cell data comes from patient PD6646 from Williams et al.[22] (Figure 6F). A *CBL* and a *JAK2* mutation give rise to two independent clones, both of which already carry a *DNMT3A* mutation (Figure 6F i). By our maximum likelihood estimate, the *DNMT3A* + *JAK2* clone is slightly more fit than the *DNMT3A* + *CBL* clone (Figure 6F ii). Both Phylofit and our Internal lengths method also estimate a higher growth rate for the *JAK2* clone. While this patient was undergoing treatment in this time period and the trajectory does not appear logistic, the *JAK2* clone increases in variant allele frequency while the *CBL* clone decreases, consistent with our estimate suggesting that the *JAK2* clone is more fit. There is an important caveat in this example because the specific interactions between clone/mutation and treatment may be responsible for the increasing/decreasing VAF, which is not accounted for by our coalescent estimate of fitness that characterizes early growth before treatment would have begun.

Discussion

We developed new methods using coalescent theory to estimate rates of clonal expansion (and clone age) at greatly reduced computational expense. Leveraging previous work[30], we validated our methods using efficient computational realizations of phylogenies resulting from birth-death branching

processes. We then applied our methods to single cell resolution data from blood, showing that our growth rate estimates are both meaningful and consistent in biological and clinical contexts. We found general consistency of estimates with a previously published MCMC-based approach, Phylofit[22] ($R^2 = 0.94$-0.96). Where possible, we validated our estimates using single cell data from multiple timepoints, and also show that our estimates are consistent with and generally more precise than orthogonal estimates of net growth rate derived from longitudinal bulk data. Because they are based on analytical results, our methods for estimating growth rates from phylogenetic reconstruction are simple and run quickly without sacrificing accuracy. For future datasets with a higher number of sampled cells n and larger numbers of patients and clones, near instantaneous runtime at any tree size may be a critical feature separating our methods from MCMC or ABC-based alternatives. We provide a simple and easy to use R package, *cloneRate*, which will allow other researchers to estimate growth rates with their own input data.

For testing model performance on simulated data, we use results of Harris, Johnston, and Roberts[29] and Lambert[30] to reconstruct the exact genealogy of a sample of size n from a birth-death process at time T, conditional on the population size being at least n at time T. This method avoids the need to simulate the entire large clonal population starting from a single cell as is commonly performed in other methods. From a mathematical point of view, the idea of using the coalescent point process to obtain results about statistics such as the site frequency spectrum and the allele frequency spectrum goes back to Lambert[42] and was later developed further[34,43,44], and then was recently applied to cancer modeling by Dinh et. al.[45]. Here we combine these ideas with the results from Lambert[30] to obtain asymptotic results for quantities that can be derived from a large sample from a birth-death process. By taking advantage of the independence that is inherent in the coalescent point process, we are able to apply the m-dependent Central Limit Theorem to show that the total internal branch length, which can be used to estimate the growth rate of the process, has an asymptotic normal distribution. This observation allows us to obtain an asymptotically valid confidence interval for the growth rate. Finally, this is a unifying method for growth rate estimation that is applicable to many biologically relevant models assumed in previous works for clonal dynamics in blood[19,20,22–24].

Acknowledging that this is both a limitation and a strength, our methods estimate only the growth rate in the early expansion phase, when growth is approximately exponential. Growth rates following the initial expansion phase may change over time in unpredictable ways and this does not affect our results. In focusing only on the early growth rate, our methods do not rely on assumptions of the overall growth trajectory. Additionally, we have shown that early growth rates are relevant to the greater context of clonal and malignant hematopoiesis. For example, we found that higher growth rates are associated with shorter time from clone initiation to MPN diagnosis. The association between MPN diagnosis and growth rate suggests a possible avenue for early detection by predicting which patients are more likely to remain asymptomatic and which are more likely to undergo malignant transformation. Understanding the role of evolutionary dynamics to predict risk of progression in clonal hematopoiesis and provide prognostic information in hematological malignancies has been noted as a top clinical priority[46,47].

Further, multi-driver clones show significantly increased rates of expansion, suggesting possible cumulative and/or synergistic effects of driver mutations. We found wide heterogeneity of fitness effects for *JAK2* clones, consistent with previous findings[19], and relatively low fitness effects with smaller variation for *DNMT3A* clones. In the context of clonal hematopoiesis, single hit drivers with lower growth rates may increase risk for MPN by increasing the reservoir of cells at risk of additional stochastic mutations, thus initiating multi-hit driver clones with potentially additive fitness effects. There are other possible benefits to knowing the early rate of expansion. For example, early expansion rates are affected by fewer outside pressures such as treatment[48] and may be more consistent across patients.

Our findings also provide guidance to experimental researchers designing single-cell DNA sequencing experiments that aim to determine clone fitness. The number of sampled cells required for reliable estimates of growth rates is roughly 20, depending on desired accuracy (see Figure 2). Bulk whole genome sequencing performed prior to single-cell experiments could provide variant allele frequency information that can be used to estimate the cell fraction of clones of interest. Then, the total number of cells sequenced can be decided in a way that ensures enough sampled cells from clones of interest are included, while reducing overall costs.

One limitation is that current methods rely on the manual annotation of clones from a phylogenetic tree. While this is currently a fairly easy task given the relatively small size of single cell DNA sequencing datasets, it may become more challenging for expected increases in throughput[49]. An automated way to detect clonal expansions and distinguish normal cell turnover from expansions may be required to effectively scale the application of our methods. Such an automated algorithm would likely have to leverage not just the distribution of coalescence times, but also measures of tree balance.

With our methods, phylogenetic reconstruction can become an even more powerful tool to infer the past evolutionary dynamics of a population of cells. It has been hypothesized that individuals at high risk of developing myeloid malignancies can be identified before presenting with any symptoms[21]. Knowing which drivers are associated with more aggressive expansions will provide clinicians with better tools to direct treatment and/or prevention strategies. Additionally, clonal expansions without known drivers can provide mechanistic and biological insight. While blood is currently the most convenient medium for creation of single cell-derived DNA sequencing data and validation of these methods, age-related clonal expansions are also a feature of somatic evolution in tissues with spatial organization. Selection of the same drivers are found at similar burdens in solid tissues across patients, and thus accurate phylogenetic reconstruction in solid tissues may allow our method to be applicable in a variety of disease types. More comprehensive methods and datasets leading to the construction of more accurate phylogenetic trees, when combined with the methods presented here, will enable researchers and clinicians to quickly draw conclusions about net growth rate from mutational data.

Methods

Simulating the exact genealogy

We present here Lambert's construction of the exact genealogy of a sample of size n at time T from a birth-death process[30]. The idea is to describe the genealogical tree of n individuals from $n-1$ random variables $H_1, \ldots, H_{n-1}$ which represent coalescence times. To reconstruct the tree from the coalescence times, we begin by drawing a vertical line of height T. We then draw vertical lines of heights $H_1, \ldots, H_{n-1}$ and, at the top of each vertical line, draw a horizontal line to the left, stopping when it hits a vertical branch. The resulting tree is ultrametric, meaning that the root to tip distance is the same for all tips. See figure 1C-D. This construction is known as a coalescent point process and goes back to the work of Popovic[50] and of Aldous and Popovic[51] in the setting of critical branching processes.

By building on earlier work of Stadler[52] and Lambert and Stadler[53], who considered the case in which each individual in the population is sampled with some fixed probability y, Lambert[30] showed that we obtain exact the genealogy of a sample of size n from a birth-death process at time T, conditional on the population size at time T being at least n, if we choose $H_1, \ldots, H_{n-1}$ by the following two-step procedure:

1. Choose a random variable Y with probability density function on $(0, 1)$ given by

$$f_Y(y) = \frac{n\delta_T y^{n-1}}{(y + \delta_T - y\delta_T)^{n+1}}, \qquad \delta_T = \frac{re^{-rT}}{\lambda(1 - e^{-rT}) + re^{-rT}}. \tag{9}$$

2. Conditional on $Y = y$, let the random variables $H_1, \ldots, H_{n-1}$ be i.i.d. with probability density function on $(0, T)$ given by

$$f_{H_i|Y=y}(t) = \frac{y\lambda + (r - y\lambda)e^{-rT}}{y\lambda(1 - e^{-rT})} \cdot \frac{y\lambda r^2 e^{-rt}}{(y\lambda + (r - y\lambda)e^{-rt})^2}. \tag{10}$$

Note that the formula for the density of Y comes from equation (12) in Lambert[30], and that δ_T here is $1-a$ in Lambert[30]. The density for H_i comes from equation (7) in Lambert[30]. One can check that the resulting joint density for $H_1, \ldots, H_{n-1}$ matches the joint density for the ordered coalescence times given in Proposition 19 of Harris, Johnston, and Roberts[29].

While Lambert's construction is only exact when the birth and death rates are constant over time, leading to a population which grows exponentially at a constant rate, Cheek[54] has shown that under certain conditions, the construction remains approximately valid even when the growth rate of the population slows over time, provided that the population is still growing superlinearly at the time T when the sample is taken. For example, this method should give a good approximation in certain models of logistic population growth, provided that the sample is taken before the population reaches a fraction x of its carrying capacity, where $0 < x < 1$[54]. Consequently, we believe that our methods, while derived in the case of constant birth and death rates, may be more broadly applicable.

Internal and external branch lengths

Here, we state our main limit theorem, which describes the lengths of the internal and external branches for the genealogical tree of a birth-death process when T and n are large. The asymptotic distribution of the internal branch lengths can be used to estimate the net growth rate of a clone, while the ratio of external to internal branch lengths could provide an estimate of the clone age when the growth rate is known, as detailed further in Estimating the clone age.

We call a branch of the genealogical tree internal if it is ancestral to between 2 and $n-1$ of the n leaves (red edges in Figure 1C-D) and external if it is ancestral to only one of the n leaves (blue edges in Figure 1C-D). Note that a mutation along an internal branch will be inherited by more than one of the sampled individuals, while a mutation along an external branch will be unique to one of the sampled individuals. Therefore, one can estimate the internal and external branch lengths from the number of shared and private mutations respectively.

The site frequency spectrum and the allele frequency spectrum have previously been studied for populations whose genealogy is described by a coalescent point process[34,42–45]. Because the internal and external branch lengths are closely related to the site frequency spectrum, our methods for using the coalescent point process to understand the internal and external branch lengths are similar to the methods used in these previous works. However, these earlier results are applicable when we are interested in the site frequency spectrum of the entire population, or when each individual is sampled independently with some probability p. Our result pertains to the case of a sample of fixed size n from a much larger population, leading to a star-shaped genealogical tree with long external branches in which most of the coalescence occurs near the root of the tree.

The asymptotic distribution for the internal and external branch lengths was obtained for the classical Kingman's coalescent[55] and for for coalescents with multiple mergers[56,57]. Recently, an asymptotic result for external branch lengths in Yule trees was proved[58]. However, as far as we know, such results have not previously been established for a sample of size n from a birth-death tree.

To state our theorem, we need to consider a sequence of birth-death processes indexed by the sample size n, and the time at which the sample is taken, which we will now denote by T_n, must tend to infinity with n. We will also write λ_n, μ_n, r_n, and ν_n for the birth, death, growth, and mutation rates respectively to emphasize that we allow them to depend on n. We will let L_n^{in} and L_n^{ex} denote the total length of all internal and external branches respectively in the genealogical tree.

Theorem 1. *Suppose*

$$\lim_{n\to\infty} ne^{-r_n T_n} = 0. \tag{11}$$

Then, using $\xrightarrow{P}$ *to denote convergence in probability as* $n \to \infty$*, we have*

$$\frac{r_n L_n^{in}}{n} \xrightarrow{P} 1. \tag{12}$$

Furthermore, suppose instead we have

$$\lim_{n\to\infty} n^{3/2}(\log n)e^{-r_n T_n} = 0. \tag{13}$$

Let Z have a standard normal distribution, and let W have an exponential distribution with mean 1, independent of Z. Then

$$\left(\frac{r_n}{\sqrt{n}}\left(L_n^{in} - \frac{n}{r_n}\right), \frac{r_n}{n}L_n^{ex} - r_n T_n + \log n + 1\right) \Rightarrow (Z, \log W), \tag{14}$$

where $\Rightarrow$ denotes convergence in distribution as $n \to \infty$.

Recall that the expected population size at time T_n is $e^{r_n T_n}$, so the condition (11) means that the sample size n must be much smaller than the population size. Under this condition, the total internal branch length L_n^{in} is close to n/r_n with high probability, which means the growth rate estimate (4) should be accurate. Under the stronger condition (13), the distribution of the total internal branch length L_n^{in} is approximately normal with mean n/r_n and standard deviation $\sqrt{n}/r_n$, which we denote by

$$L_n^{in} \sim \mathcal{N}(n/r_n,\ \sqrt{n}/r_n)\,. \tag{15}$$

This means that the confidence interval in (5) should be accurate.

Because $\mathbb{E}[\log W] = -\gamma$, where $\gamma \approx .577$ is Euler's constant, Theorem 1 also suggests that for the total external branch length,

$$\mathbb{E}[L_n^{ex}] \approx nT_n - \frac{n}{r_n}(\log n + 1 - \mathbb{E}[\log W]) = nT_n - \frac{n}{r_n}(\log n + 1 + \gamma). \tag{16}$$

Shared and private mutations

Let M_n^{in} denote the number of mutations that appear on two or more of the sampled individuals, and let M_n^{ex} denote the number of mutations that appear on only one of the sampled individuals. Because we are assuming that mutations occur along each lineage at rate ν_n, the conditional distribution of M_n^{in} given L_n^{in} is Poisson with mean $\nu_n L_n^{in}$, and likewise for M_n^{ex}. In particular, we have

$$\mathbb{E}[M_n^{in}] = \nu_n \mathbb{E}[L_n^{in}] \approx \frac{n\nu_n}{r_n}. \tag{17}$$

and, using the conditional variance formula,

$$\mathrm{Var}(M_n^{in}) = \mathbb{E}[\mathrm{Var}(M_n^{in}|L_n^{in})] + \mathrm{Var}(\mathbb{E}[M_n^{in}|L_n^{in}]) = \nu_n \mathbb{E}[L_n^{in}] + \nu_n^2 \mathrm{Var}(L_n^{in}).$$

Note that the approximation for $\mathbb{E}[M_n^{in}]$ is consistent with (8) because $\sum_{k=2}^{\infty} 1/(k(k-1)) = 1$. The following corollary to Theorem 1 shows that M_n^{in} has an asymptotically normal distribution.

Corollary 2. *Suppose that* (13) *holds and that*

$$\lim_{n\to\infty} \frac{\nu_n n}{r_n} = \infty. \tag{18}$$

Let

$$\sigma_n^2 = n\left(\frac{\nu_n}{r_n} + \frac{\nu_n^2}{r_n^2}\right).$$

Let Z have a standard normal distribution. Then

$$\frac{1}{\sigma_n}\left(M_n^{in} - \frac{n\nu_n}{r_n}\right) \Rightarrow Z,$$

where $\Rightarrow$ denotes convergence in distribution as $n \to \infty$.

Also, using (16) we have the approximation

$$\mathbb{E}[M_n^{ex}] \approx n\nu_n T_n - \frac{n\nu_n}{r_n}(\log n + 1 + \gamma). \tag{19}$$

A similar formula was derived by Durrett[37]. For the private mutations M_n^{ex}, depending on the mutation rate, the dominant source of fluctuations could either be the Gaussian fluctuations from the mutations process or the non-Gaussian fluctuations from the random variable W.

Comparing analytical estimates to those using Phylofit

As described by Williams et al., Phylofit is "an efficient MCMC approach that models selection / growth by directly fitting the three parameter deterministic phase population trajectory using the joint probability density of coalescence times given the population size trajectory"[22]. The three parameters mentioned are the growth rate, the total number of hematopoietic stem cells, and the midpoint time of a deterministic logistic growth population trajectory. The likelihood function for the coalescence times is based on Equation 1 in Lan et. al[59]. In this sense, it is similar to our approach, leveraging the information provided by coalescence times. However, there are two ways to run Phylofit, only one of which is directly comparable to our methods.

Phylofit optionally incorporates Aberrant Cell Fraction (ACF) into its calculation of the likelihood. Aberrant Cell Fraction is simply the number of sampled cells within a clone divided by the total number of sampled cells. We refer to it alternatively as "sampled clonal fraction" in Application to human blood datasets, but use "Aberrant Cell Fraction" here for consistency with the terminology in Williams et al.[22]. In this sense, it is analogous to the variant allele frequency that would be observed in bulk whole genome sequencing. When incorporating ACF, Phylofit assumes that the population trajectory of the clone is logistic, with a carrying capacity for the clone equal to the total number of hematopoietic stem cells. While coalescence events provide information from the early history of the clone, when the total clone size is on the order of the sample size n, the ACF provides information on the clone size at the time of sampling. Therefore, while the assumptions for our methods simply require an early expansion phase with constant birth and death rates, Phylofit with ACF assumes a

logistic population size trajectory with a carrying capacity indicating that a clone becoming completely dominant in the blood (ACF=1). This is a much stronger assumption than is necessary for our methods or when using Phylofit without incorporating ACF. If the logistic population size trajectory is not the actual population size trajectory of a clone, the estimated growth rate from Phylofit with ACF will be affected. The population size trajectory of a clone at the time of sampling is likely to have been affected by an ACF carrying capacity below 1, treatment, and/or other competing expansions. In fact, deceleration of clonal expansion rates diverging from the logistic trajectory is observed in the recent work published by Mitchell et al.[24] and Fabre et al.[19]. Therefore, we compare our estimates to Phylofit without ACF, as this method, like our methods, estimates the growth rate during the expansion phase of the clone. However, we note that the published results in[22] do use ACF, so our Phylofit results differ from those which they present. For more details on ACF and clonal deceleration, see Supplementary section 3.1.

Estimating the clone age

While we did not find a dataset appropriate for applying our method of estimating the clone age (i.e., time from clone initiation to time of sampling), we show here how it can be done when the growth rate of a clone is known and the mutation rate is unknown. If the mutation rate is known, then estimating the clone age is straightforward because we can simply divide the average number of mutations on the sampled individuals by the mutation rate. We therefore focus on how to estimate the clone age when the growth rate is known but the mutation rate is unknown. Note first that it is not possible to make such an estimate by using only the shared mutations. To see this, consider the figure below in which the dots represent mutations. Because the genealogical tree is nearly star-shaped, we will see the same shared mutations regardless of whether we take the sample at time $T/2$ or time T. The only difference is that if the sample is taken at time T, we will see more private mutations. We therefore estimate the tumor age by comparing the number of shared mutations to the number of private mutations. From (17) and (19), a natural estimate of the age of the tumor is

$$\hat{T} = \frac{M_n^{ex}}{rM_n^{in}} + \frac{\log n + 1 + \gamma}{r}.$$

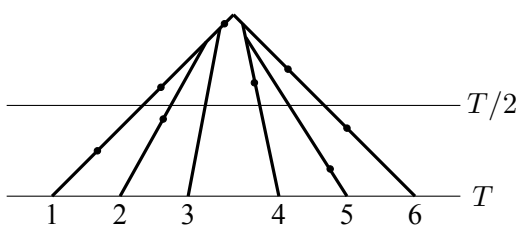


Figure 7: Six lineages sampled at times $T/2$ and T

The result below establishes some asymptotic properties of this estimate, and shows how to obtain a confidence interval for T. Condition (20) is needed to ensure that the Gaussian fluctuations from the lengths of the internal branches and the mutations process are the dominant source of fluctuations,

rather than the non-Gaussian fluctuations from the random variable W that measures the initial growth of the branching process.

Corollary 3. *Suppose the following conditions hold:*

$$\lim_{n\to\infty} \frac{\log n}{r_n T_n} = 0, \qquad \lim_{n\to\infty} \frac{\nu_n n}{r_n} = \infty, \qquad \lim_{n\to\infty} \frac{n\nu_n}{r_n^2 (r_n + \nu_n) T_n^2} = 0. \tag{20}$$

Define

$$\hat{T}_n = \frac{M_n^{ex}}{r_n M_n^{in}} + \frac{\log n + 1 + \gamma}{r_n},$$

as introduced above. Let Z have a standard normal distribution. Then

$$\frac{1}{T_n}\sqrt{\frac{n\nu_n}{r_n + \nu_n}}\left(\hat{T}_n - T_n\right) \Rightarrow Z.$$

From this result, one can show that for $0 < \alpha < 1$, an asymptotically valid $100(1-\alpha)\%$ confidence interval for T_n can be obtained by

$$\left[\frac{\hat{T}_n}{1 + \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{1 + \frac{n}{M_n^{in}}}}, \frac{\hat{T}_n}{1 - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{1 + \frac{n}{M_n^{in}}}}\right].$$

Longitudinal modeling

Longitudinal clone inclusion criteria

In longitudinal data from individuals without hematological malignancies[19], sequential data typically shows a lower but increasing VAF, making it ideal for our estimates of early growth rate. However, in longitudinal data from individuals with MPN[22], increased clonal competition and treatment gives rise to more complicated dynamics. Many clones which at first appear advantageous are outcompeted by clones with higher fitness or knocked down by treatment, leading to clones occasionally decreasing in VAF over the sampled time period. To avoid these external effects, which are not representative of the growth rate during the early expansion phase, we did not consider longitudinal data from clones decreasing in size. Any data points featuring more than a 20% decrease in Variant Allele Frequency (VAF) from a previous timepoint were removed. Removal due to a 20% drop was only applied if the previous VAF was ≥ 0.05, to avoid removing data due to small fluctuations at low VAF. After removing any decreasing data, we required 4 data points in total and at least 2 with a VAF > 0 and ≤ 0.25. Finally, because we are looking for expanding clones, we remove any clones which do not increase in VAF by at least 0.05.

As noted in Longitudinal validation of clone growth estimates, this filtering leaves us with 60 clones, with 56 from Fabre et al.[19] and 4 from Williams et al.[22]. Of these, one clone from Williams et al.[22] and two clones from Fabre et al.[19] have sufficient matched single cell data to make growth rate estimates from both data modalities. We note that the only MPN clone with matched single-cell and longitudinal

data (PD9478: *JAK2* + *DNMT3A*) is also from a patient that is essentially untreated, with the only intervention being venesection (bleeding). Further, the *JAK2* + *DNMT3A* clone in this patient appears to be the only large expansion, and there are 11 longitudinal samples from the same cell type, making this clone an ideal candidate for validation.

From the remaining 57 clones without matched single cell data, we identified 17 clones with mutations in the same driver gene as a single cell clone. We then excluded the longitudinal data from two MPN clones from Williams et al.[22], as treatment likely affected the VAF in both cases. Further, we removed one *JAK2* clone where there was a competing clone at high VAF, as this was likely to affect the growth rate. After fitting the logistic growth rate model to each of the remaining clones, as detailed in Logistic modeling of longitudinal data, we removed one DNMT3A clone where the fit failed to converge. After all filtering steps, there were 13 longitudinal clones matching to 10 single cell clones. *DNMT3A* mutant clones were the most abundant, with 7 *DNMT3A* clones having longitudinal data and 5 having single cell data. There were 3 longitudinal *JAK2* clones and 4 single cell *JAK2* clones. There were 3 longitudinal *PPM1D* clones and 1 single cell *PPM1D* clone.

Logistic modeling of longitudinal data

We used the *nls()* function from the *stats* package in R, with the port algorithm, to perform fitting to the following logistic growth equation, which models the Variant Allele Frequency (VAF) over time:

$$\text{VAF}(t) = \frac{K}{1 + e^{-(rt+\phi)}} \tag{21}$$

Bounds for K, representing the carrying capacity, are $[0, 0.5]$, as all mutations are in diploid regions. Bounds for r are $[0, 5]$ per year, and bounds for ϕ are $[-500, 0]$. 95% confidence intervals are found by assuming normality of the parameter estimate, using $r \pm 1.96 *$ stdError to calculate the bounds. Midpoint time of the logistic curve is given by $t_m = -\phi/r$. See Data and code availability for exact implementation and for steps to reproduce.

Importantly, we do not assume a carrying capacity K equal to a variant allele frequency of 0.5, instead allowing the carrying capacity to be fit simultaneously with the growth rate and midpoint time. This is a distinction from the inherent assumptions in the Phylofit with ACF approach, which is discussed in Comparing analytical estimates to those using Phylofit and further in Supplementary section 3.1. Our decision to fit K rather than fixing it at 0.5 is motivated by data[19] showing growth rates slowing more than would be expected by a logistic fit with $K = 0.5$, even in the absence of treatment. In fitting to longitudinal data from 16 clones (3 from Figure 6A-C and 13 from Figure 6E) 9 have a fit VAF carrying capacity below 0.4, consistent with the claim that clones do not always saturate at an allele frequency of 0.5. A possible explanation for this lower carrying capacity is lineage bias which may lead to clonal dominance in only a subset of the blood progenitors. For example, *JAK2* mutants were found at higher clonal fractions in megakaryocyte and erythroid progenitors[23]. Variant allele frequency in whole blood may saturate below 0.5, even when a clone has become completely dominant within a

specific type of progenitor (i.e. Megakaryocyte Erythroid Progenitors (MEP) in *JAK2* mutants). In fact, single cell clonal fraction data from the 13 MPN patients we have analyzed shows no example of a somatic clone that is completely dominant, despite matched stromal[23] or buccal[22] normal cell samples distinguishing between somatic and germline variants.

It should be noted that a logistic growth model with any carrying capacity may not be a good model for clones in the blood, especially in the presence of other clones and/or treatment[48]. Our coalescent methods avoid this dependence on a particular model by assuming only that exponential growth occurs immediately following the initiation of a clone, while its population size is still on the order of the sample size, n. However, longitudinal validation requires the choice of a particular growth model in order to estimate a growth rate. Based on previous work exploring practical parameter identifiability in sigmoid growth models[41], the logistic model outperformed Gompertz and Richards' models, which is why we use it for longitudinal validation.

Data and code availability

We have created an open source R package called *cloneRate* to perform growth rate estimation using ultrametric or mutation-based phylogenetic trees as input data (`https://github.com/bdj34/cloneRate/`). *cloneRate* also includes methods for rapid generation of exact sampled trees from supercritical birth-death processes based on the work of Lambert[30]. Users can apply our methods to new data and also recreate our results herein using published data[19,22–24], which is included in the package for convenience. Articles containing the details needed to reproduce our analyses are available from the package website, along with other vignettes to guide users (`https://bdj34.github.io/cloneRate/`).

References

[1] Lee-Six, H.; Olafsson, S.; Ellis, P.; Osborne, R. J.; Sanders, M. A.; Moore, L.; Georgakopoulos, N.; Torrente, F.; Noorani, A.; Goddard, M., et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **2019**, *574*, 532–537.

[2] Jonason, A. S.; Kunala, S.; Price, G. J.; Restifo, R. J.; Spinelli, H. M.; Persing, J. A.; Leffell, D. J.; Tarone, R. E.; Brash, D. E. Frequent clones of p53-mutated keratinocytes in normal human skin. *Proceedings of the National Academy of Sciences* **1996**, *93*, 14025–14029.

[3] Martincorena, I.; Fowler, J. C.; Wabik, A.; Lawson, A. R.; Abascal, F.; Hall, M. W.; Cagan, A.; Murai, K.; Mahbubani, K.; Stratton, M. R., et al. Somatic mutant clones colonize the human esophagus with age. *Science* **2018**, *362*, 911–917.

[4] Martincorena, I.; Roshan, A.; Gerstung, M.; Ellis, P.; Van Loo, P.; McLaren, S.; Wedge, D. C.; Fullam, A.; Alexandrov, L. B.; Tubio, J. M., et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **2015**, *348*, 880–886.

[5] Jaiswal, S.; Fontanillas, P.; Flannick, J.; Manning, A.; Grauman, P. V.; Mar, B. G.; Lindsley, R. C.; Mermel, C. H.; Burtt, N.; Chavez, A., et al. Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine* **2014**, *371*, 2488–2498.

[6] Suda, K.; Nakaoka, H.; Yoshihara, K.; Ishiguro, T.; Tamura, R.; Mori, Y.; Yamawaki, K.; Adachi, S.; Takahashi, T.; Kase, H., et al. Clonal expansion and diversification of cancer-associated mutations in endometriosis and normal endometrium. *Cell reports* **2018**, *24*, 1777–1789.

[7] Yokoyama, A.; Kakiuchi, N.; Yoshizato, T.; Nannya, Y.; Suzuki, H.; Takeuchi, Y.; Shiozawa, Y.; Sato, Y.; Aoki, K.; Kim, S. K., et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **2019**, *565*, 312–317.

[8] Steensma, D. P.; Ebert, B. L. Clonal hematopoiesis as a model for premalignant changes during aging. *Experimental hematology* **2020**, *83*, 48–56.

[9] Warren, J. T.; Link, D. C. Clonal hematopoiesis and risk for hematologic malignancy. *Blood* **2020**, *136*, 1599–1605.

[10] Scott, J.; Marusyk, A. Somatic clonal evolution: a selection-centric perspective. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **2017**, *1867*, 139–150.

[11] Greaves, M.; Maley, C. C. Clonal evolution in cancer. *Nature* **2012**, *481*, 306–313.

[12] Shen, Y. J.; Mishima, Y.; Shi, J.; Sklavenitis-Pistofidis, R.; Redd, R. A.; Moschetta, M.; Manier, S.; Roccaro, A. M.; Sacco, A.; Tai, Y.-T., et al. Progression signature underlies clonal evolution and dissemination of multiple myeloma. *Blood* **2021**, *137*, 2360–2372.

[13] van Zeventer, I. A.; de Graaf, A. O.; Wouters, H. J.; van der Reijden, B. A.; van der Klauw, M. M.; de Witte, T.; Jonker, M. A.; Malcovati, L.; Jansen, J. H.; Huls, G. Mutational spectrum and dynamics of clonal hematopoiesis in anemia of older individuals. *Blood* **2020**, *135*, 1161–1170.

[14] Geiger, H.; De Haan, G.; Florian, M. The ageing haematopoietic stem cell compartment. *Nature Reviews Immunology* **2013**, *13*, 376–389.

[15] Schenz, J.; Rump, K.; Siegler, B. H.; Hemmerling, I.; Rahmel, T.; Thon, J. N.; Nowak, H.; Fischer, D.; Hafner, A.; Tichy, L., et al. Increased prevalence of clonal hematopoiesis of indeterminate potential in hospitalized patients with COVID-19. *Frontiers in Immunology* **2022**, *13*.

[16] Tall, A. R.; Fuster, J. J. Clonal hematopoiesis in cardiovascular disease and therapeutic implications. *Nature Cardiovascular Research* **2022**, *1*, 116–124.

[17] Sano, S.; Horitani, K.; Ogawa, H.; Halvardson, J.; Chavkin, N. W.; Wang, Y.; Sano, M.; Mattisson, J.; Hata, A.; Danielsson, M., et al. Hematopoietic loss of Y chromosome leads to cardiac fibrosis and heart failure mortality. *Science* **2022**, *377*, 292–297.

[18] Gillis, N. K.; Ball, M.; Zhang, Q.; Ma, Z.; Zhao, Y.; Yoder, S. J.; Balasis, M. E.; Mesa, T. E.; Sallman, D. A.; Lancet, J. E., et al. Clonal haemopoiesis and therapy-related myeloid malignancies in elderly patients: a proof-of-concept, case-control study. *The lancet oncology* **2017**, *18*, 112–121.

[19] Fabre, M. A.; de Almeida, J. G.; Fiorillo, E.; Mitchell, E.; Damaskou, A.; Rak, J.; Orrù, V.; Marongiu, M.; Chapman, M. S.; Vijayabaskar, M., et al. The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* **2022**, 1–8.

[20] Watson, C. J.; Papula, A.; Poon, G. Y.; Wong, W. H.; Young, A. L.; Druley, T. E.; Fisher, D. S.; Blundell, J. R. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **2020**, *367*, 1449–1454.

[21] Abelson, S.; Collord, G.; Ng, S. W.; Weissbrod, O.; Mendelson Cohen, N.; Niemeyer, E.; Barda, N.; Zuzarte, P. C.; Heisler, L.; Sundaravadanam, Y., et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **2018**, *559*, 400–404.

[22] Williams, N.; Lee, J.; Mitchell, E.; Moore, L.; Baxter, E. J.; Hewinson, J.; Dawson, K. J.; Menzies, A.; Godfrey, A. L.; Green, A. R., et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **2022**, *602*, 162–168.

[23] Van Egeren, D.; Escabi, J.; Nguyen, M.; Liu, S.; Reilly, C. R.; Patel, S.; Kamaz, B.; Kalyva, M.; DeAngelo, D. J.; Galinsky, I., et al. Reconstructing the lineage histories and differentiation trajectories of individual cancer cells in myeloproliferative neoplasms. *Cell stem cell* **2021**, *28*, 514–523.

[24] Mitchell, E.; Spencer Chapman, M.; Williams, N.; Dawson, K. J.; Mende, N.; Calderbank, E. F.; Jung, H.; Mitchell, T.; Coorens, T. H.; Spencer, D. H., et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **2022**, 1–8.

[25] Kingman, J. F. C. The coalescent. *Stochastic processes and their applications* **1982**, *13*, 235–248.

[26] Griffiths, R. C.; Tavare, S. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **1994**, *344*, 403–410.

[27] Slatkin, M.; Hudson, R. R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **1991**, *129*, 555–562.

[28] Karcher, M. D.; Palacios, J. A.; Lan, S.; Minin, V. N. phylodyn: an R package for phylodynamic simulation and inference. *Molecular ecology resources* **2017**, *17*, 96–100.

[29] Harris, S. C.; Johnston, S. G.; Roberts, M. I. The coalescent structure of continuous-time Galton–Watson trees. *The Annals of Applied Probability* **2020**, *30*, 1368–1414.

[30] Lambert, A. The coalescent of a sample from a binary branching process. *Theoretical Population Biology* **2018**, *122*, 30–35.

[31] Lee-Six, H.; Øbro, N. F.; Shepherd, M. S.; Grossmann, S.; Dawson, K.; Belmonte, M.; Osborne, R. J.; Huntly, B. J.; Martincorena, I.; Anderson, E., et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **2018**, *561*, 473–478.

[32] Moeller, M. E.; Pere, N. V. M.; Werner, B.; Huang, W. Measures of genetic diversification in somatic tissues at bulk and single cell resolution. *bioRxiv* **2022**,

[33] Ignatieva, A.; Hein, J.; Jenkins, P. A. A characterisation of the reconstructed birth–death process through time rescaling. *Theoretical Population Biology* **2020**, *134*, 61–76.

[34] Delaporte, C.; Achaz, G.; Lambert, A. Mutational pattern of a sample from a critical branching population. *Journal of Mathematical Biology* **2016**, *73*, 627–664.

[35] Henningsen, A.; Toomet, O. maxLik: A package for maximum likelihood estimation in R. *Computational Statistics* **2011**, *26*, 443–458.

[36] Antle, C.; Klimko, L.; Harkness, W. Confidence Intervals for the Parameters of the Logistic Distribution. *Biometrika* **1970**, *57*, 397–402.

[37] Durrett, R. Population genetics of neutral mutations in exponentially growing cancer cell populations. *Annals of Applied Probability* **2013**, *23*, 230.

[38] Gunnarsson, E. B.; Leder, K.; Foo, J. Exact site frequency spectra of neutrally evolving tumors: A transition between power laws reveals a signature of cell viability. *Theoretical Population Biology* **2021**, *142*, 67–90.

[39] Williams, M. J.; Werner, B.; Barnes, C. P.; Graham, T. A.; Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature genetics* **2016**, *48*, 238–244.

[40] Bozic, I.; Gerold, J. M.; Nowak, M. A. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. *PLOS Computational Biology* **2016**, *12*, e1004731.

[41] Simpson, M. J.; Browning, A. P.; Warne, D. J.; Maclaren, O. J.; Baker, R. E. Parameter identifiability and model selection for sigmoid population growth models. *Journal of theoretical biology* **2022**, *535*, 110998.

[42] Lambert, A. The allelic partition for coalescent point processes. *Markov Processes and Related Fields* **2009**, *15*, 359–386.

[43] Champagnat, N.; Lambert, A. Splitting trees with neutral Poissonian mutations I: Small families. *Stochastic Processes and their Applications* **2012**, *122*, 1003–1033.

[44] Champagnat, N.; Henry, B. Moments of the frequency spectrum of a splitting tree with neutral Poissonian mutations. *Electronic Journal of Probability* **2016**, *21*, 1–34.

[45] Dinh, K. N.; Jaksik, R.; Kimmel, M.; Lambert, A.; Tavaré, S. Statistical Inference for the Evolutionary History of Cancer Genomes. *Statistical Science* **2020**, *35*, 129–144.

[46] Stahl, M.; Abdel-Wahab, O.; Wei, A. H.; Savona, M. R.; Xu, M. L.; Xie, Z.; Taylor, J.; Starczynowski, D.; Sanz, G. F.; Sallman, D. A., et al. An agenda to advance research in myelodysplastic syndromes: A TOP 10 Priority List from the first international workshop in MDS. *Blood Advances* **2023**,

[47] Savona, M. R.; Malcovati, L.; Komrokji, R.; Tiu, R. V.; Mughal, T. I.; Orazi, A.; Kiladjian, J.-J.; Padron, E.; Solary, E.; Tibes, R., et al. An international consortium proposal of uniform response criteria for myelodysplastic/myeloproliferative neoplasms (MDS/MPN) in adults. *Blood, The Journal of the American Society of Hematology* **2015**, *125*, 1857–1865.

[48] Bolton, K. L.; Ptashkin, R. N.; Gao, T.; Braunstein, L.; Devlin, S. M.; Kelly, D.; Patel, M.; Berthon, A.; Syed, A.; Yabe, M., et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nature genetics* **2020**, *52*, 1219–1226.

[49] Evrony, G. D.; Hinch, A. G.; Luo, C. Applications of single-cell DNA sequencing. *Annual review of genomics and human genetics* **2021**, *22*, 171.

[50] Popovic, L. Asymptotic genealogy of a critical branching process. *The Annals of Applied Probability* **2004**, *14*, 2120–2148.

[51] Aldous, D.; Popovic, L. A critical branching process model for biodiversity. *Advances in Applied Probability* **2005**, *37*, 1094–1115.

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

[53] Lambert, A.; Stadler, T. Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theoretical Population Biology* **2013**, *90*, 113–128.

[54] Cheek, D. The coalescent tree of a Markov branching process with generalised logistic growth. *Journal of Mathematical Biology* **2022**, *84*.

[55] Dahmer, I.; Kersting, G. The internal branch lengths of the Kingman coalescent. *Annals of Applied Probability* **2015**, *25*, 1325–1348.

[56] Dahmer, I.; Kersting, G.; Wakolbinger, A. The total external branch length of beta-coalescents. *Combinatorics, Probability, and Computing* **2015**, *23*, 1010–1027.

[57] Birkner, M.; Dahmer, I.; Diehl, C.; Kersting, G. The joint fluctuations of the lengths of the Beta$(2-\alpha,\alpha)$-coalescents. 2020; ArXiv Preprint 2009.13642.

[58] Disanto, F.; Fuchs, M. Distribution of external branch lengths in Yule trees. 2022; ArXiv Preprint 2208.04804.

[59] Lan, S.; Palacios, J. A.; Karcher, M.; Minin, V. N.; Shahbaba, B. An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. *Bioinformatics* **2015**, *31*, 3282–3289.

[60] Johnston, S. G. The genealogy of Galton-Watson trees. *Electronic Journal of Probability* **2019**, *24*, 1–35.

[61] Thorisson, H. *Coupling, Stationarity, and Regeneration*; Springer: New York, 2000.

[62] Balakrishnan, N. *Handbook of the Logistic Distribution*; Marcel Dekker: New York, 1992.

[63] Diananda, P. The central limit theorem for m-dependent variables. Mathematical Proceedings of the Cambridge Philosophical Society. 1955; pp 92–95.

[64] Durrett, R. *Probability: Theory and Examples*, 5th ed.; Cambridge University Press: Cambridge, 2019.

[65] Sethuraman, J. Some Limit Theorems for Joint Distributions. *The Indian Journal of Statistics, Series A* **1961**, *23*, 379–386.

[66] Steensma, D. P. Clinical consequences of clonal hematopoiesis of indeterminate potential. *Hematology 2014, the American Society of Hematology Education Program Book* **2018**, *2018*, 264–269.