RESEARCH ARTICLE

Inferring parameters of cancer evolution in chronic lymphocytic leukemia

Nathan D. Lee[1], **Ivana Bozic**[1,2]*

1 Department of Applied Mathematics, University of Washington, Seattle, Washington, United States of America, **2** Herbold Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America

* ibozic@uw.edu

Abstract

As a cancer develops, its cells accrue new mutations, resulting in a heterogeneous, complex genomic profile. We make use of this heterogeneity to derive simple, analytic estimates of parameters driving carcinogenesis and reconstruct the timeline of selective events following initiation of an individual cancer, where two longitudinal samples are available for sequencing. Using stochastic computer simulations of cancer growth, we show that we can accurately estimate mutation rate, time before and after a driver event occurred, and growth rates of both initiated cancer cells and subsequently appearing subclones. We demonstrate that in order to obtain accurate estimates of mutation rate and timing of events, observed mutation counts should be corrected to account for clonal mutations that occurred after the founding of the tumor, as well as sequencing coverage. Chronic lymphocytic leukemia (CLL), which often does not require treatment for years after diagnosis, presents an optimal system to study the untreated, natural evolution of cancer cell populations. When we apply our methodology to reconstruct the individual evolutionary histories of CLL patients, we find that the parental leukemic clone typically appears within the first fifteen years of life.

Author summary

By the time a patient's cancer is diagnosed, it has been growing undetected for years, or even decades. A cancer's initiation, development, and progression are driven by a sequence of driver mutations, genetic alterations that confer a fitness advantage to the cells containing them. As a cancer expands, it also accumulates many neutral mutations that don't confer a growth advantage. As a result, tumors are highly heterogeneous, made up of different genetically distinct populations, or subclones, of cancer cells. Most cancers will require immediate treatment upon diagnosis, making study of their natural progression over time difficult. However, the blood cancer chronic lymphocytic leukemia (CLL) often does not require immediate treatment and is closely monitored for years, which makes it ideal for studying cancer evolution before treatment radically alters the cancer's dynamics. We make use of the complex tumor heterogeneity to reconstruct the timing of key driver events in the tumor's development, showing that the initial leukemic clone

Citation: Lee ND, Bozic I (2022) Inferring parameters of cancer evolution in chronic lymphocytic leukemia. PLoS Comput Biol 18(11): e1010677. https://doi.org/10.1371/journal.pcbi.1010677

Editor: Jasmine Foo, University of Minnesota, UNITED STATES

Received: May 26, 2022

Accepted: October 24, 2022

Published: November 4, 2022

Data Availability Statement: All code for analysis and simulations can be found at https://github.com/nathanlee543/Cancer_Inf_Sims. The manuscript is a computational study and no new data – besides simulated data – has been generated. Simulated data is available in the Github repository. CLL data analyzed is from previously published work (Gruber, M., Bozic, I., Leshchiner, I. et al. Growth dynamics in naturally progressing chronic lymphocytic leukaemia. Nature 570, 474-479 (2019). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6630176/) and is publicly available in that work's supplementary tables.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

often appears early in life. Additionally, we estimate mutation rate, subclone growth rates, and fitness advantage provided by driver mutations.

Introduction

When a cell accrues a sequence of driver mutations—genetic alterations that provide a proliferative advantage relative to surrounding cells—it can begin to divide uncontrollably and eventually develop the complex features of a cancer [1–3]. Thousands of specific driver mutations have been implicated in carcinogenesis, with individual tumors harboring from few to dozens of drivers, depending on the cancer type [4]. Mutations that don't have a significant effect on cellular fitness also arise, both before and after tumor initiation [5]. These neutral mutations, or "passengers", can reach detectable frequencies by random genetic drift or the positive selection of a driver mutation in the same cell [6–9]. Mutational burden detectable by bulk sequencing reveals tens to thousands of passengers per tumor [10, 11].

Genome sequencing technologies have revealed the heterogeneous, informative genetic profiles produced by the evolutionary process driving carcinogenesis [12, 13]. These genetic profiles have been used to obtain insight into specific features of the carcinogenic process operating in individual patients. For example, the molecular clock feature of passenger mutations has been employed to measure timing of early events in tumor formation, as well as identify stages of tumorigenesis and metastasis [14–22]. Other studies have estimated mutation rates [5, 23, 24], selective growth advantages of cancer subclones [25–28], and the effect of spatial structure on cancer evolution [29–31]. We note that previous approaches typically only estimate one or a few parameters of cancer evolution. In addition, many state-of-the-art methods make use of computationally expensive approaches [24, 30, 32] or simplifying assumptions, such as approximating tumor expansion as deterministic or ignoring cell death [27, 32]. Our approach relies on analytic formulas and sampling, which for realistic numbers of subclones and time points is efficient, and does not require simulation of tumor growth or computationally expensive model fitting.

Mathematical models of cancer progression, especially when used in conjunction with experimental and clinical data, can provide important insights into the evolutionary history of cancer [9, 19, 33–37]. Branching processes—a type of a stochastic process—can be used to model how different populations of dividing, dying, and mutating cells in a tumor evolve over time [38]. Their theory and applications have been well developed to model the multistage nature of cancer development [25, 29, 35, 38–40]. Here we use a branching process model of carcinogenesis to derive a comprehensive reconstruction of an individual tumor's evolution.

Tumors can grow for many years, even decades, before they reach detectable size [16]. Typically, tumor samples used for sequencing would be obtained at the end of the tumor's natural, untreated progression. More recently, longitudinal sequencing, where a tumor is sequenced at multiple times during its development, has provided better resolution of tumor growth dynamics and evolution in various cancer types [27, 41–44]. Chronic lymphocytic leukemia (CLL) is an ideal system for studying cancer evolution because it can be monitored, via peripheral blood samples, without treatment until disease progression [45].

We establish that two longitudinal bulk sequencing and tumor size measurements are sufficient to reconstruct virtually all parameters (mutation rate, growth rates, times of appearance of driver mutations, and time since the driver mutation) of cancer evolution in individual patients. Our analytic approach yields simple formulas for the parameters; thus, estimation of the parameters governing cancer growth is not computationally intensive, regardless of tumor size. Our framework makes possible a personalized, high-resolution reconstruction of a cancer's timeline of selective events and quantitative characterization of the evolutionary dynamics of the subclones making up the cancer cell population.

Results

Model

We consider a multi-type branching process of tumor expansion (Fig 1A). Tumor growth is started with a single initiated cell at time 0. Initiated tumor cells divide with rate b and die with rate d. These cells already have the driver mutations necessary for expansion, so we assume $b > d$. The population of initiated cells can go extinct due to stochastic fluctuations, or survive stochastic drift and start growing (on average) exponentially with net growth rate $r = b - d$. We will focus only on those populations that survived stochastic drift.

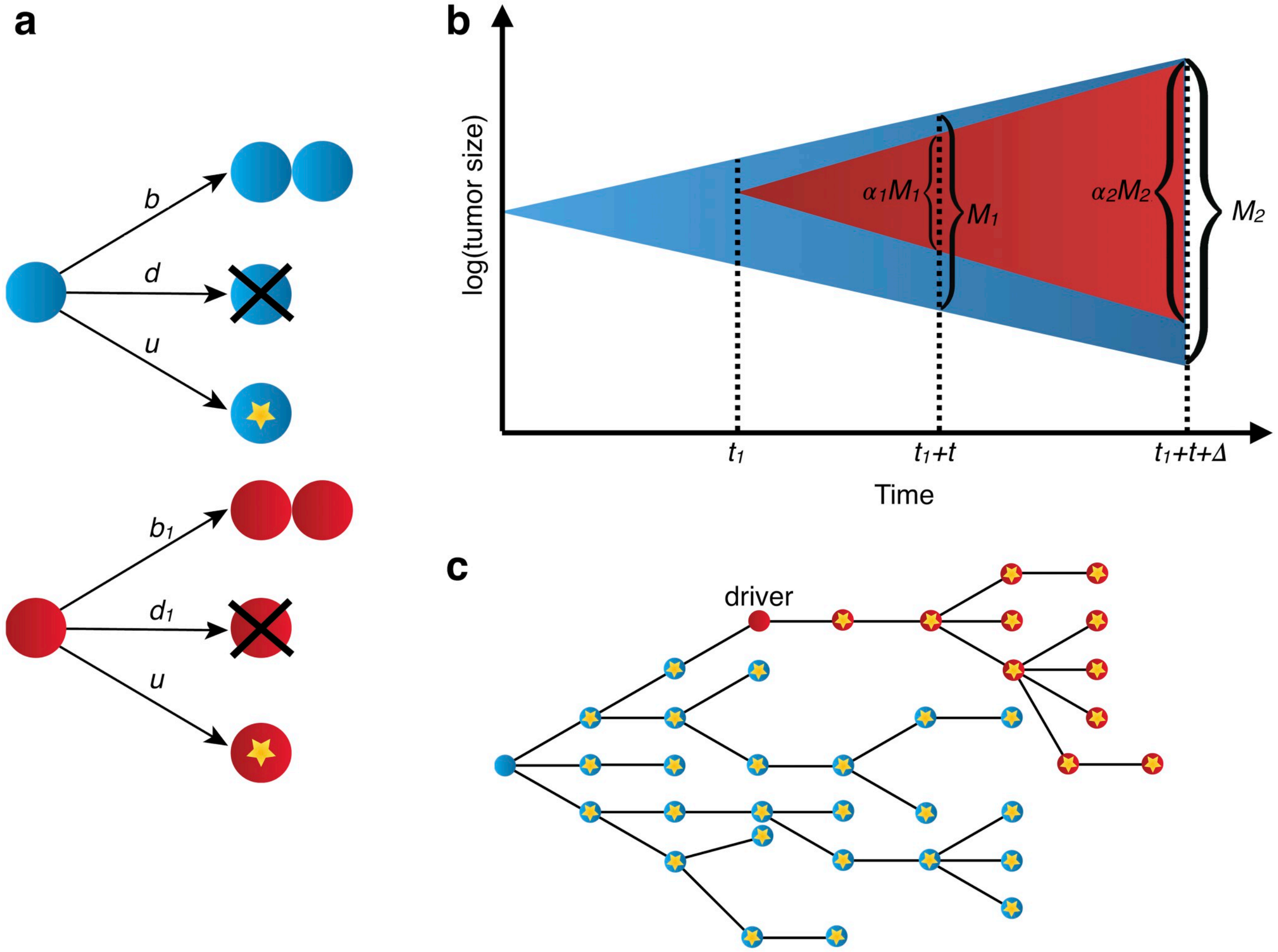

Fig 1. Stochastic branching process model of tumor evolution. (a) Stochastic branching process model for tumor expansion. Initiated tumor cells (blue) divide with birth rate b, die with death rate d, and accrue passenger mutations with mutation rate u. Type-1 cells, which carry the driver mutation, divide with birth rate b_1, die with death rate d_1, and accrue passenger mutations with mutation rate u. (b) The initiated tumor, or type-0, (blue) population growth is initiated from a single cell. A driver mutation occurs in a single type-0 cell at time t_1, starting the type-1 population (red). The tumor sample is collected and bulk sequenced at times $t_1 + t$ and $t_1 + t + \Delta$, where the driver fraction is α_1 and α_2, respectively. Tumor size (in number of cells) is M_1 and M_2 at first and second sample collection dates. (c) By the time the tumor is observed, it has a high level of genetic heterogeneity due to the mutations that have accrued in both type-0 (blue) and type-1 populations (red). Each yellow star represents a different passenger mutation.

https://doi.org/10.1371/journal.pcbi.1010677.g001

At some time $t_1 > 0$ a new driver mutation occurs in a single initiated tumor cell, starting a new independent birth-death process, with birth rate b_1 and death rate d_1 (Fig 1B). Net growth rate of cells with the new driver is $r_1 = b_1 - d_1$. The new driver increases the rate of growth, i.e., $r_1 > r$. We define the driver's selective growth advantage by $g = (r_1/r - 1)$. In addition, both populations of cells (with and without the driver) accrue passenger mutations with rate u (Fig 1C).

After the driver mutation occurs, an additional time t passes before the tumor is observed. Type-0 cells are original initiated tumor and type-1 cells contain the driver mutation. In Materials and methods we also analyze the more general case of two nested or sibling driver mutations, as well as the fully generalized case of any clonal structure that might arise during tumor expansion.

Parameter estimates from two longitudinal measurements

We demonstrate that with two longitudinal bulk sequencing measurements, it is possible to accurately estimate net growth rates, time of appearance of a driver mutation, time between a driver mutation and observation, and mutation rate in the tumor. The tumor is first sequenced at time of observation, $t_1 + t$, where both time of driver mutation, t_1, and time from driver mutation to observation, t, are yet unknown (Fig 1B). A second bulk sequencing is performed at $t_1 + t + \Delta$, a known Δ time units after the tumor is first observed (Fig 1B). Later, we apply our method to the CLL data from Ref. [27], where the average size of Δ for all the pre-treatment samples sequenced is 1.8 years (0.6–4.9 years). In general, we expect that in the case of smaller Δ values measurement errors would have a larger effect on the estimated growth rates, due to an expected smaller change in cancer cell count and subclonal structure during a smaller time interval. From the bulk sequencing data, the fraction of cells carrying the driver mutation, α_1 and α_2, can be measured at the time points $t_1 + t$ and $t_1 + t + \Delta$, respectively. We denote total number of cells in the tumor at the two bulk sequencing time points as M_1 and M_2. For liquid cancers, cell counts of the relevant cancer cell population serve as indicators of cancer progression. In the case of CLL, white blood cell (WBC) count is useful as a measure of tumor burden in peripheral blood, as it is routinely taken and includes the cancerous cell population. More precise estimates of tumor burden would include absolute lymphocyte count (ALC) and number of B lymphocytes. Both ALC and WBC counts can suffer from inaccuracies due to the prevalence of smudge cells in CLL, often resulting in an underestimate of these counts [46].

Equating expected values of the sizes of the type-0 and type-1 population at the two bulk sequencing time points with the measured numbers of cells present in clones 0 and 1, we obtain estimates of the net growth rates of the two subclones:

$$r = \frac{1}{\Delta} \log\left(\frac{(1-\alpha_2)M_2}{(1-\alpha_1)M_1}\right) \tag{1}$$

$$r_1 = \frac{1}{\Delta} \log\left(\frac{\alpha_2 M_2}{\alpha_1 M_1}\right). \tag{2}$$

From the growth rate estimates and subclone sizes, we can approximate the expected value of the time a population in a branching process takes to reach an observed size [38]. This yields an estimate of the time t from the appearance of driver mutation until observation:

$$t = \frac{1}{r_1} \log\left(M_1 \alpha_1\right). \tag{3}$$

Using the bulk sequencing data from the second time point, γ, the number of subclonal passengers between the specified frequencies f_1 and f_2, can be measured. Using results from previous work [47], we derive the expected value of γ (Materials and methods), which can be used to estimate the mutation rate u:

$$u = \frac{f_1 f_2 r r_1 \gamma}{(f_2 - f_1)(\alpha_2 r + r_1(1 - \alpha_2))}. \tag{4}$$

The m passenger mutations that were present in the original type-1 cell when the driver mutation occurred (Fig 1C) are present in all type-1 cells. m can be estimated from bulk sequencing data and used to estimate time of appearance of the driver. We maximize the likelihood function $P(m|t_1)$ with respect to time of appearance of the driver, t_1, (see Materials and methods) to obtain the maximum likelihood estimate

$$t_1 = \frac{m}{u}. \tag{5}$$

Using formulas (4) and (5), we can now estimate t_1.

Estimates verified in simulated tumors

To assess the accuracy of the parameter estimates for several modes of tumor evolution, we simulate tumor growth by performing a Monte Carlo simulation, which simulates the birth, death, and accumulation of mutations in the individual cells that make up a tumor. This simulation generates the mutation frequency and tumor size data used by the estimates (see Methods section for details of simulation). We simulate three different types of tumors (slow growing, fast growing, and no cell death), with a high and a low mutation rate for each (S1 Table).

In a simulation of a fast-growing tumor with a single subclonal driver mutation that confers a strong selective growth advantage of 100%, we can accurately estimate growth rates, mutation rate, time of driver event, and time since driver event (Fig 2A and 2B). Growth rates of both initiated tumor and driver subclones can be estimated with a high degree of accuracy, achieving mean percentage error (MPE) of 0.03% and -0.07% for the lower mutation rate (u = 1) scenario. The mutation rate u and estimates for time of driver appearance, t_1, and time since driver, t, can also be estimated accurately, with MPEs of -0.9%, 3.8%, and -0.4%, respectively. Estimates for u, t_1, and t have a somewhat greater degree of variation compared to the growth rate estimates, due to the inherent randomness of the number of mutations and time to reach the observed size that occur in each realization of the stochastic process.

For the parameter regime with no cell death and the regime for a slow-growing tumor, we again achieve high accuracies for the net growth rates (S1(A), S1(B), S2(A) and S2(B) Figs). In the lower mutation rate (u = 1) scenario, parameter estimates for the mutation rate u and time of driver appearance t_1 can be accurately estimated for both regimes, with MPEs of -1.3% and 4.9% for the no cell death case, and MPEs of -3% and 3.7% for the slow-growing tumor.

We note that the estimator for t (time since driver event) is biased, with the extent depending on the ratio of birth rate to net growth rate, and the tumor size. The underlying cause of the bias is due to a simplifying assumption in the estimator's derivation (see Methods, "Derivation of estimates of evolutionary parameters"), and this bias decreases as tumor size increases and as the ratio of growth and division rate gets closer to 1. For the three main modes of growth in our study, we performed additional Monte Carlo simulations to precisely quantify the effect of death:birth ratio and tumor size on the estimator's accuracy (S5 Fig). For all three modes of growth, we observe a monotonic decrease in error as tumor size increases to more

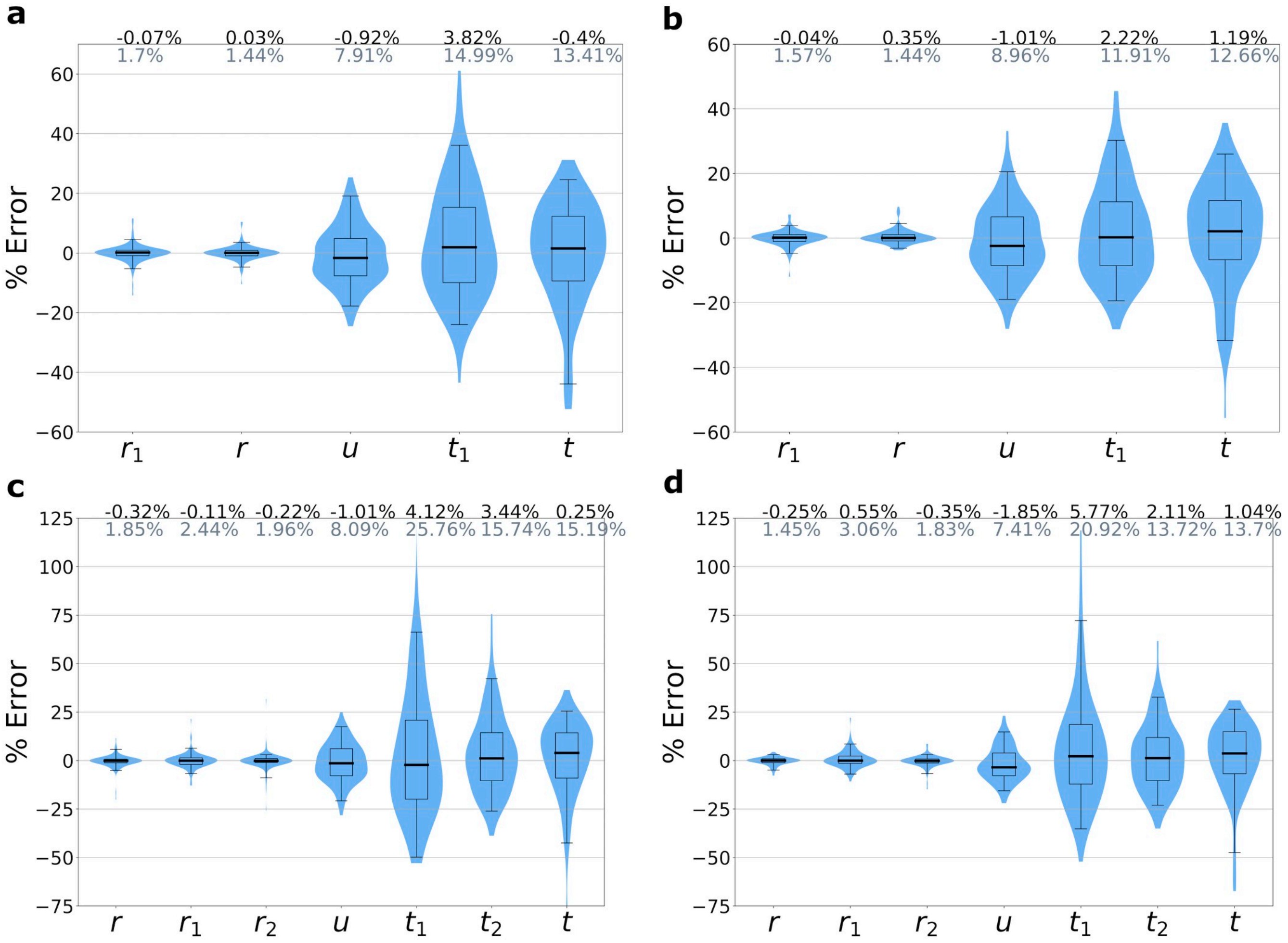

Fig 2. Accuracy of parameter inferences from simulated data. We simulated tumor growth by performing a Monte Carlo simulation, which simulates the birth, death, and accumulation of mutations in the individual cells that make up a tumor, and generates the mutation frequency and tumor size data used by the estimates. Simulations are of fast-growing tumors with (a) single driver subclone and mutation rate $u = 1$, (b) single driver subclone and $u = 3$, (c) two nested driver subclones with $u = 1$, and (d) two sibling driver subclones with $u = 1$. Mean percent errors (MPEs) of estimates are shown in black above the plots, and mean absolute percent errors (MAPEs) are shown in gray. Boxes contain 25th-75th quartiles, with median indicated by thick horizontal black line. Whiskers of boxplots indicate 2.5 and 97.5 percentiles. Violins are smoothed density estimates of the percent error data points. Complete parameter values and number of runs are included in S1 Table.

https://doi.org/10.1371/journal.pcbi.1010677.g002

clinically realistic sizes. For a tumor size of 10^9, all modes of growth have a MPE of less than 4%, so for a clinically realistic cancer size—10^{11} for the CLL dataset—we expect an even better accuracy.

We also perform Monte Carlo simulations for the more complex cases of two nested and two sibling driver subclones (see Methods for derivations of estimators) for the same three modes of cancer growth used for the single driver subclone case above: fast growth (Fig 2C and 2D), no cell death (S1(C) and S1(D) Fig), and slow growth (S2(C) and S2(D) Fig). For two nested driver subclones, the second driver subclone also carries its parental subclone's driver mutation (S4(A) Fig). For two sibling driver subclones, the drivers occur in separate subclones (S4(B) Fig). The growth rate estimates show good agreement with the ground truth values, with MPEs close to 0. The mutation rate estimates also have good accuracy, with the absolute values of their MPEs all ≤4%. As for the single subclone cases already discussed, the time estimates for the nested and sibling subclone simulations have a greater variance. The estimate for t—time between the last driver mutation and diagnosis—shows good accuracy for the fast-growing tumors, but larger errors for the no cell death and slow growth cases. For both the nested and sibling simulations, the estimates for the times of driver mutations 1 and 2 (t_1 and t_2, respectively) have MPEs less than 6%.

Correcting mutation counts observed from genome sequencing data

We note that in our estimate for the time of appearance of the driver, t_1 (see formula (5)), used for comparison to simulated data, we employed a correction to m, the number of mutations that were present in the founder type-1 cell at t_1. From sequencing data, these m mutations are indistinguishable (Fig 3A) from mutations that occurred after t_1 in type-1 cells and reached fixation in the type-1 population [47]. Thus, the value of m observed from sequencing data, m_{obs},

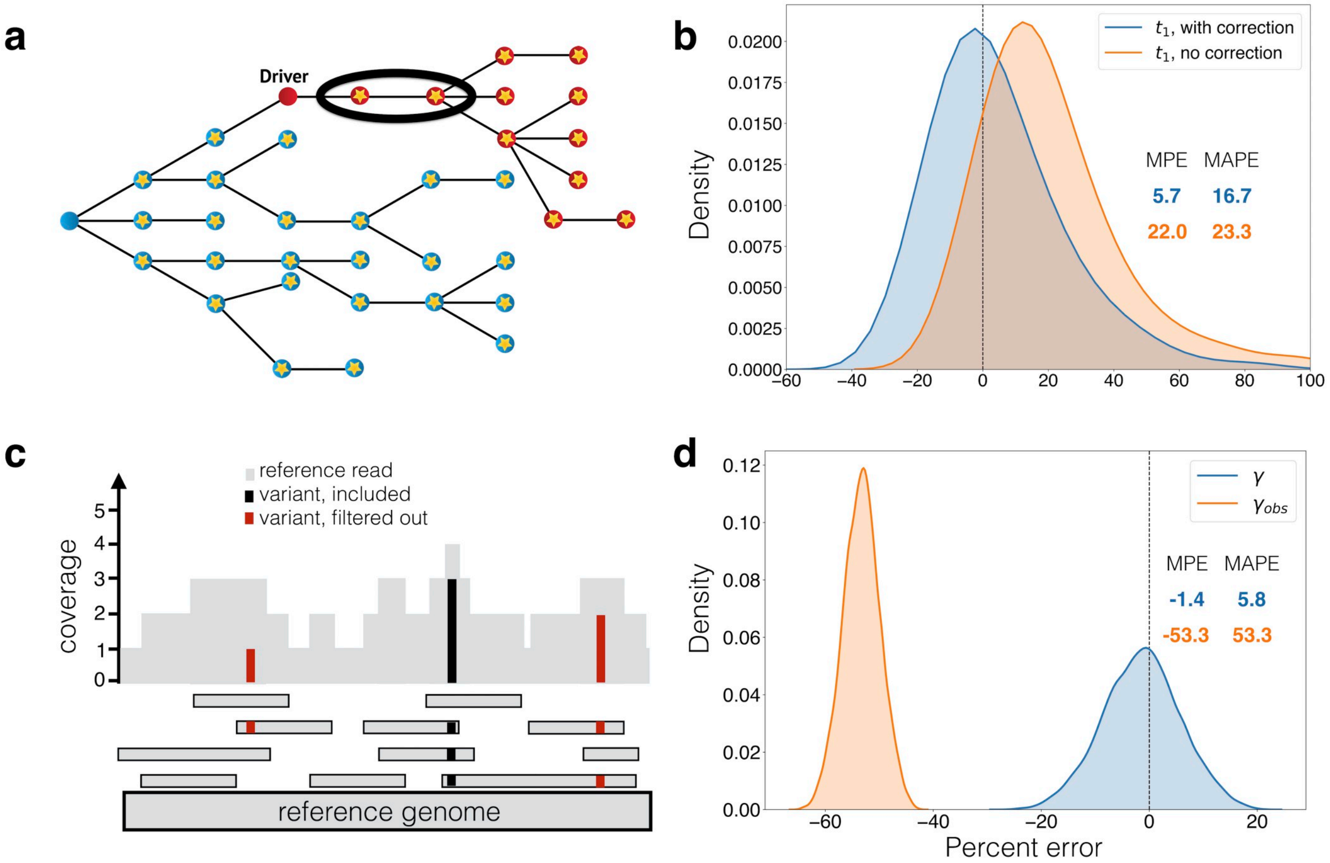

Fig 3. Corrections for observed mutation counts. (a) If passenger mutations (circles with stars) that occur after the driver reach fixation in the driver population (red), then they are indistinguishable from the passengers that were present in the first cell with the driver, which accrued in the type-0 population (blue). The estimate of when the driver occurred needs to account for these mutations (circled). In (b), we compare percent errors of parameter estimates for time from tumor initiation until appearance of a driver subclone, t_1, with and without this correction (Eq (6)). Errors for estimate with correction are shown in blue, and for estimate without correction (Eq. (5)) in orange. Errors are plotted as a kernel density estimate for Monte Carlo simulations of slow-growing tumor with mutation rate $u = 5$. Mean percent errors (MPEs) and mean absolute percent errors (MAPEs) are listed. (c) Mutations present on two or fewer variant reads (red) are filtered out in post-processing. Mutations with more than two variant reads (black) are included. The number of subclonal mutations between frequencies f_1 and f_2, γ, which is used in the mutation rate estimate, must be corrected for mutations that are filtered out. In (d), the percent errors for the observed (orange) and corrected (blue) γ (Eq (7)) are plotted as kernel density estimates. Observed mutations are those that passed post-processing, i.e. those that have more than $L = 2$ mutant reads. True mutation frequencies were generated from 135 surviving runs of a Monte Carlo simulation of a fast-growing tumor with mutation rate $u = 1$, from which sequencing reads were simulated with 200x average coverage (see Materials and methods). Percent errors are calculated relative to the true γ measured from the true mutation frequencies.

https://doi.org/10.1371/journal.pcbi.1010677.g003

will overestimate the true m. In Materials and methods we show that the expected value of the number of passengers that occurred after t_1 and reached fixation in the type-1 population is u/r_1. We subtract this correction factor from m_{obs}:

$$m = m_{obs} - u/r_1. \tag{6}$$

The correction for the m mutations present in the original type-1 cell (6) at time t_1 improves the accuracy of the estimate for time of appearance of driver mutation t_1. For the fast-growing tumor with mutation rate $u = 1$ (S3(A) Fig), the correction lowers the mean percent error (MPE) of the t_1 estimate from 14.0% to 3.8%. For the slow-growing tumor with mutation rate $u = 5$ (Fig 3B), the correction lowers the MPE of the t_1 estimate from 22.0% to 5.7% (Fig 3B).

Another issue arises from obtaining mutation count γ, number of mutations with frequency between f_1 and f_2, from genome sequencing data. When sequencing data is post-processed by filtering out mutations with L or fewer variant reads, low-frequency mutations will be difficult to detect [35] (Fig 3C). For a sample with average sequencing coverage of R and tumor purity p, mutations with mutant allele frequency below $L/(pR)$ will typically not be observable. As a result, since mutations with frequencies between f_1 and f_2 count towards γ, if $f_1 \leq 2L/(pR)$, the observed number of subclonal mutations between frequencies f_1 and f_2, γ_{obs}, will underestimate the true value, γ. For cancers with low mutational burden, such as CLL, we set a relatively low f_1 (1%) to have sufficient resolution to infer mutation rate. Consequently, some mutations with frequency above f_1 will likely be filtered out, and we account for this by correcting for the expected number of such subclonal mutations present at cancer cell frequencies (CCFs) between f_1 and $2L/(pR)$ (see Materials and methods):

$$\gamma = \gamma_{obs}\left(\frac{\frac{1}{f_1} - \frac{1}{f_2}}{\frac{pR}{2L} - \frac{1}{f_2}}\right). \tag{7}$$

Before applying our methodology to patient sequencing data, we estimated the validity of the above correction applied to observed simulated mutation counts. When we simulate sequencing reads from simulated mutation frequencies (see Materials and methods) and post-process by removing mutations with $L = 2$ or fewer variant reads, the adjustment we derived for mutation count γ (7) is critical, even for average sequencing coverage of 200x (Fig 3D). Without any correction, the observed γ has MPE of -53.3% compared to true γ, but with the correction, the computed γ has MPE of -1.4%. When average coverage is 100x, this correction becomes even more important, as many of the low-frequency mutations are discarded (S3(B) Fig). Without any correction, the observed γ has MPE of -79.7%. With the correction the computed γ has MPE of -3.4%. The accuracy of the γ measurement affects our estimate of the mutation rate (4).

Estimating parameters for individual patients with CLL

We use our formulas to infer the patient-specific parameters of cancer evolution for four patients with CLL whose growth patterns and clonal dynamics were analyzed in [27]. These CLLs had peripheral WBC counts measured and whole exome sequencing (WES) performed at least twice before treatment. We consider patients whose WBC counts were classified as having an exponential-like growth pattern, with average $\gamma_{obs} > 2$, and with 3 or fewer macroscopic subclones (i.e. subclones with cancer cell fractions of 20% or greater for at least one pre-treatment time point). Our framework is designed specifically to study naturally evolving

Table 1. Inferred parameters for CLL patients with exponential growth patterns, for which there are at least two longitudinal bulk sequencing measurements before treatment. Estimates are computed from tumor size measurements and mutation frequencies from whole exome sequencing. Mutation rates are for the exome only. The time estimates are in terms of the patient's age in years.

Parameter	Pt. 3	Pt. 6	Pt. 9	Pt. 21
r (/yr)	0.51	0.68	0.28	0.79
r_1 (/yr)	0.85	0.41	-0.40	1.52
r_2 (/yr)		0.46	0.67	
r_3 (/yr)		1.09	0.63	
u (mut/yr)	0.48	0.15	0.36	0.20
MRCA (yr)	14.6	2.8	4.9	6.4
t_1 (yr)	33.5	35.4	18.8	19.6
t_2 (yr)		46.7	21.3	
t_3 (yr)		45.9	24.8	
age at diagnosis (yr)	63	58	54	35
age at leukocytosis (yr)	61.9	65.7	51.8	34.4

https://doi.org/10.1371/journal.pcbi.1010677.t001

cancer dynamics, unperturbed by treatment, which will drastically alter the cancer's dynamics and size. For calculation of the γ_{obs} mutations between frequencies f_1 and f_2, we set f_1 = 1% due to the difficulty of detecting low frequency variants <1% [48, 49]. We set f_2 to 20% to minimize overlap with potential driver mutations of the macroscopic subclones. The average γ_{obs} for the four analyzed patients ranges from 2.5 to 19.3, with a median of 5.2. As in Ref. [27], we perform subclonal reconstruction for each patient using PhylogicNDT [43]. To obtain confidence intervals for our parameter estimates, we utilize a sampling procedure to account for model and measurement uncertainties, including uncertainties in subclone frequencies, fitted growth curves, and the Poisson process for mutation accumulation (see Materials and methods). For each patient's tumor, we compute estimates of the growth rate of each clone, exome mutation rate, the times that each subclone arose, and how long each subclone expanded before the tumor was detected (Tables 1 and 2). We also estimate what time the cancer was clinically detectable, by sampling from the distribution of fitted growth parameters and solving the resulting root-finding problem for time to reach detectable size under our growth model (see Materials and methods). For CLL specifically, we compute time of leukocytosis—an abnormally high WBC count. We reconstruct these histories for tumors with various clonal structures.

Table 2. Confidence intervals for inferred parameters for CLL patients with exponential growth patterns, for which there are at least two longitudinal bulk sequencing measurements before treatment. Estimates are computed from tumor size measurements and mutation frequencies from whole exome sequencing. Mutation rates are for the exome only. The time estimates are in terms of the patient's age in years.

Parameter	Pt. 3	Pt. 6	Pt. 9	Pt. 21
r (/yr)	[0.20, 0.85]	[0.15, 1.30]	[0.17, 0.42]	[0.30, 1.14]
r_1 (/yr)	[0.65, 1.04]	[0.08, 0.73]	[-0.45, -0.19]	[1.01, 2.04]
r_2 (/yr)		[0.08, 0.85]	[0.49, 0.94]	
r_3 (/yr)		[0.65, 1.78]	[0.39, 0.86]	
u (mut/yr)	[0.39, 0.59]	[0.12, 0.19]	[0.35, 0.37]	[0.19, 0.23]
MRCA (yr)	[1.4, 26.8]	[0.1, 13.2]	[1.2, 10.8]	[0.3, 16.7]
t_1 (yr)	[24.1, 39.2]	[21.7, 46.1]	[8.8, 35.1]	[10.8, 24.0]
t_2 (yr)		[25.6, 57.5]	[7.7, 31.7]	
t_3 (yr)		[31.3, 54.6]	[10.3, 37.6]	
age at leukocytosis (yr)	[60.3, 62.4]	[64.2, 67.1]	[51.6,51.9]	[32.8,34.6]

https://doi.org/10.1371/journal.pcbi.1010677.t002

Patients 3 and 21 are examples of a CLL with a single subclone (Fig 4). For Patient 3, Clone 0, the most recent common ancestor (MRCA) of this patient's CLL, was initiated when the patient was 14.6 [1.4, 26.8] years old (median and [95% confidence interval] of estimate). Clone 0 grew with a net growth rate of 0.51 [0.20, 0.85] per year. Approximately two decades later, Clone 1 was initiated when the patient was 33.5 [24.1, 39.2] years old. Clone 1 expanded with a growth rate of 0.85 [0.65, 1.04] per year (corresponding to a selective growth advantage of 68.7% over Clone 0), and the patient was diagnosed approximately three decades later at age 63.

For patient 21, we estimate that the parental clone (MRCA, Clone 0) of this patient's CLL was initiated when the patient was 6.4 [0.3, 16.7] years old, and grew with a net growth rate of 0.79 [0.30, 1.14] per year. Clone 1 appeared when the patient was 19.6 [10.8, 24.0] years old, and grew more quickly than Clone 0, with a selective growth advantage of ~90% over Clone 0). Clone 1 contained a FGFR1 mutation, which might have been acting as a driver of the increased net proliferation. Clone 1 then grew for ~15 years before the patient was diagnosed at age 35.

Patients 6 and 9 present more complex clonal structures (Fig 4). Clone 0, the parental clone of the CLL of Patient 9, arose when the patient was 4.9 [1.2, 10.8] years old, and had a growth rate of 0.28 [0.17, 0.42] per year. Clone 1 arose when the patient was 18.8 [8.8, 35.1] years old. Interestingly, during clinical observation between diagnosis and treatment, Clone 1 was declining in size, with a growth rate of -0.40 [-0.45, -0.19] per year. In line with recent findings [50], we found that sometimes the estimated growth rate during the period of observation, such as the negative growth rate of Clone 1, is smaller than the minimal possible growth rate necessary to reach the observed clone size. In that case, for calculating mutation rate, time of the driver(s), time of detectability, and time between driver(s) and diagnosis we use the minimal growth rate. Clone 2, containing a KRAS mutation, had the largest net growth rate of the three clones (0.67 [0.49, 0.94] per year), corresponding to a selective growth advantage of 140.9% over the parental clone. Clone 2 arose when the patient was 21.3 [7.7, 31.7] years old.

We estimate that the CLL of Patient 6 was initiated when the patient was 2.8 [0.1, 13.2] years old. The leukemic parental clone, Clone 0, then grew at a rate of 0.68 [0.15, 1.30] per year. Approximately 33 years after the appearance of Clone 0, when the patient was 35.4 [21.7, 46.1] years old, the first subclone, Clone 1 appeared. Clone 3 arose from within Clone 1 when the patient was 45.9 [31.3, 54.6] years old. Clone 3 harbored a driver mutation in ASXL1 and had selective growth advantage of 60.8% over Clone 0. The patient was diagnosed at age 58, eventually needing treatment 12.0 years after diagnosis.

The average mutation rate in the four CLL patients we analyze is 0.30 mutations/year. This rate is over the exome, which accounts for ~1% of the human genome. Our average estimated mutation rate in CLL exomes is similar to the measured rate of accumulation of mutations in human tissues of 40 mutations per year over the entire genome [51]. Other recent work has estimated a mutation rate of 17 mutations per year in human haematopoietic stem cell/multipotent progenitors [52]. Our estimated mutation rates during CLL progression are on par or higher than the recent estimates in healthy hematopoietic cells [52], in line with the expectation that mutation rates may be increased in cancer. The estimated times of appearance of CLL subclones are very long, on the order of 10 years or more. This finding is in agreement with results from Gruber et al. [27], who find few new CLL subclones over years to a decade of evolution. We observe that CLL initiation occurred early in most patients, within the first fifteen years of their lives, consistent with recent work in other cancer types [19, 36]. We find that CLL patients reach leukocytosis an average of 1.5 years before the first timepoint at which cancer genome sequencing was performed. For three of the patients, our estimated time of leukocytosis was before diagnosis, on average 1.3 years prior to diagnosis.

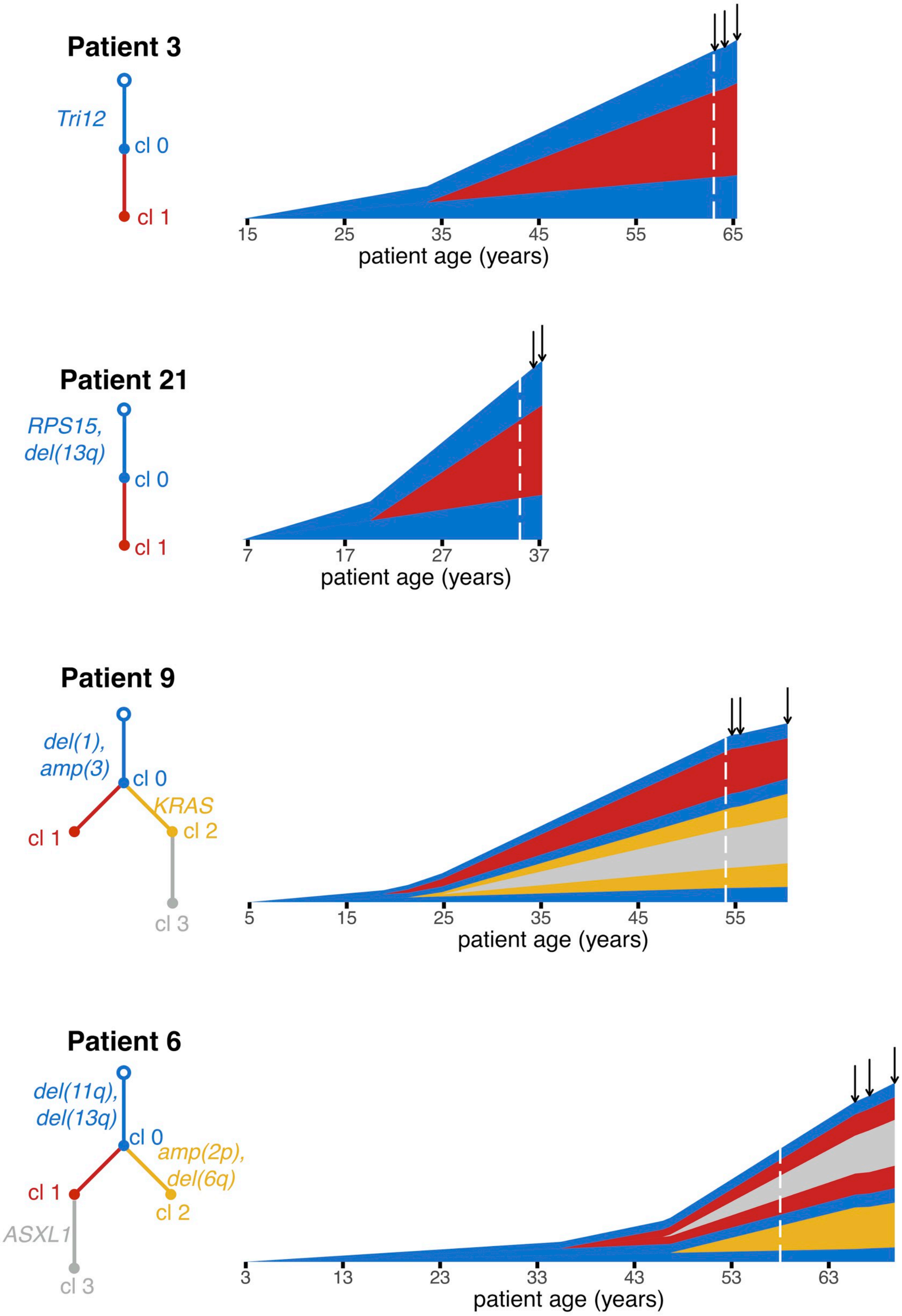

Fig 4. Reconstructing the timeline of CLL evolution in patients. We applied our methodology to estimate subclonal growth rates, mutation rates and evolutionary timelines in CLL tumors from Ref. [27]. Vertical height of a clone represents its log$_{10}$-scaled size. Mutations were clustered into clones and phylogenetic trees were inferred using PhylogicNDT [43]. Tree edges are colored by clone number and are labeled with driver mutations, if any. For each patient, we show estimates for patient age at CLL initiation and times of appearance of CLL subclones. Dashed white line indicates when the patient was diagnosed. Solid black arrows indicate times of bulk sequencing measurements.

https://doi.org/10.1371/journal.pcbi.1010677.g004

Discussion

We use a stochastic branching process model to reconstruct the timing of driver events and quantify the evolutionary dynamics of different subclonal populations of cancer cells. We estimate growth rates of tumor subclones, selective growth advantage of individual driver mutations, mutation rate in the tumor, time between tumor initiation and appearance of a subclonal driver mutation, and time between driver mutation and tumor observation. Together, this allows us to estimate the age of the patient at tumor initiation, as well as the age at appearance of a subclonal driver.

Previous work has computed relative order of driver events [18, 21, 53], while other studies have given estimates for scaled mutation rates and time of events [24, 32]. However, we present estimates for absolute, unscaled mutation rates and times, which are easily interpretable and don't implicitly depend on unknown parameters. We assume that mutations accrue with time, which simplifies derivations and is supported by recent experimental data that shows that non-dividing cells may accrue mutations at a similar rate as dividing cells [54]. Other potential assumptions regarding mutation accumulation include mutations occurring at cell division [55] or assuming mutation rate is proportional to the copy number state [56]. For example, recent work reported that some mutational signatures in human cancers are generated during mitosis [55]. Other work has shown that the rate of accumulation of somatic single nucleotide variants is proportional to copy number [56]. We further assume that all cancer subpopulations have the same passenger mutation rate. In the case that mutations occur predominantly at cell division, assuming that the rate of cell division is comparable across all tumor subclones, our estimates would still be valid. In the case of a subclone that has an elevated mutation rate (e.g. due to a chromosomal amplification, mutation in a DNA repair pathway gene or an increased cell division rate), we would underestimate the mutation rate and overestimate the time of driver mutation(s) in that subclone. In the other subclones, the opposite would be true.

For individual CLLs that underwent bulk sequencing at two time points [27], we infer growth rates of individual subclones, mutation rate in the tumor, the times when cancer subclones began growing, the time between driver mutations and the patient's diagnosis, and time when the cancer is clinically observable. Our inferences are limited by the relatively low number of mutations present in CLL, as well as sequencing coverage [27], so we set a minimum passenger mutation count when selecting specific cases to analyze. The accuracy of estimates presented here is expected to be higher with whole genome sequencing available, with higher sequencing coverage, or in cancer types with more mutations, with some important limitations. Exponential growth—the mean behavior of our branching process model—has been well documented in vivo [27, 57–59], but tumors can also often exhibit sigmoidal growth (e.g. logistic, Gompertz models), where initial exponential growth is followed by a deceleration in growth [58, 60–63]. Our estimators should only be used for cancers exhibiting exponential growth; for other modes of growth, such as the logistic-growing class of CLL patients in Ref. [27], the parameter estimates would have to be derived specifically for the particular mode of growth observed. Exponential growth is the simplest common cancer growth pattern, and yet, estimating the exponential growth rates requires at least two longitudinal timepoints. To fit all parameters for patients with more complex growth dynamics, additional longitudinal samples will be needed; this type of analysis would be further limited due to the scarcity of longitudinal pre-treatment samples in many cancer types. In the case of solid tumors, the number of cells can be estimated from measurements of tumor volume [64], however multiple biopsies would potentially be needed to fully account for the existing genetic heterogeneity. Furthermore, a solid tumor's spatial structure, mode of evolution, and biopsy collection influence how well selection and mutation spectra can be observed [30, 31, 65]. Recent modeling and computational work, in combination with careful multi-region sequencing and single cell sequencing, have begun to disentangle these confounding factors [26, 29, 30].

Our model and derivations assume a fixed mutation rate u after transformation and fixed growth rates of cancer subclones, similar to previous approaches [24, 30, 35]. Some individual cancer subclones (such as Clone 1 from Pt. 9) not only do not grow exponentially, they actually decline in absolute cell numbers, even if the overall tumor is undergoing expansion. This phenomenon has been previously observed [27, 66], and could be caused by the declining subclone getting outcompeted by more fit subclones. Sudden genomic instability events, or a change in cancer mutation and/or growth rate over time could also introduce errors into our parameter inferences. Recent sequencing data points to mutational processes that change over time during cancer evolution [20, 67]; incorporating possible changes in the mutation and/or growth rate into the model would require much higher density of sequencing and clinical data [37], as would employing a more complex growth model (e.g. boundary-driven or sigmoidal growth).

Materials and methods

Branching process model of tumor evolution

We employ a continuous, multi-type branching process model of cancer evolution. For the case of a single driver subclone, there are two cell types, type-0 and type-1. Tumor expansion is initiated by a single type-0, or initiated tumor cell. Type-0 cells divide with rate b and die with rate d, yielding a net growth rate of $r = b - d$. At time t_1, a single driver mutation is introduced into a randomly selected cell in the type-0 population, founding a new type-1 population of cells. This type-1 population undergoes its own independent branching process. They divide with rate b_1, die with rate d_1, and have net growth rate $r_1 = b_1 - d_1$. If the driver mutation gives type-1 cells a selective growth advantage over the type-0 population, then $r_1 > r$. With the ratios of the growth rates denoted as $s = r_1/r$, the growth advantage can be quantified as $g = (s - 1) \cdot 100\%$. In the case of neutral evolution, $g = 0$. If there is a selective advantage, $g > 0$. Neutral mutations, or passengers, have no effect on the cell's fitness, and accrue according to a Poisson process with rate u. We assume an infinite alleles model such that there is no back mutation and an infinite sites model such that every new passenger mutation is unique. Only surviving populations are considered. All derivations below will condition on survival. The type-0 and type-1 populations at time t will be denoted as $X_0(t)$ and $X_1(t)$, respectively.

Measurements sufficient to determine evolutionary history

Here we derive estimates for parameters describing the carcinogenic process for a single driver subclone, using measurements taken from two time points late in the tumor's development. We require sequencing of the tumor at the two time points, when the tumor is first observed at the unknown time $t_1 + t$ and a specified Δ later, at $t_1 + t + \Delta$. From these two bulk sequencing measurements, we obtain measurements of α_1 and α_2, the fraction of cells carrying the driver mutation at $t_1 + t$ and $t_1 + t + \Delta$, respectively. In addition, from the bulk sequencing at $t_1 + t + \Delta$, we obtain measurements of m, the number of mutations present in the founder type-1 cell, as well as γ, the number of mutations with frequency between the specified f_1 and f_2. The total population size at these times, M_1 and M_2, is also measured.

Expected value of γ, number subclonal mutations

For a population consisting of a single clone with birth and death rates b and d, the expected number of subclonal mutations present at a frequency larger than f is shown to be [47]

$$\frac{\bar{u}(1-f)}{(1-\delta)f} \tag{8}$$

where $\delta = d/b$ and $\bar{u}$ is the probability that a daughter cell gains a new passenger mutation at cell division. In this paper, we allow mutations to occur at any point in time and consider the absolute mutation rate per cell, u, which is equal to $\bar{u}b$. Then the expected number of subclonal mutations between f_1 and f_2, $\mathbb{E}\gamma$, is

$$\mathbb{E}\gamma = \frac{u(1-f_1)}{b(1-\delta)f_1} - \frac{u(1-f_2)}{b(1-\delta)f_2} \tag{9}$$

$$= \frac{u}{r}(1/f_1 - 1/f_2) \tag{10}$$

where $r = b - d > 0$.

Now we derive $\mathbb{E}\gamma$ in the case of clones 0 through k, each clone with growth rate $r_i > 0$ and fraction α_i^c. Each clone i has $\alpha_i^c \frac{u}{r_i}(1/f_1 - 1/f_2)$ expected subclonal passengers between frequencies f_1 and f_2. Thus, the total expected number of passengers with frequencies between f_1 and f_2 is

$$\mathbb{E}\gamma = (1/f_1 - 1/f_2)\sum_{i=0}^{k}\frac{u\alpha_i^c}{r_i}. \tag{11}$$

For the simplest case we consider, a tumor with a single driver mutation occurring in the initiated tumor population, there is a type-0 population with growth rate r and a type-1 population with growth rate r_1. Eq (11) reduces to

$$\mathbb{E}\gamma = \left(\frac{u\alpha}{r_1} + \frac{u(1-\alpha)}{r}\right)\left(\frac{1}{f_1} - \frac{1}{f_2}\right) \tag{12}$$

where α is the fraction of cells having the driver mutation.

Derivation of estimates of evolutionary parameters for single driver subclone

With the cancer bulk sequenced at the two time points $t_1 + t$ and $t_1 + t + \Delta$, we are able to derive estimates for t_1, t, r, r_1, and u. First we solve for r and r_1, based on the estimated cell counts at $t_1 + t$ and $t_1 + t + \Delta$. The observed type-i cell count is equated to the expected value of the type-i population size, conditioned on survival. For a birth-death process started with a single type-i cell at time 0, we have $\mathbb{E}[X_i(t)] = e^{r_i t}$. That process has extinction probability d_i/b_i [38]. Then,

$$\mathbb{E}[X_i(t)] = \mathbb{E}[\mathbb{E}[X_i(t)|I_{X_i(t)>0}]] \tag{13}$$

$$\approx \mathbb{E}[X_i(t)|X_i(t) = 0](d_i/b_i) + \mathbb{E}[X_i(t)|X_i(t) > 0](1 - d_i/b_i) \tag{14}$$

$$= \mathbb{E}[X_i(t)|X_i(t) > 0](1 - d_i/b_i) \tag{15}$$

where $I_{X_i(t)>0}$ is a random variable and indicator function defined as

$$I_{X_i(t)>0} = \begin{cases} 0 & \text{if } X_i(t) = 0 \\ 1 & \text{if } X_i(t) > 0 \end{cases}.$$

Thus, from (15), for large enough time t,

$$\mathbb{E}[X_i(t)|X_i(t) > 0] \approx \frac{1}{1 - d_i/b_i} e^{r_i t} = \frac{b_i}{r_i} e^{r_i t}. \tag{16}$$

It then follows that for the type-0 population,

$$\mathbb{E}[X_0(t_1 + t)|X_0(t_1 + t) > 0] = \frac{b}{r} e^{r(t_1+t)} = (1 - \alpha_1)M_1 \tag{17}$$

$$\mathbb{E}[X_0(t_1 + t + \Delta)|X_0(t_1 + t + \Delta) > 0] = \frac{b}{r} e^{r(t_1+t+\Delta)} = (1 - \alpha_2)M_2. \tag{18}$$

Proceeding similarly for the type-1 population, we obtain

$$r_1 = \frac{1}{\Delta} \log\left(\frac{\alpha_2 M_2}{\alpha_1 M_1}\right) \tag{19}$$

$$r = \frac{1}{\Delta} \log\left(\frac{(1 - \alpha_2)M_2}{(1 - \alpha_1)M_1}\right). \tag{20}$$

The expected value of the first time a population of type-1 cells in a branching process reaches the observed size $\alpha_1 M_1$ is [38]

$$\mathbb{E}t = \frac{1}{r_1} \log\left(\frac{\alpha_1 M_1 r_1}{b_1}\right) - \frac{1}{r_1}\int_0^\infty e^{-z} \log z dz \tag{21}$$

$$= \frac{1}{r_1} \log\left(\frac{\alpha_1 M_1 r_1}{b_1}\right) + \frac{0.5772}{r_1} \tag{22}$$

$$= \frac{1}{r_1}(\log(\alpha_1 M_1) + \log(r_1/b_1) + 0.5772) \tag{23}$$

$$\approx \frac{1}{r_1} \log(\alpha_1 M_1). \tag{24}$$

The last approximation is justified because for realistic cell counts, the first term in (23) dominates the other two, which is also evident in simulation studies (S5 Fig). For example, if $r_1 = \frac{1}{2} b_1$, then the second term $\log(r_1/b_1) = -0.69$, compared to the first term $\log(\alpha_1 M_1) = 19.11$. Even if r_1 is as low as $0.1 b_1$, the second term is -2.30. In this case, the percent error of the approximation (24) is 7.3%. In general, the accuracy increases with increased tumor size.

With the measurement of γ, the number of subclonal passengers with frequency between f_1 and f_2, we can estimate the mutation rate u. In the previous section we derive the expected value of γ as

$$\mathbb{E}\gamma = \left(\frac{u\alpha}{r_1} + \frac{u(1 - \alpha)}{r}\right)\left(\frac{1}{f_1} - \frac{1}{f_2}\right). \tag{25}$$

Using the estimates of r and r_1 from (19) and (20), and the measured value of γ from the second bulk sequencing, Eq (25) can be solved for the mutation rate u,

$$u = \frac{f_1 f_2 r r_1 \gamma}{(f_2 - f_1)(\alpha_2 r + r_1(1 - \alpha_2))}. \tag{26}$$

When estimating mutation rate for the CLL patients from Ref. [27], for which there is bulk sequencing at two or more time points, we average the mutation rate calculated at each of these time points. (26) is applied for each time point with the respective CCFs and observed γ values for each time point.

To derive the maximum likelihood estimates of t_1, we consider the likelihood function $P(m|t_1)$. The number of passenger mutations present in the founder type-1 cell that appeared at time t_1 is a Poisson process with rate u. Thus,

$$P(m|t_1) \propto \frac{(ut_1)^m e^{-ut_1}}{m!}. \tag{27}$$

Maximizing the logarithm of the likelihood function with respect to t_1 yields a MLE for t_1 in terms of estimated or measured quantities:

$$t_1 = m/u. \tag{28}$$

Estimating number of unobserved subclonal mutations from sequencing data

When sequencing data is post-processed by filtering out any mutations with L or fewer variant reads, the number of mutations between f_1 and f_2 will likely be underestimated if $2L/(Rp) > f_1$, where R is average sequencing coverage and p is tumor purity. Define γ_{obs} as the observed number of mutations between frequencies f_1 and f_2, after post-processing has been performed that filtered out any mutations with L or fewer variant reads. The expected number of subclonal mutations between frequencies f_1 and x is given by

$$\gamma(x) = c(1/f_1 - 1/x) \tag{29}$$

where c is a constant that will vary depending on the patient and sample. It can be fit on the sequencing data by noting

$$\gamma_{obs} = \gamma(f_2) - \gamma(2L/(Rp)) \tag{30}$$

$$= c(Rp/(2L) - 1/f_2). \tag{31}$$

Therefore, c can be estimated from the sequencing data as

$$c = \frac{\gamma_{obs}}{Rp/(2L) - 1/f_2}. \tag{32}$$

Then, we can estimate γ as

$$\gamma = \gamma_{obs}\left(\frac{\frac{1}{f_1} - \frac{1}{f_2}}{\frac{Rp}{2L} - \frac{1}{f_2}}\right). \tag{33}$$

Number of passengers reaching fixation after t_1

We estimate the number of passengers that occurred after t_1 and reached fixation in the type-1 population in order to adjust the m_{obs} mutation count. From [47], when mutations occur at cell division, the expected number of clonal passengers is $\delta\bar{u}/(1-\delta)$. $\bar{u}$ is the probability that a daughter cell gains a new passenger mutation at cell division, so the mutation rate is $u = \bar{u}b_1$. For the type-1 population, $\delta = d_1/b_1 < 1$. When mutations accrue over time, and not only at divisions, the expected number of clonal passengers is thus

$$\bar{u}/(1-\delta) = u/r_1. \tag{34}$$

Similarly, for a clone i, the expected number of passengers that occur after time t_i and reach fixation is

$$u/r_i \tag{35}$$

where $r_i = b_i - d_i > 0$.

Simulation of tumor evolution and sequencing data

To assess the accuracy of the analytic results, we perform a continuous time Monte Carlo simulation to model tumor evolution and collection of sequencing data with an implementation of the Gillespie algorithm [68]. Simulations are written in C/C++.

The type-j population has division rate b_j, death rate d_j, and mutation rate u. Mutations can occur at any point of the cell cycle, not just during division. z_n is the number of type-j cells with passenger n as their most recent passenger mutation. The type-0 population is initiated with a single cell at time 0, and the type-j population for $k \geq j > 0$ is initiated with a single cell at time t_j. Let a be the vector recording the ancestor of new mutations. Element a_i is the subclonal ancestor of the ith passenger mutation. For each $j \in 0, 1, \ldots, k$, repeat 1–4 while time is less than $t_k + t + \Delta$.

1. Set $\Gamma = N_j(b_j + d_j + u)$. Time increment to next event time is randomly sampled from Exp $[\Gamma]$.
   - If $j < k$, if time is greater than or equal to t_{j+1} for first time, randomly select type-j subclone i to have driver mutation, remove one cell from type-j population count, and set $N_{j+1} = 1$. Record the true value of m_{j+1}, the number of passenger mutations present in the founder type-$(j + 1)$ cell.
2. Randomly select cell, with most recent passenger mutation i, to have the event.
3. Determine which type of event and update population and mutation frequencies. Sample Y from Uniform$[0, \Gamma]$ to determine event type:
   1. $y \in (0, b_j) \to$ birth. $N_j \mathrel{+}= 1, z_i \mathrel{+}= 1$.
   2. $y \in (b_j, b_j + d_j) \to$ death. $N_j \mathrel{-}= 1, z_i \mathrel{-}= 1$.
   3. $y \in (b_j + d_j, b_j + d_j + u) \to$ passenger mutation. Suppose it's the pth passenger, $z_i \mathrel{-}= 1$, $z_p = 1$. Update ancestor: $a_p = i$.
4. For $j = 0$, if time is less than t_1 and population goes extinct, restart simulation. For $j \geq 1$, if time is greater than t_j and population goes extinct, restart type-j simulation at t_j with a single cell.

5. Reindex to remove extinct passenger mutations, and traverse back through ancestor vector a to sum total number of cells with each passenger.

Measurements are taken at bulk sequencing times $t_k + t$ and $t_k + t + \Delta$. If time is greater than or equal to $t_k + t$, we measure $M_1 = \sum_{j=0}^{k} N_j$ and CCF of clone j as N_j/M_1. Then an additional bulk sequencing measurement is taken at the final time $t_k + t + \Delta$, where we measure $M_2 = \sum_{j=0}^{k} N_j$ and the CCF of clone j as N_j/M_2. At $t_k + t + \Delta$, we measure γ, the number of mutations with frequency between f_1 and f_2.

To measure $m_{j,obs}$, the observed number of passengers in the founder type-j cell, we count the number of passengers present in all type-j cells. We also save the true value of m_j.

For when we calculate a percent error of corrected and observed γ values in Fig 3D and S3 (B) Fig, we simulate sequencing data by sampling from the mutation frequencies obtained in the Monte Carlo simulation, outlined above, using the approach of [35]. Define average sequencing coverage as R, number of cells at time of sequencing as M, Z_i as the number of cells with mutation i, R_i as read coverage, and χ_i as the true mutation frequency from Monte Carlo simulation. For each saved Monte Carlo simulation run, repeat the following 100 times:

1. Generate read coverage: $R_i \sim \text{Binomial}[M, R/M]$.
2. Generate number of cells carrying mutation i: $Z_i \sim \text{Binomial}[R_i, \chi_i/2]$.
3. Post-processing. If there are $L = 2$ or fewer variant reads, discard mutation.
4. Measure γ_{obs}, the observed number of subclonal mutations between frequencies f_1 and f_2: $\gamma_{obs} = \Sigma_i I(f_1 \leq 2Z_i/R \leq f_2, Z_i > L)$.
5. Calculate the truth, γ_{true}, from the true mutation frequencies: $\gamma_{true} = \Sigma_i I(f_1 \leq \chi_i \leq f_2)$.

Parameter values for simulations

For the simulations we consider three parameter sets corresponding to three modes of tumor evolution: a fast-growing tumor, slow-growing tumor, and tumor with no cell death, each with multiple mutation rates. We simulate three clonal structures: single driver subclone, two nested driver subclones, and two sibling driver subclones. All parameter values are listed in S1 Table. Mutation rate parameter values lie within observed genome wide point mutation rates per day [69]. For simulation of parental clone and subclone, the fast-growing tumor dynamics are from [34]. The slower growing tumor parameter regime has a reduced net growth of $r = 0.025$, compared to the fast-growing tumor's net growth rate of $r = 0.07$.

Subclonal reconstruction of CLL sequencing data

The sequencing data from all CLLs analyzed is from Ref. [27], Supplementary Tables 2–4. As in that publication, we use PhylogicNDT [43] to perform subclonal reconstruction. We run the Cluster and BuildTree modules of PhylogicNDT on the longitudinal mutation data from Supplementary Table 3 of [27], using mutation alternate/reference counts, copy number, and tumor purity at all pre-treatment time points. Then for each patient, PhylogicNDT outputs a clonal reconstruction, which includes a phylogenetic tree of the subclones and posterior distributions of subclone CCFs. Additionally, it clusters mutations and assigns them to clones. We directly use subclone assignments and posteriors generated from PhylogicNDT. In our analysis we focus on estimating timing and growth rates of macroscopic subclones whose CCFs are greater than 20% for at least one pre-treatment time point.

Accounting for uncertainties in subclone frequencies and growth rates

Our estimates for parameters of cancer evolution require as input the information on the number of subclonal populations in the tumor, their CCFs and their phylogenetic relationships. In order to obtain this information, we use PhylogicNDT [43], which performs subclonal reconstruction of longitudinal cancer sequencing data. The uncertainty in subclone CCFs reported by PhylogicNDT affects our estimates for subclone growth rates, which in turn affect the estimates of mutation rate and time t between driver(s) and diagnosis. We account for this uncertainty by drawing from the CCF posterior distributions that are output by PhylogicNDT. Using these sampled CCF values, we then calculate growth rates, mutation rate u, and time t between driver(s) and diagnosis, thereby generating confidence intervals for these parameters due to CCF uncertainty.

To estimate subclonal growth rates, we fit an exponential growth curve to subclonal sizes measured at two or more time points. This regression yields fitted values for each clone's growth rate and age. To account for uncertainty in the curve fit (in the case of more than two longitudinal samples), we sample the growth rates and age of clone from a bivariate normal distribution with mean equal to the fitted parameters and variance equal to the covariance matrix of the fitted parameters. When the estimated growth rate during the period of observation—including negative growth rates—is smaller than the minimal possible growth rate necessary to reach the observed clone size, we use the minimal growth rate for calculating mutation rate, time of the driver(s), time between driver(s) and diagnosis, and time of detectability.

Estimating time of cancer detectability

The time a cancer is detectable is the time at which the cancer exceeds the minimum observable size. For the CLL data, we estimate the time that the patients first exhibited an abnormally high WBC count, or leukocytosis, characterized by a WBC count of 11,500/μL [70], or approximately 5.75 x 10^{10} total WBCs, assuming a total blood volume of 5 L. In the previous section, we describe how we fit the growth dynamics for the CLL data and obtain a distribution of the fitted growth parameters. Here, we sample from the distribution of the fitted parameters 10,000 times (using the minimal growth rate in the case of a growth rate too low to give rise to the observed WBC count), and numerically solve for the time at which the total WBC count was equal to 5.75 x 10^{10}. i.e., we numerically find the root with respect to t_i of

$$f(\hat{\theta}_i, t_i) - 5.75 \text{ x } 10^{10} = 0 \tag{36}$$

where t_i is the ith estimated time out of 10,000 estimates, $f(\cdot)$ is the exponential function describing the mean cancer growth, and $\hat{\theta}_i$ is the ith random sample from the fitted growth parameters (intercept and growth rate).

Accounting for model uncertainty

The largest source of model uncertainty is the Poisson process for how mutations accumulate, which is used to estimate the time t_1 of the driver mutation. In the fast-growing tumor simulation experiments, the time t_1 had the largest error and variation (Fig 2). The estimate for t_1 depends on the m mutations present in all cells in the driver subclone. The observed m is a single random sample from a Poisson distribution. To account for the uncertainty in t_1 arising from m in the CLLs analyzed, we sample t_1 from the posterior distributions $P(t_1|m)$. This source of model uncertainty due to the Poisson process will be most significant for cancers like CLL with a smaller number of mutations.

The time t between driver mutation and diagnosis is a random variable due to the stochasticity of cancer cell growth, and will naturally have a certain amount of variation. Time between driver event and diagnosis in a branching process follows a Gumbel distribution [38] and will have a constant variance. The mean, however, will increase with the logarithm of the cancer cell counts, which for the CLLs analyzed are $\sim 10^{11}$. The simulations of cancer evolution grow to smaller tumor sizes ($\sim 10^5$) and, as a result, the estimate for t has a significant amount of uncertainty (Fig 2). However, for time scales necessary to generate a tumor, the estimate for t will be quite accurate. For commonly observed tumor sizes, the stochastic fluctuations in the time for the cancer to reach that size will be smaller relative to the magnitude of the time. For a cancer with cell count $\sim 10^{11}$, the standard deviation of the time t will be less than 5% of its expected value.

Tumor with two nested driver subclones

Here we consider the case where there are two nested driver subclones (S4(A) Fig). "Nested" means that all cells carrying the second driver mutation also carry the first. Type-0, or initiated tumor, cells have birth rate b_0, death rate d_0, and net growth rate $r_0 = b_0 - d_0$. Type-1 cells, which only have the first driver, have birth rate b_1, death rate d_1, and net growth rate $r_1 = b_1 - d_1$. Type-2 cells, which carry both drivers, have birth rate b_2, death rate d_2, and net growth rate $r_2 = b_2 - d_2$. The first driver occurred in a type-0 cell at time t_1. The second driver occurred in a type-1 cell at $t_2 = t_1 + t_2'$. The mutation rate u is the same for all subclones.

At times $t_1 + t_2' + t$ and $t_1 + t_2' + t + \Delta$, the tumor is bulk sequenced. The bulk sequencing allows the measurement of the fraction of cells with driver 1 at time $t_1 + t_2' + t$, α_1; the fraction of cells with driver 2 at $t_1 + t_2' + t$, α_2; fraction of cells with driver 1 at time $t_1 + t_2' + t + \Delta$, β_1; the fraction of cells with driver 2 at $t_1 + t_2' + t + \Delta$, β_2; and the observed number of subclonal passenger mutations between frequencies f_1 and f_2, γ_{obs}. Note that the fraction of the population that is a type-1 cell at the two times is $\alpha_1 - \alpha_2$ and $\beta_1 - \beta_2$. The fraction of type-0 cells at the two bulk sequencing time points are $1 - \alpha_1$ and $1 - \beta_1$. The total number of cells at bulk sequencing time points are M_1 and M_2. We then equate the estimated cell counts to the expected value of the type-i population size X_i, conditioned on survival.

$$\mathbb{E}[X_i(t_1 + t_2' + t) | X_i(t_1 + t_2' + t) > 0] = \begin{cases} \frac{b_0}{r_0} e^{r_0(t_1+t_2'+t)} & i = 0 \\ \frac{b_1}{r_1} e^{r_1(t_2'+t)} & i = 1 \\ \frac{b_2}{r_2} e^{r_2 t} & i = 2 \end{cases} \tag{37}$$

$$= \begin{cases} (1 - \alpha_1) M_1 & i = 0 \\ (\alpha_1 - \alpha_2) M_1 & i = 1 \\ \alpha_2 M_1 & i = 2 \end{cases} \tag{38}$$

$$\mathbb{E}[X_i(t_1 + t_2' + t + \Delta) | X_i(t_1 + t_2' + t + \Delta) > 0] = \begin{cases} \frac{b_0}{r_0} e^{r_0(t_1+t_2'+t+\Delta)} & i = 0 \\ \frac{b_1}{r_1} e^{r_1(t_2'+t+\Delta)} & i = 1 \\ \frac{b_2}{r_2} e^{r_2(t+\Delta)} & i = 2 \end{cases} \tag{39}$$

$$= \begin{cases} (1-\beta_1)M_2 & i=0 \\ (\beta_1-\beta_2)M_2 & i=1 \\ \beta_2 M_2 & i=2 \end{cases} \tag{40}$$

Solving the above equations for r_i, we obtain the growth rate estimates:

$$r_0 = \frac{1}{\Delta}\log\left(\frac{(1-\beta_1)M_2}{(1-\alpha_1)M_1}\right) \tag{41}$$

$$r_1 = \frac{1}{\Delta}\log\left(\frac{(\beta_1-\beta_2)M_2}{(\alpha_1-\alpha_2)M_1}\right) \tag{42}$$

$$r_2 = \frac{1}{\Delta}\log\left(\frac{\beta_2 M_2}{\alpha_2 M_1}\right). \tag{43}$$

The expected value of the first time a population of type-2 cells in a branching process reaches the observed size $\alpha_2 M_1$ [38],

$$\mathbb{E}t = \frac{1}{r_2}\log\left(\frac{\alpha_2 M_1 r_2}{b_2}\right) - \frac{1}{r_2}\int_0^\infty e^{-z}\log z dz \tag{44}$$

$$= \frac{1}{r_2}\log\left(\frac{\alpha_2 M_1 r_2}{b_2}\right) + \frac{0.5772}{r_2} \tag{45}$$

$$\approx \frac{1}{r_2}\log(\alpha_2 M_1) \tag{46}$$

where the approximation in (46) is justified as for (24).

By (11),

$$\mathbb{E}\gamma = u\left(\frac{1-\beta_1}{r_0} + \frac{\beta_1-\beta_2}{r_1} + \frac{\beta_2}{r_2}\right)\left(\frac{1}{f_1} - \frac{1}{f_2}\right). \tag{47}$$

Using the estimates for r_0, r_1, and r_2 from (41)–(43), and setting (47) equal to the value of γ obtained from (33) and the second bulk sequencing, u can be estimated:

$$u = \frac{f_1 f_2 \gamma}{(f_2 - f_1)\left(\frac{1-\beta_1}{r_0} + \frac{\beta_1-\beta_2}{r_1} + \frac{\beta_2}{r_2}\right)}. \tag{48}$$

When estimating mutation rate for the CLL patients from Ref. [27], for which there is bulk sequencing at two or more time points, we average the mutation rate calculated at each of these time points. (48) is applied for each time point with the respective CCFs and observed γ values for each time point.

Every type-1 cell carries the m_1 passenger mutations that were present in the original type-1 cell when the first driver mutation occurred at t_1. Similarly, every type-2 cell carries the m_2 passengers that were present in the founder type-2 cell when the second driver mutation occurred at t_2. Note, none of the m_1 mutations are counted towards m_2. Now we consider the likelihood function

$$P(m_1, m_2|t_1, t'_2). \tag{49}$$

$$P(m_1, m_2|t_1, t'_2) \propto P(m_1|t_1)P(m_2|t'_2) \tag{50}$$

$$\propto \frac{(ut_1)^{m_1} e^{-ut_1}}{m_1!} \frac{(ut'_2)^{m_2} e^{-ut'_2}}{m_2!} \tag{51}$$

Now, maximizing the logarithm of (51) with respect to t_1 and t'_2,

$$t_1 = \frac{m_1}{u} \tag{52}$$

$$t'_2 = \frac{m_2}{u}. \tag{53}$$

The number of passengers present in the founder type-i cell cannot be directly observed, but we can measure $m_{i\ obs}$, the number of passengers present in all type-i cells. An expected u/r_1 passengers occurring after t_1 in type-1 cells and reaching fixation in the type-1 subclone will be incorrectly included in $m_{1\ obs}$, rather than in $m_{2\ obs}$ (see Methods). Similarly, an expected u/r_2 passengers occurring after t_2 in type-2 cells and reaching fixation in the type-2 subclone will be incorrectly included in $m_{2\ obs}$. Thus,

$$m_1 = m_{1\ obs} - u/r_1 \tag{54}$$

$$m_2 = m_{2\ obs} - u/r_2 + u/r_1. \tag{55}$$

Tumor with two sibling driver subclones

Here we consider a tumor with two "sibling" driver mutations (S4(B) Fig). Sibling driver mutations are drivers that occur in separate subclones. In this case, cells are either initiated tumor cell (type-0), carry driver 1 (type-1), or carry driver 2 (type-2). No cells contain both drivers. Driver 1 occurred in a type-0 cell at time t_1. Driver 2 occurred in a type-0 cell at t_2. Type-0 cells have birth rate b_0, death rate d_0, and net growth rate $r_0 = b_0 - d_0$. Type-1 cells, which carry driver 1, have birth rate b_1, death rate d_1, and net growth rate $r_1 = b_1 - d_1$. Type-2 cells, which carry driver 2, have birth rate b_2, death rate d_2, and net growth rate $r_2 = b_2 - d_2$. The mutation rate u is the same for all subclones.

Suppose time τ_i elapses between driver mutation i and tumor observation. Bulk sequencing of the tumor is performed at $t_1 + \tau_1$ (or equivalently $t_2 + \tau_2$), and a known Δ later. Sequencing the tumor allows the measurement of the fraction of cells with driver 1 at the first sequencing, α_1; the fraction of cells with driver 2 at the first sequencing, α_2; fraction of cells with driver 1 at the second sequencing, β_1; the fraction of cells with driver 2 at the second sequencing, β_2; and the number of subclonal passenger mutations between frequencies f_1 and f_2, γ. The fraction of type-0 cells at the two bulk sequencing time points are $1 - \alpha_1 - \alpha_2$ and $1 - \beta_1 - \beta_2$. The total number of cells at the two sequencing time points are M_1 and M_2.

We then equate the estimated cell counts to the expected value of the type-i population size X_i, conditioned on survival.

$$\mathbb{E}[X_i(t_i+\tau_i)|X_i(t_i+\tau_i)>0] = \begin{cases} \frac{b_0}{r_0}e^{r_0(t_1+\tau_1)} & i=0 \\ \frac{b_i}{r_i}e^{r_i(\tau_i)} & i=1,2 \end{cases} \tag{56}$$

$$= \begin{cases} (1-\alpha_1-\alpha_2)M_1 & i=0 \\ \alpha_i M_1 & i=1,2 \end{cases} \tag{57}$$

$$\mathbb{E}[X_i(t_i+\tau_i+\Delta)|X_i(t_i+\tau_i+\Delta)>0] = \begin{cases} \frac{b_i}{r_i}e^{r_i(t_1+\tau_1+\Delta)} & i=0 \\ \frac{b_i}{r_i}e^{r_i(\tau_i+\Delta)} & i=1,2 \end{cases} \tag{58}$$

$$= \begin{cases} (1-\beta_1-\beta_2)M_2 & i=0 \\ \beta_i M_2 & i=1,2 \end{cases} \tag{59}$$

Solving the above equations for r_i, we obtain

$$r_0 = \frac{1}{\Delta}\log\left(\frac{(1-\beta_1-\beta_2)M_2}{(1-\alpha_1-\alpha_2)M_1}\right) \tag{60}$$

$$r_i = \frac{1}{\Delta}\log\left(\frac{\beta_i M_2}{\alpha_i M_1}\right) \quad i=1,2 \tag{61}$$

The expected value of the first time a population of type-i cells in a branching process reaches the observed size $\alpha_i M_1$ is [38]

$$\mathbb{E}\tau_i = \frac{1}{r_i}\log\left(\frac{\alpha_i M_1 r_i}{b_i}\right) - \frac{1}{r_i}\int_0^\infty e^{-z}\log z dz \tag{62}$$

$$= \frac{1}{r_i}\log\left(\frac{\alpha_i M_1 r_i}{b_i}\right) + \frac{0.5772}{r_i} \tag{63}$$

$$\approx \frac{1}{r_i}\log(\alpha_i M_1) \quad i=1,2 \tag{64}$$

where the approximation in (64) is justified as for (24).

By (11),

$$\mathbb{E}\gamma = u\left(\frac{1-\beta_1-\beta_2}{r_0} + \frac{\beta_1}{r_1} + \frac{\beta_2}{r_2}\right)\left(\frac{1}{f_1} - \frac{1}{f_2}\right) \tag{65}$$

Using the estimates for r_0, r_1, and r_2 from (60) and (61), and setting (65) equal to the value of γ obtained from (33) and the second bulk sequencing, u can be estimated.

$$u = \frac{f_1 f_2 \gamma}{(f_2 - f_1)\left(\frac{1-\beta_1-\beta_2}{r_0} + \frac{\beta_1}{r_1} + \frac{\beta_2}{r_2}\right)} \tag{66}$$

When estimating mutation rate for the CLL patients from Ref. [27], for which there is bulk sequencing at two or more time points, we average the mutation rate calculated at each of these time points. (66) is applied for each time point with the respective CCFs and observed γ values for each time point.

Every type-1 cell carries the m_1 passenger mutations that were present in the original type-1 cell when the first driver mutation occurred at t_1. Similarly, every type-2 cell carries the m_2 passengers that were present in the founder type-2 cell when the second driver mutation occurred at t_2. We assume that m_1 and m_2 don't contain any shared mutations. In the CLL dataset we use, this is true. We consider the likelihood function $P(m_1, m_2|t_1, t_2)$

$$P(m_1, m_2|t_1, t_2) \propto P(m_1|t_1)P(m_2|t_2) \tag{67}$$

$$\propto \frac{(ut_1)^{m_1} e^{-ut_1}}{m_1!} \frac{(ut_2)^{m_2} e^{-ut_2}}{m_2!}. \tag{68}$$

Maximizing the logarithm of (68) with respect to t_1 and t_2 yields the maximum likelihood estimates:

$$t_1 = \frac{m_1}{u} \tag{69}$$

$$t_2 = \frac{m_2}{u}. \tag{70}$$

Using the same approach as in the case of a single driver, we obtain the corrections for the observed number of mutations present in all cells of each subclone:

$$m_1 = m_{1\ obs} - u/r_1 \tag{71}$$

$$m_2 = m_{2\ obs} - u/r_2. \tag{72}$$

Fully generalized estimates for any phylogeny of *k* drivers

Here we derive estimates for a completely general tumor phylogeny. Suppose a tumor has k driver mutations. In this general case, define a type-i cell as a cell where its most recent driver mutation was driver i. Note that a type-i cell can have between 0 and $k - 1$ other driver mutations. A phylogenetic reconstruction of the k driver mutations is necessary for the completely general case. From this phylogenetic tree, the ancestor of each subclone can be obtained. Define the function $a(i)$ as the ancestor of the type-i population. That is, if all driver mutations contained in the type-i population are ordered, $a(i)$ gives the driver mutation that occurred prior to i. Define t_i as the time between when driver i occurred and when the type-i cells' previous driver mutation occurred. At time of observation, assume the type-i population has κ_i total driver mutations, where $1 \leq \kappa_i \leq k$ for all $1 \leq i \leq k$. Denote the time between the type-i's κ_i, or last, driver mutation and when the tumor is observed as τ_i. This is the time between the founder type-i cell's birth and tumor observation. Then the tumor is first observed and bulk sequenced at $T_1 \equiv (\sum_{j=0}^{\kappa_i-1} t_{a^j(i)}) + \tau_i$ (equivalently τ_0 for $i = 0$), where we denote a^j as the jth iterate of the function a:

$$a^0(i) \equiv i \tag{73}$$

$$a^j(i) \equiv a(a^{j-1}(i)) \quad \forall j \geq 1. \tag{74}$$

The tumor is also bulk sequenced at $T_2 \equiv (\sum_{j=0}^{\kappa_i-1} t_{a^j(i)}) + \tau_i + \Delta$ (equivalently $\tau_0 + \Delta$ for $i = 0$). These assumptions allow for any subclone phylogeny, including combinations of the previously discussed sibling and nested subclone types.

The bulk sequencing allows the measurement of the fraction of cells with driver i at T_1, α_i; the fraction of cells with driver i at time T_2, β_i; and the number of subclonal passenger mutations between frequencies f_1 and f_2, γ. Again, the total number of cells at measurement times T_1 and T_2 are M_1 and M_2. To write the type-i frequencies, α_i^c and β_i^c, in terms of the driver frequencies, we subtract the fraction of cells descending from type-i cells but gaining additional driver mutation(s) after i, from the fraction of cells containing driver i:

$$\alpha_i^c = \begin{cases} \alpha_i - \sum_{j=1}^{k} \delta_{i,a(j)} \alpha_j & 1 \leq i \leq k \\ 1 - \sum_{j=1}^{k} \alpha_j^c & i = 0 \end{cases} \tag{75}$$

$$\beta_i^c = \begin{cases} \beta_i - \sum_{j=1}^{k} \delta_{i,a(j)} \beta_j & 1 \leq i \leq k \\ 1 - \sum_{j=1}^{k} \beta_j^c & i = 0 \end{cases} \tag{76}$$

where $\delta_{i,a(j)}$ is the Kronecker delta, defined as

$$\delta_{i,a(j)} = \begin{cases} 0 & \text{if } i \neq a(j) \\ 1 & \text{if } i = a(j) \end{cases}.$$

We equate the estimated cell counts at the first bulk sequencing time point to the expected value of the type-i population size X_i, conditioned on survival.

$$\begin{aligned} \mathbb{E}[X_i(T_1)|X_i(T_1) > 0] &= \frac{b_i}{r_i} e^{r_i \tau_i} \\ &= \alpha_i^c M_1 \end{aligned} \tag{77}$$

And similarly, at the second bulk sequencing time point,

$$\mathbb{E}[X_i(T_2)|X_i(T_2) > 0] = \frac{b_i}{r_i} e^{r_i(\tau_i + \Delta)} \tag{78}$$

$$= \beta_i^c M_2. \tag{79}$$

Solving the above equations for r_i, we obtain

$$r_i = \frac{1}{\Delta} \log \left(\frac{\beta_i^c M_2}{\alpha_i^c M_1} \right) \quad \forall i = 0, 1, \ldots, k. \tag{80}$$

By (11)

$$\mathbb{E}\gamma = \left(u\sum_{i=0}^{k}\frac{\beta_i^c}{r_i}\right)\left(\frac{1}{f_1}-\frac{1}{f_2}\right). \tag{81}$$

Now, using the growth rate estimates r_i and the subclone sizes, we can estimate each τ_i. The expected value of the first time a population of type-i cells in a branching process reaches the observed size $\alpha_i^c M_1$ is [38]

$$\mathbb{E}\tau_i = \frac{1}{r_i}\log\left(\frac{\alpha_i^c M_1 r_i}{b_i}\right) - \frac{1}{r_i}\int_0^\infty e^{-z}\log z dz \tag{82}$$

$$= \frac{1}{r_i}\log\left(\frac{\alpha_i^c M_1 r_i}{b_i}\right) + \frac{0.5772}{r_i} \tag{83}$$

$$\approx \frac{1}{r_i}\log\left(\alpha_i^c M_1\right) \tag{84}$$

where the approximation in (84) is justified as for (24).

Using the $(k + 1)$ r_i estimates from (80), and setting (81) equal to the value of γ obtained at the second bulk sequencing from (33), u can be estimated:

$$u = \frac{f_1 f_2 \gamma}{(f_2 - f_1)\left(\sum_{i=0}^{k}\frac{\beta_i^c}{r_i}\right)}. \tag{85}$$

When estimating mutation rate for the CLL patients from Ref. [27], for which there is bulk sequencing at two or more time points, we average the mutation rate calculated at each of these time points. (85) is applied for each time point with the respective CCFs and observed γ values for each time point.

The number of passengers present in the original type i founder cell cannot be directly observed, but we can measure m_i, the number of clonal passengers present in the type i population, only including passengers not present in other clones. We will assume that the m_i don't contain any shared mutations, which is true for the CLL dataset we consider. The likelihood function $P(m_1, \ldots, m_k|t_1, \ldots, t_k)$ is proportional to

$$\prod_{i=1}^{k} P(m_i|t_i) \propto \prod_{i=1}^{k}\frac{(ut_i)^{m_i}e^{-ut_i}}{m_i!}. \tag{86}$$

Then, maximizing the logarithm of (86) with respect to $t_1, t_2, \ldots, t_k$,

$$t_i = \frac{m_i}{u} \quad \forall i = 1, \ldots, k. \tag{87}$$

The observed clonal passengers in the founder type-i cell will incorrectly include passengers that reached fixation in the type-i population after driver mutation i occurred, instead of correctly being counted toward the descendant of clone i. As a result, we again correct for the expected number of these passengers, u/r_i. That is,

$$m_i = m_{i,\,obs} - u/r_i + u/r_{a(i)} \quad \forall i = 1, \ldots, k. \tag{88}$$

Supporting information

S1 Fig. Percent errors (PEs) for case with no death. Accuracy of parameter inferences for Monte Carlo simulation of tumor with no cell death for (a) single driver subclone with mutation rate $u = 1$, (b) single driver subclone with $u = 10$, (c) two nested subclones with $u = 1$, and (d) two sibling subclones with $u = 1$. Mean percent error (MPEs) are the black numbers above the plots, and mean absolute percent errors (MAPEs) are the grey numbers below the MPEs. Boxes contain 25th-75th quartiles, with median indicated by thick horizontal black line. Whiskers of boxplots indicate 2.5 and 97.5 percentiles. Violins are smoothed density estimates of the percent error datapoints. Complete parameter values and number of runs are included in S1 Table.
(PDF)

S2 Fig. Percent errors (PEs) for slow-growing tumor. Accuracy of parameter inferences for surviving Monte Carlo simulation runs of slow-growing tumor for (a) single subclone with mutation rate $u = 1$, (b) single subclone with $u = 5$, (c) two nested subclones with $u = 1$, and (d) two sibling subclones with $u = 1$. Mean percent error (MPEs) are the black numbers above the plots, and mean absolute percent errors (MAPEs) are the grey numbers below the MPEs. Boxes contain 25th-75th quartiles, with median indicated by thick horizontal black line. Whiskers of boxplots indicate 2.5 and 97.5 percentiles. Violins are smoothed density estimates of the percent error data points. Complete parameter values and number of runs are included in S1 Table.
(PDF)

S3 Fig. Corrections for observed mutation counts. (a) We compare percent errors of parameter estimates for time from tumor initiating until appearance of a driver subclone, t_1, with and without the correction for passengers that occur after the driver and reach fixation in the driver population (Eq (6), main text). Errors for estimate with correction are shown in blue, and for estimate without correction (Eq (5), main text) in orange. Errors are plotted as a kernel density estimate for Monte Carlo simulations of fast-growing tumor with mutation rate $u = 1$. Mean percent errors (MPEs) and mean absolute percent errors (MAPEs) are listed. (b) The percent errors for the observed (orange) and corrected (blue) number of subclonal mutations between frequencies f_1 and f_2, γ, (Eq (7), main text) are plotted as kernel density estimates. Observed mutations are those that passed post-processing, i.e. those that have more than $L = 2$ mutant reads. True mutation frequencies were generated from 135 surviving runs of a Monte Carlo simulation of a fast-growing tumor with mutation rate $u = 1$, from which sequencing reads were simulated with 100x average coverage (see Materials and methods). Percent errors are calculated relative to the true γ measured from the true mutation frequencies.
(PDF)

S4 Fig. Model for tumor expansion with two driver mutations. (a) Two nested driver subclones. Initiated tumor (type-0) cells in blue, cells with driver 1 (type-1) in red, and cells with both drivers (type-2) in orange. A driver mutation occurs in a type-0 cell at t_1. A second driver mutation occurs in a type-1 cell at $t_1 + t_2'$. Tumor is bulk sequenced at $t_1 + t_2' + t$ and $t_1 + t_2' + t + \Delta$. (b) Two sibling driver subclones. Type-0 cells (in blue). A driver mutation occurs in a type-0 cell at t_1. A second driver mutation occurs in a different type-0 cell at t_2. Tumor is bulk sequenced at $t_1 + \tau_1$ (or, equivalently $t_2 + \tau_2$) and $t_1 + \tau_1 + \Delta$ (equivalently $t_2 + \tau_2 + \Delta$).
(PDF)

S5 Fig. Accuracy for *t* estimate increases with tumor size. A Monte Carlo simulation of a birth-death process was performed for (a) fast-growing, (b) slow-growing, and (c) no cell death parameter regimes. For each of the 100 surviving simulated tumors, the percent error of the *t* estimate (Eq (3)) was calculated when the tumor first reached the specified tumor sizes. Means are indicated by red points and lines, ± one standard deviation is shown by the red region, and individual data points for each simulation run are shown as the grey points (with horizontal jitter for visibility).
(PDF)

S1 Table. Parameter values. Parameter values and number of surviving runs for Monte Carlo simulations. For all simulations $f_1 = 0.01$, $f_2 = 0.20$, $L = 2$.
(XLSX)

S1 Methods. Unbiasedness of growth rate.
(PDF)

Author Contributions

Conceptualization: Nathan D. Lee, Ivana Bozic.

Formal analysis: Nathan D. Lee.

Methodology: Nathan D. Lee, Ivana Bozic.

Software: Nathan D. Lee.

Supervision: Ivana Bozic.

Visualization: Nathan D. Lee.

Writing – original draft: Nathan D. Lee, Ivana Bozic.

Writing – review & editing: Nathan D. Lee, Ivana Bozic.

References

1. Nowell PC. The Clonal Evolution of Tumor Cell Populations. Science. 1976; 194(4260):23–28. https://doi.org/10.1126/science.959840 PMID: 959840
2. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458(7239):719–724. https://doi.org/10.1038/nature07943 PMID: 19360079
3. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011; 144(5):646–674. https://doi.org/10.1016/j.cell.2011.02.013 PMID: 21376230
4. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018; 173(2):371–385.e18. https://doi.org/10.1016/j.cell.2018.02.060 PMID: 29625053
5. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nature Genetics. 2016; 48(3):238–244. https://doi.org/10.1038/ng.3489 PMID: 26780609
6. Kimura M. Evolutionary Rate at the Molecular Level. Nature. 1968; 217(5129):624–626. https://doi.org/10.1038/217624a0 PMID: 5637732
7. Kimura M. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles*. Genetics Research. 1968; 11(3):247–270. https://doi.org/10.1017/S0016672300011459
8. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genetics Research. 1974; 23(1):23–35. https://doi.org/10.1017/S0016672300014634 PMID: 4407212
9. Turajlic S, Sottoriva A, Graham T, Swanton C. Resolving genetic heterogeneity in cancer. Nature Reviews Genetics. 2019; 20(7):404–416. https://doi.org/10.1038/s41576-019-0114-6 PMID: 30918367

10. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013; 339(6127):1546–1558. https://doi.org/10.1126/science.1235122 PMID: 23539594

11. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499(7457):214–218. https://doi.org/10.1038/nature12213 PMID: 23770567

12. Merlo LMF, Pepper JW, Reid BJ, Maley CC. Cancer as an evolutionary and ecological process. Nature Reviews Cancer. 2006; 6(12):924–935. https://doi.org/10.1038/nrc2013 PMID: 17109012

13. Pepper JW, Findlay CS, Kassen R, Spencer SL, Maley CC. SYNTHESIS: Cancer research meets evolutionary biology. Evolutionary Applications. 2009; 2(1):62–70. https://doi.org/10.1111/j.1752-4571.2008.00063.x PMID: 25567847

14. Tsao JL, Yatabe Y, Salovaara R, Järvinen HJ, Mecklin JP, Aaltonen LA, et al. Genetic reconstruction of individual colorectal tumor histories. Proceedings of the National Academy of Sciences. 2000; 97(3): 1236–1241. https://doi.org/10.1073/pnas.97.3.1236 PMID: 10655514

15. Jones S, Chen Wd, Parmigiani G, Diehl F, Beerenwinkel N, Antal T, et al. Comparative lesion sequencing provides insights into tumor evolution. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(11):4283–4288. https://doi.org/10.1073/pnas.0712345105 PMID: 18337506

16. Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature. 2010; 467(7319):1114–1117. https://doi.org/10.1038/nature09515 PMID: 20981102

17. Naxerova K, Brachtel E, Salk JJ, Seese AM, Power K, Abbasi B, et al. Hypermutable DNA chronicles the evolution of human colon cancer. Proceedings of the National Academy of Sciences. 2014; 111(18):E1889–E1898. https://doi.org/10.1073/pnas.1400179111 PMID: 24753616

18. McGranahan N, Favero F, Bruin ECd, Birkbak NJ, Szallasi Z, Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Science Translational Medicine. 2015; 7(283):283ra54–283ra54. https://doi.org/10.1126/scitranslmed.aaa1408 PMID: 25877892

19. Mitchell TJ, Turajlic S, Rowan A, Nicol D, Farmery JHR, O'Brien T, et al. Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. Cell. 2018; 173(3):611–623.e17. https://doi.org/10.1016/j.cell.2018.02.020 PMID: 29656891

20. PCAWG Evolution & Heterogeneity Working Group, PCAWG Consortium, Gerstung M, Jolly C, Leshchiner I, Dentro SC, et al. The evolutionary history of 2,658 cancers. Nature. 2020; 578(7793):122–128. https://doi.org/10.1038/s41586-019-1907-7 PMID: 32025013

21. Sundermann LK, Wintersinger J, Rätsch G, Stoye J, Morris Q. Reconstructing tumor evolutionary histories and clone trees in polynomial-time with SubMARine. PLOS Computational Biology. 2021; 17(1): e1008400. https://doi.org/10.1371/journal.pcbi.1008400 PMID: 33465079

22. PCAWG Evolution and Heterogeneity Working Group, PCAWG Consortium, Rubanova Y, Shi R, Harrigan CF, Li R, et al. Reconstructing evolutionary trajectories of mutation signature activities in cancer using TrackSig. Nature Communications. 2020; 11(1):731. https://doi.org/10.1038/s41467-020-14352-7 PMID: 32024834

23. Tomasetti C, Bozic I. The (not so) immortal strand hypothesis. Stem Cell Research. 2015; 14(2):238–241. https://doi.org/10.1016/j.scr.2015.01.005 PMID: 25700960

24. Werner B, Case J, Williams MJ, Chkhaidze K, Temko D, Fernández-Mateos J, et al. Measuring single cell divisions in human tissues from multi-region sequencing data. Nature Communications. 2020; 11(1):1–9. https://doi.org/10.1038/s41467-020-14844-6 PMID: 32098957

25. Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, et al. Accumulation of driver and passenger mutations during tumor progression. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107(43):18545–18550. https://doi.org/10.1073/pnas.1010978107 PMID: 20876136

26. Sun R, Hu Z, Sottoriva A, Graham TA, Harpak A, Ma Z, et al. Between-region genetic divergence reflects the mode and tempo of tumor evolution. Nature Genetics. 2017; 49(7):1015–1024. https://doi.org/10.1038/ng.3891 PMID: 28581503

27. Gruber M, Bozic I, Leshchiner I, Livitz D, Stevenson K, Rassenti L, et al. Growth dynamics in naturally progressing chronic lymphocytic leukaemia. Nature. 2019; 570(7762):474–479. https://doi.org/10.1038/s41586-019-1252-x PMID: 31142838

28. Salichos L, Meyerson W, Warrell J, Gerstein M. Estimating growth patterns and driver effects in tumor evolution from individual samples. Nature Communications. 2020; 11(1):1–14. https://doi.org/10.1038/s41467-020-14407-9 PMID: 32024824

29. Noble R, Burri D, Le Sueur C, Lemant J, Viossat Y, Kather JN, et al. Spatial structure governs the mode of tumour evolution. Nature Ecology & Evolution. 2021. https://doi.org/10.1038/s41559-021-01615-9 PMID: 34949822

30. Chkhaidze K, Heide T, Werner B, Williams MJ, Huang W, Caravagna G, et al. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. PLOS Computational Biology. 2019; 15(7):e1007243. https://doi.org/10.1371/journal.pcbi.1007243 PMID: 31356595

31. Fu X, Zhao Y, Lopez JI, Rowan A, Au L, Fendler A, et al. Spatial patterns of tumour growth impact clonal diversification in a computational model and the TRACERx Renal study. Nature Ecology & Evolution. 2021.

32. Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, et al. Quantification of subclonal selection in cancer from bulk sequencing data. Nature Genetics. 2018; 50(6):895. https://doi.org/10.1038/s41588-018-0128-6 PMID: 29808029

33. Avanzini S, Kurtz DM, Chabon JJ, Moding EJ, Hori SS, Gambhir SS, et al. A mathematical model of ctDNA shedding predicts tumor detection size. Science Advances. 2020; 6(50):eabc4308. https://doi.org/10.1126/sciadv.abc4308 PMID: 33310847

34. Bozic I, Reiter JG, Allen B, Antal T, Chatterjee K, Shah P, et al. Evolutionary dynamics of cancer in response to targeted combination therapy. eLife. 2013; 2:e00747. https://doi.org/10.7554/eLife.00747 PMID: 23805382

35. Dinh KN, Jaksik R, Kimmel M, Lambert A, Tavaré S. Statistical Inference for the Evolutionary History of Cancer Genomes. Statistical Science. 2020; 35(1):129–144. https://doi.org/10.1214/19-STS7561

36. Lahouel K, Younes L, Danilova L, Giardiello FM, Hruban RH, Groopman J, et al. Revisiting the tumorigenesis timeline with a data-driven generative model. Proceedings of the National Academy of Sciences. 2020; 117(2):857–864. https://doi.org/10.1073/pnas.1914589117 PMID: 31882448

37. Bozic I, Wu CJ. Delineating the evolutionary dynamics of cancer from theory to reality. Nature Cancer. 2020; 1(6):580–588. https://doi.org/10.1038/s43018-020-0079-6 PMID: 35121980

38. Durrett R. Branching Process Models of Cancer. In: Durrett R, editor. Branching Process Models of Cancer. Mathematical Biosciences Institute Lecture Series. Cham: Springer International Publishing; 2015. p. 1–63. Available from: https://doi.org/10.1007/978-3-319-16065-8_1.

39. Tavaré S. The linear birth-death process: an inferential retrospective. Advances in Applied Probability. 2018; 50(A):253–269. https://doi.org/10.1017/apr.2018.84

40. Heyde A, Reiter JG, Naxerova K, Nowak MA. Consecutive seeding and transfer of genetic diversity in metastasis. Proceedings of the National Academy of Sciences. 2019; 116(28):14129–14137. https://doi.org/10.1073/pnas.1819408116

41. Griffith M, Miller C, Griffith O, Krysiak K, Skidmore Z, Ramu A, et al. Optimizing Cancer Genome Sequencing and Analysis. Cell Systems. 2015; 1(3):210–223. https://doi.org/10.1016/j.cels.2015.08.015 PMID: 26645048

42. Haber DA, Velculescu VE. Blood-Based Analyses of Cancer: Circulating Tumor Cells and Circulating Tumor DNA. Cancer Discovery. 2014; 4(6):650–661. https://doi.org/10.1158/2159-8290.CD-13-1014 PMID: 24801577

43. Leshchiner I, Livitz D, Gainor JF, Rosebrock D, Spiro O, Martinez A, et al. Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. Bioinformatics; 2018. Available from: http://biorxiv.org/lookup/doi/10.1101/508127.

44. Myers MA, Satas G, Raphael BJ. CALDER: Inferring Phylogenetic Trees from Longitudinal Tumor Samples. Cell Systems. 2019; 8(6):514–522.e5. https://doi.org/10.1016/j.cels.2019.05.010 PMID: 31229560

45. Hallek M, Cheson BD, Catovsky D, Caligaris-Cappio F, Dighiero G, Döhner H, et al. iwCLL guidelines for diagnosis, indications for treatment, response assessment, and supportive management of CLL. Blood. 2018; 131(25):2745–2760. https://doi.org/10.1182/blood-2017-09-806398 PMID: 29540348

46. Marionneaux SM, Keohane EM, Lamanna N, King TC, Mehta SR. Smudge Cells in Chronic Lymphocytic Leukemia: Pathophysiology, Laboratory Considerations, and Clinical Significance. Laboratory Medicine. 2021; 52(5):426–438. https://doi.org/10.1093/labmed/lmaa119 PMID: 33527134

47. Bozic I, Gerold JM, Nowak MA. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. PLOS Computational Biology. 2016; 12(2):e1004731. https://doi.org/10.1371/journal.pcbi.1004731 PMID: 26828429

48. Kim J, Kim D, Lim JS, Maeng JH, Son H, Kang HC, et al. The use of technical replication for detection of low-level somatic mutations in next-generation sequencing. Nature Communications. 2019; 10(1):1047. https://doi.org/10.1038/s41467-019-09026-y PMID: 30837471

49. Song P, Chen SX, Yan YH, Pinto A, Cheng LY, Dai P, et al. Selective multiplexed enrichment for the detection and quantitation of low-fraction DNA variants via low-depth sequencing. Nature Biomedical Engineering. 2021; 5(7):690–701. https://doi.org/10.1038/s41551-021-00713-0 PMID: 33941896

50. Fabre MA, de Almeida JG, Fiorillo E, Mitchell E, Damaskou A, Rak J, et al. The longitudinal dynamics and natural history of clonal haematopoiesis. Nature. 2022; 606(7913):335–342. https://doi.org/10.1038/s41586-022-04785-z PMID: 35650444

51. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016; 538(7624):260–264. https://doi.org/10.1038/nature19768 PMID: 27698416

52. Mitchell E, Spencer Chapman M, Williams N, Dawson KJ, Mende N, Calderbank EF, et al. Clonal dynamics of haematopoiesis across the human lifespan. Nature. 2022; 606(7913):343–350. https://doi.org/10.1038/s41586-022-04786-y PMID: 35650442

53. Auslander N, Wolf YI, Koonin EV. In silico learning of tumor evolution through mutational time series. Proceedings of the National Academy of Sciences. 2019; 116(19):9501–9510. https://doi.org/10.1073/pnas.1901695116 PMID: 31015295

54. Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing SV, Ellis P, et al. Somatic mutation landscapes at single-molecule resolution. Nature. 2021; 593(7859):405–410. https://doi.org/10.1038/s41586-021-03477-4 PMID: 33911282

55. PCAWG Mutational Signatures Working Group, PCAWG Consortium, Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, et al. The repertoire of mutational signatures in human cancer. Nature. 2020; 578(7793):94–101. https://doi.org/10.1038/s41586-020-1943-3 PMID: 32025018

56. Wang Z, Xia Y, Mills L, Nikolakopoulos AN, Maeser N, Sheltzer JM, et al. Evolving copy number gains promote tumor expansion and bolster mutational diversification. Genomics; 2022. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.06.14.495959.

57. Friberg S, Mattson S. On the growth rates of human malignant tumors: implications for medical decision making. Journal of Surgical Oncology. 1997; 65(4):284–297. https://doi.org/10.1002/(SICI)1096-9098(199708)65:4%3C284::AID-JSO11%3E3.0.CO;2-2 PMID: 9274795

58. Rodriguez-Brenes IA, Komarova NL, Wodarz D. Tumor growth dynamics: insights into evolutionary processes. Trends in Ecology & Evolution. 2013; 28(10):597–604. https://doi.org/10.1016/j.tree.2013.05.020 PMID: 23816268

59. Talkington A, Durrett R. Estimating Tumor Growth Rates In Vivo. Bulletin of Mathematical Biology. 2015; 77(10):1934–1954. https://doi.org/10.1007/s11538-015-0110-8 PMID: 26481497

60. Norton L. A Gompertzian model of human breast cancer growth. Cancer Research. 1988; 48(24 Pt 1):7067–7071. PMID: 3191483

61. Spratt JA, von Fournier D, Spratt JS, Weber EE. Decelerating growth and human breast cancer. Cancer. 1993; 71(6):2013–2019. https://doi.org/10.1002/1097-0142(19930315)71:6%3C2013::AID-CNCR2820710615%3E3.0.CO;2-V PMID: 8443753

62. Gerlee P. The Model Muddle: In Search of Tumor Growth Laws. Cancer Research. 2013; 73(8):2407. https://doi.org/10.1158/0008-5472.CAN-12-4355 PMID: 23393201

63. Vaghi C, Rodallec A, Fanciullino R, Ciccolini J, Mochel JP, Mastri M, et al. Population modeling of tumor growth curves and the reduced Gompertz model improve prediction of the age of experimental tumors. PLOS Computational Biology. 2020; 16(2):e1007178. https://doi.org/10.1371/journal.pcbi.1007178 PMID: 32097421

64. Carlsson G, Gullberg B, Hafström L. Estimation of liver tumor volume using different formulas?An experimental study in rats. Journal of Cancer Research and Clinical Oncology. 1983; 105(1):20–23. https://doi.org/10.1007/BF00391826 PMID: 6833336

65. West J, Schenck RO, Gatenbee C, Robertson-Tessi M, Anderson ARA. Normal tissue architecture determines the evolutionary course of cancer. Nature Communications. 2021; 12(1):2060. https://doi.org/10.1038/s41467-021-22123-1 PMID: 33824323

66. Marusyk A, Tabassum DP, Altrock PM, Almendro V, Michor F, Polyak K. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. Nature. 2014; 514(7520):54–58. https://doi.org/10.1038/nature13556 PMID: 25079331

67. Petljak M, Alexandrov LB, Brammeld JS, Price S, Wedge DC, Grossmann S, et al. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell. 2019; 176(6):1282–1294.e20. https://doi.org/10.1016/j.cell.2019.02.012 PMID: 30849372

68. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry. 1977; 81(25):2340–2361. https://doi.org/10.1021/j100540a008

69. Bozic I, Paterson C, Waclaw B. On measuring selection in cancer from subclonal mutation frequencies. PLOS Computational Biology. 2019; 15(9):e1007368. https://doi.org/10.1371/journal.pcbi.1007368 PMID: 31557163

70. Keohane EM, Smith LJ, Walenga JM, editors. Rodak's hematology: clinical principles and applications. Fifth edition ed. St. Louis, Missouri.: Elsevier/Saunders; 2016.