Inferring parameters of cancer evolution from sequencing and clinical data

Nathan Lee[1] and Ivana Bozic[1*]

[1]Department of Applied Mathematics, University of Washington, Seattle, WA, USA

*Correspondence to 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. 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. We apply our methodology to reconstruct the individual evolutionary histories of chronic lymphocytic leukemia patients. Fitting our model to longitudinal patient data reveals strengths and weaknesses of using an exponential model of cancer growth with constant mutation rate to estimate parameters of cancer evolution.

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, 2, 3]. The positive selection of advantageous variants, as well as neutral evolutionary dynamics, have been shown to govern the carcinogenic process [4]. Genome sequencing technologies have revealed the heterogeneous, informative genetic profiles produced by this evolutionary process driving carcinogenesis [5, 6]. 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 [7].

Mutations that don't have a significant effect on cellular fitness also arise, both before and after tumor initiation [4]. 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 [8, 9, 10, 11]. Mutational burden detectable by bulk sequencing reveals tens to thousands of passengers per tumor [12, 13]. The accumulation of neutral mutations over time can serve as a "molecular evolutionary clock" [14]. The molecular clock feature of passengers has been employed to measure timing of early events in tumor formation, well before detectability, as well as identify stages of tumorigenesis and metastasis [15, 16, 17, 18, 19, 20, 21, 22, 23].

Tumors can grow for many years, even decades, before they reach detectable size [17]. 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 structure in various cancer types [24, 25, 26, 27, 28].

Cancer genome sequencing, along with computational and mathematical modeling, has provided insights into parameters of the carcinogenic process operating in individual patients. Many approaches limit their focus to specific features of cancer evolution. For example, some studies have estimated mutation rates [4, 29], selective growth advantages within cancer subclones [30, 31, 28, 32], rate parameters of birth-death processes [33, 34], and the effect of spatial structure on cancer evolution [35, 36]. Many state of the art methods make use of computationally expensive approaches [37, 36, 29] or simplifying assumptions, such as approximating tumor expansion as deterministic or ignoring cell death [37, 28]. We use a stochastic branching process model for cancer growth, which includes cell death, to derive a comprehensive reconstruction of an individual tumor's evolution.

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). 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 tumor's timeline of selective events and quantitative characterization of the evolutionary dynamics of the subclones making up the tumor.

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.

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. Cells containing i new driver mutations, where i is either 0 or 1, will be referred to as type-i cells or simply, clone i. In the Supplementary Methods we present similar estimates for the more general case of two nested or sibling driver mutations, as well as the fully generalized estimates for any clonal structure that might arise during tumor expansion.

Results

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). From the bulk sequencing data, the fraction of cells carrying the driver mutation, α_1 and α_2, can be measured at the timepoints $t_1 + t$ and $t_1 + t + \delta$. We denote total number of cells in the tumor at the two bulk sequencing timepoints as M_1 and M_2. Number of cells in the tumor can be estimated from measurements of tumor volume [38].

Equating expected values of the sizes of 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 [39]. This yields an estimate of the time t from the appearance of driver mutation until observation:

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

Using the bulk sequencing data from the second timepoint, γ, the number of subclonal passengers between the specified frequencies f_1 and f_2, can be measured. Using results from previous work [40], we derive the expected value of γ (see Supplementary 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 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 Supplementary 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.

Finally, we include the following correction to formula (5) above. The adjustment involves 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 from mutations that occurred after t_1 in type-1 cells, and reached fixation in the type-1 population [40]. Thus, the value of m observed from sequencing data, m_{obs}, will overestimate m. In the Supplementary 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}:

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

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 Supplementary 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.

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. 2). Growth rates of both initiated tumor and driver subclones can be estimated with a high degree of accuracy, achieving mean percentage error (MPE) of -0.07% and 0.03% for the lower mutation rate ($u = 1$). The mutation rate u and time estimates t_1 and t can also be estimated accurately, with MPEs of -.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 (Fig. S1, Fig. S2). For the lower mutation rate ($u = 1$), parameter estimates for the mutation rate u and time estimate 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. The t estimates have somewhat higher errors, with MPE of -6.3% for the no cell death case, and MPE of 30.3% for the slow-growing tumor.

We note that our correction for the m mutations present in the original type-1 cell (6) improves the accuracy of the estimate for time of appearance of driver mutation t_1 (Fig. 3ab). For the fast growing tumor with mutation rate $u = 1$ (Fig. 3a), 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%.

We note that the availability of single cell sequencing measurements, in addition to the bulk sequencing, would additionally improve the accuracy of the estimates. In that case, time between tumor initiation and tumor observation can be computed from the number of passenger mutations observed in single cells.

Correcting mutation counts observed from genome sequencing data

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 [41]. For a sample with average sequencing coverage of ρ, mutations with mutant allele frequency below L/ρ will typically not be observable.

As a result, since mutations with frequencies between f_1 and f_2 count towards γ, if $f_1 \leq 2L/\rho$, the observed number of subclonal mutations between frequencies f_1 and f_2, γ_{obs}, will underestimate the true value, γ. In the Supplementary Methods, we derive a correction for γ, based on the expected value of the number of subclonal mutations present at cancer cell frequencies between f_1 and $2L/\rho$:

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

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 Supplementary 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 (Fig. 3c). 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).

In sum, the parameter estimates (including corrections to mutations counts m and γ) we will apply to patient clinical and sequencing data are:

$$r = \frac{1}{\delta}\log\left(\frac{(1-\alpha_2)M_2}{(1-\alpha_1)M_1}\right) \tag{8}$$

$$r_1 = \frac{1}{\delta}\log\left(\frac{\alpha_2 M_2}{\alpha_1 M_1}\right) \tag{9}$$

$$t = \frac{1}{r_1}\log(M_1\alpha_1) \tag{10}$$

$$u = \frac{r r_1 \gamma_{obs}}{(\frac{\rho}{2L} - \frac{1}{f_2})(\alpha_2 r + r_1(1-\alpha_2))} \tag{11}$$

$$t_1 = \frac{m_{obs}}{u} - 1/r_1 \tag{12}$$

Estimating parameters for individual patients with CLL

We use our formulas to infer the patient-specific parameters of cancer evolution for four patients with chronic lymphocytic leukemia (CLL) whose growth patterns and clonal dynamics were analyzed in [28]. These CLLs had peripheral white blood cell (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. We chose patients with $\gamma_{obs} > 3$ and examined subclones that had cancer

cell fractions (CCFs) of $\sim$ 20% or greater for both time points. 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 (Table S1). We reconstruct these histories for tumors with various clonal structures.

Patients 3 and 21 are examples of a CLL with a single subclone (Fig. 4). For patient 21, we estimate that Clone 0, most recent common ancestor (MRCA) of this patient's CLL, was initiated when the patient was less than a year old, and grew with a net growth rate of 0.82 per year. Clone 1 appeared 18.5 years later, and grew more quickly than Clone 0, with a growth rate of 1.46 per year (corresponding to selective growth advantage of 78.4% 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 $\sim$16 years before the patient was diagnosed at age 35. We estimate that this patient's CLL had an exome mutation rate of 0.16 mutations per year.

The parental clone (MRCA, Clone 0) of the CLL of Patient 3 was initiated when the patient was 29 years old. Clone 0 grew with a net growth rate of 0.61 per year. 9 years later, Clone 1 was initiated when the patient was 38 years old. Clone 1 expanded with a growth rate of 1.02 per year (corresponding to a selective growth advantage of 66.8% over Clone 0), and the patient was diagnosed 24 years later at age 63. We find that the CLL exome mutation rate was 0.59 mutations per year in this patient.

Patients 9 and 13 present more complex clonal structures. CLL of Patient 13 contains two nested subclones (Clones 1 and 2) within MRCA Clone 0. We estimate that the CLL was initiated when the patient was $\sim$10 years old. Clone 0 then grew at a rate of 0.23 per year. 15.3 years after the appearance of Clone 0, when the patient was 25 years old, the first nested subclone, Clone 1, appeared. At the time of sequencing, Clone 1 had a net growth rate of -0.22 per year, and was decreasing in size. Clone 2, a nested subclone inside Clone 1, began growing 12.3 years after Clone 1 when the patient was 38 years old. Clone 2, with an ATM driver mutation, had the largest net growth rate of the three clones (1.19 per year). This corresponds to a strong selective growth advantage of 409% over clone 0. Clone 2 became the dominant clone, making up 57% of the cancer cells. 18.9 years after clone 2 initiation, the patient was diagnosed at age 57. In Patient 13, we estimate a CLL exome mutation rate of 0.56 mutations per year.

CLL of Patient 9 contains two sibling subclones, Clones 1 and 2, in addition to the parental population, Clone 0. Growth rate of Clone 0 is 0.21 per year. At the time of sequencing, Clone 1 had a negative growth rate of -0.44 (/year). A second sibling subclone, Clone 2, containing a KRAS mutation, had the largest net growth rate of the three clones (0.43 per year), resulting in a selective growth advantage of 98.8%. We estimate that the CLL exome mutation rate of Patient 9 is 0.139 mutations/year. The estimates of the timing of subclonal events are unrealistic for this patient. The large estimate for time between MRCA and

t_2 is mostly due to the high number of mutations present in the founder cell of Clone 2 (16), compared to a median of 7 mutations in typical CLL subclones. Contributing factors are the relatively low estimated mutation rate in this CLL (so that on average a long time is required for accumulating mutations), and the relatively slow growth rate of the most fit subclone, Clone 2, leading to a high estimate for the time from appearance of Clone 2 to diagnosis.

The example of Patient 9 reveals potential weaknesses of using an exponential model of cancer growth with constant mutation rate to estimate parameters of cancer evolution: some cancer subclones (such as Clone 1 from Pt. 9) not only do not grow exponentially, they actually decline in absolute cell numbers. As we are using longitudinal data, unlike many previous approaches [19, 4, 37, 23], we are able to obtain absolute time estimates of events in tumor evolution. However, the obtained time estimates are sometimes incompatible with the known age of the patient, which could be due to sudden genomic instability events, or a change in cancer mutation and/or growth rate over time. These issues would not have been uncovered if using sequencing data from single timepoints.

The average mutation rate in the four CLL patients we analyze is 0.37 mutations/year. This rate is over the exome, which accounts for $\sim 1\%$ of the human genome. Our average estimated mutation rate in CLL exomes is very close to the measured rate of accumulation of mutations in human tissues of 40 mutations per year over the entire genome [42]. The estimated times of appearance of CLL subclones are very long, on the order of 10 years or more. This finding is agreement with results from Gruber et al. [28], who find few new CLL subclones over years to a decade of evolution. We observe that CLL initiation occurred early in most patients, often within the first decade of their lives, consistent with recent work for other cancers as well [20, 43].

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 can accurately estimate the 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 [19, 44, 22], while other studies have given estimates for scaled mutation rates and time of events [37, 29]. 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, and not only at cell divisions, which simplifies derivations and is supported by recent experimental data [42].

For individual CLLs that underwent bulk sequencing at two time points [28], we infer growth rates of individual subclones, mutation rate in the tumor, the times when cancer subclones began growing, and the time between driver mutations and the patient's diagnosis. Our inferences are limited by the relatively low number of mutations present in CLL, as well as sequencing coverage [28]. The accuracy of estimates presented here is expected to be even higher in cancer types with more mutations or with higher sequencing coverage.

Our model and derivations assume a fixed mutation rate u after transformation and fixed growth rates of cancer subclones, similar to other approaches [36, 29, 41]. While this assumption yields reasonable time estimates for three out of the four CLL patients we studied, the time estimates for the fourth patient (Patient 9) are inconsistent with known patient age. This discrepancy is likely due to the change in mutation or growth rates experienced by the CLL of this patient during its evolution. The example of Patient 9 demonstrates the potential weaknesses of assuming fixed mutation and growth rates during tumor evolution, which are only uncovered since we are able to estimate absolute mutation rates and timelines of tumor evolution. In fact, recent sequencing data points to mutational processes that change over time during cancer evolution [45, 21]; incorporating possible changes in the mutation and/or growth rate into the model would require much higher density of sequencing and clinical data [46].

Availability of data and materials

All simulated data generated during this study are included in this published article and its supplementary information files. CLL data analyzed is publicly available in Supplementary Tables from Ref. [28].Code for the simulations will be made available before publication.

Figures

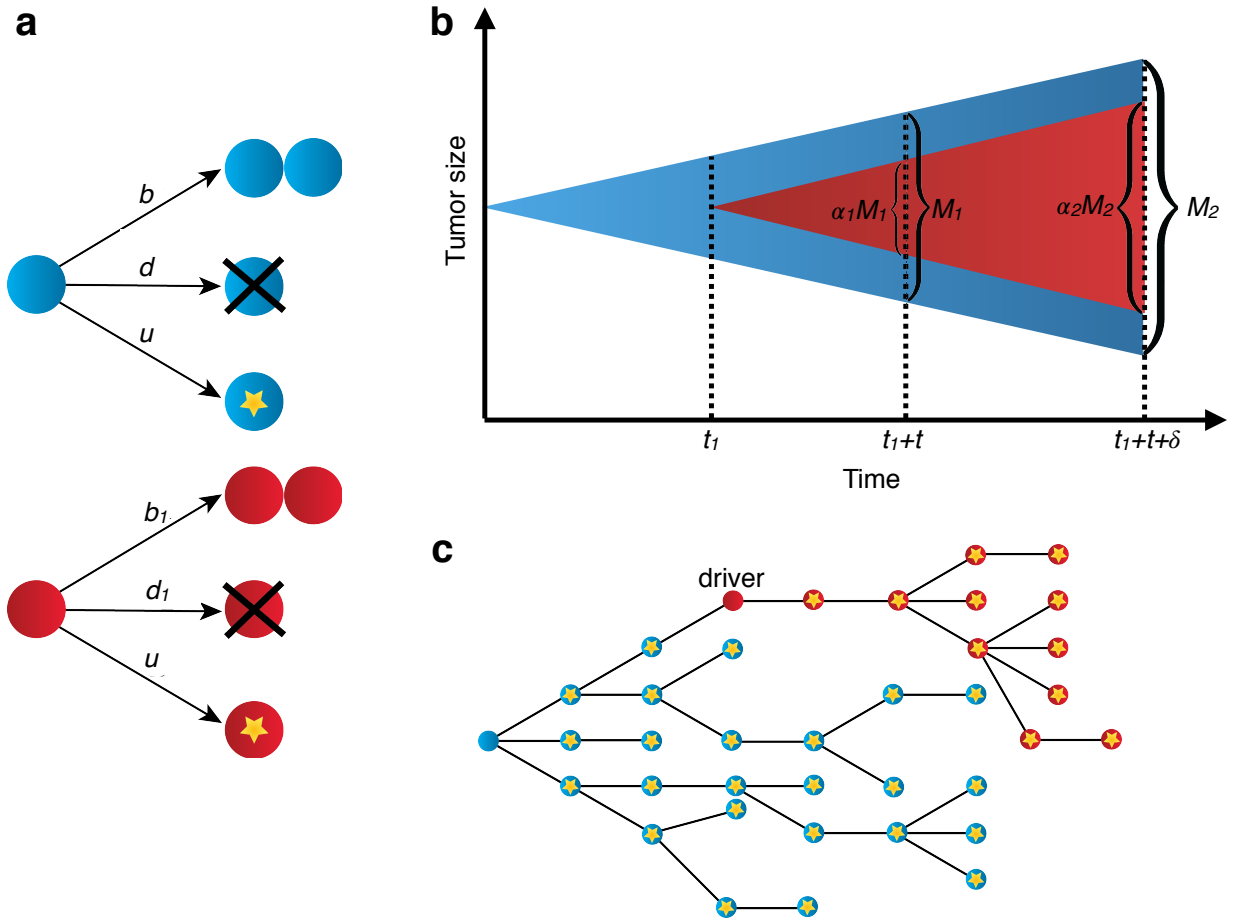

Figure 1: **Stochastic branching process model of tumor evolution.** (a) Stochastic branching process model for tumor expansion. Initiated tumor cells (blue) divide with birth rate b, die with death rate d, and accrue passenger mutations with mutation rate u. Type-1 cells, which carry the driver mutation, 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 is bulk sequenced at times $t_1 + t$ and $t_1 + t + \delta$. (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.

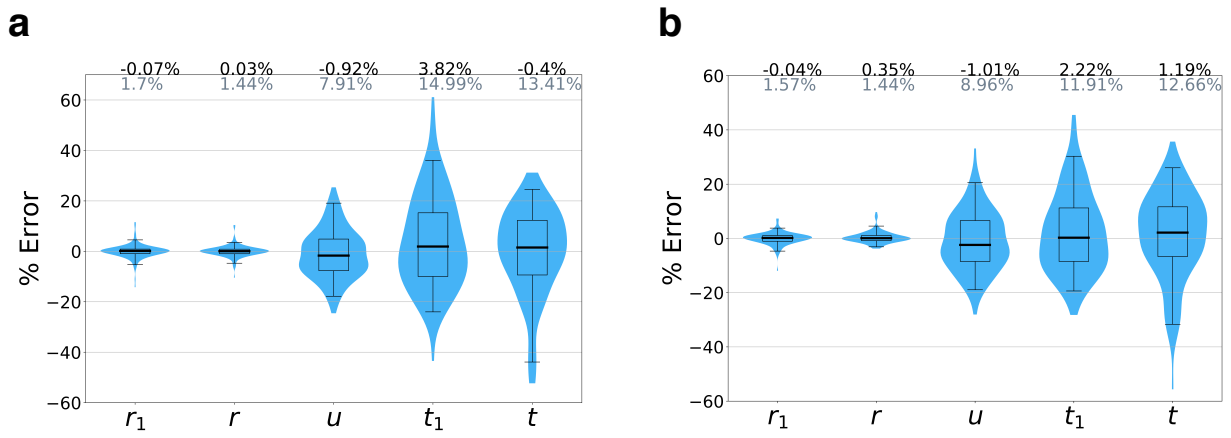

Figure 2: **Accuracy of parameter inferences from simulated data**. We simulated tumor growth by performing a Monte Carlo simulation, which simulates the birth, death, and accumulation of mutations in the individual cells that make up a tumor, and generates the mutation frequency and tumor size data used by the estimates. 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. Ground truth parameter set: $b = b_1 = 0.25$, $d = 0.18$, $d_1 = 0.11$, $t_1 = 70$, $t = 50$, $\delta = 20$, $f_1 = 1\%$, and $f_2 = 20\%$. Mutation rate (a) $u = 1$, (b) $u = 3$. At least 100 Monte Carlo simulation runs with a surviving tumor performed for each parameter combination.

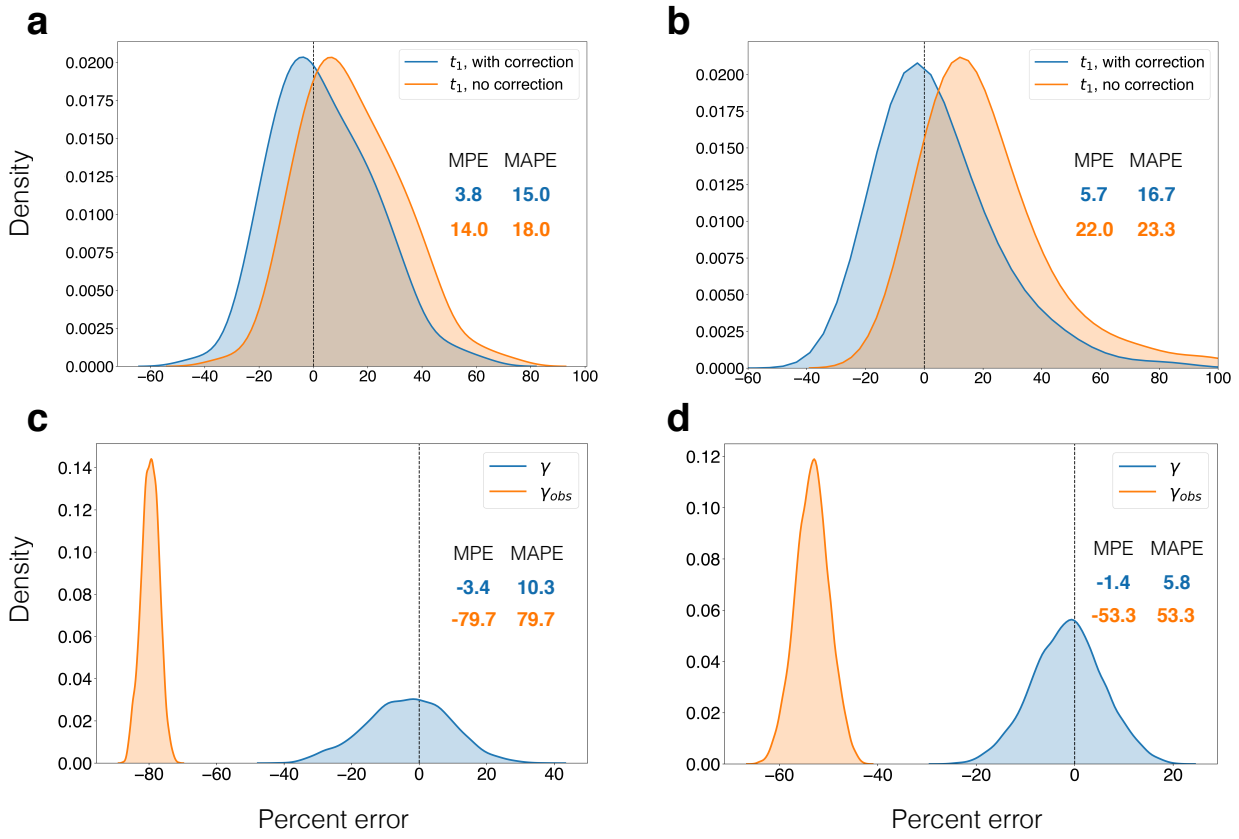

Figure 3: **Corrections for observed mutation counts.** In (a) and (b), we compare percent errors of parameter estimates for time from tumor initiating until appearance of a driver subclone, t_1, with and without the correction for the number of passengers that occurred after t_1 and reached fixation in the type-1 population (Eq. (6)). Errors for estimate with correction (Eq. (12)) 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 (a) fast growing tumor with mutation rate $u = 1$ and (b) slow growing tumor with mutation rate $u = 5$. Mean percent errors (MPEs) and mean absolute percent errors (MAPEs) are listed. In (c) and (d), the percent errors for the observed (orange) and corrected (blue) number of subclonal mutations between frequencies f_1 and f_2, 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 mutatation rate $u = 1$, from which sequencing reads were simulated with (c) 100x and (d) 200x average coverage (see Supplementary Methods). Percent errors are calculated relative to the true γ measured from the true mutation frequencies.

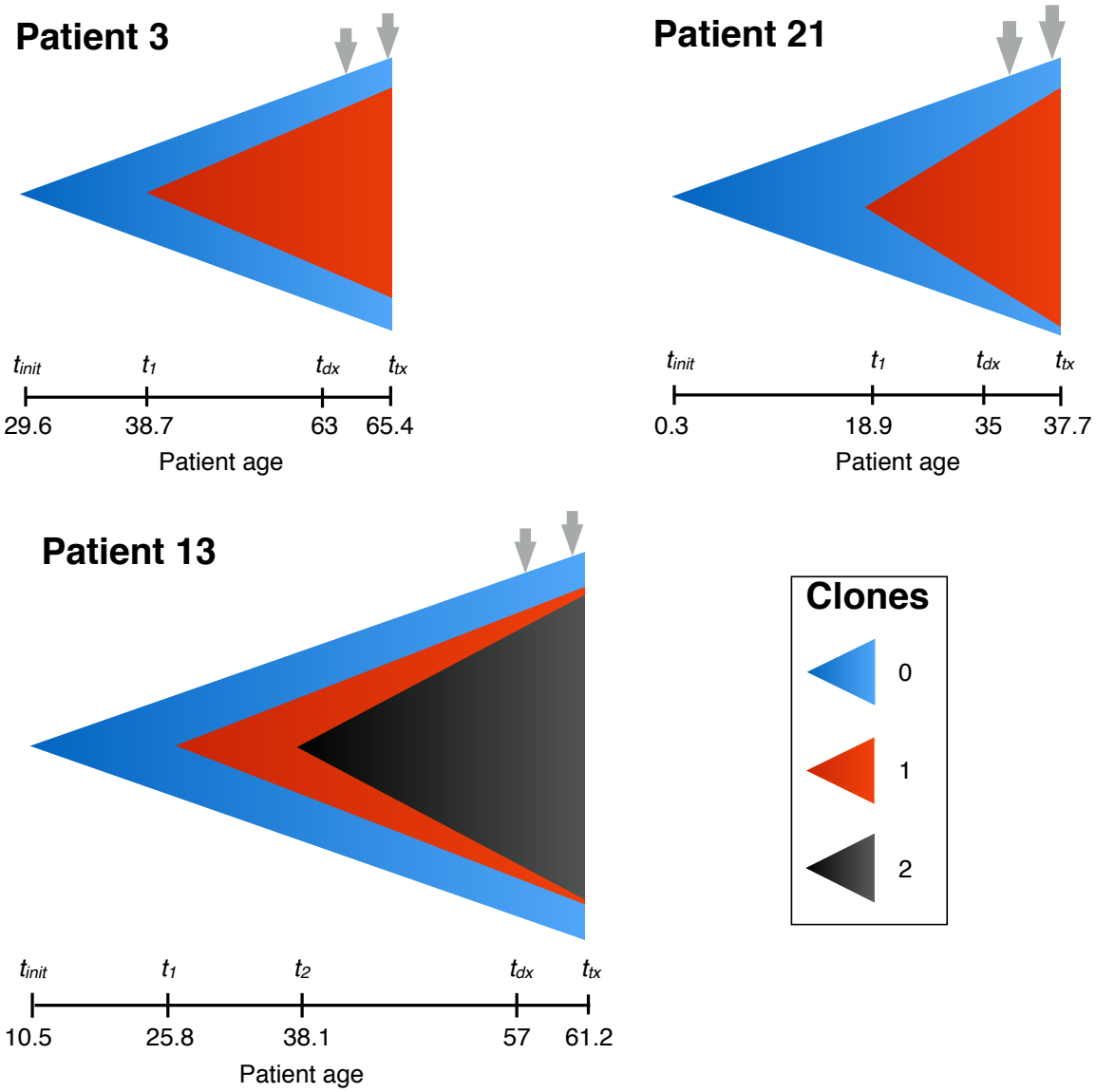

Figure 4: **Reconstructing the timeline of CLL evolution in patients.** We applied our methodology to estimate subclonal growth rates, mutation rates and evolutionary timelines in CLL tumors from Ref. [28]. Horizontal axis represents time and vertical height of a clone represents its size. For each patient, we list estimates for patient age at CLL initiation t_{init}, times of appearance of CLL subclones, patient's age at diagnosis t_{dx} and patient's age when they received treatment t_{tx}. Grey arrows indicate times of the two bulk sequencing measurements.

References

[1] Nowell, P. C. The Clonal Evolution of Tumor Cell Populations. *Science* **194**, 23–28 (1976). URL http://www.jstor.org/stable/1742535.

[2] Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009). URL https://www.nature.com/articles/nature07943.

[3] Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011). URL https://www.cell.com/fulltext/S0092-8674(11)00127-9.

[4] Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature Genetics* **48**, 238–244 (2016). URL https://www.nature.com/articles/ng.3489.

[5] Merlo, L. M. F., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer* **6**, 924–935 (2006). URL https://www.nature.com/articles/nrc2013.

[6] Pepper, J. W., Findlay, C. S., Kassen, R., Spencer, S. L. & Maley, C. C. SYNTHESIS: Cancer research meets evolutionary biology. *Evolutionary Applications* **2**, 62–70 (2009). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1752-4571.2008.00063.x.

[7] Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385.e18 (2018). URL https://www.cell.com/cell/abstract/S0092-8674(18)30237-X.

[8] Kimura, M. Evolutionary Rate at the Molecular Level. *Nature* **217**, 624–626 (1968). URL https://www.nature.com/articles/217624a0.

[9] Kimura, M. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles*. *Genetics Research* **11**, 247–270 (1968). URL http://www.cambridge.org/core/journals/genetics-research/article/genetic-variability-maintained-in-a-finite-population-due-to-mutational-production-of-neutral-and-A74BD3A5D72ED2C52444FD99DFE483EF.

[10] Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genetics Research* **23**, 23–35 (1974). URL http://www.cambridge.org/core/journals/genetics-research/article/hitchhiking-effect-of-a-favourable-gene/918291A3B62BD50E1AE5C1F22165EF1B.

[11] Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics* **20**, 404–416 (2019). URL https://www.nature.com/articles/s41576-019-0114-6.

[12] Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339**, 1546–1558 (2013). URL https://science.sciencemag.org/content/339/6127/1546.

[13] Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013). URL https://www.nature.com/articles/nature12213.

[14] Zuckerkandl, E. & Pauling, L. Evolutionary Divergence and Convergence in Proteins. In Bryson, V. & Vogel, H. J. (eds.) *Evolving Genes and Proteins*, 97–166 (Academic Press, 1965). URL http://www.sciencedirect.com/science/article/pii/B9781483227344500176.

[15] Tsao, J.-L. *et al.* Genetic reconstruction of individual colorectal tumor histories. *Proceedings of the National Academy of Sciences* **97**, 1236–1241 (2000). URL https://www.pnas.org/content/97/3/1236.

[16] Jones, S. *et al.* Comparative lesion sequencing provides insights into tumor evolution. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 4283–4288 (2008). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2393770/.

[17] Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010). URL https://www.nature.com/articles/nature09515.

[18] Naxerova, K. *et al.* Hypermutable DNA chronicles the evolution of human colon cancer. *Proceedings of the National Academy of Sciences* **111**, E1889–E1898 (2014). URL http://www.pnas.org/cgi/doi/10.1073/pnas.1400179111.

[19] McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine* **7**, 283ra54–283ra54 (2015). URL https://stm.sciencemag.org/content/7/283/283ra54.

[20] Mitchell, T. J. *et al.* Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. *Cell* **173**, 611–623.e17 (2018). URL http://www.sciencedirect.com/science/article/pii/S0092867418301648.

[21] PCAWG Evolution & Heterogeneity Working Group *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020). URL http://www.nature.com/articles/s41586-019-1907-7.

[22] Sundermann, L. K., Wintersinger, J., Rätsch, G., Stoye, J. & Morris, Q. Reconstructing tumor evolutionary histories and clone trees in polynomial-time with SubMARine. preprint, Bioinformatics (2020). URL http://biorxiv.org/lookup/doi/10.1101/2020.06.11.146100.

[23] PCAWG Evolution and Heterogeneity Working Group *et al.* Reconstructing evolutionary trajectories of mutation signature activities in cancer using TrackSig. *Nature Communications* **11**, 731 (2020). URL http://www.nature.com/articles/s41467-020-14352-7.

[24] Griffith, M. *et al.* Optimizing Cancer Genome Sequencing and Analysis. *Cell Systems* **1**, 210–223 (2015). URL http://www.sciencedirect.com/science/article/pii/S2405471215001131.

[25] Haber, D. A. & Velculescu, V. E. Blood-Based Analyses of Cancer: Circulating Tumor Cells and Circulating Tumor DNA. *Cancer Discovery* **4**, 650–661 (2014). URL https://cancerdiscovery.aacrjournals.org/content/4/6/650.

[26] Leshchiner, I. *et al.* Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. preprint, Bioinformatics (2018). URL http://biorxiv.org/lookup/doi/10.1101/508127.

[27] Myers, M. A., Satas, G. & Raphael, B. J. CALDER: Inferring Phylogenetic Trees from Longitudinal Tumor Samples. *Cell Systems* **8**, 514–522.e5 (2019). URL http://www.sciencedirect.com/science/article/pii/S2405471219301917.

[28] Gruber, M. *et al.* Growth dynamics in naturally progressing chronic lymphocytic leukaemia. *Nature* **570**, 474–479 (2019). URL https://www.nature.com/articles/s41586-019-1252-x.

[29] Werner, B. *et al.* Measuring single cell divisions in human tissues from multi-region sequencing data. *Nature Communications* **11**, 1–9 (2020). URL https://www.nature.com/articles/s41467-020-14844-6. Number: 1 Publisher: Nature Publishing Group.

[30] Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 18545–18550 (2010). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2972991/.

[31] Sun, R. *et al.* Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nature Genetics* **49**, 1015–1024 (2017). URL https://www.nature.com/articles/ng.3891.

[32] Salichos, L., Meyerson, W., Warrell, J. & Gerstein, M. Estimating growth patterns and driver effects in tumor evolution from individual samples. *Nature Communications* **11**, 1–14 (2020). URL https://www.nature.com/articles/s41467-020-14407-9.

[33] Tavaré, S. The linear birth–death process: an inferential retrospective. *Advances in Applied Probability* **50**, 253–269 (2018). URL https://www.cambridge.org/core/product/identifier/S0001867818000848/type/journal_article.

[34] Roney, J. P., Ferlic, J., Michor, F. & McDonald, T. O. ESTIpop: A computational tool to simulate and estimate parameters for continuous-time Markov branching processes. *Bioinformatics* btaa526 (2020). URL https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa526/5840725.

[35] Noble, R., Burri, D., Kather, J. N. & Beerenwinkel, N. Spatial structure governs the mode of tumour evolution. preprint, Cancer Biology (2019). URL http://biorxiv.org/lookup/doi/10.1101/586735.

[36] Chkhaidze, K. *et al.* Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *PLOS Computational Biology* **15**, e1007243 (2019). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007243.

[37] Williams, M. J. *et al.* Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics* **50**, 895 (2018). URL https://www.nature.com/articles/s41588-018-0128-6.

[38] 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* **105**, 20–23 (1983). URL http://link.springer.com/10.1007/BF00391826.

[39] Durrett, R. Branching Process Models of Cancer. In Durrett, R. (ed.) *Branching Process Models of Cancer*, Mathematical Biosciences Institute Lecture Series, 1–63 (Springer International Publishing, Cham, 2015). URL https://doi.org/10.1007/978-3-319-16065-8_1.

[40] Bozic, I., Gerold, J. M. & Nowak, M. A. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. *PLOS Computational Biology* **12**, e1004731 (2016). URL http://dx.plos.org/10.1371/journal.pcbi.1004731.

[41] Dinh, K. N., Jaksik, R., Kimmel, M., Lambert, A. & Tavaré, S. Statistical Inference for the Evolutionary History of Cancer Genomes. *Statistical Science* **35**, 129–144 (2020). URL https://projecteuclid.org/euclid.ss/1583226033.

[42] Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016). URL http://www.nature.com/articles/nature19768.

[43] Lahouel, K. *et al.* Revisiting the tumorigenesis timeline with a data-driven generative model. *Proceedings of the National Academy of Sciences* **117**, 857–864 (2020). URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1914589117`.

[44] Auslander, N., Wolf, Y. I. & Koonin, E. V. In silico learning of tumor evolution through mutational time series. *Proceedings of the National Academy of Sciences* **116**, 9501–9510 (2019). URL `https://www.pnas.org/content/116/19/9501`.

[45] Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282–1294.e20 (2019). URL `https://linkinghub.elsevier.com/retrieve/pii/S0092867419301618`.

[46] Bozic, I. & Wu, C. J. Delineating the evolutionary dynamics of cancer from theory to reality. *Nature Cancer* **1**, 580–588 (2020). URL `http://www.nature.com/articles/s43018-020-0079-6`.