

# Exact site frequency spectra of neutrally evolving tumors: A transition between power laws reveals a signature of cell viability

Einar Bjarki Gunnarsson<sup>a,\*</sup>, Kevin Leder<sup>a</sup>, Jasmine Foo<sup>b</sup>

<sup>a</sup> Department of Industrial and Systems Engineering, University of Minnesota, Twin Cities, MN 55455, USA

<sup>b</sup> School of Mathematics, University of Minnesota, Twin Cities, MN 55455, USA

## ARTICLE INFO

### Article history:

Received 23 February 2021

Available online 21 September 2021

### Keywords:

Mathematical modeling  
Branching processes  
Exponentially growing populations  
Site frequency spectrum  
Infinite sites model  
Cancer evolution

## ABSTRACT

The site frequency spectrum (SFS) is a popular summary statistic of genomic data. While the SFS of a constant-sized population undergoing neutral mutations has been extensively studied in population genetics, the rapidly growing amount of cancer genomic data has attracted interest in the spectrum of an exponentially growing population. Recent theoretical results have generally dealt with special or limiting cases, such as considering only cells with an infinite line of descent, assuming deterministic tumor growth, or taking large-time or large-population limits. In this work, we derive exact expressions for the expected SFS of a cell population that evolves according to a stochastic branching process, first for cells with an infinite line of descent and then for the total population, evaluated either at a fixed time (fixed-time spectrum) or at the stochastic time at which the population reaches a certain size (fixed-size spectrum). We find that while the rate of mutation scales the SFS of the total population linearly, the rates of cell birth and cell death change the shape of the spectrum at the small-frequency end, inducing a transition between a  $1/j^2$  power-law spectrum and a  $1/j$  spectrum as cell viability decreases. We show that this insight can in principle be used to estimate the ratio between the rate of cell death and cell birth, as well as the mutation rate, using the site frequency spectrum alone. Although the discussion is framed in terms of tumor dynamics, our results apply to any exponentially growing population of individuals undergoing neutral mutations.

© 2021 Elsevier Inc. All rights reserved.

## 1. Introduction

The study of genetic variation driven by neutral mutations has a long history in population genetics (Kimura, 1968). Usually, the population is assumed to be of a large constant size  $N$ , and reproduction follows either the Wright–Fisher model (nonoverlapping generations) or the Moran model (overlapping generations) (Durrett, 2008). Neutral mutations occur at rate  $u$  per individual per time unit, and each new mutation is assumed to be unique (the *infinite-sites model* of Kimura (1969)). This framework gives rise to a sample-based theory of tracing genealogies of extant individuals backwards in time via the *coalescent* (Kingman, 1982a,b). A popular summary statistic of genomic data is the *site frequency spectrum* (SFS), which records the frequencies of mutations in a population or population sample. Under the Moran model with neutral mutations, the expected number of mutations found in  $j$  cells of a sample of size  $n$  is  $\mathbb{E}[\xi_j] = (Nu)(1/j)$  (Durrett, 2008), and any linear combination of the form  $\sum_{j=1}^{n-1} j c_j \xi_j$  with  $\sum_{j=1}^{n-1} c_j = 1$  is an unbiased estimator of  $\theta := Nu$ , the population-scaled mutation

rate (Zeng et al., 2006; Achaz, 2009).<sup>1</sup> Prominent estimators of this form include Watterson's  $\theta_W$  (Watterson, 1975), Tajima's  $\theta_\pi$  (Tajima, 1989), Fu and Li's  $\xi_1$  (Fu and Li, 1993) and Fay and Wu's  $\theta_H$  (Fay and Wu, 2000), and these estimators form the basis of several statistical tests of neutral evolution vs. evolution under selection (Zeng et al., 2006; Achaz, 2009). In this way, the site frequency spectrum has provided a simple means of understanding the evolutionary history of populations from genomic data.

Cancer can be viewed as its own evolutionary process, operating at the somatic level. Cancer initiation is usually understood to be a series of mutational events that culminates in malignant cells able to proliferate uncontrollably (Armitage and Doll, 1954, 1957; Knudson, 1971; Nowell, 1976). Such “driver” mutations are complemented by more frequent neutral or “passenger” mutations (Tomasetti et al., 2013; Bozic et al., 2010), that have no functional role in the evolution to malignancy, but contribute to the genetic diversity characteristic of cancer (Burrell et al., 2013; Vogelstein et al., 2013; McGranahan and Swanton, 2017). The dominant paradigm of tumor progression has been that of sequential clonal expansion of driver mutations. However, several

\* Corresponding author.

E-mail addresses: [gunna042@umn.edu](mailto:gunna042@umn.edu) (E.B. Gunnarsson), [lede0024@umn.edu](mailto:lede0024@umn.edu) (K. Leder), [jyfoo@umn.edu](mailto:jyfoo@umn.edu) (J. Foo).

<sup>1</sup> In Theorem 1.33 of Durrett (2008), the result is given as  $\mathbb{E}[\xi_j] = (2Nu)(1/j)$  for a population of size  $2N$ .

**Table 1**  
Notation used in the paper.

Symbol	Description	Definition
$r_0$	Division rate of tumor cells (per unit time)	Section 2.1
$d_0$	Death rate of tumor cells (per unit time)	Section 2.1
$p_0$	Extinction probability of the tumor (and of a single-cell derived clone)	(1) in Section 2.1
$q_0$	Survival probability of the tumor (and of a single-cell derived clone)	(2) in Section 2.1
$w$	Mutation rate (expected number of mutations per cell division)	Section 2.3
$\tilde{t}_N$	Fixed time at which the skeleton subpopulation has expected size $N$	(10) in Section 3.2
$\tilde{\tau}_N$	Stochastic time at which the skeleton subpopulation first reaches size $N$	(11) in Section 3.2
$t_N$	Fixed time at which a surviving tumor cell population has expected size $N$	(18) in Section 4.1
$\tau_N$	Stochastic time at which the tumor cell population first reaches size $N$	(19) in Section 4.1

recent works suggest that a neutral evolution model, under which all driver mutations are already present in the tumor-initiating cell, is sufficient to explain the intratumoral heterogeneity in many cancers, see e.g. Sottoriva et al. (2015), Ling et al. (2015) and Williams et al. (2016), and the reviews by Venkatesan and Swanton (2016) and Davis et al. (2017). A simple test of neutral tumor evolution based on the site frequency spectrum was proposed in Williams et al. (2016), which has since generated debate e.g. surrounding its significance level and statistical power (see e.g. McDonald et al. (2018), Tarabichi et al. (2018), Bozic et al. (2019), with author responses in Werner et al. (2018), Heide et al. (2018)). The authors of Williams et al. (2016) subsequently suggested a Bayesian framework for detecting tumor subclones evolving under selection (Williams et al., 2018), and more recent approaches to that problem include Dinh et al. (2020) and Caravagna et al. (2020). These works and the surrounding debate are indicative both of the fact that increased attention is being paid to the role of neutral evolution in cancer, and that efforts are just underway to develop robust methods of inferring the evolutionary history of tumors (Turajlic et al., 2019).

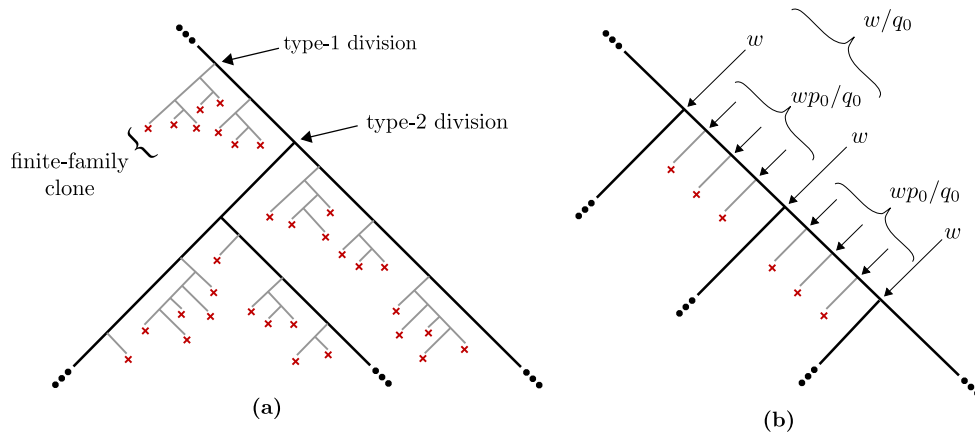
While the constant-sized models of population genetics are appropriate for understanding early cancer development in small tissue compartments, exponential growth models are more relevant for understanding long-run tumor progression (Durrett, 2015; Ohtsuki and Innan, 2017). In this work, we will employ a stochastic branching process model in which tumor cells divide at rate  $r_0$  and die at rate  $d_0$ , with net birth rate  $\lambda_0 := r_0 - d_0 > 0$ , and  $w$  neutral mutations accumulate on average per cell division. Let  $p_0 := d_0/r_0$  be the *extinction probability* of the tumor, and let  $q_0 := 1 - p_0 = \lambda_0/r_0$  be its *survival probability* (Section 2.1). We will show in (14) of Section 3.2 that if we only consider cells with an infinite line of descent, then at the time the number of cells becomes  $N$ , the expected number of mutations found in  $j$  cells is  $\xi_j = (w/q_0)N \cdot 1/(j(j+1))$  for  $2 \leq j \leq N-1$ . This SFS differs from the SFS of the constant-sized Moran model of population genetics in two important ways: The spectrum now follows a  $1/j^2$  power law as opposed to a  $1/j$  law, and it now depends on the growth parameters  $r_0$  and  $d_0$  via the survival probability  $q_0$ . Cumulative versions of the  $1/j^2$  spectrum have previously been established by Durrett (2013, 2015), Bozic et al. (2016) and Williams et al. (2016), as we outline in more detail in Section 6. In addition, Williams et al. (2016), Bozic et al. (2016) and Ling et al. (2015) have used the  $1/j^2$  spectrum to infer the ratio  $w/q_0$  from tumor data, but extracting information about the mutation parameter  $w$  and the growth parameter  $q_0$  separately seemingly requires different tools. In a recent work by Werner et al. (2020), the authors measured pairwise mutational differences between the ancestors of spatially separated tumor bulk samples, and they developed a coalescent-based approach for estimating  $w$  and a function of  $p_0$  given by (5).

Tumor evolution is commonly characterized by low cell viability, i.e. a large extinction probability  $p_0$ . For example, in a modeling study of cancer recurrence, Avanzini and Antal (2019) collected clinical estimates of the tumor volume doubling time

and the time between cell divisions in metastatic breast cancer, colorectal cancer, head & neck cancer, lung cancer and prostate cancer. Based on the collected data, they computed a typical net growth rate ( $\lambda_0$ ) and division rate ( $r_0$ ) for metastases of each cancer type, which lead them to estimate  $p_0 = 1 - \lambda_0/r_0$  as 0.90, 0.97, 0.95, 0.97 and 0.76, respectively. Similarly, in an investigation of targeted combination therapy, Bozic et al. (2013) estimated an average net growth rate of  $\lambda_0 = 0.01$  per day for 21 melanoma lesions, which they combined with a typical division rate of  $r_0 = 0.14$  per day (Rew and Wilson, 2000) to compute a typical death rate of  $d_0 = 0.13$  per day. These estimates suggest a typical extinction probability of  $p_0 = 0.93$  for the melanoma lesions. Finally, Bozic et al. (2016) used their cumulative version of the  $1/j^2$  spectrum to estimate  $w/q_0$  from the SFS of mutations at cell frequency 24%–50% in colorectal cancer.<sup>2</sup> For microsatellite stable (MSS) tumor samples, they combined their median estimate of  $w/q_0$  with an independent estimate of the mutation rate (Jones et al., 2008) to obtain  $p_0$  as 0.997. It should be emphasized that even for a given cancer type, there is substantial heterogeneity between individual tumors (as the clinical data collected in Table 1 of Avanzini and Antal (2019) indicate), so these values should only be taken as rough estimates. However, these simple estimates do suggest that low cell viability is broadly relevant to tumor evolution. Low cell viability, and the corresponding high cell turnover, induces a large mutational burden and high genetic diversity, which enhances the adaptability of the tumor under treatment. It is therefore important to understand how low-viability tumors behave, and to explore how they can potentially be identified from genomic data.

Prior theoretical works on the expected SFS of an exponentially growing tumor population offer only a limited understanding of how  $p_0$  affects the spectrum, as these works generally consider only cells with an infinite line of descent, or they consider special cases such as deterministic growth of the tumor bulk or no cell death, which is reasonable when  $p_0$  is small. Moreover, many prior results are given in the large-time or large-population limit, and for practical reasons, the focus is often on mutations of frequency 10% and higher. These results are discussed in more detail in Section 6. Our goal in this work is to gain a more complete understanding of the SFS of an exponentially growing tumor with neutral mutations. We seek to understand how the spectrum behaves both at small and large frequencies, for all values of  $p_0$ , and for any population size  $N$ . We obtain separate results for cells with an infinite line of descent and for the total population, evaluated either at a fixed time or at the stochastic time at which the population reaches a certain size, each of which is relevant to tumor data analysis depending on the context. We observe that while the SFS of cells with an infinite line of descent

<sup>2</sup> Bozic et al. (2016) estimated  $w/q_0$  from the *variant allele frequency* (VAF) spectrum, which records the proportion of chromosomes carrying the mutations, as opposed to the proportion of cells carrying the mutations. They considered allele frequencies 12%–25%, which translates to cell frequencies 24%–50% in the simplified setting where all cells are diploid and no cell is mutated at the same site on both chromosomes.



**Fig. 1.** Categorization of cells into skeleton cells and finite-family cells, and mutation accumulation on the skeleton. **(a)** The skeleton subpopulation, which consists of cells with an infinite line of descent, can be thought of as forming the trunk and scaffold branches of the genealogical branching tree (bold branches). Each skeleton cell divides into one skeleton cell and one finite-family cell at rate  $2r_0p_0 = 2d_0$  (type-1 skeleton division, see (3) in Section 3.1), in which case a finite-family clone grows out from the skeleton as a lateral branch (light gray branches). A skeleton cell divides into two skeleton cells at rate  $r_0q_0 = \lambda_0$  (type-2 skeleton division, see (4) in Section 3.1), in which case another scaffold branch is added. **(b)** In between two type-2 skeleton divisions, the expected number of mutations that accumulate due to type-1 divisions is  $wp_0/q_0$  by (8). Each type-2 division adds  $w$  mutations on average and starts a new skeleton population size level. Thus, the expected number of mutations that accumulate on the skeleton per size level is  $w + wp_0/q_0 = w/q_0$ .

depends on the mutation rate  $w$  and the extinction probability  $p_0$  only via the ratio  $w/q_0$ , the two parameters decouple in the SFS of the total population. In fact, as  $p_0$  increases from 0 to 1, the small-frequency end of the spectrum transitions from the  $1/j^2$  power law characteristic of pure-birth exponential growth to the  $1/j$  power law characteristic of a constant-sized population. We investigate simple metrics that quantify this transition, and use one of them to propose a simple estimator for  $p_0$ , which we subsequently evaluate using idealized synthetic single-cell sequencing data.

The rest of the paper is organized as follows. In Section 2, we formulate our branching process model with neutral mutations, define the *skeleton subpopulation* of cells with an infinite line of descent, and establish relevant notation. In Section 3, we analyze the SFS of skeleton cells, and in Section 4, we analyze the SFS of the total cell population. In Section 5, we use our theoretical results to propose and evaluate a simple estimator for  $p_0$ . In Section 6, we summarize our results, and discuss in detail how they relate to the existing literature. The proofs of all of our theoretical results are found in appendices at the end of the paper.

## 2. Model description

### 2.1. Branching process dynamics

We assume that the tumor evolution follows a branching process model in continuous time. Cells divide into two cells at rate  $r_0 > 0$  per unit time and die at rate  $d_0 \geq 0$  per unit time, which means that in a small time interval of length  $\Delta t$ , a cell divides with probability  $r_0\Delta t$  and dies with probability  $d_0\Delta t$ . Assume  $r_0 > d_0$  and define  $\lambda_0 := r_0 - d_0 > 0$  as the net growth rate. Let  $Z_0(t)$  denote the size of the tumor population at time  $t$  and assume  $Z_0(0) = 1$ , i.e. the tumor expands from a single tumor-initiating cell. Define

$$\Omega_\infty := \{Z_0(t) > 0 \text{ for all } t > 0\}$$

as the event that the tumor does not go extinct, and

$$p_0 := \mathbb{P}(\Omega_\infty^c) = \mathbb{P}(Z_0(t) = 0 \text{ for some } t > 0) \quad (1)$$

as the *extinction probability* of the tumor. This probability can be computed as  $p_0 = d_0/r_0$  with  $0 \leq p_0 < 1$ , see e.g. Section 3

of Durrett (2015). Note that any clone derived from a single tumor cell gives rise to its own branching process with the same growth parameters  $r_0$  and  $d_0$  and the same extinction probability. We also define

$$q_0 := 1 - p_0 = \lambda_0/r_0 \quad (2)$$

as the *survival probability* of the tumor or of a single-cell derived clone.

### 2.2. Decomposition into skeleton cells and finite-family cells

On the nonextinction event  $\Omega_\infty$ , the cells alive at time  $t > 0$  can be split into two categories, one consisting of cells with an infinite line of descent, i.e. cells that start clones that do not go extinct, and the other consisting of cells whose descendants eventually go extinct. We refer to the former cells as *skeleton cells* and the latter as *finite-family cells*. An arbitrary tumor cell is a skeleton cell with probability  $q_0$ , so in the long run, the proportion of skeleton cells in the population is  $q_0$ . We can think of skeleton cells as forming the trunk and scaffold branches of the genealogical branching tree, with finite-family clones growing out from the skeleton as lateral branches, see Fig. 1a.

### 2.3. Mutation accumulation

We next add neutral mutations under the infinite-sites model. Prior to a cell division, each parental DNA molecule is unwound and separated into two complementary strands. Each parental strand serves as a template for the construction of a new complementary daughter strand. The end result is two copies of the DNA molecule, each consisting of one parental and one daughter strand. Errors in nucleotide pairing during this process can result in one or more point mutations per daughter strand. We assume that these errors amount to  $w/2$  mutations on average per daughter strand, for a total of  $w$  mutations on average per cell division. Note that the only assumptions we make on the distribution of the number of mutations is that it is nonnegative and integer-valued with a finite mean. The point mutation rate has been estimated as  $5 \cdot 10^{-10}$  per base pair per cell division (Jones et al., 2008), and it is commonly higher in cancer due to genomic instability (Burrell et al., 2013). Since the number of base pairs is of order  $10^7$  in the exome and  $10^9$  in the genome, it makes sense to allow  $w$  to be any positive number, i.e.  $w \in (0, \infty)$ . In

many works, the convention is to allow at most one mutation per cell division, introducing a probability  $u \in (0, 1)$  of a new mutation. Since our analysis only depends on the mean number of mutations  $w$  per cell division, it includes this case with  $w := u$ . We assume that the mutation rate is constant throughout tumor evolution, which ignores e.g. the possibility of an elevated mutation rate over time due to genomic instability.

While our focus in this work is on discrete mutation accumulation, we will also present all of our results in terms of continuous mutation accumulation, another common and biologically relevant assumption. In the continuous model, neutral mutations occur at rate  $\nu > 0$  per cell per unit time, at any time throughout the lifetime of the cell. In other words, each cell undergoes a neutral mutation in a small time interval of length  $\Delta t$  with probability  $\nu \Delta t$ . The continuous model differs from the discrete model in that at most one mutation occurs at a time, and this mutation occurs in between cell divisions with probability 1. However, as we will show, the mean behavior of the two models is similar when  $\nu = wr_0$ , since in the discrete model, each cell accumulates  $wr_0 \Delta t$  mutations on average in a small time interval of length  $\Delta t$ .

#### 2.4. Clonal and subclonal mutations

Before proceeding, we need to make a distinction between *clonal* and *subclonal* mutations. A mutation is clonal if it is shared by all tumor cells, while it is subclonal if there is at least one tumor cell without it. As an example, say the tumor-initiating cell divides into two cells, *A* and *B*, and that cell *A* acquires a new mutation. This mutation is initially subclonal. However, if the clone started by cell *B* dies out, the mutation in cell *A* becomes clonal from that point onward. While this example demonstrates how clonal mutations can arise post-tumor-initiation, all mutations that accumulate prior to initiation, as the cancer precursor cell evolves to malignancy, also become clonal. For this reason, the clonal mutations usually tell us more about the events preceding cancer than the dynamics post-initiation, and they can in fact outnumber the subclonal mutations (Tomasetti et al., 2013). Nevertheless, clonal mutations do appear in the SFS of mutations post-initiation, and they play distinct roles in the fixed-time and fixed-size spectrum, which is why we pay them special attention below.

### 3. Site frequency spectrum of skeleton cells

In this section, we establish the expected fixed-time and fixed-size spectrum of skeleton cells. The reason we are interested in analyzing skeleton cells separately is twofold:

- When  $p_0 = 0$  (no cell death), all cells are skeleton cells, and the SFS of the total population is the SFS of the skeleton. More generally, when  $p_0$  is small, the total population spectrum is well-approximated by the simpler skeleton spectrum.
- When the extinction probability  $p_0$  is large, finite-family cells affect the SFS of the total population at the small-frequency end. However, the large-frequency end is still characterized by the skeleton spectrum, as we demonstrate in Section 4.

#### 3.1. Effective rates of cell division and mutation

When the tumor is conditioned on nonextinction, the probability that the tumor-initiating cell divides during the first  $\Delta t$

units of time, and that exactly one of the two daughter cells survives, i.e. starts a clone that does not go extinct, is

$$\begin{aligned} & \mathbb{P}(\text{division in } [0, \Delta t], \text{ one offspring survives}) \\ &= \frac{\mathbb{P}(\Omega_\infty)}{1 - p_0} = \frac{r_0 \Delta t \cdot 2p_0(1 - p_0)}{1 - p_0} = 2r_0 p_0 \Delta t, \end{aligned} \quad (3)$$

and the probability of a division in  $[0, \Delta t]$  where both daughter cells survive is

$$\begin{aligned} & \mathbb{P}(\text{division in } [0, \Delta t], \text{ both offspring survive}) \\ &= \frac{\mathbb{P}(\Omega_\infty)}{1 - p_0} = \frac{r_0 \Delta t \cdot (1 - p_0)^2}{1 - p_0} = r_0 q_0 \Delta t. \end{aligned} \quad (4)$$

Since each skeleton cell starts a clone that does not go extinct, we can conclude that a skeleton cell divides into one skeleton cell and one finite-family cell at rate  $2r_0 p_0 = 2d_0$  per unit time (type-1 skeleton division), and it divides into two skeleton cells at rate  $r_0 q_0 = \lambda_0$  per unit time (type-2 skeleton division). The probability that a skeleton division is type-2 is

$$\kappa_0 := \frac{r_0 q_0}{r_0 q_0 + 2r_0 p_0} = \frac{1 - p_0}{1 + p_0}. \quad (5)$$

This probability can also be computed directly as follows:  $(1 - p_0)^2$  is the probability that both daughter cells survive, and  $1 - p_0^2$  is the probability that at least one of them does, so  $(1 - p_0)^2 / (1 - p_0^2) = (1 - p_0) / (1 + p_0)$  is the probability that a skeleton division is type-2.

Let  $\tilde{Z}_0(t)$  denote the number of skeleton cells at time  $t$ , conditional on the nonextinction event  $\Omega_\infty$ . Since type-1 divisions do not affect the size of the skeleton, we can think of the type-2 divisions as the “effective” divisions. More precisely,  $(\tilde{Z}_0(t))_{t \geq 0}$  is a pure-birth exponential growth process, known as a *Yule process*, with birth rate  $\lambda_0$  and mean size  $\mathbb{E}[\tilde{Z}_0(t)] = e^{\lambda_0 t}$  at time  $t$  (Durrett, 2015; O’Connell, 1993). Type-1 divisions do contribute to neutral mutation accumulation however. Indeed, each type-1 division adds  $w/2$  mutations on average to the skeleton, and each type-2 division adds  $w$  mutations on average. The rate at which mutations accumulate on the skeleton is then

$$(w/2) \cdot 2r_0 p_0 + w \cdot r_0 (1 - p_0) = wr_0 \quad (6)$$

per skeleton cell per unit time, which equals the mutation rate for the original, unconditioned process  $(Z_0(t))_{t \geq 0}$ . The mutation rate per type-2 division, or the *effective* mutation rate, is on the other hand

$$wr_0 / \lambda_0 = w / q_0 \quad (7)$$

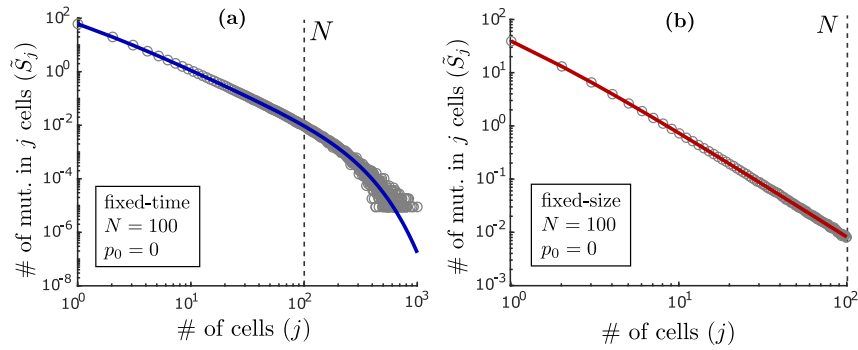
per unit time. We can also think of mutations as accumulating across skeleton population size levels as follows. A type-2 division increases the size of the skeleton population by one, and it adds  $w$  mutations on average. Upon the type-2 division, the number of type-1 divisions before the next type-2 division has the geometric distribution with support  $\{0, 1, 2, \dots\}$  and success probability  $\kappa_0$  given by (5). It follows that in between the two type-2 divisions, the expected number of mutations that accumulate on the skeleton is

$$(w/2) \cdot (1/\kappa_0 - 1) = (w/2) \cdot 2p_0/q_0 = wp_0/q_0. \quad (8)$$

At each population size level, the skeleton therefore accumulates  $w$  mutations on average due to the type-2 division that starts the level, and  $wp_0/q_0$  mutations on average due to type-1 divisions that occur before the next type-2 division that changes levels. We thus obtain

$$w + wp_0/q_0 = w/q_0 \quad (9)$$





**Fig. 2.** Comparison between the expected SFS of the skeleton, as derived in Proposition 1, and simulation results. (a) The expected fixed-time spectrum (12) of Proposition 1 (solid blue line) shows good agreement with the average spectrum of simulated tumors (gray dots). In this example, the extinction probability is  $p_0 = 0$ , the mutation rate is  $w = 1$ , and the expected size of the skeleton is  $N = 100$ , which is also the expected size of the tumor since  $p_0 = 0$ . We generated  $10^5$  tumors with  $p_0 = 0$  and  $w = 1$  and stopped each simulation at the fixed time  $\tilde{t}_N$  with  $N = 100$  as defined by (10). At each cell division, the number of mutations acquired by each daughter cell was generated as a Poisson random variable with mean  $w/2$ . (b) The expected fixed-size spectrum (14) of Proposition 1 (solid red line) shows good agreement with the average spectrum of simulated tumors (gray dots). We again generated  $10^5$  tumors with  $p_0 = 0$  and  $w = 1$ , but this time, we stopped each simulation when the tumor reached size  $N = 100$ , i.e. at the stochastic time  $\tilde{\tau}_N$  defined by (11). The fundamental difference between the fixed-time and fixed-size spectrum is that the skeleton size at time  $\tilde{t}_N$  is variable, while it is always  $N$  at time  $\tilde{\tau}_N$ . As a result, the fixed-size spectrum is restricted to  $j = 1, \dots, N$ , while the fixed-time spectrum has nonzero mass at values  $j > N$ .

mutations per level (Fig. 1b). The effective mutation rate  $w/q_0$  plays a key role in the SFS of the skeleton, with the continuous-time viewpoint in (7) applying to the fixed-time spectrum, and the population-size-level viewpoint in (9) applying to the fixed-size spectrum.

### 3.2. Expected fixed-time and fixed-size skeleton spectrum

Let  $\tilde{S}_j(t)$  denote the number of mutations that are found in  $j \geq 1$  skeleton cells at time  $t$ , conditional on the nonextinction event  $\Omega_\infty$ . This is the site frequency spectrum of skeleton cells. For any integer  $N \geq 1$ , define

$$\tilde{t}_N := \log(N)/\lambda_0 \quad (10)$$

as the (fixed) time at which the skeleton has expected size  $N$ , i.e.  $e^{\lambda_0 \tilde{t}_N} = N$ , and define

$$\tilde{\tau}_N := \inf\{t \geq 0 : \tilde{Z}_0(t) = N\} \quad (11)$$

as the (stochastic) time at which the skeleton reaches size  $N$ . In Proposition 1, we provide the expected SFS of the skeleton evaluated both at time  $\tilde{t}_N$  (fixed-time spectrum) and at time  $\tilde{\tau}_N$  (fixed-size spectrum). Both the fixed-time and fixed-size spectrum can be relevant to tumor data analysis depending on the context. For example, *in vitro* cell culture experiments and *in vivo* mouse experiments are often conducted over a fixed time period, in which case the fixed-time spectrum would apply. In the clinic, however, the size of a tumor sample is more readily estimated than its age, in which case the fixed-size spectrum is more relevant (Komarova et al., 2007). It is therefore useful to understand both spectra and to what extent they differ.

#### Proposition 1.

(1) Define  $\tilde{t}_N$  as in (10). Then, for any  $N \geq 1$  and any  $j \geq 1$ ,

$$\begin{aligned} \mathbb{E}[\tilde{S}_j(\tilde{t}_N)] &= (w/q_0)N \cdot \int_0^{1-1/N} (1-y)^{j-1} dy \\ &= (w/q_0)N \cdot \left(1 - \frac{1}{N}\right)^j \left(\frac{1}{j(j+1)} + \frac{1}{N} \frac{1}{j+1}\right). \end{aligned} \quad (12)$$

For fixed  $j \geq 1$ , then as  $N \rightarrow \infty$ ,

$$\mathbb{E}[\tilde{S}_j(\tilde{t}_N)] \sim (w/q_0)N \cdot 1/(j(j+1)), \quad (13)$$

where  $f(y) \sim g(y)$  as  $y \rightarrow \infty$  means  $\lim_{y \rightarrow \infty} f(y)/g(y) = 1$ .

(2) Define  $\tilde{\tau}_N$  as in (11). Then, for any  $N \geq 2$ ,

$$\mathbb{E}[\tilde{S}_j(\tilde{\tau}_N)]$$

$$= \begin{cases} (w/q_0)N \cdot 1/(j(j+1)) - (wp_0/q_0)\delta_{1,j}, & 1 \leq j \leq N-1, \\ wp_0/q_0 = w/q_0 - w, & j = N, \end{cases} \quad (14)$$

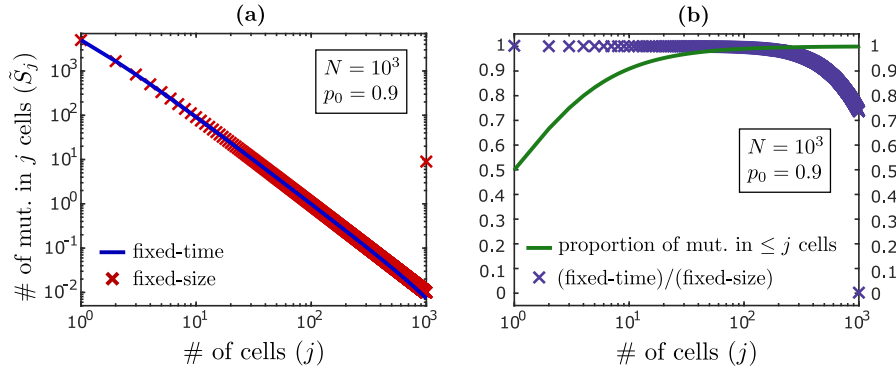
where  $\delta_{\ell,m} = 1$  if  $\ell = m$  and  $\delta_{\ell,m} = 0$  otherwise.

**Proof.** Appendix A.  $\square$

Analogous results for continuous mutation accumulation are presented in Appendix C. In Fig. 2, we compare our fixed-time (12) and fixed-size (14) results with simulation results for  $w = 1$ ,  $p_0 = 0$  and  $N = 100$ . In this example, there are no clonal mutations, since  $p_0 = 0$ . The fundamental difference between the fixed-time and fixed-size spectrum is that the skeleton size at time  $\tilde{t}_N$  is variable, while it is always  $N$  at time  $\tilde{\tau}_N$ . The fixed-time spectrum therefore has nonzero mass at  $j > N$ , due to instances in which the skeleton is larger than  $N$  at time  $\tilde{t}_N$ . It is however natural to ask how the fixed-time spectrum restricted to  $j = 1, \dots, N$  relates to the fixed-size spectrum. By (14), the fixed-size spectrum follows the power law  $(w/q_0)N \cdot 1/(j(j+1))$  exactly on  $j = 2, \dots, N-1$ , and asymptotically as  $N \rightarrow \infty$  for  $j = 1$ . By (13), the fixed-time spectrum converges to the same power law for fixed  $j \geq 1$  as  $N \rightarrow \infty$ , which means that it follows this power law when  $N$  is large and  $j \ll N$ . In Fig. 3, we compare the fixed-time (12) and fixed-size (14) spectrum for  $N = 10^3$  and  $p_0 = 0.9$ . As expected, the two spectra agree on  $j \ll N$ , while the fixed-time spectrum deviates from the fixed-size spectrum at the very largest frequencies (for  $j$  of order  $N$ ). In Fig. 3b, we show that almost all mutations are found on  $j \ll N$ . In this example, the difference between the two spectra is within 1% on the range  $j = 1, \dots, 150$ , on which 99.3% of mutations are found.

We next note the sharp discontinuity at  $j = N$  in the fixed-size spectrum of Fig. 3a, which does not appear in the fixed-time spectrum. This is due to the distinct ways in which clonal mutations manifest in the two spectra. In the fixed-time spectrum, clonal mutations can appear at any value of  $j$ , depending on the skeleton size at time  $\tilde{t}_N$ . In the fixed-size spectrum, all these mutations are concentrated at  $j = N$ , which creates a significant point mass at  $j = N$ . Note that by (14), the expected number of mutations found in  $j = N-1$  skeleton cells is given by  $(w/q_0)(1/(N-1))$  in the fixed-size spectrum, which is of order  $1/N$  as  $N \rightarrow \infty$ , while clonal mutations are given by the constant  $wp_0/q_0$ , independently of  $N$ .

The observed difference between the fixed-time and fixed-size spectrum at the very largest frequencies reflects the unbounded



**Fig. 3.** Comparison between the expected fixed-time spectrum (12) and fixed-size spectrum (14) of the skeleton as derived in Proposition 1. (a) The fixed-time spectrum (12) of the skeleton (blue curve) is a good approximation of the fixed-size spectrum (14) (red crosses) for mutations at frequencies  $j \ll N$ , given parameters  $N = 10^3$ ,  $p_0 = 0.9$  and  $w = 1$ . The two spectra diverge at the very largest frequencies, and the difference is substantial at  $j = N$ , since clonal mutations are concentrated at  $j = N$  in the fixed-size spectrum, while they are scattered in the fixed-time spectrum. (b) Here, we show the ratio of the fixed-time spectrum to the fixed-size spectrum (purple crosses), and the proportion of mutations found in  $\leq j$  skeleton cells in the fixed-size spectrum (green curve). The fixed-time and fixed-size spectrum are virtually the same on  $j \ll N$ , where almost all mutations are found.

range of the fixed-time spectrum. When  $N$  is large, this difference can be alleviated by computing the SFS of mutations found in a given *proportion* of cells (as opposed to a given *number* of cells), as this normalizes both spectra to the frequency range  $[0, 1]$ . To justify this claim, we need to combine our results with results previously obtained by Durrett (2013, 2015) and Bozic et al. (2016), as we discuss in detail in Section 6.

### 3.3. Proportion of mutations found in one cell

We conclude this section by computing two simple and important metrics derived from the site frequency spectrum of the skeleton. Note first that by Proposition 1, the expected number of mutations found in a single skeleton cell is

$$\begin{aligned} \mathbb{E}[\tilde{S}_1(\tilde{\tau}_N)] &= (1/2)(w/q_0)(N^2 - 1)/N \sim (1/2)(w/q_0)N, \\ \mathbb{E}[\tilde{S}_1(\tilde{\tau}_N)] &= (1/2)(w/q_0)N - wp_0/q_0 \sim (1/2)(w/q_0)N, \end{aligned} \quad (15)$$

as  $N \rightarrow \infty$ . Next, let  $\tilde{M}_j(t) := \sum_{k \geq j} \tilde{S}_k(t)$  denote the number of mutations found in  $\geq j$  skeleton cells at time  $t$ . Again, by Proposition 1, the expected total number of mutations on the skeleton is

$$\begin{aligned} \mathbb{E}[\tilde{M}_1(\tilde{\tau}_N)] &= (w/q_0)(N - 1) \sim (w/q_0)N, \\ \mathbb{E}[\tilde{M}_1(\tilde{\tau}_N)] &= (w/q_0)(N - 1) \sim (w/q_0)N, \end{aligned} \quad (16)$$

as  $N \rightarrow \infty$ . Expressions (15) and (16) suggest that for large  $N$ , half the mutations discovered at time  $\tilde{\tau}_N$  or  $\tilde{\tau}_N$  are found in only one cell. This is a consequence of the pure-birth exponential growth of the skeleton. Indeed, note that if we only consider the effective type-2 skeleton divisions, the total number of divisions required to reach generation  $k$  is  $\sum_{j=0}^{k-1} 2^j = 2^k - 1$ . The expected total number of mutations in generation  $k$  is then  $(w/q_0)(2^k - 1)$ , which is (16) with  $N = 2^k$  the number of cells in generation  $k$ . An additional  $2^k$  divisions are required to reach generation  $k + 1$ , so each generation roughly doubles the total number of mutations. Of course, our model is stochastic, it operates in continuous time, and generations may overlap, but this simple discrete argument gives intuition as to why half the mutations are found in one cell, and more generally why most mutations are found at the smallest frequencies.

## 4. Site frequency spectrum of total population and transition between power laws

When the extinction probability  $p_0$  is small, the SFS of the total population  $(Z_0(t))_{t \geq 0}$  is well-approximated by the SFS of the

skeleton  $(\tilde{Z}_0(t))_{t \geq 0}$ . However, tumor evolution is commonly characterized by a large extinction probability, as was discussed in the introduction. In this section, we investigate the expected fixed-time and fixed-size spectrum of the total population  $(Z_0(t))_{t \geq 0}$  for all values of  $p_0$ . We show that as  $p_0$  increases, the small-frequency end of the spectrum starts to deviate from the skeleton spectrum of Section 3, and that as  $p_0$  approaches 1, it transitions to the spectrum of a constant-sized population.

### 4.1. Expected fixed-time and fixed-size total population spectrum

Let  $S_j(t)$  denote the number of mutations found in  $j \geq 1$  cells at time  $t$ , the site frequency spectrum of the total cell population. We wish to compute the mean of  $S_j(t)$  conditioned on the tumor surviving to time  $t$ . We can compute the probability of this survival event as

$$\mathbb{P}(Z_0(t) > 0) = q_0 e^{\lambda_0 t} / (e^{\lambda_0 t} - p_0), \quad t \geq 0,$$

see (B.2) of Appendix B, and the expected size of a tumor that survives to time  $t$  as

$$\mathbb{E}[Z_0(t) | Z_0(t) > 0] = (e^{\lambda_0 t} - p_0) / q_0, \quad t \geq 0. \quad (17)$$

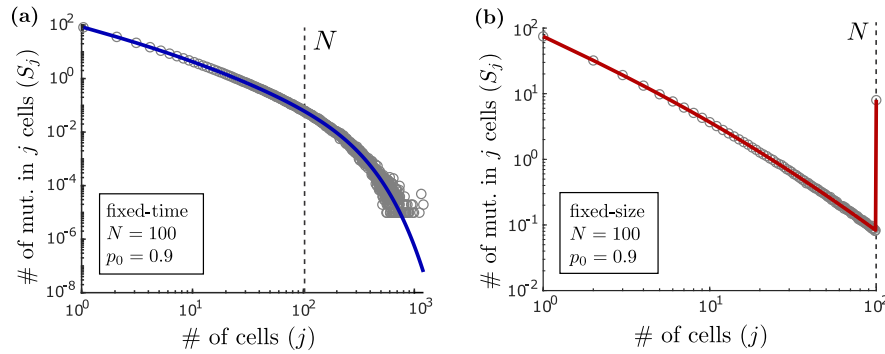
Note that  $\mathbb{E}[Z_0(t) | Z_0(t) > 0] \sim e^{\lambda_0 t} / q_0$  as  $t \rightarrow \infty$ , which means that the long-run expected growth of a tumor conditioned on survival is exponential, and the initial value of the exponential growth function is given by  $1/q_0$ . An interesting consequence of the conditioning on survival is that if  $r_0$  and  $d_0$  are increased by the same amount (so that  $\lambda_0 = r_0 - d_0$  stays fixed), the survival probability  $q_0 = \lambda_0 / r_0$  will decrease, while the size of a tumor conditioned on survival will increase. Now, for any integer  $N \geq 1$ , define

$$t_N := \log(q_0 N + p_0) / \lambda_0 \quad (18)$$

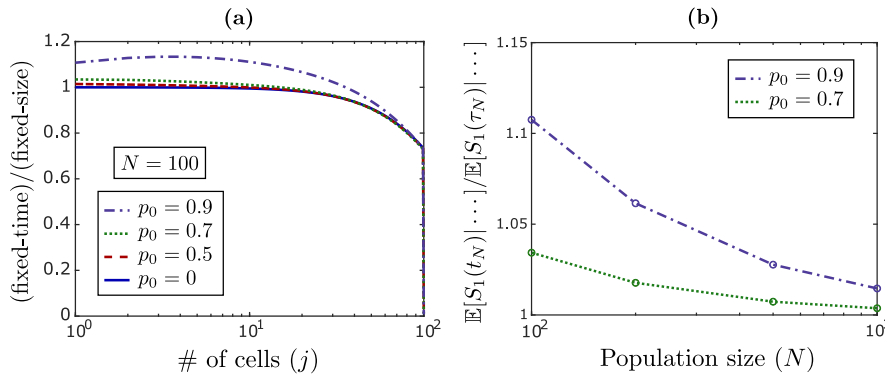
as the (fixed) time at which a surviving tumor has expected size  $N$ , i.e.  $(e^{\lambda_0 t_N} - p_0) / q_0 = N$ , and define

$$\tau_N := \inf\{t \geq 0 : Z_0(t) = N\} \quad (19)$$

as the (stochastic) time at which the tumor reaches size  $N$ , with  $\inf \emptyset = \infty$ . In Proposition 2, we provide the expected SFS of  $(Z_0(t))_{t \geq 0}$  at time  $t_N$  and  $\tau_N$ , conditioned on the survival events  $\{Z_0(t_N) > 0\}$  and  $\{\tau_N < \infty\}$ , respectively. In this case, we cannot obtain an explicit expression for the fixed-size spectrum, and provide instead a computational expression. Note however that for the case  $p_0 = 0$  of no cell death, the fixed-size spectrum follows the power law  $wN \cdot 1/(j(j+1))$  exactly on  $j = 1, \dots, N-1$ , by (14) of Proposition 1.



**Fig. 4.** Comparison between the expected SFS of the total population, as derived in Proposition 2, and simulation results. (a) The expected fixed-time spectrum (20) of Proposition 2 (solid blue line) shows good agreement with the average spectrum of simulated tumors (gray dots). In this example, the extinction probability is  $p_0 = 0.9$ , the mutation rate is  $w = 1$ , and the expected tumor size is  $N = 100$ . We generated  $10^5$  tumors with  $p_0 = 0.9$  and  $w = 1$  and stopped each simulation at the fixed time  $t_N$  with  $N = 100$  as defined by (18). At each cell division, the number of mutations acquired by each daughter cell was generated as a Poisson random variable with mean  $w/2$ . (b) The expected fixed-size spectrum (22) of Proposition 2 (solid red line) shows good agreement with the average spectrum of simulated tumors (gray dots). We again generated  $10^5$  tumors with  $p_0 = 0.9$  and  $w = 1$ , but this time, we stopped each simulation when the tumor reached size  $N = 100$ , i.e. at the stochastic time  $\tau_N$  defined by (19). Note the discontinuity in the fixed-size spectrum at  $j = N$ , which is due to clonal mutations.



**Fig. 5.** Comparison between the expected fixed-time spectrum (20) and fixed-size spectrum (22) of the total population as derived in Proposition 2. (a) Here, we show the ratio between the fixed-time spectrum (20) and the fixed-size spectrum (22) for  $N = 100$ ,  $p_0 \in \{0, 0.5, 0.7, 0.9\}$  and  $w = 1$ . The fixed-time spectrum is a good approximation of the fixed-size spectrum on  $j \ll N$  for smaller values of  $p_0$ , but the two spectra start to diverge as  $p_0$  increases. (b) Here, we show the ratio between the expected number of mutations found in one cell for the fixed-time spectrum and the fixed-size spectrum, as a function of the population size  $N$ , for  $p_0 \in \{0.7, 0.9\}$  and  $w = 1$ . As  $N$  increases, the difference between the two spectra reduces both for  $p_0 = 0.7$  and  $p_0 = 0.9$ .

### Proposition 2.

(1) Define  $t_N$  as in (18). Then, for any  $N \geq 1$  and any  $j \geq 1$ ,

$$\mathbb{E}[S_j(t_N) | Z_0(t_N) > 0] = wN \int_0^{1-1/N} (1 - p_0 y)^{-1} (1 - y) y^{j-1} dy. \quad (20)$$

For fixed  $j \geq 1$ , then as  $N \rightarrow \infty$ ,

$$\begin{aligned} \mathbb{E}[S_j(t_N) | Z_0(t_N) > 0] &\sim wN \cdot \int_0^1 (1 - p_0 y)^{-1} (1 - y) y^{j-1} dy \\ &= wN \cdot \sum_{k=0}^{\infty} \frac{p_0^k}{(j+k)(j+k+1)}, \end{aligned} \quad (21)$$

where  $f(y) \sim g(y)$  as  $y \rightarrow \infty$  means  $\lim_{y \rightarrow \infty} f(y)/g(y) = 1$ .

(2) Define  $\tau_N$  as in (19), let  $\mathcal{S} := \{(\ell, m) : \ell, m \geq 0 \text{ and } \ell + m \leq N\}$  and  $\mathcal{A} := \{(0, 0)\} \cup \{(r, s) : r, s \geq 0 \text{ and } r + s = N\}$ . Then, for any  $N \geq 2$  and any  $1 \leq j \leq N$ ,

$$\mathbb{E}[S_j(\tau_N) | \tau_N < \infty] = (w/q_0) \cdot \sum_{k=1}^{N-1} (1 - p_0^{N-k}) \cdot h_{(1,k)}^{(j, N-j)}, \quad (22)$$

where for each  $(r, s) \in \mathcal{A}$ , the vector  $(h_{(\ell, m)}^{(r, s)})_{(\ell, m) \in \mathcal{S}}$  solves the system

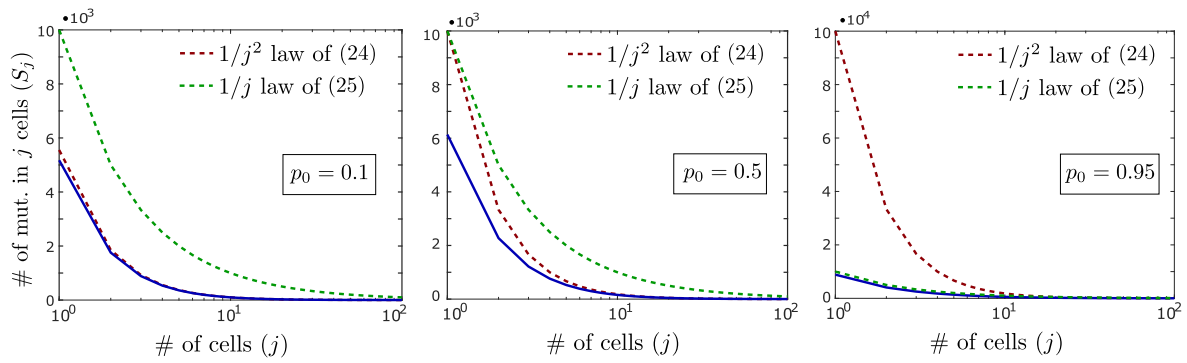
$$\begin{aligned} (\ell + m)(1 + p_0)h_{(\ell, m)}^{(r, s)} \\ = \ell h_{(\ell+1, m)}^{(r, s)} + \ell p_0 h_{(\ell-1, m)}^{(r, s)} + m h_{(\ell, m+1)}^{(r, s)} + m p_0 h_{(\ell, m-1)}^{(r, s)} \end{aligned} \quad (23)$$

for  $\ell, m \geq 1$  and  $\ell + m < N$ , with boundary conditions given by (B.10) in Appendix B.

**Proof.** Appendix B.  $\square$

Analogous results for continuous mutation accumulation are presented in Appendix C. In Fig. 4, we compare our fixed-time (20) and fixed-size (22) results with simulation results for  $w = 1$ ,  $p_0 = 0.9$  and  $N = 100$ . The fundamental difference between the fixed-time and fixed-size spectrum is the same as we observed in Section 3.2. In Fig. 5a, we compare the two spectra in more detail for  $N = 100$  and  $p_0 \in \{0, 0.5, 0.7, 0.9\}$ . The fixed-time spectrum is a good approximation of the fixed-size spectrum on  $j \ll N$  for all but  $p_0 = 0.9$ , in which case there is a significant difference even on  $j \ll N$ . In Fig. 5b, we show that this difference reduces as  $N$  increases, with the expected number of mutations found in one cell being 1.46% off for  $p = 0.9$  and  $N = 1000$ . The number of cells in  $1 \text{ cm}^3$  of tumor tissue is around  $10^7$ – $10^9$  (Del Monte, 2009), in which case the fixed-time spectrum can generally be expected to be a good approximation of the fixed-size spectrum on  $j \ll N$ . The two spectra may diverge, however, when applied to smaller tumor samples (e.g.  $1 \text{ mm}^3$  or smaller) or when  $p_0$  is very close to 1.

In Appendix E, we present heuristic calculations that indicate how the fixed-size spectrum can be approximated by the fixed-time spectrum on  $j \ll N$  as  $N \rightarrow \infty$ . The key insight is that once



**Fig. 6.** The small-frequency end of the total population spectrum transitions between two power laws as  $p_0$  increases from 0 to 1. When  $p_0$  is small, the expected fixed-time spectrum (20) of Proposition 2 (solid blue line) approximately follows the  $1/j^2$  law of (24) (dotted red line). As  $p_0$  increases to 1, the small-frequency end of the spectrum transitions to the  $1/j$  law of (25) (dotted green line), while the  $1/j^2$  law increases according to  $(w/q_0)N \cdot 1/(j(j+1))$ . Note that the y-axis is on a linear scale, and that the scale of the rightmost panel is ten times larger than the scale of the other two. Also note that the  $1/j$  law of (25) is fixed as a function of  $p_0$  and is therefore the same curve in all panels. Parameters are  $N = 10^4$  and  $w = 1$ , and the spectrum is shown only at the smallest frequencies  $j = 1, \dots, 100$ .

the tumor has reached a large size, its growth becomes essentially deterministic with exponential rate  $\lambda_0$ , which allows us to approximate the probabilities  $h_{(1,k)}^{(j,N-j)}$  in (22) by continuous-time probabilities following Iwasa et al. (2006). For added intuition, we refer to our discussion in Section 4.5 below and accompanying calculations in Appendix H, where we conjecture a law of large numbers for the fixed-time and fixed-size spectrum, whose limits agree in the mean with the asymptotic expected fixed-time spectrum (21).

For each  $j = 1, \dots, N$ , computing  $\mathbb{E}[S_j(\tau_N) | \tau_N < \infty]$  according to (22) requires solving a linear system of the form (23), which has order  $N^2/2$  equations. A more general version of this system arises in the study of the number of wild-type and mutant cells under a two-type (Luria-Delbrück) population model stopped at a certain size, see e.g. Komarova et al. (2007). The coefficient matrix of the system is sparse and banded, and it has a certain structure which allows one to solve it in  $O(N^3)$  arithmetic operations (George, 1973), compared to  $O((N^2)^3) = O(N^6)$  for Gaussian elimination. Solving the system directly is still not computationally feasible for the largest values of  $N$ , in which case one must develop an approximate solution. See Komarova et al. (2007) for a partial differential equations approach. Our observations suggest another simple approach: For  $N$  sufficiently large, one can approximate the fixed-size spectrum by the fixed-time spectrum on  $j \ll N$ . For  $j$  of order  $N$ , one can apply the  $1/j^2$  skeleton law of (14), since as we discuss next, the skeleton will be responsible for the large-frequency mutations.

#### 4.2. Transition between power laws

By Proposition 1 of Section 3.2, the SFS of the skeleton depends on the mutation rate  $w$  and the extinction probability  $p_0$  only through the effective mutation rate  $w/q_0$ . By Proposition 2, however, the two parameters decouple in the SFS of the total population. The mutation rate  $w$  scales the total population spectrum linearly, and the same is true of a large population size  $N$ , but the dependence on  $p_0$  is more complex. To better understand how  $p_0$  affects the spectrum, recall first that by (21), the asymptotic expected fixed-time spectrum is given by

$$\mathbb{E}[S_j(t_N) | Z_0(t_N) > 0] \sim wN \cdot \sum_{k=0}^{\infty} \frac{p_0^k}{(j+k)(j+k+1)}, \quad N \rightarrow \infty.$$

In Appendix F, we show that for fixed  $0 < p_0 < 1$ , sending  $j \rightarrow \infty$  in this expression yields

$$wN \cdot \sum_{k=0}^{\infty} \frac{p_0^k}{(j+k)(j+k+1)} \sim (w/q_0)N \cdot 1/(j(j+1)), \quad j \rightarrow \infty, \quad (24)$$

which is the  $1/j^2$  power law given in (13)–(14) of Proposition 1. We also show that for fixed  $j \geq 1$ , sending  $p_0 \rightarrow 1$  in the same expression yields

$$wN \cdot \sum_{k=0}^{\infty} \frac{p_0^k}{(j+k)(j+k+1)} \sim wN \cdot 1/j, \quad p_0 \rightarrow 1, \quad (25)$$

which is the  $1/j$  power law of the constant-sized Moran model of population genetics, see Theorem 1.33 of Durrett (2008). When referring to results from Durrett (2008), note that there, the population is assumed to have size  $2N$ , which is a common convention in population genetics.

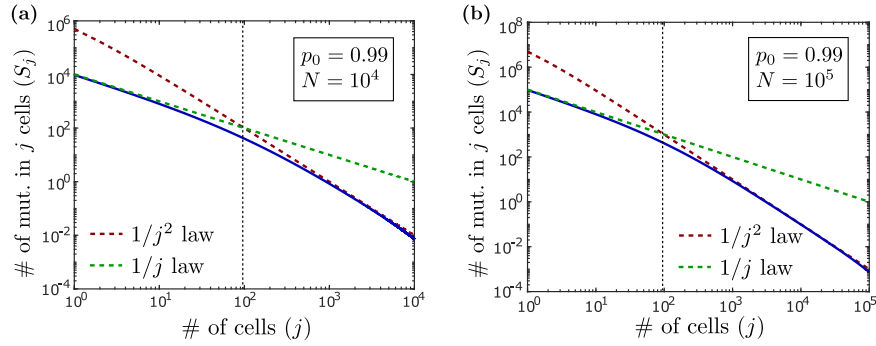
Recall that when  $p_0 = 0$ , the SFS of the total population  $(Z_0(t))_{t \geq 0}$  is the SFS of the skeleton  $(\tilde{Z}_0(t))_{t \geq 0}$ . Expressions (24) and (25) suggest that when  $p_0 > 0$ , the SFS of  $(Z_0(t))_{t \geq 0}$  continues to follow the  $1/j^2$  skeleton law at the large-frequency end, while a deviation starts to occur at the small-frequency end. In fact, as  $p_0$  approaches 1, the small-frequency end transitions to the  $1/j$  law of the constant-sized Moran model. This is illustrated in Fig. 6 for  $p_0 \in \{0.1, 0.5, 0.95\}$  and  $N = 10^4$ . Note that the y-axis is on a linear scale, and that the scale of the rightmost panel is ten times larger than the scale of the other two. Also note that the  $1/j$  law in (25) is independent of  $p_0$ , so it is the same curve in all panels. For small  $p_0$ , the fixed-time spectrum (20) is well-approximated by the  $1/j^2$  law (24), and each lies below the  $1/j$  law (25). To see why, note that the  $1/j$  law gives  $wN$  as the number of mutations found in one cell, while the  $1/j^2$  law gives  $(1/2)(w/q_0)N$ , which is  $(1/2)wN$  for  $p_0 = 0$ . As  $p_0$  increases to 1, the fixed-time spectrum (20) transitions to the  $1/j$  law, while the  $1/j^2$  law increases according to  $(w/q_0)N \cdot 1/(j(j+1))$ , diverging further and further from the fixed-time spectrum. In Fig. 7a, we show how for a fixed, large value of  $p_0$  ( $p_0 = 0.99$ ), the fixed-time spectrum (20) transitions from the  $1/j^2$  law at the large-frequency end to the  $1/j$  law at the small-frequency end. Note that the power laws of (24) and (25) intersect at

$$j = 1/q_0 - 1,$$

which gives an indication of the frequency at which the transition occurs. This intersecting point is independent of  $N$ , as is illustrated in Fig. 7b (dotted vertical line).

To understand this transition between power laws, note that mutations that occur early in the evolution of a large tumor are only detected if they occur on the skeleton, which is why mutations at large frequencies in a large tumor follow the  $1/j^2$  skeleton spectrum. For mutations that occur late, we need to consider both skeleton cells and finite-family cells. When  $p_0$  is large, most cells are finite-family cells (their long-run proportion is  $p_0$ ), and as  $p_0$  approaches 1, finite-family clones start to behave like a critical branching process with net growth rate 0. This





**Fig. 7.** For a fixed, large value of the extinction probability  $p_0$ , the expected fixed-time spectrum (20) of Proposition 2 (solid blue line) transitions from the  $1/j^2$  power law of (24) (dotted red line) at the large-frequency end to the  $1/j$  power law of (25) (dotted green line) at the small-frequency end. The dotted vertical line shows the intersecting point  $j = 1/q_0 - 1$  of the two power laws in (24) and (25). In (a), the parameters are  $N = 10^4$ ,  $p_0 = 0.99$  and  $w = 1$ , and in (b), we increase the expected tumor size to  $N = 10^5$ . Note that the transition between power laws occurs around the same value of  $j$  in both cases,  $j = 1/q_0 - 1$ .

is why late mutations follow the  $1/j$  law of a constant-sized population. Thus, even though the branching process dynamics of cell division and cell death are the same throughout the evolution of the tumor, from the perspective of mutation accumulation, the tumor effectively behaves like a pure-birth exponential growth process initially, and more like a constant-sized process at the end, assuming a large extinction probability  $p_0$ . The transition between power laws is the transition between these two growth regimes.

Note finally that as we observed for the skeleton spectrum in Section 3.2, the fixed-time spectrum (20) deviates from the  $1/j^2$  law at the very largest frequencies (for  $j$  of order  $N$ ) in Fig. 7, due to the variability in tumor size at time  $t_N$ . It is therefore more correct to say that for fixed large  $p_0$ , the fixed-time spectrum transitions from the skeleton spectrum at the large-frequency end to the constant-sized spectrum at the small-frequency end, without reference to the power laws. It remains true, however, that as  $p_0$  increases from 0 to 1, the small-frequency end of the spectrum transitions between the two power laws.

#### 4.3. Total mutational burden of the tumor

We next wish to quantify how  $p_0$  affects overall mutation accumulation. To this end, we derive in Proposition 3 the expected total mutational burden of the tumor, both under the fixed-time and fixed-size spectrum. This quantity indicates the genetic diversity of the tumor, which has implications e.g. for its adaptability under treatment. It also enables us to compute a normalized version of the SFS, which can be useful for parameter estimation, as is discussed further in Section 5. Before stating the proposition, we define  $M_j(t) := \sum_{k \geq j} S_k(t)$  as the number of mutations found in  $\geq j$  cells at time  $t$ .

#### Proposition 3.

- (1) For  $0 < p_0 < 1$ , the expected total number of mutations in the fixed-time spectrum is given by

$$\begin{aligned} \mathbb{E}[M_1(t_N) | Z_0(t_N) > 0] \\ = -wN \cdot (1/p_0) \log(q_0 + p_0/N) \\ \sim -wN \cdot \log(q_0)/p_0, \quad N \rightarrow \infty. \end{aligned} \quad (26)$$

For  $p_0 = 0$ ,  $\mathbb{E}[M_1(t_N)] = w(N-1) \sim wN$  as  $N \rightarrow \infty$  by (16).

- (2) Define  $S$  and  $A$  as in Proposition 2. The expected total number of mutations in the fixed-size spectrum is given by

$$\begin{aligned} \mathbb{E}[M_1(t_N) | \tau_N < \infty] \\ = (w/q_0) \cdot \sum_{k=1}^{N-1} (1 - p_0^{N-k}) (1 - h_{(1,k)}^{(0,N)} - h_{(1,k)}^{(0,0)}), \end{aligned} \quad (27)$$

where for each  $(r, s) \in A$ , the vector  $(h_{(\ell, m)}^{(r, s)})_{(\ell, m) \in S}$  solves the linear system (23) of Proposition 2. For  $p_0 = 0$ ,  $\mathbb{E}[M_1(t_N)] = w(N-1) \sim wN$  as  $N \rightarrow \infty$  by (16).

**Proof.** Appendix D.  $\square$

Now, for ease of notation, write  $\bar{M}_1 := \mathbb{E}[M_1(t_N) | Z_0(t_N) > 0]$  for the expected total mutational burden under the fixed-time spectrum. We are interested in comparing  $\bar{M}_1$  with the expected number of mutations under the  $1/j^2$  skeleton law of (24). To this end, define

$$\hat{M}_1 := (w/q_0)N, \quad (28)$$

following (16). This simple estimate has been used e.g. in Ling et al. (2015), where the authors estimate the total number of mutations in a hepatocellular carcinoma (HCC) tumor under a few different assumptions on tumor evolution. The ratio between  $\bar{M}_1$  and  $\hat{M}_1$  is, for  $0 < p_0 < 1$ ,

$$\begin{aligned} \bar{M}_1/\hat{M}_1 &= -(q_0/p_0) \cdot \log(q_0 + p_0/N) \\ &\sim -(q_0/p_0) \cdot \log(q_0), \quad N \rightarrow \infty. \end{aligned} \quad (29)$$

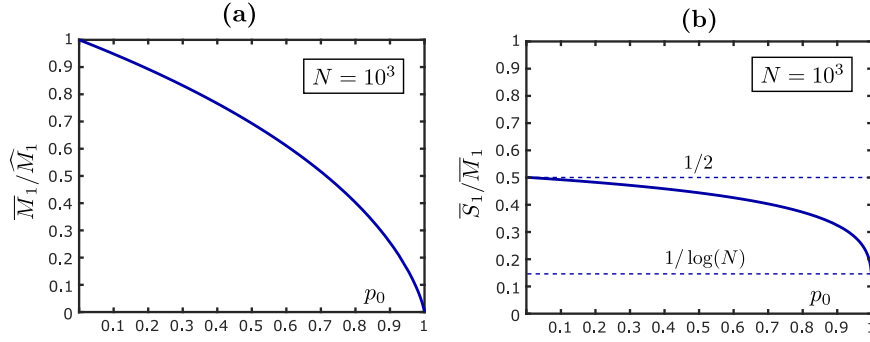
In Fig. 8a, we show this ratio as a function of  $p_0$  for  $N = 1000$ . The ratio is decreasing in  $p_0$ , it converges to 1 as  $p_0 \rightarrow 0$ , and it converges to 0 as  $p_0 \rightarrow 1$ . To give some examples, in the  $N \rightarrow \infty$  limit,  $\bar{M}_1/\hat{M}_1 = 0.46$  for  $p_0 = 0.75$  and  $\bar{M}_1/\hat{M}_1 = 0.047$  for  $p_0 = 0.99$ . Thus, if one is interested in estimating the total number of mutations in an exponentially growing tumor, using the simple expression (28) implied by the skeleton spectrum will result in a significant overestimate when  $p_0$  is large. Indeed, as  $p_0$  increases, finite-family cells start to dominate the population, and they accumulate mutations less efficiently than skeleton cells.

#### 4.4. Proportion of mutations found in one cell

At the extremes  $p_0 = 0$  and  $p_0 = 1$ , the small-frequency end of the SFS is characterized by the  $1/j^2$  skeleton law and the  $1/j$  constant-sized law, respectively. To better understand how the small-frequency end behaves for intermediate values of  $p_0$ , we next determine the relative proportion of mutations found at the very smallest frequency, i.e. in one cell. This metric quantifies the transition between the  $1/j^2$  and  $1/j$  power laws at the small-frequency end, and it enables us to propose a simple estimator for  $p_0$  in Section 5 below.

In Appendix G, we show that in the fixed-time spectrum, the expected number of mutations found in one cell is given by, for  $0 < p_0 < 1$ ,

$$\mathbb{E}[S_1(t_N) | Z_0(t_N) > 0]$$



**Fig. 8.** Graphs of the metrics  $\bar{M}_1/\hat{M}_1$  and  $\bar{S}_1/\bar{M}_1$  given by (29) and (31), as a function of  $p_0$ . (a) Ratio between the expected total number of mutations in the fixed-time spectrum,  $\bar{M}_1 := \mathbb{E}[M_1(t_N)|Z_0(t_N) > 0]$  given by (26), and the simple estimate  $\hat{M}_1$  of (28), derived from the  $1/j^2$  skeleton law of (24), as a function of  $p_0$  for  $N = 10^3$ . The two estimates  $\bar{M}_1$  and  $\hat{M}_1$  agree for  $p_0 = 0$ , but  $\hat{M}_1$  becomes a significant overestimate of  $\bar{M}_1$  as  $p_0$  increases. (b) The proportion of mutations found in one cell,  $\bar{S}_1/\bar{M}_1 = \varphi_N(p_0)$  as given by (31), quantifies the transition between the  $1/j^2$  and  $1/j$  power laws at the small-frequency end of the spectrum. Note that the transition between power laws accelerates as  $p_0$  increases.

$$\begin{aligned} &= wN \cdot (1/p_0)(1 - 1/N + (q_0/p_0)\log(q_0 + p_0/N)) \\ &\sim wN \cdot (1/p_0)(1 + (q_0/p_0)\log(q_0)), \quad N \rightarrow \infty. \end{aligned} \quad (30)$$

As in Section 4.3, we write  $\bar{S}_1 := \mathbb{E}[S_1(t_N)|Z_0(t_N) > 0]$  for ease of notation. We then define  $\varphi_N(p_0) := \bar{S}_1/\bar{M}_1$  as the proportion of mutations found in one cell. By (26) and (30),

$$\begin{aligned} \varphi_N(p_0) &= -(1 - 1/N + (q_0/p_0)\log(q_0 + p_0/N)) / \log(q_0 + p_0/N) \\ &\sim -(1 + (q_0/p_0)\log(q_0)) / \log(q_0), \quad N \rightarrow \infty. \end{aligned} \quad (31)$$

In Fig. 8b, we show  $\varphi_N = \bar{S}_1/\bar{M}_1$  as a function of  $p_0$  for  $N = 1000$ . The function is strictly decreasing in  $p_0$ , it converges to 0.50 as  $p_0 \rightarrow 0$ , and it converges to  $(1 - 1/N)/\log(N)$  as  $p_0 \rightarrow 1$ , which is of order  $1/\log(N)$  for  $N$  large. In the  $p_0 \rightarrow 0$  regime, the SFS of the total population  $(Z_0(t))_{t \geq 0}$  is the SFS of the skeleton  $(\tilde{Z}_0(t))_{t \geq 0}$ , in which case half the mutations are found in one cell by Section 3.3. In the  $p_0 \rightarrow 1$  regime, the SFS of  $(Z_0(t))_{t \geq 0}$  is the SFS of the constant-sized Moran model, in which case the expected number of mutations is of order  $wN \sum_{j=1}^N 1/j \sim wN \log(N)$  for  $N$  large, and the proportion of mutations found in one cell is of order  $1/\log(N)$ . Note that the rate of change of  $\varphi_N(p_0)$  increases as  $p_0$  increases (Fig. 8b). This implies that the deviation from the  $1/j^2$  skeleton law at the small-frequency end is initially slow for small values of  $p_0$ , but it accelerates as  $p_0$  increases and transitions quickly to the  $1/j$  law for large values of  $p_0$ . It also implies that  $\bar{S}_1/\bar{M}_1$  is more useful for distinguishing larger values of  $p_0$  than smaller values, as will become more apparent in Section 5 below.

#### 4.5. Spectra of individual large tumors (laws of large numbers)

The results of Proposition 2 hold in expectation, meaning that they apply to an average SFS computed over a large number of tumors. If we want to use these results to understand the evolutionary history of individual tumors, we need to know more about how well they apply on a tumor-by-tumor basis. It is well-known that conditional on the nonextinction event  $\Omega_\infty$ ,  $Z_0(t) \sim Ye^{\lambda_0 t}$  as  $t \rightarrow \infty$  almost surely, where  $Y$  follows the exponential distribution with mean  $1/q_0$  (Theorem 1 of Durrett (2015)). In other words, the tumor population  $Z_0(t)$  eventually grows at exponential rate  $\lambda_0$ , but the initial value of the exponential growth function is random and depends on the individual tumor. We can use this fact to formulate laws of large numbers for the fixed-time and fixed-size spectrum, which we state formally as conjectures. In Appendix H, we present simple calculations in support of these conjectures, and we also prove an analogous result (H.1) for a simplified, semideterministic version of our model.

#### Conjecture.

- (1) Define  $t_N$  as in (18). Then, there exists an exponential random variable  $X$  with mean 1 so that for fixed  $j \geq 1$ , conditional on  $\Omega_\infty$ ,

$$S_j(t_N) \sim X \cdot wN \cdot \int_0^1 (1 - p_0 y)^{-1} (1 - y) y^{j-1} dy \quad (32)$$

as  $N \rightarrow \infty$  almost surely.

- (2) Define  $\tau_N$  as in (19). Then, for fixed  $j \geq 1$ , conditional on  $\Omega_\infty$ ,

$$S_j(\tau_N) \sim wN \cdot \int_0^1 (1 - p_0 y)^{-1} (1 - y) y^{j-1} dy \quad (33)$$

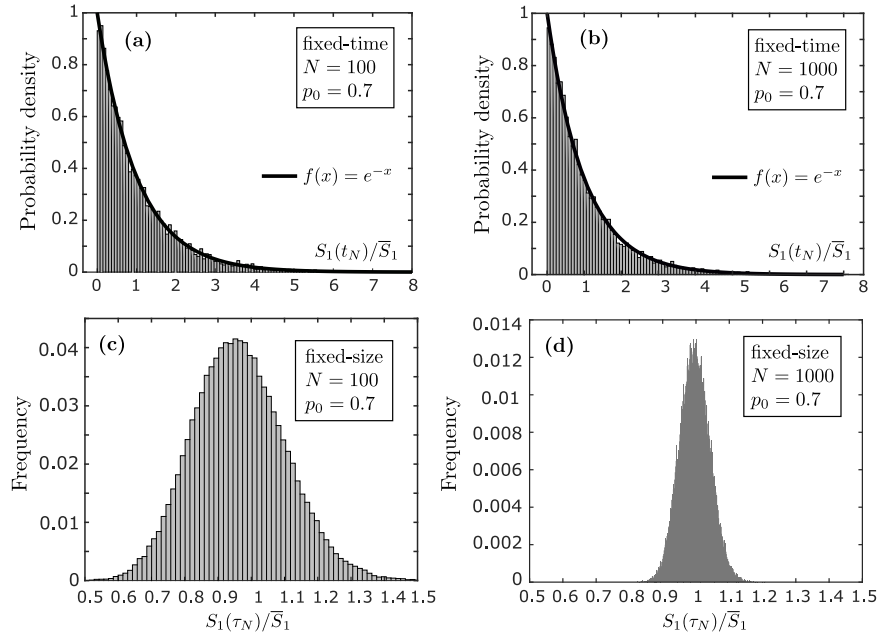
as  $N \rightarrow \infty$  almost surely.

Both the fixed-time conjecture (32) and the fixed-size conjecture (33) agree with simulation results, see Figs. 9 and 10. The main difference between (32) and (33) is that the right-hand-side of (32) is stochastic, while the right-hand side of (33) is a constant. The former expression has a random scaling factor  $X$ , which captures the variability in tumor size at time  $t_N$ , whereas the fact that the tumor size is always  $N$  at time  $\tau_N$  eliminates this variability in the latter expression. Note that since  $\mathbb{E}[X] = 1$ , the right-hand sides of (32) and (33) agree in the mean, and this mean agrees with the right-hand side of the asymptotic spectrum (21) of Proposition 2. Importantly, the scaling factor  $X$  in (32) is independent of  $j$ , which indicates that the proportion of mutations found in  $j$  cells is the same in (21), (32) and (33). In other words, according to these conjectures, if we normalize the SFS with the total number of mutations,  $S_j(t)/M_1(t)$ , the fixed-time and fixed-size spectrum of an individual large tumor will be completely characterized by the asymptotic expected spectrum (21) of Proposition 2 (Fig. 10). This can be useful for parameter estimation, as we discuss next.

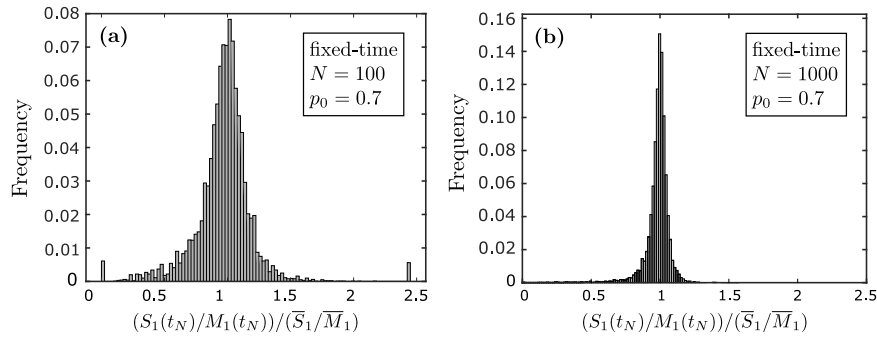
#### 5. Signatures of cell viability

In this section, we use our theoretical results to propose a simple estimator for the extinction probability  $p_0$ , based on extracting one or more spatially separated subclones from a tumor. By a subclone, we mean all currently living descendants of a given common ancestor, i.e. all leaves of the branching tree started by a given tumor cell. Since every clone or subclone derived from a single tumor cell obeys the same branching process dynamics as the overall tumor, all of our previous results can be applied to individual subclones.

We make the strong assumption that each cell in each sampled subclone can be single-cell sequenced so that all its mutations are captured, even at the smallest frequencies, which is beyond current sequencing technology. Our main purpose with this section



**Fig. 9.** Simulation results support the conjectured laws of large numbers (32) and (33) for the fixed-time and fixed-size spectrum respectively. **(a)** Histogram of  $S_1(t_N)/\bar{S}_1$  over  $10^4$  simulation runs with  $N = 100$ ,  $p = 0.7$  and  $w = 1$ , where  $t_N$  is defined by (18), and  $\bar{S}_1 := \mathbb{E}[S_1(t_N) | Z_0(t_N) > 0]$  as given by (30). The y-axis is normalized so as to approximate the density of the underlying probability distribution. By comparison with  $x \mapsto e^{-x}$ , we see that  $S_1(t_N)/\bar{S}_1$  appears to be a mean-1 exponential random variable, which is consistent with the conjectured law of large numbers (32). **(b)** When the population size is increased to  $N = 1000$ ,  $S_1(t_N)/\bar{S}_1$  retains the mean-1 exponential distribution, consistent with (32). **(c)** Histogram of  $S_1(\tau_N)/\bar{S}_1$  over  $10^4$  simulation runs with  $N = 100$ ,  $p = 0.7$  and  $w = 1$ , where  $\tau_N$  is defined by (19). **(d)** Same as in (c), except now  $N = 1000$ . Together, (c) and (d) indicate that the ratio  $S_1(\tau_N)/\bar{S}_1$  concentrates around 1 as  $N$  increases, which is consistent with the conjectured law of large numbers (33).



**Fig. 10.** Simulation results indicate that if the SFS is normalized by the total number of mutations, the fixed-time and fixed-size spectrum obey the same law of large numbers. **(a)** Histogram of  $(S_1(t_N)/M_1(t_N))/(\bar{S}_1/\bar{M}_1)$  over  $10^4$  simulation runs with  $N = 100$ ,  $p = 0.7$  and  $w = 1$ , where  $t_N$  is defined by (18), and the expected ratio  $\bar{S}_1/\bar{M}_1$  is given by (31). Note the point masses at 0 and 2.44, which represent simulation runs where  $S_1(t_N)/M_1(t_N) = 0$  and  $S_1(t_N)/M_1(t_N) = 1$ , respectively. **(b)** Same as in (a), except now  $N = 1000$ . Together, (a) and (b) indicate that as  $N$  increases,  $S_1(t_N)/M_1(t_N)$  concentrates around  $\bar{S}_1/\bar{M}_1$ . This in turn suggests that if the fixed-time spectrum is normalized by the total number of mutations  $M_1(t_N)$ , it obeys the same law of large numbers as the normalized version of the fixed-size spectrum.

is to show how the information contained in the small-frequency end of the SFS can in principle be used to decouple the mutation rate  $w$  and the extinction probability  $p_0$ . We discuss practical considerations and potential alternative approaches in Section 6 below.

Say that we sample a subclone of size  $n$ . For  $1 \leq j \leq n-1$ , let  $s_j$  be the number of mutations found in  $j$  cells of the subclone, and let  $m_1 := \sum_{j=1}^{n-1} s_j$  be the total number of mutations. Here, we ignore mutations found in all cells of the subclone, since they include (i) mutations that accumulate prior to tumor initiation, (ii) mutations that occur post-tumor-initiation but prior to initiation of the subclone, and (iii) mutations that occur post-subclone-initiation but still end up in all subclone cells. Let  $\bar{s}_1$  be the expected number of mutations found in one subclone cell under the fixed-time spectrum, and let  $\bar{m}_1$  be the expected total number of mutations under the fixed-time spectrum. By (31), we

can write

$$\bar{s}_1/\bar{m}_1 = \varphi_n(p_0),$$

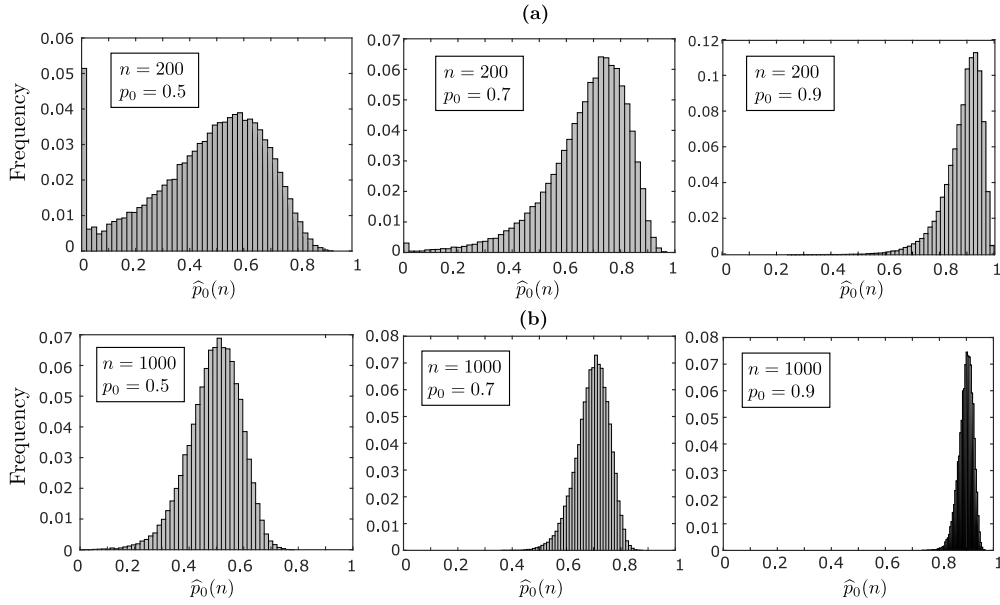
where  $\varphi_n(p_0)$  is continuous and strictly decreasing in  $p_0$ . In particular,  $\varphi_n(p_0)$  is invertible. This implies that given  $\bar{s}_1$  and  $\bar{m}_1$ ,  $p_0$  can be recovered from this expression via

$$p_0 = \varphi_n^{-1}(\bar{s}_1/\bar{m}_1).$$

For the sampled values  $s_1$  and  $m_1$ , this suggests the following estimator for  $p_0$ :

$$\hat{p}_0(n) := \varphi_n^{-1}(s_1/m_1). \quad (34)$$

Of course, if the subclone is sampled at a certain size, it makes more sense to use the fixed-size spectrum than the fixed-time spectrum. In addition,  $m_1$  excludes mutations found in all subclone cells, while  $\bar{m}_1$  includes some of these mutations. These



**Fig. 11.** Histogram of the estimator  $\hat{p}_0(n)$  defined in (34) computed across  $10^5$  synthetic subclone samples of size  $n$ , given true values  $p_0 \in \{0.5, 0.7, 0.9\}$  and  $w = 1$ . In (a), the subclone size is  $n = 200$ , and in (b), the size is  $n = 1000$ . Together, (a) and (b) indicate that as the size of the subclone increases, the estimator  $\hat{p}_0(n)$  concentrates around the true value of  $p_0$ , which in turn indicates statistical consistency of the estimator.

potential sources of error are minor and can easily be resolved if necessary, as we discuss in more detail below.

Note that the ratio of expected values  $\bar{s}_1/\bar{m}_1$  takes values in  $[(1 - 1/n)/\log(n), 1/2]$  by Section 4.4, whereas due to stochasticity, the sampled ratio  $s_1/m_1$  can take any value in  $[0, 1]$ . To complete the estimator in (34), we therefore extend the definition of  $\varphi_n^{-1}$  by setting

$$\varphi_n^{-1}(x) := \begin{cases} 0, & 1/2 \leq x \leq 1, \\ 1, & 0 \leq x \leq (1 - 1/n)/\log(n). \end{cases} \quad (35)$$

For example, if we observe a ratio  $s_1/m_1$  larger than  $1/2$ , we default to the estimate  $\hat{p}_0(n) = 0$ , since the expected ratio  $\bar{s}_1/\bar{m}_1$  is largest (and equal to  $1/2$ ) for  $p_0 = 0$ . Once  $p_0$  has been estimated, an estimate for the mutation rate  $w$  can be obtained from (30) or (26).

The estimator  $\hat{p}_0(n)$  has the benefit of being simple to define and to compute. However, as was mentioned above, it may make more sense to use the fixed-size spectrum than the fixed-time spectrum, and  $\bar{m}_1$  includes clonal mutations that arise post-subclone-initiation, whereas  $m_1$  excludes these mutations. The second potential source of error is minor, and it can easily be removed simply by subtracting from  $\bar{m}_1$  the contribution from mutations shared by all subclone cells. The first potential source of error is also likely to be insignificant when  $n$  is large, since by the conjectured laws of large numbers in Section 4.5, the normalized spectrum of an individual large subclone is robust to whether it is observed at a fixed time or a fixed size. For smaller values of  $n$ , one can replace  $\bar{s}_1$  and  $\bar{m}_1$  by the corresponding quantities (22) and (27) for the fixed-size spectrum, which we denote here by  $\bar{s}_1$  and  $\bar{m}_1$ . It remains true that we can write  $\bar{s}_1/\bar{m}_1 = \psi_n(p_0)$  for some function  $\psi_n$  of  $p_0$ , which allows us to define an estimator for  $p_0$  as before. However, the fixed-size estimator has to be obtained numerically, e.g. by precomputing  $\psi_n(p_0)$  over a grid of values for  $p_0$ , and minimizing the error between the observed ratio  $s_1/m_1$  and the expected ratio  $\bar{s}_1/\bar{m}_1$  over the grid.

To evaluate  $\hat{p}_0(n)$ , we use computer simulations to generate multiple independent subclones of size  $n$  with true extinction probability  $p_0$ , and for each generated subclone, we compute the estimate  $\hat{p}_0(n)$ . In Fig. 11a, we show a histogram for  $\hat{p}_0(n)$

across  $10^5$  synthetic subclone samples of size  $n = 200$  with true extinction probabilities  $p_0 \in \{0.5, 0.7, 0.9\}$ . In Table 2, we show performance metrics for  $\hat{p}_0(n)$  computed across the  $10^5$  samples. For  $p_0 = 0.5$ , the estimator defaults to  $\hat{p}_0(n) = 0$  for 4.6% of the subclone samples, for  $p_0 = 0.7$ , it defaults to 0 in 0.3% of cases, and for  $p_0 = 0.9$ , it never defaults to 0. For all values of  $p_0$ , the median estimate of  $\hat{p}_0(n)$  accurately recovers the true value, and as  $p_0$  increases, the quality of the estimate improves in terms of standard deviation. In Fig. 11b, we increase the subclone size to  $n = 1000$  and observe a marked improvement in the quality of  $\hat{p}_0(n)$ . This indicates statistical consistency of the estimator, meaning that  $\hat{p}_0(n) \rightarrow p_0$  in probability as  $n \rightarrow \infty$ , which would also be a direct consequence of the conjectured laws of large numbers (32)–(33) and the continuous mapping theorem. In other words, the estimator appears to recover the true value of  $p_0$  with arbitrarily high precision given a sufficiently large subclone. Note that the size of  $n$  required to return a high-precision estimate becomes smaller as  $p_0$  increases, making  $\hat{p}_0(n)$  especially useful when  $p_0$  is large. Indeed, as we remarked in Section 4.4, the rate of change of the expected ratio  $\bar{s}_1/\bar{m}_1$  increases as  $p_0$  increases, making it more useful for distinguishing between larger values of  $p_0$  than smaller values.

Whenever it is possible to do multi-region sampling, there may be benefits to extracting multiple small, spatially separated subclones over a single large one. In this more general setting, we sample  $K \geq 1$  subclones of size  $n$ . For  $1 \leq j \leq K$ , let  $s_j^k$  be the number of mutations found in  $j$  cells of subclone number  $k$ , and let  $m_1^k := \sum_{j=1}^{n-1} s_j^k$  be the total number of mutations in subclone  $k$ . We replace  $s_1$  and  $m_1$  in the definition of  $\hat{p}_0(n)$  in (34) by the sums  $\sum_{k=1}^K s_1^k$  and  $\sum_{k=1}^K m_1^k$  to obtain the estimator

$$\hat{p}_0(n, K) := \varphi_n^{-1}(\sum_{k=1}^K s_1^k / \sum_{k=1}^K m_1^k). \quad (36)$$

Of course,  $\hat{p}_0(n, 1) = \hat{p}_0(n)$ . In Fig. 12, we show a histogram for  $\hat{p}_0(n, K)$  evaluated across  $10^5$  synthetic samples, each sample consisting of  $K = 5$  independent subclones of size  $n = 200$ . In Table 3, we show performance metrics computed across the  $10^5$  samples. Qualitatively, the histograms in Fig. 12 are very similar to the histograms of Fig. 11b, and quantitatively, the performance metrics in Table 3 mimic those for the  $n = 1000$  case in Table 2. In other words, the quality of the estimate of  $p_0$



**Table 2**

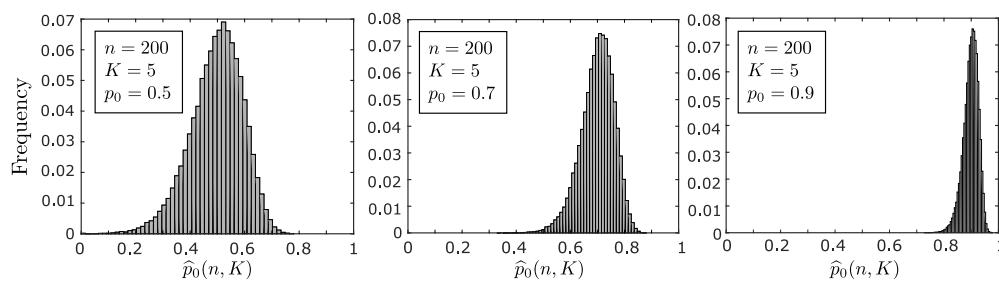
Performance metrics for the estimator  $\hat{p}_0(n)$  computed from the data that underlies Fig. 11. Both for  $n = 200$  and  $n = 1000$ , the median of  $\hat{p}_0(n)$  accurately recovers the true value of  $p_0$ . In addition, the estimator improves in terms of standard deviation both as  $n$  increases and as  $p_0$  increases.

$n$	$p_0$	Mean	Median	St. dev.	Defaults to 0 (%)	Defaults to 1 (%)
200	0.5	0.4694	0.5041	0.2099	0.0459	0
	0.7	0.6755	0.7045	0.1505	0.0026	0
	0.9	0.8839	0.9001	0.0725	0	0.00002
1000	0.5	0.4921	0.5000	0.0948	0.00002	0
	0.7	0.6957	0.7010	0.0603	0	0
	0.9	0.8979	0.9009	0.0266	0	0

**Table 3**

Performance metrics for the estimator  $\hat{p}_0(n, K)$  computed from the data that underlies Fig. 12. Note the similarity between these metrics and the lower half of Table 2.

$K$	$n$	$p_0$	Mean	Median	St. dev.	Defaults to 0 (%)	Defaults to 1 (%)
5	200	0.5	0.4965	0.5050	0.0957	0.00013	0
		0.7	0.7000	0.7055	0.0612	0	0
		0.9	0.8983	0.9014	0.0279	0	0



**Fig. 12.** Histogram of the estimator  $\hat{p}_0(n, K)$  of (36) computed across  $10^5$  synthetic samples of  $K = 5$  subclones of size  $n = 200$ , given true values  $p_0 \in \{0.5, 0.7, 0.9\}$  and  $w = 1$ . Note the similarity between these histograms and the histograms in (b) of Fig. 11, which indicates that sampling five subclones of size 200 gives a comparable estimate of  $p_0$  to sampling a single subclone of size 1000 in these examples.

obtained from sampling one subclone of size 1000 is comparable to the one obtained from sampling five subclones of size 200. In this scenario, it may make more sense to extract multiple small subclones than one large one, since it is impossible to tell from a single subclone sample alone whether the tumor as a whole can be considered as evolving neutrally. Should there be differences in the subclone dynamics, a multiregion sample may tease this out, and should the dynamics be the same, the estimate one obtains for  $p_0$  will be of comparable quality to the single large subclone case.

## 6. Discussion

In this work, we have established exact expressions for the expected site frequency spectrum of a tumor, or more generally any population, that evolves according to a branching process with neutral mutations under the infinite-sites assumption of population genetics. We first considered the skeleton subpopulation, consisting of cells with an infinite line of descent, and obtained explicit expressions for the SFS of the skeleton evaluated both at a fixed time and a fixed size. We then examined the total population, deriving an explicit expression for the fixed-time spectrum and a computational expression for the fixed-size spectrum. Our results apply to mutations at small and large frequencies, to tumor tissue samples and tumor subclones of any size, and to all values of the extinction probability  $p_0$ , even values as large as  $p_0 = 0.90$  and above, which are broadly relevant for cancer. We now discuss in detail how our results relate to results previously obtained in the literature.

We begin by stating skeleton results established by Durrett (2013, 2015), Bozic et al. (2016) and Williams et al. (2016). The result in Bozic et al. (2016) says that for fixed  $0 < f < 1$ , the

expected number of subclonal mutations found in a proportion  $\geq f$  of cells at time  $t$  is, as  $t \rightarrow \infty$ ,

$$(w/q_0)(1/f - 1). \quad (37)$$

Note that this is a cumulative spectrum of mutations with frequency at least  $f$ . A similar result appears in Williams et al. (2016) under a deterministic growth model. Durrett's result (Durrett, 2013, 2015), which preceded the other two, is given under continuous mutation accumulation, and it includes clonal mutations, which by (C.1) of Appendix C requires adding  $v/\lambda_0$  ( $= w/q_0$ ) mutations to (37). This yields  $(v/\lambda_0)(1/f)$  as the cumulative spectrum, see Theorem 1 of Durrett (2013) and Theorem 2 of Durrett (2015).<sup>3</sup> Under discrete mutation accumulation, the number of clonal mutations is  $w p_0/q_0$  by (14) of Proposition 1, which yields the result  $(w/q_0)(1/f) - w$  including clonal mutations. The difference of  $w$  reflects the difference in the number of clonal mutations between the discrete and continuous model, see Appendix C.

To compare the  $1/f$  law in (37) with our results of Proposition 1, recall that by (14), the fixed-size skeleton spectrum can be written as  $\mathbb{E}[\tilde{S}_j(\tilde{\tau}_N)] = (w/q_0)N \cdot 1/(j(j+1)) - (w p_0/q_0)\delta_{1,j}$  for  $j = 1, \dots, N-1$ . Under this spectrum, it is easy to compute the expected number of subclonal mutations found in a proportion  $\geq f$  of skeleton cells as

$$\begin{aligned} & \mathbb{E}\left[\sum_{j=\lceil Nf \rceil}^{N-1} \tilde{S}_j(\tilde{\tau}_N)\right] \\ &= (w/q_0)N \cdot \left(\sum_{j=\lceil Nf \rceil}^{N-1} 1/(j(j+1))\right) - (w p_0/q_0)\delta_{1,\lceil Nf \rceil} \end{aligned}$$

<sup>3</sup> There is an apparent typo in Theorem 1 of Durrett (2013). The result is written as  $(v/q_0)(1/f)$  in our notation, but should be  $(v/\lambda_0)(1/f)$ . This is corrected in Theorem 2 of the later notes Durrett (2015) by the same author.

$$\begin{aligned}
&= (w/q_0)N \cdot (1/\lceil Nf \rceil - 1/N) - (wp_0/q_0)\delta_{1,\lceil Nf \rceil} \\
&\sim (w/q_0)(1/f - 1), \quad N \rightarrow \infty.
\end{aligned} \tag{38}$$

The  $1/f$  law in (37) can therefore be viewed as a cumulative version of our  $1/j^2$  law. Note that (37) is established in the fixed-time regime in the above cited works, while the calculations in (38) show that the  $1/f$  law also holds in the fixed-size regime.

To summarize, in the fixed-size regime, we have established exact adherence to the  $1/j^2$  law on  $j = 2, \dots, N - 1$  for any skeleton size  $N \geq 2$ , and our result implies the cumulative  $1/f$  law as  $N \rightarrow \infty$ . In the fixed-time regime, we have established the exact expression (12), which converges to the  $1/j^2$  law for  $j \ll N$  as  $N \rightarrow \infty$ . The  $1/f$  result in (37) complements our fixed-time result, since it confirms that if we compute the spectrum of mutations found in a certain *proportion* of cells, rather than in a certain *number* of cells, the fixed-time spectrum converges to the same  $1/f$  law as the fixed-size spectrum as  $N \rightarrow \infty$ , including at the very largest frequencies. It should be emphasized that the  $1/f$  law, which has been extensively cited in the literature, is an asymptotic result established for the skeleton. For the total population, including finite-family cells, the  $1/f$  law is applicable only to mutations at large frequencies in a large tumor. This is particularly the case for low-viability tumors.

The expressions (20) and (21) for the expected fixed-time spectrum of Proposition 2 have been previously obtained by Ohtsuki and Innan (2017) under the assumption of deterministic growth of the tumor bulk and stochastic growth of mutant subclones, which is a reasonable approximation when  $p_0$  is small. We have established these expressions for the fully stochastic model and for all values of  $p_0$ , by evaluating the spectrum at the fixed time  $t_N$  at which a surviving tumor has expected size  $N$ , defined in (18), and conditioning on the survival event  $\{Z_0(t_N) > 0\}$ . We have also conjectured the laws of large numbers (32) and (33) for the fixed-time and fixed-size spectrum, supported by heuristic calculations and simulations. We have proved a law of large numbers (H.1) for the semideterministic model of Ohtsuki and Innan (2017), see Appendix H, and we plan to prove the fully stochastic results (32) and (33) in a future work. Lambert (2009) has proved a similar result in the context of a coalescent point process (CPP), a framework under which an extant population is endowed with a coalescent structure that specifies how lineages coalesce when traced backwards in time. Lambert's result, see his Theorem 2.3, deals with a ranked sample of individuals from a CPP in the large-sample limit, and it has the same form as our conjectured fixed-size law of large numbers (33), the latter applying to our forwards-in-time branching process stopped at a certain size. The expected fixed-size spectrum (22) of Proposition 2 is new as far as we know, as well as expressions (26) and (27) of Proposition 3 for the expected total mutational burden of the tumor.

All of the above results hold under the infinite-sites model of population genetics. Cheek and Antal (2018) have recently examined the SFS of an exponentially growing population without this assumption, citing single-cell sequencing results of Kuipers et al. (2017) as motivation. They observe that if recurrent mutations are allowed (but no back mutations), and there are  $S$  sites in the genome, the expected SFS at time  $t$  can be computed as  $S \cdot \mathbb{P}(Y(t) = j)$ , where  $Y(t)$  is the number of mutants at time  $t$  in a two-type model of wild-type and mutant cells, each growing at the same rate. Then, to compute the SFS, one needs the distribution of  $Y(t)$ , which has been obtained under various assumptions e.g. by Antal and Krapivsky (2011), Kessler and Levine (2013, 2015), and Keller and Antal (2015). Cheek and Antal obtain SFS results under limits of large time/size and small mutation rate, and their results obey  $1/j^2$  power laws at large frequencies. However, when the mutation rate is sufficiently large compared to the population size, their small-frequency behavior

diverges from ours, see e.g. their Fig. 1. Yet other authors have substituted the infinite-sites model with the *infinite-alleles* model, under which each new mutation creates a new type of individual, see e.g. Griffiths and Pakes (1988), Pakes (1989), Champagnat et al. (2012), Wu and Kimmel (2013). Under this model, the site frequency spectrum is usually replaced by an *allele frequency spectrum*, which tracks frequencies of genetically distinct individuals, known as *haplotypes* (Lambert, 2009).

Our complete theoretical results give rise to several important insights. First of all, whereas the fixed-time and fixed-size skeleton spectrum depends on the mutation rate  $w$  and the extinction probability  $p_0$  only through the effective mutation rate  $w/q_0$ , the two parameters decouple in the total population spectrum. The mutation rate  $w$  scales the spectrum linearly, whereas the extinction probability  $p_0$  changes its shape at the small-frequency end. In fact, as  $p_0$  increases from 0 to 1, the small-frequency end of the spectrum transitions from the  $1/j^2$  power law characteristic of pure-birth exponential growth to the  $1/j$  law characteristic of constant-sized populations. We examined the simple metrics  $\bar{M}_1/\hat{M}_1$  and  $\bar{S}_1/\bar{M}_1$  that quantify this transition, where  $\hat{M}_1$  is the expected total mutational burden under the  $1/j^2$  power law spectrum. We saw that  $\bar{M}_1/\hat{M}_1 \rightarrow 0$  as  $p_0 \rightarrow 1$ , which suggests that the simple estimate  $\hat{M}_1$  of the total number of mutations, applied e.g. in Ling et al. (2015), is a significant overestimate of the actual expected number of mutations  $\bar{M}_1$  when  $p_0$  is large. We finally used the metric  $\bar{S}_1/\bar{M}_1$  to propose a simple estimator for  $p_0$ , based on sampling one or more spatially separated subclones from a tumor. This estimator accurately recovers the true value of  $p_0$  from synthetic single-cell sequencing data, and it is most accurate when  $p_0$  is large.

Our proposed estimator is currently of more theoretical than practical significance. It assumes that complete subclones of a given size can be reliably extracted from a tumor sample, and that each cell in each subclone can be single-cell sequenced so that all of its mutations are captured. We have proposed the estimator mainly to emphasize the information contained in the small-frequency end of the spectrum, and to show how it can in principle be used to decouple  $w$  and  $p_0$  using the SFS alone. Whether and how this decoupling can be achieved under current and foreseeable limitations of genomic data warrants further investigation. For example, one can derive an estimator based on  $s_j/m_j$  for  $j > 1$ , where  $m_j := \sum_{k=j}^{n-1} s_k$ , which excludes mutations found at the smallest frequencies. One can also design a more elaborate estimating procedure, which retains some of the smallest frequencies, but explicitly models the sequencing error. As it becomes easier to distinguish small-frequency mutations from sequencing errors, our simple estimator may at the very least provide quick and easy identification of low-viability tumors, and complement other more involved techniques.

We finally note that the estimation of  $w$  and  $p_0$  relies on several assumptions on tumor evolution. First of all, our model assumes that all mutations are selectively neutral. While this assumption is likely reasonable for smaller tumor subclones, it may be hard to verify that a subclone estimate is representative for the tumor as a whole. Our model also assumes exponential growth throughout tumor evolution, whereas e.g. due to spatial constraints and nutrient availability, the growth may be subexponential both during the early and late stages. Finally, our model assumes that the mutation rate is constant over time, and that the infinite-sites assumption holds for the tumor sample being analyzed. We note that the SFS-based estimates of  $w/q_0$  obtained by Williams et al. (2016) and Bozic et al. (2016) are derived from the large-frequency end of the SFS, which reflects early tumor dynamics. The same is true of the Werner et al. (2018) method of decoupling  $w$  and  $p_0$  by tracing genealogies of spatially separated bulk samples, since the common ancestors of these samples are

likely to have existed early in tumor evolution. Conversely, our suggested approach of utilizing the small-frequency end of the SFS will more reflect late tumor evolution. On the one hand, in the evolving discussion of tumor evolutionary history inference, it is important to acknowledge that any given estimation procedure may only give a temporally or spatially constrained picture of the dynamics. On the other hand, utilizing different parts of the SFS, or combining SFS estimates with other estimates, may allow one to glean insights into the dynamics at different stages of tumor evolution, and to possibly assess the validity of any modeling assumptions.

## Acknowledgments

The authors would like to thank Anna Kraut, the three anonymous reviewers and the editor for their valuable comments and suggestions. EBG and KL were supported in part by NSF, USA grant CMMI-1552764. JF was supported in part by NSF, USA grant DMS-1349724. KL and JF were supported in part by Research Council of Norway R&D Grant 309273. EBG was supported in part with funds from the Norwegian Centennial Chair Program, USA.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Proof of Proposition 1

In this section, we prove Proposition 1 on the expected SFS of the skeleton. We first present a brief outline of the proof.

In part (1), the SFS is observed at the fixed time  $\tilde{t}_N$ . To compute the expected number of mutations that end up in  $j \geq 1$  skeleton cells at time  $\tilde{t}_N$ , we decompose the time interval  $[0, \tilde{t}_N]$  into infinitesimal intervals  $[t, t + \Delta t]$ . We can compute how many mutations occur during each small interval using the expected mutation rate (6) in the main text. Then, to only count the mutations that end up in  $j$  skeleton cells at time  $\tilde{t}_N$ , we multiply this number by the probability that a single-cell derived skeleton clone has size  $j$  at time  $\tilde{t}_N$ , which has a known expression. We finally integrate over time to add up the contributions of the infinitesimal intervals.

In part (2), the SFS is observed at the stochastic time  $\tilde{t}_N$ , at which the skeleton reaches size  $N$ . In this case, we decompose into population size levels instead of into small time intervals. We know how many mutations accumulate on population size level  $k$  by (9) in the main text. Then, to get the expected number of mutations that accumulate on level  $k$  and end up in  $j \geq 1$  skeleton cells at time  $\tilde{t}_N$ , we need to compute the probability that starting from one skeleton cell carrying a particular mutation and  $k - 1$  cells without it,  $j$  cells carry the mutation when the skeleton reaches size  $N$ . To that end, we define a Markov chain that keeps track of how many skeleton cells carry the mutation as the skeleton increases in size, and we compute its hitting probabilities. We finally sum over  $k$  to add up the contributions of each population size level.

### Proof of Proposition 1.

- (1) By (6) in the main text, the expected mutation rate per skeleton cell per unit time is  $wr_0$ . To stratify mutations based on their frequencies at time  $\tilde{t}_N$ , we define

$$\tilde{p}_j(s) := \mathbb{P}(\tilde{Z}_0(s) = j | \tilde{Z}_0(0) = 1), \quad j \geq 1, s \geq 0,$$

as the size-distribution at time  $s$  of a single-cell derived skeleton clone. Since  $(\tilde{Z}_0(t))_{t \geq 0}$  is a Yule process with birth rate  $\lambda_0$ , this distribution has an explicit expression,

$$\tilde{p}_j(s) = (1/e^{\lambda_0 s})(1 - 1/e^{\lambda_0 s})^{j-1}, \quad j \geq 1,$$

which is the geometric distribution with support  $\{1, 2, \dots\}$  and success probability  $1/e^{\lambda_0 s}$ , see e.g. Section 3 of [Durrett \(2015\)](#) (the support does not include 0 since skeleton clones do not go extinct). For  $0 \leq t \leq \tilde{t}_N$ , let  $\tilde{S}_{j, \tilde{t}_N}(t)$  denote the number of mutations that accumulate in the time interval  $[0, t]$  and are found in  $j \geq 1$  skeleton cells at time  $\tilde{t}_N$ . We write  $\tilde{S}_j(\tilde{t}_N) := \tilde{S}_{j, \tilde{t}_N}(\tilde{t}_N)$  for the site frequency spectrum of the skeleton at time  $\tilde{t}_N$ . If a mutation occurs during an infinitesimal time interval  $[t, t + \Delta t]$ , the clone started by the cell carrying the mutation has size  $j$  at time  $\tilde{t}_N$  with probability  $\tilde{p}_j(\tilde{t}_N - t) + O(\Delta t)$ , where  $f(x) = O(x)$  means that there exists  $C > 0$  so that  $|f(x)| \leq Cx$  for sufficiently small  $x > 0$ . The expected number of mutations that accumulate in  $[t, t + \Delta t]$  and are present in  $j \geq 1$  skeleton cells at time  $\tilde{t}_N$  is therefore

$$\mathbb{E}[\tilde{S}_{j, \tilde{t}_N}(t + \Delta t)] - \mathbb{E}[\tilde{S}_{j, \tilde{t}_N}(t)] = wr_0 \Delta t \cdot e^{\lambda_0 t} \cdot \tilde{p}_j(\tilde{t}_N - t) + o(\Delta t),$$

where we use that  $\mathbb{E}[\tilde{Z}_0(t)] = e^{\lambda_0 t}$  is the mean skeleton size at time  $t$ , and  $f(x) = o(x)$  means that  $f(x)/x \rightarrow 0$  as  $x \rightarrow 0$ . This calculation is somewhat heuristic in that we have simply multiplied an expected mutation rate by an expected population size, which is in turn multiplied by the probability that a particular mutation ends up in  $j$  cells at time  $\tilde{t}_N$ . In our proof of part (1) of [Proposition 2](#) for the total population, we present a more detailed argument which can be used to obtain this expression more rigorously.

Integrating over time, and using that  $q_0 = \lambda_0/r_0$  by expression (2) and  $N = e^{\lambda_0 \tilde{t}_N}$  by expression (10), we obtain

$$\begin{aligned} \mathbb{E}[\tilde{S}_j(\tilde{t}_N)] &= \int_0^{\tilde{t}_N} wr_0 \tilde{p}_j(\tilde{t}_N - t) e^{\lambda_0 t} dt \\ &= (w/q_0)N \cdot \int_0^{\tilde{t}_N} (e^{\lambda_0 t}/N)(1 - e^{\lambda_0 t}/N)^{j-1} \cdot \lambda_0 (e^{\lambda_0 t}/N) dt. \end{aligned}$$

Substituting  $y := 1 - e^{\lambda_0 t}/N$ ,  $dy = -\lambda_0 (e^{\lambda_0 t}/N) dt$ , this implies

$$\begin{aligned} \mathbb{E}[\tilde{S}_j(\tilde{t}_N)] &= (w/q_0)N \cdot \int_0^{1-1/N} (1-y)^{j-1} dy \\ &= (w/q_0)N \cdot \left(1 - \frac{1}{N}\right)^j \left(\frac{1}{j(j+1)} + \frac{1}{N} \frac{1}{j+1}\right), \end{aligned}$$

the desired result. Clearly, for fixed  $j \geq 1$ , then as  $N \rightarrow \infty$ ,

$$\mathbb{E}[\tilde{S}_j(\tilde{t}_N)] \sim (w/q_0)N \cdot 1/(j(j+1)). \quad (\text{A.1})$$

The asymptotic expression (A.1) can also be derived more heuristically as follows, which gives another way of interpreting the expression. If the skeleton is observed at a large time  $t$ , the age  $s$  of an arbitrary mutation has approximate density  $\lambda_0 e^{-\lambda_0 s}$ . A mutation with age  $s$  at time  $t$  is found in  $j \geq 1$  skeleton cells at time  $t$  with probability  $\tilde{p}_j(s)$ . The probability that an arbitrary mutation is found in  $j \geq 1$  skeleton cells at time  $t$  is therefore, as  $t \rightarrow \infty$ ,

$$\int_0^\infty \tilde{p}_j(s) \cdot \lambda_0 e^{-\lambda_0 s} ds = \int_0^1 (1-y)^{j-1} dy = 1/(j(j+1)),$$

using the substitution  $y := 1 - e^{-\lambda_0 s}$ ,  $dy = \lambda_0 e^{-\lambda_0 s} ds$ . Next, we can compute the expected total number of mutations up until time  $\tilde{t}_N$  via

$$\int_0^{\tilde{t}_N} wr_0 e^{\lambda_0 s} ds = (w/q_0)(N - 1) \sim (w/q_0)N,$$

which is given as (16) in the main text. Finally, we can obtain (A.1) as the expected total number of mutations up

until time  $\tilde{t}_N$  multiplied by the probability  $1/(j(j+1))$  of finding an arbitrary mutation in  $j$  skeleton cells at time  $\tilde{t}_N$  as  $N \rightarrow \infty$ .

The distribution  $j \mapsto 1/(j(j+1))$  is a special case of the *Yule–Simon distribution*, which was originally computed by [Yule \(1925\)](#) as the distribution of the number of species within a genus, where a species mutates to a new species within the same genus at some rate  $s$ , and a genus mutates to a new genus at some rate  $g$ . In the previous paragraph, we have adapted Yule’s basic argument to our setting with  $g = s = \lambda_0$ . We refer to [Simkin and Roychowdhury \(2011\)](#) for a comprehensive discussion of how the Yule–Simon distribution and variants thereof have appeared in a wide variety of scientific contexts since its original conception.

- (2) In (8) of Section 3.1, we showed that on average,  $w p_0/q_0$  mutations accumulate on type-1 divisions in between two type-2 divisions, while type-2 divisions add  $w$  mutations and change the skeleton population size level. Since the type-1 mutations on level  $k = 1$  are the clonal mutations, the expected number of clonal mutations is  $w p_0/q_0$ , which is the  $j = N$  case of the desired result. For  $k = 2, \dots, N - 1$ , the expected number of mutations on level  $k$  is  $w + w p_0/q_0 = w/q_0$ , which includes the type-2 division that starts the level.

For  $1 \leq j \leq N - 1$ , let  $\tilde{h}_{(1,k-1)}^j$  be the probability that starting with one skeleton cell carrying a particular mutation and  $k - 1$  cells without it,  $j$  cells carry the mutation when the skeleton reaches size  $N$ . Since for levels  $k = 2, \dots, N - 1$ , there are  $w/q_0$  mutations on average per level, and each mutation on level  $k$  contributes  $\tilde{h}_{(1,k-1)}^j$  to the expected number of mutations found in  $j$  skeleton cells at level  $N$ , we obtain for  $1 \leq j \leq N - 1$ ,

$$\mathbb{E}[\tilde{S}_j(\tilde{\tau}_N)] = (w/q_0) \sum_{k=2}^{N-1} \tilde{h}_{(1,k-1)}^j + w \delta_{1,j} \\ = (w/q_0) \sum_{k=1}^{N-2} \tilde{h}_{(1,k)}^j + w \delta_{1,j},$$

where the extra  $w \delta_{1,j}$  term is due to mutations that occur on the final type-2 division that changes levels from  $N - 1$  to  $N$ , each of which is found in one skeleton cell. As was the case for the proof of part (1), the fact that for each level  $k$ , we can simply multiply the expected number of mutations with the probability that each particular mutation ends up in  $j$  cells when the skeleton reaches size  $N$  can be justified more rigorously using an argument similar to the one we present in part (2) of [Proposition 2](#) in [Appendix B](#).

It remains to compute the probabilities  $\tilde{h}_{(1,k)}^j$ . To this end, define a two-dimensional discrete-time Markov chain on the state space  $\{(\ell, m) : \ell, m \geq 1, \ell + m \leq N\}$ , where  $\ell$  is the number of skeleton cells carrying a particular mutation and  $m$  is the number of cells without it. Since each skeleton cell, with or without the mutation, divides into two cells at rate  $\lambda_0$ , the transition probabilities for this chain are given by

$$(\ell, m) \rightarrow (\ell + 1, m) \quad \text{w.p.} \quad \ell/(\ell + m),$$

$$(\ell, m) \rightarrow (\ell, m + 1) \quad \text{w.p.} \quad m/(\ell + m),$$

for  $\ell, m \geq 1$  and  $\ell + m < N$ . The states  $(\ell, N - \ell)$  for  $1 \leq \ell \leq N - 1$  are absorbing. A diagram for this Markov chain is shown in [Fig. A.13](#).

Let  $\tilde{h}_{(\ell,m)}^r$  denote the probability that the above chain is absorbed in state  $(r, N - r)$  when started from state  $(\ell, m)$ . It is immediate that  $\tilde{h}_{(\ell,m)}^r = 0$  if  $\ell > r$  or  $m > N - r$ . For  $(\ell, m)$  with  $\ell \leq r, m \leq N - r$  and  $\ell + m < N$ , by conditioning on whether the first transition out of state  $(\ell, m)$  is to  $(\ell + 1, m)$  or  $(\ell, m + 1)$ , we obtain the following recursion for  $\tilde{h}_{(\ell,m)}^r$ :

$$(\ell + m) \tilde{h}_{(\ell,m)}^r = \ell \tilde{h}_{(\ell+1,m)}^r + m \tilde{h}_{(\ell,m+1)}^r. \quad (\text{A.2})$$

The boundary conditions are  $\tilde{h}_{(\ell,N-\ell)}^r = \delta_{\ell,r}$  for  $1 \leq \ell \leq N - 1$ . It is actually possible to compute  $\tilde{h}_{(\ell,m)}^r$  directly as the sum of probabilities of all possible paths from  $(\ell, m)$  to  $(r, N - r)$  without using the above recursion. By noting that there are  $\binom{N-(\ell+m)}{r-\ell}$  possible paths, and that each path has the same probability, we can obtain

$$\tilde{h}_{(\ell,m)}^r = \binom{N-(\ell+m)}{r-\ell} \cdot \prod_{n=0}^{r-\ell-1} \frac{\ell+n}{\ell+m+n} \cdot \prod_{n=0}^{N-m-r-1} \frac{m+n}{r+m+n}, \quad (\text{A.3})$$

with  $\prod_{\emptyset} := 1$ . As verification, it is straightforward to check that (A.3) solves (A.2).

To obtain  $\tilde{h}_{(1,k)}^j$  for  $1 \leq k \leq N - 2$  and  $1 \leq j \leq N - 1$ , note first that  $\tilde{h}_{(1,k)}^j = 0$  for  $k > N - j$ . For  $1 \leq k \leq \min(N - j, N - 2)$ , we can simplify (A.3) as follows:

$$\begin{aligned} \tilde{h}_{(1,k)}^j &= \binom{N-(k+1)}{j-1} \cdot \prod_{n=0}^{j-2} \frac{1+n}{k+1+n} \cdot \prod_{n=0}^{N-k-j-1} \frac{k+n}{k+j+n} \\ &= \frac{(N-k-1)!}{(j-1)!(N-k-j)!} \cdot \frac{(j-1)!}{(k+1) \cdots (k+j-1)} \cdot \frac{k \cdots (N-j-1)}{(k+j) \cdots (N-1)} \\ &= k \cdot \frac{(N-k-1)!}{(N-k-j)!} \cdot \frac{(N-j-1)!}{(N-1)!} \\ &= \binom{N-j-1}{k-1} \binom{N-1}{j}^{-1}. \end{aligned}$$

Thus, for  $1 \leq j \leq N - 1$ ,

$$\begin{aligned} \mathbb{E}[\tilde{S}_j(\tilde{\tau}_N)] &= (w/q_0) \sum_{k=1}^{N-2} \tilde{h}_{(1,k)}^j + w \delta_{1,j} \\ &= (w/q_0) \sum_{k=1}^{\min(N-j, N-2)} \binom{N-j-1}{k-1} \binom{N-1}{j}^{-1} + w \delta_{1,j}. \end{aligned} \quad (\text{A.4})$$

For  $j = 1$ , it is easy to compute

$$\begin{aligned} \mathbb{E}[\tilde{S}_1(\tilde{\tau}_N)] &= (w/q_0)(1/(N-1))(\sum_{k=1}^{N-2} k) + w \\ &= (1/2)(w/q_0)(N-2) + w \\ &= (1/2)(w/q_0)N - w p_0/q_0. \end{aligned}$$

To simplify (A.4) for  $2 \leq j \leq N - 1$ , we first note that for any positive integers  $a$  and  $b$ ,

$$\int_0^1 t^{a-1} (1-t)^{b-1} dt = \frac{(a-1)!(b-1)!}{(a+b-1)!},$$

see e.g. Theorem 1.1.4 of [Andrews et al. \(1999\)](#). By observing that

$$N \int_0^1 t^k (1-t)^{N-1-k} dt = \binom{N-1}{k}^{-1},$$

we can rewrite (A.4) as follows:

$$\begin{aligned} \mathbb{E}[\tilde{S}_j(\tilde{\tau}_N)] &= (w/q_0)N \cdot \int_0^1 (\sum_{k=1}^{N-j} \binom{N-j-1}{k-1} t^k (1-t)^{N-1-k}) dt \\ &= (w/q_0)N \\ &\quad \cdot \int_0^1 t(1-t)^{j-1} (\sum_{k=1}^{N-j} \binom{N-j-1}{k-1} t^{k-1} (1-t)^{(N-j-1)-(k-1)}) dt. \end{aligned}$$

The sum inside the integral is the total probability mass ( $= 1$ ) of the binomial distribution with number of trials  $N - j - 1$  and success probability  $t$ . Therefore, for  $2 \leq j \leq N - 1$ ,

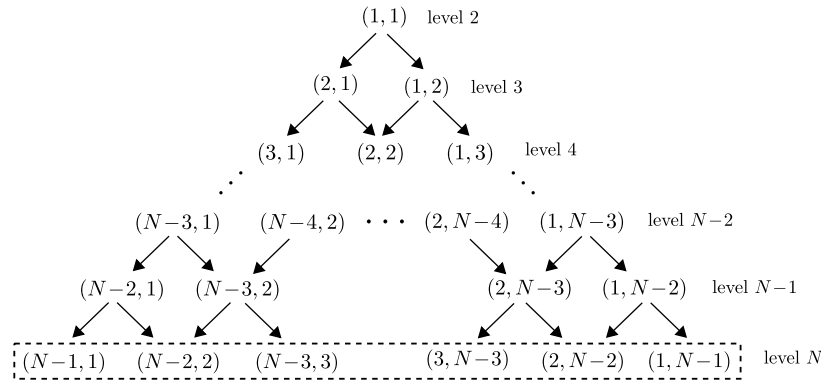
$$\begin{aligned} \mathbb{E}[\tilde{S}_j(\tilde{\tau}_N)] &= (w/q_0)N \cdot \int_0^1 t(1-t)^{j-1} dt \\ &= (w/q_0)N \cdot 1/(j(j+1)). \end{aligned}$$

This concludes the proof.  $\square$

## Appendix B. Proof of [Proposition 2](#)

In this section, we prove [Proposition 2](#) on the expected SFS of the total population. The proof strategy is the same as in the proof of [Proposition 1](#) for the skeleton. There are some added complications, however, since the tumor as a whole follows a birth–death process, whereas the skeleton subpopulation follows a pure-birth process.





**Fig. A.13.** A diagram of the discrete-time Markov chain on  $\{(\ell, m) : \ell, m \geq 1, \ell + m \leq N\}$ , where  $\ell$  is the number of skeleton cells carrying a particular mutation and  $m$  is the number of cells without it. Each type-2 division increases the skeleton population size level by one, and since skeleton cells do not die, the chain never returns to the lower levels. The states  $(\ell, N - \ell)$  with  $1 \leq \ell \leq N - 1$  are absorbing (dashed box), since we are only interested in the evolution up until level  $N$ .

In part (1), we wish to compute the expected number of mutations that accumulate in an infinitesimal time interval and end up in  $j \geq 1$  tumor cells at the fixed time  $t_N$ . In this case, we need to condition on survival up until time  $t_N$ , which was not necessary in the same computation for the skeleton. In the proof below, we first give an informal argument for how to handle the conditioning on survival, and then present detailed calculations at the end of the proof. Another complication is that the size distribution of a single-cell derived clone now has a more complex form than for the skeleton, which translates into more work simplifying the integral that results from adding up the infinitesimal interval contributions.

In part (2), we decompose into population size levels. To compute how many mutations accumulate on population size level  $k$ , we first compute how many times the population hits this level. This computation was not necessary for the skeleton, since the skeleton only increases in size. We then need to compute the probability that starting from one cell carrying a particular mutation and  $k-1$  cells without it,  $j$  cells carry the mutation when the population reaches size  $N$ . This probability can be computed as a hitting probability of a Markov chain, as in the proof of Proposition 1, but the Markov chain now has a more complicated structure. This time, we are not able to compute the hitting probabilities explicitly, and we instead provide a linear system which determines them.

### Proof of Proposition 2.

(1) We begin by defining

$$p_j(s) := \mathbb{P}(Z_0(s) = j | Z_0(0) = 1), \quad j \geq 0, s \geq 0,$$

as the size-distribution at time  $s$  of a single-cell derived clone. This distribution has an explicit expression: Setting

$$g(t) := \frac{p_0(e^{\lambda_0 t} - 1)}{e^{\lambda_0 t} - p_0} \quad \text{and} \quad h(t) := \frac{e^{\lambda_0 t} - 1}{e^{\lambda_0 t} - p_0},$$

we can write

$$p_0(t) = \mathbb{P}(Z_0(t) = 0 | Z_0(0) = 1) = g(t),$$

and for  $j \geq 1$ ,

$$\begin{aligned} p_j(t) &= \mathbb{P}(Z_0(t) = j | Z_0(0) = 1) \\ &= (1 - g(t))(1 - h(t))(h(t))^{j-1}, \end{aligned}$$

see e.g. (8) of Durrett (2015). Simplifying, we obtain

$$\begin{aligned} p_0(t) &= \frac{p_0(e^{\lambda_0 t} - 1)}{e^{\lambda_0 t} - p_0}, \\ p_j(t) &= \frac{q_0^2 e^{\lambda_0 t}}{(e^{\lambda_0 t} - p_0)^2} \cdot \left( \frac{e^{\lambda_0 t} - 1}{e^{\lambda_0 t} - p_0} \right)^{j-1}, \quad j \geq 1. \end{aligned} \quad (\text{B.1})$$

The probability that a single-cell derived clone is still alive at time  $t$  is then given by

$$\mathbb{P}(Z_0(t) > 0 | Z_0(0) = 1) = 1 - p_0(t) = q_0 e^{\lambda_0 t} / (e^{\lambda_0 t} - p_0). \quad (\text{B.2})$$

For  $0 \leq t \leq t_N$ , let  $S_{j,t_N}(t)$  denote the number of mutations that accumulate in  $[0, t]$  and are found in  $j \geq 1$  cells at time  $t_N$ . We write  $S_j(t_N) := S_{j,t_N}(t_N)$  for the site frequency spectrum at time  $t_N$ . Say a cell division occurs in an infinitesimal time interval  $[t, t + \Delta t]$ . The division results in  $w$  mutations on average, each assigned to one of the two daughter cells, and the clone started by this cell has size  $j \geq 1$  cells at time  $t_N$  with probability  $p_j(t_N - t) + o(\Delta t)$ . We wish to show that on the event  $\{Z_0(t_N) > 0\}$  of survival of the population up until time  $t_N$ , the expected number of mutations that accumulate in  $[t, t + \Delta t]$  and are found in  $j \geq 1$  cells at time  $t_N$  is

$$\begin{aligned} &\mathbb{E}[(S_{j,t_N}(t + \Delta t) - S_{j,t_N}(t)) \mathbf{1}_{\{Z_0(t_N) > 0\}}] \\ &= wr_0 \Delta t \cdot e^{\lambda_0 t} \cdot p_j(t_N - t) + o(\Delta t), \end{aligned} \quad (\text{B.3})$$

where we use that  $\mathbb{E}[Z_0(t)] = e^{\lambda_0 t}$ . It will then follow from (B.3) that

$$\begin{aligned} &\mathbb{E}[S_{j,t_N}(t + \Delta t) - S_{j,t_N}(t) | Z_0(t_N) > 0] \\ &= wr_0 e^{\lambda_0 t} \Delta t \cdot p_j(t_N - t) / (1 - p_0(t_N)) + o(\Delta t). \end{aligned} \quad (\text{B.4})$$

Note that (B.3) clearly holds for the semideterministic model in which the tumor bulk grows deterministically at rate  $\lambda_0$ , mutant clones arise at stochastic rate  $wr_0$ , and mutant clones grow stochastically. It is not obvious that (B.3) also holds for our fully stochastic model, since including the event  $\{Z_0(t_N) > 0\}$  of survival up until time  $t_N$  should presumably affect the expected population size at time  $t \leq t_N$ . The key is to observe that if a mutation occurs on a cell division at time  $t \leq t_N$  and ends up in  $j \geq 1$  cells at time  $t_N$ , the population is automatically alive at time  $t_N$ . The relevant survival event in (B.3) is therefore  $\{Z_0(t) > 0\}$ , and the relevant population size factor is  $\mathbb{E}[Z_0(t) \mathbf{1}_{\{Z_0(t) > 0\}}] = \mathbb{E}[Z_0(t)] = e^{\lambda_0 t}$ . Another potential concern in establishing (B.3) for our model is that we allow multiple mutations to occur per cell division. To not distract further from the main calculations, we assume that the reader is willing to accept (B.3) as true for the moment, and we provide a detailed mathematical argument for this expression at the end of the proof.

Using (B.4), we can integrate over time to obtain

$$\begin{aligned} \mathbb{E}[S_j(t_N)|Z_0(t_N) > 0] \\ = (1 - p_0(t_N))^{-1} \cdot \int_0^{t_N} w r_0 e^{\lambda_0 t} p_j(t_N - t) dt. \end{aligned} \quad (\text{B.5})$$

Focusing on the integral, we write

$$\begin{aligned} \int_0^{t_N} w r_0 e^{\lambda_0 t} p_j(t_N - t) dt \\ = (w/q_0) \cdot \int_0^{t_N} \frac{q_0^2 e^{\lambda_0 t_N} e^{\lambda_0 t}}{(e^{\lambda_0 t_N} - p_0 e^{\lambda_0 t})^2} \cdot \left( \frac{e^{\lambda_0 t_N} - e^{\lambda_0 t}}{e^{\lambda_0 t_N} - p_0 e^{\lambda_0 t}} \right)^{j-1} \cdot \lambda_0 e^{\lambda_0 t} dt \\ = w q_0 e^{\lambda_0 t_N} \cdot \int_0^{t_N} \frac{e^{\lambda_0 t}}{(e^{\lambda_0 t_N} - p_0 e^{\lambda_0 t})^2} \cdot \left( \frac{e^{\lambda_0 t_N} - e^{\lambda_0 t}}{e^{\lambda_0 t_N} - p_0 e^{\lambda_0 t}} \right)^{j-1} \cdot \lambda_0 e^{\lambda_0 t} dt. \end{aligned}$$

Set  $L := e^{\lambda_0 t_N}$ . Using the substitution  $x := e^{\lambda_0 t}$ ,  $dx = \lambda_0 e^{\lambda_0 t} dt$ , we obtain

$$\int_0^{t_N} w r_0 e^{\lambda_0 t} p_j(t_N - t) dt = w q_0 L \cdot \int_1^L \frac{x}{(L - p_0 x)^2} \cdot \left( \frac{L - x}{L - p_0 x} \right)^{j-1} dx.$$

We again change variables, this time  $y := (L - x)/(L - p_0 x)$ , in which case

$$\begin{aligned} x &= L(1 - y)/(1 - p_0 y), \\ dx &= -(q_0 L/(1 - p_0 y)^2) dy, \\ L - p_0 x &= q_0 L/(1 - p_0 y), \end{aligned}$$

and  $y = (L - 1)/(L - p_0) = 1 - q_0/(L - p_0)$  for  $x = 1$  and  $y = 0$  for  $x = L$ , which implies

$$\begin{aligned} \int_0^{t_N} w r_0 e^{\lambda_0 t} p_j(t_N - t) dt \\ = w L \cdot \int_0^{1 - q_0/(L - p_0)} (1 - p_0 y)^{-1} (1 - y) y^{j-1} dy. \end{aligned} \quad (\text{B.6})$$

We now apply (B.2) and (B.5) to see that

$$\begin{aligned} \mathbb{E}[S_j(t_N)|Z_0(t_N) > 0] \\ = w \cdot \frac{e^{\lambda_0 t_N} - p_0}{q_0} \cdot \int_0^{1 - q_0/(e^{\lambda_0 t_N} - p_0)} (1 - p_0 y)^{-1} (1 - y) y^{j-1} dy, \end{aligned}$$

and the desired result (20) follows from the fact that  $(e^{\lambda_0 t_N} - p_0)/q_0 = N$  by the definition of  $t_N$  in (18). It also follows that for fixed  $j \geq 1$ ,

$$\mathbb{E}[S_j(t_N)|Z_0(t_N) > 0] \sim w N \cdot \int_0^1 (1 - p_0 y)^{-1} (1 - y) y^{j-1} dy$$

as  $N \rightarrow \infty$ . To write the last expression as a sum, note that

$$(1 - p_0 y)^{-1} = \sum_{k=0}^{\infty} p_0^k y^k,$$

which is valid for all  $0 \leq p_0 < 1$  and  $0 \leq y \leq 1$ . It follows that

$$\begin{aligned} \int_0^1 (1 - p_0 y)^{-1} (1 - y) y^{j-1} dy &= \sum_{k=0}^{\infty} p_0^k \left( \int_0^1 (1 - y) y^{j+k-1} dy \right) \\ &= \sum_{k=0}^{\infty} \frac{p_0^k}{(j+k)(j+k+1)}. \end{aligned}$$

We conclude by establishing (B.3), which was

$$\begin{aligned} \mathbb{E}[(S_{j,t_N}(t + \Delta t) - S_{j,t_N}(t)) 1_{\{Z_0(t_N) > 0\}}] \\ = w r_0 \Delta t \cdot e^{\lambda_0 t} \cdot p_j(t_N - t) + o(\Delta t). \end{aligned}$$

We decompose according to population size at time  $t$ . Assume that  $Z_0(t) = k$  with  $k \geq 1$ , i.e. there are  $k$  cells at time  $t$ . Let  $D_{t,\Delta t}$  denote the event that exactly one of the  $k$  cells divides in the infinitesimal time interval  $[t, t + \Delta t]$ , and enumerate the  $k + 1$  cells after the cell division as  $Y_t^1, \dots, Y_t^{k+1}$ , where  $Y_t^1$  and  $Y_t^2$  are the two new cells. Let  $W$  denote the number of mutations that occur on the cell division, where  $W$  is a nonnegative integer-valued random variable with  $\mathbb{E}[W] = w$ , independent of  $(Z_0(t))_{t \geq 0}$ . For  $\ell \geq 1$ , let  $B_\ell$  be i.i.d. with  $\mathbb{P}(B_\ell = 1) = \mathbb{P}(B_\ell = 2) = 1/2$ , independent of  $(Z_0(t))_{t \geq 0}$  and  $W$ , and assign mutation number  $\ell$  to cell number  $B_\ell$  for  $1 \leq \ell \leq W$ . Finally, let

$Y_t^m(s)$  be the number of descendants of cell  $Y_t^m$  at time  $t + s$ , with  $Y_t^m(0) = 1$ . With this notation, define

$$\begin{aligned} A_{j,k,\ell}(t) &:= \{Z_0(t) = k\} \cap D_{t,\Delta t} \cap \{\ell \leq W\} \\ &\quad \cap \{Y_t^{B_\ell}(t_N - t) = j\} \cap \{Z_0(t_N) > 0\}. \end{aligned}$$

This is the event that the tumor survives to time  $t_N$ , that it consists of  $k$  cells at time  $t$ , that exactly one of the  $k$  cells divides in  $[t, t + \Delta t]$ , that at least  $\ell$  mutations occur on this division, and that mutation number  $\ell$  is found in  $j$  cells at time  $t_N$ . The reason we are interested in this event is that we can write

$$\begin{aligned} \mathbb{E}[(S_{j,t_N}(t + \Delta t) - S_{j,t_N}(t)) 1_{\{Z_0(t_N) > 0\}}] \\ = \sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} \mathbb{P}(A_{j,k,\ell}(t)) + o(\Delta t), \end{aligned}$$

where the  $o(\Delta t)$  term captures the possibility of more than one cell division in  $[t, t + \Delta t]$ .

To compute  $\mathbb{P}(A_{j,k,\ell}(t))$ , note first that

$$\begin{aligned} \mathbb{P}(A_{j,k,\ell}(t)) \\ = \mathbb{P}(\{Z_0(t) = k\} \cap D_{t,\Delta t} \cap \{\ell \leq W\} \cap \{Y_t^{B_\ell}(t_N - t) = j\}), \end{aligned}$$

since the survival event  $\{Z_0(t_N) > 0\}$  is implied by the other events. By independence,

$$\begin{aligned} \mathbb{P}(A_{j,k,\ell}(t)) \\ = \mathbb{P}(\ell \leq W) \cdot \mathbb{P}(\{Z_0(t) = k\} \cap D_{t,\Delta t} \cap \{Y_t^{B_\ell}(t_N - t) = j\}). \end{aligned}$$

To analyze the latter probability, note that since  $B_\ell$  is independent of  $(Z_0(t))_{t \geq 0}$ , and  $((Z_0(s))_{s \leq t}, (Y_t^1(s))_{s \geq 0}) \stackrel{d}{=} ((Z_0(s))_{s \leq t}, (Y_t^2(s))_{s \geq 0})$ , we can write

$$\begin{aligned} \mathbb{P}(\{Z_0(t) = k\} \cap D_{t,\Delta t} \cap \{Y_t^{B_\ell}(t_N - t) = j\}) \\ = \mathbb{P}(\{Z_0(t) = k\} \cap D_{t,\Delta t} \cap \{Y_t^1(t_N - t) = j\}). \end{aligned}$$

Using the Markov property, we can calculate the latter probability as

$$\begin{aligned} \mathbb{P}(\{Z_0(t) = k\} \cap D_{t,\Delta t} \cap \{Y_t^1(t_N - t) = j\}) \\ = \mathbb{P}(Z_0(t) = k) \cdot \mathbb{P}(D_{t,\Delta t} | Z_0(t) = k) \\ \cdot \mathbb{P}(Y_t^1(t_N - t) = j | Z_0(t) = k, D_{t,\Delta t}) \\ = \mathbb{P}(Z_0(t) = k) \cdot e^{-kr_0 \Delta t} k r_0 \Delta t \cdot (p_j(t_N - t) + O(\Delta t)). \end{aligned}$$

Combining the above, we obtain

$$\begin{aligned} \mathbb{E}[(S_{j,t_N}(t + \Delta t) - S_{j,t_N}(t)) 1_{\{Z_0(t_N) > 0\}}] \\ = \sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} \mathbb{P}(A_{j,k,\ell}(t)) + o(\Delta t) \\ = r_0 p_j(t_N - t) \Delta t \cdot \left( \sum_{\ell=1}^{\infty} \mathbb{P}(W \geq \ell) \right) \\ \cdot \left( \sum_{k=1}^{\infty} k \mathbb{P}(Z_0(t) = k) \right) + o(\Delta t) \\ = w r_0 \Delta t \cdot e^{\lambda_0 t} \cdot p_j(t_N - t) + o(\Delta t), \end{aligned}$$

where we use  $\sum_{\ell=1}^{\infty} \mathbb{P}(W \geq \ell) = \mathbb{E}[W] = w$  and  $\sum_{k=1}^{\infty} k \mathbb{P}(Z_0(t) = k) = \mathbb{E}[Z_0(t)] = e^{\lambda_0 t}$ . This concludes the proof.

- (2) Let  $(X_n)_{n \geq 0}$  denote the discrete-time jump process embedded in  $(Z_0(t))_{t \geq 0}$  that only keeps track of changes in population size. More precisely, if  $\sigma_n$  is the time of the  $n$ th jump of  $(Z_0(t))_{t \geq 0}$  for  $n \geq 1$ , then  $X_0 = 1$  and  $X_n = Z_0(\sigma_n)$  for  $n \geq 1$ . Since cells divide at rate  $r_0$  and die at rate  $d_0$ ,  $(X_n)_{n \geq 0}$  is a simple random walk, absorbed at 0, which moves up with probability  $a := r_0/(r_0 + d_0) = 1/(1 + p_0)$  and down with probability  $b = 1 - a = p_0/(1 + p_0)$ . Since we are only interested in what happens until the population either goes extinct or reaches level  $N$ , we treat  $N$  as an absorbing state. Define

$$T_k := \inf\{n \geq 0 : X_n = k\}, \quad 0 \leq k \leq N, \quad (\text{B.7})$$

as the (discrete) time at which the random walk first hits level  $k$ , with  $\inf \emptyset = \infty$ . Let  $\mathbb{P}_j$  denote the probability measure of  $(X_n)_{n \geq 0}$  when started at  $X_0 = j$ . By the gambler's ruin formula,

$$\mathbb{P}_j(T_k < T_0) = (1 - p_0^j)/(1 - p_0^k), \quad 0 \leq j \leq k. \quad (\text{B.8})$$

For  $1 \leq k \leq N - 1$ , let  $\Lambda_{k,k+1}$  denote the number of transitions from  $k$  to  $k + 1$ ,

$$\Lambda_{k,k+1} := \sum_{j=0}^{\infty} 1_{\{Z_j=k, Z_{j+1}=k+1\}},$$

and let  $\Lambda_k$  denote the number of visits to  $k$ ,

$$\Lambda_k := \sum_{j=0}^{\infty} 1_{\{Z_j=k\}}.$$

By the strong Markov property and (B.8), we can write

$$\mathbb{E}_1[\Lambda_k] = \mathbb{P}_1(T_k < T_0) \cdot \mathbb{E}_k[\Lambda_k] = \frac{q_0}{1-p_0^k} \cdot \mathbb{E}_k[\Lambda_k].$$

When the chain leaves state  $k$ , it moves up with probability  $1/(1+p_0)$  and down with probability  $p_0/(1+p_0)$ . Starting from  $k+1$ , the probability that the chain does not return to  $k$  (probability it is absorbed at  $N$ ) is  $q_0/(1-p_0^{N-k})$  by (B.8), and starting from  $k-1$ , the probability it does not return to  $k$  (probability it is absorbed at 0) is  $1 - (1-p_0^{k-1})/(1-p_0^k)$  again by (B.8). Thus, starting from  $k$ ,  $\Lambda_k$  has the geometric distribution with support  $\{1, 2, \dots\}$  and success probability

$$\frac{1}{1+p_0} \cdot \frac{q_0}{1-p_0^{N-k}} + \frac{p_0}{1+p_0} \cdot \left(1 - \frac{1-p_0^{k-1}}{1-p_0^k}\right) = \frac{q_0(1-p_0^N)}{(1+p_0)(1-p_0^k)(1-p_0^{N-k})}.$$

It follows that

$$\begin{aligned} \mathbb{E}_1[\Lambda_k] &= \frac{(1+p_0)(1-p_0^{N-k})}{1-p_0^N}, \\ \mathbb{E}_1[\Lambda_{k,k+1}] &= \frac{1}{1+p_0} \cdot \mathbb{E}_1[\Lambda_k] = \frac{1-p_0^{N-k}}{1-p_0^N}. \end{aligned} \quad (\text{B.9})$$

For  $1 \leq k \leq N - 1$ , define  $T_{k,k+1}^i$  as the (discrete) time of the  $i$ th transition from  $k$  to  $k + 1$  inductively by

$$T_{k,k+1}^i := \inf\{n > T_{k,k+1}^{i-1} : X_{n-1} = k, X_n = k + 1\}, \quad i \geq 1,$$

with  $T_{k,k+1}^0 := 0$  and  $\inf \emptyset = \infty$ . A transition from  $k$  to  $k + 1$  in  $(X_n)_{n \geq 0}$  occurs due to one of the  $k$  cells in the original process  $(Z_0(t))_{t \geq 0}$  dividing. Assume  $W_{i,k}$  mutations occur on the  $i$ th such transition, where  $W_{i,k}$  are i.i.d. nonnegative integer-valued random variables with  $\mathbb{E}[W_{i,k}] = w$ , independent of  $(X_n)_{n \geq 0}$ . Enumerate the cells at time  $T_{k,k+1}^i$  as  $Y_{i,k}^1, \dots, Y_{i,k}^{k+1}$ . By the same argument as laid out in part (1) above, we can assume that each mutation is assigned to the first cell. Let  $Y_{i,k}^m(n)$  be the number of descendants of cell  $Y_{i,k}^m$  at time step  $T_{k,k+1}^i + n$ , with  $Y_{i,k}^m(0) = 1$ . Then define the event

$$A_{j,k,i,\ell} := \{T_N < T_0, T_{k,k+1}^i < \infty, \ell \leq W_{i,k}, Y_{i,k}^1(T_N - T_{k,k+1}^i) = j\}.$$

This is the event that the random walk eventually hits level  $N$ , that it transitions at least  $i$  times from  $k$  to  $k + 1$  before doing so, that at least  $\ell$  mutations occur on the  $i$ th such transition, and that the  $\ell$ -th mutation is found in  $j$  cells at level  $N$ . We can then write

$$\begin{aligned} \mathbb{E}[S_j(\tau_N) | \tau_N < \infty] &= (\mathbb{P}_1(T_N < T_0))^{-1} \cdot \sum_{k=1}^{N-1} \sum_{i=1}^{\infty} \sum_{\ell=1}^{\infty} \mathbb{P}_1(A_{j,k,i,\ell}) \\ &= \frac{1-p_0^N}{q_0} \cdot \sum_{k=1}^{N-1} \sum_{i=1}^{\infty} \sum_{\ell=1}^{\infty} \mathbb{P}_1(A_{j,k,i,\ell}). \end{aligned}$$

To compute  $\mathbb{P}_1(A_{j,k,i,\ell})$ , note first that by independence,

$$\mathbb{P}_1(T_N < T_0, T_{k,k+1}^i < \infty, \ell \leq W_{i,k}, Y_{i,k}^1(T_N - T_{k,k+1}^i) = j)$$

$$= \mathbb{P}(\ell \leq W_{i,k}) \cdot \mathbb{P}_1(T_N < T_0, T_{k,k+1}^i < \infty, Y_{i,k}^1(T_N - T_{k,k+1}^i) = j).$$

By the strong Markov property,

$$\begin{aligned} \mathbb{P}_1(T_N < T_0, T_{k,k+1}^i < \infty, Y_{i,k}^1(T_N - T_{k,k+1}^i) = j) &= \mathbb{P}_1(T_N < T_0, Y_{i,k}^1(T_N - T_{k,k+1}^i) = j | T_{k,k+1}^i < \infty) \\ &\cdot \mathbb{P}_1(T_{k,k+1}^i < \infty) \\ &= \mathbb{P}_{k+1}(T_N < T_0, Y^1(T_N) = j) \cdot \mathbb{P}_1(T_{k,k+1}^i < \infty), \end{aligned}$$

where we restart the chain at the stopping time  $T_{k,k+1}^i$  with  $k + 1$  cells enumerated as  $Y^1, \dots, Y^{k+1}$ . Define  $h_{(1,k)}^{(j,N-j)} := \mathbb{P}_{k+1}(T_N < T_0, Y^1(T_N) = j)$  for the moment, i.e. the probability that starting with one cell carrying a particular mutation and  $k$  cells without it,  $j$  cells carry the mutation when the population reaches size  $N$ . We can then write

$$\begin{aligned} \mathbb{P}_1(T_N < T_0, T_{k,k+1}^i < \infty, Y_{i,k}^1(T_N - T_{k,k+1}^i) = j) &= h_{(1,k)}^{(j,N-j)} \cdot \mathbb{P}_1(T_{k,k+1}^i < \infty). \end{aligned}$$

Combining the above, and using  $\mathbb{E}_1[\Lambda_{k,k+1}] = (1 - p_0^{N-k})/(1 - p_0^N)$  by (B.9), we obtain

$$\begin{aligned} \mathbb{E}[S_j(\tau_N) | \tau_N < \infty] &= \frac{1-p_0^N}{q_0} \\ &\cdot \sum_{k=1}^{N-1} \left( \sum_{i=1}^{\infty} \left( \sum_{\ell=1}^{\infty} \mathbb{P}(W_{i,k} \geq \ell) \right) \mathbb{P}_1(T_{k,k+1}^i < \infty) \right) h_{(1,k)}^{(j,N-j)} \\ &= \frac{1-p_0^N}{q_0} \cdot \sum_{k=1}^{N-1} w \cdot \mathbb{E}_1[\Lambda_{k,k+1}] \cdot h_{(1,k)}^{(j,N-j)} \\ &= (w/q_0) \cdot \sum_{k=1}^{N-1} (1 - p_0^{N-k}) \cdot h_{(1,k)}^{(j,N-j)}, \end{aligned}$$

which is the desired result.

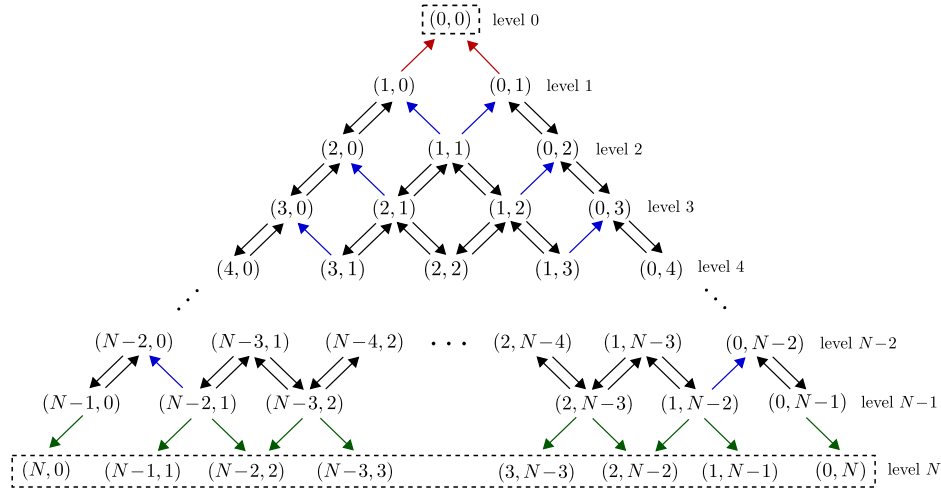
It remains to determine how the probabilities  $h_{(1,k)}^{(j,N-j)}$  for  $1 \leq j \leq N$  can be computed. As in the proof of (1) of Proposition 1, one can view  $h_{(r,N-r)}^{(r,N-r)}$  as the probability of absorption in state  $(r, N - r)$ , starting from state  $(\ell, m)$ , for a Markov chain on the state space  $\mathcal{S} := \{(\ell, m) : \ell, m \geq 0 \text{ and } \ell + m \leq N\}$ , where  $\ell$  is the number of cells carrying a particular mutation and  $m$  is the number of cells without it. The difference is that now, cells can die, so population level changes can be both up and down. The transition probabilities are therefore more complex in this case, and given by

$$\begin{aligned} (\ell, m) &\rightarrow (\ell + 1, m) \quad \text{w.p.} \quad \ell r_0 / ((\ell + m)(r_0 + d_0)) \\ &= \ell / ((\ell + m)(1 + p_0)), \\ (\ell, m) &\rightarrow (\ell - 1, m) \quad \text{w.p.} \quad \ell d_0 / ((\ell + m)(r_0 + d_0)) \\ &= \ell p_0 / ((\ell + m)(1 + p_0)), \\ (\ell, m) &\rightarrow (\ell, m + 1) \quad \text{w.p.} \quad m r_0 / ((\ell + m)(r_0 + d_0)) \\ &= m / ((\ell + m)(1 + p_0)), \\ (\ell, m) &\rightarrow (\ell, m - 1) \quad \text{w.p.} \quad m d_0 / ((\ell + m)(r_0 + d_0)) \\ &= m p_0 / ((\ell + m)(1 + p_0)), \end{aligned}$$

for  $(\ell, m)$  with  $\ell, m \geq 0$  and  $0 < \ell + m < N$ . The states  $(0, 0)$  and  $(\ell, N - \ell)$  with  $0 \leq \ell \leq N$  are absorbing. A diagram of the Markov chain is shown in Fig. B.14.

For given  $(r, s) \in A := \{(0, 0)\} \cup \{(r, s) : r, s \geq 0 \text{ and } r + s = N\}$  and  $(\ell, m)$  with  $\ell, m \geq 0$  and  $0 < \ell + m < N$ , by conditioning on the first transition out of state  $(\ell, m)$ , we can derive the following recursion for  $h_{(\ell,m)}^{(r,s)}$ :

$$\begin{aligned} (\ell + m)(1 + p_0) h_{(\ell,m)}^{(r,s)} &= \ell h_{(\ell+1,m)}^{(r,s)} + \ell p_0 h_{(\ell-1,m)}^{(r,s)} + m h_{(\ell,m+1)}^{(r,s)} + m p_0 h_{(\ell,m-1)}^{(r,s)}. \end{aligned}$$



**Fig. B.14.** A diagram of the discrete-time Markov chain on  $\mathcal{S} := \{(\ell, m) : \ell, m \geq 0, \ell + m \leq N\}$ , where  $\ell$  is the number of cells carrying a particular mutation and  $m$  is the number of cells without it. Contrary to the chain for the skeleton process of Proposition 1 (Fig. A.13), we now incorporate cell death, which means both that population level changes can be up and down, and that we add states of the form  $(\ell, m)$  with  $\ell = 0$  or  $m = 0$ . The states  $(0, 0)$  and  $\{(\ell, N - \ell)\}_{0 \leq \ell \leq N}$  are absorbing (dashed boxes), and the states  $\{(\ell, m) : \ell, m \geq 1, \ell + m < N\}$ ,  $\{(\ell, 0)\}_{1 \leq \ell \leq N-1}$  and  $\{(0, m)\}_{1 \leq m \leq N-1}$  form their respective communicating classes. Colored arrows indicate transitions out of communicating classes.

By the gambler's ruin formula (B.8), the boundary conditions are

$$\begin{aligned}
 h_{(\ell, 0)}^{(N, 0)} &= 1 - h_{(\ell, 0)}^{(0, 0)} = (1 - p_0^\ell)/(1 - p_0^N), \quad 0 \leq \ell \leq N, \\
 h_{(\ell, 0)}^{(r, s)} &= 0, \quad 0 \leq \ell \leq N, \quad (r, s) \notin \{(0, 0), (N, 0)\}, \\
 h_{(0, m)}^{(0, N)} &= 1 - h_{(0, m)}^{(0, 0)} = (1 - p_0^m)/(1 - p_0^N), \quad 0 \leq m \leq N, \\
 h_{(0, m)}^{(r, s)} &= 0, \quad 0 \leq m \leq N, \quad (r, s) \notin \{(0, 0), (0, N)\}, \\
 h_{(\ell, N-\ell)}^{(r, N-r)} &= \delta_{r, \ell}, \quad 1 \leq \ell \leq N-1, \\
 h_{(\ell, N-\ell)}^{(r, s)} &= 0, \quad 1 \leq \ell \leq N-1, \quad (r, s) \in \{(0, 0), (N, 0), (0, N)\}.
 \end{aligned} \tag{B.10}$$

This is the desired linear system.  $\square$

### Appendix C. Results for continuous mutation accumulation

In this section, we discuss how to derive the expected SFS of the skeleton and the total population under the continuous model of mutation accumulation (see Section 2.3). This requires minor modifications to the proofs of Propositions 1 and 2.

#### C.1. Skeleton spectrum

Under continuous mutation accumulation, the expected fixed-time spectrum of the skeleton is given by

$$\mathbb{E}[\tilde{S}_j(\tilde{t}_N)] = (v/\lambda_0)N \cdot \int_0^{1-1/N} (1-y)y^{j-1} dy,$$

which is the same as (12) of Proposition 1 with  $w/q_0$  replaced by  $v/\lambda_0$ , the effective mutation rate in the continuous-time model. We can use the same proof as in part (1) of Appendix A, simply replacing the mutation rate  $wr_0$  by  $v$  and noting that  $q_0 = \lambda_0/r_0$ . However, the expected fixed-size spectrum of the skeleton becomes

$$\begin{aligned}
 \mathbb{E}[\tilde{S}_j(\tilde{\tau}_N)] &= \begin{cases} (v/\lambda_0)N \cdot 1/(j(j+1)) - (v/\lambda_0)\delta_{1,j}, & 1 \leq j \leq N-1, \\ v/\lambda_0, & j = N. \end{cases} \tag{C.1}
 \end{aligned}$$

In the continuous model, mutations occur at rate  $v$  per unit time, and the effective type-2 cell divisions occur at rate  $\lambda_0$  per unit time. Thus, for  $1 \leq k \leq N-1$ , the number of mutations

that accumulate on skeleton population size level  $k$ , prior to the type-2 division that changes levels to  $k+1$ , has the geometric distribution with support  $\{0, 1, 2, \dots\}$  and success probability

$$k\lambda_0/(k\lambda_0 + kv) = \lambda_0/(\lambda_0 + v).$$

The expected number of mutations per level is therefore

$$(\lambda_0 + v)/\lambda_0 - 1 = v/\lambda_0,$$

which applies to all levels  $k$  with  $1 \leq k \leq N-1$ . In particular, there are  $v/\lambda_0$  ( $= w/q_0$ ) clonal mutations in the continuous model, as opposed to  $w p_0/q_0 = w/q_0 - w$  clonal mutations in the discrete model. Recall that in the latter model, mutations coincide with cell divisions, and clonal mutations come from the type-1 divisions that occur before the first type-2 division in the process. The first type-2 division adds  $w$  mutations, but it also changes levels, so mutations occurring on this division are not clonal. In the continuous model, all mutations occur in between cell divisions, which is why there is no such boundary effect. Similarly, in the discrete model, the very last type-2 division that changes the skeleton size to  $N$  adds  $w$  mutations on average that each ends up in one cell. Since this does not occur in the continuous model, the extra term  $-v/\lambda_0$  ( $= -w/q_0$ ) that appears for  $j=1$  in (C.1) differs from the extra term  $-w p_0/q_0 = -w/q_0 + w$  in the discrete model by  $w$ . These key differences between mutation accumulation in the two models are diagrammed in Fig. C.15.

#### C.2. Total population spectrum

Under continuous mutation accumulation, the expected fixed-time spectrum of the total population is given by

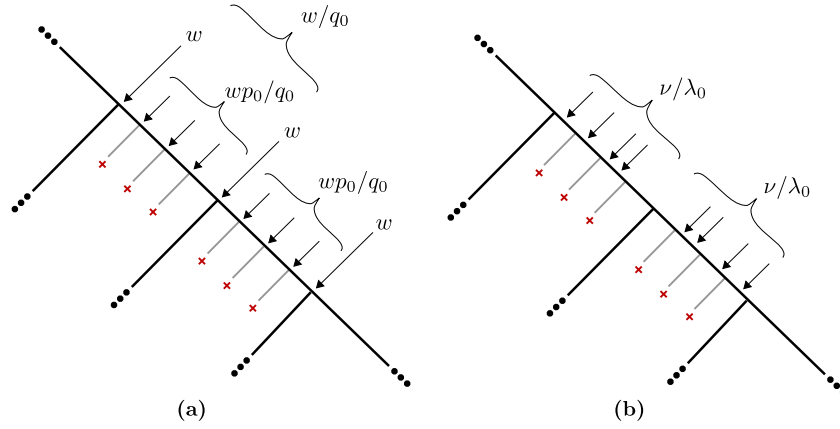
$$\mathbb{E}[S_j(t_N) | Z_N(t_N) > 0] = (v/r_0)N \cdot \int_0^{1-1/N} (1-p_0 y)^{-1} (1-y)y^{j-1} dy,$$

which is the same as (20) of Proposition 2 with  $w$  replaced by  $v/r_0$ . We can use the same proof as in part (1) of Appendix B, replacing the mutation rate  $wr_0$  by  $v$ . However, the expected fixed-size spectrum of the total population becomes, for  $1 \leq j \leq N$ ,

$$\mathbb{E}[S_j(\tau_N) | \tau_N < \infty] = (v/\lambda_0) \cdot \sum_{k=1}^{N-1} (1 - p_0^{N-k}) \cdot h_{(1, k-1)}^{(j, N-j)}.$$

In the continuous model, mutations no longer coincide with changes in population level, so instead of counting level changes





**Fig. C.15.** Mutation accumulation on the skeleton in the discrete vs. continuous model of mutation. **(a)** In the discrete model, mutations coincide with cell divisions. On average,  $wp_0/q_0$  mutations accumulate on type-1 divisions in between two type-2 divisions, and  $w$  mutations are added on each type-2 division. This results in  $wp_0/q_0 + w = w/q_0$  mutations on average per skeleton population size level, for all but the first level. **(b)** In the continuous model, all  $\nu/\lambda_0 (= w/q_0)$  mutations per level accumulate in between cell divisions. The differences between the discrete and continuous model result in slightly different behavior at the boundary values  $j = 1$  and  $j = N$  in the fixed-size spectrum (14) of Proposition 1 and fixed-size result (C.1) for the continuous model.

$\Lambda_{k,k+1}$  as we did in the proof of part (2) of Proposition 2, we need to compute the expected time spent at level  $k$ . We already know from (B.9) that the number of visits to level  $k$  in the embedded discrete-time chain  $(X_n)_{n \geq 0}$  has expected value

$$\mathbb{E}_1[\Lambda_k] = \frac{(1+p_0)(1-p_0^{N-k})}{1-p_0^N}.$$

During each visit to state  $k$  in the discrete-time chain, the time spent at population level  $k$  in the continuous-time process  $(Z_0(t))_{t \geq 0}$  is exponentially distributed with rate  $k(r_0 + d_0) = kr_0(1 + p_0)$ . It follows that the mean time spent on level  $k$  for  $1 \leq k \leq N-1$  is

$$\frac{1}{kr_0(1+p_0)} \cdot \frac{(1+p_0)(1-p_0^{N-k})}{1-p_0^N} = \frac{1-p_0^{N-k}}{kr_0(1-p_0^N)}.$$

Since mutations occur at rate  $k\nu$  per unit time on level  $k$ , we obtain

$$\begin{aligned} \mathbb{E}[S_j(\tau_N) | \tau_N < \infty] &= \frac{1-p_0^N}{q_0} \cdot \sum_{k=1}^{N-1} k\nu \cdot \frac{1-p_0^{N-k}}{kr_0(1-p_0^N)} \cdot h_{(1,k-1)}^{(j,N-j)} \\ &= (\nu/\lambda_0) \cdot \sum_{k=1}^{N-1} (1-p_0^{N-k}) \cdot h_{(1,k-1)}^{(j,N-j)}, \end{aligned} \quad (\text{C.2})$$

where we use that  $q_0 = \lambda_0/r_0$ . We now have  $h_{(1,k-1)}^{(j,N-j)}$  in the sum instead of  $h_{(1,k)}^{(j,N-j)}$  in (22) of Proposition 2 since mutations no longer coincide with population level changes. The first equality in (C.2) can be obtained rigorously following the same line of reasoning as in the proof of part (2) of Proposition 2.

#### Appendix D. Proof of Proposition 3

In this section, we prove Proposition 3 on the total mutational burden of the tumor, both under the fixed-time and fixed-size spectrum. The result follows from Proposition 2 using simple calculations.

##### Proof of Proposition 3.

- (1) Define  $M_j(t) := \sum_{k \geq j} S_k(t)$  as the cumulative number of mutations found in  $\geq j$  cells at time  $t$ . For fixed  $j \geq 1$ , by (20) of Proposition 2 and Fubini's theorem, the expected cumulative fixed-time spectrum can be written as

$$\begin{aligned} \mathbb{E}[M_j(t_N) | Z_0(t_N) > 0] \\ = wN \cdot \sum_{k=j}^{\infty} \left( \int_0^{1-1/N} (1-p_0 y)^{-1} (1-y) y^{k-1} dy \right) \end{aligned}$$

$$\begin{aligned} &= wN \cdot \int_0^{1-1/N} (1-p_0 y)^{-1} (1-y) \left( \sum_{k=j}^{\infty} y^{k-1} \right) dy \\ &= wN \cdot \int_0^{1-1/N} (1-p_0 y)^{-1} y^{j-1} dy \\ &= wN \cdot \sum_{k=0}^{\infty} p_0^k \left( \int_0^{1-1/N} y^{j+k-1} dy \right) \\ &= wN \cdot \sum_{k=0}^{\infty} \frac{p_0^k}{j+k} \left( 1 - \frac{1}{N} \right)^{j+k}. \end{aligned} \quad (\text{D.1})$$

To obtain the desired result, set  $j = 1$  in (D.1) and use that  $\sum_{k=1}^{\infty} x^k/k = -\log(1-x)$ .

- (2) By (22) of Proposition 2,

$$\begin{aligned} \mathbb{E}[M_1(\tau_N) | \tau_N < \infty] \\ = \sum_{j=1}^N \left( (w/q_0) \cdot \sum_{k=1}^{N-1} (1-p_0^{N-k}) \cdot h_{(1,k)}^{(j,N-j)} \right) \\ = (w/q_0) \cdot \sum_{k=1}^{N-1} (1-p_0^{N-k}) \cdot \left( \sum_{j=1}^N h_{(1,k)}^{(j,N-j)} \right), \end{aligned}$$

and the result follows from the fact that  $\sum_{(r,s) \in A} h_{(1,k)}^{(r,s)} = 1$ .  $\square$

#### Appendix E. Fixed-time vs. fixed-size total population spectrum

Here, we present a simple heuristic argument for why the fixed-size spectrum of the total population can be approximated by the fixed-time spectrum on  $j \ll N$  when  $N$  is sufficiently large. As we discussed in Section 4.5 of the main text, conditional on the nonextinction event  $\Omega_{\infty}$ , the tumor eventually grows at exponential rate  $\lambda_0$ . If  $s$  is the time it takes to go from population level  $k$  to population level  $N$ , and we assume that  $s$  can be treated as deterministic, we can write  $ke^{\lambda_0 s} = N$  i.e.  $e^{\lambda_0 s} = N/k$ , following e.g. Iwasa et al. (2006). We can then make the approximation  $h_{(1,k)}^{(j,N-j)} \approx p_j(s)$ , where  $h_{(1,k)}^{(j,N-j)}$  is defined as in the proof of part (2) of Proposition 2, and  $(p_j(s))_{j \geq 0}$  is the size-distribution at time  $s$  for a single-cell derived clone. Applying (B.1) with  $e^{\lambda_0 s} = N/k$ , we obtain

$$h_{(1,k)}^{(j,N-j)} \approx p_j(s) = \frac{q_0^{2Nk}}{(N-p_0k)^2} \cdot \left( \frac{N-k}{N-p_0k} \right)^{j-1}.$$

Then, observing that  $1 - p_0^{N-k} \approx 1$  when  $k \ll N$ , we can write

$$\begin{aligned} (w/q_0) \cdot \sum_{k=1}^{N-1} (1-p_0^{N-k}) \cdot h_{(1,k)}^{(j,N-j)} \\ \approx (w/q_0) \cdot \int_1^N \frac{q_0^{2Nk}}{(N-p_0k)^2} \cdot \left( \frac{N-k}{N-p_0k} \right)^{j-1} dk. \end{aligned}$$

Using the substitution  $y := (N-k)/(N-p_0k)$  and writing  $N-p_0 \approx N$ , this becomes the fixed-time spectrum (20) of Proposition 2.

## Appendix F. Derivation of expressions (24) and (25)

Here, we establish the asymptotic expressions (24) and (25) in the main text. To establish (24), fix  $0 < p_0 < 1$  and set

$$f_j(k) := p_0^k \cdot \frac{j(j+1)}{(j+k)(j+k+1)}, \quad j \geq 1, k \geq 0.$$

Clearly,  $f_j(k) \leq p_0^k$  for all  $j \geq 1$  and  $k \geq 0$ . Since  $\sum_{k=0}^{\infty} p_0^k = 1/q_0 < \infty$ , it follows from the dominated convergence theorem that

$$\lim_{j \rightarrow \infty} \sum_{k=0}^{\infty} f_j(k) = 1/q_0,$$

from which it follows that

$$wN \cdot \sum_{k=0}^{\infty} \frac{p_0^k}{(j+k)(j+k+1)} \sim (w/q_0)N \cdot 1/(j(j+1)), \quad j \rightarrow \infty.$$

To establish (25), fix  $j \geq 1$  and set

$$f_{p_0}(k) := \frac{p_0^k}{(j+k)(j+k+1)}, \quad 0 < p_0 < 1, k \geq 0.$$

Clearly,  $f_{p_0}(k) \leq 1/((j+k)(j+k+1))$  for all  $0 < p_0 < 1$  and  $k \geq 0$ . Since

$$\sum_{k=0}^{\infty} \frac{1}{(j+k)(j+k+1)} = \sum_{k=0}^{\infty} \left( \frac{1}{j+k} - \frac{1}{j+k+1} \right) = 1/j < \infty,$$

it follows from the dominated convergence theorem that

$$\lim_{p_0 \rightarrow 1} \sum_{k=0}^{\infty} f_{p_0}(k) = 1/j,$$

from which it follows that

$$wN \cdot \sum_{k=0}^{\infty} \frac{p_0^k}{(j+k)(j+k+1)} \sim wN \cdot 1/j, \quad p_0 \rightarrow 1.$$

## Appendix G. Derivation of expression (30)

Here, we establish expression (30) in the main text. By (20) of Proposition 2 and Fubini's theorem, we can write

$$\begin{aligned} \mathbb{E}[S_1(t_N)|Z_0(t_N) > 0] &= wN \cdot \int_0^{1-1/N} (1-p_0y)^{-1}(1-y)dy \\ &= wN \cdot \sum_{k=0}^{\infty} p_0^k \left( \int_0^{1-1/N} y^k(1-y)dy \right) \\ &= wN \cdot \sum_{k=0}^{\infty} p_0^k \left( \frac{1}{k+1} \left(1 - \frac{1}{N}\right)^{k+1} - \frac{1}{k+2} \left(1 - \frac{1}{N}\right)^{k+2} \right) \\ &= wN \cdot \sum_{k=0}^{\infty} \frac{p_0^k}{k+1} \left(1 - \frac{1}{N}\right)^{k+1} - wN \cdot \sum_{k=0}^{\infty} \frac{p_0^k}{k+2} \left(1 - \frac{1}{N}\right)^{k+2}. \quad (\text{G.1}) \end{aligned}$$

Using that  $\sum_{k=1}^{\infty} x^k/k = -\log(1-x)$ , the former term can be computed as

$$wN \cdot \sum_{k=0}^{\infty} \frac{p_0^k}{k+1} \left(1 - \frac{1}{N}\right)^{k+1} = -wN \cdot (1/p_0) \log(q_0 + p_0/N),$$

and the latter term can be computed as

$$\begin{aligned} wN \cdot \sum_{k=0}^{\infty} \frac{p_0^k}{k+2} \left(1 - \frac{1}{N}\right)^{k+2} &= wN \cdot (1/p_0^2) \sum_{k=0}^{\infty} \frac{p_0^{k+2}}{k+2} \left(1 - \frac{1}{N}\right)^{k+2} \\ &= wN \cdot (1/p_0^2) \left( -\log(q_0 + p_0/N) - p_0(1 - 1/N) \right). \end{aligned}$$

Combining with (G.1), we obtain

$$\begin{aligned} \mathbb{E}[S_1(t_N)|Z_0(t_N) > 0] &= wN \cdot (1/p_0) \left( 1 - 1/N + (q_0/p_0) \log(q_0 + p_0/N) \right), \end{aligned}$$

the desired result.

## Appendix H. Laws of large numbers

Here, we present simple calculations in support of the conjectured laws of large numbers (32) and (33) of the main text. As stated in the main text, conditional on the nonextinction event  $\Omega_{\infty}$ , we have  $Z_0(t) \sim Ye^{\lambda_0 t}$  as  $t \rightarrow \infty$  almost surely, where  $Y$  follows the exponential distribution with mean  $1/q_0$  (Theorem 1 of Durrett (2015)). For the fixed-time spectrum, the number of

mutations that accumulate in  $[0, t_N]$  and are found in  $j \geq 1$  cells at time  $t_N$  is then approximately

$$S_j(t_N) \approx \int_0^{t_N} wr_0 \cdot Ye^{\lambda_0 t} \cdot p_j(t_N - t) dt.$$

From (B.6) in the proof of Proposition 2, we know that

$$\begin{aligned} \int_0^{t_N} wr_0 e^{\lambda_0 t} p_j(t_N - t) dt &= we^{\lambda_0 t_N} \cdot \int_0^{1-q_0/(e^{\lambda_0 t_N} - p_0)} (1-p_0y)^{-1}(1-y)y^{j-1} dy \\ &\sim wq_0N \cdot \int_0^1 (1-p_0y)^{-1}(1-y)y^{j-1} dy, \quad N \rightarrow \infty, \end{aligned}$$

where we use that  $e^{\lambda_0 t_N} = q_0N + p_0$  by the definition of  $t_N$  in (18). This implies that

$$S_j(t_N) \approx q_0Y \cdot wN \cdot \int_0^1 (1-p_0y)^{-1}(1-y)y^{j-1} dy$$

for large  $N$ . Since  $Y$  has the exponential distribution with mean  $1/q_0$ ,  $q_0Y$  has the exponential distribution with mean 1. This suggests (32) in the main text.

For the fixed-size spectrum, note that if  $N$  is large, then at time  $\tau_N - t$ , we can write  $Z_0(\tau_N - t) \approx Ne^{-\lambda_0 t}$ . The number of mutations that accumulate in  $[0, \tau_N]$  and are found in  $j \geq 1$  cells at time  $\tau_N$  is then approximately

$$\begin{aligned} S_j(\tau_N) &\approx \int_0^{\tau_N} wr_0 \cdot Ne^{-\lambda_0 t} \cdot p_j(t) dt \\ &= Ne^{-\lambda_0 \tau_N} \cdot \int_0^{\tau_N} wr_0 e^{\lambda_0 t} p_j(\tau_N - t) dt. \end{aligned}$$

Again using (B.6) from the proof of Proposition 2, we can write

$$\begin{aligned} \int_0^{\tau_N} wr_0 e^{\lambda_0 t} p_j(\tau_N - t) dt &= we^{\lambda_0 \tau_N} \cdot \int_0^{1-q_0/(e^{\lambda_0 \tau_N} - p_0)} (1-p_0y)^{-1}(1-y)y^{j-1} dy, \end{aligned}$$

from which it follows that

$$\begin{aligned} S_j(\tau_N) &\approx wN \cdot \int_0^{1-q_0/(e^{\lambda_0 \tau_N} - p_0)} (1-p_0y)^{-1}(1-y)y^{j-1} dy \\ &\approx wN \cdot \int_0^1 (1-p_0y)^{-1}(1-y)y^{j-1} dy \end{aligned}$$

for  $N$  large. This suggests (33) in the main text.

We finally mention that it is straightforward to prove a law of large numbers for a simplified version of our model, where the tumor bulk grows deterministically ( $Z_0(t) = e^{\lambda_0 t}$ ), mutant clones arise at stochastic rate  $wr_0$ , and mutant clones grow stochastically. To state the result, let  $\hat{S}_j(\tilde{t}_N)$  denote the number of mutations found in  $j \geq 1$  cells at time  $\tilde{t}_N$  under the simplified model, where  $\tilde{t}_N$  is given by (10), i.e.  $e^{\lambda_0 \tilde{t}_N} = N$ . We want to show that

$$\hat{S}_j(\tilde{t}_N) \sim wN \cdot \int_0^1 (1-p_0y)^{-1}(1-y)y^{j-1} dy \quad (\text{H.1})$$

as  $N \rightarrow \infty$  almost surely. Note that the limit is a constant since we assume deterministic growth of the tumor bulk. For  $0 \leq t \leq \tilde{t}_N$ , let  $\hat{N}_{j,\tilde{t}_N}(t)$  denote the number of mutant clones created in  $[0, t]$  that have size  $j \geq 1$  at time  $\tilde{t}_N$ . Then  $(\hat{N}_{j,\tilde{t}_N}(t))_{0 \leq t \leq \tilde{t}_N}$  is an inhomogeneous Poisson process with rate function  $\hat{\lambda}(t) = wr_0 e^{\lambda_0 t} p_j(\tilde{t}_N - t)$  and mean function

$$\hat{m}(t) = \int_0^t \hat{\lambda}(s) ds, \quad 0 \leq t \leq \tilde{t}_N.$$

Set  $\hat{N}_j(\tilde{t}_N) := \hat{N}_{j,\tilde{t}_N}(\tilde{t}_N)$ . By (B.6) in the proof of Proposition 2, and the fact that  $e^{\lambda_0 \tilde{t}_N} = N$ ,

$$\begin{aligned} \hat{m}(\tilde{t}_N) &= wN \cdot \int_0^{1-q_0/(N-p_0)} (1-p_0y)^{-1}(1-y)y^{j-1} dy \\ &\sim wN \cdot \int_0^1 (1-p_0y)^{-1}(1-y)y^{j-1} dy, \quad N \rightarrow \infty. \end{aligned}$$

Then, by a simple Poisson concentration inequality, see Theorem 1 of Cannone (2017),

$$\begin{aligned} \mathbb{P}(|\hat{N}_j(\tilde{t}_N)/\hat{m}(\tilde{t}_N) - 1| > (\hat{m}(\tilde{t}_N))^{-1/3}) &= \mathbb{P}(|\hat{N}_j(\tilde{t}_N) - \hat{m}(\tilde{t}_N)| > (\hat{m}(\tilde{t}_N))^{2/3}) \\ &\leq 2 \exp(-(\hat{m}(\tilde{t}_N))^{1/3} / (2(1 + (\hat{m}(\tilde{t}_N))^{-1/3}))). \end{aligned}$$

Since  $\hat{m}(\tilde{\epsilon}_N)$  is of order  $N$  as  $N \rightarrow \infty$ , it follows from the Borel–Cantelli lemma that  $\hat{N}_j(\tilde{\epsilon}_N)/\hat{m}(\tilde{\epsilon}_N) \rightarrow 1$  as  $N \rightarrow \infty$  almost surely. Since  $\hat{S}_j(\tilde{\epsilon}_N) = \hat{N}_j(\tilde{\epsilon}_N)$ , we have the result.

## References

- Achaz, G., 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183 (1), 249–258.
- Andrews, G.E., Askey, R., Roy, R., 1999. *Special Functions*. 71, Cambridge University Press.
- Antal, T., Krapivsky, P., 2011. Exact solution of a two-type branching process: models of tumor progression. *J. Stat. Mech. Theory Exp* 2011 (08), P08018.
- Armitage, P., Doll, R., 1954. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer* 8 (1), 1.
- Armitage, P., Doll, R., 1957. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *Br. J. Cancer* 11 (2), 161.
- Avanzini, S., Antal, T., 2019. Cancer recurrence times from a branching process model. *PLoS Comput. Biol* 15 (11), e1007423.
- Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K.W., Vogelstein, B., Nowak, M.A., 2010. Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci. USA* 107 (43), 18545–18550.
- Bozic, I., Gerold, J.M., Nowak, M.A., 2016. Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLoS Comput. Biol* 12 (2), e1004731.
- Bozic, I., Paterson, C., Waclaw, B., 2019. On measuring selection in cancer from subclonal mutation frequencies. *PLoS Comput. Biol* 15 (9), e1007368.
- Bozic, I., Reiter, J.G., Allen, B., Antal, T., Chatterjee, K., Shah, P., Moon, Y.S., Yaqubie, A., Kelly, N., Le, D.T., et al., 2013. Evolutionary dynamics of cancer in response to targeted combination therapy. *ELife* 2, e00747.
- Burrell, R.A., McGranahan, N., Bartek, J., Swanton, C., 2013. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501 (7467), 338–345.
- Cannone, C., 2017. A short note on Poisson tail bounds. <http://www.cs.columbia.edu/~ccanonne/files/misc/2017-poissonconcentration.pdf>.
- Caravagna, G., Heide, T., Williams, M.J., Zapata, L., Nichol, D., Chkhaidze, K., Cross, W., Cresswell, G.D., Werner, B., Acar, A., et al., 2020. Subclonal reconstruction of tumors by using machine learning and population genetics. *Nat. Genet* 52 (9), 898–907.
- Champagnat, N., Lambert, A., Richard, M., 2012. Birth and death processes with neutral mutations. *Int. J. Stoch. Anal* 2012.
- Cheek, D., Antal, T., 2018. Mutation frequencies in a birth–death branching process. *Ann. Appl. Probab* 28 (6), 3922–3947.
- Davis, A., Gao, R., Navin, N., 2017. Tumor evolution: Linear, branching, neutral or punctuated? *Biochim. Biophys. Acta Rev. Cancer* 1867 (2), 151–161.
- Del Monte, U., 2009. Does the cell number 109 still really fit one gram of tumor tissue? *Cell Cycle* 8 (3), 505–506.
- Dinh, K.N., Jaksik, R., Kimmel, M., Lambert, A., Tavaré, S., et al., 2020. Statistical inference for the evolutionary history of cancer genomes. *Stat. Sci* 35 (1), 129–144.
- Durrett, R., 2008. *Probability Models for DNA Sequence Evolution*. Springer Science & Business Media.
- Durrett, R., 2013. Population genetics of neutral mutations in exponentially growing cancer cell populations. *Ann. Appl. Probab* 23 (1), 230.
- Durrett, R., 2015. Branching process models of cancer. In: *Branching Process Models of Cancer*. Springer, pp. 1–63.
- Fay, J.C., Wu, C.-I., 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155 (3), 1405–1413.
- Fu, Y.-X., Li, W.-H., 1993. Statistical tests of neutrality of mutations. *Genetics* 133 (3), 693–709.
- George, A., 1973. Nested dissection of a regular finite element mesh. *SIAM J. Numer. Anal* 10 (2), 345–363.
- Griffiths, R.C., Pakes, A.G., 1988. An infinite-alleles version of the simple branching process. *Adv. Appl. Probab* 489–524.
- Heide, T., Zapata, L., Williams, M.J., Werner, B., Caravagna, G., Barnes, C.P., Graham, T.A., Sottoriva, A., 2018. Reply to ‘Neutral tumor evolution?’. *Nat. Genet* 50 (12), 1633–1637.
- Iwasa, Y., Nowak, M.A., Michor, F., 2006. Evolution of resistance during clonal expansion. *Genetics* 172 (4), 2557–2566.
- Jones, S., Chen, W.-d., Parmigiani, G., Diehl, F., Beerenwinkel, N., Antal, T., Traulsen, A., Nowak, M.A., Siegel, C., Velculescu, V.E., et al., 2008. Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl. Acad. Sci. USA* 105 (11), 4283–4288.
- Keller, P., Antal, T., 2015. Mutant number distribution in an exponentially growing population. *J. Stat. Mech. Theory Exp* 2015 (1), P01011.
- Kessler, D.A., Levine, H., 2013. Large population solution of the stochastic Luria–Delbrück evolution model. *Proc. Natl. Acad. Sci. USA* 110 (29), 11682–11687.
- Kessler, D.A., Levine, H., 2015. Scaling solution in the large population limit of the general asymmetric stochastic Luria–Delbrück evolution process. *J. Stat. Phys* 158 (4), 783–805.
- Kimura, M., 1968. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet. Res* 11 (3), 247–270.
- Kimura, M., 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61 (4), 893.
- Kingman, J.F., 1982a. On the genealogy of large populations. *J. Appl. Probab* 19, 27–43.
- Kingman, J.F.C., 1982b. The coalescent. *Stoch. Process. their Appl* 13 (3), 235–248.
- Knudson, A.G., 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. USA* 68 (4), 820–823.
- Komarova, N.L., Wu, L., Baldi, P., 2007. The fixed-size Luria–Delbrück model with a nonzero death rate. *Math. Biosci* 210 (1), 253–290.
- Kuipers, J., Jahn, K., Raphael, B.J., Beerenwinkel, N., 2017. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res* 27 (11), 1885–1894.
- Lambert, A., 2009. The allelic partition for coalescent point processes. *Markov Process. Relat. Fields* 15 (3), 359–386.
- Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., Chen, K., Dong, L., Cao, L., Tao, Y., et al., 2015. Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proc. Natl. Acad. Sci. USA* 112 (47), E6496–E6505.
- McDonald, T.O., Chakrabarti, S., Michor, F., 2018. Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution. *Nat. Genet* 50 (12), 1620–1623.
- McGranahan, N., Swanton, C., 2017. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* 168 (4), 613–628.
- Nowell, P.C., 1976. The clonal evolution of tumor cell populations. *Science* 194 (4260), 23–28.
- O’Connell, N., 1993. Yule process approximation for the skeleton of a branching process. *J. Appl. Probab* 30 (3), 725–729.
- Ohtsuki, H., Innan, H., 2017. Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. *Theor. Popul. Biol* 117, 43–50.
- Pakes, A.G., 1989. An infinite alleles version of the Markov branching process. *J. Aust. Math. Soc* 46 (1), 146–169.
- Rew, D., Wilson, G., 2000. Cell production rates in human tissues and tumours and their significance. Part II: clinical data. *Eur. J. Surg. Oncol* 26 (4), 405–417.
- Simkin, M.V., Roychowdhury, V.P., 2011. Re-inventing Willis. *Phys. Rep* 502 (1), 1–35.
- Sottoriva, A., Kang, H., Ma, Z., Graham, T.A., Salomon, M.P., Zhao, J., Marjoram, P., Siegmund, K., Press, M.F., Shibata, D., et al., 2015. A big bang model of human colorectal tumor growth. *Nat. Genet* 47 (3), 209–216.
- Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123 (3), 585–595.
- Tarabichi, M., Martincorena, I.n., Gerstung, M., Leroi, A.M., Markowitz, F., Spellman, P.T., Morris, Q.D., Lingjærde, O.C., Wedge, D.C., Van Loo, P., 2018. Neutral tumor evolution? *Nat. Genet* 50 (12), 1630–1633.
- Tomasetti, C., Vogelstein, B., Parmigiani, G., 2013. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl. Acad. Sci. USA* 110 (6), 1999–2004.
- Turajlic, S., Sottoriva, A., Graham, T., Swanton, C., 2019. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet* 20 (7), 404–416.
- Venkatesan, S., Swanton, C., 2016. Tumor evolutionary principles: how intratumor heterogeneity influences cancer treatment and outcome. *Am. Soc. Clin. Oncol. Educ. Book* 36, e141–e149.
- Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Kinzler, K.W., 2013. Cancer genome landscapes. *Science* 339 (6127), 1546–1558.
- Watterson, G., 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol* 7 (2), 256–276.
- Werner, B., Case, J., Williams, M.J., Chkhaidze, K., Temko, D., Fernández-Mateos, J., Cresswell, G.D., Nichol, D., Cross, W., Spiteri, I., et al., 2020. Measuring single cell divisions in human tissues from multi-region sequencing data. *Nat. Commun* 11 (1), 1–9.
- Werner, B., Williams, M.J., Barnes, C.P., Graham, T.A., Sottoriva, A., 2018. Reply to ‘Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution’. *Nat. Genet* 50 (12), 1624–1626.
- Williams, M.J., Werner, B., Barnes, C.P., Graham, T.A., Sottoriva, A., 2016. Identification of neutral tumor evolution across cancer types. *Nat. Genet* 48 (3), 238.

- Williams, M.J., Werner, B., Heide, T., Curtis, C., Barnes, C.P., Sottoriva, A., Graham, T.A., 2018. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet* 50 (6), 895–903.
- Wu, X., Kimmel, M., 2013. Modeling neutral evolution using an infinite-allele Markov branching process. *Int. J. Stoch. Anal* 2013, 1–10.
- Yule, G.U., 1925. II.—A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. *Philos. Trans. R. Soc. Lond., B, Biol. Sci* 213 (402–410), 21–87.
- Zeng, K., Fu, Y.-X., Shi, S., Wu, C.-I., 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174 (3), 1431–1439.