Heterogeneity in evolutionary models of tumor progression

Rick Durrett, Jasmine Foo, Kevin Leder, John Mayberry and Franziska Michor

July 14, 2010

sec-intro

1 Introduction

All cancers and tumor types display a striking variability among the cancer cells within a single tumor. Tumors are composed of a heterogeneous mixture of cell types with distinct genotypes, morphologies, metabolic activities and behaviors such as proliferation rate, antigen expression, drug response and metastatic potential [16, 20]. For example, molecular and phenotypic analysis of cells from breast carcinomas reveal defined subpopulations with distinct gene expression and (epi)genetic profiles [33]. Heterogeneity and subpopulations within single tumors have also been demonstrated via flow cytometry in cervical cancers and lymph node metastases [27]. Cytogenetic heterogeneity such as variation in ploidy and chromosomal structure has been discovered among cells within breast tumors and leukemias [35]. Genetic clonal diversity has also been observed in individual pre-malignant lesions in Barrett's esophagus, a condition associated with increased risk of developing esophageal adenocarcinoma [21, 24]. Virtually every major type of human cancer and biological subtype has been shown to contain distinct cell subpopulations with differing heritable alterations [20, 16, 25].

Tumor heterogeneity has direct clinical implications on disease classification and prognosis, as well as on treatment efficacy and drug target identification [25, 20]. The degree of genetic clonal diversity in Barrett's esophagus has been correlated to clinical progression to esophageal cancer [24]. In prostate carcinomas, tumor heterogeneity has been cited as a key factor in pretreatment underestimation of tumor aggressiveness and incorrect assessment of DNA ploidy status of tumors [34, 15]. Heterogeneity has long been implicated in the development of resistance to cancer therapies after an initial period of response [13, 25], as well as in the development of metastases [11]. In addition, tumor heterogeneity has been shown to hamper the precision of microarray-based analyses of gene expression patterns, which are currently widely used for the identification of genes associated with specific tumor types [28]. These issues underscore the importance of obtaining a more detailed understanding of the origin and temporal evolution of heterogeneity during tumorigenesis.

The clonal evolution model of carcinogenesis states that tumors are monoclonal, i.e. originating from a single abnormal cell, and that over time the descendants of this ancestral

cell acquire various combinations of mutations [20, 25]. Under this model, genetic drift and natural selection drive the progression and diversity of the tumor. As the tumor progresses, genetic instability and uncontrolled proliferation allow the production of cells with additional alterations which may confer characteristics such as drug resistance, resistance to cell death signaling, metastatic proclivity and increased mutation and proliferation rates. The resulting distinct subpopulations evolve, in turn produce new mutations and subpopulations as time progresses.

Mathematical models of tumor heterogeneity can be found in the literature, many of which consider the dynamics of distinct subpopulations evolving under selective pressure [26, 4, 5, 18, 14, 19, 31, 10, 3]. In many of these models, the subpopulations represent predefined phenotypes, such as the drug -sensitive and -resistant cells of a tumor, and the growth dynamics of these populations are explored. In this work we consider a stochastic evolutionary model of tumorigenesis in an exponentially growing population, wherein genetic alterations confer random fitness changes. Under this model, we address questions surrounding the extent of genetic diversity in tumor subpopulations and its evolution in time. This work is an extension of a previous paper in which we investigated the effect of the random mutational fitness distribution on the growth kinetics of the tumor [8]. The paper is organized as follows: in section 2 we introduce the model and discuss the different types of subpopulations that emerge in simulations. In section 3, we state some useful results from [10, 8] and use these results to investigate heterogeneity between populations which have accumulated varying numbers of mutations. We also derive useful approximations for (i) the size of the subpopulation of all individuals with k mutations and (ii) the waiting time until we see the first individual with k mutations appear. In section 4, we analyze the effects of our model parameters on the degree of genotypic heterogeneity within the population of cells with a single genetic alteration by considering two quantitative measures of heterogeneity: Simpson's index and the proportion of cells which come from the largest family of genotypically identical cells. We conclude in Section 5 with a discussion of our results.

2 Model Description

sec-model

We consider a multi-type branching process model of tumorigenesis in which mutations confer an additive change to the birth rate of the cell. This additive change is drawn according to a probability distribution ν which we refer to as the fitness distribution. In our terminology, type-i cells have accumulated $i \geq 0$ mutations. The initial population consists entirely of type-0 cells that give birth at rate a_0 to new type-0 cells and produce type-1 cells at rate u_1. We refer to u_1 as the mutation rate for type-0 cells. We assume that all cells in the population die at rate $b_0 < a_0$, and that the population of type-0 cells starts at a sufficiently large population V_0 so that we can approximate its size by $Z_0(t) = V_0 e^{\lambda_0 t}$, where $\lambda_0 = a_0 - b_0$. When a type-0 cell produces a type-1 cell, the new cell gives birth to type-1 cells at rate $a_0 + X$, where $X \geq 0$ is drawn according to the distribution ν and produces type-2 cells at rate u_2. In general, a type-$(k-1)$ cell with birth rate a produces a new type-k cell at rate u_k and the new type-k cell assumes an increased birth rate $a + x$ where $x \geq 0$ is drawn

according to ν. We suppose that each type-k cell produced by a type-$(k-1)$ cell starts a genetically distinct lineage of cells and we refer to the set of all of its type-k descendants as its family. We also let $Z_k(t)$ denote the total number of type-k cells in the population at time t and when we refer to the kth wave or generation of mutants, we mean the set of all type-k cells. We denote the total population at time t by $Z(t)$.

We will restrict our attention in this paper to fitness distributions concentrated on $[0, b]$ for some $b > 0$. If the fitness distribution is unbounded with tail $\nu(x, \infty) \sim Kx^\beta e^{-\gamma x^\alpha}$, then it is shown in [8] that the growth rate of the number of first generation mutants is super-exponential and hence this choice of distribution is unrealistic for modeling tumorigenesis. We shall discuss two distinct classes of distributions:

(i) ν is discrete and assigns mass g_i to a finite number of values $b_1 < b_2 < \cdots < b_N = b$.

(ii) ν is continuous with a bounded density $g(x)$ that is continuous and positive at b.

The special case of discrete distributions with $N = 1$ (i.e. deterministic fitness advances) was first studied in [17, 14] and asymptotic results for this model were obtained in [10]. A similar discrete time branching process model is considered in [3]. There, cells can accumulate "driver" and "passenger" mutations. The former decrease the death rate of cells while the latter provide no selective advantage. This essentially corresponds to a special case of (i) above with $N = 2$ and $b_1 = 0$. Asymptotic results for general continuous distributions were obtained in [8]. Theorem 1 below contains a summary of results which are relevant for our current discussion.

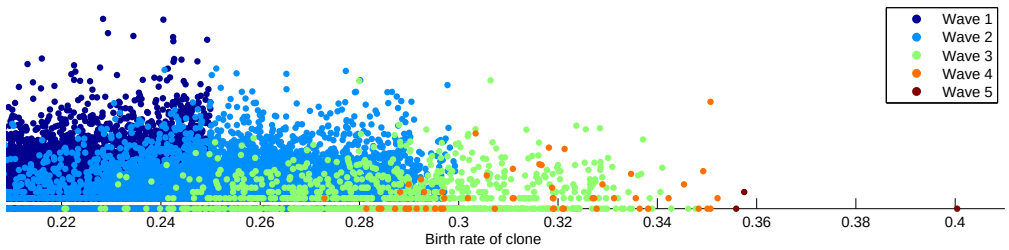
Figure 1 shows a snapshot of the population decomposition in a sample simulation of our model in case (ii) with $\nu \sim$ Uniform[0,0.05]. Here we have started with a single type-0 cell that has birth rate $a_0 = 0.2$ and death rate $b_0 = 0.1$. The type-0 population is not shown, but has reached $O(10^6)$ cells at the time of the snapshot. This figure illustrates the complex genotypic variability present in the population of cells produced by our model. We observe that there are two sources of heterogeneity present in the population: variability in the number of mutations per cell (heterogeneity between generations) and genotypic variation between members of the same generation (heterogeneity within a generation). Generations appear in waves with a large number of different mutant families making significant contributions to the generation size. Our goal here is to discuss these two sources of heterogeneity and derive analytic results that quantify the relationship between the amount of genotypic variation present in the population and our model parameters. Heterogeneity between generations is discussed in Section 3 while heterogeneity within a generation is discussed in Section 4.

3 Heterogeneity between generations

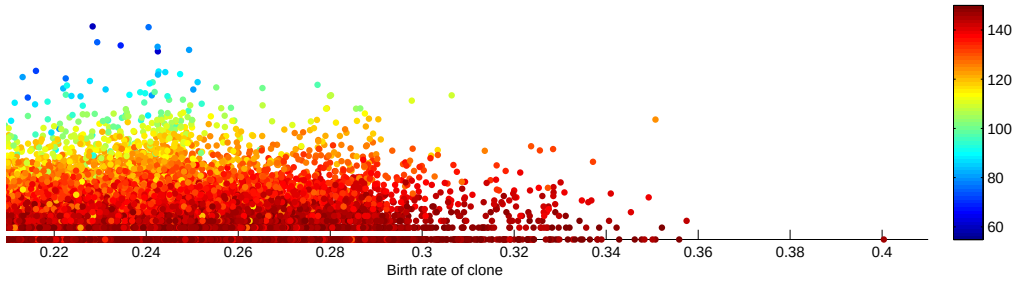
sec-btwn

We begin with a summary of some previous results. We shall assume that $u_i \equiv u$, define

$$\lambda_k = \lambda_0 + kb$$



(a)



(b)

Figure 1: A sample cross-section of tumor heterogeneity. (a) Each mutant family present in the tumor at time $t = 150$ is represented with a circle, positioned on the horizontal axis according to fitness (birth rate) and on the vertical axis according to family size. Colors delineate families with differing numbers of mutational alterations (wave k cells have k alterations). (b) Colorscale depicts time of creation of each family. In this simulation, $a_0 = 0.2, b_0 = 0.1, \nu \sim U([0, 0.05]), u = 0.001$.

ave_sample2

to be the maximum growth rate that can be attained by generation k mutants, and let

$$p_k = -k + \sum_{j=0}^{k-1} \frac{\lambda_k}{\lambda_j}.$$

Throughout this paper, we shall also use "$\Rightarrow$" to denote convergence in distribution.

th2B **Theorem 1.** *If ν satisfies* **(i)** *above, then*

$$(1/u)^{(k+p_k)} e^{-\lambda_k t} Z_k(t) \Rightarrow V_{d,\,k}$$

where $V_{d,\,k}$ has Laplace transform

$$\exp(-d_k(\lambda_0, b) V_0 \theta^{\lambda_0/\lambda_k})$$

for all $\theta \geq 0$. If ν satisfies **(ii)** *above, then*

$$(t/u)^{k+p_k} e^{-\lambda_k t} Z_k(t) \Rightarrow V_{c,\,k}$$

where $V_{c,\,k}$ has Laplace transform

$$\exp(-c_k(\lambda_0, b) V_0 \theta^{\lambda_0/\lambda_k})$$

for all $\theta \geq 0$. Here, $d_k(\lambda_0, b)$ and $c_k(\lambda_0, b)$ are constants which depend on the indicated parameters. Explicit formulas can be found in [8], Section 4.

The "d" and "c" in the constants and subscripts stand for discrete and continuous. In case (ii), Theorem 1 follows from [8], Theorem 4. In case (i) when $N = 1$, Theorem 1 follows from the discussion in [8], Section 4 (see also [10], Theorem 5 for a similar result when the number of type-0 individuals is random), but these results can easily be extended to general finite distributions because the exponential growth of generation k implies that one can ignore the contribution of mutations which advance the fitness by $b_i < b$. Note that in case (ii), there is a polynomial correction to the exponential growth of generation k, but the limiting behavior of the two systems is otherwise similar. This slowdown is due to the fact that when the fitness distribution is continuous, it takes an additional amount of time until the birth of the first individuals in generation k with near maximal growth rates (i.e. growth rates near λ_k - see [8], Section 1.1).

3.1 Theoretical results: small mutation limit

Theorem 1 implies that in case (i), for example, we have the approximation

$$\log Z_k(t) \approx \lambda_k t - (k + p_k) \log(1/u) + \log V_{d,k} \tag{3.1}$$ Vkapprox

when t is large. Dividing both sides of this equation by $L = \log(1/u)$ and looking at the process on a time-scale of order L, we can see that the log size of generation k approaches a deterministic, linear limit as the mutation rate $u \to 0$.

linlim **Theorem 2.** *As $u \to 0$,*

$$(1/L)\log^+ Z_k(Lt) \to z_k(t) = [\lambda_k t - (k + p_k)]^+ = \lambda_k(t - \beta_k)^+$$

in probability where

$$\beta_k = \frac{k + p_k}{\lambda_k} = \sum_{j=0}^{k-1} \frac{1}{\lambda_j}.$$

The convergence is uniform on compact subsets of $[0, \infty)$.

The uniform convergence is a consequence of the continuity and monotonicity of the limits. Note that the limiting process depends on λ_0, the growth rate of type-0's, and b, the maximum attainable fitness increase, but is otherwise independent of the particular choice of fitness distribution. An example of the limiting process is shown in Figure 2. In Section 3.2, we shall compare the limiting approximation given by Theorem 2, the approximation given by the righthand side of (3.1), and the results of simulations.

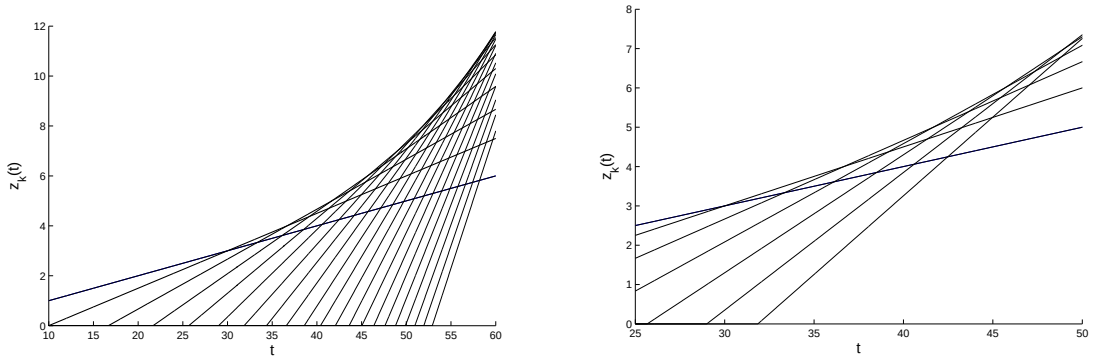


Figure 2: t vs. $z_k(t)$: $\lambda_0 = .1$, $b = .05$. Left: First 20 waves started at $t = 10 = 1/\lambda_0$, the time that 1's begin to be born. Right: Closer look at the firstseven waves showing the changes in the dominant type. fig:bpwa

As a consequence of Theorem 2, we obtain the following important corollary regarding the birth time of type-k's.

Corollary 1. *Let $T_k = \inf\{t \geq 0 : Z_k(t) > 0\}$ be the first time a type-k individual is born. Then as $\mu \to 0$,*

$$T_k/L \to \beta_k$$

in probability for all $k \geq 0$.

From the definition of β_k, it is clear that

$$\beta_k - \beta_{k-1} = \frac{1}{\lambda_k}$$

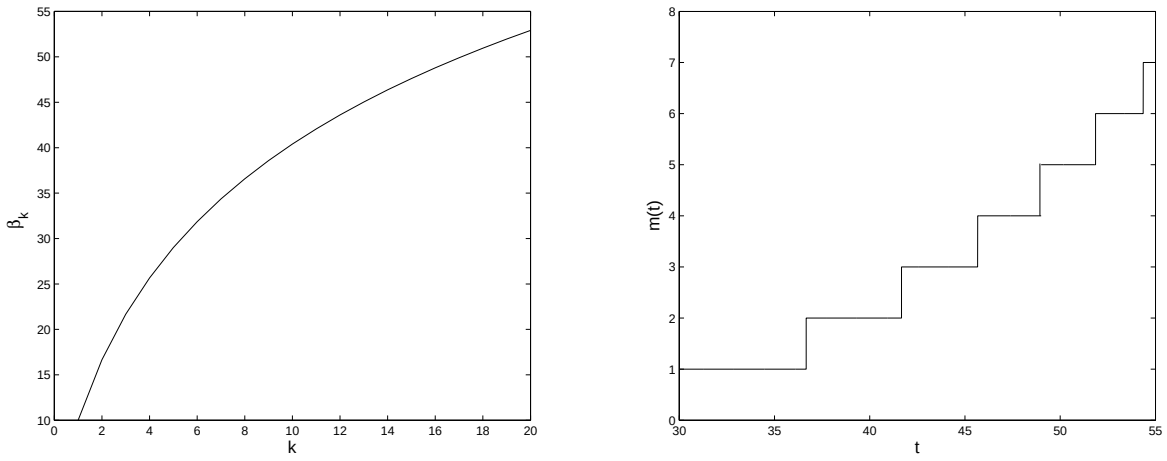


Figure 3: Left: Birth times for the first 20 generations plotted as a function of generation number. Right: Dominant type as a function of time. Same parameters as Figure 2.

fig:dom

is decreasing so that the increments between the birth times for successive generations decrease as k increases, leading to an acceleration in the rate at which new mutations are accumulated. This acceleration can be seen on the left side of Figure 3 and occurs regardless of our particular choice of fitness distribution although since λ_k^{-1} is inversely proportional to b, distributions which allow for larger fitness increases will tend to exhibit shorter increments.

In [3], the authors observe a similar acceleration of waves, but based on approximations in [2], they suggest that this acceleration is an artifact of the presence of both passenger and driver mutations and does not occur when only driver mutations which confer a fixed selective advantage are allowed (i.e. when the fitness increments are deterministic). In our model, the cause of the acceleration of waves is due to the difference in growth rates between the successive generations: type-k's are born when generation $k-1$ reaches size $O(1/u)$ and since the asymptotic growth rate of generation k is larger than the asymptotic growth rate of generation $k-1$, generation $k+1$ will reach size $O(1/u)$ faster than generation k. Another example of this phenomenon can be found in [9] where the authors study a related stochastic model of tumor growth in which the population of cells grows at a fixed exponential rate and subpopulations of different cell types compete for space. There, the cause of the acceleration is again related to growth rates: later generations take longer to achieve dominance in the expanding population of cells and hence, new types are born with a higher fitness advantage over the population bulk, allowing them to reach size $O(1/u)$ more rapidly.

We conclude this section with a second useful consequence of Theorem 2: an asymptotic result for the time at which type-k individuals become dominant in the population.

Corollary 2. *Let* $S_k = \inf\{t \geq 0 : Z_k(t) > Z_j(t), \forall\, j \neq k\}$ *be the first time that type-k individuals become the dominant type. Then*

$$S_k/L \to t_k = b^{-1} + \beta_k$$

in probability as $\mu \to 0$ *for all* $k \geq 1$.

The limit t_k is the solution to

$$\lambda_k(t_k - \beta_k) = \lambda_{k-1}(t_k - \beta_{k-1})$$

i.e. the time when $z_k(t)$ first overtakes $z_{k-1}(t)$. The right side of Figure 3 shows an example of how the index $m(t)$ of the largest generation at time t, defined by

$$z_{m(t)}(t) = \max\{z_k(t) : k \geq 0\},$$

changes over time. The transitions between periods of dominance are only sharp in the small mutation limit and an interesting question for future investigations would be to determine the time scale on which these transitions occur. Note also that because of the log scale, at any given time the population consists primarily of members in the current dominant generation, i.e.

$$(1/L)\log Z(Lt) \to z_{m(t)}(t)$$

as $\mu \to 0$. Therefore, for small mutation rates, the amount of genetic heterogeneity present in the population is determined by the amount of heterogeneity present in the dominant generation.

3.2 Positive mutation rates: numerical simulations

ec-simvtheo

In this section we numerically investigate the heterogeneity between waves in the case of positive mutation rates and deterministic fitness distributions. Given that there are approximately 3 billion $= 3 \times 10^9$ base pairs in the human genome and assuming a mutation rate per base pair of $O(10^{-8}) - O(10^{-10})$, we expect that point mutations occur at a rate of between .3 and 30 per cell division. However we assume that advantageous mutations constitute only a fraction of the possible mutations and therefore that the mutation rate per cell division for advantageous mutations should be in the approximate range of $O(10^{-5}) - O(10^{-2})$. We shall use $u = 10^{-5}$ in our investigations below (which also corresponds to the value used by the authors in [3]).

Generation	Mean	Standard Deviation
1	4.7638	1.3738
2	10.6010	2.2434
3	17.0519	3.0282

Table 1: Means and standard deviations for $\log V_{d,k}$, $k = 1, 2, 3$.

table:Vc

In Figure 5 (a) and (b), we compare the average log size of the kth generation in simulations with the limiting approximation given by Theorem 2. We observe qualitatively similar behavior, however the limiting approximation consistently underestimates the times at which

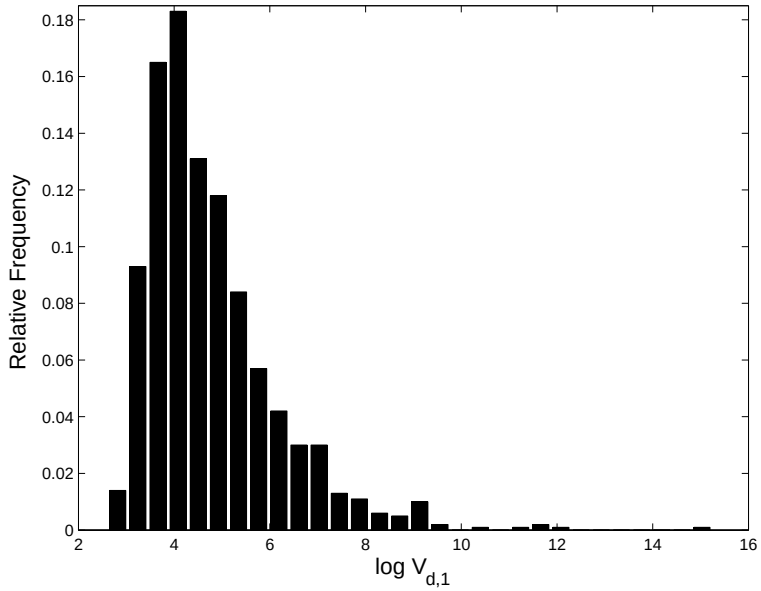


Figure 4: Relative frequency histogram of 1000 random samples from the distribution of $\log V_{d,1}$. fig:logV

new waves appear. To explain the source of this bias, we use the alternative approximation given by righthand side of (3.1) which can be rewritten as

$$\hat{z}_k^L(t) = Lz_k(t/L) + \log V_{d,k} \tag{3.2}$$ zhat

where $L = \log(1/u)$. Using the expression for the Laplace transform of $V_{d,k}$ and the numerical algorithm of [30], we simulate 1000 random draws from the distribution of $V_{d,k}$. Table 1 shows the sample mean and standard deviation for $\log V_{d,k}$, $k = 1, 2, 3$. The distribution of $\log V_{d,k}$ has a positive mean and is skewed right (see Figure 4 for an example with $k = 1$) implying that the limit in Theorem 2 will in general underestimate the size of generation k for positive mutation rates. The approximation obtained by replacing $\log V_{d,k}$ with the sample mean of $\log V_{d,k}$ is plotted in Figure 5 (c). We note that after an initial period in which the number of type-k individuals is small, this picture closely resembles the plot in (a). We also note that the variance of $\log V_{d,k}$ tends to increase with k and hence, we should expect to see an increasing amount of variability in simulations in the time when type-k individuals are born. In Figure 5 (d), we plot the right hand side of (3.2) replacing $\log V_{d,k}$ with the value two standard deviations above its mean to illustrate an extreme scenario.

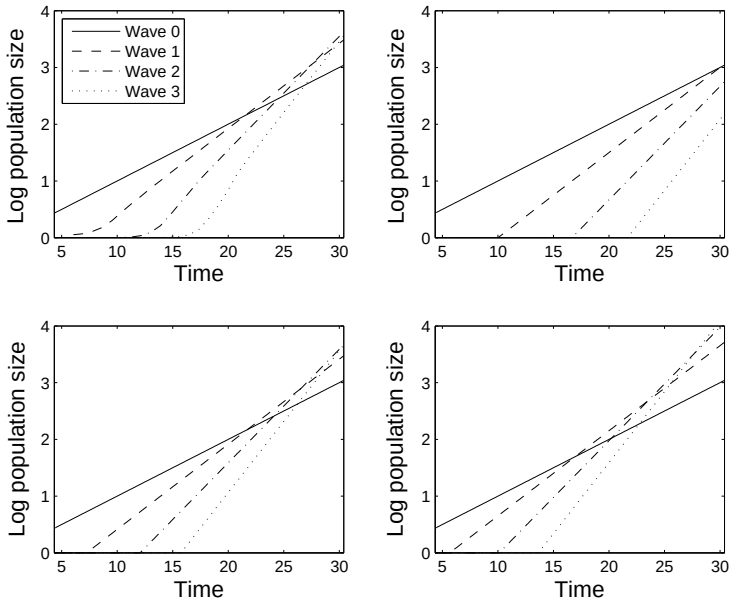


Figure 5: The log size of generations 1 through 4 as a function of time. Note that both time and space are plotted in units of $L = \log(1/u)$ **(a)** (Top left) Average values over 10^6 simulations. **(b)** (Top right) Limiting approximation from Theorem 2. **(c)** (Lower left) Approximation from (3.2) using the mean of $\log V_{d,k}$. **(d)** (Lower right) Approximation from (3.2) using the value two standard deviations above the mean of $\log V_{d,k}$. The parameters used in the simulation are $u = 10^{-5}$, $a_0 = 0.2$, $b_0 = 0.1$, and $b = .05$.

fig:het

4 First generation heterogeneity

sec-first

In this section, we study within generation heterogeneity by examining the amount of genotypic heterogeneity present in the first generation of individuals. We use two statistical measures to assess heterogeneity: (i) Simpson's Index, which gives the probability that two randomly chosen individuals from the first generation come from the same clone, and (ii) the fraction of individuals in the first generation which come from the largest family of individuals. To obtain these results, we will rely on Theorem 3 below which gives us insight into how the limit in Theorem 1 comes about: the limit is the sum of points in a nonhomogeneous Poisson process. Each point in the limiting process represents the contribution of a different mutant lineage to $Z_1(t)$ so that it suffices to calculate (i) and (ii) for the limiting process. Before stating this result, we need to introduce some terminology. Here and in what follows, we use $|A|$ to denote the number of points in the set A. We say that Λ is a Poisson process on $(0,\infty)$ with mean measure μ if Λ is a random set of points in $(0,\infty)$ with the following properties:

(i) For any $A \subset (0,\infty)$, $N(A) = |\Lambda \cap A|$ is a Poisson random variable with mean $\mu(A)$.

(ii) For any $k \geq 1$, if $A_1, ..., A_k$ are disjoint subsets of $(0,\infty)$, then $N(A_i)$, $1 \leq i \leq k$ are independent.

We also let $\alpha = \lambda_0/\lambda_1 \in (0,1)$ denote the ratio of the growth rate of 0's to the maximal growth rate of 1's and note that $1 + p_1 = 1/\alpha$.

th1C **Theorem 3.** *Let Λ be a Poisson process on $(0,\infty)$ with mean measure*

$$\mu(A) = \int_A \alpha z^{-(\alpha+1)} dz$$

and let S denote the sum of the points in Λ. Then there exist positive constants $A_d, A_c = A_d(\lambda_0, b), A_c(\lambda_0, b)$ which depend on the indicated parameters so that in case **(i)** *as $t \to \infty$*

$$(A_d u V_0)^{-(1+p_1)} e^{-\lambda_1 t} Z_1(t) \Rightarrow S,$$

and in case **(ii)** *as $t \to \infty$*

$$(A_c u V_0))^{-(1+p_1)} t^{1+p_1} e^{-\lambda_1 t} Z_1(t) \Rightarrow S.$$

For more details, see [8], Theorem 3 and [10], Corollary to Theorem 3. Note that the mean measure for Λ has tail $\mu(x,\infty) = x^{-\alpha}$.

Let X_n denote the n^{th} largest point in Λ, and let $S_n = \sum_{i=1}^n X_i$ denote the sum of the n largest points. To determine the dependence of X_n on n we first note that if we define $\Lambda' = f(\Lambda)$ where $f(x) = x^{-\alpha}$, then Λ' is a Poisson process and after making the change of variables $y = x^{-\alpha}$, we can see that the mean measure is

$$\mu'(A) = \int_{f^{-1}(A)} \alpha x^{-(\alpha+1)} dx = \int_A dy = |A|.$$

In other words, Λ' is a homogeneous Poisson process with constant intensity and hence, the spacings between points are independent exponentials with mean 1. If we let T_n denote the time of the n^{th} arrival in Λ', then the law of large numbers implies that $T_n \sim n$ as $n \to \infty$. Since $X_n = T_n^{-1/\alpha}$, we obtain $X_n \sim n^{-1/\alpha}$ as $n \to \infty$. This already suggests how the behavior of X_n depends on α: smaller α corresponds to a quicker decay in X_n and hence, less heterogeneity. In addition we have the following Lemma. Although this result holds for general $\alpha > 0$, we recall that we are assuming here and throughout this section that $\alpha \in (0, 1)$.

meansum **Lemma 1.**

$$EX_n = \Gamma(n - 1/\alpha)/\Gamma(n).$$

Furthermore, if we define $S_\infty = \sum_{i=1}^{\infty} X_i$, *then*

$$ES_\infty < \infty.$$

Proof. Since T_n has a Gamma(n,1) distribution, we have $EX_n = ET_n^{-1/\alpha} = \Gamma(n-1/\alpha)/\Gamma(n)$. Stirling's approximation implies that $\Gamma(n - 1/\alpha)/\Gamma(n) \sim n^{-1/\alpha}$ and the second conclusion follows. □

4.1 Simpson's Index

One common measure of heterogeneity in a population is Simpson's index which is the probability that two randomly selected individuals come from the same family. Recalling that X_i is the contribution of the ith largest family of generation 1 individuals to the total size of generation, we define Simpson's index for our point process by

$$R = \frac{\sum_{i=1}^{\infty} X_i^2}{(S_\infty)^2} = \sum_{i=1}^{\infty} \left(\frac{X_i}{S_\infty} \right)^2.$$

The formula for the mean ER is remarkably simple.

simpmean **Theorem 4.** $ER = 1 - \alpha$

This result shows that the average amount of heterogeneity present in the first generation depends only on α, the ratio of the growth rate of type-0's to the maximum attainable growth rate of type-1 individuals. The key to our proof is a result in [12] which considers

$$R_n = \sum_{i=1}^{n} \left(\frac{Y_i}{S_n} \right)^2$$

where Y_i are iid random variables in the domain of attraction of a stable law with index α and $S_n = Y_1 + \cdots + Y_n$ and shown that

$$\lim_{n \to \infty} ER_n = 1 - \alpha.$$

To explain the connection between the two results, we note that if we have $P(Y_i > x) = x^{-\alpha}$, for $x \geq 1$ and let $Y_{n,i} = Y_i/n^{1/\alpha}$, then

$$nP(Y_{n,i} \in A) \to \mu(A).$$

This implies that if we let $\Delta_n = \{Y_{n,i} : i \leq n\}$ be the point process associated with the $Y_{n,i}$ and define the measures $\xi_n \equiv \sum_{x \in \Lambda_n} \delta_x$ and $\xi = \sum_{x \in \Lambda} \delta_x$, then we have

$$\xi_n \Rightarrow \xi.$$

so that we should expect ER to agree with $\lim ER_n$. We make this argument rigorous in Appendix A.

In Figure 6, we plot the sample mean of Simpson's index of Z as a function of time for different values of α and compare with $1 - \alpha$, the expected value of Simpson's index for the limiting point process. We observe that initially, the sample mean tends towards $1 - \alpha$. For larger values of b, the sample mean overshoots the limiting value. Our theory guarantees that eventually the values of the sample mean will converge, however we were not able to simulate far enough out in time to see this occur.

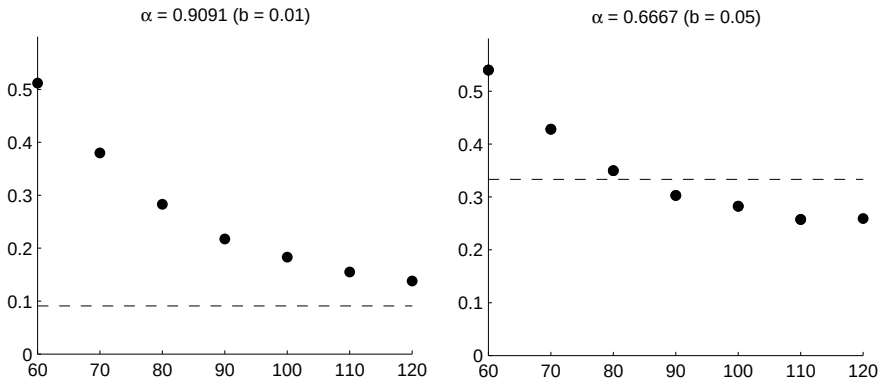


Figure 6: Expected value of Simpson's Index for the first wave. We compare the sample mean (dots) of Simpson's index at times $t = 60$ to 120 with the expected value of Simpson's Index for the limiting point process (lines), for two different values of α. Parameters: $\lambda_0 = 0.2, a_0 = 0.1, \nu \sim U([0, b])$, where $b = 0.01$ (left) and $b = 0.05$(right). fig:SI_c

We can further exploit the previously discussed connection between Δ_n and Λ to gain additional insight into the distribution of Simpson's index. For example, one can obtain expressions for higher moments of R (see [1]). In [22], it is established that as $n \to \infty$,

$$S_n(2) = R^{-1/2} = \frac{\sum_{i=1}^n Y_i}{(\sum_{i=1}^n Y_i^2)^{1/2}}$$

has a limiting distribution with a density f that satisfies

$$f(y) \sim ae^{-by^2}, \text{ as } y \to \infty,$$

for some constants $a, b > 0$ (see [22], equation (5.7) and [32], Theorem 6.1 for more details). Therefore, after making the change of variables $x = y^{-1/2}$, we can see that the density of R near the origin has the form

$$g(x) \sim ax^{-3/2}e^{-bx^{-1}}, \text{ as } x \to 0, .$$

In Figure 7, we perform simulations of Simpson's index for the first wave mutants in the branching process Z. We use parameters $u = 10^{-3}, \lambda_0 = 0.2, a_0 = 0.1$ and mutations confer an additive fitness change drawn according to $\nu \sim U([0, 0.01])$. In the simulations we observe convergence of the empirical distribution of Simpson's Index to the distribution for the limiting point process. Curiously, in the plot at $t = 70$, we observe a few noticeable bumps in the distribution (e.g. around 0.5). These bumps can be explained by a few families with similar size dominating the population; this occurs with enough probability to be apparent in the density.

4.2 Largest clones

sec-largest

To further investigate heterogeneity properties of the point process, let

$$S_n = \sum_{i=1}^{n} X_i$$

be the total contribution from the largest n points. Let

$$V_n = X_1/S_n$$

be the fraction of individuals descended from the largest family of first generation mutants. Our next result shows that $1/V_n$ converges to a non-trivial limit and provides us with an explicit formula for the characteristic function of the limit.

charfunc **Theorem 5.** *As $n \to \infty$, $V_n^{-1} \Rightarrow W$ where W has characteristic function ψ satisfying $\psi(0) = 1$ and*

$$\psi(t) = \frac{e^{it}}{f_\alpha(t)}$$

for all $t \neq 0$ with

$$f_\alpha(t) = 1 + \alpha \int_0^1 (1 - e^{itu})u^{-(\alpha+1)}du.$$

The form of the characteristic function is the same as the characteristic function for $\lim_{n\to\infty} T_n/Y_{(1)}$ where the Y_i are iid random variables with power law tails, $Y_{(1)} = \max_{i\leq n} Y_i$, and $T_n = \sum_{i=1}^{n} Y_i$ (see, for example, [6]). Again, this agreement is a consequence of the previously discussed connection between Δ_n and the limiting Poisson point process.

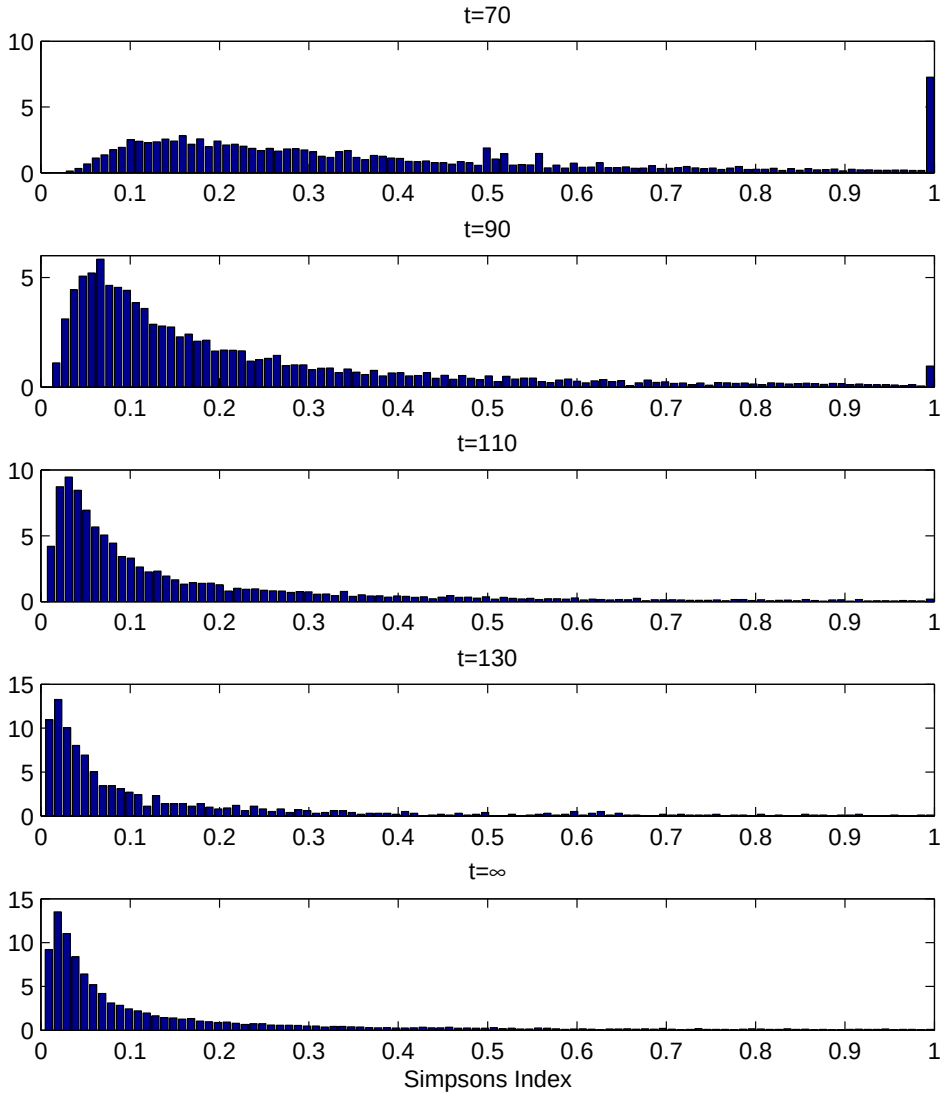


Figure 7: Empirical distribution of Simpson's Index for the first wave. We include plots of Simpson's Index for the branching process at times $t = 70, 90, 110, 130$ along with Simpson's Index for the limiting point process ($t = \infty$). The histograms show the average over 1000 simulations. For the limiting point process, we approximate Simpson's index by looking at the largest 10^4 points in the process. Parameters: $\lambda_0 = 0.2, a_0 = 0.1, \nu \sim U([0, 0.01])$. fig:SI_c

It is interesting to note that the characteristic function in Theorem 5 is not integrable. The problem is that the density of V_n^{-1} blows up near 1. As an explanation for this, we note that with probability

$$\exp(-(1-x^{-\alpha}))\exp(-x^{-\alpha})x^{-\alpha} = e^{-1}x^{-\alpha}$$

there is a point in the process bigger than x and no points in $[1,x)$. When this happens,

$$V_n^{-1} = S_n/X_1 \le 1 + n/x$$

and so

$$F_n(y) = P(V_n^{-1} \le y) \ge e^{-1}n^{-\alpha}(y-1)^{\alpha}.$$

If we had $F_n(y) \sim (y-1)^{\alpha}$, then the density would blow up like $(y-1)^{\alpha-1}$ as $y \to 1$. We will confirm that this gives the right asymptotic by providing an explicit formula for the density of W.

Vdensity **Corollary 3.** *W has a density on $(1,\infty)$ given by*

$$f(y) = \lim_{M\to\infty} \int_{-M}^{M} \frac{e^{it(1-y)}}{f_\alpha(t)} dt.$$

Note that integral expression above does not converge absolutely so part of the proof will consist of showing that the limit exists. If we apply the change of variable $s = t(y-1)$ in the definition of f, we see that

$$f(y) = (y-1)^{\alpha-1} \int_{-\infty}^{\infty} \frac{e^{-it}}{(y-1)^{\alpha} + \int_0^{(y-1)^{-1}} \frac{1-e^{iut}}{u^{\alpha+1}} du}$$

thus confirming the intuition that the density blows up like $(y-1)^{\alpha-1}$ as y approaches 1.

Differentiating ψ leads to simple expressions for the mean and variance of the limit.

Vmean **Corollary 4.** *$EW = \frac{1}{1-\alpha}$ and var$(W) = \frac{2}{(1-\alpha)^2(2-\alpha)}$.*

Figure 8 suggests that the rate of convergence is slow for α close to 1.

Returning to the study of $V_n = X_1/S_n$, we note that Theorem 5 implies that V_n converges to a non-trivial limit $V = W^{-1}$ and Jensen's inequality applied to the strictly convex function $1/x$ implies that $E(\lim X_1/S_n) > 1-\alpha$. Simulations suggest that despite this fact, deviations of the mean from $1-\alpha$ are small as illustrated in Figure 8.

5 Discussion

sec-disc

In this work we have investigated the evolution of diversity in a stochastic model of tumor expansion incorporating random mutational advances. In Section 3 we considered heterogeneity between tumor subpopulations with varying numbers of mutations. In the limit as

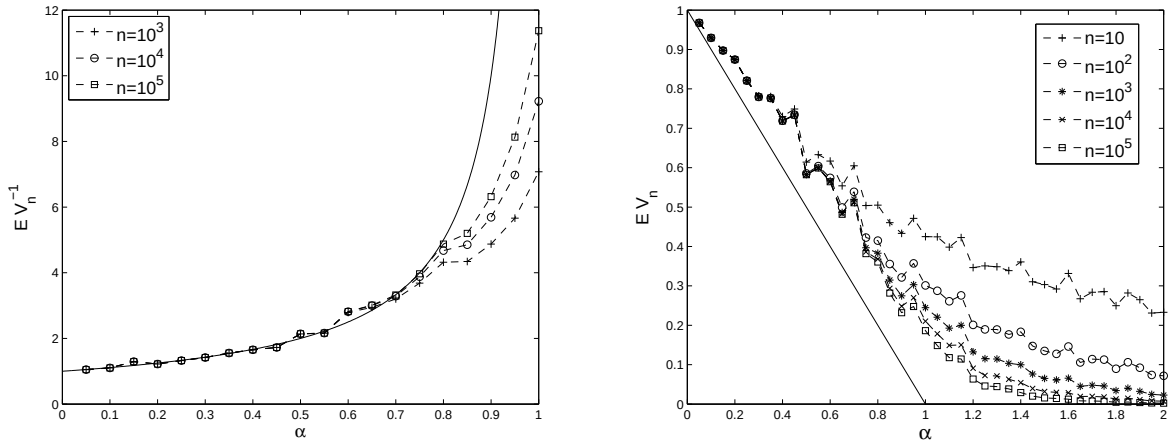


Figure 8: Left: Comparison of Monte Carlo estimates for EV_n^{-1} and the limit $(1-\alpha)^{-1}$. Right: Comparison of the Monte Carlo estimates for EV_n and the curve $(1-\alpha)^{+}$. Number of simulations = 100.

fig:snm

the mutation rate approaches zero, we obtained estimates of the contribution of each wave of mutants to the total population. These limiting approximations depend on the maximum attainable fitness advance for mutants, but not on the specific form of the fitness distribution. We also obtained estimates of the time of arrival of the first cell of each wave which showed that the accumulation of new genetic alterations will accelerate over time due to the increasing growth rates of successive generations. Simulations demonstrate that for small, but positive mutation rates, the behavior of the system is qualitatively similar to the limiting predictions (see Figure 5). These simulations also suggest that as time increases, multiple waves of mutants coexist without a single, largely dominant wave.

In Section 4 we investigated the genotypic diversity within the first wave of mutants. In particular, we considered two measures of diversity: the Simpson's Index and the fraction of individuals in the first generation which come from the largest family of individuals. We analyzed these diversity measures in a limiting regime where the first generation of mutants can be described by an inhomogeneous point process. For this point process, we obtained the exact mean of Simpson's Index as well as the form of its density near the origin. Interestingly, the mean of Simpson's index is given by the quantity $1-\alpha$, where α is the ratio between the fitness of the unmutated ancestral population and the maximum possible fitness of the first wave of mutants. Next, we observed that as time increases the mass of the distribution of Simpson's Index moves closer to 0, indicating higher levels of diversity in the tumor at later times (see Figure 7). This is also observed via direct numerical simulation of the branching process; we observed convergence of the distribution and mean of Simpson's Index to the predicted limiting values.

In Section 4.2 we investigated the ratio between the total population size of the first wave of mutants and the size of the largest family. We show that this ratio can be approximated

by a random variable with mean $(1-\alpha)^{-1}$. The density of this random variable is given in Corollary 3. We note that as α approaches 1 the mean of this ratio grows to infinity. In other words, as α approaches 1 the largest family constitutes a vanishing proportion of the total wave one population.

Our results indicate that tumor diversity is strongly dependent upon the age of the tumor and the quantity $1-\alpha$, but is otherwise unaffected by changes to the fitness distribution. Investigations of Simpson's Index and the ratio S_n/X_1 indicate that if α is close to one (i.e. fitness advances are small) we expect the tumor population to have a higher diversity. In addition, our results on Simpson's Index and inter-wave heterogeneity indicate that a longer lived tumor will have a higher level of diversity. An open problem which could lead to additional insights into how tumor heterogeneity changes over time is to quantify the amount of heterogeneity present in later generations by obtaining an explicit formula for the mean of Simpson's Index for generation $k \geq 2$. We conjecture that the mean of Simpson's Index will be a decreasing function of the generation number, indicating higher levels of diversity in later generations and a further increase in the total amount of heterogeneity present in the tumor at later times.

A Proofs of results on first generation heterogeneity

sec-proofs

We will use the following notation in this appendix. For a real number t, we define the function

$$sgn(t) = \begin{cases} -1, & t < 0 \\ 0, & t = 0 \\ 1, & t > 0. \end{cases}$$

For a complex number z we denote the real part of z by $Re[z]$ and its imaginary part by $Im[z]$.

A.1 Simpson's Index

To prove Theorem 4, let

$$R_n = \frac{\sum_{i=1}^n Y_{n,i}^2}{\left(\sum_{i=1}^n Y_{n,i}\right)^2}$$

denote the value of Simpson's index for the point process Λ_n. The following lemma is a restatement of Theorem 5.3 in [12] applied to the Y_i.

iidsimp **Lemma 2.** *$ER_n \to 1-\alpha$ as $n \to \infty$.*

We will prove Theorem 4 by showing that we also must have $ER_n \to ER$. To this end, let

$$R_n(\epsilon) = \frac{\sum_{i=1}^n Y_{n,i}^2 \mathbf{1}_{Y_{n,i}>\epsilon}}{\left(\sum_{i=1}^n Y_{n,i}^2 \mathbf{1}_{Y_{n,i}>\epsilon}\right)^2}$$

denote the truncated value of Simpson's index for Λ_n and

$$R(\epsilon) = \frac{\sum_i^\infty X_i^2 \mathbf{1}_{X_i>\epsilon}}{(\sum_{i=1}^\infty X_i \mathbf{1}_{X_i>\epsilon})^2}$$

denote the truncated value for Λ. Then for any $\epsilon > 0$, we have

$$|ER_n - ER| \le E|R_n - R_n(\epsilon)| + E|R_n(\epsilon) - R(\epsilon)| + E|R(\epsilon) - R| \tag{A.1}$$ tribnd

We will complete the proof by deriving appropriate bounds for each of the three terms on the righthand side of (A.1). For the first term, we have

Rnerror **Lemma 3.**

$$\limsup_{n\to\infty} E\,|R_n - R_n(\epsilon)| \le h_\epsilon$$

where $h_\epsilon \to 0$ *as* $\epsilon \to 0$.

Proof. Let $\epsilon > 0$ and write $A_{n,k} = \sum_{i=1}^n Y_{n,i}^k$, $A_{n,k}(\epsilon) = \sum_{i=1}^n Y_{n,i}^k \mathbf{1}_{Y_{n,i}>\epsilon}$, and $\bar{A}_{n,k}(\epsilon) = A_{n,k} - A_{n,k}(\epsilon)$ for $k = 1, 2$. Since

$$EY_1^k \mathbf{1}_{Y_1 \le \epsilon n^{1/\alpha}} = \int_1^{\epsilon n^{1/\alpha}} k y^{k-1} y^{-\alpha} dy \le C\epsilon^{k-\alpha} n^{k/\alpha - 1}$$

for $k = 1, 2$, we have the bound

$$E\bar{A}_{n,k}(\epsilon) \le C\epsilon^{k-\alpha}. \tag{A.2}$$ errorbnd

After noting that $A_{n,2} \le A_{n,1}^2$, $A_{n,2}(\epsilon) \le A_{n,1}^2(\epsilon) \le A_{n,1}^2$, and $\bar{A}_{n,1}(\epsilon) + A_{n,1}(\epsilon) = A_{n,1}$, we have for any $\delta > 0$,

$$\begin{aligned}
E|R_n - R_n(\epsilon)| &= E\left|\frac{\bar{A}_{n,2}(\epsilon)}{A_{n,1}^2} + R_n(\epsilon)\left(\frac{A_{n,1}^2(\epsilon) - A_{n,1}^2}{A_{n,1}^2}\right)\right| \\
&\le \delta^{-2} E\bar{A}_{n,2}(\epsilon) + P(A_{n,1} \le \delta) \\
&\quad + \delta^{-1} E\left(\frac{|2A_{n,1}(\epsilon)\bar{A}_{n,1}(\epsilon) + \bar{A}_{n,1}^2(\epsilon)|}{A_{n,1}}\right) + P(A_{n,1} \le \delta) \\
&= \delta^{-2} E|\bar{A}_{n,2}(\epsilon)| + \delta^{-1} E\left(\left|\bar{A}_{n,1}(\epsilon)\left(2 + \frac{A_{n,1}(\epsilon)}{A_{n,1}}\right)\right|\right) + 2P(A_{n,1} \le \delta) \\
&\le \delta^{-2} E\bar{A}_{n,2}(\epsilon) + 3\delta^{-1} E\bar{A}_{n,1}(\epsilon) + 2P(A_{n,1} \le \delta) \\
&\le \epsilon^{2-\alpha}/\delta^2 + 3\epsilon^{1-\alpha}/\delta + 2P(A_{n,1} \le \delta)
\end{aligned} \tag{A.3}$$ Rndiff

where we have used (A.2) in the last line. To control the third term on the right, let ϕ denote the Laplace transform of Y_1. Then

$$\begin{aligned}
1 - \phi(t) &= \alpha \int_1^\infty (1 - e^{-ty}) y^{-(\alpha+1)} dy \\
&= \alpha t^\alpha \int_t^\infty (1 - e^{-x}) x^{-(\alpha+1)} dx \sim Ct^\alpha
\end{aligned}$$

as $t \to 0$ since $1 - e^{-x} \sim x$ as $x \to 0$ implies that $\int_0^\infty (1 - e^{-x}) x^{-(\alpha+1)} dx < \infty$. We can conclude that

$$E \exp(-tA_{n,1}) = (1 - (1 - \phi(t/n^{1/\alpha}))^n \to \exp(-Ct^\alpha)$$

as $n \to \infty$. In particular, $A_{n,1} \Rightarrow A_1$, where A_1 has the above Laplace transform. Since

$$1 - \exp(-Ct^\alpha) \to 1$$

as $t \to \infty$, we have $P(A_1 = 0) = 0$ so that taking $\delta = \epsilon^{(1-\alpha)/2}$ in (A.3) yields the result. □

To bound the second term on the righthand side of (A.1), we need some notation. Let M_p denote the class of all point measures on $(0, \infty)$. In a slight abuse of notation, we write $v \in \nu$ when $\nu \in M_p$ and $v \in \text{supp}(\nu)$. We shall equip M_p with the topology of vague convergence (see, for example, Section 3.4 in Resnick [29]) and take as our σ-algebra the one generated by open sets in this topology. Associated with any random set of points, we can associate a measure ξ which is a random variable with values in M_p. We will write $\Lambda_n \Rightarrow \Lambda$ to mean that the associated random measures $\xi_n \Rightarrow \xi$.

Lemma 4. *$\Lambda_n \Rightarrow \Lambda$ and if we define the maps $F_{k,\epsilon} : M_p \to [0, \infty)$ by*

$$F_{k,\epsilon}(\mu_n) = \sum_{x \in \mu_n} x^k \mathbf{1}_{x > \epsilon}$$

for $k = 1, 2$, then

$$(F_{1,\epsilon}(\Lambda_n), F_{2,\epsilon}(\Lambda_n)) \Rightarrow (F_{1,\epsilon}(\Lambda), F_{2,\epsilon}(\Lambda)).$$

Proof. Since

$$nP(Y_{n,i} \in A) = n \int_{n^{1/\alpha} A} \alpha y^{-(\alpha+1)} dy = \int_A \alpha x^{-(\alpha+1)} dx = \mu(A)$$

for all Borel sets A, the first claim follows from Proposition 3.21 in [29]. The second claim follows from the Continuous Mapping Theorem (see, for example, page 152 in [29]), the fact that $F_{k,\epsilon}$ is continuous away from measures ν with $\epsilon \notin \nu$, and the fact that the random measure associated with Λ has no point masses with probability 1. □

As a consequence of this lemma, the fact that $R_n(\epsilon) \leq 1$, and the bounded convergence theorem, we have

Rep **Corollary 5.**

$$E\,|R_n(\epsilon) - R(\epsilon)| \to 0$$

as $n \to \infty$ for any $\epsilon > 0$

It thus remains to establish the following

Rerror **Lemma 5.**

$$\limsup_{\epsilon \to 0} E\,|R(\epsilon) - R| = 0.$$

Proof. We can establish this result using the same results as in the proof of Lemma 3. In particular if we define $A_k = \sum_{n=1}^{\infty} X_n^k$ and $\bar{A}_k(\epsilon) = \sum_{n=1}^{\infty} X_n^k \mathbf{1}_{X_n < \epsilon}$ for $k = 1, 2$. Then following the display in A.3 we have for any $\delta > 0$

$$E\,|R - R(\epsilon)| \leq \delta^{-2} E\bar{A}_2(\epsilon) + 3\delta^{-1} E\bar{A}_1(\epsilon) + 2P(A_1 \geq \delta).$$

It is obvious that $P(A_1 = 0) = 0$ and $E\bar{A}_2(\epsilon) \leq E\bar{A}_1(\epsilon)$ for $\epsilon < 1$, so it only remains to establish that

$$E\bar{A}_1(\epsilon) \to 0,$$

as $\epsilon \to 0$. However this result follows immediately from Lemma 1. Therefore taking $\delta = \left(E\bar{A}_1(\epsilon)\right)^{1/4}$ completes the proof. □

We can now complete the proof of Theorem 4 by letting $n \to \infty$ and then $\epsilon \to 0$ in (A.1) and applying Lemmas 3 and 5 and Corollary 5. □

A.2 Largest Clones

We begin with the

Proof of Theorem 5. Theorem 5.1 in [6] implies that we have

$$E \exp\left(itT_n/Y_{(1)}\right) \to \psi(t)$$

as $n \to \infty$ where as in Section 4, $Y_{(1)} = \max_{i \leq n} Y_i$ and $T_n = \sum_{i=1}^{n} Y_i$. To conclude that $T_n/Y_{(1)} \Rightarrow V$, we need to show that ψ is continuous at 0. To establish this fact, we make the change of variables $v = tu$ to conclude that

$$f_\alpha(t) = 1 + \alpha \int_0^1 (1 - e^{itu}) u^{-(\alpha+1)} du = 1 + \alpha |t|^\alpha \int_0^{|t|} (1 - e^{iv\mathrm{sgn}(t)}) v^{-(\alpha+1)} dv. \tag{A.4}$$ altpsi

Since $1 - \exp(iv) \sim -iv$ as $v \to 0$, the integral on the right-hand side of (A.4) is finite and hence,

$$\psi(t) = e^{it} f_\alpha^{-1}(t) \to 1$$

as $t \to 0$. Since $T_n/Y_{(1)} \Rightarrow V$, the fact that $S_n/X_1 \Rightarrow V$ follows from the arguments in the previous section. □

Proof of Corollary 3. We first establish that there are no point masses in the distribution of V. By the inversion formula we have for any $a \in \mathbb{R}$,

$$\begin{aligned} P(V = a) &= \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} e^{-iat} \psi(t) dt \\ &= \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \frac{e^{it(1-a)}}{f_\alpha(t)} dt. \end{aligned}$$

If we focus on the positive axis and use the change of variable $s = t/T$,

$$\frac{1}{2T}\int_0^T \frac{e^{it(1-a)}}{f_\alpha(t)}dt = \frac{1}{2}\int_0^1 \frac{e^{isT(1-a)}}{f_\alpha(sT)}ds.$$

From display (A.4) it follows that for every $s \in (0,1)$ we have $e^{isT(1-a)}/f_\alpha(sT) \to 0$ as $T \to \infty$. Note that

$$Re[f_\alpha(t)] = 1 + \alpha\int_0^1 \frac{1-\cos ut}{u^{\alpha+1}}du > 1,$$

which implies $|f_\alpha(t)| \geq 1$ for all t. Therefore $\left|e^{isT(1-a)}/f_\alpha(t)\right| \leq 1$ for all t and it follows via the Dominated Convergence Theorem that

$$\lim_{T\to\infty}\frac{1}{2}\int_0^1 \frac{e^{isT(1-a)}}{f_\alpha(sT)}ds = 0.$$

A similar result holds for the integral on the negative axis and we conclude that

$$P(V = a) = 0.$$

We can therefore conclude for $x > 1$ and $h > 0$ via the inversion formula (see [7], (3.2)) and Fubini's Theorem that

$$\begin{aligned} P\left(V \in (x, x+h)\right) &= \lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\int_x^{x+h} e^{-ity}\psi(t)dydt \\ &= \lim_{T\to\infty}\frac{1}{2\pi}\int_x^{x+h}\int_{-T}^{T} e^{-ity}\psi(t)dtdy. \end{aligned}$$

Therefore in order to establish the result we need to show that

$$\lim_{T\to\infty}\frac{1}{2\pi}\int_x^{x+h}\int_{-T}^{T} e^{-ity}\psi(t)dtdy = \frac{1}{2\pi}\int_x^{x+h}\int_{-\infty}^{\infty} e^{-ity}\psi(t)dtdy.$$

This follows if we show that $\lim_{T\to\infty}\int_{-T}^{T} e^{-ity}\psi(t)dt$ is a convergent integral, and that there exists a bounded function h defined on $(x, x+h)$ such that

$$|h_T(y)| = \left|\int_{-T}^{T} e^{-ity}\psi(t)dt\right| \leq h(y).$$

We first use integration by parts to see

$$h_T(y) = \int_{-T}^{T}\frac{e^{it(1-y)}}{f_\alpha(t)}dt = \frac{i}{1-y}\left(\frac{e^{iT(1-y)}}{f_\alpha(T)} - \frac{e^{-iT(1-y)}}{f_\alpha(-T)} + \int_{-T}^{T}\frac{e^{it(1-y)}f'_\alpha(t)}{f_\alpha(t)^2}dt\right).$$

Recalling that $|f_\alpha(T)| \to \infty$ as $T \to \pm\infty$, it follows that if we establish that $\frac{f'_\alpha(t)}{f_\alpha(t)^2}$ is integrable on $(-\infty, \infty)$, then the convergence of the integral and the existence of a bounded

dominating function will be established. Since f_α is bounded away from 0, it suffices to check that the function decays fast enough. Recalling the definition of f_α

$$f'_\alpha(t) = -i\alpha t^{\alpha-1} \int_0^t \frac{e^{iv}}{v^\alpha} dv,$$

which follows by passing the derivative inside the integral in the definition of f_α. We can establish that

$$\sup_{T<\infty} \left| \int_0^T \frac{e^{iv}}{v^\alpha} dv \right| < \infty$$

by observing

$$\int_0^\infty \frac{e^{iv}}{v^\alpha} dv = e^{-i\pi(1-\alpha)/2}\Gamma(1-\alpha)$$

which can be found in many places, e.g. [23]. Thus,

$$|f'_\alpha(t)| \leq \alpha t^{\alpha-1} \sup_{T<\infty} \left| \int_0^T \frac{e^{iv}}{v^\alpha} dv \right| \leq C_0 t^{\alpha-1}.$$

We can similarly establish that for t sufficiently large

$$|f_\alpha(t)|^2 \geq C_1 t^{2\alpha},$$

for a positive finite constant C_1. Thus for t sufficiently large

$$\left| \frac{e^{it(1-y)} f'_\alpha(t)}{f_\alpha(t)^2} \right| \leq \frac{C}{t^{\alpha+1}},$$

establishing the result. □

Proof of Corollary 4. Using the Taylor series expansion of $\exp(iu)$ about 0 in (A.4) above implies that

$$1 + f_\alpha(t) = 1 - \sum_{n=1}^\infty \frac{\alpha\,(it)^n}{(n-\alpha)n!}.$$

and therefore,

$$f_\alpha^{(k)}(t) = \sum_{n=k}^\infty \frac{\alpha i^n t^{n-k}}{(n-\alpha)(n-k)!}$$

so that in particular,

$$f_\alpha^{(k)}(0) = \frac{i^k \alpha}{k-\alpha}$$

for all $k \geq 1$. Let $S(t) = \log \psi(t) = it - \log f_\alpha(t)$. Then dropping the α subscript on f_α, we have

$$S'(t) = (i - f'(t)/f(t)) = (i - (\log f(t))')$$

which yields the desired result for the mean:

$$EY = iS'(0) = i(i - f'(0)) = \frac{1}{1-\alpha}.$$

Now

$$S''(t) = -(\log f(t))'' = -\frac{f''(t)f(t) - (f'(t))^2}{f^2(t)}$$

so

$$var(Y) = S''(0) = -f''(0) + (f'(0))^2 = \frac{\alpha}{2-\alpha} + \frac{\alpha^2}{(1-\alpha)^2} = \frac{\alpha}{(1-\alpha)^2(2-\alpha)}$$

completing the proof □

References

AT07 [1] H. Albrecher and J. L. Teugels. Asymptotic analysis of a measure of variation. *Teor. Ĭmovīr. Mat. Stat.*, (74):1–9, 2006.

Beer07 [2] N. et al. Beerenwinkel. Genetic progression and the wait- ing time to cancer. *PLoS Comp. Biol. 3*, 3(11):e225, 2007.

Boz09 [3] I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. Kinzler, B. Vogelstein, and M. Nowak. Accumulation of driver and passenger mutations during tumor progression. *arXiv:0912.1627v1*, 2009.

CoGo85 [4] A. Coldman and J. Goldie. Role of mathematical modeling in protocol formulation in cancer chemotherapy. *Cancer Treat. Rep.*, 49:1041–1045, 1984.

CoGo86 [5] A. Coldman and J. Goldie. A stochastic model for the origin and treatment of tumors containing drug-resistant cells. *Bull. Math. Biol.*, 48(3/4):279–292, 1986.

DAR [6] D.A. Darling. The role of the maximum term in the sum of independent random variables. *Trans. AMS*, 72:85–107, 1952.

DURPTE [7] R. Durrett. *Probability: Theory and Examples, Third Edition*. Duxbury Press, Belmont, California., 2005.

DEA [8] R. Durrett, J. Foo, K. Leder, J. Mayberry, and F. Michor. Evolutionary dynamics of tumor progression with random fitness values. *Theoretical Population Biology*, in review, 2010.

DMAY [9] R. Durrett and J. Mayberry. Traveling waves of selective sweeps. *Ann. Appl. Prob.*, to appear, 2010.

DM [10] R. Durrett and S. Moseley. Evolution of resistance and progression to disease during clonal expansion of cancer. *Theoretical Population Biology*, 77(1):42–48, 2010.

Fidler78 [11] I. Fidler. Tumor heterogeneity and the biology of cancer invasion and metastasis1. *Cancer Research*, 38:2651–2660, 1978.

FEA [12] A. Fuches, A. Joffe, and J. Teugels. Expectation of the ratio of the sum of squares to the square of the sum: exact and asymptotic results. *Theory Probab. Appl.*, 46(2):243–255, 2001.

GeRo02 [13] J. Geisler, S. Rose, H. Geisler, G. Miller, and M. Wiemann. Drug resistance and tumor heterogeneity. *CME Journal of Gynecologic Oncology*, 7:25–28, 2002.

HaIwMi07 [14] H. Haeno, Y. Iwasa Y, and F. Michor. The evolution of two mutations during clonal expansion. *Genetics*, 177:2209–2221, 2007.

HaAu05 [15] L. Haumlggarth, G. Auer, C. Busch, M. Norberg, M. Haumlggman, and L. Egevad. The significance of tumor heterogeneity for prediction of dna ploidy of prostate cancer. *Scandinavian Journal of Urology and Nephrology*, 39(5):387–392, 2005.

He84 [16] G.H. Heppner. Tumor heterogeneity. *Cancer Research*, 44:2259–2265, 1984.

IwNoMi06 [17] Y. Iwasa, M.A. Nowak, and F. Michor. Evolution of resistance during clonal expansion. *Genetics*, 172:2557–2566, 2006.

KaTo00 [18] A.R. Kansal, S. Torquato, E.A. Chiocca, and T.S. Deisboeck. Emergence of a subpopulation in a computational model of tumor growth. *Journal of Theoretical Biology*, 207:431–441, 2000.

Ko06 [19] N.L. Komarova. Stochastic modeling of drug resistance in cancer. *Journal of Theoretical Biology*, 239:351–366, 2006.

CaPo07 [20] K. Polyak L. Campbell. Breast tumor heterogeneity: Cancer stem cells or clonal evolution? *Cell Cycle*, 6(19):2332–2338, 2007.

LaPa07 [21] L. Lai and et al. Increasing genomic instability during premalignant neoplastic progression revealed through high-resolution array-cgh. *Genes Chromosomes Cancer*, 46:532–542, 2007.

LOG [22] B.F. Logan, C.L. Mallows, S.O. Rice, and L.A. Shepp. Limit distributions of self-normalized sums. *Ann. Probab.*, 1:788–809, 1973.

LOY [23] P. Loya. Dirichlet and fresnel integrals via integrated integration. *Mathematics Magazine*, 78:63–67, 2005.

MaGa06 [24] C. Maley and et al. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature Genetics*, 38:468–473, 2006.

MePeRe06 [25] L. Merlo and et al. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*, 6:924–935, 2006.

Michelson89 [26] S. Michelson, K. Ito, H. Tran, and J. Leith. Stochastic models for subpopulation emergence in heterogeneous tumors. *Bull. Math. Biology*, 51(6):731–747, 1989.

NgSe06 [27] H. Nguyen and et al. Evidence of tumor heterogeneity in cervical cancers and lymph node metastases as determined by flow cytometry. *Cancer*, 71:2543–2550, 2006.

OsBu05 [28] M. O'Sullivan, V. Budhraja, Y. Sadovsky, and J. Pfeifer. Tumor heterogeneity affects the precision of microarray analysis. *Diagnostic Molecular Pathology*, 14(2):65–71, 2005.

SR [29] S. Resnick. *Extreme Values, Regular Variation, and Point Processes*. Springer-Verlag, 1987.

Ridout [30] M. S. Ridout. Generating random numbers from a distribution specified by its laplace transform. *Statistics and Computing*, 19(4):439–450, 2009.

Sc08 [31] J. Schweinsberg. The waiting time for m mutations. *Electron. J. Probability*, 13, 2008.

SHA [32] Q.M. Shao. Self-normalized large deviations. *Ann. Probab.*, 25(285-328), 1997.

Shipitsin07 [33] M. Shipitsin and et al. Molecular definition of breast tumor heterogeneity. *Cancer Cell*, 11(259-273), 2007.

WaWi00 [34] N. Wang, C. Willkin, A. Bcking, and B. Tribukait. Evaluation of tumor heterogeneity of prostate carcinoma by flow- and image dna cytometry and histopathological grading. *Analytical Cellular Pathology*, 20(1):49–62, 2000.

Wolman86 [35] S. Wolman. Cytogenetic heterogeneity: Its role in tumor evolution. *Cancer Genet. Cytogenet.*, 19:129–140, 1986.