Adv. Appl. Prob. **44,** 1052–1083 (2012)
Printed in Northern Ireland

A HIERARCHICAL PROBABILITY MODEL OF COLON CANCER

MICHAEL KELLY,* *University of California, San Diego*

Abstract

We consider a model of fixed size $N = 2^l$ in which there are l generations of daughter cells and a stem cell. In each generation i there are 2^{i-1} daughter cells. At each integral time unit the cells split so that the stem cell splits into a stem cell and generation 1 daughter cell and the generation i daughter cells become two cells of generation $i + 1$. The last generation is removed from the population. A stem cell acquires first and second mutations at rates u_1 and u_2, and a daughter cell acquires first and second mutations at rates v_1 and v_2. We find the distribution for the time it takes to acquire two mutations as N goes to ∞ and the mutation rates go to 0. The mutation rates may tend to 0 at different speeds. We also find the distribution for the locations of the mutations. In particular, we determine whether or not the mutations occur on a stem cell and if not, at what generation in the daughter cells they occur. Several outcomes are possible, depending on how fast the rates go to 0. The model considered has been proposed by Komarova (2007) as a model for colon cancer.

Keywords: Cancer; mutation; Poisson process; population model

2010 Mathematics Subject Classification: Primary 60J99
Secondary 92D99

1. Introduction

In the 1950s Armitage and Doll [1] proposed that cancer may be the end result of an accumulation of two or more cell mutations. Later, Cairns [2] first raised the question of how stem cells affect the development of cancer. We are interested in a particular model in which stem cells play a central role. Komarova [12] discussed three mathematical models which may be used to model the mutations that lead to cancer. The first is the Moran model, which may be used to model cancers in liquids such as leukemia. In this model there is a fixed population of size N. Each of the cells acquires mutations independently at rate μ. Each cell in the population dies at rate 1 and is replaced by any individual in the population, including itself, with equal probability. The second is a spatial model which may be used to model cancers in solid tissues. This model is similar to the Moran model except that the cells are given spatial locations and, when they die, they are only replaced by nearby cells. The third model, the one we focus on in this paper, is referred to as the hierarchical model in [12]. The difference between this model and the other two is that we consider the difference between stem cells and daughter cells. This model was proposed in [12] as a model for colon cancer.

As discussed in [12], many cells in the human body, including those in the colon, go through a three-step process. It begins with a stem cell which will stay in the population for a long time and have many descendants. Some of these descendants will also be stem cells, but others will be differentiated progenitor cells. The progenitor cells, or what we will refer to as daughter

Received 1 April 2011; revision received 11 May 2012.
* Current address: School of Mathematics, University of Minnesota, 127 Vincent Hall, 206 Church St. SE, Minneapolis, MN 55455, USA. Email address: mbkelly@math.ucsd.edu

cells in this paper, will split into more daughter cells. The number of times these cells split is dependent upon what organ of the body they are in. We will refer to the number of splits that a daughter cell has undergone as the generation of the cell. Once the cells split enough times they reach maturity and are swept out of the population in a biological process called apoptosis. The colon is lined with crypts that contain pockets of cells. The cells in the colon, as described by Komarova [14], are such that stem cells reside at the bottom of the crypt and the daughters migrate up the crypt so that the higher-generation daughter cells are near the top.

One can find many conjectures on the number of mutations necessary to cause cancer. In the original model proposed in [12], cancer is the result of two mutations. The reason for two mutations is that it represents the inactivation of two alleles in a tumor suppressor gene. Knudson claimed that retinoblastoma is the result of two mutations in [10] and [11]. For other sources on two-mutation models of cancer, we refer the reader to [8], [15], and [17]. We also model cancer as a result of two mutations. In the hierarchical model there are three ways in which the mutations may occur. Stem cells may acquire both mutations so that cancer is a result of mutations of stem cells only. It is possible that a stem cell receives the first mutation and a daughter cell receives the second, or a daughter cell and one of its descendants will each receive mutations before they are swept from the crypt. In [12] these cases are abbreviated as ss, sd, and dd, respectively.

The hierarchical model will be referred to as H_1. This model has a fixed population of size $N = 2^l$, where l is the number of generations of daughter cells in the crypt. At all times $t \geq 0$ there is one stem cell and, for $k \in \{1, 2, \ldots, l\}$, there are 2^{k-1} daughter cells of generation k. We start with a full crypt and no mutations. At each integral time unit all of the cells split in the following way.

- The stem cell splits into a stem cell and a generation 1 daughter cell.
- For each generation k with $1 \leq k \leq l-1$, a daughter cell of generation k will split into two cells of generation $k+1$.
- The daughter cells of generation l undergo apoptosis and are swept from the population.

Note that the generations are a constant size throughout time. The cells will accumulate mutations via Poisson processes. A cell with 0, 1, or 2 mutations is called a type-0, type-1, or type-2 cell, respectively. A mutation which occurs on a type-0 or type-1 cell is called a type-1 or type-2 mutation, respectively. This terminology is used so that a mutation that makes a cell type 2 is called a type-2 mutation. Once a type-2 mutation occurs the colon is assumed to have cancer. The cells will each have two Poisson processes marking them, one which will cause type-1 mutations and one which will cause type-2 mutations. The first Poisson process that marks a cell will only cause a type-1 mutation if the cell is type 0. If a mark of the Poisson process occurs while the cell is not type 0 then nothing happens. Likewise, the second Poisson process only causes mutations on type-1 cells. If a mark from this Poisson process occurs on a cell while it is type 1 then the cell becomes type 2, but if the cell is not type 1 then nothing happens. All of the Poisson processes are independent. The mutations are passed to the descendants when a cell splits. It is sometimes convenient to think of the cells as fixed in a binary tree and the mutations as traveling through the tree in a direction which takes them from the root to the leaves. Because of this we will often refer to the sequence of stem cells as the stem cell line and we fix the Poisson processes that are marking the cells on particular locations in the tree.

We should mention that several other very similar models have been used to study how stem cells affect the development of cancer. In [13], Komarova and Cheng considered the effects of

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

the development of cancer based on the quantity of stem cells in the population. In [6], Frank *et al.* considered a model in which the stem cells split only finitely many times.

For our model, the rates at which stem cells acquire type-1 and type-2 mutations are u_1 and u_2, respectively. The rates at which the daughter cells acquire type-1 and type-2 mutations are v_1 and v_2, respectively. All of the rates are functions of N and will approach 0 as N approaches ∞. We will always consider what happens as N goes to ∞. All limits will be assumed as taking N to ∞ unless otherwise stated.

A type-1 mutation to a cell is called successful if that cell or one of its descendants receives a type-2 mutation. A type-1 mutation to a stem cell is always successful and a type-1 mutation to a daughter cell is successful if the daughter cell has a type-2 descendent before its progeny is eliminated from the population. We will call the successful type-1 mutation whose type-2 descendant is the first type 2 to occur the cancer causing type-1 mutation. Note that being the cancer-causing type-1 mutation is not equivalent to being the first successful type-1 mutation.

We prove the theorem by coupling various models. This motivates us to define the following functions.

- $\tau'(A)$ is the time at which the cancer-causing type-1 mutation occurs in model A.
- $\tau(A)$ is the first time that any cell acquires a type-2 mutation in model A.
- $\sigma(A) := j/l$ when the cancer-causing type-1 mutation occurs in generation j in model A. If the cancer-causing type-1 mutation occurs on a stem cell in model A then $\sigma(A) = 0$.
- $\rho(A) := j/l$ when the first type-2 mutation occurs in generation j in model A. If the first type-2 mutation occurs on a stem cell in model A then $\rho(A) = 0$.

One of the two goals of this paper is to find the asymptotic distribution of $\tau(H_1)$ as N approaches ∞. Similar work has been done for the Moran model by Schweinsberg [18] and Durrett *et al.* [5], in which more general results have already been found, and for the spatial model by Durrett and Moseley [4]. In [12], Komarova made the following connection between the Moran model and the hierarchical model. In the Moran model a mutation may undergo fixation, meaning that it spreads throughout the entire population through the birth–death process and all of the cells are the same type. Because the last generation is always removed in the hierarchical model, the only way to get fixation is if a stem cell acquires a mutation. These are the cases ss and sd. In these cases the mutation will spread throughout the population in l time units. In the Moran model it is also possible that the progeny of mutated cells undergo what is called stochastic tunneling. This is when multiple mutations are acquired before they fixate. This is analogous to daughter cells acquiring two mutations before a stem cell aquires one mutation in the hierarchical model. This is the dd case and can also happen in the sd case if the second mutation occurs before the first has time to fixate (in particular, the second mutation occurs in less than l time units).

The rate at which daughter cells acquire successful type-1 mutations is given in [12] to be approximately

$$\sum_{i=1}^{l} v_1 2^{i-1}(1 - \mathrm{e}^{-v_2(2^{l-i+1}-2)}). \tag{1}$$

To see this, suppose that all the cells are type 0. When all of the cells in generation i are type 0, then type-1 mutations occur on this generation at rate $v_1 2^{i-1}$. Each of the cells will have $2^{l-i+1} - 2$ descendants. Every descendant lives for one time unit and acquires type-2

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

mutations at rate v_2. This gives the probability of success of a type-1 mutation in generation i to be approximately $1 - \mathrm{e}^{-v_2(2^{l-i+1}-2)}$. Then we sum over all generations.

Our second goal is to determine the limiting distributions of $\sigma(H_1)$ and $\rho(H_1)$. The location of the mutations can be essential to the treatment of cancer. As an example, studies of the effects of the drug imatinib on chronic myeloid leukemia have shown that leukemic stem cells will most likely not cause tumors but rather that a tumor is a result of a mutation on one of the daughter cells; see [3] and [16]. Imatinib treats leukemic daughter cells but not leukemic stem cells. While using imatinib problems arising from cancer are prevented, but patients cannot stop treatment because the leukemic stem cells will continue producing new leukemic daughter cells. Therefore, the location of where the mutations occur may play a pivotal role in determining how to treat the cancer.

We do not find the limiting distribution of $\tau'(H_1)$ as there seems to be no motivation to do so. We only make the definition $\tau'(A)$ because it will occasionally be useful for achieving the two goals described above.

We have established most of the notation above, but some more will be included here. For any real number a, we define $a^+ = a \vee 0$. For functions $f(x)$ and $g(x)$ we will denote the limits $f(x)/g(x) \to 0$, $f(x)/g(x) \to 1$, and $f(x)/g(x) \to \infty$ as $x \to \infty$ by $f \ll g$, $f \sim g$, and $f \gg g$, respectively. To reduce the number of subscripts, we will use $\log x$ for $\log_2 x$. Note that, with this notation, $l = \log N$. We will use '$\xrightarrow{\mathrm{D}}$' to denote convergence in distribution and '$\xrightarrow{\mathrm{P}}$' to denote convergence in probability. We make the following assumptions throughout most of the paper.

Assumption 1. *There exist constants $\alpha, \beta > 0$ such that $v_2 \sim \beta N^{-\alpha}$.*

Assumption 2. *The mutation rates satisfy $u_1 \le u_2$ and $v_1 \le c v_2$ for some $c > 0$.*

We do not allow $\alpha = 0$ so as to reduce the number of cases to be considered. As a result of Assumption 1, the probability that the cancer causing type-1 mutation occurs on a daughter cell in generation $i < l(1-\alpha)^+$ tends to 0. According to Komarova [13], Assumption 2 agrees with almost all of the biologically relevant cases. We let X be an exponentially distributed random variable with mean 1, and we let Y be a random variable with the Rayleigh distribution so that $\mathrm{P}(Y \le t) = 1 - \mathrm{e}^{-t^2/2}$ for any $t > 0$.

The following theorem is the goal of this paper.

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Recall that all limits are taken as N goes to ∞.*

1. *If $v_1 v_2 \ll 1/(N(\log N)^2)$ and $v_1 v_2 N \log N \gg u_1$, $(\alpha \wedge 1) v_1 v_2 N (\log N) \tau(H_1) \xrightarrow{\mathrm{D}} X$. The distribution of $\sigma(H_1)$ converges to the uniform distribution on $((1-\alpha)^+, 1]$ and $\rho(H_1)$ converges in probability to 1.*
2. *If $1/(N(\log N)^2) \ll v_1 v_2 \ll 1/N$ and $v_1 v_2 \gg u_1^2/N$, then $\sqrt{v_1 v_2 N}\,\tau(H_1) \xrightarrow{\mathrm{D}} Y$. Both $\sigma(H_1)$ and $\rho(H_1)$ converge in probability to 1.*
3. *If $v_1 v_2 \gg 1/N$ then $\sqrt{v_1 v_2 N}\,\tau(H_1) \xrightarrow{\mathrm{D}} Y$. Both $\sigma(H_1)$ and $\rho(H_1)$ converge in probability to 1.*
4. *Assume that the following two conditions hold.*
   - *Either $v_1 v_2 \ll 1/(N(\log N)^2)$ and $u_1 \gg v_1 v_2 N \log N$ or $1/(N(\log N)^2) \ll v_1 v_2 \ll 1/N$ and $u_1 \gg \sqrt{v_1 v_2 N}$.*
   - *Both $u_2 \ll 1/\log N$ and $u_2 \ll v_2 N$.*

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

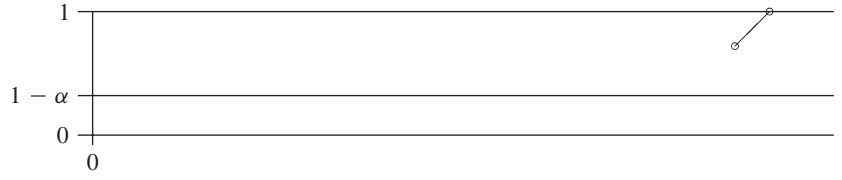

FIGURE 1: Case 1 of Theorem 1.

Then $u_1\tau(H_1) \xrightarrow{\mathrm{D}} X$. *The probability that the first mutation occurs on the stem cell line converges to* 1 *and* $\rho(H_1)$ *converges in probability to* $\alpha \wedge 1$.

5. *Assume that the following two conditions hold.*

   - *Either* $v_1v_2 \ll 1/(N(\log N)^2)$ *and* $u_1 \gg v_1v_2N\log N$ *or* $1/(N(\log N)^2) \ll v_1v_2 \ll 1/N$ *and* $u_1 \gg \sqrt{v_1v_2N}$.
   - *Either* $u_2 \gg 1/\log N$ *or* $u_2 \gg v_2N$.

   Then the probability that both mutations occur on the stem cell line converges to 1. *If* $u_1 \ll u_2$ *then* $u_1\tau(H_1) \xrightarrow{\mathrm{D}} X$ *and if* $u_1 \sim Au_2$ *for some* $A > 0$ *then* $u_1\tau(H_1) \xrightarrow{\mathrm{D}} X + Z$, *where* Z *is an exponentially distributed random variable with mean* A *which is independent of* X.

The first three cases of Theorem 1 are the dd regime. Case 4 is the sd regime and case 5 is the ss regime.

In case 1 the condition $v_1v_2 \ll 1/(N(\log N)^2)$ indicates that, with probability tending to 1, the first successful type-1 mutation on a daughter cell will occur after $\log N$ time. The condition $v_1v_2N\log N \gg u_1$ indicates that a type-2 mutation will occur on a daughter cell before a type-1 mutation occurs on a stem cell with probability tending to 1. Because the amount of time that can pass between a successful type-1 mutation and a type-2 mutation is bounded by $\log N$, the time it takes for the type-2 mutation to occur is negligible in the limit. This is why the distribution of $\tau(H_1)$ converges to an exponential distribution.

There is a useful picture to keep in mind. We will graph time scaled by $1/\log N$ on the horizontal axis and generation scaled by $1/\log N$ on the vertical axis. A mutation on a cell in generation i at time t will be represented by a circle at $(t/l, i/l)$. We represent only the successful type-1 and type-2 mutations. When a successful type-1 mutation is marked, the following type-2 mutation will be connected to it by a line. Figure 1 is an illustration of case 1.

The distribution of $\sigma(H_1)$ arises from a balance between the large number of cells in the later generations versus the large number of descendants of cells in the earlier generations as discussed above. The reasoning used to derive (1) shows that generation i acquires mutations at a rate of approximately

$$v_1 2^{i-1}(1 - \mathrm{e}^{-v_2(2^{l-i+1})}) \approx v_1v_2N.$$

Note that the approximate rate is independent of i. This balance causes the distribution of the marks of the successful type-1 mutations to converge to a uniform Poisson process on $[0, \infty) \times ((1-\alpha)^+, 1)$. The probability that the second mutation occurs in the later generations is just a result of the bulk of the population being concentrated in the later generations.

In case 2 the condition $1/(N(\log N)^2) \ll v_1v_2$ indicates that a daughter cell will acquire a successful type-1 mutation before $\log N$ time with probability tending to 1. The condition $v_1v_2 \ll 1/N$ indicates that the time it takes for a successful type-1 mutation to occur on a

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

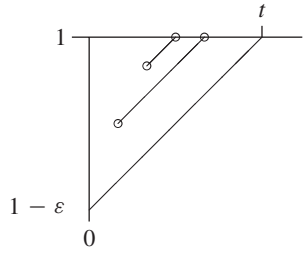

FIGURE 2: Case 2 of Theorem 1—a magnified image of the top-left corner.

daughter cell tends to ∞. The condition $v_1 v_2 \gg u_1^2/N$ indicates that the cancer-causing type-1 mutation will occur on a daughter cell with probability tending to 1. As in case 1, $\rho(H_1) \xrightarrow{P} 1$ because most of the cells are in the later generations. Because cells split at rate 1 it takes $O(\log N)$ time units before a significant number of an individual's progeny is realized. In this case the type-2 mutation will occur much faster than $\log N$ time with probability tending to 1. Therefore, an individual's progeny does not play such an important role. For this reason, the cancer-causing type-1 mutation is approximately equally likely to occur on any cell. Most of the cells are in the later generations so $\sigma(H_1)$ tends to 1. We illustrate this case in Figure 2.

In Figure 2, a type-2 mutation will occur by time t if a successful type-1 mutation has occurred in the triangle beneath time t. Note that Figure 2 illustrates an example in which the first successful type-1 mutation is not the cancer-causing type-1 mutation. Because the marks of the type-1 mutations are converging to a uniform Poisson process in the triangle, the distribution of $\tau(H_1)$ will converge to the Rayleigh distribution.

In case 3 the condition $v_1 v_2 \gg 1/N$ indicates that some cell will receive two mutations before time 1 with probability tending to 1. Any daughter cell is equally likely to acquire the two mutations and because $u_1 \to 0$ the probability that the stem cell acquires the two mutations tends to 0. This causes $\sigma(H_1)$ and $\rho(H_1)$ to tend to 1 in probability since the bulk of the population is concentrated in the later generations. The waiting time for the first individual to acquire two mutations has a Rayleigh distribution, which gives the result for $\tau(H_1)$. The results hold for this case when $\alpha = 0$.

We now explain the assumptions of case 4 which ensure that the sd regime occurs with probability tending to 1. If stem cells could not mutate and $v_1 v_2 \ll 1/(N(\log N)^2)$, then, according to case 1, $(\alpha \wedge 1) v_1 v_2 N \log N \tau(H_1) \xrightarrow{D} X$. The condition $u_1 \gg v_1 v_2 N \log N$ indicates that a type-1 mutation occurs on the stem cell line before a type-2 mutation occurs on a daughter cell when the mutation rates of the daughter cells satisfy $v_1 v_2 \ll 1/N(\log N)^2$. Likewise, if the stem cell could not mutate and $1/(N(\log N)^2) \ll v_1 v_2 \ll 1/N$, then, according to case 2, $\sqrt{v_1 v_2 N}\tau(H_1) \xrightarrow{D} Y$. The condition $u_1 \gg \sqrt{v_1 v_2 N}$ indicates that the stem cell line acquires a type-1 mutation before the daughter cells acquires a type-2 mutation when the mutation rates of the daughter cells satisfy $1/(N(\log N)^2) \ll v_1 v_2 \ll 1/N$. The condition $u_2 \gg 1/\log N$ or $u_2 \gg v_2 N$ indicates that the first type-2 mutation occurs on a daughter cell rather than the stem cell line.

In case 4 the time at which the type-1 mutation occurs on the stem cell line is much larger than $\log N$ with probability tending to 1. Therefore, the time it takes for the first type-2 mutation to occur is negligible. This implies that the type-1 mutation that occurs on the stem cell line is the cancer-causing type-1 mutation with probability tending to 1 and illustrates why $u_1 \tau(H_1)$ is converging to an exponential distribution. Once a stem cell acquires a type-1 mutation the daughter cells inherit the type-1 mutation at an exponential rate. For any $\varepsilon > 0$, the probability

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

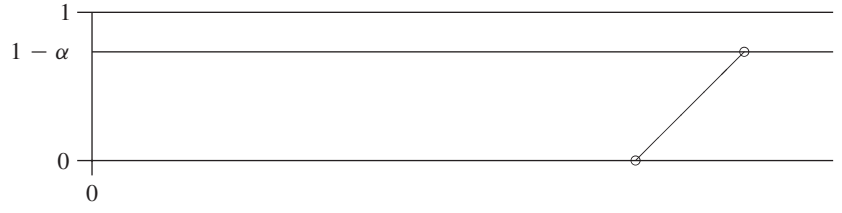

FIGURE 3: Case 4 of Theorem 1—stem cell mutations occur on $[0, \infty) \times \{0\}$.

that the first type-2 mutation will occur when the type-1 mutation has spread to generation i for some $i \in ((\alpha \wedge 1 - \varepsilon) \log N, (\alpha \wedge 1 + \varepsilon) \log N)$ tends to 1. This is why $\rho(H_1) \xrightarrow{\mathrm{P}} (\alpha \wedge 1)$. Figure 3 gives an illustration of this case.

The first condition in case 5 is the same as the first condition in case 4. Under this condition, the probability that the first successful type-1 mutation occurs on the stem cell line tends to 1. The second condition in case 5 implies that the first type-2 mutation occurs on the stem cell line with probability tending to 1.

The results for $\tau(H_1)$ are similar to the results when waiting for two mutations in the Moran model. In particular, when the mutation rates are slow in the Moran model, the time until two mutations converges to the exponential distribution and when the rates are faster, the waiting time converges to the Rayleigh distribution. The original results can be found in [8] and [19], and they are also a special case of the results in [18].

There are many boundary cases and most of them are not included in this paper, where we use the term boundary case to refer to the boundary between two of the conditions. That is, if $v_1 \ll 1/N$ gives one result and $v_1 \gg 1/N$ gives another, we would consider $v_1 \sim A/N$ for some constant A to be a boundary case. If included, the boundary cases would make up the bulk of this paper. One reason for this is that our variables $\{v_1, v_2, u_1, u_2\}$ span a four-dimensional space, so the regions will have many boundaries. Moreover, sometimes three regions intersect in the same place. It does not seem that there would be any special difficulties in computing most of these boundary cases using the same methods used in this paper.

We call H_1 the null model when all of the mutation rates are the same. The following proposition gives the results for the null model, including results for the boundary cases.

Proposition 1. *Let $\mu = u_1 = u_2 = v_1 = v_2$. Suppose that Assumption 1 holds, so that there exist constants $\beta, \alpha > 0$ such that $\mu \sim \beta N^{-\alpha}$.*

1. *If $\mu \ll 1/(N \log N)$ then $\mu\tau(H_1) \xrightarrow{\mathrm{D}} X$. The probability that the first successful type-1 mutation occurs on the stem cell line converges to 1 and $\rho(H_1)$ converges in probability to 1.*

2. *If $\mu \sim A/(N \log N)$ then $(1 + A)\mu\tau(H_1) \xrightarrow{\mathrm{D}} X$. Let ξ be a Bernoulli random variable such that $\mathrm{P}(\xi = 1) = A/(1 + A)$ and $\mathrm{P}(\xi = 0) = 1/(1 + A)$. Let U be a random variable, independent of ξ, with the uniform distribution on $[0, 1]$. Then*

$$\sigma(H_1) \xrightarrow{\mathrm{D}} U\xi$$

and

$$\rho(H_1) \xrightarrow{\mathrm{P}} 1.$$

3. *If $1/(N \log N) \ll \mu \ll 1/(\sqrt{N} \log N)$ then $(\alpha \wedge 1)\mu^2 N(\log N)\tau(H_1) \xrightarrow{\mathrm{D}} X$. The distribution of $\sigma(H_1)$ converges to a uniform distribution on $((1 - \alpha)^+, 1]$ and $\rho(H_1)$ converges in probability to 1.*

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

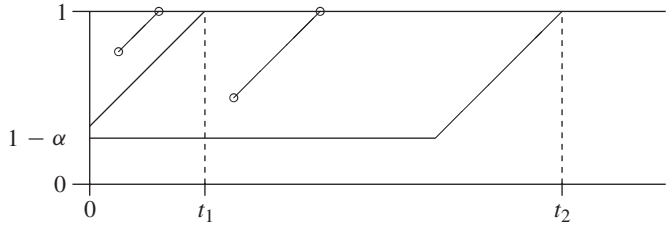

FIGURE 4: Case 4 of Proposition 1.

4. *If* $\mu \sim A/(\sqrt{N}\log N)$ *then*

$$\lim \mathrm{P}\left(\frac{\tau(H_1)}{\log N} \le t\right) = (1 - \mathrm{e}^{-A^2t^2/2})\,\mathbf{1}_{[0,1/2]}(t) + (1 - \mathrm{e}^{-A^2t/2+A^2/8})\,\mathbf{1}_{(1/2,\infty)}(t).$$

Let Z *be a random variable with density*

$$f(x) = \left(\int_{1-x}^{1/2} A^2\mathrm{e}^{-A^2t^2/2}\,\mathrm{d}t + 2\mathrm{e}^{-A^2/8}\right)\mathbf{1}_{[1/2,1]}(x).$$

As N *goes to* ∞, $\sigma(H_1)$ *converges in distribution to* Z *and* $\rho(H_1)$ *converges in probability to* 1.

5. *If* $1/(\sqrt{N}\log N) \ll \mu \ll 1/\sqrt{N}$ *then* $\mu\sqrt{N}\tau(H_1) \xrightarrow{\mathrm{D}} Y$. *Both* $\sigma(H_1)$ *and* $\rho(H_1)$ *converge in probability to* 1.

6. *If* $\mu \sim A/\sqrt{N}$ *then, for each fixed time* $t > 0$, *there exist constants* c *and* C *such that* $\liminf \mathrm{P}(\tau(H_1) \le t) \ge c > 0$ *and* $\limsup \mathrm{P}(\tau(H_1) \le t) \le C < 1$. *Both* $\sigma(H_1)$ *and* $\rho(H_1)$ *converge in probability to* 1.

7. *If* $1/\sqrt{N} \ll \mu$ *then* $\mu\sqrt{N}\tau(H_1) \xrightarrow{\mathrm{D}} Y$. *Both* $\sigma(H_1)$ *and* $\rho(H_1)$ *converge in probability to* 1.

Parts 1, 3, 5, and 7 of Proposition 1 follow directly from Theorem 1. Parts 2, 4, and 6, the boundary cases, will be proved in Section 6.

In part 2 the cancer-causing type-1 mutation may occur on a stem cell or a daughter cell. The event $\xi = 1$ corresponds to the cancer causing type-1 mutation occurring on a daughter cell and the event $\xi = 0$ corresponds to the cancer-causing type-1 mutation occurring on the stem cell line.

In part 4 the mutations occur in $O(\log N)$ time units. Figure 4 is an illustration for this case.

Note that the exponents in the limiting distribution for $\tau(H_1)$ in part 4 correspond to the area of a triangle or quadrilateral. This is because the cancer-causing type-1 mutation will occur in $O(\log N)$ time units. Let t_1 and t_2 be the times marked in Figure 4. The probability that a type-2 mutation has occurred by time $t_1/\log N$ is the probability that a mark indicating a successful type-1 mutation has occurred in the triangle associated with t_1 in Figure 4. Likewise, the probability that a type-2 mutation has occurred by time $t_2/\log N$ is the probability that a mark indicating a successful type-1 mutation has occurred in the quadrilateral associated with t_2 in Figure 4.

The main result of part 6 is that, when $\mu \sim A/\sqrt{N}$, the time until two mutations is $O(1)$. The results are therefore affected by the discreteness of the model.

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

In the next section we introduce a new model which will be coupled with H_1. Theorem 1 will be proved with this new model in place of H_1 and the coupling will give the results for H_1. The third section of this paper is devoted to obtaining results about the dd regime. The fourth section is on results about the sd and ss regimes. In Section 5 we prove Theorem 1. The last section is a discussion of the boundary cases in the null model and a proof of Proposition 1. In Appendix A we give a glossary of the notation and of the descriptions of the auxiliary models used throughout the paper.

2. A useful model

In this section we define a new model, H_2, which will be useful to compare with H_1. In model H_2 there is one stem cell and, for each integer i, there are 2^{i-1} generation i daughter cells for all times $t \geq 0$. The cells in model H_2 split at each integral time unit in the same way that the cells in model H_1 split. Just as in model H_1, the stem cells in model H_2 receive type-1 and type-2 mutations at rates u_1 and u_2, respectively, and the daughter cells receive type-1 and type-2 mutations at rates v_1 and v_2, respectively. The difference between the models is how the cells accumulate type-1 mutations. In model H_1 all type-1 mutations have the same behavior. A type-1 mutation proposed to occur on a type-1 cell in H_1 is rejected because the cell is already a type 1. In model H_2 the behavior of type-1 mutations differ depending on whether or not the mutation occurred on a stem cell. If a type-1 mutation occurs on a stem cell, it has the same behavior as in model H_1. The mutation will eventually be passed to all other cells in the population and any type-1 mutation proposed to occur on a type-1 stem cell or a daughter cell that is the progeny of a type-1 stem cell is rejected. However, all type-1 daughter cells which are type-1 cells as a result of a type-1 mutation occurring on a daughter cell are able to accumulate type-1 mutations. If a type-1 mutation is proposed to occur on such a daughter cell with one type-1 mutation, then the mutation is accepted and the cell now carries two type-1 mutations. Type-1 mutations to type-0 daughter cells result in cells that are allowed to carry any number of type-1 mutations, and when a cell has k type-1 mutations, it receives type-2 mutations at rate kv_2. Because the type-1 mutations on daughter cells do not change the rate at which type-1 mutations occur, (1) is more accurate for model H_2.

We now give an alternate description of model H_2 which will allow us to make a coupling between models H_1 and H_2. Consider the daughter cells as fixed in a tree and consider the mutations as moving to the higher-generation daughter cells at each integral time unit in model H_1. Label the daughter cells $D_1, D_2, \ldots, D_{N-1}$.

In model H_2 each daughter cell D_i has a counter C_i starting at 0 and is acted on by a sequence of Poisson processes $\{P_n^i\}_{n=1}^{\infty}$, each having rate v_2, which determine the type-2 mutations. All of the Poisson processes are independent of one another. When a type-1 mutation occurs on a daughter cell D_i, it increases the counter C_i by 1. This is considered to be a type-1 mutation. If a type-1 mutation increases the counter to n, it is the nth type-1 mutation on the cell. When the counter C_i has reached n, all type-2 mutations that would occur according to the Poisson processes $P_1^i, P_2^i, \ldots, P_n^i$ are accepted as type-2 mutations on cell D_i. All type-2 mutations that would occur according to the Poisson processes $P_{n+1}^i, P_{n+2}^i, \ldots$ are rejected. If a type-2 mutation occurs on cell D_i as a result of the Poisson process P_n^i, then the nth type-1 mutation according to C_i is considered to be successful. If the first type-2 mutation on a cell is a result of the Poisson process P_n^i, then the nth type-1 mutation according to C_i is the cancer-causing type-1 mutation. Rather than the mutations moving up the tree, at each integral time unit the daughter cells in generations $i \geq 2$ will inherit the counter number from their ancestor in the previous generation. The daughter cell in generation 1 will reset its counter to 0 at each

integral time unit. However, a type-1 mutation on a stem cell does not have a counter. Once a type-1 mutation has spread from a stem cell to a daughter cell the daughter cell can no longer accumulate type-1 mutations and the model is the same as model H_1.

We couple H_1 and H_2 as follows.

- The Poisson processes that mark the stem cells are the same.
- If a daughter cell has inherited a type-1 mutation from a stem cell then the Poisson processes marking type-2 mutations on the cell are the same in each model.
- The Poisson processes marking type-1 mutations on daughter cells are the same.
- The Poisson processes marking type-2 mutations on daughter cells in model H_1 are the same as the Poisson processes P_1^i in model H_2 so long as the daughter cells did not inherit their type-1 mutations from a stem cell.

There are no analogous Poisson processes in model H_1 for the $N-1$ sequences of Poisson processes $P_2^i, P_3^i, \ldots$ in model H_2.

Lemma 1. *Let the Poisson processes in models H_1 and H_2 be coupled as described above. Then* $\mathrm{P}(\tau(H_1) = \tau(H_2))$, $\mathrm{P}(\rho(H_1) = \rho(H_2))$, *and* $\mathrm{P}(\sigma(H_1) = \sigma(H_2))$ *all converge to* 1.

Proof. A type-2 mutation which occurs in model H_2 but not in H_1 is a result of the rejection of the type-1 mutation in model H_1 that has led to the type-2 mutation in H_2. This type-1 mutation could only be rejected in model H_1 because the cell on which it was supposed to occur was already a type-1 cell. Type-1 mutations on the stem cell line will occur at the same time in both models. If we consider a type-1 mutation that occurs on a daughter cell in model H_2, the probability that it also occurs in model H_1 is the probability that the cell is a type 0. Because the differentiated cells will be removed from the population after $\log N$ time, if we propose a type-1 mutation at a time t on any cell that has not inherited a type-1 mutation from a stem cell, then the probability that the cell has a type-1 mutation is at most $1 - \mathrm{e}^{-v_1 \log N}$. Therefore, if a type-1 mutation occurs in model H_2 at time t, with probability at least $\mathrm{e}^{-v_1 \log N}$, it will also occur in model H_1. We show that the same will be true of the cancer-causing type-1 mutation.

We number the positions of the cells $1, 2, \ldots, N$ and let 1 be the position of the stem cell line. Let $\bar{N} = \{1, 2, \ldots, N\}$ and $L = [0, l] \cup \{\infty\}$. First we note that the Poisson processes marking the daughter cells in model H_2 induce a Poisson process on the space $[0, \infty) \times \bar{N} \times L$. A point (t, i, s) is marked to indicate that a type-1 mutation occurred at time t on the cell at location i and at time $s+t$ the type-1 mutation became successful. If the type-1 mutation is not successful then $s = \infty$. One may note that this is a Poisson process by two applications of the marking theorem (see [9, p. 55]). Type-1 mutations occur according to a Poisson process on $[0, \infty)$ at rate $v_1(N-1) + u_1$. Each daughter cell has probability $v_1/(v_1(N-1)+u_1)$ of being the cell that receives the type-1 mutation and the stem cell has probability $u_1/(v_1(N-1)+u_1)$ of being the cell that receives the type-1 mutation. By a first application of the marking theorem this gives us a Poisson process on $[0, \infty) \times \bar{N}$. The probability that a type-1 mutation is successful can be determined from the associated point (t, i) which tells us at what time and on what cell the type-1 mutation occurred. Each one of these points has an associated value s that indicates when, and if, the type-1 mutation becomes successful. This gives the Poisson process on $[0, \infty) \times \bar{N} \times L$.

Let Z be the random variable which indicates the value in $[0, \infty) \times \bar{N} \times L$ that corresponds to the time of the cancer-causing type-1 mutation, the cell on which it occurred, and the time of

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

the first type-2 mutation. If we condition on the event $Z = (t_0, i_0, s_0)$ for some i_0 in generation j which is not the stem cell line, then there can be no marks in subset

$$\{(t, i, s) : s < t_0 + s_0 - t\} \cup \{(t, i, s) : (\lceil t \rceil - i, \{1\}, s + t - \lceil t \rceil)\}$$

of $[0, \infty) \times \bar{N} \times L$. The marks that occur outside of this subset occur independently of the marks that occur within. Conditioning does not change the probability that a mark outside of this set has occurred by time t_0. This only reduces the rate at which type-1 mutations occur before time t_0. Therefore, $\mathrm{P}(\tau(H_1) \neq \tau(H_2) \mid Z = (t_0, i_0, s_0)) \leq 1 - \mathrm{e}^{-v_1 \log N}$. Let P_Z be the probability measure on $[0, \infty) \times \bar{N} \times L$ induced by Z. Then

$$\begin{aligned}
\mathrm{P}(\tau(H_1) \neq \tau(H_2)) &= \int_{[0,\infty)\times\bar{N}\times L} \mathrm{P}(\tau(H_1) \neq \tau(H_2) \mid Z = x)\, \mathrm{P}_Z(\mathrm{d}x) \\
&\leq \int_{[0,\infty)\times\bar{N}\times L} (1 - \mathrm{e}^{-v_1 \log N})\, \mathrm{P}_Z(\mathrm{d}x) \\
&= 1 - \mathrm{e}^{-v_1 \log N}.
\end{aligned}$$

This shows that $\mathrm{P}(\tau(H_1) \neq \tau(H_2)) \to 0$ if $v_1 \ll 1/\log N$. It follows from Assumption 1 that $v_2 \ll 1/\log N$, and combining this with Assumption 2 we see that $v_1 \ll 1/\log N$ as well.

On the event $\tau(H_1) = \tau(H_2)$ we have $\rho(H_1) = \rho(H_2)$ and $\sigma(H_1) = \sigma(H_2)$ with probability 1. The only way these equalities can fail is if two type-2 mutations occur simultaneously in model H_2, an event whose probability is 0. Therefore, $\mathrm{P}(\rho(H_1) = \rho(H_2))$ and $\mathrm{P}(\sigma(H_1) = \sigma(H_2))$ both converge to 1 as well.

The rest of the work in proving Theorem 1 is in proving Theorem 1 with H_2 in place of H_1. Once this is done Theorem 1 follows from Lemma 1.

3. The dd regime

To understand the behavior in the dd regime, we consider a new model which is the same as H_2 except that mutations only occur on daughter cells. That is, there are no Poisson processes that mark mutations on the stem cells. This new model will be called model M_1. The purpose of this section is to prove the following proposition.

Proposition 2. 1. *If* $v_1 v_2 \ll 1/(N(\log N)^2)$ *then* $(\alpha \wedge 1) v_1 v_2 N (\log N) \tau(M_1) \xrightarrow{\mathrm{D}} X$. *The distribution of* $\sigma(M_1)$ *converges to a uniform distribution on* $((1-\alpha)^+, 1]$ *and* $\rho(M_1)$ *converges in probability to* 1.

2. *If* $1/(N(\log N)^2) \ll v_1 v_2 \ll 1/N$ *then* $\sqrt{v_1 v_2 N}\, \tau(M_1) \xrightarrow{\mathrm{D}} Y$. *Both* $\sigma(M_1)$ *and* $\rho(M_1)$ *converge in probability to* 1.

Lemma 2. *For any positive integer* $k < l$, *we have* $\mathrm{P}(\rho(M_1) \geq (l-k)/l) > 1 - 1/2^k$.

Proof. Let Z be the number of generations between the cancer-causing type-1 mutation and the first type-2 mutation. Then $Z \in \{0, 1, 2, \ldots, l\}$. Because there are only l generations, if the second mutation occurs $l - k$ generations or more after the first then it must be in the last k generations. So $\mathrm{P}(\rho(M_1) \geq (l-k)/l \mid Z \in \{l-k, l-k+1, \ldots, l\}) = 1$. If we condition on the event that $Z = j$ for some $j \leq l - k - 1$ then the probability that the cancer-causing type-1 mutation occurs on any cell in generations $1, 2, \ldots, l - j$ is equally likely. This is because the Poisson processes marking the mutations on the descendants of the cells j generations after any generation i are independent and identically distributed.

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

The last k of the $l-j$ generations always make up at least a fraction of $1-1/2^k$ cells, so we have $\mathrm{P}(\rho(M_1) \geq (l-k)/l \mid Z \in \{0,1,2,\ldots,l-k-1\}) > 1-1/2^k$, where we obtain a strict inequality because we do not count the stem cell line. The result follows.

It is important to note that Lemma 2 holds for any N and we do not require $N \to \infty$. Also, the rates at which v_1 and v_2 tend to 0 are irrelevant.

Corollary 1. *As N goes to ∞, $\rho(M_1)$ will converge to 1 in probability.*

Lemma 3. *Let $(\beta_1, \beta_2] \subset (0,1]$. Let C be a positive constant, and let $C' \in \{1,2\}$. Then*

$$\sum_{i \in \mathbb{N} \cap (l\beta_1, l\beta_2]} v_1 2^{i-1}(1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')}) \sim C(\beta_2-\beta_1 \vee (1-\alpha))^+ v_1 v_2 N \log N.$$

Proof. We will first define some notation for this proof for the sake of readability. Let $I \subset \mathbb{R}$. We define

$$I^* := I \cap (l\beta_1, l\beta_2] \cap \mathbb{N}.$$

First we consider the case when $\alpha \geq 1$. Using the upper bound $1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')} \leq Cv_2 2^{l-i+1}$, we have

$$\frac{\sum_{i \in (l\beta_1, l\beta_2]^*} v_1 2^{i-1}(1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')})}{v_1 v_2 2^l l} \leq C(\beta_2-\beta_1).$$

From the second-order Taylor expansion we obtain a lower bound of

$$1-\mathrm{e}^{-C(2^{l-i+1}-C')} \geq Cv_2(2^{l-i+1}-C') - \tfrac{1}{2}C^2 v_2^2 (2^{l-i+1}-C')^2.$$

We will break this sum into five parts:

$$2^{i-1}(1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')}) \geq Cv_2 2^l - CC'v_2 2^i - C^2 v_2^2 2^{2l-i} + C^2 C' v_2^2 2^l - C^2 (C')^2 v_2^2 2^{i-2}.$$

Computations for each of the five individual sums give

$$\sum_{i \in (l\beta_1, l\beta_2]^*} \frac{Cv_2 2^l}{v_2 2^l l} \to C(\beta_2-\beta_1),$$

$$\sum_{i \in (l\beta_1, l\beta_2]^*} \frac{CC'v_2 2^i}{v_2 2^l l} \leq \frac{CC' 2^{l+1}}{2^l l} \to 0,$$

$$\sum_{i \in (l\beta_1, l\beta_2]^*} \frac{C^2 (C')^2 v_2^2 2^{i-2}}{v_2 2^l l} \leq \frac{C^2 C'^2 v_2}{l} \to 0,$$

$$\sum_{i \in (l\beta_1, l\beta_2]^*} \frac{C^2 C' v_2^2 2^l}{v_2 2^l l} \leq C^2 C' v_2 \to 0,$$

$$\sum_{i \in (l\beta_1, l\beta_2]^*} \frac{C^2 v_2^2 2^{2l-i}}{v_2 2^l l} = \frac{C^2 v_2 2^l (\sum_{i=\lceil l\beta_1 \rceil}^{\beta_2} 2^{-i})}{l} \leq C^2 v_2 2^{l(\beta_2-\beta_1)} \to 0,$$

so long as $v_2 \ll 1/2^{l(\beta_2-\beta_1)} = N^{-(\beta_2-\beta_1)}$, which will hold since $\alpha \geq 1$ in this case.

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

So we have

$$\lim_{N\to\infty}\left(\frac{\sum_{i\in(l\beta_1,l\beta_2]^*} v_1 2^{i-1}(1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')})}{v_1v_2 2^l l}\right) = C(\beta_2-\beta_1),$$

which completes the $\alpha \geq 1$ case.

Now let $0<\alpha<1$ and let $\varepsilon>0$ be small enough so that $0<1-\alpha-\varepsilon<1-\alpha+\varepsilon<1$. We now break the sum into three blocks:

$$\frac{\sum_{i\in[1,l(1-\alpha-\varepsilon))^*\cup[l(1-\alpha-\varepsilon),l(1-\alpha+\varepsilon)]^*\cup(l(1-\alpha+\varepsilon),l]^*} 2^{i-1}(1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')})}{v_2 2^l l}.$$

We can consider each of these three sums individually.

For the middle sum, we only need the bound

$$0 \leq \frac{\sum_{i\in[l(1-\alpha-\varepsilon),l(1-\alpha+\varepsilon)]^*} 2^{i-1}(1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')})}{v_2 2^l l} \leq 2C\varepsilon,$$

which follows by the upper bound $1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')} \leq Cv_2 2^{l-i+1}$.

One can apply similar computations as in the $\alpha=1$ case to obtain

$$\frac{\sum_{i\in(l(1-\alpha+\varepsilon),l]^*} 2^{i-1}(1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')})}{v_2 2^l l} \to C(\beta_2-\beta_1\vee(1-\alpha+\varepsilon))^+.$$

For the first sum, note that $1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')}\leq 1$. This gives the bound

$$\begin{aligned}
0 &\leq \sum_{i\in[1,l(1-\alpha-\varepsilon))^*} \frac{2^{i-1}(1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')})}{v_2 2^l l} \\
&\leq \sum_{i\in[1,l(1-\alpha-\varepsilon))^*} \frac{2^{i-1}}{v_2 2^l l} \\
&\leq \frac{2^{l(1-\alpha-\varepsilon)}}{v_2 2^l l} \\
&\to 0.
\end{aligned}$$

The convergence is a result of the definition of α. In particular, $v_2 \gg N^{-\alpha-\varepsilon}(\log N)^{-1}$.

Combining the three sums yields

$$C(\beta_2-\beta_1\vee(1-\alpha+\varepsilon))^+ \leq \liminf \frac{\sum_{i\in(l\beta_1,l\beta_2]^*} v_1 2^{i-1}(1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')})}{l v_1 v_2 2^l}$$

and

$$\limsup \frac{\sum_{i\in(l\beta_1,l\beta_2]^*} v_1 2^{i-1}(1-\mathrm{e}^{-Cv_2(2^{l-i+1}-C')})}{l v_1 v_2 2^l} \leq C(\beta_2-\beta_1\vee(1-\alpha+\varepsilon))^+ + 2C\varepsilon.$$

Letting ε approach 0 gives the result.

Corollary 2. *Let T be the time at which the first successful type-1 mutation occurs. Then $(\alpha\wedge 1)v_1v_2N(\log N)T \xrightarrow{\mathrm{D}} X$.*

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

Proof. For $1 \le i \le l$, there are 2^{i-1} cells in generation i. Each of these cells acquires type-1 mutations at rate v_1. The cells in generation i have $2^{l-i+1} - 2$ descendants. If the cell splits as soon as it becomes a type 1, the probability that none of its descendants acquire a type-2 mutation is $e^{-v_2(2^{l-i+1}-2)}$. On the other hand, after a cell acquires a type-1 mutation it could live for at most 1 time unit until it splits. If this is the case then the probability that neither the cell that receives the type-1 mutation nor any of its descendants receives a type-2 mutation is $e^{-v_2(2^{l-i+1}-1)}$. If we let $R(t)$ be the rate at which the successful type-1 mutations occur at time t then, for any time t, we have

$$\begin{aligned} 1 &= \lim \frac{\sum_{i=1}^{l} v_1 2^{i-1}(1 - e^{-v_2(2^{l-i+1}-2)})}{(\alpha \wedge 1) v_1 v_2 N \log N} \\ &\le \liminf \frac{R(t)}{(\alpha \wedge 1) v_1 v_2 N \log N} \\ &\le \limsup \frac{R(t)}{(\alpha \wedge 1) v_1 v_2 N \log N} \\ &\le \lim \frac{\sum_{i=1}^{l} v_1 2^{i-1}(1 - e^{-v_2(2^{l-i+1}-1)})}{(\alpha \wedge 1) v_1 v_2 N \log N} \\ &= 1, \end{aligned}$$

where the limits are results of Lemma 3.

The successful type-1 mutations occur according to a time inhomogeneous Poisson process with an intensity measure ν, where $\nu([0,t]) = \int_0^t R(s)\,ds$. We have shown that ν satisfies

$$t \sum_{i=1}^{l} v_1 2^{i-1}(1 - e^{-v_2(2^{l-i+1}-2)}) \le \nu([0,t]) \le t \sum_{i=1}^{l} v_1 2^{i-1}(1 - e^{-v_2(2^{l-i+1}-1)})$$

for all $t \ge 0$ and all N. For any $t \ge 0$, we have

$$P\left(T \le \frac{t}{(\alpha \wedge 1) v_1 v_2 N (\log N)}\right) = 1 - e^{-\nu([0,t])/((\alpha \wedge 1) v_1 v_2 N \log N)} \to 1 - e^{-t},$$

where the limiting results follow by Lemma 3. Therefore, $(\alpha \wedge 1) v_1 v_2 N (\log N) T$ is converging in distribution to an exponentially distributed random variable with parameter 1.

The next lemma states that when $v_1 v_2 \ll 1/(N(\log N)^2)$, the probability that the first successful type-1 mutation is the cancer causing type-1 mutation tends to 1.

Lemma 4. *Let T be the time at which the first successful type-1 mutation occurs in model M_1. If $v_1 v_2 \ll 1/(N(\log N)^2)$ then $P(T = \tau'(M_1)) \to 1$.*

Proof. Let $Z = \tau(M_1) - T$ be the time it takes to acquire the first type-2 mutation after the first successful type-1 mutation has appeared, and let $\hat{T}$ be the time it takes to acquire the second successful type-1 mutation after the first.

By Corollary 2, $(\alpha \wedge 1) v_1 v_2 N (\log N) T \xrightarrow{D} X$ and $(\alpha \wedge 1) v_1 v_2 N (\log N) \hat{T} \xrightarrow{D} X$. Then, because a type-2 mutation must occur within a $\log N$ time after a successful type-1 mutation on a daughter cell, we have

$$P(\hat{T} < Z) \le P(\hat{T} < \log N) = P((\alpha \wedge 1) v_1 v_2 N (\log N) \hat{T} < (\alpha \wedge 1) v_1 v_2 N (\log N)^2) \to 0.$$

Moreover, $P(\hat{T} \ge Z) \le P(T = \tau'(M_1))$ so $P(T = \tau'(M_1)) \to 1$.

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

Lemma 5. *If* $v_1 v_2 \ll 1/(N(\log N)^2)$ *then* $(\alpha \wedge 1) v_1 v_2 N(\log N)\tau(M_1) \xrightarrow{\mathrm{D}} X$.

Proof. From Lemma 4 we know that the probability that the first successful type-1 mutation is the cancer-causing mutation is converging to 1. Combining this with Corollary 2, $(\alpha \wedge 1) \times v_1 v_2 N(\log N)\tau'(M_1) \xrightarrow{\mathrm{D}} X$.

Owing to apoptosis, $\tau(M_1) - \tau'(M_1)$ is bounded above by $\log N$ so $(\alpha \wedge 1) v_1 v_2 N(\log N) \times (\tau(M_1) - \tau'(M_1)) \xrightarrow{\mathrm{P}} 0$. Then

$$(\alpha \wedge 1) v_1 v_2 N(\log N)\tau(M_1) = (\alpha \wedge 1) v_1 v_2 N(\log N)(\tau'(M_1) + (\tau(M_1) - \tau'(M_1))) \xrightarrow{\mathrm{D}} X.$$

Lemma 6. *If* $v_1 v_2 \ll 1/(N(\log N)^2)$ *then the distribution of* $\sigma(M_1)$ *converges to the uniform distribution on* $((1-\alpha)^+, 1]$.

Proof. By Lemma 4, the first successful type-1 mutation will be the cancer-causing type-1 mutation with probability tending to 1. Therefore, to find the limiting results on $\sigma(M_1)$, it is enough to find the depth at which the first successful type-1 mutation occurs as N tends to ∞.

Each generation i with $1 \le i \le l$ acquires successful type-1 mutations independently at a rate bounded between $v_1 2^{i-1}(1 - \mathrm{e}^{-v_2(2^{l-i+1}-2)})$ and $v_1 2^{i-1}(1 - \mathrm{e}^{-v_2(2^{l-i+1}-1)})$ for any time t. Therefore, for a fixed N and i, the probability that the first successful type-1 mutation occurs on generation i is between

$$\frac{v_1 2^{i-1}(1 - \mathrm{e}^{-v_2(2^{l-i+1}-2)})}{\sum_{j=1}^{l} v_1 2^{j-1}(1 - \mathrm{e}^{-v_2(2^{l-j+1}-1)})}$$

and

$$\frac{v_1 2^{i-1}(1 - \mathrm{e}^{-v_2(2^{l-i+1}-1)})}{\sum_{j=1}^{l} v_1 2^{j-1}(1 - \mathrm{e}^{-v_2(2^{l-j+1}-2)})}.$$

Let $\beta \in [0, 1]$. Using the notation and result from Lemma 3,

$$\limsup \mathrm{P}(\sigma(M_1) \le \beta) \le \limsup \frac{\sum_{i \in (0, l\beta]^*} v_1 2^{i-1}(1 - \mathrm{e}^{-v_2(2^{l-i+1}-1)})}{\sum_{j \in (0,l]^*} v_1 2^{j-1}(1 - \mathrm{e}^{-v_2(2^{l-j+1}-2)})} = \frac{(\beta - (1-\alpha)^+)^+}{\alpha \wedge 1}$$

and

$$\liminf \mathrm{P}(\sigma(M_1) \le \beta) \ge \liminf \frac{\sum_{i \in (0, l\beta]^*} v_1 2^{i-1}(1 - \mathrm{e}^{-v_2(2^{l-i+1}-2)})}{\sum_{j \in (0,l]^*} v_1 2^{j-1}(1 - \mathrm{e}^{-v_2(2^{l-j+1}-1)})} = \frac{(\beta - (1-\alpha)^+)^+}{\alpha \wedge 1}.$$

Combining the results of Corollary 1 and Lemmas 5 and 6 we have part 1 of Proposition 2. For the next two proofs, we note that Corollary 1 already tells us that $\rho(M_1)$ converges to 1 in probability.

Proof of part 2 of Proposition 2. For the slower mutation rates, it was enough to note that a cell in generation i has $2^{l-i+1} - 2$ descendants. Under these conditions, the mutation rates are fast enough that we will need to consider how many descendants a cell in generation i has at a time before its progeny undergoes apoptosis. For each $k \in \mathbb{N} \cup \{0\}$, let $C_{i,k}$ be the collection of cells in generation i during time $[k, k+1)$. If $t \ge l - i + k$, the number of descendants of each one of the cells in $C_{i,k}$ will be $2^{i-1}(2^{l-i+1} - 2)$ and their progeny will no longer be in the population. For $k < t < l - i + k$, the number of descendants of each cell in $C_{i,k}$ will be between 2^{t-1-k} and 2^{t+1-k}. This will allow us to give upper and lower bounds on the number

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

of cells in or descended from cells in generation i by time t. If we consider a time $t < l - i$ then the descendants of the cells in $C_{i,0}$ will not yet have undergone apoptosis. Therefore, at time $t < l - i$ the number of cells that have been in generation i and their descendants is between

$$\sum_{j=0}^{\lfloor t \rfloor} 2^{t-1-j} \geq 2^t - 1 \quad \text{and} \quad \sum_{j=0}^{\lfloor t \rfloor} 2^{t+1-j} \leq 2^{t+2} - 1.$$

If $t \geq l - i$ then some of the cells that have descended from generation i cells will have undergone apoptosis. The total number of cells that have been in or descended from generation i cells at time t, including those that have undergone apoptosis, will be between

$$\sum_{j=0}^{l-i} 2^{l-i-j-1} + (t - l + i)(2^{l-i+1} - 2) = 2^{l-i} - 1 + (t - l + i)(2^{l-i+1} - 2)$$

and

$$\sum_{j=0}^{l-i} 2^{l-i-j+1} + (t - l + i)(2^{l-i+1} - 2) = 2^{l-i+2} - 1 + (t - l + i)(2^{l-i+1} - 2).$$

Recall that there are always 2^{i-1} cells in generation i which are acquiring type-1 mutations at rate v_1. We can once again multiply the rate of type-1 mutations on generation i by the bounds on the probability that such a mutation is successful to find bounds on the rate of successful type-1 mutations in generation i. We find that successful type-1 mutations occur on generation i according to a Poisson process that has intensity measure between

$$2^{i-1} v_1 (1 - \mathrm{e}^{-v_2(2^t - 1)}) \quad \text{and} \quad 2^{i-1} v_1 (1 - \mathrm{e}^{-v_2(2^{t+2} - 1)})$$

if $t < l - i$, and

$$2^{i-1} v_1 (1 - \mathrm{e}^{-v_2(2^{l-i} - 1 + (t-l+i)(2^{l-i+1} - 2))}) \quad \text{and} \quad 2^{i-1} v_1 (1 - \mathrm{e}^{-v_2(2^{l-i+2} - 1 + (t-l+i)(2^{l-i+1} - 2))})$$

if $t \geq l - i$.

We now use the bounds on the rates of successful type-1 mutations in each generation i to find the limiting distribution of $\tau(M_1)$. For large enough N, we will have $t < \sqrt{v_1 v_2 N} \log N$ for any real number t by the hypothesis $1/(N(\log N)^2) \ll v_1 v_2$. Let $t/\sqrt{v_1 v_2 N} < l$. Then

$$\mathrm{P}\left(\tau(M_1) \leq \frac{t}{\sqrt{v_1 v_2 N}} \right) = 1 - \mathrm{e}^{-f(N,t)},$$

where, by summing over the generations and using the fact that $1 - \mathrm{e}^{-x} \leq x$, we obtain

$$\begin{aligned} f(N,t) \leq & \sum_{0 \leq i < l - t/\sqrt{v_1 v_2 N}} 2^{i-1} v_1 (1 - \mathrm{e}^{-v_2(2^{t/\sqrt{v_1 v_2 N} + 2} - 1)}) \\ & + \sum_{l - t/(v_1 v_2 N) \leq i \leq l} 2^{i-1} v_1 (1 - \mathrm{e}^{-v_2(2^{l-i+2} - 1 + (t/\sqrt{v_1 v_2 N} - l + i)(2^{l-i+1} - 2))}) \\ \leq & \sum_{0 \leq i < l - t/\sqrt{v_1 v_2 N}} 2^{i-1} (2^{t/\sqrt{v_1 v_2 N} + 2} - 1) v_1 v_2 \\ & + \sum_{l - t/(v_1 v_2 N) \leq i \leq l} 2^{i-1} \left(2^{l-i+2} - 1 + \left(\frac{t}{\sqrt{v_1 v_2 N}} - l + i \right) (2^{l-i+1} - 2) \right) v_1 v_2. \end{aligned}$$

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

For the first sum,

$$\begin{aligned}\sum_{0\le i<l-t/\sqrt{v_1v_2N}} 2^{i-1}(2^{t/\sqrt{v_1v_2N}+2}-1)v_1v_2 &\le \frac{1}{2}(2^{t/\sqrt{v_1v_2N}+2}-1)(2^{l-t/\sqrt{v_1v_2N}+1}-1)v_1v_2\\ &\le 2^{l+2}v_1v_2\\ &\to 0.\end{aligned}$$

For the second sum, we first compute

$$\sum_{l-t/\sqrt{v_1v_2N}\le i\le l} 2^{i-1}(2^{l-i+2}-1)v_1v_2 \le 2^{l+2}v_1v_2\frac{t}{\sqrt{v_1v_2N}} \to 0.$$

Lastly,

$$\begin{aligned}&\sum_{l-t/\sqrt{v_1v_2N}\le i\le l} 2^{i-1}\left(\frac{t}{\sqrt{v_1v_2N}}-l+i\right)(2^{l-i+1}-2)v_1v_2\\ &\quad\le 2^l v_1v_2 \sum_{l-t/\sqrt{v_1v_2N}\le i\le l}\left(\frac{t}{\sqrt{v_1v_2N}}-l+i\right)\\ &\quad\le \frac{2^l v_1v_2}{2}\left(\frac{t}{\sqrt{v_1v_2N}}+1\right)^2\\ &\quad\to \frac{t^2}{2}.\end{aligned}$$

Therefore, $\limsup \mathrm{P}(\sqrt{v_1v_2N}\,\tau(M_1)\le t)\le 1-\mathrm{e}^{-t^2/2}$.

For the lower bound, we have

$$\begin{aligned}f(N,t) &\ge \sum_{0\le i<l-t/\sqrt{v_1v_2N}} 2^{i-1}v_1(1-\mathrm{e}^{-v_2(2^{t/\sqrt{v_1v_2N}}-1)})\\ &\quad+\sum_{l-t/\sqrt{v_1v_2N}\le i\le l} 2^{i-1}v_1(1-\mathrm{e}^{-v_2(2^{l-i}-1+(t/\sqrt{v_1v_2N}-l+i)(2^{l-i+1}-2))})\\ &\ge \sum_{l-t/\sqrt{v_1v_2N}\le i\le l} 2^{i-1}v_1(1-\mathrm{e}^{-v_2(t/\sqrt{v_1v_2N}-l+i)(2^{l-i+1}-2)}).\end{aligned}$$

Using the bound $1-\mathrm{e}^{-x}\ge x-x^2/2$, we have

$$\sum_{l-t/\sqrt{v_1v_2N}\le i\le l} 2^{i-1}v_1(1-\mathrm{e}^{-v_2(t/\sqrt{v_1v_2N}-l+i)(2^{l-i+1}-2)})$$

will be greater than or equal to the sum over $i\in[l-t/\sqrt{v_1v_2N},l]$ of

$$2^{i-1}v_1\left(v_2\left(\frac{t}{\sqrt{v_1v_2N}}-l+i\right)(2^{l-i+1}-2)-v_2^2\left(\frac{t}{\sqrt{v_1v_2N}}-l+i\right)^2\frac{(2^{l-i+1}-2)^2}{2}\right).$$

First consider

$$\sum_{l-t/\sqrt{v_1v_2N}\le i\le l} 2^{i-1}v_1v_2^2\left(\frac{t}{\sqrt{v_1v_2N}}-l+i\right)^2\frac{(2^{l-i+1}-2)^2}{2}.$$

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

This sum is bounded between 0 and $\sum_{l-t/\sqrt{v_1v_2N}\le i\le l} v_2t^2 2^{l-i}$. Let $0 < \varepsilon < \alpha$. For large enough N, we have $t < \sqrt{v_1v_2N}l(\alpha-\varepsilon)$, which is equivalent to $l(1-\alpha-\varepsilon) < l - t/\sqrt{v_1v_2N}$. So, for large enough N, we have

$$\sum_{l-t/\sqrt{v_1v_2N}\le i\le l} v_2t^2 2^{l-i} \le \sum_{l(1-\alpha+\varepsilon)\le i\le l} v_2t^2 2^{l-i} \le l v_2 N^{\alpha-\varepsilon} \to 0.$$

It remains to show that

$$\liminf \sum_{l-t/\sqrt{v_1v_2N}\le i\le l} 2^{i-1}v_1v_2\left(\frac{t}{\sqrt{v_1v_2N}} - l + i\right)(2^{l-i+1}-2) \ge \frac{t^2}{2}.$$

Let $j \in \mathbb{N}$ and $t > 0$. For large enough values of N, we will have $j < t/\sqrt{v_1v_2N} < \log N$. Note that if $i \le l - j$ then $2^{l-i+1} - 2 \ge (1-2^{-j})2^{l-i+1}$, so

$$\begin{aligned}&\sum_{l-t/\sqrt{v_1v_2N}\le i\le l} 2^{i-1}v_1v_2\left(\frac{t}{\sqrt{v_1v_2N}} - l + i\right)(2^{l-i+1}-2)\\ &\quad\ge \sum_{l-t/\sqrt{v_1v_2N}\le i\le l-j} 2^{i-1}v_1v_2\left(\frac{t}{\sqrt{v_1v_2N}} - l + i\right)(1-2^{-j})2^{l-i+1}.\end{aligned}$$

Because j is fixed we have

$$\sum_{l-j\le i\le l} 2^{i-1}v_1v_2\left(\frac{t}{\sqrt{v_1v_2N}} - l + i\right)(1-2^{-j})2^{l-i+1} \to 0,$$

since each of the summands converges to 0. Therefore, we can add this sum without changing the limit. This gives us a lower bound of

$$\liminf \sum_{l-t/\sqrt{v_1v_2N}\le i\le l} 2^{l}v_1v_2\left(\frac{t}{\sqrt{v_1v_2N}} - l + i\right)(1-2^{-j}) \ge \frac{t^2}{2}(1-2^{-j}).$$

We chose j to be any natural number, so $\liminf \mathrm{P}(\sqrt{v_1v_2N}\tau(M_1) \le t) \ge 1 - \mathrm{e}^{-t^2/2}$.

The above two bounds establish that $\mathrm{P}(\sqrt{v_1v_2N}\tau(M_1) \le t) \to 1 - \mathrm{e}^{-t^2/2}$ for any $t \ge 0$. It remains to show that $\sigma(M_1)$ converges in probability to 1. First note that, for any $\varepsilon > 0$, we have

$$\mathrm{P}(\tau(M_1) \le \varepsilon \log N) = \mathrm{P}(\sqrt{Nv_1v_2}\tau(M_1) \le \sqrt{Nv_1v_2}\varepsilon \log N) \to 1,$$

which follows because the distribution of $\sqrt{Nv_1v_2}\tau(M_1)$ is converging to the Rayleigh distribution and $\sqrt{Nv_1v_2}\varepsilon \log N$ is converging to ∞. Let $\delta > 0$. By Corollary 1 we know that $\rho(M_1)$ converges in probability to 1, so, as N goes to ∞, $\mathrm{P}(\rho(M_1) > 1-\delta) \to 1$. If $\sigma(M_1) < 1-2\delta$ and $\rho(M_1) > 1-\delta$, then $\tau(M_1) > \delta \log N$. Because $\mathrm{P}(\tau(M_1) > \delta \log N) \to 0$ we must also have $\mathrm{P}(\sigma(M_1) < 1-2\delta) \to 0$, where $\delta > 0$ was arbitrary. Then $\mathrm{P}(1-\sigma(M_1) > 2\delta) \to 0$ for any $\delta > 0$, so $\sigma(M_1) \xrightarrow{\mathrm{P}} 1$.

4. The sd and ss regimes

In this section we need two different models. The first one is the same as model H_2 except that only stem cells receive type-1 mutations and only daughter cells receive type-2 mutations. The second is the same as H_2 except that only stem cells receive mutations. These will be referred to as models M_2 and M_3, respectively.

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

Proposition 3. 1. *If* $u_1 \ll 1/\log N$ *and* $u_1 \ll Nv_2$, *then* $u_1\tau(M_2) \xrightarrow{\mathrm{D}} X$ *and* $\rho(M_2) \xrightarrow{\mathrm{P}} (\alpha \wedge 1)$.

2. *If* $u_1 \ll u_2$ *then* $u_1\tau(M_3) \xrightarrow{\mathrm{D}} X$.

3. *Let* $A > 0$, *and let* Z *be an exponentially distributed random variable with mean* A *that is independent of* X. *If* $u_1 \sim Au_2$ *then* $u_1\tau(M_3) \xrightarrow{\mathrm{D}} X + Z$.

The goal of this section is to prove Proposition 3. It will be shown later that the conditions used in Proposition 3 for the sd regime are the only relevant conditions.

Lemma 7. *For time* $t \le \log N$ *after a stem cell receives a type-1 mutation, we have*

$$\mathrm{e}^{-2^{t+2}v_2} \le \mathrm{P}(\tau(M_2) - \tau'(M_2) > t) \le \mathrm{e}^{-(2^{t-2}-2)v_2}.$$

Proof. Let $Z = \tau(M_2) - \tau'(M_2)$. First we establish the upper bound. After the stem cell line receives the first mutation it takes at most one time unit until the mutation is passed along to the first generation daughter cell. Assuming that it does take one time unit until the first generation daughter cell inherits the mutation, we can obtain an upper bound on $\mathrm{P}(Z > t)$. Let time $t = 0$ denote the time at which the stem cell line receives the type-1 mutation. There are no mutations being acquired by the daughter cells for time $t \in [0, 1)$. For time $t \in [1, 2)$, the generation 1 daughter cell is the only type-1 daughter cell. So, for $t \in [1, 2)$, we have $\mathrm{P}(Z > t) = \mathrm{e}^{-(t-1)v_2}$. For time $t \in [2, 3)$, the first two generations have the mutation which is a total of three cells. Therefore, for $t \in [2, 3)$, we have $\mathrm{P}(Z > t) = \mathrm{e}^{-(3(t-2)v_2+v_2)}$, where the v_2 is added because of the probability of having a mutation before time 2. Extending this inductively gives us

$$\mathrm{P}(Z > t) \le \exp\left[-\left[(2^{\lfloor t\rfloor} - 1)(t - \lfloor t\rfloor) + \sum_{i=2}^{\lfloor t\rfloor}(2^{i-1} - 1)\right]v_2\right] \le \mathrm{e}^{(-2^{t-2}-1)v_2}$$

for any $t \le \log N$.

For the lower bound, we use the same reasoning as above except that we assume that it takes zero time for the generation 1 daughter cell to become type 1 after the stem cell line is type 1. This yields

$$\mathrm{P}(Z > t) \ge \exp\left[-\left[(2^{\lceil t\rceil} - 1)(t - \lfloor t\rfloor) + \sum_{i=1}^{\lfloor t\rfloor}(2^{i} - 1)\right]v_2\right] \ge \mathrm{e}^{-2^{t+2}v_2}.$$

Lemma 8. *The location of the second mutation satisfies* $\rho(M_2) \xrightarrow{\mathrm{P}} \alpha \wedge 1$.

Proof. Let $Z = \tau(M_2) - \tau'(M_2)$. By Lemma 7 we have $\mathrm{P}(Z > \log N) \ge \mathrm{e}^{-4Nv_2}$. If $\alpha > 1$ then $\mathrm{P}(Z > \log N) \to 1$ and the mutation will spread throughout the entire crypt. If this is the case then any cell is equally likely to have the second mutation. Therefore, $\mathrm{P}(\rho(M_2) \le \beta) \le (2^{\beta l} - 1)/(2^l - 1)$ for any $\beta \in [0, 1)$, so $\rho(M_2) \xrightarrow{\mathrm{P}} 1$.

Now suppose that $\alpha \le 1$. Let $\varepsilon > 0$, so that $\alpha - \varepsilon > 0$. Then, by Lemma 7,

$$\mathrm{P}(Z > l(\alpha - \varepsilon)) \ge \mathrm{e}^{-2^{l(\alpha-\varepsilon)+2}v_2}.$$

Because $4N^{\alpha-\varepsilon}v_2 \to 0$ we obtain the convergence $\mathrm{P}(Z > l(\alpha - \varepsilon)) \to 1$. By time $l(\alpha - \varepsilon)$ the mutation will have spread to the first $\lfloor l(\alpha - \varepsilon)\rfloor$ generations, so, for times after $l(\alpha - \varepsilon)$,

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

we know that at least $2^{\lfloor l(\alpha-\varepsilon)\rfloor}$ cells have the type-1 mutation. Therefore,

$$\mathrm{P}(\{\rho(M_2) \le \beta\} \cap \{Z > l(\alpha-\varepsilon)\}) \le \frac{2^{\beta l}-1}{2^{(\alpha-\varepsilon)l-1}-1}.$$

Thus, for any $\beta < \alpha - \varepsilon$,

$$\mathrm{P}(\rho(M_2) \le \beta) < \frac{2^{\beta l}-1}{2^{(\alpha-\varepsilon)l-1}-1} + \mathrm{P}(X_2 \le l(\alpha-\varepsilon)) \to 0.$$

Hence, $\mathrm{P}(\rho(M_2) \ge \alpha - \varepsilon) \to 1$. Because ε may be arbitrarily small we have completed the $\alpha = 1$ case.

Suppose that $\alpha < 1$, and let $\varepsilon > 0$, so that $\alpha + \varepsilon \le 1$. Then, by Lemma 7,

$$\mathrm{P}(Z > l(\alpha+\varepsilon)) \le \mathrm{e}^{-(2^{l(\alpha+\varepsilon)-2}-1)v_2}.$$

Because $N^{\alpha+\varepsilon}v_2/4 \to \infty$ we have $\mathrm{P}(Z > l(\alpha+\varepsilon)) \to 0$. By time $l(\alpha+\varepsilon)$ the mutation has only spread to the first $l(\alpha+\varepsilon)$ generations, so $\mathrm{P}(\rho(M_2) > \alpha+\varepsilon) \to 0$, where ε is arbitrarily small.

Lemma 9. *If $u_1 \ll 1/\log N$ and $u_1 \ll N v_2$, then $u_1\tau(M_2) \xrightarrow{\mathrm{D}} X$.*

Proof. Since the stem cell line acquires mutations according to a Poisson process at rate u_1, $u_1\tau'(M_2)$ is an exponentially distributed random variable with mean 1. It remains to show that $u_1(\tau(M_2) - \tau'(M_2)) \xrightarrow{\mathrm{P}} 0$.

Suppose that we consider a new model M_2' which is the same as model M_2 except that the type-2 mutations can only occur on daughter cells a $\log N$ time after the stem cell line has a type-1 mutation. We can couple models M_2 and M_2' so that the same Poisson processes are marking the mutations on the cells in each model but that any proposed type-2 mutation is rejected in model M_2' until a $\log N$ time after the stem cell line is type 1. Under the coupling, $\tau'(M_2) = \tau'(M_2')$. Also, if we let $Z = \tau(M_2') - \tau'(M_2')$ then $Z \ge \tau(M_2) - \tau'(M_2)$. Therefore, it is enough to show that $u_1 Z \xrightarrow{\mathrm{P}} 0$.

If we wait a $\log N$ time after the stem cell line receives a type-1 mutation then all of the daughter cells will be type 1. Thus, for any fixed N, we have

$$\mathrm{P}(Z > t) = \mathbf{1}_{[0,\log N]}(t) + \mathrm{e}^{-v_2(N-1)(t-\log N)}\,\mathbf{1}_{(\log N,\infty]}(t).$$

Let $\varepsilon > 0$. Then

$$\mathrm{P}(u_1 Z > \varepsilon) = \mathbf{1}_{[0,\log N]}\left(\frac{\varepsilon}{u_1}\right) + \mathrm{e}^{-v_2(N-1)(\varepsilon/u_1-\log N)}\,\mathbf{1}_{(\log N,\infty]}\left(\frac{\varepsilon}{u_1}\right).$$

By our assumptions, $u_1 \log N \to 0$, so, for large enough N, this becomes

$$\mathrm{P}(u_1 Z > \varepsilon) = \mathrm{e}^{-v_2(N-1)(\varepsilon/u_1-\log N)}.$$

Also, by our assumptions, $-v_2(N-1)(\varepsilon/u_1 - \log N) \sim -v_2 N\varepsilon/u_1 \to -\infty$, so

$$\mathrm{P}(u_1 Z > \varepsilon) \to 0.$$

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

Proof of Proposition 3. Combining Lemmas 8 and 9 we obtain part 1 of Proposition 3.

Note that $u_1\tau(M_3)$ has the exponential distribution with mean 1. To prove part 2 of Proposition 3, we need to show that $u_1(\tau(M_3) - \tau'(M_3)) \xrightarrow{\mathrm{P}} 0$. Let $\varepsilon > 0$. Then

$$\mathrm{P}(u_1(\tau(M_3) - \tau'(M_3)) > \varepsilon) = \mathrm{P}\left((\tau(M_3) - \tau'(M_3)) > \frac{\varepsilon}{u_1}\right) = \mathrm{e}^{-\varepsilon u_2/u_1}.$$

Since $u_2/u_1 \to \infty$, we have $\mathrm{P}(u_1(\tau(M_3) - \tau'(M_3)) > \varepsilon) \to 0$.

Lastly, we prove part 3 of Proposition 3. In model M_3 both mutations occur on the stem cell line. In this case $u_1\tau'(M_3)$ and $u_2(\tau(M_3) - \tau'(M_3))$ are both exponentially distributed with mean 1. Because $u_1(\tau(M_3) - \tau'(M_3)) = (u_1/u_2)u_2(\tau(M_3) - \tau'(M_3))$, $u_1(\tau(M_3) - \tau'(M_3))$ is exponentially distributed with mean u_1/u_2. By assumption, $u_2/u_1 \to 1/A$, so $u_1(\tau(M_3) - \tau'(M_3))$ converges in distribution to Z. The random variables $\tau'(M_3)$ and $\tau(M_3) - \tau'(M_3)$ are independent for each N, so

$$u_1\tau(M_3) = u_1\tau'(M_3) + u_1(\tau(M_3) - \tau'(M_3)) \xrightarrow{\mathrm{D}} X + Z.$$

5. Proof of Theorem 1

Proof of part 3 of Theorem 1. We make use of the following well-known fact. If $\{a_n\}_{n=1}^{\infty}$ is a sequence of real numbers such that $a_n \to a$ then

$$\lim_{n\to\infty}\left(1 - \frac{a_n}{n}\right)^{n-1} = \mathrm{e}^{-a}.$$

Before time 1 the cells never split and there is no apoptosis. Let H_1' be the same as model H_1 except that stem cells never receive mutations. Note that H_1' differs from M_1 because daughter cells cannot accumulate type-1 mutations in model H_1'. If we ignore the splitting and apoptosis and consider how long it takes for a cell to acquire two mutations under the mutation mechanism alone, then we have $N-1$ daughter cells acquiring mutations independently. For any individual cell, the time it takes to acquire two mutations will have the same distribution as the sum of two independent exponentially distributed random variables with means $1/v_1$ and $1/v_2$. If we denote the time until cell i has a type-2 mutation by T_i and assume that $v_1 \neq v_2$, then

$$\mathrm{P}(T_i \le t) = 1 - \frac{v_2\mathrm{e}^{-v_1 t} - v_1\mathrm{e}^{-v_2 t}}{v_2 - v_1}.$$

There are $N-1$ cells independently acquiring mutations, so, for $t \le 1$, we have

$$\mathrm{P}(\tau(H_1') \le t) = 1 - \left(\frac{v_2\mathrm{e}^{-v_1 t} - v_1\mathrm{e}^{-v_2 t}}{v_2 - v_1}\right)^{N-1},$$

or, equivalently,

$$\mathrm{P}(\sqrt{v_1 v_2 N}\,\tau(H_1') \le t) = 1 - \left(\frac{v_2\mathrm{e}^{-\sqrt{v_1/(v_2 N)}t} - v_1\mathrm{e}^{-\sqrt{v_2/(v_1 N)}t}}{v_2 - v_1}\right)^{N-1}.$$

Note that $N\sqrt{v_1^3/v_2 N^3} = v_1^2/\sqrt{v_1 v_2 N} \to 0$ and $N\sqrt{v_2^3/v_1 N^3} = v_2^2/\sqrt{v_1 v_2 N} \to 0$. For large enough N, we can apply the third-degree Taylor expansion of the exponential function to

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

obtain the bounds

$$1-\frac{t^2}{2N}-\sqrt{\frac{v_1^3}{v_2N^3}}\frac{t^3}{6}\le\frac{v_2e^{-\sqrt{v_1/(v_2N)}t}-v_1e^{-\sqrt{v_2/(v_1N)}t}}{v_2-v_1}\le1-\frac{t^2}{2N}+\sqrt{\frac{v_2^3}{v_1N^3}}\frac{t^3}{6}.$$

For any fixed t, we have

$$\left(1-\frac{t^2}{2N}-\sqrt{\frac{v_1^3}{v_2N^3}}\frac{t^3}{6}\right)^{N-1}\to e^{-t^2/2}$$

and

$$\left(1-\frac{t^2}{2N}+\sqrt{\frac{v_2^3}{v_1N^3}}\frac{t^3}{6}\right)^{N-1}\to e^{-t^2/2}.$$

If $v_1=v_2$ and we ignore splitting and apoptosis, then the probability that one cell has two mutations by time t is $1-e^{-v_1t}-v_1te^{-v_1t}$. The probability that one of the N cells has two mutations by time t is $1-(e^{-v_1t}-v_1te^{-v_1t})^N$. By applying the same techniques as above we obtain $P(\sqrt{v_1v_2N}\tau(H_1')\le t)\to1-e^{-t^2/2}$ when $v_1=v_2$.

Combining the two results above we have $P(\sqrt{v_1v_2N}\tau(H_1')\le t)\to1-e^{-t^2/2}$ when ignoring splitting and apoptosis. Then $P(\tau(H_1')<1)=P(\sqrt{v_1v_2N}\tau(H_1')<\sqrt{v_1v_2N})\to1$. Therefore, the probability that two mutations occur before time 1 is converging to 1, so we may ignore splitting and apoptosis in this case. This gives the desired result for $\tau(H_1')$.

Stem cells acquire type-1 mutations at rate $u_1\to0$ in model H_1. Let T be the first time the stem cell line acquires a mutation in model H_1. Then $P(T<1)\to0$. We can couple models H_1 and H_1' so that the same Poisson processes are marking the mutations on the daughter cells. Then $P(\tau(H_1)=\tau(H_1'))\ge P(\{T\ge1\}\cap\{\tau(H_1')<1\})\to1$, which gives the results for model H_1.

Because any cell is equally likely to acquire the two mutations, it is clear that $\sigma(H_1)$ and $\rho(H_1)$ both converge in probability to 1.

This gives the result for part 3 of Theorem 1 even if $\alpha=0$.

The following lemma, whose proof is elementary, will be used several times in this section.

Lemma 10. *Let $\{\alpha_n\}_{n=1}^\infty$ and $\{\beta_n\}_{n=1}^\infty$ be sequences of positive numbers which converge to 0. Let $\{X_n\}_{n=1}^\infty$ and $\{Y_n\}_{n=1}^\infty$ be independent sequences of random variables, and let X and Y be positive random variables such that $\alpha_nX_n\xrightarrow{D}X$ and $\beta_nY_n\xrightarrow{D}Y$ as $n\to\infty$. If $\alpha_n\ll\beta_n$ then $P(X_n\ge Y_n)\to1$ as $n\to\infty$.*

We will couple the models H_2, M_1, M_2, and M_3 so that the Poisson processes used in models M_1, M_2, and M_3 are the appropriate subcollections of Poisson processes which are used in model H_2. Let T be the time that a type-1 mutation occurs on the stem cell line in model H_2. Note that because stem cells cannot inherit type-1 mutations the coupling implies that $T=\tau'(M_2)=\tau'(M_3)$.

Lemma 11. *Suppose that $v_1v_2\ll1/(N(\log N)^2)$. If $u_1\ll v_1v_2N\log N$ then $P(\tau(M_1)<T)\to1$. If $u_1\gg v_1v_2N\log N$ then $P(\tau(M_3)<\tau(M_1))\to1$.*

Proof. By part 1 of Proposition 2, $(\alpha\wedge1)v_1v_2N(\log N)\tau(M_1)\xrightarrow{D}X$. Mutations to the stem cell line occur at rate u_1, so $u_1T\xrightarrow{D}X$. Because the Poisson processes that mark the mutations in model M_1 are independent of the Poisson process that marks the mutations on the stem cell line, if $u_1\ll v_1v_2N\log N$ then $P(\tau(M_1)<T)\to1$ by Lemma 10.

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

On the other hand, suppose that $u_1 \gg v_1 v_2 N \log N$. We assume that $u_1 \leq u_2$, so we could decrease $\mathrm{P}(\tau(M_3) < \tau(M_1))$ by decreasing u_2 to u_1. Then the distribution of $u_1\tau(M_3)$ is the distribution of the sum of two independent exponentially distributed random variables. By Lemma 10, $\mathrm{P}(\tau(M_3) < \tau(M_1)) \to 1$.

Lemma 12. *Suppose that* $1/(N(\log N)^2) \ll v_1 v_2 \ll 1/N$. *If* $u_1 \ll \sqrt{v_1 v_2 N}$ *then* $\mathrm{P}(\tau(M_1) < T) \to 1$. *If* $u_1 \gg \sqrt{v_1 v_2 N}$ *then* $\mathrm{P}(\tau(M_3) < \tau(M_1)) \to 1$.

Proof. Let $u_1 \ll \sqrt{v_1 v_2 N}$. By part 2 of Proposition 2 we have $\sqrt{v_1 v_2 N}\tau(M_1) \xrightarrow{\mathrm{D}} Y$. The stem cell line acquires mutations at rate u_1, so $u_1 T \to X$. The Poisson processes that are marking the mutations in model M_1 are independent of the Poisson process that marks mutations on the stem cell line, so the result follows by Lemma 10.

If $u_1 \gg v_1 v_2 N \log N$ then the proof follows by the same reasoning as used in the proof of Lemma 11 when $u_1 \gg v_1 v_2 N \log N$.

Lemma 13. *If* $u_2 \ll 1/\log N$ *and* $u_2 \ll N v_2$, *then* $\mathrm{P}(\tau(M_2) < \tau(M_3)) \to 1$.

Proof. By the coupling, $\tau'(M_2) = \tau'(M_3)$. After time $\tau'(M_2)$ the Poisson processes marking the mutations in models M_2 and M_3 are independent. Let $T_2 = \tau(M_2) - \tau'(M_2)$ and $T_3 = \tau(M_3) - \tau'(M_3)$. Then $\mathrm{P}(\tau(M_2) < \tau(M_3)) = \mathrm{P}(T_2 < T_3)$.

Consider again the model M_2' that was introduced in the proof of Lemma 9 which is the same as model M_2 except that the type-2 mutations can only occur on daughter cells $\log N$ time units after the stem cell line has a type-1 mutation. We can couple models M_2 and M_2' as we did before so that the time at which the stem cell line acquires a mutation is the same in models M_2 and M_2'. In particular, $\tau'(M_2') = \tau'(M_2) = \tau'(M_3)$. Let $T_2' = \tau(M_2') - \tau'(M_2')$. Then $T_2' \geq T_2$, so it is enough to show that $\mathrm{P}(T_2' < T_3) \to 1$.

If we wait a $\log N$ time after the stem cell line receives a type-1 mutation then all of the daughter cells will be type 1 and the $N-1$ daughter cells acquire type-2 mutations at rate v_2. Thus, for any fixed N, we have

$$\mathrm{P}(T_2' > t) = \mathbf{1}_{[0,\log N]}(t) + \mathrm{e}^{-v_2(N-1)(t-\log N)}\,\mathbf{1}_{(\log N,\infty]}(t).$$

Let $\varepsilon > 0$. Then

$$\begin{aligned}\mathrm{P}(T_2' < T_3) = {} & \mathrm{P}(T_2' < T_3 \mid T_3 < \log N)\,\mathrm{P}(T_3 < \log N) \\ & + \mathrm{P}(T_2' < T_3 \mid T_3 \geq \log N)\,\mathrm{P}(T_3 \geq \log N).\end{aligned}$$

Because $u_2 \ll 1/\log N$ and $u_2 T_3$ has the exponential distribution with mean 1, we have $\mathrm{P}(T_3 \geq \log N) \to 1$. The memoryless property of the exponential distribution gives

$$\mathrm{P}(T_2' < T_3 \mid T_3 \geq \log N) = \frac{v_2(N-1)}{v_2(N-1) + u_2} \to 1,$$

which completes the proof.

Lemma 14. *If* $u_2 \gg 1/\log N$ *or* $u_2 \gg N v_2$, *then* $\mathrm{P}(\tau(M_3) < \tau(M_2)) \to 1$.

Proof. By the coupling, $\tau'(M_2) = \tau'(M_3)$. After time $\tau'(M_2)$ the Poisson processes marking the mutations in models M_2 and M_3 are independent. Let $T_2 = \tau(M_2) - \tau'(M_2)$ and $T_3 = \tau(M_3) - \tau'(M_3)$. Then $\mathrm{P}(\tau(M_3) < \tau(M_2)) = \mathrm{P}(T_3 < T_2)$.

Suppose that $u_2 \gg 1/\log N$. By Lemma 8 we know that $\rho(M_2) \xrightarrow{\mathrm{P}} \alpha \wedge 1$. If $0 < \delta < (\alpha \wedge 1)$ then $\mathrm{P}(\rho(M_2) > (\alpha \wedge 1) - \delta) \to 1$. If $\rho(M_2) > (\alpha \wedge 1) - \delta$ then the second mutation occurs

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

on a generation higher than $((\alpha \wedge 1) - \delta)l$. Since only stem cells acquire type-1 mutations in model M_2, we have $T_2 \geq \lfloor((\alpha \wedge 1) - \delta)l\rfloor$ because it takes at least that much time for the type-1 mutation to spread to the generation $\lceil((\alpha \wedge 1) - \delta)l\rceil$ daughter cells. On the other hand, in model M_3 the second mutation occurs at rate u_2, so u_2T_3 is exponentially distributed with mean 1. Then $\mathrm{P}(T_3 < K \log N) = \mathrm{P}(u_2T_3 < u_2K \log N) \to 1$ for any positive number K since $u_2 \log N \to \infty$. Therefore, $\mathrm{P}(T_3 < T_2) \to 1$.

Suppose that $u_2 \gg Nv_2$. The rate at which type-2 mutations occur in model M_2 is always bounded by $(N-1)v_2$. Suppose that we consider a new model M_2'' which is the same as M_2 except that once the stem cell line has a type-1 mutation, all of the daughter cells also have a type-1 mutation instantaneously. Models M_2 and M_2'' can be coupled so that after the stem cell line acquires a type-1 mutation then any type-2 mutation proposed by a Poisson process on a daughter cell is accepted in model M_2''. Let $T_2'' = \tau(M_2'') - \tau'(M_2'')$. Then $(N-1)v_2T_2''$ has the exponential distribution with mean 1. By Lemma 10, $\mathrm{P}(T_3 < T_2'') \to 1$. Because $T_2 \geq T_2''$, we have the desired result.

Proof of Theorem 1. From the coupling we have $\tau(H_2) = \tau(M_1) \wedge \tau(M_2) \wedge \tau(M_3)$ because any type-2 mutation which occurs in model H_2 must occur in at least one of the models M_i for some i, and if a mutation occurs in model M_i then it will also occur in model H_2.

Suppose that $\mathrm{P}(\tau(M_1) < T) \to 1$. Before time T only stem cells acquire type-1 mutations in models M_2 and M_3. Therefore, models M_2 and M_3 only have type-0 cells before time T and $\mathrm{P}(\tau(M_1) < \tau(M_2) \wedge \tau(M_3)) \to 1$.

- By Lemma 11, if $v_1v_2 \ll 1/(N(\log N)^2)$ and $u_1 \ll v_1v_2N \log N$, then $\mathrm{P}(\tau(M_1) < T) \to 1$, so, by part 1 of Proposition 2 and the coupling of H_2 with M_1, we have $(\alpha \wedge 1) \times v_1v_2N(\log N)\tau(H_2) \xrightarrow{\mathrm{D}} X$. Also, by Lemma 11, the distribution of $\sigma(H_2)$ converges to a uniform distribution on $((1-\alpha)^+, 1]$ and $\rho(H_2)$ converges in distribution to 1.
- By Lemma 12, if $1/(N(\log N)^2) \ll v_1v_2 \ll 1/N$ and $u_1 \ll \sqrt{v_1v_2N}$, then $\mathrm{P}(\tau(M_1) < T) \to 1$, so, by part 2 of Proposition 2 and the coupling of H_2 with M_2, we have $\sqrt{v_1v_2N}\tau(H_2) \xrightarrow{\mathrm{D}} Y$. Also, by Lemma 12, both $\sigma(H_2)$ and $\rho(H_2)$ converge in distribution to 1.

If either $v_1v_2 \ll 1/(N(\log N)^2)$ and $u_1 \gg v_1v_2N \log N$ or $1/(N(\log N)^2) \ll v_1v_2 \ll 1/N$ and $u_1 \gg \sqrt{v_1v_2N}$, then $\mathrm{P}(\tau(M_3) < \tau(M_1)) \to 1$ by Lemmas 11 and 12, respectively. Therefore, $\mathrm{P}(\tau(M_2) \wedge \tau(M_3) < \tau(M_1)) \to 1$, which implies that the cancer-causing type-1 mutation occurs on the stem cell line in model H_2 with probability converging to 1. Given these four conditions, it only remains to compare $\tau(M_2)$ and $\tau(M_3)$.

- By Lemma 13, if $u_2 \ll 1/\log N$ and $u_2 \ll Nv_2$, then $\mathrm{P}(\tau(M_2) < \tau(M_3)) \to 1$. Because $u_1 \leq u_2$, the hypotheses are true for u_1 as well. Therefore, by the coupling of H_2 with M_2 and part 1 of Proposition 3, $u_1\tau(H_2) \xrightarrow{\mathrm{D}} X$ and $\rho(H_2)$ converges in probability to $\alpha \wedge 1$.
- By Lemma 14, if $u_2 \gg 1/\log N$ or $u_2 \gg Nv_2$, then $\mathrm{P}(\tau(M_3) < \tau(M_2)) \to 1$. If $u_1 \ll u_2$ then, by the coupling of H_2 with M_3 and part 2 of Proposition 3, we have $u_1\tau(H_2) \xrightarrow{\mathrm{D}} X$. If $u_1 \sim Au_2$ then, by the coupling of H_2 with M_3 and part 3 of Proposition 3, we have $u_1\tau(H_2) \xrightarrow{\mathrm{D}} X + Z$, where Z is an exponentially distributed random variable with mean A that is independent of X.

By Lemma 1, the results hold for model H_1 as well.

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

6. The null model

In this section, we always have $u_1 = u_2 = v_1 = v_2 = \mu$ and we prove Proposition 1 for model H_2. Then Proposition 1 will hold for model H_1 as well by Lemma 1. We begin this section by pointing out that the conditions of part 5 of Theorem 1 always fail in the null model. The two conditions in the first conjunction become $\mu \ll 1/(N \log N)$. Of the two conditions in the second conjunction, one becomes $\sqrt{N} \ll 1$, which always fails. This reduces all of the conditions in the first bullet point to $\mu \ll 1/(N \log N)$. The conditions in the second bullet point become $\mu \gg 1/\log N$ or $1 \gg N$, so the conditions in part 5 are reduced to $\sqrt{N} \ll 1$, $1 \gg N$, or $1/\log N \ll \mu \ll 1/(N \log N)$, which all fail.

This shows that the probability that the first type-2 mutation occurs on the stem cell line converges to 0. For this reason, we will never consider model M_3 in this section.

Proof of part 2 of Proposition 1. We can couple model H_2 with models M_1 and M_2 such that the Poisson processes marking model M_1 are independent of the Poisson processes marking model M_2. Before time $\tau'(M_2)$, the Poisson processes marking model M_1 are also marking the daughter cells in model H_2 and the Poisson process that marks the stem cell in model M_2 is also marking the stem cell line in model H_2. After time $\tau'(M_2)$, the Poisson processes marking the cells in model M_1 are only marking the daughter cells in model H_2 that have not yet inherited the type-1 mutation from the stem cell. All of the Poisson processes marking type-2 mutations on cells in model M_2, meaning that those cells have inherited the type-1 mutation from the stem cell, also mark the corresponding cells in model H_2. After time $\tau'(M_2) + \log N$, only the Poisson processes marking model M_2 are marking model H_2.

Let T be the time at which the first successful type-1 mutation occurs in model M_1, and let Z be the time at which the first successful type-1 mutation occurs in model H_2. By Corollary 2 we have $A\mu T \xrightarrow{\mathrm{D}} X$. Because the stem cell acquires type-1 mutations at rate μ and every type-1 mutation on the stem cell is successful, we have $(A+1)\mu Z \xrightarrow{\mathrm{D}} X$. If the first successful type-1 mutation occurs on a daughter cell then the type-2 mutation must occur within $\log N$ time of Z since after this time the progeny of the cell will no longer be in the population. Let Y_2 be the time it takes to obtain the second successful type-1 mutation after the first has occurred. If the first successful type-1 mutation occurs on the stem cell then all of the cells will be type 1 within $\log N$ time. Therefore, if the first successful type-1 mutation occurs on the stem cell and there is not another successful type-1 mutation within $\log N$ time, $Y_2 = \infty$ since there can be no more type-1 mutations. We have $\limsup \mathrm{P}((1+A)\mu Y_2 \le t) \le 1 - \mathrm{e}^{-t}$. Therefore,

$$\begin{aligned}
\limsup \mathrm{P}(Y_2 < (\tau(H_2) - Z)) &\le \limsup \mathrm{P}(Y_2 < \log N) \\
&= \limsup \mathrm{P}((1+A)\mu Y_2 < (A+1)\mu \log N) \\
&\le 1 - \mathrm{e}^{-(1+A)\mu \log N} \\
&\to 0.
\end{aligned}$$

Similarly to the result given in Lemma 4, we have $\mathrm{P}(Z = \tau'(H_2)) \to 1$. Hence, it is enough to find the distribution of the time of the first successful type-1 mutation.

We have established that $(A+1)\mu Z \xrightarrow{\mathrm{D}} X$ and $\mathrm{P}(Z = \tau'(H_2)) \to 1$, which imply that $(1+A)\mu\tau'(H_2) \xrightarrow{\mathrm{D}} X$. Let A_1 be the event that the first successful type-1 mutation occurs on a daughter cell, and let A_2 be the event that the first successful type-1 mutation occurs on the stem cell. If the first successful type-1 mutation occurs on a daughter cell then, due to apoptosis, $\tau(H_2) - Z$ is bounded above by $\log N$. Therefore,

$$\mathrm{P}(\{A\mu(\tau(H_2) - Z) > \varepsilon\} \cap A_1) \to 0.$$

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

If the first successful type-1 mutation occurs on a stem cell then in $\log N$ time all of the cells will be type 1, and type-2 mutations will occur at rate μN. Let $\hat{Z}$ be an exponentially distributed random variable with mean $1/\mu N$. Then we have

$$\mathrm{P}(\{A\mu(\tau(H_2) - Z) > \varepsilon\} \cap A_2) \le \mathrm{P}(A\mu(\log N + \hat{Z}) > \varepsilon) \to 0.$$

Since either A_1 or A_2 must occur, we have $A\mu(\tau(H_2) - Z) \xrightarrow{\mathrm{P}} 0$. Then

$$(1 + A)\mu\tau(H_2) = A\mu(Z + (\tau(M_1) - Z)) \xrightarrow{\mathrm{D}} X.$$

By the coupling, before time $\tau'(M_2)$ the daughter cells in model H_2 acquire successful type-1 mutations at the same rate as the daughter cells in model M_1. We know from the proof of Lemma 6 that each generation i with $1 \le i \le l$ acquires successful type-1 mutations independently at a rate bounded between $\mu 2^{i-1}(1 - \mathrm{e}^{-\mu(2^{l-i+1}-2)})$ and $\mu 2^{i-1}(1 - \mathrm{e}^{-\mu(2^{l-i+1}-1)})$ for any time t in model M_1. Therefore, these bounds also hold for the rate at which daughter cells acquire successful type-1 mutations in model H_2 before time $\tau'(M_2)$. Let $\beta \in [0, 1]$. Using the notation and result from Lemma 6 and the fact that the stem cell line acquires type-1 mutations at rate μ,

$$\begin{aligned}\limsup \mathrm{P}(\sigma(H_2) \le \beta) &\le \limsup \frac{\mu + \sum_{i\in(0,l\beta]^*} \mu 2^{i-1}(1 - \mathrm{e}^{-\mu(2^{l-i+1}-1)})}{\mu + \sum_{i\in(0,l]^*} \mu 2^{j-1}(1 - \mathrm{e}^{-\mu(2^{l-j+1}-2)})} \\ &= \frac{1}{1+A} + \frac{A}{1+A}\beta\end{aligned}$$

and

$$\begin{aligned}\liminf \mathrm{P}(\sigma(H_2) \le \beta) &\ge \liminf \frac{\mu + \sum_{i\in(0,l\beta]^*} \mu 2^{i-1}(1 - \mathrm{e}^{-\mu(2^{l-i+1}-2)})}{\mu + \sum_{i\in(0,l]^*} \mu 2^{j-1}(1 - \mathrm{e}^{-\mu(2^{l-j+1}-1)})} \\ &= \frac{1}{1+A} + \frac{A}{1+A}\beta.\end{aligned}$$

Lemma 1 gives the result for $\sigma(H_1)$.

Because $\rho(M_1)$ and $\rho(M_2)$ both converge in probability to 1, we will have $\rho(H_2) \xrightarrow{\mathrm{P}} 1$ as well. Lemma 1 then implies that $\rho(H_1) \xrightarrow{\mathrm{P}} 1$.

Let $\mathcal{N}$ be the set of Radon measures ν on a Polish space $(\Psi, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-field such that $\nu(\{x\}) \in \mathbb{N} \cup \{0, \infty\}$ for all $x \in \Psi$. For the next proof, we will consider a point process to be a random variable taking on elements of $\mathcal{N}$. We consider $\nu(\{x\})$ to be the number of times the point x has been marked. For a Poisson point process whose intensity measure has no atoms, $\nu(\{x\})$ is 0 or 1 for all x and $\{x \in \Psi : \nu(\{x\}) > 0\}$ is discrete with probability 1.

Let $\Psi = [0, \infty) \times [0, 1]$. The Poisson point process of successful type-1 mutations in model M_1 induces a point process on Ψ, where if a successful type-1 mutation occurs at time t on a cell in generation i in model M_1 then there is a point of Ψ at $(t/l, i/l)$. We will call this point process P_M.

Lemma 15. *If $\mu \sim A/(\sqrt{N}\log N)$ then the limiting distribution of P_M is a Poisson point process P_∞ which has intensity measure $\nu' = A^2(\lambda \times \lambda_{[1/2,1]})$, where λ is the Lebesgue measure and $\lambda_{[1/2,1]}$ is the measure defined by $\lambda_{[1/2,1]}(B) = \lambda(B \cap [\frac{1}{2}, 1])$ for any Lebesgue measurable set B.*

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

Proof. We let $C_C(\Psi, [-1, 0])$ be the set of continuous functions $h\colon \Psi \to [-1, 0]$ such that the set $\{\psi \in \Psi\colon h(\psi) \neq 0\}$ is precompact. Recall that a point process X has an associated generating functional $\mathfrak{F}\colon C_C(\Psi, [-1, 0]) \to \mathbb{R}$ defined by

$$\mathfrak{F}(h) = \mathrm{E}\bigg[\prod_{\psi\in\Psi} (h(\psi) + 1)^{\nu(\psi)}\bigg],$$

where ν is a Radon measure on Ψ as described above. Probability generating functionals uniquely determine the distribution of point processes (see Theorem 14 of [7, Section 29.5]). Moreover, a sequence of point processes converges in distribution to a point process if and only if the corresponding sequence of generating functionals converges pointwise to a functional $\mathfrak{F}$ that satisfies the following. If h_m is in the domain of $\mathfrak{F}$ for each m, $\bigcup_{m=1}^{\infty}\{\psi\colon h_m(\psi) \neq 0\}$ is relatively compact, and $h_m(\psi) \to 0$ as $m \to \infty$ for each ψ, then $\mathfrak{F}(h_m) \to 1$ as $m \to \infty$. In this case $\mathfrak{F}$ is the probability generating functional of the limiting point process (see Theorem 20 of [7, Section 29.7]).

Note that, for any N, the points marked in Ψ will all have coordinates (x, y), where y takes values in $\{1/\log N, 2/\log N, \ldots, 1\}$. We know from the proof of Lemma 6 that the rate at which mutations occur along generation i is bounded between $2^{i-1}\mu(1 - \mathrm{e}^{-\mu(2^{l-i+1}-2)})$ and $2^{i-1}\mu(1 - \mathrm{e}^{-\mu(2^{l-i+1}-1)})$. Therefore, if we look at the points that are marked in Ψ whose second coordinate is fixed at $i/\log N$, the rate at which the marking will occur will be between $(\log N)2^{i-1}\mu(1 - \mathrm{e}^{-\mu(2^{l-i+1}-2))}$ and $(\log N)2^{i-1}\mu(1 - \mathrm{e}^{-\mu(2^{l-i+1}-1)})$, where the $\log N$ appears because time is scaled by $1/\log N$. This observation will allow us to work with time homogeneous Poisson point processes.

Let $\mathfrak{F}$ denote the generating functional associated with P_M. Let $\mathfrak{F}_1$ be the generating functional associated with the Poisson process on Ψ which marks points at rate $(\log N)2^{i-1} \times \mu(1 - \mathrm{e}^{-\mu(2^{l-i+1}-2)})$ on $y = i/l$, and let $\mathfrak{F}_2$ be the generating functional associated with the Poisson process on Ψ which marks points at rate $(\log N)2^{i-1}\mu(1 - \mathrm{e}^{-\mu(2^{l-i+1}-1)})$ on $y = i/l$. Call the time homogeneous Poisson point processes P_1 and P_2, respectively. Because the intensity measure of P_M is always between the intensity measures of P_1 and P_2 we have the bounds $\mathfrak{F}_1 \leq \mathfrak{F} \leq \mathfrak{F}_2$.

Let X be a Poisson process with intensity measure ν. It is known that the probability generating functional associated with X is

$$\mathfrak{P}(h) = \exp\bigg[-\int_{\Psi} h \,\mathrm{d}\nu\bigg].$$

To show that a sequence of Poisson processes $\{X_n\}_{n=0}^{\infty}$ with intensity measures $\{\nu_n\}_{n=0}^{\infty}$ converges in distribution to a Poisson process X with intensity measure ν, it is enough to show that $\{\nu_n\}_{n=0}^{\infty}$ converges weakly to ν. That is, for each $h \in C_C(\Psi, [-1, 0])$, we need $\int_{\Psi} h \,\mathrm{d}\nu_n \to \int_{\Psi} h \,\mathrm{d}\nu$ as $n \to \infty$. Let ν_N^1 be the intensity measure of P_1 when there are N cells in the population, and let ν_N^2 be the intensity measure of P_2 when there are N cells in the population. The goal is to show that ν_N^1 and ν_N^2 both converge weakly to ν'. Then the limiting distribution of P_M will be P_∞.

Let $R = (a, b] \times (c, d] \subset \Psi$. Then

$$\nu_N^1(R) = (b-a)(\log N) \sum_{i\in(lc,ld]} 2^i \mu(1 - \mathrm{e}^{-\mu(2^{l-i+1}-2)}) \to A^2\bigg(d - c \vee \frac{1}{2}\bigg)^+ (b-a) = \nu'(R)$$

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

by Lemma 3 and the assumption that $\mu \sim A/(\sqrt{N}\log N)$ which implies that $\mu^2 N \log N \sim A^2/\log N$. Now let O be any open subset of Ψ. We can write $O = \bigcup_{n=1}^{\infty} R_n$, where each R_n is a half open rectangle in the same form as R above and the sets $\{R_n\}_{n=1}^{\infty}$ are pairwise disjoint. Then

$$\liminf_{N\to\infty} \nu_N^1(O) = \liminf_{N\to\infty} \sum_{j=1}^{\infty} \nu_N^1(R_j) \geq \sum_{j=1}^{\infty} \nu'(R_j) = \nu'(O),$$

where the inequality follows by Fatou's lemma. By the same reasoning, $\liminf \nu_N^2(O) \geq \nu'(O)$ for any open subset O of Ψ. It follows by the Portmanteau theorem that both ν_N^1 and ν_N^2 converge weakly to ν' as N goes to ∞. Hence, the limiting distribution of P_M is P_∞.

The notation used in Lemma 15 will also be used in the following proof.

Proof of part 4 of Proposition 1. Note that this is the boundary between two cases that are determined by model M_1. By Corollary 1 we know that $\rho(M_1) \xrightarrow{\text{P}} 1$ for all the conditions that we consider. Therefore, $\rho(H_1) \xrightarrow{\text{P}} 1$ in this case.

The strategy is to define functions g and h on the set of Radon measures that are continuous everywhere except a set of measure 0. Then we will apply the continuous mapping theorem to obtain the desired convergence in distribution. Let D be the subset of $\mathcal{N}$ such that $\nu \in D$ if there exists $(x, y) \in \Psi$ and $t \in \mathbb{R}$ such that $\nu(x, y) > 0$ and $\nu(x+t, y+t) > 0$. For all $t \geq 0$, define the sets $T_t = \{(x, y)\colon \frac{1}{2} \leq y \leq 1 \text{ and } 0 \leq x \leq y + t - 1\} \subset \Psi$. These sets correspond the triangles and quadrilaterals that were shown in Figure 4. Let $V = \{(x, y) \in \Psi\colon \nu(x, y) > 0\}$, and define $t_0 = \inf\{t\colon V \cap T_t \neq \varnothing\}$. Define

$$g(\nu) = \lim_{\varepsilon\to 0} \sup\{y\colon (x, y) \in V \cap T_{t_0+\varepsilon} \text{ for some } x\}$$

and $h(\nu) = t_0$.

Given a Poisson point process P on Ψ whose intensity has no atoms, we can project the points of P onto the line $y = -x$ in $\mathbb{R}^2$ along perpendicular angles of $\pi/4$. With probability 1, no two points of P will be mapped to the same point under the projection. That is, under the law of P, D has probability 0. Moreover, with probability 1, there will be no limit points under the projection. Therefore, under the intensity measure $A^2(\lambda_{[1/2,1]} \times \lambda)$, there exists a unique point $(x_0, y_0) \in V \cap T_{t_0}$ and an $\varepsilon > 0$ such that $V \cap T_{t_0+\varepsilon} = \{(x_0, y_0)\}$ with probability 1. By definition, $g(P) = y_0$. We claim that g and h are continuous at any Radon measure $\nu \in \mathcal{N} \setminus D$.

Let $\nu \in \mathcal{N} \setminus D$, and let $\{\nu_n\}_{n=1}^{\infty}$ be a sequence of Radon measures that converges weakly to ν. Let $\varepsilon > 0$, and let (x_0, y_0) be the unique point of $T_{t_0+\varepsilon}$ such that $\nu(x_0, y_0) > 0$. For each point $(x', y') \in \Psi$ and every natural number m, define a function

$$f_{(x',y'),m}(x, y) = \begin{cases} -1 & \text{if } |(x, y) - (x', y')| < \varepsilon/m, \\ -\left(2 - \dfrac{m|(x, y) - (x', y')|}{\varepsilon}\right) & \text{if } \varepsilon/m \leq |(x, y) - (x', y')| \leq 2\varepsilon/m, \\ 0 & \text{otherwise.} \end{cases}$$

For large enough m, we have $\int_\Psi f_{(x_0,y_0),m}(x, y)\,\mathrm{d}\nu = -1$, so $\int_\Psi f_{(x_0,y_0),m}(x, y)\,\mathrm{d}\nu_n \to -1$ as $n \to \infty$ for large enough values of m. Because we can make m arbitrarily large, there must be a sequence of points $\{(x_n, y_n)\}_{n=1}^{\infty}$ such that $\nu_n(x_n, y_n) = 1$ for all n and $(x_n, y_n) \to (x_0, y_0)$ as $n \to \infty$. Likewise, for any point $(x', y') \in T_{t_0+\varepsilon}$, there exists a large enough m such that $\int_\Psi f_{(x',y'),m}(x, y)\,\mathrm{d}\nu = 0$, so $\int_\Psi f_{(x',y'),m}(x, y)\,\mathrm{d}\nu_n \to 0$ as $n \to \infty$. This shows that, for large enough n, the Radon measures ν_n will assign measure 0 to all points in a ball of radius ε/m

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

about (x', y'). From this, it is easy to conclude that $g(\nu_n) \to g(\nu)$ and $h(\nu_n) \to h(\nu)$. Therefore, g and h are both continuous on $\mathcal{N} \setminus D$. By Lemma 15 and the continuous mapping theorem, $g(P_M)$ converges in distribution to $g(P_\infty)$ and $h(P_M)$ converges in distribution to $h(P_\infty)$.

The next goal is to show that $g(P_M) - \sigma(M_1) \xrightarrow{\mathrm{P}} 0$ and $h(P_M) - \tau(M_1)/\log N \xrightarrow{\mathrm{P}} 0$. Then we will have $\sigma(M_1) \xrightarrow{\mathrm{D}} g(P_\infty)$ and $\tau(M_1)/\log N \xrightarrow{\mathrm{D}} h(P_\infty)$. To achieve this, we will first show that the probability that (x_0, y_0) corresponds to the cancer-causing type-1 mutation converges in probability to 1. Suppose that (x_0, y_0) does not correspond to the cancer-causing type-1 mutation, and let (x_1, y_1) denote the point in Ψ corresponding to the cancer causing type-1 mutation in M_1. Let $\varepsilon > 0$, and suppose that $(x_1, y_1) \notin T_{t_0+\varepsilon}$. The point $(x_0, y_0) \in T_{t_0}$ corresponds to a successful type-1 mutation in model M_1, and by the way that model M_1 marks points in Ψ, there will be a type-2 mutation in model M_1 that corresponds to a point in T_{t_0}. The ray starting at (x_1, y_1) with an angle of $\pi/4$ will represent all of the descendants of the cancer-causing type-1 mutation. The point on this line whose first coordinate is t_0 will be (t_0, y''), where $y'' \le 1 - \varepsilon$. In this case $\rho(M_1) = y'' \le 1 - \varepsilon$.

Let E_1 be the event that (x_0, y_0) is the point in Ψ that corresponds to the cancer-causing type-1 mutation, and let E_2 be the event that two or more points occur in $T_{t_0+\varepsilon}$. On E_1^C, let (x_1, y_1) be the point in Ψ corresponding to the cancer-causing type-1 mutation. We know that P_M converges in distribution to P_∞ by Lemma 15, so

$$\begin{aligned}\limsup \mathrm{P}(E_1^C) &= \limsup(\mathrm{P}(E_1^C \cap \{(x_1, y_1) \in T_{t_0+\varepsilon}\}) + \mathrm{P}(E_1^C \cap \{(x_1, y_1) \notin T_{t_0+\varepsilon}\})) \\ &\le \limsup \mathrm{P}(E_2) + \limsup \mathrm{P}(\rho(M_1) < 1 - \varepsilon) \\ &\le \frac{A^2}{2}\varepsilon,\end{aligned}$$

where the last line follows because $\rho(M_1) \xrightarrow{\mathrm{P}} 1$ and $\mathrm{P}(E_2) \le \mathrm{P}(V \cap (T_{t_0+\varepsilon} \setminus T_{t_0}) \ne \varnothing)$. Because $\varepsilon > 0$ was chosen arbitrarily, we have $\lim \mathrm{P}(E_1^C) = 0$.

The above has established that $\lim \mathrm{P}(E_1) = 1$. By the definitions of $\sigma(M_1)$ and $g(P_M)$, it is clear that

$$\mathrm{P}(\sigma(M_1) - g(P_M) = 0 \mid E_1) = 1$$

because $\sigma(M_1) = g(P_M) = y_0$. Conditional on the event E_1, we also know that $\tau'(M_1) = (\log N)x_0$. Let (x_0', y_0') be the point in Ψ that corresponds to the type-2 mutation in M_1, so that $\rho(M_1) = y_0'$. Let ν be the Radon measure of points in Ψ induced by M_1, and consider the fact that the descendants of the cancer-causing type-1 mutation will lie on a line starting at (x_0, y_0) with angle $\pi/4$. It is clear that $h(\nu) = t_0 = x_0 + 1 - y_0$ and $\rho(M_1) = y_0 + \tau(M_1)/\log(N) - x_0$. Thus, if $h(\nu) - \tau(M_1)/\log N > \varepsilon$ then $1 - \rho(M_1) > \varepsilon$, or, equivalently, $\rho(M_1) < 1 - \varepsilon$. Therefore, because $\mathrm{P}(E_1) \to 1$,

$$\mathrm{P}\left(h(P_M) - \frac{\tau(M_1)}{\log N} > \varepsilon \;\middle|\; E_1\right) = \mathrm{P}(\rho(M_1) < 1 - \varepsilon \mid E_1) \to 0.$$

Again, using the fact that $\mathrm{P}(E_1) \to 1$, we obtain the desired result.

Now we are left to show that $g(P_\infty)$ and $h(P_\infty)$ have the distributions that are stated in part 4 of Proposition 1. We have $\mathrm{P}(h(P_\infty) \le t)$ is the probability that a point of the Poisson process with intensity $A^2(\lambda_{[1/2,1]} \times \lambda)$ has been marked in T_t. For $t \le \frac{1}{2}$, this is $1 - \mathrm{e}^{-A^2t^2/2}$ and, for $t > \frac{1}{2}$, this is $1 - \mathrm{e}^{-A^2t/2 + A^2/8}$. Therefore,

$$\mathrm{P}\left(\frac{\tau(M_1)}{\log N} \le t\right) \to (1 - \mathrm{e}^{-A^2t^2/2})\,\mathbf{1}_{[0,1/2]}(t) + (1 - \mathrm{e}^{-A^2t/2 + A^2/8})\,\mathbf{1}_{(1/2,\infty)}(t).$$

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

To find the distribution of $g(P_\infty)$, we will use the joint density function of $g(P_\infty)$ and $h(P_\infty)$. From the above computation, it is clear that the density of $h(P_\infty)$ is

$$f_h(t) = A^2 t \mathrm{e}^{-A^2t^2/2}\,\mathbf{1}_{[0,1/2]}(t) + \frac{A^2}{2}\mathrm{e}^{-A^2t/2+A^2/8}\,\mathbf{1}_{(1/2,\infty)}(t).$$

Conditioned on the event that $h(P_\infty) = t$, we know that $g(P_\infty)$ will have uniform distribution. If $t \le \frac{1}{2}$ then $g(P_\infty)$ is uniformly distributed on the interval $[1-t, 1]$. If $t > \frac{1}{2}$ then $g(P_\infty)$ is uniformly distributed on $[\frac{1}{2}, 1]$. This gives us the conditional density function

$$f_{g|h}(s|t) = \begin{cases} \dfrac{1}{t} & \text{if } 1-t \le s \le 1 \text{ and } 0 \le t \le \frac{1}{2}, \\ 2 & \text{if } \frac{1}{2} \le s \le 1 \text{ and } t > \frac{1}{2}. \end{cases}$$

Therefore, the joint density function of $g(P_\infty)$ and $h(P_\infty)$ is

$$f(s,t) = A^2\mathrm{e}^{-A^2t^2/2}\,\mathbf{1}_{[0,1/2]}(t)\,\mathbf{1}_{[1-t,1/2]}(s) + A^2\mathrm{e}^{-A^2t/2+A^2/8}\,\mathbf{1}_{(1/2,\infty)}(t)\,\mathbf{1}_{[1/2,1]}(s).$$

Integrating over t we find that the density of $g(P_\infty)$ is

$$f_g(s) = \left(\int_{1-s}^{1/2} A^2\mathrm{e}^{-A^2t^2/2}\,\mathrm{d}t + 2\mathrm{e}^{-A^2/8}\right)\mathbf{1}_{[1/2,1]}(s).$$

This gives the desired limiting distribution for model M_1. By the usual coupling arguments, the results will hold for model H_1 as well.

Proof of part 6 of Proposition 1. Note that, under these conditions, both mutations occur on daughter cells with probability tending to 1. First we consider a model M_1' in which only generation $l-1$ will acquire type-1 mutations and generation l will acquire type-2 mutations. Also, assume that only one of the daughters will keep a mutation when the cells split so that if a type-1 cell splits it has a type-0 daughter and a type-1 daughter. The rate at which the type-1 mutations occur will be $\mu N/4$ since there are $N/4$ cells in generation $l-1$. Note that $\mu N/4 \sim A\sqrt{N}/4$. The probability that a type-1 mutation will have a type-2 descendant is $1 - \mathrm{e}^{\mu t} \sim \mu t \sim At/\sqrt{N}$. Therefore, the type-2 mutations occur according to a Poisson process whose intensity measure ν satisfies $\nu([0,t]) \ge (A\sqrt{N}/4)(At/\sqrt{N}) = A^2t/4$. We may have to wait up to two time units for the type-2 mutation to occur after the successful type-1 mutation appears. For the sake of a lower bound, we will always assume that it takes two time units after a successful type-1 mutation until the type-2 mutation occurs. By coupling model M_1' with model M_1 in the obvious way we have $\liminf \mathrm{P}(\tau(M_1) \le t) \ge 1 - \mathrm{e}^{-2-A^2t/4}$.

For the upper bound, we consider a model M_1'' in which type-1 cells never undergo apoptosis. There are $N-1$ cells that acquire type-1 mutations, so the type-1 mutations occur at rate $\mu(N-1) \sim A\sqrt{N}$. If we wait t time units after a type-1 mutation has occurred on a cell then the cell will have at most 2^t descendants. If the type-1 mutation had occurred at time 0 and all of the descendants had existed since the type-1 mutation occurred, then the probability that one of the cells had acquired a type-2 mutation would be $t2^{\lfloor t\rfloor}\mu \le t2^t\mu \sim t2^tA/\sqrt{N}$. Because the type-1 mutation may occur after time 0 and there have not been 2^t descendants with the type-1 mutation since the mutation occurred, this is an upper bound on the probability that a type-2 mutation has occurred by time t. Therefore, the type-2 mutations occur according to a Poisson process with intensity $\nu([0,t]) \le (A\sqrt{N})(t2^tA/\sqrt{N}) = t2^tA^2$. By coupling model M_1'' with

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

model M_1 in the obvious way we have $\limsup \mathrm{P}(\tau(M_1) \le t) \le 1 - \mathrm{e}^{-A^2 2^t t}$. This shows part 6 of Proposition 1 with $c = 1 - \mathrm{e}^{-2-A^2 t/4}$ and $C = 1 - \mathrm{e}^{-A^2 2^t t}$.

By Corollary 1 we know that $\rho(M_1) \to 1$. By the definitions of $\sigma(M_1)$ and $\rho(M_1)$ for any $\varepsilon > 0$, if $\rho(M_1) - \sigma(M_1) > \varepsilon$ then $\tau(M_1) > \varepsilon \log N$. Therefore,

$$\mathrm{P}(\rho(M_1) - \sigma(M_1) > \varepsilon) \le \mathrm{P}(\tau(M_1) > \varepsilon \log N) \le \mathrm{e}^{-A^2 2^{\delta \log N}(\delta \log N)} \to 0.$$

Let $\varepsilon > 0$ and $\delta > 0$, and choose N large enough so that $\mathrm{P}(1 - \rho(M_1) > \varepsilon/2) < \delta/2$ and $\mathrm{P}(\rho(M_1) - \sigma(M_2) > \varepsilon/2) < \delta/2$. Then

$$\begin{aligned} \mathrm{P}(1 - \sigma(M_1) > \varepsilon) &= \mathrm{P}(1 - \rho(M_1) + \rho(M_1) - \sigma(M_1) > \varepsilon) \\ &\le \mathrm{P}\big(1 - \rho(M_1) > \tfrac{1}{2}\varepsilon\big) + \mathrm{P}\big(\rho(M_1) - \sigma(M_1) > \tfrac{1}{2}\varepsilon\big) \\ &< \delta. \end{aligned}$$

Therefore, $\sigma(M_1) \xrightarrow{\mathrm{P}} 1$.

By the usual coupling arguments we obtain the same results for H_1.

Appendix A

A.1. Notation

N	The size of the population.
l	Equals $\log N$.
u_1	The rate at which the stem cell line acquires type-1 mutations.
u_2	The rate at which the stem cell line acquires type-2 mutations.
v_1	The rate at which the daughter cells acquire type-1 mutations.
v_2	The rate at which the stem cell line acquires type-1 mutations.
$\tau'(A)$	The time at which the cancer-causing type-1 mutation occurs.
$\tau(A)$	The first time that any cell acquires a type-2 mutation in model A.
$\sigma(A)$	Equals j/l when the cancer-causing type-1 mutation occurs in generation j in model A. If the cancer-causing type-1 mutation occurs on the stem cell then $\sigma(A) = 0$.
$\rho(A)$	Equals j/l when the first type-2 mutation occurs in generation j in model A. If the first type-2 mutation occurs on the stem cell then $\rho(A) = 0$.
α	The number satisfying $\lim_{N\to\infty} v_2 N^\alpha = \beta$ for some $\beta > 0$.
X	A exponentially distributed random variable with mean 1.
Y	A random variable with the Rayleigh distribution. Namely, $\mathrm{P}(Y \le t) = 1 - \mathrm{e}^{-t^2/2}$ for $t > 0$.
μ	The rate at which cells acquire mutations when $u_1 = u_2 = v_1 = v_2$.

A.2. Auxiliary models

H_2	The same as model H_1 except daughter cells may accumulate multiple type-1 mutations.
M_1	The same as model H_2 except no mutations occur on the stem cell line.
M_2	The same as model H_2 except that only stem cells receive type-1 mutations and only daughter cells receive type-2 mutations.
M_3	The same as model H_2 except that only stem cells receive mutations.

https://doi.org/10.1239/aap/1354716589 Published online by Cambridge University Press

Acknowledgements

I would like to thank Jason Schweinsberg for patiently helping me work through various parts of the problem and for helping to revise the first drafts of the paper. I would also like to thank the referees for their helpful comments on a previous draft of the paper.

References

[1] Armitage, P. and Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer* **8,** 1–12.
[2] Cairns, J. (2006). Cancer and the immortal strand hypothesis. *Genetics* **174,** 1069–1072.
[3] Dingli, D. and Michor, F. (2006). Successful therapy must eradicate cancer stem cells. *Stem Cells* **24,** 2603–2610.
[4] Durrett, R. and Moseley, S. (2010). A spatial model for tumor growth. Preprint.
[5] Durrett, R., Schmidt, D. and Schweinsberg, J. (2009). A waiting time problem arising from the study of multi-stage carcinogenesis. *Ann. Appl. Prob.* **19,** 676–718.
[6] Frank, S. A., Iwasa, Y. and Nowak, M. A. (2003). Patterns of cell divisions and the risk of cancer. *Genetics* **163,** 1527–1532.
[7] Fristedt, B. and Gray, L. (1997). *A Modern Approach to Probability Theory.* Birkhäuser, Boston, MA.
[8] Iwasa, Y., Michor, F., Komarova, N. L. and Nowak, M. A. (2005). Population genetics of tumor suppressor genes. *J. Theoret. Biol.* **233,** 15–23.
[9] Kingman, J. F. C. (1993). *Poisson Processes.* Oxford University Press, New York.
[10] Knudson, A. G., Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proc. Nat. Acad. Sci. USA* **68,** 820–823.
[11] Knudson, A. G., Jr. (1978). Retinoblastoma: a prototypic hereditary neoplasm. *Semin. Oncol.* **5,** 57–60.
[12] Komarova, N. L. (2007). Loss- and gain-of-function mutations in cancer: mass-action, spatial and hierarchical models. *J. Statist. Phys.* **128,** 413–446.
[13] Komarova, N. L. and Cheng, P. (2006). Epithelial tissue architecture protects against cancer. *Math. Biosci.* **200,** 90–117.
[14] Komarova, N. L. and Wang, L. (2004). Initiation of colorectal cancer: where do the two hits hit? *Cell Cycle* **3,** 1558–1565.
[15] Komarova, N. L., Sengupta, A. and Nowak, M. A. (2003). Mutation-selection networks of cancer initiation: tumor suppressor genes and chromosomal instability. *J. Theoret. Biol.* **233,** 433–450.
[16] Michor, F. (2007). Chronic myeloid leukemia blast crises arises from progenitors. *Stem Cells* **25,** 1114–1118.
[17] Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life.* Belknap Press, Cambridge, MA.
[18] Schweinsberg, J. (2008). The waiting time for m mutations. *Electron. J. Prob.* **13,** 1442–1478.
[19] Wodarz, D. and Komarova, N. L. (2005). *Computational Biology of Cancer: Lecture Notes and Mathematical Modeling.* World Scientific Publishing, London.