



How mutation accumulation depends on the structure of the cell lineage treeImre Derényi ^{*}*ELTE Eötvös University, Department of Biological Physics, Pázmány Péter Sétány 1A, H-1117 Budapest, Hungary
and MTA-ELTE Statistical and Biological Physics Research Group, Pázmány Péter Sétány 1A, H-1117 Budapest, Hungary*Márton C. Demeter , Mario Pérez-Jiménez , and Dániel Grajzel*ELTE Eötvös University, Department of Biological Physics, Pázmány Péter Sétány 1A, H-1117 Budapest, Hungary
and MTA-ELTE “Lendület” Evolutionary Genomics Research Group, Pázmány Péter Sétány 1A, H-1117 Budapest, Hungary*Gergely J. Szöllősi [†]*ELTE Eötvös University, Department of Biological Physics, Pázmány Péter Sétány 1A, H-1117 Budapest, Hungary;
MTA-ELTE “Lendület” Evolutionary Genomics Research Group, Pázmány Péter Sétány 1A, H-1117 Budapest, Hungary;
HUN-REN Centre for Ecological Research, Institute of Evolution, H-1113 Budapest, Hungary;
and Model-Based Evolutionary Genomics Unit, Okinawa Institute of Science and Technology Graduate University, 904-0412 Okinawa, Japan*

(Received 15 March 2023; accepted 8 March 2024; published 12 April 2024)

All the cells of a multicellular organism are the product of cell divisions that trace out a single binary tree, the so-called cell lineage tree. Because cell divisions are accompanied by replication errors, the shape of the cell lineage tree is a key determinant of how somatic evolution, which can potentially lead to cancer, proceeds. Carcinogenesis requires the accumulation of a certain number of driver mutations. By mapping the accumulation of mutations into a graph theoretical problem, we present an exact numerical method to calculate the probability of collecting a given number of mutations and show that for low mutation rates it can be approximated with a simple analytical formula, which depends only on the distribution of the lineage lengths, and is dominated by the longest lineages. Our results are crucial in understanding how natural selection can shape the cell lineage trees of multicellular organisms and curtail somatic evolution.

DOI: [10.1103/PhysRevE.109.044407](https://doi.org/10.1103/PhysRevE.109.044407)**I. INTRODUCTION**

Cells of multicellular organisms, regardless of the ultimate complexity of the organism they are the part of, trace their history back to a single cell. They are related by a single binary tree. Because every cell division is accompanied by DNA replication, which is an error-prone process, the shape of this tree (referred to as the *cell lineage tree*) determines the rate at which mutations accumulate and, as a result, the tempo and mode of somatic evolution. Early in development, when cells of non-renewing tissues (for example, primary oocytes in the female germline [1,2]) and the initial population of tissue-specific stem cells in self-renewing tissues (for example, the hematopoietic stem cells [3,4] or the spermatogonia of the male germline [1],[2]) are produced, cell lineage trees closely follow a perfect binary tree, which minimizes the number of cell divisions and, consequently, the accumulation of mutations [5].

Complex multicellular organisms are, however, defined by differentiated cells that make up their tissues, most of which must be continually renewed. The sustained supply of differentiated cells required during an organism’s lifetime is produced along differentiation hierarchies, which also have a second function central to the maintenance of multicellularity: to limit somatic evolution [5–8]. A fundamental question for understanding how tissues can limit somatic evolution (and its consequences such as aging and cancer) is how the shape of cell lineage trees [5,8–10] determines mutation accumulation.

Cancer is a disease of multicellular organisms, which occurs when a somatic cell, after going through a number of genetic and epigenetic changes, starts to proliferate uncontrollably [11]. It was Armitage and Doll [12,13] who observed that the incidence of various types of cancers in humans grows as a power function of age and proposed a multistage model of cancer, where the exponent of the power function increased by unity (which typically falls between 5 and 7) corresponds to the number m of driver mutations required for cancer initiation. Although we now know much more about carcinogenesis, the fundamental concept of the accumulation of a few critical mutations remains widely accepted [14–16].

After every cell division new driver mutations can occur randomly, the number of which in each daughter cell is assumed to follow a Poisson distribution with a mean value denoted as μ and referred to as the *driver mutation rate per cell division*. This rate corresponds to the product of the number of

^{*}Corresponding author: derenyi@elte.hu[†]Corresponding author: ssolo@elte.hu

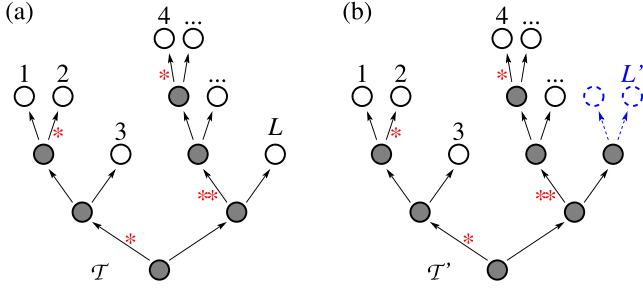


FIG. 1. Illustration of a cell lineage tree with mutation accumulation. (a) A cell lineage tree \mathcal{T} with L leaf nodes (white circles) and $L - 1$ internal nodes (gray circles). The lineage length (or divisional load) D_i corresponding to leaf node i is the number of edges (cell divisions, denoted by arrows) leading from the root (bottom-most) node to the leaf node. Mutations are indicated by red stars. (b) Cell lineage tree \mathcal{T}' obtained by making leaf node L in tree \mathcal{T} divide (indicated by blue dashed arrows and circles).

driver genes, the average mutational target size per gene, and the base-pair mutation rate per cell division. Estimates for the driver mutation rate per cell division [17–19] vary in the range of $\mu = 10^{-6}$ to 10^{-4} , reflecting tissue specific uncertainties in each factor. What is common in these estimates is that the driver mutation rate is much smaller than unity ($\mu \ll 1$), i.e., driver mutations occur very rarely.

We make the simplifying assumption that mutations (except for deleterious ones, which terminate the respective lineages) do not affect the dynamics of cell proliferation and, as a result, do not change the structure of the cell lineage tree until a critical number of drivers are accumulated. It is unclear (and might depend on the type of tissue or tumor) to what extent this assumption of neutrality holds for individual somatic mutations (including putative driver mutations). While there are clear exceptions [20], the above assumption of neutrality is consistent with the fact that the majority of cancers arise without a histologically discernible premalignant phase, and also with recent timing analyses, which suggest that driver mutations often precede diagnosis by many years, if not decades [21]. These observations indicate strong cooperation between driver mutations, suggesting that major histological changes that would significantly alter the structure of the lineage tree may not take place until the full repertoire of driver mutations is acquired [22].

Here we consider the accumulation of a fixed number of mutations m along a generic cell lineage tree \mathcal{T} [demonstrated in Fig. 1(a)], representing the full history of cell divisions of an organism until a given moment in time. On this tree the leaf nodes correspond to the cells that were either present in the organism at that time or had been lost by then, while the internal nodes correspond to the cells that had already gone through a cell division. Similar lineage trees can also be drawn for tissues or tissue units with the root node being the founder cell of the tissue or tissue unit. Note that at any particular moment the entire lineage tree (and not only the subtree leading to the living cells at a particular point in time) needs to be considered. This is because wherever a cell collects the critical number of driver mutations, cancer is initiated and an uncontrolled growth begins. Therefore, even if none of

the descendants of a cell survive along the original (unaltered) lineage tree, it had a chance to initiate cancer at its birth.

Assuming that driver mutations occur at a uniform rate μ per cell division (indicated by red stars in Fig. 1) and that they are neutral until at least m are accumulated, we explore the fundamental question of how the probability of accumulating m mutations can be calculated for a generic cell lineage tree \mathcal{T} .

II. RESULTS

The above question can be translated into a mathematical (graph theoretical) problem: Given a binary tree \mathcal{T} , what is the probability $P_{\mathcal{T}}(\mu, m)$ that a lineage (i.e., a path from the root node to a leaf node) with at least m mutations appears, if mutations are dropped to the edges independently with an expected value of μ per edge? In Fig. 1(a), e.g., the lineage belonging to leaf node 4 has three mutations (red stars). Because mutations occur independently, their number (j) follows a Poisson distribution:

$$p(\mu, j) = \frac{\mu^j}{j!} e^{-\mu}. \quad (1)$$

Although no general analytical solution is known for this problem, an exact numerical answer can be given for any particular tree by following the procedure outlined below. Any binary tree can be constructed by iteratively merging the subtrees of sister nodes at their parent node, starting from the leaves. At each internal node (gray nodes in Fig. 1) the two subtrees of its descendant nodes (daughter cells) are merged. The last merger at the root node of the lineage tree completes the merging sequence.

For a merger event, let us denote the two sibling subtrees to be merged by \mathcal{A} and \mathcal{B} . Let us denote the probability that the lineage with the largest number of mutations in subtree \mathcal{A} has exactly j mutations by $\tilde{P}_{\mathcal{A}}(\mu, j)$, and in subtree \mathcal{B} by $\tilde{P}_{\mathcal{B}}(\mu, j)$. After extending each subtree by adding the edge that leads to their shared parent node, the probability that the lineage with the largest number of mutations has exactly k mutations in the extended subtree \mathcal{A} is

$$\tilde{P}_{\mathcal{A}}^+(\mu, k) = \sum_{j=0}^k \tilde{P}_{\mathcal{A}}(\mu, k-j) p(\mu, j) \quad (2)$$

and that in the extended subtree \mathcal{B} is

$$\tilde{P}_{\mathcal{B}}^+(\mu, k) = \sum_{j=0}^k \tilde{P}_{\mathcal{B}}(\mu, k-j) p(\mu, j). \quad (3)$$

After joining the two extended sibling subtrees at their parent node into subtree $(\mathcal{A}\mathcal{B})$, the probability that the lineage with the largest number of mutations in this newly merged subtree has exactly m mutations is

$$\begin{aligned} \tilde{P}_{(\mathcal{A}\mathcal{B})}(\mu, m) &= \tilde{P}_{\mathcal{A}}^+(\mu, m) \sum_{k=0}^{m-1} \tilde{P}_{\mathcal{B}}^+(\mu, k) \\ &+ \tilde{P}_{\mathcal{B}}^+(\mu, m) \sum_{k=0}^{m-1} \tilde{P}_{\mathcal{A}}^+(\mu, k) \\ &+ \tilde{P}_{\mathcal{A}}^+(\mu, m) \tilde{P}_{\mathcal{B}}^+(\mu, m). \end{aligned} \quad (4)$$

The initial condition for the merger sequence is that, when the subtree (e.g., \mathcal{A}) is a leaf node (i.e., a trivial graph consisting of a single node and no edges), $\tilde{P}_{\mathcal{A}}(\mu, j) = 0$ for $j > 0$ and $\tilde{P}_{\mathcal{A}}(\mu, 0) = 1$.

After the merging process is completed the probability that the entire lineage tree \mathcal{T} has a lineage with at least m mutations can be obtained as

$$P_{\mathcal{T}}(\mu, m) = 1 - \sum_{m'=0}^{m-1} \tilde{P}_{\mathcal{T}}(\mu, m'). \quad (5)$$

No closed-form formula can exist for this result, because it depends on the detailed structure of the tree. For low mutation rates, however, simplifying approximations can be made. If μ is small enough such that $P_{\mathcal{T}}(\mu, m) \ll 1$, then whenever a lineage with at least m mutations appears, it appears practically alone, leading from the root node to one or a few closely related leaf nodes. This suggests that $P_{\mathcal{T}}(\mu, m)$ can be expressed as a simple sum,

$$P_{\mathcal{T}}(\mu, m) \approx \sum_{i=1}^L F(i, \mu, m), \quad (6)$$

for all the leaf nodes from 1 to L . Furthermore, because the only quantity that is specific to a leaf node is its lineage length (or divisional load, denoted by D_i), $F(i, \mu, m)$ should depend on i only through D_i . It is also expected to be dominated by its leading-order term in μ , which is proportional to μ^m . Altogether, it can be approximated as $F(i, \mu, m) \approx \mu^m f(D_i, m)$. Thus, for low enough mutation rates μ , the probability that a lineage with at least m mutations exists in the cell lineage tree should take the form

$$P_{\mathcal{T}}(\mu, m) = \sum_{i=1}^L \mu^m f(D_i, m) \quad (7)$$

in leading order of μ .

The function $f(D_i, m)$ can be determined by introducing an additional division to the lineage tree [for which we chose leaf node L , as indicated by blue dashed lines in Fig. 1(b)]. The probability for the new tree \mathcal{T}' can be written in two ways. First,

$$\begin{aligned} P_{\mathcal{T}'}(\mu, m) &= \sum_{i=1}^{L-1} \mu^m f(D_i, m) + 2\mu^m f(D_L + 1, m) \\ &= P_{\mathcal{T}}(\mu, m) + \mu^m [2f(D_L + 1, m) - f(D_L, m)]. \end{aligned} \quad (8)$$

Second,

$$P_{\mathcal{T}'}(\mu, m) = P_{\mathcal{T}}(\mu, m) + [1 - P_{\mathcal{T}}(\mu, m)]G_{\mathcal{T}}(L, \mu, m), \quad (9)$$

where the first term accounts for the possibility that a lineage with at least m mutations has already existed before the introduction of the new division, and the second term corresponds to the scenario that no such lineage has existed, but with the elongations at least one of the two new lineages reaches the necessary number of mutations. The probability of this latter event is denoted by $G_{\mathcal{T}}(L, \mu, m)$. In the leading order of μ , the probability $P_{\mathcal{T}}(\mu, m)$ can be neglected in the second term

and $G_{\mathcal{T}}(L, \mu, m)$ can be expressed as

$$G_{\mathcal{T}}(L, \mu, m) = 2 \left\{ \frac{[\mu(D_L + 1)]^m}{m!} - \frac{[\mu D_L]^m}{m!} \right\}, \quad (10)$$

where the two terms between the braces describe the probabilities that m mutations occur along $D_L + 1$ and D_L long lineages, respectively, their difference corresponds to the excess probability conferred by a single elongation, and the factor 2 stands for the number of elongations. Here $\mu D_L \ll 1$ has been assumed, which follows from the condition that $P_{\mathcal{T}}(\mu, m) \ll 1$. After plugging this formula into Eq. (9) we arrive at

$$P_{\mathcal{T}'}(\mu, m) = P_{\mathcal{T}}(\mu, m) + \mu^m g(D_L, m), \quad (11)$$

where

$$g(D, m) = 2 \frac{(D + 1)^m - D^m}{m!}. \quad (12)$$

Comparing the two expressions (8) and (11) for $P_{\mathcal{T}'}(\mu, m)$ confirms the validity of our expectations on the form of $P_{\mathcal{T}}(\mu, m)$ in Eq. (7) and leads to the recursion relation

$$f(D + 1, m) = \frac{f(D, m) + g(D, m)}{2} \quad (13)$$

for the function $f(D, m)$, with the initial values $f(0, m) = 0$ for any $m > 0$.

Expanding the recursion results in

$$f(D, m) = \sum_{l=1}^D \frac{g(D-l, m)}{2^l} = \sum_{l=1}^D \frac{(D-l+1)^m - (D-l)^m}{2^{l-1} m!}, \quad (14)$$

which cannot be simplified further. However, because in most real cell lineage trees the lineages [especially the longest ones that dominate the sum in Eq. (7)] are much longer than unity, $f(D, m)$ can be well approximated by keeping its highest-order term in D :

$$f(D, m) \approx 2 \frac{D^{m-1}}{(m-1)!}. \quad (15)$$

Plugging this approximation back into Eq. (7) leads to our second main result that the probability of accumulating at least m mutations along a lineage can be well approximated with

$$P_{\mathcal{T}}^{\text{appr}}(\mu, m) = \frac{2\mu^m}{(m-1)!} \sum_{i=1}^L D_i^{m-1} \quad (16)$$

in the leading order of μ and for the highest order of the lineage lengths. This approximation is very accurate for low enough mutation rates μ [such that $P_{\mathcal{T}}^{\text{appr}}(\mu, m) \ll 1$, which is a valid assumption for the cancer risks of independently maintained units of most human tissues], as demonstrated in Fig. 2 for a series of lineage trees that interpolate between the most skewed binary tree (a linear chain of all the internal nodes) and the most balanced one (the perfect binary tree) with $L = 2^{16}$ leaf nodes.

III. DISCUSSION

We derived our results for a cell lineage tree, where the root node is either the zygote or the founder cell of a tissue or

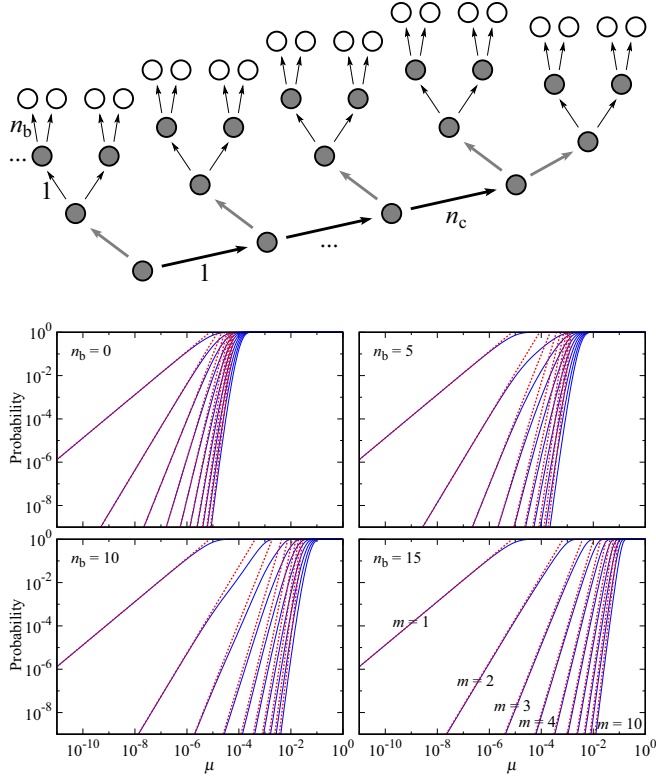


FIG. 2. Examples for the probability of accumulating at least m mutations in a series of cell lineage trees. The top panel illustrates how a series of lineage trees that interpolate between the most skewed binary tree (a linear chain of all the internal nodes, $n_b = 0$) and the most balanced one (the perfect binary tree, $n_c = 0$) is generated: Identical perfect binary subtrees (indicated with thin arrows) of depths (lineage lengths) n_b are joined (with thick gray arrows) to an n_c long linear chain (indicated with thick black arrows). The number of leaf nodes can be expressed as $L = (n_c + 2) 2^{n_b}$. The four plots in the bottom panel show the exact probabilities $P_{\mathcal{T}}(\mu, m)$ (blue lines, obtained from the merging process) that the lineage trees (for $L = 2^{16}$ and for four different values of $n_b = 0, 5, 10,$ and 15) have a lineage with at least m mutations (ranging between $m = 1$ and 10 from left to right) as a function of the mutation rate μ , as well as their approximation $P_{\mathcal{T}}^{\text{appr}}(\mu, m)$ (red dashed lines) using Eq. (16). The approximation is very accurate as long as $P_{\mathcal{T}}^{\text{appr}}(\mu, m) \ll 1$.

a unit of tissue. Most tissue units (such as the colonic crypts), however, are sustained by a population of stem cells, rather than a single stem cell. Because our main formula (16) is a simple sum over all the leaves, it is also valid for a collection of trees, such as those generated by an initial set of stem cells of an independently maintained tissue unit.

Let us demonstrate our results by comparing two basic population dynamics models for N cells (or individuals, in general) in discrete time, as illustrated in Fig. 3 (where each lost cell is explicitly marked with a cross). One of them is an ensemble of N parallel linear chains (as introduced in Fig. 2 with $n_b = 0$). At each time step a randomly selected cell divides, and one of its daughter cells becomes lost (to maintain the population size). The other model is the Moran process, in which at each time step two cells are chosen randomly, one for division and one for loss (with $1/N$ chance the same cell is chosen for both events, which means that after division one of its daughters becomes lost). Lost cells are explicitly marked with a cross.

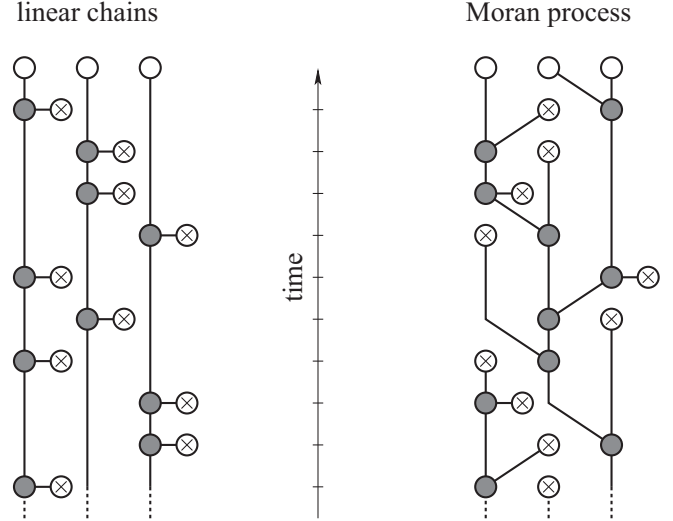


FIG. 3. Examples for two basic population dynamics models of $N = 3$ cells in discrete time. The left panel shows an ensemble of N parallel linear chains. At each time step a randomly selected cell divides, and one of its daughter cells becomes lost. The right panel shows a Moran process, in which at each time step two cells are chosen randomly, one for division and one for loss (with $1/N$ chance the same cell is chosen for both events, which means that after division one of its daughters becomes lost). Lost cells are explicitly marked with a cross.

chosen for both events, which means that after division one of its daughters becomes lost). In both models as time progresses the number of past divisions (i.e., the lineage length) of each cell increases linearly (with slightly different rates in the two models); therefore, the lineage lengths of the lost cells follow a uniform distribution on average, and the two models behave similarly. In both models, time is measured in units of the time step and the generation time is defined as N time steps, and we are interested in the value of $P_{\mathcal{T}}^{\text{appr}}(\mu, m)$ after a large number of generations G have passed.

In the model of linear chains, leaf node t (lost at time t , where $0 < t < NG$) has a lineage length of $D_t = t/N$ on average (because at each time step every cell divides with a probability of $1/N$). Thus, lost cells have a contribution of $\sum_{t=1}^{NG-1} (t/N)^{m-1} \approx \int_0^{NG} (t/N)^{m-1} dt = NG^m/m$ to the summation in Eq. (16). The contribution of the N surviving cells (NG^{m-1} with each lineage having a length of G on average) is negligible for $G \gg 1$. From these the value of $P_{\mathcal{T}}^{\text{appr}}(\mu, m)$ for this model can be approximated as

$$P_{\mathcal{T}}^{\text{appr,chain}}(\mu, m) \approx \frac{2N(\mu G)^m}{m!}. \quad (17)$$

An intuitive argument for the $2N$ factor is that the last mutation can occur in either of the two daughter cells of any of the N chains, while all the previous $m - 1$ mutations must have occurred along the ancestral lineage, involving only the surviving daughter cells.

In the Moran model the lineage lengths (numbers of past divisions) also grow linearly in time, but somewhat faster (as is apparent in Fig. 3). This can be understood by realizing that at each time step the sum of the lineage lengths for all the N

cells of the population increases by $\lambda = 2(1 - 1/N) + 1/N = 2 - 1/N$ rather than by unity on average, where the first term describes the increment produced by the two daughter cells both surviving with a chance of $1 - 1/N$, while the second term corresponds to the event when only one of the daughters survives with a chance of $1/N$. Thus, after each generation the lineage length of every cell increases by λ . Again, the main contribution to the summation in Eq. (16) comes from the lost cells. With leaf node t (lost at time t , where $0 < t < NG$) having a lineage length of $D_t = \lambda t/N$ on average, this contribution is $\sum_{t=1}^{NG-1} (\lambda t/N)^{m-1} \approx \int_0^{NG} (\lambda t/N)^{m-1} dt = N\lambda^{m-1} G^m/m$, and the value of $P_{\mathcal{T}}^{\text{appr}}(\mu, m)$ can be approximated as

$$P_{\mathcal{T}}^{\text{appr.Moran}}(\mu, m) \approx \frac{2N(\mu G)^m}{m!} \lambda^{m-1} \approx P_{\mathcal{T}}^{\text{appr.chain}}(\mu, m) \lambda^{m-1}. \quad (18)$$

Although the Moran model behaves similarly to the model of linear chains, it produces longer lineages (by a factor of λ) and, therefore, has a higher chance of accumulating m mutations after the same number of generations by a factor of λ^{m-1} .

These examples also highlight the significance of our result that for the calculation of the risk of cancer it is enough to determine the lineage length distribution of the cell lineage tree, and no further information is required from its structural details, as long as the risk is much smaller than unity. To see how realistic this assumption is, let us make an estimation for human stem cells. Bases on the latest counts [23] an adult has around 3×10^{13} cells, out of which of the order of 10^{10} are tissue specific stem cells. Because the lifetime risk of cancer is certainly smaller than unity, the cancer risk carried by the lineage tree of a single stem cell is smaller than 10^{-10} . Even if there are (orders of magnitude) errors in this estimation, the cancer risk per stem cell should be much smaller than unity, indicating that the parameter range where our approximate formula is accurate is the biologically relevant one.

There is previous work dealing with the probability of collecting a given number of neutral mutations [24,25]. However, these studies are restricted to particular population dynamics models, use stochastic simulations and, as a consequence, cannot explore the regime of very small, biologically relevant probabilities. Analytical considerations [25] take only the genealogy (ancestry) of the population into account and ignore all cells with no extant descendants. Here we emphasize that all the cells (no matter if they have any extant descendants) have a contribution to the total probability and, therefore, the entire cell lineage tree must always be considered.

The above studies [24,25] point out that stem cell populations with symmetric divisions (corresponding to the Moran model) accumulate a given number of neutral mutations with lower probabilities than populations with asymmetric divisions (corresponding to our model of linear chains), although the ratio of the two probabilities seems to converge to unity for very small probabilities. This appears to contradict our result of λ^{m-1} for this ratio, which is clearly larger than unity if $N > 1$ and $m > 1$. The resolution of this paradox is that, when two different cells are selected for division and lost in a time step of the Moran model (with a chance of $1 - 1/N$), then the lost cell is considered to produce two

progenitor cells (symmetrically); i.e., these cells undergo one more round of cell divisions. Thus, in each generation not N , but $2(1 - 1/N) + 1/N = 2 - 1/N$ progenitors are produced, which coincides with the value of λ . Consequently, λ cancels in the per progenitor increment of the lineage length (number of past divisions), which makes the probability of accumulating a given number of mutations identical in the two cases. This can be viewed as a direct application of our results, which shows decisively that for very small (biologically relevant) cancer risks the symmetric divisions have no advantage over the asymmetric ones, despite that stochastic simulations at higher probabilities indicate otherwise.

Our approach considers only rare driver mutations associated with cell divisions. In reality other types of more frequent mutations (e.g., ones conferring small fitness advantages or disadvantages, or lethal ones) also occur. Their effects, however, are assumed to be incorporated into the structure of the lineage tree, along which the driver mutations can accumulate, and therefore, they do not need to be treated together with the drivers. Some driver mutations can also occur without cell divisions. Although they do not fit directly into our framework, we can mimic them by decorating the original cell lineage tree with unit-long phantom edges (i.e., by adding phantom cell divisions, where one of the daughter cells dies immediately).

We note that our results can also be applied to other evolutionary problems (e.g., maintaining cell cultures) where the divisions of the individuals are accompanied by mutations, and the accumulation of a given number of rare critical mutations has serious consequences.

In summary, the main results of our work are that assuming neutral mutations (i) the probability of accumulating a given number of mutations can be calculated exactly with an iterative method based on subtree merging [Eqs. (2)–(5)] for an arbitrary tree; and (ii) in the case of rare mutations (which is typical for the driver mutations necessary for tumor initiation) it can be approximated with a simple analytical formula [Eq. (16)], which depends only on the lineage length distribution of the leaf nodes. Because the approximate formula is a sum of relatively high powers of the lineage lengths, it is dominated by the longest lineages. This explains why the minimization of the longest lineages (e.g., through hierarchical differentiation [5]) is crucial in minimizing somatic evolution, in general, and cancer risk, in particular. The approximate formula is the simplest one that is theoretically possible, because it involves those and only those properties of the cell lineage tree that matter (the lineage length distribution); therefore, it cannot be reduced any further. An important feature of our approach is that it considers the cell lineage tree fixed, along which mutations accumulate. This makes it possible to calculate the probability of mutation accumulation for arbitrary mutation rates (with either the exact method or the approximate method), without the need for stochastic simulations of the generation of mutations.

In the context of cancer incidence, our results imply that the cumulative cancer incidence for any particular tissue should grow as a power function of age with exponent m , if the lengths of its longest lineages grow linearly in time [cf. Eq. (17) for long linear chains]. Indeed, for most large self-renewing tissues stem cells are known to divide at least several times a year (e.g., every 25 to 50 weeks [26] or every

2 to 20 months [9] for blood, and every 4 days [27,28] for the colon) and, therefore, are expected to produce dominant linearly growing lineages.

It is also clear, however, that differences across individuals and between tissues of different individuals can influence both the mutation rate (through differential exposure to mutagenic influences of either endogenous or exogenous origin) and the shape of cell lineage trees (through, e.g., increased cell proliferation caused by chronic inflammation in the tissue). Cancer risk is, thus, influenced by multiple factors, and the structure of the wild-type cell lineage tree is one of them. The significance of our work is that our analytical results provide a tool to determine the contribution of the structure of the lineage tree to the risk of cancer. This contribution is not expected to be minimized by evolution, but only to be kept below some small value (blending into the contributions of other relevant factors) such that the organism will have a high chance of not dying of cancer during its natural lifetime.

We foresee direct applications of our analytical results as more extensive cell lineage tracing data become available with emerging new single-cell techniques (such as single-cell RNA sequencing or cytometry by time of flight [29]), at both the organism level [30,31] and the tissue level [32]. Our results are also readily applicable for quantifying cancer susceptibility in theoretical models of tissue development and maintenance.

ACKNOWLEDGMENTS

The authors are thankful to Ágnes Backhausz for her valuable comments and advice. G.J.S. received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 714774 and Grant No. GINOP-2.3.2.-15-2016-00057.

I.D. and G.J.S. contributed equally to this work.

-
- [1] Z. Gao, M. J. Wyman, G. Sella, and M. Przeworski, Interpreting the dependence of mutation rates on age and time, *PLoS Biol.* **14**, e1002355 (2016).
- [2] J. F. Crow, The origins, patterns and implications of human spontaneous mutation, *Nat. Rev. Genet.* **1**, 40 (2000).
- [3] K. Busch, K. Klapproth, M. Barile, M. Flossdorf, T. Holland-Letz, S. M. Schlenner, M. Reth, T. Höfer, and H.-R. Rodewald, Fundamental properties of unperturbed haematopoiesis from stem cells *in vivo*, *Nature (London)* **518**, 542 (2015).
- [4] B. Werner, F. Beier, S. Hummel, S. Balabanov, L. Lassay, T. Orlikowsky, D. Dingli, T. H. Brümmendorf, and A. Traulsen, Reconstructing the *in vivo* dynamics of hematopoietic stem cells from telomere length distributions, *Elife* **4**, e08687 (2015).
- [5] I. Derényi and G. J. Szöllősi, Hierarchical tissue organization as a general mechanism to limit the accumulation of somatic mutations, *Nat. Commun.* **8**, 14545 (2017).
- [6] M. A. Nowak, F. Michor, and Y. Iwasa, The linear process of somatic evolution, *Proc. Natl. Acad. Sci. USA* **100**, 14966 (2003).
- [7] J. W. Pepper, K. Sprouffske, and C. C. Maley, Animal cell differentiation patterns suppress somatic evolution, *PLoS Comput. Biol.* **3**, e250(2007).
- [8] M. Demeter, I. Derényi, and G. J. Szöllősi, Trade-off between reducing mutational accumulation and increasing commitment to differentiation determines tissue organization, *Nat. Commun.* **13**, 1666 (2022).
- [9] H. Lee-Six, N. F. Øbro, M. S. Shepherd, S. Grossmann, K. Dawson, M. Belmonte, R. J. Osborne, B. J. Huntly, I. Martincorena, E. Anderson *et al.*, Population dynamics of normal human blood inferred from somatic mutations, *Nature (London)* **561**, 473 (2018).
- [10] E. Mitchell, M. Spencer Chapman, N. Williams, K. J. Dawson, N. Mende, E. F. Calderbank, H. Jung, T. Mitchell, T. H. Coorens, D. H. Spencer *et al.*, Clonal dynamics of haematopoiesis across the human lifespan, *Nature (London)* **606**, 343 (2020).
- [11] L. Nunney, C. C. Maley, M. Breen, M. E. Hochberg, and J. D. Schiffman, Peto's paradox and the promise of comparative oncology, *Philos. Trans. R. Soc., B* **370**, 20140177 (2015).
- [12] P. Armitage and R. Doll, The age distribution of cancer and a multi-stage theory of carcinogenesis, *Br. J. Cancer* **8**, 1 (1954).
- [13] R. Doll, The age distribution of cancer: Implications for models of carcinogenesis, *J. R. Stat. Soc. A* **134**, 133 (1971).
- [14] S. H. Moolgavkar, Commentary: Fifty years of the multistage model: Remarks on a landmark paper, *Int. J. Epidemiol.* **33**, 1182 (2004).
- [15] N. Neu, B. Ploier, and C. Ofner, Cardiac myosin-induced myocarditis. heart autoantibodies are not involved in the induction of the disease, *J. Immunol.* **145**, 4094 (1990).
- [16] P. M. Altrock, L. L. Liu, and F. Michor, The mathematics of cancer: Integrating quantitative models, *Nat. Rev. Cancer* **15**, 730 (2015).
- [17] I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein, and M. A. Nowak, Accumulation of driver and passenger mutations during tumor progression, *Proc. Natl. Acad. Sci. USA* **107**, 18545 (2010).
- [18] C. D. McFarland, L. A. Mirny, and K. S. Korolev, Tug-of-war between driver and passenger mutations in cancer and other adaptive processes, *Proc. Natl. Acad. Sci. USA* **111**, 15138 (2014).
- [19] K. Lahouel, L. Younes, L. Danilova, F. M. Giardiello, R. H. Hruban, J. Groopman, K. W. Kinzler, B. Vogelstein, D. Geman, and C. Tomasetti, Revisiting the tumorigenesis timeline with a data-driven generative model, *Proc. Natl. Acad. Sci. USA* **117**, 857 (2020).
- [20] L. Vermeulen, E. Morrissey, M. Van Der Heijden, A. M. Nicholson, A. Sottoriva, S. Buczaccki, R. Kemp, S. Tavaré, and D. J. Winton, Defining stem cell dynamics in models of intestinal tumor initiation, *Science* **342**, 995 (2013).
- [21] M. Gerstung, C. Jolly, I. Leshchiner, S. C. Dentro, S. Gonzalez, D. Rosebrock, T. J. Mitchell, Y. Rubanova, P. Anur, K. Yu *et al.*, The evolutionary history of 2,658 cancers, *Nature (London)* **578**, 122 (2020).
- [22] I. Martincorena and P. J. Campbell, Somatic mutation in cancer and normal cells, *Science* **349**, 1483 (2015).
- [23] I. A. Hattori, E. D. Galbraith, N. S. Merleau, T. P. Miettinen, B. M. Smith, and J. A. Shander, The human cell count and

- size distribution, *Proc. Natl. Acad. Sci. USA* **120**, e2303077120 (2023).
- [24] P. T. McHale and A. D. Lander, The protective role of symmetric stem cell division on the accumulation of heritable damage, *PLoS Comput. Biol.* **10**, e1003802 (2014).
- [25] P. Greulich and B. D. Simons, Extreme value statistics of mutation accumulation in renewing cell populations, *Phys. Rev. E* **98**, 050401(R) (2018).
- [26] S. N. Catlin, L. Busque, R. E. Gale, P. Guttorp, and J. L. Abkowitz, The replication rate of human hematopoietic stem cells in vivo, *Blood* **117**, 4460 (2011).
- [27] O. Basak, M. van de Born, J. Korving, J. Beumer, S. van der Elst, J. H. van Es, and H. Clevers, Mapping early fate determination in $Lgr5^+$ crypt stem cells using a novel *Ki67-RFP* allele, *EMBO J.* **33**, 2057 (2014).
- [28] H. Gehart and H. Clevers, Tales from the crypt: New insights into intestinal stem cells, *Nat. Rev. Gastroenterol. Hepatol.* **16**, 19 (2019).
- [29] M. H. Spitzer and G. P. Nolan, Mass cytometry: Single cells, many features, *Cell* **165**, 780 (2016).
- [30] A. Alemany, M. Florescu, C. S. Baron, J. Peterson-Maduro, and A. Van Oudenaarden, Whole-organism clone tracing using single-cell sequencing, *Nature (London)* **556**, 108 (2018).
- [31] L. S. Ludwig, C. A. Lareau, J. C. Ulirsch, E. Christian, C. Muus, L. H. Li, K. Pelka, W. Ge, Y. Oren, A. Brack *et al.*, Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics, *Cell* **176**, 1325 (2019).
- [32] Y. Reizel, N. Chapal-Ilani, R. Adar, S. Itzkovitz, J. Elbaz, Y. E. Maruvka, E. Segev, L. I. Shlush, N. Dekel, and E. Shapiro, Colon stem cell and crypt dynamics exposed by cell lineage reconstruction, *PLoS Genet.* **7**, e1002192 (2011).