Cancer Risk and the Somatic Cell Lineage Tree

Imre Derényi^{a,b,1,2}, Márton C. Demeter^{b,c}, and Gergely J. Szöllősi^{b,c,d,1}

^aDepartment of Biological Physics, Eötvös University, Pázmány P. stny. 1A, H-1117 Budapest, Hungary; ^bMTA-ELTE Statistical and Biological Physics Research Group, Pázmány P. stny. 1A, H-1117 Budapest, Hungary; ^cMTA-ELTE "Lendület" Evolutionary Genomics Research Group, Pázmány P. stny. 1A, H-1117 Budapest, Hungary; ^dEvolutionary Systems Research Group, Centre for Ecological Research, Hungarian Academy of Sciences, 8237 Tihany, Hungary

This manuscript was compiled on July 13, 2020

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 one of the key determinants of how somatic evolution, which can potentially lead to cancer, proceeds. Cancer initiation usually requires the accumulation of a certain number of driver mutations. By mapping the accumulation of driver mutations into a graph theoretical problem, we show that in leading order of the mutation rate the probability of collecting a given number of driver mutations depends only on the distribution of the lineage lengths (irrespective of any other details of the cell lineage tree), and we derive a simple analytical formula for this probability. Our results are crucial in understanding how natural selection can shape the cell lineage trees of multicellular organisms in order to reduce their lifetime risk of cancer. In particular, our results highlight the significance of the longest cell lineages. Our analytical formula also provides a tool to quantify cancer susceptibility in theoretical models of tissue development and maintenance, as well as for empirical data on cell lineage trees.

Somatic Evolution | Cell Lineage Tree | Mutation accumulation | Physics of Cancer

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 (1). It was Armitage and Doll (2, 3) who noticed 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 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 unchallenged (4–6).

All the cells of an organism are derived from a single cell along a single binary tree (as demonstrated in Fig. 1(a)). The extant cells (i.e., the leaf nodes) of a snapshot of the tree taken at a given moment of time are the ones that were either present in the organism at that time or had been lost by then. The internal nodes correspond to the cells that had already gone through a cell division. Similar lineage trees can be drawn for tissues or parts of tissues with the root node being the founder cell of the tissue or tissue part.

It is unclear (and might depend on the type of cancer) if individual driver mutations change the structure of the lineage tree or if they are neutral until a sufficient number is accumulated. The 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 (7).

Fig. 1. Illustration of a cell lineage tree with mutation accumulation. (a) A cell lineage tree 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 T' obtained by making leaf node L in tree T divide (indicated by blue dashed arrows and circles).

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 mutations is acquired (8).

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 all the m required for cancer initiation are accumulated, we explore the fundamental question of how the structure of a cell lineage tree affects its risk of cancer.

Significance Statement

A series of cell divisions starting from a single cell produce and maintain tissues of multicellular organisms. Somatic evolution, including the development of cancer, takes place along the *cell lineage tree* traced out by these cell divisions. A fundamental question in cancer research is how the lifetime risk of cancer depends on the properties of an arbitrary cell lineage tree. Here we show that for small mutation rates (which is the case in reality) the distribution of the lineage lengths alone determines cancer risk, and that this risk can be described by a simple analytical formula. Our results have far-reaching implications not only for cancer research, but also for evolutionary biology in general.

I.D. and G.J.S. designed the study and wrote the paper, I.D. performed the analytical calculations, M.C.D. carried out numeric simulations.

No conflict of interests.

¹I.D. and G.J.S. contributed equally to this work.

²To whom correspondence should be addressed. E-mail: derenyi@elte.hu

1. 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 driver mutations appears, if driver mutations are dropped to the edges uniformly with a frequency of μ per edge? In Fig. 1(a), e.g., the lineage belonging to leaf node 4 has 3 driver mutations (red stars).

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 as outlined below. Every binary tree can be built up by subsequently merging its subtrees at their parent nodes, starting from the leaves. Each internal node (gray nodes in Fig. 1) is used for merging the two subtrees of its child nodes (daughter cells). The last merging at the root node of the entire lineage tree completes the merging sequence.

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

$$\tilde{P}_{\mathcal{A}}^+(\mu, k) = \sum_{j=0}^k \tilde{P}_{\mathcal{A}}(\mu, k-j) p(\mu, j) \quad [1]$$

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

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

where

$$p(\mu, j) = \frac{\mu^j}{j!} e^{-\mu} \quad [3]$$

is the Poisson distribution with parameter μ .

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

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

The initial condition for the merging 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 driver mutations can be obtained as

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

The reason that no closed formula exists for $P_{\mathcal{T}}(\mu, m)$ is the nonlinear nature of Eq. [4]. For small enough values of the driver mutation rate μ (such that $P_{\mathcal{T}}(\mu, m) \ll 1$), however, the dominant contributions of Eq. [4] arise when the summation factors (including the lineages with small numbers of mutations) are close to unity. Note that the very last term can be incorporated into either the first or the second summation. In this case Eqs. [1] to [4] become linear, and $P_{\mathcal{T}}(\mu, m)$ is expected to take the form

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

where index i runs over 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), $F(i, m)$ should depend on i only through D_i . For small enough values of μ , $F(i, m)$ is also expected to be dominated by its leading order term in μ , which is proportional to μ^m , therefore, $F(i, m) \approx \mu^m f(D_i, m)$.

Thus, for small enough driver mutation rates μ the risk of cancer (i.e., probability that a lineage with at least m driver mutations exists in the cell lineage tree) takes 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 up 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,

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

where the first term accounts for the possibility that a lineage with at least m mutations has already existed before adding 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 driver mutations. The probability of this latter event is denoted by $G_{\mathcal{T}}(\mu, m, L)$. In leading order of μ the probability $P_{\mathcal{T}}(\mu, m)$ can be neglected from the second term and $G_{\mathcal{T}}(\mu, m, L)$ can be expressed as

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

where the difference between the braces describes the excess probability conferred by a single elongation and the factor of 2 stands for the number of elongations. Combining the above, $P_{\mathcal{T}'}(\mu, m)$ can be expressed as

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

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 recursive relationship

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

for the functions $f(D, m)$, with the initial values of $f(0, m) = 0$ 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 only its highest order terms in D . For the two highest order terms

$$f(D, m) \approx 2 \frac{D^{m-1}}{(m-1)!} - 3 \frac{D^{m-2}}{(m-2)!}, \quad [15]$$

which is also equivalent to

$$f(D, m) \approx 2 \frac{(D-1.5)^{m-1}}{(m-1)!}. \quad [16]$$

Plugging this approximation back into Eq. [7] leads to our main result that the risk of cancer, i.e., the probability of accumulating at least m driver mutations along a lineage can be well approximated by

$$P_{\mathcal{T}}^{\text{appr}}(\mu, m) = \frac{2\mu^m}{(m-1)!} \sum_{i=1}^L (D_i - 1.5)^{m-1} \quad [17]$$

in leading order of μ and for the two highest orders of the lineage lengths. This approximation is very accurate for small 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 independent units of most human tissues), as demonstrated in Fig. 2 for a series of lineage trees that interpolate between the most skewed tree (linear chain with unit long side branches) and the most compact one (the perfect binary tree) with $L = 2^{16}$ leaf nodes.

2. 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 a unit of tissue. Most tissue units (such as the colonic crypts), however, are sustained a set of stem cells, rather than a single one. Because our main formula [17] 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.

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.

The main message of our work is that assuming rare neutral driver mutations, the risk of cancer depends only on the

Fig. 2. Examples for the probability of cancer in a series of cell lineage trees. The top panel illustrates how a series of lineage trees that interpolate between the most skewed tree (linear chain with unit long side branches, $n_b = 0$) and the most compact 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 driver 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. 17. The approximation is very accurate as long as $P_{\mathcal{T}}^{\text{appr}}(\mu, m) \ll 1$.

distribution of the lineage lengths (irrespective of any other details of the lineage tree) and can be accurately approximated by a simple analytical formula. Because this formula is a sum of relatively high powers of the lineage lengths it is dominated by the longest lineages, which quantitatively explains why the minimization of the longest lineages (e.g., through hierarchical differentiation (9)) is crucial in minimizing cancer risk and somatic evolution in general.

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. 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 (10) or every 2 to 20 months (11) for blood, and every 4 days (12, 13) for the colon) and, therefore, are expected to produce dominant linearly growing lineages.

We foresee applications of our method as more extensive cell lineage tracing data becomes available, both at the organism (14, 15) and the tissue level (16).

ACKNOWLEDGMENTS. The authors are thankful to Ágnes Backhausz for her valuable comments and advice. GJSz received funding from the European Research Council under the European

Union's Horizon 2020 research and innovation programme under grant agreement no. 714774 and the grant GINOP-2.3.2.-15-2016-00057.

- Leonard Nunney, Carlo C. Maley, Matthew Breen, Michael E. Hochberg, and Joshua D. Schiffman. Peto's paradox and the promise of comparative oncology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1673):20140177, 2015. .
- P Armitage and R Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer*, 8:1–12, 1954. .
- Richard Doll. The age distribution of cancer: Implications for models of carcinogenesis. *Journal of the Royal Statistical Society. Series A (General)*, 134(2):133–166, 1971. .
- Suresh H Moolgavkar. Commentary: Fifty years of the multistage model: remarks on a landmark paper. *International Journal of Epidemiology*, 33(6):1182–1183, 08 2004. ISSN 0300-5771. . URL <https://doi.org/10.1093/ije/dyh288>.
- N Neu, B Ploier, and C Ofner. Cardiac myosin-induced myocarditis. heart autoantibodies are not involved in the induction of the disease. *The Journal of Immunology*, 145(12):4094–4100, 1990.
- Philipp M Altrock, Lin L Liu, and Franziska Michor. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, 15(12):730–745, 2015.
- Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J Mitchell, Yulia Rubanova, Pavana Anur, Kaixian Yu, et al. The evolutionary history of 2,658 cancers. *Nature*, 578(7793):122–128, 2020.
- Iñigo Martincorena and Peter J Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015.
- Imre Derényi and Gergely J Szöllösi. Hierarchical tissue organization as a general mechanism to limit the accumulation of somatic mutations. *Nature communications*, 8(1):1–8, 2017.
- Sandra N Catlin, Lambert Busque, Rosemary E Gale, Peter Gutter, and Janis L Abkowitz. The replication rate of human hematopoietic stem cells in vivo. *Blood, The Journal of the American Society of Hematology*, 117(17):4460–4466, 2011.
- Henry Lee-Six, Nina Friesgaard Øbro, Mairi S Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J Osborne, Brian JP Huntly, Inigo Martincorena, Elizabeth Anderson, et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561(7724):473–478, 2018. .
- Onur Basak, Maaik van de Born, Jeroen Korving, Joep Beumer, Stefan van der Elst, Johan H van Es, and Hans Clevers. Mapping early fate determination in lgr5+ crypt stem cells using a novel ki67-rfp allele. *The EMBO journal*, 33(18):2057–2068, 2014.
- Helmuth Gehart and Hans Clevers. Tales from the crypt: new insights into intestinal stem cells. *Nature Reviews Gastroenterology & Hepatology*, 16(1):19–34, 2019.
- Anna Alemany, Maria Florescu, Chloé S Baron, Josi Peterson-Maduro, and Alexander Van Oudenaarden. Whole-organism clone tracing using single-cell sequencing. *Nature*, 556(7699):108–112, 2018.
- Leif S Ludwig, Caleb A Lareau, Jacob C Ulirsch, Elena Christian, Christoph Muus, Lauren H Li, Karin Pelka, Will Ge, Yaara Oren, Alison Brack, et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell*, 176(6):1325–1339, 2019.
- Yitzhak Reizel, Noa Chapal-Ilani, Rivka Adar, Shalev Itzkovitz, Judith Elbaz, Yosef E Maruvka, Elad Segev, Liran I Shlush, Nava Dekel, and Ehud Shapiro. Colon stem cell and crypt dynamics exposed by cell lineage reconstruction. *PLoS Genet*, 7(7):e1002192, 2011.