Trade-off between reducing mutational load and increasing commitment to differentiation determines tissue organization

Márton Demeter
MTA-ELTE "Lendület" Evolutionary Genomics Research Group, Pázmány P. stny. 1A, H-1117 Budapest, Hungary and Department of Biological Physics, Eötvös University, Pázmány P. stny. 1A, H-1117 Budapest, Hungary

Imre Derényi*
Department of Biological Physics, Eötvös University, Pázmány P. stny. 1A, H-1117 Budapest, Hungary and MTA-ELTE Statistical and Biological Physics Research Group, Pázmány P. stny. 1A, H-1117 Budapest, Hungary

Gergely J. Szöllősi†
MTA-ELTE "Lendület" Evolutionary Genomics Research Group, Pázmány P. stny. 1A, H-1117 Budapest, Hungary Department of Biological Physics, Eötvös University, Pázmány P. stny. 1A, H-1117 Budapest, Hungary and Evolutionary Systems Research Group, Centre for Ecological Research, Klebelsberg Kunó u. 3, H-8237 Tihany, Hungary

Abstract

Species-specific differences control cancer risk across orders of magnitude variation in body size and lifespan, e.g., by varying the copy numbers of tumor suppressor genes. It is unclear, however, how different tissues within an organism can control somatic evolution despite being subject to markedly different constraints but sharing the same genome. Hierarchical differentiation, characteristic of self-renewing tissues, can restrain somatic evolution both by limiting divisional load, thereby reducing mutation accumulation, and by increasing the cells' commitment to differentiation, which can "wash out" mutants. Here, we explore the organization of hierarchical tissues that have evolved to limit their lifetime risk of cancer to a tissue-specific level. Analytically estimating the likelihood of cancer, we demonstrate that a trade-off exists between mutation accumulation and the strength of washing out. This result explains the differences in the organization of widely different hierarchically differentiating tissues, such as the colon and the blood.

* derenyi@elte.hu
† ssolo@elte.hu

INTRODUCTION

Cancer is a disease of multicellular organisms, which occurs when individual cells fail to contribute to normal tissue function and instead divide selfishly, resulting in uncontrolled local growth, metastasis, and often death[1]. Multicellular organisms have evolved both species- and tissue-specific mechanisms to suppress somatic evolution and, thus, delay aging and the emergence of cancers. The most striking evidence for the evolution of cancer suppression originates with a prediction of the multistage model[2,3], which was succinctly expressed by Peto[4]. He observed that even though humans are around 1000 times larger than mice and live about 30 times longer, the overall incidence of cancer in the two species is very similar, a sign of evolutionary fine-tuning[5].

Recent data on human tissues[6] support[7,8] theoretical predictions[9–11] that tissues within an individual that are large and rapidly dividing should also exhibit increased cancer resistance. In particular, Tomasetti and Vogelstein[6] gathered information on the lifetime cancer risk and the total number of divisions of healthy self-replicating cells (i.e., stem cells) for 31 different tissues. Their data display a striking tendency: the dependence of cancer incidence on the number of stem cell divisions is sub-linear, i.e., a hundred-fold increase in the number of divisions only results in a ten-fold increase in incidence[8]. This observation indicates that tissues with more stem cell divisions (typically larger ones with rapid turnover, e.g., the colon) are relatively less prone to develop cancer, which by analogy, we may call Peto's paradox for tissues.

There are clear examples of how species-specific differences can control cancer risk, e.g., by increasing the copy number of tumor suppressor genes[12]. It is, however, much less clear how different tissues subject to different constraints but sharing the same genome, can control somatic evolution.

It is well established that hierarchical differentiation, characteristic of self-renewing tissues that must generate a large number of cells during an individual's lifetime and in which cancers typically arise, can restrain somatic evolution. Somatic evolution is kept under control at two levels of the somatic evolutionary process: mutation accumulation and selection. At the level of mutation accumulation, hierarchical organization can limit the mutational burden of maintaining tissues[8,13,14]. At the level of selection, even mutations that provide a significant proliferative advantage can be "washed out" as a result of differentiation, which drives the cells towards the terminally differentiated state and permanent loss of proliferative ability[15–18]. It is, in contrast, not well understood what combination of these mechanisms different tissues employ. It is also not clear to what extent mutational load can be minimized and washing out maximized in a differentiation hierarchy.

To explore these questions, we consider a minimal generic model of hierarchically organized, self-sustaining tissue with cells arranged into $n+1$ hierarchical levels based on their differentiation state (Fig. 1a). The bottom level (level 0) is comprised of tissue-specific stem cells, while higher levels (levels k, where $0 < k < n$) contain progressively more differentiated progenitors, and the top level (level n) corresponds to the terminally differentiated cells. The stem cell level produces differentiated cells at a rate of δ_0, while the differentiation rates of higher levels (denoted by δ_k for level k) are progressively larger. The increasing tendency of the differentiation rates of the progenitor levels ($0 < k < n$) is specified by the level-specific amplification factors $\gamma_k = \delta_k/\delta_{k-1}$, which relate the differentiation rate of a progenitor level to that of the level below it (cf. Fig. 1).

Using the same generic model Derényi and Szöllősi[8] showed that hierarchical organization provides a robust and nearly ideal mechanism to limit the lifetime divisional load (the number of divisions along the longest cell lineages over the lifetime of the individual). Crucially, as long as a sufficient number of progressively faster differentiating cell types are present, the theoretical minimum number of cell divisions can be very closely approached. In optimal self-sustaining differentiation hierarchies N_0 stem cells can produce N terminally differentiated cells during an organism's lifetime with no more than $\log_2(N/N_0)+2$ cell divisions along any lineage. Achieving this optimal reduction in divisional load requires $n_{\mathrm{D}}^* \approx \log_2(N/N_0)$ levels, corresponding to $n_{\mathrm{D}}^* \approx 36$ for the hematopoietic system and $n_{\mathrm{D}}^* \approx 20$ for the epithelial tissue of the colon.

In real tissues, however, we do not expect to see hierarchies that fully minimize the lifetime divisional load. The situation is analogous to DNA-replication fidelity. Current evidence indicates that DNA-replication fidelity is not limited by physiological constraints, but it is set by a balance between selection and genetic drift[19,20]. Peto's paradox for tissues, described above, provides evidence for the existence of a similar "drift-barrier" in the optimality of different tissues in suppressing somatic evolution. This is manifested by the fact that smaller tissues are much less protected against cancer than larger ones (scaled to the same size).

For lifetime divisional load, this implies that it can only be minimized by selection to the extent that the selective advantage achieved is sufficiently large to overcome drift. In the context of the hematopoietic system, for instance, $n \approx 14$ levels are already sufficient to reduce the divisional load to twice the optimal value, but having only 6 levels would correspond to a tenfold increase[8]. Detailed modeling of human hematopoiesis has provided estimates of between 17 and 31 levels[21].

Aside from divisional load, the rate of somatic evolution also depends on the strength of "washing out". Washing out can be quantified by the "proliferative disadvantage" of cells, a quantity (formally defined below) that is proportional to the difference between the per level rate of cell loss (via symmetric differentiation or cell death) and the rate of self-renewal. In healthy tissues, stem cells are lost and self-renewed at the same rate, and they have no proliferative disadvantage. Progenitor cells at higher levels of the hierarchy, however, always have an inherent proliferative disadvantage as some cells arrive by differentiation from below, and self-renewal replenishes only a fraction of removed cells. As a result, the descendants of progenitors are eventually "washed out" of the tissue by cells differentiating from lower levels of the hierarchy.

Here, we explore the organizational properties of hierarchical tissues needed to keep the lifetime risk of cancer below a *thresh-*

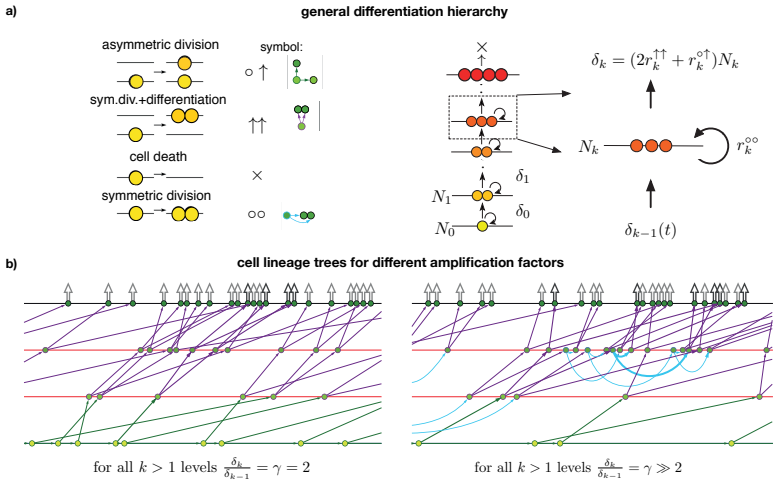

FIG. 1. **Minimal generic model of hierarchically differentiating tissues and corresponding cell lineage trees.** a) cells are organized into $n+1$ hierarchical levels based on their differentiation state. The bottom level (level 0) corresponds to tissue-specific stem cells, higher levels represent progressively differentiated progenitor cells, and the top level (level n) is comprised of terminally differentiated cells. Five microscopic events can occur with a cell: (i) symmetric cell division with differentiation, (ii) asymmetric cell division, (iii) symmetric cell division without differentiation, and (iv) cell death. The symbols used for each event in the mathematical notation and in the cell lineage trees below are also shown. Each level k (except for the terminally differentiated one) provides the next level with newly differentiated cells at a rate δ_k, and self-renewal occurs at a per cell rate of $r_k^{\circ\circ}$. Terminally differentiated cells cannot divide and are destined to wear away (i.e., leave the tissue). The number of cells at level k in fully developed tissue under normal homeostatic conditions is denoted by N_k. b) Cell lineage trees are shown for two different values of a uniform amplification factor $\gamma_k = \gamma$. Given the same rate of production of terminally differentiated cells (terminal tips of the cell linage tree) larger values of γ correspond to a steeper decline in cell division rates towards lower levels, leading to slower dividing stem cells. But at the same time, they also correspond to increasing self-renewal (symmetric cell division events in blue) and, equivalently, decreasing proliferative disadvantage of progenitor cells.

old value, determined by the "drift-barrier", i.e. the balance between selection and genetic drift. We show that under general conditions, the lifetime divisional load and, therefore, the mutational burden increase as the amplification factors decrease. At the same time, the strength of washing out also increases, making it necessary to accumulate either more or stronger driver mutations, or both, to overcome the increasing proliferative disadvantage. As a result, under most conditions, there is a trade-off between mutation accumulation and the proliferative disadvantage of cells, which can lead to a nontrivial evolutionary optimum in the risk of cancer.

RESULTS

The proliferative disadvantage of cells is determined by the amplification factor

We consider a general differentiation hierarchy $\mathcal{H}$ described by the per cell rates of symmetric differentiation ($r_k^{\uparrow\uparrow}$), asymmetric differentiation ($r_k^{\circ\uparrow}$), symmetric cell division ($r_k^{\circ\circ}$), and cell death ($r_k^{\times}$) for each level (k). Using this notation the per cell rate of net cell production of level k can be expressed as

$$R_k = r_k^{\uparrow\uparrow} + r_k^{\circ\uparrow} + r_k^{\circ\circ} - r_k^{\times}, \tag{1}$$

while the per cell rate at which cells are depleted ("washed out") from their level is

$$W_k = r_k^{\uparrow\uparrow} - r_k^{\circ\circ} + r_k^{\times}. \tag{2}$$

Their dimensionless ratio

$$\pi_k = \frac{W_k}{R_k} \tag{3}$$

can be defined as the proliferative disadvantage of cells at level k.

The total rate at which differentiated cells are produced by level k is

$$\delta_k = \left(2r_k^{\uparrow\uparrow} + r_k^{\circ\uparrow}\right) N_k, \tag{4}$$

where N_k is the number of cells of level k. Homeostasis implies that on any particular level (except for the stem cell level) the rate at which cells arrive via differentiation is equal to the rate at which cells are depleted:

$$\delta_{k-1} = W_k N_k. \tag{5}$$

Because the stem cell level do not receive cells from other levels, its depletion rate must be zero ($W_0 = 0$) in homeostasis. Formally this is equivalent to setting $\delta_{-1} = 0$. Subtracting equation (5) from equation (4):

$$\delta_k - \delta_{k-1} = \left(r_k^{\uparrow\uparrow} + r_k^{\circ\uparrow} + r_k^{\circ\circ} - r_k^{\times}\right) N_k = R_k N_k \tag{6}$$

illustrates that under homeostasis the total net cell production rate of level k accounts for the difference between the outgoing and incoming cell differentiation rates. Dividing the above expression for δ_{k-1} by that for $\delta_k - \delta_{k-1}$ leads to a simple relationship

$$\pi_k = \frac{1}{\gamma_k - 1} \tag{7}$$

between the proliferative disadvantage π_k and amplification factor $\gamma_k = \delta_k/\delta_{k-1}$ of each progenitor level $0 < k < n$.

Necessary conditions for cancer

In the context of our hierarchical tissue model, carcinogenesis occurs when a mutant population starts to grow exponentially. More formally, on level $k < n$ the necessary condition is that the proliferative disadvantage of a mutant cell becomes negative (i.e., it acquires a proliferative advantage). This can occur as a result of accumulating "driver" mutations that (i) increase the rate of self-proliferation ($r_k^{\circ\circ}$) or (ii) decrease the rates of symmetric differentiation ($r_k^{\uparrow\uparrow}$) or cell death ($r_k^{\times}$). The terminally differentiated level ($k = n$), where only cell death is assumed to occur, but not cell division, is unaffected by driver mutations.

Here, we assume that driver mutations increase $r_k^{\circ\circ}$ or decrease either $r_k^{\uparrow\uparrow}$ or $r_k^{\times}$ by a fraction s of the net cell production rate R_k. In this case on levels $0 < k < n$ carcinogenesis requires the accumulation of

$$d_k(s, \gamma_k) = \left\lceil \frac{\pi_k}{s} \right\rceil = \left\lceil \frac{1}{s(\gamma_k - 1)} \right\rceil \tag{8}$$

driver mutations, where $\lceil x \rceil$ denotes the ceiling function, the value of which corresponds to the smallest integer that is equal to or larger then x.

The stem cell level ($k = 0$) at the very bottom of the hierarchy, which must completely renew itself, constitutes an important exception. As a result of the necessity of complete self-renewal, stem cells must have a proliferative disadvantage of zero, i.e., $\pi_0 = 0$. Formally, this implies that even a single driver mutation will, if it is not lost, lead to an exponential, albeit potentially very slow expansion of the stem cell pool. The differentiated descendants of these mutant stem cells, however, will still be at a proliferative disadvantage. As a result, these mutants will be washed out from higher levels $0 < k$ of the hierarchy, unless a sufficient number of drivers ($\geq \pi_k/s$) is accumulated to overcome the proliferative disadvantage π_k.

In the following we make two simplifying assumptions: (i) we assume all amplification factors are equal, i.e., $\gamma_k = \gamma$, which implies that $\pi_k = \pi$ and $d_k(s, \gamma_k) = d(s, \gamma) = \lceil 1/s(\gamma - 1) \rceil$ for all progenitor levels $0 < k < n$ (cf. Fig. 2a), and (ii) we assume that drivers are neutral until a sufficient number $m \geq d(s, \gamma)$ is accumulated for carcinogenesis to occur at progenitor levels. This later assumption also means that expansion of the stem cell pool is considered to be negligible for mutants with $m < d(s, \gamma)$ driver mutations.

The assumption of uniform amplification factors, which corresponds to division rates increasing exponentially along the hierarchy[13,14], is motivated by both mathematical convince and the optimality of identical γ_k values in reducing the lifetime divisional load[8].

The assumption that divers are neutral until a sufficient number is accumulated, while clearly not true in general, is consistent with the fact that the majority of cancers arise without a histologically discernible premalignant phase and recent timing analyses that suggest that driver mutations often precede diagnosis by many years, if not decades[22]. These observations indicate strong cooperation between driver mutations, suggesting that major histological changes may not take place until the full repertoire of mutations is acquired[23].

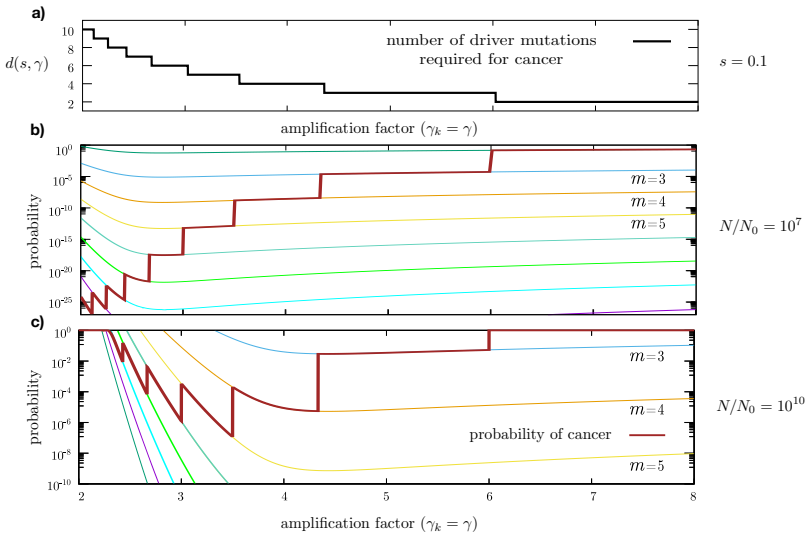

FIG. 2. **Trade-off between mutation accumulation and proliferative disadvantage.** a) the number of driver mutations $d(s,\gamma)$ necessary for carcinogenesis for $s = 0.1$ as a function of uniform amplification factors $\gamma_k = \gamma$. b) and c) Two hierarchies differing only in the number of terminally differentiated cells produced per stem cell during the lifetime of the tissue, respectively, $N/N_0 = 10^7$ and $N/N_0 = 10^{10}$, are shown. The probability $P(\mu, m, \mathcal{H})$ of accumulating $m = 2, 3, \ldots$ mutations are shown with color lines for both. In addition, the probability of accumulating a sufficient number of mutations for carcinogenesis to occur, $P(\mu, d(s,\gamma), \mathcal{H})$ is shown with a thick red line for each. In both plots $N_0 = 1$, $n = 15$, $s = 0.1$ and $\mu = 10^{-5}$.

The probability of accumulating m mutations

Derényi et al.[24] recently showed that the probability of accumulating m neutral mutations on an arbitrary cell lineage tree $\mathcal{T}$ with L leaves (e.g., the lineage tree in Fig. 1b with its terminally differentiated cells as leaves), each of which has undergone $D_1(\mathcal{T}), \ldots, D_L(\mathcal{T})$ divisions (referred to as their divisional loads) can, as long as this probability is much smaller than unity, be very accurately approximated as

$$P(\mu, m, \mathcal{T}) \approx \frac{2\mu^m}{(m-1)!} \sum_{i=1}^{L} [D_i(\mathcal{T}) - 1.5]^{m-1}. \tag{9}$$

This formula, because it is a simple sum over all the leaves, is also valid for a collection of trees, such as those generated by the initial set of stem cells of a tissue. For the collection of lineage trees $\mathcal{H}$ of our hierarchical tissue model the expected lineage lengths (divisional loads) of terminally differentiated cells produced at time t (and destined to wear away, as indicated by white arrows in Fig. 1b can be expressed as

$$D(t, \mathcal{H}) = \frac{\delta_0}{N_0}(t - \tau_{\mathrm{tr}}) + \sum_{l=1}^{n-1} (\gamma_l - 1) + 1, \tag{10}$$

where τ_{tr} is the transient time necessary for the initial build up of the differentiation hierarchy from the stem cells[8].

Combining equations (9) and (10) we can derive the probability $P(\mu, m, \mathcal{H})$ of accumulating m mutations during the lifetime of a tissue hierarchy $\mathcal{H}$ by replacing the linage lengths $D_1(\mathcal{T}), \ldots, D(\mathcal{T})_L$ of terminally differentiated cells produced at a constant rate δ_n with their expected values in time $D(t, \mathcal{H})$:

$$\begin{aligned} P(t_{\mathrm{life}}, \mu, m, \mathcal{H}) &\approx \frac{2\mu^m}{(m-1)!} \int_{\tau_{\mathrm{tr}}}^{t_{\mathrm{life}}} [D(t, \mathcal{H}) - 1.5]^{m-1} \, \delta_n \mathrm{d}t \\ &= N_0 \left(\frac{2\mu^m}{m!} \right) \{ [D(t_{\mathrm{life}}, \mathcal{H}) - 1.5]^m - [D(\tau_{\mathrm{tr}}, \mathcal{H}) - 1.5]^m \} \prod_{l=1}^{n-1} \gamma_l, \end{aligned} \tag{11}$$

where $t_{\rm life} = \tau_{\rm tr} + N/\delta_n$ is the lifetime of the tissue, during which N terminally differentiated cells are generated, while

$$D(\tau_{\rm tr}, \mathcal{H}) = \sum_{l=1}^{n-1}(\gamma_l - 1) + 1, \quad \text{and} \tag{12}$$

$$D(t_{\rm life}, \mathcal{H}) = \frac{N}{N_0}\prod_{l=1}^{n-1}\frac{1}{\gamma_l} + \sum_{l=1}^{n-1}(\gamma_l - 1) + 1 \tag{13}$$

are respectively, the transient and lifetime divisional loads of the tissue.

From the above we can see that aside from the mutation rate μ and the number of mutations m, $P(\mu, m, \mathcal{H})$ also depends on parameters of the tissue hierarchy $\mathcal{H}$. In particular, the number of hierarchical levels n, the amplification factors γ_k (with $\gamma_k = \gamma$ for a uniform hierarchy), the number of stem cells N_0 and the number of terminally differentiated cells produced per stem cell N/N_0 during the lifetime of the tissue. All other parameters being the same $P(\mu, m, \mathcal{H})$ is proportional to the number of stem cells, i.e., tissue size, and increases approximately with the m-th power of N/N_0. It is also apparent that $P(\mu, m, \mathcal{H})$ behaves similarly to the lifetime divisional load $D(t_{\rm life}, \mathcal{H})$: as long as $n \leq \log_2(N/N_0)$ it has a nontrivial minimum close to that of the lifetime divisional load at $\gamma^*_{\rm D} = (N/N_0)^{1/n}$[8]and its value at this minimum decreases with increasing n.

Trade-off between mutation accumulation and proliferative disadvantage

The above results allow us to derive the risk of cancer during the lifetime of a tissue hierarchy $\mathcal{H}$ by calculating the probability of accumulating $m = d(s, \gamma)$ mutations as

$$P_{\rm cancer}(t_{\rm life}, s, \mu, n, \gamma, N_0, N/N_0) = P(t_{\rm life}, \mu, d(s, \gamma), \mathcal{H}). \tag{14}$$

This is one of our main results.

Similar to $P(\mu, m, \mathcal{H})$, all other parameters being the same, $P_{\rm cancer}(s, \mu, n, \gamma, N_0, N/N_0)$ is proportional to N_0, increases with increasing N/N_0 and, for realistic tissues, decreases with increasing n. As illustrated by the red lines in Figs. 2 b and c, however, in contrast to the probability of accumulating a fixed m number of mutations the minimum of the probability of cancer as a function of the amplification factor γ is not, in general, close to the value $\gamma^*_{\rm D} = (N/N_0)^{1/n}$ that minimizes the lifetime divisional load of the tissue. Instead, the amplification factor $\gamma^*_{\rm cancer}$ that minimizes the probability of cancer is determined by a trade-off between the proliferative disadvantage along the hierarchy, reflected in increasing $d(s, \gamma)$ for decreasing γ as shown in Fig. 2 a, and mutation accumulation, which is minimized near $\gamma^*_{\rm D} = (N/N_0)^{1/n}$[8], as illustrated by the color lines in Figs. 2 b and c.

Only for a fully optimal hierarchy with $n^*_{\rm D} = \log_2(N/N_0)$ levels, where $\gamma^*_{\rm D} = 2$ does the minimum of the lifetime divisional load coincide with maximal proliferative disadvantage along the hierarchy.

The organization of hierarchical tissues that have evolved to limit somatic evolution

To explore the implications of this trade-off for real-life tissues that have evolved to keep the lifetime risk of cancer below a threshold value set by the "drift-barrier", we consider two human tissues the hierarchical organization of which are best understood. The hematopoietic system, where $N_0 = 10^4$ stem cells produce approximately $N = 10^{15}$ terminally differentiated cells, and the colon, where 10^8 stem cells produce 10^{14} terminally differentiated cells during a person's lifetime.

A fully optimal hierarchy for the hematopoietic system that minimizes the lifetime divisional load, while at the same time maximizing the proliferative disadvantage along the hierarchy, would require $n^*_{\rm D} \approx 36$ hierarchical levels, while the colon would require $n^*_{\rm D} \approx 20$. In addition, stem cells at the bottom of both hierarchies would only divide twice during an entire lifetime[8].

Detailed modeling of human hematopoiesis has provided estimates of between 17 and 31 hierarchical levels[21], and long term hematopetic stem cells are thought to divide at most a few times a year (estimates of every 25 to 50 weeks[25] and every 2 to 20 months[26] have been proposed). The number of hierarchical levels in colonic crypts is less clear, but stem cells are known to divide approximately every 4 days[27,28].

From these data it is obvious that neither tissue appears to possess a fully optimal hierarchy, despite evidence that large and rapidly dividing human tissues have evolved increased cancer resistance[6–8]. This observation is consistent with the existence of a "drift-barrier", i.e., that selection can only optimize tissues to the extent that the selective advantage achieved is sufficiently large to overcome genetic drift.

To model the existence of a drift-barrier we consider the least complex tissue, i.e., the one with the smallest number of hierarchical levels, that can keep the probability of cancer below a threshold value. We consider the number of stem cells N_0 and

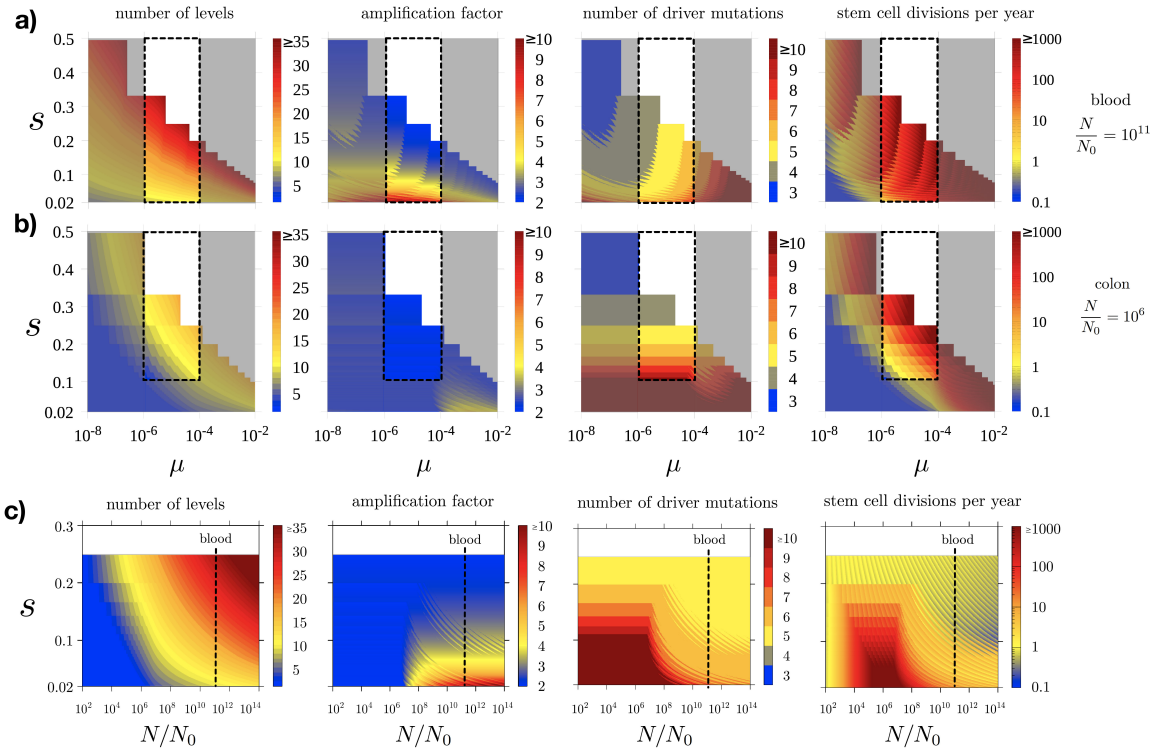

FIG. 3. **The organization of hierarchical tissues that have evolved to limit somatic evolution**. We consider the least complex tissue, i.e., the one with the smallest number of hierarchical levels that can keep the lifetime risk of cancer below a specific value set by the "drift-barrier" (see Methods for details). **a)** For the hematopoietic system, where $N_0 = 10^4$ stem cells produce approximately $N = 10^{15}$ mature cells during an individual's lifetime we consider a threshold lifetime risk of about 2%. **b)** For the colon, where $N_0 = 10^8$ and $N = 10^{14}$ we consider a threshold of about 4%. For both tissues, the dashed lines indicate realistic limits for μ and s based on the literature, as discussed in the main text. **c)** We keep $N = 10^{15}$, $\mu = 10^{-5}$ and the maximum acceptable lifetime risk of 2% fixed, but change the ratio N/N_0.

the number of terminally differentiated cells produced N as fixed by external constraints, and vary the rate of driver mutations per cell division μ and their strength s.

We determined the minimum number of levels n and the corresponding uniform amplification factor γ necessary to keep the lifetime risk of cancer below the threshold value of 2% for cancers of the hematopoietic system and about 4% for colorectal cancer[29] (see Methods for details). In Figs. 3 a and b we show results for the number of levels n and the amplification factor γ, together with the number of drivers (which is determined by s and γ, cf. equation (8)) and the stem cell division time (determined by n, γ, and N_0).

Estimates for the rate of driver mutations per cell division[30–32] vary over $\mu = 10^{-6}$ to 10^{-4} reflecting, potentially tissue specific, uncertainty in both the number of mutational targets and the somatic mutation rate per cell division. For the average selective advantage of driver mutations estimates range from $s \approx 10^{-3}$[30] to $s > 10^{-1}$[31,33]. For the colon empirical measurements[34] and theoretical arguments suggest that $s < 10^{-1}$ in unlikely, but for blood the entire range of values is plausible.

The unshaded areas bounded by the dashed lines in Figs. 3 a and b show the ranges of μ and s values consistent with the above estimates. For the hematopoietic system we find that the number of hierarchical levels ranges between $n = 15$ and 30, and the amplification factor between $\gamma = 2$ and 6, broadly consistent with estimates[21] based on available *in vivo* data. The number of drivers falls between $d = 4$ and 6, while stem cells divide a few times per year. For the colon, we find a significantly lower number of levels between $n = 5$ and 15 and an amplification factor of $\gamma = 2$, corresponding to maximal washing out, again consistent with our understanding of the organization of the colorectal epithelium[35–37].

The organization of hierarchical tissues based on age-incidence data

Above we postulate that the complexity of hierarchical tissue organization is set by a drift-barrier effect, i.e., it corresponds to the smallest number of hierarchical levels that can limit the lifetime risk of cancer to below a tissue specific value (2% and

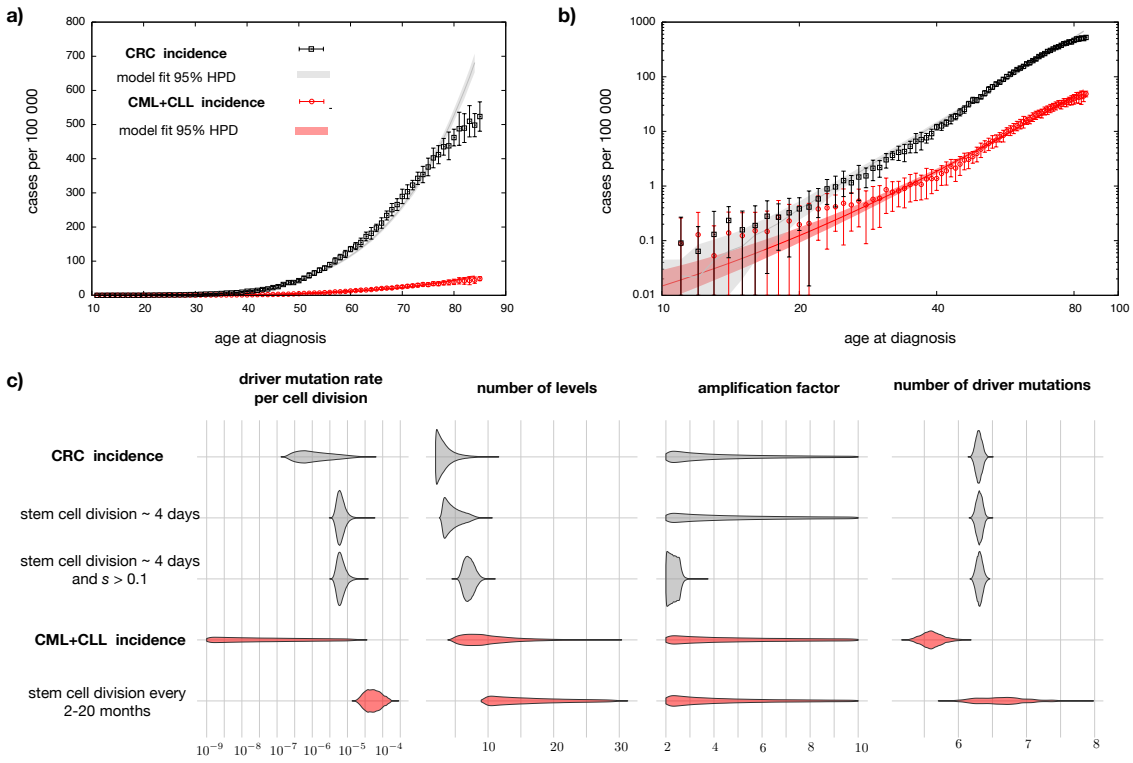

FIG. 4. **The organization of hierarchical tissues based on age-incidence data**. Model fit for SEER age-incidence data on linear a), and logarithmic scales b). Colorectal cancer incidence is shown in black, and chronic leukemias in red, along with the 95% highest posterior density of the model fit. c) Distribution of model parameters under different prior assumptions. "CRC incidence" and "CML+CLL incidence" rows correspond to no additional prior information (fits shown in parts a and b), further rows correspond to results under informative priors on stem cell division rate and for CRC on the selection strength of driver mutations.

4% for cancers of blood and the colon, respectively). In addition to tissue specific lifetime risk, however, data on the incidence of different cancers at different ages is readily available and has been used in previous studies[30–32] for model validation and parameter inference.

To establish our model's ability to reproduce age-incidence data and, at the same time, test the assumption that the complexity of hierarchical tissue organization, in terms of the number of hierarchical levels, is determined by a drift-barrier effect, we used age-incidence data from the SEER database[29]. In particular, we used cancer incidence for different ages given by equation 14 together with a heuristic Bayesian model-fitting approach (see Methods for details) to estimate the posterior distribution of the parameters s, μ, n and γ that best fit the age-incidence measurements.

As shown in Figs. 4a and b, good fit was achieved for a broad range of parameters (Fig. 4c). Setting informative priors on the stem cell division rate (approximately 4 days for CRC[27,28] and 2 to 20 months for CML+CLL[26]), as well as on the selection coefficient of driver mutations in CRC (being larger than 0.1[18,34]) narrowed the range of parameter values.

In addition, while we did not make any assumptions about the number of hierarchical levels n while fitting age-incidence, we none-the-less find a lower posterior mean of 3.4 (95% CI: $(2.07, 6.3)$) for n based on age-incidence for CRC, compared to 10.4 (95% CI: $(5.3, 19.4)$) based on CML+CLL age-incidence data. If we also include information on driver mutation strength s (for CRC) and the stem cell division rate (for both CRC and CML+CLL) the estimates of n become sharper and the difference between tissues more pronounced with 7.0 (95% CI: $(5.8, 8.4)$) and 16.7 (95% CI: $(9.9, 27.3)$) for CRC and CML+CLL, respectively. More generally, the distribution of model parameters shown in Fig. 4c is broadly consistent with the results presented in Fig. 3.

Interestingly, age-incidence based parameter estimates provide relatively well-defined driver mutation rates, especially when informative priors are used on stem cell division rates. For colorectal cancer based on incidence alone the posterior mean driver mutation rate per cell division is 1.3×10^{-6} (95% CI: $(2.6 \times 10^{-7}, 1.1 \times 10^{-5})$), while with an informative prior on stem cell division rates it is 6.3×10^{-6} (95% CI: $(4.6 \times 10^{-6}, 1.1 \times 10^{-5})$). For chronic leukemias based on incidence alone the driver mutation rates are consistent with a very broad range with 95% CI $(1.4 \times 10^{-9}, 1.0 \times 10^{-5})$, constraining the stem cell division rate to be between 2 and 20 months, however, leads to a relatively well defined posterior with a mean of 5.9×10^{-5} (95% CI: $(2.3 \times 10^{-5}, 2.5 \times 10^{-4})$).

I. DISCUSSION

Animals have been evolving mechanisms to suppress cancer ever since the origin of multicellularity. The existence of species level adaptations, as exemplified by the near irrelevance of mammalian body size and lifespan to lifelong cancer risk, has been clear for several decades[4,5]. The realization that rapidly renewing tissues of long-lived animals, such as humans, must also have evolved tissue specific protective mechanisms also dates back several decades[9,10]. Evidence for tissue specific adaptations is, however, more recent[6–8].

In the above we developed an analytical approximation for the lifetime risk of cancer in a hierarchically differentiating self-renewing tissue based on recent mathematical result for estimating the probability of mutations on arbitrary cell lineage trees[24]. Using this result we determine the organizational properties of hierarchical tissues that have evolved to limit somatic evolution by keeping the lifetime risk of cancer below a maximum acceptable value. We find that the optimal tissue organization is determined by a trade-off between two competing mechanism, reduced mutation accumulation[8], and increased "washing out" through the progression of increasingly differentiated cell types[15].

We show that such a trade-off exists as long as differentiation hierarchies are not fully optimal in reducing divisional load. This is likely the case in most tissues of most species, as fully optimal hierarchies require complex hierarchies with a large number of levels incompatible with current empirical evidence[6,8]. Such complex hierarchies are also unlikely to have evolved according to the "drift-barrier" hypothesis[19,20,38] which, in contrast to the view that natural selection fine-tunes every aspect of organisms, predicts that genetic drift, resulting from finite population sizes, can limit the power of selection and constrain the degree to which phenotypes can be optimized by selection.

The trade-off occurs in the tempo of increase of the cell production rate along the differentiation hierarchy, which we parametrize by the amplification factor. The amplification factor corresponds to the ratio of the rate at which adjacent levels produce differentiated cells, i.e., hierarchies where the acceleration of differentiation rates along the hierarchy is faster have a higher amplification factor. As show in Fig. 2, tissues with a smaller amplification factor experience increased mutational burden, however, at the same time exhibit increased washing out, resulting in a trade-off between the two.

We demonstrate that based on the lifetime number of the terminally differentiated cells produced per stem cell, our theoretical description (Fig. 3a and b) provides realistic predictions for the organization of the human hematopoietic system and the epithelial tissue of the colon. In particular, the hematopoietic differentiation hierarchy is predicted to have a relatively larger number of levels with a relatively high amplification rate ensuring low mutational load from cell divisions, in agreement with previous results[21]. The colorectal epithelium, the paradigmatic model of differentiation induced proliferative disadvantage[15,18], in contrast, has a near minimal amplification factor and few differentiation levels ensuring strong washing out and requiring a fast stem cell turnover rate in agreement with experimental data[27,28]. Our results have also been validated by explicitly fitting the age-incidence data for both tissues (Fig. 4).

In summary, based on the trade-off between mutation accumulation and washing-out we provide a general analytical tool for predicting the organization (including the cell differentiation rates and the number of hierarchical levels) of tissues of various sizes (N_0 and N) based on the rate (μ) and strength (s) of driver mutations. An immediate consequence of our predictions is the explanation of the surprisingly fast turnover rate of the stems cells of the colonic crypts.

II. METHODS

Calculating the minimum number of levels n and the corresponding amplification factor

For specific values of N, N_0, μ and s, to determine the minimum number of levels n and the corresponding uniform amplification factor γ, starting with $n = 1$ we determine the minimum of the lifetime cancer risk (defined by equation (14)) as a function of γ. If this minimum is above the threshold value of 2% for cancers of the hematopoietic system and about 4% for colorectal cancer[29], we increase n by one, otherwise we stop the procedure.

Fitting age-incidence data

To estimate the posterior distribution of tissue hierarchies consistent with the SEER data[29] we used a heuristic composite likelihood. To define the likelihood we calculated for ages $t_{\rm age} = 10$ to $t_{\rm age} = 85$ years, per year of age, the mean $m(t_{\rm age})$ and variance $\sigma^2(t_{\rm age})$ of SEER incidence data over time (see details below). The likelihood of the age-incidence data was then defined as the product per year over ages $t_{\rm age} = 10$ to $t_{\rm age} = 85$ of the probability of the incidence calculated using equation

(14):

$$\mathcal{L}\left(s,\mu,n,\gamma,N_0,N/N_0 \mid \text{SEER data}\right) = \prod_{t_{\text{age}}=10}^{85} f\left[P_{\text{cancer}}(t_{\text{age}},s,\mu,n,\gamma,N_0,N/N_0) \mid m(t_{\text{age}}),\ \sigma^2(t_{\text{age}})\right], \tag{15}$$

where $f[x \mid m,\sigma^2]$ denotes the probability density function of the normal distribution with mean m and variance σ^2.

For N and N_0 we used the same values as in Fig. 3; for μ we specified a non-informative uniform prior over $10^{-9} < \mu < 1$; for s either a non-informative uniform prior over $0 < s < 10$ or an informative prior (see main text) uniform over $0.1 < s < 10$, for n and γ either a non-informative uniform prior over $2 \leq n < 30$ and a non-informative uniform prior over $2 < \gamma < 10$, or an informative prior reflecting stem cell division rate r_0, which can be expressed as $r_0 = \delta_0/N_0 = N/t_{\text{life}}\gamma^{1-n}/N_0$ (see main text).

We estimated the mean and variance of the incidence across the year of diagnosis for colorectal cancer (CRC, defined as all cancers with SEER site codes C18, C19 or C20) and chronic myeloid and lymphocytic leukemias (CML and CLL, defined as morphology codes T9823, T9863 and T9876). For CRC we took years 1973-1985, as a significant shift was apparent in more recent years, possibly due to wider-spread screening. For the incidence of CML+CLL we took all years 1973-2011.

To generate samples from the posterior we used a custom Markov chain Monte Carlo sampler.

III. ACKNOWLEDGEMENT

GJSz received funding from the European Research Council under the European Unions Horizon 2020 research and innovation programme under grant agreement no. 714774 and the grant GINOP-2.3.2.-15-2016-00057.

1. Nunney, L., Maley, C. C., Breen, M., Hochberg, M. E. & Schiffman, J. D. Peto's paradox and the promise of comparative oncology. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20140177 (2015).
2. Armitage, P. & Doll, R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer* **8**, 1–12 (1954).
3. Doll, R. The age distribution of cancer: Implications for models of carcinogenesisy. *Journal of the Royal Statistical Society. Series A (General)* **134**, 133–166 (1971).
4. Peto, R. Epidemiology, multistage models, and short-term mutagenicity tests. *Cold Spring Harbor Conferences on Cell Proliferation* **4**, 1403–1428 (1977).
5. Peto, R. Quantitative implications of the approximate irrelevance of mammalian body size and lifespan to lifelong cancer risk. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20150198 (2015).
6. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81 (2015).
7. Noble, R., Kaltz, O. & Hochberg, M. E. Peto's paradox and human cancers. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20150104 (2015).
8. Derényi, I. & Szöllősi, G. J. Hierarchical tissue organization as a general mechanism to limit the accumulation of somatic mutations. *Nature communications* **8**, 1–8 (2017).
9. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197–200 (1975).
10. Albanes, D. & Winick, M. Are cell number and cell proliferation risk factors for cancer? *JNCI: Journal of the National Cancer Institute* **80**, 772–775 (1988).
11. Nunney, L. Lineage selection and the evolution of multistage carcinogenesis. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **266**, 493–498 (1999).
12. Sulak, M. *et al.* Tp53 copy number expansion is associated with the evolution of increased body size and an enhanced dna damage response in elephants. *Elife* **5**, e11994 (2016).
13. Werner, B., Dingli, D., Lenaerts, T., Pacheco, J. M. & Traulsen, A. Dynamics of mutant cells in hierarchical organized tissues. *PLoS computational biology* **7** (2011).
14. Werner, B., Dingli, D. & Traulsen, A. A deterministic model for the occurrence and dynamics of multiple mutations in hierarchically organized tissues. *Journal of The Royal Society Interface* **10**, 20130349 (2013).
15. Nowak, M. A., Michor, F. & Iwasa, Y. The linear process of somatic evolution. *Proceedings of the national academy of sciences* **100**, 14966–14969 (2003).
16. Pepper, J. W., Sprouffske, K. & Maley, C. C. Animal cell differentiation patterns suppress somatic evolution. *PLoS computational biology* **3** (2007).
17. Hindersin, L., Werner, B., Dingli, D. & Traulsen, A. Should tissue structure suppress or amplify selection to minimize cancer risk? *Biology direct* **11**, 41 (2016).
18. Grajzel, D., Derényi, I. & Szöllősi, G. J. A compartment size-dependent selective threshold limits mutation accumulation in hierarchical tissues. *Proceedings of the National Academy of Sciences* (2020).

19. Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G. & Lynch, M. Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences* **109**, 18488–18492 (2012).
20. Burgess, D. J. Knowing when to stop. *Nature Reviews Genetics* **17**, 501 (2016).
21. Dingli, D., Traulsen, A. & Pacheco, J. M. Compartmental architecture and dynamics of hematopoiesis. *PloS one* **2** (2007).
22. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
23. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
24. Derényi, I., Demeter, M.Cs. & Szöllősi, G. J. Cancer Risk and the Somatic Cell Lineage Tree *bioRXiv* **https://doi.org/10.1101/2020.07.13.201004** , 201004 (2020).
25. Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P. & Abkowitz, J. L. The replication rate of human hematopoietic stem cells in vivo. *Blood, The Journal of the American Society of Hematology* **117**, 4460–4466 (2011).
26. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
27. Basak, O. *et al.* Mapping early fate determination in lgr5+ crypt stem cells using a novel ki67-rfp allele. *The EMBO journal* **33**, 2057–2068 (2014).
28. Gehart, H. & Clevers, H. Tales from the crypt: new insights into intestinal stem cells. *Nature Reviews Gastroenterology & Hepatology* **16**, 19–34 (2019).
29. Howlader, N. *et al.* Seer cancer statistics review, 1975–2012, national cancer institute. bethesda, md (2015).
30. Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences* **107**, 18545–18550 (2010).
31. McFarland, C. D., Mirny, L. A. & Korolev, K. S. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proceedings of the National Academy of Sciences* **111**, 15138–15143 (2014).
32. Lahouel, K. *et al.* Revisiting the tumorigenesis timeline with a data-driven generative model. *Proceedings of the National Academy of Sciences* **117**, 857–864 (2020).
33. Williams, M. J. *et al.* Quantification of subclonal selection in cancer from bulk sequencing data. *Nature genetics* **50**, 895–903 (2018).
34. Vermeulen, L. *et al.* Defining stem cell dynamics in models of intestinal tumor initiation. *Science* **342**, 995–998 (2013).
35. Nicolas, P., Kim, K.-M., Shibata, D. & Tavaré, S. The stem cell population of the human colon crypt: analysis via methylation patterns. *PLoS computational biology* **3** (2007).
36. Bravo, R. & Axelrod, D. E. A calibrated agent-based computer model of stochastic cell dynamics in normal human colon crypts useful for in silico experiments. *Theoretical Biology and Medical Modelling* **10**, 66 (2013).
37. Ritsma, L. *et al.* Intestinal crypt homeostasis revealed at single-stem-cell level by in vivo live imaging. *Nature* **507**, 362–365 (2014).
38. Lynch, M. & Walsh, B. *The origins of genome architecture*, vol. 98 (Sinauer Associates Sunderland, MA, 2007).