Extreme value statistics of mutation accumulation in renewing cell populations

Philip Greulich[1, 2] and Benjamin D. Simons[3, 4, 5]

[1] *Mathematical Sciences, University of Southampton, Southampton, UK*
[2] *Insititute for Life Sciences, University of Southampton, Southampton, UK*
[3] *Cavendish Laboratory, University of Cambridge, Cambridge, UK*
[4] *Wellcome Trust-CRUK Gurdon Institute, University of Cambridge, Cambridge, UK*
[5] *Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK*

(Dated: September 28, 2018)

The emergence of a predominant phenotype within a cell population is often triggered by the chance accumulation of a sequence of rare genomic DNA mutations within a single cell. For example, tumors may be initiated by a single cell in which multiple mutations cooperate to bypass their natural defense mechanism. The risk of such an event is thus determined by the extremal accumulation of mutations across tissue cells. To address this risk, here we study the statistics of the maximum mutation numbers in a generic, but tested, model of a renewing cell population. By drawing an analogy between the genealogy of a cell population and the theory of branching random walks, we obtain analytical estimates for the probability of exceeding a threshold number of mutations to trigger a proliferative advantage of a cell over its neighbors, and determine how the statistical distribution of maximum mutation numbers scales with age and cell population size.

PACS numbers: x

Over the lifetime of an organism, its constituent cells continuously accumulate DNA mutations, which can affect the pathways that control cell proliferation and survival. Yet, due to gene multiplicity or functional redundancy [1–6], disruptions of such pathways may often be tolerated within a homeostatic tissue cell population. Evidence from studies of the cancer genome [3, 7, 8] suggest that the accumulation of a critical number of individually "neutral" or "near-neutral" mutations may, in many cases, be necessary to trigger a selective survival advantage on cycling cells – a process known as "genetic" or "epistatic buffering" [3–5, 9–11]. The resulting selective advantage of mutated cells confers clonal dominance [12, 13] which, if sustained long-term [14], constitutes a potential tumor-initiating event. Crucially, since one cell within a tissue cell population is sufficient to trigger such an event [15], the risk of this occurring is naturally dominated by the statistics of rare events – in this case the extreme accumulation of a multiplicity of mutations within a cell, rather than by the cell population averages reaching some level of mutational burden. The statistics of extreme mutation accumulation represents, therefore, a question of both academic and practical interest.

The normal maintenance of adult renewing tissues, such as the skin epidermis or the gut epithelium, relies on the activity of stem cells, which divide to replenish functional differentiated cells lost through exhaustion or cell death [16, 17]. Alongside asymmetric divisions, which leave the stem cell number unchanged [18, 19], in many tissues, frequent stochastic stem cell loss is compensated by symmetrical division of neighbors so that the stem cell number is maintained only at the population level [20, 21]. It is on this background, that these long-lived cells acquire mutations that may lead, in turn, to a selective growth advantage.

Historically, efforts to model how the serial acquisition of mutations can initiate a selective advantage and drive the expansion of asexual populations – such as tumor cells – have either considered single lineages, neglecting the potential effects of clonal competition [22–25], or have focused on cases of non-neutral individual mutations [26–30]. Apart from studies of specialized models that consider a "double-hit" condition for tumor initiation [24, 31–33], nonetheless, the competition of multiple mutant lineages (clonal interference [28, 29, 34]) and the impact of stochastic cell fate dynamics on the statistics of rare mutational signatures have remained under-explored. Only recently, numerical studies of mutation accumulation [35], and theoretical analyses of the double-hit scenario [33], have hinted how symmetric cell divisions can protect cell populations from extreme mutational acquisition events. Nonetheless, the statistical basis of cancer risk on rare event phenomena in renewing tissues remains poorly defined. Here, we present a generic theory for how properties of the extreme mutation number distribution scale with age and cell population size, and how this determines the risk of accumulating a critical number of mutations. In particular, we quantify how epistatic buffering can delay the transformation towards tumor growth, and elucidate how drift dynamics of the renewing cell population moderates the strength of fluctuations, diminishing the frequency of rare events.

To model the long-term accumulation of mutations in a renewing cell population, we consider a stochastic model similar to Refs. [24, 36] in which cells replicate through division, are lost, and acquire mutations stochastically, while the total number of cells N is maintained constant (the condition of homeostasis). For simplicity, we assume that mutations occur at a constant rate, μ, and assign indices $i = 1, ..., N$ to the cells, where cell i is characterized by the number of mutations it has acquired, m_i. When cell i is lost, at rate λ, another cell j simultaneously divides symmetrically, producing a copy with the same mutational signature which replaces the lost cell (assuming its index i) – a Moran process/voter model [37][38, 39]. To keep the analysis conceptually simple, we initially assume that a cell can be replaced by division of any other cell $j = 1, ..., N$. Later, and in detail in the Supplemental Material, we will also consider the case in which only neighboring cells can replace each other, which more faithfully mimics the behavior of self-renewing populations in many cycling tissues.

In the following, we consider a mutation rate that is independent of stem cell loss/replacement. This choice acknowledges that some stem cell divisions may lead to asymmetric fate outcome [19], allowing mutations to arise without the loss and replacement of a stem cell. Furthermore, we explicitly consider the situation of epistatic buffering, where an accumulation of the critical complement of mutations is necessary to change the cell dynamics. In this phase, the mutations' effect is neutral. In summary, the model dynamics can be written as the process

$$m_i \xrightarrow{\lambda} m_j, \quad m_i \xrightarrow{\mu} m_i + 1 \ , \tag{1}$$

where indices i and j are chosen randomly.

In the following, we will address the risk $\bar{P}_N(m_c, T)$ that at least one cell in a population of N cells acquires a critical number of mutations m_c after time $t = T$. This corresponds to the probability that the maximal mutation number across the population, $m^* := \max(m_1, ..., m_N)$, reaches or exceeds m_c which is related to the cumulative distribution function (CDF) of m^*, $P^*_N(m_c, T) := \text{Prob}(m^*(T) < m_c) = 1 - \bar{P}_N(m_c, T)$, whose properties we study here.

Before addressing the dynamics of the general model, as a benchmark, we first consider the case $\lambda = 0$. In this case, cells accumulate mutations independently, corresponding to N independently distributed Poisson processes. Although, strictly, this case does not admit a simple scaling form for the distribution of extremes [40], it can be well-approximated for large μT by the extreme value distribution of normally distributed random variables with mean and variance μT (see Supplemental Material). From this it follows that, at large μT, the difference between maximum mutation number m^* and the population mean, $\Delta m^* := m^* - \langle m_i \rangle$ has a CDF, $P^*_N(\Delta m_c) := \text{Prob}(\Delta m^* < \Delta m_c)$, according to a Gumbel distribution [41],

$$P^*_N(\Delta m_c) \simeq e^{-e^{-X}}, \ \ X = \frac{\Delta m_c - \tilde{m}}{\sigma_N} - \ln \ln 2 \ , \tag{2}$$

with median $\tilde{m}$ and scaling width σ_N given by

$$\tilde{m} \simeq \sqrt{2\mu T \ln N}, \quad \sigma_N \approx \sqrt{\frac{\mu T}{2 \ln N}} \ . \tag{3}$$

The scaling estimate for the mean value $\langle \Delta m^* \rangle$ coincides with that of $\tilde{m}$ (see Supplemental Material). Thus, the CDF's front becomes tight for large T and N around $\Delta m^* \simeq (2\mu T \ln N)^{1/2}$.

In the case of a non-zero cell loss/replacement rate with $\lambda > 0$, any two cells may have a common ancestor and thus do not accumulate mutations independently. It is then instructive to consider the *genealogy* of the cell population, as illustrated in Fig. 1a. The genealogy describes the mutational history of all ancestors of cells at time $t = T$ and has the form of a binary tree, where branches connect daughter cells with their mothers [42]. It contains all mutational paths that start at $t = 0$ and reach the present. In considering the mutational statistics at time $t = T$, it is therefore sufficient to consider only mutations that occur on the genealogy [42, 43].

The tree structure of the genealogy is characterized by its branching times t_k, at which the branch number changes from $k-1$ to k (see Fig. 1a), i.e. during the period $t_k < t < t_{k+1}$, the genealogy consists of k branches. The branching times can be determined by following the genealogy backwards in time $\hat{t} = T - t$; a coalescent process [43–45]. This results in branching times whose intervals $\Delta t_k := t_{k+1} - t_k$ are exponentially distributed, with $\text{Prob}(\Delta t_k) = \langle \Delta t_k \rangle^{-1} e^{-\Delta t_k / \langle \Delta t_k \rangle}$ and mean branching times (see Supplemental Material)

$$\langle \Delta t_k \rangle = \frac{N}{k\,(k-1)} \frac{1}{\lambda} \ . \tag{4}$$

Importantly, the accumulation of mutations along a single branch follows a simple, independent Poisson process [43] and thus the mean mutation number is simply $\langle m_i \rangle(T) = \mu T$.

For times T large enough, one can trace the genealogy back to its root at $\hat{t} = \hat{T}_{LCA} := T - t_2$, the *last common ancestor (LCA)* (see Fig. 1a). Thus, an LCA exists whenever $T > \hat{T}_{\text{LCA}}$, which is on average $\langle \hat{T}_{\text{LCA}} \rangle = \sum_{k=2}^{N} \langle \Delta t_k \rangle \approx N/\lambda$ (for $N \gg 1$). In that case, before the time $T_{\text{LCA}} = T - \hat{T}_{\text{LCA}} = t_2$, the genealogy corresponds to the mutational path of a single cell for which the maximum m^* equals the mutation number m. Hence, it follows that $\Delta m^* = m^* - \langle m_i \rangle > 0$ only for times larger than T_{LCA}, such that the statistics of Δm^* does become independent of the total time T for $T > \hat{T}_{\text{LCA}}$ [46]. Indeed, Monte Carlo simulations of the model confirm this conjecture, as is illustrated in Fig. 1b for physiological parameters, $\lambda \approx 6000\mu$, according to Refs. [7, 8], where a plateau is reached for $\langle \Delta m^* \rangle(T)$ around $T \approx \langle \hat{T}_{\text{LCA}} \rangle$. This is in contrast to the case $\lambda = 0$, for which $\langle \Delta m^* \rangle \sim (\mu T)^{1/2}$.

To support this finding for $T > \hat{T}_{\text{LCA}}$ quantitatively, we note that the branching times in the genealogy are

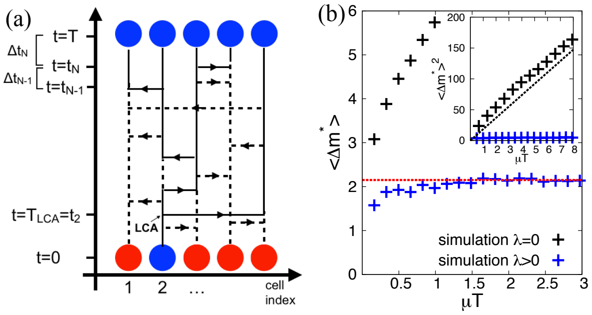

Figure 1. The genealogy and its implications. (a) Illustration of the history of mutation accumulation. Vertical lines represent mutational paths; horizontal arrows mark loss/replacement events. Dashed lines are mutational paths that are lost, while bold lines are paths that survive until time $t = T$, constituting the genealogy. If T is large enough, the genealogy possesses a last common ancestor (LCA). (b) Mean difference between maximum mutation number and the population mean, $\langle \Delta m^* \rangle = \langle m^* \rangle - \langle m_i \rangle$, as a function of μT, for fixed $N = 10,000$. Blue pluses are results from Monte Carlo simulations for parameters from human eyelid epidermis stem cells $\lambda = 6067\mu$ (corresponding to $\mu = 0.27/(63\,\text{years})$ and $\lambda = 0.5/\text{week}$ [7, 8]) so that $\langle \hat{T}_{\text{LCA}} \rangle \approx N/\lambda = 1.65/\mu$. The red dashed line illustrates the saturation asymptote. Black pluses are corresponding results for $\lambda = 0$ and the black dashed line marks $\langle \Delta m^* \rangle \simeq (\mu T)^{1/2}(2 \ln N)^{1/2}$.

random and exponentially distributed – a Markov process – which corresponds to a branching process with initially two branches at time T_{LCA}, and branching rate *per branch* $\nu_k := 1/\langle \Delta t_k \rangle k$. By approximating the random accumulation of mutations along each branch (Poisson process) by diffusive random walks in the variables $m_i - \mu T$ (valid for $\mu T \gg 1$), the mutation accumulation of the genealogy becomes an unbiased *branching random walk (BRW)*. For unit branching rate, it has been shown [47] that the CDF of the maximum Δm^* of the BRW, $P_N^*(\Delta m_c, \tau) = \text{Prob}(\Delta m^*(\tau) < \Delta m_c)$, follows a Fisher-KPP-type equation [48, 49]

$$\partial_\tau P_N^* = D \frac{\partial^2 P_N^*}{\partial \Delta m_c^2} - P_N^*[1 - P_N^*] \ , \tag{5}$$

with the dimensionless time $\tau = \nu t$ measured in units of the constant branching time ν^{-1}, and D, the diffusion constant of the random walk. The solution of this equation has the form

$$P_N^*(\Delta m_c, \tau) = f(\Delta m_c - \tilde{m}(\tau)) \ , \tag{6}$$

with the median of P_N^*, $\tilde{m}(\tau) = 2\sqrt{D}\,\tau + O(\ln \tau)$ [47].

On the genealogy, the branching rate ν_k is not constant. However, by aggregating time in units of branching times, in a step-wise manner, we can define a rescaled time $\tau(\{t_k\})$ (see Supplemental Material) and map the genealogy on a unit branching process with effective diffusion constant $D_k := \mu/2\nu_k$. While D_k does not explicitly depend on time, we can take the ensemble average over the branch numbers k, $D(\tau) := \langle D_k \rangle_k|_\tau = \mu N/(2\lambda) \times \langle (k-1)^{-1} \rangle_k|_\tau$, to get an effective time-dependent diffusion constant. According to Ref. [50] (see also Supplemental Material), for a diffusion constant $D(\tau')$ that decreases over time $\tau' < \tau = \tau(T)$, the CDF of the maximum of a BRW has the form of a Fisher-KPP wave, according to (6), but with

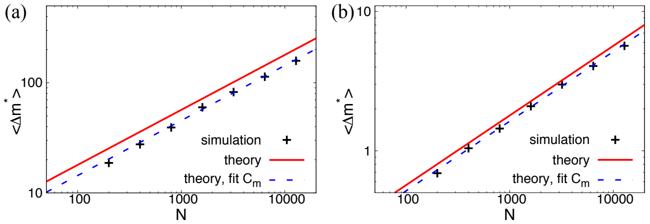

Figure 2. Mean maximum mutation number ahead of the mean, $\langle \Delta m^* \rangle$, as a function of N, for $T = 10\,N/\lambda$ such that $T > \hat{T}_{\text{LCA}}$. Shown are the results of Monte Carlo simulations (pluses), and theoretical predictions from the BRW approximation, Eq. (7), for fitted numerical constant $\mathcal{C}_{\tilde{m}}^{\text{fit}}$ (blue dashed line) and theoretically estimated value (solid red line), $\mathcal{C}_{\tilde{m}}^{\text{th}} = 1.79$ for (a) $\mu = \lambda$ ($\mathcal{C}_{\tilde{m}}^{\text{fit}} = 1.43$) and (b) $\mu = 0.001\lambda$ ($\mathcal{C}_{\tilde{m}}^{\text{fit}} = 1.63$) .

$$\tilde{m} = \left[\int_0^\tau 2\sqrt{D(\tau')}\, d\tau' \right] (1 - O(\tau^{-\frac{2}{3}})) \approx \mathcal{C}_{\tilde{m}} \sqrt{\frac{\mu N}{\lambda}} \ , \tag{7}$$

where $\mathcal{C}_{\tilde{m}} = \int_0^\infty \sqrt{2\langle (k-1)^{-1} \rangle_k|_{\tau'}}\, d\tau'$ is independent of the parameters N, T, λ and μ. On the right hand side of the equation, we assumed that $\tau \gg 1$, which is valid for large N, since for a unit branching process (in rescaled time τ) the branch number is $k(T) = N \approx e^\tau$. Thereby, terms of $O(\tau^{-\frac{2}{3}})$ are omitted, and the integral becomes independent of N. A numerical evaluation (see Supplemental Material) yields $\mathcal{C}_{\tilde{m}} \approx 1.79$.

As anticipated (Fig. 1b), we find that $\tilde{m}$ becomes independent of T for $T \gtrsim \hat{T}_{\text{LCA}}$ (The mean, $\langle \Delta m^* \rangle$, follows the same scaling in N and T [47]). Notably, $\tilde{m}$ scales with N as the standard deviation σ_0 of the distribution of individual intra-population differences, $m_i - m_j$ (derived in Ref. [36]), which is in contrast to independent Poisson processes for which $\tilde{m} \sim \sigma_0\sqrt{2 \ln N}$. In Fig. 2, we compare the theoretical results from Eq. (7) with results for $\langle \Delta m^* \rangle$ from Monte Carlo simulations for $T > \hat{T}_{\text{LCA}}$, as a function of N. The theory, with fitted constant $\mathcal{C}_{\tilde{m}}$, shows excellent agreement with simulations, while the calculated value $\mathcal{C}_{\tilde{m}}^{\text{th}} = 1.79$ shows small deviations which originate from contributions with small τ' in the approximation of Eq. (7). Notably, our theory is also valid for $\mu T \sim 1$ as shown in Fig. 2b for $\mu = 0.001\lambda$.

While the nonlinear form of the Fisher-KPP equation does not admit an exact solution, the CDF's upper tail with $\bar{P}_N = 1 - P_N^*(\Delta m_c, \tau) \ll 1$ can be mapped onto a simple diffusion equation with time-varying diffusion constant (see Supplemental Material). Since variances add linearly in this case, for $T > \hat{T}_{LCA}$ and $\Delta m_c \gg \tilde{m}(N)$, the CDFs tail is that of a non-normalized Normal distribution,

$$P_N^*(\Delta m_c) \simeq 1 - \frac{N\sigma_{\text{eff}}\, e^{-\frac{\Delta m_c^2}{2\sigma_{\text{eff}}^2}}}{\sqrt{2\pi}\,\Delta m_c}\ , \tag{8}$$

with $\sigma_{\text{eff}}(\tau) = 2(\int_0^\tau D(\tau')\,d\tau')^{1/2} \approx \mathcal{C}_\sigma\,(\frac{\mu N}{\lambda})^{1/2}$ and $\mathcal{C}_\sigma \approx (\int_0^\infty 2\langle(k-1)\rangle_k|_{\tau'}\,d\tau')^{1/2} \approx 0.76$ (see Supplemental Material).

For fixed T and large N, the population may not possess an LCA. In this case the genealogy fragments into k independent sub-genealogies $l = 1, ..., k$. Each subgenealogy, however, can again be approximated by a branching random walk, with a CDF, $P_l(\Delta m_c)$, having a Gaussian tail according to Eq. (8). Therefore, and since the subpopulations accumulate mutations independently from each other, the CDF of the whole population, $P_N^*(\Delta m_c)$, is approximated for large N by a Gumbel distribution according to Eq. (2) [41], scaled by σ_{eff}, and with effective number of independently distributed random variables, $k \approx N/\lambda T$ (see Supplemental Material). This CDF has then a median and scaling width,

$$\tilde{m} \simeq \mathcal{C}_\sigma\sqrt{2\,\mu T \ln\frac{N}{\lambda T}},\quad \sigma_N \simeq \mathcal{C}_\sigma\sqrt{\frac{\mu T}{2\ln\frac{N}{\lambda T}}}, \tag{9}$$

where $\mathcal{C}_\sigma$ is defined according to Eq. (8). The same applies to $\langle\Delta m^*\rangle$. Thus, $\langle\Delta m^*\rangle$ always stays below the corresponding value for $\lambda = 0$, $\langle\Delta m^*\rangle \simeq \sqrt{2\mu T\ln N}$. Figure 3 shows Monte Carlo simulations of $\langle\Delta m^*\rangle$ together with theory, with fitted $\mathcal{C}_\sigma$, which shows a good agreement in the given range of N, for both large and small mutation rates, $\mu = \lambda$ and $\mu = 0.001\lambda$. Deviations from the theoretically approximated value $\mathcal{C}_\sigma^{\text{th}} = 0.76$ are expected, for the same reasons as for $\mathcal{C}_{\tilde{m}}$ before, and furthermore for very large N and small μT, since then the tail of the extreme value distribution, P_l, differs from the Gaussian approximation, Eq. (8). We note that in contrast to the intra-population standard deviation, $\sigma_0 \approx (\mu T)^{1/2}$, for fixed T and large N [36], $\langle\Delta m^*\rangle$ does scale with N.

Until now, we have considered a process in which any cell may replace any other in the cell population. However, in most tissues, cell fate is regulated locally, resulting in stem cell loss and replacement correlated between neighboring cells, taking place in tubular (one-dimensional), epithelial (two-dimensional) or volumnar (three-dimensional) settings [51]. Such situations can be modeled by embedding cells on a finite d-dimensional regular lattice, allowing replacement only between neighboring cells [52]. In this case, our general theory remains valid, based on the mapping of the genealogy on a BRW (for details, see Supplemental Material); only the distribution of branching times, Δt_k, differs. Nonetheless, it is only for $d = 1$ that we observe a significantly different scaling compared to the infinite-dimensional case, with $\tilde{m} \sim N(\mu/\lambda)^{1/2}$ and $\tilde{m} \sim (\mu T \ln\frac{N}{\sqrt{\pi\lambda T}})^{1/2}$ for $T \gtrsim \hat{T}_{\text{LCA}}$ and $T < \hat{T}_{\text{LCA}}$ respectively.

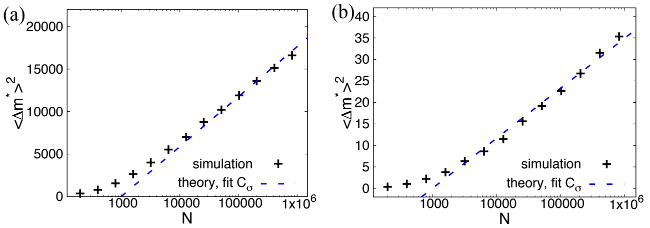

Figure 3. Squared mean maximum mutation number ahead of the population mean, $\langle\Delta m^*\rangle^2$, as a function of N, for fixed $T = 1000/\lambda$. Shown are the results of Monte Carlo simulations (pluses) and theory, Eq. (9), with fit parameter $\mathcal{C}_\sigma^{\text{fit}}$ (dashed line) for (a) $\mu = \lambda$ ($\mathcal{C}_\sigma = 1.13$) and (b) $\mu = 0.001\lambda$ ($\mathcal{C}_\sigma = 1.59$).

Finally, we consider the risk of accumulating a critical mutation number m_c, the 'first hit' probability of m_c, $\tilde{P}_{m_c}(T)$. The distribution of first hit times is related to the CDF, P_N^*, and can in principle be determined through its Laplace transform [53]. However, a closed analytical expression for this distribution is not available. Thus, we must rely on numerical solutions. In Figure 4, results from stochastic simulations for $\tilde{P}_{m_c}(T)$ are shown with parameters chosen to match physiological conditions of human epidermal stem cells [7, 8], comparing a situation of stochastic stem cell loss/replacement (*symmetric stem cell divisions*), $\lambda > 0$ with the case of *asymmetric stem cell divisions* only, $\lambda = 0$. Panel 4a shows $\tilde{P}$ as function of time T for $m_c = 6$ [22, 54], and indicates an earlier and more abrupt rise in risk for asymmetric divisions than for symmetric divisions. This is because the front of $\tilde{P}$, estimated by $\tilde{m}$, is significantly ahead for asymmetric divisions, as our theory predicts (see also Fig. 1b). In this regime, $\tilde{P}$ is more than an order of magnitude larger for asymmetric divisions than for symmetric ones, as shown in Fig. 4a top. This provides a theoretical foundation for claims that the risk of tumor initiation is decreased by symmetric divisions [33, 35]. In Figure 4b, we show how $\tilde{P}$ depends on the threshold number of mutations m_c. We see that, even for mean mutation numbers much lower than unity, a threshold of multiple mutations is likely to be accumulated in some cells. This demonstrates that the acquisition of a critical number of mutations is indeed dominated by extreme values, and illustrates the importance of epistatic buffering – quantified by $m_c > 1$ – to reduce the risk of triggering a selective advantage.

To summarise, we have shown that, on average, the maximum number of neutral mutations among cells in a renewing cell population is substantially lower if cells replace each other when dividing (symmetric divisions, replacement rate $\lambda > 0$) compared to non-replacing cell populations (only asymmetric divisions, $\lambda = 0$), despite

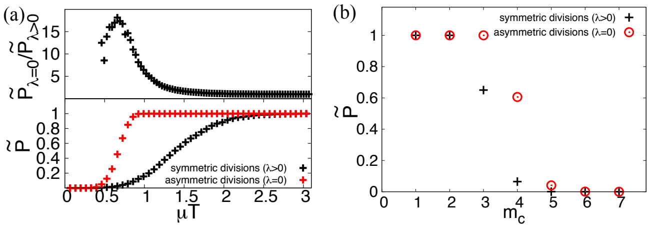

Figure 4. Probability $\tilde{P}_{m_c}$ of accumulating a critical number of mutations, m_c, in a single cell amongst $N = 10^5$ cells for $\lambda = 6000\mu$ (black points) which, for illustrative purposes, corresponds to an estimated area of skin epidermis of 0.1 cm^2 [7, 8] (see also Fig. 1b) and for $\lambda = 0$ (red points). (a) $\tilde{P}_{m_c}$ as function of time and threshold $m_c = 6$ [22, 54]. The top panel shows the ratio of the latter two cases, $\tilde{P}_{\lambda=0}/\tilde{P}_{\lambda>0}$. (b) $\tilde{P}_{m_c}$ as a function of the threshold mutation number m_c for fixed time $\mu T = 0.27$, corresponding to 63 human years of skin turnover [8]

bearing the same mean number of mutations. For $\lambda = 0$, the headway of the maximum mutation number above the mean mutation number, Δm^*, diverges with time T as $(\mu T)^{1/2}$, while for $\lambda > 0$, Δm^* saturates to a constant, which scales as $(\mu N/\lambda)^{1/2}$ ($\mu =$ mutation rate), for all dimensions $d > 1$. This can be understood by the mapping to branching random walks, and by considering that any divergence of population features may emerge only after the last common ancestor which exists at a time point independent of T. For finite (fixed) T, a different scaling $\Delta m^* \sim (\mu T)^{1/2}(\ln N_{\text{eff}})^{1/2}$ is observed, where N_{eff} may depend on the dimension (e.g. $N_{\text{eff}} \propto N/\lambda T$ for $d > 2$). These results are of importance for estimating tumor incidence rates under epistatic buffering, since usually a single cell with maximum accumulation of mutations is triggering a tumor-initiating event. We thus conclude that at intermediate time scales, the risk of tumor initiation is substantially higher for asymmetric than for symmetric stem cell fate.

We thank Steffen Rulands for valuable discussions and for help with literature research on genealogies and the coalescent process. We further are thankful for the support by a EPSRC Critical Mass Grant and a DFG Research Fellowship.

[1] A. G. Knudson, Proc. Natl. Acad. Sci. **68**, 820 (1971).
[2] A. G. Knudson, Nature Reviews **1**, 157 (2001).
[3] A. Ashworth, C. J. Lord, and J. S. Reis-Filho, Cell **145**, 30 (2011).
[4] J. L. Hartman, B. Garvik, and L. Hartwell, Science **291**, 1001 (2001).
[5] J. A. G. M. de Visser, et al., Evolution **57**, 1959 (2003).
[6] Z. Gu, et al., Nature **421**, 63 (2003).
[7] B. D. Simons, Proc. Natl. Acad. Sci. **113**, 128 (2016).
[8] I. Martincorena, et al., Science **348**, 880 (2015).
[9] J. H. Moore, Nature Genetics **37**, 13 (2005).
[10] L. Jasnos and R. Korona, Nature Genetics **39**, 550 (2007).
[11] N. L. Komarova, Proc. Natl. Acad. Sci. **111**, 10789 (2014).
[12] L. Wagstaff, G. Kolahgar, and E. Piddini, Trends in Cell Biology **23**, 160 (2013).
[13] M. P. Alcolea, et al., Nature Cell Biology **16**, 615 (2014).
[14] S. Brown, et al., Nature (2017).
[15] G. M. Cooper, in *The Cell: A Molecular Approach. 2nd edition.* (Sinauer Associates;, 2000).
[16] L. Alonso and E. Fuchs, Proc. Natl. Acad. Sci. **100**, 11830 (2003).
[17] N. Barker, et al., Nature **449**, 1003 (2007).
[18] C. S. Potten, Cell Tissue Kinet. **7**, 77 (1974).
[19] S. J. Morrison and J. Kimble, Nature **441**, 1068 (2006).
[20] E. Clayton, et al., Nature **446**, 185 (2007).
[21] C. Lopez-Garcia, A. M. Klein, B. D. Simons, and D. J. Winton, Science **330**, 822 (2010).
[22] P. Armitage and R. Doll, British Journal of Cancer **8**, 1 (1954).
[23] P. Armitage and R. Doll, British Journal of Cancer **11**, 161 (1957), 209.
[24] D. B. Weissman, M. M. Desai, D. S. Fisher, and M. W. Feldman, Theoretical Population Biology **75**, 286 (2009).
[25] D. M. Weinreich and L. Chao, Evolution **59**, 1175 (2005).
[26] I. Bozic, et al., Proc. Natl. Acad. Sci. **107**, 18545 (2010).
[27] F. Michor, Y. Iwasa, C. Lengauer, and M. A. Nowak, Seminars in Cancer Biology **15**, 484 (2005).
[28] M. M. Desai and D. S. Fisher, Genetics **176**, 1759 (2007), 0612016.
[29] I. M. Rouzine, J. Wakeley, and J. M. Coffin, Proceedings of the National Academy of Sciences **100**, 587 (2003).
[30] L. S. Tsimring, H. Levine, and D. A. Kessler, Physical Review Letters **76**, 4440 (1996).
[31] S. H. Moolgavkar and A. G. Knudson, Journal of the National Cancer Institute **66**, 1037 (1981).
[32] Y. Iwasa, F. Michor, and M. A. Nowak, Genetics **1579**, 1571 (2004).
[33] L. Shahriyari and N. L. Komarova, PLoS ONE **8**, e76195 (2013).
[34] S.-C. Park and J. Krug, Proceedings of the National Academy of Sciences **104**, 18135 (2007), arXiv:0711.1989.
[35] P. T. McHale and A. D. Lander, PLoS computational biology **10**, e1003802 (2014).
[36] B. H. Good and M. M. Desai, Theoretical Population Biology **85**, 86 (2013).
[37] Note1, the voter model and Moran process are equivalent for regular lattices and random pairing as considered here.
[38] P. A. P. Moran, Mathematical Proceedings of the Cambridge Philosophical Society **54**, 60 (1958).
[39] P. Clifford and A. Sudbury, Biometrika **60**, 581 (1973).
[40] M. R. Leadbetter, G. Lindgren, and H. Rootzén, *Extremes and related properties of random sequences and processes* (Springer, New York, 1983).
[41] J.-P. Bouchaud and M. Marc, J. Phys. A: Math. Gen. **30**, 7997 (1997).
[42] J. Kingman, Journal of Applied Probability **19**, 27 (1982).
[43] R. R. Hudson, in *Oxford Surveys in Evolutionary Biology*, edited by D. Futuyama and J. Antonovics (1991), p. 1, 7th ed.
[44] J. Kingman, Stochastic Processes and their Applications **13**, 235 (1982).

[45] J. Wakely, in *Coalescent Theory: An Introduction* (W. H. Freeman, 2008).

[46] Note2, in fact, any statistical measure that depends on population variation becomes independent of T, such as the population's variance [36].

[47] M. D. Bramson, Communications on Pure and Applied Mathematics **31**, 531 (1978).

[48] R. A. Fisher, Annals of Eugenics **7**, 355 (1937).

[49] A. Kolmogorov, I. Petrovskii, and N. Piscunov, Byul. Moskovskogo Gos. Univ. **1**, 1 (1937).

[50] M. Fang and O. Zeitouni, J. Stat. Phys. **149**, 1 (2012).

[51] A. M. Klein and B. D. Simons, Development **138**, 3103 (2011).

[52] A. M. Klein, D. P. Doupé, P. H. Jones, and B. D. Simons, Phys. Rev. E **77**, 031907 (2008).

[53] M. Nyberg, T. Ambjörnsson, and L. Lizana, New J. Phys. **18**, 063019 (2016).

[54] I. P. M. Tomlinson, M. R. Novelli, and W. F. Bodmer, Proc. Natl. Acad. Sci. **93**, 14800 (1996).