Extreme value statistics of mutation accumulation in renewing cell populations

Philip Greulich\textsuperscript{1,2} and Benjamin D. Simons\textsuperscript{3,4,5}

\textsuperscript{1}Mathematical Sciences, University of Southampton, Southampton, UK
\textsuperscript{2}Institute for Life Sciences, University of Southampton, Southampton, UK
\textsuperscript{3}Cavendish Laboratory, University of Cambridge, Cambridge, UK
\textsuperscript{4}Wellcome Trust-CRUK Gurdon Institute, University of Cambridge, Cambridge, UK
\textsuperscript{5}Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK

(Dated: February 22, 2022)

The emergence of a predominant phenotype within a cell population is often triggered by a rare accumulation of DNA mutations in a single cell. For example, tumors may be initiated by a single cell in which multiple mutations cooperate to bypass a cell’s defense mechanisms. The risk of such an event is thus determined by the extremal accumulation of mutations across tissue cells. To address this risk, 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 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 \cite{11,12}, disruptions of such pathways may often be tolerated within a homeostatic tissue cell population. Evidence from studies of the cancer genome \cite{2,6,7} 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 called “genetic” or “epistatic buffering” \cite{2,4,8,10}. The resulting proliferative advantage of mutated cells confers clonal dominance \cite{11,12} which, if sustained long-term \cite{13}, constitutes a potential tumour-initiating event. Crucially, since one cell within a tissue cell population is sufficient to trigger such an event \cite{14}, 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 tissue, 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 \cite{15,16}. In addition to asymmetric divisions, which leave the stem cell population unchanged \cite{17}, in most of these tissues the frequent, stochastic loss of stem cells is compensated by replacement via neighbors that divide symmetrically \cite{18,19}. 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 drive tumor progression have focused predominantly on population means or have neglected the potential for epistatic buffering \cite{20,23}. The impact of stochastic cell fate dynamics on the statistics of rare mutational signatures has remained under-explored. However, recently, numerical studies have shown how maintenance mechanisms reliant on stochastic stem cell self-renewal can protect cell populations from extreme mutational acquisition events \cite{24}. These findings have been reinforced by analytical studies based on a specific model of tumor-initiation involving a “double-hit” \cite{25}. However, 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 number, and how this determines the risk of accumulating a critical number of mutations. In particular, we elucidate how drift dynamics of the renewing cell population moderates the strength of fluctuations, diminishing the frequency of rare events. Besides its relevance for assessing the risk of tumor initiation, this theory also generically elucidates how a predominant phenotype can emerge in a cell population (e.g. bacteria) via the epistatic cooperation of individually (near-)neutral mutations.

To model the long-term accumulation of mutations in a renewing cell population we consider a stochastic model closely related to the Moran process in population genetics \cite{25}. In this model, cells replicate through division, acquire mutations and are lost stochastically while the total number of cells $N$ is maintained constant (the condition of homeostasis). Therefore, we assign a fixed number of ‘sites’ $i = 1, ..., N$ to the cells, where a cell at site $i$ is char-
characterized by mutation number, \( m_i \). When a cell at site \( i \) is lost, at rate \( \lambda \), another cell at a random site \( j \) simultaneously divides symmetrically, producing a copy with the same mutational signature, to replace the lost cell on site \( i \). In addition, any cell \( i \) with \( m_i \) mutations can acquire stochastically an additional mutation at a constant rate \( \mu \). Note that in stem cell populations, asymmetric cell divisions, where one of the daughter cells commits to terminal differentiation and loss, leave the configuration of mutations across cells invariant, and so need not to be considered explicitly. Their potential to effect additional mutations through division is incorporated as the mutation rates are decoupled from the loss/replacement rate. For simplicity, we do not distinguish between the loci of mutations in the genome, an approximation that is valid for low net mutational burden. Furthermore, we consider the scenario before transformation into a hyper-proliferative state, in which the mutations’ effect on proliferation is neutral. The model dynamics can be written as the process

\[
\begin{align*}
\lambda m_i & \rightarrow m_j, \\
\mu m_i & \rightarrow m_i + 1,
\end{align*}
\]

(1)

(2)

where sites \( i \) 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 \( m_c \), which is related to the cumulative distribution function (CDF) of \( m^* \), \( P^*_N(m_c, T) := \text{Prob}(m^*(T) \leq m_c) = 1 - \bar{P}_N(m_c, T) \). In particular, we will study the dependence of the CDF’s median and mean on \( N \) and \( T \).

Before addressing the dynamics of mutation accumulation on the background of stochastic cell loss/replacement, as a benchmark, we first consider the case \( \lambda = 0 \) in which cells accumulate mutations independently. In this case, the model describes \( N \) independently distributed Poisson processes and an expression for the extreme mutation number distribution can be determined straightforwardly. Although this expression does not admit a simple scaling form \cite{ref27}, the dependence on \( N \) can be well-approximated for large \( \mu T \) by normally distributed random variables with a mean and variance \( \mu T \) (see Supplemental Material). From this correspondence it follows that, at large \( N \) and \( \mu T \), the difference between maximum mutation number \( m^* \) and the population mean, \( \Delta m^* := m^* - \langle m_i \rangle \) has a CDF \( P^*_N(\Delta m_c) = \text{Prob}(\Delta m^* < m_c) \) which follows a Gumbel distribution,

\[
P^*_N(\Delta m_c) = e^{-e^{-X}} \text{ with } X = \frac{\Delta m_c - \tilde{m}}{\sigma_N} - \ln 2,
\]

(3)

with median \( \tilde{m} \) and scaling width \( \sigma_N \) given by \cite{ref28}

\[
\tilde{m} \approx \sqrt{2\mu T \ln N}, \quad \sigma_N \approx \frac{\sqrt{\mu T}}{2 \ln N}.
\]

(4)

The scaling estimate for the mean value \( \langle \Delta m^* \rangle \) coincides with that of \( \tilde{m} \) (see Supplemental Material). Thus, the CDF becomes narrowly peaked for large \( T \) and \( N \) around \( \Delta m^* \sim (2\mu T \ln N)^{1/2} \).

In the case of a non-zero cell loss/replacement rate, \( \lambda > 0 \), 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 \cite{ref29}. 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 \cite{ref29, ref30}.

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. 1h), 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 \cite{ref30, ref32}. This results in branching times whose intervals \( \Delta t_k := t_{k+1} - t_k \) are exponentially distributed, with \( \text{Prob}(\Delta t_k) = (\Delta t_k)^{-1}e^{-\Delta t_k/(\Delta t_k)} \) and mean branching
times (see Supplemental Material)

\[
\langle \Delta t_k \rangle = \frac{N}{k(k-1)} \frac{1}{\lambda}.
\]

(5)

Importantly, the accumulation of mutations along a single branch follows a simple, independent Poisson process \(\mathcal{P}(\lambda t)\) and thus the mean mutation number is simply \(\langle m_k \rangle(T) = \mu T\).

For times \(T\) large enough, if one follows the coalescent process far backwards in time, the number of branches reduces until \(t = T_{\text{LCA}} := T - t_2\), beyond which there is only a single one left, the last common ancestor (LCA) (see Fig. 1a). Thus, an LCA exists whenever \(T > T_{\text{LCA}}\), which is on average \(\langle T_{\text{LCA}} \rangle = \sum_{k=2}^{N} \langle \Delta t_k \rangle \approx N/\lambda\) (for \(N \gg 1\)). In that case, before the time \(T_{\text{LCA}} = T - T_{\text{LCA}} = t_2\), 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 \(T_{\text{LCA}}\), such that the statistics of \(\Delta m^*\) does not explicitly depend on the total time \(T\) if \(T > T_{\text{LCA}}\). Indeed, Monte Carlo simulations of the model confirm this conjecture for \(\langle \Delta m^* \rangle(T)\), as is illustrated in Fig. 1b for a high mutation rate \(\mu = \lambda\) (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 (7), but with

\[
\hat{m}(\tau) = \left[ \int_0^\tau 2\sqrt{D(\tau')} \, d\tau' \right] (1 - O(\tau^{-\frac{1}{2}})) \approx \hat{C}_m \sqrt{\frac{\mu N}{\lambda}} ,
\]

(8)

where \(\hat{C}_m = \int_0^\infty \sqrt{2((k-1)^{-1})} \, k \, d\tau'\) is independent of the model parameters \(N, T, \lambda\) and \(\mu\). Here, we took the limit \(\tau(T) \to \infty\), which is valid for large \(N\), as this is a unit rate branching process with branch number \(k(T) = N \approx e^\tau\). Thereby, terms of \(O(\tau^{-\frac{1}{2}})\) are omitted, and the integral becomes independent of \(N\). A numerical evaluation (see Supplemental Material) yields \(\hat{C}_m \approx 1.79\).

As expected, \(\hat{m}\) becomes independent of \(T\) if an LCA exists. The mean, \(\langle \Delta m^* \rangle\), follows the same scaling in \(N\) and \(T\), since due to the wave property of the CDF it only differs by a constant. This confirms our previous conjecture and the simulation results for \(\langle \Delta m^* \rangle(T)\) (Fig. 1b). In Fig. 2, we compare the theoretical results from Eq. (8) with results for \(\langle \Delta m^* \rangle\) from Monte Carlo simulations, as a function of \(N\). Here, \(T\) was scaled with \(T = 10 N / \lambda\) to assure that \(T > T_{\text{LCA}}\). The theory with fitted numerical constant \(C_{\text{th}}^m (C_{\text{th}}^m = 1.79, \text{blue dashed line})\) shows excellent agreement with simulations, while the calculated value \((C_{\text{th}}^m = 1.79, \text{red line})\) shows some deviation. These deviations are expected due to contributions with small \(\tau'\) in \(\hat{C}_m\) (whose approximation is valid for large \(\tau'\)). Remarkably, the predictions of the approximation are 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 tail with 

\[ P_N^*(\Delta m_c) \sim 1 - \frac{N\sigma_{\text{eff}} e^{-\frac{\Delta m_c^2}{2\sigma_{\text{eff}}^2}}}{\sqrt{2\pi} \Delta m_c} \]  

(9)

with 

\[ \sigma_{\text{eff}}(\tau) = 2\sqrt{\int_0^\tau D(\tau') d\tau'} \approx C_\sigma \sqrt{\frac{\mu N}{\lambda}}, \]  

(10)

where \( C_\sigma \approx \sqrt{\int_0^\infty 2(1-k)k\ln(\lambda\tau)} d\tau \approx 0.76 \) (see Supplemental Material).

For \( T < T_{\text{LCA}} \) the population may not possess an LCA. In this case the genealogy fragments into \( k \) sub-genealogies \( l = 1, \ldots, k \), each representing a subpopulation of \( n_l \) cells, while \( N = \sum_{l=1}^k n_l \). Each sub-genealogy, 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. (7). 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. (3) [28], scaled by \( \sigma_{\text{eff}} \), and with effective number of independently distributed random variables, \( k \approx N/\lambda T \) (see Supplemental Material). This CDF has then median and scaling width,

\[ \hat{m} \approx C_\sigma \left( 2 \ln \left( \frac{N}{\lambda T} \right) \right) \mu T, \quad \sigma_N \approx C_\sigma \left( 2 \ln \left( \frac{N}{\lambda T} \right) \right)^{1/2}, \]  

(11)

where \( C_\sigma \) is according to Eq. (10). The same scaling applies to \( \langle \Delta m^* \rangle \).

Thus, we expect the scaling \( \Delta m^* \sim \sqrt{\ln N} \), as for \( \lambda = 0 \), but with a negative offset \( -2C_\sigma^2 \ln(\lambda T)\mu T \) under the square root. Figure 3 shows Monte Carlo simulations of \( \langle \Delta m^* \rangle \) together with the theory, with fitted \( C_\sigma^2 \), which shows a good agreement in the shown range of \( N \), for both large and small mutation rates, \( \mu = \lambda \) and \( \mu = 0.001\lambda \). Deviations from the theoretically approximated value \( C_\sigma^2 \) are expected, for the same reasons as for \( C_\bar{m} \) before, and furthermore for very large \( N \), since the extreme value distribution of Poisson variables differs from the Gaussian approximation in this limit.

Finally, we consider the risk of accumulating a critical number of mutations \( m_c \), \( P_N(m_c, T) = 1 - P_N^*(m_c, T) \). In Figure 4 results from stochastic simulations for \( \hat{P}_N(T) \) are shown for \( m_c = 6 \) [20,37] with parameters chosen to match physiological conditions of human epidermal stem cells [6,7] (see figure caption), comparing a model of stochastic stem cell loss and replacement, \( \lambda > 0 \), with the hypothetical case of asymmetric stem cell divisions, \( \lambda = 0 \). As the front of \( P_N^*(m_c) \) moves with a speed \( \mu + \Delta T \), which is larger for \( \lambda = 0 \) (see Figure 4b), it reaches the critical value \( m_c = 6 \) earlier than for \( \lambda > 0 \), resulting in an earlier increase of \( P_N(m_c, T) \) in Figure 4b. Remarkably, the threshold of \( m_c = 6 \) mutations is already frequently exceeded for mean mutation numbers \( \mu T < 1 \), which demonstrates that the acquisition of a critical number of mutations is indeed dominated by extreme values. In Figure 4b, the risk ratio between the risk \( P_N|_{\lambda=0} \) for asymmetric divisions and that for symmetric divisions, \( P_N|_{\lambda>0} \), is shown. One observes a non-monotonic behavior: for intermediate times the ratio is large, while it approaches one for large times. The initial growth of the ratio is due to the smaller slope in the tail for \( \lambda > 0 \), \( 1/\sigma_N^{\lambda=0} \sim \sqrt{\ln(\mu T/\lambda T)} \) compared to the slope \( 1/\sigma_N^{\lambda=0} \sim \sqrt{\ln N} \) for \( \lambda = 0 \). Once \( P_N|_{\lambda=0} \) has saturated, however, \( P_N|_{\lambda>0} \) catches up with the latter until both approach \( P_N \approx 1 \). This provides a theoretical foundation for claims that the risk of tumor initiation is decreased by stochastic stem cell loss and replacement [24,25], yet this advantage is limited to intermediate time scales.

Until now, we have considered an “infinite-dimensional” process in which any cell may replace another. However, the majority of biological tissues are low-dimensional, where stem cell loss and replacement occur between neighboring cells in tubular (one-dimensional), epithelial (two-dimensional) or voluminar (three-dimensional) settings. Such situations can be modeled by embedding cells on a \( d \)-dimensional regular lattice, allowing replacement only between neighboring cells [38,39]. In this case, our general theory remains valid, based on the mapping of the genealogy on a BRW; only the distribution of branching times, \( \Delta t_k \), differs (see Supplemental Material). Nonetheless, only for \( d = 1 \) and \( T > T_{\text{LCA}} \), a significantly different scaling with \( \hat{m} \sim N \sqrt{\mu \lambda} \) is observed, compared to the infinite-dimensional case.

In summary, we have studied the asymptotic behav-
ior, with time $T$ and cell number $N$, of the maximum mutation number statistics in a renewing cell population, in which cells may be stochastically lost and replaced (Moran process). This is of importance if multiple neutral or near-neutral mutations can cooperate through epistasis to trigger hyper-proliferation, a potential tumor initiating event. We showed that, for a non-zero cell replacement rate, $\lambda > 0$, the difference between the average maximum mutation number and the population mean, $\langle \Delta m^* \rangle$, saturates to a constant value when scaled with $T$. Using an analogy to branching random walks, we showed that the value of this constant scales as $\langle \mu N / \lambda \rangle^{1/2}$ for $\lambda > 0$, while in the absence of symmetric loss/replacement, $\lambda = 0$ (e.g. for asymmetric stem cell divisions only), $\langle \Delta m^* \rangle$ scales with time and cell number as $\langle \mu T \ln(N) \rangle^{1/2}$. If $T$ is fixed, $\langle \Delta m^* \rangle$ scales as $(\mu T \ln(N/\lambda T))^{1/2}$ for $\lambda > 0$ and large $N \gg \lambda T$. From this result, it follows that at intermediate time scales, the risk of triggering hyper-proliferation is higher for asymmetric than for symmetric stem cell divisions, yet these probabilities converge for large times. Crucially, our theory also applies in a low-dimensional setting, where cell loss and replacement occurs between neighbors, albeit with a different scaling dependence for one-dimensional structures.

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. Ashworth, C. J. Lord, and J. S. Reis-Filho, Cell 145, 30 (2011).
[3] J. L. Hartman, B. Garvik, and L. Hartwell, Science 291, 1001 (2001).
[4] J. A. G. M. de Visser, et al., Evolution 57, 1959 (2003).
[5] Z. Gu, et al., Nature 421, 63 (2003).
[6] B. D. Simons, Proc. Natl. Acad. Sci. 113, 128 (2016).
[7] I. Martincorena, et al., Science 348, 880 (2015).
[8] J. H. Moore, Nature Genetics 37, 13 (2005).
[9] L. Jasnos and R. Korona, Nature Genetics 39, 550 (2007).
[10] N. L. Komarova, Proc. Natl. Acad. Sci. 111, 10789 (2014).
[11] L. Wagstaff, G. Kolahgar, and E. Piddini, Trends in Cell Biology 23, 160 (2013).
[12] M. P. Alcolea, et al., Nature Cell Biology 16, 615 (2014).
[13] S. Brown, et al., Nature (2017).
[14] G. M. Cooper, in The Cell: A Molecular Approach. 2nd edition. (Sinauer Associates;, 2000).
[15] L. Alonso and E. Fuchs, Proc. Natl. Acad. Sci. 100, 11830 (2003).
[16] N. Barker, et al., Nature 449, 1003 (2007).
[17] C. S. Potten, Cell Tissue Kinet. 7, 77 (1974).
[18] E. Clayton, et al., Nature 446, 185 (2007).
[19] C. Lopez-Garcia, A. M. Klein, B. D. Simons, and D. J. Winton, Science 330, 822 (2010).
[20] P. Armitage and R. Doll, British Journal of Cancer 8, 1 (1954).
[21] P. Armitage and R. Doll, British Journal of Cancer 11, 161 (1957), 209.
[22] C. J. Portier, A. Kopp-Schneider, and C. D. Sherman, Mathematical Biosciences 135, 129 (1996).
[23] I. Bozic, et al., Proc. Natl. Acad. Sci. 107, 18545 (2010).
[24] P. T. McHale and A. D. Lander, PLoS computational biology 10, e1003802 (2014).
[25] L. Shahrivari and N. L. Komarova, PLoS ONE 8, e76195 (2013).
[26] P. A. P. Moran, Mathematical Proceedings of the Cambridge Philosophical Society 54, 60 (1958).
[27] M. R. Leadbetter, G. Lindgren, and H. Rootzén, Extremes and related properties of random sequences and processes (Springer, New York, 1983).
[28] J.-P. Bouchaud and M. Marc, J. Phys. A: Math. Gen. 30, 7997 (1997).
[29] J. Kingman, Journal of Applied Probability 19, 27 (1982).
[30] R. R. Hudson, in Oxford Surveys in Evolutionary Biology, edited by D. Futuyama and J. Antonovics (1991), p. 1, 7th ed.
[31] J. Kingman, Stochastic Processes and their Applications 13, 235 (1982).
[32] J. Wakeley, in Coalescent Theory: An Introduction (W. H. Freeman, 2008).
[33] M. D. Branson, Communications on Pure and Applied Mathematics 31, 531 (1978).
[34] R. A. Fisher, Annals of Eugenics 7, 355 (1937).
[35] A. Kolmogorov, I. Petrovskii, and N. Piscunov, Byul. Moskovskogo Gos. Univ. 1, 1 (1937).
[36] M. Fang and O. Zeitouni, J. Stat. Phys. 149, 1 (2012).
[37] I. P. M. Tomlinson, M. R. Novelli, and W. F. Bodmer, Proc. Natl. Acad. Sci. 93, 14800 (1996).
[38] A. M. Klein, D. P. Doupé, P. H. Jones, and B. D. Simons, Phys. Rev. E 77, 031907 (2008).
[39] A. M. Klein and B. D. Simons, Development 138, 3103 (2011).