# Context-dependent selection as the keystone in the somatic evolution of cancer

**Vibishan B.**[1,2] **and Milind G. Watve**[1,2,*]

[1] Department of Biology, Indian Institute of Science Education and Research (IISER), Pune

This manuscript was compiled on December 21, 2018

**Somatic evolution of cancer involves a series of mutations and accompanying genomic, epigenomic and physiological changes in one or more clones of cells. Whether the mutations accumulate by chance alone ("bad luck" hypothesis) or necessarily involve selection on intermediate mutants leading to clonal expansion is unclear. An implicit assumption in clonal expansion is that any mutation leading to partial loss of regulation of cell proliferation will give a selective advantage to the mutant. However, a number of experiments show that an intermediate pre-cancer mutant has only a conditional selective advantage; given that tissue microenvironmental conditions are differ across individual organisms, the selective advantage to a mutant will be widely distributed over the population of organisms. We comparatively evaluate the three models, namely "bad luck", context-independent selection and context-dependent selection with respect to their ability to predict patterns in total incidence, age-specific incidence, and their ability to explain Petos and related paradoxes. Results show that the number of stem cells and mutation rates are unlikely to be rate limiting in human cancers, and that context dependence is necessary and sufficient to explain all the observed epidemiological patterns. This implies that the susceptibility to cancer can be substantially different across individuals, and cancer is not sheer bad luck. A wide range of physiological, genetic and behavioural factors influence the tissue micro-environment, and could therefore be the source of this context dependence in somatic evolution of cancer. The identification and targeting of these micro-environmental factors that influence the dynamics of selection offer new possibilities for cancer prevention.**

Somatic evolution | Mutation accumulation | Epidemiology | Cancer etiology |

Over the past 60 years or so, ideas in the field of cancer epidemiology have evolved significantly, with the first models starting from the Armitage-Doll multi-stage models (1), which predicted a power law relationship of cancer risk with age. The connection between the multiple stages and sequential genetic mutation events was established definitively for retinoblastoma with the two-hit hypothesis (2). This finding has since directed a lot of attention to mutational processes and genetic instability within cells as fundamental forces in cancer, and their signatures in population-level datasets; Tomasetti et al. have made the argument for cancer risk being largely determined by random mutations (3, 4), and others have studied the impact of selectively-neutral or deleterious passenger mutations on the expansion and progression of advantageous mutant clones (5), mutation accumulation rates across tissue types (6), or dependencies between mutations (7). On the other hand, an old debate in the theory of evolution is how the simple process of random mutations and natural selection can lead to compled structures such as the eye, that need coordinated action of several genes. This is often perceived as a monkey-on-a-typewriter paradox (8)-how likely is it that a monkey sitting at a typewriter and hitting keys at random would end up typing a meaningful sentence? The problem of cancer is qualitatively similar to this, but quantitatively even more difficult, as no single mutation is known to make a cell cancerous. All cancers are necessarily a combination of different types of genomic changes including point mutations, aneuploidy, and other chromosomal aberrations. The cancer phenotype has a large number of distinguishing characters, encapsulated by the notion of the "hallmarks" of cancer (9–11), and the wide range of characterisitics that these hallmarks include make it astonishing that so many alterations in cell properties come together in cancers purely out of chance, especially since most cancers must evolve independently in each individual organism.

Selection was clearly required to explain the complex cancer phenotype, clonal expansion provided the first such instance of a selective evolutionary process. Every component mutation on the way to a cancerous phenotype causes the mutant clone to expand, and as the mutant population increases, the probability of a second component mutation increases proportionately (12). Implicit in this theory is the assumption that every component mutation has a selective advantage over the normal cell. Since most changes involved in carcinogenesis relate to evading growth regulatory mechanisms, it is considered logical that any mutation that allows for such evasion will have a natural selective advantage. However, evidence has been accumulating over the past few years that the fitness advantage of a mutant is largely dependent on the tissue micro-environment (13, 14). Studies in mice (15) and humans (16) have demonstrated the effect of contingent factors, such as behavioural profiles and lifestyle parameters,

## Significance Statement

As opposed to the contemporary mutation-cenrtic perspective, selective forces, at the level of both the organism and the population, play important roles in cancer progression. We construct models of non-selective and selective forces, and compare their predictions of population-level cancer incidence to available epidemiological data from US populations. We find that incorporating selection, particularly selection that varies across organisms in a heterogenous population, is necessary to explain observed trends in cancer incidence. We posit that this heterogenous selection stems from factors that affect the tumour micro-environment that determine the competitive advantage to cancer-causing somatic mutations.

The authors declare that there are no conflicts of interest

[2] B.V. and M.G.W. contributed equally to this work.

[*] To whom correspondence should be addressed. E-mail: milind@iiserpune.ac.in

www.pnas.org/cgi/doi/10.1073/pnas.XXXXXXXXXX

on cancer progression. Such findings provide clear indications that the selective forces which determine mutant clone fitness can vary considerably across individual organisms, leading to *context-dependent clonal expansion* of potentially oncogenic mutants. This aspect of cancer progression has seen some modeling effort in the past few years (17–19), but a part of the legacy from Armitage-Doll has been partly lost. The first models in cancer etiology involved comparison of model predictions with epidemiologically-observed trends in incidence. This framework allowed for comparison of competing theories of carcinogenesis based on how closely they agreed with epidemiological data, and we construct a similar framework here, to evaluate competing causal factors in cancer etiology based on comparisons with data from the US population (20).

Across biological levels, we identify at least three processes that seem to play a major role in cancer progression and etiology, two of which, explicitly include selection at different levels: (1) random mutagenesis, or the "bad luck" hypothesis, (2) expansion of mutant clones within organisms, and (3) context-dependent selection acting across organisms. Since well-curated data is available for human cancer incidence patterns, we develop models of these three processes, and compare their predictions with the epidemiological picture of cancer in the human population. We briefly summarize the essential epidemiological features that must be part of any comparative modeling framework:

1. Total and age-specific incidence: Across cancer types, total population-level incidence lies around 20 to 30%, while age-specific and cumulative incidence patterns show more variations (20). Interestingly, recent analyses have shown for several cancer types that the age-specific incidence rates decline late in life, in contrast with general model predictions of a power law increase in incidence with age (21). The late-life decline causes the cumulative incidence to saturate with age at a small percentage of the population size, which is an important detail. No matter the lifespan, the proportion of cancer in the population can never reach 100%, which represents a finite limit that is not determined by time.

2. Incidence vs cell number: As mentioned earlier, the relationship between cancer risk and cell number (as the lifetime number of cell divisions, *lscd*) has been kept in the spotlight by recent work by Tomasetti et al. (3, 4). On the whole, the linearity of the relationship between *lscd* and cancer risk is still under debate, although an explanation of this non-linearity remains incomplete. We think that the relationship is non-linear, and possibly saturating, as is clear from an examination of the slopes of the cancer risk-*lscd* association. Linear or not, there are still real-life data for the relationship with which model predictions may be compared; as we show later, the veracity of the linearity claim is not directly relevant to the conclusions we draw from our models.

3. Incidence vs mutation rate: Empirical data on this relationship are less common, as mutation rates are difficult to measure reliably, although some efforts have been made (22). However, a general notion exists that higher mutation rates increases cancer risk, and this remains to tested rigorously, barring theoretical work dealing with the effect of mutagens on patterns in incidence data (23).

4. Non-mutagenic carcinogens: There are several agents, including hormones and growth factors, that increase cancer risk without affecting the basal mutation rate [citations]. The activity of these agents, and their signature in epidemiological patterns, are both important in building a complete framework of explaining cancer etiology.

5. Peto's paradox, and similar observations: This relates to the incidence-cell number relationship, as cancer risk is seen not to scale with body size or cell number across species (17), with does correlate with the latter within a species (24). A wide range of explanations have been offered for this observation (25), but a comprehensive explanation has been elusive. Nevertheless, it remains an key feature of cancer etiology with regard to the modeling effort.

## The "bad luck" model

This hypothesis assumes that the required set of driver mutations accumulate in a cell by chance alone. This may happen over a period of time, or in a single large-scale event, like chromothripsis (26). Regardless of whether mutations accumulate sequentially or otherwise, the "bad luck" model does not assume selection of any kind on mutations over the course of somatic evolution.

Consider an organism with a population of $n$ stem cells, each with a mutation rate per cellular generation per locus, $p$. The probability that at least one cell acquires one mutation at a given point of time can be given as $1 - (1 - p)^n$. If $k$ such mutations are requried for cancer onset, the probability of cancer according to the bad luck model can be given as below, based on an algebraic formulation (27):

$$p_{can} = 1 - (1 - p^k)^n \qquad [1]$$

Given the probability of cancer per unit time, $p_{can}$ from equation 1, the cumulative incidence of cancer for age, $A$, can be expressed as below:

$$p_A = 1 - (1 - p_{can})^A \qquad [2]$$

From equation 1, it is clear that the probability of cancer has a threshold relationship with both $n$ and $p$, such that incidence rises from near zero to 100% over a small range of $p$ and $n$, as shown in Figure 1. Within this narrow range, total incidence as reflected by the cumulative probability, $p_{can}$, occasionally lies in the realistic range of about 30%, but these are the exceptions.

From equation 2, the relationship of $p_{can}$ with age is a monotonically increasing function with a maximum at one. Figure 2 shows this relationship across the entire parameter range of $n$ and $p$, for which this prediction holds. Cancer probability increases monotonically with age, and only saturates at 100% incidence, which stands in stark contrast to the observed late-life decline in age-specific rates.

For the sake of simplicity, we have ignored the cost of lethal and/or passenger mutations, and assume that the mutations occur together at any given point of time; although the model predictions change marginally, we see that this assumption does not affect the general inferences we draw. Taken together, this formulation of the "bad luck" model predicts a sharp threshold relationship of cancer probability with both $n$ and $p$, and a monotonically increasing relationship with age, both

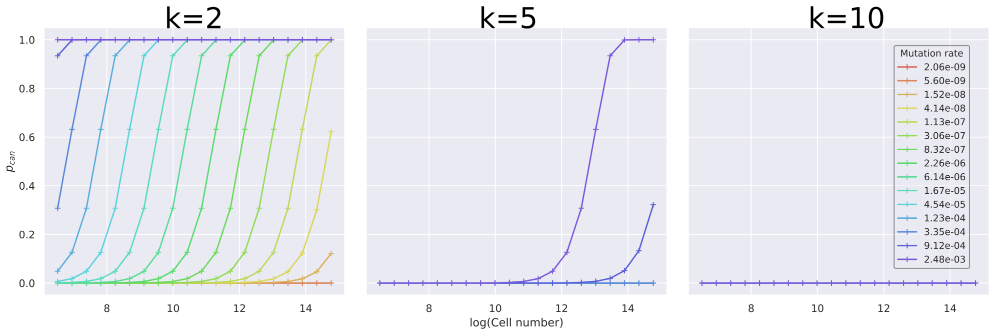

**Fig. 1.** Cancer probability, $p_{can}$ vs mutation rate and cell number for the "bad luck" model; $p_{can}$ remains near zero in part of the parameter range, and rises to one over a narrow region of the corresponding parameter. This proability is cumulative and therefore reflects total incidence in the population. For the "bad luck" model, the total incidence is rarely in the observed range of around 30%. The number of oncogenic mutations required for cancer onset ($k$) does not affect the existence of a threshold with $n$ and $p$, but does affect where the threshold occurs in the parameter space; for $k = 10$, the threshold does not occur within the tested range of $n$ and $p$. The legend shows values of $p$ for each curve.

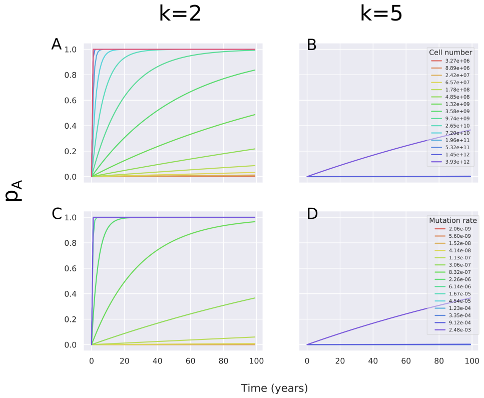

**Fig. 2.** Cancer probability vs age, as given by equation 2, (A and B) over the range of $n$, and (C and D) over the range of $p$; inset legends give the corresponding values. Cancer probability increases monotonically with age, saturating only at one in most cases. Where the probability does approach realistic values of total incidence, it still does not reflect the late-life decline in incidence rates observed epidemiologically. As in Figure 1, the number of oncogenic mutations required for cancer onset does not change the nature of the incidence-age relationship. For A and C, $k = 2$, and for B and D, $k = 5$.

of which can be falsified based on a simple comparison with published epidemiological data.

While there are ways of incorporating additional complexities within the formulation presented here, we choose to explore the effects of clonal expansion and context-dependent selection through a simulation-based framework. As we show below, it allows for easier exploration of the parameters under study, and creates a logical progression in terms of biological scale; the "bad luck" model deals with cell-level properties, clonal expansion is a tissue- or organ-level phenomenon, while context-dependent selection occurs between organisms at the level of the population.

## Models with selection on mutants

We use a linear process to model the sequential accumulation of mutations in a population of stem cells. We begin by considering the development of a generalized tissue compartment in each organism starting from one stem cell, with mutation rate per cell generation per locus, $p$, growing logistically to a carrying capacity, $n$, following the discrete logistic equation below:

$$m_{i,t} = m_{i,t-1} + m_{i,t-1} * g_i * (\frac{n - \sum_{i=0}^{k} m_{i,t-1}}{n}) - m_{i,t-1} * d \quad [3]$$

Here, $m_{i,t}$ is the size of the $i$th mutant population at time, $t$, with $i = 0$ being the non-mutant cell population, $g_i$ is the corresponding logistic growth rate, $d$ is the common death rate, and $k$ is the threshold number of oncogenic mutations required for cancer onset. As the organism develops into an adult, net growth in the stem cell compartment saturates, but reaches a dynamic equilibrium between cell death and renewal. The stem cell population can be reduced, either by death of stem cells, or assymetric division to produce differentiated cells. The death rate in equation 3 is this constant rate of cell removal from the stem cell population. Assuming a common death rate for all cell populations, the replacement of the lost cells by either mutants or non-mutants is a function of their growth rates. We simulate new mutation events stochastically; the probability of at least one cell accumulating a mutation is given by $1 - (1 - p)^{m_{i,t}}$, and if this probability exceeds a random number between 0 and 1, a new $(i+1)$th cell population is initiated. We assume that each new oncogenic mutation offers a growth advantage over older cell populations, leading to successive cycles of clonal expansion in which the newer population gradually replaces older cells through competitive exclusion. We simulate this linear evolution process until $k$ mutations have been accumulated, which is the assumed threshold for cancer onset. Death of the organism occurs either at cancer onset when the $k$th mutation occurs, or at the end of the natural lifespan of 100 years, whichever happens first. This simulation is repeated independently for 10000 organisms, and the population-level cancer incidence is recorded, along with the age of onset.

**Choice of parameter range.** In order to standardize the discrete logistic simulation, we assume the time unit to be one day per logistic growth step. Most human organs complete development and maturation wihtin the first 10-20 years of the lifespan, and the final carrying capacity achieved is the adult stem cell number, ranging between $10^6$ and $10^{11}$ across different tissues. Given the final population size and the time taken to reach it, a simple calculation based on the logistic equation shows the required growth rate for a non-mutant stem cell to be in the range of 0.00383-0.0131 [details in the supplement]. Starting from the non-mutant growth rate, $g_0$, growth rates are assumed to increase linearly for each subsequent mutant population. For all simulations, we assume $g_0 = 0.007$ to obtain cancer incidence within an organism's lifespan. Ranges of $n$ and $p$ are retained as in the "bad luck" model.

**The context-independent selection case.** While the clonal expansion theory introduced the notion of selective advantages to oncogenic mutants, it makes the implicit assumption that identical mutations have the same selective advantage in every organism in which they occur; stated otherwise, individual organisms do not differ in their propensity for mutant clonal expansion. To capture this in the context-independent selection case, we use the same linear progression of growth rates for all organisms in the simulation.

**The context-dependent selection case.** As argued earlier, it is becoming increasingly clear that the competitive outcomes of identical mutations can depend strongly on the microenvironmental context in which cell competition occurs. In order for selection on mutants to be context-dependent in our model, we randomize the progression of growth rates during mutation accumulation. Each organism begins with the same $g_0$, but the progression of growth rates is randomized across individual organisms, such that organisms with large $g$s would progress faster towards cancer onset, while those with small, or negative values of mutant $g$s would never progress to a cancerous state as the mutant gets selected against. This produces variation across organisms for cancer propensity, leading to a more realistic model.

**Predictions from the selection models.** As Figure 1 shows, under the assumption of context-independent selection, the incidence of cancer shows a strong threshold relationship with age, where cancer is unlikely or rare up to a certain age, and increases rapidly to 100% with a relatively short span of time. This is indicated by the fact that the age-specific crude incidence falls sharply to zero at some point in the lifespan, when the cumulative incidence also reaches 100%. As with the "bad luck" model, incidence has a threshold relationship wtih both $n$ and $p$, in terms of the age at half-maximum incidence (Figure 3C and G) and the maximum cumulative incidence, $I_{max}$ (Figure 3D and H). Moreover, saturation of incidence occurs only at 100%, which is identical to the "bad luck" model, despite the inherently stochastic implementation of mutation occurrence. Where the incidence of cancer is near the realistic range, for small values of $p$ and $n$, the late-life decrease in incidence is still not reflected in the context-independent selection case. The prediction of 100% incidence is due to the fact that all organisms in the population share the same growth rate progression for mutants. No other parameter in the model inherently precludes the accumulation of all k oncogenic mutations in some organisms, and the dynamics of accumulation is entirely a function of $p$ and $n$, along with stochastic variation. The context-independent selection model thus predicts that with increasing $p$ and/or $n$, cancer incidence increases, and reaches saturation only at 100% incidence. Therefore, clonal expansion accounts for some physiologically-relevant phenomena, like competitive growth of mutants, its description of cancer incidence and the effects of model parameters are either unrealistic, or incomplete.

As opposed to the context-independent selection model, the context-dependent model produces a saturating trend in cumulative incidence that begins to saturate at a level much lower than 100%. As Figure 4 shows, the saturation limit for many values of n and p is quite close to the epidemiological estimate of cancer risk (20-30%). This is an important feature of the context-dependent model, as it allows the model to generate more realistic patterns in age-specific cancer incidence. The realistic saturation of population incidence can be explained by the fact that propensity for clonal expansion varies across organisms, such that cancer progression occurs very quickly in some organisms, and not at all in others. Of the three models analysed so far, only the context-dependent model captures a trend similar to the late-life decline observed in many cancers in humans as the crude incidence curves in Figure 4A and E show, which suggests alternative explanations for trends in cancer incidence, independent of possible evovled cancer defenses.

## Sensitivity of predictions

For all the simulations so far, we have assumed $k = 5$ under all conditions, randomizing only the other parameters. Real populations, on the other hand, could potentially be a heterogenous mix of $k$ across organisms, which might affect the model predictions. To explore this possibility, we co-randomized $k$ with either $n$ or $p$ while maintaining the same growth rate progression, and looked at the association of $log(n)$ or $log(p)$ with the time taken for cancer onset. The time taken to cance is a useful parameter as it describes temporal mutation accumulation dynamics directly, while allowing for some limited inference regarding total incidence in the population. If time taken to cancer is largely small, population incidence is likely to be large in the given parameter space. Broadly, we find that cancer incidence is significant only up to maximum $k = 10$. Increasing $k$ also reduces total incidence and shifts the observed time to cancer to later in life (Figure 5), as expected based on mutation accumulation. We also find that the magnitude of $k$ could modulate the strength of the association $n$, and to a lesser extent, with $p$. Doing the same for the context-dependent case, we randomized $g$ as explained earlier, along with either $k$ and $n$, or $k$ and $p$. Remarkably, introducing $g$ as a random variable leads to most of the variance in time to cancer being explained by $g$, and to some extent, $n$. Together, this indicates that independent of the kind of selection assumed in the model, $k$ influences the expected effect of $n$, $p$ and $g$ on time to cancer, and that $g$, when randomized, is a stronger predictor than $p$ or $n$ alone.

We have also assumed a normal distribution for the growth rate progression in the simulations so far. Realistic estimates of the distribution of somatic mutant growth rates are currently unknown, but heavy-tailed distributions like the Gumbel distribution have been considered a possibility (28). We have reevaluated the predictions of the context-dependent selection model, for two distributions of $g$, Gumbel and uniform. We find that the shape of the distribution does not affect the nature of the relationships predicted by the model, although some quantitative differences ocur as expected (Figure 7.

## Discussion

On the whole, the better prediction profile of the context-dependent model stems from the distribution of $g$ in the pop-

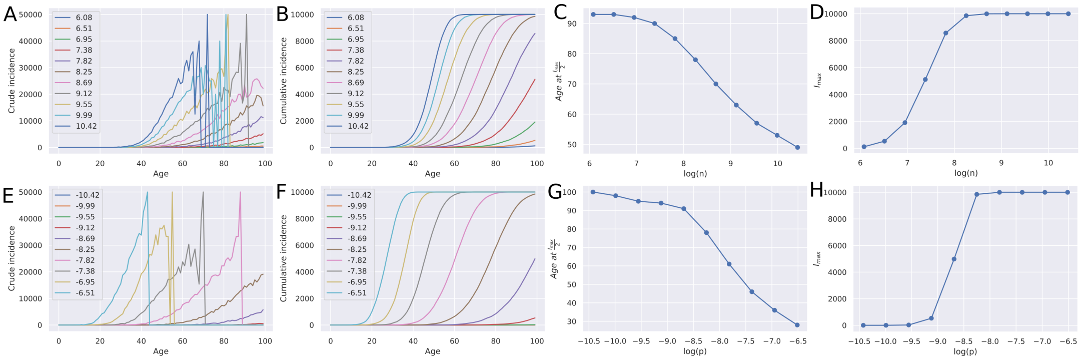

**Fig. 3.** Incidence patterns from the context-independent selection model over the range of (A-D) $n$, and (E-H) $p$. From left to right in each row, the plots are of (A, E) age-specific crude incidence per 100000 vs age, (B, F) cumulative incidence for the simulated population vs age, (C, G) age at which half the maximum incidence is reached vs $log(n)$ or $log(p)$, and (D, H) the maximum cumulative incidence, $I_{max}$ vs $n$ or $p$. Inset legends for the age curves are $log(n)$ and $log(p)$ in the top and bottom row respectively. For A-D, $p = 5.603 * 10^{-9}$, and for C-H, $n = 1.785 * 10^{8}$. Insets for the incidence curves show $log(n)$ and $log(p)$ in the top and bottom rows respectively. Growth rates progress linearly in the general form, $g_i = 0.007 * (i + 1)$, where $i = 0, ..., k$ and $k = 5$.

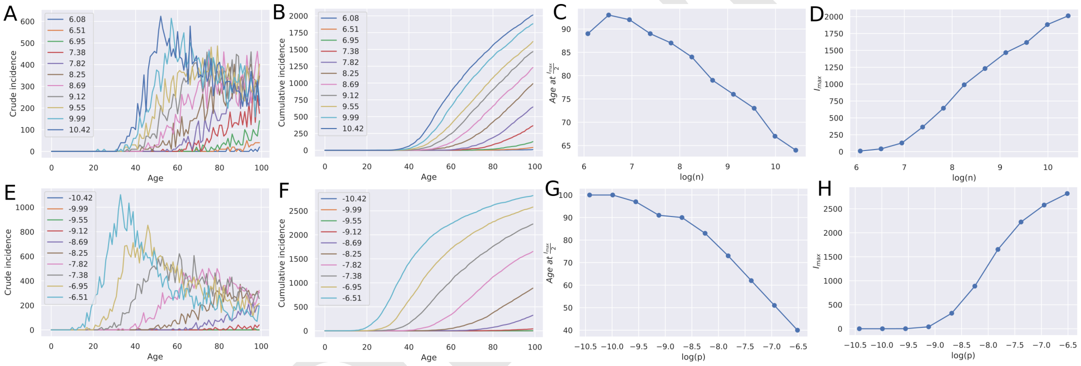

**Fig. 4.** Incidence patterns from the context-dependent selection model over the range of (A-D) $n$, and (E-H) $p$. From left to right in each row, the plots are of (A, E) age-specific crude incidence per 100000 vs age, (B, F) cumulative incidence for the simulated population vs age, (C, G) age at which half the maximum incidence is reached vs $log(n)$ or $log(p)$, and (D, H) the maximum cumulative incidence, $I_{max}$ vs $n$ or $p$. Inset legends for the age curves are $log(n)$ and $log(p)$ in the top and bottom row respectively. For A-D, $p = 5.603 * 10^{-9}$, and for C-H, $n = 1.785 * 10^{8}$. Insets for the incidence curves show $log(n)$ and $log(p)$ in the top and bottom rows respectively. Growth rates progress linearly from $g_0 = 0.007$ to $g_k = 0.007 * \mu$, where $\mu$ is normally-distributed random variable with $\overline{\mu} = 0$ and $\sigma = 3$, and $k = 5$.

**Table 1. Summary of model predictions vs observed trends**

| **Epidemiological observation** | "Bad luck" | Context-independent selection | Context-dependent selection |
|---|---|---|---|
| **Total incidence across cancer types does not exceed 30%** | Saturation at 100% only | Saturation at 100% only | Saturation<100% possible |
| **Age-specific incidence decreases in old age for several cancers** | No late-life decline | No late-life decline | Late-life decline predicted |
| **Incidence vs $n$** | Strong threshold saturating at 100% | Strong threshold saturating at 100% | Progressive increase & saturation<100% |
| **Incidence increases progressively with $p$, but not indefinitely** | Strong threshold saturating at 100% | Strong threshold saturating at 100% | Progressive increase & saturation<100% |
| **Non-mutagenic carcinogens increase cancer risk, but not the mutation rate** | Incompatible | Incompatible | Explained based on $g$ distribution |
| **Peto's paradox** | Insufficient | Insufficient | Cancer risk saturates with $n$ |

ulation, and this has interesting implications for the kind of causal factors that are important in explaining cancer etiology. One of these implications concern the mutation-centric thinking that characterizes a good portion of current opinion in cancer biology. For instance, several growth factors and hormones are known to increase cancer risk, not the least of which is insulin, without increasing the basal mutation rate or the cell number. The action of such "non-mutagenic carcinogens" is not compatible with a mutation-centric approach to carcinogenesis. We propose instead that parameters like $g$, which reflect context-dependent selection on mutants, offer better explanative scope. Similarly, late-life incidence in the model comes from a slow growth rate progression across mutations, and a late-life decline can therefore be described in terms of the distribution of the growth rate progression. In a population where slow growth rate progression for mutants is rare, organisms with active mutagenesis progress to cancer relatively earlier in life, and cancer is rarer later in life. It becomes important therefore, to consider the local or evolutionary factors underlying such a temporal pattern of mutation accumulation. This line of thinking can also be extended to address Peto's paradox. It is not a new argument that Peto's paradox can be explained by invoking selection (18, 24, 25). The results from our model in fact reinforce the notion that the processes underlying Peto's paradox, and other related observations, are largely selection-limited, which explains why cancer risk does not increase in proportion with cell number and/or body size. Moreover, our analysis introduces the additional dimension of population-level variation in cancer propensity as part of the explanation of Peto's paradox, and accounting for this variation that lies on top of evolved cancer defences leads to a more nuanced view of the paradox. This selection is imposed by the micro-environment in the tumour or pre-cancerous niche, and includes all the factors that determine the selective advantage of oncogenic or precancerous mutant. As mentioned earlier, empirical evidence of such context dependence has been accumulating on multiple fronts. Breast cancer literature is particularly rich in this regard; levels of estrogen and progesterone affect production of growth factors and ECM components by cells (29–31), and important observations have also been made suggesting synergistic action between hormonal and cytokine-based regulation of cell growth (32). The substantial scope of these interactions has been comprehensively reviewed, both for breast cancer particularly (33), and more generally across cancer types (13, 14, 34). For instance, the prohibitive effect of estradiol in ER-negative tumours (35) points to the possibility that estradiol reduces the selective advantage of an ER-negative phenotype. More recently, experiments have been reported in which behaviourally-enriched environments or physical exercise seemed to show cancer-suppressive effects (15, 16) that correlate with secretion of particular growth factors. Independently, the growth factor composition of culture media is also found to markedly alter selective outcomes of mutant phenotypes in cell competition *in vitro* (36), and growth characteristics of pre-cancerous cell lines across cancer stages (37). Taken together, these observations offer potential explanations for context-dependent selection on oncogenic mutations, and suggest interesting avenues of transalational importance for cancer prevention and therapy.

A composite view of cancer etiology requires not only the incorporation of these, and other complexities in models, but also the comparative testing framework that continues to be rare in cancer literature, barring a few efforts (23). The value of such a framework is immense, as it allows for falsification of factors that make predictions contrary to observed data; this falsification is frequently more robust and informative than an indirect confirmation of potential causal factors. With our analysis, we hope to bring this framework back into the mainstream at a time when the availaibility of large-scale data spanning levels of biological organization is better than ever before, from the population to the various cellular "-omes". A comparative framework could prove more powerful now, in the context of robust data and computational techniques, and should therefore become a major focus for the cancer modeling effort.

1. Armitage P, Doll R (1954) The Age Distribution of Cancer and a Multi-Stage Theory of Carcinogenesis. *British journal of cancer* 8(1):1–12.
2. Knudson AG (1971) Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences* 68(4):820–823.
3. Tomasetti C, Vogelstein B (2015) Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* 347(6217):78–81.
4. Tomasetti C, Li L, Vogelstein B (2017) Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* 355(6331):1330–1334.
5. McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA (2013) Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences* 110(8):2910–2915.
6. Blokzijl F, et al. (2016) Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* 538(7624):260–264.
7. Mina M, et al. (2017) Conditional Selection of Genomic Alterations Dictates Cancer Evolution and Oncogenic Dependencies. *Cancer Cell* 32(2):155–168.e6.
8. Dawkins R (1996) *The blind watchmaker : why the evidence of evolution reveals a universe without design.* (W. W. Norton & Company), pp. xvii, 358.
9. Hanahan D, Weinberg R (2000) The hallmarks of cancer. *Cell* 100(1):57–70.
10. Schäfer M, Werner S (2008) Cancer as an overhealing wound: An old hypothesis revisited. *Nature Reviews Molecular Cell Biology* 9(8):628–638.
11. Hanahan D, Weinberg R (2011) Hallmarks of Cancer: The Next Generation. *Cell* 144(5):646–674.
12. Nowell P (1976) The clonal evolution of tumor cell populations. *Science* 194(4260):23–28.
13. Hanahan D, Coussens LM (2012) Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. *Cancer Cell* 21(3):309–322.
14. Pietras K, Östman A (2010) Hallmarks of cancer: Interactions with the tumor stroma. *Experimental Cell Research* 316(8):1324–1331.
15. Cao L, et al. (2010) Environmental and Genetic Activation of a Brain-Adipocyte BDNF/Leptin Axis Causes Cancer Remission and Inhibition. *Cell* 142(1):52–64.
16. Rundqvist H, et al. (2013) Effect of Acute Exercise on Prostate Cancer Cell Growth. *PLoS ONE* 8(7):e67579.
17. Nagy JD, Victor EM, Cropper JH (2007) Why don't all whales have cancer? A novel hypothesis resolving Peto's paradox. *Integrative and Comparative Biology* 47(2):317–328.
18. Caulin AF, Maley CC (2011) Peto's Paradox: Evolution's prescription for cancer prevention.
19. Hochberg ME, Noble RJ (2017) A framework for how environment contributes to cancer risk. *Ecology Letters* 20(2):117–134.
20. American Cancer Society (2016) Cancer Facts & Figures 2016. *Cancer Facts & Figures 2016* pp. 1–9.
21. Harding C, Pompei F, Wilson R (2012) Peak and decline in cancer incidence, mortality, and prevalence at old ages. *Cancer* 118(5):1371–1386.
22. Hao D, Wang L, Di Lj (2016) Distinct mutation accumulation rates among tissues determine the variation in cancer risk. *Nature Publishing Group* (January):1–5.
23. Frank SA (2007) *Dynamics of Cancer. Incidence, Inheritance, and Evolution.* (Princeton University Press), pp. 1–378.
24. Noble R, Kaltz O, Hochberg ME (2015) Peto's paradox and human cancers. *Philosophical Transactions of the Royal Society B: Biological Sciences* 370(1673):20150104.
25. Tollis M, Boddy AM, Maley CC (2017) Peto's Paradox: How has evolution solved the problem of cancer prevention? *BMC Biology* 15(1):60.
26. Stephens PJ, et al. (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144(1):27–40.
27. Calabrese P, Shibata D (2010) A simple algebraic cancer equation: calculating how cancers may arise with normal mutation rates. *BMC Cancer* 10(1):3.
28. Durrett R, Foo J, Leder K, Mayberry J, Michor F (2010) Evolutionary dynamics of tumor progression with random fitness values. *Theoretical Population Biology* 78(1):54–66.
29. Haslam SZ, Woodward TL (2001) Tumour-stromal interactions Reciprocal regulation of extracellular matrix proteins and ovarian steroid activity in the mammary gland. *Breast Cancer Research* 3(6):365.
30. Woodward TL, Xie J, Fendrick JL, Haslam SZ (2000) Proliferation of Mouse Mammary Epithelial Cells in Vitro: Interactions among Epidermal Growth Factor, Insulin-Like Growth Factor I, Ovarian Hormones, and Extracellular Matrix Proteins 1. *Endocrinology* 141(10):3578–3586.
31. Dickson RB, Lippman ME (1987) Estrogenic Regulation of Growth and Polypeptide Growth Factor Secretion in Human Breast Carcinoma. *Endocrine Reviews* 8(1):29–43.

32. Freund A, et al. (2003) IL-8 expression and its possible relationship with estrogen-receptor-negative status of breast cancer cells. *Oncogene* 22(2):256–265.
33. Hansen R, Bissell MJ (2000) Tissue architecture and breast cancer: the role of extracellular matrix and steroid hormones. *Endocrine Related Cancer* 7(2):95–113.
34. Cabarcas SM, Mathews LA, Farrar WL (2011) The cancer stem cell niche-there goes the neighborhood? *International Journal of Cancer* 129(10):2315–2327.
35. Garcia M, Derocq D, Freiss G, Rochefort H (1992) Activation of estrogen receptor transfected into a receptor-negative breast cancer cell line decreases the metastatic and invasive potential of the cells. *Proceedings of the National Academy of Sciences of the United States of America* 89(23):11538–42.
36. Archetti M, Ferraro DA, Christofori G (2015) Heterogeneity for IGF-II production maintained by public goods dynamics in neuroendocrine pancreatic cancer. *Proc Natl Acad Sci U S A* 112(6):1–6.
37. Chan MT, et al. (2014) Effects of insulin on human pancreatic cancer progression modeled in vitro. *BMC Cancer* 14(1):814.

## Supporting Information (SI)

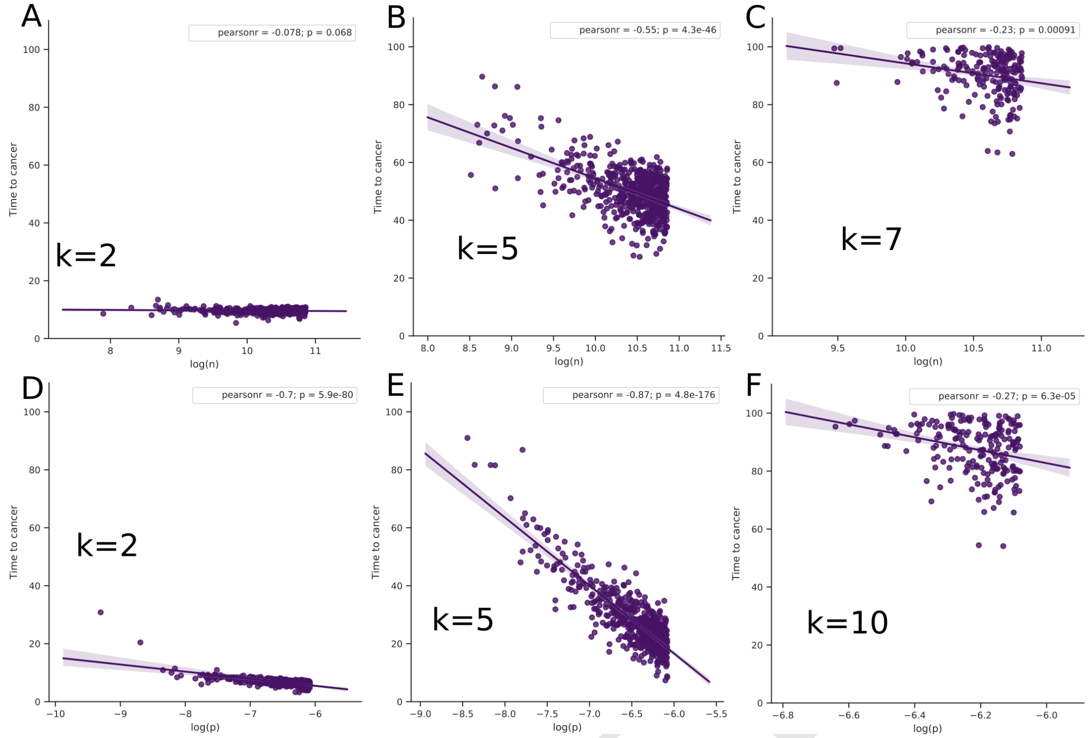

**Fig. 5.** Effect of $k$ in the context-independent selection case. The plots are time to cancer onset against $log(n)$ or $log(p)$, with $k$ randomized with (A-C) $n$, or (D-F) $p$; value of $k$ in the inset corresponds to the number of threshold oncogenic mutations assumed for the corresponding points. From A to C, for higher threshold of oncogenic mutations, the effect of $n$ on time to cancer gets stronger, as shown by the improvement in the association. For small $k$ however, $n$ does not affect the age of cancer onset. On the other hand, $p$ has a strong effect on the time to cancer at every value of $k$ considered. $k$, $n$ and $p$ were uniformly-distributed random variables with ranges $[0, 20]$, $[1.203 * 10^6, 2.649 * 10^{10}]$, and $[3.775 * 10^{-11}, 3.059 * 10^{-7}]$ respectively. For (A-C), $p = 5.603 * 10^{-9}$. For (D-F), $n = 1.785 * 10^8$.

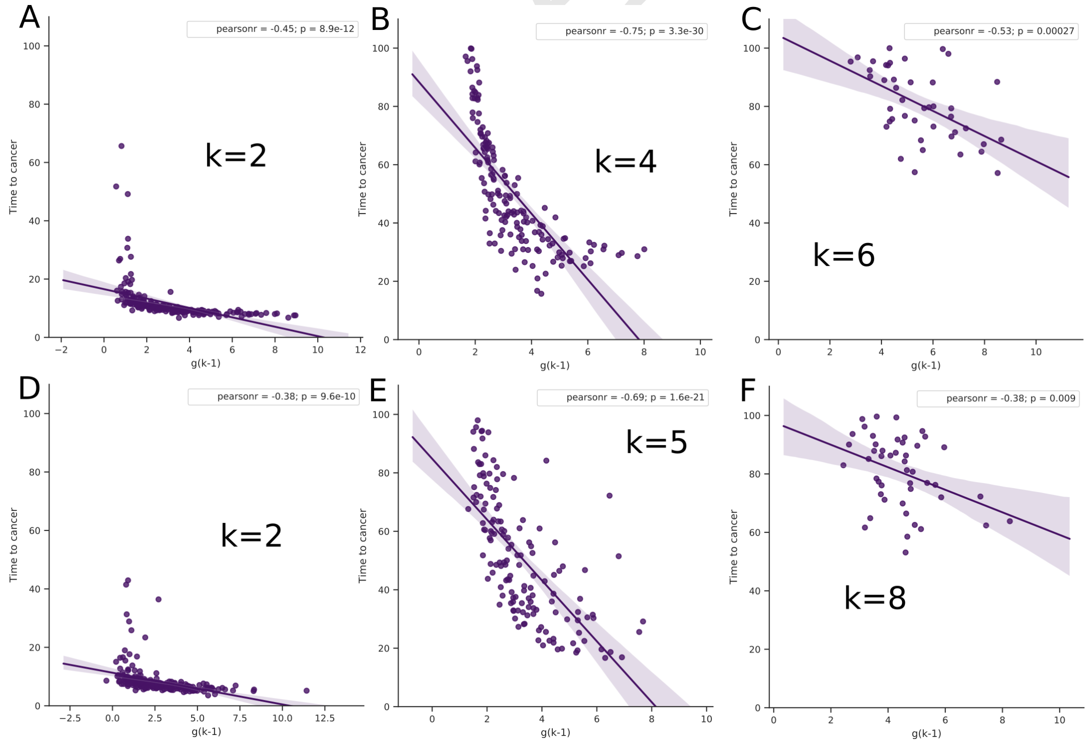

**Fig. 6.** Effect of $k$ in the context-dependent selection case. The plots are time to cancer onset against $g(k-1) = 0.007 * \mu$ for the last oncogenic mutation, with $k$ randomized with (A-C) $n$, or (D-F) $p$; value of $k$ in the inset corresponds to the number of threshold oncogenic mutations assumed for the corresponding points. Compared to Figure 5, $g(k-1)$ explains variance in time to cancer much better than either $n$ or $p$. This is true of both (A-C) when $n$ and $k$ are also randomized, and (D-F) when $p$ and $k$ are also randomized. The effect of $g(k-1)$ is nevertheless modulated by the required $k$. Ranges of $k$, $n$ and $p$ are the same as in Figure 5. For (A-C), $p = 5.603 * 10^{-9}$. For (D-F), $n = 1.785 * 10^8$.

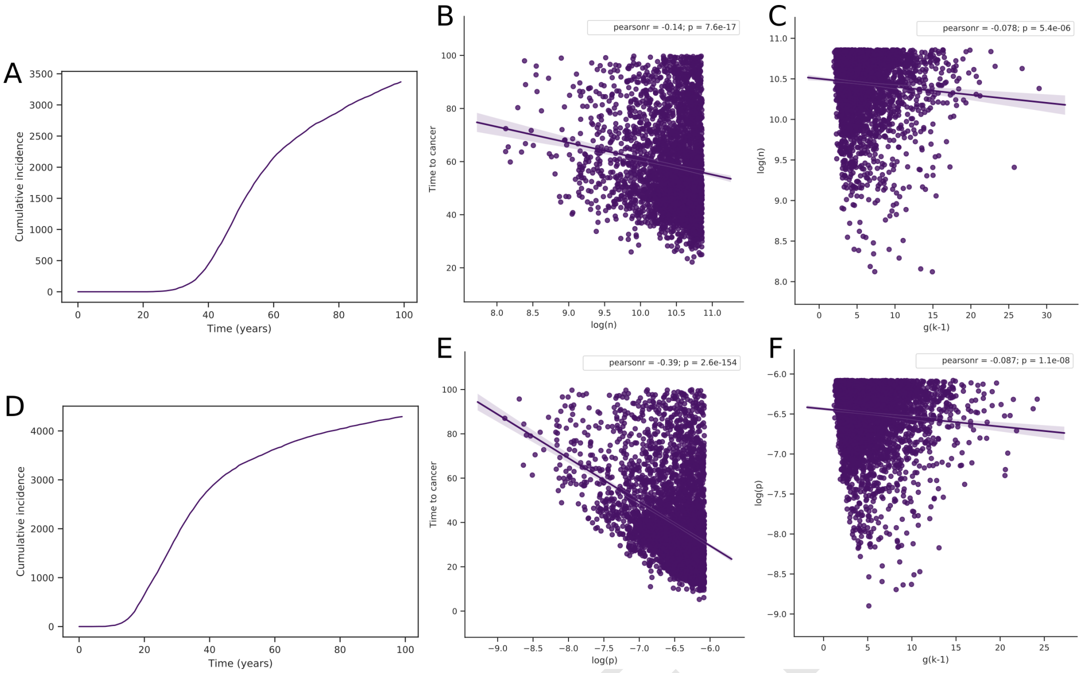

**Fig. 7.** $g$ dsitributed as a Gumbel distribution with $\overline{\mu} = 0$ and $\sigma = 3$; (A and D) cumulative incidence for the simulated population vs age, (B) time to cancer vs $log(n)$, (C) $log(n)$ vs $g(k-1) = 0.007 * \mu$ for the last mutation, (E) time to cancer vs $log(p)$, and (F) $log(p)$ vs $g(k-1)$. For all cases, $k = 5$.

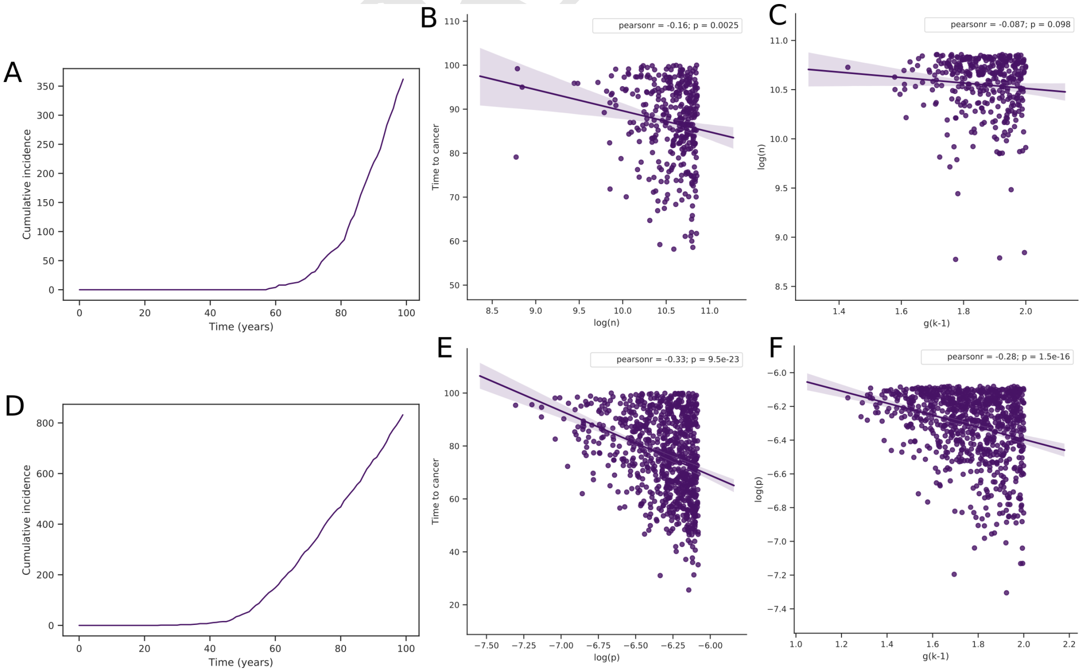

**Fig. 8.** $g$ dsitributed as a uniform distribution with range $[-3, 3]$; (A and D) cumulative incidence for the simulated population vs age, (B) time to cancer vs $log(n)$, (C) $log(n)$ vs $g(k-1) = 0.007 * \mu$ for the last mutation, (E) time to cancer vs $log(p)$, and (F) $log(p)$ vs $g(k-1)$. For all cases, $k = 5$.