ARTICLE

doi:10.1038/nature16166

Substantial contribution of extrinsic risk factors to cancer development

Song Wu[1,2], Scott Powers[1,2,3], Wei Zhu[1,2] & Yusuf A. Hannun[2,3,4,5]

Recent research has highlighted a strong correlation between tissue-specific cancer risk and the lifetime number of tissue-specific stem-cell divisions. Whether such correlation implies a high unavoidable intrinsic cancer risk has become a key public health debate with the dissemination of the 'bad luck' hypothesis. Here we provide evidence that intrinsic risk factors contribute only modestly (less than ~10–30% of lifetime risk) to cancer development. First, we demonstrate that the correlation between stem-cell division and cancer risk does not distinguish between the effects of intrinsic and extrinsic factors. We then show that intrinsic risk is better estimated by the lower bound risk controlling for total stem-cell divisions. Finally, we show that the rates of endogenous mutation accumulation by intrinsic processes are not sufficient to account for the observed cancer risks. Collectively, we conclude that cancer risk is heavily influenced by extrinsic factors. These results are important for strategizing cancer prevention, research and public health.

Cancers were once thought to originate from mature tissue cells that underwent dedifferentiation in response to cancer progression[1]. Today, cancers are proposed to originate from the malignant transformation of normal tissue progenitor and stem cells[2,3], although this is not wholly accepted[4]. Nevertheless, recent research has highlighted a strong correlation of 0.81 between tissue-specific cancer risk and the lifetime population size in cumulative number of cell divisions of tissue-specific stem cells[5]. However, there has been controversy regarding the conclusion that this correlation implies a very high unavoidable risk for many cancers that is due solely to the intrinsic baseline population size of tissue-specific stem cells[6,7]. Many arguments against the 'bad luck' hypothesis have been made[5–13], yet none of these have offered specific alternatives to quantitatively evaluate the contribution of extrinsic risk factors in cancer development. Applying several distinct modelling approaches, here we provide strong evidence that unavoidable intrinsic risk factors contribute only modestly (less than ~10–30%) to the development of many common cancers.

We made the conservative and yet conventional assumption that errors occurring during the division of cells, being routes of malignant transformation, can be influenced by both intrinsic processes as well as extrinsic factors (Fig. 1). 'Intrinsic processes' include those that result in mutations due to random errors in DNA replication, whereas 'extrinsic factors' are environmental factors that affect mutagenesis rates (such as ultraviolet (UV) radiation, ionizing radiation and carcinogens). For example, radiation can cause DNA damage, which would primarily result in deleterious mutations with functional consequences on cancer development only after cell division. Therefore, extrinsic factors may act through the accumulation of genetic alterations during cell division to increase cancer risk. Accordingly, cancer risk would result from those apparently uncontrollable intrinsic processes (Fig. 1, arrow 1) as well as from those highly modifiable and thus preventable extrinsic factors (Fig. 1, arrow 2).

Correlation cannot differentiate risks

According to the above hypothesis, both intrinsic and extrinsic factors can impart cancer risk through the accumulation of these errors, especially the 'driver mutations' (Fig. 1, arrow 3). As such, a correlational analysis between cancer risk and cell division, for either stem or non-stem cells, is unable to differentiate between the contributions of intrinsic and extrinsic factors. This is best illustrated through a thought experiment where we consider a hypothetical scenario of a sudden global emergence of a very potent mutagen, such as a strong radiation burst from a nuclear fallout, which quadruples the lifetime risks for all cancers. In this scenario, it transpires that the proportion of cancer risk caused by intrinsic random errors would be small (at most one-quarter if we assume all of the original risk was due to intrinsic processes). However, if we conduct regression analyses on either the new hypothetical cancer risks or the current cancer risks as reported, against the number of stem-cell divisions[5], the correlations from both cases would be 0.81 (Fig. 2). This thought experiment negates the ability of the correlation to detect solely the contribution of intrinsic factors as it cannot distinguish between intrinsic and extrinsic factors. Thus, it argues against the implication that around two-thirds of variation could be explained by division-related random intrinsic errors.

Lower bound intrinsic risk line

The above conclusion then raises the question of what proportion of total cancer risk is due to extrinsic versus intrinsic factors. In a data-driven approach, we first re-examined the quantitative relationship between the observed lifetime cancer risk and the divisions of the

Figure 1 | Schematic showing how intrinsic processes and extrinsic factors relate to cancer risks through stem-cell division. This hypothesis maintains the strong role of stem-cell division in imparting cancer risk, but it also illustrates the potential contributions of both intrinsic and extrinsic factors operating through stem-cell division. Other effects, for example, through division of non-stem cells, are considered later in this analysis.

[1]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York 11794, USA. [2]Stony Brook Cancer Center, Stony Brook University, Health Sciences Center, Stony Brook, New York 11794, USA. [3]Department of Pathology, Stony Brook University, Health Sciences Center, Stony Brook, New York 11794, USA. [4]Department of Medicine, Stony Brook

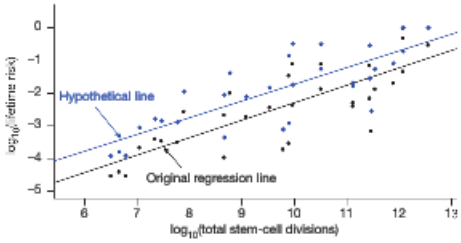


Figure 2 | Correlation analysis of stem-cell division and cancer risk does not distinguish contribution of extrinsic versus intrinsic factors to cancer risk. The black dots are data from figure 1(also shown in supplementary table 1) of Tomasetti & Vogelstein[5], and the black line shows their original regression line. The blue diamonds represent the hypothesized quadrupled cancer risks due to hypothetical exposure to an extrinsic factor such as radiation. The blue regression line for the hypothetical risk data maintains the same correlation as the original black line, albeit reflecting a much higher contribution of extrinsic factors to cancer risk.

normal tissue stem cells as reported[5], with a distinct alternative method. Our rationale was that intrinsic risk, or indeed its upper bound, can be better estimated by the lowest boundary on the plots of cancer risk versus total tissue stem-cell divisions (Fig. 3a, red 'intrinsic' risk line), meaning that intrinsic cancer risk should be determined by the cancer incidence for those cancers with the least risk in the entire group controlling for total stem-cell divisions (Fig. 3a, red dots). The argument here is that cancers with the same number of stem-cell divisions should share the same base of intrinsic cancer risk (if the relationship is causal); if one or more cancers would feature a much higher cancer incidence, for example, lung cancer among smokers versus non-smokers, then this probably reflects additional (and probably extrinsic) risk factors (smoking in this case). One could argue that the low-incidence tumour types may have lower incidences because of additional genetic repair mechanisms that restrict evolving malignant cells from accumulating sufficient numbers of genetic alterations required to become fully tumorigenic; however, without more specific data on the operation of repair mechanisms, these could drive the risk up or down, depending on whether they are less or more efficient in any particular tissue. According to our hypothesis, intrinsic risk from stem-cell divisions would define the lowest bound for a given number of stem-cell divisions, therefore we define an 'intrinsic' risk line for stem-cell divisions by regressing the smallest cancer risks on any given number of stem-cell divisions (Fig. 3a, red line). The 'intrinsic' risk lines themselves are still probably overestimates for the intrinsic risk; however, we should suspect that any cancer risk above that line implies additional biologic determinants, on the basis of which we can compute the percentage of cancer risk not explained by intrinsic 'randomness'. As shown in Fig. 3a, most cancer types have very high excess risks relative to the 'intrinsic' risk line, indicating large proportions of risks that are unaccounted for by the intrinsic factors, typically larger than 90%. Moreover, these estimated excess risks are very robust: with plausible measurement errors added to the total stem-cell divisions, the resulting excess risks remain essentially intact (Extended Data Table 1).

Extrinsic risks by tissue cell turnover

Although we performed the initial analysis from a 'stem-cell theory' point of view, we wanted to evaluate if our results are dependent on this specific theory or independent of it. Furthermore, the lack of reliable data on human tissue stem-cell dynamics is a notable concern (see Supplementary Information), rendering the analysis in Fig. 3a less determinate. Thus, we separately collected data for the total number of tissue cell divisions that is based on homeostatic tissue cell numbers and their turnover rates (see Supplementary Information), and analysed the relationship of cancer risk versus total tissue cell divisions (Fig. 3b). This approach allows for every dividing cell to be a potential cancer-initiating cell, which would be an application of another cell-of-origin theory of cancer whereby tumours may originate from a hierarchy of cells, from stem cells to committed progenitor cells to differentiated cells[4]. Mathematically, this can also be considered as an extreme form of stem-cell theory where the fraction of stem cells is 1 (this latter formulation then provides an upper bound of the effects of the size of the stem-cell population on cancer risk and the role of extrinsic factors). The regression analysis between cancer risk and total tissue cell division shows a high correlation of 0.75, establishing a strong quantitative relationship between cancer risk and total cell division. To dissect the extrinsic versus intrinsic risks, we applied the same rationale and regressed the smallest cancer risks on any given number of cell divisions (Fig. 3b, red line). Although we could only find reliable turnover data for a subset of tissues, it is remarkable that the conclusion drawn here is nearly identical to that in Fig. 3a; that is, large proportions of risks that may not be attributable to intrinsic factors are mostly higher than 90%. It is important to note that here we included breast and prostate cancers—two high-incidence cancers missing in the original stem-cell analysis[5]. Again, plausible measurement errors have been added to the total cell divisions, and the excess risks remained almost identical (Extended Data Table 1). In summary, irrespective of whether a subpopulation or all dividing cells contribute to cancer, these results indicate that intrinsic factors do not play a major causal role.

Epidemiological evidence

In parallel, numerous epidemiological studies have established strong evidence that many cancers have substantial risk proportions attributed to environmental exposures (Extended Data Table 2). Particularly, for breast and prostate cancers, it has long been observed that large international geographical variations exist in their incidence rates (for example, Western Europe has the highest incidence of breast cancer, which is almost 5 times higher than areas such as Eastern Asia or Middle Africa; Australia/New Zealand has the highest incidence of prostate cancer, which is almost 25 times higher than areas such as South-Central Asia)[14], and immigrants moving from countries with lower cancer incidence to countries with higher cancer rates soon acquire the higher risk of their new country[15,16]. While several risk factors have been identified for these cancers, no single one can account for their substantial extrinsic risk proportions, suggesting complex mechanisms for their aetiologies. Colorectal cancer is a high-incidence cancer that is widely considered to be an environmental disease[17], with an estimated 75% or more of colorectal cancer risk attributable to diet[18]. For many other cancers, known environmental risk factors have also been identified. For example, for melanoma the risk ascribed to sun exposure is around 65–86%[19], and for non-melanoma basal and squamous skin cancers ~90% is attributable to UV radiation[20]. At least 75% of oesophageal cancer, or head and neck cancer, is caused by tobacco and alcohol[21,22]. It is also well known that certain pathogens may markedly increase the risk of cancers. For instance, human papilloma virus may cause ~90% of cervical cancer cases[23], ~90% of anal cancer cases[24] and ~70% of oropharyngeal cancer cases[25]; hepatitis B and C may account for ~80% of hepatocellular carcinoma cases[26]; and *Helicobacter pylori* may be responsible for 65–80% of gastric cancer cases[27]. These, along with many other reports, provide direct evidence that environmental factors play important roles in cancer incidence and they are modifiable through lifestyle changes and/or vaccinations.

Additionally, analyses of data from the Surveillance, Epidemiology, and End Results Program (SEER) in the USA between 1973–2012 demonstrate that while many cancers have declining or maintain relatively consistent age-adjusted incidence rates (for example, cervical, gallbladder and oesophageal cancers, Extended Data Fig. 1), incidences of some cancers (including melanoma, thyroid, kidney, liver, thymus,

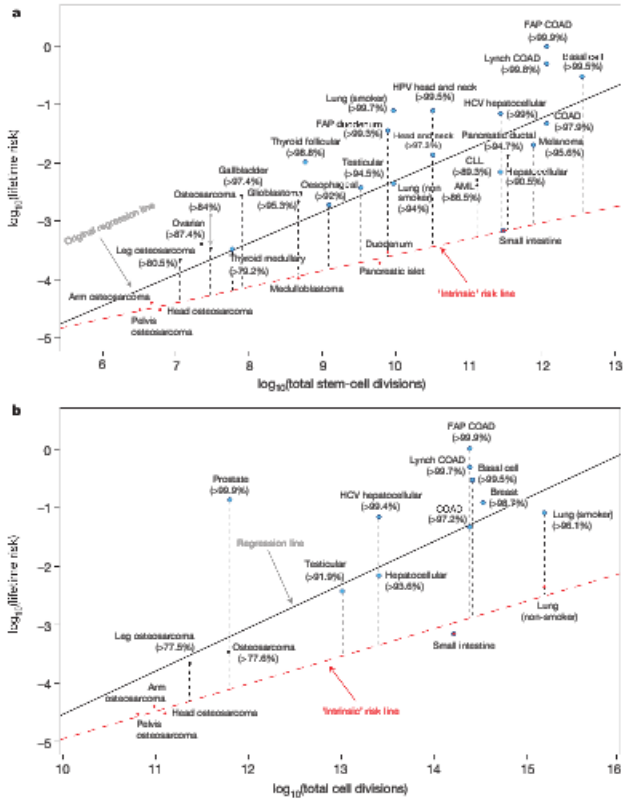


Figure 3 | Estimation of the proportion of lifetime cancer risk that is not due entirely to 'bad luck'. a, b, Estimations based on total tissue stem-cell divisions originally reported in Tomasetti & Vogelstein[5] (**a**) and total tissue cell divisions (**b**). Red dots are cancers used to compute the 'intrinsic' risk linear regression lines (red dashed lines). Blue dots are cancers known to have substantial extrinsic risks from epidemiology studies. The numbers in parentheses are the estimated percentages of cancer risks that are due to factors other than intrinsic risks.

small intestine, extranodal non-Hodgkin lymphoma, testicular, anal and anorectal cancers) have been steadily increasing, and their current incidences are substantially higher than their historical minima in the past 40 years[28] (Extended Data Fig. 1). Moreover, the mortality trend of lung cancer from 1930–2011 (ref. 29), which usually mirrors its incidence trend, shows a more than 15-fold increase for lung cancer risk. These substantial increases in incidence suggest that large risk proportions are attributable to changing environments (for example, smoking and air pollutants and their role in the risk of developing lung cancer). Collectively, nearly all major cancers have been covered in these epidemiological studies, further supporting the hypothesis of substantial extrinsic risks for most cancers. Notably, most of these cancers from the epidemiological and SEER results, except for small intestine, are located above the red 'intrinsic' risk lines in Fig. 3a, b (blue points). Accounting for the external factors would move them closer to the proposed 'intrinsic' line, further supporting the conjecture that the intrinsic line is mainly defined by cancers without compelling known epidemiological risk, whereas those above are at higher risks owing to extrinsic factors.

Analysis of mutational signatures

In addition to epidemiological studies, we evaluated recent studies on mutational signatures in cancer. These are regarded as 'fingerprints' left on cancer genomes by different mutagenic processes[30], revealing ~30 distinct signatures among various cancers[31]. Analysis of these signatures was therefore used to shed light on the proportion of intrinsic versus extrinsic origins of cancer. Two signature mutations, 1A/1B (see ref. 31), demonstrated strong positive correlations with age in the majority of cancers, suggesting that they are acquired at a relatively constant rate over the lifetime of cancer patients and thus probably result from intrinsic processes; however, all other signature mutations (~30) lack the consistent correlations with age, suggesting that they are acquired at different rates in life and thus are probably a consequence of extrinsic carcinogen exposures[31]. Indeed, several mutational signatures have been linked to known factors such as UV radiation and smoking[31]. We therefore categorized the signatures into intrinsic (type 1A/1B) and extrinsic mutations with known or unknown factors, and summarized their corresponding percentages in Extended Data Table 3. Notably, many cancers have substantial extrinsic mutations with

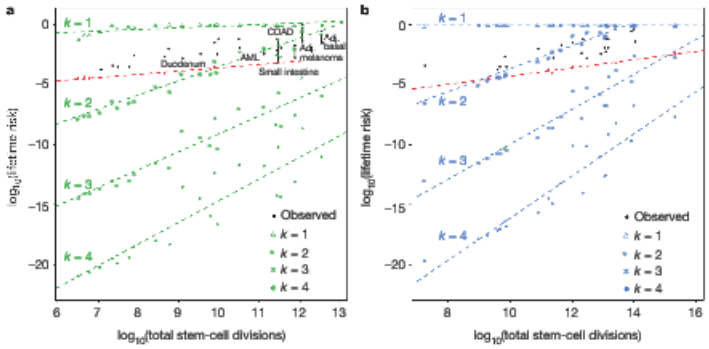


Figure 4 | Theoretical lifetime intrinsic risks (tLIR) for cancers based on different number of hits (k) required for cancer onset. a, b, The green (**a**) and blue (**b**) dashed lines are the 'intrinsic' risk lines estimated on the basis of total reported stem-cell numbers and total homeostatic tissue cells, respectively. The intrinsic stem-cell mutation rate (r) is assumed to be 1×10^{-8} per cell division. The red dashed lines are the 'intrinsic' risk lines estimated on the basis of the observed data using the same mechanism as Fig. 3a. Adjusted (adj.) basal and adjusted melanoma represent cancer risks after adjusting for the effect of sun exposure and UV radiation. AML, acute myeloid leukaemia.

known factors. More importantly, cancers known to have substantial environmental risk proportions, for example, breast cancer[15], prostate cancer[16], colorectal cancer[18], melanoma[19], head and neck cancer[21], oesophageal cancer[22], cervical cancer[23], liver cancer[26] and stomach cancer[27], all harbour large percentages of total extrinsic mutational signatures. This suggests that the percentages of total extrinsic mutational signatures can serve as a good surrogate for extrinsic cancer risks. While a few cancers have relatively large proportions of intrinsic mutations (>50%), the majority of cancers have large proportions of extrinsic mutations, for example, ~100% for myeloma, lung and thyroid cancers and ~80–90% for bladder, colorectal and uterine cancers, indicating substantial contributions of carcinogen exposures in the development of most cancers.

Modelling theoretical lifetime intrinsic risk

Finally, in another independent model-driven approach to dissecting the risk contribution of the intrinsic processes, we modelled the potential lifetime cancer risk due to intrinsic stem-cell mutation errors by varying the number of hits (that is, driver gene mutations), denoted by k, required for cancer onset. We derived the probability distribution of the propagation of driver gene mutations from one generation to the next, and subsequently established the theoretical relationship between cell divisions and the degree of lifetime cancer risk due to intrinsic cell mutation errors alone, which we refer to as the theoretical lifetime intrinsic risk (tLIR). To overcome the limitation of inaccurate estimation in the reported stem-cell numbers[5], we calculated tLIR using both the reported stem-cell number (tLIRsc) and the total tissue cell number (tLIRtt). The latter is equivalent to assuming all homeostatic tissue cells to be stem cells, representing an extreme overestimation of tissue stem cells, which consequently leads to a conservative estimation of the upper bounds in tLIR. The somatic mutation rate in tumours is estimated to be 5×10^{-10} per nucleotide site per cell division[32–34]. On this basis, in our initial calculation we used an intrinsic mutation rate (r) of 1×10^{-8} per cell division, which is equivalent to approximately 20 mutable nucleotide sites for each driver gene where the driver gene will mutate if at least one site mutates. As shown in Fig. 4a, b, if only one hit (that is, mutation of one designated driver gene) is required to develop cancer—that is, $k = 1$—the lifetime risk for almost all cancers is close to 100%. This confirms that one mutation is not enough for cancer onset (otherwise everyone would theoretically acquire each type of cancer). If two driver gene mutations are needed, $k = 2$, the modelled intrinsic risk becomes small for cancers with a small total number of stem-cell divisions; however, it is still very large for those with higher stem-cell divisions, and even unreasonably large for some cancers by surpassing the corresponding observed total lifetime cancer risks (adjusted basal cell carcinoma, colon adenocarcinoma, adjusted melanoma, small intestine cancer, acute myeloid leukaemia and duodenal cancer; Fig. 4a). It is therefore unlikely that, at least in these cancers, two hits will suffice to induce cancer. As shown in Fig. 4, if we consider the more reasonable case where three mutations are required[35], $k = 3$, almost all modelled intrinsic risks (both tLIRsc and tLIRtt) drop well below our earlier 'intrinsic' risk lines estimated conservatively from the observed data alone (red dashed lines, estimated based on observed data following the same mechanism as Fig. 3a). The lifetime risk drops even further for $k = 4$ and beyond. The extrinsic risks based on the tLIRsc and tLIRtt are further summarized in Extended Data Table 4. This modelling approach demonstrates that cancer risk due to intrinsic stem-cell mutation errors alone is low for almost all cancers that require over two mutations, indeed it is lower than the relatively conservative estimate based on data alone (red lines, Fig. 4). As the driver gene mutation rate in stem-cell division is a key parameter, we further conducted sensitivity analyses with different rates ($r = 1 \times 10^{-10}$ to 1×10^{-6}) to examine how this may affect the tLIR (Extended Data Figs 2 and 3). The results show that for $k = 3$, when $r < 1 \times 10^{-7}$ (~200 sites for each driver gene hit), almost all modelled intrinsic risks are below the observed 'intrinsic' risk line (red lines); when $r = 1 \times 10^{-6}$ (~2,000 sites for each driver gene hit), the majority of modelled intrinsic risks are still well below the observed 'intrinsic' risk lines, particularly those with small total number of divisions (Extended Data Fig. 2). For $k = 4$, when $r < 1 \times 10^{-6}$, almost all modelled intrinsic risks are below the observed 'intrinsic' risk lines estimated through the data-driven approach (Extended Data Fig. 3). These sensitivity analyses demonstrate that our conclusions are highly robust, and that the attribution of intrinsic mutations to lifetime cancer risk through stem-cell divisions, particularly for those cancers with low risk, is rather small, even using widely different intrinsic mutation rates.

In summary, we find that a simple regression analysis cannot distinguish between intrinsic and extrinsic factors. We have provided a new framework to quantify the lifetime cancer risks from both intrinsic and extrinsic factors on the basis of four independent approaches that are data-driven and model-driven, with and without using the stem-cell estimations. Importantly, these four approaches

provide a consistent estimate of contribution of extrinsic factors of >70–90% in most common cancer types. This is consistent with the overall conclusion regarding the role of extrinsic factors in cancer development.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 April; accepted 23 October 2015.
Published online 16 December 2015.

1. Sell, S. Stem cell origin of cancer and differentiation therapy. *Crit. Rev. Oncol. Hematol.* **51,** 1–28 (2004).
2. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414,** 105–111 (2001).
3. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255,** 197–200 (1975).
4. Visvader, J. E. Cells of origin in cancer. *Nature* **469,** 314–322 (2011).
5. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347,** 78–81 (2015).
6. Ashford, N. A. *et al.* Cancer risk: role of environment. *Science* **347,** 727 (2015).
7. Wild, C. *et al.* Cancer risk: role of chance overstated. *Science* **347,** 728 (2015).
8. Potter, J. D. & Prentice, R. L. Cancer risk: tumors excluded. *Science* **347,** 727 (2015).
9. Gotay, C., Dummer, T. & Spinelli, J. Cancer risk: prevention is crucial. *Science* **347,** 728 (2015).
10. Song, M. & Giovannucci, E. L. Cancer risk: many factors contribute. *Science* **347,** 728–729 (2015).
11. O'Callaghan, M. Cancer risk: accuracy of literature. *Science* **347,** 729 (2015).
12. Tomasetti, C. & Vogelstein, B. Cancer risk: accuracy of literature—response. *Science* **347,** 729–731 (2015).
13. Altenberg, L. Statistical problems in a paper on variation in cancer risk among tissues, and new discoveries. Preprint at http://arxiv.org/abs/1501.04605 (2015).
14. Torre, L. A. *et al.* Global cancer statistics, 2012. *CA Cancer J. Clin.* **65,** 87–108 (2015).
15. Gray, J., Evans, N., Taylor, B., Rizzo, J. & Walker, M. State of the evidence: the connection between breast cancer and the environment. *Int. J. Occup. Environ. Health* **15,** 43–78 (2009).
16. Shimizu, H. *et al.* Cancers of the prostate and breast among Japanese and white immigrants in Los Angeles County. *Br. J. Cancer* **63,** 963–966 (1991).
17. Haggar, F. A. & Boushey, R. P. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clin. Colon Rectal Surg.* **22,** 191–197 (2009).
18. Johnson, I. T. & Lund, E. K. Review article: nutrition, obesity and colorectal cancer. *Aliment. Pharmacol. Ther.* **26,** 161–181 (2007).
19. Parkin, D. M., Mesher, D. & Sasieni, P. 13. Cancers attributable to solar (ultraviolet) radiation exposure in the UK in 2010. *Br. J. Cancer* **105** (Suppl 2), S66–S69 (2011).
20. Koh, H. K., Geller, A. C., Miller, D. R., Grossbart, T. A. & Lew, R. A. Prevention and early detection strategies for melanoma and skin cancer. Current status. *Arch. Dermatol.* **132,** 436–443 (1996).
21. Blot, W. J. *et al.* Smoking and drinking in relation to oral and pharyngeal cancer. *Cancer Res.* **48,** 3282–3287 (1988).
22. Kamangar, F., Chow, W.-H., Abnet, C. & Dawsey, S. Environmental causes of esophageal cancer. *Gastroenterology Clin. North Am.* **38,** 27–57 (2009).
23. Bosch, F. X. *et al.* Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. *J. Natl. Cancer Inst.* **87,** 796–802 (1995).
24. Frisch, M. *et al.* Sexually transmitted infection as a cause of anal cancer. *N. Engl. J. Med.* **337,** 1350–1358 (1997).
25. Chaturvedi, A. K. *et al.* Human papillomavirus and rising oropharyngeal cancer incidence in the United States. *J. Clin. Oncol.* **29,** 4294–4301 (2011).
26. El-Serag, H. B. Epidemiology of viral hepatitis and hepatocellular carcinoma. *Gastroenterology* **142,** 1264–1273 (2012).
27. *Helicobacter* and Cancer Collaborative Group. Gastric cancer and *Helicobacter pylori*: a combined analysis of 12 case control studies nested within prospective cohorts. *Gut* **49,** 347–353 (2001).
28. Surveillance, Epidemiology, and End Results (SEER) Program. SEER*Stat Database: Incidence—SEER 9 Regs Research Data, Nov 2014 Sub (1973–2012) (National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, 2015).
29. Dela Cruz, C. S., Tanoue, L. T. & Matthay, R. A. Lung cancer: epidemiology, etiology, and prevention. *Clin. Chest Med.* **32,** 605–644 (2011).
30. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24,** 52–60 (2014).
31. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
32. Jones, S. *et al.* Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl Acad. Sci. USA* **105,** 4283–4288 (2008).
33. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl Acad. Sci. USA* **110,** 1999–2004 (2013).
34. Bozic, I. & Nowak, M. A. Unwanted evolution. *Science* **342,** 938–939 (2013).
35. Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G. & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc. Natl Acad. Sci. USA* **112,** 118–123 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank L. Obeid for constructive comments. This work was supported in part by NCI grants 97132 and 168409 and Stony Brook NYSTEM award C026716.

Author Contributions Y.A.H. formulated the hypothesis. S.W. and Y.A.H. designed the research. S.W. and W.Z. performed mathematical and statistical analysis. S.W., S.P., W.Z. and Y.A.H. performed research. S.W., S.P., W.Z. and Y.A.H. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.W. (Song.Wu@stonybrook.edu) or Y.A.H. (Yusuf.Hannun@sbumed.org).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Derivation of the probability of possessing *k* hits after *n* cell divisions for one cell. On the basis of the theory of the clonal stem-cell origin of cancer, in a given tissue the stem cell would first go through m rounds of symmetric divisions (for each division, each stem cell would divide into two daughter stem cells) to reach a total of S stem cells ($S=2^m$) at the steady state. Subsequently, these S stem cells would go through a rounds of asymmetric divisions (for each division, each stem cell would yield only one daughter stem cell) throughout the lifetime of the tissue. This means that the total rounds of lifetime stem-cell divisions per generation is $n=m+a$. Information on the total rounds of symmetric and asymmetric divisions as well as the total number of stem cells in the steady state for various tissues discussed in this work has been extracted from supplementary table 1 of Tomasetti & Vogelstein[5]. With k hits (mutations of k predetermined driver genes) on a stem cell required for cancer onset, the number of possible cell states of a given stem-cell generation would be $k+1$, including a zero state with no hit. If we assume that once a hit occurs it cannot be reversed and therefore be carried to all progeny cells, then a cell state may only transition from lower to higher or equal levels from generation to generation. In Extended Data Fig. 4, we demonstrate with $k=3$ the state transitions of accumulating driver gene mutations. Let X_g denote the number of driver gene mutations accumulated at generation g, and r be the intrinsic driver gene mutation rate due to random errors during DNA replication; the transition probabilities to generation $g+1$ with i mutations from the previous generation g with $j \le i$ mutations are derived as follows:

$$P(X_{g+1}=i)=\sum_{j=0}^{i} P(X_{g+1}=i|X_g=j)P(X_g=j)$$
$$=\sum_{j=0}^{i}\binom{k-j}{i-j}r^{i-j}(1-r)^{k-i}P(X_g=j)$$

In particular, for the emission state $i=0$:

$$P(X_{g+1}=0)=(1-r)^k P(X_g=0)$$

For the absorbing state $i=k$:

$$P(X_{g+1}=k)=\sum_{j=0}^{k} r^{k-j}P(X_g=j)$$

Based on these, the computing algorithm is derived as follows:
Set the initial cell state at generation 0:

$$P(X_0=0)=1;\ P(X_0=1)=0;\ \ldots;\ P(X_0=k)=0$$

For $g=1,\ldots,n$ and $0\le i\le k$, we compute the following probabilities iteratively:

$$P(X_g=i)=\sum_{j=0}^{i}\binom{k-j}{i-j}r^{i-j}(1-r)^{k-i}P(X_{g-1}=j)$$

where n is the total number of divisions that one stem cell may experience during its lifetime.

Derivation of the theoretical lifetime intrinsic risk (tLIR) of cancer for a given tissue. As mentioned previously, we assume stem cells in a specific tissue undergo two phases of divisions (Extended Data Fig. 5): (1) a total of m symmetric divisions before full tissue development, and (2) a total of a asymmetric divisions for normal tissue turnovers. So in a fully developed tissue, there is a total of $S=2^m$ stem cells. For each stem cell, the probability of possessing all k hits for cancer onset after $n=m+a$ rounds of divisions is $P(X_n=k)$, which can be calculated from the previous part. Therefore, the theoretical lifetime intrinsic risk (tLIR) of developing cancer—that is, the probability of at least one stem cell containing k hits during its lifetime—can be expressed as:

$$\text{tLIR}=1-[1-P(X_n=k)]^S$$

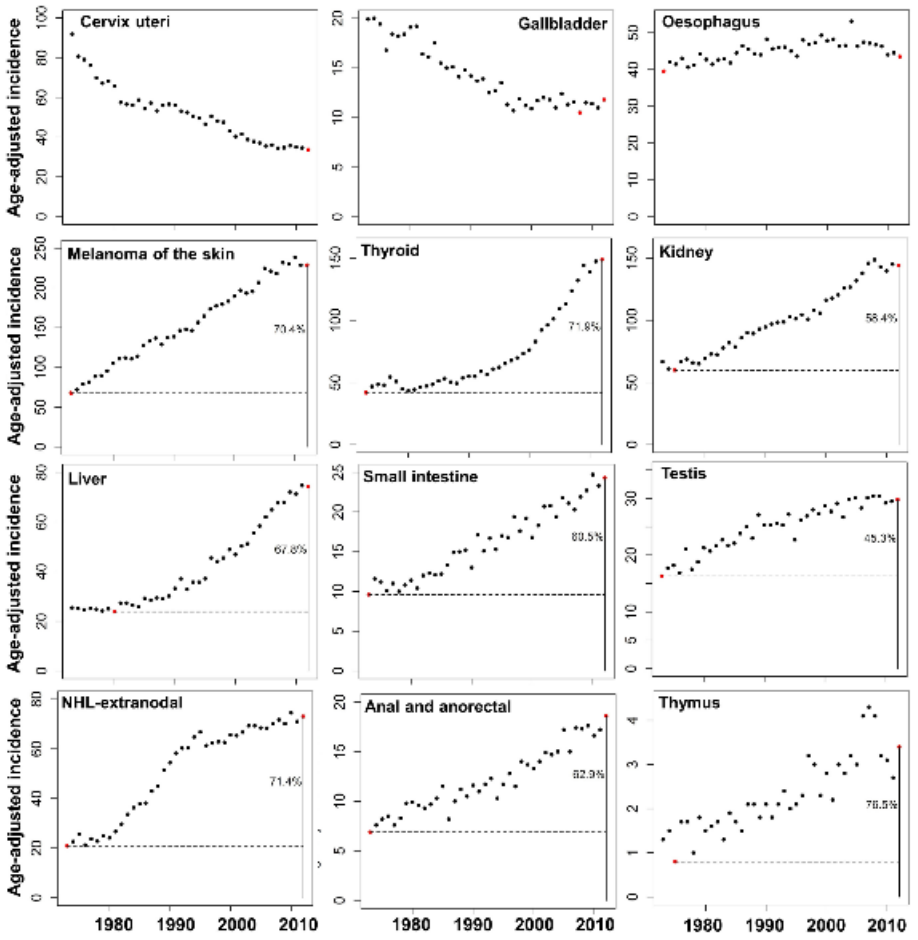
Estimating cancer risk for different tissues. The rounds of symmetric and asymmetric divisions for different tissues were adopted from supplementary table 1 of Tomasetti & Vogelstein[5]. In particular, the rounds of symmetric divisions, m, is equal to the integer part of $\log_2 S$, where S is the number of normal stem cells in the tissue of origin (data from ref. 5), and the rounds of asymmetric divisions a was the column labelled 'd' in supplementary table 1 of ref. 5. Sensitivity analyses have been conducted for scenarios with a broad range of mutation rates, from 1×10^{-10} to 1×10^{-6}, and several required hits ($k=1, 2, 3, 4$).

Lower-bound estimates of extrinsic risks with the SEER data. As a program of the National Cancer Institute (NCI), SEER (Surveillance, Epidemiology, and End Results Program) is a source of information on cancer incidence and survival in the USA (http://seer.cancer.gov/). The age-adjusted cancer incidences were extracted from the database 'SEER 9 Regs Research Data, Nov 2014 Sub (1973–2012) <Katrina/Rita Population Adjustment>' using the SEER*Stat 8.2.1 (ref. 28). For several cancers, it has been observed that their incidence rates have increased markedly during the past 40 years (Extended Data Fig. 1). For these cancers, it is reasonable to assume that anything above the historical minimum incidence should be attributed to some environmental/extrinsic factors. Therefore, we can establish the following inequality:

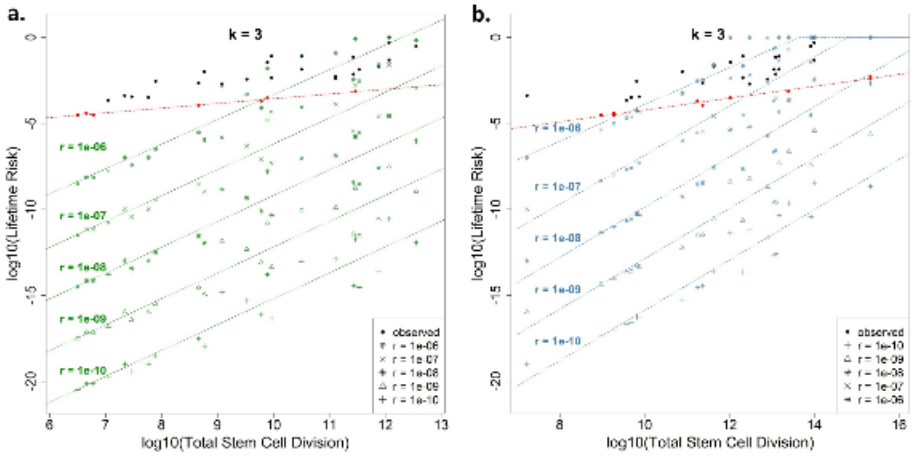
Extrinsic risk > (1 − historical minimum incidence rate/incidence rate in 2012).

Correspondingly, the lower bounds of contributions by extrinsic factors for these cancers can be calculated. As shown in Extended Data Fig. 1, some cancers show substantial contributions from extrinsic factors.

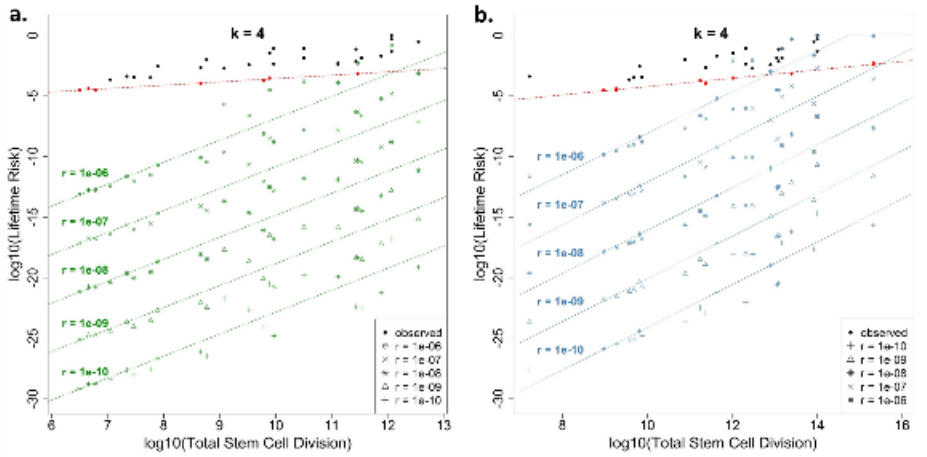
Data and statistical analysis. The observed lifetime cancer risks and the cumulative number of divisions (n) of all stem cells per lifetime are adopted from supplementary table 1 of Tomasetti & Vogelstein[5]. The total tissue cell divisions are from our evaluation of the data (Supplementary Information). For the robustness analysis of Fig. 3 as shown in Extended Data Table 1, error terms following the normal distribution with mean 0 and standard deviations of 1 or 0.4 were added to the $\log_{10}$(total stem-cell division) or $\log_{10}$(total cell division). These allow the number of total stem-cell and cell divisions to vary approximately within a range of ~1/100–100-fold or ~1/5–5-fold, respectively. On the basis of the new data set with measurement errors, the excess risks for each cancer were quantified. This process is repeated 1,000 times, and from this the mean, the 2.5 and the 97.5 percentiles (namely the 95% confidence intervals) of the excess risk for each cancer are tabulated. In calculating the percentage of intrinsic versus extrinsic mutations based on mutational signatures from cancer genome, we define the intrinsic mutation as those with signatures 1A/1B, and extrinsic mutation as all other mutational signatures (2–21, R1–R3, U1 and U2). The corresponding data were obtained from supplementary figures 59–88 of ref. 31. All statistical analyses and mathematical calculations were performed using R (version 3.1.2).



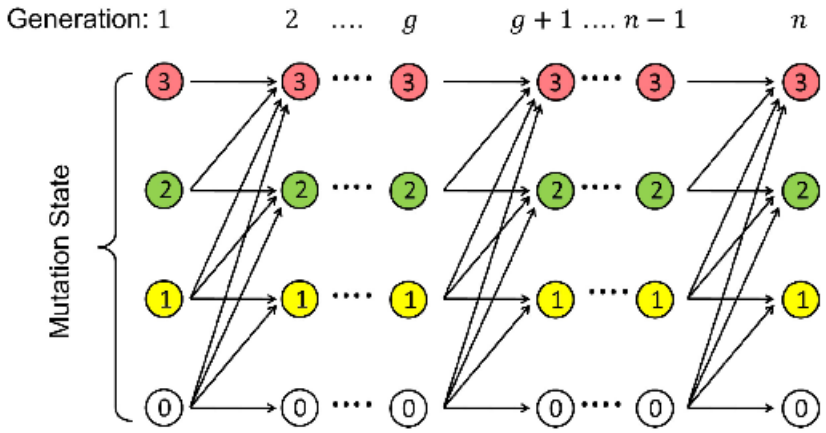
Extended Data Figure 1 | Examples of increased cancer incidence trends from 1973–2012 in SEER data. The cancer types include melanoma, thyroid cancer, kidney cancer, liver cancer, small intestine cancer, testicular cancer, non-Hodgkin lymphoma (NHL), anal and anorectal cancer and thymus cancer. The horizontal dashed lines indicate the historical minimal incidence. The vertical solid lines indicate the most recent year. The numbers represent the minimal percentage of extrinsic risk. The cervix uteri cancer, gallbladder cancer and oesophageal cancer are examples with declining or consistent incidence trend. The incidence rate is per 100,000 people.



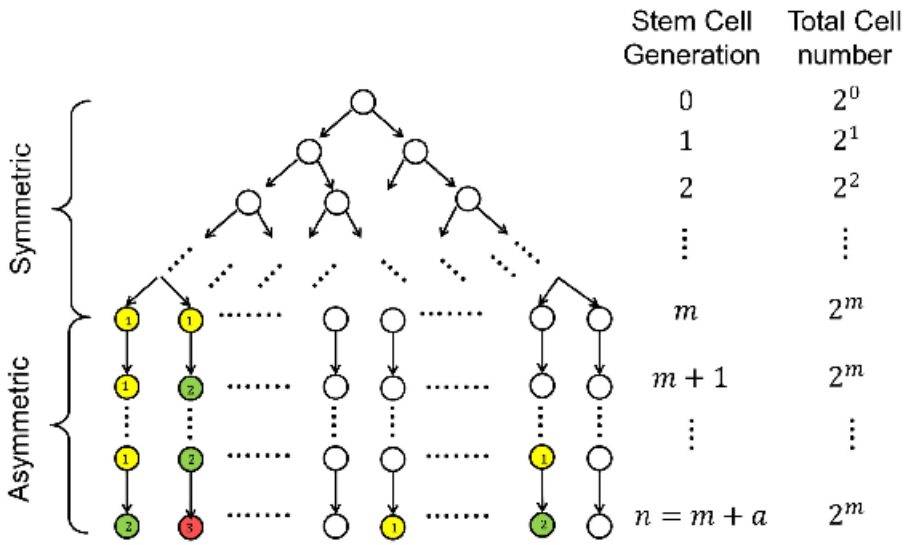
Extended Data Figure 2 | Sensitivity analysis of different mutation rates on tLIR when the number of hits (k) required is 3. a, b, Theoretical intrinsic lifetime risks (tLIR) for cancers have been calculated based on five different mutation rates: $r = 1 \times 10^{-10}$, 1×10^{-9}, 1×10^{-8}, 1×10^{-7}, 1×10^{-6}. The red dashed lines are the 'intrinsic' risk lines based on the observed data following the same estimation mechanism as the intrinsic risk line in Fig. 3a. The green (**a**) and blue (**b**) dashed lines are the 'intrinsic' risk lines estimated based on total reported stem-cell numbers and total homeostatic tissue cells, respectively.



Extended Data Figure 3 | Sensitivity analysis of different mutation rates on tLIR when the number of hits (k) required is 4. a, b, Theoretical intrinsic lifetime risks (tLIR) for cancers have been calculated based on five different mutation rates: $r = 1 \times 10^{-10}$, 1×10^{-9}, 1×10^{-8}, 1×10^{-7}, 1×10^{-6}. The red dashed lines are the 'intrinsic' risk lines based on the observed data following the same estimation mechanism as the intrinsic risk line in Fig. 3a. The green (**a**) and blue (**b**) dashed lines are the 'intrinsic' risk lines estimated based on total reported stem-cell numbers and total homeostatic tissue cells, respectively.



Extended Data Figure 4 | Intrinsic cancer risk modelling. Part 1 of 2: propagation diagram of driver gene mutation states between generations in one stem cell, from which the stem-cell mutation transition probabilities from one generation to the next are computed.



Extended Data Figure 5 | Intrinsic cancer risk modelling. Part 2 of 2: schema of stem-cell divisions and driver gene mutations, from which the theoretical lifetime intrinsic risks (tLIR) for cancer due to k driver gene mutations are computed. Each coloured circle represents the mutation of a new driver gene in the given stem cell (yellow, first mutation; green, second mutation; red, third mutation). If the mutation of 3 designated driver genes would induce a cancerous stem cell ($k = 3$), then this diagram shows a cancer occurrence as the second stem cell in the last generation (generation n) that has accumulated all 3 driver gene mutations.

Extended Data Table 1 | Robustness analysis on total stem-cell divisions and cell divisions estimates in Fig. 3

		Total stem-cell divisions (Fig. 3A)			Total cell divisions (Fig. 3B)		
Name	Observed Risk	Log10 (divisions)	Excess risk	Excess risk 95% CI*	Log10 (divisions)	Excess risk	Excess risk 95% CI*
AML	0.0041	11.11	>0.871	[0.623 , 0.962]	NA	NA	NA
Basal cell	0.3	12.55	>0.996	[0.985 , 0.999]	14.42	>0.995	[0.99 , 0.998]
Breast	0.123	NA	NA	NA	14.54	>0.987	[0.974 , 0.994]
CLL	0.0052	11.11	>0.899	[0.701 , 0.973]	NA	NA	NA
COAD	0.048	12.07	>0.980	[0.934 , 0.995]	14.40	>0.971	[0.943 , 0.986]
FAP COAD	1	12.07	>0.999	[0.997 , 1.000]	14.40	>0.999	[0.997 , 0.999]
Lynch COAD	0.5	12.07	>0.998	[0.994 , 1.000]	14.40	>0.997	[0.994 , 0.999]
Duodenum†	3.00E-04	9.89	-	-	NA	NA	NA
FAP Duodenum	0.035	9.89	>0.993	[0.980 , 0.998]	NA	NA	NA
Esophageal	0.00194	9.08	>0.906	[0.748 , 0.975]	NA	NA	NA
Gallbladder	0.0028	7.89	>0.967	[0.922 , 0.991]	NA	NA	NA
Glioblastoma	0.00219	8.43	>0.943	[0.868 , 0.984]	NA	NA	NA
Head & neck	0.0138	10.50	>0.973	[0.921 , 0.992]	NA	NA	NA
HPV Head & neck	0.07935	10.50	>0.995	[0.985 , 0.999]	NA	NA	NA
Hepatocellular	0.0071	11.43	>0.906	[0.720 , 0.975]	13.41	>0.932	[0.872 , 0.969]
HCV Hepatocellular	0.071	11.43	>0.991	[0.969 , 0.998]	13.41	>0.993	[0.986 , 0.997]
Lung (nonsmoker)†	0.0045	9.97	>0.938	[0.835 , 0.982]	15.2	-	-
Lung (smoker)	0.081	9.97	>0.997	[0.990 , 0.999]	15.20	>0.958	[0.905 , 0.982]
Medulloblastoma†	0.00011	8.43	-	-	NA	NA	NA
Melanoma	0.0203	11.88	>0.960	[0.872 , 0.990]	NA	NA	NA
Osteosarcoma	0.00035	7.47	>0.790	[0.459 , 0.947]	11.79	>0.762	[0.568 , 0.887]
Arms osteosarcoma†‡	4.00E-05	6.66	-	-	10.99	-	-
Head osteosarcoma†‡	3.02E-05	6.78	-	-	11.1	-	-
Legs osteosarcoma	0.00022	7.05	>0.727	[0.306 , 0.930]	11.37	>0.761	[0.537 , 0.889]
Pelvis osteosarcoma†‡	3.00E-05	6.50	NA	NA	10.81	-	-
Ovarian germ cell	0.000411	7.34	>0.832	[0.573 , 0.958]	NA	NA	NA
Pancreatic ductal	0.013589	11.54	>0.948	[0.805 , 0.987]	NA	NA	NA
Pancreatic islet†	0.000194	9.78	-	-	NA	NA	NA
Prostate	0.14	NA	NA	NA	11.81	>0.999	[0.999 , 1]
Small intestine†‡	7.00E-04	11.47	-	-	14.22	-	-
Testicular	0.0037	9.53	>0.942	[0.843 , 0.984]	13.02	>0.914	[0.835 , 0.959]
Thyroid follicular	0.01026	8.77	>0.986	[0.964 , 0.996]	NA	NA	NA
Thyroid medullary	0.000324	7.77	>0.731	[0.308 , 0.928]	NA	NA	NA

Measurement errors were added to $\log_{10}$(divisions) and 1,000 simulations were carried out to calculate the mean and 95% confidence interval (CI) of the excess risks. See Methods for details. NA: d not available.
*Confidence interval.
†Cancers used to compute the 'intrinsic' risk line based on total stem-cell divisions.
‡Cancers used to compute the 'intrinsic' risk line based on total cell divisions.

Extended Data Table 2 | Epidemiological studies on the extrinsic risks of various cancers

Cancer Types	Extrinsic risk	Examples of potential extrinsic risk factors*
Breast	substantial	Oral contraceptive, hormone replacement therapy, lifestyle (diet, smoking, alcohol, weight)
Prostate	substantial	Diet, obesity, smoking
Lung	>90%	Smoking; air pollutant
Colorectal	>75%	Diet, smoking, alcohol, obesity
Melanoma	65-86%	Sun exposure
Basal cell	~90%	UV
Hepatocellular	~80%	HBV, HCV
Gastric	65-80%	H. pylori
Cervical	~90%	HPV
Head & Neck	~75%	Tobacco, alcohol
Esophageal	>75%	Smoking, alcohol, obesity, diet
Oropharyngeal	~70%	HPV
Thyroid	>72%	Diet low in iodine, radiation
Kidney	>58%	Smoking, obesity, workplace exposures
Thymus	>77%	Largely unclear
Small intestine	>61%	Diet, smoking, alcohol
Extranodal non-Hodgkin's lymphoma (NHL)	>71%	Chemicals, radiation, immune system deficiency
Testis	>45%	Largely unclear
Anal and anorectal cancers	>63%	HPV, smoking

*http://www.cancer.org/cancer.

Extended Data Table 3 | Percentages of intrinsic versus extrinsic MS with known and unknown causes in different cancer types

	Intrinsic MS	Extrinsic MS - Known	Extrinsic MS - Unknown	Extrinsic MS - Total
ALL	65.8	34.2	0	34.2
AML	100	0	0	0
Bladder	14.2	71.2	14.6	85.8
Breast	35.5	60.1	4.4	64.5
Cervical	25.3	74.7	0	74.7
CLL	76.7	23.3	0	23.3
Colorectal	17.1	66	16.9	82.9
Esophageal	48	25.3	26.7	52
Glioblastoma	53.8	0	46.2	46.2
Glioma-Low Grade	9.2	2.8	88	90.8
Head & Neck	24.9	75.1	0	75.1
Kidney Chromophobe	17.4	37.5	45.1	82.6
Kidney Clear Cell	66.5	4.1	29.4	33.5
Kidney Papillary	0	15.7	84.3	100
Liver	10.9	21.3	67.8	89.1
Lung Adenocarcinoma	9.1	73.8	17.1	90.9
Lung - Small Cell	0	92.8	7.2	100
Lung-Squamous	0	47	53	100
Lymphoma B-cell	46.3	33.4	20.3	53.7
Medulloblastoma	48.4	0	51.6	51.6
Melanoma	7.2	90.9	1.9	92.8
Myeloma	0	19.9	80.1	100
Neuroblastoma	53.2	0	46.8	46.8
Ovarian	36.6	63.4	0	63.4
Pancreatic	49.9	50.1	0	50.1
Pilocytic Astrocytoma	82.5	0	17.5	17.5
Prostate	32.2	10.2	57.6	67.8
Stomach	22.3	6.1	71.6	77.7
Thyroid	0	39.7	60.3	100
Uterine	10.7	65.5	23.8	89.3

Intrinsic mutational signatures (MS) includes signatures 1A/B, and extrinsic MS includes signatures 2-21, R1-R3, U1 and U2, excluding signature 11 for Temozolomide, an alkylating agent used for chemotherapy. The blue, yellow and red colours highlight cancers that are have substantial extrinsic risk proportions based on epidemiological data, MS with known causes and MS with unknown causes, respectively. Data from the supplementary figs 59-88 in ref. 31.

Extended Data Table 4 | Percentages of extrinsic risks based on the reported stem-cell estimates and total homeostatic tissue cells, as shown in Fig. 4

Extrinsic Risks	Based on stem cell estimates				Based on total homeostatic tissue cells			
Cancer Type	k=1	k=2	k=3	k=4	k=1	k=2	k=3	k=4
AML	H.T.O.	H.T.O.	1.000	1.000	H.T.O.	H.T.O.	0.465	1.000
Basal cell	H.T.O.	0.462	1.000	1.000	H.T.O.	H.T.O.	1.000	1.000
CLL	H.T.O.	H.T.O.	1.000	1.000	H.T.O.	H.T.O.	0.578	1.000
COAD	H.T.O.	H.T.O.	0.999	1.000	H.T.O.	H.T.O.	0.928	1.000
FAP COAD	H.T.O.	0.630	1.000	1.000	H.T.O.	H.T.O.	0.997	1.000
Lynch COAD	H.T.O.	0.260	1.000	1.000	H.T.O.	H.T.O.	0.993	1.000
Duodenum	H.T.O.	H.T.O.	1.000	1.000	H.T.O.	H.T.O.	0.986	1.000
FAP Duodenum	H.T.O.	0.977	1.000	1.000	H.T.O.	H.T.O.	1.000	1.000
Esophageal	H.T.O.	0.946	1.000	1.000	H.T.O.	H.T.O.	0.997	1.000
Gallbladder	H.T.O.	1.000	1.000	1.000	H.T.O.	0.974	1.000	1.000
Glioblastoma	H.T.O.	0.995	1.000	1.000	H.T.O.	H.T.O.	1.000	1.000
Head & neck	H.T.O.	0.631	1.000	1.000	H.T.O.	H.T.O.	0.997	1.000
HPV Head & neck	H.T.O.	0.936	1.000	1.000	H.T.O.	H.T.O.	0.999	1.000
Hepatocellular	H.T.O.	0.572	1.000	1.000	H.T.O.	H.T.O.	1.000	1.000
HCV Hepatocellular	H.T.O.	0.957	1.000	1.000	H.T.O.	H.T.O.	1.000	1.000
Lung (nonsmoker)	H.T.O.	0.971	1.000	1.000	H.T.O.	H.T.O.	1.000	1.000
Lung (smoker)	H.T.O.	0.998	1.000	1.000	H.T.O.	0.388	1.000	1.000
Medulloblastoma	H.T.O.	0.904	1.000	1.000	H.T.O.	H.T.O.	1.000	1.000
Melanoma	H.T.O.	0.444	1.000	1.000	H.T.O.	0.444	1.000	1.000
Osteosarcoma	H.T.O.	1.000	1.000	1.000	H.T.O.	0.624	1.000	1.000
Arms osteosarcoma	H.T.O.	0.999	1.000	1.000	H.T.O.	0.269	1.000	1.000
Head osteosarcoma	H.T.O.	0.999	1.000	1.000	H.T.O.	0.032	1.000	1.000
Legs osteosarcoma	H.T.O.	1.000	1.000	1.000	H.T.O.	0.718	1.000	1.000
Pelvis osteosarcoma	H.T.O.	1.000	1.000	1.000	H.T.O.	0.542	1.000	1.000
Ovarian germ cell	H.T.O.	0.999	1.000	1.000	H.T.O.	0.999	1.000	1.000
Pancreatic ductal	H.T.O.	0.806	1.000	1.000	H.T.O.	H.T.O.	1.000	1.000
Pancreatic islet	H.T.O.	0.611	1.000	1.000	H.T.O.	H.T.O.	1.000	1.000
Small intestine	H.T.O.	H.T.O.	0.998	1.000	H.T.O.	H.T.O.	0.684	1.000
Testicular	H.T.O.	0.973	1.000	1.000	H.T.O.	H.T.O.	0.999	1.000
Thyroid follicular	H.T.O.	1.000	1.000	1.000	H.T.O.	0.866	1.000	1.000
Thyroid medullary	H.T.O.	0.999	1.000	1.000	H.T.O.	0.785	1.000	1.000

Extrinsic risk:= 1 – (tLIRsc or tLIRtt)/observed risk. H.T.O., higher than the observed.