eAppendix

Detailed methods

Tomasetti and Vogelstein (ref 1 of main text) collected data on all tissues for which good data on stem cell divisions were available. From these data, they obtained 31 data points. Here we describe how our data set derives from these 31 points.

The original data points include multiple cancers that derive from the same tissue. For instance, included are overall risk of colorectal adenocarcinoma, but also risk of colorectal adenocarcinoma for FAP patients and Lynch syndrome patients. As Potter and Prentice (2015) point out, including such special syndromes creates artificially high lifetime risk for the same number of stem cell divisions, potential distorting the outcome. Notice that no cancers for special syndromes are included for tissues with a low number of stem cell divisions. These considerations subtracted six data points.

AML and CLL derive from the same tissue. In the hypothesis of Tomasetti and Vogelstein, these should therefore be treated as one cancer. These two cancers were therefore merged to one, giving one data point fewer. Age distributions were weighed by lifetime risk.

Osteosarcoma is included both as a group and for separate organs. However, decomposing osteosarcoma as a group into osteosarcoma for separate organs necessarily gives the reported relationship: lifetime risk of osteosarcoma in separate organs will never exceed, and will in practice be significantly lower than, overall lifetime risk of osteosarcoma. Similarly, the number of stem cell divisions is partitioned across organs. Thus, if both the overall group and the subgroups are included in the analysis, the reported relationship will be supported by such decomposition, whether it exists or not. Finally, as others (Potter and Prentice 2015) have remarked, including multiple points for the same cancer but in different organs may put too much weight on dynamics specific for osteosarcoma. Thus, osteosarcoma is included as a group only, leading to four fewer data points.

For the remaining cancers, we followed Tomasetti and Vogelstein in consulting the SEER database, this time for age distributions. These databases depend strongly on reporting. Yet, it is unlikely that the effect we found would be much different in other databases, as it is fairly well known that some cancers, e.g. colorectal adenocarcinomas, hardly occur before age 45, whereas other cancers, e.g. medulloblastoma, are reputed to be childhood diseases.

For pancreas and thyroid cancer, age distributions were available only for the organ as a whole, leading to two fewer data points.

For basal cell carcinoma no age distributions were reported, leading to the loss of one data point, but these would certainly be in line with the rest of our results: basal cell carcinoma results from UV-light exposure accumulated over time.

An age distribution was reported on the website for ovarian cancer overall only. Since ovarian germ cell cancer makes up only ~3% of ovarian cancer, this age distribution could not be used, leading to the loss of one data point.

For the duodenum, no age distributions were found, leading to one fewer data point.

We favored age distributions published on the same website as the one used in the original analysis, i.e. http://seer.cancer.gov/statfacts. If, however, a reliable publication was available based on the SEER data, we did allow these data to be included. When necessary, we used an ungrouping algorithm that relies on modest assumptions (Rizzi et al. 2015) to be able to use the same age groups as the SEER website. This was the case for osteosarcoma (Duong and Richardson 2013), medulloblastoma (Smoll and Drummong 2012) and glioblastoma (Kozak et al. 2009). For all other cancers age distributions were readily available on seer.cancer.gov.

This leaves 15 data points and all 16 other data points accounted for. Even though we were able to use only 15 data points, the correlation between lifetime cancer risk and stem cell divisions up to age 80 as reported by Tomasetti and Vogelstein [ref 1 of main text], was found (Figure 1). This demonstrates that the reduction in the number of points was not the reason why the correlation could not be found at younger ages.

Additional references

Duong LM, Richardson LC. Descriptive Epidemiology of Malignant Primary Osteosarcoma Using Population-based Registries, United States, 1999-2008. J Registry Manag. 2013; 40(2): 59–64.

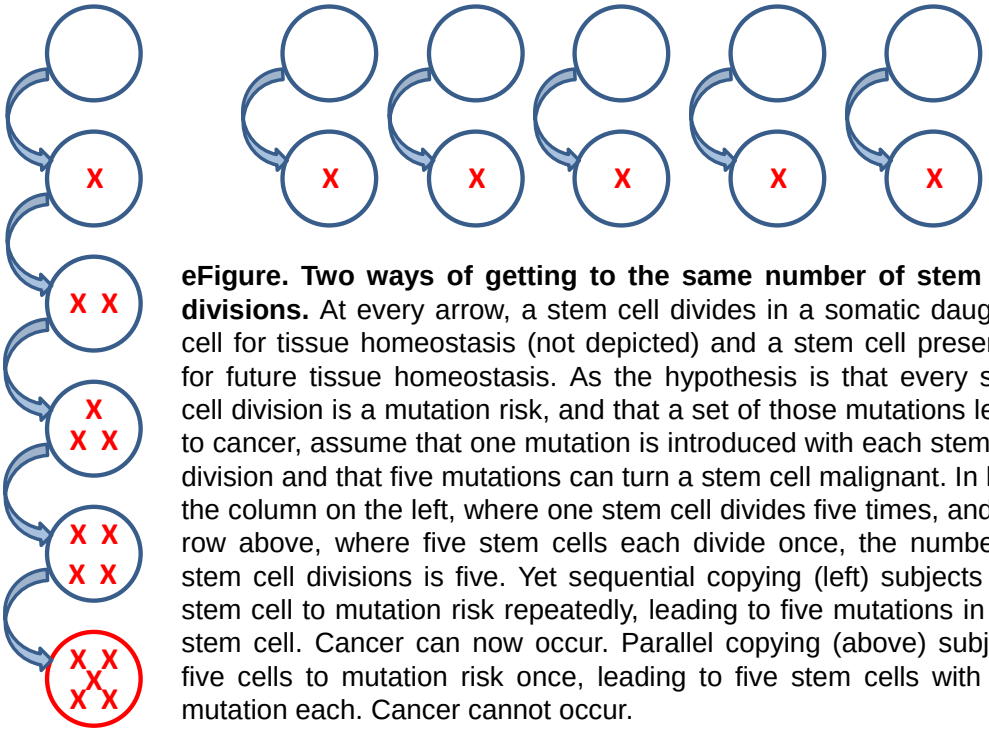
Kozak KR, Mahadevan A, Moody JS. Adult gliosarcoma: epidemiology, natural history, and factors associated with outcome. Neuro Oncol. 2009 Apr;11(2):183-91. doi: 10.1215/15228517-2008-076. Epub 2008 Sep 9.

Potter JD, Prentice RL. Cancer risk: tumors excluded. Science. 2015 Feb 13;347(6223):727. doi: 10.1126/science.aaa6507. Epub 2015 Feb 5.

Rizzi S, Gampe J, Eilers PHC. Efficient Estimation of Smooth Distributions From Coarsely Grouped Data. Am J Epidemiol. 2015 Jul 15;182(2):138-47. doi: 10.1093/aje/kwv020. Epub 2015 Jun 16.

Smoll NR, Drummond KJ. The incidence of medulloblastomas and primitive neurectodermal tumours in adults and children. Journal of Clinical Neuroscience 19 (2012) 1541–1544.

eFigure



eFigure. Two ways of getting to the same number of stem cell divisions. At every arrow, a stem cell divides in a somatic daughter cell for tissue homeostasis (not depicted) and a stem cell preserved for future tissue homeostasis. As the hypothesis is that every stem cell division is a mutation risk, and that a set of those mutations leads to cancer, assume that one mutation is introduced with each stem cell division and that five mutations can turn a stem cell malignant. In both the column on the left, where one stem cell divides five times, and the row above, where five stem cells each divide once, the number of stem cell divisions is five. Yet sequential copying (left) subjects one stem cell to mutation risk repeatedly, leading to five mutations in one stem cell. Cancer can now occur. Parallel copying (above) subjects five cells to mutation risk once, leading to five stem cells with one mutation each. Cancer cannot occur.