Title: Epidemiological correlation between cancer risk and tumor genomic mutation rate confirms the predominant contribution of somatic mutation

Dapeng Hao[1], Li Wang [1,2], Li-jun Di[1*]

Affiliations:

1. Cancer Center, Faculty of Health Sciences, University of Macau, Macau, SAR of China

2. Metabolomics Core, Faculty of Health Sciences, University of Macau, Macau, SAR of China

*Correspondence to: lijundi@umac.mo; Tel: 853-8822-4497;

Fax: 853-8822-2314,

Keywords: Cancer; Genomics; Somatic mutation; Cancer risk; Epidemiology

Abstract:

Cancer is believed to be a result of accumulated mutations. However, this concept has not been fully confirmed owing to the impossibility of tracking down the ancestral somatic cell with mutation accumulation before it gives rise to a detectable tumor. We sought to verify the concept by exploring the correlation between cancer risk and "consensus" mutation rate in bulk tumor of different tissues. We collected a comprehensive list of "consensus" mutation rates revealed by bulk tumor sequencing of 53 studies, and investigated its correlation with cancer risk to mirror the correlation between mutation rate in somatic cells and cancer risk. This revealed a 1:1 relationship between mutation rate and cancer risk in 41 different cancer types based on the sequencing data of 5,542 patients. The correlation was extremely robust even against the variation of estimations of mutation rate. Moreover, the correlation establishes a baseline to evaluate the effect of non-mutagenic carcinogens on cancer risk. Since the mutations obtained from bulk tumor sequencing are largely inherited from somatic mutations of an ancestral cell that gives rise to tumor, our mathematic modeling provides the evidence to reinforce that cancer risk is predominantly determined by the first rate-limiting mutation.

Introduction

The variation in the number of mutations across different cancer types is widely noticed[1]. Identification of these mutations is traditionally according to the genomic sequencing data of bulk tumors[2]. Notable among the most frequently mutated cancers are basal cell carcinoma (BCC) and melanoma, which contain ~2,200 and ~800 mutations in the coding region[3, 4]. On the other side, some pediatric cancers such as rhabdoid cancer contain less than 10 mutations per tumor[5]. The variation is also seen across different cancers with similar involvement of environmental mutagens. For example, Glioblastoma multiforme (GBM) has ~5 times as many mutations as medulloblastoma[6, 7]. Interestingly, BCC, melanoma and GBM are among the common human cancers, whereas rhabdoid cancer and medulloblastoma are relatively rare, suggesting the hypothetical existence of a correlation between the mutation rate in tumors and the cancer risk.

Meanwhile, the accumulation of mutations in somatic cells is hypothesized to be the fundamental reason for tumorigenesis[1]. However, the correlation between mutation accumulation in somatic cells and cancer risk has never been worked out because of the technical limitation in obtaining the somatic mutation rate of any tissue[8]. A recent finding that cancer risk is correlated with the number of stem cell divisions highlights the hypothesis by suggesting that cancer risk is a result of accumulated genomic changes occurring by chance during DNA replication[9]. However, this study didn't take many common human cancers into account such as prostate cancer and breast cancer probably because of the lacking of data regarding the number of stem cell divisions in

these tissues. Furthermore, solely attributing stem cell division to apparently higher rate of lung cancer in smokers versus non-smokers and of colorectal cancer in inherited mismatch repair deficiency patients versus normal colorectal cancer patients, is in against with the general realization that smoking and inherited mismatch repair deficiency increase the mutation rate without strong influence on the cell division[10-12]. Therefore, there should be factors beyond stem cell divisions that contribute to mutation accumulation.

The availability of the whole genome sequencing data of bulk tumor tissues[13], however, presents an opportunity to evaluate the mutation rate of the ancestral somatic cell of each tumor. The "consensus mutations" detected in bulk tumor sequencing outcompete the random mutations present in individual cells because the random mutations are masked by sequencing millions of cells simultaneously, as reflected by the undetectable mutation in normal control samples in cancer genomics studies[6, 14, 15] but revealed in single cells[16]. Therefore, the "consensus" mutations revealed by bulk tumor sequencing are largely a reflection of the mutations accumulated in the ancestral cell that gives rise to tumor (**see Supplementary Text for more discussions**). This supposition is strongly supported by the finding that half or more of the mutations detected in tumor bulk occur prior to tumor initiation, that is, during the growth of normal cells[1, 17]. Thus, investigating the correlation between tumor mutation rate and cancer risk can be an alternative way to mirror the correlation between mutation rate in somatic cells and cancer risk.

Results and Discussion

Correlation between cancer risk among tissues and mutation rate in tumor bulk

First of all, the mutation rates of 41 different cancer types from the data of 5,542 human tumors detected by whole genome/exome sequencing were collected (**see Methods and Supplementary Text**). Then the lifetime risk of cancers as a function of the mutation rate of the corresponding cancers was plotted (**Figure 1**). A strong correlation was observed between the two different parameters, with a Pearson correlation coefficient 0.72 ($p < 1.4\times10^{-7}$) in linear (x, y) coordinates and 0.77 ($p < 4.5\times10^{-9}$) in logarithmic (log(x), log(y)) coordinates. Spearman rank correlation is also significant (Spearman's rho = 0.75; $p < 2.1\times10^{-8}$). This correlation strongly supports the hypothesis that there is statistic significant association between tumor mutation rate and cancer risk and at least 50~60% of the variation of cancer risk is due to the difference of mutation rate. To overcome the potential bias by using mixed data sources to estimate the mutation rate, we selected 2,736 tumors across 29 cancer types that were sequenced using a uniform experimental pipeline and then analyzed using the same analytical pipeline from quality control, data processing and mutation calling[13]. This new dataset reveals a similar correlation between lifetime risk and mutation rate (**Supplementary Figure 1**; $r = 0.66$, $p < 3\times10^{-5}$). A strong correlation was observed between the mutation rates in this new dataset and the mutation rates by our estimation ($r = 0.95$ in log-log scale, $p < 2.5\times10^{-14}$) suggesting our estimates of mutation rate are highly robust against the data collection process from various sources.

This correlation applies to cancers across different tissues, associated with different environmental exposures and hereditary factors. For instance, when the mutation rate of

the same cancer type is increased by mutagens (i.e., lung cancer patients as smokers vs. non-smokers) or hereditary defects (i.e., HNPCC vs. MSI-Colorectal cancer，MSS-stomach cancer vs. MSI-stomach cancer), the lifetime incidence rises in a corresponding rate. Another example is the most common type of human cancers -- skin cancers including melanoma, squamous and basal cell carcinoma[18], for which our result suggests that the differences in cancer incidence match the variation of mutation rate. Importantly, the correlation is extremely robust even when the estimates of mutation rates were allowed to vary significantly (**see Methods**).

Of noting is that the slope of the regression line in log-log scale is 1.09 (0.80-1.39,95% CI) in Figure 1 and 1.01 (0.55-1.46, 95% CI) in the new dataset, indicating a nearly perfect 1:1 relationship between mutation rate and cancer risk. Under an ideal condition, if the "consensus" mutations in bulk sequencing represent the mutations of the ancestral cell, both the mutations accumulated before its fundamental change toward preneoplastic growth and the first rate-limiting mutation to initiate the preneoplastic growth, are included. Therefore, the mutation rate in our measurements should correlate with the rate of mutation accumulation before the preneoplastic growth, as well as the rate of the first rate-limiting mutation. We speculate that such a correlation between measured mutation rate and the rate of the first rate-limiting mutation may well explain the 1:1 relationship between mutation rate and cancer risk. In the other words, the first rate-limiting mutation may decide the cancer risk predominantly. Our mathematical modeling further supports the 1:1 relationship (**see Methods and Supplementary Text**), and shows that the theoretical connection between the "consensus" mutation and

the first-rate limiting step is surprisingly consistent with some important behaviors of cancer.

Evaluating the effect of mutation rate on cancer risk

To distinguish the effect of mutation rate on cancer incidence from other potential non-mutagenic carcinogens, we computed the ratio of cancer risk to mutation rate for different cancers (**Figure 2**). Higher the ratio is, more important role the non-mutagenic carcinogens may play in that cancer's incidence. Most cancers have relatively low level of ratio, with a median value of 0.002, suggesting that each ~30 mutations in the coding region (1 mutation per Mb) are associated with a 0.2% increase of lifetime incidence. Interestingly, there are four cancers having obviously higher ratios (**Figure 2**), including two hormone-related cancers (prostate cancer and breast cancer) and two virus-related cancers (liver cancer with Hepatitis C infection and head&neck cancer with HPV-16 infection). The ratios of prostate cancer and breast cancer are 76 and 26 times higher than the median ratio of cancers respectively, which indicates that non-mutagenic factors (i.e., hormones) have more power in increasing the lifetime incidence of these two cancers than what would be expected by the number of accumulated mutations in the genome. Other two cancers, including endometrial cancer and ovary cancer whose risk is associated with excessive estrogen exposure, also show relatively higher ratio (**Figure 2**). The ratios of liver cancer and head & neck cancer infected with virus are 10 times and 9 times higher than the median value respectively, which is consistent with the increased incidence ratio of these two cancers infected with virus versus those not infected[19, 20]. These data suggest that the viruses present in these cancers increase

the cancer incidence in a non-mutagenic way. Importantly, cancers with high-level involvement of environmental exposure appear to show relatively higher ratios (i.e., lung cancers of smokers versus nonsmokers), suggesting non-mutagenic effects of environmental exposure contribute to the increased cancer risk.

Conclusion

Our analyses support that tumor mutation rate is a reliable predictor for cancer incidence in most of the human cancers, probably because that the tumor mutation rate mirrors the normal somatic mutation rate in all the analyzed tissues. Indeed the mutations present in a tumor bulk provide a lifetime record of the mutation accumulation contributed by stem cell division over the full course of self-renewal and tissue specific differentiation, as well as anything else such as being exposure to environmental or inherited factors. And it has been shown that tumor likely arises from cells with a normal mutation rate[21], based on the fact they outnumber the cells with aberrant mutation rate so much.

In addition, our finding of hormones and viruses related cancers showing significant increase in the ratio of life time risk to mutation rate indicates non-mutagenic effect associated with hormones and viruses can be important player in increasing the cancer incidence. Nevertheless, the majority of cancers are strongly influenced by mutation rate.

Methods

Tumor Samples and Cancer Risk

We included in our analyses a total of 41 different cancers from 5,542 samples obtained

from 53 previous studies (**Supplementary Table1-2**). All the mutation rates are based on results of whole genome sequencing (WGS) or whole exome sequencing (WES). The average mutation rates of most cancers were collected from literatures directly, or evaluated using the data form the literatures. Cancers not included in this study were largely due to the lack of data or too few samples of that cancer that were detected by WGS/WES.

When available, cancer lifetime incidences were obtained from Surveillance, Epidemiology and End Results (SEER) database (www.seer.cancer.gov)[22] and generated by their software DevCan[23], or obtained directly from a previous study[9]. If the data were not available this way, we using the epidemiological statistics to estimate the lifetime incidence for a specific cancer. Details of data collection and processing for each cancer subtype are provided in the **Supplementary Text** in separate sections.

Robustness Analysis

Mutation rates can vary markedly across patients within a cancer type [1, 13]. To estimate the robustness of the correlation between mutation rate and cancer risk, we first estimated the standard deviations (std.) of the mutation rate by bootstrap for cancers with mutation rates available for individual patients (n = 2,175 in total). The std. of mutation rate was similar among cancers benchmarked against the average (avg.) of mutation, and overall was about ~5% of the avg. of mutation rate in the magnitude. Then, to allow the estimates of mutation rates to vary significantly, we sampled from a normal distribution with std. equal to 20% of the avg. of mutation rate in magnitude for each of the 41 cancers. This setting allowed the estimates for the mutation rate to vary

significantly. For example, in melanoma with a mutation rate of ~28 per Mb, this allows a (28·0.2)·(28·0.2) ≈ 31 per Mb mutation rate variation in either direction. After 10,000 iterations, the average Pearson correlation was 0.7 (range: 0.45-0.87) in linear (x, y) coordinates and 0.76 (range: 0.70-0.82) in logarithmic (log(x), log(y)) coordinates. The average Spearman correlation was 0.73 (range: 0.64-0.80). The correlation was significant in all iterations, with highest p value of Pearson correlation <0.003 in linear coordinates and $<4\cdot10^{-7}$ in logarithmic coordinates, and highest p value of Spearman correlation < $6\cdot10^{-6}$. All the statistical analyses were performed in MATLAB, version 2014a.

Mathematical Modeling

Based on the classic Armitage-Doll model [24], assume that for a progenitor cell evolving to a clinically meaningful tumor, the first rate-limiting step (driver mutation) and n ensuing independent steps are required, the cancer incidence can be given by

$$\log\big(I(t)\big) = \log(\mu t) + \log Lp_1p_2\cdots p_n\frac{(t-1)^n}{n!},$$

where μ represents the mutation rate per unit interval of time before the first rate-limiting step and $(p_1\cdots p_n)$ represent the probability of ensuing steps $(p_1\cdots p_n)$ per unit time interval during clonal evolution. Here, μt, given t is representative of lifetime, would be the accumulated mutation rate in the ancestral somatic cell of tumor that contributes the majority of the mutation rate revealed by sequencing tumor bulk. This modeling well explains the 1:1 relationship between mutation rate in tumor bulk and cancer incidence in log-log coordinates. More details of the mathematical modeling process are provided in the **Supplementary Text.**

Acknowledgements:

D.H. conceived the research, analyzed data, led the research and drafted the manuscript. L.W. initiated the funded part of the research, analyzed the data and drafted the manuscript. L.D. conceived the research, interpreted the data and drafted the manuscript. L.D. and L.W. acknowledge support from Science and Technology Development Fund, Macao S.A.R (FDCT) (FDCT/025/A1 and FDCT/088/2014/A2) and support from University of Macau (MYRG2015-00037-FHS, MYRG2015-00167-FHS; MRG022/DLJ/2015/FHS; MRG023/DLJ/2015/FHS).

The authors declare no conflict of interest.

Figure Legend

Figure 1. The correlation between the lifetime risk of cancer and the mutation rate in tissue bulk of that cancer.
Values and cancer names corresponding to the abbreviations in the figure are shown in Supp. Table S1.

Figure 2. Ratio of lifetime cancer risk to mutation rate across cancers.
Cancers are ranked by alphabetical order in the x-axis. Cancers with ratio higher than two times of inter-quartile deviation of the data above the top quartile are denoted as red nodes.

Supplementary Figure 1. The correlation between the lifetime risk of cancer and the mutation rate in tissue bulk of that cancer, using the data generated by a uniform pipeline.

Reference

1. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
2. Tripathy, D., et al., *Next generation sequencing and tumor mutation profiling: are we ready for routine use in the oncology clinic?* BMC Med, 2014. **12**(1): p. 140.
3. Jayaraman, S.S., et al., *Mutational landscape of basal cell carcinomas by whole-exome sequencing.* J Invest Dermatol, 2014. **134**(1): p. 213-20.
4. Hodis, E., et al., *A landscape of driver mutations in melanoma.* Cell, 2012. **150**(2): p. 251-63.
5. Lee, R.S., et al., *A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers.* J Clin Invest, 2012. **122**(8): p. 2983-8.
6. Brennan, C.W., et al., *The somatic genomic landscape of glioblastoma.* Cell, 2013. **155**(2): p. 462-77.
7. Parsons, D.W., et al., *The genetic landscape of the childhood cancer medulloblastoma.* Science, 2011. **331**(6016): p. 435-9.
8. Greenman, C., et al., *Patterns of somatic mutation in human cancer genomes.* Nature, 2007. **446**(7132): p. 153-8.
9. Tomasetti, C. and B. Vogelstein, *Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions.* Science, 2015. **347**(6217): p. 78-81.
10. Imielinski, M., et al., *Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing.* Cell, 2012. **150**(6): p. 1107-20.
11. Govindan, R., et al., *Genomic landscape of non-small cell lung cancer in smokers and never-smokers.* Cell, 2012. **150**(6): p. 1121-34.
12. Seshagiri, S., et al., *Recurrent R-spondin fusions in colon cancer.* Nature, 2012. **488**(7413): p. 660-4.
13. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.
14. Cancer Genome Atlas Research, N., *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-15.
15. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
16. Xu, X., et al., *Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor.* Cell, 2012. **148**(5): p. 886-95.
17. Tomasetti, C., B. Vogelstein, and G. Parmigiani, *Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation.* Proc Natl Acad Sci U S A, 2013. **110**(6): p. 1999-2004.
18. Leiter, U., T. Eigentler, and C. Garbe, *Epidemiology of skin cancer.* Adv Exp Med Biol, 2014. **810**: p. 120-40.
19. Davila, J.A., et al., *Hepatitis C infection and the increasing incidence of hepatocellular carcinoma: a population-based study.* Gastroenterology, 2004. **127**(5): p. 1372-80.
20. Dahlstrom, K.R., et al., *Human papillomavirus type 16 infection and squamous cell*

carcinoma of the head and neck in never-smokers: a matched pair analysis. Clin Cancer Res, 2003. **9**(7): p. 2620-6.

21. Tomlinson, I.P., M.R. Novelli, and W.F. Bodmer, *The mutation rate and cancer.* Proc Natl Acad Sci U S A, 1996. **93**(25): p. 14800-3.
22. *National Cancer Institute. Surveillance, Epideniology and End Results Program;* http://www.seer.cancer.gov.
23. *DevCan: probability of developing or dying of cancer software, version 6.7.2. Statistical research and application branch, National Cancer Institute, 2007.*
24. Armitage, P. and R. Doll, *The age distribution of cancer and a multi-stage theory of carcinogenesis.* Br J Cancer, 1954. **8**(1): p. 1-12.

Pearson correlation coefficient $r = 0.77$

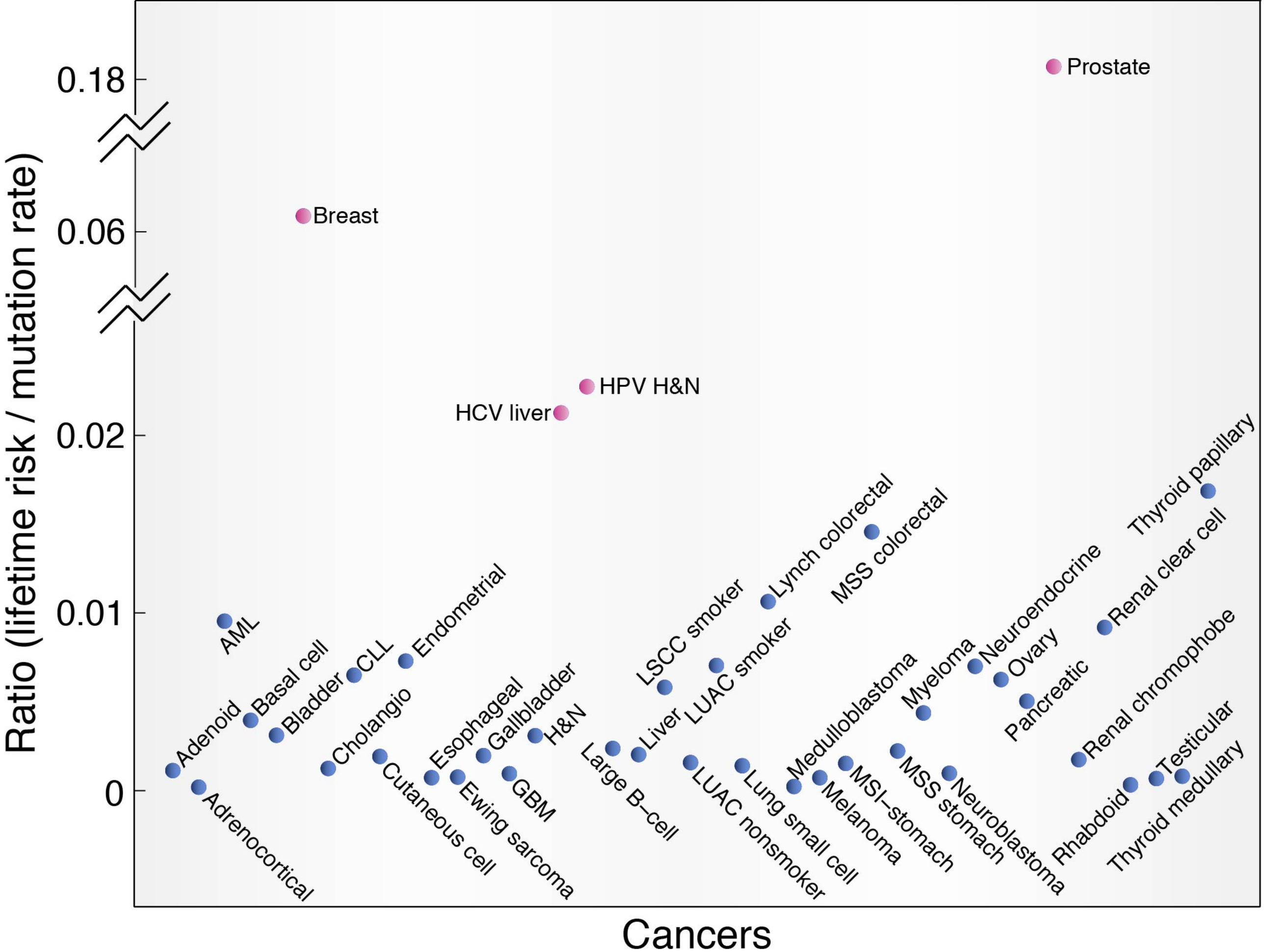
Lifetime risk

10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}

Lynch colorectal
Basal cell
Prostate
Breast
HPV H&N
Cutaneous cell
LUAC smoker
MSI–stomach
MSS colorectal
HCV liver
LSCC smoker
Pancreatic
Endometrial
Bladder
Melanoma
Ovary
Renal clear cell
Lung small cell
H&N
Thyroid papillary
MSS stomach
Liver
Large B-cell
Myeloma
CLL
AML
GBM
Testicular
Gallbladder
LUAC nonsmoker
Esophageal
Cholangio
Renal chromophobe
Neuroblastoma
Neuroendocrine
Adenoid
Thyroid medullary
Ewing sarcoma
Adrenocortical
Medulloblastoma
Rhabdoid

10^{-1}
10^{0}
10^{1}
10^{2}

Mutation rate (per Mb)


Ratio (lifetime risk / mutation rate)
0.18
0.06
0.02
0.01
0
Cancers
Prostate
Breast
HPV H&N
HCV liver
Thyroid papillary
Lynch colorectal
MSS colorectal
LSCC smoker
LUAC smoker
Renal clear cell
Neuroendocrine
Ovary
Pancreatic
Myeloma
Medulloblastoma
Renal chromophobe
Testicular
Rhabdoid
Thyroid medullary
Neuroblastoma
MSS stomach
MSI–stomach
Melanoma
Lung small cell
LUAC nonsmoker
Large B–cell
Liver
H&N
Gallbladder
Esophageal
GBM
Ewing sarcoma
Cutaneous cell
Cholangio
Endometrial
CLL
Bladder
Basal cell
Adenoid
AML
Adrenocortical