Cancer Biology
Evolutionary Biology

The theory of massively repeated evolution and full identifications of Cancer Driving Nucleotides (CDNs)

Reviewed Preprint
v2 • October 25, 2024
Revised by authors

Reviewed Preprint
v1 • September 3, 2024

Lingjie Zhang, Tong Deng, Zhongqi Liufu, Xueyu Liu, Bingjie Chen, Zheng Hu, Chenli Liu, Miles E Tracy, Xuemei Lu, Haijun Wen, Chung-I Wu

State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China • State Key Laboratory of Genetic Resources and Evolution/Yunnan Key Laboratory of Biodiversity Information, Kunming Institute of Zoology, The Chinese Academy of Sciences, Kunming, China • GMU-GIBH Joint School of Life Sciences, Guangzhou Medical University, Guangzhou, China • CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China • Department of Ecology and Evolution, University of Chicago, Chicago, USA

eLife Assessment

This **important** paper introduces a theoretical framework and methodology for identifying Cancer Driving Nucleotides (CDNs), primarily based on single nucleotide variant (SNV) frequencies. A variety of **solid** approaches indicate that a mutation recurring three or more times is more likely to reflect selection rather than being the consequence of a mutation hotspot. The method is rigorously quantitative, though the requirement for larger datasets to fully identify all CDNs remains a noted limitation. The work will be of broad interest to cancer geneticists and evolutionary biologists.

https://doi.org/10.7554/eLife.99340.2.sa3

Abstract

Tumorigenesis, like most complex genetic traits, is driven by the joint actions of many mutations. At the nucleotide level, such mutations are Cancer Driving Nucleotides (CDNs). The full sets of CDNs are necessary, and perhaps even sufficient, for the understanding and treatment of each cancer patient. Currently, only a small fraction of CDNs is known as most mutations accrued in tumors are not drivers. We now develop the theory of CDNs on the basis that cancer evolution is massively repeated in millions of individuals. Hence, any advantageous mutation should recur frequently and, conversely, any mutation that does not is either a passenger or deleterious mutation. In the TCGA cancer database (sample size n = 300 - 1000), point mutations may recur in i out of $\mathbf{n}$ patients. This study explores a wide range of mutation characteristics to determine the limit of recurrences (i^*) driven solely by neutral evolution. Since no neutral mutation can reach $i^* = 3$, all mutations recurring at $i \geq 3$ are CDNs. The theory shows the feasibility of identifying almost all CDNs if $\mathbf{n}$ increases to 100,000 for each cancer type. At present, only < 10% of CDNs have been identified. When the full sets

of CDNs are identified, the evolutionary mechanism of tumorigenesis in each case can be known and, importantly, gene targeted therapy will be far more effective in treatment and robust against drug resistance.

Introduction

Cancers are complex genetic traits with multiple mutations that interact to yield the ensemble of tumor phenotypes. The ensemble has been characterized as “cancer hallmarks” that include sustaining growth signaling, evading growth suppression, resisting apoptosis, achieving immortality, executing metastasis and so on (Hanahan and Weinberg 2000; Hanahan and Weinberg 2011; Hanahan 2022). It seems likely that each of the 6-10 cancer hallmarks is governed by a set of mutations. Most, if not all, of these mutations are jointly needed to drive the tumorigenesis.

In the genetic sense, cancers do not differ fundamentally from other complex traits whereby multiple mutations are simultaneously needed to execute the program. A well-known example is the genetics of speciation whereby interspecific hybrids are either sterile or infertile even though they do not have deleterious genes (Wu and Ting 2004; Wang et al. 2022; Wu 2022). A recent example is SARS-CoV-2.

The early onset of COVID-19 requires all four mutations of the D614G group and the later Delta strain has 31 mutations accrued in three batches (Ruan, Wen, et al. 2022; Ruan, Hou, et al. 2022; Cao et al. 2023; Ruan et al. 2023) While cancer research has often proceeded one mutation at a time, each of the mutations has been shown to be insufficient for tumorigenesis until many (Ortmann et al. 2015; Takeda 2021; Hodis et al. 2022) are co-introduced.

We now aim for the identification of all (or at least most) of the driver mutations in each patient. Both functional tests and treatments demand such identifications. The number of key drivers has been variously estimated to be 6 – 10 (Martincorena et al. 2017; Anandakrishnan et al. 2019; Campbell et al. 2020). Although cancer driving “point mutations”, referred to as Cancer Driving Nucleotides (or CDNs), are not the only drivers, they are indeed abundant (Campbell et al. 2020). Furthermore, CDNs, being easily quantifiable, may be the only type of drivers that can be fully identified (see below). Here, we will focus on the clonal mutations present in all cells of the tumor without considering within-tumor heterogeneity for now (Ling et al. 2015; Turajlic et al. 2019; Black and McGranahan 2021; B. Chen et al. 2022; Zhai et al. 2022; Bian et al. 2023; Zhu et al. 2023).

Since somatic evolution proceeds in parallel in millions of humans, point mutations can recur multiple times as shown in **Fig. 1**. The recurrences should permit the detection of advantageous mutations with unprecedented power. The converse should also be true that mutations that do not recur frequently are unlikely to be advantageous. **Fig. 1** depicts organismal evolution and cancer evolution. While both panels **A** and **B** show 7 mutations, there can be two patterns for cancer evolution - the pattern in **(C)** where all mutations are at different sites is similar to the results of organismal evolution whereas the pattern in **(D)** is unique in cancer evolution. The hotspot of recurrences shown in **(D)** holds the key to finding all CDNs in cancers.

In the literature, hotspots of recurrent mutations have been commonly reported (Gartner et al. 2013; Chang et al. 2016; Cannataro et al. 2018; Buisson et al. 2019; Hess et al. 2019; Stobbe et al. 2019; Juul et al. 2021; Nesta et al. 2021; Zhao et al. 2021; Bergstrom et al. 2022; Sherman et al. 2022; Wong et al. 2022; Zeng and Bromberg 2022). A hotspot, however, could be either a mutational or functional hotspot. Mutational hotspots are the properties of the mutation machinery that would include nucleotide composition, local chromatin structure, timing of replication, etc. (Stamatoyannopoulos et al. 2009; Pleasance et al. 2010;

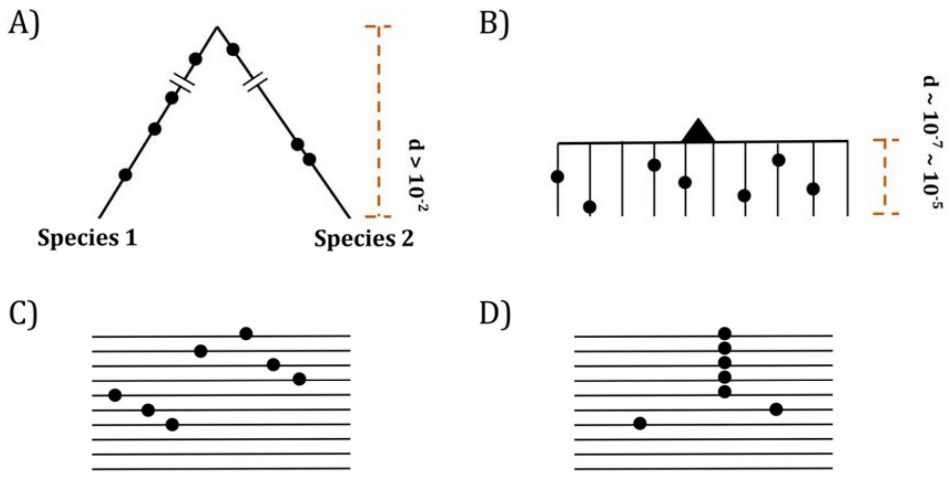


Figure 1

Two modes of DNA sequence evolution. **(A)** A hypothetical example of DNA sequences in organismal evolution. **(B)** Cancer evolution that experiences the same number of mutations as in **(A)** but with many short branches. **(C)** A common pattern of sequence variation in cancer evolution. **(D)** In cancer evolution, the same mutation at the same site may occasionally be seen in multiple sequences. The recurrent sites could be either mutational or functional hotspots, their distinction being the main objective of this study.

Makova and Hardison 2015; Polak et al. 2015; Martincorena et al. 2017). In contrast, functional hotspots are CDNs under positive selection. CDN evolution is akin to "convergent evolution" that repeats itself in different taxa (He, Xu, and Shi 2020; He, Xu, Zhang, et al. 2020; Wu et al. 2020; Pan, Zhang, et al. 2022; Wu 2023) and is generally considered the most convincing proof of positive selection.

While many studies conclude that sites of mutation recurrence are largely mutational hotspots (Buisson et al. 2019; Hess et al. 2019; Stobbe et al. 2019; Nesta et al. 2021; Bergstrom et al. 2022), others deem them functional hotspots, driven by positive selection (Gartner et al. 2013; Chang et al. 2016; Bailey et al. 2018; Cannataro et al. 2018; Juul et al. 2021; Zhao et al. 2021; Zeng and Bromberg 2022). In the attempt to distinguish between these two hypotheses, studies make assumptions, often implicitly, about the relative importance of the two mechanisms in their estimation procedures. The conclusions naturally manifest the assumptions made when extracting information on mutation and selection from the same data (Elliott and Larsson 2021).

This study consists of three parts. First, the mutational characteristics of sequences surrounding CDNs are analyzed. Second, a rigorous probability model is developed to compute the recurrence level at any sample size. Above a threshold of recurrence, all mutations are CDNs. Third, we determine the necessary sample sizes that will yield most, if not all, CDNs. In the companion study, the current cancer genomic data are analyzed for the characteristics of CDNs that have already been discovered (Zhang et al. 2024). Together, these two studies show how full functional tests and precise target therapy can be done on each cancer patient.

Results

In **PART I**, we search for the general mutation characteristics at and near high recurrence sites by both machine learning and extensive sequence comparisons. In **PART II**, we develop the mathematical theory for the maximal level of recurrences of neutral mutations (designated i^*). CDNs are thus defined as mutations with $\geq i^*$ recurrences in $\mathbf{n}$ cancer samples. We then expand in **PART III** the theory to very large sample sizes ($\mathbf{n} \geq 10^5$), thus making it possible to identify all CDNs.

To carry out the analyses, we first compile from the TCGA database the statistics of multiple hit sites (i hits in $\mathbf{n}$ samples) in 12 cancer types. This study focuses on the mutational characteristics pertaining to recurrence sites. We often present the three cancer types with the largest sample sizes (lung, breast and CNS cancers), while many analyses are based on pan-cancer data. Analyses of every of the 12 cancer types individually will be done in the companion paper. The TCGA database is used as it is well established and covers the entire coding region (Weinstein et al. 2013). Other larger databases (Cerami et al. 2012; Tate et al. 2019; de Bruijn et al. 2023) are employed when the whole exon analyses are not crucial.

The compilation of multi-hit sites across all genes in the genome

Throughout the study, $\boldsymbol{S_i}$ denotes the number of synonymous sites where the same nucleotide mutation occurs in i samples among $\mathbf{n}$ patients. $\boldsymbol{A_i}$ is the equivalent of $\boldsymbol{S_i}$ for non-synonymous (amino acid altering) sites. **Table 1** presents the numbers from lung cancer for demonstration. It also shows the $\boldsymbol{A_i}$ and $\boldsymbol{S_i}$ numbers with CpG sites filtered out. There are 22.5 million nonsynonymous sites, among which ~ 0.2 million sites have one hit ($\boldsymbol{A_1}$ = 195,958) in 1035 patients. The number then decreases sharply as i increases. Thus, $\boldsymbol{A_2}$ = 2946 (number of 2-hit sites), $\boldsymbol{A_3}$ = 99, and $\boldsymbol{A_4}$ + $\boldsymbol{A_5}$ + = 79. We also note that the $\boldsymbol{A_i/S_i}$ ratio increases from 2.89, 2.82, 3.04 to 4.71 and so on.

	All sites			CpG sites removed		
i	A_i	S_i	A_i / S_i	A_i	S_i	A_i / S_i
0	22540623	7804281	2.89	21375384	7014012	3.04
1	195958	69393	2.82	168371	56821	2.96
2	2946	969	3.04	2188	643	3.4
3	99	21	4.71	68	16	4.25
4	23	1	23	17	1	17
5	16	0	16 : 0	9	0	9 : 0
6	10	0	10 : 0	6	0	6 : 0
7	5	0	5 : 0	5	0	5 : 0
8	8	0	8 : 0	6	0	6 : 0
9	4	0	4 : 0	3	0	3 : 0
≥3	178	22	8.09	122	17	7.18
≥4	79	1	79	54	1	54
[10-20]	7	1	7	4	0	4 : 0
≥20	6	0	6 : 0	4	0	4 : 0

Note – The ratio of A_i / S_i is provided as a measure of selection strength.

Table 1

An example of A_i and S_i (from lung cancer, n = 1035).

Fig. 2 shows the average of A_i and S_i among the 12 cancer types (see **Methods**). The salient features are shown by differences between the solid and dotted lines. As will be detailed in **PART II**, the dotted lines, extending linearly from $i = 0$ to $i = 1$ in logarithmic scale, should be the expected values of A_i and S_i, if mutation rate is the sole driving force. In the actual data, A_1 and S_1 decrease to ~0.002 of A_0 and S_0, the step being least affected by selection (see **PART II** later). For A_2 and S_2, the decrease is only ~0.01 of A_1 and S_1. The decrease from i to i+1 becomes smaller and smaller as i increases, suggesting that the process in not entirely neutral. Furthermore, the lower panel of **Fig. 2** shows that A_i/S_i continues to rise as i increases. These patterns again suggest a stronger positive selection at higher i values. The extrapolation lines shown in **Fig. 2** roughly define $i = 3$ as a cutoff where the expected A_3 falls below 1 (see **PART II** for details). The precise model of **PART II** will define high recurrence sites ($i \geq 3$) as CDNs.

PART I - The mutational characteristic of high recurrence sites

In this part, the analyses are done in two different ways. The sequence-feature approach is to examine the mutation characteristics of sequence features (say, 3 mers, 5 mers, etc.) across patients. The patient-feature approach is to examine patients for their mutation signatures and mutation loads.

1. The sequence-feature approach

The simplest and best-known sequence feature associated with high mutation rate is CpG sites. In mammals, methylation and de-amination would enhance the mutation rate from CpG to TpG or CpA by 5~10-fold (Hodgkinson and Eyre-Walker 2011; Ségurel et al. 2014). As the CpG site mutagenesis has been extensively reported, we only present the confirmation in the Supplement (**Fig. S1**). Indeed, CpG sites account for ~6.5% of the coding sequences but contributing ~22% among the mutated sites in **Fig. 2**. Hence, the 7-fold increase in the CpG mutation rate should contribute more to A_i and S_i as i increases. **Table 1** has shown the effects of filtering out CpG sites in the counts of recurrences. Clearly, CpG sites do contribute disproportionately to the recurrences but, even when they are separately analyzed, the conclusion is unchanged. As shown later in **PART II**, every increment of i should decrease the site number by ~0.002 in the TCGA database. Thus, even with a 10-fold increase in mutation rate, the decrease rate would still be 0.02. In the theory sections, CpG site mutations are incorporated into the model.

In this section, we aim to find out how extreme the mutation mechanisms must be to yield the observed recurrences. If these mechanisms seem implausible, we may reject the mutational-hotspot hypothesis and proceed to test the functional hotspot hypothesis.

The analyses of mutability variation by Artificial Intelligence (AI)

The variation of mutation rate at site level could be shaped by multiple mutational characteristics. Epigenomic features, such as chromatin structure and accessibility, could affect regional mutation rate at kilobase or even megabase scale (Stamatoyannopoulos et al. 2009; Makova and Hardison 2015), while nucleotide biases by mutational processes typically span only a few base pairs around the mutated site (Roberts et al. 2013; Haradhvala et al. 2018; Herzog et al. 2021). AI-powered multi-modal integration offers a new tool to quantify the joint effect of various factors on mutation rate variability (Luo et al. 2019; Sherman et al. 2022; Song et al. 2023). Here we explore the association between the mutation recurrence (i) and site-level mutation rate predicted by AI.

Fig. 3A shows the mutation rate landscape across all recurrence sites in breast, CNS and lung cancer using the deep learning framework **Dig** (Sherman et al. 2022). In this approach, the mutability of a focus site is calculated based on both local stretch of DNA and broader scale of epigenetic features. The X-axis shows all mutated sites with $i > 0$ scanned by **Dig**. While the

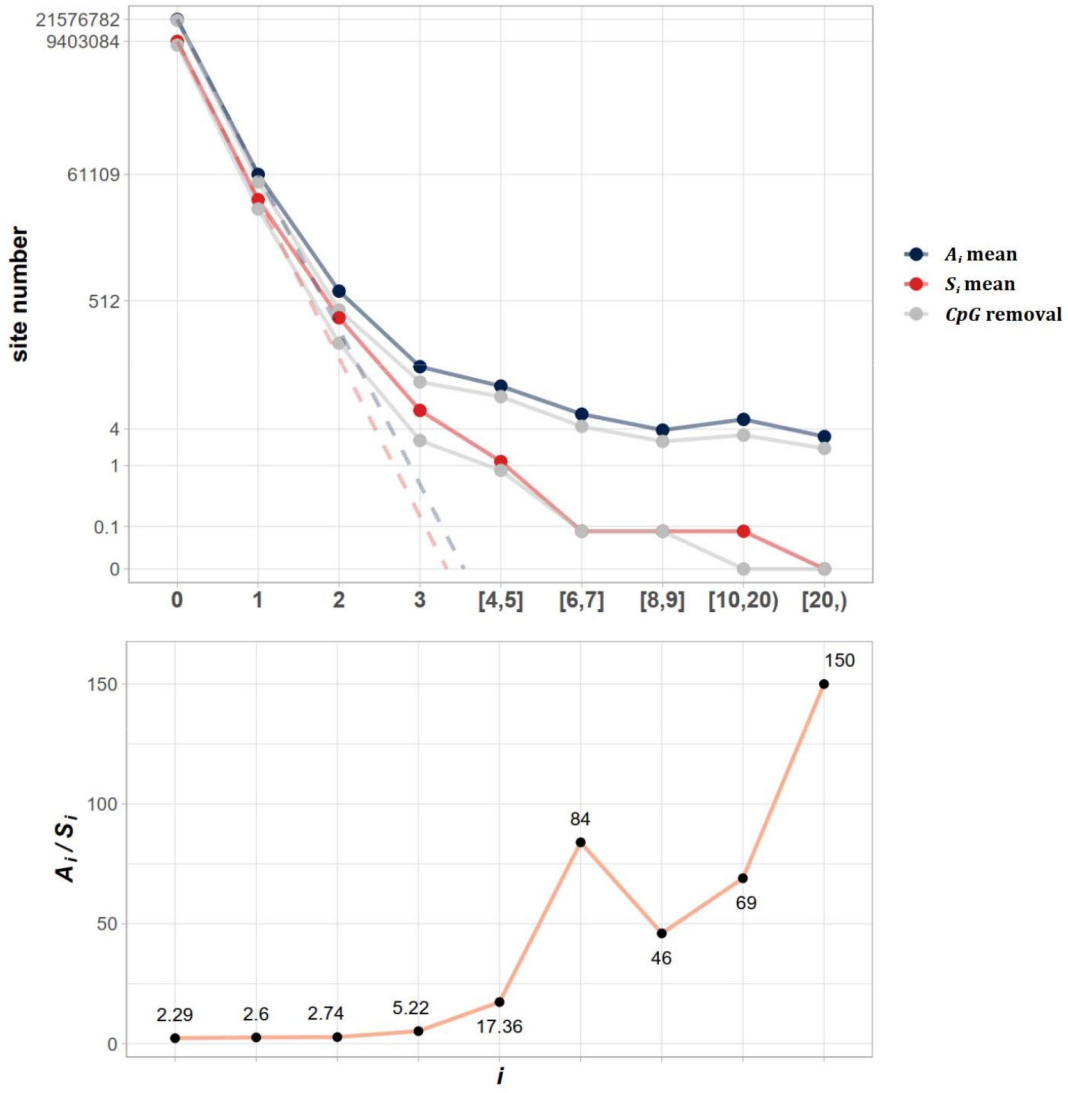


Figure 2

The average A_i and S_i values across different *i* ranges (X-axis). **Top:** The average of A_i and S_i in the log scale. Color lines - full data; gray lines - CpG sites removed. The dash lines are linear extrapolations. **Bottom: The A_i / S_i** ratio as a function of *i*. The drop of A_i / S_i ratio at *i* [8, 9] is due t**o** the potential synonymous CDNs, see Supplement File S2.

mutation rate fluctuates around the average level, we detect no significant difference in mutation rate as a function of ***i*** (**Methods**). In CNS, two sites exhibited exceptional mutability at $i = 6$, surpassing the average by 10-fold. Unsurprisingly, these two are CpG sites and correspond to amino acid change of V774M and R222C in EGFR (Supplement File S2), which are canonical actionable driver mutations in glioma target therapy. In other words, the two sites called by AI for possible high mutability appear to be selection driven.

In **Fig. 3B**, we take a closer look at how CDN is situated against the background of mutation rate variation, using the example of *PAX3* (Paired Box Homeotic Gene 3) (Wang et al. 2008; Li et al. 2019). In this typical example, *Dig* predicts site mutability to vary from site to site. In the lower panel is an expanded look at a stretch of 200 bps. In this stretch, about 8% of sites are 5 to 10-fold more mutable than the average. But none of them are mutated in the data ($i = 0$). There are indeed three sites with $i > 0$ in this DNA segment including a CDN site C1271T (marked by the red star). This CDN site is estimated to have a 2-fold elevation in mutability, which is less than 1/50 of the necessary mutability to reach $i \geq 3$. The other two mutated sites, marked by the green star, are also indicated.

Other AI methods have also been used in the mutability analysis (Fang et al. 2022), reaching nearly the same conclusion. Overall, while AI often suggests sequence context to influence the local variation in mutation rate, the reported variation does not correspond to the distribution of CDNs. In the next subsection, we further explore the local contexts for potential biases in mutability.

The conventional analyses of local contexts - from 3-mers to 101-mers

Since the AI analyses suggest the dominant role of local sequence context in mutability, we carry out such conventional analyses in depth. Other than the CpG sites, local features such as the TCW (W = A or T) motif recognized by APOBEC family of cytidine deaminases (Burns et al. 2013; Roberts et al. 2013), would have impacts as well. We first calculate the mutation rate for motifs of 3-mer, 5-mer and 7-mer, respectively, with 64, 1024, 16384 in number, (see **Methods**). The pan-cancer analyses across the 12 cancer types are shown in **Fig. 4A**. We use **α** to designate the fold change between the most mutable motif and the average. Since the number of motifs increases 16-fold between each length class, the **α** value increases from 4.7 to 8.8 and then to 11.5. Nevertheless, even the most mutable 7-mer, TAACGCG, which has a CpG site at the center, is only 11.52-fold higher than the average. This spread is insufficient to account for the high recurrences, which decrease to ~0.002 for each increment of *i* (see **PART II** below).

A more direct approach is given in **Fig. 4B** explained in the legends. The absence of a trend shows that the high recurrence sites are not associated with the mutability of the motif. In **Fig. 4C**, the analysis extends to longer motifs of 21 bp, 41 bp, 61 bp, 81 bp and 101 bp surrounding the high-recurrence sites. For example, the motif of 101 bp may be (10, 90), (20, 80) and so on either side of a recurrence site. We then compute the pairwise differences in sequences of the motifs among recurrence sites. The logic is that, if certain motifs dictate high mutation rates, we may observe unusually high sequence similarity in the pairwise comparisons. As can be seen in all 5 panels of **Fig. 4C**, there are no outliers in the tail of the distribution. In other words, the sequences surrounding the high-recurrence sites appear rather random. Detailed motif analysis of CDNs within individual cancer types using deep learning models (ResNet, LSTM and GRU) further supports this conclusion.

In conclusion, the analyses by the sequence approach do not find any association between high recurrence sites ($i \geq 3$) and the mutability of the local sequences.

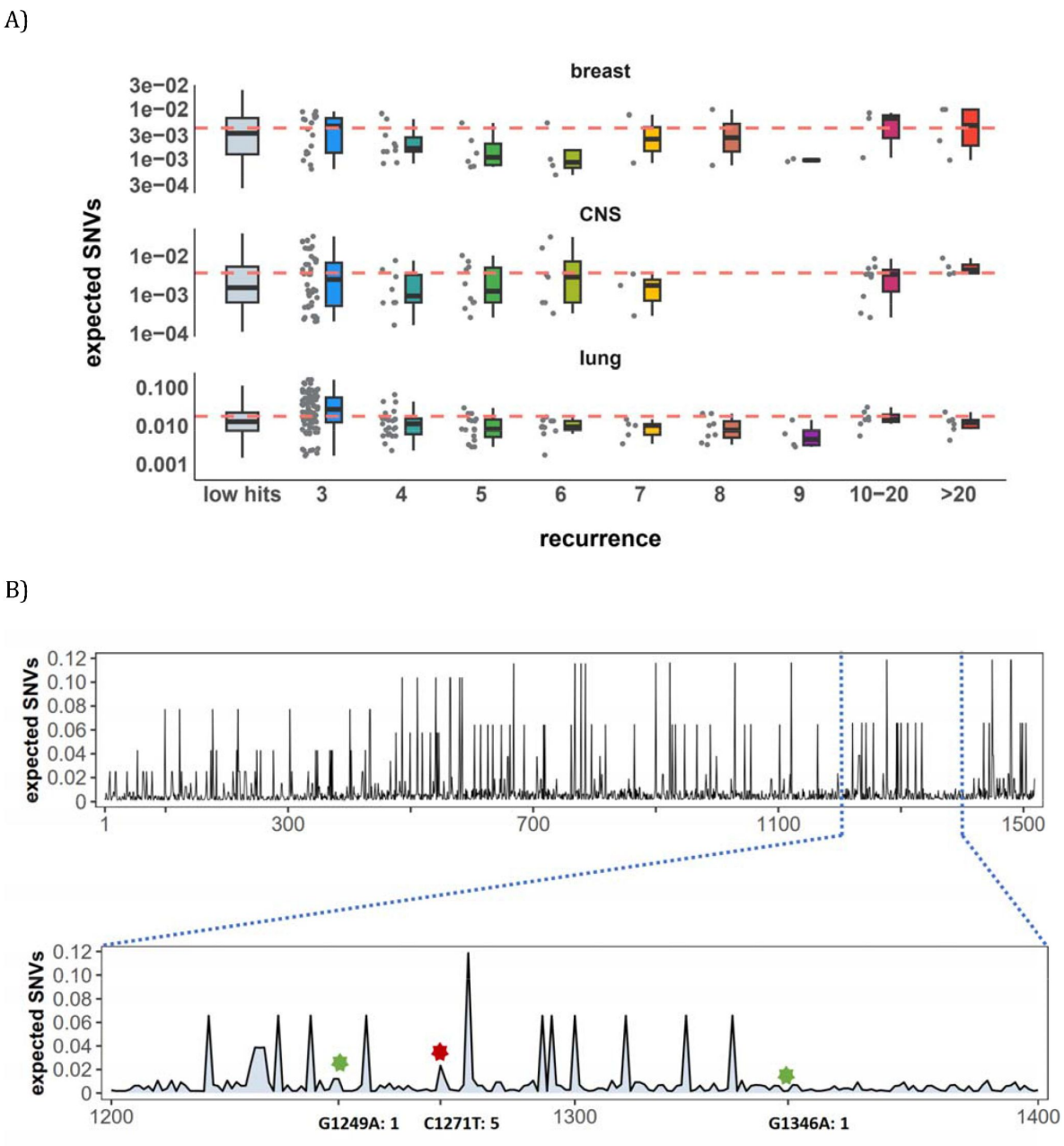


Figure 3

Site-level mutation rate variation obtained from *Dig* (Sherman et al. 2022), a published AI tool. **(A)** Each dot represents the expected SNVs (Y-axis) at a site where missense mutations occurred i times in the corresponding cancer population. The boxplot shows the overall distribution of mutability at *i*, with the red dashed line denoting the average. There is no observable trend that sites of higher *i* are more mutable (The blank areas are due to the absence of CDNs with mutation recurrence counts of 8 or 9 in CNS cancer mutation data, see Supplement File S2). **(B)** A detailed look at the coding region of PAX3 gene in colon cancer. The expected mutability of sites in the 200 bp window is plotted. The three mutated sites in this window, marked by green and red (a CDN site) stars, are not particularly mutable. Overall, the mutation rate varies by about 10-fold as is generally known for CpG sites.

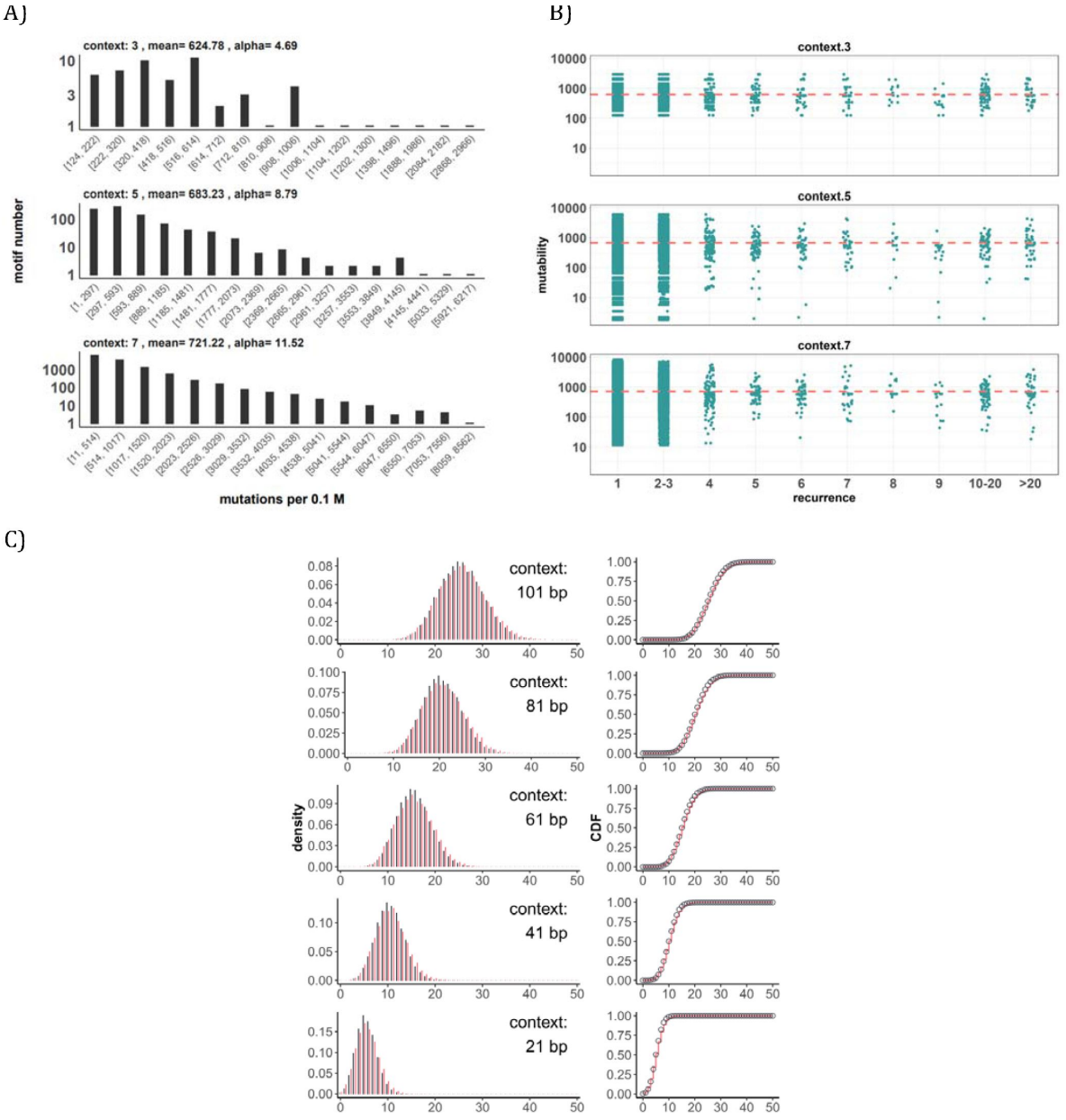


Figure 4

Conventional analyses of local contexts at recurrence sites. **(A)** From top panel down - For the 64 (4^3) 3-mer motifs, their mutational rates are shown on the X-axis. The most mutable motif over the average mutability (α) is 4.69. For the 1024 (=4^5) 5-mer and 16,384 (4^7) 7-mer motifs, the α values are, respectively, 8.79 and 11.52. The most mutable motifs, as expected, are dominated by CpG's. **(B)** Each dot represents the motif surrounding a high-recurrence site. The recurrence number is shown on the X-axis and the mutability of the associated motif's mutability (mutations per 0.1) is shown on the Y-axis. The average mutation rate across all motifs of given length category is indicated by a red horizontal dashed line. The absence of a trend indicates that the high recurrence sites are not associated with the mutability of the motif. **(C)** The analysis is extended to longer motifs surrounding each CDN (21, 41, 61, 81, and 101 bp). For each length group, all pairwise comparisons are enumerated. The observed distributions (black bars and points) are compared to the expected Poisson distributions (red bars and curves) and no difference is observed. Thus, local sequences of CDNs do not show higher-than-expected similarity.

2. The patient-feature approach

In this second approach, we examine the mutation characteristics among patients across sequence features. The first question is whether high recurrence sites tend to happen in patients with higher mutation loads. **Fig. 5A** depicts the distribution of mutation loads among patients harboring a CDN of recurrence *i*. Hence, a patient's load may appear several times in the plot, each appearance corresponding to one CDN in the patient's data. For the comparison across *i* values, the mutation load is normalized by a z-score within each cancer population to equalize the 3 cancer types. The overall trend shows consistently that patients with recurrence sites do not bias toward high mutation loads. The presence of recurrence sites in patients with low mutation loads suggests that overall mutation burden is not a determining factor of recurrence.

With the results of **Fig. 5A**, we then ask a related question: whether these high recurrence mutations are driven by factors that affect mutation characteristics. Such influences have been captured by the analyses of "mutational signatures" (Alexandrov et al. 2013; Alexandrov et al. 2020). Each signature represents a distinct mutation pattern (e.g., high rate of TCT -> TAT and other tri-nucleotide changes) associated with a known factor, such as smoking or an aberrant mutator (aristolochic acid, for example). Each patient's mutation profile can then be summarized by the composite of multiple mutational signatures.

The issue is thus whether a patient's CDNs can be explained by the patient's composite mutational signatures. **Fig. 5B** reveals that in lung cancer, the signature compositions among patients with different recurrence cutoffs are statistically indistinguishable (**Methods**). Smoking (signature SBS4) consistently emerges as the predominant mutational process across all levels of recurrences. In breast cancer, while SBS2 and SBS13 exhibit some differences in the bins of $i^* = 2$ and $i^* = 3$, the profiles remain rather constant for all bins of $i^* \geq 3$. The two lowest bins, not unexpectedly, are also different from the rest in the total mutation load (see **Fig. 5A**). **In table S1**, we provide a comprehensive review of supporting literature on genes with recurrence sites of $i \geq 3$ for breast cancer. In CNS cancer, SBS11 appears significantly different across bins, in particular, $i \geq 20$. This is a signature associated with Temozolomide treatment and should be considered a secondary effect. In short, while there are occasional differences in mutational signatures across i^* bins, none of such differences can account for the recurrences (see **PART II**).

To conclude **PART I**, the high-recurrence sites do not stand out for their mutation characteristics. Therefore, the variation in mutation rate across the whole genome can reasonably be approximated mathematically by a continuous distribution, as will be done below.

PART II - The theory of Cancer Driving Nucleotides (CDNs)

We now develop the theory for $\boldsymbol{S_i}$ and $\boldsymbol{A_i}$ under neutrality where *i* is the recurrence of the mutation at each site. We investigate the maximal level of neutral mutation recurrences (i^*), above which the expected values of $\boldsymbol{S_i}$ and $\boldsymbol{A_i}$ are both < 0. Since no neutral mutations are expected to reach i^*, every mutation with the recurrence of i^* or larger should be non-neutral. Importantly, given that the expected $\boldsymbol{S_i}$ and $\boldsymbol{A_i}$ is a function of $\boldsymbol{U^i}$ where $\mathbf{U} = \mathbf{nE(u)}$ is in the order of 10^{-2} and 10^{-3}, i^* is insensitive to a wide range of mutation scenarios. For that reason, the conclusion is robust.

The mutation rate of each nucleotide (**u**) follows a Gamma distribution with a scale parameter θ and a shape parameter **k**. Gamma distribution is often used for its flexibility and, in this context, models the waiting time required to accumulate **k** mutations. Its mean ($=\mathbf{k}\theta$) and variance ($=\mathbf{k}\theta^2$) are determined by both parameters but the shape (skewness and kurtosis) is determined only by **k**. In particular, the Gamma distribution has a long tail suited to modeling a small number of sites with very high mutation rate.

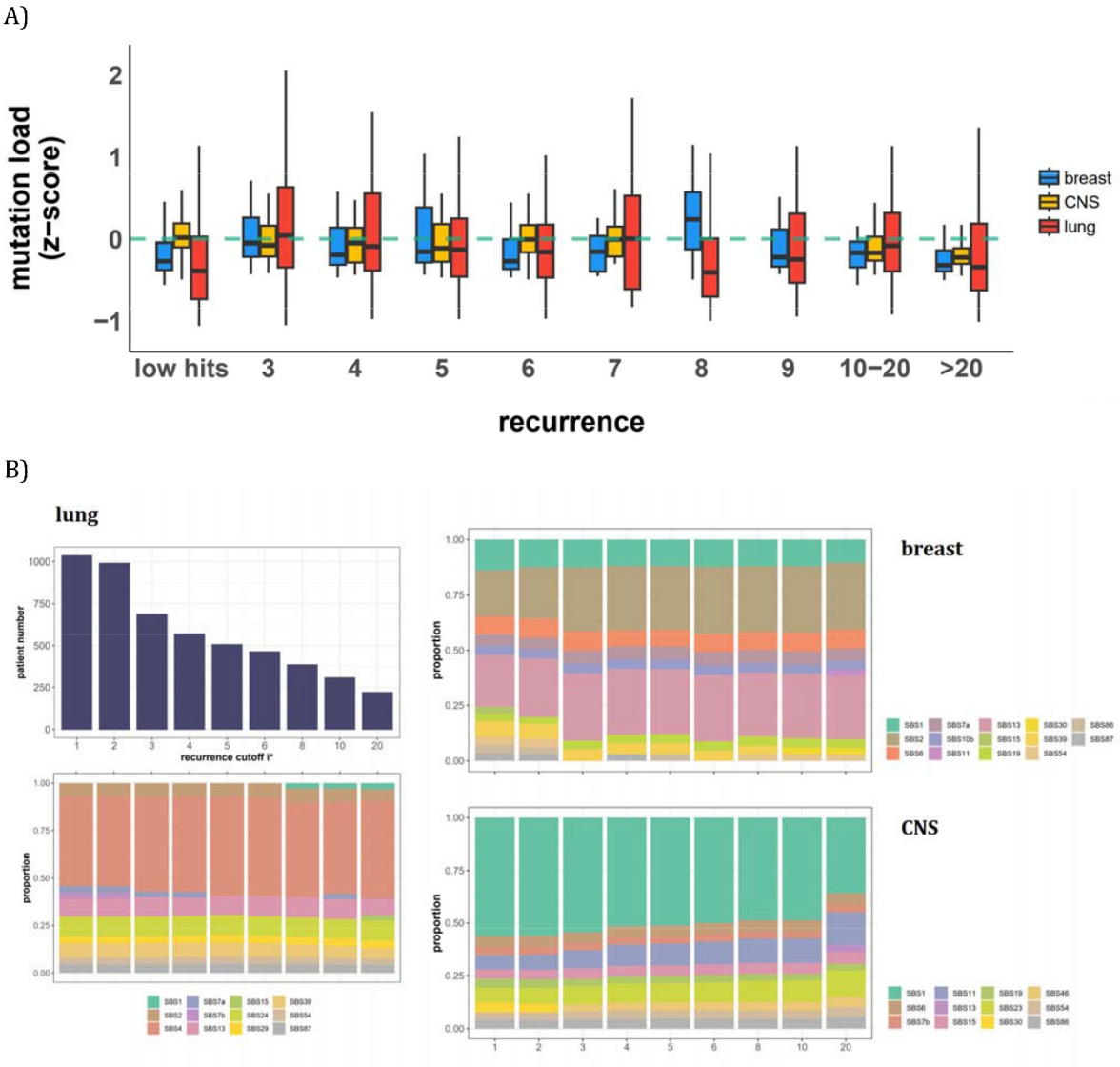


Figure 5

Patient level analysis for mutation load and mutational signatures. **(A)** Boxplot depicting the distribution of mutation load among patients with recurrent mutations. The X-axis denotes the count of recurrent mutations, while the Y-axis depicts the normalized z-score of mutation load (see **Methods**). The green dashed line indicates the mean mutation load. In short, the mutation load does not influence the mutation recurrence among patients. **(B)** Signature analysis in patients with mutations of recurrences ≥ i^* (X-axis). For lung cancer (left), the upper panel presents the number of patients for each group, while the lower panel depicts the relative contribution of mutational signatures. For breast cancer, APOBEC-related signatures (SBS2 and SBS13) are notably elevated in all groups of patients with $i^* \geq 3$, while patients with mutations of recurrence ≥ 20 in CNS cancer exhibit an increased exposure to SBS11 (Blough et al. 2011; Lin et al. 2021; Noeuveglise et al. 2023). Again, patients with higher mutation recurrences do not differ in their mutation signatures.

We now use synonymous ($\boldsymbol{S_i}$) mutations as the proxy for neutrality. Hence, in **n** samples,

$$S_i = \sum_{l=1}^{L_S} C_n^i u(l)^i [1-u(l)]^{n-i} \sim C_n^i L_S E(u^i)$$

Eq. 1

where $\boldsymbol{L_S}$ is the total number of synonymous sites and **u(*l*)** is the mutation rate of site ***l***. In the equation above, the term [1-*u*(*l*)]$^{n-i}$ is dropped. We note that $[1-u(l)]^{n-i} \sim e^{[-u(l)(n-i)]} \sim 1$ as $\boldsymbol{u}$ is in the order of 10^{-6} and **n** is in the order of 10^2 from the TCGA data. With the gamma distribution of $\boldsymbol{u}$ whereby the i^{th} moment is given by

$$E(u^i) = \frac{\Gamma(k+i)}{\Gamma(k)}\theta^i = \frac{\Gamma(k+i)}{\Gamma(k)\cdot k^i}E(u)^i$$

we obtain:

$$S_i = L_S \cdot g(i,k)[nE(u)]^i$$

Where:

$$g(i,k) = C_{k+i-1}^i \cdot \frac{1}{k^i}$$

In a condensed form,

$$S_i = G \cdot [nE(u)]^i$$

Eq. 2

where $G=L_S \cdot g(i,k)$,. Similarly, if nonsynonymous mutations are assumed neutral, then,

$$A_i = L_A \cdot g(i,k)[nE(u)]^i \sim 2.3 \cdot S_i$$

Eq. 3

The number of 2.3 is roughly the ratio of the total number of nonsynonymous over that of synonymous sites (Hartl and Clark 1989; Li 1997; Chen et al. 2019). This number would vary moderately among cancers depending on their nucleotide substitution patterns.

E(u) of Eqs. 2-3 is generally (1∼5) × 10^{-6} per site in cancer genomic data and **n** is generally between 300 and 1000. Hence, **nE(u)** is the total mutation rate summed over all **n** patients and is generally between 0.001 and 0.005 in the TCGA data. Given **nE(u)** is in the order of 10^{-3}, $\boldsymbol{S_i}$ and $\boldsymbol{A_i}$ would both decrease by 2∼3 orders of magnitude with each increment of *i* by 1. We note that the total number of synonymous sites, $\mathbf{L}_S$, is ∼0.9 × 10^7 and $\boldsymbol{L}_A$ is ∼2.3 times larger. Therefore, $\boldsymbol{S}_3 < 1$ and $\boldsymbol{A}_3 < 1$. When *i* reaches 4, $\boldsymbol{S_{i\geq 4}}$ and $\boldsymbol{A_{i\geq 4}}$ would both be ⍰ 1 when averaged over cancer types.

For each cancer type, the conclusion of $\boldsymbol{S_3} < 1$ and $\boldsymbol{A_3} < 1$ is valid with the actual value of $\boldsymbol{S_3}$ and $\boldsymbol{A_3}$ ranging between 0.01 and 1. In other words, with **n** < 1000, neutral mutations are unlikely to recur 3 times or more in the TCGA data ($i^* = 3$). While $\boldsymbol{S_{i\geq 3}}$ and $\boldsymbol{A_{i\geq 3}}$ sites are high-confidence CDNs, the value of i^* is a function of **n**. At **n** ≤ 1000 for the TCGA data, i^* should be 3 but, when **n** reaches 10,000, i^* will be 6. The benefits of large **n**'s will be explored in **PART III**.

Possible outliers to the distribution of mutation rate

Although we have explored extensively the sequence contexts, other features beyond DNA sequences could still lead to outliers to mathematically distributions. These features may include DNA stem loops (Buisson et al. 2019) or unusual epigenetic features (Zheng et al. 2014;

Makova and Hardison 2015; Supek and Lehner 2015). We therefore expand the model by assuming a small fraction of sites (**p**) to be hyper-mutable that is **α** fold more mutable than the genomic mean. Most likely, **α** and **p** are the inverse of each other. For the bulk of sites (1-**p**) of the genome, we assume that their mutations follow the Gamma distribution. (Nevertheless, the bulk can be assumed to have a fixed mutation rate of **E(u)** without affecting the conclusion qualitatively.)

We let **p** range from 10^{-2} to 10^{-5} and **α** up to 1000. As no stretches of DNA show such unusually high mutation rate (1000-fold higher than the average), such sites are assumed to be scattered across the genome and are rare. With the parameter space of defined above, we choose the (***p***, **α**) pairs that agree with the observed values of $\boldsymbol{S_1}$ to $\boldsymbol{S_3}$ which are sufficiently large for estimations. **Table 2** presents the value range and standard deviation for **p** and **α** across the 6 cancer types that have >500 patient samples. Among the 6 cancer types, the lung cancer data do not conform to the constraints and we set **p** = 0. With observed values for $\boldsymbol{S_{1\sim3}}$ as constraints, $\mathbf{S_4}$ rarely exceeds 1. Hence, even with the purely conjectured existence of outliers in mutation rate, $i^* = 4$ is already too high a cutoff.

Table 2 suggests that ***p*** has to be smaller than 10^{-5} and **α** > 1000 to yield $\mathbf{S_4} > 1$. Since the coding region has 3×10^7 sites, $\boldsymbol{p} < 10^{-5}$ would mean that the outliers are at most in the low hundreds. In other words, the number of high recurrence sites projected by the theory is close to the observed numbers. Therefore, there are really no unknown outlier sites of high mutation rate. Positive selection would be a more straightforward explanation, explained below.

The influence of selection on mutation occurrences

We now show that, although the mutational bias alone cannot account for the high occurrences, selection can easily do so. We assume a fraction, ***f***, of $\boldsymbol{A_i}$'s to be under positive selection. The fraction should be small, probably ≪ 0.01, and will be labeled. The rest, labeled $\boldsymbol{A_i}$ is considered neutral. Hence, $\boldsymbol{A_i}$ is proportional to $[nE(u)]^i$ and $\boldsymbol{A_i}/\boldsymbol{S_i} = \boldsymbol{L_A}/\boldsymbol{L_S} \sim 2.3$. Like $\boldsymbol{S_{i\geq4}}$, $\boldsymbol{A_{i\geq4}} \ll 1$. In contrast,

$$A_i^* = G^*[w \cdot nE(u)]^i \qquad \text{Eq.4}$$

$$\frac{A_i^*}{S_i} = \frac{G^*}{G} w^i \qquad \text{Eq.5}$$

where $\boldsymbol{G^*}$ is also a constant but its value depends on ***f***. The crucial parameter, ***w*** (= 2Ns), is the selective advantage (*s*) scaled by the population size of progenitor cancer cell (*N*). Since ***w*** can easily be >10, even at $i = 3$, A_3^*/S_3 would be >100 as large as A_1^*/S_1. In other words, observed mutation recurrences at $i \geq 3$ for advantageous mutation should not be uncommon. **Eq. 4** also shows that ***w*** and **E(u)** jointly affect the recurrence; therefore, CpG sites (many of which fall in functional sites) are expected to be strongly represented among high recurrence sites.

PART III. The theory of large samples ($n > 10^5$) and identification of all CDN's

Using the theory of CDN developed above, the companion paper shows that each sequenced cancer genome in the current databases, on average, harbors only 1∼2 CDNs. The number varies in this range depending on the cancer type (Zhang et al. 2024). For comparison, tumorigenesis may require at least 5∼10 driver mutations as estimated by various criteria (Armitage and Doll 1954; Hanahan and Weinberg 2011; Belikov 2017; Martincorena et al. 2017; Anandakrishnan et al. 2019; Campbell et al. 2020). The results show that there are many more

Cancer Type	S_3	p	α	S_4	S_5
Lung*	--	*0.0*	--	--	--
breast	0.12	8.75E-04 (8.21E-04)	88.6 (32.0)	0.102 (0.068)	0.004 (0.004)
CNS	0.02	2.73E-04 (1.09E-04)	295.1 (57.0)	0.448 (0.173)	0.026 (0.015)
kidney	0.03	3.03E-05 (2.98E-05)	304.1 (108.0)	0.067 (0.056)	0.005 (0.006)
Upper-AD tract	0.47	0.002 (0.001)	48.9 (10.7)	0.174 (0.078)	0.005 (0.003)
Large intestine	1.03	0.009 (0.001)	51.6 (1.4)	0.998 (0.087)	0.026 (0.003)

Note – For each cancer type, p stands for the proportion of highly mutable sites, with mutation rate being α-fold of the average. S_3 gives the expected number without mutable outliers ($p = 0$). S_4 and S_5 denote the expected number with the best (p, α) pairs with the standard deviation in parentheses. For lung cancer, S_2 and S_3 do not fit the outlier model (Supplement File S1); therefore, we set $p = 0$.

Table 2

Summary for modeling outlier sites in 6 cancer types

CDNs that have not been discovered. This is not unexpected since most CDNs are found in <1% of patients. If each CDN is observed in 1% of patients and each patient has 5 CDNs, then there should be at least 500 CDNs for each cancer type.

In the companion study that uses the **A/S** ratios of **Fig. 1**, the estimated number of CDNs ranges from 500 to 2000, whereas the current estimates based on $\boldsymbol{A_{i\geq 3}}$ sites is only 50~100 (Zhang et al. 2024). Where, then, are the undiscovered CDNs and how to find them? Since all $\boldsymbol{A_{i\geq 3}}$ sites are concluded to be CDNs, the bulk of CDNs must be among $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ sites. The best way to identify the CDNs hidden in $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ is to increase the sample size, **n**, dramatically.

We hence extend Eq. 1 for $\boldsymbol{S_i}$ and $\boldsymbol{A_i}$ to large **n**'s. Note that Eq. 1 drops the term $(1\text{-}u)^{n\text{-}i}\sim e^{-u(n\text{-}i)}$ of as it is ~ 1, when **nE(u)** ≫ 1. With a large **n** when $e^{u(n\text{-}i)}$ is not near 1, the recurrence of mutations would follow a Poisson distribution with the expected value of **nE(u)**. Assuming that $\boldsymbol{u}$ follows a gamma distribution with a shape parameter of $\boldsymbol{k}$, the probability of observing i mutations would follow the negative binomial distribution as shown below:

$$f\left(i|k,n,E(u)\right)=\frac{\Gamma(i+k)}{\Gamma(i+1)\Gamma(k)}k^k[nE(u)]^i[k+nE(u)]^{-k-i}$$

Eq. 6

The cumulation density function for Eq. 6 is then:

$$F(i\leq i^*)=\sum_{i=1}^{i^*}f\left(i|k,n,E(u)\right)$$

Eq. 7

Eq. 7 Then, by definition, $\boldsymbol{A_{i\geq i^*}}$ should be ≫ 1 so that mutations with recurrences $i > i^*$ could be defined as CDN. Thus:

$$F(i\leq i^*)=1-\frac{\varepsilon}{L_A}$$

Eq. 8

Where $\varepsilon = \boldsymbol{A_{i\geq i^*}}$ denotes the number of sites with mutation recurrence $\geq i^*$ under the sole influence of mutational force. $\frac{\varepsilon}{L_A}$ could then be regarded as significance of i^* since it controls the overall false positive rate of CDNs.

Specifically, with $\boldsymbol{k}$ = 1, the probability function of mutation recurrence of a given site would transform to a geometric distribution with $\boldsymbol{p}$ = 1 / (1 + **nE(u)**), the cumulative density function (CDF) is then:

$$F(i\leq i^*)=1-\left(1-\frac{1}{1+nE(u)}\right)^{i^*}$$

Eq. 9

Combined with Eq. 5, the mutation recurrence cutoff i^* of being a CDN could be expressed as:

$$i^*=\frac{log\left(\frac{\varepsilon}{L_A}\right)}{log\left(\frac{nE(u)}{1+nE(u)}\right)}$$

Eq. 10

For very large **n**, 1 / **nE(u)** is small and $i^*/\mathbf{n}$ can be approximated as

$$i^*/\boldsymbol{n} = log(\boldsymbol{L_A}) \cdot \boldsymbol{E}(\boldsymbol{u}) \sim 5 \times 10^{-5} \quad \text{Eq. 11}$$

Eq. 11 shows that $i^*/\mathbf{n}$ would approach asymptotically as **n** increases. This asymptotic value is attained when n reaches $\sim 10^6$.

Fig. 6 shows the range of i^* for **n** up to 10^6. As expected, i^* increases by small increments while **n** increases in 10-fold jumps. For example, when **n** increases by 3 orders of magnitude, from 100 to 100,000, i^* only doubles from 3 to 12. The disproportional increment between i^* and **n** explains why we use the actual number of i^* for the cutoff, instead of the ratio $i^*/\mathbf{n}$. As shown in the inset of **Fig. 2**, the ratio of $i^*/\mathbf{n}$ would approach the asymptote at $\mathbf{n} \sim 10^6$, where an advantageous mutation only needs to rise to 0.00006 to be detected. With **n** reaching this level, we shall be able to separate most CDNs apart from the mutation background.

When **n** approaches 10^5, the number of CDNs will likely increase more than 10-fold as conjectured in **Fig. 7A**. In that case, every patient would have, say, 5 CDNs that can be subjected to gene targeting. (The companion study shows that, at present, an average cancer patient would have fewer than one targetable CDN.) Before the project is realized, it is nevertheless possible to test some aspects of it using the GENIE data of targeted sequencing. Such screening for mutations in the roughly 700 canonical genes serves the purpose of diagnosis with **n** ranging between 10~17 thousands for the breast, lung and CNS cancers. Clearly, GENIE efforts did not engage in discovering new mutations although they would discover additional CDNs in the canonical genes. In the companion study, we demonstrate that the analysis of CDNs identifies a potential set of 1.6 times more driver genes than those detected by whole gene selection signal calls (Zhang et al. 2024).

Fig. 7A assumes that the prevalence, $\boldsymbol{i}/\mathbf{n}$, should not be much affected by **n**, but the cutoff for CDNs, $\boldsymbol{i}^*/\mathbf{n}$, would decrease rapidly as **n** increases. **Fig. 7B** shows that the $\boldsymbol{i}/\mathbf{n}$ ratios indeed correspond well between TCGA and GENIE, which differ by 10 to 20-fold in sample size. Importantly, as predicted in **Fig. 6**, the number of CDNs increases by 3 to 5-fold (**Fig. 7C - E**). Many of these newly discovered CDNs from GENIE are found in the $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ classes of TCGA while many more are found in the $\boldsymbol{A}_0$ class in TCGA. In conclusion, increasing **n** by one to two orders of magnitude would be the simplest means of finding all CDNs.

Discussion

The nature of high-recurrence mutations has been controversial. Many authors have argued for mutational hotspots (Hess et al. 2019; Stobbe et al. 2019; Nesta et al. 2021; Bergstrom et al. 2022; Wong et al. 2022) but just as many have contended that they are CDNs driven by selection (Gartner et al. 2013; Chang et al. 2016; Bailey et al. 2018; Cannataro et al. 2018; Juul et al. 2021; Zhao et al. 2021; Zeng and Bromberg 2022). While the two views co-exist, they are in fact incompatible. If the mutational hotspot hypothesis is correct, the selection hypothesis, and the determination of CDNs, would not be needed.

We believe that this study is the first to comprehensively test of the null hypothesis of mutational hotspots. In **PART I**, the mutational characteristics near all putative CDNs are examined and **PART II** presents the probability theory based on the analyses. The conclusion is that it is possible to reject the null hypothesis for recurrences as low as 3 in the TCGA data. The main reason for the high sensitivity is shown in Eqs. 2 and 3 where $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$ is proportional to $[\mathbf{nE}(\boldsymbol{u})]^i$ or, roughly 0.002^i. We recognized that the conclusion is based on what we currently know about

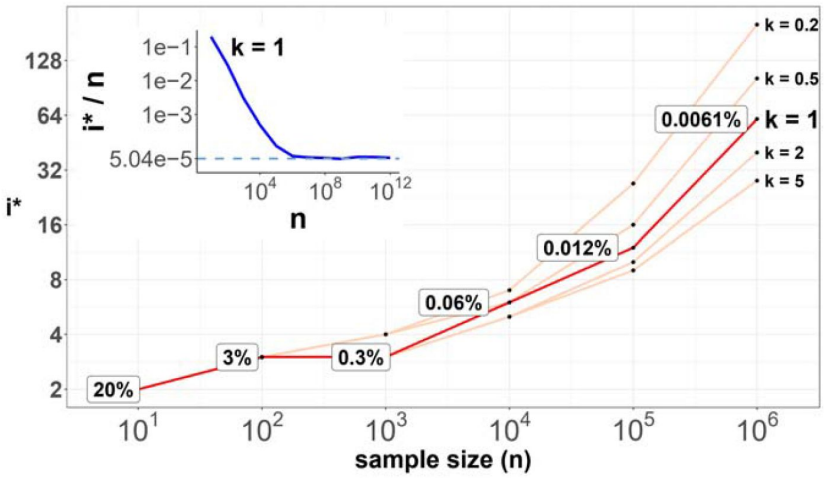


Figure 6

i^* values (Y-axis, log scale) against sample sizes (n, X-axis) across different shape parameter k's. The Y axis presents the i^* values under different sample sizes (n of the X-axis) in log scale. Five shape parameters (k) of the gamma-Poisson model are used. In the literatures on the evolution of mutation rate, k is usually greater than 1. The inset figure illustrates how $i^*/\mathbf{n}$ (prevalence) would decrease with increasing sample sizes. The prevalence would approach the asymptotic line of [ ] when **n** reaches 10^6. In short, more CDNs (those with lower prevalence) will be discovered as n increases. Beyond $\mathbf{n} = 10^6$, there will be no gain.

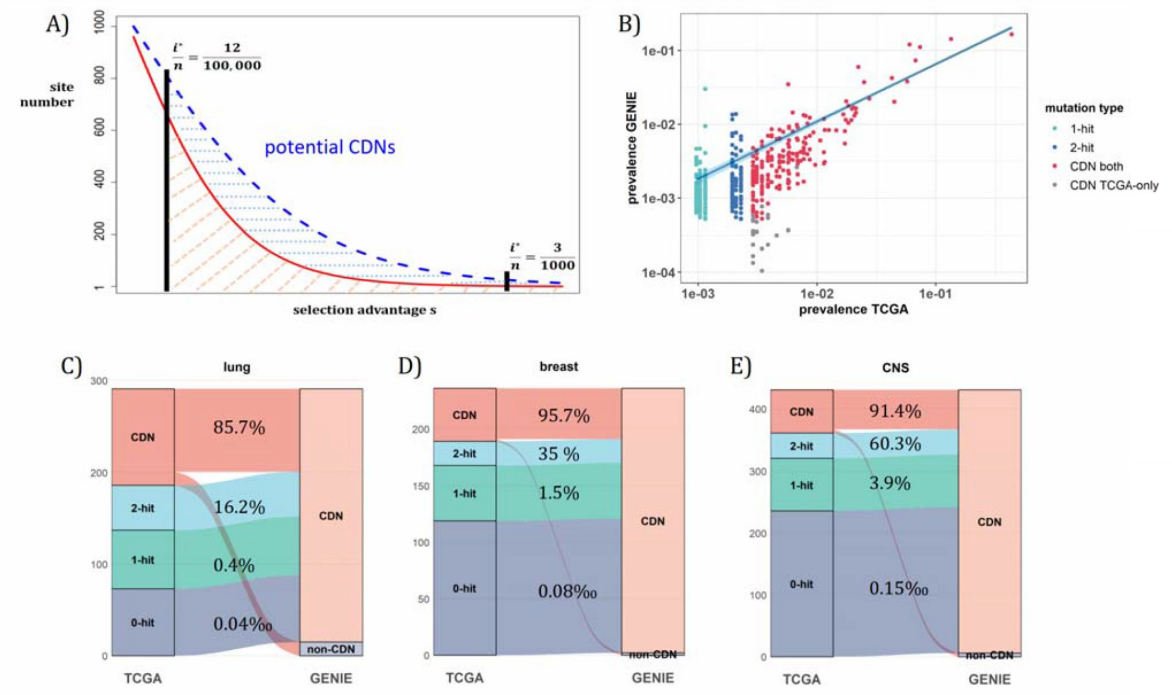


Figure 7

Analysis of CDNs with expanded sample set in GENIE. **(A)** Schematic illustrating the impact of sample size expansion on the number of discovered CDNs. The two vertical lines show the cutoffs of i^* /**n** at (3/1000) vs. (12/100,000). The Y axis shows that the potential number of site would decrease with i^* /**n**, which is a function of selective advantage. The area between the two cutoffs below the line represent the new CDNs to be discovered when n reaches 100,000. The power of **n** = 100,000 is even larger if the distribution follows the blue dashed line. **(B)** The prevalence (i^*/n) of sites is well correlated between datasets of different **n** (TCGA with **n** < 1000 and GENIE with n generally 10-fold higher), as it should be. Sites are displayed by color. "1-hit": CDNs identified in GENIE but remain in singleton in TCGA, "2-hit": CDNs identified in GENIE but present in doubleton in TCGA. "CDN both": CDNs identified in both databases. **(C-E)** CDNs discovered in GENIE (**n** > 9000) but absent in TCGA (**n** < 1000). The newly discovered CDNs may fall in TCGA as 0 - 2 hit sites. The numbers in the middle column show the percentage of lower recurrence (non-CDN) sites in TCGA that are detected as CDNs in the GENIE database, which has much larger **n**'s.

mutation mechanisms. In a sense, the theory developed here can help the search for such unknown mutation mechanisms, if they do exist. Finally, the theory developed would permit the explorations in several new fronts when the sample size, **n**, expands to 10^5.

The first front is to identify (nearly) all CDNs. When **n** reaches 10^5, any point mutation with a prevalence higher than 12/100,000 would be a CDN, which is 25-fold more sensitive than in the TCGA data (3/1000). The companion analysis suggests that CDNs with lower prevalence, say 12/100,000 may still be highly tumorigenic in patients with the said mutation. If prevalence and potence are indeed poorly correlated, the search for lower prevalence CDNs by increasing **n** to 10^5 is equivalent to searching for less common but still potent cancer driving mutations.

The second one is functional tests in patient-derived cell lines. When we have all CDNs identified, a patient can be expected to have multiple (≥5; Zhang et al. 2024) CDNs. These mutations will be the basis of in vitro test, as well as in animal model experiments, by gene editing, as shown recently (Hodis et al. 2022). Targeting multiple mutations simultaneously is necessary and may even be sufficient.

The third front, and arguably the most important one, is cancer treatment by targeted therapy (Dang et al. 2017; Danesi et al. 2021; Waarts et al. 2022; Lin et al. 2023; Zhou et al. 2023). When multiple CDN mutations in the same patient can be simultaneously targeted, the efficacy should be high. No less crucial, resistance to treatment should be diminished since it would be harder on cancer cells to evolve multiple escape routes to evade multiple drugs. Moreover, CDN analysis is crucial for stratifying patients for targeted therapy, as only targeting genes that are positively selected during cancer evolution can truly achieve therapeutic effects.

There will be other fronts to explore with the full set of CDNs. A large database will facilitate the detection of negative selection which has eluded detection (Q. Chen et al. 2022). Chen et al. have analyzed a curious phenomenon in somatic evolution, which they term “quasi-neutral evolution” (Chen et al. 2019). It will also be possible to study the evolution of mutation mechanisms in cancer cells based on such large samples (Jackson and Loeb 1998; Ruan et al. 2020). This last topic is addressed in the Supplement. Finally, at the center of evolutionary genetics is the multi-genic interactions that control complex phenotypes such as human diseases (e.g. diabetes) (Vujkovic et al. 2020; Lagou et al. 2023; Xue et al. 2023; Suzuki et al. 2024), genetics of speciation (Q. Chen et al. 2022; Pan, Zhang, et al. 2022) and the emergence of viral strains (Deng et al. 2022; Pan, Liu, et al. 2022). Cancers may be the first such complex genetic systems that can be unraveled thanks to the massively repeated evolution. As cancer genomics is increasingly adopted in cancer treatment, these benefits should become apparent when **n** reaches 10^5 for most cancer types.

Methods

Data Collection

Single nucleotide variation (SNV) data for the TCGA cohort was downloaded from the GDC Data Portal (*https://portal.gdc.cancer.gov/*, data version 2022-02-28). Only mutations identified by at least two pipelines were included in this study. Mutations were further filtered based on their population frequency recorded in the Genome Aggregation Database (**gnomAD**, version v2.1.1), with an upper threshold of 1‰. We focused on coding region mutations of missense, nonsense, and synonymous types on autosomes. The mutation load for each patient was defined as the sum of these three types of mutations. Patients with a mutation load exceeding 3000 were identified as having a mutator phenotype and were excluded from our analysis. In total, 7369 samples representing 12 cancer types were included for mutation analysis.

Additional mutation data was acquired from the AACR Project GENIE Consortium via cBioPortal. Due to the prevalence of targeted sequencing within the dataset, filtering was implemented to ensure the inclusion of samples with sequencing assays encompassing all exonic regions of target genes. Furthermore, sample-level filtering was performed to guarantee a unique sequencing sample per patient. Germline filtering was applied to the resulting point mutations, removing mutations with SNP frequencies exceeding 0.0005 in any subpopulation annotated by gnomAD. To exclude patients exhibiting hypermutator phenotypes, mutation loads were scaled to the whole exon level with $\widetilde{n_L} = n_l \cdot \frac{L}{l}$ where n_i and l represent the mutation load and genomic length of target sequencing region, respectively. L denotes the genome-wide whole coding region length. Patients with $\widetilde{n_L} > 3000$ were subsequently excluded, consistent with the threshold employed for the TCGA dataset.

Calculation of missense and synonymous site number (L_A and L_S)

The idea of missense or synonymous sites originates from the question that how many missense or synonymous mutations would be expected if each site of the genome were mutated once given background mutation patterns. Here, the background mutation patterns refer to intrinsic biases in the mutational process, such as the over-representation of C>T (or G>A) mutations at CpG sites due to spontaneous deamination of 5-methylcytosine. In coding regions, four-fold degenerate sites are generally considered neutral, as any mutation path would not alter the encoded amino acid sequence. This analysis follows established methods at the single-base level to infer the expected number of missense and synonymous sites across the genome (Gojobori et al. 1982; Wu and Maeda 1987; Hartl and Clark 1989; Martincorena et al. 2017).

To illustrate the calculation process, we provide an example of synonymous site estimation. At four-fold degenerate sites throughout the genome, we tally the number of mutations from base $\boldsymbol{m}$ to base $\boldsymbol{v}$ as $n_{m>v}$ ($m,v \in \{A,C,G,T\}, m \neq v$). The likelihood of observing a mutation from reference base m to variant base $\boldsymbol{v}$ will be $r_{m>v} = \frac{n_{m>v}}{N_m}$, where N_m represents the number of four-fold degenerate site with reference base m. We then normalize all likelihoods by $R_{m>v} = \frac{r_{m>v}}{\sum r}$, where $\sum r$ represents the sum of likelihoods across 12 possible mutation paths, $R_{m>v}$ thus describes the relative probability of an occurred mutation to be $m>v$ at any site. For a given genomic coding region of length $\boldsymbol{L}$, the synonymous site number will be:

$$L_S = \sum_L \delta_{m>v}^{syn} R_{m>v} \qquad \text{Eq. S1}$$

With $\delta_{m>v}^{syn}$ being a Kronecker delta function where:

$$\delta_{m>v}^{syn} = 1 \text{ if } m > v \text{ is synonymous}$$

$$\delta_{m>v}^{syn} = 0 \text{ otherwise}$$

Similarly, the expected number of missense sites $\boldsymbol{L_A}$, is calculated as follows:

$$L_A = \sum_L \delta_{m>v}^{mis} R_{m>v} \qquad \text{Eq. S2}$$

Calculation of A_i and S_i

For each mutated site, we track its number of recurrent mutations i ($i > 0$) across two mutation categories: missense and synonymous. Subsequently, we aggregate across entire coding region to count the number of sites harboring i missense mutations ($\boldsymbol{A_i}$) and synonymous mutations ($\boldsymbol{S_i}$). For

i = 0, we define $\boldsymbol{A_0}$ and $\boldsymbol{S_0}$ as the estimated number of potential missense and synonymous sites thatremain unmutated within the current sample size. A_0=L_A-$\sum_{i>0}A_i$, S_0=L_S-$\sum_{i>0}S_i$.

AI-based mutation rate analysis

To capture the complex interplay of genomic and epigenomic factors influencing mutation susceptibility, we employed pre-trained artificial intelligence (AI) models from *Dig*, an aggregated tool combining deep learning and probabilistic models (Sherman et al. 2022). Downloaded from the *Dig* data portal (*http://cb.csail.mit.edu/cb/DIG/downloads/*), these models leverage a rich set of features encompassing both kilobase-scale epigenomic context (replicating timing, chromatin accessibility, etc.) and fine-grained base-pair level information (such as sequence context biases) to predict the site level mutation rate. For each cancer type, we re-fitted the pre-trained models with mutations analyzed in our study. The mutation rate for each site, scaled by population size, was obtained via the *elementDriver* function within Dig, and was represented by *EXP_SNV* from the final results. For a closer look at mutation rate landscape of *PAX3*, we re-fitted the AI-model with point mutations from large intestine cancer. The mutation rates were generated site-by-site for the coding regions of *PAX3*.

Given the scarcity of mutated sites with recurrence $i \geq 3$ (comprising only 0.15% of all mutated sites), a rigorous statistical approach was adopted to assess the significance of mutability differences between these high-hit groups and low-hit groups. We implemented the procedure as follows:

1. Raw significance level: For each recurrence group i (containing $\boldsymbol{A_i}$ sites), a one-sided Kolmogorov-Smirnov (K-S) test was employed to calculate a raw significance level (denoted as p_0) against the low-hit group.
2. Resampling for Significance Pool: we resample $\boldsymbol{A}_i$ sites from the entire pool of mutated sites with missense mutations. The significance $\{p_j, j=1,2 ..,100,000\}$ from one-sided K-S test is calculated against the low hit group. The resampling process was repeated 100,000 times, generating a distribution of resampled significance levels, denoted as.
3. Adjusted Significance Level: the raw significance p_0 was then compared to the resampled significance pool $\{p_j, j=1,2 ..,100,000\}$. The proportion of $p_0 \leq p_j$ was then calculated as the adjusted significance level that accounted for potential sampling effects.

Motif based mutability

For a given nucleotide base, we extended the sequence to each side by 1, 2, and 3 base pairs, producing sets of 3-mer, 5-mer, and 7-mer motifs, respectively. We then pooled point mutations from 12 cancer types to create a comprehensive dataset. For 3-mer and 5-mer motifs, we utilized synonymous mutations as the reference for mutation rate calculations. For 7-mer motifs, the vast number of possible sequence combinations (4^7 = 16,384) posed a challenge, as synonymous mutations alone might not adequately cover all potential contexts. To address this, we employed all singleton mutations (1-hit mutations) from both missense and synonymous categories for 7-mer motif analysis. This decision was based on the assumption that singleton mutations are less affected by selective pressures, supported by the genome-wide observation that the ratio of missense to synonymous singletons ($\boldsymbol{A_1}/\boldsymbol{S_1}$) approximates the ratio of unmutated missense to synonymous sites ($\boldsymbol{A_0}/S_0$). The site number for 3-mer and 5-mer motifs of a given context $\boldsymbol{c}$ was calculated as follows:

$$L_{c,S} = \sum_{L} \delta_{c,m>v}^{syn} \cdot R_{m>v}$$

Which is an extension of Eq. S1, with $\delta_{c,m>v}^{syn}$ being a Kronecker delta function where:

$$\delta_{c,m>v}^{syn} = 1 \text{ if base change of } \boldsymbol{m} \text{ to } \boldsymbol{v} \ (\boldsymbol{m} > \boldsymbol{v}) \text{ is synonymous under sequence context } c.$$

$$\delta_{c,m>v}^{syn} = 0 \text{ otherwise.}$$

The mutation rate then could be expressed as:

$$\mu_c = \frac{n_{c,m>v}^{syn}}{L_{c,S}}$$

Eq. S3

for 7-mer motifs, the calculation is:

$$L_c = L_{c,S} + L_{c,A} = \sum_L \delta_{c,m>v}^{syn} \cdot R_{m>v} + \sum_L \delta_{c,m>v}^{mis} \cdot R_{m>v}$$
$$\mu_c = \frac{n_{c,m>v}^{syn} + n_{c,m>v}^{mis}}{L_c}$$

Eq. S4

Where for Eqs. S3 and S4, $n_{c,m>v}^{syn}$ and $n_{c,m>v}^{mis}$ represent the mutation numbers with $\boldsymbol{m} > \boldsymbol{v}$ being synonymous and missense under sequence context c, respectively. The mutation rate is then scaled as the expected mutation number per 10^5 corresponding sequence motifs for better presentation.

Significance for motif enrichment (**Fig. 4B**) mirrored the AI analysis. For each $i \geq 3$ site, we calculated raw K-S p-values against motif mutabilities (denoted as p_0). These were then compared to a resampled significance pool $\{p_j, j=1,2 ..,100,000\}$, with the proportion of $p_0 \leq p_j$ employed as the final p-value, depicting enrichment significance for highly mutable motifs in recurrence group i against low hits.

Consensus length comparison

To explore potential sequence motifs associated with recurrent mutations ($i \geq 3$), we employ a sliding window of 10 bp stride to extract the local context from reference genome (**Fig. S5**). We examined diverse window sizes (21, 41, 61, 81, and 101 bp) to capture potential motifs of varying lengths and distances to the mutated site. Consensus length of local contexts was measured by Hamming distance in pairwise comparisons of aligned windows (with same stride) between mutated sites.

To prioritize sequence similarities likely driven by mutational mechanisms rather than functional constraints or gene structure, we restricted consensus comparisons to non-homologous genes. This approach effectively mitigated potential biases arising from homologous genes (e.g., *KRAS* and *NRAS*) or repeated domains within a single gene (e.g., *FBXW7*). The statistical significance of observed consensus lengths was assessed using the K-S test, which compared the empirical distribution of consensus lengths against a Poisson distribution, with mean of **λ** set to one-quarter of the window size, which reflects the expected distribution under random scenarios.

Mutational signature analysis

The mutation load of each patient could be further decomposed to several known mutational processes, which is represented by mutational signatures. In general, each mutational signature embodies the relative mutabilities across distinct mutational contexts. Leveraging single base substitution (SBS) signatures from COSMIC (v3.3), we employed the *SignatureAnalyzer* tool to quantify the contribution of each signature to individual mutational loads (Kim et al. 2016). For

composition analysis in **Fig. 5B**, we focused on signatures contributing at least 2% to the total mutations within a given cancer type, given that there are 79 mutational signatures in use for deconvolution.

To assess signature contribution changes across recurrence cutoffs (i), we grouped patients with mutations of recurrence $i \geq i^*$ and scaled signature contributions to 1 to cancel out the population size effect. Pairwise K-S test between different i^*s is employed to determine whether signature contributions are significantly different under each i^*.

Outlier model

The purpose of the outlier model is to investigate if high-hit sites could be explained by a fraction (p) of highly mutable sites (with mutability **α**-fold higher). By definition, we have:

$$S_i = (1-p) \cdot L_S \cdot [nE(u)]^i + p \cdot L_S \cdot [\alpha \cdot nE(u)]^i \quad \text{Eq. S5}$$

With n represents the population size and **E(u)** denotes the average mutation rate per site per patient.

We let p range from 10^{-5} to 10^{-2} and **α** from 1.1 to 1000. For each (p, **α**) pair, we solve Eq. S5 based on observed $\boldsymbol{S_1}$ (i = 1) to obtain **E(u)**. Then, we calculate expected $\boldsymbol{S}_i$ with i = 2, 3. The (p, **α**) pairs are filtered by imposing constraints grounded in observed $\boldsymbol{S_2}$ and $\boldsymbol{S_3}$ values. Specifically, we retained only those pairs whose expected $\boldsymbol{S_2}$ and $\boldsymbol{S_3}$ values resided within the 95% quantile range of a Poisson distribution with **λ** set to the observed values. This filtering process yielded biologically plausible (p, **α**) pairs that were then used to derive $\boldsymbol{S_4}$ and $\boldsymbol{S_5}$. Finally, we computed the mean and standard deviation for p, **α**, $\boldsymbol{S}_4$, and $\boldsymbol{S}_5$ across all filtered pairs to capture their central tendencies and variability.

CDN analysis in GENIE

To circumvent potential biases in **E(u)** estimation stemming from the varying target gene coverage across sequencing panels within the GENIE dataset, we leveraged **E(u)** values derived from the corresponding cancer types in the TCGA dataset. Specifically, we focused on 1-hit synonymous mutations within coding regions, as these are generally considered to be the least influenced by selective pressures in coding regions. Based on **Eq. 1** from the main text, we have:

$$S_1 = L_S \cdot nE(u)e^{-(n-1)E(u)} \quad \text{Eq. S6}$$

Where $e^{-(n-1)E(u)}$ comes from approximation of $[1-E(u)]^{n-1}$ from binomial distribution, and **n** is the population size in TCGA. The calculation for the threshold i^*, based on **Eq. 10** from the main text, is:

$$i^* = \frac{log\left(\frac{\varepsilon}{L_A}\right)}{log\left(\frac{n_e E(u)}{1+n_e E(u)}\right)}$$

Eq. S7 The only difference in Eq. S7 is that we use $\mathbf{n}_e$ to represent the number of patients sequenced for a target gene in GENIE, considering the overlapping between different assays in use. In essence, we will have for each gene a CDN threshold i^*.

The comparison of CDNs between TCGA and GENIE are restricted to genes sequenced by GENIE panels. For CDNs identified in GENIE using Eq. S7, we investigated their hit information and CDN identity within TCGA dataset. The CDN flow proportion depicted in **Fig. 7 C-E** represents the

ratio of sites identified as CDN in GENIE to A_i^* (i = 0, 1, 2) of TCGA. A_i^* mirrors $\boldsymbol{A_i}$ but specifically considers the coding region sequenced by the GENIE panel. For CDNs identified in TCGA, the flow ratio is just the proportion of sites being identified as CDNs in GENIE. Notably, for CDNs identified in GENIE but lacking mutations in TCGA (i = 0), A_0^* is obtained using Eq. S2 with $\boldsymbol{L}$ being the length of coding region sequenced in GENIE.

Data Availability

The key scripts used in this study are available at *https://gitlab.com/ultramicroevo/cdn_v1*. A subset of key example files for breast cancer analysis can be found in the "/example_data_files" directory. The complete list of CDNs analyzed in this study is provided in Supplement File S2.

Acknowledgements

The authors gratefully acknowledge the following for their support in the initiation of the Cancer Driving Nucleotide (CDN) project: the First Affiliated Hospital, the Seventh Affiliated Hospital of Sun Yat-sen University; Cancer Center of Clifford Hospital, Jinan University; Cancer Hospital Chinese Academy of Medical Sciences, Shenzhen Center; Guangdong Academy of Medical Sciences, Guangdong Provincial People's Hospital. We thank the Kunming Institute of Zoology for valuable discussions on the CDN concept. We are also grateful to Weiwei Zhai, Qianfei Wang, and Weini Huang for their insightful comments and suggestions. Finally, we acknowledge the American Association for Cancer Research (AACR) and The Cancer Genome Atlas (TCGA) project for providing invaluable datasets and resources that have significantly advanced our understanding of cancer biology and improved patient outcomes. This work was supported by the National Natural Science Foundation of China (32150006, 32293193/32293190 to C.I.W., 82341092 to HJ Wen, and 32200493 to Y.R.), the National Key Research and Development Projects of the Ministry of Science and Technology of China (2021YFC2301300, 2021YFC0863400), and Guangdong Key Research and Development Program (No. 2022B1111030001).

Declaration of interests

The authors declare no competing interests.

References

Alexandrov LB *et al.* (2020) **The repertoire of mutational signatures in human cancer** *Nature* **578**:94–101

Alexandrov LB *et al.* (2013) **Signatures of mutational processes in human cancer** *Nature* **500**:415–421

Anandakrishnan R, Varghese RT, Kinney NA, Garner HR (2019) **Estimating the number of genetic mutations (hits) required for carcinogenesis based on the distribution of somatic mutations** *PLOS Computational Biology* **15**

Armitage P, Doll R. (1954) **The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis** *Br J Cancer* **8**:1–12

Bailey MH *et al.* (2018) **Comprehensive Characterization of Cancer Driver Genes and Mutations** *Cell* **173**:371–385

Belikov AV (2017) **The number of key carcinogenic events can be predicted from cancer incidence** *Sci Rep* **7**

Bergstrom EN *et al.* (2022) **Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA** *Nature* **602**:510–517

Bian S, Wang Y, Zhou Y, Wang W, Guo L, Wen L, Fu W, Zhou X, Tang F. (2023) **Integrative single-cell multiomics analyses dissect molecular signatures of intratumoral heterogeneities and differentiation states of human gastric cancer** *National Science Review* **10**

Black JRM, McGranahan N. (2021) **Genetic and non-genetic clonal diversity in cancer evolution** *Nat Rev Cancer* **21**:379–392

Blough MD, Beauchamp DC, Westgate MR, Kelly JJ, Cairncross JG (2011) **Effect of aberrant p53 function on temozolomide sensitivity of glioma cell lines and brain tumor initiating cells from glioblastoma** *J Neurooncol* **102**:1–7

de Bruijn I *et al.* (2023) **Analysis and Visualization of Longitudinal Genomic and Clinical Data from the AACR Project GENIE Biopharma Collaborative in cBioPortal** *Cancer Res* **83**:3861–3867

Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, Lawrence MS (2019) **Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features** *Science* **364**

Burns MB, Temiz NA, Harris RS (2013) **Evidence for APOBEC3B mutagenesis in multiple human cancers** *Nat Genet* **45**:977–983

Campbell PJ *et al.* (2020) **Pan-cancer analysis of whole genomes** *Nature* **578**:82–93

Cannataro VL, Gaffney SG, Townsend JP (2018) **Effect Sizes of Somatic Mutations in Cancer** *JNCI: Journal of the National Cancer Institute* **110**:1171–1177

Cao Y *et al.* (2023) **Was Wuhan the early epicenter of the COVID-19 pandemic?—A critique** *National Science Review* **10**

Cerami E *et al.* (2012) **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data** *Cancer Discov* **2**:401–404

Chang MT *et al.* (2016) **Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity** *Nat Biotechnol* **34**:155–163

Chen B, Shi Z, Chen Q, Shen X, Shibata D, Wen H, Wu C-I. (2019) **Tumorigenesis as the Paradigm of Quasi-neutral Molecular Evolution** *Mol Biol Evol* **36**:1430–1441

Chen B, Wu X, Ruan Y, Zhang Y, Cai Q, Zapata L, Wu C-I, Lan P, Wen H. (2022) **Very large hidden genetic diversity in one single tumor: evidence for tumors-in-tumor** *Natl Sci Rev* **9**

Chen Q, Yang H, Feng X, Qingjian Chen, Shi S, Wu C-I, He Z. (2022) **Two decades of suspect evidence for adaptive molecular evolution—negative selection confounding positive-selection signals** *National Science Review* **9**

Danesi R *et al.* (2021) **Druggable targets meet oncogenic drivers: opportunities and limitations of target-based classification of tumors and the role of Molecular Tumor Boards** *ESMO Open* **6**

Dang CV, Reddy EP, Shokat KM, Soucek L. (2017) **Drugging the "undruggable"** *cancer targets. Nat Rev Cancer* **17**:502–508

Deng S, Xing K, He X. (2022) **Mutation signatures inform the natural host of SARS-CoV-2** *National Science Review* **9**

Elliott K, Larsson E. (2021) **Non-coding driver mutations in human cancer** *Nat Rev Cancer* **21**:500–509

Fang Y, Deng S, Li C. (2022) **A generalizable deep learning framework for inferring fine-scale germline mutation rate maps** *Nat Mach Intell* **4**:1209–1223

Gartner JJ *et al.* (2013) **Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma** *Proceedings of the National Academy of Sciences* **110**:13481–13486

Gojobori T, Li W-H, Graur D. (1982) **Patterns of nucleotide substitution in pseudogenes and functional genes** *J Mol Evol* **18**:360–369

Hanahan D. (2022) **Hallmarks of Cancer: New Dimensions** *Cancer Discovery* **12**:31–46

Hanahan D, Weinberg RA (2000) **The Hallmarks of Cancer** *Cell* **100**:57–70

Hanahan D, Weinberg RA (2011) **Hallmarks of Cancer: The Next Generation** *Cell* **144**:646–674

Haradhvala NJ *et al.* (2018) **Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair** *Nat Commun* **9**

Hartl DL, Clark AG (1989) **Principles of population genetics** Sunderland, Mass: Sinauer

He Z, Xu S, Shi S. (2020) **Adaptive convergence at the genomic level—prevalent, uncommon or very rare?** *National Science Review* **7**:947–951

He Z, Xu S, Zhang Z, Guo W, Lyu H, Zhong C, Boufford DE, Duke NC, Consortium The International Mangrove, Shi S. (2020) **Convergent adaptation of the genomes of woody plants at the land–sea interface** *National Science Review* **7**:978–993

Herzog M, Alonso-Perez E, Salguero I, Warringer J, Adams DJ, Jackson SP, Puddu F. (2021) **Mutagenic mechanisms of cancer-associated DNA polymerase IZ alleles** *Nucleic Acids Research* **49**:3919–3931

Hess JM, Bernards A, Kim J, Miller M, Taylor-Weiner A, Haradhvala NJ, Lawrence MS, Getz G. (2019) **Passenger Hotspot Mutations in Cancer** *Cancer Cell* **36**:288–301

Hodgkinson A, Eyre-Walker A. (2011) **Variation in the mutation rate across mammalian genomes** *Nat Rev Genet* **12**:756–766

Hodis E *et al.* (2022) **Stepwise-edited, human melanoma models reveal mutations' effect on tumor and microenvironment** *Science* **376**

Jackson AL, Loeb LA (1998) **The Mutation Rate and Cancer** *Genetics* **148**:1483–1490

Juul RI, Nielsen MM, Juul M, Feuerbach L, Pedersen JS (2021) **The landscape and driver potential of site-specific hotspots across cancer genomes** *npj Genom. Med* **6**:1–15

Kim J *et al.* (2016) **Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors** *Nat Genet* **48**:600–606

Lagou V *et al.* (2023) **GWAS of random glucose in 476,326 individuals provide insights into diabetes pathophysiology, complications and treatment stratification** *Nat Genet* **55**:1448–1461

Li M, Wei Chen, Sun X, Zhipeng Wang, Zou X, Wei H, Zhan Wang, Wansheng Chen (2019) **Metastatic colorectal cancer and severe hypocalcemia following irinotecan administration in a patient with X-linked agammaglobulinemia: a case report** *BMC Med Genet* **20**

Li W-H (1997) **Molecular evolution** Sunderland, Mass: Sinauer Associates

Lin J *et al.* (2023) **YTHDF2-mediated regulations bifurcate BHPF-induced programmed cell deaths** *National Science Review* **10**

Lin L, Cai J, Tan Z, Meng X, Li R, Li Y, Jiang C. (2021) **Mutant IDH1 Enhances Temozolomide Sensitivity via Regulation of the ATM/CHK2 Pathway in Glioma** *Cancer Res Treat* **53**:367–377

Ling S *et al.* (2015) **Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution** *Proc Natl Acad Sci USA* **112**:E6496–E6505

Luo P, Ding Y, Lei X, Wu F-X. (2019) **deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks** *Frontiers in Genetics [Internet]* **10**

Makova KD, Hardison RC (2015) **The effects of chromatin organization on variation in mutation rates in the genome** *Nat Rev Genet* **16**:213–223

Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ (2017) **Universal Patterns of Selection in Cancer and Somatic Tissues** *Cell* **171**:1029–1041

Nesta AV, Tafur D, Beck CR (2021) **Hotspots of Human Mutation** *Trends in Genetics* **37**:717–729

Noeuveglise A *et al.* (2023) **Impact of EGFRA289T/V mutation on relapse pattern in glioblastoma** *ESMO Open* **8**

Ortmann CA *et al.* (2015) **Effect of Mutation Order on Myeloproliferative Neoplasms** *N Engl J Med* **372**:601–612

Pan Y, Liu P, Wang F, Wu P, Cheng F, Jin X, Xu S. (2022) **Lineage-specific positive selection on ACE2 contributes to the genetic susceptibility of COVID-19** *National Science Review* **9**

Pan Y *et al.* (2022) **Genomic diversity and post-admixture adaptation in the Uyghurs** *National Science Review* **9**

Pleasance ED *et al.* (2010) **A comprehensive catalogue of somatic mutations from a human cancer genome** *Nature* **463**:191–196

Polak P *et al.* (2015) **Cell-of-origin chromatin organization shapes the mutational landscape of cancer** *Nature* **518**:360–364

Roberts SA *et al.* (2013) **An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers** *Nat Genet* **45**:970–976

Ruan Y, Hou M, Tang X, He X, Lu X, Lu J, Wu C-I, Wen H. (2022) **The Runaway Evolution of SARS-CoV-2 Leading to the Highly Evolved Delta Strain** *Molecular Biology and Evolution* **39**

Ruan Y, Wang H, Chen B, Wen H, Wu C-I. (2020) **Mutations Beget More Mutations—Rapid Evolution of Mutation Rate in Response to the Risk of Runaway Accumulation** *Mol Biol Evol* **37**:1007–1019

Ruan Y, Wen H, Hou M, He Z, Lu X, Xue Y, He X, Zhang Y-P, Wu C-I. (2022) **The twin-beginnings of COVID-19 in Asia and Europe—one prevails quickly** *National Science Review* **9**

Ruan Y, Wen H, Hou M, Zhai W, Xu S, Lu X. (2023) **On the epicenter of COVID-19 and the origin of the pandemic strain** *National Science Review* **10**

Ségurel L, Wyman MJ, Przeworski M. (2014) **Determinants of Mutation Rate Variation in the Human Germline** *Annual Review of Genomics and Human Genetics* **15**:47–70

Sherman MA, Yaari AU, Priebe O, Dietlein F, Loh P-R, Berger B. (2022) **Genome-wide mapping of somatic mutation rates uncovers drivers of cancer** *Nat Biotechnol* **1**

Song Q *et al.* (2023) **DeepAlloDriver: a deep learning-based strategy to predict cancer driver mutations** *Nucleic Acids Research* **51**:W129–W133

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR (2009) **Human mutation rate associated with DNA replication timing** *Nat Genet* **41**:393–395

Stobbe MD, Thun GA, Diéguez-Docampo A, Oliva M, Whalley JP, Raineri E, Gut IG (2019) **Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer** *PLOS Computational Biology* **15**

Supek F, Lehner B. (2015) **Differential DNA mismatch repair underlies mutation rate variation across the human genome** *Nature* **521**:81–84

Suzuki K *et al.* (2024) **Genetic drivers of heterogeneity in type 2 diabetes pathophysiology** *Nature* **627**:347–357

Takeda H. (2021) **A Platform for Validating Colorectal Cancer Driver Genes Using Mouse Organoids** *Front Genet* **12**

Tate JG *et al.* (2019) **COSMIC: the Catalogue Of Somatic Mutations In Cancer** *Nucleic Acids Research* **47**:D941–D947

Turajlic S, Sottoriva A, Graham T, Swanton C. (2019) **Resolving genetic heterogeneity in cancer** *Nat Rev Genet* **20**:404–416

Vujkovic M *et al.* (2020) **Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis** *Nat Genet* **52**:680–691

Waarts MR, Stonestrom AJ, Park YC, Levine RL (2022) **Targeting mutations in cancer** *J Clin Invest* **132**

Wang Q, Fang W-H, Krupinski J, Kumar S, Slevin M, Kumar P. (2008) **Pax genes in embryogenesis and oncogenesis** *Journal of Cellular and Molecular Medicine* **12**:2281–2294

Wang X *et al.* (2022) **Extensive gene flow in secondary sympatry after allopatric speciation** *National Science Review* **9**

Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM (2013) **The Cancer Genome Atlas Pan-Cancer analysis project** *Nat Genet* **45**:1113–1120

Wong JKL, Aichmüller C, Schulze M, Hlevnjak M, Elgaafary S, Lichter P, Zapatka M. (2022) **Association of mutation signature effectuating processes with mutation hotspots in driver genes and non-coding regions** *Nat Commun* **13**

Wu C-I (2022) **What are species and how are they formed?** *National Science Review* **9**

Wu C-I. (2023) **The genetics of race differentiation—should it be studied?** *National Science Review* **10**

Wu C-I, Maeda N. (1987) **Inequality in mutation rates of the two strands of DNA** *Nature* **327**:169–170

Wu C-I, Ting C-T. (2004) **Genes and speciation** *Nat Rev Genet* **5**:114–122

Wu C-I, Wang G-D, Xu S. (2020) **Convergent adaptive evolution—how common, or how rare?** *National Science Review* **7**:945–946

Xue D *et al.* (2023) **Functional interrogation of twenty type 2 diabetes-associated genes using isogenic human embryonic stem cell-derived β-like cells** *Cell Metabolism* **35**:1897–1914

Zeng Z, Bromberg Y. (2022) **Inferring Potential Cancer Driving Synonymous Variants** *Genes* **13**

Zhai W *et al.* (2022) **Dynamic phenotypic heterogeneity and the evolution of multiple RNA subtypes in hepatocellular carcinoma: the PLANET study** *National Science Review* **9**

Zhang L *et al.* (2024) **On the discovered Cancer Driving Nucleotides (CDNs) –Distributions across genes, cancer types and patients** *eLife* **13**

Zhao W *et al.* (2021) **CanDriS: posterior profiling of cancer-driving sites based on two-component evolutionary model** *Briefings in Bioinformatics* **22**

Zheng CL *et al.* (2014) **Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes** *Cell Reports* **9**:1228–1234

Zhou Y, Litfin T, Zhan J. (2023) **3 = 1 + 2: how the divide conquered de novo protein structure prediction and what is next?** *National Science Review* **10**

Zhu H *et al.* (2023) **Proteomics of adjacent-to-tumor samples uncovers clinically relevant biological events in hepatocellular carcinoma** *National Science Review* **10**

Editors

Reviewing Editor
Detlef Weigel
Max Planck Institute for Biology Tübingen, Tübingen, Germany

Senior Editor
Detlef Weigel
Max Planck Institute for Biology Tübingen, Tübingen, Germany

Reviewer #1 (Public review):

The authors developed a rigorous methodology for identifying all Cancer Driving Nucleotides (CDNs) by leveraging the concept of massively repeated evolution in cancer. By focusing on mutations that recur frequently in pan-cancer, they aimed to differentiate between true driver mutations and neutral mutations, ultimately enhancing the understanding of the mutational landscape that drives tumorigenesis. Their goal was to call a comprehensive catalogue of CDNs to inform more effective targeted therapies and address issues such as drug resistance.

Strengths

(1) The authors introduced a concept of using massively repeated evolution to identify CDNs. This approach recognizes that advantageous mutations recur frequently (at least 3 times) across cancer patients, providing a lens to identify true cancer drivers.

(2) The theory showed the feasibility of identifying almost all CDNs if the number of sequenced patients increases to 100,000 for each cancer type.

Weaknesses

(1) No novel true driver mutations were identified in this study.

(2) Different cancer types have unique mutational landscapes. The methodology, while robust, might face challenges in uniformly identifying CDNs across various cancers with distinct genetic and epigenetic contexts.

(3) The statement "In other words, the sequences surrounding the high-recurrence sites appear rather random.". Since it was a pan-cancer analysis, the unique patterns of each cancer type could be strongly diluted in the pan-cancer data.

https://doi.org/10.7554/eLife.99340.2.sa2

Reviewer #2 (Public review):

Summary:

The authors propose that cancer driver mutations can be identified by Cancer Driving Nucleotides (CDNs). CDNs are defined as SNVs that occur frequently in genes. There are many ways to define cancer driver mutations, and strengths and weaknesses are the reliance of statistics to define them.

Strengths:

There are many well-known approaches and studies that have already identified many canonical driver mutations. A potential strength is that mutation frequencies may be able to identify, as yet, unrecognized driver mutations. They use of a previously developed method to estimate mutation hotspots across the genome (Dig, Sherman et al 2022). This publication has already used cancer sequence data to infer driver mutations based on higher than expected mutation frequencies. The advance here is to further illustrate that recurrent mutations (estimated at 3 or more mutations (CDNs) at the same base) are more likely to be the result of selection for a driver mutation (Fig 3). Further analysis indicates that mutation sequence context (Fig 4) or mutation mechanisms (Fig 5) are unlikely to be major causes for recurrent point mutations. Finally, they calculate (Fig 6) that most driver mutations identifiable by the CDN approach could be identified with about 100,000 to one million tumor coding genomes.

Weaknesses:

The manuscript does provide specific examples where recurrent mutations identify known driver mutations, but does not identify "new" candidate driver mutations. Driver mutation validation is difficult and at least clinically, frequency (ie observed in multiple other cancer samples) is indeed commonly used to judge if a SNV has driver potential. The method would miss alternative ways to trigger driver alterations (translocations, indels, epigenetic, CNVs). Nevertheless, the value of the manuscript is its quantitative analysis of why mutation frequencies can identify cancer driver mutations.

https://doi.org/10.7554/eLife.99340.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***eLife assessment*** *This valuable paper reports a theoretical framework and methodology for identifying Cancer Driving Nucleotides (CDNs), primarily based on single nucleotide*

> *variant (SNV) frequencies. A variety of solid approaches indicate that a mutation recurring three or more times is more likely to reflect selection rather than being the consequence of a mutation hotspot. The method is rigorously quantitative, though the requirement for larger datasets to fully identify all CDNs remains a noted limitation. The work will be of broad interest to cancer geneticists and evolutionary biologists.*

The key criticism "the requirement for larger datasets to fully identify all CDNs remains a noted limitation" that is also found in both reviews. We have clarified the issue in the main text, the relevant parts, from which are copied below. The response below also addresses many comments in the reviews. In addition, Discussion of eLife-RP-RA-2024-99341 has been substantially expanded to answer the questions of Reviewer 2.

We shall answer the boldface comment in three ways. First, it can be answered using GENIE data. Fig. 7 of the main text (eLife-RP-RA-2024-99340) shows that, when n increases from ~ 1000 to ~ 9,000, the numbers of discovered CDNs increase by 3 – 5 fold, most of which come from the two-hit class. Hence, the power of discovering more CDNs with larger datasets is evident. By extrapolation, a sample size of 100,000 should be able to yield 90% of all CDNs, as calculated here. (Fig. 7 also addresses the queries of whether we have used datasets other than TCGA. We indeed have used all public data, including GENIE and COSMIC.)

Second, the power of discovering more cancer driver genes by our theory is evident even without using larger datasets. Table 3 of the companion study (eLife-RP-RA-2024-99341) shows that, averaged across cancer types, the conventional method would identify 45 CDGs while the CDN method tallies 258 CDGs. The power of the CDN method is demonstrated. This is because the conventional approach has to identify CDGs (cancer driver genes) in order to identify the CDNs they carry. However, many CDNs occur in non-CDGs and are thus missed by the conventional approach. In Supplementary File S2, we have included a full list of CDNs discovered in our study, along with population allele frequency annotations from *gnomAD*. The distribution patterns of these CDNs across different cancer types show their pan-cancer properties as further explored in the companion paper.

Third, while many, or even most CDNs occur in non-CDGs and are thus missed, the conventional approach also includes non-CDN mutations in CDGs. This is illustrated in Fig. 5 of the companion study (eLife-RP-RA-2024-99341) that shows the adverse effect of misidentifications of CDNs by the conventional approach. In that analysis, the gene-targeting therapy is effective if the patient has the CDN mutations on *EGFR*, but the effect is reversed if the *EGFR* mutations are non-CDN mutations.

> ***Reviewer #1 (Public Review):***
>
> *The authors developed a rigorous methodology for identifying all Cancer Driving Nucleotides (CDNs) by leveraging the concept of massively repeated evolution in cancer. By focusing on mutations that recur frequently in pan-cancer, they aimed to differentiate between true driver mutations and neutral mutations, ultimately enhancing the understanding of the mutational landscape that drives tumorigenesis. Their goal was to call a comprehensive catalogue of CDNs to inform more effective targeted therapies and address issues such as drug resistance.*
>
> *Strengths*
>
> *(1) The authors introduced a concept of using massively repeated evolution to identify CDNs. This approach recognizes that advantageous mutations recur frequently (at least 3 times) across cancer patients, providing a lens to identify true cancer drivers.*
>
> *(2) The theory showed the feasibility of identifying almost all CDNs if the number of sequenced patients increases to 100,000 for each cancer type.*

> *Weaknesses*
>
> *(1) The methodology remains theoretical and no novel true driver mutations were identified in this study.*

We now address the weakness criticism, which is gratefully received.

The second part of the criticism (no novel true driver mutations were identified in this study) has been answered in the long responses to eLife assessment above. The first part "*The methodology remains theoretical*" is somewhat unclear. It might be the lead to the second part. However, just in case, we interpret the word "theoretical" to mean "the lack of experimental proof" and answer below.

As Reviewer #1 noted, a common limitation of theoretical and statistical analyses of cancer drivers is the need to validate their selective advantage through *in vitro* or *in vivo* functional testing. This concern is echoed by both reviewers in the companion paper (eLife-RP-RA-2024-99341), prompting us to consider the methodology for functional testing of potential cancer drivers. An intuitive approach would involve introducing putative driver mutations into normal cells and observing phenotypic transformation in vitro and in vivo. In a recent stepwise-edited human melanoma model, Hodis et al. demonstrated that disease-relevant phenotypes depend on the "correct" combinations of multiple driver mutations (Hodis et al. 2022). Other high-throughput strategies can be broadly categorized into two approaches: (1) introducing candidate driver mutations into pre-malignant model systems that already harbor a canonical mutant driver (Drost and Clevers 2018; Grzeskowiak et al. 2018; Michels et al. 2020) and (2) introducing candidate driver mutations into growth factor-dependent cell models and assessing their impact on resulting fitness (Bailey et al. 2018; Ng et al. 2018). The underlying assumption of these strategies is that the fitness outcomes of candidate driver mutations are influenced by pre-existing driver mutations and the specific pathways or cancer hallmarks being investigated. This confines the functional test of potential cancer driver mutations to conventional cancer pathways. A comprehensive identification of CDNs is therefore crucial to overcome these limitations. In conjunction with other driver signal detection methods, our study aims to provide a more comprehensive profile of driver mutations, thereby enabling the functional testing of drivers involved in non-conventional cancer evolution pathways.

> *(2) Different cancer types have unique mutational landscapes. The methodology, while robust, might face challenges in uniformly identifying CDNs across various cancers with distinct genetic and epigenetic contexts.*

We appreciate the comment. Indeed, different cancer types should have different genetic and epigenetic landscapes. In that case, one may have expected CDNs to be poorly shared among cancer types. However, as reported in Fig. 4 of the companion study, the sharing of CDNs across cancer types is far more common than the sharing of CDGs (Cancer Driving Genes). We suggest that CDNs have a much higher resolution than CDGs, whereby the signals are diluted by non-driver mutations. In other words, despite that the mutational landscape may be cancer-type specific, the pan-cancer selective pressure may be sufficiently high to permit the detection of CDN sharing among cancer types.

Below, we shall respond in greater details. Epigenetic factors, such as chromatin states, methylation/acetylation levels, and replication timing, can provide valuable insights when analyzing mutational landscapes at a regional scale (Stamatoyannopoulos et al. 2009; Lawrence et al. 2013; Makova and Hardison 2015; Baylin and Jones 2016; Alexandrov et al. 2020; Abascal et al. 2021; Sherman et al. 2022). However, at the site-specific level, the effectiveness of these covariates in predicting mutational landscapes depends on their integration into a detailed model. Overemphasizing these covariates could lead to false

negatives for known driver mutations (Hess et al. 2019; Elliott and Larsson 2021). In figure 3B of the main text, we illustrate the discrepancy between the mutation rate predictions from *Dig* and empirical observation. Ideally, no covariates would be needed under extensive sample sizes, where each mutable genomic sites would have sufficient mutations to yield a statistic significance and consequently, synonymous mutations would be sufficient for the characterization of mutational landscape. In this sense, the integration of mutational covariates represents a compromise under current sample size. In our study, the effect of unique mutational landscapes is captured by E(u), the mean mutation rate for each cancer type. We further accounted for the variability of site-level mutability using a gamma distribution. The primary goal of our study is to determine the upper limit of mutation recurrences under mutational mechanisms only. While selection force acts blindly to genomic features, mutational hotspots should exhibit common characteristics determined by their underlying mechanisms. In the main text, we attempted to identify such shared features among CDNs. Until these mutational mechanisms are fully understood, CDNs should be considered as potential driver mutations.

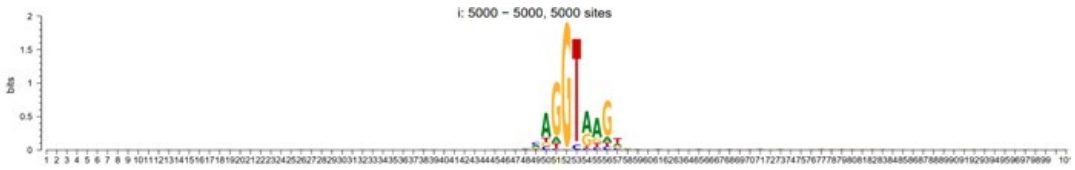
> *(3) L223, the statement "In other words, the sequences surrounding the high-recurrence sites appear rather random.". Since it was a pan-cancer analysis, the unique patterns of each cancer type could be strongly diluted in the pan-cancer data.*

We now state that the analyses of mutation characteristic have been applied to the individual cancer types and did not find any pattern that deviates from randomness. Nevertheless, it may be argued that, with the exception of those with sufficiently large sample sizes such as lung and breast cancers, most datasets do not have the power to reject the null hypothesis. To alleviate this concern, we applied the ResNet and LSTM/GRU methods for the discovery of potential mutation motifs within each cancer type. All methods are more powerful than the one used but the results are the same – no cancer type yields a mutation pattern that can reject the null hypothesis of randomness (see below).

As a positive control, we used these methods for the discovery of splicing sites of human exons. When aligned up with splicing site situated in the center (position 51 in the following plot), the sequence motif would look like:

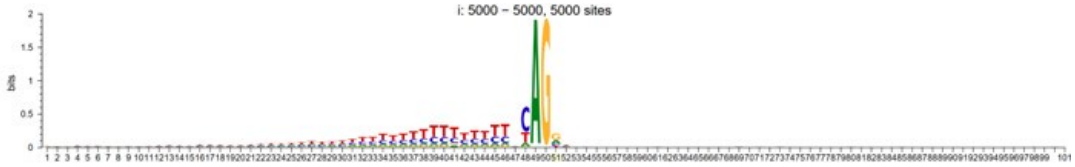
Author response image 1.

5-prime



Author response image 2.

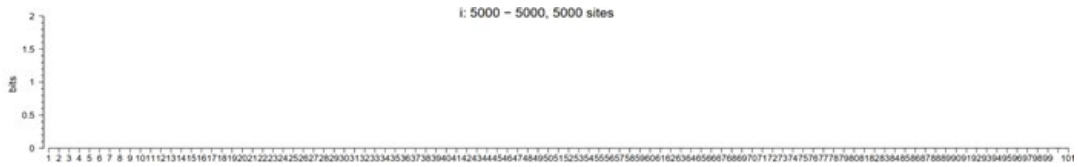
3-prime



However, To account for the potential influence of distance from the mutant site in motif analysis, we randomly shuffled the splicing sites within a specified window around the alignment center, and their sequence logo now looks like:

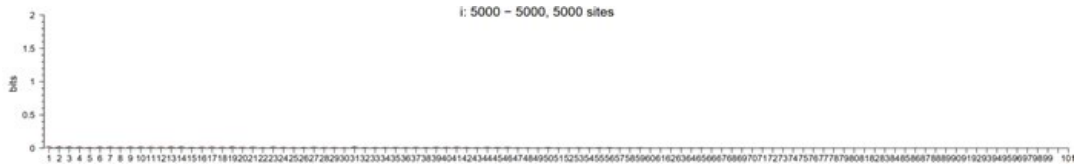
Author response image 3.

5-prime shuffled



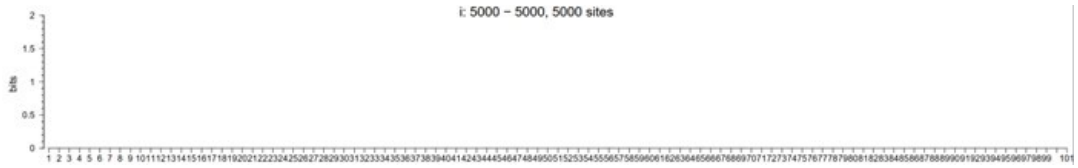
Author response image 4.

3-prime shuffled



Author response image 5.

random sequences from coding regions



The classification results of the shuffled 5-prime (donner), 3-prime (acceptor) and random sequences from coding regions (Random CDS) are presented in the Author response table 1 (The accuracy for the aligned results, which is approximately 99%, is not shown here).

Author response table 1.

			TRAIN accuracy				TEST accuracy			
Model	# layers	# parameters	ALL	donner	acceptor	Random CDS	ALL	donner	acceptor	random CDS
resNet1	46	3,721	0.77	0.77	0.76	0.79	0.71	0.72	0.69	0.72
resNet2	165	64,445	0.96	0.96	0.95	0.95	0.76	0.77	0.77	0.73
deepGRU	23	144,925	0.86	0.85	0.84	0.89	0.79	0.79	0.76	0.8
deepLSTM	6	24,445	0.85	0.82	0.87	0.86	0.79	0.76	0.82	0.78

With the positive results from these positive controls (splicing site motifs) validating our methodology, we applied the same model structure to the train and test of potential

mutational motifs of CDN sites. All models achieved approximately 50% accuracy in CDN motif analysis, suggesting that the sequence contexts surrounding CDN sites are not significantly different from other coding regions of the genome. This further implies that the recurrence of mutations at CDN sites is more likely driven by selection rather than mutational mechanisms.

Note that this preliminary analysis may be limited by insufficient training data for CDN sites. Future studies will require larger sample sizes and more sophisticated models to address these limitations.

> *(4) To solidify the findings, the results need to be replicated in an independent dataset.*

Figure 7 validates our CDN findings using the GENIE dataset, which primarily consists of targeted sequencing data from various panels. By focusing on the same genomic regions sequenced by GENIE, we observed a 3-5 fold increase in the number of discovered CDNs as sample size increased from approximately 1000 to 9000. Moreover, the majority of CDNs identified in TCGA were confirmed as CDNs in GENIE.

> *(5) The key scripts and the list of key results (i.e., CDN sites with i{greater than or equal to}3) need to be shared to enable replication, validation, and further research. So far, only CDN sites with i{greater than or equal to}20 have been shared.*

We have now updated the “Data Availability” section in the main text, the corresponding scripts for key results are available on Gitlab at: https://gitlab.com/ultramicroevo/cdn_v1.

> *(6) The versions of data used in this study are not clearly detailed, such as the specific version of gnomAD and the version and date of TCGA data downloaded from the GDC Data Portal.*

The versions of data sources have now been updated in the revised manuscript.

> ***Recommendations For The Authors:***
>
> *(1) L119, states "22.7 million nonsynonymous sites," but Table 1 lists the number as 22,540,623 (22.5 million). This discrepancy needs to be addressed for consistency.*
> *(2) Figure 2B, there is an unexplained drop in the line at i = 6 and 7 (from 83 to 45). Clarification is needed on why this drop occurs.*
> *(3) Figure 3A, for the CNS type, data for recurrence at 8 and 9 are missing. An explanation should be provided for this absence.*
> *(4) L201, the title refers to "100-mers," but L218 mentions "101-mers." This inconsistency needs to be corrected to ensure clarity and accuracy.*
> *(5) Figures 6 and 7 currently lack titles. Titles should be added to these figures to improve readability.*

Thanks. All corrections have been incorporated into the revised manuscript.

> ***Reviewer #2 (Public Review):***
> *Summary:*
> *The authors propose that cancer-driver mutations can be identified by Cancer Driving Nucleotides (CDNs). CDNs are defined as SNVs that occur frequently in genes. There are many ways to define cancer driver mutations, and the strengths and weaknesses are the reliance on statistics to define them.*
> *Strengths:*
> *There are many well-known approaches and studies that have already identified many canonical driver mutations. A potential strength is that mutation frequencies may be*

able to identify as yet unrecognized driver mutations. They use a previously developed method to estimate mutation hotspots across the genome (Dig, Sherman et al 2022). This publication has already used cancer sequence data to infer driver mutations based on higher-than-expected mutation frequencies. The advance here is to further illustrate that recurrent mutations (estimated at 3 or more mutations (CDNs) at the same base) are more likely to be the result of selection for a driver mutation (Figure 3). Further analysis indicates that mutation sequence context (Figure 4) or mutation mechanisms (Figure 5) are unlikely to be major causes for recurrent point mutations. Finally, they calculate (Figure 6) that most driver mutations identifiable by the CDN approach could be identified with about 100,000 to one million tumor coding genomes.
Weaknesses:
The manuscript does provide specific examples where recurrent mutations identify known driver mutations but do not identify "new" candidate driver mutations. Driver mutation validation is difficult and at least clinically, frequency (ie observed in multiple other cancer samples) is indeed commonly used to judge if an SNV has driver potential. The method would miss alternative ways to trigger driver alterations (translocations, indels, epigenetic, CNVs). Nevertheless, the value of the manuscript is its quantitative analysis of why mutation frequencies can identify cancer driver mutations.

Recommendations For The Authors
Whereas the analysis of driver mutations in WES has been extensive, the application of the method to WGS data (ie the noncoding regions) would provide new information.

We appreciate that Reviewer #2 has suggested the potential application of our method to noncoding regions. Currently, the background mutation model is based on the site level mutations in coding regions, which hinders its direct applications in other mutation types such as CNVs, translocations and indels. We acknowledge that the proportion of patients with driver event involving CNV (73%) is comparable to that of coding point mutations (76%) as reported in the PCAWG analysis (Fig. 2A from Campbell et al., 2020). In future studies, we will attempt to establish a CNV-based background mutation rate model to identify positive selection signals driving tumorigenesis.

References

Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing SV, Ellis P, Russell AJC, Alcantara RE, Baez-Ortega A, Wang Y, et al. 2021. Somatic mutation landscapes at single-molecule resolution. *Nature*:1–6.

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, et al. 2020. The repertoire of mutational signatures in human cancer. *Nature* 578:94–101.

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* 173:371-385.e18.

Baylin SB, Jones PA. 2016. Epigenetic Determinants of Cancer. *Cold Spring Harb Perspect Biol* 8:a019505.

Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, Perry MD, Nahal-Bose HK, Ouellette BFF, Li CH, et al. 2020. Pan-cancer analysis of whole genomes. *Nature* 578:82–93.

Drost J, Clevers H. 2018. Organoids in cancer research. *Nat Rev Cancer* 18:407–418.

Elliott K, Larsson E. 2021. Non-coding driver mutations in human cancer. *Nat Rev Cancer* 21:500–509.

Grzeskowiak CL, Kundu ST, Mo X, Ivanov AA, Zagorodna O, Lu H, Chapple RH, Tsang YH, Moreno D, Mosqueda M, et al. 2018. In vivo screening identifies GATAD2B as a metastasis driver in KRAS-driven lung cancer. *Nat Commun* 9:2732.

Hess JM, Bernards A, Kim J, Miller M, Taylor-Weiner A, Haradhvala NJ, Lawrence MS, Getz G. 2019. Passenger Hotspot Mutations in Cancer. *Cancer Cell* 36:288-301.e14.

Hodis E, Triglia ET, Kwon JYH, Biancalani T, Zakka LR, Parkar S, Hütter J-C, Buffoni L, Delorey TM, Phillips D, et al. 2022. Stepwise-edited, human melanoma models reveal mutations' effect on tumor and microenvironment. *Science* 376:eabi8175.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499:214–218.

Makova KD, Hardison RC. 2015. The effects of chromatin organization on variation in mutation rates in the genome. *Nat Rev Genet* 16:213–223.

Michels BE, Mosa MH, Streibl BI, Zhan T, Menche C, Abou-El-Ardat K, Darvishi T, Członka E, Wagner S, Winter J, et al. 2020. Pooled In Vitro and In Vivo CRISPR-Cas9 Screening Identifies Tumor Suppressors in Human Colon Organoids. *Cell Stem Cell* 26:782-792.e7.

Ng PK-S, Li J, Jeong KJ, Shao S, Chen H, Tsang YH, Sengupta S, Wang Z, Bhavana VH, Tran R, et al. 2018. Systematic Functional Annotation of Somatic Mutations in Cancer. *Cancer Cell* 33:450-462.e10.

Sherman MA, Yaari AU, Priebe O, Dietlein F, Loh P-R, Berger B. 2022. Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. *Nat Biotechnol* 40:1634–1643.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* 41:393–395.

https://doi.org/10.7554/eLife.99340.2.sa0