Shan et al. *Cell Discovery* (2020)6:48
https://doi.org/10.1038/s41421-020-0177-8

Cell Discovery
www.nature.com/celldisc

ARTICLE Open Access

Defining relative mutational difficulty to understand cancer formation

Lin Shan[1,2], Jiao Yu[1,2], Zhengjin He[1,2], Shishuang Chen[1,2], Mingxian Liu[1,2], Hongyu Ding[1,2], Liang Xu[1,2], Jie Zhao[1,2], Ailing Yang[1,2] and Hai Jiang[1,2]

Abstract

Most mutations in human cancer are low-frequency missense mutations, whose functional status remains hard to predict. Here, we show that depending on the type of nucleotide change and the surrounding sequences, the tendency to generate each type of nucleotide mutations varies greatly, even by several hundred folds. Therefore, a cancer-promoting mutation may appear only in a small number of cancer cases, if the underlying nucleotide change is too difficult to generate. We propose a method that integrates both the original mutation counts and their relative mutational difficulty. Using this method, we can accurately predict the functionality of hundreds of low-frequency missense mutations in p53, PTEN, and INK4A. Many loss-of-function p53 mutations with dominant negative effects were identified, and the functional importance of several regions in p53 structure were highlighted by this analysis. Our study not only established relative mutational difficulties for different types of mutations in human cancer, but also showed that by incorporating such a parameter, we can bring new angles to understanding cancer formation.

Introduction

Gene mutation is a major cause of tumorigenesis. Certain mutations on important cancer genes such as KRAS and p53 drive cancer formation[1,2]. As a result, such mutations are enriched in cancer, and are found in numerous cancer samples. It is generally perceived that if a mutation occurs in higher number of cancer cases, it is more likely to be a driver mutation[3,4]. However, most mutations in cancer only occurs in very small number of cancer cases, and the functional impacts of these mutations are hard to predict.

To address this problem, it is necessary to consider that the chance of observing a mutation in cancer cases is influenced by at least two major aspects: (1) how difficult it is to generate the mutation; and (2) whether the mutation promotes cancer, therefore it will be selectively enriched in cancer cases. If different mutations are initially generated at significantly different rates, it will greatly impact the mutational distribution in cancer genome database such as Catalog of Somatic Mutations in Cancer (COSMIC). Certain cancer-driving, but too-hard-to generate mutations may appear exceedingly rare in cancer database, yet certain passenger-type mutations may pile up in greater numbers, if the underlying mutations are too easy to occur.

At nucleotide level, there are 12 routes of interchanges between A/G/T/C for single nucleotide substitutions, which underly most cancer mutations. The chances of generating each kind of mutations are certainly not equal. Many factors contribute to such phenomenon. First, different endogenous and exogenous mutagenic events lead to different types of nucleotide substitutions[5–9]. Second, the abilities to recognize, repair, and tolerate different types of mutations are also different[10,11]. Third, although difficult to predict, different nucleotide sequences surrounding the mutation site may cause local variances, which may physically or chemically affect the chance of mutagenesis. In addition, certain sequences are also more

Correspondence: Hai Jiang (hai@sibcb.ac.cn)
[1]State Key Laboratory of Cell Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
These authors contributed equally: Lin Shan, Jiao Yu

prone to be edited by enzymes such as APOBEC[12,13]. Therefore, different flanking nucleotide sequences can also affect mutation rate[7,14,15].

Taken together, the probability to generate different types of nucleotide change may vary greatly. If two mutations both change the functional status of an important gene and promote cancer, they should be found in multiple cancer samples. However, if one of such mutation is too difficult to generate at nucleotide level, the number of cancer cases carrying that mutation will decrease significantly. Considering this, if we can define the relative difficulty to generate each type of nucleotide mutations in cancer, we will be able to better estimate the functional importance of cancer mutations.

Although mutational signatures for ageing, UV, APOBEC, smoking, and other cancer causes have been established[16], it is difficult to predict what percentage of cancers are influenced by each signature, and to what extent. Moreover, some of the major contributors to cancer, including nitrosamine, have not been assigned a Sanger mutation signature. Therefore, the relative difficulty to generate different types of mutation in cancer has not been adequately established. In this report, through analysis of mutational data from 26,000 cancer genomes, we established the relative mutational difficulty for different types of cancer mutations and showed that it can help accurately interpret functional importance of cancer mutations.

Results

Defining relative mutational difficulties in human cancer

Given the complexity of mutagenesis in cancer, it is very difficult to construct a mathematical model that could weight in all relevant factors to forwardly predict how much more difficult it is to generate one type of mutation versus the other. However, such differences do factually exist, and they collectively determined the mutation distributions in human cancer. Based on this notion, we argue that by analyzing large human cancer genome dataset, we can reversely derive the relative difficulties for each type of mutation (Supplementary Fig. S1).

We retrieved mutation dataset for all human coding genes from the COSMIC database. From the approximately 26,000 cancer samples (Supplementary Fig. S2) that were subjected to exome or whole-genome sequencing, more than 3 million single nucleotide mutations were identified on protein coding sequences (Fig. 1a). Considering that some mutations such as KRAS G12D and BRAF V600E are selectively enriched during cancer development, which could skew our estimation of mutation tendency, we excluded mutational events that occur in more than five cancer samples (see "Methods" for further discussion). This eliminated about 2% of mutations (Fig. 1a) and the remaining mutations were collated into different groups.

Overall, the number of C→T mutations and its complementary G→A mutations constitute more than half of mutations in cancer (Fig. 1b). The rate of C→T mutation is 14-folds more than T→G mutation, demonstrating that the chances to generate each type of mutations do vary significantly (Fig. 1b).

Importantly, to reach a systematic view of how neighboring sequences might affect mutational tendency, we performed an extensive analysis, in which nucleotides at -2, -1, $+1$, and $+2$ position were all taken into consideration. Consequently, mutations were divided into 3072 groups (Supplementary Table S1).

For example, the most likely to occur cancer mutation is C→T mutation on TTCGT sequences, which appeared 10,563 times. There are approximately 575 million TTCGT sequences in 26,000 coding genomes. Therefore, the chance of a C→T cancer mutation on TTCGT sequences can be calculated as 1.85×10^{-5} (=10,563/575,000,000), which is about 200-folds more than the probability of A→C mutation on an ACATC sequence (Fig. 1c). In other words, it is 200 times more "difficult" to generate the latter mutation in human cancer. Similarly, such "difficulty" indexes were generated for all 3072 types of nucleotide substitutions, which showed a wide distribution (Fig. 1c, Supplementary Fig. S3, Table S1). Analysis of these difficulty indexes showed that in addition to nucleotides on -1 and $+1$ positions (Fig. 1d), the nucleotides on $+2$ and -2 positions can also exert significant impacts on mutational tendency (Fig. 1e, Supplementary Fig. S4). This indicates that it is important to incorporate the flanking nucleotide sequences into analysis when assessing individual mutations.

Our analysis shows that different types of mutations are generated at remarkably different rates (Fig. 1c). Given that the chance to generate different types of mutations can vary by several 100-folds, it strongly suggests the need to reassess human cancer mutations and our dataset (Supplementary Table S2) will provide a useful tool.

To more precisely evaluate individual cancer mutation, we also took into consideration that certain types of human malignancies such as melanoma, endometrial and colorectal cancers exhibit significantly higher mutation rates than other types of human cancer[17]. Therefore, the same type of mutation may be generated at significantly different rates in different types of cancer. Considering this, we generated cancer type-specific mutational difficulty indexes with similar method (Fig. 1f, Supplementary Fig. S5, Table S2). which will enable precise assessment of cancer mutations.

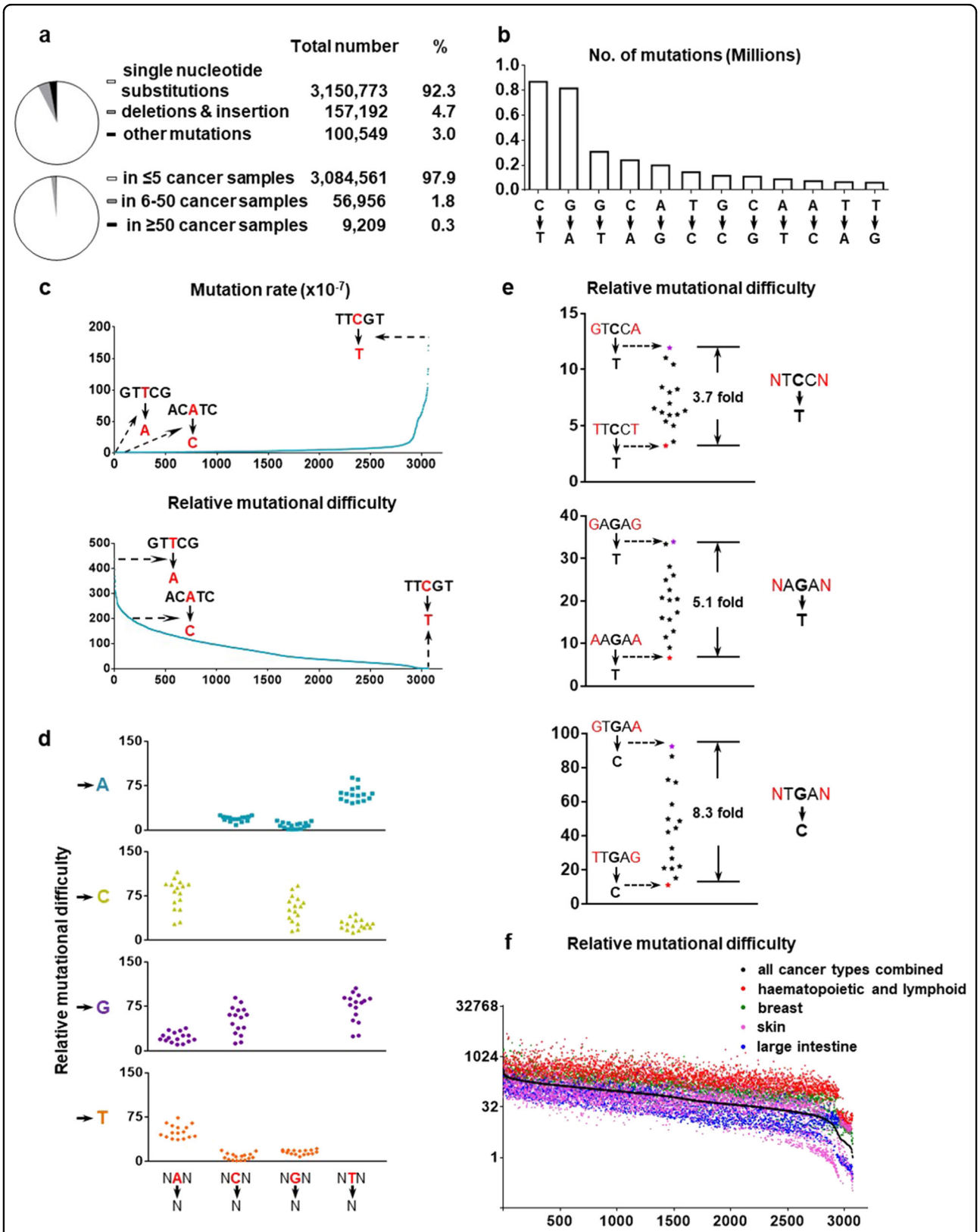

Fig. 1 Relative mutational difficulty in human cancer. a, **b** Overview and classification of coding mutations from about 26,154 cancer genomes. **c** Rates and relative difficulties of different types of mutations based on 26,154 cancer genomes. Depending on the type of nucleotide substitution and the surrounding sequences, mutations are divided into 3072 groups. **d**, **e** The impact of flanking nucleotides on relative mutational difficulty. **f** Cancer type-specific relative mutational difficulty.

Incorporating mutational difficulty to predict loss of function p53 mutations

We hypothesize that these "difficulty" indexes can serve as a valuable tool to predict the functional importance of cancer mutations. For example, if an A→C mutation on an ACATC sequence, despite the high difficulty, is still strongly selected for and appears in noticeable number of cancer samples, it could indicate that such a mutation is significantly enriched during cancer development. Therefore, such mutations may be crucial for cancer development.

We applied this method to assess the functional impact of p53 missense mutations. Several well-established p53 hotspot mutations account for about 27% of all p53 missense mutations and are known to abolish gene function. Most of the less frequent p53 missense mutations, although constituting the majority, are hard to predict in terms of their functional impact. We factored in the aforementioned "mutational difficulty" to estimate the functional importance of each mutation. For example, the M133R mutation is caused by T→G substitution on a GATGT sequence, whose difficulty index is 233. This mutation appeared in only 11 cancer samples in the COSMIC database. Given our argument, the frequency of this M133R mutation may have been severely penalized by the high mutational difficulty. Considering this, we designated M133R's original count as 11 and revised count as 2563 ($= 11 \times 233$). Notably, the revised count for this mutation is comparable to that of the hotspot R282W mutation (original count 609, difficulty index 3.31, revised count 2017), suggesting M133R is also a deleterious mutation despite its low frequency in cancer database.

To more precisely assess these p53 mutations, we also took into consideration that the same type of mutations is generated at different rates in different cancer types (Fig. 1f). Therefore, in all our analysis we used cancer type-specific mutational difficulty indexes to calculate the revised mutation count for each p53 mutation (see "Methods") (Supplementary Table S3).

The global view of p53 missense mutations is provided in Fig. 2a. The map of p53 original mutation count is characterized by seven high peaks at R248 and R273, which are crucial for interaction with DNA, as well as R175, Y220, G245, R249, and R282, which are crucial for maintaining p53 structure. In the revised mutation count map, many more such high peaks appeared, suggesting that other portions of p53 also contain numerous amino acid residues that are essential for p53 function (Fig. 2a). Importantly, judging from original counts, only a few p53 missense mutations occur more frequently than the hotspot R282W mutation (Fig. 2b). After considering the mutational difficulty, more than 130 of p53 missense mutations exhibit a higher revised count than R282W (Fig. 2b), suggesting that many more p53 missense mutations potentially abolish gene function.

To establish a cut-off value that could help identify p53 mutants that still retain wild-type function, we compiled revised count values for all p53 synonymous mutations and found them to be mostly below 700 (Supplementary Fig. S6). Therefore, a revised count below 700 may suggest wild type function for p53 mutants. We also estimated that a revised count over 900 might suggest loss of function. We constructed more than 80 low-frequency p53 mutants with various revised count values to test such hypothesis. The human osteosarcoma cell line Saos-2 carries homozygous deletion of p53. It could tolerate hotspot p53 mutants but not wild-type p53 (Fig. 2c). Twenty-two p53 mutants with revised count lower than 700 were cloned and tested. Consistent with our hypothesis, they all behaved like wild type p53 in this assay (Fig. 2c, Supplementary Fig. S7), suggesting they do retain gene function as predicted by our method. For example, the R282Q mutation (original count 36) is located on the functionally essential amino acid residue R282. However, the underlying mutation is relatively easy to occur, and with a revised count of 186, this mutant retained wild-type p53 function. We also noticed that the highly frequent R158H mutation, although observed in 95 cancer samples, is an easy-to-occur mutation. With a revised count lower than 700, this mutation also retained wild type function. The high number of cancer samples carrying this R158H mutation may be more of a result of the easiness to generate the underlying mutation.

In contrast, certain high difficulty p53 mutations, although many of which only occur in less than 10 cancer samples, are predicted to be loss of function mutations with revised counts over 900. We examined 61 such p53 mutations, and they were all well tolerated by Saos-2 cells, confirming their loss-of-function status (Fig. 2c, Supplementary Fig. S7).

We also noticed that, on P177, the P177L mutation (original count 24, revised count 262) retains wild-type function (Fig. 2d). Interestingly, on the same residue is another mutation P177R, which exhibit lower original count but much higher revised count (original count = 18, revised count = 2887). Despite it being less frequent than P177L, it is actually a loss-of-function mutation (Fig. 2d). Importantly, in our analysis we observed multiple such cases that even on the same residue, less frequent mutations could be loss-of-function, yet mutants with higher original counts retain wild-type function. Examples include R282Q/P, M160I/L, L130F/R, R158H/L, and others (Fig. 2d). Such a reverse phenomenon can be explained by their revised mutation counts, again demonstrating the validity of our method.

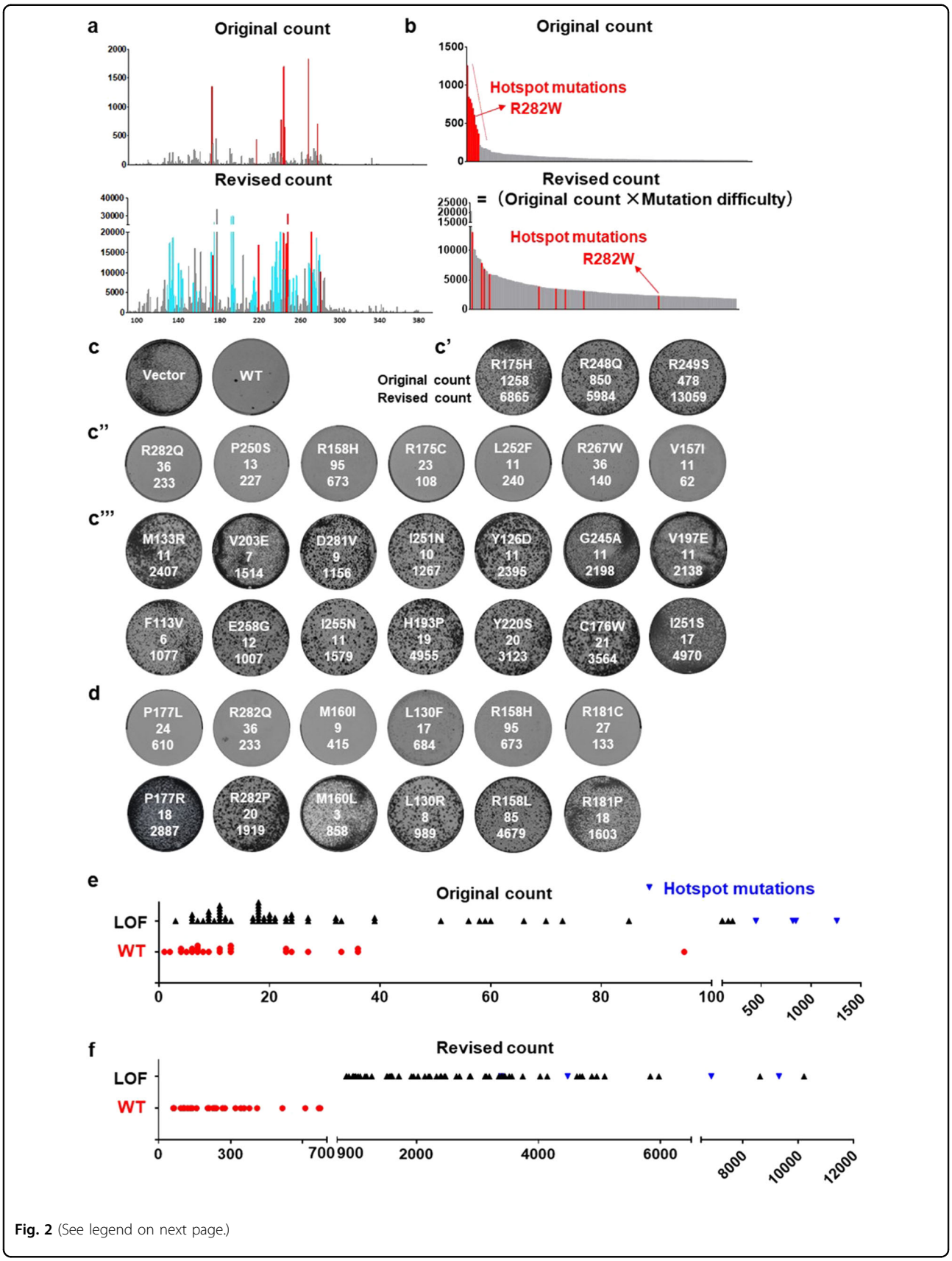

Fig. 2 (See legend on next page.)

(see figure on previous page)

Fig. 2 Integrating relative mutational difficulty to predict the functional status of p53 mutations. a p53 mutation histogram based on original and revised counts. Different types of mutations on the same amino acid residue (e.g., R273H and R273C) are combined to make this graph. Red lines indicate hotspot mutation sites such as G245 and R282. In the lower panel, anmino acid residues with low original counts but high revised counts are marked in blue. **b** The original and revised counts of p53 cancer mutations. Red lines indicate hotspot mutations such as R282W. **c** Expression of wild-type p53 suppresses the growth of Saos-2 cells. Genes were delivered to cells via retroviral infection. For all colony formation assays in this study, cells were infected with low MOI such that 30–50% of cells were infected with virus. **c′** p53 hotspot mutants are well-tolerated by Saos-2 cells. The original and revised counts are listed below each mutant. **c″** p53 mutants with revised count lower than 700 behave like wild-type p53 and suppresses Saos-2 growth. **c‴** p53 mutants with revised count higher than 900 are loss of function mutants and are well-tolerated by Saos-2 cells. All colony formation assay in this study were done in three independent biological repeats. **d** Pairs of p53 mutations on the same amino acid. Shown are examples of high-difficulty mutations, although appearing in lower number of cancer samples, are loss of function mutations instead. **e** Original mutation counts do not correlate with functional status of p53 mutants. **f** Revised mutation counts correctly predict the functional status of p53 mutants.

To examine the biochemical function of these p53 mutants, we introduced them into HCT116 p53−/− cell line, and tested whether DNA damage drugs can still induce the expression of p21, a well-established p53 transcriptional target[18]. Quantitative polymerase chain reaction (qPCR) analysis showed that these p53 mutants were expressed at similar levels (Supplementary Fig. S9a). Again, p53 mutants with revised counts lower than 700 behaved similarly to wild p53, whereas p53 mutants with revised mutation counts higher than 900 behaved similarly to hotspot mutants, failing to upregulate p21 mRNA expression upon DNA damage (Supplementary Fig. S9b).

Summarized in Fig. 2e, despite the common perception that high-impact mutations appear more frequently in cancer database, the original mutation count is not a reliable predictor of functional status. p53 mutants with original counts less than 100 can either be loss of function mutants or retain wild-type function. In contrast, the functional status of p53 mutants are accurately predicted by their revised mutation counts (Fig. 2f). This shows that by defining relative mutational difficulty, we can provide novel tools to accurately assess cancer mutations.

Dominant negative effects of p53 mutants

It is known that human p53 hotspot mutations also exert dominant negative effect over wild-type p53[19]. To explore whether such dominant negative effect also exists for other p53 mutations, and whether our method could predict such dominant negative effect, we established an experimental system using the Eμ-Myc p19Arf−/− mouse lymphoma cell line. This cell line retains wild type p53, which can be activated by DNA damage to induce cell death[20]. Expression of hotspot p53 mutant together with green flourescent protein (GFP) was achieved in this cell line via retroviral vectors. Hotspot p53 mutant exerts dominant negative effects over endogenous wild-type p53, and cells could not efficiently elicit cell death when treated with DNA damaging drugs. As a result, the percentage of GFP-positive, hotspot p53 mutant-expressing cells will increase after drug treatment (Fig. 3a). In contrast, expression of wild-type p53 in this system moderately will sensitize cells to DNA damage drugs (Fig. 3a).

We cloned the murine versions of various human p53 mutants and tested whether they exhibit dominant negative function. Importantly, among the 83 constructs we tested, all p53 mutants with revised counts lower than 700 behaved like wild-type p53 (Fig. 3b), whereas all p53 mutants with revised count higher than 900 exhibited dominant negative effect (Fig. 3b). Again, as a predictor of dominant negative effect, revised count performed significantly better than original mutation count (Fig. 3c, d).

We further tested whether our method could predict cancer-promoting abilities of p53 mutants in vivo. Using a tail-vein hydrodynamic injection method, together with a transposon system[21] and CRISPR gene editing[22], Myc overexpression and PTEN knockout was achieved in liver cells of wild-type FVB mice. Under such condition, no mice developed liver tumor at 3 weeks. Addition of R246S murine p53 mutant, which mimics the human R249S hotspot mutation, overrode endogenous wild-type p53 in mice liver and caused massive tumors (Fig. 3e). Using this setting, we tested eight p53 mutations. Among them, R175C and R282Q are moderately recurring mutations on amino acid residues that are essential for p53 function. R158H is highly recurrent with an original count of 95. Our method and in vitro results (Fig. 2c″) suggest that despite their crucial location and/or high frequency, they all retain wild-type p53 function. On the other hand, K132Q, H193P, I251S, and I255F are predicted to be loss of function mutations based on their revised counts. These mutations are located at several β-strands, structures that are understudied in terms of their importance to p53 function. The cDNAs of these p53 mutants were introduced together with Myc cDNA and sgPTEN to analyze their ability to promote liver cancer in vivo.

Figure 3e showed images of whole mouse livers from this experiment. Consistent with our prediction, four mutants with revised counts lower than 700 all behaved like wild-type p53 and caused no tumors. In contrast, four p53 mutants with high revised counts all caused massive

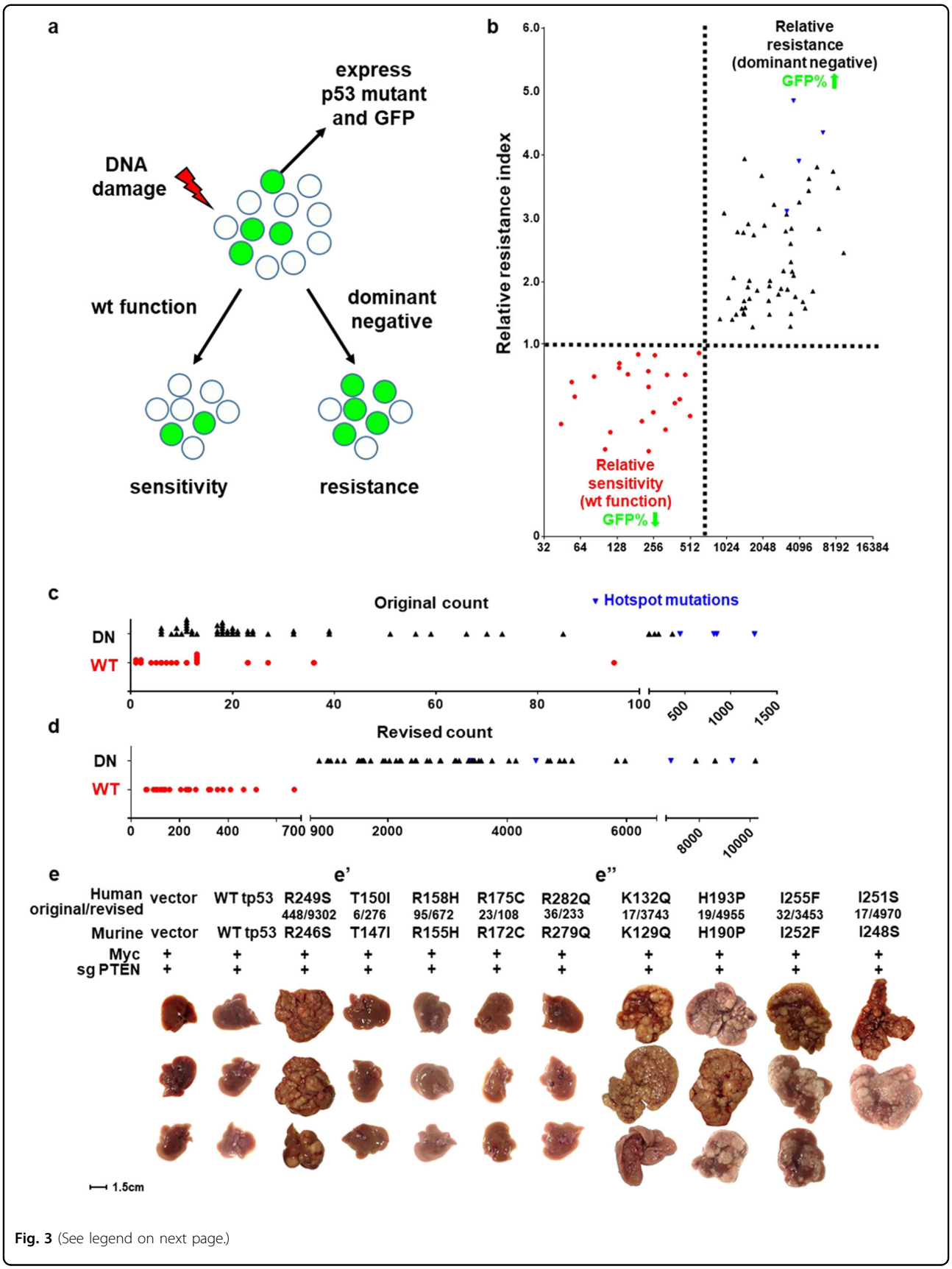

Fig. 3 (See legend on next page.)

(see figure on previous page)

Fig. 3 Integrating relative mutational difficulty to predict dominant negative effects of p53 mutations. a An experimental system to analyze dominant negative effects of p53 mutations. A murine lymphoma cell line that retain wild-type p53 is partially infected by retrovirus that express p53 mutants and GFP. If the p53 mutants exert dominant negative effect over endogenous wild-type p53, it will render cells more resistant to DNA damage induced by doxorubicin, and the rate of GFP-positive cells increases in surviving cells. Expression of wild-type p53 will moderately sensitize cells to doxorubicin treatment. **b**–**d** Revised counts, but not original counts of p53 mutants correctly predicts whether such mutants exhibit dominant negative effects. Murine p53 mutants corresponding to human p53 mutants were used in these experiments. **e** Revised counts correctly predict whether p53 mutants can promote liver cancer formation in vivo. The original and revised counts are listed below each mutant, separated by a "/" mark. Murine p53 mutants corresponding to human p53 mutants were used in this experiment. Mice were sacrificed 30 days after hydrodynamic delivery of genes in vivo. $n = 3$ for each experimental group, except I248S for which one of the injected mice did not recover from hydrodynamic injection. Shown are images of whole liver from each mouse.

liver tumors in mice (Fig. 3e), indicating they were able to override endogenous wild-type p53 to promote cancer.

Functional importance of several regions on p53 structure highlighted by our method

Our results suggest there are many low frequency p53 mutations with significant functional impacts. Their locations are marked by blue lines in Fig. 2a. To better understand their general distribution, we mapped these high-impact mutations on the three-dimensional structure of p53. Known p53 hotspot mutations are located on the interfaces that are crucial for p53 structure and interaction with DNA. We first noticed that many residues adjacent to hotspot sites, such as V173, H178, M246, V274, and A276, although with rather low original mutation counts, showed very high revised counts (Fig. 4a, Supplementary Fig. S10a). This observation suggests that many residues surrounding hotspot sites are in fact also essential for p53 function. These crucial sites are rarely mutated in cancer because their mutation frequencies are severely penalized by high mutational difficulty.

In addition to these residues, several other regions of p53 stood out with high revised mutation counts. One of such rarely mutated, high impact region is residues 130–138. These residues form a β strand and loop structure that lays closely to a β strand–loop–helix domain (amino acids 270–282), which host several hotspot mutations and are responsible for DNA-interaction[23,24] (Fig. 4b). Other high impact amino acids identified by this method are five β strands that formed the central β-barrel of p53. On the three-dimensional structure of p53, such amino acids are also very close to AA173–179 and AA244–249, both hosting hotspot mutations[25] (Fig. 4c, Supplementary Fig. S10b). Colony formation assays in Saos-2 cells confirmed that many rare, but difficult-to-generate mutations on these sites disrupt p53 function (Supplementary Fig. S7). Such mutations were also tested in the Eμ-Myc p19Arf−/− system and exhibited dominant negative effects over wild-type p53 (Fig. 3b). Taken together, our method could regroup p53 mutations by integrating mutational difficulty, and points to additional regions that are crucial for the function of p53.

Of note, previous studies established that p53 has five conserved domains[26,27]. We complied revised counts for these five conserved domains (Fig. 4d). Many residues on conserved domain II–V showed high revised count values, indicating that these domains are indeed crucially involved in cancer formation. The exception is conserved domain I (AA 13–20), which contains two Serine residues (S15 and S20) that are phosphorylated upon DNA damage. These two residues may be functionally overlapping, and mutation of either S15 or S20 alone may not be enough to allow for prompt tumorigenesis. Consistent with this, mouse models in which these two serine residues were mutated either developed no tumors[28], or showed much weaker tumor phenotypes compared with p53 hotspot mutant mice[29,30].

Functional landscape of p53 mutations in human cancer

Our results suggest that, in addition to hotspot mutation sites, numerous other amino acid residues are also crucial for p53 function. Based on our analysis, we estimate that out of the 1219 types of missense p53 mutations in COMIC database, 27% are loss of function mutations and 70% retain wild-type function. In addition, out of the 19598 p53-mutated cancer samples in COSMIC database, 83% samples contain loss of function p53 mutation and 15% samples retain wild-type p53 function. Experimental analysis (Figs. 2 and 3) showed that our method can accurately predict the functional status of p53 mutations. However, the experimental work is limited by the fact that we have not conducted an unbiased screen of all possible mutations, and therefore the true-positive rate cannot be extrapolated across to other mutations. To address this, we compared our functional assessment of p53 mutations with other methods.

The Functional Analysis through Hidden Markov Models (FATHMM) method[31], which estimates functional impact based on sequence conservation and the overall tolerance of the protein/domain to mutations, has been commonly used to predict cancer-driving mutations. Such a method is used by COSMIC to annotate cancer

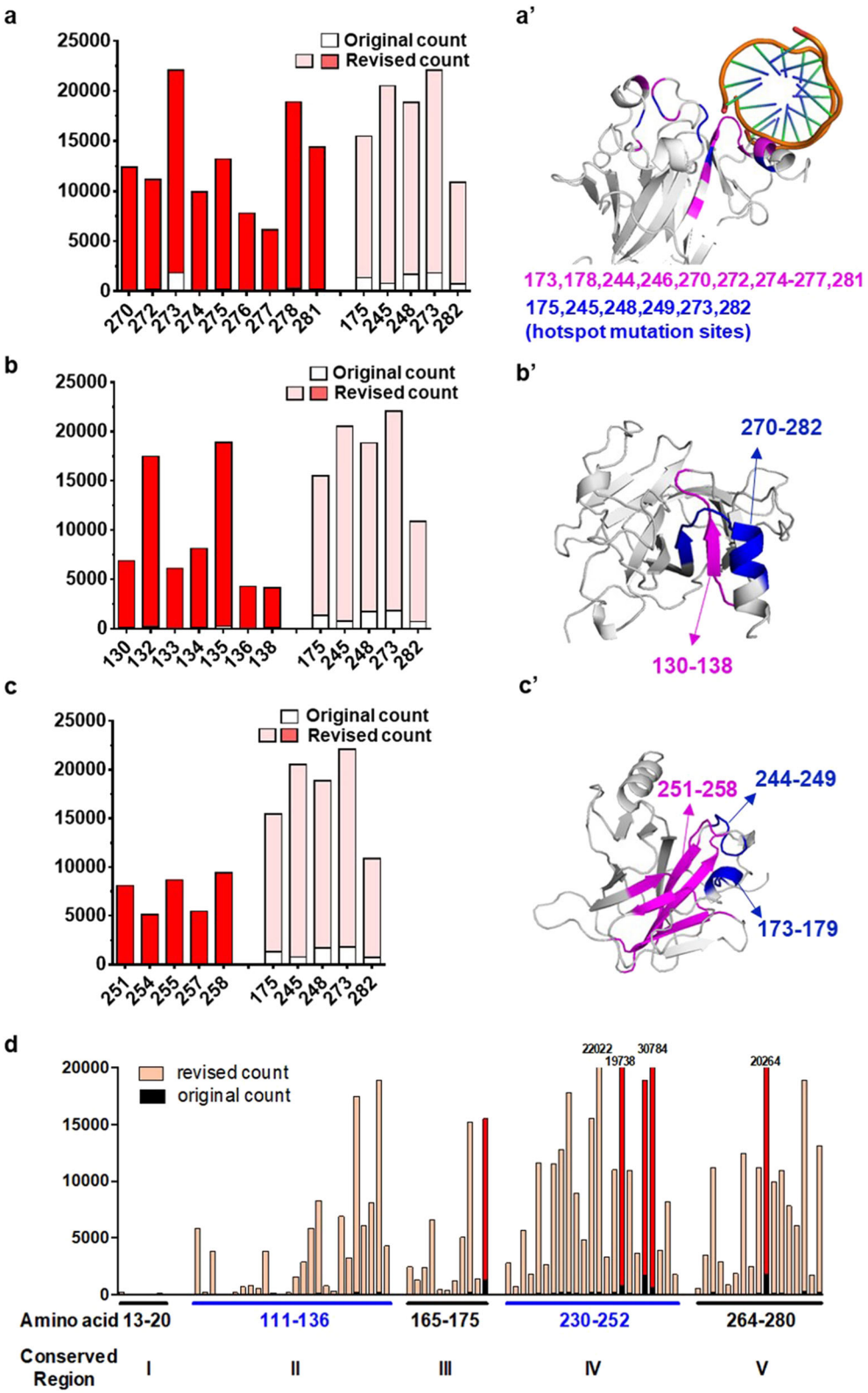

Fig. 4 (See legend on next page.)

(see figure on previous page)
Fig. 4 Functionally important amino acid residues and regions in p53. a–c On the left panels, original mutation counts of listed amino acid residues are indicated by white boxes, whereas revised counts are indicated by red or pink boxes. Hotspot mutation sites such as R175 and G245 are included as controls. On the right panels, **a′** functionally important amino acid residues near hotspot mutation sites are labeled in purple. Hotspot mutation sites are labeled in blue. **b′**, **c′** Additional regions crucial for p53 function are labeled in purple. Regions that mediate DNA binding such as AA270–282 and AA173–179 are labeled in blue. **d** The original and revised mutation counts for amino acid residues on the five conserved regions of p53. Columns in red indicate amino acid residues that host hotspot mutations.

mutations. With regards to p53 mutations, comparison of our prediction results with the FATHMM method showed only 50% overlap (Fig. 5, Supplementary Table S4).

Interestingly, in a recent publication by Giacomelli et al., thousands of different types of p53 mutants were introduced to cancer cells, and the functional status of these p53 mutations were assessed by whether these mutations were tolerated by cells under different conditions[32]. Such a dataset (PHANTM) provides direct experimental readout of p53 mutations. Our predictions of the functionality of p53 mutations are highly consistent with their experimental results, with an 88% overlap (Fig. 5). The 12% p53 mutations that are differently predicted are listed in Supplementary Table S5. They may have resulted from wrong prediction of our method, or from small inaccuracies associated with the pool-based large-scale studies used in Giacomelli et al.

We further looked at how other bioinformatic tools such as PROVEN, SIFT, Polyphen predict the functionality of p53 mutants. In comparison, our method showed the highest percentage of consistency with the experimental results of Giacomelli et al.[32] (Fig. 5).

Predicting the functional status of PTEN and INK4A mutations

Next, we asked whether this method could be applied to other established cancer genes such as PTEN and INK4A. We cloned about 20 low-frequency PTEN and INK4A mutations and expressed them in PTEN or INK4A deficient cancer cell lines to see whether such mutations abolish gene function. As predicted by our method, those mutations with low revised counts retained wild-type function, whereas those mutations with high revised counts caused gene loss of function (Fig. 6a–c).

For example, high difficulty mutations including PTEN Y27N, C124S, and I135K, although each only occurring in five cancer samples in COSMIC, could not suppress AKT signaling, proving that they all abolish PTEN function (Fig. 6a). In contrast, the PTEN R173C and R173H mutations, despite being the fifth and sixth most common PTEN mutations and occurring in 47 and 36 cancer samples, both retained wild-type function (Fig. 6a). According to our method, both are low-difficulty mutations, which explains why they did not disrupt PTEN function. This observation, together with finding that p53 R158H mutation (original count 95, revised count 673) also retains wild type function, demonstrate that our method not only help identify rare mutations that promotes cancer, it can also point out high frequency, passenger-type mutations in cancer database.

Discussion

Significant differences of mutational difficulties in human cancer

The functional importance of a mutation to cancer can be reflected by its selective enrichment in cancer samples. However, due to the lack of understanding of relative mutational difficulty in cancer, most studies use mutation frequency in cancer database to directly calculate selective pressure. Our analysis shows that, depending on the type of nucleotide substitution and the surrounding sequences, the chances of generating different types of mutations can vary by as much as 400-folds (Fig. 1c). Such a drastic difference highlights the need to reapproach how we interpret the functional importance of cancer mutations.

Many factors contributed to the fact that different types of nucleotide substitutions are created at rather different rates[9]. In this report, through analysis of large number of human cancer genomes, we reversely derived the relative difficulties for each type of mutation. We also established such numbers in a cancer type-specific manner. Such a dataset (Supplementary Table S2) will be a useful tool to understanding cancer genome.

For most genes, close to 30,000 cancer samples have been analyzed and deposited in the COSMIC database as of January 2018. Certain easy-to-occur mutations may simply accumulate in numerous cancer samples without providing advantages for cancer development. In the future, when increased number of cancer genomes are deposited to the COSMIC database, it is expected that more and more such easy-to-occur passenger mutations will pile up on the mutation histogram. Without considering relative mutational difficulty, these seemingly "mutational peaks" may lead to erroneous assumptions that they are cancer-driving mutations.

To functionally estimate the importance of novel cancer mutations, the cancer types that host such mutations should also be taken into consideration. As illustrated in Supplementary Fig. S5, for many types of mutations, it is

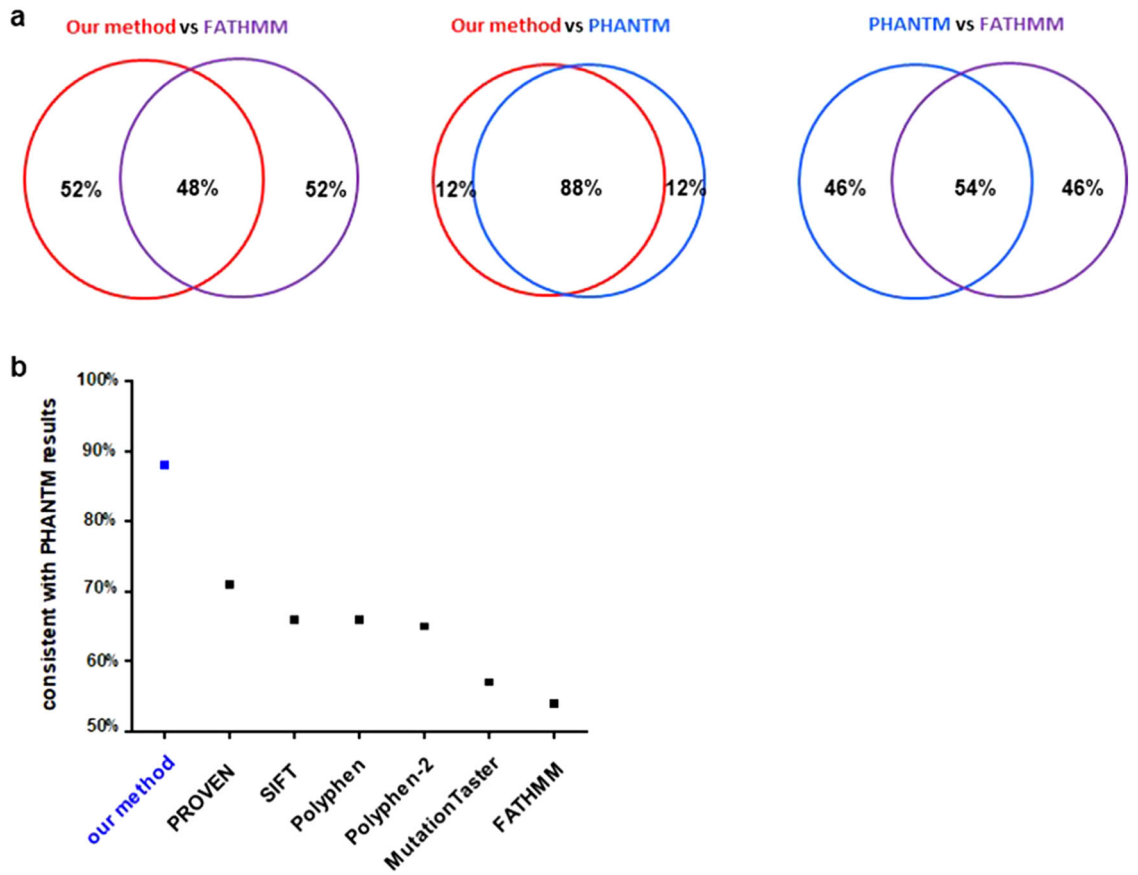

Fig. 5 Comparison of functional predictions based on different methods. a FATHMM annotation of p53 mutants are retrieved from individual p53 mutation page on COSMIC site. PHANTM, phenotypic annotation of TP53 mutations (http://mutantp53.broadinstitute.org/) provides functional assessment of individual p53 mutant based on the experimental results of Giacomelli et al.[32]. Prediction by our method are highly consistent with the PHANTM results, with an 88% overlap. **b** Comparison of our method with other methods with regard to p53 mutant functional annotation.

much easier to generate them in skin, colorectal and endometrial cancers. At the same time, some types of mutations are relatively more difficult to generate in these cancer types (Supplementary Fig. S5). Therefore, if the original count of a cancer mutation is primarily contributed by skin, colorectal, and endometrial cancer samples, such mutations should be viewed with caution. However, they should not be automatically overlooked either.

A previous study by Cannataro et al.[33] used notions similar to ours. Selective pressure for a cancer mutation was calculated through their observed frequency and mutation possibility. However, our method differs significantly with Cannataro et al. in how we calculate mutation tendency. In the Cannataro study, mutation possibility was calculated first by assigning mutations in a cancer sample to Sanger mutation signatures. However, for a number of cancer-causing factors including nitrosamine and drinking hot water, Sanger mutation signatures have not been established. In addition, in the Cannataro study, only cancer samples with more than 50 mutations were used to calculate mutation possibility. This would exclude about half of cancer samples in the COSMIC database. For example, about 67% of breast cancer samples and 90% of leukemia and lymphoma samples carry less than 50 mutations per sample. Our method also differs with Cannataro et al. in two additional aspects. First, we excluded mutations that have been strongly enriched during cancer formation (Fig. 1a), as they could significantly skew our estimation of background mutation tendency (Supplementary Fig. S3). Second, in our analysis we also considered the effect of +2 and −2 nucleotide, which could significantly impact mutation difficulty[34] (Fig. 1e).

Potential implications for cancer prevention

In addition to establishing a method that could help evaluate individual cancer mutations in a sequence- and cancer type-specific manner, we also asked whether our method could help understand cancer etiology in general.

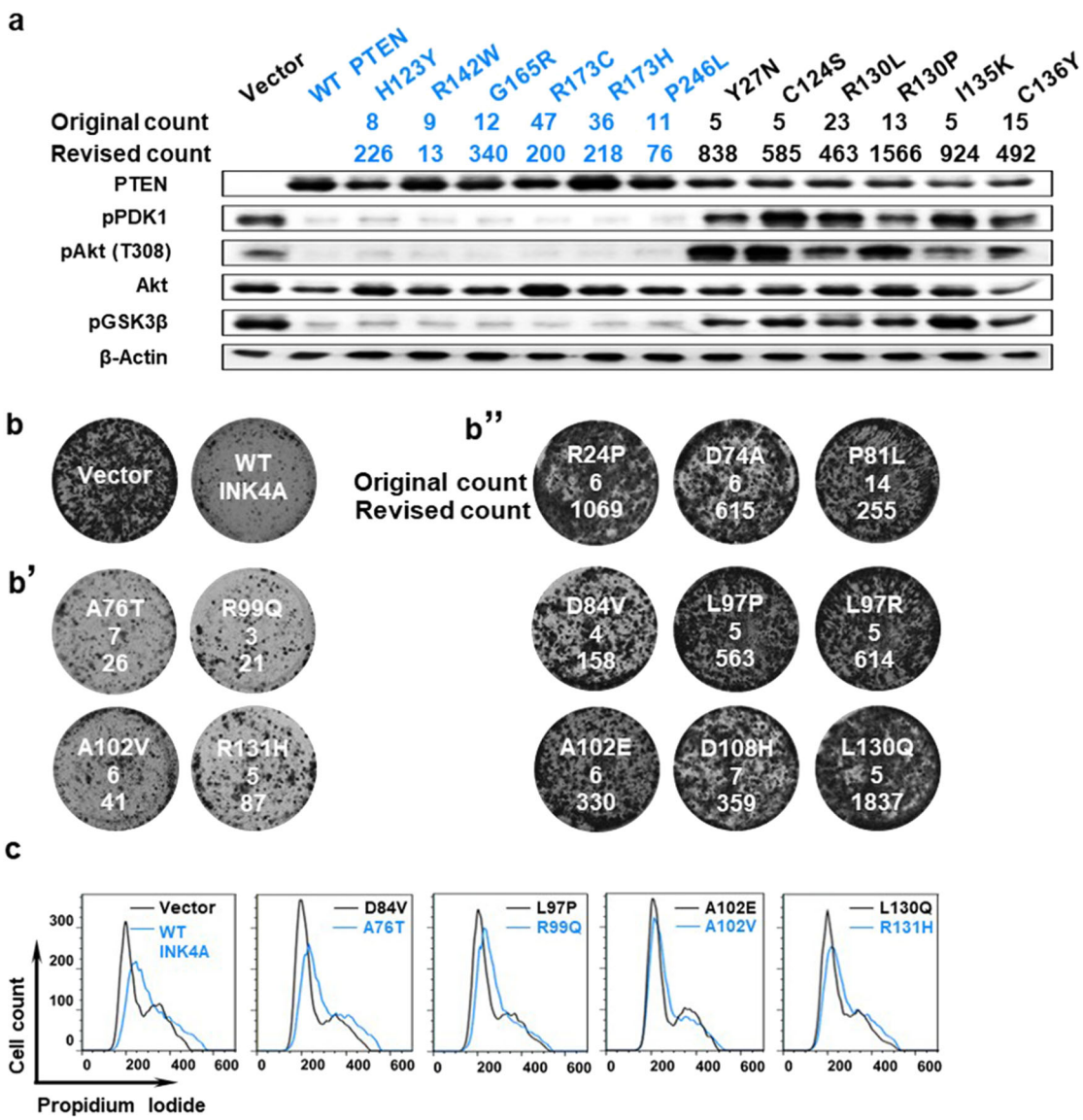

Fig. 6 Integrating relative mutational difficulty to assess the functional status of PTEN and INK4A mutations. a Original and revised counts of indicated PTEN mutants. Expression of wild-type PTEN suppresses AKT signaling in 786-O cells, which do not express PTEN. PTEN Mutants with low revised counts (in blue) behave like wild-type PTEN and suppress downstream AKT signaling. PTEN mutants with high revised counts (in black) lose gene function and cannot suppress AKT signaling. The original and revised counts are listed below each mutant. **b** Expression wild-type INK4A suppress growth of U251 cells, in which the INK4A locus is deleted. **b'** INK4A mutants with low revised counts retain wild-type gene function and suppressed U251 growth. **b''** CDKN2A mutants with high revised counts are defective in gene function and are well tolerated by U251 cells. The original and revised counts are listed below each mutant. **c** INK4A mutants with high revised counts (in black) cannot induce cell cycle arrest in U251 cells. CDKN2A mutants with low revised counts (in blue) behave like wild-type CDKN2A and induce G1/early S phase arrest in U251 cells.

In our analysis of human cancer mutations, we noticed that the vast majority of common mutations on tumor suppressors, such as p53, PTEN, FBXW7, and SMAD4 are low-difficulty mutations (Supplementary Fig. S11a). Totally, 77% of these common, low difficulty mutations on tumor suppressors are C→T or its complementary G→A mutations on CpG sequence, which could be the results of spontaneous deamination, one of the most frequent type of DNA damage in cells[9,35,36]. Our analysis also showed such kind of C→T or G→A mutations on CpG sequence are the easiest to occur types of mutations in human cancer (Supplementary Fig. S4). From a cancer prevention point of view, it will be rather difficult to prevent such type of deleterious events on tumor suppressors.

In contrast, except for IDH1/2 and AKT1, common mutations on oncogenes such as KRAS, BRAF, CTNNB1,

PI3KCA, and JAK2 are typically high-difficulty mutations (Supplementary Fig. S11a'). Only 14% of common cancer-promoting mutations on oncogenes are C→T or G→A mutations on CpG sequences. About half of common cancer-promoting mutations on oncogenes require purine to pyrimidine changes or vice versa. Such type of drastic changes is unlikely to be caused by simple chemical reactions such as deamination of the nucleobase. Rather, exogenous carcinogenic events are potentially needed to damage DNA to eventually create such types of mutation[10]. Limiting the exposure to environmental carcinogens, as well as managing long-term inflammation, among many applicable measures, may significantly reduce the chance of obtaining activating mutations on oncogenes in general. This will significantly deplete the driving force of cancer and impede cancer development. Therefore, even though two-third of mutations in human cancer are caused by spontaneous events[37], avoiding environmental carcinogenic factors holds great promises to significantly reduce the incidence of many types of cancer.

On the other hand, certain types of cancers will still be hard to prevent. For example, the driver mutations on the oncogenes IDH1 and IDH2 are both low-difficulty, C→T or G→A mutations (Supplementary Fig. S11a'). Therefore, cancer cases associated with such mutations, including certain subtypes of glioblastoma, cholangiocarcinoma, and acute myeloid leukemia[38,39] may be hard to prevent. Similar to the argument by Tomasetti et al.[37], for these types of cancer early detection still holds more promise than cancer prevention methods.

Lastly, it is apparent that low difficulty, spontaneous mutations on tumor suppressors contribute to human cancer, most significantly through several easy-to-mutate hotspot sites on p53 (Supplementary Fig. S11a). From a pure theoretical point of view, it is possible to introduce synonymous mutations to these sites to render them more resistant to deleterious mutations. We analyzed the potential benefits of changing the nucleotide coding sequence on p53 hotspot sites. For example, changing the p53 R273 sequence from R(CGT) to R(AGA) will reduce the chance of generating loss of function mutations on this site by seven folds (Supplementary Fig. S11b). For four other mutational hotspots on p53, similar codon changes can also significantly reduce the chance of generating loss of function mutations (Supplementary Fig. S11b) and are projected to greatly reduce cancer cases involving these hotspot sites (Supplementary Fig. S12a, b). Our analysis also showed that such codon changes will have minimal impact on the chance of generating LOF mutations at amino acid residues surrounding hotspot sites (Supplementary Fig. S12c). If theoretically, spontaneous low-difficulty mutations on p53 can be limited by such measures, and high-difficulty mutations on oncogenes and other sites of p53 can be thwarted by avoiding environmental carcinogens, it may dramatically reduce cancer incidence.

Methods and material

Data acquisition

Mutation data from 26,154 cancer genomes were retrieved from COSMIC website in January 2018. If a gene has multiple isoforms, only the major form was included in our analysis such that mutations on the same sites are not counted multiple times. For 19,940 genes, 3,101,161 single-nucleotide substitutions were identified. In order to calculate the natural mutational tendency, we first eliminated mutational events that occur more than five times in the dataset. These mutations may have been selectively enriched during cancer development and could skew our calculation of nature mutational tendency.

To assess the influence of neighboring sequences on mutational tendency, sequences of the coding genome corresponding to the 19,940 genes were downloaded from Ensemble (GRCh38.p10). For each nucleotide mutation, -2, -1, $+1$, $+2$ nucleotides were extracted from the corresponding coding sequence.

Types of mutations

At the central position, there are 12 routes of interchange between A/G/T/C. The permutations at -2, -1, $+1$, $+2$ nucleotides amount to 4^4. Therefore, we collated all mutations into $12 \times 4^4 = 3072$ groups.

Calculation of mutational tendency

For the aforementioned 3072 groups, we first counted how many mutations from the 26,154 cancer genomes belong to each group. Next, we counted how many times each penta-nucleotide sequence appears in coding sequences of the 19,940 genes. For example, there are 10,389 C→T mutations on TTCGT sequences in 26,154 coding genomes. There are 21446 TTCGT sequences per coding genome. Therefore, the mutational tendency of C→T on TTCGT is approximately $10389/(21446 \times 26154) = 1.85 \times 10^{-5}$, which is the highest amongst all 3072 combinations. We set the "difficulty" score for such a mutation as 1. The mutational tendency of A→C mutation in a CGATG sequence is 0.93×10^{-7}, and its relative difficulty score is calculated as $1.85 \times 10^{-5}/0.93 \times 10^{-7} = 200$. Difficulty scores for all other combinations were generated using the similar method.

Of note, the numbers of different penta-nucleotides in the coding genome vary greatly. For example, in the coding genome there are 3001 TAGCG sequences and more than 100,000 TGGAG sequences. Therefore, it is necessary to divide the number of mutations by the number of available sites to accurately understand the relative mutational difficulty.

The necessity to remove highly recurrent mutations

In the above analysis, we aim to estimate the mutational tendency for each type of mutation in human cancer. Certain cancer-promoting mutations on genes such as KRAS and BRAF are strongly selected for during cancer formation. The number of such mutations are significantly increased in the dataset, not because they are easy to generate, but because they are strongly enriched by the tumorigenesis process. Therefore, their presence in the dataset may skew our estimation of the natural mutational tendency for each type of mutation. Considering this, in the above calculation, we excluded mutations that occur in more than five cancer sample, in order to achieve a closer estimate of mutational tendency. Of note, about 2% all of mutations in the 26,000 cancer genomes (Fig. 1a) occur in more than 5 cancer samples and were excluded in our analysis.

Supplementary Fig. S3 shows the comparison of mutational difficulties calculated with and without excluding such recurrent mutations. In Supplementary Fig. S3a, if no mutations are excluded, the mutational difficulty scores for KRAS G12R, BRAF V600E, and HIF1A K213Q, among many others, will be significantly lower. In Supplementary Fig. S3b, if only excluding mutations that occur in more than 20 samples, the mutational difficulty scores for TP53 V157G, NOTCH1 D573A, CDKN2A A36G, among others, will still be significantly lower. In Supplementary Fig. S3c, if only excluding mutations that occur in more than ten samples, the mutational difficulty scores for TP53 Y126D, KDM6A T794P, PIK3CA V344G, among others, will still be noticeably lower. Based on this, we calculated mutational difficulty after excluding mutations that occur in more than five samples.

Cancer type-specific mutational difficulty

To generate cancer type-specific mutational difficulty scores, mutations were first grouped by cancer types, from which mutation rates were calculated using similar methods. For example, we observed 1195 C→T mutations on TTCGT sequences in 296 endometrial cancer samples, and the mutational tendency of C→T on TTCGT is approximately 1.88×10^{-4} in endometrial cancer. Since in previous calculation we set the mutational difficulty score as 1 for a mutation rate of 1.85×10^{-5}, we can calculate the relative mutational difficulty for C→T on TTCGT as 0.1 ($= 1.85 \times 10^{-5}/1.88 \times 10^{-4}$) in endometrial cancer.

Analysis of p53, PTEN, and INK4A mutations

Mutational data for p53, PTEN, and INK4A was last acquired from COSMIC on January 2018. At the time, p53 mutation data were from 130,448 cancer samples, PTEN from 72,199 samples and INK4A from 72,566 samples. Current numbers in COSMIC database have slightly increased due to website updates.

Of note, the CDKN2A locus contains two genes, INK4A and ARF. Previous studies showed that recurrent mutations on the CDKN2A locus do not change the function of the ARF gene[40]. In addition, U251 cells, which deleted the CDKN2A locus, could tolerate ARF expression, but not INK4A expression. Therefore, for later experimental validation, we cloned and analyzed INK4A mutants in this study.

For each mutation, we first extracted the pentanucleotide sequence surrounding the mutation site and matched it with relative mutational difficulty scores. For mutational sites that are adjacent to intron–exon junctions, the genomic sequence was used to extract the nucleotide sequences surrounding the mutational site.

Next, we calculated the revised mutational count based on original mutation count and cancer type-specific relative mutational difficulty. For example, if a p53 mutation occurs in 10 colorectal cancers and 5 lung cancers, and the relative mutational difficulties for the mutation is 1 in colorectal and 3 in lung cancer, the revised count for such a mutation can be calculated as $10 \times 1 + 5 \times 3 = 25$. Revised counts calculated using this method were used to predict the functional impact of p53, PTEN, and INK4A mutations in our study.

Supplementary Fig. S6 shows the distribution of revised mutation counts for all p53 synonymous mutation based on COSMIC data. For most of these synonymous mutations, the revised counts are below 700. Therefore, we estimate that those missense p53 mutations with revised count below 700 retain wild-type p53 function, which were later validated with functional experiments. In Supplementary Fig. S6, we also observed that the revised counts of several p53 synonymous mutations exceeded 700. This is because certain seemingly synonymous mutations abolish p53 function. For example, the p53 T125T mutation (c.375G to A/C/T) disrupts the adjacent intron–exon splice site, and abolishes gene function[41]. The revised count for T125T is 1353 and is predicted to be a loss of function mutation by our method.

Supplementary Table S3 listed the original and revised counts of p53 mutations based on COSMIC database. If a mutation's revised count is lower than 700, it is predicted to retain wild-type function. If a mutation's revised count is higher than 900, it is predicted to be loss of function mutation. A few exceptions exist and are explained below.

In the COSMIC database, the S149F mutation on p53 is caused by single nucleotide substitution in 5 samples, and the revised count is lower than 700. However, in one additional cancer sample, a CC to TT nucleotide change also caused the S149F mutation. Because of the rarity of such double mutations, we did not assign relative mutational difficulty score to such double mutations. Therefore, we cannot make functional prediction for this mutant, and an "*" is marked in the "revised count"

column for S149F. Such phenomenon also occurred for S166L, V218M, and R158C, and these mutations are labeled similarly with an "*" in Supplementary Table S3. Several other mutations (e.g., S127F) also exhibited such double nucleotide substitution, however, their revised counts calculated from single-nucleotide substitutions already exceeded 900. Therefore, such mutations are predicted to be loss of function mutations in Supplementary Table S3.

GFP-based cell survival competition assay to determine sensitivity change caused by p53 mutants

The experiment was carried out with a protocol modified from[42]. Briefly, Eμ-Myc p19Arf−/− cells are infected with retrovirus that express GFP and mutant p53, such that 20–50% of cells are GFP positive. Cells are treated with DNA damage drug at doses that would kill 80–90% of uninfected Eμ-Myc p19Arf−/− cells. In this assay, if p53 mutant exerts dominant negative effects on endogenous wild-type p53, after DNA damage drug treatment the GFP positive, p53 mutant-expressing cells will be relatively more resistant than GFP-negative cells that only express wild-type p53. At 72 h, treated and untreated cells are analyzed by flow cytometry. GFP percentages of live (PI-negative) cells are recorded and used to calculate relative resistance index (RI).

Calculation of relative resistance/sensitivity from GFP-based cell survival competition assay

The value of relative RI can be calculated as $RI = (G2 - G1 \times G2)/(G1 - G1 \times G2)$. G1 means how many percentages of cells are GFP positive before drug treatment. G2 means how many percentages of cells are GFP positive after drug treatment. The explanation for such calculation was provided in ref. [43].

Relative RI larger than 1 means the corresponding p53 mutant displayed dominant negative effect, protected cells from DNA damage, and the rate of GFP+ cells in surviving cells increased after drug treatment. Relative RI smaller than 1 means the corresponding p53 mutant displayed wild-type function, sensitized cells to DNA damage, and the rate of GFP+ cells in surviving cells decreased.

Cell lines and drugs

Eμ-Myc p19Arf−/− cell was cultured in B-cell medium (45% Dulbecco's modified Eagle's medium and 45% Iscove's modified Dulbecco's media, supplemented with 10% fetal bovine serum (FBS), L-glutamate, and 5 μM β-mercaptoenthanol). Phoenix, HCT116 p53−/−, Saos-2, U251, A549, 293T, and 293A were cultured in Dulbecco's modified Eagle's medium supplemented with glutamate and 10% (v/v) FBS. 786-O cell was cultured in RPMI medium supplemented with glutamate and 10% (v/v) FBS.

Saos-2, HCT116 p53−/−, U251, 786-O cells were obtained from the Cell Bank, China Academy of Sciences (Shanghai, China). Doxorubicin was purchased from Selleck.

Antibodies

Antibodies against Phospho-Akt (Thr308) (D25E6) (Cell signaling, #13038), Akt (pan) (C67E7) (Cell signaling, #4691), Phospho-GSK-3β (Ser9) (D84E12) (Cell signaling, #5558), Phospho-PDK1 (Ser241) (C49H2) (Cell signaling, #3438), and PTEN (pan) (Y184) (Abcam, #32199) were used for Western blot analysis.

Cloning of p53, PTEN, and INK4A mutants

Wild-type p53, INK4a, and PTEN expression vectors were constructed as follows. The full-length open reading frame of p53, INK4a, and PTEN cDNAs were amplified by PCR using KOD plus neo DNA polymerase (Code No. KOD-401 Lot No. 646300) and a pair of primers with *EcoR*I and *Xho*I sites. The PCR product was cloned into the *EcoR*I/*Xho*I sites of the pMSCV-IRES-GFP vector. cDNAs with missense mutations were constructed by overlap extension PCR. All mutation constructs were sequenced to confirm that the appropriate mutations had been incorporated and that no additional mutations were generated.

All p53, INK4a, and PTEN mutants tested in this study are listed in Supplementary Table S6.

Expression of mutants in cells

To test the functional status of p53 mutants, retrovirus that expresses p53 mutants, puromycin resistance gene and GFP was used to infect Saos-2 cells or HCT116 p53−/− cells. Cells are infected with similar virus MOI such that 30–50% of cells are GFP positive for all experimental groups. We were able to determine the expression level of p53 mutants in HCT116 p53−/− cells since they can tolerate p53 mutants that retain wild-type gene function. The results showed that under such infection protocol, expression levels of different p53 mutants were comparable (Supplementary Fig. S9).

Colony formation assay

Forty-eight hour after infection, 5000 of GFP-positive Saos-2 cells were resuspended in medium containing 10% FBS and plated in 6-well plates. After 24 hours, they were treated with 2 μg/ml puromycin. Twenty-four hours later, puromycin-containing medium was replaced with fresh complete culture medium. Five days later, 2 μg/ml puromycin was again used to treat cells for 24 h before removal. Cells are cultured for an additional 10 days. Colonies were then fixed with 4% paraformaldehyde and stained with 0.1% crystal violet for 30 min. Stained cell colonies were washed with phosphate-buffered saline

(PBS) for three times and dried. Images were obtained by a digital camera. Similar protocols were used to test INK4a mutants in U251 cells.

Mouse liver cancer model

The mouse liver cancer model was performed using published protocol as published in Chen et al.[21]. Two microlitre of plasmid solutions were injected into tail vein of a mouse in about 7 s. This creates immediate high pressure at the liver portal vein, and plasmids will enter liver cells under such conditions. In this experiment, sgRNA targeting PTEN was used to inactivate PTEN to facilitate liver cancer development, similar to Yang et al.[22]. cDNAs of c-Myc and different p53 mutants were cloned into a sleeping beauty system, so that upon entering liver cells, they will be stably integrated into host cell chromosomes. In this experiment, p53 mutants were not generated by CRISPR, but were introduced by cDNA to avoid uncertainty with CRISPR mutagenesis.

Specifically, Myc cDNA and p53 mutants were cloned into a transposon system using the PT3 vector[21]. Such plasmids were mixed with sgPTEN–Cas9 plasmid[22], together with Sleeping Beauty transposase-expressing plasmid in PBS. Gene mixture was delivered to mouse by tail vain hydrodynamic injection. Concentrations of Sleeping Beauty transposase and Myc-expressing plasmids were at 0.5 and 1.25 μg/ml, respectively. Other plasmids or corresponding empty vectors were used at 5 μg/ml. The experiments were all done in female FVB mice at 7 weeks of age. The mice were hosted in SPF housing condition. The experimental was approved by the institutional animal care and use committee.

Cell cycle analysis

U251 cells expressing wild type or mutant forms of INK4A were analyzed. When grew to proper density (about 70–80%), cells were collected and fixed overnight in 70% ethanol. Cells were then treated with 0.2% Triton X-100, 50 μg/ml propidium iodide and 100 μg/ml RNase A for 40 min, then analyzed by FACS.

Quantitative real-time PCR assay

RNA was purified using GeneJET RNA Purification Kit (thermo scientific) and qPCR was performed on a StepOne real-time PCR machine (BIO-RAD) using SYBR Green PCR master mix (Promega). mRNA level of actin was used as control. Primers used for qPCR analysis are listed in Supplementary Table S8.

Statistics

Differences of event frequency between two groups were analyzed using Student's unpaired two-tailed *t* test. *p* Values < 0.01 were marked as *** in figures, *p* values < 0.05 were marked as ** in figures.

Acknowledgements

This work was supported by the major scientific research project (2017YFA0504503) from the Ministry of Science and Technology of China, "Strategic Priority Research Program" of the Chinese Academy of Sciences (XDB19000000), and National Natural Science Foundation of China (81972600). We thank Drs. Dangsheng Li, Zhaocai Zhou, Ning Gao, Jinqiu Zhou, and Yun Zhao for discussions and helpful comments, and Animal Core Facility and Core Facility for Cell Biology at SIBCB.

Author contributions

H.J. conceived the study and wrote the paper. J.Y. and L.S. performed the COSMIC data analysis with help from H.D., M.L., L.X., J.Z., Z.H., S.C., and A.Y. L.S. performed the studies of p53. J.Y. performed the studies of PTEN and INK4A. L. S. and J.Y. prepared the figures and analyzed the data. All authors discussed the results and commented on the paper. H.J. supervised the study.

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary Information accompanies the paper at (https://doi.org/10.1038/s41421-020-0177-8).

Received: 15 November 2019 Accepted: 6 May 2020
Published online: 21 July 2020

References

1. Kastenhuber, E. R. & Lowe, S. W. Putting p53 in Context. *Cell* **170**, 1062–1078 (2017).
2. Cheok, C. F., Verma, C. S., Baselga, J. & Lane, D. P. Translating p53 into the clinic. *Nat. Rev. Clin. Oncol.* **8**, 25–37 (2011).
3. Brosh, R. & Rotter, V. When mutants gain new powers: news from the mutant p53 field. *Nat. Rev. Cancer* **9**, 701–713 (2009).
4. Baugh, E. H., Ke, H., Levine, A. J., Bonneau, R. A. & Chan, C. S. Why are there hotspot mutations in the TP53 gene in human cancers? *Cell Death Differ.* **25**, 154–160 (2018).
5. Farazi, P. A. & DePinho, R. A. Hepatocellular carcinoma pathogenesis: from genes to environment. *Nat. Rev. Cancer* **6**, 674–687 (2006).
6. Alexandrov, L. B. et al. Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).
7. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
8. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836 (2019).
9. Boyle, J. Molecular biology of the cell, 5th edition by B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Biochem. Mol. Biol. Educ.* 296–297 (2008).
10. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15**, 585–598 (2014).
11. Frigola, J. et al. Reduced mutation rate in exons due to differential mismatch repair. *Nat. Genet.* **49**, 1684–1692 (2017).
12. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
13. Buisson, R. et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, eaaw2872 (2019).
14. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12**, 756–766 (2011).
15. Ma, L., Zhang, T., Huang, Z., Jiang, X. & Tao, S. Patterns of nucleotides that flank substitutions in human orthologous genes. *BMC Genomics* **11**, 416 (2010).
16. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
17. Schneider, G., Schmidt-Supprian, M., Rad, R. & Saur, D. Tissue-specific tumorigenesis: context matters. *Nat. Rev. Cancer* **17**, 239–253 (2017).

18. el-Deiry, W. S. et al. WAF1/CIP1 is induced in p53-mediated G1 arrest and apoptosis. *Cancer Res.* **54**, 1169–1174 (1994).
19. Willis, A., Jung, E. J., Wakefield, T. & Chen, X. Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. *Oncogene* **23**, 2330–2338 (2004).
20. Stott, F. J. et al. The alternative product from the human CDKN2A locus, p14 (ARF), participates in a regulatory feedback loop with p53 and MDM2. *EMBO J.* **17**, 5001–5014 (1998).
21. Chen, X. et al. Distinct pathways of genomic progression to benign and malignant tumors of the liver. *Proc. Natl Acad. Sci. USA* **104**, 14771–14776 (2007).
22. Yang, G. et al. CRISPR-mediated direct mutation of cancer genes in the mouse liver. *Nature* **514**, 380–384 (2014).
23. Follis, A. V. et al. The DNA-binding domain mediates both nuclear and cytosolic functions of p53. *Nat. Struct. Mol. Biol.* **21**, 535–543 (2014).
24. Eldar, A., Rozenberg, H., Diskin-Posner, Y., Rohs, R. & Shakked, Z. Structural studies of p53 inactivation by DNA-contact mutations and its rescue by suppressor mutations via alternative protein-DNA interactions. *Nucleic Acids Res.* **41**, 8748–8759 (2013).
25. Walker, D. R. et al. Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features. *Oncogene* **18**, 211–218 (1999).
26. Cossman, J. & Schlegel, R. P53 in the diagnosis of human neoplasia. *J. Natl Cancer Inst.* **83**, 980–981 (1991).
27. Soussi, T. & May, P. Structural aspects of the p53 protein in relation to gene evolution: a second look. *J. Mol. Biol.* **260**, 623–637 (1996).
28. Chao, C., Herr, D., Chun, J. & Xu, Y. Ser18 and 23 phosphorylation is required for p53-dependent apoptosis and tumor suppression. *EMBO J.* **25**, 2615–2622 (2006).
29. Wijnhoven, S. W. P. et al. Dominant-negative but not gain-of-function effects of a p53.R270H mutation in mouse epithelium tissue after DNA damage. *Cancer Res.* **67**, 4648–4656 (2007).
30. MacPherson, D. et al. Defective apoptosis and B-cell lymphomas in mice with p53 point mutation at Ser 23. *EMBO J.* **23**, 3689–3699 (2004).
31. Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N. M. & Gaunt, T. R. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* **29**, 1504–1510 (2013).
32. Giacomelli, A. O. et al. Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.* **50**, 1381–1387 (2018).
33. Cannataro, V. L., Gaffney, S. G. & Townsend, J. P. Effect sizes of somatic mutations in cancer. *J. Natl Cancer Inst.* **110**, 1171–1177 (2018).
34. Chan, K. et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.* **47**, 1067–1072 (2015).
35. Rideout, W. M., Coetzee, G. A., Olumi, A. F. & Jones, P. A. 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science* **249**, 1288–1290 (1990).
36. Walsh, C. P. & Xu, G. L. Cytosine methylation and DNA repair. *Curr. Top. Microbiol. Immunol.* **301**, 283–315 (2006).
37. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations,. cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (2017).
38. Dang, L., Jin, S. & Su, S. M. IDH mutations in glioma and acute myeloid leukemia. *Trends Mol. Med.* **16**, 387–397 (2010).
39. Borger, D. R. et al. Frequent mutation of isocitrate dehydrogenase (IDH)1 and IDH2 in cholangiocarcinoma identified through broad-based tumor genotyping. *Oncologist* **17**, 72–79 (2011).
40. Quelle, D. E., Cheng, M., Ashmun, R. A. & Sherr, C. J. Cancer-associated mutations at the INK4a locus cancel cell cycle arrest by p16INK4a but not by the alternative reading frame protein p19ARF. *Proc. Natl Acad. Sci. USA* **94**, 669–673 (1997).
41. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**, 1324–1335 (2014).
42. Bruno, P. M. et al. A subset of platinum-containing chemotherapeutic agents kills cells by inducing ribosome biogenesis stress. *Nat. Med.* **23**, 461–471 (2017).
43. Ding, H. et al. Systematic analysis of drug vulnerabilities conferred by tumor suppressor loss. *Cell Rep.* **27**, 3331–3344 (2019).