Bin *et al. BMC Medical Genetics* 2019, **20**(Suppl 2):190
https://doi.org/10.1186/s12881-019-0926-4

BMC Medical Genetics

RESEARCH Open Access

An analysis of mutational signatures of synonymous mutations across 15 cancer types

Yannan Bin[†], Xiaojuan Wang[†], Le Zhao, Pengbo Wen and Junfeng Xia[*]

From 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference
Wuhan and Shanghai, China. 15-18 August 2018, 3-4 November 2018

Abstract

Background: Synonymous mutations have been identified to play important roles in cancer development, although they do not modify the protein sequences. However, relatively little research has specifically delineated the functionality of synonymous mutations in cancer.

Results: We investigated the nucleotide-based and amino acid-based features of synonymous mutations across 15 cancer types from The Cancer Genome Atlas (TCGA), and revealed novel driver candidates by identifying hotspot mutations. Firstly, synonymous mutations were analyzed between TCGA and 1000 Genomes Project at nucleotide and amino acid levels. We found that C:G → T:A transitions were the most frequent single-base substitutions, and leucine underwent the largest number of synonymous mutations in TCGA due to prevalent C → T transition, which induced the transformation between optimal and non-optimal codons. Next, 97 synonymous hotspot mutations in 86 genes were nominated as candidate drivers with potential cancer risk by considering the mutational rates across different sequence contexts. We observed that non-CpG-island GC transition sequence context was positively selected across most of cancer types, and different sequence contexts under which hotspot mutations occur could be significance for genetic differences and functional features. We also found that the hotspots were more conserved than neutral mutations of hotspot-mutation-containing-genes and frequently happened at leucine. In addition, we mapped hotspots, neutral and non-hotspot mutations of hotspot-mutation-containing-genes to their respective protein domains and found ion transport domain was the most frequent one, which could mediate the cell interaction and had relevant implication for tumor therapy. And the signatures of synonymous hotspots were qualitatively similar with those of harmful missense variants.

Conclusions: We illustrated the preferences of cancer associated synonymous mutations, especially hotspots, and laid the groundwork for understanding the synonymous mutations act as drivers in cancer.

Keywords: Cancer, Synonymous mutations, Hotspot, Driver

* Correspondence: jfxia@ahu.edu.cn
[†]Yannan Bin and Xiaojuan Wang contributed equally to this work.
Institutes of Physical Science and Information Technology, School of Computer Science and Technology, Anhui University, Hefei 230601, Anhui, China

Background

Synonymous mutations, which occur in the gene-coding regions without changing the encoded amino acids, have long been supposed to be silent for the fitness of organisms and neutral during evolution [1]. However, this conservative concept begins to be rebutted by two lines of evidence: first one is the understanding of synonymous mutational effect on protein synthesis and folding; second, codon usage bias reveals that synonymous codons are under evolutionary pressure [2]. Because of the degeneracy of the genetic codons, synonymous mutations don't changing the encoded amino acids, but change the DNA and RNA sequence. Nevertheless, there are growing evidences that the significant impact of synonymous mutations on RNA splicing, stability and folding [3, 4], translation or co-translational protein folding [5–8]. Chen et al. conducted a broad survey of 21,429 disease-related single nucleotide polymorphisms (SNPs) to indicate that synonymous SNPs and non-synonymous SNPs showed similar probability and effect size for human diseases [9]. In addition, some studies have identified that synonymous mutations frequently act as driver mutations in human cancers [10, 11] and can affect clinical outcome or treatment response [12–14].

As a complex genetic disease, cancer was affected by a large number of variants. But to date, the targets of drugs and treatments associated with cancer are limited on a few genes, therefore, it is difficult to achieve cures for cancer. Next-generation sequencing technology has enabled the systemic analyses of huge variants in large cohorts of cancer cases, e.g., The Cancer Genome Atlas (TCGA) [15] and International Cancer Genome Consortium [16]. Cancer genomes not only contain cancer-causing driver mutations, but also many additional accumulated passenger mutations without direct relation to the tumor phenotype. It is a key step to identify driver mutations for understanding cancer biology and evolving targeted treatments. There were several methods focused on predicting driver mutations, such as E-Driver [17], MuSiC [18] and OncodriveCLUST [19]. Nevertheless, those studies mainly focused on the missense mutations and ignored the potential functions of synonymous mutations. Although Silent Variant Analyzer [20] is a tool for the annotation and prediction of pathogenic synonymous mutations, the small datasets for training and validation restrict its applicability. TCGA, using the latest sequencing and analysis methods to identify somatic variants across thousands of tumors, is found to meet the data needs in this work [15, 21]. Additionally, more recent studies have indicated that different substitution types, codon usage bias and hotspot mutational positions in base sequence could be associated with different biological processes and cancer types [2, 22, 23]. The hotspot mentioned in this work is not the hotspot in protein-protein interfaces [24], and is defined as the mutation that occurs significantly more frequently than the background frequency characterized by genes, cancer types and mutation subtypes.

In this study, we documented the full repertoire of cancer associated synonymous mutations, especially synonymous hotspot mutations, to investigate the mutational signatures in cancer. To acquire insight into the characters of pathogenic and neutral synonymous mutations between cancer and benign samples, the differences of synonymous mutations at nucleotide and amino acid levels (such as nucleotide substitutions, mutational positions of codon and distribution of amino acids at which synonymous mutations happened) were investigated between datasets in TCGA and the 1000 Genomes Project (1000G) (as neutral samples) [16]. And then, we nominated synonymous hotspot mutations as candidate drivers based on the mutational rates across different sequence contexts and investigated the features (such as conservation, distribution of amino acids and protein domains undergo mutations) of hotspots, neutral synonymous mutations and non-hotspots in the hotspot-mutation-containing-genes (HMCGs). For the comprehensiveness of analysis, this study not only highlights the nucleotide level preferences, but also amino acids level, and especially hotspot mutations. The observation could add perspective to understand cancer-related synonymous mutations. The procedure is illustrated in Additional file 1: Figure S1.

Material and methods

Synonymous mutation dataset

The cancer related synonymous mutations in TCGA were downloaded from COSMIC v79 (Catalogue of Somatic Mutations in Cancer) [25]. We got 373,434 cancer related synonymous mutations obtained from 5749 tumor samples across 15 types of cancer: breast cancer (BRCA), central nervous system tumor (CNST), cervical adenocarcinoma (CEAD), endometrial adenocarcinoma (ENAD), haematopoietic and lymphoid tumor (HLTU), kidney carcinoma (KICA), large intestine adenocarcinoma (INAD), liver carcinoma (LICA), lung adenocarcinoma (LUAD), ovarian carcinoma (OVCA), prostate adenocarcinoma (PRAD), skin cancer (SKCA), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA) and urinary tract carcinoma (UTCA).

The aim of 1000G is to discover variants with a frequency of occurrence > 1% in multiple human populations worldwide. In this study, 21,121 putatively benign synonymous mutations were derived from 1000G (the phase 3 version 5b, 20,130,502) and this dataset will be referred to as the neutral synonymous mutations dataset for the comparative analysis of cancer related synonymous mutations.

Statistical analyses

The majority of statistical analyses in this work were completed by using R (https://www.r-project.org/), e.g., the distributions of synonymous variants across different cancer types, nucleotide substitutions and amino acids. Other statistical analyses were performed by GraphPad Prism 5 (GraphPad Software). A p-value < 0.05 was considered statistically significant.

Hotspot mutations identification

Here we used the Hot-Driver package [26] to identify the hotspot mutations that are positively correlated with the number of mutations across all cancer samples for all 15 cancer types. The Hot-Driver suite assigns mutations to six subtypes: AT transition (ATts), AT transversion (ATtv), non-CpG-island GC transition (NC_GCts), non-CpG-island GC transversion (NC_GCtv), CpG-island GC transition (C_GCts) and CpG-island GC transversion (C_GCtv). Based on mutational significance of each mutation subtype on amino acid position, a statistical method combines the significance level of different mutation subtypes to calculate the overall p-values (Poisson distribution and Fisher's test). We reported hotspot mutations in amino acid position with adjusted p-values < 0.05 corrected by false discovery rate. Lastly, in this study, to avoid the bias of background number of passenger mutations, we only selected the hotspot mutations that predicted as the pathogenic mutations by Functional Analysis through Hidden Markov Models [25].

To further investigate the mutational signatures of hotspot mutations, neutral synonymous mutations of HMCGs in 1000G and the non-hotspot mutation of HMCGs in TCGA were used as control datasets. These three datasets were applied for the further functional analyses, for example, conservation, amino acids and protein domains under which mutations occurred.

Conservation comparison

The conservation of nucleotide sequence for each gene was assessed by rejected substitution (RS) score, computed by GERP++ [27]. In this work, RS scores were extracted from the nucleotide bases that belong to hotspot mutations, neutral synonymous mutations of HMCGs in 1000G and non-hotspot mutations of HMCGs in TCGA, respectively. Single-tailed unpaired t-test was used to test significantly difference between hotspot mutations, neutral synonymous mutations and non-hotspot mutations.

Protein domain annotation

We mapped the hotspot mutations to conserved protein domains obtained from Pfam-A (version 29.0), a database of protein domain families [28], and manually curated data were used in this work. Since genes that shared a common domain are more likely to share related functions, the important mutations in certain genes tend to cluster in close proximity within functional domains [18, 29, 30].

Results and discussion

Synonymous mutation distribution across cancer types

From TCGA, we obtained 373,434 synonymous mutations of 5749 tumor samples from 15 types of cancer. As seen in the upper panel of Fig. 1, with regard to synonymous mutation proportion (blue bar), SKCA is the largest and HLTU the smallest. For the proportion of tumor samples (yellow bar), BRCA is the most and UTCA the fewest. Moreover, the average numbers of variants per sample across 15 cancer types are different with each other (the lower panel of Fig. 1), and exhibit many more or many less synonymous mutations per sample than the average number of 15 cancer types. Notable among these outliers are SKCA, ENAD, STAD and LUAD, which contain more than 100 synonymous mutations per sample. These larger numbers of mutations reflect the participation of potential factors (ultraviolet light, hyperestrogenism, *Helicobacter pylori* infection and cigarette smoke, respectively) in the pathogenesis of these cancer types [22, 31–33]. Due to ultraviolet light and deamination processes, the majority of SKCA mutations are C:G → T:A transitions [22]. Additionally, it has been reported that the mutations occur at methylated CpG dinucleotide, majority of which are C:G → T:A transitions, would significantly cause human genetic diseases [34]. Studies have shown that nucleotide substitutions, including synonymous mutations, could be related to carcinogen exposures and DNA repair processes [35–37].

Synonymous mutational nucleotide substitutions

Distribution of 12 possible mutational nucleotide substitution patterns of synonymous mutations across the 15 cancer types was shown in Fig. 2. The greatest frequently occurring is C → T transitional substitution (with average proportions of 44.01% across 15 cancer types), which is possible to associate with the aberrant DNA methylation [38–40]. SKCA contains the largest proportion of C → T transitions than other cancer types, owing to the signatures of ultraviolet light exposure and deamination processes [22]. Among transversions, the C → A substitution is the most frequent one (6.18%). As a result of signature smoke exposure, LUAD has an increased C → A transversions [33]. At 5-methylcytosine in CpG dinucleotides, C:G → T:A transitions and C → A transversion are associated with the most common epigenetic modifications of DNA [34, 35]. Moreover, due to the overabundance of synonymous sites involved in CpG dinucleotides, the mutation rate in exons is 30~60% higher than that in the non-coding regions [41]. It is

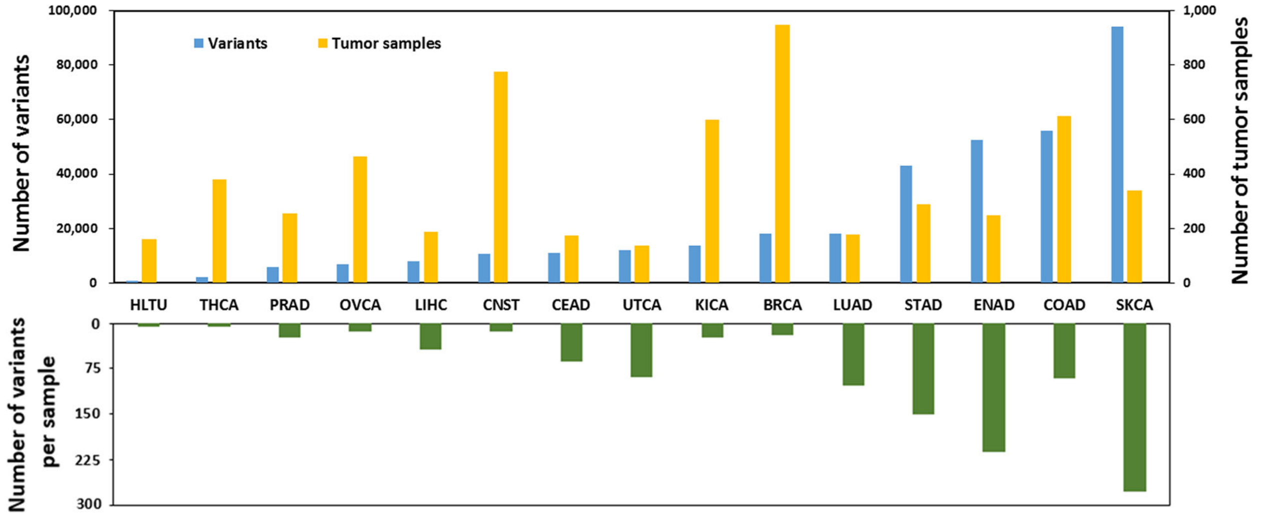

Fig. 1 Distribution of synonymous variants, tumor samples across different cancer types. Upper panel: blue and yellow bars represent the proportions of synonymous variants and corresponding tumor samples, respectively. Lower panel: Green bar represents the average numbers of variants per sample. The x-axis represents the cancer types

found that the percentage of transitions between C and T preceded that between A and G, which is known to be a general property of DNA sequence change and evolution [42]. Moreover, the most frequently substituted bases are C and G, and the most frequently mutated to bases are T and A. Cancer associated synonymous mutations have the tendency to become A/T-rich. Previous study has proposed that special A/T-rich sequence binding protein acts as a global chromatin organizer for metastatic activity by controlling gene expression [43].

Comparisons between TCGA and 1000G datasets

At nucleotide level, owing to the degeneracy of genetic codons, the nucleotide substitutions of synonymous codons occur at the third codon position (pos3), except some L and R codons (only T↔C transition and A↔C transversion) vary at the first codon position (pos1) (Additional file 1: Table S1). Similar with the distributions of mutational types in TCGA dataset, the most frequency mutational nucleotide changes in 1000G dataset are also C → T and G → A transitions (Fig. 3a).

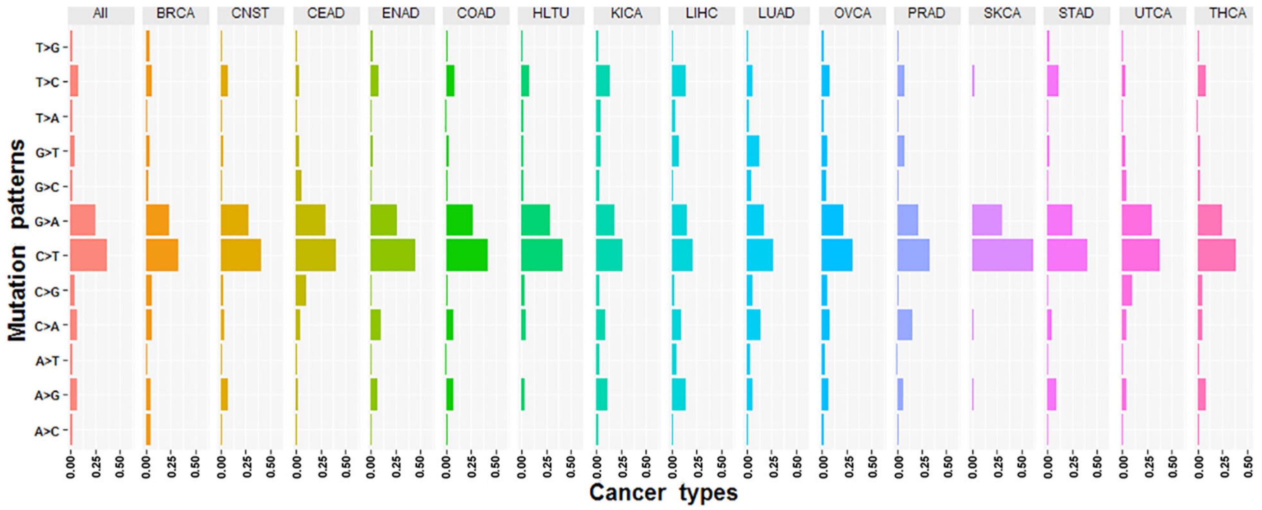

Fig. 2 Distribution of 12 possible mutation patterns of synonymous mutations in different cancer types. The x-axis represents the cancer types and the y-axis represents the 12 mutation patterns. Each bar represents the percentage of mutational types. 'All' in this figure represents the average proportions of mutational types across 15 cancer types

However, there are some differences for proportions of mutational nucleotide changes between TCGA and 1000G datasets. Firstly, we investigated the differences by performing a one-sample t-test, and the average proportion of each substitution in 1000G dataset was used as hypothetical value. A p-value < 0.05 is considered to be significant. The distributions of single-base mutational nucleotide changes in TCGA dataset are significant different from those in 1000G datasets (p-value < 0.001) except T → G transversion at pos3 (p-value = 0.8066). The non-significant different may be due to the less effect of T → G transversion on the transformation between efficient codon and lower efficient codon, which can affect protein production. Secondly, based on the significant differences, the single-base substitutions in TCGA dataset are apt to G:C → A:T transitions, in contrast, these substitutions in 1000G are T:A → C:G transitions. Synonymous mutations in TCGA are prone to become A/T-rich. In a previous study, it was reported that A/T-rich sequence could affect gene expression and be more important for cancer development [43].

Besides nucleotide level, the analysis of synonymous mutations was also performed at the amino acid level. Except Met (M) and Trp (W), all amino acids are encoded by two or more codons. The correlation between the percentages of mutation and the codon numbers of amino acids in TCGA ($r^2 = 0.67$) is stronger than that in 1000G ($r^2 = 0.54$) (Additional file 1: Figure S2). DNA sequences of diverse organisms have shown that synonymous codons of amino acids are used with unequal frequency [44]. The codon usage bias is related to various biological processes, such as gene expression level, protein structure, mutation frequency and GC composition [45]. However, it is probable that the uses of synonymous codons in TCGA tend to be at equal frequencies, and are less affected by codon usage bias than those in 1000G. Furthermore, the efficient codon replacing with a less efficient one could affect protein synthesis. Because the abundance of cognate tRNAs involved in preferred codons are available within the cell, the use of efficient codons could increase the gene expression [6]. Therefore, it is proposed that the cancer related synonymous mutations prefer to influence the gene expression and are more pathogenic than neutral ones in TCGA.

In TCGA and 1000G datasets, the mutational proportions of 18 amino acids with two, three, four or six-fold degenerate codons are different with each other (Fig. 3b). In 1000G, Ala is the most frequently mutated amino acid due to G → A transition. And this transition is associated with two of four Ala's codons, but independent of the transformation between optimal and non-optimal codons [10]. In TCGA, synonymous mutations are dominated by Leu (L) due to prevalent C:G → T:A transitions. It is similar with the character of pathogenic missense mutations, the substitutions under L are also the most frequent [46]. It is notable that among the three amino acids with six synonymous codons, Arg (R) shows the fewest number of mutations not only in TCGA but also in 1000G, which may be associated with the synonymous codon usage bias (R has only one optimal codon while Ser and L both have two optimal codons). In summary, it is possible that synonymous mutations under L may have more important effect on gene expression and protein production than the mutations of other amino acids during biological processes.

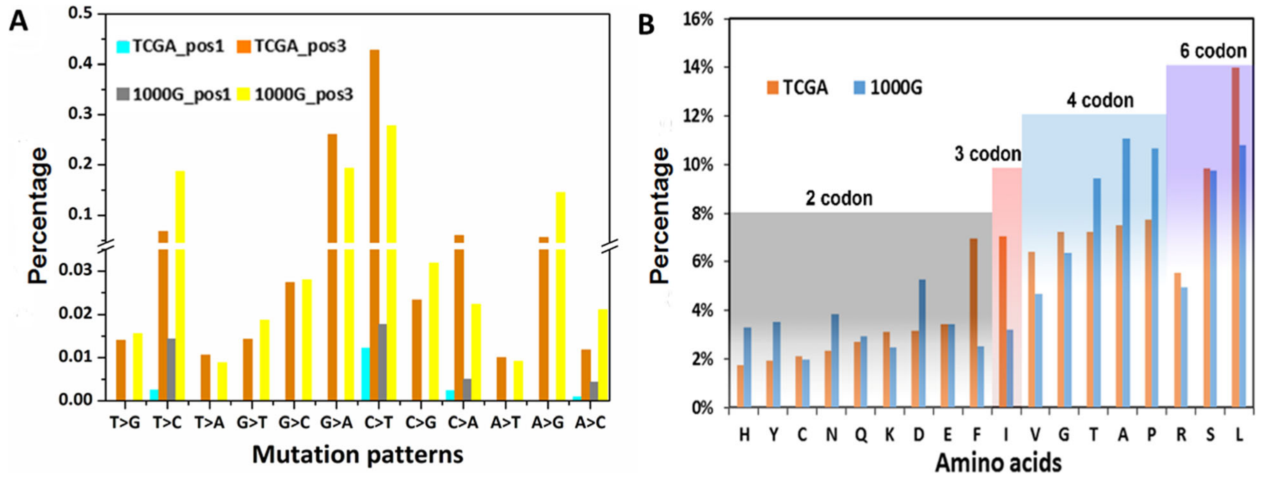

Fig. 3 Mutational signatures of synonymous mutations in TCGA and 1000G. **a** Distribution of synonymous mutations for the 12 mutation patterns at pos1 and pos3 of codons in TCGA and 1000G. **b** Proportions of 18 amino acids with two, three, four or six-fold degenerate codons under which synonymous mutations occurred. Orange bar represents the proportion of mutations in TCGA and grey bar represents that in 1000G. 2, 3, 4, and 6 codon represent the numbers of synonymous codons, respectively

Synonymous hotspot mutations for cancer

In this work, the hotspot mutation is defined as the mutation that occurs significantly more frequently than the background frequency characterized by genes, cancer types and mutation subtypes. In this study, we identified 97 hotspot mutations in 86 HMCGs associated with 14 cancer types (Additional file 1: Table S2). There is none hotspot mutation in HLTU due to the lowest mutation frequency (Fig. 1). To investigate the differences across cancers, we compared the number of synonymous hotspot mutations in different cancer types. From the distribution of hotspot mutations across cancers (Additional file 1: Figure S3A), it was found that the number of hotspot mutations varies largely from one cancer to another, many more or many fewer mutations than average. For example, INAD has the largest number of hotspots (39 hotspots), but ENAD and STAD only have 13 hotspots with the smallest number, and HLTU has none. The enrichment of hotspot mutations reflects the genetic heterogeneity of INAD, which has been discussed in previous research [47]. And heterogeneity may affect the progression of INAD from the early to the advanced stages, drive phenotypic variations and present a significant challenge to personalized medicine. By contrast, HLTU has none hotspot and THCA has only one hotspot mutation in *ILF3* (N192). In addition, to estimate the important synonymous mutations for pan-cancer, the distribution of hotspot mutations across different genes was analyzed (Additional file 1: Figure S3B). T125 in *TP53* is the most prevalently occurred mutation in nine different cancer types. It has been identified to be pathogenic for its detrimental role in *TP53* splicing [48]. It is also found that most hotspots are unique for only one cancer type. The common and diverse mutational signatures of hotspots across different cancer types may promote the understanding of the positive selection in the human genome, and facilitate the cancer target therapy [26].

The mutational characters of 97 hotspot mutations across 14 cancer types were also investigated under different sequence contexts, including ATts, ATtv, C_GCts, C_GCtv, NC_GCts and NC_GCtv six subtypes. From the distribution of hotspot mutations under different subtypes of sequence contexts (Fig. 4a), the number of hotspot mutations under NC_GCts sequence context (consists of C:G → T:A transitions) is the largest one compared with the other types, and that under ATtv (consists of A:T → C:G and A↔T transversions) is the least one. This phenomenon is due to the most frequency of C → T and G → A transitions and least frequency of A:T → C:G and A↔T transversions (Fig. 3a), which correspond to the NC_GCts and ATtv sequence contexts, respectively. As seen in Fig. 4b, the most widespread sequence context undergoes hotspots is NC_GCts sequence context, which presents in 14 cancer types. And it is also the most prevalent sequence context in nine cancer types (CEAD, CNST, ENAD, INAD, KICA, PRAD, SKCA, STAD and THCA). Moreover, in CEAD and THCA, the hotspots are only enriched in NC_GCts. However, in LUAD, the hotspot mutations are enriched in NC_GCtv sequence context. In LICA, OVCA and UTCA, the sequence contexts under which hotspot mutations occur are equal. In summary, NC_GCts sequence context is positively selected across most cancer types, and different sequence contexts on which hotspots happen are significant for considering their genetic differences and functional features.

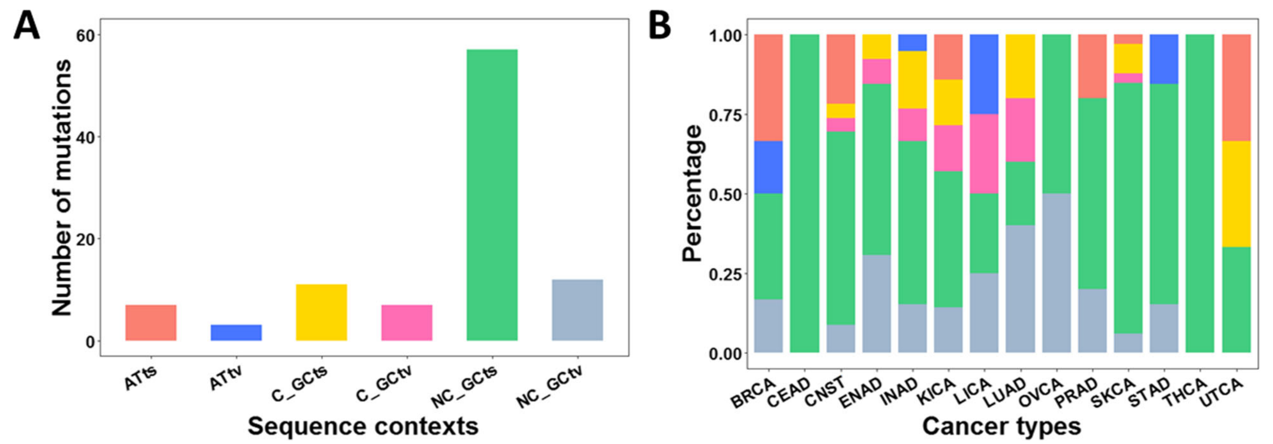

Fig. 4 Mutational signatures of synonymous hotspot mutation in 15 cancer typers. **a** Histogram of the number of hotspot mutations under different sequence contexts. **b** Bar diagram where the value 100% represents the total number of mutations under different sequence contexts distributed in each cancer type. ATts, AT transition (coral bar in **a** and **b**); ATtv, AT transversion (blue bar in **a** and **b**); C_GCts, CpG-island GC transition (yellow bar in **a** and **b**); C_GCtv, CpG-island GC transversion (rose bar in **a** and **b**); NC_GCts, non-CpG-island GC transition (green bar in **a** and **b**); NC_GCtv, non-CpG-island GC transversion (grey bar in **a** and **b**)

Conservation comparison

It is customary for mutations with important functional and evolutionary implications located in highly conservative region and protein domains. To evaluate the conservation of hotspot mutations, neutral synonymous mutations of 86 HMCGs in 1000G (235 neutral synonymous mutations were shown in Additional file 2: Table S3) and non-hotspot mutations of HMCGs in TCGA (1358 non-hotspot mutations were shown in Additional file 3: Table S4), we computed their RS scores to estimate the evolutionary constraints across different genome sites. As shown in Fig. 5, the RS scores of hotspot mutations are significantly higher than those of neutral synonymous mutations in 1000G (p-value < 2e-16). In contrast, there is no significantly different between hotspot and non-hotspot mutations (p-value = 0.93), it is possible that the non-hotspot in TCGA might influence cancer processes, but their harmfulness is less than that of hotspots. The result suggests that the sites which hotspots occur on are more conservative than those of neutral synonymous mutations. As potential driver mutations, these hotspots may be more important for cancer development.

Amino acids analysis

To further investigate the difference among hotspot mutations, neutral synonymous mutations and non-hotspot mutations of HMCGs, the distributions of amino acids under which the mutations occurred were investigated (Fig. 6). There is none synonymous mutation under Met and Trp due to the lack of synonymous codons. Clearly, the distributions of amino acids are different for the three synonymous mutation datasets. In hotspot mutation dataset, L and Phe (12.37 and 11.34%, respectively) are the most mutated amino acids due to prevalent C → T transition, which is involved the transformation between optimal and non-optimal codons. As an important amino acid of leucine-rich repeats, L is associated with a versatile structural framework for the formation of protein-protein interactions [49]. However, for neutral mutation dataset, the most frequent substitutions are under Ala (14.44%), which is largely tolerated outside functional site as the smallest residues can be fitted into structures easily. Among non-hotspot mutations, the most frequent substitutions are under L (14.51%). The distributions of the hotspot and neutral synonymous mutations are quantitatively similar to those of missense mutations published previously [46].

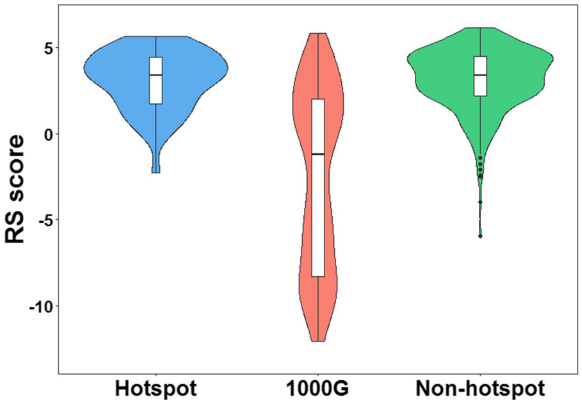

Fig. 5 Comparison of RS scores between the datasets of hotspots and neutral synonymous mutations, non-hotspots. Hotspot represents hotspot mutations; 1000G represents neutral synonymous mutations of HMCGs in 1000G dataset; Non-hotspot represents non-hotspot mutations of HMCGs in TCGA dataset. The RS score of hotspot mutation dataset is significant higher than that neutral synonymous mutation dataset (p-value <2e-16). But there is no significant different between hotspot and non-hotspot datasets (p-value = 0.93) for RS score

Domain characterization

We also investigated the domain compositions of the proteins, under which hotspot mutations, neutral synonymous mutations of HMCGs in 1000G and non-hotspot mutations of HMCGs in TCGA occurred. Thirty-five different Pfam domains were detected in the proteins under which the hotspot mutations occurred, whereas 29 and 91 protein domains under which neutral synonymous and non-hotspot mutations happened, respectively. It is found that nine domain types are common to hotspot, neutral synonymous and non-hotspot mutations (Fig. 7a), including 7 transmembrane receptor domain (7tm_1, PF00001), cytidine and deoxycytidylate deaminase zinc-binding region domain (dCMP_cyt_deam_1 domain, PF00383), nucleotide-binding domain (cobW domain, PF02492), Hsp70 protein domain (HSP70 domain, PF00012), ion transport protein domain (Ion_trans domain, PF00520), immunoglobulin I-set domain (I-set domain, PF07679), laminin N-terminal domain (Laminin_N domain, PF00055), membrane-bound O-acyltransferase family domain (MBOAT domain, PF03062) and transmembrane protein 67 domain (Meckelin domain, PF09773). The functions of these domains are different from each other. As rhodopsin-like receptor, 7tm_1 domain comprises the group of G protein-coupled receptor and encompasses a wide range of functions such as various autocrine, paracrine, and endocrine processes. dCMP_cyt_deam_1 domain is the cytidine and deoxycytidylate deaminase zinc-binding region, which is associated with the catalytic activity of cytidine deaminase. cobW domain contains a nucleotide-binding loop and a histidine-rich region that plays an important role in metal binding. HSP70 domain is strongly upregulated by heat stress and toxic chemicals, particularly heavy metals. Ion_trans domain contains sodium, potassium and calcium ion channels, and

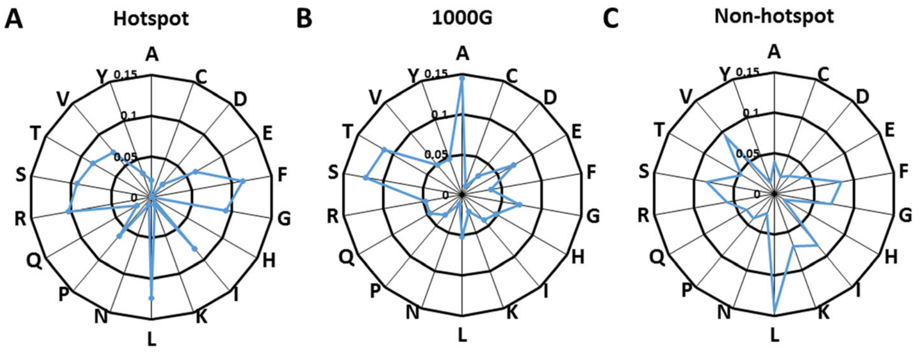

Fig. 6 Radar diagrams showing the distributions of amino acid at which mutations occurred. **a** Hotspot mutations; **b** Neutral synonymous mutations of HMCGs in 1000G dataset; **c** Non-hotspot mutations of HMCGs in TCGA dataset

a loop flanked by two helices determines ion selectivity. I-set domain is not only frequent in cell adhesion protein, but also appears in many other types of proteins [50]. Laminin_N domain is extracellular matrix molecule and MBOAT domain contains various acyltransferase enzymes. Then we analyzed the distributions of hotspot, neutral synonymous and non-hotspot mutations under the nine common protein domains. Among these nine domains, the Ion_trans domain is the most frequent one (28 items), but the proportion of hotspots is less than those of neutral synonymous mutations and non-hotspots (Fig. 7b). As an important target for tumor therapy [51], Ion_trans domain is critical for cell-to-cell communication and regulates multiple biological processes. However, the analysis of synonymous mutation distribution in Ion_trans domain is opposite with the previous analysis of missense mutations [46], which may be due to the different pathogenic mechanisms of synonymous and non-synonymous mutations. 7tm_1 domain is the highest proportion of hotspot mutations and consists of the group of G protein-coupled receptor, which could promote cancer metastasis [52].

Conclusions

In this study, we not only investigated the distribution and mutational nucleotide changes of synonymous mutations across 15 cancer types, but also made the comparison of synonymous mutational signatures between TCGA and 1000G at nucleotide and amino acid levels. Meanwhile, we nominated 97 hotspot mutations in 86

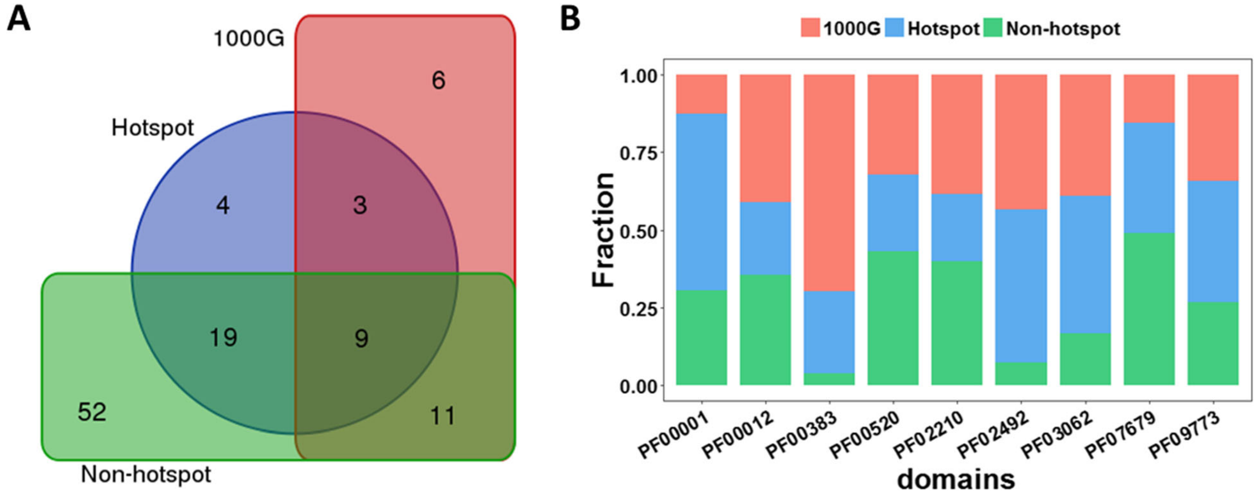

Fig. 7 Analysis of mutation distribution in protein domains. **a** Venn diagram showing the number of domain types for the multiple intersections among hotspot mutations, neutral synonymous mutations and non-hotspot mutations. **b** Bar diagram where the value 100% represents the total number of mutations in the nine protein domains common to hotspot, neutral synonymous and non-hotspot mutations. Blue, salmon and green sections on the bar represent which proportions of the total mutations are hotspot, neutral synonymous and non-hotspot mutations, respectively. PF00001, 7tm_1 domain; PF00383, dCMP_cyt_deam_1 domain; PF02492, cobW domain; PF00012, HSP70 domain; PF00520, Ion_trans domain; PF07679, I-set domain; PF00055, Laminin_N domain; PF03062, MBOAT domain; PF09773, Meckelin domain

genes in TCGA as potential drivers by considering their mutational rates across different mutational subtypes. And the common and diverse mutational signatures among hotspot mutations, neutral synonymous mutations of HCMGs in 1000G and non-hotspot mutations of HCMGs in TCGA were also detected. The result indicated that there were significant differences in conservation, amino acids and domain characterization between hotspots and neutral synonymous mutations. But there are some limitations in this study. Firstly, it needs more experimental work to investigate the effects of these hotspots on protein folding, RNA splicing, stability and folding, and whether they are drivers in cancers, and the relationship with cancer clinical outcome or treatment response. Secondly, the consistent patterns and specificity of hotspots in individual cancer are important and should be explored. But in this study, we just performed a pan-cancer analyzed of the hotspots. We will attack these problems in our further work. In summary, the present study would help to better understand the function of synonymous mutations in different cancer types and depicting their roles in carcinogenesis.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12881-019-0926-4.

Additional file 1: Figure S1. Illustration of analysis procedure of cancer associated synonymous mutations. **Figure S2.** Correlation between percentages of synonymous mutations and codon numbers of amino acids in TCGA and 1000G. **Figure S3.** Distribution of synonymous hotspot mutations across cancer types and genes. **Table S1.** Synonymous codons of amino acids with optimal and non-optimal codons for human genome. **Table S2.** Hotspot synonymous mutations across different cancer types in TCGA dataset.

Additional file 2: Table S3. List of neutral synonymous mutations of HMCGs in 1000G dataset.

Additional file 3: Table S4. List of non-hotspot synonymous mutations of HMCGs in TCAGA dataset.

Abbreviations

1000G: 1000 Genomes Project; ATts: AT transition; ATtv: AT transversion; BRCA: Breast cancer; C_GCts: CpG-island GC transition; C_GCtv: CpG-island GC transversion; CEAD: Cervical adenocarcinoma; CNST: Central nervous system tumor; ENAD: Endometrial adenocarcinoma; HLTU: Haematopoietic and lymphoid tumor; HMCGs: Hotspot-mutation-containing-genes; INAD: Large intestine adenocarcinoma; KICA: Kidney carcinoma; LICA: Liver carcinoma; LUAD: Lung adenocarcinoma; NC_GCts: Non-CpG-island GC transition; NC_GCtv: Non-CpG-island GC transversion; OVAD: Ovarian carcinoma; pos1: The first codon position; pos3: The third codon position; PRAD: Prostate adenocarcinoma; RS: Rejected substitution; SKCA: Skin cancer; SNPs: Single nucleotide polymorphisms; STAD: Stomach adenocarcinoma; TCGA: The Cancer Genome Atlas; THCA: Thyroid carcinoma; UTCA: Urinary tract

Acknowledgements

The authors thank all members of our laboratory for their valuable discussions.

About this supplement

This article has been published as part of *BMC Medical Genetics Volume 20 Supplement 2, 2019: Proceedings of the 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference: medical genetics*. The full contents of the supplement are available online at https://bmcmedgenet.biomedcentral.com/articles/supplements/volume-20-supplement-2.

Authors' contributions

YB performed the analysis and drafted the manuscript. XW, LZ and PW helped perform the analysis. JX designed the study, performed the analysis, and drafted the manuscript. All authors read and approved the final manuscript.

Funding

Publication costs are funded by the grants from the National Natural Science Foundation of China (61672037 and 21601001), the Anhui Provincial Outstanding Young Talent Support Plan (gxyqZD2017005), the Young Wanjiang Scholar Program of Anhui Province, the China Postdoctoral Science Foundation Grant (2018 M630699) and the Anhui Provincial Postdoctoral Science Foundation Grant (2017B325).

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Published: 9 December 2019

References

1. Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF. DNA sequence evolution: the sounds of silence. Philos T R Soc B. 1995;349:241–7.
2. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet. 2011;12:683–91.
3. Parmley JL, Chamary J, Hurst LD. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol. 2005;23:301–9.
4. Chamary J, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet. 2006;7:98–108.
5. Nackley AG, Shabalina S, Tchivileva I, Satterfield K, Korchynskyi O, Makarov S, Maixner W, Diatchenko L. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science. 2006;314:1930–3.
6. Gustafsson C, Govindarajan S, Minshull J. Codon bias and heterologous protein expression. Trends Biotechnol. 2004;22:346–53.
7. Soussi T, Taschner PE, Samuels Y. Synonymous somatic variants in human cancer are not infamous: a plea for full disclosure in databases and publications. Hum Mutat. 2017;38:339–42.
8. Diederichs S, Bartsch L, Berkmann JC, Frose K, Heitmann J, Hoppe C, Iggena D, Jazmati D, Karschnia P, Linsenmeier M, et al. The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. EMBO Mol Med. 2016;8:442–57.
9. Chen R, Davydov EV, Sirota M, Butte AJ. Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. PLoS One. 2010;5:e13574.
10. Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. Cell. 2014;156:1324–35.
11. Gartner JJ, Parker SCJ, Prickett TD, Dutton-Regester K, Stitzel ML, Lin JC, Davis S, Simhadri VL, Jha S, Katagiri N, et al. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. P Natl Acad Sci Usa. 2013;110:13481–6.
12. Ma F, Sun T, Shi Y, Yu D, Tan W, Yang M, Wu C, Chu D, Sun Y, Xu B, et al. Polymorphisms of EGFR predict clinical outcome in advanced non-small-cell

lung cancer patients treated with Gefitinib. Lung Cancer-j Iaslc. 2009;66:114–9.

13. Griseri P, Bourcier C, Hieblot C, Essafi-Benkhadir K, Chamorey E, Touriol C, Pages G. A synonymous polymorphism of the Tristetraprolin (TTP) gene, an AU-rich mRNA-binding protein, affects translation efficiency and response to Herceptin treatment in breast cancer patients. Hum Mol Genet. 2011;20:4556–68.
14. Schutz FA, Pomerantz MM, Gray KP, Atkins MB, Rosenberg JE, Hirsch MS, McDermott DF, Lampron ME, Lee GS, Signoretti S, et al. Single nucleotide polymorphisms and risk of recurrence of renal-cell carcinoma: a cohort study. Lancet Oncol. 2013;14:81–7.
15. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455:1061–8.
16. Genomes Project Consortium. A map of human genome variation from population scale sequencing. Nature. 2010;467:1061–73.
17. Porta-Pardo E, Garcia-Alonso L, Hrabe T, Dopazo J, Godzik A. A pan-cancer catalogue of cancer driver protein interaction interfaces. PLoS Comput Biol. 2015;11:e1004518.
18. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012;22:1589–98.
19. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods. 2013;10:1081–2.
20. Buske OJ, Manickaraj A, Mital S, Ray PN, Brudno M. Identification of deleterious synonymous variants in human genomes. Bioinformatics. 2013;29:1843–50.
21. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502:333–9.
22. Sanchez MI, Grichnik JM. Melanoma's high C>T mutation rate: is deamination playing a role? Exp Dermatol. 2014;23:551–2.
23. Cheung LWT, Yu S, Zhang D, Li J, Ng PKS, Panupinthu N, Mitra S, Ju Z, Yu Q, Liang H, et al. Naturally occurring neomorphic PIK3R1 mutations activate the MAPK pathway, dictating therapeutic response to MAPK pathway inhibitors. Cancer Cell. 2014;26:479–94.
24. Xia JF, Zhao XM, Song J, Huang DS. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinformatics. 2010;11:174.
25. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2014;43:D805–D11.
26. Chen T, Wang Z, Zhou W, Chong Z, Meric-Bernstam F, Mills GB, Chen K. Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types. BMC Genomics. 2016;17:249–62.
27. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6:e1001025.
28. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44:D279–D85.
29. Deng SP, Huang DS. SFAPS: an R package for structure/function analysis of protein sequences based on informational spectrum method. Methods. 2014;69:207–12.
30. Huang D-S, Zhao X-M, Huang G-B, Cheung Y-M. Classifying protein sequences using hydropathy blocks. Pattern Recogn. 2006;39:2293–300.
31. Zanetta GM, Webb MJ, Li H, Keeney GL. Hyperestrogenism: a relevant risk factor for the development of cancer from endometriosis. Gynecol Oncol. 2000;79:18–22.
32. Parsonnet J, Friedman GD, Orentreich N, Vogelman H. Risk for gastric cancer in people with CagA positive or CagA negative helicobacter pylori infection. Gut. 1997;40:297–301.
33. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455:1069–75.
34. Cooper DN, Youssoufian H. The CpG dinucleotide and human genetic disease. Hum Genet. 1988;78:151–5.
35. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene. 2002;21:7435.
36. Pfeifer GP, You YH, Besaratinia A. Mutations induced by ultraviolet light. Mutat Res. 2005;571:19–31.
37. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149:979–93.
38. Tao MH, Freudenheim JL. DNA methylation in endometrial cancer. Epigenetics. 2010;5:491–8.
39. Etcheverry A, Aubry M, de Tayrac M, Vauleon E, Boniface R, Guenot F, Saikali S, Hamlat A, Riffaud L, Menei P, et al. DNA methylation in glioblastoma: impact on gene expression and clinical outcome. BMC Genomics. 2010;11:701.
40. Deng SP, Cao S, Huang DS, Wang YP. Identifying stages of kidney renal cell carcinoma by combining gene expression and DNA methylation data. IEEE/ACM Trans Comput Biol Bioinform. 2017;14:1147–53.
41. Subramanian S, Kumar S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. 2003;13:838–44.
42. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993;10:512–26.
43. Li Q, Chen Z, Xu J, Cao X, Chen Q, Liu X, Xu Z. Overexpression and involvement of special AT-rich sequence binding protein 1 in multidrug resistance in human breast carcinoma cells. Cancer Sci. 2010;101:80–6.
44. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011;12:32–42.
45. Wan X, Xu D, Kleinhofs A, Zhou J. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. BMC Evol Biol. 2004;4:1–11.
46. Schaafsma GC, Vihinen M. Large differences in proportions of harmful and benign amino acid substitutions between proteins and diseases. Hum Mutat. 2017;38:839–48.
47. Losi L, Baisse B, Bouzourene H, Benhattar J. Evolution of intratumoral genetic heterogeneity during colorectal cancer progression. Carcinogenesis. 2005;26:916–22.
48. Soussi T. Locuss-pecific databases in cancer: what future in a post-genomic era? The TP53 LSDB paradigm. Hum Mutat. 2014;35:643–53.
49. Kobe B, Kajava AV. The leucine-rich repeat as a protein recognition motif. Curr Opin Struc Biol. 2001;11:725–32.
50. Freigang J, Proba K, Leder L, Diederichs K, Sonderegger P, Welte W. The crystal structure of the ligand binding module of axonin-1/TAG-1 suggests a zipper mechanism for neural cell adhesion. Cell. 2000;101:425–33.
51. Becchetti A, Munaron L, Arcangeli A. The role of ion channels and transporters in cell proliferation and cancer. Front Physiol. 2013;4:312.
52. Tang X, Jin R, Qu G, Wang X, Li Z, Yuan Z, Zhao C, Siwko S, Shi T, Wang P, et al. GPR116, an adhesion G-protein-coupled receptor, promotes breast cancer metastasis via the Galphaq-p63RhoGEF-rho GTPase pathway. Cancer Res. 2013;73:6206–18.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.