Biochemistry and Biophysics Reports 37 (2024) 101597

Contents lists available at ScienceDirect

Biochemistry and Biophysics Reports

journal homepage: www.elsevier.com/locate/bbrep

Prognostic analysis of mutated genes and insight into effects of DNA damage and repair on mutational strand asymmetries in gastric cancer

Yangyang Shen [a,b], Kai Shi [a], Dongfeng Li [a], Qiang Wang [c,**], Kangkang Wu [d,***], Chungang Feng [a,*]

[a] College of Animal Science and Technology, Nanjing Agricultural University, Nanjing, China
[b] Institute of Animal Science, Jiangsu Academy of Agriculture Science, Nanjing, China
[c] Department of Urology, Peking University People's Hospital, Beijing, China
[d] Department of Infectious Disease, Children's Hospital of Nanjing Medical University, Nanjing, China

ARTICLE INFO

Keywords:
Gastric cancer
Survival analysis
Strand asymmetry
DNA damage and repair
Non-coding regulatory elements

ABSTRACT

Gastric cancer (GACA) is a complex and multifaceted disease influenced by a variety of environmental and genetic factors. Somatic mutations play a major role in its development, and their characteristics, including the asymmetry between two DNA strands, are of great interest and appear as a signal of information and guidance, revealing mechanisms of DNA damage and repair. Here, we analyzed the impact of High-frequency mutated genes on patient prognosis and found that the effect of expression levels of tumor protein p53 (TP53) and lysine methyltransferase 2C (KMT2C) genes remained high throughout the development of GACA, with similar expression patterns. After investigating mutation asymmetry across mutagenic processes, we found that transcriptional asymmetry was dominated by T > G mutations under the influence of transcription couples repair and damage. The apolipoprotein B mRNA editing enzyme catalytic polypeptide like (APOBEC) enzyme that induces mutations during DNA replication has been identified here and we identified a replicative asymmetry, which was dominated by C > A mutations in left-replicating. Strand bias in different mutation classes at transcription factor binding sites and enhancer regions were also confirmed, which implies the important role of non-coding regulatory elements in the occurrence of mutations. This work systematically describes mutational strand asymmetries in specific genomic regions, shedding light on the DNA damage and repair mechanisms underlying somatic mutations in cohorts of GACA patients with gastric cancer.

1. Introduction

Gastric cancer is one of the most common malignant tumors in China and is also a highly prevalent malignant tumor throughout Asia [1,2]. In 2020, there were about 1.09 million new cases of gastric cancer worldwide, which ranked 5th in the incidence of malignant tumors [3]. The number of deaths from stomach cancer was about 769,000, which ranked fourth in the number of deaths from malignant tumors; 43.9 % of the cases and 48.6 % of the deaths occurred in China [4]. The global incidence of gastric cancer varies greatly geographically, with a 15 to 20-fold difference between high- and low-prevalence areas. The regions with the highest incidence of gastric cancer are Northeast Asia, South and Central America, and Eastern Europe. In Northeast Asia, gastric cancer is one of the most diagnosed cancers among Japanese and Korean men [5,6]. Most patients are already in the progressive stage of gastric cancer when diagnosed, which makes it an important public health problem [7,8].

It is well known that cancer is caused by mutations. Cancer occurs due to the accumulation of a large number of genetic mutations in somatic cells over a long period of time, and these genetic mutations encourage the formation of cancer [9]. The somatic mutant spectrum in humans suggests that tissue-specific mutant characteristics and co-associations play an important role in human aging and disease studies. Over the last decade, Next generation sequencing (NGS) as enabled application of clinical genomics to the diagnosis and treatment of cancers. Cai et al. [10] drew a comprehensive mutational landscape of

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: wq301135@163.com (Q. Wang), wkkanhui@163.com (K. Wu), fengchungang@njau.edu.cn (C. Feng).

https://doi.org/10.1016/j.bbrep.2023.101597
Received 9 August 2023; Received in revised form 26 November 2023; Accepted 27 November 2023
Available online 7 December 2023

153 gastric tumors and demonstrated utility of massively parallel DNA sequencing of tumors to guide clinical management. Pan et al. [11] detected somatic mutation profiles of 45 cases of gastric cancer by whole exon sequencing, which suggested that MLL4, ERBB3, FBXW7, MLL3, MTOR, NOTCH1, PIK3CA, KRAS, ERBB4, and EGFR were mutated in gastric cancer. The DNA repair gene TP53 was the gene most frequently mutated, and patients with TP53 mutations had a significantly higher number of mutations, which indicated that the TP53 gene may serve a crucial role in the maintenance of genome integrity and stability.

A comprehensive understanding of mutation density and patterns in cancer genomes is very important for the study of mutation mechanisms, the establishment of models for the evolution of cancer genomes, and the identification of cancer genes [12,13]. Pan-Cancer Analysis of Whole Genomes (PCAWG), which is the most comprehensive atlas of cancer genomes to date, contains sequences and analyses of the genomes of 2658 primary cancers and their matching normal tissue samples from 38 different tumor types. This atlas revealed the extensive role of large-scale mutations in cancer, identified previously unknown cancer-associated mutations in gene regulatory regions, and elucidated the interaction between somatic mutations and the transcriptome [14]. Mutations leave a mutational signature on the DNA strand. In the cancer genome, somatic mutations show heterogeneity in terms of mutation density along the genome and mutation spectrum among cancer types [15,16]. Thus far, about 30 different mutation characteristics have been reported, each with a corresponding inducing factor. For example, in lung cancer, smoking-related damage is mainly G:C>T:A [17]. In skin cancer, C:G>T:A transitions are mainly associated with ultraviolet (UV) radiation exposure [18].

Mutations result from DNA damage and unsuccessful repair processes. DNA damage encountered on the transcription ("template") strand prevents the progression of RNA polymerase, which leads to the recruitment of nucleotide excise repair (NER) complexes that correct the damage [19–21]. In addition to the background activity of these processes, the distribution of mutations is influenced by genomic, epigenomic, and cellular physiological factors, such as replication and transcription. Single nucleotide substitutions occur at different rates on both DNA strands due to inherent asymmetries in the replication and transcription processes. Mutational strand asymmetries in cancer genomes, which appear as a signal of information and guidance, reveal mechanisms of DNA damage and repair.

Haradhvala et al. [22] revealed widespread asymmetries across mutagenic processes, where transcriptional (“T-class”) asymmetry dominated UV-, smoking-, and liver-cancer-associated mutations, and replicative (“R-class”) asymmetry dominated POLE-, APOBEC-, and MSI-associated mutations; this was based on the analysis of whole-genome sequences of 590 tumors from 14 different cancer types. They reported remarkable phenomena of transcription coupled damage (TCD) on non-transcribed DNA strands and transcription coupled repair (TCR) on transcribed strands, and provided evidence that APOBEC mutagenesis occurred on the lagging-strand template during DNA replication. Strand asymmetry has been well studied in the context of transcription. As the most representative example of transcriptional strand bias in somatic mutations in several cancers, skin cancer has higher C > T mutations in the non-transcriptional strand than in the transcriptional strand [23]. Lung cancer has higher G > T mutations in the non-transcriptional strand than in the transcriptional strand [24]. Liver cancer has higher A > G mutations in the non-transcriptional strand than in the transcriptional strand [25].

While mutational strand asymmetries in human liver, skin, and lung cancers had been extensively studied, gastric cancer, in comparison, remained relatively under-explored. Given the high prevalence of gastric cancer in Asian populations, understanding its unique mutational characteristics was of paramount importance. To address this gap, our study delved into the mutational strand asymmetries within the gastric cancer genome, focusing particularly on Chinese patients. By harnessing somatic mutation data from the International Cancer Genome Consortium database, we aimed to provide a comprehensive analysis of mutation density patterns around specific genomic regions, including Transcription Factor Binding Sites (TFBS) and Enhancers. Notably, we placed significant emphasis on Transcription Start Sites (TSS) and replication origin Initiation Sites (IS). Additionally, we conducted a comparative analysis using data from Japanese patients, seeking both commonalities and distinctions in mutational strand asymmetries. Through this endeavor, we aspired to shed light on the unique mutational profile of gastric cancer, offering valuable insights into its underlying mechanisms, particularly within the context of Asian populations.

2. Materials and methods

2.1. Data collection and processing

Data from GACA-CN (Gastric Cancer-China) and GACA-JP (Gastric Cancer-Japan) were downloaded from the International Cancer Genome Consortium (ICGC) portal through the “ICGC DCC DATA RELEASES” website (https://dcc.icgc.org/releases/current/Projects/GACA-CN). In GACA-CN, the total number of donors is 145, containing 42 from PCAWG and 123 from DCC. In GACA-CN, the total number of donors is 585 from DCC. Subsequently, we conducted a meticulous screening based on chromosomal count and mutation type, ultimately selecting 22 autosomal single nucleotide variants (SNVs) for in-depth analysis. All computations and assessments were conducted utilizing the hg19 human genome assembly. It is important to note that the mutations in this study refer to a collection of all potential pathogenicity or uncertainty in terms of clinical significance, including possible pathogenicity, pathogenicity, pathogenicity/probable pathogenicity, conflicting interpretations of pathogenicity and uncertain significance.

2.2. Waterfall plot of somatic mutations

Following the annotation process, the R package GenVisR (1.34.0) is utilized to create the waterfall plot. It’s crucial to bear in mind that GenVisR generates the waterfall chart by extracting information from three specific columns in the provided dataset. Therefore, ensure that the columns are appropriately named according to the required format. These three columns include: sample name, gene symbol, and mutation type.

2.3. Survival analysis

Overall survival (OS) or disease free survival (DFS, also called relapse-free survival and RFS) analysis based on gene expression were performed on GEPIA (http://gepia.cancer-pku.cn/). It uses Log-rank test, a.k.a the Mantel–Cox test, for hypothesis test. The group cutoff we chose here is Median, and percentage of cutoff high and low is 50 %. Pathological Stage Plot was used $\log_2$(TPM+1) for log-scale.

2.4. GO and KEGG enrichment analysis

GO and KEGG functional enrichment analyses were performed using Rstudio. Before the code run, we should install the following packages: org.Hs.eg.db (3.18.0), DOSE (3.28.1), topGO (2.54.0), clusterProfiler (4.10.0), pathview (1.42.0) and DO.db (2.9). Gene annotation of these mutated sites was recorded using the ICGC database, and we obtained 19,976 protein-coding genes. In GO and KEGG functional enrichment analysis, $p < 0.05$ was considered statistically significant.

2.5. Identifying patterns of mutation density

MutDens is an R application used to investigate patterns of mutational density for any specific genomic region (https://github.com/hui-sheen/MutDens) [26]. The location information for TSS, IS, TFBS,

and Enhancer regions was sourced from MutDens, a software package equipped with a data folder containing specific area location files. This data also encompassed information for other species. Prior to initiation, we prepared two input files: one for somatic mutations and another for focal genomic positions. We then tailored parameters in the optFile. It's worth noting that MutDens doesn't automatically verify software dependencies. As a result, we manually installed rmarkdown (2.25), GenomicRanges (1.52.1), knitr (1.44), MASS (7.3–60), and the species-specific Bioconductor annotation package. Subsequently, we followed the instructions outlined in the software's operation guide (README.md) to execute the process.

By default, mutation density was normalized to "mutations per megabyte" (MPM) to account for the ratio of a specific nucleotide type. However, in direct comparisons of two sample cohorts or two focal regions, the overall mutation burden per cohort/regions was considered. Otherwise, the statistical test result would reflect the difference in genome-wide mutation burden, not necessarily the situation within the vicinity of focal positions. Consequently, the mutation density was assessed using the metric of "mutations per kilo total mutations per megabase" (MPKM), rather than MPM [26].

2.6. Mutation signature analysis

The mutation signature analysis in this study was conducted using the deconstructSigs-R package (1.8.0). Initially, the data was organized into five columns (ID, chr, pos, ref, alt) information. If working with a MAF file from TCGA, extracting this data is straightforward. The mut.to.sigs.input function was then employed to construct the input file necessary for calculating the signature, yielding the 96 tribase types for each sample. Subsequently, the signature composition was deduced, and the percentage of three-base sequences in the tumor was generated after computing the weight of the mutation signature. Finally, the plotSignatures command was used for visualization.

APOBEC mutation signature analysis was extracted from the percentage of tribase sequences in the tumor. APOBEC activation in cancer results in elevated levels of genomic C-to-U deamination events, manifested as C-to-T switching or C-to-G switching in the TC*w* (w = A or T) trinucleotide environment. Combined mutations in the TC*w* background, including: TCA to TTA or TGA, and TCT to TTT or TGT, can represent APOBEC mutation counts [27].

2.7. Analysis of regulatory elements of genes

EpiRegio is a web page about find genomic regulatory elements online tool [28]. First enter the website: https://epiregio.de/geneQuery/(accessed it last time: 11/2023). The entry of the gene is simple and accepts either "gene symbol" or "ensembl ID". After completing gene input, click Query Database to display the number of regulatory elements on the gene, the start and end locations of regulatory elements, whether the function of regulatory elements is to inhibit or activate, and in what tissues it plays a regulatory role.

2.8. Statistical methods

All statistics were done in R4.2.1 and GraphPad Prism9 software. Plotting was done using R packages ggplot2, GraphPad Prism9, and Adobe Illustrator 2022 software. Wilcoxon signed-rank tests were performed using MutDens. A p-value of $p < 0.05$ indicated a statistically significant difference. A $p < 0.001$ was an extremely significant difference.

3. Results

3.1. Survival analysis

Survival Analysis is a statistical method to study the survival phenomenon and response time data and their statistical rules. It is an important means to associate phenotype with patient prognosis. Here in GACA-CN dataset, we found that the survival probability and survival time of male patients were significantly lower than that of female patients during the follow-up period (Log-rank test, $p = 0.046$) (Fig. 1A). At the same time, the survival analysis based on the primary tumor of the donor showed that the larger the tumor, the lower the survival probability (Log-rank test, $p = 0.003$). When the stage is T4 at diagnosis, the survival probability of patients is less than 50 % (Fig. 1B).

3.2. Distribution of somatic mutations on gastric cancer genome

To display the distribution of somatic mutations in the gastric cancer genome on 22 autosomes visually, we drew a mutation density map under a 1 Mb block. Ideally, nucleotide markers present uniform density in the whole genome. The distribution of high-depth resequencing mutation markers in the whole genome can be displayed with different colors that represent the mutation density in this region (Fig. 1C). The distribution of chromosomal locations of somatic mutations showed significant high-frequency mutation regions on chromosomes 2 and 10 in the Chinese gastric cancer data. In healthy people, the shape of each chromosome is basically constant, but there are a few chromosomes with small variations. In the metaphase of cell division, the centromere of some chromosomes is located at one end of the chromosome with a very short arm, that is, the proximal centromere chromosome, such as in human chromosomes 13–15. However, there were no mutations on the broken arms of these chromosomes, that is, there were no obvious mutations in genomic desert areas (Fig. 1C).

The waterfall diagram graphically shows the map of the top 20 high-frequency mutant genes in the GACA-CN dataset. The leftmost part of the waterfall diagram represents the gene and gene mutation frequency. The top is the gene mutation load and the effect of mutation on amino acids. On the far right, different colors represent different mutation types. The middle shows the mutation of each sample. From the mutation characteristics of these datasets, it is intuitive to find that most of the mutations occur in the intronic region and intergenic region (green part), and the mutation load variability in the dataset is large (Fig. 1D). It can also be seen that mutations in exon regions are basically non-synonymous mutations, which may have an important impact on the function of genes. At the same time, TP53, KMT2C and low density lipoprotein receptor-related protein 1B (LRP1B) are the top three mutated genes of gastric cancer (Fig. 1D).

3.3. High-frequency mutated gene analysis of gastric cancer

As reported in recent studies [29], because TP53, LRP1B, and KMT2C are the top three mutated genes of gastric cancer. However, Survival analysis and Pathological Stage of those top 3 mutated cancer genes remain to know. First, we investigated the impact of gene expression levels on patient prognosis, including overall Survival and disease-free Survival. TP53 is the highest frequently mutated gene, almost 49 % of donors were affected. Here in overall survival analysis (Fig. 2A), it can be seen that both high or low expression of the TP53 gene resulted in a similar overall risk of death, with a survival rate of approximately 20 % after 70 months (Log rank test, p = 0.46). The survival rate in LRP1B low expression group was significantly higher than that in LRP1B high expression group (Log rank test, p = 0.023) (Fig. 2B). Also, both high or low expression of the KMT2C gene resulted in a similar overall survival rate, with a survival rate of approximately 30 % after 80 months (Log rank test, p = 0.71) (Fig. 2C).

Disease Free Survival (DFS) is the time from randomization to the first recurrence/metastasis or death from any cause. It is also not concerned with the cause of death and is usually used as the main outcome measure after radical surgery. As is shown in Fig. 2E, the DFS survival rate in LRP1B low expression group was significantly higher than that in LRP1B high expression group (Log rank test, p = 0.02). However, the

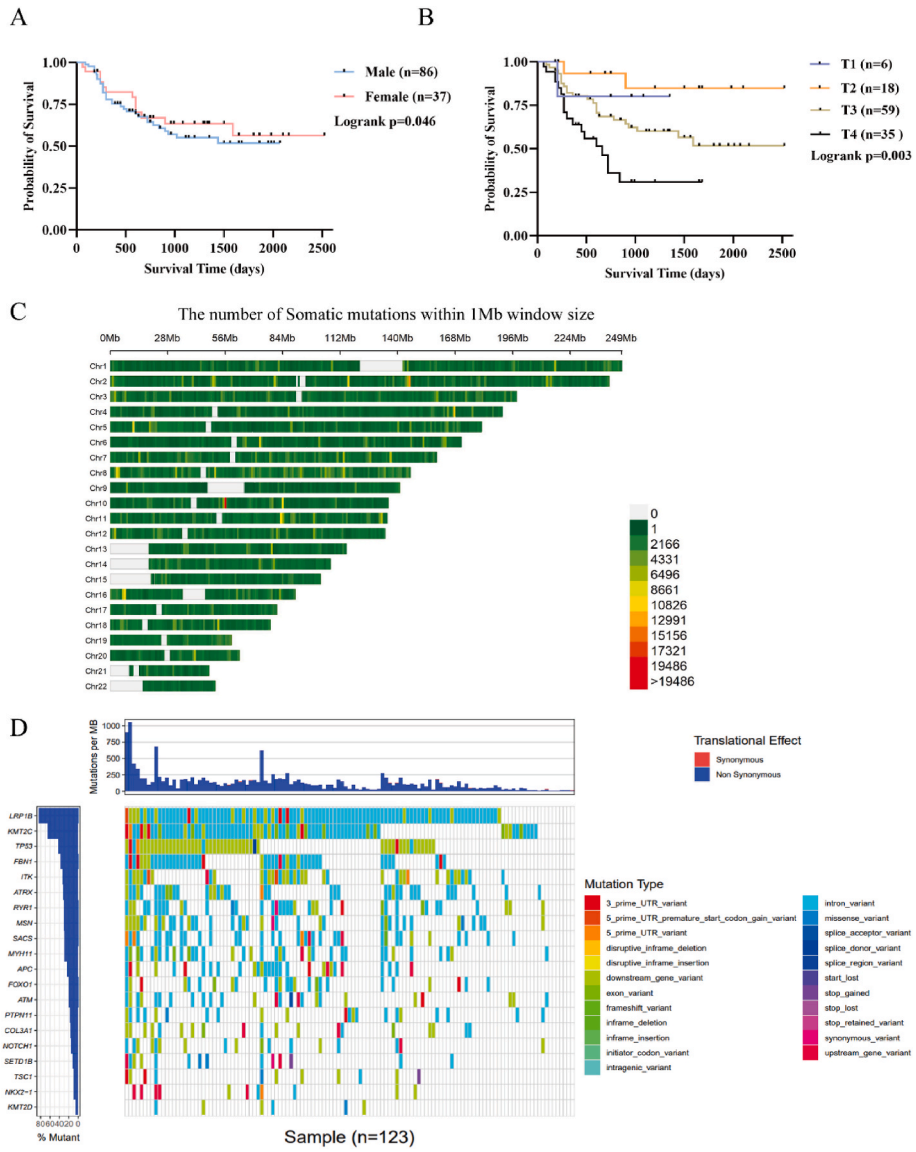

Fig. 1. Survival analysis by gender (A) and donor tumor diagnosis stage (B). Heat map of the distribution density of somatic mutations on gastric cancer autosomes (C). Different colors indicate the number of somatic mutations contained within 1 Mb Windows. The waterfull plot of the top 20 high-frequency mutant genes in the GACA-CN dataset (D). In the waterfall diagram of mutations, the horizontal axis represents different samples, the vertical axis is the gene, the filling represents the mutation of the gene, and different colors represent different mutations. The bar chart above is the statistics of the mutation for each sample.

DFS survival rate were not affected by the expression level of TP53 (Log rank test, p = 0.054) (Fig. 2D) and KMT2C genes (Log rank test, p = 0.74) (Fig. 2F), which may indicate that these two genes may have similar gene expression patterns during the development of gastric cancer. What's more, Pathological Stage Plot shows that TP53 and KMT2C genes expression level remained high throughout sub stages of gastric cancer development (Fig. 2G). These findings have implications for clinical diagnosis, treatment strategies, and further research in gastric cancer. The identification of significant differences in gene expression across stages helps in identifying potential biomarkers for diagnosis and prognosis, as well as therapeutic targets for personalized medicine approaches.

3.4. GO and KEGG enrichment analysis

Since many reported genes were related to the gastric cancer occurrences, we wanted to know what biological processes in which these genes that are related to gastric cancer are involved and what major signaling pathways in which they are enriched. In a cohort of Chinese gastric cancer patients, we acquired a total of 4,169,143 single nucleotide point mutations on 22 human autosomes. Gene annotation of these sites was recorded using the ICGC database, and the names of protein-coding genes were downloaded for further GO and KEGG enrichment analysis. We obtained 19,976 protein-coding genes. In the biological process (BP), genes related to gastric cancer were mainly enriched in mononuclear cell differentiation, signal release, embryonic organ development, and gland development (Fig. 3A). In the cellular component (CC), they were mainly enriched in neuronal cell bodies, collagen-containing extracellular matrices, cell–substrate junctions, and focal adhesions (Fig. 3B). In the molecular function (MF), genes related to gastric cancer were mainly enriched in GTPase regulator activity, nucleoside−triphosphatase regulator activity, DNA–binding transcription factor binding, and DNA−binding transcription activator activity (Fig. 3C). KEGG results showed that the enrichment mainly occurred in the following pathways: the PI3K-Akt signaling pathway, human papillomavirus infection, cytokine-cytokine receptor interaction, and the MAPK signaling pathway (Fig. 3D).

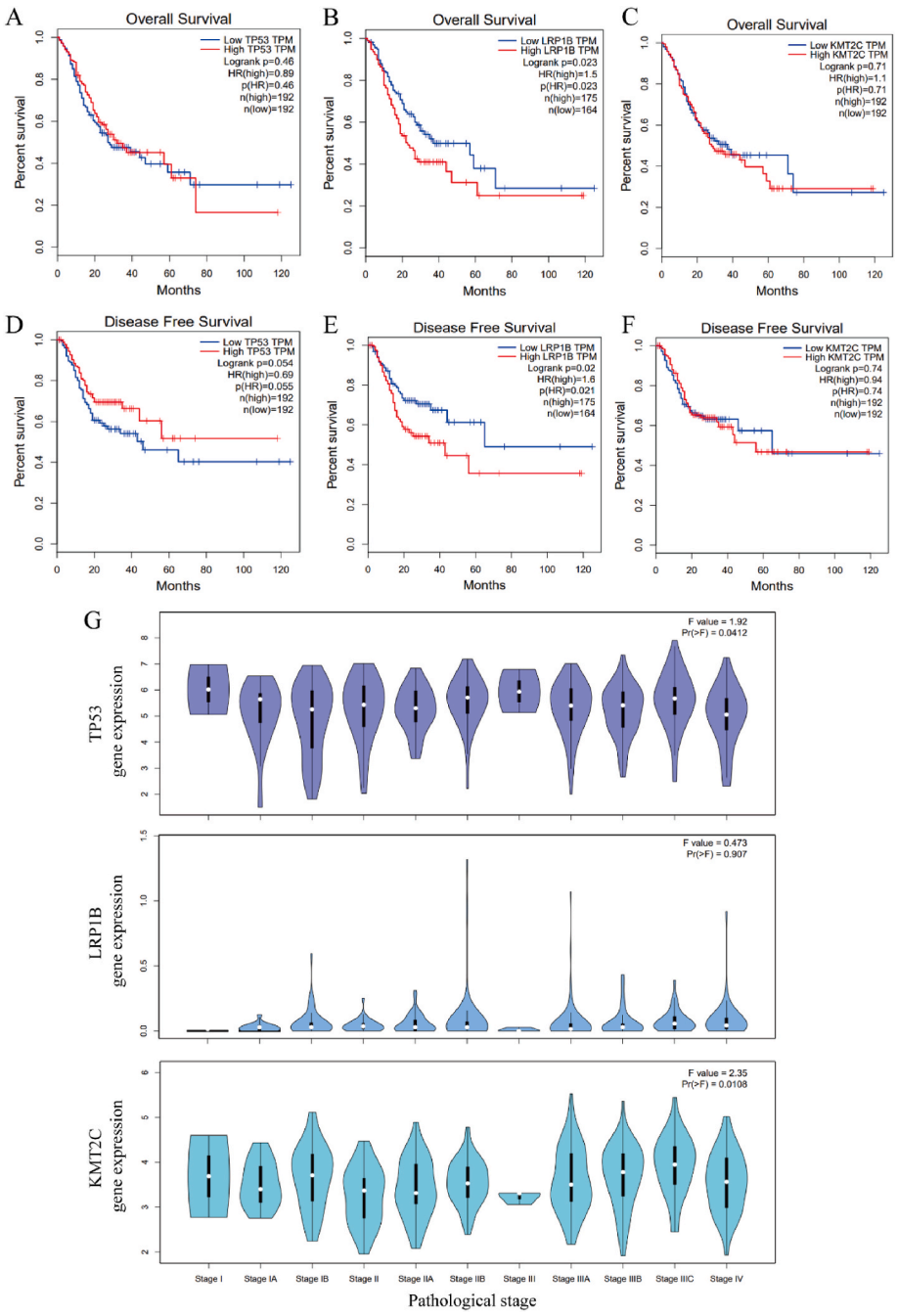

Fig. 2. Disease Free Survival (DFS) analysis and Pathological Stage of top 3 mutated cancer genes. Overall survival curve of TP53 (A), LRP1B (B) and KMT2C (C) genes. Disease free survival curve of TP53 (D), LRP1B (E) and KMT2C (F)genes. All those survival analyses were used by Log-rank test. Violin plots of gene expression generated according to the pathological stage of the patient (G). The method for differential gene expression analysis is one-way ANOVA, using pathological stage as variable for calculating differential expression: Gene expression ~ pathological stage. The expression data are first $\log_2$(TPM+1) transformed for differential analysis.

3.5. Base composition in focal regions

Base composition is a key component of genome organization and illustrating the details of base composition is an important step in elucidating the evolutionary significance and potential biological functions of other aspects of genome organization [30]. By using an advanced and unique R application named MutDens (https://github.com/hui-sheen/MutDens), we extracted the mutation type in specific genomic regions, which included TSS, IS, TFBS, and Enhancer. From the distribution of six nucleotide mutation types, we found that C > T accounted for about 28.4 %, T > C 21.8 %, C > A 16.5 %, C > G 4.6 %, T > A 9.9 %, and T > G 18.8 % in the whole genome (Fig. 4A). In those four focal regions, the average C > T accounted for about 41.7 %, T > C 18 %, C > A 15.4 %, C > G 5.9 %, T > A 7.1 %, and T > G 11.8 % (Fig. 4B). There was an excess of C > T mutations in the TSS and TFBS regions, which was significantly higher than that of the whole genome (Chi-square test, $p < 2.2 \times 10^{-16}$) (Fig. 4C and E). Nevertheless, transitions from T > C in the TSS and TFBS regions were significantly lower than that of whole genome (Chi-square test, $p < 2.2 \times 10^{-16}$). For transversions from T:A > A:T and A:T > C:G, those types in the TSS and TFBS regions were both significantly lower than that of the whole genome (Chi-square test, $p < 2.2 \times 10^{-16}$).

It is interesting that those differences in the six mutation types between focal regions like TSS and TFBS and the whole genome also

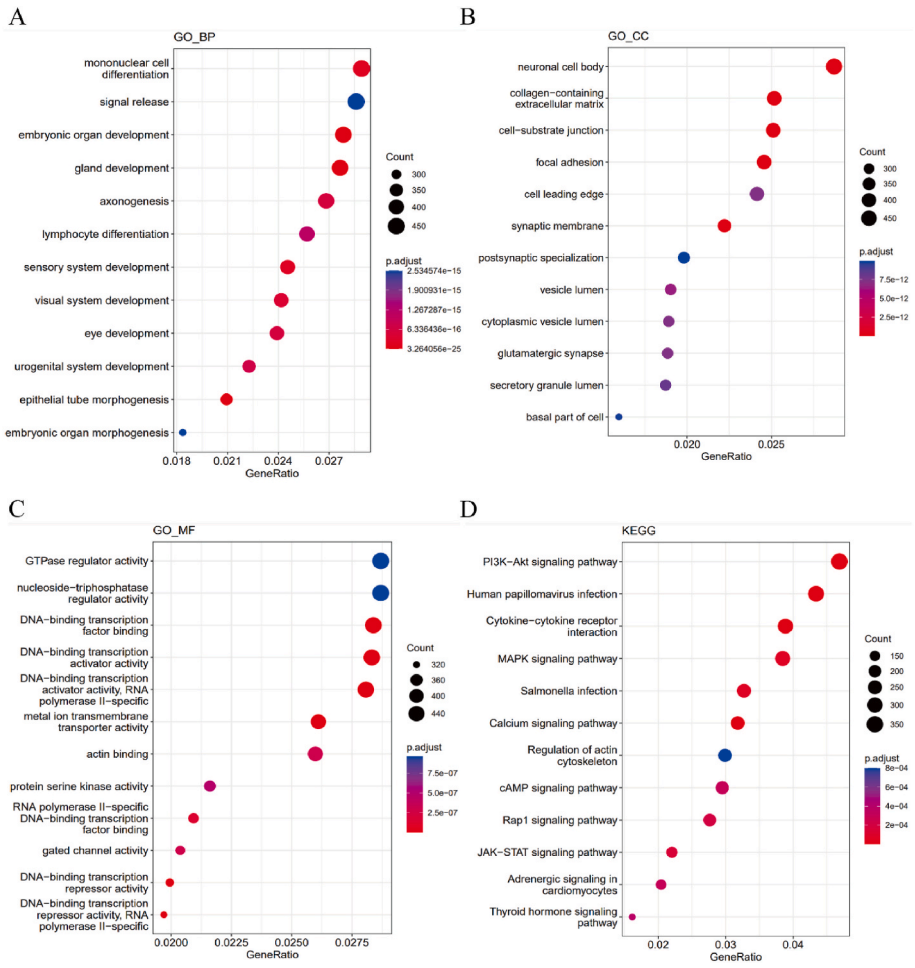

Fig. 3. The GO enrichment analysis of protein coding genes in GACA-CN dataset. Bubble plot (A) based on GO enrichment analysis of genetic related biological processes (BP). Bubble plot (B) were genetically related to cell components (CC). Bubble plot (C) related to molecular function (MF). In the bubble plot, the X-axis is the proportion of genes, and the size of the circle represents the number of genes that are enriched in each GO gene. The larger the circle, the more genes are enriched in that GO gene. The color of the circle represents the significance of the enrichment, and the redder the circle, the more significantly the gene is enriched on this GO. Bubble plot (D) based on KEGG enrichment analysis. The ordinate is the name of the signal pathway.

existed in the IS and Enhancer regions (Fig. 4D and F). We speculated that mutations tended to occur on the flanks of these four focal regions. To confirm this hypothesis, we calculated the mutation burden in MPKM (mutations per kilo total mutations per megabase) in specific regions of the gastric cancer genome. A high value of six mutation forms was found in the TSS, TFBS, IS, and Enhancer regions compared with the whole genome (Supplementary Table 1). At the same time, we also found that the value of the mutation burden of C > T was the highest and C > G was the lowest among those four focal regions, which is a widely accepted pattern of mutation distribution.

3.6. *Effects of TCD and TCR on transcriptional strand bias*

For a particular class of mutations, a clear divergence between the two coupled mutation density curves indicates transcriptional/replicative strand bias [26]. For TSS, we observed that the mutation density of the C > A form was drastically different between coding and template strands upstream from TSS (Wilcoxon signed-rank test, $p = 0.0003$ for TSS's left half range) (Fig. 4G). The mutation densities of the C > G and T > C forms on the coding strand were much higher than the template strand in GACA-CN dataset (Wilcoxon signed-rank test, $p = 0.013$ and 0.039 for TSS's left half range, respectively) (Fig. 4H and I), but no difference in GACA-JP dataset (Fig. 4K). We also observed that the mutation density of the C > A (Fig. 4J) and T > C (Fig. 4L) forms on the coding strand was drastically higher than on the template strand (Wilcoxon signed-rank test, p = 0.0049 and p = 0.019 for TSS's left half range, respectively) in GACA-JP dataset. It is worth noting that we found a mutation hot region in the 1 Kb downstream region of TSS and central peaks of origins in GACA-JP dataset, with mutation density about 5x higher than that in the upstream region, and TSS-coincident mutational spikes were commonly seen in six different mutation types.

If the mutation density curve for the coding strand was higher than the other curve for the template strand downstream from TSS, transcriptional strand bias would be postulated. Our scan of Chinese gastric cancer cohort for all six mutational classes indeed revealed mutational strand bias in many cases. For example, we find these transcriptional strand bias of T > G in gastric cancer data here (Fig. 4M) (Wilcoxon signed-rank test, $p = 0.04$). In addition, the mutation densities of the T > G forms on the coding strand were drastically less than the template strand in the upstream from TSS (Fig. 4N) (Wilcoxon signed-rank test, $p = 0.0038$), which was also confirmed in GACA-JP dataset.

The most important finding to emerge from this study was the description of mutation density in vicinity regions of TSS, that is, the transcriptional strand bias. GACA-CN and GACA-JP cohorts both showed transcriptional strand bias by presenting more T > G mutations on the non-transcribed strand. We believe that the strong T > G transcription strand bias characteristic in the genome of gastric cancer patients is caused by a combination of the higher transcription-coupled repair (TCR) efficiency on the template strand and transcription-coupled damage (TCD) suffered by the coding strand in the single-stranded DNA

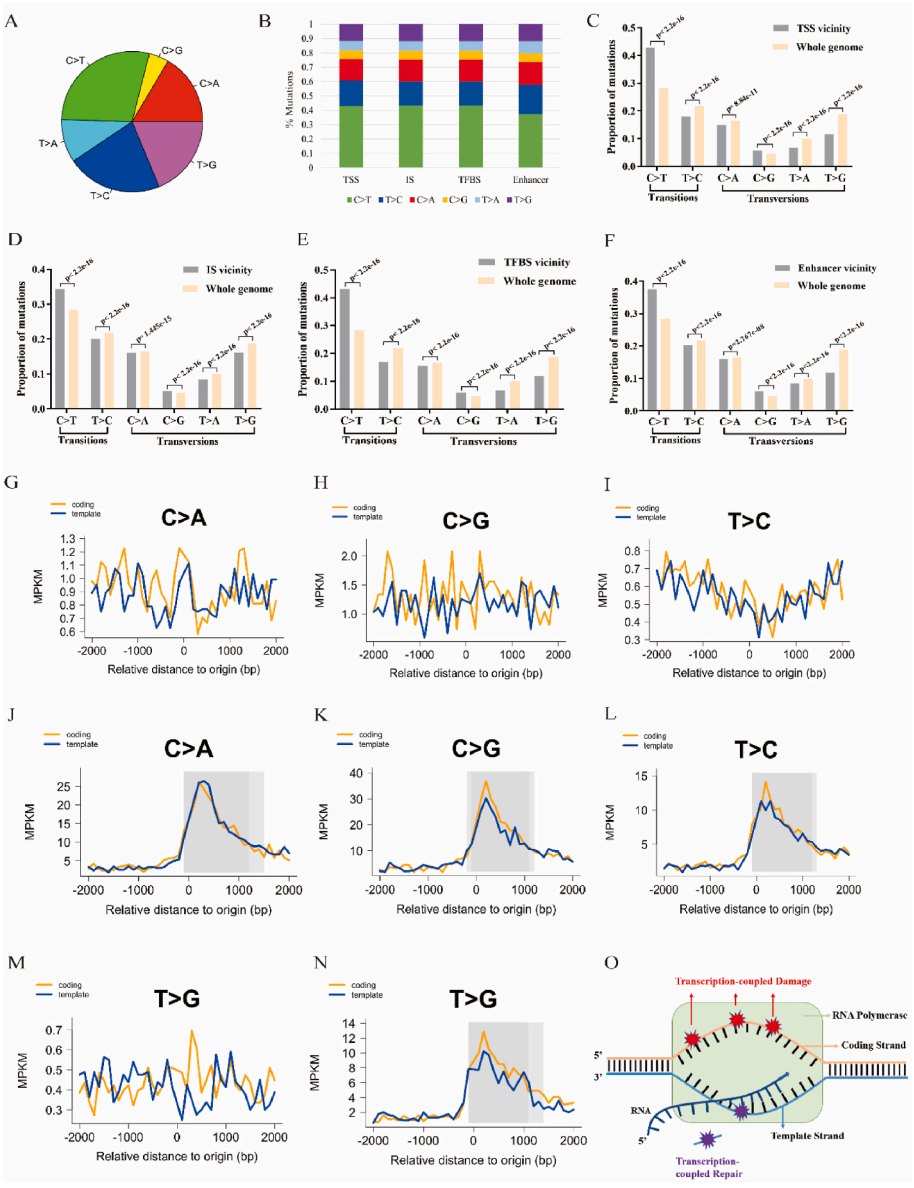

Fig. 4. Distribution of mutational classes was analyzed in both the whole genome (A) and in proximity to Transcription Start Sites (TSS), Intronic Splice (IS) sites, Transcription Factor Binding Sites (TFBS), and Enhancer regions (B). Comparison of the mutational base spectrum was conducted between the entire genome (in yellow) and mutations in the vicinity of (C) TSS, (D) IS, (E) TFBS, and (F) Enhancer regions (in gray). Significance was determined using the Chi-square test in RStudio. The plots were generated using MutDens with the GACA-CN dataset for case studies (G, H, I, M). Mutation density curves for (G) C > A, (H) C > G, (I) T > C, and (M) T > G mutations were analyzed with respect to the specific genomic position: Transcription Start Site (TSS). Plots for the GACA-JP dataset are presented in (J, K, L, N). Mutation density curves for (J) C > A, (K) C > G, (L) T > C, and (N) T > G mutations were examined. Thymine-Cytosine Dimer (TCD) induces damage on the non-transcribed strand, which is exposed as single-stranded DNA (ssDNA) during transcription. Thymine-Cytosine Repair (TCR) mends the transcribed strand. Both processes contribute to T-class asymmetry (O).

state during transcription (Fig. 4O). This is a complex mutation formation mechanism that produces extreme mutations of strand imbalance.

3.7. Replication strand bias during DNA replication and mutation signatures characterizing

Replication strand bias was probable if the mutation density dominance flipped from one mutation form to another on the left and right near the replication origin. In GACA-CN dataset, a visual graph in the replication origin regions showed that the mutation density of the C > A form was drastically different between the two strands (Wilcoxon signed-rank test, $p = 0.033$ for the whole range of the replication origin initiation site and $p = 0.009$ for the initiation site's left half range) (Fig. 5A). When we took the GACA-JP dataset into consideration, a strong mutational strand asymmetry of C > A was found drastically different between the two strands (Fig. 5B) (Wilcoxon signed-rank test, $p = 0.033$ for the whole range of the replication origin initiation site and $p = 0.009$ for the initiation site's left half range). We also found an origin-coincident and mutational off-center peak in the C > T class, but did not find the mutation strand bias characteristic of C > T during DNA replication (Fig. 5C). And C>T mutational strand asymmetry was found in the whole range of the replication origin initiation site in GACA-JP dataset (Fig. 5D), which was not exist in GACA-CN dataset.

Different mutation base types are combined to form an inherent mutation pattern, which is the so-called mutation signature. The identification of mutational markers has helped to deepen the understanding of environmental or endogenous factors in cancer development. In the GACA-CN dataset, we found that there were a variety of mutation

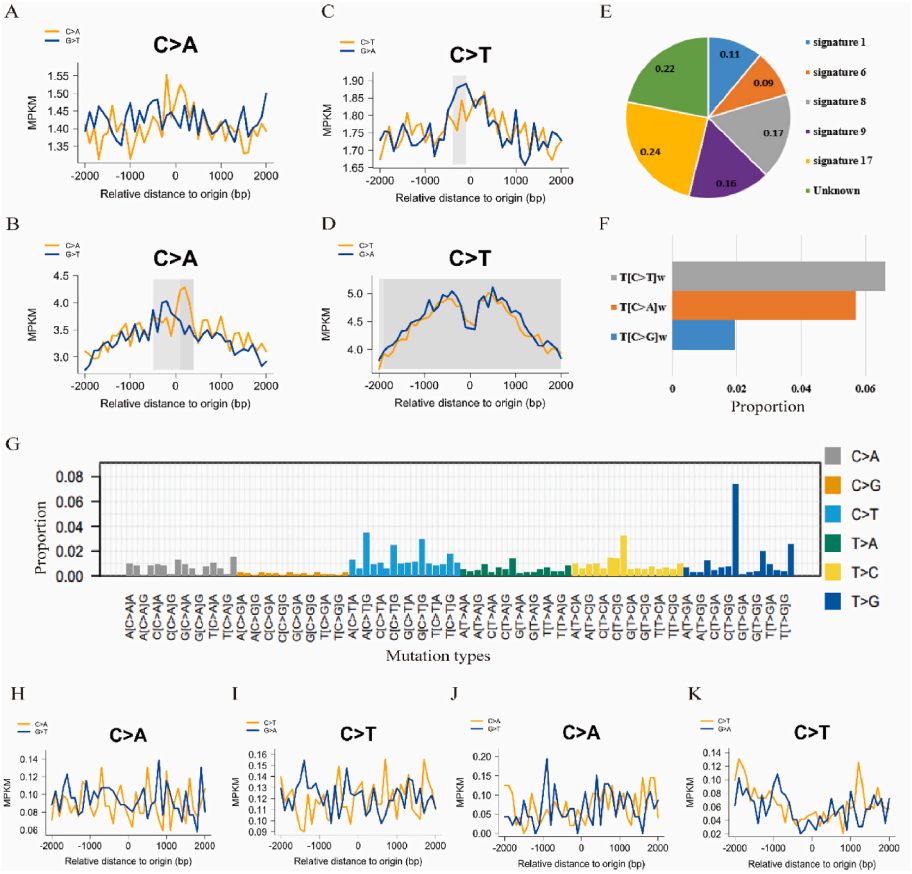

Fig. 5. Plots generated by MutDens in replication origin Initiation Site (IS) vicinity regions of the GACA-CN dataset (A, C) and the GACA-JP dataset (B, D). Distinct mutation density patterns were identified and marked in gray rectangles, which include off-center peak of origins (B,C,D). Compositional pie chart of mutational signatures in the gastric cancer genome (E). The proportion of 3 substitution types in APOBEC-characteristic mutations (F). Mutation signature distribution map using deconstructSigs-R package after considering the context sequence of mutation sites. The mutation position was added by one base before and after to form a three-base pattern, and then 96 (6*4*4) mutation combinations were counted (G). Representative plots of Enhancer vicinity regions generated by MutDens in case studies of GACA-CN dataset (H, I) and the GACA-JP dataset (J, K).

features in the gastric cancer genome (Fig. 5E), including the spontaneous deamination mark of cytosine methylation signature-1, accounting for about 11 %; Signature-6, which is characterized by DNA mismatch repair, accounted for about 9 %; Signature-8, characterized by nucleotide excision repair (NER) defects, accounted for about 17 %; Signature-9, characterized by hypermutation induced by polymerase eta during replication, accounted for about 16 %; Signature-17, which is marked by ROS damage, accounted for about 24 %. The rich mutational signature in the gastric cancer genome hints at the complexity of somatic mutation formation.

In humans, APOBEC mutagenesis primarily occurs on the lagging strand template during DNA replication. The APOBEC signature shows strong R-class asymmetry, with a higher rate of C > G and C > T mutations in right-replicating regions, where reference-strand DNA is predicted to be replicated as the lagging-strand, exposed as ssDNA between Okazaki segments. In this study, we also found that the APOBEC induction pattern in the gastric cancer genome was mainly T[C>T]*w* and T[C>A]*w*, and less T[C>G]*w* (*w* = A or T) (Fig. 5F). Fig. 5G shows the composition of 96 specific mutation signatures, which provides a reference for APOBEC feature analysis.

3.8. *Mutation asymmetry patterns in enhancer and TFBS vicinity regions*

In recent years, enhancers, which are *cis*-type elements that affect the transcription regulation of proximal or distal genes, have received extensive attention. Studying enhancers and the different patterns of mutations around them can help us to understand more about how evolution happens. In addition to driving species evolution, enhancers have been linked to disease: mutations in enhancers have been linked to more than 80 % of human diseases [31]. After visualization of six different mutant classes in Enhancer regions, we found that the mutation density of C > A and C > T forms were significantly less than G > T (Wilcoxon signed-rank test, $p = 0.0113$) and G > A (Wilcoxon signed-rank test, $p = 0.0279$) upstream of the enhancer region in GACA-CN dataset, respectively (Fig. 5H and I). However, the mutational asymmetry in the region near the enhancer was not found in the GACA-JP dataset (Fig. 5J and K).

Transcription factors are protein molecules that bind to genes with specific sequences to ensure the expression of target genes in specific time and space, and these factors control chromatin and transcription by recognizing specific DNA sequences. If the mutation occurs in the promoter region, it may affect the binding of the gene by the transcription factor, which in turn affects gene expression. Therefore, we investigated the mutation density pattern in the proximal region of TFBS. In GACA-CN dataset, the mutation density of C > A was significantly less than G > T upstream of TFBS (Wilcoxon signed-rank test, $p = 0.0187$), but higher in the downstream of TFBS (Wilcoxon signed-rank test, $p = 2.03e-04$) (Fig. 6A). The mutation density of C > G was significantly less than G > C upstream of TFBS (Fig. 6B) (Wilcoxon signed-rank test, $p = 3.09e-05$).

The mutation density of C > T was significantly less than G > An upstream of TFBS (Wilcoxon signed-rank test, $p = 1.0e-07$), but higher downstream of TFBS (Wilcoxon signed-rank test, $p = 3.22e-04$). We also found an origin-coincident, mutational off-center peak in the C > T class

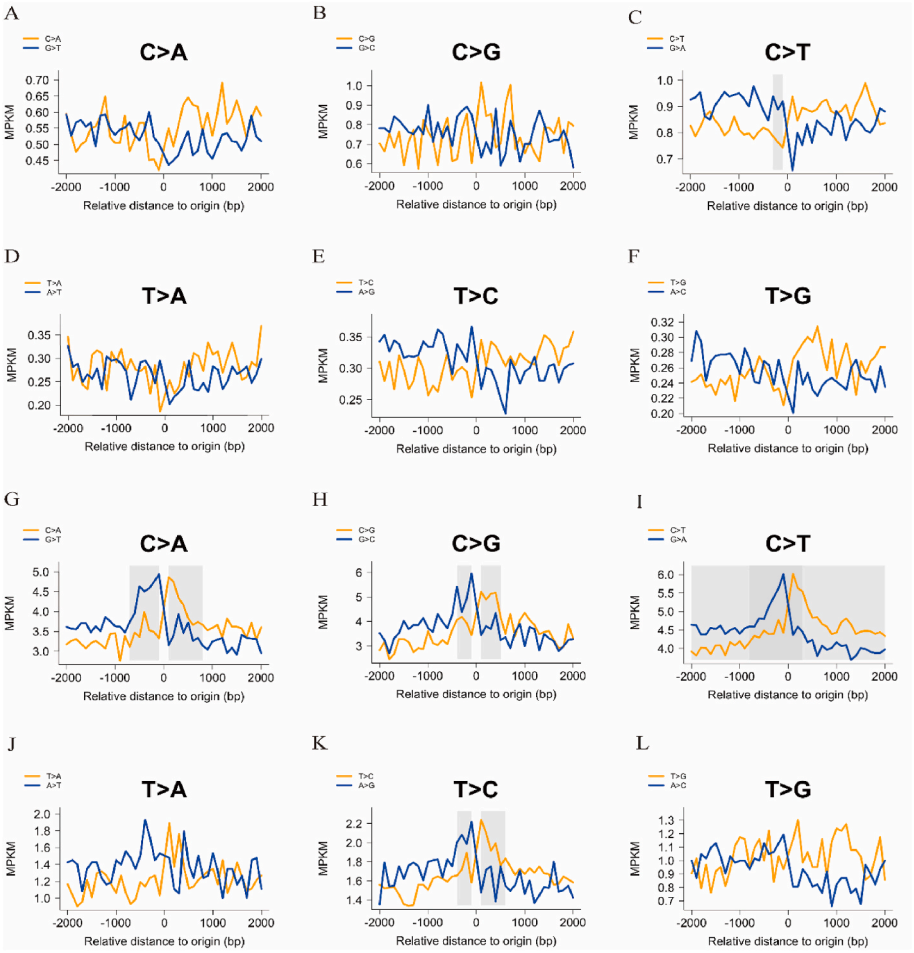

Fig. 6. Mutation density curves of (A) C > A, (B) C > G, (C) C > T, (D) T > A, (E) T > C, and (F) T > G, analyzed against the special genomic position sets: Transcription factor binding site (TFBS) and generated by MutDens in case studies of GACA-CN dataset. Mutation density curves of (G) C > A, (H) C > G, (I) C > T, (J) T > A, (K) T > C, and (L) T > G, analyzed against the special genomic position sets: Transcription factor binding site (TFBS) in GACA-JP dataset. Distinct mutation density patterns were identified and marked in gray rectangles, including off-center peak of origins.

(Fig. 6C). These similar waves were also found in the density of T > C, which was less than A > G upstream of TFBS (Wilcoxon signed-rank test, p = 1.0e-07) and higher downstream of TFBS (Wilcoxon signed-rank test, p = 1.91e-05) (Fig. 6E). The mutation density of the T > A form was drastically higher than A > T (Wilcoxon signed-rank test, p = 4.83e-04 (Fig. 6D) in the GACA-CN dataset, but lower in the GACA-JP dataset (Fig. 6J). We also found that the mutation density of T > G was significantly less than A > C upstream of TFBS (Wilcoxon signed-rank test, p = 1.13e-05), but higher downstream of TFBS (Wilcoxon signed-rank test, p = 8.2e-05) (Fig. 6F). These results are further confirmed in the GACA-JP dataset. For examples, the mutation density of C > A (Fig. 6G), C > G (Fig. 6H), C > T (Fig. 6I) were significantly less than G > T, G > C, G > An upstream of TFBS but higher in the downstream of TFBS.

4. Discussion

Gastric cancer is one of the most common cancers in China, and it is characterized by high incidence and poor prognosis. Most cancer patients cannot be cured due to a late diagnosis. It is essential to analyze changes in the genome to find biomarkers for early detection of gastric cancer and to present an accurate prognosis. Given the high prevalence of gastric Cancer in China and the possible interaction of environmental factors, Peking Cancer Hospital initiated the Gastric Cancer-China (GACA-CN) project and shared somatic mutation data with the International Cancer Genome Consortium (ICGC), which was the data source for our analysis. In addition, to reveal the mutation characteristics of the Asian population, we combined the Gastric Cancer somatic mutation data from Japanese people from the Gastric Cancer-Japan (GACA-JP) Project released in ICGC to find the common points and differences in the asymmetry of gastric cancer mutations.

The results of location distribution of somatic mutation chromosomes showed that chromosome 2 was prone to mutation in the GACA-CN and GACA-JP data. This was consistent with the published characteristics of somatic mutations in gastric cancer [10] that reflect its heterogeneity [32]. According to the location of chromosome mutations in the GACA-JP data, no mutations occurred on the broken arm of chromosomes 13–15, but high-frequency mutations occurred on chromosomes 2 and 17 (Fig. S1). These mutation hotspots were located at the positions of TTN and TP53 genes. TTN has a high frequency of mutations in a variety of tumors [33,34], and TP53 also has a high frequency. Most functional studies have found that after a TP53 mutation, the cancer inhibitory function was lost [35,36], and even the effect of inhibiting cancer was changed to promoting cancer [37]. Interestingly, we also found a high mutation density on chromosome 19 in the GACA-JP dataset (Fig. S1), which was not found in the GACA-CN dataset (Fig. 1C).

TP53 and KMT2C are high frequency genes in gastric cancer and can be used as prognostic markers for immunotherapy. This study further analyzed the survival effect of genes on patient prognosis, and confirmed that the gene expression of TP53 and KMT2C mutations was higher throughout the cancer development cycle, and the survival rate of patients after 70 months was low. However, the survival rate in LRP1B low expression group was significantly higher than that in LRP1B

high expression group. LRP1B is known to be a putative tumor suppressor and a member of the low-density lipoprotein (LDL) receptor family. The LDL receptor family has a role associated with extracellular ligand clearance and is thought to be involved in extracellular signaling, as evidenced by LRP1B silencing and downregulation observed in renal cell carcinoma and thyroid carcinoma.

In this study, we also found that among the characteristics of somatic mutations in four specific regions, transitions accounted for about 60 % and transversions accounted for about 40 %, among which C > T was the main mutation type (Fig. 4B, Fig. S2). In the subsequent calculation of mutation burden in MPKM, there was a high mutation load near these four regions, especially TSS and TFBS (Supplementary Table 1), which also suggested that mutations might be more prone to occur in the proximal regions of TSS, TFBS, IS, and Enhancer. This further reflected the regional heterogeneity of somatic mutations in gastric cancer, the high mutation load and mutation uncertainty of the genome, and the diversity of the genome at the molecular level. Therefore, personalized precision medicine is required to diagnose and to treat for gastric cancer.

Our results highlight the widespread mutational strand asymmetries observed in cancer genomes, mediated by DNA replication, RNA transcription, and their associated repair pathways. Here in GACA-CN and GACA-JP cohorts, we first identified the transcriptional strand bias of T>G transversion. It suggests that the RNA polymerase on the transcription strand is blocked in the transcription process, and the blocked RNA polymerase recruit's nucleotide shear repair related factors to repair the damaged nucleotides to avoid mutation. Because the whole genome is replicated every time a cell divides, replication direction has the potential to exert larger asymmetries in mutational data. For IS, we found a different result between GACA-JP and GACA-CN cohorts. In GACA-CN, we saw predominantly C > A mutations in left-replicating regions and G > T in right-replicating regions (Fig. 5A). This was consistent with the recent report on POLE tumors: that tumors carried functional mutations in the proofreading exonuclease domain of POLE [38]. In humans, Haradhvala et al. suggested that apolipoprotein B mRNA editing enzyme catalytic polypeptide like (APOBEC) mutagenesis primarily occurred on the lagging-strand template during DNA replication, and the APOBEC signature showed strong R-class asymmetry. There was a higher rate of C > G and C > T mutations in right-replicating regions, where reference-strand DNA was predicted to be replicated as the lagging-strand template where it was exposed as ssDNA between Okazaki segments. The magnitude of this asymmetry increased with enrichment of the APOBEC signature [22].

Enhancers are genomic sequences that play a key role in regulating tissue-specific gene expression levels. An increasing number of diseases that are associated with impaired enhancer function through chromosomal rearrangement, genetic variation within enhancers, or epigenetic regulation have been discovered recently [39]. In this study, we found that the mutation density of C > A (Fig. 5H) and C > T (Fig. 5I) forms were significantly less than G > T and G > An upstream of enhancer regions, respectively. Enhancers are characterized by specific chromatin modifications like H3K4me1 and H3K27ac [40–42]. We speculate that the mutation bias in the region around enhancers may be influenced by chromatin status. For example, the frequency of base mutations correlated negatively with H3K27ac modifications in germline mutations in autistic individuals and in somatic mutations in cancer [43,44].

In addition to these mutant signatures associated with specific mutation processes, such as APOBEC mutations, nucleotide mismatch repair, or various carcinogens, more recently, nucleotide excisational repair (NER) associated with a mutant signature has been associated with specific mutation patterns within TFBS in the cancer genome [45]. The mutation asymmetries in TFBS regions in the GACA-JP cohort was consistent with the GACA-CN cohort, but the GACA-JP cohort had a more significant mutation strand bias in the upstream and downstream regions of TFBS. This was especially true for those TFBS-coincident mutational, off-center peaks in the C > A, C > G, C > T, and T > C classes (Fig. 6G, H, I and K). Given the differences in DNA binding specificity between TFs, we hypothesized that mutational signatures that were specific to gastric cancer types affected TFBSs differentially across TF families.

Gene expression is regulated mainly at the transcriptional level by the binding of TF to promoters (i.e., *cis*-regulatory regions that surround genes' transcription start sites, TSS) and enhancers (i.e., *cis*-regulatory regions distal to genes) at TF binding sites (TFBS) [46–48]. After comparing the number of regulatory elements between high mutational impact genes and low mutational impact genes, we found the regulatory elements in high mutational impact genes is significantly higher than low mutational impact genes (Fig. S3), indicating the important role of the regulatory elements during mutation occurrence.

Last but not least, a limitation of this analysis lies in its static nature. While our donor pool encompasses various stages of cancer (Stage I to IV), it does not take into account other recognized important prognostic factors, such as stage, N ratio, and HER positivity. This is one of the shortcomings of our study. Furthermore, cancer onset is a dynamic stochastic process, during which cell populations acquire mutations driving tumor initiation, immune evasion, expansion, invasive infiltration, and resistance to treatment [49]. Meanwhile, we conducted KEGG and GO analyses on all identified genetic variants to gain a preliminary understanding of the potential functional implications associated with these variants. However, we acknowledge a critical point regarding the necessity of considering the proven impact of genetic variants on gene function or expression level. As rightly pointed out, not every somatic variant is inherently pathogenic or functional. The work by He et al. [50] underscores this point, highlighting that among a set of somatic variants, only a fraction was identified as strong or potentially clinically actionable. This prompts us to acknowledge the need for caution in interpreting the functional significance of all identified variants solely based on computational predictions. In light of these considerations, we emphasize that our study represents an initial step, and its findings should be viewed in the context of this limitation. Future research directions should involve more targeted analyses, including functional characterization studies, to delineate the true impact of specific variants on gene function or expression. The importance of distinguishing between pathogenic and non-pathogenic variants will be a key aspect of our future investigations.

The emergence and advancement of next-generation sequencing technologies have greatly facilitated the quantification of cancer evolution dynamics. In the future, a combined analysis using targeted sequencing, genomic analysis, single-cell sequencing methods, and emerging multi-omics technologies should be employed. By sequencing and analyzing circulating tumor cells or tumor DNA, a minimally invasive approach holds promise for monitoring and capturing the dynamic mutations in cancer [51].

5. Conclusion

Somatic mutation is the main factor in cancer. Rather than studying the characteristics of individual mutations, studying mutations in conjunction with specific genomic regions provides a unique perspective on tumorigenesis and history. Mutation strand bias, in contrast to mutation load and mutation signature, represents another holistic view of genome-wide mutation and appears to be a promising approach to tumorigenesis. The analysis in this series of studies compares many single base substitutions between two complementary strands of DNA. The mutation mechanism can be revealed only by checking the mutation form and number, which eliminates the mutation annotation step used commonly in another research workflow. Starting from specific regions of the genome (TSS, IS, TFBS, and Enhancer regions), we present mutation density patterns in those four focal regions, which is an important supplement to the conventional studies on the mechanism of somatic mutations; it is significant to reveal the characteristics of somatic mutations in gastric cancer. Novel mutations and repair processes are emerging constantly from studies of cancer genome sequencing, and

looking at them through the lens of the asymmetry of mutation strands provides immediate insight into their molecular mechanisms. This work systematically describes the mutational strand asymmetries and helps to reveal the underlying biological mechanism of somatic mutations in cohorts of Asian patients with gastric cancer.

Data availability

The data obtained and analyzed in the current study are available in the ICGC database (https://icgc.org/).

Code availability (software application or custom code)

Not applicable.

Ethics approval

Ethical review/approval and written informed consent were not required for the study due to the computational nature of the work.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

CRediT authorship contribution statement

Yangyang Shen: Conceptualization, Data curation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Kai Shi:** Conceptualization, Data curation, Formal analysis, Methodology, Software. **Dongfeng Li:** Formal analysis, Investigation, Resources, Software, Validation. **Qiang Wang:** Conceptualization, Formal analysis, Investigation, Resources, Validation. **Kangkang Wu:** Data curation, Formal analysis, Project administration, Resources, Validation, Visualization. **Chungang Feng:** Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the General Program of National Natural Science Foundation of China (No. 82070765), the National Key R&D Program of China (No. 2021YFD1300100), the Guangxi Key R&D Program (No. AB21220005), and the Revitalization Program of Biological Breeding of Jiangsu Province (No. JBGS(2021)109). Thanks to Thomas A. Gavin (Cornell University, U.S.) and Cassie Zhao (GemPharmatech Co., Ltd.) for improving the English of this manuscript. We acknowledge International Cancer Genome Consortium (ICGC) for data sharing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbrep.2023.101597.

References

[1] W.Q. Chen, R.S. Zheng, P.D. Baade, S.W. Zhang, X.Q. Yu, Cancer Statistics in China, 2016.
[2] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2016, CA A Cancer J. Clin. 66 (2016).
[3] M. Alsina, V. Arrazubi, M. Diez, J. Tabernero, Current developments in gastric cancer: from molecular profiling to treatment strategy, Nat. Rev. Gastroenterol. Hepatol. 20 (2023) 155–170, https://doi.org/10.1038/s41575-022-00703-w.
[4] J.A. Ajani, T.A. D'Amico, D.J. Bentrem, J. Chao, D. Cooke, C. Corvera, P. Das, P. C. Enzinger, T. Enzler, P. Fanta, F. Farjah, H. Gerdes, M.K. Gibson, S. Hochwald, W. L. Hofstetter, D.H. Ilson, R.N. Keswani, S. Kim, L.R. Kleinberg, S.J. Klempner, J. Lacy, Q.P. Ly, K.A. Matkowskyj, M. McNamara, M.F. Mulcahy, D. Outlaw, H. Park, K.A. Perry, J. Pimiento, G.A. Poultsides, S. Reznik, R.E. Roses, V.E. Strong, S. Su, H.L. Wang, G. Wiesner, C.G. Willett, D. Yakoub, H. Yoon, N. McMillian, L. A. Pluchino, Gastric cancer, version 2.2022, NCCN clinical practice guidelines in oncology, J. Natl. Compr. Cancer Netw. 20 (2022) 167–192, https://doi.org/10.6004/jnccn.2022.0008.
[5] L.A. Torre, R.L. Siegel, E.M. Ward, A. Jemal, Global cancer incidence and mortality rates and trends–an update, Cancer Epidemiol. Biomarkers Prev. 25 (2016) 16–27, https://doi.org/10.1158/1055-9965.EPI-15-0578.
[6] G.B.D.S.C. Collaborators, The global, regional, and national burden of stomach cancer in 195 countries, 1990-2017: a systematic analysis for the Global Burden of Disease study 2017, Lancet Gastroenterol Hepatol 5 (2020) 42–54, https://doi.org/10.1016/S2468-1253(19)30328-0.
[7] J.K. Zhao, M. Wu, C.H. Kim, Z.Y. Jin, J.Y. Zhou, R.Q. Han, J. Yang, X.F. Zhang, X. S. Wang, A.M. Liu, X. Gu, M. Su, X. Hu, Z. Sun, G. Li, L. Li, L. Mu, Z.F. Zhang, Jiangsu Four Cancers Study: a large case-control study of lung, liver, stomach, and esophageal cancers in Jiangsu Province, China, Eur. J. Cancer Prev. 26 (2017) 357–364, https://doi.org/10.1097/CEJ.0000000000000262.
[8] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, Ca - Cancer J. Clin. 68 (2018) 394–424, https://doi.org/10.3322/caac.21492.
[9] I.R. Watson, K. Takahashi, P.A. Futreal, L. Chin, Emerging patterns of somatic mutations in cancer, Nat. Rev. Genet. 14 (2013) 703–718, https://doi.org/10.1038/nrg3539.
[10] H. Cai, C. Jing, X. Chang, D. Ding, T. Han, J. Yang, Z. Lu, X. Hu, Z. Liu, J. Wang, L. Shang, S. Wu, P. Meng, L. Lin, J. Zhao, M. Nie, K. Yin, Mutational landscape of gastric cancer and clinical application of genomic profiling based on target next-generation sequencing, J. Transl. Med. 17 (2019) 189, https://doi.org/10.1186/s12967-019-1941-0.
[11] X. Pan, X. Ji, R. Zhang, Z. Zhou, Y. Zhong, W. Peng, N. Sun, X. Xu, L. Xia, P. Li, J. Lu, J. Tu, Landscape of somatic mutations in gastric cancer assessed using next-generation sequencing analysis, Oncol. Lett. 16 (2018) 4863–4870, https://doi.org/10.3892/ol.2018.9314.
[12] E.D. Pleasance, R.K. Cheetham, P.J. Stephens, D.J. McBride, S.J. Humphray, C. D. Greenman, I. Varela, M.L. Lin, G.R. Ordonez, G.R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R.J. Carter, L. Chen, A.J. Cox, S. Edkins, P.I. Kokko-Gonzales, N.A. Gormley, R.J. Grocock, C.D. Haudenschild, M.M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L.J. Mudie, Z. Ning, T. Royce, O.B. Schulz-Trieglaff, A. Spiridou, L.A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M.T. Ross, P.J. Campbell, D.R. Bentley, P. A. Futreal, M.R. Stratton, A comprehensive catalogue of somatic mutations from a human cancer genome, Nature 463 (2010) 191–196, https://doi.org/10.1038/nature08658.
[13] L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, S.A. Aparicio, S. Behjati, A.V. Biankin, G.R. Bignell, N. Bolli, A. Borg, A.L. Borresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H.R. Davies, C. Desmedt, R. Eils, J.E. Eyfjord, J.A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jager, D. T. Jones, D. Jones, S. Knappskog, M. Kool, S.R. Lakhani, C. Lopez-Otin, S. Martin, N.C. Munshi, H. Nakamura, P.A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J.V. Pearson, X.S. Puente, K. Raine, M. Ramakrishna, A.L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T.N. Schumacher, P.N. Span, J.W. Teague, Y. Totoki, A.N. Tutt, R. Valdes-Mas, M.M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L.R. Yates, I. Australian Pancreatic Cancer Genome, I.B. C. Consortium, I.M.-S. Consortium, I. PedBrain, J. Zucman-Rossi, P.A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S.M. Grimmond, R. Siebert, E. Campo, T. Shibata, S.M. Pfister, P.J. Campbell, M.R. Stratton, Signatures of mutational processes in human cancer, Nature 500 (2013) 415–421, https://doi.org/10.1038/nature12477.
[14] I.T.P.-C.A.o.W.G. Consortium, Pan-cancer analysis of whole genomes, Nature 578 (2020) 82–93, https://doi.org/10.1038/s41586-020-1969-6.
[15] E.D. Pleasance, P.J. Stephens, S. O'Meara, D.J. McBride, A. Meynert, D. Jones, M. L. Lin, D. Beare, K.W. Lau, C. Greenman, I. Varela, S. Nik-Zainal, H.R. Davies, G. R. Ordonez, L.J. Mudie, C. Latimer, S. Edkins, L. Stebbings, L. Chen, M. Jia, C. Leroy, J. Marshall, A. Menzies, A. Butler, J.W. Teague, J. Mangion, Y.A. Sun, S. F. McLaughlin, H.E. Peckham, E.F. Tsung, G.L. Costa, C.C. Lee, J.D. Minna, A. Gazdar, E. Birney, M.D. Rhodes, K.J. McKernan, M.R. Stratton, P.A. Futreal, P. J. Campbell, A small-cell lung cancer genome with complex signatures of tobacco exposure, Nature 463 (2010) 184–190, https://doi.org/10.1038/nature08629.
[16] M.S. Lawrence, P. Stojanov, P. Polak, G.V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C.H. Mermel, S.A. Roberts, A. Kiezun, P.S. Hammerman, A. McKenna, Y. Drier, L. Zou, A.H. Ramos, T.J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M.L. Cortes, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D.I. Heiman,

T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A.M. Dulak, J. Lohr, D. A. Landau, C.J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S.B. Gabriel, C.W.M. Roberts, J.A. Biegel, K. Stegmaier, A.J. Bass, L.A. Garraway, M. Meyerson, T.R. Golub, D.A. Gordenin, S. Sunyaev, E.S. Lander, G. Getz, Mutational heterogeneity in cancer and the search for new cancer-associated genes, Nature 499 (2013) 214–218, https://doi.org/10.1038/nature12213.

[17] G.P. Pfeifer, M.F. Denissenko, M. Olivier, N. Tretyakova, S.S. Hecht, P. Hainaut, Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers, Oncogene 21 (2002) 7435–7451, https://doi.org/10.1038/sj.onc.1205803.

[18] I. Martincorena, A. Roshan, M. Gerstung, P. Ellis, P. Van Loo, S. McLaren, D. C. Wedge, A. Fullam, L.B. Alexandrov, J.M. Tubio, L. Stebbings, A. Menzies, S. Widaa, M.R. Stratton, P.H. Jones, P.J. Campbell, Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin, Science 348 (2015) 880–886, https://doi.org/10.1126/science.aaa6806.

[19] G. Spivak, A.K. Ganesan, The complex choreography of transcription-coupled repair, DNA Repair 19 (2014) 64–70, https://doi.org/10.1016/j.dnarep.2014.03.025.

[20] M. Fousteri, L.H. Mullenders, Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects, Cell Res. 18 (2008) 73–84, https://doi.org/10.1038/cr.2008.6.

[21] P.C. Hanawalt, G. Spivak, Transcription-coupled DNA repair: two decades of progress and surprises, Nat. Rev. Mol. Cell Biol. 9 (2008) 958–970, https://doi.org/10.1038/nrm2549.

[22] N.J. Haradhvala, P. Polak, P. Stojanov, K.R. Covington, E. Shinbrot, J.M. Hess, E. Rheinbay, J. Kim, Y.E. Maruvka, L.Z. Braunstein, A. Kamburov, P.C. Hanawalt, D.A. Wheeler, A. Koren, M.S. Lawrence, G. Getz, Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair, Cell 164 (2016) 538–549, https://doi.org/10.1016/j.cell.2015.12.050.

[23] L.B. Alexandrov, S. Nik-Zainal, D.C. Vvedge, S. Aparicio, Signatures of mutational processes in human cancer, Nature 500 (2013) 415–421. A413.

[24] J.E. Kucab, X. Zou, S. Morganella, M. Joel, A.S. Nanda, E. Nagy, C. Gomez, A. Degasperi, R. Harris, S.P. Jackson, V.M. Arlt, D.H. Phillips, S. Nik-Zainal, A compendium of mutational signatures of environmental agents, e816, Cell 177 (2019) 821–836, https://doi.org/10.1016/j.cell.2019.03.001.

[25] E. Letouze, J. Shinde, V. Renault, G. Couchy, J.F. Blanc, E. Tubacher, Q. Bayard, D. Bacq, V. Meyer, J. Semhoun, P. Bioulac-Sage, S. Prevot, D. Azoulay, V. Paradis, S. Imbeaud, J.F. Deleuze, J. Zucman-Rossi, Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis, Nat. Commun. 8 (2017) 1315, https://doi.org/10.1038/s41467-017-01358-x.

[26] H. Yu, S. Ness, C.I. Li, Y. Bai, P. Mao, Y. Guo, Surveying mutation density patterns around specific genomic features, Genome Res. 32 (2022) 1930–1940, https://doi.org/10.1101/gr.276770.122.

[27] S. Vural, J. Krushkal, R. Simon, Analysis of APOBEC3A and APOBEC3B mutational signatures using next-generation sequencing data from cancer cell lines, 2590-2590, Cancer Res. 77 (2017).

[28] N. Baumgarten, D. Hecker, S. Karunanithi, F. Schmidt, M. List, M.H. Schulz, EpiRegio: analysis and retrieval of regulatory elements linked to genes, Nucleic Acids Res. 48 (2020) W193–W199, https://doi.org/10.1093/nar/gkaa382.

[29] G. Mendiratta, E. Ke, M. Aziz, D. Liarakos, M. Tong, E.C. Stites, Cancer gene mutation frequencies for the U.S. population, Nat. Commun. 12 (2021) 5961, https://doi.org/10.1038/s41467-021-26213-y.

[30] M. Krasovec, A. Eyre-Walker, S. Sanchez-Ferandin, G. Piganeau, Spontaneous mutation rate in the smallest photosynthetic eukaryotes, Mol. Biol. Evol. 34 (2017) 1770–1779, https://doi.org/10.1093/molbev/msx119.

[31] T. Fuqua, J. Jordan, M.E. van Breugel, A. Halavatyi, C. Tischer, P. Polidoro, N. Abe, A. Tsai, R.S. Mann, D.L. Stern, J. Crocker, Dense and pleiotropic regulatory information in a developmental enhancer, Nature 587 (2020) 235–239, https://doi.org/10.1038/s41586-020-2816-5.

[32] J. Cui, Y. Yin, Q. Ma, G. Wang, V. Olman, Y. Zhang, W.C. Chou, C.S. Hong, C. Zhang, S. Cao, X. Mao, Y. Li, S. Qin, S. Zhao, J. Jiang, P. Hastings, F. Li, Y. Xu, Comprehensive characterization of the genomic alterations in human gastric cancer, Int. J. Cancer 137 (2015) 86–95, https://doi.org/10.1002/ijc.29352.

[33] C. Chauveau, J. Rowell, A. Ferreiro, A rising titan: TTN review and mutation update, Hum. Mutat. 35 (2014) 1046–1059, https://doi.org/10.1002/humu.22611.

[34] X. Cheng, H. Yin, J. Fu, C. Chen, J. An, J. Guan, R. Duan, H. Li, H. Shen, Aggregate analysis based on TCGA: TTN missense mutation correlates with favorable prognosis in lung squamous cell carcinoma, J. Cancer Res. Clin. Oncol. 145 (2019) 1027–1035, https://doi.org/10.1007/s00432-019-02861-y.

[35] H. Ikeda, T. Kukitsu, W. Johmen, H. Nakamura, N. Yamauchi, K. Ishikawa, T. Saikawa, S. Noda, T. Saitoh, Y. Ueno, Y. Noda, S. Yamazaki, Y. Kuroda, S. Koshiko, Y. Sasagawa, Gastric invasive micropapillary carcinoma with intestinal phenotypes harboring a TP53 R175H mutation, Case Rep. Oncol. 7 (2014) 611–620, https://doi.org/10.1159/000367583.

[36] F. Meric-Bernstam, X. Zheng, M. Shariati, S. Damodaran, C. Wathoo, L. Brusco, M. E. Demirhan, C. Tapia, A.K. Eterovic, R.K. Basho, N.T. Ueno, F. Janku, A. Sahin, J. Rodon, R. Broaddus, T.B. Kim, J. Mendelsohn, K.R. Mills Shaw, D. Tripathy, G. B. Mills, K. Chen, Survival outcomes by TP53 mutation status in metastatic breast cancer, JCO Precis Oncol 2018 (2018), https://doi.org/10.1200/PO.17.00245.

[37] B.M. Lowenthal, T.W. Chan, J.A. Thorson, K.J. Kelly, T.J. Savides, M.A. Valasek, Gastric medullary carcinoma with sporadic mismatch repair deficiency and a TP53 R273C mutation: an unusual case with wild-type BRAF, Case Rep Pathol 2017 (2017), 3427343, https://doi.org/10.1155/2017/3427343.

[38] E. Shinbrot, E.E. Henninger, N. Weinhold, K.R. Covington, A.Y. Goksenin, N. Schultz, H. Chao, H. Doddapaneni, D.M. Muzny, R.A. Gibbs, C. Sander, Z. F. Pursell, D.A. Wheeler, Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication, Genome Res. 24 (2014) 1740–1750, https://doi.org/10.1101/gr.174789.114.

[39] A. Claringbould, J.B. Zaugg, Enhancers in disease: molecular basis and emerging treatment strategies, Trends Mol. Med. 27 (2021) 1060–1073, https://doi.org/10.1016/j.molmed.2021.07.012.

[40] S. Spicuglia, L. Vanhille, Chromatin signatures of active enhancers, Nucleus 3 (2012) 126–131, https://doi.org/10.4161/nucl.19232.

[41] C.T. Ong, V.G. Corces, Enhancer function: new insights into the regulation of tissue-specific gene expression, Nat. Rev. Genet. 12 (2011) 283–293, https://doi.org/10.1038/nrg2957.

[42] J. Ernst, M. Kellis, ChromHMM: automating chromatin-state discovery and characterization, Nat. Methods 9 (2012) 215–216, https://doi.org/10.1038/nmeth.1906.

[43] J.J. Michaelson, Y. Shi, M. Gujral, H. Zheng, D. Malhotra, X. Jin, M. Jian, G. Liu, D. Greer, A. Bhandari, W. Wu, R. Corominas, A. Peoples, A. Koren, A. Gore, S. Kang, G.N. Lin, J. Estabillo, T. Gadomski, B. Singh, K. Zhang, N. Akshoomoff, C. Corsello, S. McCarroll, L.M. Iakoucheva, Y. Li, J. Wang, J. Sebat, Whole-genome sequencing in autism identifies hot spots for de novo germline mutation, Cell 151 (2012) 1431–1442, https://doi.org/10.1016/j.cell.2012.11.019.

[44] B. Schuster-Bockler, B. Lehner, Chromatin organization is a major influence on regional mutation rates in human cancer cells, Nature 488 (2012) 504–507, https://doi.org/10.1038/nature11273.

[45] R. Sabarinathan, L. Mularoni, J. Deu-Pons, A. Gonzalez-Perez, N. Lopez-Bigas, Nucleotide excision repair is impaired by binding of transcription factors to DNA, Nature 532 (2016) 264–267, https://doi.org/10.1038/nature17661.

[46] J.A. Castro-Mondragon, M.R. Aure, O.C. Lingjaerde, A. Langerod, J.W.M. Martens, A.L. Borresen-Dale, V.N. Kristensen, A. Mathelier, Cis-regulatory mutations associate with transcriptional and post-transcriptional deregulation of gene regulatory programs in cancers, Nucleic Acids Res. 50 (2022) 12131–12148, https://doi.org/10.1093/nar/gkac1143.

[47] W.W. Wasserman, A. Sandelin, Applied bioinformatics for the identification of regulatory elements, Nat. Rev. Genet. 5 (2004) 276–287, https://doi.org/10.1038/nrg1315.

[48] S.A. Lambert, A. Jolma, L.F. Campitelli, P.K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T.R. Hughes, M.T. Weirauch, The human transcription factors, Cell 172 (2018) 650–665, https://doi.org/10.1016/j.cell.2018.01.029.

[49] R. Li, L. Di, J. Li, W. Fan, Y. Liu, W. Guo, W. Liu, L. Liu, Q. Li, L. Chen, Y. Chen, C. Miao, H. Liu, Y. Wang, Y. Ma, D. Xu, D. Lin, Y. Huang, J. Wang, F. Bai, C. Wu, A body map of somatic mutagenesis in morphologically normal human tissues, Nature 597 (2021) 398–403, https://doi.org/10.1038/s41586-021-03836-1.

[50] M.M. He, Q. Li, M. Yan, H. Cao, Y. Hu, K.Y. He, K. Cao, M.M. Li, K. Wang, Variant Interpretation for Cancer (VIC): a computational tool for assessing clinical impacts of somatic variants, Genome Med. 11 (2019) 53, https://doi.org/10.1186/s13073-019-0664-4.

[51] I. Bozic, C.J. Wu, Delineating the evolutionary dynamics of cancer from theory to reality, Nat. Can. (Ott.) 1 (2020) 580–588, https://doi.org/10.1038/s43018-020-0079-6.