SCIENTIFIC REPORTS

OPEN

Received: 4 December 2018
Accepted: 17 May 2019
Published online: 03 June 2019

Integrative Analysis of Somatic Mutations in Non-coding Regions Altering RNA Secondary Structures in Cancer Genomes

Funan He[1], Ran Wei[1], Zhan Zhou[2], Leihuan Huang[1], Yinan Wang[1], Jie Tang[1], Yangyun Zou[1], Leming Shi[1,3], Xun Gu[4], Melissa J. Davis[5] & Zhixi Su[1,6]

RNA secondary structure may influence many cellular processes, including RNA processing, stability, localization, and translation. Single-nucleotide variations (SNVs) that alter RNA secondary structure, referred to as riboSNitches, are potentially causative of human diseases, especially in untranslated regions (UTRs) and noncoding RNAs (ncRNAs). The functions of somatic mutations that act as riboSNitches in cancer development remain poorly understood. In this study, we developed a computational pipeline called SNIPER (riboSNitch-enriched or depleted elements in cancer genomes), which employs MeanDiff and EucDiff to detect riboSNitches and then identifies riboSNitch-enriched or riboSNitch-depleted non-coding elements across tumors. SNIPER is available at github: https://github.com/suzhixi/SNIPER/. We found that riboSNitches were more likely to be pathogenic. Moreover, we predicted several UTRs and lncRNAs (long non-coding RNA) that significantly enriched or depleted riboSNitches in cancer genomes, indicative of potential cancer driver or essential noncoding elements. Our study highlights the possibly neglected importance of RNA secondary structure in cancer genomes and provides a new strategy to identify new cancer-associated genes.

The RNA secondary structure plays a crucial role in gene regulation through affecting RNA localization, stability, splicing and translation efficiency. As most of the human genome is transcribed[1], the structure of RNA may profoundly influence the process of post-transcriptional regulation and translation efficiency[2]. Thus, analysis of RNA secondary structure might help us to better understand its molecular and biological role in regulation.

A number of computational and experimental methods have been developed to predict RNA secondary structures or tertiary structures, which also help to identify mutations associated with RNA structures[3–7]. For example, the advent of transcriptome RNA structure probing has enabled researchers to perform transcriptome-wide characterization of RNA secondary structure[8–10]. Through the parallel analysis of RNA structure, a recent study has identified nearly 15% of all transcribed SNVs in a trio family (father, mother, and child) as riboSNitches, which altered local RNA structures. As such mutations are heritable, riboSNitches in human genome should be quite prevalent[9]. The distribution of riboSNitches varies among different elements of the transcripts. While genome-wide studies have demonstrated that riboSNitches were significantly depleted near RNA regulatory elements such as miRNA and protein binding sites[11], recently studies have also revealed that mutations which alter local RNA secondary structures of RNA binding protein (RBP) binding sites may influence their affinity with corresponding RBP[12–16]. These findings suggest that SNVs altering RNA secondary structure may play pivotal roles in gene regulation.

[1]Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, 200433, China. [2]Institute of Drug Metabolism and Pharmaceutical Analysis and Zhejiang Provincial Key Laboratory of Anti-Cancer Drug Research, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, 310058, China. [3]Shanghai Cancer Center and Cancer Institute, Fudan University, Shanghai, 200032, China. [4]Department of Genetics, Development and Cell Biology, Iowa State University, Ames, Iowa, 50011, USA. [5]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC, 3052, Australia. [6]Present address: Singlera Genomics Inc, Shanghai, China. Correspondence and requests for materials should be addressed to Z.S. (email: suzhixi@gmail.com)

RiboSNitches potentially contribute to human diseases, including cancers. It has been suggested that mutations responsible for hyperferritinemia cataract syndrome and retinoblastoma may disrupt gene expression by altering RNA secondary structure[17–19]. RiboSNitches have also been found in RNase MRP lncRNAs and these mutations may be relevant to human cartilage-hair hypoplasia[20]. Recently, researchers also found that a riboSNitch in 3′UTR of *FKBP5* could mediate susceptibility to chronic post-traumatic pain through altering the binding of miR-320a to this gene[21]. In a word, SNVs that disrupt key structural elements of a RNA can result in its dysfunction and cause human disease. As for cancer, some cancer-associated riboSNitches have been identified in non-small cell lung cancers, especially in UTRs and around miRNA binding sites[22], and in retinoblastoma in *RB1* 5′UTR[17].

Many previous studies have been able to discover cancer driver noncoding elements, especially in regulatory regions such as promoters and enhancers[23–27]. A recent pioneer study has predicted the functional impact of mutations based on RNA structural alterations and CADD (Combined Annotation Dependent Depletion) prediction to detect cancer-driver lncRNAs, suggesting that it might be a useful approach to detect driver noncoding elements leveraging the impact of mutations on the RNA secondary structure[24]. Compared with the secondary structure of RNA, sequence conservation is low, and may not be an effective indication of the functional importance of noncoding regions. For instance, although the sequence conservation of lncRNAs is relatively weak in primates, their secondary and tertiary structures are highly conserved[28–30]. Thus, a mutation near such structurally conserved regions is likely to disrupt biological function by altering the local structure. The identification of riboSNitch-enriched or depleted noncoding elements might facilitate the discovery of relevant genes and ncRNAs in cancer and in other diseases as well.

The role of riboSNitches in cancer genomes remains largely unexplored. Therefore, we developed the pipeline SNIPER (riboSNitch-enriched or depleted elements in cancer genomes) to predict riboSNitches and used an empirical substitution model to simulate neutral mutation processes to identify riboSNitch-enriched or depleted noncoding elements in cancer genomes. We only focused on UTRs and lncRNAs in the current study, because of the multiple indistinguishable functional effects of coding region mutations and our limited server computing power. We used this pipeline to conduct a genome-wide analysis to explore the prevalence and the possible function of noncoding riboSNitches in cancer genomes and in tumorigenesis.

Results

MeanDiff and EucDiff are effective approaches to detect riboSNitches. We developed a method to detect riboSNitches. For each SNV, we replaced the corresponding reference allele with the alternative allele to generate a mutated or altered transcript (Fig. 1). Then, the RNA structure predictor were employed to reference and altered transcripts respectively, and by comparing the structural differences between the two transcripts, the impact of this SNV on RNA structure could be estimated. Rather than minimum free energy approaches, we chose the BPPM-based (Base Pairing Probability Matrix) algorithm RNAplfold to predict RNA conformation, as recommended by previous studies[31]. Here, two different methods, MeanDiff and EucDiff, were introduced to detect riboSNitches by calculating the correlation between base pair probabilites of reference and those of mutated transcripts based on RNAplfold (Fig. 1; details in the methods).

To evaluate the performance of our methods, a benchmark dataset of 2,116 SNV-transcript pairs was used, including 1,058 sequences with riboSNitches and 1,058 sequences with non-riboSNitches. Each SNV and its flanking 50 bp sequence was considered as standard input for folding prediction, i.e. 101 base pairs in total[9,31]. Top and bottom 2.5% results were regarded as riboSNitches and non-riboSNitches respectively, as recommended by previous study[31]. For each method, we tested a range of window size (from 2 bp to 50 bp) when calculating BPPM value for both reference and mutated sequence. The maximum window size was set to 50 bp since the input sequences were only 101 bp. We found continuous improvements in area under the ROC curve (AUC) with increasing window size for MeanDiff and EucDiff, and the two methods showed comparable performance (Fig. 2A,B).When the window size was set to 50 bp, the ROC curves illustrated a slightly better performance of MeanDiff (AUC = 0.76) than EucDiff (AUC = 0.75). Comparing with a previous study using SNPfold (AUC = 0.736 at 5% tails)[31], our methods with a window size of 50 bp showed a little improvement (Table 1).

Considering DNA transcription is accompanied by RNA folding, we enlarged the window size to 200 bp to allow for better measurement of possible influence of RNA structure on transcription[32]. In the following analysis, to improve specificity, we determined mutations in both top 2.5% results of MeanDiff and EucDiff as riboSNitches and similarly in both bottom 2.5% results of MeanDiff and EucDiff as non-riboSNitches. Using this criterion, the AUC value increased to 0.774 (Fig. 2C), and it was used in the following analysis.

RiboSNitches are more likely to be pathogenic variants in cancer genomes. To determine whether riboSNitches and non-riboSNitches have different functional impacts, we firstly detected riboSNitches in somatic mutations that we collected from TCGA (The Cancer Genome Atlas), ICGC (International Cancer Genome Consortium) and other previous publications[33,34] (the results are summarized in Supplementary Table S1). Then, we predicted the functional impact scores of somatic mutations using FATHMM-MKL[35]. Given that predicted scores are continuous, we divided them into 5 bins, which were benign, likely benign, potentially pathogenic, likely pathogenic and pathogenic (higher score indicates more pathogenic). We found that the predicted scores of riboSNitches and non-riboSNitches were distributed in all 5 bins with the most significant distribution differences in benign and pathogenic bins (Fig. 3A,B). Overall, the FATHMM scores of riboSNitches are higher than those of non-riboSNitches (Fig. 3C,D). These results imply that somatic mutations that alter RNA secondary structure are more likely to be pathogenic.

To validate this conclusion, we collected a total of 91,183 pathogenic variants and 79,090 benign variants from the ClinVar[36], UniProt[37] and Human Gene Mutation Database (HGMD)[38], to determine whether the riboSNitches constitute a larger proportion of pathogenic variants than non-riboSNitches. As most of the mutations

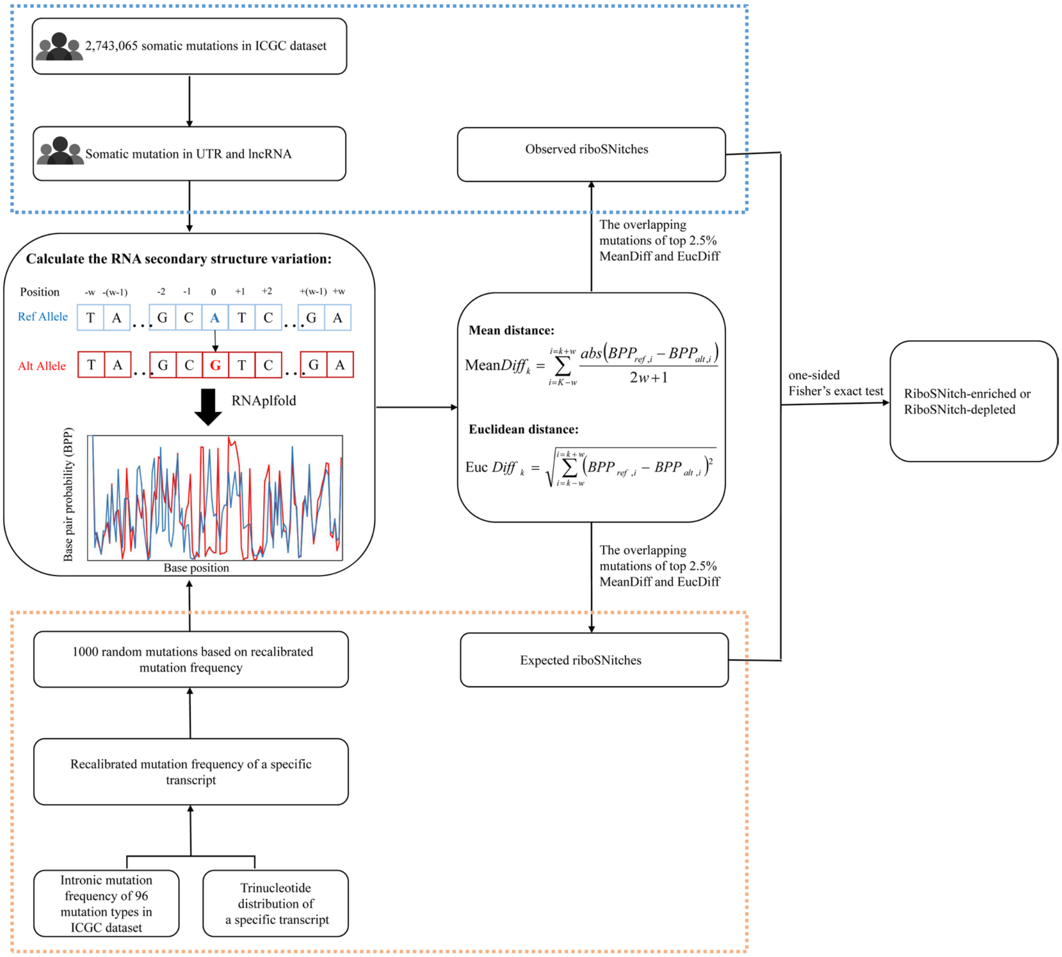
2,743,065 somatic mutations in ICGC dataset
Somatic mutation in UTR and lncRNA
Observed riboSNitches
Calculate the RNA secondary structure variation:
Position -w -(w-1) -2 -1 0 +1 +2 +(w-1) +w
Ref Allele T A . . . G C A T C . . . G A
Alt Allele T A . . . G C G T C . . . G A
RNAplfold
Base pair probability (BPP)
Base position
The overlapping mutations of top 2.5% MeanDiff and EucDiff
Mean distance:
$MeanDiff_k = \sum_{i=K-w}^{i=k+w} \frac{abs(BPP_{ref,i} - BPP_{alt,i})}{2w+1}$
Euclidean distance:
$Euc\ Diff_k = \sqrt{\sum_{i=k-w}^{i=k+w} (BPP_{ref,i} - BPP_{alt,i})^2}$
one-sided Fisher's exact test
RiboSNitch-enriched or RiboSNitch-depleted
The overlapping mutations of top 2.5% MeanDiff and EucDiff
1000 random mutations based on recalibrated mutation frequency
Expected riboSNitches
Recalibrated mutation frequency of a specific transcript
Intronic mutation frequency of 96 mutation types in ICGC dataset
Trinucleotide distribution of a specific transcript

Figure 1. The framework of SNIPER. First, RNA secondary structure was calculated using RNAplfold for ICGC dataset and 1000 randomizations data based on intronic mutation frequency of 96 mutation types and trinucleotide distribution, separately. Then, MeanDiff and EucDiff were used to calculate the structure differences between reference and mutated sequences. Next, mutations in the top 2.5% of both MeanDiff and EucDiff were defined as riboSNitch, and in the bottom 2.5% of both MeanDiff and EucDiff were defined as non-riboSNitch. By comparing the number of observed and expected riboSNitches, riboSNitch-enriched or depleted elements can be detected.

were from normal human samples, we used the MeanDiff and EucDiff cutoff of the 1000 Genome Project to identify riboSNitches and non-riboSNitches in this dataset. As shown in Fig. 3E, pathogenic variants tended to be riboSNitches, and benign variants tended to be non-riboSNitches ($P = 2.87E\text{-}05$, Chi-squared test). Thus, we used these variations which were already known as benign or pathogenic to reconfirm the conclusion that riboSNitches were more likely to be pathogenic.

Features of riboSNitches in the cancer genomes. After dividing somatic mutations into the six substitution subtypes (C > A, C > G, C > T, T > A, T > C, T > G), we found that the value of MeanDiff or EucDiff varied between them, suggesting that different substitution subtypes might have different impacts on the RNA secondary structure (Supplementary Figs S1–S4). Interestingly, C > G mutations were inclined to have a greater impact on the RNA secondary structure (according to the results of MeanDiff and EucDiff) compared with the other mutation types in both the ICGC and TCGA datasets. As for the location, mutations in 5′UTRs had overall higher scores than mutations in 3′UTRs, lncRNAs and protein-coding regions ($P < 2.2e\text{-}16$ for all comparisons, Mann-Whitney test). Intriguingly, more than 80% of the somatic mutations in 5′UTRs were found to be substitutions of GC pairs. One possible explanation is that there are more conserved RNA structures in 5′UTRs, which may imply more functional elements. Because the structure of GC pairs are more stable than AT pairs, this result also suggested that a variety of 5′UTR mutations of a GC pair in the cancer genomes may disrupt the stability of the local RNA structure.

To determine whether riboSNitches are depleted in functional regions in cancer genome, we collected RBP binding data from CLIPdb[39] and miRNA binding targets with high confidence from the TargetScan and miRanda dataset[40,41]. In comparison to non-riboSNitches, riboSNitches were enriched around miRNA binding sites ($P = 5E\text{-}21$, one-sided Fisher's exact test), suggesting that miRNA binding targets might be under positive selection in cancer through altering the RNA secondary structure. In contrast, riboSNitches were significantly

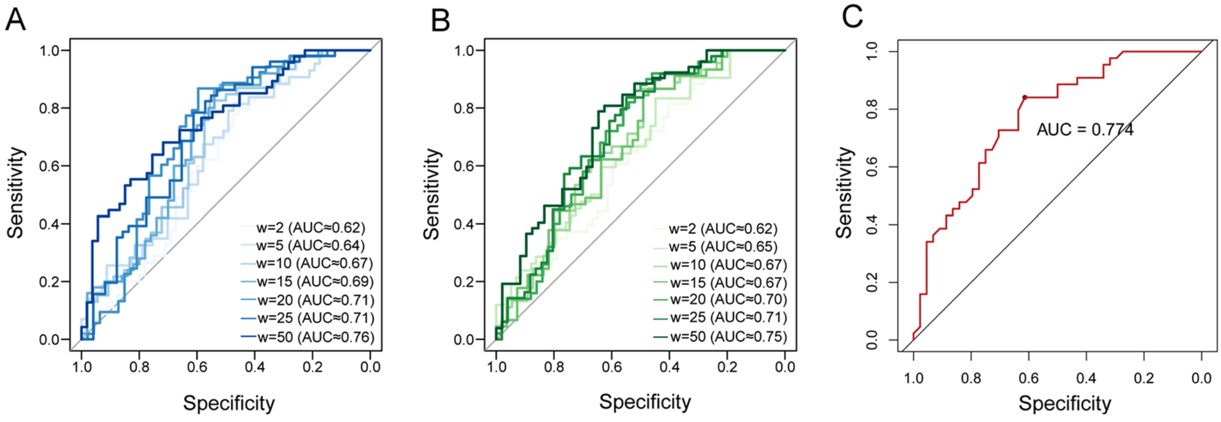

Figure 2. Performance of MeanDiff and EucDiff. ROC curves and AUC values were calculated for benchmark data at 5% tails of MeanDiff (**A**) and EucDiff (**B**) prediction. The color of the curves was shifted from light to dark to represent different window sizes. (**C**) The ROC curve and AUC values of the intersection of the top 2.5% MeanDiff mutations and EucDiff mutations.

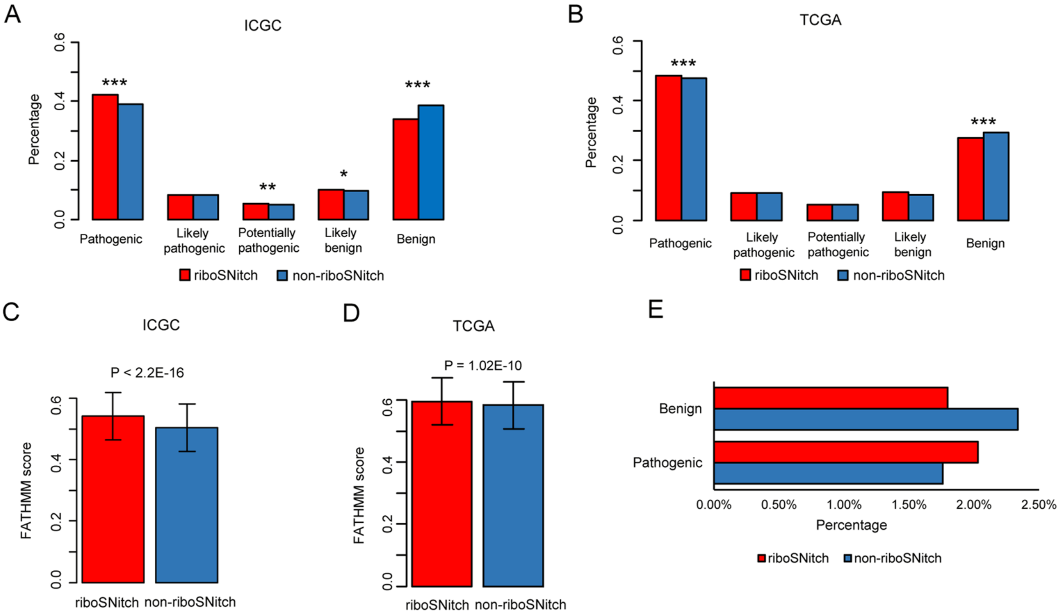

Figure 3. The different functional effects of riboSNitches and non-riboSNitches. (**A**) Functional consequences of riboSNitch (red) and non-riboSNitch (blue) in the ICGC dataset. (**B**) Functional consequences of riboSNitch (red) and non-riboSNitch (blue) in the TCGA dataset. We divided all the mutations into 5 categories based on FATHMM scores. The *P* value was calculated by the Chi-square test. (**C,D**) FATHMM score distribution of riboSNitches and non-riboSNitches in the ICGC and TCGA dataset. The *P* value was calculated by the Mann-Whitney test. (**E**) The different functional effects of riboSNitch and non-riboSNitch in benign and pathogenic variants. The *P* value was calculated by the Chi-square's test.

	w = 2	w = 5	w = 10	w = 15	w = 20	w = 25	w = 50
MeanDiff	0.62	0.64	0.67	0.69	0.71	0.71	0.76
EucDiff	0.62	0.65	0.67	0.67	0.70	0.71	0.75
SNPfold	NA	NA	NA	NA	NA	NA	0.736

Table 1. The AUC values in different window size using MeanDiff, EucDiff and SNPfold. *NA represents the results not provided by Corley *et al*[31].

depleted around RBP binding sites ($P = 1.79E\text{-}07$, one-sided Fisher's exact test) in the cancer genome, consistent with a previous study of a trio family[9], suggesting that RBP binding targets were under purifying selection in cancer genomes to maintain a constrained structure. Those results are consistent with the knowledge that RNA secondary structure plays an essential role in RNA regulation, especially in the interactions with miRNAs and RNA-binding proteins. Therefore, we suggest that further research should be conducted to discover whether the expression of these riboSNitch-enriched or depleted genes is regulated by altered RNA secondary structure in cancer.

A computational framework for predicting riboSNitch-enriched or riboSNitch-depleted genes.

To explore the potential biological function of riboSNitches in cancer development, we developed a computational pipeline called SNIPER to identify riboSNitch-enriched or depleted genes based on the dataset of cancer somatic mutations in noncoding region (Fig. 1). We hypothesized that riboSNitch-enriched genes would display an excess of riboSNitches compared with expectation, demonstrating that these genes have undergone positive selection for RNA shape during cancer progression, which could be regarded as an evolutionary process. To find those riboSNitch-enriched genes, we used MeanDiff and EucDiff to tally the number of riboSNitches in ICGC and previous studies at first. Then, we constructed a neutral mutation model based on the cancer intronic mutation profile and the tri-nucleotide context of each transcript to calculate the expected number of riboSNitches as described in Methods. Finally, through the comparison of observed riboSNitch ratio with the simulated result, we can identify riboSNitch-enriched or depleted genes (Fig. 1). Unlike previous methods, we used intronic instead of exonic mutation rate to simulate expected number of riboSNitches, and thus riboSNitch-enriched genes detected by SNIPER could be regarded as positively selected genes in cancer. As shown above, riboSNitches were more likely to be pathogenic, so we conjectured that these riboSNitch-enriched genes might play a nonnegligible role in cancer progression. On the other hand, riboSNitch-depleted genes could be structurally conserved genes in cancer, namely essential genes and could play a fundamental role in cancer process. Therefore, our approach may help us to identify potentially cancer-associated genes involved in tumorigenesis and development.

Identification of riboSNitch-enriched elements to find cancer-related genes.

To detect cancer-related candidate genes, our method was applied to the ICGC dataset in UTRs to comprehensively discover riboSNitch-enriched or depleted elements of protein-coding genes. The alteration of structure of UTRs might impact gene expression through altering miRNA or RBP binding, and thus contributes to oncogenesis. In our analysis, both MeanDiff and EucDiff were used to measure the impact of mutations on RNA secondary structure based on the results computed by RNAplfold, and the parameter was set as recommended[32]. RiboSNitches were identified as the intersection of the top 2.5% mutations of MeanDiff and EucDiff. Finally, statistical significance was calculated with Fisher's exact test and corrected using the Benjamini and Hochberg method[42]. For protein-coding genes, we only detected cancer-related functional candidate UTRs because the RNA structure of the 3′UTRs and 5′UTRs are crucial for gene regulation and translational stability. Coding region mutations were excluded for their complex effects, and the effect of altering the RNA secondary structure might be masked by other influences.

Firstly, we applied our method to 5′UTRs to discover putative cancer driver elements. Here, for the genes with riboSNitch-enriched 5′UTRs, we found 224-fold enrichment of Cancer Gene Census (CGC) genes over random expectation ($P < 2.2e\text{-}16$, Chi-squared test). In this case, the 5′UTRs of *KAT6A* and *NOTCH2* were identified at a q-value cutoff of 0.05, and both genes were discovered in the CGC, and are regarded as known cancer genes (Fig. 4A). *NOTCH2* is regarded as an oncogene, and plays an essential role in cancer signaling pathways[43]. *KAT6A* is a lysine acetyltransferase gene that has been shown to be involved in cell growth of luminal breast cancer in a previous study[44]. In addition, *RALGPS2* was identified as a putative driver when reducing the cutoff of q-value to 0.2, and *RALGPS2* is involved in cell survival and associated with the cell cycle in lung cancer cells[45]. We found that *NOTCH2*, *KAT6A* and *RALGPS2* are all potential cancer driver genes, demonstrating that this method can help us to find RNA structure-related cancer driver elements.

Next, we also employed our approach to identify riboSNitch-enriched elements in 3′UTRs. As in the case of 5′UTRs, 7 genes with riboSNitch-enriched 3′UTRs were identified using SNIPER, including *CLCNKB*, *CYP4B1*, *SLC9B1*, *CCDC104*, *POLR2M*, *ACAD11* and *DIO1*, at the q-value cutoff of 0.05 (Fig. 4B). *CYP4B1* is a cytochrome enzyme, which has been found to be highly expressed in bladder tumor patients[46]. *SLC9B1* is a Na+/H+ transporter, which contributes to the maintenance of cellular homeostasis[47]. *POLR2M*, the RNA polymerase II subunit M, plays a crucial role in gene transcription, and known as a candidate driver gene involved in the progression of prostate cancer[48]. *ACAD11* is a gene in the acyl-dehydrogenase family, which is involved in cell survival and plays a key role in the pro-survival function of *TP53*[49]. The *DIO1* gene encodes a type I iodothyronine deiodinase, an important regulator of cell proliferation, differentiation and metabolism[50]. Additionally, the 3′UTRs of *SMO*, *SRPK1*, *FOXD4* and *DBP* were also identified as riboSNitch-enriched elements when reducing the cutoff of q-value to 0.2. Of these, *SRPK1* has been shown to have a tumor-suppressive effect and is a candidate driver gene[51,52].

To determine whether the elements identified above are cancer-specific riboSNitch-enriched elements, we compared the observed elements in the cancer genomes to those in the germline dataset. We found that two riboSNitch-enriched 5′UTRs were cancer-specific, which were *KAT6A* and *RALGPS2*. All the 3′UTRs identified above were found to be cancer-specific riboSNitch-enriched elements, except for *CLCNKB* and *SMO*. These results suggest that cancer-specific riboSNitch-enriched elements might have the potential to be cancer-related functional elements or putative driver elements.

Last, for lncRNAs, at a q-value cutoff of 0.05, only *USP30-AS1* showed a significant enrichment for riboSNitches using our method, but it is not a cancer-specific lncRNA. When relaxing the q-value cutoff to 0.1, *LINC01365*, *ZNF503-AS1* and *LINC00689* were identified as riboSNitch-enriched, among which *ZNF503-AS1*

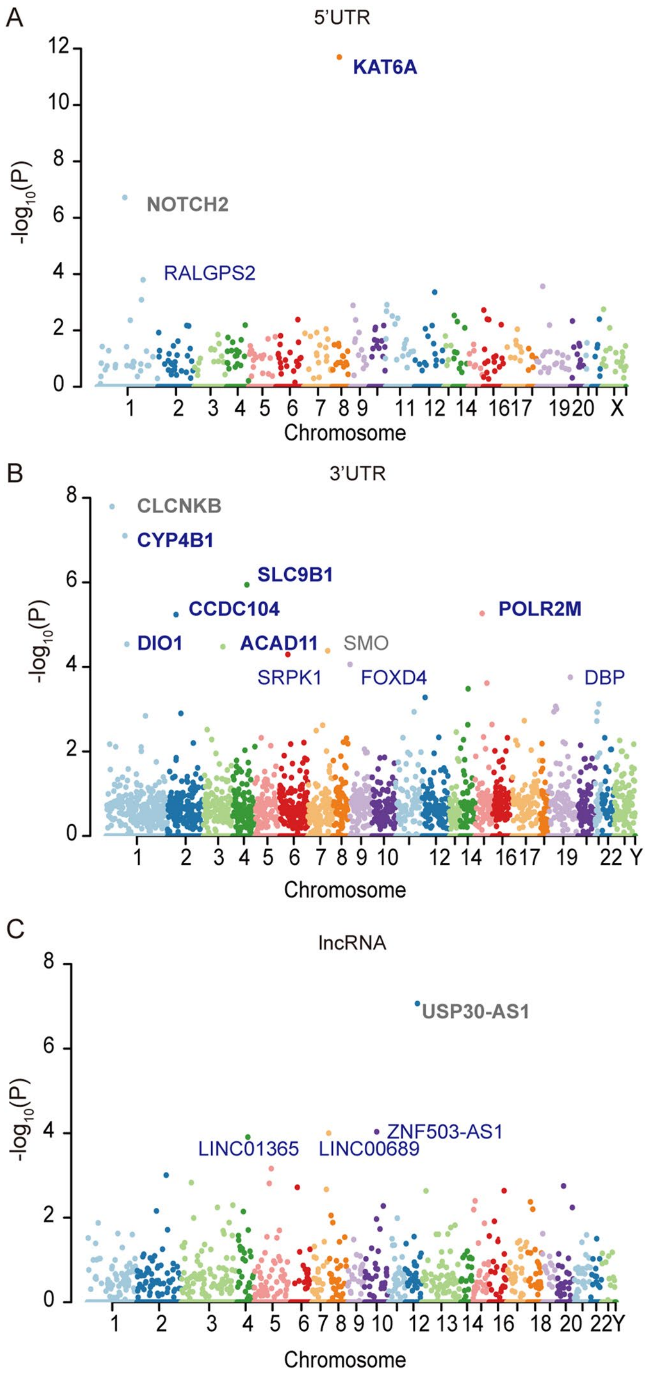

Figure 4. RiboSNitch-enriched elements in the cancer genome. A Manhattan plot representing RiboSNitch-enriched 5′UTRs (**A**), 3′UTRs (**B**), and lncRNAs (**C**) with the most significant *P* values. All the identified elements with an FDR < 0.2 are listed in the plot. Genes in bold represent an FDR < 0.05. Genes in blue indicate that this gene was identified as a cancer-specific enriched element.

and *LINC00689* were cancer-specific (Fig. 4C). As for *ZNF503-AS1*, it is intriguing that it can promote the proliferation and migration of pigment epithelium by regulating *ZNF503*, and expression of *ZNF503-AS1* has been shown to be a prognostic indicator for lung squamous cell carcinoma[53,54]. From FuncPred, a website for the prediction of lncRNA function[55], we found that three of the lncRNAs (*USP30-AS1*, *ZNF503-AS1* and *LINC00689*) predicted above were related to the cancer process with FDR < 0.05[55].

Additionally, considering that less than 2% of somatic mutations were identified as riboSNitches under the criterion of the intersection of the top 2.5% mutations of MeanDiff and EucDiff, we relaxed the cutoff of MeanDiff and EucDiff to 5%, 10% and 20%. All the riboSNitch-enriched elements identified by our approach at different cutoffs were listed in Supplementary Table S2.

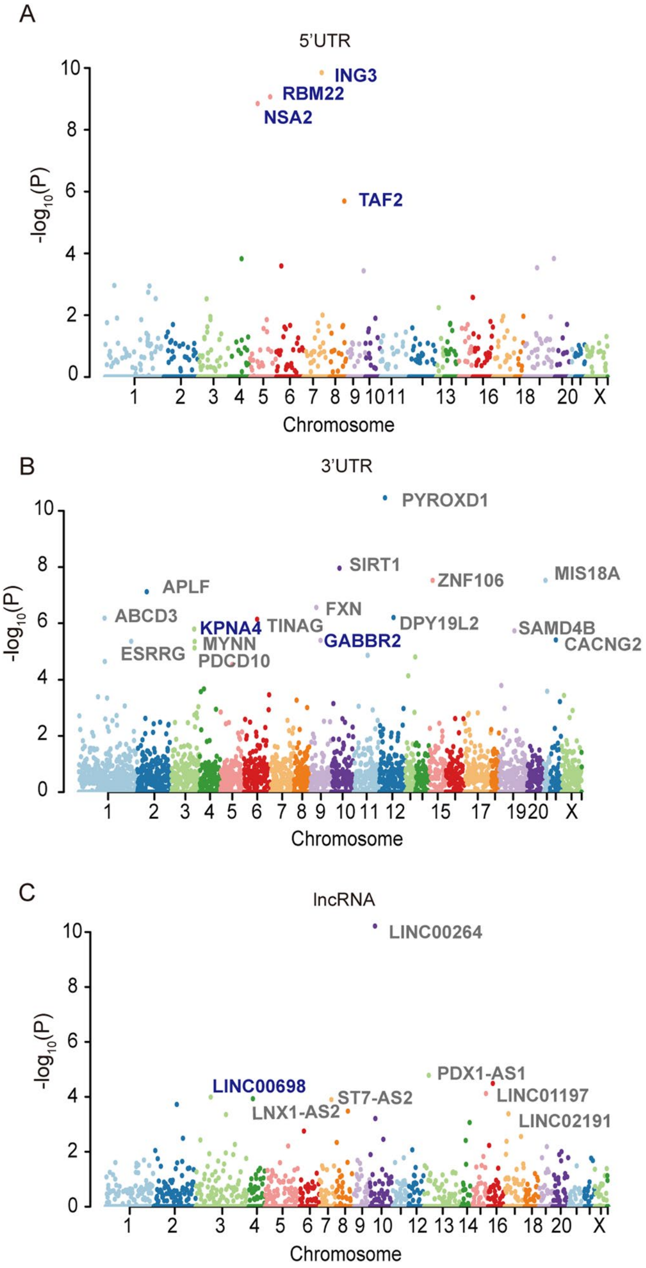

Figure 5. RiboSNitch-depleted elements in the cancer genome. A Manhattan plot representing RiboSNitch-depleted 5′UTRs (**A**), 3′UTRs (**B**), and lncRNAs (**C**) with the most significant *P* values. All the identified elements with an FDR < 0.05 are listed in the plot. Genes in blue indicate that this gene was identified as a cancer-specific depleted element.

Identification of riboSNitch-depleted elements to find structural conserved elements in cancer. Our approach can also be used to predict riboSNitch-depleted elements in cancer genomes. These elements might be essential for cancer due to their constrained RNA secondary structure. We identified riboSNitch-depleted elements, including 5′UTRs, 3′UTRs and lncRNAs (Supplementary Table S3). Four riboSNitch-depleted 5′UTRs (*ING3*, *RBM22*, *NSA2* and *TAF2*) were detected at the q-value cutoff of 0.05, and all of these elements were cancer-specific (Fig. 5A). We also discovered 22 riboSNitch-depleted 3′UTRs, but only two of them (*KPNA4* and *GABBR2*) were identified as cancer-specific (Fig. 5B). Among the six cancer-specific riboSNitch-depleted elements, five (*ING3*, *RBM22*, *NSA2*, *TAF2* and *KPNA4*) were found to be conditionally essential in cancer cell lines in the OGEE v2 database[56]. Additionally, seven riboSNitch-depleted lncRNAs were identified at the q-value cutoff of 0.05, but only *LINC00698* was cancer-specific (Fig. 5C). Our results imply that cancer-specific riboSNitch-depleted regions might be potentially cancer essential elements.

Discussion

Our study provides a new strategy to investigate cancer somatic mutations in noncoding region from the perspective of RNA shape, and it suggests the importance of RNA secondary structure in cancer possibly by regulating the expression of genes. To our knowledge, this is the first comprehensive study to analyze riboSNitches in somatic mutations across cancer genomes based on the two main international cancer genomics repositories. In our analyses across mutations in cancer genomes, different mutation subtypes exhibited different impacts on RNA structure, and the riboSNitches were enriched around binding targets of miRNAs and depleted around those of RBPs. Those results suggest that some somatic mutations have the potential to profoundly influence the RNA secondary structures, and may further regulate gene expression[11,13,21].

A myriad of experimental technologies has been developed to probe and analyze RNA structure, yet accurately characterizing RNA structure is still extremely difficult for the high dynamics of RNA *in vivo* and *in vitro*. Moreover, diverse computational approaches have been designed to predict the impact of a single mutation by integrating empirical data. It should be noted that predicting riboSNitches basing on experimental data was more accurate than computational methods[57]. However, for lack of experimental data of the cancer genome, it is acceptable to use computational methods to predict riboSNitches among cancer somatic mutations. In SNIPER pipeline, we detected riboSNitches using RNAplfold and a combination of MeanDiff and EucDiff, and found that the performance of our *in silico* method was better than SNPfold, which has been proven as one of the best methods listed in a previous study[31]. Though it is viable to use a computational approach to determine riboSNitches in the cancer genome, further experiments are needed to explore the functional relevance of variant-induced RNA structural alterations and their biological mechanisms responsible for disease phenotype.

In the present study, we found that riboSNitches were more likely to be pathogenic and may increase the risk of disease. Based on this, we proposed a new pipeline SNIPER to detect riboSNitches and to identify riboSNitch-enriched or depleted noncoding elements in cancer genome. First, we constructed an expected riboSNitch distribution under a neutral mutation model based on the cancer intronic mutation profile and the tri-nucleotide context of each transcript. Then, by comparing the observed riboSNitch ratio with expected, riboSNitch-enriched or riboSNitch-depleted elements could be predicted. Given the use of intronic mutation rate as a neutral background model in our method, we could detect elements under positive selection in cohorts of tumors. As shown in our results, riboSNitch-enriched UTRs could be potential cancer drivers, and riboSNitch-depleted UTRs were prone to be essential genes in cancer, especially cancer-specific ones. In addition, we identified several cancer-related lncRNAs, and the actual function of these lncRNAs in cancer progression needs to be further investigated. Above all, our method allowed us to find riboSNitch-enriched and riboSNitch-depleted noncoding elements in the cancer genome and could help us to identify more cancer driver and essential genes.

Previous methods that have been designed to identify cancer driving elements in noncoding regions tended to discover positive selection signals by comparing mutation rates between target sequences and corresponding flanking regions in noncoding sequences, especially for regulatory elements[24]. In our study, we developed a new method to identify positive selection signals by detecting the relative impact of somatic mutations located in noncoding regions on the RNA secondary structure. Although most of the human genome can be transcribed, we only focused on changes of RNA secondary structure of UTRs and lncRNAs, resulting in only a small number of mutations. Despite this limitation, SNIPER was able to identify regions significantly enrich or deplete riboSNitches in cancer genomes, and we effectively identified several candidate driver and essential elements.

Next-generation sequencing technologies have enabled genome-wide analysis of variations in human genomes, greatly enhancing our understanding of RNA structure-related variations. With the amount of cancer genome sequencing data accumulates, we can further analyze riboSNitch-enriched or depleted elements for each cancer type. In the meanwhile, although many studies have been conducted to predict potentially functional lncRNAs[55,58,59], the molecular functions of these lncRNAs remain to be explored.

Given the complexity of coding region mutations and the limited computing power of our servers, our analysis mainly focused on the noncoding regions, namely UTRs and lncRNAs. Providing an initial analysis of riboSNitches in the cancer genome, we found that they are more likely to be pathogenic variants. Additionally, we also pave the way to explore potentially functional noncoding regions in the cancer genome from the perspective of RNA secondary structure. Although we have highlighted the potential effect of riboSNitches in the cancer cohort, it remains a challenge to validate whether such mutations are involved in tumorigenesis or play a role in post-transcriptional regulation and gene translation in cancer.

Materials and Methods

Datasets. The majority of cancer somatic mutation data used in this study were obtained from the ICGC and TCGA data portal. We also retrieved somatic mutations of 25 whole-genome sequenced melanomas[33] and 100 whole-genome sequenced gastric cancers from previous studies[34]. The germline mutation dataset was retrieved from the 1000 Genome Project phase 3 data[60].

First, we excluded all small insertions and deletions, and only single-nucleotide variations were retained for further analysis. Then, all the coordinates of mutations based on hg38 were lifted to hg19 using the UCSC liftOver toolkit[61]. To remove SNVs with low confidence, both cancer somatic mutations and germline mutations from the 1000 Genome Project were filtered against the hg19 signal artifact blacklist; the merged blacklist was downloaded from the Broad Institute (https://personal.broadinstitute.org/anshul/projects/encode/rawdata/blacklists). This list is a comprehensive collection of signal artifact blacklist regions in hg19, which included regions with abnormal mapping, repeat elements, and an enrichment of ultra-high frequency artifacts.

We used FATHMM-MKL[35] to predict the potential effects of riboSNtiches and non-riboSNitches. According to the pathogenic score, we divided all the variants into 5 categories: benign (score $\in$ [0, 0.2], likely benign (score ∈ (0.2, 0.4]), potentially pathogenic (score ∈ (0.4, 0.6]), likely pathogenic (score ∈ (0.6, 0.8]) and pathogenic (score ∈ (0.8, 1]).

The miRNA-target interactions were obtained from the TargetScan (release 7.1) and miRanda-mirSVR (August 2010 release)[40,41]. Only binding sites with high confidence were selected in our analysis. For TargetScan, conserved targets or targets of conserved miRNA families with PCT ≥ 90 were used. For miRanda, the conserved miRNAs with a high mirSVR score (cutoff was set to 1) were used. In addition, we collected all the CLIP-seq data for the HeLa cell line from CLIPdb[39], which included the interaction regions with different RBPs detected by PiRaNhA[62].

The known cancer genes were retrieved from CGC of COSMIC (Catalogue of Somatic Mutations in Cancer) tier 2[63], and the cancer-related lncRNAs were downloaded from the Lnc2Cancer database[64].

Gene annotation. The coordinates of all transcripts were obtained from the ENCODE website (https://www.encodeproject.org/); specifically, GENCODE[65] release 19 annotation in GTF format was downloaded at https://www.gencodegenes.org. From the GTF file, genes were assigned to protein coding genes (PCGs), pseudogenes, lncRNAs and other small noncoding genes by the "gene_type" catalogue. For more accurate annotation, we obtained the human protein-coding gene list and long noncoding RNA list from HGNC[66]. Finally, a total of 19,035 PCGs and 3,435 lncRNAs were included in our analysis.

It is noteworthy that one gene may generate multiple transcripts by alternative splicing with different RNA secondary structures. To reduce such uncertainty, one principal transcript of each gene was selected for RNA structure prediction. To obtain the principal transcript, we carried out the following steps. (1) The multi-transcript protein-coding genes were first ranked by annotating the principal splice isoform (APPRIS) level, which provides a reliable classification scheme for the transcript isoforms of the alternatively spliced genes in the human genome[67]. The APPRIS level ranges from 1 to 5, with 1 being the most reliable. (2) If transcripts of a gene were ranked equivalently, then the transcript with the CCDS ID was regarded as the more reliable one[68]. (3) In addition, the annotation level of transcripts was ranked from 1 to 3, with 1 being the most stable. (4) If the principal transcript could not be selected from all the above-described methods, the longest was regarded as the principal transcript. Thus, the principal transcripts were generated in the following priority order: APPRIS > CCDS > transcript level > transcript length.

After selecting the principal transcript of each gene, the mutations that passed the filtering were annotated. Additionally, we only considered the RNA secondary structure of mature transcript in the present study. Finally, we found 3,332,314 unique cancer somatic mutations from TCGA, ICGC and previous studies[33,34] and 1,917,818 germline mutations from the 1000 Genome dataset.

RNA secondary structure prediction and riboSNitch detection. RNAplfold (http://www.tbi.univie.ac.at/RNA/) can generate an average base pair probability for each site for a given sequence. With the output base pairing probability matrices, it is straightforward to detect the difference in RNA structure between the wild-type and mutant. Therefore, we used RNAplfold, which is a locally stable secondary structure prediction toolkit of the ViennaRNA package[69], to calculate local RNA secondary structures.

RiboSNitches are defined as SNVs that have a great impact on the local RNA secondary structure[9]. To identify the impact of a given somatic mutation, we can calculate the RNA structure alteration between the tumor sequence and paired normal sequence. For lack of raw sequencing data and control the effects of other confounding factors such as genetic variations in each sample, we assumed that the normal sequences are identical to the reference genome and the tumor sequences are the same except for the mutation. Reference sequences of each transcript were extracted by BEDTools getfasta using the GENCODE v19 annotation[70]. The corresponding tumor sequences were obtained by replacing the reference base with the mutated one. Then, RNAplfold was applied to both normal and cancer sequences to predict the base pair probabilities at each site. In our study, we only predicted the RNA secondary structure of mature transcripts, and intron sequences were excluded.

After calculating the RNA secondary structure variation, we computed the differences in base pair probabilities between reference and mutated sequences using MeanDiff and EucDiff (Fig. 1). Of note, the structural alterations were not restricted to a single base, and thus we calculated the alteration of base pair probability in a w bp window size around the mutation site. The equations for the MeanDiff and EucDiff are as follows:

$$MeanDiff_k = \sum_{i=k-w}^{i=k+w} \frac{abs(\mathrm{BPP}_{ref,i} - \mathrm{BPP}_{alt,i})}{2\mathrm{w} + 1} \tag{1}$$

$$EucDiff_k = \sqrt{\sum_{i=k-w}^{i=k+w} (\mathrm{BPP}_{\mathrm{ref},i} - \mathrm{BPP}_{\mathrm{alt},i})^2} \tag{2}$$

where k is the position of the mutation in the transcript, w is the window size, and $\mathrm{BPP}_{\mathrm{ref,\,i}}$ and $\mathrm{BPP}_{\mathrm{alt,\,i}}$ represent the i th base pair probability of the reference and mutated sequence, which were computed using RNAplfold. According to previous study[32], we set the parameter of window size as 200 bp to predict the probabilities of local RNA secondary structure. In order to improve the accuracy of prediction, we defined riboSNitch as the SNVs belonging to the intersection of top 2.5% MeanDiff and EucDiff, and non-riboSNitch as it belonging to both bottom 2.5%.

We further mapped all the riboSNitches and non-riboSNitches to the binding sites of miRNA and RBPs. In our analysis, mutations located in or around binding sites (±~20 bp) were determined to have the potential to influence the binding of miRNA and RBP.

Evaluation of the performance of MeanDiff and EucDiff using the benchmark dataset. To evaluate our methods and determine a proper way to compute RNA structure alterations due to somatic mutations, we used a benchmark dataset of identified riboSNitch and non-riboSNitch sequences, which were previously detected by parallel analysis of RNA structure method. The dataset, including 1058 riboSNitch and 1058 non-riboSNitch with sequences of 101 bp, was retrieved from a previous study[31]. RNAplfold was used to calculate the RNA structures for all sequences, and the differences in base pair probabilities of mutations were calculated using MeanDiff and EucDiff with different window size of 2 bp, 5 bp, 10 bp, 15 bp, 20 bp, 25 bp and 50 bp (the maximum window size allowed for the given sequences). As recommended, the tail 5% of MeanDiff or EucDiff were regarded as riboSNitches and non-riboSNitches[31]. The receiver operating characteristic (ROC) curve was calculated based on the benchmark dataset to evaluate MeanDiff and EucDiff using different window sizes. The ROC curve plot was computed with the R package "pROC"[71].

Detection of riboSNitch-enriched or depleted elements. Cancer development is an evolutionary process, and as a vast number of somatic mutations are neutral mutations, we used mutation-profile-based random mutation procedure to simulate neutral mutations in cancer as the expected mutations. The neutral mutation rate can be calculated from intergenic or intron regions because such regions are less likely to be under the selective pressure than transcribed regions. Hence, in our study, we firstly computed the intronic mutation rate of each mutation type as the background mutation profile. As there were insufficient intronic mutations in the TCGA dataset, we calculated the intronic mutation frequency only in the ICGC dataset.

Considering different transcripts with different contexts, we also calculated the tri-nucleotide context of each transcript. To simulate neutral mutation in cancer, we used the cancer intronic mutation profile to represent the neutral mutation rate in cancer genome. Considering the mutation rate is related to mutation types and the sequence context, we taken the 5′ base and 3′base into account. Thus, we computed 96 possible mutation profile of intron (6 mutation types $*$ 4 types of 5′ base $*$ 4 types of 3′ base) and 32 corresponding tri-nucleotide composition of each transcript (2 types of base $*$ 4 types of 5′ base $*$ 4 types of 3′ base). Based on the intronic mutation frequency of 96 mutation types in ICGC dataset and trinucleotide distribution, we recalibrated mutation frequency of a specific transcript. After that, through simulating random sampling, we generated 1000 random mutations of each transcript. Then, we obtained the expected number of riboSNitches after using RNAplfold to calculate the RNA secondary structure and using MeanDiff and EucDiff to computed the difference of base pair probability between reference and mutated sequences. Similarly, the intersection of the top 2.5% mutations were considered as riboSNitches.

The number of riboSNitches in these 1000 randomizations was regarded as an expected value. In addition, we tally the riboSNitches and total mutations in the cancer mutation dataset as observed value. Mutations that occurred at the same position in different patients were counted independently. After the observed riboSNitches and expected number of riboSNitches were calculated, a one-sided Fisher's exact test was conducted to identify genes that were significantly enriched or depleted the riboSNitches compared with the expected riboSNitches, and the resulting *P* value were corrected using the Benjamini and Hochberg method[42] (Fig. 1). All statistical analyses were conducted in R. The results with a significant *P* value after the correction for the false discovery rate were regarded as putative riboSNitch-enriched or depleted candidates. Considering the complexity of coding region mutations, we only applied our approach to 5′UTRs, 3′UTRs as well as lncRNAs to discover riboSNitch-enriched or depleted candidates.

Cancer-specific riboSNitch-enriched or depleted elements. To search for cancer-specific elements, we compared the riboSNitches in the cancer genome to in the 1000 Genome dataset. A one-sided Fisher's exact test was used to determine whether riboSNitches were enriched in regions of the cancer genome. All elements with *P* value less than 10^{-3} were regarded as cancer-specific elements.

References

1. Wong, G. K. S., Passey, D. A. & Yu, J. Most of the human genome is transcribed. *Genome Res* **11**, 1975–1977, https://doi.org/10.1101/gr.202401 (2001).
2. Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* **15**, 469–479, https://doi.org/10.1038/nrg3681 (2014).
3. Ouyang, Z. Q., Snyder, M. P. & Chang, H. Y. SeqFold: Genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res* **23**, 377–387, https://doi.org/10.1101/gr.138545.112 (2013).
4. Yao, J., Reinharz, V., Major, F. & Waldispuhl, J. RNA-MoIP: prediction of RNA secondary structure and local 3D motifs from sequence data. *Nucleic acids research* **45**, W440–W444, https://doi.org/10.1093/nar/gkx429 (2017).
5. Underwood, J. G. *et al.* FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Methods* **7**, 995–U981, https://doi.org/10.1038/Nmeth.1529 (2010).
6. Sabarinathan, R. *et al.* RNAsnp: Efficient Detection of Local RNA Secondary Structure Changes Induced by SNPs. *Hum Mutat* **34**, 546–556, https://doi.org/10.1002/humu.22273 (2013).
7. Lucks, J. B. *et al.* Multiplexed RNA structure characterization with selective 2 ′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *P Natl Acad Sci USA* **108**, 11063–11068, https://doi.org/10.1073/pnas.1106501108 (2011).
8. Mustoe, A. M. *et al.* Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. *Cell* **173**, 181–195, https://doi.org/10.1016/j.cell.2018.02.034 (2018).
9. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–709, https://doi.org/10.1038/nature12946 (2014).
10. Bai, Y. H., Dai, X. Z., Harrison, A., Johnston, C. & Chen, M. Toward a next-generation atlas of RNA secondary structure. *Brief Bioinform* **17**, 63–77, https://doi.org/10.1093/bib/bbv026 (2016).
11. Luo, Z., Yang, Q. & Yang, L. RNA Structure Switches RBP Binding. *Mol Cell* **64**, 219–220, https://doi.org/10.1016/j.molcel.2016.10.006 (2016).
12. Lambert, N. *et al.* RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell* **54**, 887–900, https://doi.org/10.1016/j.molcel.2014.04.016 (2014).

13. Taliaferro, J. M. *et al*. RNA Sequence Context Effects Measured *In Vitro* Predict *In Vivo* Protein Binding and Regulation. *Mol Cell* **64**, 294–306, https://doi.org/10.1016/j.molcel.2016.08.035 (2016).
14. Lewis, C. J., Pan, T. & Kalsotra, A. RNA modifications and structures cooperate to guide RNA-protein interactions. *Nature reviews. Molecular cell biology* **18**, 202–210, https://doi.org/10.1038/nrm.2016.163 (2017).
15. Dominguez, D. *et al*. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell* **70**, 854–867 e859, https://doi.org/10.1016/j.molcel.2018.05.001 (2018).
16. Taylor, K. *et al*. MBNL splicing activity depends on RNA binding site structural context. *Nucleic acids research* **46**, 9119–9133, https://doi.org/10.1093/nar/gky565 (2018).
17. Kutchko, K. M. *et al*. Multiple conformations are a conserved and regulatory feature of the RB1 5 ′ UTR. *Rna* **21**, 1274–1285, https://doi.org/10.1261/rna.049221.114 (2015).
18. Martin, J. S. *et al*. Structural effects of linkage disequilibrium on the transcriptome. *Rna-a Publication of the Rna Society* **18**, 77–87, https://doi.org/10.1261/rna.029900.111 (2012).
19. Halvorsen, M., Martin, J. S., Broadaway, S. & Laederach, A. Disease-Associated Mutations That Alter the RNA Structural Ensemble. *Plos Genet* **6**, https://doi.org/10.1371/journal.pgen.1001074 (2010).
20. Rogler, L. E. *et al*. Small RNAs derived from lncRNA RNase MRP have gene-silencing activity relevant to human cartilage-hair hypoplasia. *Hum Mol Genet* **23**, 368–382, https://doi.org/10.1093/hmg/ddt427 (2014).
21. Linnstaedt, S. D. *et al*. A Functional riboSNitch in the 3 ′Untranslated Region of FKBP5 Alters MicroRNA-320a Binding Efficiency and Mediates Vulnerability to Chronic Post-Traumatic Pain. *J Neurosci* **38**, 8407–8420, https://doi.org/10.1523/Jneurosci.3458-17.2018 (2018).
22. Sabarinathan, R. *et al*. Transcriptome-Wide Analysis of UTRs in Non-Small Cell Lung Cancer Reveals Cancer-Related Genes with SNV-Induced Changes on RNA Secondary Structure and miRNA Target Sites. *Plos One* **9**, https://doi.org/10.1371/journal.pone.0082699 (2014).
23. Lanzos, A. *et al*. Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. *Sci Rep-Uk* **7**, https://doi.org/10.1038/srep41544 (2017).
24. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol* **17**, https://doi.org/10.1186/s13059-016-0994-0 (2016).
25. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature genetics* **46**, 1258–1263, https://doi.org/10.1038/ng.3141 (2014).
26. Ashouri, A. *et al*. Pan-cancer transcriptomic analysis associates long non-coding RNAs with key mutational driver events. *Nature communications* **7**, https://doi.org/10.1038/ncomms13197 (2016).
27. Abeshouse, A. *et al*. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025, https://doi.org/10.1016/j.cell.2015.10.025 (2015).
28. Novikova, I. V., Hennelly, S. P. & Sanbonmatsu, K. Y. Sizing up long non-coding RNAs: do lncRNAs have secondary and tertiary structure? *Bioarchitecture* **2**, 189–199, https://doi.org/10.4161/bioa.22592 (2012).
29. Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. *Cell* **147**, 1537–1550, https://doi.org/10.1016/j.cell.2011.11.055 (2011).
30. Seemann, S. E. *et al*. The identification and functional annotation of RNA structures conserved in vertebrates. *Genome Res* **27**, 1371–1383, https://doi.org/10.1101/gr.208652.116 (2017).
31. Corley, M., Solem, A., Qu, K., Chang, H. Y. & Laederach, A. Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. *Nucleic acids research* **43**, 1859–1868, https://doi.org/10.1093/nar/gkv010 (2015).
32. Lange, S. J. *et al*. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic acids research* **40**, 5215–5226, https://doi.org/10.1093/nar/gks181 (2012).
33. Berger, M. F. *et al*. Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485**, 502–506, https://doi.org/10.1038/nature11071 (2012).
34. Wang, K. *et al*. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature genetics* **46**, 573–582, https://doi.org/10.1038/ng.2983 (2014).
35. Shihab, H. A. *et al*. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536–1543, https://doi.org/10.1093/bioinformatics/btv009 (2015).
36. Landrum, M. J. *et al*. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **42**, D980–D985, https://doi.org/10.1093/nar/gkt1113 (2014).
37. Apweiler, R. *et al*. UniProt: the Universal Protein knowledgebase. *Nucleic acids research* **32**, D115–D119, https://doi.org/10.1093/nar/gkh131 (2004).
38. Stenson, P. D. *et al*. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**, 1–9, https://doi.org/10.1007/s00439-013-1358-4 (2014).
39. Yang, Y. C. T. *et al*. CLIPdb: a CLIP-seq database for protein-RNA interactions. *Bmc Genomics* **16**, https://doi.org/10.1186/s12864-015-1273-2 (2015).
40. Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, https://doi.org/10.7554/eLife.05005 (2015).
41. Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* **11**, https://doi.org/10.1186/gb-2010-11-8-r90 (2010).
42. Hochberg, Y. & Benjamini, Y. More Powerful Procedures for Multiple Significance Testing. *Stat Med* **9**, 811–818, https://doi.org/10.1002/sim.4780090710 (1990).
43. Hayashi, T. *et al*. Not all NOTCH Is Created Equal: The Oncogenic Role of NOTCH2 in Bladder Cancer and Its Implications for Targeted Therapy. *Clinical cancer research: an official journal of the American Association for Cancer Research* **22**, 2981–2992, https://doi.org/10.1158/1078-0432.CCR-15-2360 (2016).
44. Turner-Ivey, B. *et al*. KAT6A, a Chromatin Modifier from the 8p11-p12 Amplicon is a Candidate Oncogene in Luminal Breast Cancer. *Neoplasia* **16**, 644–655, https://doi.org/10.1016/j.neo.2014.07.007 (2014).
45. Santos, A. O., Parrini, M. C. & Camonis, J. RalGPS2 Is Essential for Survival and Cell Cycle Progression of Lung Cancer Cells Independently of Its Established Substrates Ral GTPases. *Plos One* **11**, https://doi.org/10.1371/journal.pone.0154840 (2016).
46. Imaoka, S. *et al*. CYP4B1 is a possible risk factor for bladder cancer in humans. *Biochem Bioph Res Co* **277**, 776–780, https://doi.org/10.1006/bbrc.2000.3740 (2000).
47. Chintapalli, V. R. *et al*. Transport proteins NHA1 and NHA2 are essential for survival, but have distinct transport modalities. *P Natl Acad Sci USA* **112**, 11720–11725, https://doi.org/10.1073/pnas.1508031112 (2015).
48. Schinke, E. N. *et al*. A novel approach to identify driver genes involved in androgen-independent prostate cancer. *Mol Cancer* **13**, https://doi.org/10.1186/1476-4598-13-120 (2014).
49. Jiang, D. *et al*. Analysis of p53 Transactivation Domain Mutants Reveals Acad11 as a Metabolic Target Important for p53 Pro-Survival Function. *Cell Rep* **10**, 1096–1109, https://doi.org/10.1016/j.celrep.2015.01.043 (2015).
50. Poplawski, P. *et al*. Restoration of type 1 iodothyronine deiodinase expression in renal cancer cells downregulates oncoproteins and affects key metabolic pathways as well as anti-oxidative system. *Plos One* **12**, https://doi.org/10.1371/journal.pone.0190179 (2017).

51. Gammons, M. V. *et al*. Targeting SRPK1 to control VEGF-mediated tumour angiogenesis in metastatic melanoma. *Brit J Cancer* **111**, 477–485, https://doi.org/10.1038/bjc.2014.342 (2014).
52. Mavrou, A. *et al*. Serine-arginine protein kinase 1 (SRPK1) inhibition as a potential novel targeted therapeutic strategy in prostate cancer. *Oncogene* **34**, 4311–4319, https://doi.org/10.1038/onc.2014.360 (2015).
53. Tang, R. X. *et al*. Identification of a RNA-Seq based prognostic signature with five lncRNAs for lung squamous cell carcinoma. *Oncotarget* **8**, 50761–50773, https://doi.org/10.18632/oncotarget.17098 (2017).
54. Chen, X. *et al*. LncRNA ZNF503-AS1 promotes RPE differentiation by downregulating ZNF503 expression. *Cell Death Dis* **8**, https://doi.org/10.1038/cddis.2017.382 (2017).
55. Perron, U., Provero, P. & Molineris, I. In silico prediction of lncRNA function using tissue specific and evolutionary conserved expression. *Bmc Bioinformatics* **18**, https://doi.org/10.1186/s12859-017-1535-x (2017).
56. Chen, W. H., Lu, G. T., Chen, X., Zhao, X. M. & Bork, P. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic acids research* **45**, D940–D944, https://doi.org/10.1093/nar/gkw1013 (2017).
57. Woods, C. T. & Laederach, A. Classification of RNA structure change by 'gazing' at experimental data. *Bioinformatics* **33**, 1647–1655, https://doi.org/10.1093/bioinformatics/btx041 (2017).
58. Baytak, E. *et al*. Whole transcriptome analysis reveals dysregulated oncogenic lncRNAs in natural killer/T-cell lymphoma and establishes MIR155HG as a target of PRDM1. *Tumor Biol* **39**, https://doi.org/10.1177/1010428317701648 (2017).
59. Li, Y. S. *et al*. LncRNA ontology: inferring lncRNA functions based on chromatin states and expression patterns. *Oncotarget* **6**, 39793–39805, https://doi.org/10.18632/oncotarget.5794 (2015).
60. Sudmant, P. H. *et al*. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75, https://doi.org/10.1038/nature15394 (2015).
61. Rosenbloom, K. R. *et al*. The UCSC Genome Browser database: 2015 update. *Nucleic acids research* **43**, D670–D681, https://doi.org/10.1093/nar/gku1177 (2015).
62. Uren, P. J. *et al*. Site identification in high-throughput RNA-protein interaction data. *Bioinformatics* **28**, 3013–3020, https://doi.org/10.1093/bioinformatics/bts569 (2012).
63. Forbes, S. A. *et al*. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research* **43**, D805–811, https://doi.org/10.1093/nar/gku1075 (2015).
64. Ning, S. W. *et al*. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic acids research* **44**, D980–D985, https://doi.org/10.1093/nar/gkv1094 (2016).
65. Harrow, J. *et al*. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760–1774, https://doi.org/10.1101/gr.135350.111 (2012).
66. Yates, B. *et al*. Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic acids research* **45**, D619–D625, https://doi.org/10.1093/nar/gkw1033 (2017).
67. Rodriguez, J. M. *et al*. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic acids research* **41**, D110–D117, https://doi.org/10.1093/nar/gks1058 (2013).
68. Pruitt, K. D. *et al*. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19**, 1316–1323, https://doi.org/10.1101/gr.080531.108 (2009).
69. Hofacker, I. L. RNA secondary structure analysis using the Vienna RNA package. *Current protocols in bioinformatics* Chapter 12, Unit12 12 https://doi.org/10.1002/0471250953.bi1202s26 (2009).
70. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842, https://doi.org/10.1093/bioinformatics/btq033 (2010).
71. Robin, X. *et al*. pROC: an open-source package for R and S plus to analyze and compare ROC curves. *Bmc Bioinformatics* **12**, https://doi.org/10.1186/1471-2105-12-77 (2011).

Acknowledgements

All primary data were accessed from The Cancer Genome Atlas Research Network (cancergenome.nih.gov) and International Cancer Genome Consortium (dcc.icgc.org). We would like to thank Wei Huang from Gu lab for his assistance in setting up high performance computing infrastructure. This work was supported by grants from the National Key Research and Development Program of China (2017YFC0908600), the National Natural Science Foundation of China (31771408).

Author Contributions

F.H. and Z.S. designed the study and drafted the manuscript. F.H., R.W., Z.Z. and L.H. performed the analysis and interpreted the results. Y.W., J.T. and Y.Z. collected the data. L.S., X.G. and M.J.D. assisted in writing the manuscript. All authors read, edited, and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-44489-5.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.