Published online 11 September 2021 *Nucleic Acids Research, 2021, Vol. 49, No. 18* **10369–10381**
https://doi.org/10.1093/nar/gkab772

Comprehensive characterization of somatic variants associated with intronic polyadenylation in human cancers

Zhaozhao Zhao[1,†], Qiushi Xu[1,†], Ran Wei[1,2], Leihuan Huang[1], Weixu Wang[1], Gang Wei[1,3,*] and Ting Ni[1,4,*]

[1]State Key Laboratory of Genetic Engineering, Collaborative Innovation Center of Genetics and Development, Human Phenome Institute, School of Life Sciences and Huashan Hospital, Fudan University, Shanghai 200438, P.R. China, [2]Department of Pathology, Fudan University Shanghai Cancer Center, Department of Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, P.R. China, [3]MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200438, P.R. China and [4]Shanghai Engineering Research Center of Industrial Microorganisms, School of Life Sciences, Fudan University, Shanghai 200438, P.R. China

Received December 10, 2020; Revised August 16, 2021; Editorial Decision August 21, 2021; Accepted August 26, 2021

ABSTRACT

Somatic single nucleotide variants (SNVs) in cancer genome affect gene expression through various mechanisms depending on their genomic location. While somatic SNVs near canonical splice sites have been reported to cause abnormal splicing of cancer-related genes, whether these SNVs can affect gene expression through other mechanisms remains an open question. Here, we analyzed RNA sequencing and exome data from 4,998 cancer patients covering ten cancer types and identified 152 somatic SNVs near splice sites that were associated with abnormal intronic polyadenylation (IPA). IPA-associated somatic variants favored the localization near the donor splice sites compared to the acceptor splice sites. A proportion of SNV-associated IPA events overlapped with premature cleavage and polyadenylation events triggered by U1 small nuclear ribonucleoproteins (snRNP) inhibition. GC content, intron length and polyadenylation signal were three genomic features that differentiated between SNV-associated IPA and intron retention. Notably, IPA-associated SNVs were enriched in tumor suppressor genes (TSGs), including the well-known TSGs such as *PTEN* and *CDH1* with recurrent SNV-associated IPA events. Minigene assay confirmed that SNVs from *PTEN*, *CDH1*, *VEGFA*, *GRHL2*, *CUL3* and *WWC2* could lead to IPA. This work reveals that IPA acts as a novel mechanism explaining the functional consequence of somatic SNVs in human cancer.

INTRODUCTION

Single nucleotide variants (SNVs) in cancer genome can affect cancer-related genes and associated phenotypes through diverse mechanisms (1,2). SNVs in coding region could introduce amino acid changes that contribute to altered protein function while SNVs locate in regulatory regions (enhancer, promoter, 5′ or 3′ untranslated regions (UTR)) could affect the expression of cancer-associated genes (3–9). Somatic variants in canonical splice sites have also been reported to cause dysregulation of cancer-related genes by inducing different forms of abnormal splicing (10). Recent pan-cancer studies showed that SNVs near exon-intron boundaries (±30 bp) could cause aberrant intron retention and those affected genes were enriched in tumor suppression function (e.g. *TP53*) (11,12). In addition to these above-mentioned mechanisms that explain the functional consequence of somatic variants, whether other unexpected mechanisms exist is an open question.

Growing evidence has indicated that splicing and polyadenylation of messenger RNA (mRNA) are coupled events taking place co-transcriptionally (13,14). It thus raises the possibility that SNVs may also directly cause abnormal polyadenylation of affected genes. It is known that certain SNVs locating inside the polyadenylation signal can result in impaired polyadenylation and thus reduced gene expression and ultimately related diseases (15). For example, an A to G mutation in polyadenylation signal of the α2-globin coding gene *HBA2* leads to impaired *HBA2* expression and thus α-thalassaemia (16).

*To whom correspondence should be addressed. Tel: +86 21 31246627; Fax: +86 31246627; Email: tingni@fudan.edu.cn.
Correspondence may also be addressed to Gang Wei. Tel: +86 21 31246626; Fax: +86 31246626; Email: gwei@fudan.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Recent study revealed globally increased usage of intronic polyadenylation (IPA) could increase the risk of chronic lymphocytic leukemia through translating prematurely terminated mRNAs into truncated proteins, which in turn impair the tumor-suppressive function of related genes (17). A genome-wide analysis revealed that a single nucleotide polymorphism (rs11032578) reduced the mRNA expression of *ABTB2* gene by increasing the usage of its intronic poly(A) site (18). These above lines of evidence imply that intronic polyadenylation might be an unexpected mechanism in mediating somatic variant-caused cancer phenotypes.

To explore whether somatic variants could influence intronic polyadenylation in cancers, we carried out a comprehensive analysis that integrates both whole-exome sequencing and RNA sequencing (RNA-seq) datasets across 4998 samples spanning ten cancer types in The Cancer Genome Atlas (TCGA) project to detect variants that trigger abnormal intronic polyadenylation. We discovered 152 SNV-associated IPA events with strong evidence supporting their authenticity. Notably, these IPA-associated somatic variants were enriched in tumor suppressor genes. We revealed that IPA acted as a novel mechanism explaining the functional consequence of somatic SNVs in human cancer.

MATERIALS AND METHODS

Data download

The Cancer Genome Atlas (TCGA) characterizes a comprehensive list of genomic, epigenomic, and transcriptomic features in thousands of tumor samples. Here, we processed 4998 samples which have both matched whole-exome sequencing and RNA-seq data from ten cancer types including LUAD, LUSC, HNSC, LIHC, KIRC, PRAD, THCA, COAD, BRCA and SKCM (Supplementary Table S1). All the RNA-seq BAM files were downloaded from the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/). We used somatic variants called by MC3 group in TCGA (https://gdc.cancer.gov/about-data/publications/mc3-2017; v0.2.8), which were formed by the consensus of multiple variant calling algorithms in a unified pipeline (19). Only variants passed all QC metrics by the MC3 group (i.e., 'PASS' in the 'filter' column) were used for consequent analysis. Other sequencing data are available from the NCBI GEO database: accessions GSE111310 and GSE111793 (3′-seq and RNA-seq from normal immune cells and malignant B cells from patients with chronic lymphocytic leukemia) (17,20), GSE135140 (U1 inhibition in HeLa cells) (21), GSE123105 (*PCF11* knockdown in HeLa cells) (22), GSE95057 (TRENDseq data in BE(2)-C neuroblastoma cells) (23,24).

Identification of SNVs associated with abnormal splicing

We first detected SNVs associated with abnormal splicing by allele-specific splicing analysis (12). We compiled 187 067 SNVs within 30 bp of an exon-intron junction based on RefSeq annotation including 169 445 exonic SNVs and 17 622 intronic SNVs from 10 cancer types in TCGA (Supplementary Table S1). For each of these 169 445 somatic exonic SNVs, we extracted the mapped reads with a base quality $\geq$20 at the SNV locus and examined which allele (mutant or reference) they belong to. We determined whether the read supports normal splicing or abnormal splicing by the following rules. The exon containing the somatic variant, the adjacent intron, and the adjacent exon are defined as E^*, I, E, respectively. A read was considered to support (i) normal splicing if it spanned the E^*–E junction and covered $\geq$10 bp of both exons and (ii) abnormal splicing if it spanned the E^*–I junction and covered $\geq$10 bp of both E^* and I. Then, we counted the number of reads for each of the four categories depending on the splicing status (normal versus abnormal) and the SNV allele (reference versus mutant) and tested the association between the SNV and splicing using Fisher's exact test. For accurate statistical testing, SNVs had to have $\geq$3 reads in more than two categories in the Fisher's exact test. P values from Fisher's exact test were grouped by SNV position relative to the nearest splice site and corrected by the Benjamini-Hochberg multiple-testing correction method for each group to adjust for a different background distribution at each position. If the adjusted P value is $<$0.05, the abnormal splicing event is thought to be associated with the SNV. If a somatic SNV located at the intron region, the mutant allele can be observed in mRNAs retaining the intron, but the wild-type allele cannot be observed because the intron is normally spliced out. Thus, reads with intronic SNV spanning the exon-intron junction were considered to support abnormal splicing-causing SNVs. Then we examined whether these identified abnormal splicing-associated SNVs were associated with intronic polyadenylation or intron retention.

Identification of SNV-associated intronic polyadenylation

For identified SNVs associated with abnormal splicing, we apply our recently developed IPAFinder to examine whether there exists intronic poly(A) site used abnormally in the corresponding intron region (25). IPAFinder performs *de novo* identification and quantification of IPA events, without the need for any prior poly(A) site annotation. IPAFinder modeled the normalized RNA-seq read coverage at single-nucleotide resolution to identify the profound drop in coverage, which can be interpreted as evidence of intronic polyadenylation processing. IPAFinder progressively segmented the intron region into two regions with distinct mean coverage to infer the potential intronic poly(A) site, where the squared deviation decreases most from the mean coverage of the intron when dividing the segment into two regions compared to considering it as a single segment. Although IPAFinder could distinguish composite IPA and skipped IPA events, SNVs near the splice sites are likely to impair the normal splicing and thus unlikely generate skipped IPA events, we therefore only considered composite terminal exon IPA event. Of note, IPAFinder excluded alternative splicing event such as alternative 5′ splice site by recognizing junction-spanning reads. Then, we calculated the Intronic Poly(A) site Usage Index (IPUI) of SNV-associated intronic polyadenylation event, and evaluated whether the IPUI was significantly higher than expected. The background distribution of IPUIs was estimated based on RNA-seq data sets from normal tissues as well as cancer samples without related SNVs in the corresponding cancer type. For a somatic SNV to be

associated with abnormal intronic polyadenylation, we required its IPUI to be within the top 1% of the background distribution.

Identification of SNV-associated intron retention

For identified SNVs associated with abnormal splicing, we examined whether corresponding intron was retained. IR events were determined using our established intron retention index (IRI) algorithm, which defines IRI as the ratio of read density of the intronic region (RDintron) and the mean read density of flanking upstream exon (RDupexon) and flanking downstream exon (RDdownexon) (26,27).

$$\mathrm{IRI} = \frac{\mathrm{RDintron}}{(\mathrm{RDupexon} + \mathrm{RDdownexon})/2}$$

To reduce false positive, we applied the following criteria to define an intron retention event: (i) the mapped reads should cover at least 80% of the intron length, and (ii) retention ratio should be at least 0.1 (i.e. IRI $\geq$ 0.1). If the corresponding intron was retained, we think the intron retention event is associated with the SNV.

3′-seq and RNA-seq data analyses

Among the raw reads obtained from both 3′-seq and RNA-seq experiments, low-quality reads were filtered out, followed by alignment to human reference genome sequence (hg38) using STAR with default settings (28). For 3′-seq, all aligned reads from the same cell type were pooled together in order to identify peaks (corresponding to cleavage events) across the genome. We used F-seq to call peak of poly(A) sites using default parameters except that we set the parameter of feature length to 30 nt (29). We resized the polyadenylation (PA) peaks to the shortest distance that contained 95% of the reads according to our previous publication (30), and the resized PA peaks were defined as PA clusters. To filter internal priming, we removed PA clusters with continuous 6 'A' downstream of the apex of corresponding PA cluster or with 15 'A' in the 20 nucleotides region downstream of the apex of corresponding PA cluster. The number of reads aligned to each PA cluster reflects expression of an individual 3′ end isoform. To further focus the analysis on robustly expressed transcript isoforms, we kept PA clusters that accounted for at least 10% of all the tags within respective genes in at least one sample and also low-expression PA clusters containing 5% of all the tags within respective genes in the majority of samples (80%) (31). For statistical analysis, the expression level of each isoform was examined by Fisher's exact test in comparison to the other 3′ end isoform(s) expressed by the same gene. Contingency table included the number of reads of the tested isoform and total amount of reads of all the other isoforms of the same gene for the knockdown and control samples, respectively. Obtained P values were adjusted using the Benjamini-Hochberg method. The usage of specific IPA site was quantified by the number of reads mapped to the IPA site divided by the total number of reads at all PA clusters for the same gene. We defined an IPA site to be significantly differentially used if its adjusted P value < 0.05 and usage difference > 0.05. For RNA-seq datasets of HeLa cells generated from treatment of U1 Antisense Morpholino Oligonucleotide (AMO) (21), we applied IPAFinder for the analysis. All upregulated IPA events upon U1 inhibition were obtained and then were tested the overlap with SNV-associated IPA events identified in 10 cancer types as mentioned above. 330 recurrent upregulated CLL-IPA events were obtained from the previous study (17) and were tested for the overlap with the same SNV-associated IPA events.

Splice strength estimation and motif analysis

The MAXENT tool was used to estimate the donor and acceptor splicing strength of IPA-associated SNVs with the wild-type or mutant allele (32). Aside from SNVs at canonical dinucleotides (GT and AG), SNVs at positions from −3 to +6 and from −3 to +20 near the donor and acceptor SSs were used to calculate the donor and acceptor splicing strengths, respectively. Motif logos were generated using WebLogo (33). We used MEME (version 5.3.2) to *de novo* discover 6-mer motif enrichment within the 40 nt upstream of the identified SNV-associated IPA sites (34).

Tumor-suppressor genes and oncogenes

The tumor-suppressor genes and oncogenes used in this study were defined by the TUSON algorithm from genome sequencing data obtained from more than 8200 cancers (35). TUSON is a computational method that analyzes patterns of mutation in tumors and predicts the likelihood that any individual gene functions as a tumor-suppressor gene or oncogene. We ranked genes ascendingly by their TUSON prediction P values, and used $P < 0.01$ as cutoff to obtain the reference tumor-suppressor genes or oncogenes. After removing 27 genes in common, 458 tumor-suppressor genes and 468 oncogenes were used for the enrichment analysis. Finally, hypergeometric test was used to assess whether there is a significant enrichment of tumor-suppressor genes or oncogenes in genes with SNV-associated intronic polyadenylation.

Positions of truncating mutations

The positions of truncating (TR) mutation in solid cancers of TSGs affected by SNV-associated intronic polyadenylation were obtained from the MSK cbio portal (data of reference, 21 November 2020) (36,37). The position with the highest number of TR mutations for a gene was used.

Number of amino acids of full-length proteins or IPA-generated truncated proteins

To calculate the number of amino acids of full-length proteins, we used the longest RefSeq annotated mRNA isoform and obtained the total number of amino acids. To calculate the number of amino acids of the IPA-generated truncated proteins, we counted the number of amino acids from the start codon to the end of the exon located upstream of the IPA site. The amino acids translated from intronic sequence

are added to obtain the size of the IPA-generated truncated proteins. The fraction of retained coding region (CDR) is the number of amino acids retained divided by the number of amino acids calculated from the longest mRNA isoform encoding the full-length protein.

Protein domain analysis

The information about protein domains was obtained from the UCSC UniProt table. If a gene had multiple protein isoforms, then the longest isoform was used in the analysis. The protein lengths were obtained from http://www.uniprot.org/ for *Homo sapiens*.

Survival analysis

BRCA patients were separated into two groups based on the presence of SNV-associated IPA event in gene *CDH1* and statistical analysis between Kaplan–Meier curves for the two groups was performed using the log rank test.

Cell culture

HEK293T (human embryonic kidney 293T cells) and HeLa (human cervical cancer cells) cells were purchased from ATCC. HEK293T and HeLa were cultured in DMEM with 10% FBS. All cells were cultured in a humidified atmosphere of 5% CO_2 and 95% air at 37 °C.

Experimental validation using minigene assays

For each region of interest, the candidate intron and its flanking exons were amplified by PCR using the primers listed in Supplementary Table S2. In order to introduce the mutation, we used site-directed mutagenesis based on PCR. In short, genomic DNA from HEK293T cells were amplified by PCR using primers to generate two 20–25 bp overlapping fragments containing a mutant site. These fragments were subcloned into the EcoRI and BamHI sites of the pcDNA3.1 vector by the One Step Cloning Kit (Vazyme). Each Sanger sequenced construct (2 μg) was transiently transfected into HEK293T and HeLa cells using Lipofectamine 2000 (Life Technologies) in six-well tissue culture plates. Cells were harvested for RNA extraction 32 hours (hrs) after transfection. Total RNA (1 μg) was extracted with TRIzol reagent (Invitrogen) according to the manufacturer's instruction. cDNA was synthesized using FastKing RT Kit (With gDNase) (Tiangen) with oligo dT18-XbaKpnBam primer for 3′ RACE (38). 20 μl cDNA product was diluted 5-fold, and 2 μl diluted cDNA was used as the template for each reaction. 3′ RACE was carried out using the pcDNA3.1- forward primer and XbaKpnBam reverse primer to distinguish minigene RNA from endogenous RNA. The 3′ RACE PCR products were separated by gel electrophoresis through a 2% agarose gel in 1× TAE buffer. To confirm the sequence of each band, the 3′ RACE PCR products were gel purified using the Zymoclean Gel DNA Extraction kit (Zymo) and verified by Sanger sequencing. Primer sequences are listed in Supplementary Table S2.

RESULTS

Identification of IPA events associated with somatic variants

To identify somatic SNVs coupling with abnormal intronic polyadenylation, we first compiled a total of 1 391 126 somatic SNVs from 4998 cancer samples (ten cancer types) that underwent whole-exome sequencing in TCGA. Among these SNVs, 13.4% variants were within 30 bp of exon-intron junctions and 0.66% were at intronic 5′ and 3′ splice sites (Supplementary Table S1).

Assuming there is an intronic poly(A) site used in an intron, an apparent RNA-seq read coverage drop would be observed in that intron along with reads spanning the upstream exon-intron boundary. Thus, we first tested the association between SNVs and abnormal splicing in an allele-specific manner, which means that a mutant allele is likely to span the exon-intron junction while the wild-type (reference) allele tends to be normally spliced. For somatic variants close to exon-intron junctions, we were able to identify RNA-seq reads that reflected the altered splicing and indicated the mutant allele simultaneously. We applied Fisher's exact test to compare the proportion of such reads to the proportion of reads with the wild-type allele spanning exon-intron junction. The allele-specific splicing analysis provides strong evidence to support the causality of variant to inefficient splicing, which is necessary for the generation of composite IPA events (Figure 1B and E; Supplementary Figure S1A). For identified SNVs disrupting splicing, we then examined whether there existed intronic poly(A) site in the corresponding intron. We modeled the normalized RNA-seq read coverage profiles at single-nucleotide resolution and identified the drop in coverage to infer the potential intronic poly(A) sites (Figure 1A and D; Supplementary Figure S1A; see Materials and Methods for details). To be quantitative, we calculated the Intronic Poly(A) site Usage Index (IPUI) of SNV-associated abnormal intronic polyadenylation event, and evaluated whether the IPUI was significantly higher than expected. The background distribution of IPUIs was estimated on the basis of RNA-seq data sets from normal tissues as well as cancer samples without related SNVs in the corresponding cancer type (Figure 1C and F; Supplementary Figure S1A). This rigorous framework identified 152 somatic SNV-associated abnormal intronic polyadenylation events, derived from 129 genes in ten tested cancer types (Supplementary Table S3).

Reliability evaluation on identified SNV-associated IPA events

For the identified intronic poly(A) sites (IPA sites) with increased usage in the presence of somatic SNVs, 67.4% are within 100 nucleotides (nt) of the annotated poly(A) sites compiled from RefSeq, Ensembl, UCSC gene models and poly(A) site databases (39,40). In the upstream (–40 nt) of these IPA sites, canonical poly(A) signal AATAAA can be successfully identified by MEME motif enrichment analysis (34) (Supplementary Figure S1B). Drop in read coverage around the poly(A) site and reduced usage of downstream exons also supported the authenticity of these IPA

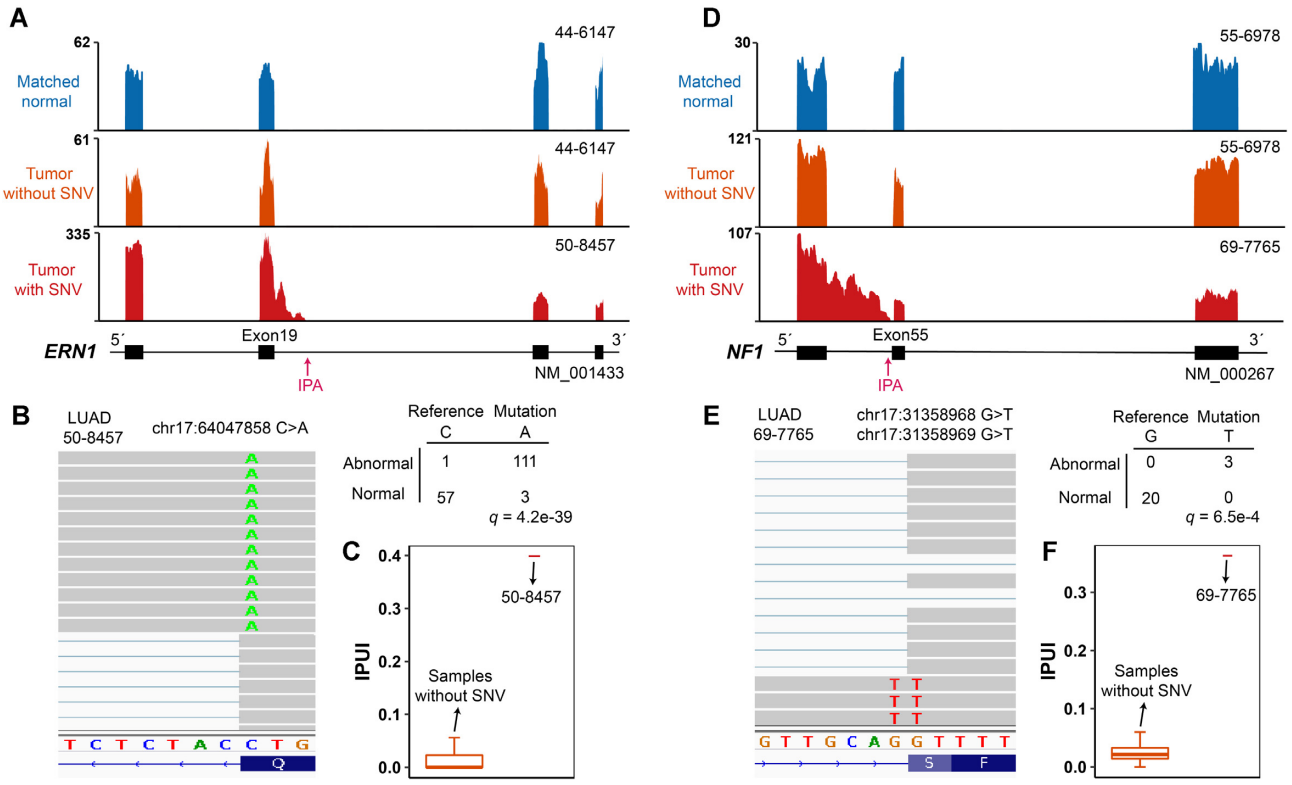

Figure 1. Identification of SNVs associated with intronic polyadenylation. (A, D) RNA-seq density plots showing that both somatic variant around 5′ splice site (**A**) and 3′ splice site (**D**) could cause abnormal intronic polyadenylation. Sample IDs are shown at the top-right corner of corresponding RNA-seq density plot. Intronic poly(A) site (IPA) is indicated by a red arrow. (B, E) IGV browser screenshots showing RNA-seq reads spanning the exon-intron junction associated with SNVs in *ERN1* (**B**) and *NF1* (**E**). Association between SNVs and splicing abnormality was tested using Fisher's exact test. (C, F) SNVs around splice sites significantly increase the usage of intronic poly(A) site in *ERN1* (**C**) and *NF1* (**F**). IPUI means intronic poly(A) site usage index.

events, as has been used for confirming IPA events by other studies (17,20). Interestingly, 13.8% of the identified intronic poly(A) sites located in introns without any annotated poly(A) sites. The authenticity of these novel IPA sites was supported by multiple lines of evidence, including mutant allele specific abnormal splicing, loss of downstream exons, and the presence of canonical poly(A) signal in the upstream of their *de novo* intronic poly(A) sites, as exemplified by *GRHL2* and *CUL3* (Supplementary Figure S1C-J). Furthermore, we experimentally validated that SNVs caused increased usage of unannotated IPA sites in four genes (*GRHL2*, *CUL3*, *WWC2* and *CDH1*) using minigene reporter assay (Supplementary Figure S2). For example, the minigene construct harboring the SNV in *GRHL2* showed a strong degree of premature intronic polyadenylation, whereas the wild-type construct showed normal splicing (Supplementary Figure S2B). Sanger sequencing further confirmed the exact sequence of mutant allele spanning the exon-intron boundary and the location of intronic poly(A) site (Supplementary Figure S2B). In addition, we also confirmed that SNV increased usage of annotated IPA site in *VEGFA* (Supplementary Figure S2G and H). These validations demonstrate that our computational analysis using paired DNA sequencing and RNA-seq data is effective in finding SNVs that causing intronic polyadenylation.

Overview of somatic SNVs associated with intronic polyadenylation

To investigate the positional effect of somatic variants associated with abnormal IPA, we evaluated the distribution of these SNVs around the exon-intron boundary, and found that IPA-associated somatic variants favored the localization near the donor splice sites compared to the acceptor splice sites (Figure 2A). Moreover, the closer the mutated base to the 5′ exon–intron boundary, the higher frequency it had (Figure 2A). To examine whether the IPA-associated SNV distribution bias between acceptor and donor splice sites is caused by the overall SNVs detection difference between these two sites, we calculated the total numbers of SNVs nearby and found these two regions showed comparable numbers and distribution of SNVs (Supplementary Figure S3A and B). This result suggests that the donor splice site preference is a unique feature for IPA-associated SNVs.

These IPA-associated variants around splice sites were found to weaken the splicing strength of corresponding donor and acceptor splice sites (SSs) (Figure 2B; Supplementary Figure S3C). The wild-type sequences around donor splice sites showed a sequence consensus, which, however, was broken by IPA-associated SNVs (Figure 2C). Interestingly, these regions correspond to the binding site of U1 small nuclear ribonucleoprotein (snRNP) (41). Previous

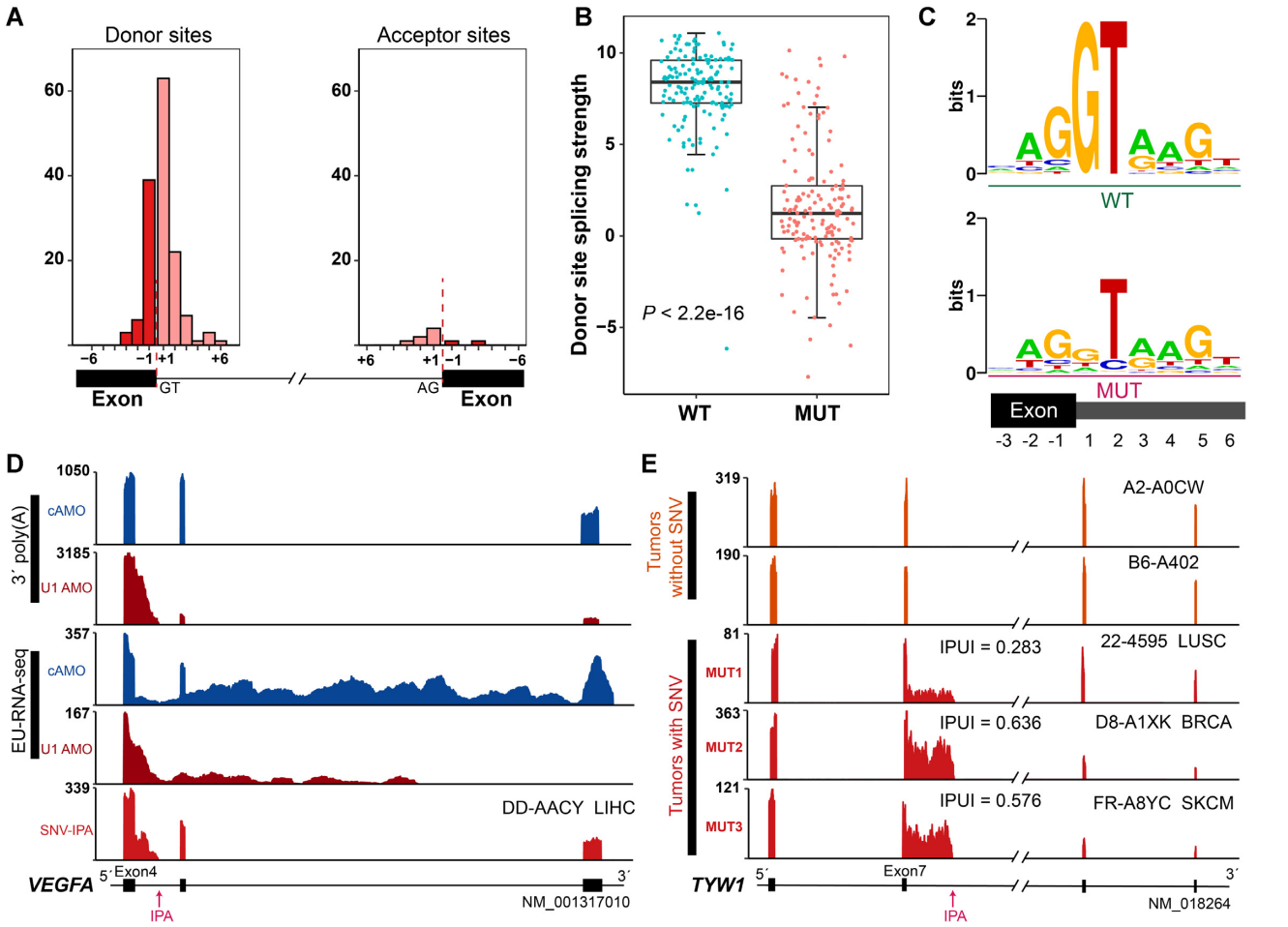

Figure 2. Overview of IPA-associated somatic SNVs. (**A**) Evaluation of position-wise numbers of abnormal intronic polyadenylation associated SNVs between the sixth exonic base (-6) and sixth intronic base (+6) for splicing donor and acceptor sites. (**B**) Differences in splicing strength of donor splice sites between the mutant (MUT) and wild-type (WT) alleles, as estimated by MAXENT (32). The two-side Wilcoxon rank sum test was used to assess the statistical significance. (**C**) Changes in sequence consensus caused by mutations near the donor splice sites. The consensus motifs were derived from the 9-mers of the authentic donor splice sites. (**D**) RNA-seq density plots showing that both U1 inhibition and SNV around the donor sites increase the intronic poly(A) site usage in *VEGFA*. Nascent RNAs were detected by RNA sequencing of 5 min pulse-labeled ethynyl-uridine RNA (EU-RNA-seq). Intronic poly(A) site is indicated by a red arrow. (**E**) RNA-seq density plots showing that somatic SNVs around the same exon-intron junction could increase the intronic poly(A) site usage at different degree. Sample IDs are shown at the top-right corner of corresponding plots.

studies revealed that U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation (PCPA) at cryptic polyadenylation signals in introns (38,42). By integrating RNA-seq data derived from HeLa cells treated with U1 Antisense Morpholino Oligonucleotide (AMO), which has been shown to pair efficiently with U1 snRNA and thereby functionally inhibit U1 snRNP (21), we found that a proportion of IPA events caused by SNV around 5′ splice site (12%, 17/142) overlapped with PCPA events triggered by U1 inhibition, as exemplified by genes *VEGFA*, *DOT1L* and *PTEN* (Figure 2D; Supplementary Figure S4). These results suggest that somatic variants near donor splice site could cause IPA likely through disrupting the base pairing between U1 snRNP and pre-mRNA.

Next, we analyzed how these IPA-associated SNVs distributed among genes and cancer types. The majority (116 out of 129) of affected genes have only one SNV-associated IPA event (Supplementary Figure S5A), and the remaining 13 genes contained multiple IPA-related SNVs derived from either different position of corresponding gene or different cancer types. For example, *DNAH5* had three location-specific IPA events coupled with three different SNVs in three cancer types (Supplementary Figure S5B). *CDH1* had seven SNV-coupled IPA events specific for breast cancer (Supplementary Figure S5A). Interestingly, we found *TYW1* had three different SNVs (one is at exonic region and others hit the GT dinucleotides) near the same exon-intron boundary but coupled with different degrees of intronic polyadenylation (Figure 2E; Supplementary Figure S6A). SNVs at highly conserved GT dinucleotides, which coupled with higher intronic poly(A) site usage, weaken the strength of donor splice site more acutely than that at exonic region (Supplementary Figure S6B). By measuring IPA site usage of *TYW1* upon perturbation of different proteins, we found that depletion of certain protein couldn't significantly increase the usage of IPA site in *TYW1* (Supplementary Figure S6C). We also analyzed the distribution of these SNV-associated IPA events among cancer types and found

different cancers had dramatic number difference. Lung cancers (LUSC and LUAD) and breast cancer (BRCA) ranked the top three regarding number of SNV-associated IPA events (Supplementary Figure S5C).

Genomic features distinguish SNV-associated intronic polyadenylation from intron retention

Previous studies showed that SNVs in exon-intron boundaries could cause intron retention (12). Our analysis above suggested that SNVs around splice sites could also cause intronic polyadenylation. Thus, we wondered what features affect the fate (retained or polyadenylated) of introns when SNVs near splice sites occur.

To address this question, we detected SNV-associated intron retention events in an allele-specific manner, as indicated by previous method (12), and obtained 709 reliable SNV-associated intron retention events in the 4,998 cancer samples (Supplementary Table S4). Consistent with previous findings (12,43), our result showed that genes harboring SNV-coupling intron retention events were enriched in tumor suppressor genes (TSGs), and *TP53* was the most frequently disrupted gene (Figure 3A; Supplementary Figure S7A and B). We also evaluated the distribution of these SNVs around the exon-intron boundary and found that the greatest effect was observed around the donor and acceptor splice sites (Figure 3B). Analysis of splicing strength revealed that intron retention-associated SNVs weaken the strength of corresponding donor and acceptor SSs (Supplementary Figure S7C). In addition, we found the well-known tumor suppressor gene *STK11*, which has been proved to form a new exon within an intron caused by splicing-associated variants (44), also had SNV-associated intron retention frequently (Figure 3C; Supplementary Figure S8A). We then examined whether certain features were different between SNV-associated retained introns and polyadenylated introns. Intron size and GC content were first compared between these two intron groups. We found that introns with increased usage of poly(A) sites in the presence of splice site variants tended to have lower GC content and longer length than retained introns (Figure 3D and E), consistent with previous findings that retained introns are significantly associated with elevated GC content and reduced length (45,46). We then compared the fraction of introns with annotated poly(A) sites between these two intron groups, and found that SNV-associated polyadenylated introns were more likely to have annotated poly(A) sites than retained introns ($P < 2.2e{-}16$; χ^2 test) (Figure 3F). These results reveal that genomic architecture and sequence composition may affect the fate of introns with SNVs disrupting splicing.

Intronic polyadenylation-associated SNVs are enriched in TSGs

To measure the relevance of SNV-associated abnormal intronic polyadenylation during cancer development, we tested whether they were enriched in oncogenes or tumor suppressor genes. We found that genes harboring SNV-associated intronic polyadenylation were enriched in TSGs but not in oncogenes (Figure 4A). Importantly, IPA-generated truncated proteins usually had a comparable number of amino acids comparing to truncated proteins generated by truncating (TR) mutations (Figure 4B), suggesting that the intronic polyadenylation caused by SNVs may phenocopy TR mutations and these corresponding IPA isoforms are probably functionally inactive. *NF1* and *PTEN* are two TSGs containing SNV-associated IPA events (Figures 1D and 4C). *NF1* encodes neurofibromin, a negative regulator of RAS signal transduction pathway by promoting the conversion from active RAS-GTP to its inactive RAS-GDP state (47). Any loss of neurofibromin functionality will lead to prolonged activation of the RAS/RAF/MAPK signaling pathway and ultimately increased cellular proliferation and loss of growth control (48–50). *PTEN* is one of the most frequently mutated genes in human cancers and cancer-associated *PTEN* mutations are found scattered over the entire *PTEN* gene (51,52). The PTEN protein consists of an N-terminal phosphatase domain, which is responsible for antagonizing the PI3 kinase/AKT pathway to function as a tumor suppressor (53), a C2 domain mediating Ca^{2+}-dependent lipid interaction, and a 50-amino-acid C-terminal tail. We found that SNVs around the end of exon 8 of *PTEN* increased the intronic poly(A) site usage and led to generation of a truncated isoform that lost C-terminal tail (Figure 4C; Supplementary Figure S8B). Minigene assay validated that two SNVs in PTEN both caused increased usage of IPA sites (Figure 4D). Recent studies demonstrated that the C-terminal sequence of PTEN is critical for its nuclear localization and the regulation of anchorage-independent growth and cell migration (54–56). This IPA-generated truncated protein has similar sequence to mutant PTEN-342 (Figure 4E), a C-terminal deletion mutant ending at codon 342, which has a significantly lower phosphatase activity and higher degradation rate compared to the wild-type PTEN (56). Therefore, somatic variants causing abnormal intronic polyadenylation may represent an important but previously unrecognized role in impairing the function of tumor suppressor genes.

SNV-associated intronic polyadenylation frequently affect *CDH1*

Among genes with SNV-associated IPA, *CDH1* was the most frequently altered gene, affecting seven samples in breast cancer (Figure 5A). E-cadherin, the protein product of the *CDH1* gene, is a calcium-dependent cell-to-cell adhesion molecule, whose deletion or deregulation is correlated with tumor invasion and metastasis in some tumor types (57,58). Furthermore, loss of E-cadherin is the key hallmark of invasive lobular carcinoma, the second most prevalent histologic subtype of invasive breast cancer, wherein mutations uniformly distributed along the coding sequence of *CDH1* and the majority of them had the potential to generate truncated proteins (59).

Systematic analysis revealed that the usage of two intronic poly(A) sites of *CDH1* in seven samples, which located in the downstream of exon 10 and exon 12 respectively, were abnormally increased in the presence of related SNVs (Figure 5A; Supplementary Figure S9A). Significant drop of RNA-seq coverage in corresponding intron region

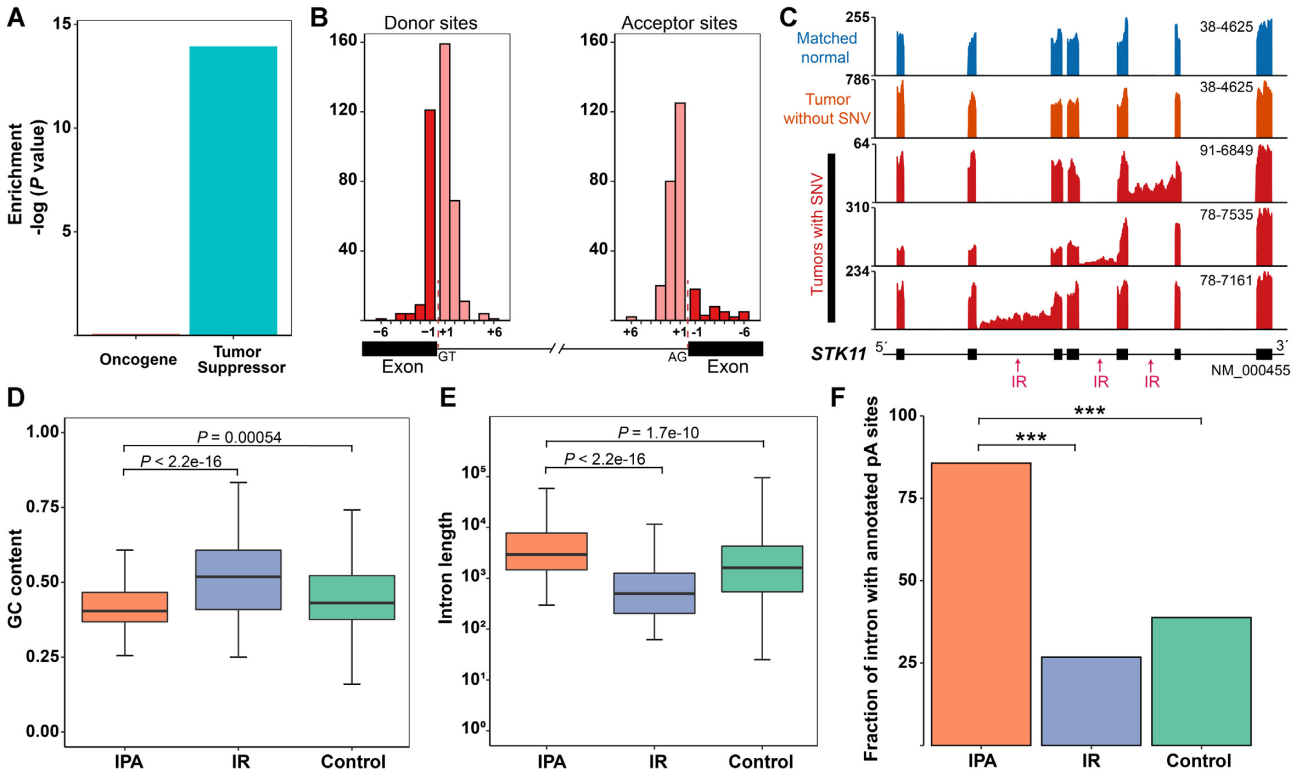

Figure 3. Comparison of features between SNV-associated polyadenylated introns and retained introns. (**A**) Enrichment of IR-associated somatic SNVs in TSGs. All genes with SNVs in the tested ten cancer types (the total number is 18093) were used as the background gene set. (**B**) Evaluation of position-wise numbers of IR-associated SNVs between the sixth exonic bases (-6) and sixth intronic base (+6) for the donor and acceptor splice sites. (**C**) RNA-seq density plots showing that SNV-associated intron retention events occur frequently in the tumor-suppressor gene *STK11*. The top two plots show that intron retention does not occur in *STK11* in two samples without SNV around the splice sites. Sample IDs are shown at the top-right corner of corresponding plots. (**D**) Box plot for GC contents of introns with SNV-associated IPA event (IPA), introns with SNV-associated intron retention event (IR) and all introns annotated by RefSeq (Control). (**E**) Box plot for lengths (bp) of introns with SNV-associated IPA event (IPA), introns with SNV-associated intron retention event (IR) and all introns annotated by RefSeq (Control). The *P* value was calculated based on two-sided Wilcoxon rank-sum test for panels D and E. (**F**) Bar plots showing the fraction of introns with SNV-associated IPA event (IPA), introns with SNV-associated intron retention event (IR) and all introns annotated by RefSeq (Control) overlapping with annotated poly(A) sites compiled from RefSeq, Ensembl, UCSC gene models and poly(A) site databases including PolyA_DB 3 and PolyASite 2.0. ***, $P < 0.001$, χ^2 test.

and reduced usage of downstream exons supported that they were genuine IPA events in *CDH1* (Figure 5A; Supplementary Figure S9A). The mutant allele was only observed in RNA-seq reads mapping to the retained part of the intron, supporting the direct association between abnormal intronic polyadenylation and somatic variants (Supplementary Figure S9B). Using minigene assays, we confirmed that the SNV increased the usage of IPA site (Figure 5B). The truncated proteins generated by abnormal intronic polyadenylation of *CDH1* may lack the tumor-suppressive function of the original full-length proteins due to the loss of C-terminal transmembrane domain and intracytoplasmic domain (Figure 5C), which are essential for directly interacting with β-catenin and p120 catenin (60,61). Together, our analysis reveals that the premature cleavage and intronic polyadenylation caused by somatic SNVs can potentially serve as another mechanism of *CDH1* inactivation.

DISCUSSION

Previous studies have revealed that somatic SNVs near the exon-intron boundaries can affect splicing (12,43). In this study, we provided evidence for the first time that somatic SNVs could cause the generation of truncated transcript by increasing the usage of intronic poly(A) sites, which showed a potency for TSG inactivation similar to traditional truncating mutations. Whether somatic SNVs lead to the choice of intron retention or intronic polyadenylation could be affected by GC content, intron size and the existence of certain regulatory cis-elements. Importantly, these SNV-associated IPA events were enriched in tumor suppressor genes but not oncogenes, indicating some kind of selection pressure existed since intronic polyadenylation usually led to functional inactivation by producing truncated protein (17,62). Also, we found that different SNVs on the same genomic positions could cause the same IPA results, such as SNVs in *PTEN* and *CDH1* (Figures 4C and 5A). Our work revealed an unknown mechanism explaining cancer-causing somatic SNVs.

Widespread upregulation of truncated mRNAs and proteins generated by intronic polyadenylation has been reported to affect genes with tumor-suppressive functions in primary chronic lymphocytic leukemia (CLL) cells (17). However, these recurrent upregulated CLL-IPA events are not associated with somatic variants (Supplementary Figure S10). Human PCF11 enhances genome-wide cleavage

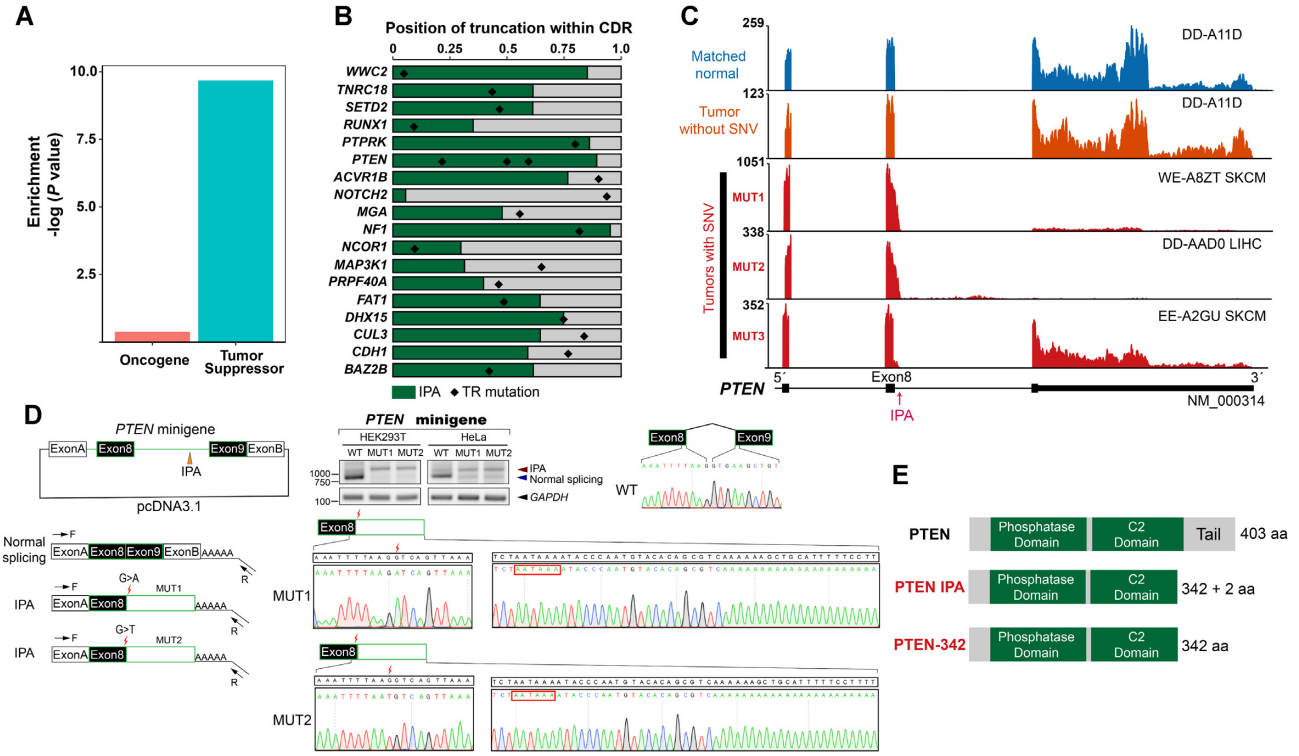

Figure 4. SNV-associated abnormal intronic polyadenylation in TSGs. (**A**) Enrichment of IPA-associated somatic SNVs in TSGs. (**B**) Tumor suppressor genes with truncating (TR) mutations and SNV-associated IPA events are shown. Dark green bars indicate the fraction of retained coding regions (CDRs) for each peptide generated by IPA isoform. Black dots indicate the positions of TR mutation. (**C**) RNA-seq density plots showing SNV-associated IPA events in the tumor-suppressor gene *PTEN*. Sample IDs are shown at the top-right corner of corresponding plots. (**D**) Experiments to validate SNV-associated intronic polyadenylation in *PTEN*. Sketch map of minigene structure and primers designed to amplify the IPA and normal splicing isoforms are shown on the left. F and R indicate forward and reverse primers, respectively. *GAPDH* serves as the internal control. Black rectangles denote exons from target genes while exon A and B are sequences from vector pcDNA3.1. Agarose gel electrophoresis for 3′ rapid amplification of cDNA ends (3′ RACE) using primer pairs is shown on the right. Both HEK293T and HeLa cells were transfected with wild-type and mutant minigene construct. 3′ RACE was performed to detect polyadenylated mRNAs using total RNA from minigene-transfected cells. Sanger sequencing results of the 3′ RACE products for the target gene are shown with the corresponding genomic sequence in black. The putative poly(A) signal is indicated in red-bordered box. (**E**) Diagrams illustrating domain information of full-length and IPA-generated truncated protein of PTEN, with known domains shown in green. For IPA isoform, the numbers of retained (342 aa) and novel amino acids (2 aa) are given. PTEN-342 denotes a PTEN C-terminal deletion mutant ending at codon 342.

and polyadenylation (22), and depletion of PCF11 leads to 3′ UTR lengthening and decreases the usage of IPA sites (63). However, we found that the usage of SNV-associated IPA sites did not have significant changes upon *PCF11* knockdown in HeLa cells, as exemplified by IPA sites in *ZNF330* and *CUL7* (Supplementary Figure S11). In the present study, we found that IPA-associated somatic variants favored the localization near the donor splice sites and disrupted the binding site of U1 snRNP. In contrast, IR-associated SNVs relatively uniformly distributed around donor splice sites and acceptor splice sites (Figure 3B). By integrating RNA-seq data derived from HeLa cells treated with U1 AMO, we found that a proportion of IPA events caused by SNV around 5′ splice site overlapped with IPA events triggered by U1 inhibition (Figure 2D; Supplementary Figure S4). These results support the notion that somatic variants near donor splice site were more likely to cause IPA may through disrupting the base-paring between U1 snRNP and pre-mRNA.

To examine whether SNV-associated IPA events are caused by expression level change of certain trans-acting factor, we downloaded and analyzed the raw sequencing data from TREND-DB (23,24). We found that reads mapped to SNV-associated unannotated IPA sites were far less than those mapped to 3′ UTR (Supplementary Figure S12A). In addition, depletion of different proteins couldn't significantly increase the usage of both SNV-associated unannotated IPA sites in *CUL3* and *WWC2* and annotated IPA sites in *TYW1* and *VEGFA* (Supplementary Figure S6C and S12B). Interestingly, we found that only knocking down SF3A1 could significantly increase the usage of SNV-associated IPA site in *PTEN* (Supplementary Figure S12C and D). However, the expression levels of *SF3A1* in all three samples containing SNV-associated IPA events were higher than median expression levels of control samples without SNV-associated IPA events (Supplementary Figure S12E). As downregulation of *SF3A1* led to increased IPA site, therefore, higher expression levels of *SF3A1* in these three samples were unlikely to contribute to the elevation of IPA event in *PTEN*. Combining with our minigene experiments (Figure 4D and Supplementary Figure S2), we conclude that SNVs serve as the major contributor to increased usage of IPA site rather than the abundance change of trans-acting factors at least in these candidate genes.

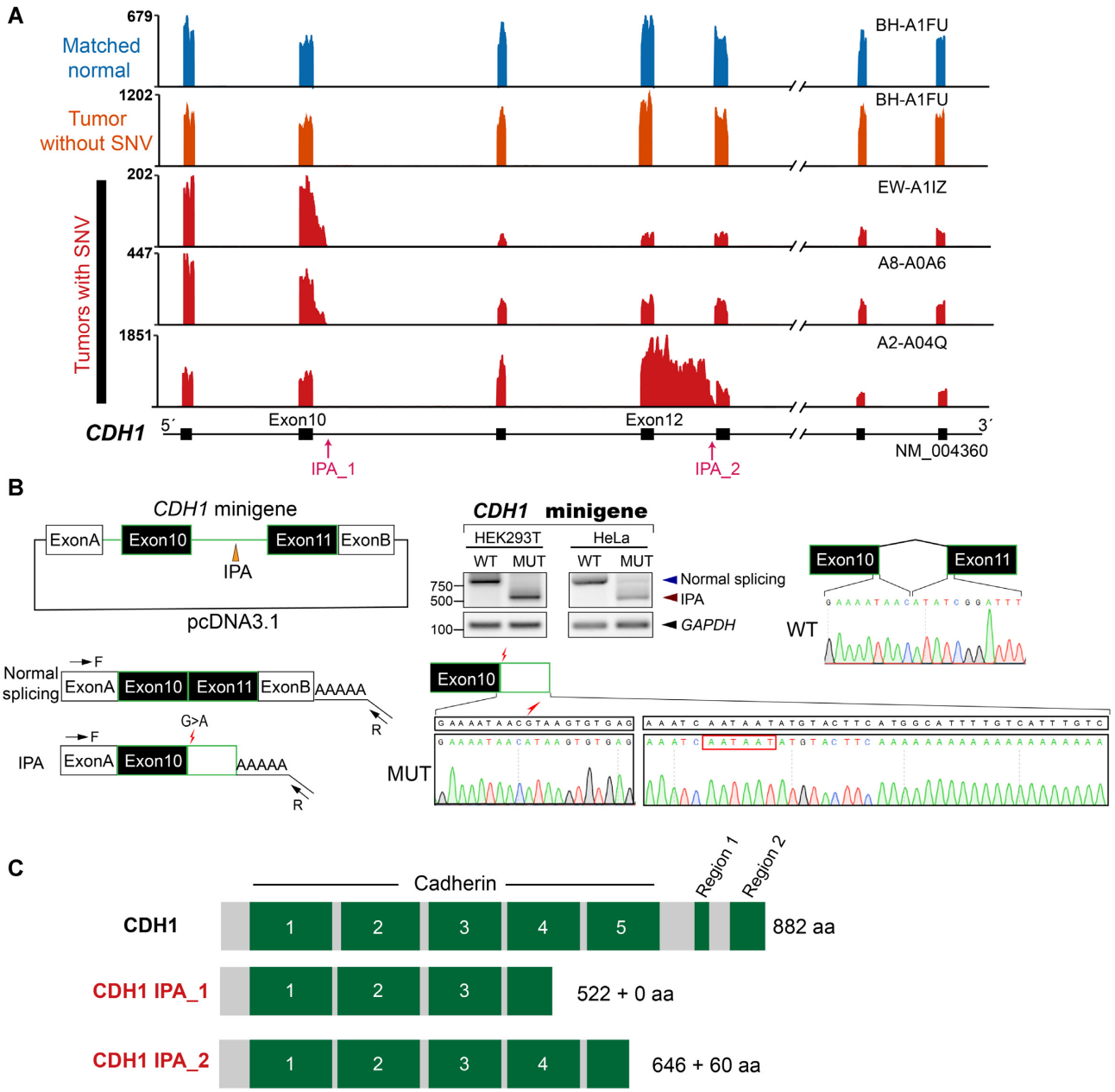

Figure 5. SNV-associated IPA events frequently affect *CDH1*. (**A**) RNA-seq density plots showing that SNV-associated IPA events frequently affect gene *CDH1*. Sample IDs are shown at the top-right corner of corresponding plots. Intronic poly(A) sites (IPA) are indicated by red arrows. (**B**) Experiments to validate SNV-associated intronic polyadenylation in *CDH1*. Sketch map of minigene structure and primers designed to amplify the IPA and normal splicing isoforms are shown on the left. F and R indicate forward and reverse primers, respectively. *GAPDH* serves as the internal control. Agarose gel electrophoresis for 3′ RACE using primer pairs is shown on the right. Sanger sequencing results of the 3′ RACE products for the target gene are shown with the corresponding genomic sequence in black. The putative poly(A) signal is indicated in red-bordered box. (**C**) Diagrams illustrating domain information of full-length and IPA-generated truncated protein of CDH1, with known domains shown in green. For the two IPA isoforms (IPA_1 and IPA_2), the numbers of retained (before +) and novel amino acids (after +) are given. Region 1 is required for binding CTNND1 and PSEN1 and region 2 is required for binding alpha, beta and gamma catenins.

Jung *et al.* showed that SNV (G > T) at last base of exon 8 in *PTEN* caused intron retention by minigene assay (12). However, we found the same SNV in *PTEN* led to intronic polyadenylation (Figure 4C and D). We speculated that the inconsistence is due to the different strategies of minigene assays. We noticed that they inserted exon 8 and 140 bp flanking intronic sequences into the pDUP4-1 vector (12), where the partial intronic sequence does not cover the intronic poly(A) site discovered in our study. In our minigene design, we cloned the fragment containing full-length sequences of exon 8, intron 8 and exon 9 into the pcDNA3.1 to validate the impact of SNV on intronic polyadenylation in *PTEN* (Figure 4D). The two studies both confirmed that the same SNV in *PTEN* could lead to abnormal splicing.

When inserting an incomplete intron sequence (140 bp), minigene assay will determine it's an intron retention event (partial intron retention at the 5′ end of intron 8). However, when inserting a full-length intron (4168 bp), minigene assay will determine it's an intronic polyadenylation event.

Somatic variants near splice sites have been reported to induce different forms of abnormal splicing, such as exon skipping, intron retention and activation of cryptic splice site (43). Here, we extended the consequence of somatic SNVs to intronic polyadenylation in addition to alternative splicing changes. These two types of events could even occur in the same gene but at different locations. For example, tumor suppressor gene *CDH1* had both SNV-associated intronic polyadenylation and SNV-associated intron retention (Figure 5A; Supplementary Figure S13). Intronic polyadenylation is one of the consequences of SNV around splice site and thus the number of IPA-related SNV is small compared with the number of mutations around splice site. *CDH1* was the most frequently altered gene among genes with SNV-associated IPA. However, BRCA patients with *CDH1* SNV-IPA (six of them are still alive until last follow up so we cannot obtain their overall survival time) don't have worse survival than other BRCA patients (Supplementary Figure S14). We speculated that the reason explaining why we could not establish association between SNV-associated IPA events and overall survival time is possibly due to the relatively few numbers of patients with SNV-associated IPA events.

Overall, our work highlights the importance of integrating transcriptome and genome sequencing data for fully understanding the functional and the clinical implications of somatic variants in human diseases.

DATA AVAILABILITY

IPAFinder is freely available at https://github.com/ZhaozzReal/IPAFinder. We provide source code and tested files to show how to detect SNV-associated IPA events and SNV-associated IR events (https://github.com/ZhaozzReal/SNV_IPA).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Prof. Liye Zhang for his valuable advices during analyzing the data.
Author contributions: T.N., Z.Z. and G.W. designed the study. Z.Z. performed the bioinformatics and statistical analyses. Q.X. performed the experiments. T.N. supervised the bioinformatics and statistical analyses. The manuscript was initially drafted by Z.Z. and Q.X. and then was revised by T.N. and G.W. All authors read and approved the final manuscript.

FUNDING

National Key Research and Development Program of China [2018YFC1003500]; National Natural Science Foundation of China [91949107, 31771336, 31521003]; Shanghai Municipal Science and Technology Major Project [2017SHZDZX01]. Funding for open access charge: National Natural Science Foundation of China.
Conflict of interest statement. None declared.

REFERENCES

1. Vogelstein,B., Papadopoulos,N., Velculescu,V.E., Zhou,S., Diaz,L.A. Jr and Kinzler,K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.
2. Martincorena,I. and Campbell,P.J. (2015) Somatic mutation in cancer and normal cells. *Science*, **349**, 1483–1489.
3. Bailey,M.H., Tokheim,C., Porta-Pardo,E., Sengupta,S., Bertrand,D., Weerasinghe,A., Colaprico,A., Wendl,M.C., Kim,J., Reardon,B. *et al.* (2018) Comprehensive characterization of cancer driver genes and mutations. *Cell*, **173**, 371–385.
4. Huang,F.W., Hodis,E., Xu,M.J., Kryukov,G.V., Chin,L. and Garraway,L.A. (2013) Highly recurrent TERT promoter mutations in human melanoma. *Science*, **339**, 957–959.
5. Li,K., Zhang,Y., Liu,X., Liu,Y., Gu,Z., Cao,H., Dickerson,K.E., Chen,M., Chen,W., Shao,Z. *et al.* (2020) Noncoding variants connect enhancer dysregulation with nuclear receptor signaling in hematopoietic malignancies. *Cancer Discov.*, **10**, 724–745.
6. Weinhold,N., Jacobsen,A., Schultz,N., Sander,C. and Lee,W. (2014) Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.*, **46**, 1160–1165.
7. Melton,C., Reuter,J.A., Spacek,D.V. and Snyder,M. (2015) Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.*, **47**, 710–716.
8. Rheinbay,E., Parasuraman,P., Grimsby,J., Tiao,G., Engreitz,J.M., Kim,J., Lawrence,M.S., Taylor-Weiner,A., Rodriguez-Cuevas,S., Rosenberg,M. *et al.* (2017) Recurrent and functional regulatory mutations in breast cancer. *Nature*, **547**, 55–60.
9. Rheinbay,E., Nielsen,M.M., Abascal,F., Wala,J.A., Shapira,O., Tiao,G., Hornshøj,H., Hess,J.M., Juul,R.I., Lin,Z. *et al.* (2020) Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, **578**, 102–111.
10. Venables,J.P. (2004) Aberrant and alternative splicing in cancer. *Cancer Res.*, **64**, 7647–7654.
11. Supek,F., Miñana,B., Valcárcel,J., Gabaldón,T. and Lehner,B. (2014) Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, **156**, 1324–1335.
12. Jung,H., Lee,D., Lee,J., Park,D., Kim,Y.J., Park,W.Y., Hong,D., Park,P.J. and Lee,E. (2015) Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.*, **47**, 1242–1248.
13. Niwa,M., Rose,S.D. and Berget,S.M. (1990) In vitro polyadenylation is stimulated by the presence of an upstream intron. *Genes Dev.*, **4**, 1552–1559.
14. Tian,B., Pan,Z. and Lee,J.Y. (2007) Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.*, **17**, 156–165.
15. Nourse,J., Spada,S. and Danckwardt,S. (2020) Emerging roles of RNA 3′-end cleavage and polyadenylation in pathogenesis, diagnosis and therapy of human disorders. *Biomolecules*, **10**, 915.
16. Higgs,D.R., Goodbourn,S.E., Lamb,J., Clegg,J.B., Weatherall,D.J. and Proudfoot,N.J. (1983) Alpha-thalassaemia caused by a polyadenylation signal mutation. *Nature*, **306**, 398–400.
17. Lee,S.H., Singh,I., Tisdale,S., Abdel-Wahab,O., Leslie,C.S. and Mayr,C. (2018) Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature*, **561**, 127–131.
18. Mittleman,B.E., Pott,S., Warland,S., Zeng,T., Mu,Z., Kaur,M., Gilad,Y. and Li,Y. (2020) Alternative polyadenylation mediates genetic regulation of gene expression. *Elife*, **9**, e57492.
19. Ellrott,K., Bailey,M.H., Saksena,G., Covington,K.R., Kandoth,C., Stewart,C., Hess,J., Ma,S., Chiotti,K.E., McLellan,M. *et al.* (2018) Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.*, **6**, 271–281.
20. Singh,I., Lee,S.H., Sperling,A.S., Samur,M.K., Tai,Y.T., Fulciniti,M., Munshi,N.C., Mayr,C. and Leslie,C.S. (2018) Widespread intronic

polyadenylation diversifies immune cell transcriptomes. *Nat. Commun.*, **9**, 1716.
21. So,B.R., Di,C., Cai,Z., Venters,C.C., Guo,J., Oh,J.M., Arai,C. and Dreyfuss,G. (2019) A complex of U1 snRNP with cleavage and polyadenylation factors controls telescripting, regulating mRNA transcription in human cells. *Mol. Cell*, **76**, 590–599.
22. Kamieniarz-Gdula,K., Gdula,M.R., Panser,K., Nojima,T., Monks,J., Wiśniewski,J.R., Riepsaame,J., Brockdorff,N., Pauli,A. and Proudfoot,N.J. (2019) Selective roles of vertebrate PCF11 in premature and full-length transcript termination. *Mol. Cell*, **74**, 158–172.
23. Marini,F., Scherzinger,D. and Danckwardt,S. (2021) TREND-DB-a transcriptome-wide atlas of the dynamic landscape of alternative polyadenylation. *Nucleic Acids Res.*, **49**, D243–D253.
24. Ogorodnikov,A., Levin,M., Tattikota,S., Tokalov,S., Hoque,M., Scherzinger,D., Marini,F., Poetsch,A., Binder,H., Macher-Göppinger,S. *et al.* (2018) Transcriptome 3′end organization by PCF11 links alternative polyadenylation to formation and neuronal differentiation of neuroblastoma. *Nat. Commun.*, **9**, 5331.
25. Zhao,Z., Xu,Q., Wei,R., Wang,W., Ding,D., Yang,Y., Yao,J., Zhang,L., Hu,Y.Q., Wei,G. and Ni,T. (2021) Cancer-associated dynamics and potential regulators of intronic polyadenylation revealed by IPAFinder using standard RNA-seq data. *Genome Res.*, https://doi.org/10.1101/gr.271627.120.
26. Ni,T., Yang,W., Han,M., Zhang,Y., Shen,T., Nie,H., Zhou,Z., Dai,Y., Yang,Y., Liu,P. *et al.* (2016) Global intron retention mediated gene regulation during CD4+ T cell activation. *Nucleic Acids Res.*, **44**, 6817–6829.
27. Yao,J., Ding,D., Li,X., Shen,T., Fu,H., Zhong,H., Wei,G. and Ni,T. (2020) Prevalent intron retention fine-tunes gene expression and contributes to cellular senescence. *Aging Cell*, **19**, e13276.
28. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
29. Boyle,A.P., Guinney,J., Crawford,G.E. and Furey,T.S. (2008) F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**, 2537–2538.
30. Ni,T., Yang,Y., Hafez,D., Yang,W., Kiesewetter,K., Wakabayashi,Y., Ohler,U., Peng,W. and Zhu,J. (2013) Distinct polyadenylation landscapes of diverse human tissues revealed by a modified PA-seq strategy. *BMC Genomics*, **14**, 615.
31. Chen,M., Lyu,G., Han,M., Nie,H., Shen,T., Chen,W., Niu,Y., Song,Y., Li,X., Li,H. *et al.* (2018) 3′ UTR lengthening as a novel mechanism in regulating cellular senescence. *Genome Res.*, **28**, 285–294.
32. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.
33. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
34. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
35. Davoli,T., Xu,A.W., Mengwasser,K.E., Sack,L.M., Yoon,J.C., Park,P.J. and Elledge,S.J. (2013) Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, **155**, 948–962.
36. Cerami,E., Gao,J., Dogrusoz,U., Gross,B.E., Sumer,S.O., Aksoy,B.A., Jacobsen,A., Byrne,C.J., Heuer,M.L., Larsson,E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.
37. Gao,J., Aksoy,B.A., Dogrusoz,U., Dresdner,G., Gross,B., Sumer,S.O., Sun,Y., Jacobsen,A., Sinha,R., Larsson,E. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal*, **6**, pl1.
38. Kaida,D., Berg,M.G., Younis,I., Kasim,M., Singh,L.N., Wan,L. and Dreyfuss,G. (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, **468**, 664–668.
39. Wang,R., Nambiar,R., Zheng,D. and Tian,B. (2018) PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res.*, **46**, D315–D319.
40. Herrmann,C.J., Schmidt,R., Kanitz,A., Artimo,P., Gruber,A.J. and Zavolan,M. (2020) PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Res.*, **48**, D174–D179.
41. Faustino,N.A. and Cooper,T.A. (2003) Pre-mRNA splicing and human disease. *Genes Dev.*, **17**, 419–437.
42. Berg,M.G., Singh,L.N., Younis,I., Liu,Q., Pinto,A.M., Kaida,D., Zhang,Z., Cho,S., Sherrill-Mix,S., Wan,L. *et al.* (2012) U1 snRNP determines mRNA length and regulates isoform expression. *Cell*, **150**, 53–64.
43. Shiraishi,Y., Kataoka,K., Chiba,K., Okada,A., Kogure,Y., Tanaka,H., Ogawa,S. and Miyano,S. (2018) A comprehensive characterization of cis-acting splicing-associated variants in human cancer. *Genome Res.*, **28**, 1111–1125.
44. Calabrese,C., Davidson,N.R., Demircioğlu,D., Fonseca,N.A., He,Y., Kahles,A., Lehmann,K.V., Liu,F., Shiraishi,Y., Soulette,C.M. *et al.* (2020) Genomic basis for RNA alterations in cancer. *Nature*, **578**, 129–136.
45. Sakabe,N.J. and de Souza,S.J. (2007) Sequence features responsible for intron retention in human. *BMC Genomics*, **8**, 59.
46. Braunschweig,U., Barbosa-Morais,N.L., Pan,Q., Nachman,E.N., Alipanahi,B., Gonatopoulos-Pournatzis,T., Frey,B., Irimia,M. and Blencowe,B.J. (2014) Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.*, **24**, 1774–1786.
47. Basu,T.N., Gutmann,D.H., Fletcher,J.A., Glover,T.W., Collins,F.S. and Downward,J. (1992) Aberrant regulation of ras proteins in malignant tumour cells from type 1 neurofibromatosis patients. *Nature*, **356**, 713–715.
48. Bollag,G., Clapp,D.W., Shih,S., Adler,F., Zhang,Y.Y., Thompson,P., Lange,B.J., Freedman,M.H., McCormick,F., Jacks,T. *et al.* (1996) Loss of NF1 results in activation of the Ras signaling pathway and leads to aberrant growth in haematopoietic cells. *Nat. Genet.*, **12**, 144–148.
49. Cichowski,K., Santiago,S., Jardim,M., Johnson,B.W. and Jacks,T. (2003) Dynamic regulation of the Ras pathway via proteolysis of the NF1 tumor suppressor. *Genes Dev.*, **17**, 449–454.
50. Johannessen,C.M., Reczek,E.E., James,M.F., Brems,H., Legius,E. and Cichowski,K. (2005) The NF1 tumor suppressor critically regulates TSC2 and mTOR. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 8573–8578.
51. Li,J., Yen,C., Liaw,D., Podsypanina,K., Bose,S., Wang,S.I., Puc,J., Miliaresis,C., Rodgers,L., McCombie,R. *et al.* (1997) PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science*, **275**, 1943–1947.
52. Song,M.S., Salmena,L. and Pandolfi,P.P. (2012) The functions and regulation of the PTEN tumour suppressor. *Nat. Rev. Mol. Cell Biol.*, **13**, 283–296.
53. Maehama,T. and Dixon,J.E. (1998) The tumor suppressor, PTEN/MMAC1, dephosphorylates the lipid second messenger, phosphatidylinositol 3,4,5-trisphosphate. *J. Biol. Chem.*, **273**, 13375–13378.
54. Leslie,N.R., Yang,X., Downes,C.P. and Weijer,C.J. (2007) PtdIns(3,4,5)P(3)-dependent and -independent roles for PTEN in the control of cell migration. *Curr. Biol.*, **17**, 115–125.
55. Terrien,E., Chaffotte,A., Lafage,M., Khan,Z., Préhaud,C., Cordier,F., Simenel,C., Delepierre,M., Buc,H., Lafon,M. *et al.* (2012) Interference with the PTEN-MAST2 interaction by a viral protein leads to cellular relocalization of PTEN. *Sci. Signal*, **5**, ra58.
56. Georgescu,M.M., Kirsch,K.H., Akagi,T., Shishido,T. and Hanafusa,H. (1999) The tumor-suppressor activity of PTEN is regulated by its carboxyl-terminal region. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 10182–10187.
57. Christofori,G. and Semb,H. (1999) The role of the cell-adhesion molecule E-cadherin as a tumour-suppressor gene. *Trends Biochem. Sci.*, **24**, 73–76.
58. (2014) Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, **513**, 202–209.
59. Ciriello,G., Gatza,M.L., Beck,A.H., Wilkerson,M.D., Rhie,S.K., Pastore,A., Zhang,H., McLellan,M., Yau,C., Kandoth,C. *et al.* (2015) Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, **163**, 506–519.
60. Nagafuchi,A. and Takeichi,M. (1988) Cell binding function of E-cadherin is regulated by the cytoplasmic domain. *EMBO J.*, **7**, 3679–3684.

61. Shimoyama,Y., Nagafuchi,A., Fujita,S., Gotoh,M., Takeichi,M., Tsukita,S. and Hirohashi,S. (1992) Cadherin dysfunction in a human cancer cell line: possible involvement of loss of alpha-catenin expression in reduced cell-cell adhesiveness. *Cancer Res.*, **52**, 5770–5774.
62. Dubbury,S.J., Boutz,P.L. and Sharp,P.A. (2018) CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature*, **564**, 141–145.
63. Li,W., You,B., Hoque,M., Zheng,D., Luo,W., Ji,Z., Park,J.Y., Gunderson,S.I., Kalsotra,A., Manley,J.L. *et al.* (2015) Systematic profiling of poly(A)+ transcripts modulated by core 3′ end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLoS Genet.*, **11**, e1005166.