Recurrent disruption of tumour suppressor genes in cancer by somatic mutations in cleavage and polyadenylation signals

Yaroslav Kainov[1,2,3], Fursham Hamid[1] and Eugene V. Makeyev[1,3]

1. Centre for Developmental Neurobiology, King's College London, London SE1 1UL, UK
2. Department of Medical and Molecular Genetics, King's College London, London SE1 9RT, UK
3. Correspondence:
   Centre for Developmental Neurobiology
   New Hunt's House,
   King's College London, London, SE1 1UL
   United Kingdom
   eugene.makeyev@kcl.ac.uk or yaroslav.kainov@kcl.ac.uk

The authors declare no potential conflicts of interest

Abstract

The expression of eukaryotic genes relies on the precise 3'-terminal cleavage and polyadenylation of newly synthesized pre-mRNA transcripts. Defects in these processes have been associated with various diseases, including cancer. While cancer-focused sequencing studies have identified numerous driver mutations in protein-coding sequences, noncoding drivers – particularly those affecting the cis-elements required for pre-mRNA cleavage and polyadenylation – have received less attention. Here, we systematically analysed cancer somatic mutations affecting 3'UTR polyadenylation signals using the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset. We found a striking enrichment of cancer-specific somatic mutations that disrupt strong and evolutionarily conserved cleavage and polyadenylation signals within tumour suppressor genes. Further bioinformatics and experimental analyses conducted as a part of our study suggest that these mutations have a profound capacity to downregulate the expression of tumour suppressor genes. Thus, this work uncovers a novel class of noncoding somatic mutations with significant potential to drive cancer progression.

Introduction

Most eukaryotic mRNAs are modified by the addition of a 5'-terminal 7-methylguanosine cap, splicing of intronic sequences, and 3'-terminal cleavage and polyadenylation. The cleavage and polyadenylation reactions are tightly coupled with transcription termination and the release of newly synthesized transcripts from RNA polymerase II. Therefore, precise cleavage and polyadenylation are critical for the production of mature mRNAs. Mechanistically, these reactions require a co-transcriptional assembly of a multisubunit protein complex at the corresponding cis-regulatory sequences near the pre-mRNA cleavage site (CS) [1,2,3]. A key cis-element guiding the assembly of the cleavage and polyadenylation machinery is the polyadenylation signal (PAS). The most common PAS sequences in mammals are AATAAA and ATTAAA hexamers, although their single nucleotide-substituted variants may function in some cases [4,5].

Earlier studies have emphasized the importance of pre-mRNA cleavage/polyadenylation in the context of human diseases. For example, alternative cleavage/polyadenylation has been proposed to modulate the expression of oncogenes and tumour suppressors in different types of cancer [6,7]. Germline mutations affecting polyadenylation signals can play a role in genetic disorders [8-10] and increase cancer susceptibility [11,12]. Notably, although the role of somatic mutations affecting polyadenylation signals has been investigated for individual

genes in a limited number of tumour samples [13,14], a systematic characterization of the role of this type of mutations in cancer has not been carried out.

Large-scale genome sequencing studies have identified numerous cancer driver and driver-like mutations within protein-coding sequences [15]. Such mutations have also been mapped to noncoding regions; however, existing research has primarily focused on promoters, enhancers, and splicing signals [16-20], rather than sequences regulating pre-mRNA cleavage and polyadenylation.

Here, we conducted a systematic genome-wide analysis of somatic single-nucleotide variants (SNVs) affecting the PAS elements in mRNA 3' untranslated regions (3'UTRs) in cancer cells. Using a large tumour whole-genome sequencing dataset, the Pan-Cancer Analysis of Whole Genomes (PCAWG) [15], we found that strong and evolutionarily conserved clavage/polyadenylation signals are often disrupted by cancer-specific SNVs. Strikingly, such mutations are significantly enriched in tumour suppressor genes. We further provide evidence that such mutations can substantially decrease the expression of tumour suppressor genes in cancer cells. Overall, our work identifies a novel class of noncoding somatic mutations with driver-like properties in cancer.

Results

Somatic mutations often disrupt cleavage and polyadenylation sequences in cancer

We first analysed SNVs neighbouring annotated human cleavage and polyadenylation positions (paSNVs) in 3'UTRs from the PolyA_DB3 database [21]. We considered two distinct cohorts: "Normal" paSNVs from a healthy human population (the 1000 Genomes phase 3 data [22]) and "Cancer" paSNVs from the whole-genome sequencing of cancer samples (PCAWG) [15].

For each paSNV, we calculated the change in cleavage/polyadenylation efficiency using the APARENT2 neural network model, which has been shown to infer this statistic more accurately than earlier approaches [23]. We additionally assessed the loss and gain of the two strongest polyadenylation signals, AATAAA and ATTAAA (referred to as AWTAAA throughout this study; Fig. 1A).

As expected, paSNVs predicted to have a strong impact on cleavage/polyadenylation were often situated immediately upstream of a CS, with most of them affecting AWTAAA hexamers (Fig. 1B). Furthermore, the presence of AATAAA or ATTAAA in polyadenylation signals tended to be associated with a high APARENT2 score (Fig. S1A), and the loss or

gain of AATAAA led to significantly stronger decrease or increase in the score, respectively, compared to other mutations (Fig. S1B). This provided internal validation of the algorithm's performance in our hands.

We then categorized all paSNVs into three groups: (1) upregulating cleavage/polyadenylation (UP-paSNVs), defined as events with a ≥1 increase in the APARENT2 score (log odds ratio, or LOR) and creating an AWTAAA hexamer; (2) downregulating cleavage/polyadenylation (DOWN-paSNVs; LOR≤-1 and disrupting an AWTAAA); and (3) the remaining annotated cleavage position-adjacent paSNVs (Fig. 1B). The latter group served as a background control (BG-paSNVs) in our subsequent analyses.

Consistent with the earlier studies [23,24,25], we observed a pronounced negative selection against the DOWN-paSNVs in the Normal dataset (Fig. 1C-D). This category showed significantly decreased allele frequencies in comparison to the BG-paSNVs and was enriched for singletons (unique variants in the analysed dataset). This effect was more evident when considering changes in both the score and the hexamer composition (Fig. S1C-D).

Notably, comparison of the Normal and Cancer datasets showed that cancer somatic mutations, on average, had a stronger effect on the polyadenylation efficiency in both the UP- and DOWN-paSNV groups (Fig. 1B; LOR sample variance 0.115 in cancer vs 0.0876 in normal). DOWN-paSNVs were significantly enriched in cancer compared to the normal population data (Fig. 2A). We also observed that mutations disrupting AWTAAA hexamers in 3'UTRs tended to occur near annotated cleavage sites in cancer (Fig. 2B).

Interestingly, cancer-specific DOWN-paSNVs affected cleavage/polyadenylation signals with higher APARENT2 scores (Fig. 2C). Furthermore, DOWN-paSNVs tended to affect more evolutionarily conserved sequences in the Cancer dataset compared to the Normal control (Fig. 2D and Fig. S2). In total, we identified 1614 distinct somatic DOWN-paSNVs in the cancer dataset affecting 1570 cleavage/polyadenylation events in 1460 genes in 602 tumours, i.e. 22.7% of all tumour samples in PCAWG (Table S1). Notably, nearly half of the DOWN-paSNVs originated from colorectal adenocarcinoma, with the remaining mutations distributed across a wide range of other cancer types (Fig. 2E).

We concluded that mutations disrupting functional cleavage/polyadenylation signals are abundant in cancer cells despite being subject to strong purifying selection in a healthy population.

Cancer-specific mutations in cleavage and polyadenylation sequences are enriched in tumour suppressor genes

There are two possible explanations for the enrichment of DOWN-paSNV events in cancer: (1) an increase in the overall mutation load and (2) positive selection for such mutations. Since the latter possibility may increase the incidence of mutations in cancer driver genes, we analysed the distribution of DOWN-paSNVs within genes from the Cancer Gene Census [26]. This revealed a remarkable over-representation of the DOWN-paSNVs in tumour suppressor genes, with the magnitude of this effect being greater than the corresponding enrichment of nonsense mutations (SNVs creating a premature translation termination codon) (Fig. 3A-B and Fig. S3).

Notably, DOWN-paSNVs were not enriched in the oncogenes (Fig. 3C), in line with the disruptive nature of such mutations under normal conditions (Fig. 1). Conversely, oncogenes but not tumour suppressors showed some enrichment for UP-paSNVs (Fig. Fig. 3A and Fig. S3).

Overall, DOWN-paSNVs were found to affect 38 tumour suppressor genes, i.e. 14.3% of all genes in this category in the Census dataset. In several cases, including *LRP1B* and *FOXO1*, which are known to act as tumour suppressors in certain cancers, the same signal/polyadenyalation signal was disrupted by the same or different mutations in more than one sample (see columns Mut_Recurrence and Signal_Recurrence in Table S1). Consistent with tumour suppressors being a major target of DOWN-paSNVs, genes with this type of mutations were significantly enriched for apoptosis-related functions (Fig. 3D; e.g. tumour suppressors *CASP9* or *FHIT*). We also detected enrichment of proteins interacting with the ubiquitination pathway (Fig. 3D-E; e.g., tumour suppressors *SMAD2*, *APC* and *AXIN1*).

To independently confirm the functional impact of DOWN-paSNVs in cancer, we compared the mutational excess of different types of somatic mutations using DigDriver [18], a neural network-based method that accounts for cancer-specific mutation rates. This analysis revealed a significantly higher observed-to-expected mutation rate for DOWN-paSNV events in cancer compared to the BG-paSNV group (Fig. S4). DOWN-paSNVs tended to be enriched in tumour suppressor genes, consistent with positive selection for these events in cancer (Fig. 4A). No such enrichment was detected in oncogenes (Fig. 4B),

Of note, our analysis of wild-type sequences showed that tumour suppressor 3'UTRs are characterized by stronger cleavage/polyadenylation signals compared to oncogenes and non-cancer genes (Fig. S5A-B). Moreover, tumour suppressors associated with hallmarks of cancer [27] in the Census dataset had stronger cleavage/polyadenylation signals than the rest of tumour suppressor genes (Fig. S5C).

According to the classical two-hit hypothesis [28], both alleles of tumour suppressor genes may acquire distinct damaging mutations in cancer. With this in mind, we analysed the co-occurrence of paSNVs with damaging non-synonymous mutations from the PCAWG collection (Non-syn. variants from the binarized gene-centric table in [20]). DOWN-paSNV-containing tumour suppressors showed a markedly increased incidence of such additional somatic mutations in the same tumour compared to the BG-paSNV control (Fig. 4C). Furthermore, the overall frequency of damaging non-synonymous mutations in tumour suppressors affected by DOWN-paSNVs in at least one sample was significantly higher than in the DOWN-paSNV-negative tumour suppressor group (Fig. 4D).

The analysis of allele copy number variation (CNV) showed that increased copy number was 4.1-times more common in the PCAWG data compared to allele loss. However, the incidence of copy number increase was substantially lower in the DOWN-paSNV group compared to the BG-paSNV control (Fig. S6). This points to a negative selection against duplications of genes affected by DOWN-paSNVs in cancer.

Taken together, these data suggest that somatic mutations disrupting cleavage and polyadenylation can facilitate the inactivation of tumour suppressors in cancer.

Somatic mutations in cleavage and polyadenylation signals can decrease the expression of tumour suppressor genes

Genetic inactivation of functional cleavage/polyadenylation sequences may negatively affect gene expression (see e.g., Ref[8]). To explore this possibility, we turned to the colorectal adenocarcinoma subset of PCAWG, as it contained most of the DOWN-paSNVs in tumour suppressors and the corresponding gene expression information [20]. We shortlisted detectably expressed tumour suppressors that contained DOWN-paSNVs and no other damaging mutations in specific cancer samples, and were wild type in other samples. Seven genes passing these filters were involved in various aspects of tumour biology, including cell survival and DNA repair (*CASP9*, *NDRG1*, and *XPA*), mTOR signalling (*TSC1*), and transcription and RNA processing (*ETV6*, *ISY1* and *SMAD2*).

Plotting pairwise gene-specific expression differences for the aggregated tumour suppressor set, we observed a significant bias towards downregulation in the samples containing DOWN-paSNVs compared to the wild-type controls (Fig. 5A; median downregulation of 1.25-fold). Remarkably, similar negative biases were detected for all seven individual genes, with median downregulation values ranging from 1.1- to 3.2-fold (Fig. 5B).

To validate the effect of DOWN-paSNVs on gene expression, we focused on a somatic mutation that disrupts the cleavage/polyadenylation signal in the tumour suppressor *XPA*. This gene has been shown to promote apoptosis in response to DNA damage, in addition to its role in nucleotide excision repair [29]. Moreover, downregulation of *XPA* has been associated with decreased patient survival in colorectal cancer [30].

The *XPA* mutation identified by our bioinformatics analyses alters the canonical AATAAA PAS hexamer to GATAAA near the terminal CS and significantly reduces the APARENT2 score (Fig. 5C). To experimentally assess the effect of this mutation on the efficiency of pre-mRNA cleavage and polyadenylation, we prepared minigene constructs where the wild-type or mutant sequences were inserted upstream of a recombinant CS (Fig. 5D).

We used the wild-type and mutant minigenes to transfect the human colorectal cancer cell line HCT-116. An RT-qPCR assay measuring the efficiency of cleavage/polyadenylation as a ratio between the CS-read-through and CS-upstream signals revealed a significant decrease in cleavage/polyadenylation efficiency in response to the *XPA* DOWN-paSNV (Fig. 5D).

To directly assess the effect of defective cleavage/polyadenylation on gene expression, the wild-type or the mutant 3'UTR sequences were inserted downstream of a luciferase reporter gene. Following the transfection of HCT-116 cells, we detected a significantly reduced production of luciferase protein from the mutant construct compared to the wild-type control (Fig. 5E).

Thus, somatic mutations disrupting polyadenylation signals in tumour suppressor genes can reduce the abundance of functional mRNA transcripts.

Discussion

We interrogated whole-genome mutation data using recently developed machine learning approaches to systematically characterize the impact of SNVs on 3'UTR polyadenylation signals (PAS) in cancer. Our analyses confirm that germline SNVs disrupting PAS are likely deleterious, as they are subjected to strong negative selection in the normal population (Fig. 1). Intriguingly, we found that somatic mutations affecting such cis-elements in cancer are more prevalent, tend to occur near stronger CSs, and target more evolutionarily conserved PAS hexamers (Fig. 2 and Fig. S2).

Importantly, these cancer somatic SNVs disrupt PAS sequences in tumour suppressor genes with a similar enrichment pattern to well-known deleterious SNVs in protein-coding regions, such as nonsense mutations (Fig. 3 and Fig. 4). Additionally, wild-type tumour suppressors

have stronger cleavage/polyadenylation signals than other groups of genes (Fig. S5), pointing to the importance of the corresponding steps of pre-mRNA processing for their expression.

Consistent with the two-hit hypothesis [28], we found that tumour suppressors with disrupted cleavage/polyadenylation signals (i.e. containing DOWN-paSNVs) are more likely to acquire other damaging somatic mutations in the same tumour (Fig. 4C-D). As long-read genomic sequencing data become increasingly available, it will be interesting to investigate whether these additional mutations occur in the same or in a different allele compared to the DOWN-paSNVs.

However, it is possible that DOWN-paSNVs can contribute to tumour progression even in the absence of other mutations. Indeed, tumour suppressors containing only DOWN-paSNVs are consistently expressed at lower levels compared to their wild-type counterparts (Fig. 5A-B). Moreover, it is currently thought that partial inactivation of many tumour suppressors can be sufficient to promote tumorigenesis [31,32].

Using the tumour suppressor gene *XPA* as an example, we directly show that a cancer-specific single-nucleotide mutation disrupting the PAS hexamer is sufficient to block pre-mRNA cleavage/polyadenylation and dampen the expression of mature mRNA (Fig. 5C-E). These results support our bioinformatics analyses and argue that SNVs targeting polyadenylation signals can have a profound effect on gene expression in cancer. Our data are also consistent with previous reports showing similar gene expression effects of PAS-specific germline SNVs [8,10].

It is expected that mutation of cleavage/polyadenylation signals should lead to the appearance of abnormal read-through transcripts that may be destabilized by either nuclear or cytoplasmic RNA quality control mechanisms [33]. Alternatively, a decrease in cleavage/polyadenylation activity might dampen transcription initiation, as these two processes are known to be interconnected [34]. Differentiating between these possibilities will be an important next step in understanding the molecular mechanisms, which may link compromised cleavage/polyadenylation and gene expression defects in cancer. Furthermore, although we focused on annotated 3'UTR CSs in this work, similar analyses of SNVs occurring in other noncoding parts of mammalian genes (e.g. introns) might reveal an even wider impact of the loss and gain of PAS-like sequences in cancer.

In conclusion, our study reveals that the genetic inactivation of cleavage and polyadenylation in tumour suppressor genes constitutes a prevalent, yet previously overlooked category of somatic cancer mutations with driver properties. These findings emphasize the importance of pre-mRNA processing in the biology of cancer and underscore the need for improved

functional annotation of single nucleotide variants in noncoding regions of the human genome.

Acknowledgements

We thank Maria Andrianova, Yegor Bazykin, Tim Hubbard, and Snezhka Oliferenko for helpful discussions. The results presented here are based in part on data generated by The Cancer Genome Atlas managed by the NCI and NHGRI (https://www.cancer.gov/ccg/research/genome-sequencing/tcga). This work was supported by the Biotechnology and Biological Sciences Research Council (grants BB/V006258/1 and BB/R001049/1 to E.V.M.).

Materials and Methods

Source data sets

Pre-mRNA cleavage site (CS) positions and the corresponding metadata were obtained from the PolyA_DB3 database [21] (release 3.2 https://exon.apps.wistar.org/PolyA_DB/v3/). The phase-3 1000 genomes vcf files were downloaded from the International Genome Sample Resource (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/). Cancer somatic SNVs and indels from whole-genome sequencing of 2,583 unique tumours (PCAWG) were downloaded from the International Cancer Genome Consortium (ICGC) data portal (https://dcc.icgc.org/) and the database of Genotypes and Phenotypes (dbGaP) (project code: phs000178). Only bona fide SNVs that differed from the reference genome at a single-nucleotide position were included in the analysis. The v97 release of the Cancer Gene Census was downloaded from https://cancer.sanger.ac.uk/cosmic/download.

Data processing

CSs located in 3'UTRs according to PolyA_DB3 were extended by 102 nt on both sides to generate 205-nt intervals. All SNVs from the 1000 genomes and the PCAWG datasets mapping to these intervals were kept for further analyses (paSNVs). FASTA files corresponding to wild-type and mutant 205-nt intervals analysed by the APARENT2[23]. For each variant we estimated log odds ratio (LOR) of mutant (mut) variant isoform abundance

with respect to the wild-type (wt) abundance (abundance were calculated by summing all cleavage probabilities mapping to 205-nt interval) as follows:

$$LOR = ln(mut) \ / \ (1 - ln \ (mut)) - ln \ (wt) \ / \ (1 - ln \ (wt))$$

Incidence of PAS hexamers was quantified using the vcountPattern function from the Biostrings R/Bioconductor package (doi:10.18129/B9.bioc.Biostrings). Evolutionary conservation was calculated for either exact SNV positions or 15-nt SNV-centred windows using the GenomicScores [35] and the phastCons100way.UCSC.hg19 [36] R/Bioconductor packages. Only unique SNV entries were kept for further analysis. In cases where a single SNV was located near more than one distinct CS, the strongest effect on cleavage/polyadenylation was used for further analyses. GO terms enrichment was analysed using the ClusterProfiler R/Bioconductor package [37].

To analyse changes in polyadenylation scores of all mutations affecting AWTAAA sequences in 3'UTRs in Fig. 2B, APARENT2 scores were calculated for all SNV-centred 205-nt intervals from both datasets located within canonical UCSC 3'UTRs. SNVs disrupting AWTAAA sequence with LOR≤-1 within 100-nt intervals centred around polyA_DB3 CSs were considered "annotated".

Cancer Census gene enrichment

Enrichment of different types of SNVs in Cancer Census genes was calculated using two-tailed Fisher's exact test. Somatic SNVs in protein-coding sequences were classified as "Nonsense", "Missense", or "Synonymous" based on the information provided in PCAWG maf files ("Variant_Classification" column). Tumour suppressors were defined as genes labelled as "TSG" but not "Oncogene" in the Census dataset. A similar stringent approach was used to define oncogenes. Genes annotated as both "Tumour suppressors'' and "Oncogenes" were excluded (most analyses), analysed as "Both" (Fig. Fig. 3A and Fig. S3), or combined with tumour suppressors to form the extended "Tumour suppressor+" group (Fig. S5B).

DigDriver enrichment analysis

We used the "Analyzing new mutation sets" mode of DigDriver to process different functional categories of somatic SNVs. Functional annotation was taken from DigPreprocess.py annotMutationFile output files. Enrichment/excess of mutations of Census cancer gene category was calculated as:

$$\frac{\frac{\sum observed\ mut\ cat1}{\sum expected\ mut\ cat1}}{\frac{\sum observed\ mut\ noncancer\ genes}{\sum expected\ mut\ noncancer\ genes}}$$

To calculate the 95% Confidence Interval (CI) of this enrichment, we performed bootstrap resampling of tumour suppressors, oncogenes and non-cancer genes in each mutation class for 1000 iterations. In each iteration, the enrichment/excess of mutations was calculated as described above. The 2.5th and 97.5th percentiles of the resampled distribution were used as the 95% confidence interval boundaries.

Gene expression analysis

To analyse possible effect of DOWN-paSNVs on transcript abundance, we selected tumour suppressor genes from the published colorectal adenocarcinoma study [20], which contained DOWN-paSNVs and no other damaging SNVs (i.e. Non-syn. variants from the binarized gene-centric table in [20]) in some samples, and no mutations in other samples. We normalized the available gene expression data (FPKM) to account for gene copy number variation and $\log_2$-tranformed them to obtain Log2(nFPKM) values. Gene-specific Log2(nFPKM) values for the wild-type samples were then subtracted from corresponding Log2(nFPKM) values for the DOWN-paSNVs samples to obtain distributions of gene expression differences (ΔLog2(nFPKM)). A one-tailed Wilcoxon signed-rank test was used to analyse the significance of a negative shift of ΔLog2(nFPKM) distributions compared to 0.

DNA constructs

To generate read-through *XPA* minigenes, 431-nt gBlock fragments (Integrated DNA Technologies) containing the human *XPA* 3'UTR in its natural context (chr9: 100436867-100437297: GRCh37/hg19) and either the wild-type or mutated PAS were cloned into the pEGFP-N3 plasmid (Clontech) at the BsrGI and NotI sites. To generate luciferase reporter plasmids, the entire *XPA* 3'UTR (chr9: 100437071-100437680; GRCh37/hg19) was amplified from HCT-116 genomic DNA using KAPA HiFi DNA polymerase HotStart ReadyMix (Roche, cat# KK2601) with MLO4220 (5'-AACGCTAGCAAATAAAGGAAATTTAGATTGGTCCT-3') and MLO4221 (5'-ATCGGTCGACTCAACAATCAGATAGTCAACCATGA-3') primers. The PCR product was gel-purified and cloned into the pGL3-control plasmid (Promega) at the XbaI and SalI sites. The cancer-specific PAS mutation was introduced using a modified Quikchange site-directed mutagenesis protocol, using the KAPA HiFi DNA polymerase HotStart ReadyMix (Roche,

cat# KK2601) with MLO4159 (5'-GCCCTAATAGCAGAGATAAACATTGAGTTG-3') and MLO4160 (5'-CAACTCAATGTTTATCTCTGCTATTAGGGC-3') primers. All constructs were verified by Sanger sequencing. Plasmid maps are available on request.

Minigenes experiments

HCT-116 cells (ATCC® CCL-247) were cultured in a humidified incubator at 37°C, 5% CO2, in DMEM containing 4.5 g/L glucose, GlutaMAX and 110 mg/L sodium pyruvate (Thermo Fisher Scientific, cat# 11360070) supplemented with 10% FBS (Hyclone, cat# SV30160.03) and 100 units/ml PenStrep (Thermo Fisher Scientific, cat# 15140122). For passaging, cells were washed with 1 × PBS and dissociated in 0.05% Trypsin-EDTA (Thermo Fisher Scientific, cat# 15400054) for 10 min at 37°C.

For read-through minigene transfection experiments, cells were typically seeded overnight in 1 mL of culture medium at 1-2 × 10^5 per well of a 12-well plate. Next morning, 1 µg of plasmid DNA was mixed with 2.5 µl of Jetprime transfection reagent in 150 µl of Jetprime transfection buffer (Polyplus, cat# 101000015), incubated for 10 min at RT and added drop-wise to the cells. Total RNAs were extracted from cells 24 hours post-transfection using TRIzol (Thermo Fisher Scientific, cat#15596026) with an additional acidic phenol-chloroform (1:1) extraction step. The aqueous phase was precipitated with an equal volume of isopropanol, washed with 70% ethanol, and dissolved in 80 µl of nuclease-free water (Thermo Fisher Scientific, cat# AM9939). RNA samples were then treated with 4-6 units of Turbo DNase (Thermo Fisher Scientific, cat# AM2238) at 37 °C for 30 min to remove the bulk of plasmid DNA contamination, extracted with an equal volume of acidic phenol–chloroform (1:1), precipitated with 3 volumes of 100% ethanol and 0.1 volume of 3 M sodium acetate (pH 5.2), washed with 70% ethanol and rehydrated in nuclease-free water. To remove any remaining traces of DNA, the RNA samples were additionally pre-treated with 2 units of RQ1-DNAse (Promega, cat# M6101) per 1 µg of RNA at 37 °C for 30 minutes. RQ1-DNAse was inactivated by adding the Stop Solution as recommended and the RNAs were immediately reverse-transcribed using SuperScript IV (Thermo Fisher Scientific, cat# 18090050) and random decamer (N10) primers at 50 °C for 30 min. cDNA samples were analysed by qPCR using a Light Cycler®96 Real-Time PCR System (Roche) and qPCR BIO SyGreen Master Mix (PCR Biosystems, cat# PB20.11-51). The RT-qPCR signals downstream of the *XPA* cleavage site (MLO944, 5'-GGCCGCGACTCTAGATCATAA-3' and MLO358, 5'-GTAACCATTATAAGCTGCAATAAACAAG-3') were normalized to those obtained using upstream primers (MLO775, 5'-AGAACGGCATCAAGGTGAAC-3' and MLO776, 5'-TGCTCAGGTAGTGGTTGTCG-3').

For luciferase minigene transfection experiments, cells were typically seeded overnight in 100 µl of culture medium at 5×10^3 per well of a 96-well plate. Next morning, 70 ng of a firefly luciferase reporter construct containing *XPA* sequences and 30 ng of the *Renilla* luciferase control (pRL-TK; Promega) were mixed with 0.2 µl of Jetprime transfection reagent in 10 µl of Jetprime transfection buffer (Polyplus, cat# 101000015), incubated for 10 min at RT and added drop-wise to the cells. Following a 24-hour incubation, transfected cells were analysed using a Dual-Glo® Luciferase Assay System kit (Promega, cat# E2920) as recommended by the manufacturer. Luminescence was measured using a Berthold Mithras LB940 plate reader.

Statistics

Unless stated otherwise, all statistical procedures were performed in R. Data were averaged from at least three experiments and shown as box plots, with box bounds representing the first and the third quartiles and whiskers extending from the first and the third quartile to the lowest and highest data points or, if there are outliers, 1.5× of the interquartile range. Data obtained from RT-qPCR and luciferase assays were compared using the two-tailed Student's t-test assuming unequal variances. Genome-wide data were analysed using Wilcoxon rank sum test or Fisher's exact test (two-tailed if not stated otherwise). Specific tests used and the *P*-values obtained are indicated in the figures and/or figure legends

References

1. Tian, B., and Manley, J.L. (2017). Alternative polyadenylation of mRNA precursors. Nature reviews. Molecular cell biology *18*, 18-30. 10.1038/nrm.2016.116.
2. Neve, J., Patel, R., Wang, Z., Louey, A., and Furger, A.M. (2017). Cleavage and polyadenylation: Ending the message expands gene regulation. RNA biology *14*, 865-890. 10.1080/15476286.2017.1306171.
3. Shi, Y., Di Giammartino, D.C., Taylor, D., Sarkeshik, A., Rice, W.J., Yates, J.R., 3rd, Frank, J., and Manley, J.L. (2009). Molecular architecture of the human pre-mRNA 3' processing complex. Molecular cell *33*, 365-376. 10.1016/j.molcel.2008.12.028.
4. Proudfoot, N. (1991). Poly(A) signals. Cell *64*, 671-674. 10.1016/0092-8674(91)90495-k.
5. Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome research *10*, 1001-1010. 10.1101/gr.10.7.1001.
6. Mayr, C., and Bartel, D.P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell *138*, 673-684. 10.1016/j.cell.2009.06.016.
7. 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. 10.1038/s41586-018-0465-8.

8. 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. 10.1038/306398a0.
9. Bogard, N., Linder, J., Rosenberg, A.B., and Seelig, G. (2019). A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. Cell *178*, 91-106 e123. 10.1016/j.cell.2019.04.046.
10. Bennett, C.L., Brunkow, M.E., Ramsdell, F., O'Briant, K.C., Zhu, Q., Fuleihan, R.L., Shigeoka, A.O., Ochs, H.D., and Chance, P.F. (2001). A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA-->AAUGAA) leads to the IPEX syndrome. Immunogenetics *53*, 435-439. 10.1007/s002510100358.
11. Stacey, S.N., Sulem, P., Jonasdottir, A., Masson, G., Gudmundsson, J., Gudbjartsson, D.F., Magnusson, O.T., Gudjonsson, S.A., Sigurgeirsson, B., Thorisdottir, K., et al. (2011). A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. Nature genetics *43*, 1098-1103. 10.1038/ng.926.
12. Li, B., Cai, Y., Chen, C., Li, G., Zhang, M., Lu, Z., Zhang, F., Huang, J., Fan, L., Ning, C., et al. (2023). Genetic variants that impact alternative polyadenylation in cancer represent candidate causal risk loci. Cancer research. 10.1158/0008-5472.CAN-23-0251.
13. Wiestner, A., Tehrani, M., Chiorazzi, M., Wright, G., Gibellini, F., Nakayama, K., Liu, H., Rosenwald, A., Muller-Hermelink, H.K., Ott, G., et al. (2007). Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. Blood *109*, 4599-4606. 10.1182/blood-2006-08-039859.
14. Shlien, A., Raine, K., Fuligni, F., Arnold, R., Nik-Zainal, S., Dronov, S., Mamanova, L., Rosic, A., Ju, Y.S., Cooke, S.L., et al. (2016). Direct Transcriptional Consequences of Somatic Mutation in Breast Cancer. Cell reports *16*, 2032-2046. 10.1016/j.celrep.2016.07.028.
15. Consortium, I.T.P.-C.A.o.W.G. (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82-93. 10.1038/s41586-020-1969-6.
16. Cao, S., Zhou, D.C., Oh, C., Jayasinghe, R.G., Zhao, Y., Yoon, C.J., Wyczalkowski, M.A., Bailey, M.H., Tsou, T., Gao, Q., et al. (2020). Discovery of driver non-coding splice-site-creating mutations in cancer. Nature communications *11*, 5573. 10.1038/s41467-020-19307-6.
17. Zhao, Z., Xu, Q., Wei, R., Huang, L., Wang, W., Wei, G., and Ni, T. (2021). Comprehensive characterization of somatic variants associated with intronic polyadenylation in human cancers. Nucleic acids research *49*, 10369-10381. 10.1093/nar/gkab772.
18. Sherman, M.A., Yaari, A.U., Priebe, O., Dietlein, F., Loh, P.R., and Berger, B. (2022). Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. Nature biotechnology *40*, 1634-1643. 10.1038/s41587-022-01353-8.
19. Rheinbay, E., Nielsen, M.M., Abascal, F., Wala, J.A., Shapira, O., Tiao, G., Hornshoj, 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. 10.1038/s41586-020-1965-x.
20. Group, P.T.C., Calabrese, C., Davidson, N.R., Demircioglu, D., Fonseca, N.A., He, Y., Kahles, A., Lehmann, K.V., Liu, F., Shiraishi, Y., et al. (2020). Genomic basis for RNA alterations in cancer. Nature *578*, 129-136. 10.1038/s41586-020-1970-0.
21. 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 research *46*, D315-D319. 10.1093/nar/gkx1000.
22. Fairley, S., Lowy-Gallego, E., Perry, E., and Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic acids research *48*, D941-D947. 10.1093/nar/gkz836.

23. Linder, J., Koplik, S.E., Kundaje, A., and Seelig, G. (2022). Deciphering the impact of genetic variation on human polyadenylation using APARENT2. Genome biology *23*, 232. 10.1186/s13059-022-02799-4.
24. Kainov, Y.A., Aushev, V.N., Naumenko, S.A., Tchevkina, E.M., and Bazykin, G.A. (2016). Complex Selection on Human Polyadenylation Signals Revealed by Polymorphism and Divergence Data. Genome biology and evolution *8*, 1971-1979. 10.1093/gbe/evw137.
25. Findlay, S.D., Romo, L., and Burge, C.B. (2022). Quantifying negative selection in human 3' UTRs uncovers constrained targets of RNA-binding proteins. bioRxiv, 2022.2011.2030.518628. 10.1101/2022.11.30.518628.
26. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic acids research *47*, D941-D947. 10.1093/nar/gky1015.
27. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674. 10.1016/j.cell.2011.02.013.
28. Knudson, A.G., Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences of the United States of America *68*, 820-823. 10.1073/pnas.68.4.820.
29. Deng, Y., Chen, Q.S., Huang, W.F., Dai, J.W., and Wu, Z.J. (2021). XPA serves as an autophagy and apoptosis inducer by suppressing hepatocellular carcinoma in a PI3K/Akt/mTOR dependent manner. Journal of gastrointestinal oncology *12*, 1797-1810. 10.21037/jgo-21-310.
30. Feng, X., Liu, J., Gong, Y., Gou, K., Yang, H., Yuan, Y., and Xing, C. (2018). DNA repair protein XPA is differentially expressed in colorectal cancer and predicts better prognosis. Cancer medicine *7*, 2339-2349. 10.1002/cam4.1480.
31. Berger, A.H., Knudson, A.G., and Pandolfi, P.P. (2011). A continuum model for tumour suppression. Nature *476*, 163-169. 10.1038/nature10275.
32. Park, S., Supek, F., and Lehner, B. (2021). Higher order genetic interactions switch cancer genes from two-hit to one-hit drivers. Nature communications *12*, 7051. 10.1038/s41467-021-27242-3.
33. Bresson, S., and Tollervey, D. (2018). Surveillance-ready transcription: nuclear RNA decay as a default fate. Open biology *8*. 10.1098/rsob.170270.
34. Mapendano, C.K., Lykke-Andersen, S., Kjems, J., Bertrand, E., and Jensen, T.H. (2010). Crosstalk between mRNA 3' end processing and transcription initiation. Molecular cell *40*, 410-422. 10.1016/j.molcel.2010.10.012.
35. Puigdevall, P., and Castelo, R. (2018). GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor. Bioinformatics *34*, 3208-3210. 10.1093/bioinformatics/bty311.
36. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research *15*, 1034-1050. 10.1101/gr.3715005.
37. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation *2*, 100141. 10.1016/j.xinn.2021.100141.

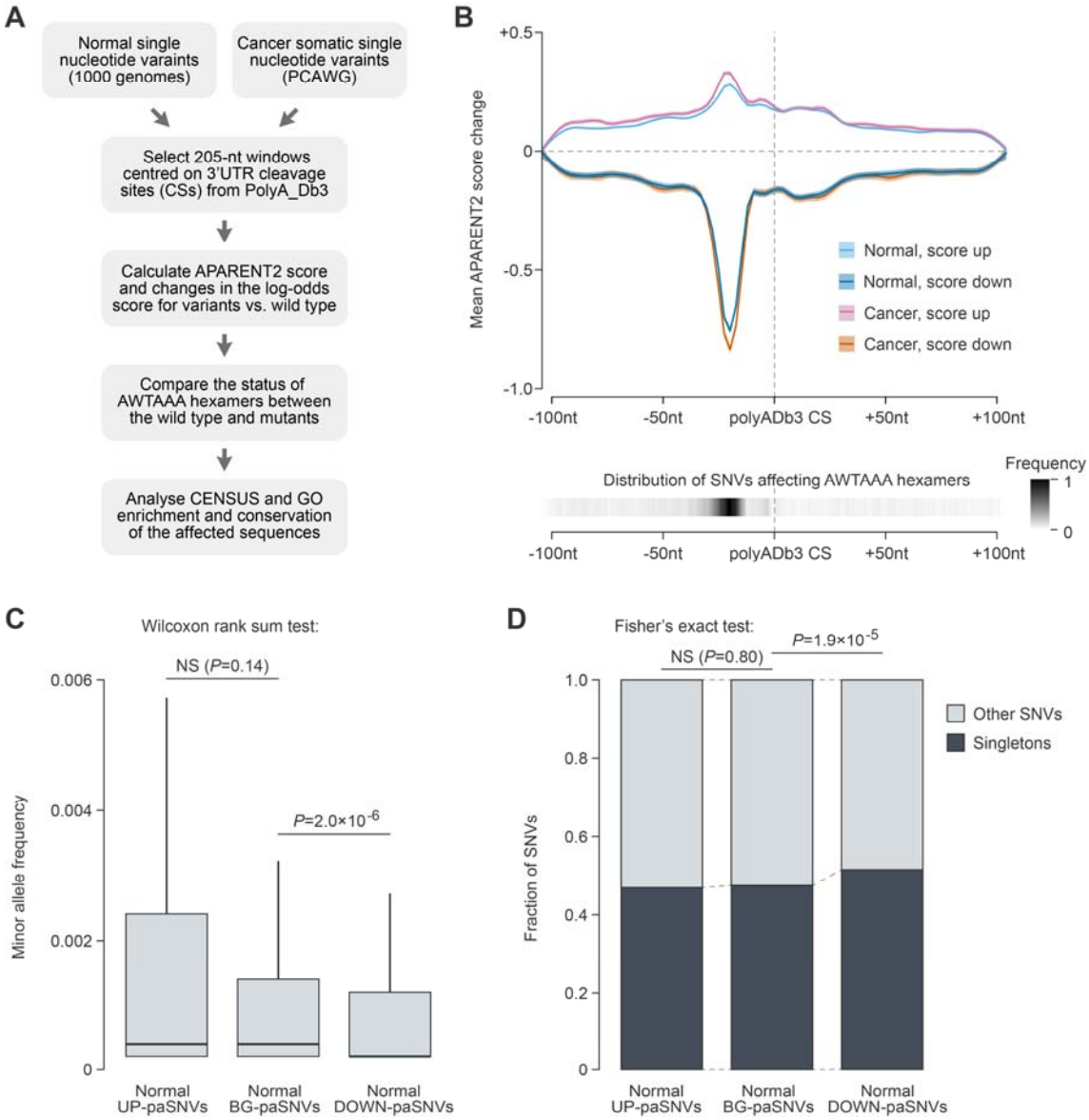

Figure 1. paSNVs disrupting cleavage/polyadenylation signals are depleted in the normal population.

(A) Bioinformatics workflow used to analyse the effect of paSNVs on pre-mRNA cleavage and polyadenylation.

(B) Top, effects of UP- and DOWN-paSNVs on the APARENT2 score (mean±SEM) as a function of their position with respect to annotated pre-mRNA cleavage sites (CSs). Bottom, combined distribution of AWTAAA-affecting paSNVs in both datasets.

(C) Box plot showing that paSNVs disrupting polyadenylation signals are significantly less frequent compared to control groups of events in normal population.

(**D)** paSNVs disrupting polyadenylation signals are enriched for singletons, consistent with purifying selection against such events in normal population.

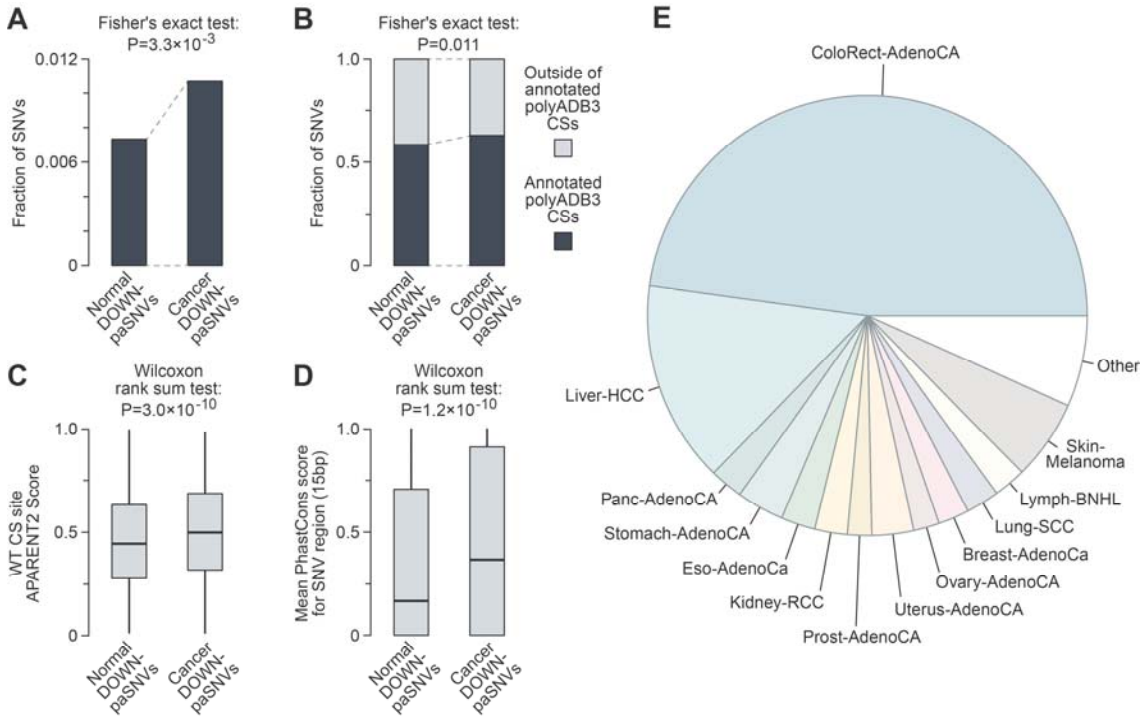

Figure 2. Cancer somatic mutations tend to disrupt functional cleavage/polyadenylation signals.

(A) Bar plot showing enrichment of paSNVs disrupting polyadenylation signals among cancer somatic mutations.

(B) Bar plot showing enrichment of SNVs affecting AWTAAA sequences in 3'UTRs close to annotated cleavage sites (CSs) among cancer somatic mutations.

(C) Box plot showing that somatic mutations disrupt stronger cleavage/polyadenylation signals in cancer.

(D) paSNVs disrupting polyadenylation signals occurs in more evolutionary conserved regions in cancer (mean PhastCons score in 15-nt window centred at SNVs).

(E) Distribution of DOWN-paSNVs across cancer types in the Pan-Cancer Analysis of Whole Genomes (PCAWG) project.

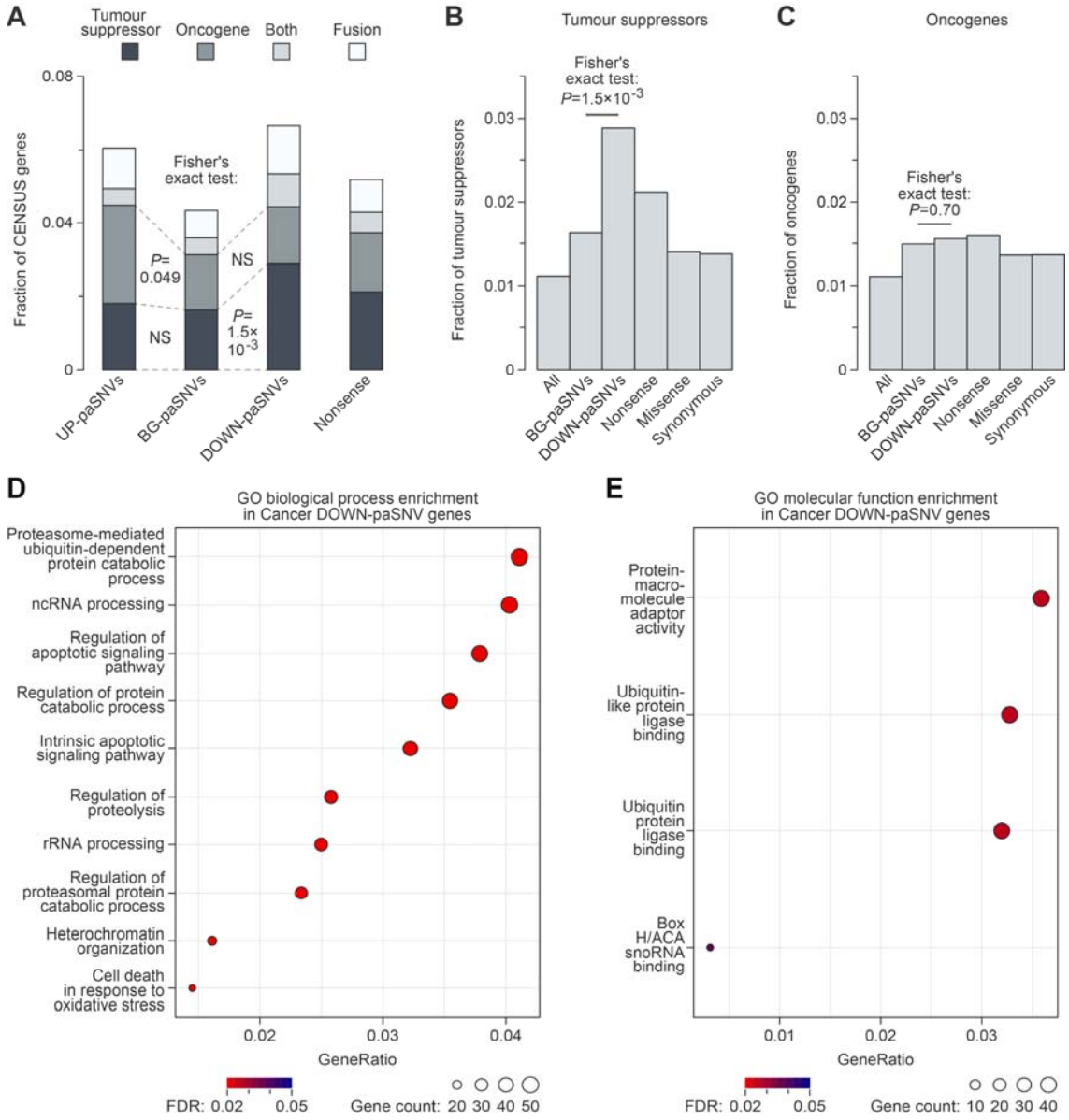

Figure 3. Somatic cancer mutations often disrupt cleavage/polyadenylation signals in tumour suppressor genes.

(A) Stacked bar plot showing enrichment of SNVs disrupting polyadenylation signals (DOWN-paSNVs) in tumour suppressors in cancer.

(B-C) Overrepresentation of (B) tumour suppressors but not (C) oncogenes among genes with cancer somatic DOWN-paSNVs, as compared to genes with cancer somatic BG-paSNVs. Fractions of tumour suppressors and oncogenes are also shown for all genes and genes containing cancer somatic nonsense (premature stop codons), missense (altered amino acid residues) and synonymous (synonymous codons) mutations. Note that the

enrichment of tumour suppressors is stronger for DOWN-paSNVs compared to nonsense mutations.

(C) Top 10 GO Biological Process terms significantly enriched in genes with cancer somatic DOWN-paSNVs. Note the enrichment of apoptosis- and cell death-related functions.

(D) GO Molecular Function terms significantly enriched in genes with cancer somatic DOWN-paSNVs.

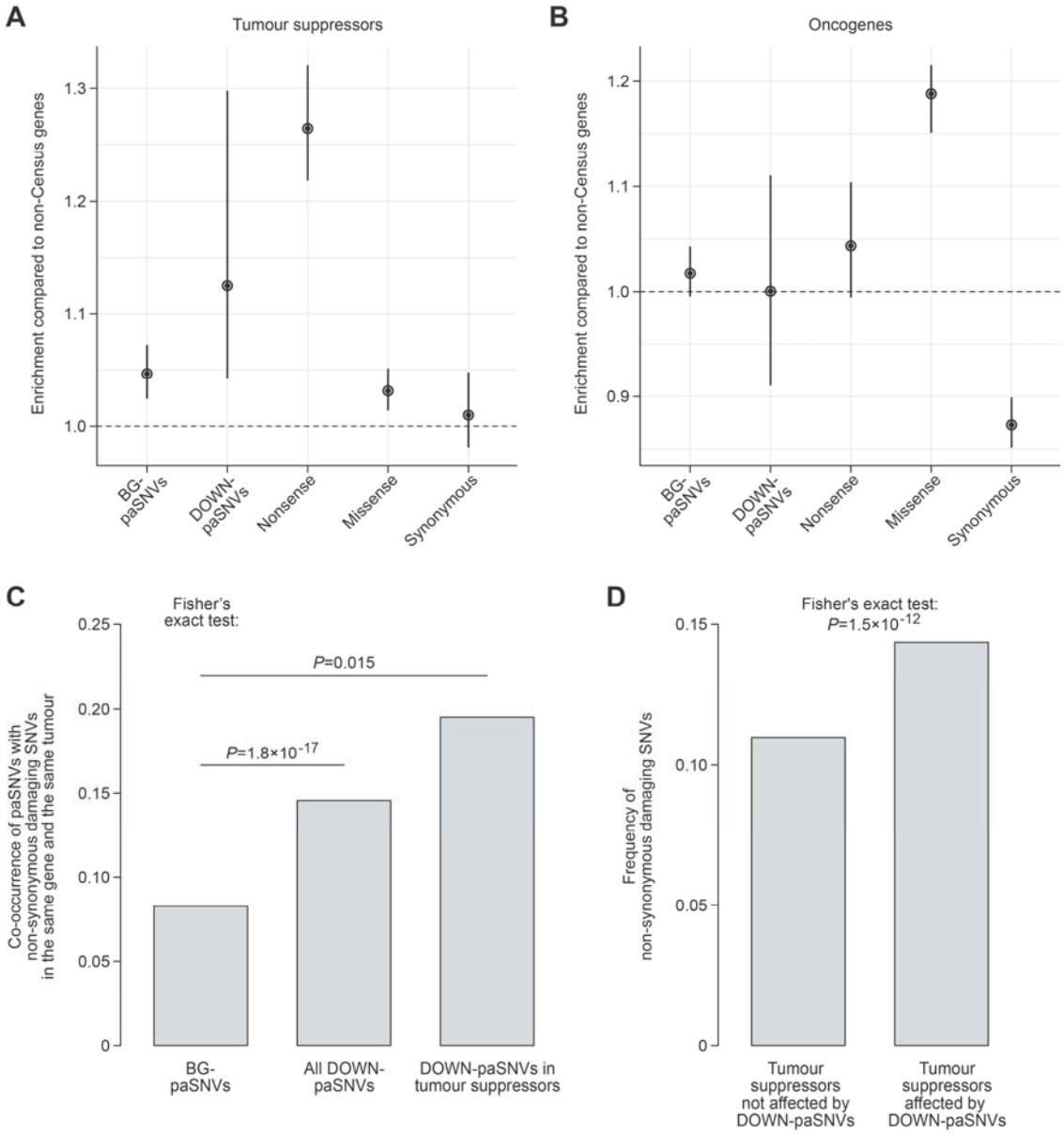

Figure 4. Disruption of cleavage/polyadenylation signals in tumour suppressors, along with other damaging mutations, may facilitate cancer progression.

(A-B) Enrichment of different groups of cancer somatic SNVs in (A) tumour suppressors and (B) oncogenes calculated using DigDriver relative to genes not listed in Cancer Census (non-Census) and presented with 95% confidence intervals. Note that DOWN-paSNVs and nonsense mutations are enriched in tumour suppressors but not in oncogenes. In contrast, oncogenes are often affected by missense mutations, as expected.

(C) Cancer somatic DOWN-paSNVs co-occur in the same tumour with non-synonymous damaging SNVs, a group of somatic mutations defined in [20], more often than BG-paSNVs. Note that the co-occurrence is particularly high for tumour suppressors.

(D) The overall frequency of non-synonymous damaging SNVs is significantly higher in the DOWN-paSNV-containing group compared to the DOWN-paSNV-lacking group of tumour suppressor genes.

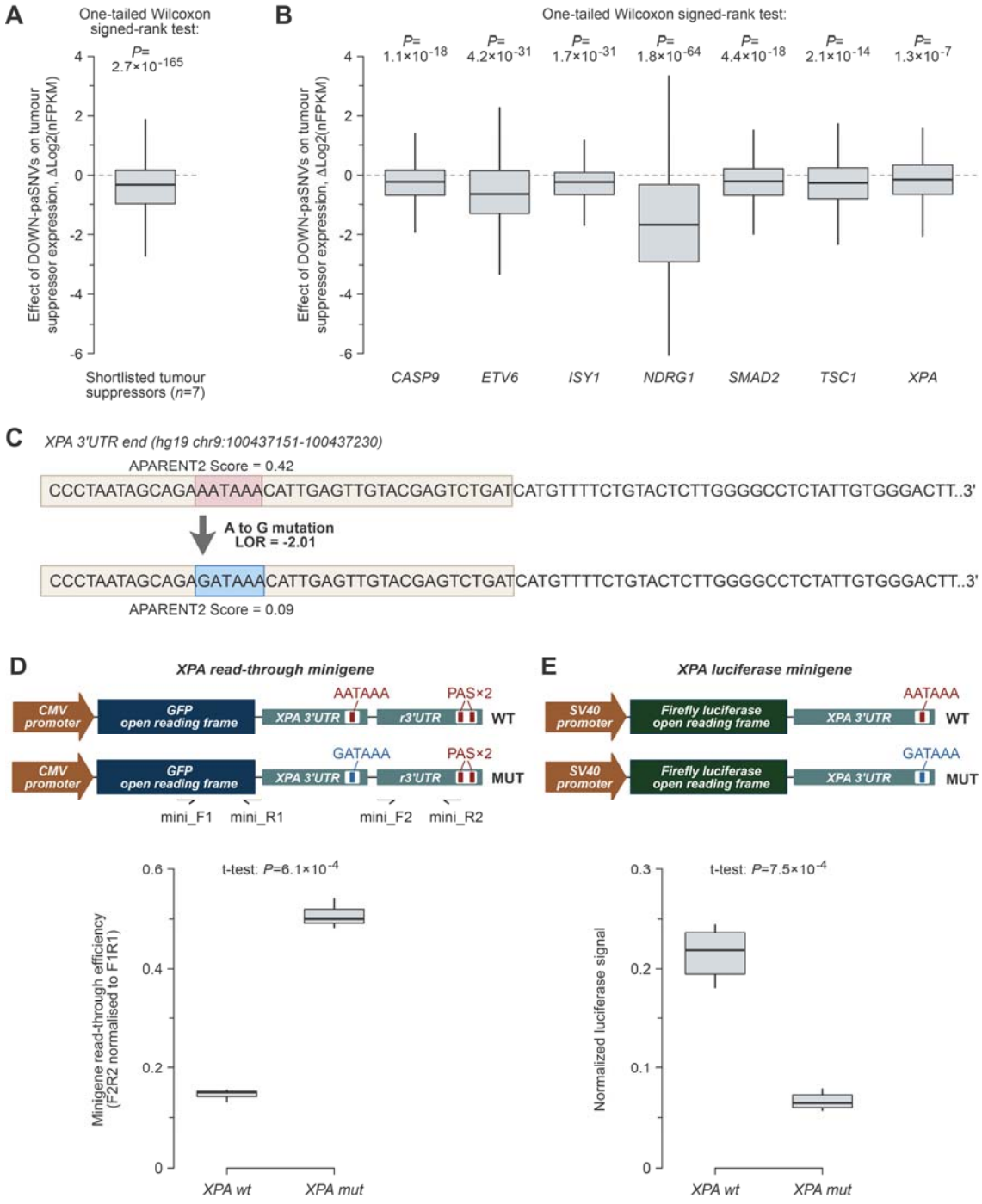

Figure 5. Somatic cancer DOWN-paSNVs are sufficient to downregulate tumour suppressor genes.

(A-B) Gene-specific expression differences between DOWN-paSNV-containing and wild-type samples (ΔLog2 of copy number variation-normalized FPKM values; see Materials and Methods) reveal a consistently negative effect of DOWN-paSNV on tumour suppressor

mRNA abundance in colorectal cancers. Box plots are shown for (A) an aggregated set of qualifying tumour suppressors and (B) individual genes from this set. Outliers are omitted for clarity.

(C) Wild-type and mutated sequences of the *XPA* tumour suppressor gene cleavage/polyadenylation signal. The PAS hexamer is enclosed within a box.

(D) Top, *XPA* cleavage site read-through minigenes and corresponding primers used for RT-qPCR analyses. Bottom, RT-qPCR data showing stronger read-through (weaker polyadenylation) in the mutant minigene.

(E) Top, luciferase expression minigenes. Bottom, luciferase assay revealing that the cancer-specific PAS mutation dampens the expression of the reporter gene.