SYSTEMS BIOLOGY

A widespread length-dependent splicing dysregulation in cancer

Sirui Zhang[1], Miaowei Mao[1], Yuesheng Lv[2], Yingqun Yang[1,3], Weijing He[4,5], Yongmei Song[6], Yongbo Wang[7], Yun Yang[1], Muthana Al Abo[8], Jennifer A. Freedman[8,9], Steven R. Patierno[8,9], Yang Wang[2]*, Zefeng Wang[1]*

Dysregulation of alternative splicing is a key molecular hallmark of cancer. However, the common features and underlying mechanisms remain unclear. Here, we report an intriguing length-dependent splicing regulation in cancers. By systematically analyzing the transcriptome of thousands of cancer patients, we found that short exons are more likely to be mis-spliced and preferentially excluded in cancers. Compared to other exons, cancer-associated short exons (CASEs) are more conserved and likely to encode in-frame low-complexity peptides, with functional enrichment in GTPase regulators and cell adhesion. We developed a CASE-based panel as reliable cancer stratification markers and strong predictors for survival, which is clinically useful because the detection of short exon splicing is practical. Mechanistically, mis-splicing of CASEs is regulated by elevated transcription and alteration of certain RNA binding proteins in cancers. Our findings uncover a common feature of cancer-specific splicing dysregulation with important clinical implications in cancer diagnosis and therapies.

INTRODUCTION

Over 95% of human genes undergo alternative splicing (AS) (*1*, *2*), generating multiple mRNA isoforms with distinct functions from a single gene (*3*). AS is tightly regulated by multiple cis-elements and trans-acting factors (*4*, *5*), and mis-regulation of AS is a common cause of human diseases (*6*). Particularly, the widespread splicing dysregulation is one of the molecular hallmarks of cancer, and the increasing evidence has suggested that the mutations in spliceosomal genes or dysregulations of splicing factors can drive various cancers (*7*–*9*). Because the aberrant splicing often affects functions of cancer-associated genes (*10*, *11*), targeting the mis-spliced genes (i.e., the genes whose splicing is significantly altered in cancer) becomes a powerful therapeutic strategy for cancers (*7*, *12*). In addition, cancer-associated AS events can serve as diagnostic biomarkers for cancer classification or prognosis (*10*, *12*). Therefore, a systematic study of AS in cancer is critical for cancer precision medicine.

Recently, the advances in high-throughput sequencing make it possible to systematically investigate the global change of AS and its regulation in cancers. In particular, the tremendous amounts of transcriptome data from thousands of cancer patients have been collected, providing a unique opportunity for systematic analyses of cancer-associated splicing alterations as well as their mechanisms and functional consequences. Several groups have analyzed the large-scale omics data from The Cancer Genome Atlas (TCGA) project for splicing changes in cancers (*13*–*15*). While these studies provided useful information, there is still a lack of general trend for cancer-associated splicing dysregulation, probably due to the cancer heterogeneity and complex mechanisms of splicing regulation (*16*, *17*). In addition, the mechanistic understanding of cancer-associated splicing dysregulation is inadequate, although several mutated or mis-regulated splicing factors were identified as oncogenes or tumor suppressors (*8*, *9*, *18*).

In this study, we conducted comprehensive analyses of AS changes using transcriptome data from thousands of patients in 18 types of cancers and found an unexpected length dependency in cancer-associated exons. Compared to typical exons, the short exons are more likely to be mis-spliced and preferentially excluded in almost all cancers. We further developed machine learning algorithms with these cancer-associated short exons (CASEs) as diagnostic markers and defined a CASE-based risk factor to accurately predict the prognosis of cancer patients. Last, we determined the possible mechanisms for such length-dependent regulation. Collectively, our results provide a deeper understanding and potential application of complex AS regulation in cancers.

RESULTS

Significant length biases in cancer-associated alternative exons

To systematically study AS in cancers, we analyzed the RNA-sequencing (RNA-seq) data from 6788 patients in TCGA project consisting of 18 cancer types. The PSI (percent-spliced-in) values of each annotated AS event across all samples were calculated, from which we identified potential events significantly altered in tumor versus matched normal tissues (Fig. 1A). The most common type of AS, known as the skipped exons (SEs), was selected for an in-depth analysis. In all types of cancers analyzed, the cancer-associated SEs tended to be shorter compared to all human exons or all SEs (Fig. 1B). Consistently, the length distributions of the cancer-associated exons versus all exons and all SEs also showed a more significant difference within

[1]CAS Key Laboratory of Computational Biology, CAS Center for Excellence in Molecular Cell Science, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China. [2]Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, China. [3]Shanghai Tech University, Shanghai 200031, China. [4]Department of Colorectal Surgery, Fudan University Shanghai Cancer Center, Shanghai 200032, China. [5]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, China. [6]State Key Laboratory of Molecular Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China. [7]Department of Cellular and Genetic Medicine, School of Basic Medical Sciences, Fudan University, Shanghai 200032, China. [8]Duke Cancer Institute, Duke University School of Medicine, Durham, NC 27710, USA. [9]Division of Medical Oncology, Department of Medicine, Duke University Medical Center, Durham, NC 27710, USA.
*Corresponding author. Email: wangzefeng@picb.ac.cn (Z.W.); yangwang@dmu.edu.cn (Yang Wang)

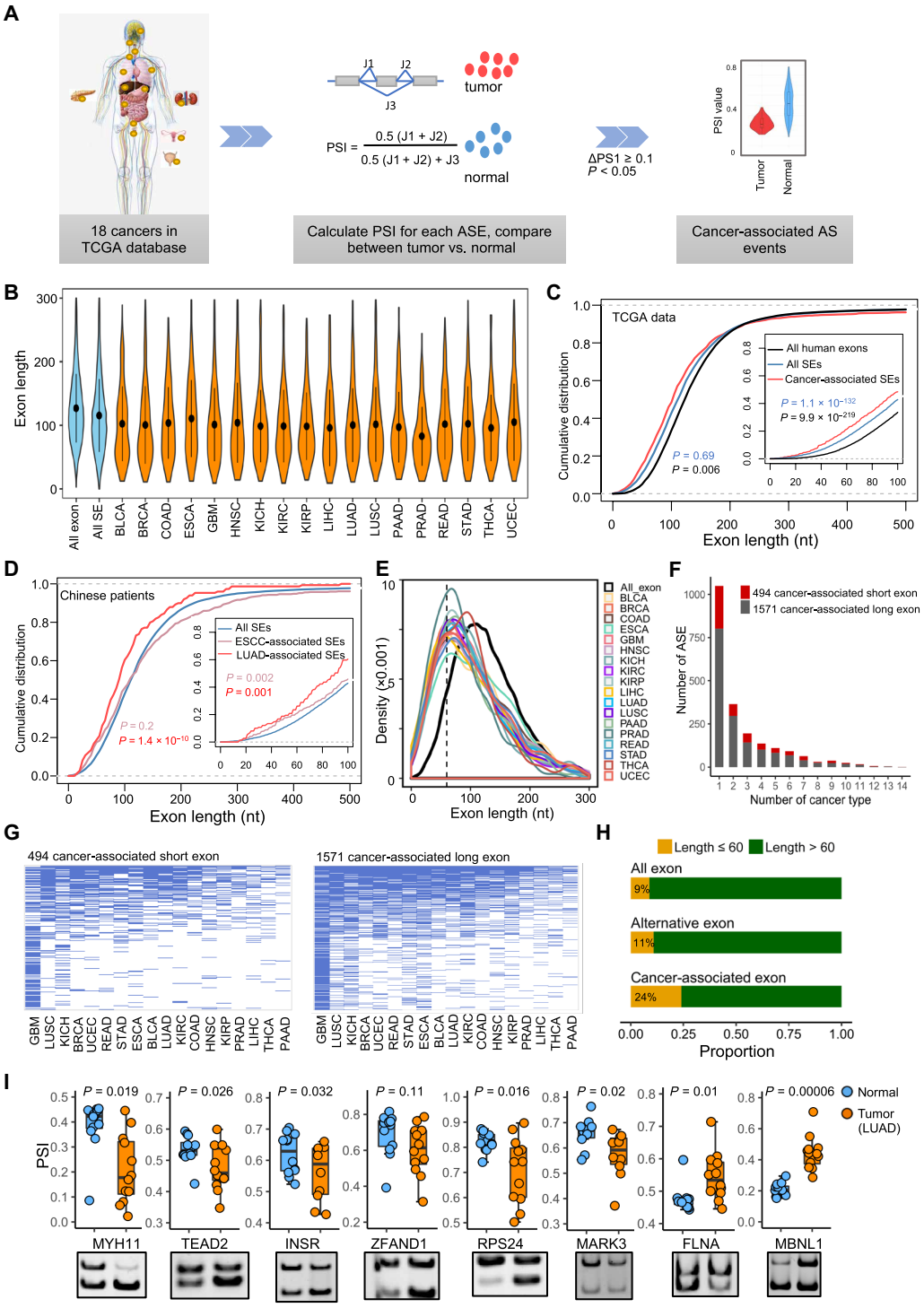

Fig. 1. The length dependency of cancer-associated exons. (**A**) Schematic diagram to identify cancer-associated ASEs. PSI values were calculated for all AS events in 18 types of cancers from TCGA project based on the junction reads. (**B**) Boxplot showing exon lengths of all human exons, all annotated SEs, and the cancer-associated SEs in each cancer type. (**C**) Cumulative distribution curves of all human exons, all annotated SEs, and all cancer-associated SEs in the public data of TCGA. The *P* values were calculated with Student's *t* test by comparing cancer-associated SEs with two other types of exons. The inset shows the distribution of exons no greater than 100 nt in length. (**D**) Cumulative distribution curves of all annotated human SEs and the cancer-associated SEs for the LUAD and ESCC from Chinese patients. The *P* values were calculated with Student's *t* test by comparing all SEs with ESCC-associated SEs or LUAD-associated SEs. The inset shows the distribution of exons with length no greater than 100 nt. (**E**) Length distribution of all human exon and cancer-associated exons in each cancer type. The dotted line represents the 60-nt length cutoff. (**F**) Numbers of CASEs and CALEs identified in different numbers of cancer types. (**G**) Distribution of CASEs and CALEs in each cancer type. (**H**) Proportion of short and long exons in all human exons, all alternative exons, and cancer-associated exons. (**I**) RT-PCR validation of the splicing of several CASEs in LUAD tumor samples and adjacent normal tissues (primer sequences listed in table S1). The experiments were carried out in 12 paired samples, with median and SD plotted above the representative gel.

the 0- to 100-nucleotide (nt) length window, suggesting that the short exons are more likely to be mis-spliced in cancer (fig. S1A).

To independently verify this phenomenon, we reanalyzed the cancer-associated SEs identified by other groups using different computational pipelines on TCGA data (*19*) and found similar length-dependent bias toward short exons (Fig. 1C). Because most patients in TCGA dataset are from Western countries, we also examined the RNA-seq samples collected from Chinese patients of lung adenocarcinoma (LUAD) and esophageal squamous cell carcinoma (ESCC) (*20*, *21*), and again, the similar length-dependent splicing changes were found in both sets of samples (Fig. 1D). Moreover, the cancer-specific SEs identified by long-read sequencing in breast cancer also showed the same length bias (fig. S1B), further confirming the length-dependent splicing changes in cancer (*22*). Together, our results revealed a previously unnoticed feature of splicing alteration in cancer, in which the short exons are more sensitive to be dysregulated regardless of the cancer types, analytic pipelines, and patient populations.

On the basis of the length difference between the cancer-associated SEs and all SEs, we defined the CASEs with a cutoff of 60 nt (equivalent to 20 amino acids) in which the difference of probability densities between cancer-associated exons and all exons reached the maximum (Fig. 1E). Using this cutoff, we identified a total of 494 CASEs (table S2), 269 of which were changed in multiple cancer types (Fig. 1, F and G), with exon 7 of *MBNL1* and exon 6 of *RPS24* being the most frequently changed across multiple cancers (in 14 and 13 cancer types, respectively; fig. S2A). Among all cancer-associated exons, 24% are short exons, whereas only 9% of all human exons and 11% of all alternative exons are short, indicating that short exons are more likely to be mis-spliced in cancers (Fig. 1H). We further validated the splicing of several CASEs using semiquantitative polymerase chain reaction (PCR) in 12 LUAD patients (Fig. 1I and fig. S2B), where most CASEs showed significant splicing changes in tumors compared to the matched normal tissues. Together, our large-scale data analyses identified CASEs as primary targets of splicing dysregulation in human cancers.

CASEs are more conserved and tend to encode in-frame short peptides, with enrichment of certain short regulatory motifs

To further determine the general characteristics of the CASEs, we first calculated the ratio of exons with increased inclusion rate (increased PSI) versus decreased PSI in cancers. Compared with the cancer-associated long exons (CALEs; length > 60 nt), the CASEs displayed a higher probability to be excluded in 15 of 18 tumors analyzed in TCGA database [except PAAD (Pancreatic adenocarcinoma), THCA (Thyroid cancer), and COAD (Colon adenocarcinoma)] (Fig. 2A) and in both cohorts of Chinese LUAD (Lung adenocarcinoma) and ESCC (Esophageal cancer) patients (Fig. 2A, in asterisks).

We then sought to further identify the specific features of the CASE-containing pre-mRNAs. We classified all alternative exons into four categories—CASEs, other short exons, CALEs, and other long exons (Fig. 2B, left)—and conducted a series of detailed comparisons. In general, the PSI values of all cancer-associated exons are lower than other exons, especially for CASEs (Fig. 2B, right). Compared to the other exons, the upstream and downstream introns adjacent to the cancer-associated exons were significantly longer, with the introns adjacent to CASEs even slightly longer than those adjacent to CALEs (Fig. 2C). Moreover, the splice sites of cancer-associated exons are significantly weaker than the other exons (Fig. 2D), which may partially underlie the increased exclusion of CASEs in cancers. The CASEs and their adjacent intronic sequences were significantly more conserved compared to other three types of exons (Fig. 2E), suggesting additional selective pressure for splicing regulatory elements.

To explore the regulation of CASE splicing, we used a statistical enrichment analysis (*23*–*25*) to identify short motifs enriched in the CASEs or adjacent introns that potentially function as regulatory cis-elements to control CASE splicing. This analysis showed that the CASEs are enriched with exonic AG-rich motifs (Fig. 2F), which are known to function as ESEs (Exonic splicing enhancers) by recruiting SR (Serine and arginine-rich) proteins that promote splicing of weak exons (*23*, *26*, *27*). The adjacent introns of CASEs were also enriched with many pyrimidine-rich elements that may bind to PTB proteins to generally inhibit splicing in many genes (*28*, *29*). A (T)GCATG motif, which has been reported to bind RBFox proteins (*30*–*33*), was strongly enriched at the downstream introns of CASEs (Fig. 2F). Enrichment of the intronic (T)GCATG motif was also identified by comparing CASEs versus CALEs and CASEs versus other short exons but not identified in CALEs versus other exons (Fig. 2F and fig. S2C), suggesting that RBFox proteins may play a previously unknown role in specifically regulating short exons in cancers. This motif is unlikely a general feature for all short exons, as it could not be identified by comparing short exons versus long exons (fig. S2C).

Another interesting feature of CASEs is that, compared to the other types of exons, CASEs are substantially enriched for exons that maintain the original reading frame (i.e., exon length is the multiples of 3 nt) (Fig. 2G). Therefore, the mis-splicing of CASEs is more likely to change the ratios of alternative protein isoforms containing or lacking an optional short peptide rather than to produce a truncated protein or cause the nonsense-mediated decay (*34*). Further analyses showed that the CASE-encoded peptides are enriched for hydrophilic residues like Lys, Ser, and Arg but depleted for hydrophobic residues like Leu, Ile, and Val (fig. S2D). Moreover, the CASE-encoded peptides were significantly enriched for intrinsic disordered regions as predicted by IUPred (Fig. 2H) (*35*). Because the intrinsic disordered regions often mediate protein-protein interactions during phase separation (*36*, *37*), the splicing of CASEs may affect important cell signaling pathways involving phase separation. In support of this notion, gene ontology (GO) analysis revealed that the CASE-containing genes are significantly enriched for functions related to guanosine triphosphatase (GTPase) regulator activity and cell adhesion (Fig. 2I and fig. S2E), implying that the CASE splicing may affect the cell proliferation and migration during cancer progression and metastasis.

CASEs can serve as diagnostic molecular markers of cancers

Because the alternative inclusion of short exons is easier to measure and the results are more reliable compared to the other alternative exons, we next seek to explore the predictive power of CASEs as potential molecular markers of cancers. Using the PSI values of CASEs as inputs, we applied two different machine learning methods [the principal components analysis (PCA) and the partial least squares discrimination analysis (PLS-DA)] to make a clear distinction between the tumors versus adjacent normal tissues (Fig. 3A and fig. S3). These separations were observed in either grouping all cancer types (Fig. 3A) or using each individual cancer type (fig. S3),

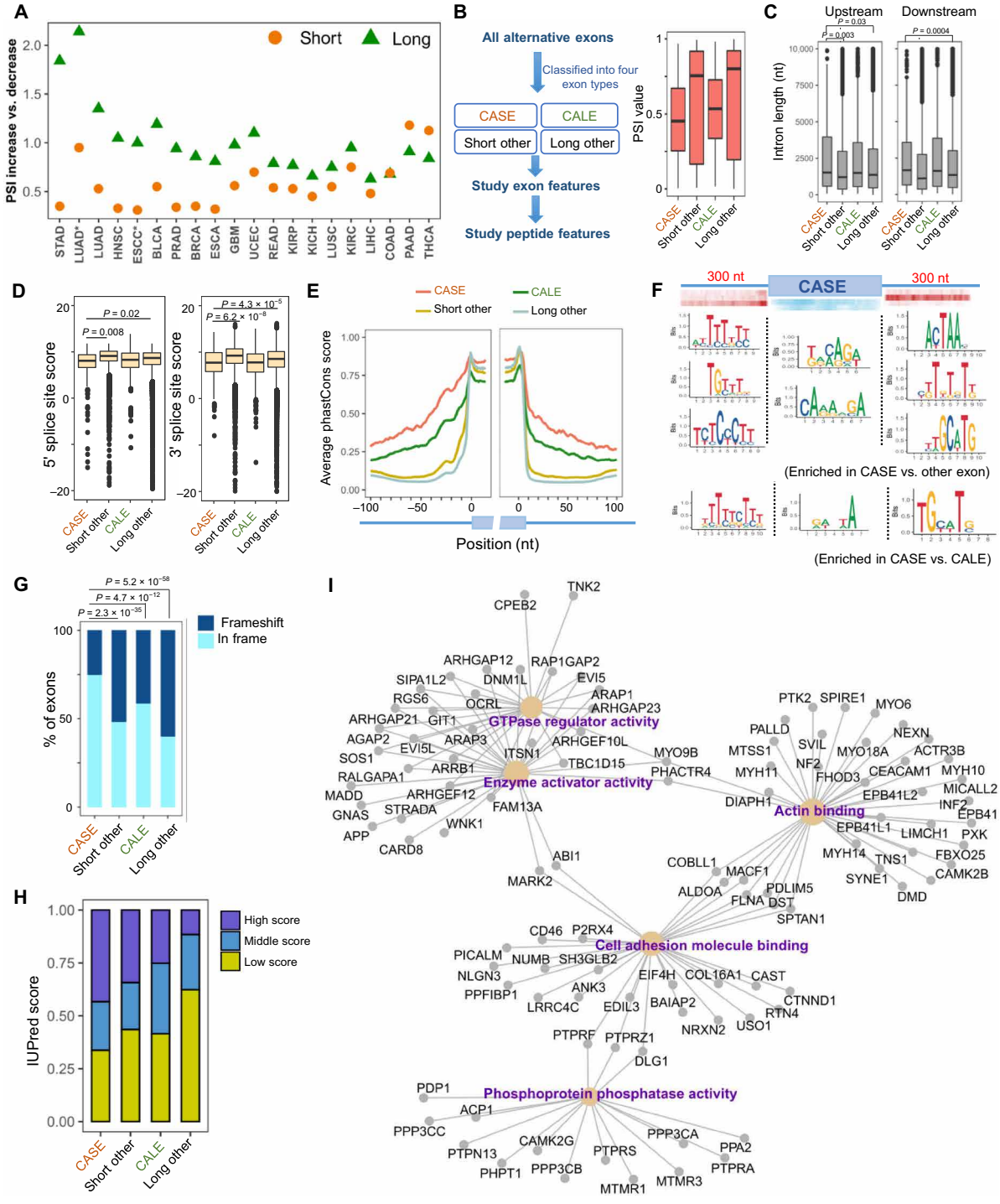

Fig. 2. Sequence features and functional implication for CASEs. (**A**) Changes of PSI for cancer-associated exons within different length groups. In each cancer type, the ratio of the SEs with increased PSI versus decreased PSI between the tumors and normal tissues was plotted. The short and long exons (in 60-nt cutoff) were plotted differently. The asterisks indicate data from Chinese patients, while the rests are from TCGA data. (**B**) Workflow to classify all human exons into four categories according to the exon length and the association with cancers (left), and distribution of PSI values for each type of exons (right). (**C**) Length of the upstream and downstream introns surrounding four types of exons. The length of CASEs (i.e., short cancer) was compared to the other three types of exons, and *P* values were calculated by Student's *t* test. (**D**) Splice site strength of four types of exons. *P* values between CASEs and other groups were calculated by Student's *t* test, with the significant differences ($P < 0.05$) being indicated. (**E**) Conservation of the sequences around the four types of exons as indicated by average phastCons scores. (**F**) Enriched motifs in CASEs and their adjacent introns presented by pictograph. For each motif, the heatmaps of the enriched site were also included at above. (**G**) The effects on the reading frame by exon inclusion/exclusion for different exon types were plotted. *P* values were calculated with Fisher's exact test. (**H**) The fractions of peptides encoded by each type of exons with a high (>0.67), mid-range (0.33 to 0.67), and low (<0.33) disorder rate were plotted using IUPred scores. (**I**) GO analysis for the CASE-containing genes, with the significantly enriched functional terms as the hubs of the gene network.

suggesting a strong predictive power of CASEs. Furthermore, we developed a random forest model to directly predict cancers using the PSI of CASEs (Fig. 3B and Materials and Methods). With a fourfold cross-validation in 100 randomly trails, the model achieved accurate prediction with the area under the curve (AUC) of 0.913 (Fig. 3C, left). This molecular diagnosis model also performed reasonably well for individual cancer type, except in pancreatic adenocarcinoma where the sample size was too small (Fig. 3C, right).

To further validate the CASE-based model, we used independent datasets from independent patient populations in the training and testing stages. Specifically, we trained the model using the LUAD datasets from TCGA that contains mostly patients of Western descendants or the ESCC dataset from Chinese patients, and tested the results with the Chinese LUAD patients or the ESCA dataset from TCGA. Our CASE-based random forest model achieved an AUC of 0.881 and 0.875, respectively (Fig. 3D), indicating that the CASEs may serve as reliable markers for molecular diagnosis of cancers.

We next examined the relative contribution of each CASE to cancer prediction by removing individual CASE from the model and calculated the reduction of accuracy. Using two independent algorithms, we identified top 10 CASEs that contributed the most to the prediction accuracy (Fig. 3E). Both algorithms identified the same set of top four CASEs, including exon 30 of *FLNA*, exon 16 of *MARK3*, exon 7 of *MBNL1*, and exon 6 of *RPS24*. The splicing of these four exons was all closely associated with patient survival

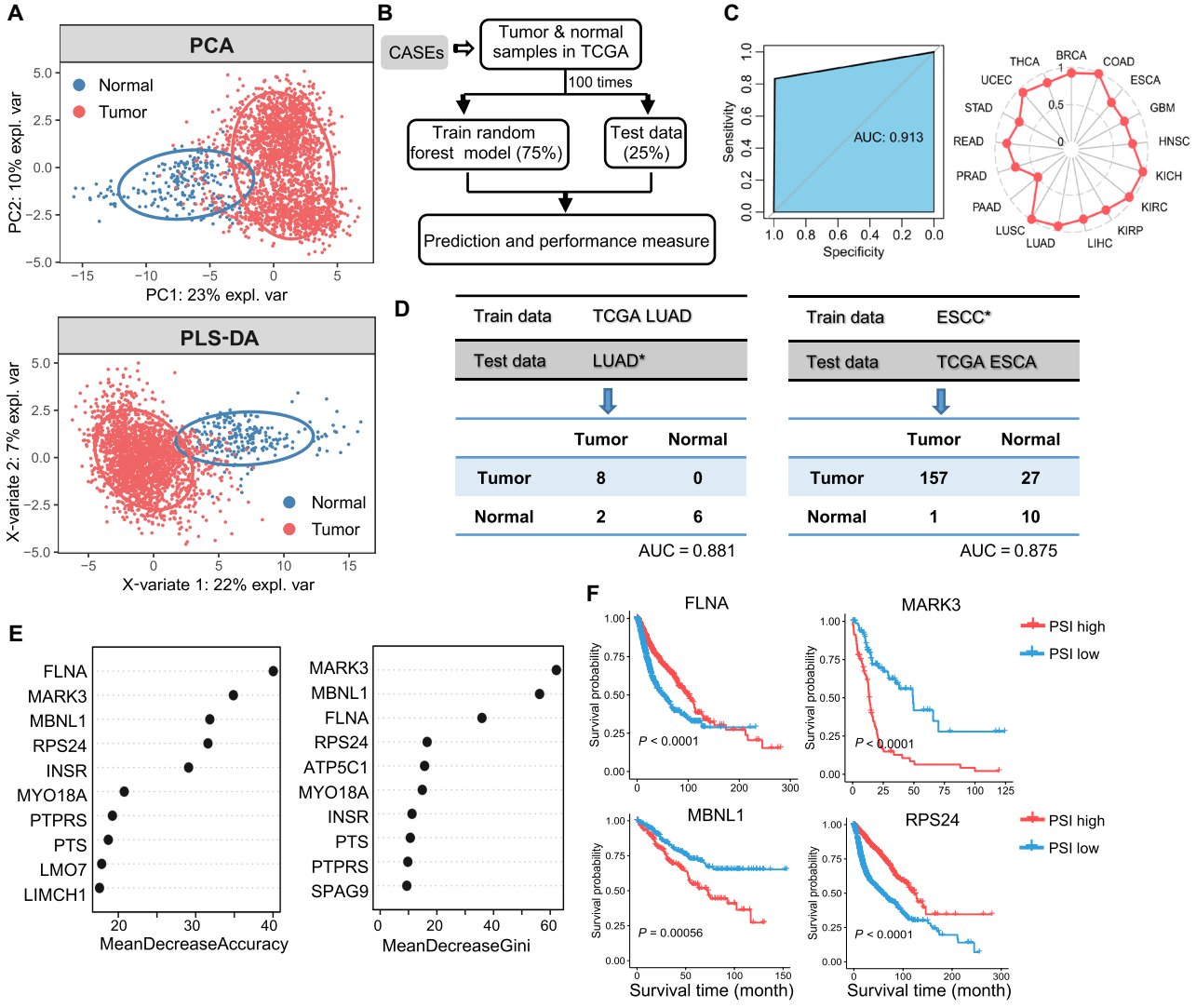

Fig. 3. Cancer prediction with CASEs. (**A**) PCA (top) and PLS-DA (bottom) analyses to separate cancers versus normal tissues based on PSI values of CASEs. The samples from all cancer types were merged for the analyses. (**B**) Training a random forest model to predict cancer using PSI values of CASEs as the features. A fourfold cross-validation in 100 randomly trails was used, where 75% of the samples were randomly selected for training and the remaining 25% for testing. (**C**) ROC curve of the random forest model (left) and AUC of the random forest model in each cancer type (right). (**D**) Performance of two additional random forest models using TCGA LUAD data as the training set and Chinese LUAD data as the testing set (left) or using Chinese ESCC data for training and TCGA ESCA data for testing (right). (**E**) Variable importance plot of the top CASEs with most impacts on the random forest model in (B). (**F**) Kaplan-Meier survival curves of all cancer patients stratified by PSI values of the top four CASEs in (E).

(Fig. 3F), raising the possibility of using CASEs as new prognostic indicators for survival of cancer patients.

Splicing of CASEs serves as a strong predictor for cancer survival

To further evaluate the potential of CASEs as predictive factors of patient survival, we combined the patients from all cancer types and performed a series of Kaplan-Meier analyses by grouping patients based on the splicing of each CASE. We found that splicing of 24% of CASEs is significantly correlated with survival with $P < 0.01$ in Cox regression (table S3), and that of 16% of CASEs (79 CASEs) is strongly associated with patient survival with $P < 0.0001$ (Fig. 4A). In comparison, a significantly smaller fraction of CALEs were associated with cancer prognosis ($P < 0.05$ by Fisher's exact test). Although both types of cancer-associated splicing events have some prognostic values, measuring the splicing of short exons is easier and more reliable, and thus, the CASEs may be a more practical choice than the CALEs as the prognosis markers. The genes containing the prognostic CASEs are functionally enriched for cell signaling pathways such as endocytosis and tight junction (Fig. 4B).

We further plotted the hazard ratios of the CASEs with the highest predictive power (Fig. 4C), among which the top four were also presented in Kaplan-Meier curves (Fig. 4D). Our results suggested that these CASEs can serve as individual prognostic marker. A major limitation in predicting patient survival of all types of cancers with an individual CASE is the "missing data problem," in which only a small number of patients can be reliably grouped and predicted using a given CASE. This problem is common to most prognostic predictions using a small set of molecular markers. To deal with this problem, we combined eight CASEs (top CASEs in Figs. 3E and 4D) into a predictive panel and defined a risk factor based on the relative PSI of these CASEs (Fig. 4E and Materials and Methods). Using this panel, it is possible to make a prediction for most patients, except for those with missing data in all the eight CASEs. We first defined this CASE-based risk factor in each individual cancer type and found an accurate prediction of patient survival in all cancer types tested (fig. S4), suggesting that CASEs may be useful as prognostic markers. Encouraged by this result, we further tested whether the CASE-based factor is robust enough to show a predictive power even when all different cancer types are combined together, which was very difficult because of the large variations between different cancer types (Fig. 4F). We grouped the patients of all cancer types according to the risk factor defined by the CASE panel and found a significant separation of the survival time for different patient groups (Fig. 4F), indicating that the CASE panel can serve as a strong pan-cancer prognostic predictor. Moreover, we also applied the CASE-based risk factor on an independent cohort of ESCC from Chinese patients and confirmed that the patients grouped according to CASE splicing have significantly different survival times (Fig. 4G). Collectively, these results indicated that CASEs could serve as a strong predictor for cancer survival.

To explore the molecular features responsible for the differential survival in cancer patients grouped by the CASE-based risk factor, we used CIBERSORT (*38*) to determine the immune cell infiltration in each cancer sample (Fig. 4H). We found that the patient group with better prognostic outcome (group 1) showed a significant reduction in the population of naïve and plasma B cells but had increased numbers of memory B cells. In addition, the T helper cells that promote B cell activation were relatively enriched in these patients, and regulatory T cells (T_{regs}) that suppress adaptive immunity are reduced (Fig. 4H). The group 1 patients also showed a reduced ratio of activated mast cells versus resting mast cells, which was found to inhibit T_{regs} (*39*). On the other hand, although the group 1 patients have more $CD4^+$ T memory cells that are responsible for long-term immune memory, the activation of T memory cells was not significant between these two groups (Fig. 4H, right). Collectively, these results showed an intriguing correlation between CASE splicing and cancer immune microenvironment, which may shed light on the functional implication of CASE splicing. It should be noted that such difference in tumor immune microenvironment is quite small and cannot be translated into an effective prognostic factor; however, it probably reflects a previously unknown clinical feature in patients with significant splicing alterations.

Splicing of CASEs is affected by elevated transcription in cancer

We further seek to determine the molecular mechanisms responsible for the mis-splicing of CASEs in cancer. Previous studies reported that most genes are spliced cotranscriptionally (*40*, *41*), and thus, the transcription rate may affect splicing by changing its time window (*42*, *43*). It is generally accepted that cancer cells have an elevated transcription due to their rapid proliferation, especially in cancers with MYC mutations (*44*). According to a simple kinetic model, fast transcription reduces the time window for splice site recognition and thus may promote skipping of short exons (*42*, *43*, *45*). Therefore, we speculate that the elevated transcription in cancer may contribute to CASE splicing.

To test this hypothesis, we analyzed the RNA-seq data from cells harboring the mutated RNA polymerase II (Pol II) with fast and slow transcription rate (Fig. 5A) (*46*). The average elongation rate of wild-type (WT) Pol II is 1.7 kb/min, whereas the R749H mutation reduced the rate to ~0.5 kb/min and the E1126G mutation increased the rate to ~1.9 kb/min (*46*). Using the RNA-seq data from cells with WT or mutated Pol II, we identified 5439 and 3041 SEs significantly affected by these two mutants, respectively (Fig. 5B). The splicing changes of SEs are highly correlated between fast and slow mutations (fig. S5A), suggesting that altered transcription elongation rates have a profound effect on alternative exon inclusion, which is consistent with a previous report (*46*). The short exons were more likely to be excluded (i.e., PSI decreased) than the long exons by altered elongation rates (Fig. 5C and fig. S5B), suggesting that the short exons are especially vulnerable to the transcriptional rate changes. Because the transcription is generally elevated in cancers, such increased sensitivity to transcription elongation might explain why the short exons are more easily to be skipped in cancer cells. Consistently, there is a significant overlap between CASEs and the short exons altered by fast Pol II mutation, suggesting that the CASE splicing is easily disrupted by increased transcription elongation (Fig. 5D). These results are consistent with a simple kinetic model, where the fast transcription reduces time windows for splice site recognition and thus may promote skipping of short exons (*47*).

To further test whether the elevated transcription in cancer contributes to CASE splicing, we examined the relationship between the transcription and CASE splicing by reanalyzing the chromatin immunoprecipitation–sequencing (ChIP-seq) data from multiple cancer cells (MCF-7, K562, HepG2, and A549) in ENCODE to examine different histone markers that reflect the transcription rates

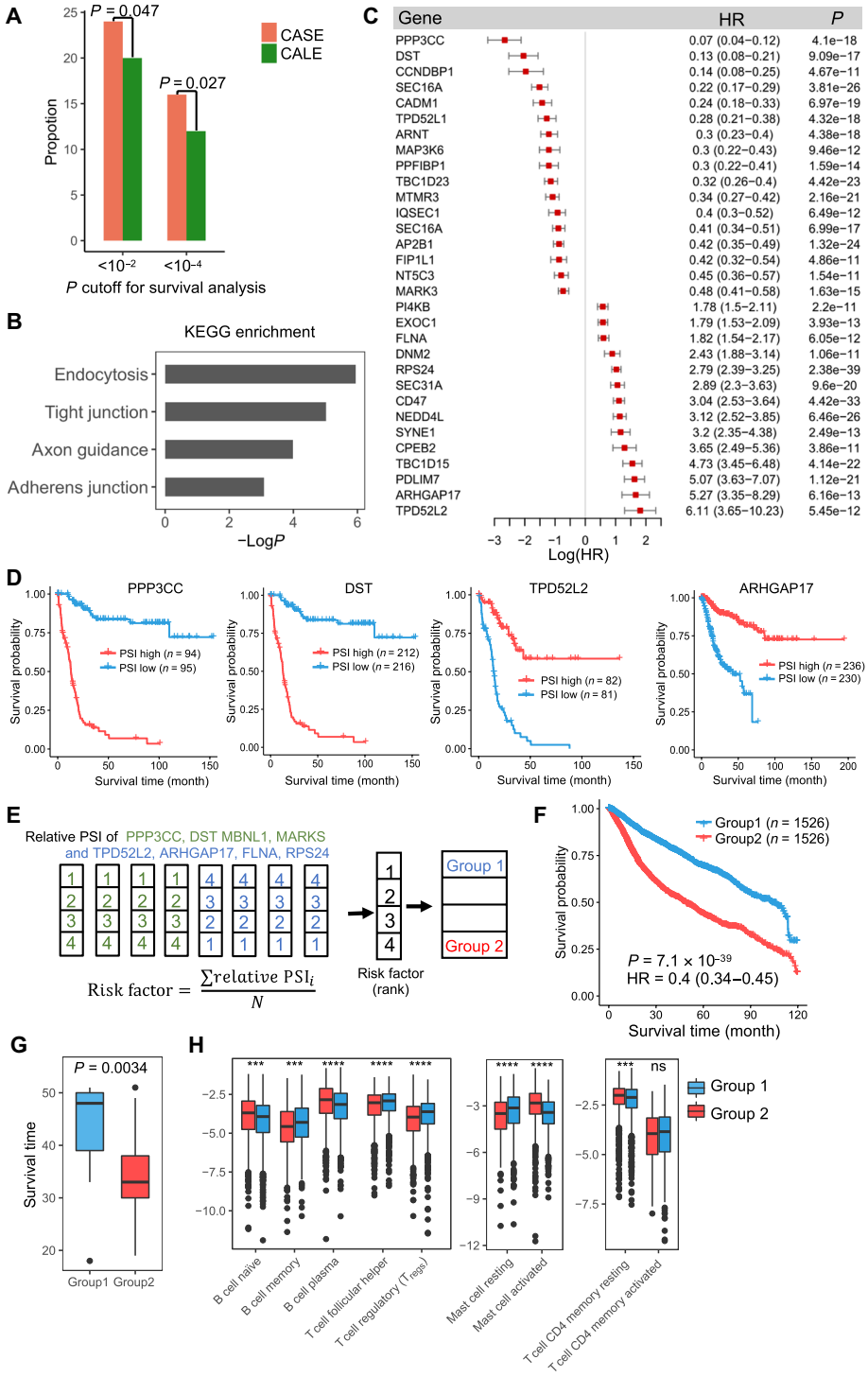

Fig. 4. Prediction of patient survival using splicing changes of CASEs. (**A**) Proportion of CASEs and CALEs that are significantly correlated with patient survival at various cutoffs (*P* values from Cox regression). The difference between the proportion of CASE and CALE was evaluated with Fisher's exact test. (**B**) KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment of the genes contained CASEs that correlated with patient survival with $P < 0.01$. (**C**) Forest plot of top CASEs significantly correlated with patient survival in all cancers ($P < 10^{-10}$, Cox regression). HR, hazard ratio. (**D**) Kaplan-Meier curves of all cancer patients stratified by PSI values of four individual CASEs with the highest predictive power in (C). The top and bottom quartiles of patients ranked by the PSI values of each CASE are grouped as PSI high and PSI low. (**E**) Pipeline to calculate the risk factor based on the PSI values of eight CASEs (see Materials and Methods). The patients with the top and bottom quartiles ranked by the risk factor were defined as group 1 and group 2. (**F**) Kaplan-Meier curve of patients grouped by the CASE-based risk factor. The *P* value was calculated by Cox regression. (**G**) Survival time distribution of two groups of patients, which was grouped using the method in (D) in Chinese ESCC cohort. The *P* value was calculated by Student's *t* test. (**H**) Immune cell infiltration defined by the relative fractions of different types of immune cells estimated using CIBERSORT in two groups of cancer patients (***$P < 0.001$ and ****$P < 0.0001$; ns, not significant, by Student's *t* test).

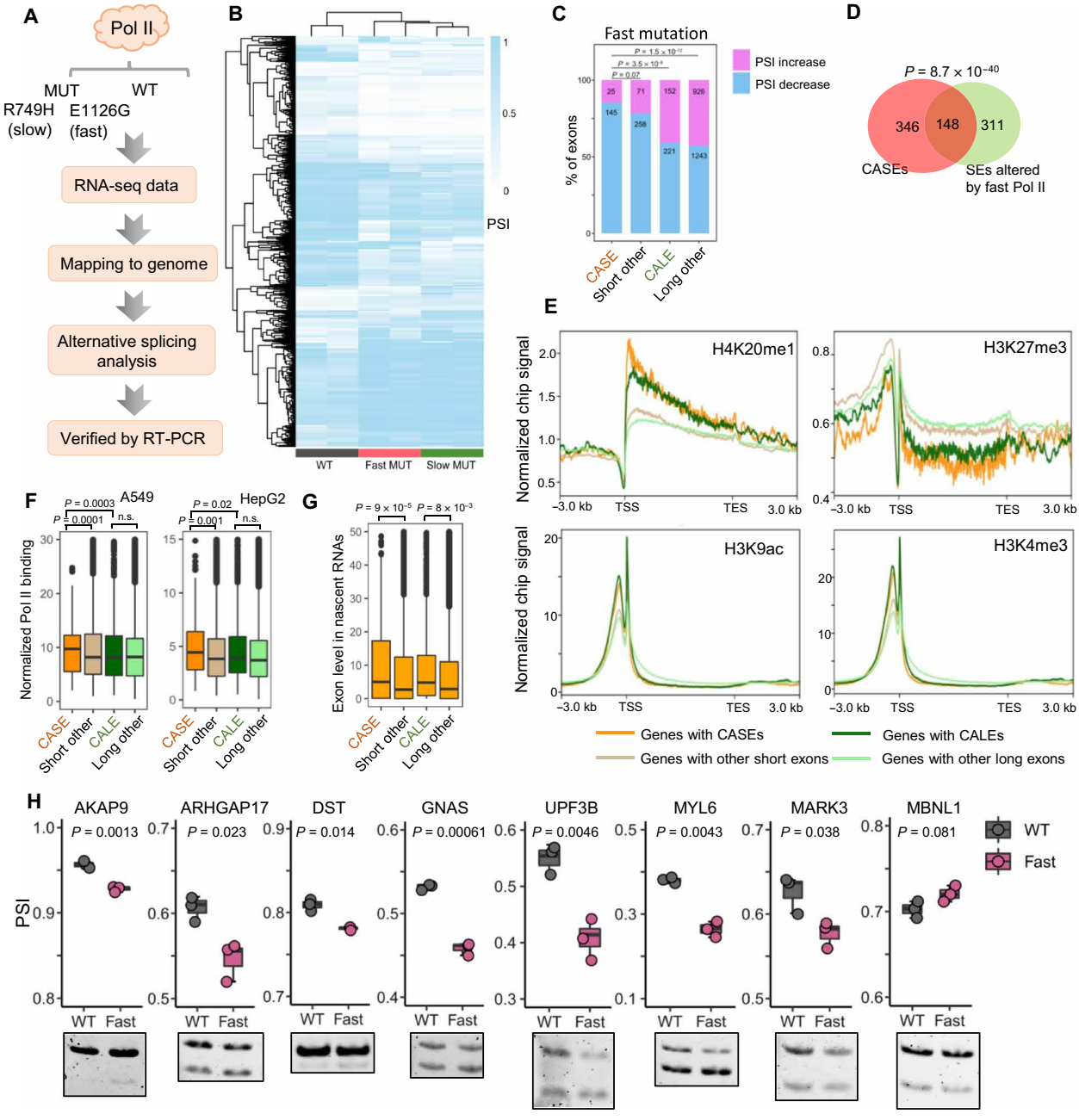

Fig. 5. Elevated transcription profoundly affects CASE splicing. (**A**) Computational pipeline to identify AS events affected by the changes of fast and slow Pol II mutation. (**B**) Heatmap and unsupervised cluster of SEs affected by the fast and slow Pol II mutation. (**C**) Number of exons with PSI increase or decrease in cells with fast Pol II mutations. The exons were divided into four types according to Fig. 2B. (**D**) Venn diagram showing the overlap between CASEs and short exons affected by fast Pol II mutation. The *P* value was calculated by Fisher's exact test. (**E**) Distribution of normalized ChIP signals for different histone modifications in genes containing four types of exons. The region from 3 kb upstream of TSS to 3 kb downstream of TES was shown. (**F**) Normalized Pol II distribution signals surrounding four different types of exons. Two different cancer cell lines (A549 and HepG2) were analyzed. (**G**) Normalized expression of four types of exons in nascent RNAs. (**H**) The splice changes of selected CASEs were measured by RT-PCR using gene-specific primers (sequences listed in table S1). The experiments were carried out in triplicates, with mean ± SD plotted above a representative gel.

(Fig. 5E and fig. S5B). We found that the CASE-containing genes had an increased level of monomethylation on H4K20 (H4K20me1) (Fig. 5E and fig. S5C), which marks an increased transcriptional elongation rate in these genes (*48*). Conversely, the transcription repression marker H3K27me3 (*49*, *50*) was significantly reduced in the genes containing cancer-associated exons (especially CASEs), again suggesting that the genes containing CASEs are transcribed in an elevated rate. Moreover, the histone modifications associated with increased transcription initiation (i.e., markers of active promotor), such as H3K4me3 and H3K9ac, also showed stronger signals in genes containing either

CASEs or CALEs across multiple cancer cells (Fig. 5E and fig. S5C). This result implied that, in addition to fast elongation, a higher transcription activity in cancer may affect CASE splicing in general.

To directly measure the effect of elevated transcription on CASE splicing, we analyzed the Pol II occupancy around different types of exons (see Materials and Methods). As expected, we found significantly more Pol II signals around the CASEs than other types of exons (Fig. 5F). Consistently, the analysis of TT-seq (transient transcriptome–sequencing) data (*51*) showed that the cancer-associated exons (CASE and CALE) have high levels of nascent RNAs (Fig. 5G), implying that the elevated transcription may cause mis-splicing of CASEs.

To experimentally validate the effect of altered transcription on CASE splicing, we generated Flp-In T-REx 293 cells stably integrated with WT or E1126G Pol II mutant with increased transcription rate (WT versus fast Pol II). We induced the expression of WT and E1126G Pol II mutant with doxycycline and inhibited the endogenous Pol II with α-amanitin. Subsequently, the splicing of CASEs was examined with semiquantitative reverse transcription PCR (RT-PCR) (Fig. 5H and fig. S5D). We found that cells with fast Pol II showed significant reduction of CASE inclusion in most CASEs tested (except exon 7 of *MBNL1*, which also has increased PSI in most cancers) (Fig. 5H and fig. S5E). This result directly supported our conclusion from computational analyses and further confirmed that the transcription rate plays a key role in regulating CASE splicing.

Identify the RNA binding proteins that regulate CASE splicing

The above results suggested that short exons are preferably affected by elevated transcription; however, it is still unclear why a certain fraction of short exons (i.e., CASEs) are mis-regulated by the increased transcription in cancer. In addition to the transcription rate, AS can generally be regulated by various trans-acting factors that specifically bind pre-mRNAs (*4*), and thus, these factors probably also contribute to the mis-splicing of specific short exons in cancers. To identify RNA binding proteins (RBPs) that potentially regulate CASE splicing, we analyzed the existing large-scale RNA-seq datasets from ENCODE consortium with knockdown of 227 RBPs. For each RBP, we identified SEs that are significantly altered upon RBP knockdown and focused on the CASEs among these identified targets (Fig. 6A).

We performed an unsupervised clustering of all CASEs according to how they were affected by RBPs and found that many CASEs were tightly correlated with each other (fig. S6A), suggesting that they are probably regulated by the same set of factors. To explore the common factors that regulate CASEs, we grouped the CASEs into three major clusters according to their splicing patterns across all cancer samples in TCGA (fig. S6B). The CASEs within the same cluster were tightly correlated in their splicing profiles (i.e., synergistically regulated) and thus were merged together for further study. We next examined the potential interactions between each RBP and its potential target RNAs using the eCLIP-seq data. For the candidate RBPs, we identified their binding sites within the cognate exons and the adjacent introns (schematic diagram in Fig. 6A). By combining all evidences from RNA interference sequencing (RNAi-seq) and eCLIP-seq, we identified RBPs that may directly regulate each cluster of CASEs, resulting in a tightly connected regulation network between RBPs and CASEs (Fig. 6B).

We next focused on the 18 RBPs that regulate all three clusters of CASEs and examined how they might affect CASE splicing in cancers. By conducting a mutation analysis of the 18 RBPs using TCGA dataset, we found that, in general, these RBP genes are not frequently mutated in cancers. In total, there are ~13% of patients with detected mutations in at least one of these genes, and the mutational status of most RBPs was not significantly correlated with CASE splicing (fig. S7A). The only exception is the core spliceosomal gene SF3B1 that is one of the most frequently mutated splicing factors (mutated in 3% of patients from the 18 cancer types analyzed). The cancers with SF3B1 mutations have significant increase of CASE skipping (fig. S7B), suggesting a minor role of RBP mutations in affecting CASE splicing. On the other hand, compared to the control RBPs, the expression of 9 of the 18 identified RBP genes is significantly more correlated with CASE splicing in TCGA samples (fig. S7C), suggesting that the altered expression of these RBPs in cancer plays a more general role in control CASE splicing.

We further focused on several RBPs that affect splicing in a length-dependent fashion (fig. S8, A to D). One of the RBPs, RNA binding fox-1 homolog 2 (RBFox2), is known to bind the (T)GCATG motif that is enriched in the downstream introns of CASEs (Fig. 2F). Knockdown of RBFox2 showed different impacts on splicing of short and long exons, and the RBFox2-sensitive short exons were significantly overlapped with CASEs (fig. S8A). RBFox2 was reported to generally promote exon inclusion when binding to the downstream introns (*52*, *53*), and the expression of RBFox2 was generally decreased in most cancers (fig. S8A), suggesting that the decrease of RBFox2 in cancers may be responsible for the skipping of many CASEs. Consistently, the CASEs with downstream (T)GCATG motif are more correlated with the RBFox2 levels than the CASEs without this motif, supporting the activity of RBFox2 to promote CASE inclusion by binding to the downstream (T)GCATG motif. As a control, in the correlation analysis using all SEs, the difference between SEs with or without this motif is very small (Fig. 6C).

In addition to RBFox2, many core components of the spliceosome, including AQR and U2AF2, were found to affect multiple clusters of CASEs. Specifically, Aquarius (AQR), a helicase-like protein that binds introns in a sequence-independent fashion (*54*), also regulates splicing in a length-dependent fashion and the AQR-regulated exons are significantly overlapped with CASEs (fig. S8B). PTBP1 and U2AF2, both preferentially binding to pyrimidine-rich sequences that are enriched around the CASEs (Fig. 2F), also affect short and long exons differentially with a significant overlap between CASEs and their targets (fig. S8, C and D). We also found that the SEs regulated by these RBPs generally have weaker splice sites (fig. S8E), which is similar to splice sites of CASEs (Fig. 2D). Because weaker exons are generally shorter, this result also suggested that the factors regulating weak exons may also play a role in regulating CASE splicing.

Last, we experimentally validated the regulation of several CASEs by these RBPs (AQR, PTBP1, RBFox2, and U2AF2) using short hairpin RNA (shRNA) knockdowns (fig. S9) and found that they played key roles in affecting CASE splicing (Fig. 6D). Together, our findings demonstrate that certain RBPs are actively involved in regulating CASE splicing in cancers, among which the activity of RBFox2, AQR, PTBP1, and U2AF2 was experimentally validated.

DISCUSSION

Splicing dysregulation in cancers has been widely regarded as a key molecular feature that plays critical roles in cancer biogenesis and progression. By systematically analyzing the big data of transcriptome

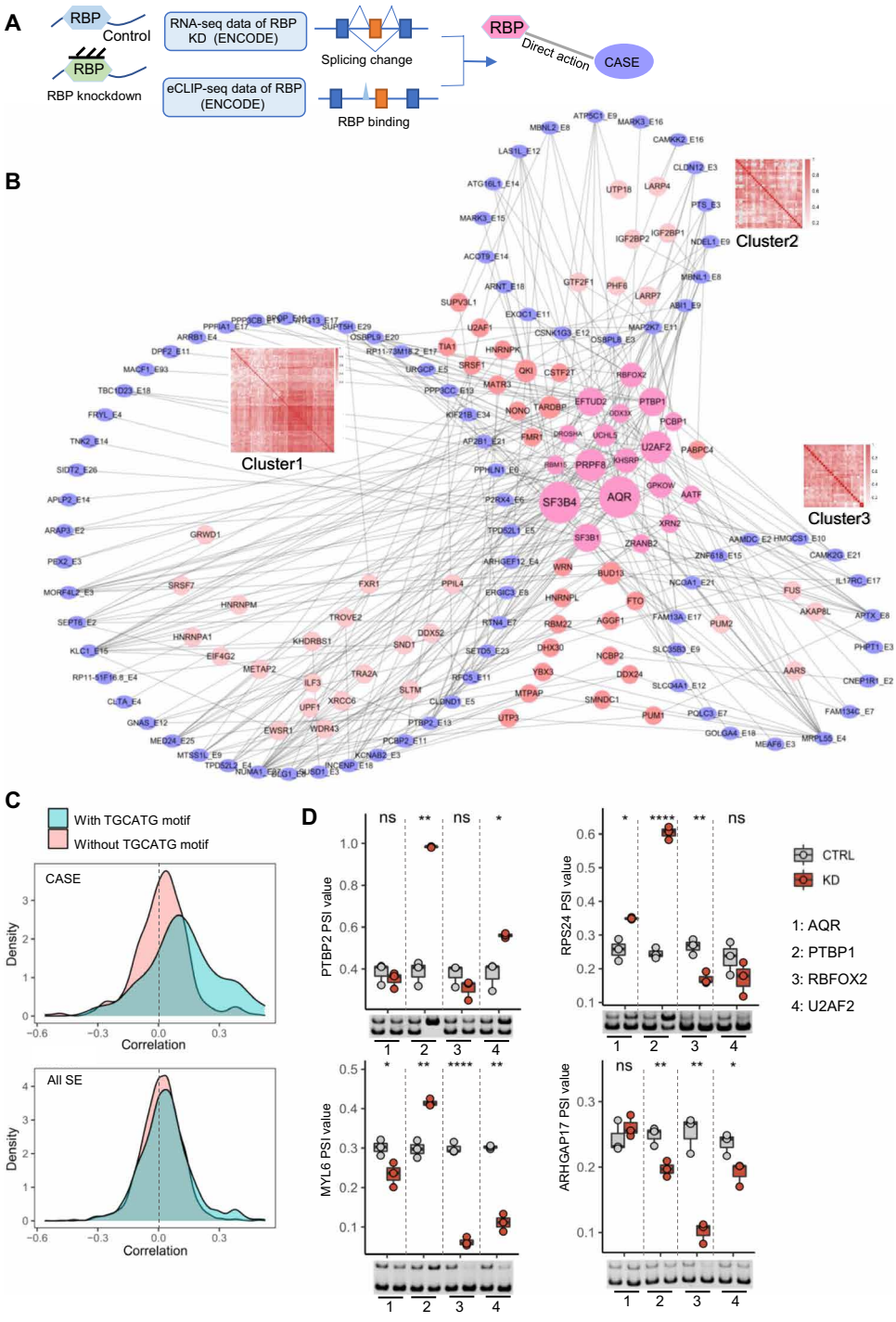

Fig. 6. Identifying RBPs that regulate CASE splicing. (**A**) Workflow to identify RBPs that directly regulate CASE splicing. The RNA-seq and eCLIP-seq data were downloaded from ENCODE, and a direct regulation was defined if the knockdown of a certain RBP affects the splicing of a specific CASE and at least one binding site of the RBP was found around that CASE. (**B**) RBP-CASE interaction network for three major clusters of CASEs. The purple circles present CASEs, and the pink circles present RBPs. The size of each circle of RBP represents the number of CASEs it regulates. (**C**) Correlation between the inclusion level of CASEs (top) and all SEs (bottom) with or without TGCATG motif in the downstream intron and the expression of RBFox2 in all cancers. (**D**) Test of the effect of several RBPs on CASE splicing. The splice changes of selected CASEs in the condition of RBP knockdown were measured by RT-PCR using gene-specific primers (*$P < 0.05$, **$P < 0.01$, and ****$P < 0.0001$; ns, not significant, by Student's *t* test in all panels).

sequencing, we uncovered a general trend of length-dependent splicing dysregulation, where the short exons are more likely to be mis-spliced and preferably skipped in most cancers. These CASEs are more conserved and tend to preserve reading frames, and the CASE-encoded peptides are more likely to contain intrinsic disordered regions that are known to affect protein phase separation, cellular signal transduction, and RNA metabolisms (*55*, *56*). Previously, a group of microexons (3 to 15 nt) was found to be specifically included in neuronal cells (*57*), and our finding of CASEs represents a more general type of length-dependent splicing regulation that is common to most tissues. Some examples of the CASE have been independently reported to have important functions in cancer pathogenesis (*20*, *58*), suggesting that this length-dependent splicing dysregulation has functional relevance in cancers. Here, we explored the clinical application of CASEs by developing a new computational pipeline to successfully predict cancer survival and also identified two potential mechanisms that affect the length-dependent splicing in cancer. Collectively, our study not only found a common feature of cancer-associated splicing dysregulation but also highlighted the clinical importance of these CASEs as prognostic markers and/or therapeutic targets.

Our findings have several important clinical implications. The powerful machine learning model we developed on the basis of CASE splicing is able to achieve >90% accuracy in predicting cancers versus normal tissues, and the CASE-based risk factor can serve as a strong prognostic predictor of cancer survival (Figs. 3 and 4). Compared to other methods that required large amount of RNA-seq data, our model only needs the splicing readouts from dozens of short exons, which are feasible and reliable in clinical measurement. By combining eight CASEs into a predictive panel, we were able to generate a predictive score for 90% of cancer patients. The large variations of survival and treatment scheme for different cancers usually make it very difficult for the prognostic prediction using a simple molecular marker. However, we found that the CASE-based score is very robust and can even generate statistically meaningful survival prediction in either individual cancer type or combined cancers. The altered splicing of an individual CASE may not make enough functional difference to affect the outcome of cancer, and thus, it is quite unexpected that a single CASE showed a strong prediction power. We speculate that the overall impact of the coordinated splicing changes in many CASE-containing genes may functionally affect the cancer survival. Consistently, we find that many CASE-containing genes are enriched for cancer-related functions (Fig. 2I). In addition, it is also possible that the preferential skipping of short exons is a result of elevated transcription rate, and thus, the aberrant splicing of a single CASE can serve as a good molecular marker by reflecting the change of transcription rate. It is likely that both the mis-splicing of CASE-containing genes and the elevation of general transcription probably affect cancer cell growth in a nonexclusive fashion.

Moreover, we found that the different CASE-defined cancer groups have distinct profiles of immune cell composition, implying that these groups may need to be treated with different regimens of chemotherapy or immunotherapy. It is worth mentioning that this results on immune infiltrates only provided a suggestive clue rather than the definitive evidence. The association of CASE-defined cancer groups with distinct immune cell infiltrates only presents a possible explanation on why the two groups showed different chances of survival, as well as some clues to guide immunotherapy. However, future tests are needed to confirm this hypothesis on precision immunotherapy.

The finding that CASEs are sensitive to altered transcription rate in cancers is very intriguing. Because most pre-mRNAs are spliced cotranscriptionally, AS can generally be affected by the transcription rate (*40*, *41*). According to a simple kinetic model, fast transcription reduces the competitive advantage of upstream splice sites and thus promotes skipping of weak exons, while a slow elongation usually promotes exon inclusion by providing a larger time window for spliceosome assembly (*42*, *43*, *45*). Therefore, an optimal transcriptional elongation rate may be required for the proper splicing of different genes (*46*). In this study, we found that the splicing of short exons is more sensitive to the transcription changes. We speculate that the change of transcription rate affects the time window by which the RBPs recognize pre-mRNAs, and thus altering the binding of certain trans-acting splicing factors like RBFox2. The elevated cancer cell proliferation requires rapid mRNA synthesis, which is usually controlled by epigenetic modifications that increase transcription efficiency. Consistent with this idea, the histone modification markers associated with transcription activation showed an enrichment in the CASE-containing genes, whereas the histone modification that correlated with transcription repression is relatively depleted in these genes (Fig. 5E and fig. S5B). However, we must acknowledge that the correlation between histone modifications and transcription elongation is a little weak [Spearman's rank correlation ~0.3 (*59*)], despite being statistically significant (*48*, *49*, *60*, *61*). This weak correlation is probably due to the large noise from big data; however, a weak correlation is still informative and can serve as a piece of indirect evidence. Future studies are needed to dissect the detailed interaction between epigenetic modifications and CASE skipping in cancers. Another caveat is the limitation from data availability: The histone modification data we used were from human cancer cells that are incomplete reflection of patient sample. Currently, the reliable datasets from patient samples are insufficient for a thorough analysis, probably due to technical difficulties in the experiments and the sample acquisition, and future studies with patient samples should provide a clearer picture.

This study also identified several RBPs that play an important role in specific regulation of CASE splicing. Our integrative analysis across various transcriptome-wide datasets was performed on three major CASE clusters rather than each individual CASE, which reduced the noises from sample heterogeneity and increase the robustness of the analysis. This analysis suggested new roles of some core spliceosomal components in regulating AS in a length-dependent fashion, such as AQR, SF3B4, PRPF8, U2AF2, and SF3B1. Because the alteration in transcription rate usually affects AS by changing the splicing time window, the specific binding and regulation by these factors may explain why some short exons are especially sensitive to changes of transcription rate. In particular, we found that RBFox2, a canonical splicing factor reduced in most cancers, plays a key role in promoting CASE inclusion. This finding may provide a possible route to restore CASE splicing in cancer. In addition, several core splicing factors were found to affect short and long exons differentially (fig. S7), probably due to the weak splice sites of CASEs, making them more sensitive to the perturbance of core splicing factors. For example, the SEs regulated by U2AF2 are generally weaker than control exons, which are similar to CASEs. The enriched motifs surrounding CASEs may also selectively bind to certain core splicing factors. For example, the poly T sequences enriched in the upstream of CASEs (Fig. 2F) were also known to specifically bind U2AF2 (*62*). It is possible that different core splicing factors may conduct

length-dependent splicing regulation with different mechanisms. Additional study on the mechanisms of length-dependent splicing will open a new window to the unknown regulatory complexity of AS.

Collectively, this study reveals a general rule for AS dysregulation in cancers, provides a simple and practical route for cancer stratification with AS, and uncovers the potential underlying mechanisms. Because the splicing dysregulation was recently found to associate with cancer immunotherapy and drug assistance (*63*, *64*), our study provides valuable information for cancer prediction with clinical implications.

MATERIALS AND METHODS

Computational pipeline to identify cancer-associated AS events

To identify cancer-related AS events, we downloaded the level 3 RNA-seq data from 6788 patients in TCGA project (https://gdac.broadinstitute.org/) consisting of 18 types of solid tumors with paired adjacent normal tissues. We combined two annotations of SEs from MISO (*65*) and vast-tools (*57*) and calculated PSI values of the SEs in the annotation in all samples using the reads number of each junction according to the following formula

$$\mathrm{PSI} = \frac{0.5(J_1 + J_2)}{0.5(J_1 + J_2) + J_3}$$

where J_1 and J_2 are the read counts of two adjacent junctions of the SE that support exon inclusion, while J_3 is the junction read count supporting exon exclusion (Fig. 1A). In each cancer type, we assume that the PSI values of a particular SE follow normal distribution, and identified cancer-associated exons (or mis-spliced exons) that are significantly altered in tumor versus normal tissues using the cutoff $|\Delta\mathrm{PSI}_{avg}| > 0.1$ and $P < 0.05$ by Student's t test.

Analysis of general features of CASEs

All alternative exons were divided into four categories: 494 CASEs, 1571 CALEs, 13,265 other short exons, and 130,875 other long exons based on the exon length and the relative changes in cancers. The sequence and functional characteristics in each group were compared. The splice site scores were calculated using the maximal entropy models (*66*). The enriched motifs of CASEs were identified based on the frequencies of each hexamer in the set of CASEs versus the control set of exons using a statistic pipeline (*23*). Two control sets, all human exons or CALEs, were used. The enriched hexamers were identified in three different regions: two intronic regions adjacent to the SEs (−300 to −20 nt in the upstream intron and 10 to 300 nt in the downstream intron) and the exonic region within SEs. All hexamers with enrichment z scores > 4 were clustered in Clustal Omega, and the motif logos of each cluster were plotted using R package "ggseqlogo."

To determine the conservation of exonic sequences and their flanking intronic regions, we used human phastCons data from alignments of 46 placental mammal, which was downloaded from UCSC (http://genome.ucsc.edu/). The average phastCons score of each site within the 100-nt intron region surrounding the exons was calculated and plotted (Fig. 2E). The GO analysis for the human genes harboring CASEs was built using R package "clusterProfiler."

To investigate features of the peptides encoded by each class of exons, we translated exons in all three frames and aligned the products to the canonical UniProt sequences to obtain the correct amino acid sequence. The intrinsic disorder region of each peptide was predicted using IUPred (*67*).

The effect of transcription elongation rate on CASEs

We downloaded the RNA-seq data of the human embryonic kidney (HEK) 293 cell line that expresses the RNA Pol II mutants with fast elongation rate from Gene Expression Omnibus (GEO) (accession number GSE63375). The elongation rate of cells with Pol II fast mutation was estimated by GRO (Global run-on sequencing)-seq in Fong *et al.*'s (*46*) study using the positions relative to the transcription start site (TSS) of three time points. Briefly, doxycycline-induced, α-amanitin–treated cells were treated with 100 mM DRB (D-ribofuranosylbenzimidazole) for 3 hours. DRB was subsequently washed away by phosphate-buffered saline three times, and the cell nuclei were harvested at t = 0, 10, and 20 min after washing for the GRO-seq experiments using BrUTP (Bromouridine-triphosphate) labeling. For the GRO-seq reads, genes were aligned at TSSs, and RPKM (reads per kilobase million) was calculated for the control sample without DRB treatment. The positions of Pol II wave relative to the TSS for the three time points were calculated and linearly fitted to estimate the elongation rate. The RNA-seq reads were aligned to the human genome (hg19) using STAR with the standard parameters, and the PSI values were calculated using rMATS (*68*). The SEs associated with altered rates of transcription elongation were defined with the cutoff $|\Delta\mathrm{PSI}_{avg}| > 0.1$ and $P < 0.05$ by comparing Pol II mutation versus WT.

Plasmid construction and generation of Flp-In cell lines

The WT and mutated forms of B10-tagged human Rpb1 were amplified from pAT7h1αAm^r (*69*) and subcloned into the Bam HI–Not I site of pcDNA5/FRT/TO. The resulting plasmids were cotransfected with pOG44 into Flp-In T-REx 293 cells with a 1:9 ratio. Forty-eight hours after the transfection, the cells were washed and split into fresh medium, incubated at 37°C overnight, and then changed with fresh medium containing hygromycin (100 μg/ml). The cells were cultured in the selective medium until the monoclonal colonies formed (fresh medium was changed every 3 days). Three to four monoclonal cells were picked and expanded for further use, and the insert sequences were confirmed by Sanger sequencing. The expression of Rpb1 was induced by adding doxycycline to a final concentration of 1 μg/ml for 2 days. All experiments were performed after treatment with α-amanitin (2.5 mg/ml) for a further 42 hours, at which time all cell lines were viable despite the fact that the endogenous Pol II was inactive.

RNA extraction and semiquantitative RT-PCR

The RNAs were extracted with the TRIzol Reagent (#15596026, Thermo Fisher Scientific) following the manufacturer's instructions. RNA (1 μg) was reverse-transcribed by the PrimeScript RT Reagent Kit with gDNA Eraser (RR047A, Takara). One-tenth of the RT production was used as the template for PCR amplification [28 cycles, labeled with trace amount of Cy5-Dctp (PA55021, GE Healthcare)], with primers corresponding to the exons of the selected genes (table S1). The PCR products were separated by 10% TBE (tris borate EDTA) polyacrylamide gel electrophoresis (PAGE) gel and stained with SYBR Safe (S33102, Invitrogen). The gels were scanned with the ChemiDoc Touch Imaging System (Image Lab, Bio-Rad). The quantification of the bands was performed by the ImageJ software described previously (*70*).

Histone modification analysis

The processed signal files (in bigwig format) from ChIP-seq of several histone modifications (H4K20me1, H3K27me3, H3K9ac, and H3K4me2) were downloaded from ENCODE consortium (https://encodeproject.org). To examine whether the genes containing cancer-associated exons have different transcription rates, we divided all human genes into four classes based on whether the genes contain CASEs, CALEs, other short exons, or none of these three types. The genomic region encompassing 3000 nt before the TSS and 3000 nt after the TES (transcription end site) of all genes was used in the analysis. The average ChIP-seq signal in each region was obtained with the script of deepTools ComputeMatrix, and the figures were plotted using deepTools plotProfile.

Analysis of Pol II occupancy and nascent RNAs

The processed Pol II signal file (in bed format) from ChIP-seq of Pol II was downloaded from ENCODE consortium (https://encodeproject.org). We normalized the peak signals to examine the Pol II occupancy in each type of exon. The average peak signals encompassing 500 nt before exon and 500 nt after exon were calculated to compare among these four types of exons.

The TT-seq data are downloaded from GEO (accession number GSE148433) (*51*), and the raw reads were aligned to the human genome (hg19) using hisat2. The level of four types of exons in TT-seq was further calculated using featureCounts.

RBP analysis

RNA-seq data from the knockdown of 227 RBPs and the eCLIP data (both in K562 cell line) were downloaded from ENCODE consortium. The sequencing reads were aligned to the human genome (hg19) using STAR with the standard parameters, and PSI values were calculated using rMATS (*68*). For each RBP, we compared the changes between RBP knockdown and paired control samples to identify the SEs with significant changes (with cutoff of $|\Delta PSI_{avg}| > 0.1$ and $P < 0.05$), which were defined as RBP-associated SEs. We focused on the RBP-associated CASEs for further studies.

For the candidate RBPs that affect CASE splicing based on the knockdown data, we searched their binding sites within the CASEs and their adjacent introns (1000 nt) and identified the RBPs that directly regulate CASEs. A tightly connected regulation network between RBPs and CASEs was obtained from the integrated analysis and built using Cytoscape.

Knockdown of RBPs with shRNAs

Human lung cancer NCI-H1299 cell lines were obtained from the American Type Culture Collection (Manassas, VA, USA). To stably knock down certain RBPs in H1299 cells, lentiviral vectors were used. We transfected 293T cells with pLKO.1-RBP (pLKO.1 empty vector as control) together with PAX2 and PMD2 according to the manufacturer's protocols. The supernatant media containing virus were collected by centrifugation to remove any cellular contaminant. Further, H1299 cells were infected with the viral particles, and the stably integrated cells were selected with puromycin (5 μg/ml) for 5 days. Then, cells were maintained in medium containing puromycin (2 μg/ml) at 37°C in a humidified incubator with 5% CO_2. All the stable cell lines were confirmed by quantitative PCR before further analysis.

PCA, PLS-DA, and the random forest model

To use CASEs as molecular markers for cancer prediction, we combined all types of cancers into a training set of 6788 tumors and 705 normal samples. We considered the 60 CASEs that are detected in at least 5500 tumor samples and 450 normal samples, and the samples with at least 20 CASEs detected were used for our analyses. The PSI values of CASEs in each sample were used as input for PCA and PLS-DA with R package "mixOmics."

We further trained a random forest model based on the PSI values of the 60 CASEs to classify tumor and normal samples, which was trained using the "randomforest" R package. We performed a four-fold cross-validation in 100 randomly trails to evaluate the prediction accuracy, where 75% of the samples were randomly selected for training and the remaining 25% for testing. We also evaluated the performance using a cross-validation between the RNA-seq data from Chinese LUAD and ESCC patients and the Western patients of TCGA. Briefly, we trained the model using LUAD samples from TCGA and tested the performance with data from Chinese LUAD patients, or trained the model based on Chinese ESCC patients and tested using ESCA samples from TCGA. The receiver operating characteristic (ROC) curve was plotted using R package "pROC."

Survival analysis using CASEs

We combined the patients of all cancer types in TCGA and downloaded the clinical data from Genomic Data Commons (https://portal.gdc.cancer.gov). For each CASE, we ranked the PSI values in all patients and grouped the patients in the top and bottom quartiles into "PSI high" and "PSI low" groups. The comparison of the overall survival between these two groups was performed using Cox regression in the R package "survival" and "survminer."

To increase the sensitivity in our prediction, eight CASEs were selected to compute a risk factor for each patient sample. We ranked the PSI values of each CASE and assign a relative score from [1,2,3,4] for each PSI quartile. A CASE-based risk factor for each sample was then defined by the mean of this relative score for the eight CASEs selected. We used the rank of this risk factor to classify the top and bottom quartiles of patients into two groups and compared the survival time and immune cell infiltration between the two groups.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abn9232

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, C. B. Burge, Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
2. Q. Pan, O. Shai, L. J. Lee, B. J. Frey, B. J. Blencowe, Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
3. J. M. Johnson, J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton, D. D. Shoemaker, Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**, 2141–2144 (2003).
4. Z. Wang, C. B. Burge, Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802–813 (2008).
5. A. G. Matera, Z. Wang, A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* **15**, 108–121 (2014).
6. G. S. Wang, T. A. Cooper, Splicing in disease: Disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* **8**, 749–761 (2007).
7. Y. Wang, Y. Bao, S. Zhang, Z. Wang, Splicing dysregulation in cancer: From mechanistic understanding to a new class of therapeutic targets. *Sci. China Life Sci.* **63**, 469–484 (2020).
8. H. Dvinge, E. Kim, O. Abdel-Wahab, R. K. Bradley, RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer* **16**, 413–430 (2016).

9. Y. Wang, D. Chen, H. Qian, Y. S. Tsai, S. Shao, Q. Liu, D. Dominguez, Z. Wang, The splicing factor RBM4 controls apoptosis, proliferation, and migration to suppress tumor progression. *Cancer Cell* **26**, 374–389 (2014).
10. A. Sveen, S. Kilpinen, A. Ruusulehto, R. A. Lothe, R. I. Skotheim, Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene* **35**, 2413–2427 (2016).
11. X. Song, Z. Zeng, H. Wei, Z. Wang, Alternative splicing in cancers: From aberrant regulation to new therapeutics. *Semin. Cell Dev. Biol.* **75**, 13–22 (2018).
12. S. C. Lee, O. Abdel-Wahab, Therapeutic targeting of splicing in cancer. *Nat. Med.* **22**, 976–986 (2016).
13. M. L. Miller, E. Reznik, N. P. Gauthier, B. A. Aksoy, A. Korkut, J. Gao, G. Ciriello, N. Schultz, C. Sander, Pan-cancer analysis of mutation hotspots in protein domains. *Cell Syst.* **1**, 197–209 (2015).
14. Y. S. Tsai, D. Dominguez, S. M. Gomez, Z. Wang, Transcriptome-wide identification and study of cancer-specific splicing events across multiple tumors. *Oncotarget* **6**, 6825–6839 (2015).
15. S. Shen, Y. Wang, C. Wang, Y. N. Wu, Y. Xing, SURVIV for survival analysis of mRNA isoform variation. *Nat. Commun.* **7**, 11548 (2016).
16. M. Chen, J. L. Manley, Mechanisms of alternative splicing regulation: Insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.* **10**, 741–754 (2009).
17. C. E. Meacham, S. J. Morrison, Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–337 (2013).
18. R. Karni, E. de Stanchina, S. W. Lowe, R. Sinha, D. Mu, A. R. Krainer, The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat. Struct. Mol. Biol.* **14**, 185–193 (2007).
19. E. Sebestyén, B. Singh, B. Miñana, A. Pagès, F. Mateo, M. A. Pujana, J. Valcárcel, E. Eyras, Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res.* **26**, 732–744 (2016).
20. S. Zhang, Y. Bao, X. Shen, Y. Pan, Y. Sun, M. Xiao, K. Chen, H. Wei, J. Zuo, D. Saffen, W. X. Zong, Y. Sun, Z. Wang, Y. Wang, RNA binding motif protein 10 suppresses lung cancer progression by controlling alternative splicing of eukaryotic translation initiation factor 4H. *EBioMedicine* **61**, 103067 (2020).
21. Y. Song, L. Li, Y. Ou, Z. Gao, E. Li, X. Li, W. Zhang, J. Wang, L. Xu, Y. Zhou, X. Ma, L. Liu, Z. Zhao, X. Huang, J. Fan, L. Dong, G. Chen, L. Ma, J. Yang, L. Chen, M. He, M. Li, X. Zhuang, K. Huang, K. Qiu, G. Yin, G. Guo, Q. Feng, P. Chen, Z. Wu, J. Wu, L. Ma, J. Zhao, L. Luo, M. Fu, B. Xu, B. Chen, Y. Li, T. Tong, M. Wang, Z. Liu, D. Lin, X. Zhang, H. Yang, J. Wang, Q. Zhan, Identification of genomic alterations in oesophageal squamous cell cancer. *Nature* **509**, 91–95 (2014).
22. D. F. T. Veiga, A. Nesta, Y. Zhao, A. D. Mays, R. Huynh, R. Rossi, T.-C. Wu, K. Palucka, O. Anczukow, C. R. Beck, J. Banchereau, A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. *Sci. Adv.* **8**, eabg6711 (2022).
23. W. G. Fairbrother, R.-F. Yeh, P. A. Sharp, C. B. Burge, Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–1013 (2002).
24. Y. Wang, Z. Wang, Systematical identification of splicing regulatory cis-elements and cognate trans-factors. *Methods* **65**, 350–358 (2014).
25. Y. Wang, M. Ma, X. Xiao, Z. Wang, Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat. Struct. Mol. Biol.* **19**, 1044–1052 (2012).
26. Z. M. Zheng, Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *J. Biomed. Sci.* **11**, 278–294 (2004).
27. S. Pandit, Y. Zhou, L. Shiue, G. Coutinho-Mansfield, H. Li, J. Qiu, J. Huang, G. W. Yeo, M. Ares Jr., X. D. Fu, Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol. Cell* **50**, 223–235 (2013).
28. R. Spellman, A. Rideau, A. Matlin, C. Gooding, F. Robinson, N. McGlincy, S. N. Grellscheid, J. Southby, M. Wollerton, C. W. Smith, Regulation of alternative splicing by PTB and associated factors. *Biochem. Soc. Trans.* **33**, 457–460 (2005).
29. K. Sawicka, M. Bushell, K. A. Spriggs, A. E. Willis, Polypyrimidine-tract-binding protein: A multifunctional RNA-binding protein. *Biochem. Soc. Trans.* **36**, 641–647 (2008).
30. Y. Jin, H. Suzuki, S. Maegawa, H. Endo, S. Sugano, K. Hashimoto, K. Yasuda, K. Inoue, A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *EMBO J.* **22**, 905–912 (2003).
31. H. Kuroyanagi, Fox-1 family of RNA-binding proteins. *Cell. Mol. Life Sci.* **66**, 3895–3907 (2009).
32. J. P. Venables, R. Klinck, C. Koh, J. Gervais-Bird, A. Bramard, L. Inkel, M. Durand, S. Couture, U. Froehlich, E. Lapointe, J. F. Lucier, P. Thibault, C. Rancourt, K. Tremblay, P. Prinos, B. Chabot, S. A. Elela, Cancer-associated regulation of alternative splicing. *Nat. Struct. Mol. Biol.* **16**, 670–676 (2009).
33. J. P. Venables, J. P. Brosseau, G. Gadea, R. Klinck, P. Prinos, J. F. Beaulieu, E. Lapointe, M. Durand, P. Thibault, K. Tremblay, F. Rousset, J. Tazi, S. Abou Elela, B. Chabot, RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. *Mol. Cell. Biol.* **33**, 396–405 (2013).
34. N. J. McGlincy, C. W. Smith, Alternative splicing resulting in nonsense-mediated mRNA decay: What is the meaning of nonsense? *Trends Biochem. Sci.* **33**, 385–393 (2008).
35. B. Mészáros, G. Erdos, Z. Dosztányi, IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018).
36. P. E. Wright, H. J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).
37. R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, M. M. Babu, Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
38. A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, A. A. Alizadeh, Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
39. S. Bulfone-Paus, R. Bahri, Mast cells as regulators of T cell responses. *Front. Immunol.* **6**, 394 (2015).
40. G. Dujardin, C. Lafaille, M. de la Mata, L. E. Marasco, M. J. Muñoz, C. Le Jossic-Corcos, L. Corcos, A. R. Kornblihtt, How slow RNA polymerase II elongation favors alternative exon skipping. *Mol. Cell* **54**, 683–690 (2014).
41. D. L. Bentley, Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* **15**, 163–175 (2014).
42. G. Dujardin, C. Lafaille, E. Petrillo, V. Buggiano, L. I. Gómez Acuña, A. Fiszbein, M. A. Godoy Herz, N. Nieto Moreno, M. J. Muñoz, M. Alló, I. E. Schor, A. R. Kornblihtt, Transcriptional elongation and alternative splicing. *Biochim. Biophys. Acta* **1829**, 134–140 (2013).
43. A. R. Kornblihtt, M. de la Mata, J. P. Fededa, M. J. Munoz, G. Nogues, Multiple links between transcription and splicing. *RNA* **10**, 1489–1498 (2004).
44. A. Baluapuri, J. Hofstetter, N. Dudvarski Stankovic, T. Endres, P. Bhandare, S. M. Vos, B. Adhikari, J. D. Schwarz, A. Narain, M. Vogt, S. Y. Wang, R. Duster, L. A. Jung, J. T. Vanselow, A. Wiegering, M. Geyer, H. M. Maric, P. Gallant, S. Walz, A. Schlosser, P. Cramer, M. Eilers, E. Wolf, MYC recruits SPT5 to RNA polymerase II to promote processive transcription elongation. *Mol. Cell* **74**, 674–687.e11 (2019).
45. G. C. Roberts, C. Gooding, H. Y. Mak, N. J. Proudfoot, C. W. Smith, Co-transcriptional commitment to alternative splice site selection. *Nucleic Acids Res.* **26**, 5568–5572 (1998).
46. N. Fong, H. Kim, Y. Zhou, X. Ji, J. Qiu, T. Saldi, K. Diener, K. Jones, X. D. Fu, D. L. Bentley, Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* **28**, 2663–2676 (2014).
47. A. Rybak-Wolf, C. Stottmeister, P. Glazar, M. Jens, N. Pino, S. Giusti, M. Hanan, M. Behm, O. Bartok, R. Ashwal-Fluss, M. Herzog, L. Schreyer, P. Papavasileiou, A. Ivanov, M. Ohman, D. Refojo, S. Kadener, N. Rajewsky, Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol. Cell* **58**, 870–885 (2015).
48. A. Veloso, K. S. Kirkconnell, B. Magnuson, B. Biewen, M. T. Paulsen, T. E. Wilson, M. Ljungman, Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* **24**, 896–905 (2014).
49. A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, K. Zhao, High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
50. B. Schuettengruber, D. Chourrout, M. Vervoort, B. Leblanc, G. Cavalli, Genome regulation by polycomb and trithorax proteins. *Cell* **128**, 735–745 (2007).
51. L. Caizzi, S. Monteiro-Martins, B. Schwalb, K. Lysakovskaia, J. Schmitzova, A. Sawicka, Y. Chen, M. Lidschreiber, P. Cramer, Efficient RNA polymerase II pause release requires U2 snRNP function. *Mol. Cell* **81**, 1920–1934.e9 (2021).
52. G. W. Yeo, N. G. Coufal, T. Y. Liang, G. E. Peng, X. D. Fu, F. H. Gage, An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat. Struct. Mol. Biol.* **16**, 130–137 (2009).
53. C. Zhang, Z. Zhang, J. Castle, S. Sun, J. Johnson, A. R. Krainer, M. Q. Zhang, Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes Dev.* **22**, 2550–2563 (2008).
54. T. Hirose, T. Ideue, M. Nagai, M. Hagiwara, M. D. Shu, J. A. Steitz, A spliceosomal intron binding protein, IBP160, links position-dependent assembly of intron-encoded box C/D snoRNP to pre-mRNA splicing. *Mol. Cell* **23**, 673–684 (2006).
55. T. Yoshizawa, R. S. Nozawa, T. Z. Jia, T. Saio, E. Mori, Biological phase separation: Cell biology meets biophysics. *Biophys. Rev.* **12**, 519–539 (2020).
56. Y. Lin, S. L. Currie, M. K. Rosen, Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs. *J. Biol. Chem.* **292**, 19110–19120 (2017).
57. M. Irimia, R. J. Weatheritt, J. D. Ellis, N. N. Parikshak, T. Gonatopoulos-Pournatzis, M. Babor, M. Quesnel-Vallieres, J. Tapial, B. Raj, D. O'Hanlon, M. Barrios-Rodiles, M. J. Sternberg, S. P. Cordes, F. P. Roth, J. L. Wrana, D. H. Geschwind, B. J. Blencowe, A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511–1523 (2014).
58. H. Chettouh, L. Fartoux, L. Aoudjehane, D. Wendum, A. Clapéron, Y. Chrétien, C. Rey, O. Scatton, O. Soubrane, F. Conti, F. Praz, C. Housset, O. Rosmorduc, C. Desbois-Mouthon,

Mitogenic insulin receptor-A is overexpressed in human hepatocellular carcinoma due to EGFR-mediated dysregulation of RNA splicing factors. *Cancer Res.* **73**, 3974–3986 (2013).
59. C. G. Danko, N. Hah, X. Luo, A. L. Martins, L. Core, J. T. Lis, A. Siepel, W. L. Kraus, Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol. Cell* **50**, 212–222 (2013).
60. M. Gerber, A. Shilatifard, Transcriptional elongation by RNA polymerase II and histone methylation. *J. Biol. Chem.* **278**, 26303–26306 (2003).
61. K. Jamieson, M. R. Rountree, Z. A. Lewis, J. E. Stajich, E. U. Selker, Regional control of histone H3 lysine 27 methylation in Neurospora. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6027–6032 (2013).
62. F. X. R. Sutandy, S. Ebersberger, L. Huang, A. Busch, M. Bach, H. S. Kang, J. Fallmann, D. Maticzka, R. Backofen, P. F. Stadler, K. Zarnack, M. Sattler, S. Legewie, J. König, In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. *Genome Res.* **28**, 699–713 (2018).
63. Z. Siegfried, R. Karni, The role of alternative splicing in cancer drug resistance. *Curr. Opin. Genet. Dev.* **48**, 16–21 (2018).
64. Y. Pan, K. E. Kadash-Edmondson, R. Wang, J. Phillips, S. Liu, A. Ribas, R. Aplenc, O. N. Witte, Y. Xing, RNA dysregulation: An expanding source of cancer immunotherapy targets. *Trends Pharmacol. Sci.* **42**, 268–282 (2021).
65. Y. Katz, E. T. Wang, E. M. Airoldi, C. B. Burge, Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
66. G. Yeo, C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
67. Z. Dosztányi, V. Csizmok, P. Tompa, I. Simon, IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434 (2005).
68. S. Shen, J. W. Park, Z. X. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, Y. Xing, rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E5593–E5601 (2014).
69. V. T. Nguyen, F. Giannoni, M. F. Dubois, S. J. Seo, M. Vigneron, C. Kédinger, O. Bensaude, In vivo degradation of RNA polymerase II largest subunit triggered by alpha-amanitin. *Nucleic Acids Res.* **24**, 2924–2929 (1996).
70. M. Mao, Y. Hu, Y. Yang, Y. Qian, H. Wei, W. Fan, Y. Yang, X. Li, Z. Wang, Modeling and predicting the activities of trans-acting splicing factors with machine learning. *Cell Syst.* **7**, 510–520.e4 (2018).

Acknowledgments: We thank A. Fiszbein for the pAT7 plasmid and X. Li for discussion and suggestions. **Funding:** This work was supported by the National Key Research and Development Program of China to Z.W. (2018YFA0107602); the National Natural Science Foundation of China to Z.W. (31730110 and 91940303), M.M. (31971367), Yun Yang (91753135 and 31870814), and Yang Wang (81830088); the NIH Basic Research in Cancer Health Disparities Award to S.R.P. and J.A.F. (R01CA220314); and the Starry Night Science Fund at Shanghai Institute for Advanced Study of Zhejiang University (SN-ZJU-SIAS-009). Z.W. was also supported by the type A CAS Pioneer 100-Talent program. M.M. was supported by the National Postdoctoral Program for Innovative Talents (BX20180336) and Shanghai Super Postdoctoral Program. Yun Yang was sponsored by the Youth Innovation Promotion Association CAS (2019267), SA-SIBS Scholarship Program, and Shanghai Science and Technology Committee Rising-Star Program (19QA1410500). **Author contributions:** S.Z. and Z.W. designed the study and analyzed the data. M.M., Y.L., and Yingqun Yang performed experiments and analyzed the data. W.H. and Y.S. provided clinical samples. Yongbo Wang, Yun Yang, M.A.A., J.A.F., S.R.P., and Yang Wang provided part of the data and interpreted the results. S.Z. and Z.W. wrote the manuscript. **Competing interests:** Z.W. and S.Z. are inventors on a patent application related to this work filed by Shanghai Institute of Nutrition and Health (filed on 20 June 2022, case reception number 374787789). The authors declare no other competing interests. **Data and materials availability:** The source data files were obtained from Gene Expression Omnibus (accession number GSE63375). RNA-seq data from the knockdown of 227 RBPs, eCLIP data of RBPs, and the processed signal files (in bigwig format) of ChIP-seq data of several histone modifications were downloaded from ENCODE consortium (https://encodeproject.org). Long-read sequencing data for breast cancer are downloaded from the European Genome Archive database (accession number EGAS00001004819). All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The source code of the analytical pipeline is available at https://github.com/Sirui724/CASE_splicing and from Zenodo at https://doi.org/10.5281/zenodo.6474460.

Submitted 31 December 2021
Accepted 1 July 2022
Published 17 August 2022
10.1126/sciadv.abn9232