Supplementary Notes

Validation of the inference of DrAS-Net

We used three independent methods to validate the inferences of the DrAS-Net (Figure S3B). First, we did a cross-validation based on 33 datasets in TCGA (Figure S3B, left panel). For each pair of these 33 datasets (in total 528 pairs), we derived the mutation-AS pairs in each dataset by the proposed method, DrAS-Net. For 216 pairs of datasets with common mutations, we investigated whether the same mutation-AS pairs were significantly overlapped with each other in each pair of dataset. As a result, we found that 83.6% pairs significantly ($p<0.05$, hypergeometric test) share common mutation-AS relationships (Figure S3C). These results suggest that the inferences derived from our study are robust across different datasets.

Second, we validated the inferences using 506 cell lines from The Cancer Cell Line Encyclopedia (CCLE) project. We obtained the somatic mutations in each cell line which were measured by hybrid capture sequencing. The mutations were mapped to TCGA tumor samples to find common mutations. In addition, we also used the RNA-seq dataset to explore whether the cell lines with mutations show perturbed expression of corresponding exons. If the mutation and perturbed expression of exon occurred in the same cell line, we regarded this as a validation (Figure S3B, middle panel). Overall, we observed a significant validation rate compared to random controls ($p<1.0e-4$). For instance, the mutation REL-R219C and the AS event MAGOHB: exon 2.1-2.2 skipping was validated in the HCT116 cell line (Figure S3D).

Third, we validated the inferences by literature mining. We used the /gene, mutation, cancer/ or /gene, splicing, cancer/ as keywords to search the entire Pubmed literature by data mining. We found that a significant fraction of mutations and alternative splicing in DrAS-Net could be validated in literature, providing evidence of their roles in cancer (Figure S3E and Table S6). Together, all these results validated the inferences of DrAS-Net.

Network-based model identifies somatic driver mutations with deleterious effects

Our analyses have demonstrated the functional importance of the identified driver genes. Several methods for assessing the effects of mutations on protein function have been developed over the years, and these methods are complementary to each other. We therefore used Combined Annotation–Dependent Depletion (CADD) (Kircher et al., 2014), which is a framework that integrates multiple annotations into one metric, to explore the functional impact of mutations in these driver genes. For each cancer type, we randomly selected the same number of mutations as background controls. We then compared the scores of our identified driver mutations with those of randomly selected mutations. In the majority (63.6%; 21/33) of cancer types, our identified driver mutations had significantly higher scores (Wilcoxon rank-sum test $p<0.05$) than controls (Figure S3F), suggesting that the identified driver mutations have deleterious effects in cancer.

Evolutionary conservation of a mutated residue has also been demonstrated to reflect the importance of the mutational event (Watson et al., 2013). We therefore explored the conservation feature of the identified mutations. We found that positions harboring driver mutations were more likely to be conserved than positions harboring randomly selected mutations in most (87.9%; 29/33) cancer types (Figure S3G). Moreover, we investigated the

proportion of driver mutations that were located in protein domains. We observed that different types of cancer exhibited variable patterns. In 48.5% (16/33) of cancer types, driver mutations were depleted in protein domains (Figure S3J). To further evaluate the functional impact of these candidate driver mutations, we examined whether they affect intrinsically disordered regions (IDRs), which had been demonstrated to play critical roles in cancer signaling regulation (Latysheva et al., 2016). We computed the probabilities of these driver mutations to reside in IDRs using the IUPred program (Dosztanyi et al., 2005). We found that they were more likely to reside in IDRs than random controls in the majority (93.9%; 31/33) of cancer types (Figure S3H). To further investigate if the driver mutations affect functional motifs within IDRs (Sebestyen et al., 2016; Tompa et al., 2014), we implemented the ANCHOR algorithm to identify protein-binding sites that reside in disordered regions. We observed that the proportion of driver mutations located in putative motifs were significantly higher ($p<0.05$, Fisher's exact test) than other mutations in most (78.8%; 26/33) cancer types (Figure S3I). Together, we have delineated a network-based framework to identify driver mutations that likely play roles across a wide range of cancer types.

To investigate whether these mutations led to specific alternative splicing outcomes, we transfected HEK293T cells with plasmids encoding for either wild-type genes or identified mutants. Analysis of *cis*-regulated splicing in *CASP6*, *BCL6*, and *PDE9A* found 7 mutants that showed significant reduction in transcription of the respective exons relative to their matching wild-type controls (Figure S3K). Furthermore, profiling of *trans*-regulated cases mediated by mutations in the FXR2 RNA-binding protein revealed both gain of exons in *TRAF2* (exon 6.2) as well as loss of exons in *PAF1* (exon 2) and the DNA damage response gene *RBBP8* (exon 19.2) (Figure S3K). Together, our results suggest that different genetic mutations likely influence distinctly different AS events in cancer.

Supplemental Experimental Procedures

Identification of the differential AS landscape across cancer types

To investigate the landscape of AS across cancer types, we obtained the genome wide AS datasets from the TCGASpliceSeq database (Ryan et al., 2016), which is a compendium of AS events in cancer. The Percent Spliced In (PSI) value was used for quantifying splicing events, which is defined as the number of reads indicating that a transcript element is present divided by the total number of reads covering the AS event. In total, the PSIs for 10,699 samples across 33 types of cancer were obtained, including 749 normal samples for 23 types of cancer (Table S1). Seven types of AS events were considered in our analysis, including exon skipping, alternative donor site, alternative acceptor site, retained intron, mutually exclusive exons, alternative terminator and alternative promoter. We required the percentage of samples with PSI values to be greater than 75% and the missing values were imputed with the mean of all samples. Next, Wilcoxon rank-sum test was used to identify the differential AS events in each cancer. We only analyzed the cancer types with more than five normal samples. P-values were corrected by BH method. The AS events with adjusted p-values less than 0.01 and the fold-changes greater than two-fold were identified as differential AS in each cancer.

Analysis of the cancer similarity based on differential AS patterns

We computed the paired similarity for 18 types of cancer based on the differential AS patterns. For each pair of cancer types, we computed the similarity score (*Sim*) based on the differential AS events as follows:

$$Sim(Ca, Cb) = \frac{\#\{DSa \cap DSb\}}{\min\left(\#\{DSa\}, \#\{DSb\}\right)}$$

where *DSa* indicates the differential AS events in cancer type *Ca* while *DSb* indicates the differential AS events in cancer type *Cb*. Cancer types were clustered based on this similarity matrix. In addition, cancer types were also clustered based on the similarity of AS subtype patterns.

Gene Set Enrichment Analysis comparing tumors with and without perturbed AS events

To compare the expression difference of tumor samples with perturbed AS events and those without such events, we applied Wilcoxon rank-sum test for RNA-seq expression profiles for each type of cancer. Genes were ranked by a score in each cancer type, which was defined as the negative log10 of the Wilcoxon rank-sum test analysis-derived false discovery rate (FDR) multiplied by the sign of the logFC (log fold change). The score was normalized to relative rank in each type of cancer and we used the average relative rank in all cancer types for Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005). GSEA was performed by applying the "weighted" enrichment statistics on the relative rank for enrichment or depletion. Here, we utilized pathways included in KEGG, REACTOME, BIOCARTA, and PID databases. The gene sets containing between 20 and 500 genes were analyzed and 1,000 permutations were performed to get the significance levels.

Immune signature score, cell cycle signature score and copy number variation analysis

We aimed to determine whether the tumor samples with perturbed AS events exhibited distinct immune features. The immune signature score, cell cycle signature score and somatic copy number alteration (SCNA) level for cancer samples were calculated as described (Davoli et al., 2017). Wilcoxon rank-sum test was used to test the difference of these scores between tumor samples with/without perturbed AS events.

Prioritization of mutated genes and genomic mutations in cancer

After assembling all the mutated gene-AS associations in each patient, we constructed a bipartite network for each cancer. Next, we prioritized the mutated genes by identifying genes with the largest extent of AS disruption in cancer. The mutated genes in each bipartite network were ranked by degree and then for each iteration, we chose a

mutated gene that covered the largest number of uncovered AS. The greedy algorithm was stopped when all the AS events were covered. Then the mutations in the identified genes in the corresponding sample were assembled as driver mutations in each cancer. This prioritization was only applied to trans-regulations. All the cis-regulated mutations were added to the driver list.

Functional analyses of the driver genes and mutations

To investigate the functional importance of the driver genes and mutations, we compared the mutation frequency, the proportion of cancer genes in their network neighbors for the driver genes. All the cancer genes were obtained from Cancer Gene Census (n=609). Moreover, functional enrichment analysis of the driver genes was carried out to investigate whether they were enriched in cancer hallmarks. The gene list of each cancer hallmark was obtained from one of the previous studies (Plaisier et al., 2012). We used hypergeometric test for exploring the statistical significance and the p-value was computed as follow:

$$p = 1 - F(x|N, K, M) = 1 - \sum_{t=0}^{x} \frac{\binom{K}{t}\binom{N-K}{M-t}}{\binom{N}{M}}$$

Where N is the number of genes in the whole genome, of which K genes were involved in the functional category under investigation (such as cancer genes and cancer hallmarks), and the number of candidate driver genes is M, of which x genes were involved in the functional category.

In addition, the CADD score was computed to evaluate the deleteriousness of single nucleotide variants as well as insertion/deletions variants identified in this study (Kircher et al., 2014). As driver mutations have been demonstrated to be conserved, we also computed conservation score for each mutation. The same number of mutations were randomly selected as background in each cancer type. This procedure was repeated for 100 times. The difference was tested by Wilcoxon rank-sum test. For Pfam domain analysis, PfamScan was used to search a FASTA sequence against a library of Pfam HMM. Somatic mutations were marked with 1 if the altered protein residues were located in protein domains, otherwise with 0.

Validation of DrAS-Net

We used three methods to validate the inferences of DrAS net. Firstly, we did a dataset cross-validation based on 33 datasets of various types of cancer. For each pair of dataset, we firstly identified the common mutations and the perturbed AS events in each dataset. Hypergeometric test was used to explore whether the same mutations were likely to mediate the same AS events in two datasets. Dataset pairs with p-value less than 0.05 were regarded as dataset cross-validated. Second, we validated the inferences based on cell lines data from The Cancer Cell Line Encyclopedia (CCLE) project (Barretina et al., 2012). Somatic mutations of cell lines were mapped to the same cancer type and the common mutations were identified. In total, we screened 506 cell lines that can be mapped to

corresponding cancer type. We next analyzed the cell lines with three driver mutations and the cancer type (COAD) with more than five cell lines was analyzed. The RNA-Seq dataset of 12 COAD cell lines were investigated to compute the expression of exons. The expression was measured as reads per million (RPM). This process was performed by bedtools (Quinlan, 2014). The exon with outlier expression were regarded as perturbed AS event. The mutation-AS pair was regarded as validated if the mutation and AS event was occurred in the same cell line. To investigate whether the validated proportion is significantly larger, we randomly selected the same number of mutation-AS pairs as DrAS-Net from all possible pairs. This process was repeated 10,000 times. Thirdly, we searched the literature to test whether the identified mutated genes and perturbed AS genes were significantly co-occurred with cancer. The same number of genes as mutated genes or AS genes were randomly selected and we also get the proportion of genes co-occurred with cancer. This process was repeated 10,000 times.

Disorder analysis of driver mutations

For each residue affected by a genomic mutation, we assessed the likelihood that the residue was located in an intrinsically disordered region of a protein. Protein sequences were subjected to the IUPred program (Dosztanyi et al., 2005) and for each mutation we obtained a predicted disorder probability. We used Wilcoxon rank-sum test to compare the difference between candidate driver mutations and background control for statistical significance. We used ANCHOR to predict disordered binding regions (Meszaros et al., 2009). If mutations were located in predicted disordered binding regions, we labeled them with 1, otherwise with 0.

Functional consequence of AS in cancer

Protein structures were downloaded from the Protein Data Bank (PDB) database (http://www.rcsb.org/) and were shown using the PyMOL tool. The protein domain and sequence annotation were obtained from the Uniprot database (The UniProt, 2017).

Site-directed mutagenesis of driver mutation candidates

To generate cancer mutations, we developed a new site-directed mutagenesis pipeline. For a given mutation, we performed two "primary PCRs" to generate gene fragments using Entry clones in human ORFeome v8.1 as template, and one "fusion PCR" to obtain the mutant allele. For the primary PCRs, two universal primers, Tag1-M13F (*GGCAGACGTGCCTCACTACTTGTAAAACGACGGCCAGT*) and Tag2-M13R (*CTGAGCTTGACGCATTGCTACAGGAAACAGCTATGACC*), and two mutation-specific primers were employed (sequences shown in Table S3). The two DNA fragments flanking the mutation of a gene were amplified using the primer pair Tag1-M13F and Mut-Rev, and the primer pair Mut-Fwd and Tag2-M13R, respectively. For the fusion PCR, the two primary PCR fragments were fused together using the primer pair Tag1 (*GGCAGACGTGCCTCACTACT*) and Tag2 (*CTGAGCTTGACGCATTGCTA*) to generate the mutant allele. To

transfer the mutant allele into the pcDNA3-EGFP destination vector, we performed a Gateway LR reaction using the mutant allele fusion PCR products. After bacterial transformation, single colonies were isolated. The correct mutant clones were verified by sequencing.

Quantification of alternative splicing by qRT-PCR

HEK 293T cells were plated twenty fours before transfection in six well plates. For transfections, 3 μg of DNA was mixed with 9 μg of polyethylenimine (PEI) and incubated for 20 minutes before application to cells. Cells were incubated with PEI/DNA mixture for 8 hours before replacing with fresh growth media. RNA was isolated 36 hours after transfection using a QIAgen RNeasy kit (Qiagen, Hilden, Germany), and cDNA was synthesized using the iScript cDNA synthesis kit (BioRad, Hercules, CA). To determine relative levels of alternative splicing, primers specific for the "gain or loss" region of the AS were designed. Quantitative reverse transcriptase PCR (qRT-PCR) was performed with Power SYBR Green Master Mix (Applied Biosytems, Foster City, CA) per manufacturer's instructions. Results were quantified using the comparative Ct (ΔΔCt) method, normalizing the splice variants to the total transcript level, and each mutant was normalized to its respective wild-type control. Experiments were conducted with two repeats. Primers are given in Table S3.

***cis*-regulation enrichment analysis**

We computed the proportion of *cis*-regulations in all cancer types, and random tests were performed to get the statistical significance of this ratio. We randomly selected the same number of interactions from the original network and then recalculated the proportion of *cis*-regulations. This process was repeated for 1,000 times, and the p-value was defined as the probability of obtaining a larger proportion than what was actually observed.

Identification of RBP target genes

UV crosslinking and immunoprecipitation (CLIP) of ribonucleoprotein complexes is a commonly used approach to identify the RNA binding sites. To identify the RBP-RNA interactions in cancer, we integrated enhanced CLIP (eCLIP) sequencing datasets with shRNA-Seq datasets in HepG2 and K562 cell line. Firstly, 136 eCLIP-Seq datasets for 68 diverse RBPs in HepG2 and 172 datasets for 86 RBPs in K562 cell lines were downloaded from the ENCODE website. Each RBP had two biological replicates. The peak files were directly downloaded. All the peaks were mapped to gene annotation. Next, we downloaded 450 shRNA-Seq experiment datasets for 225 diverse RBPs in HepG2 and 466 datasets for 233 RBPs in K562 cell line from ENCODE. Moreover, ten normal RNA-Seq datasets for HepG2 and K562 were also downloaded. Gene expression levels were measured by Fragments Per Kilobase of transcript per Million mapped reads (FPKM). Only genes with average FPKM greater than one in normal and RBP knockout datasets were used for further analysis. Our hypothesis is that if an RBP binds to a specific RNA, it may affect the expression of that gene. We thus identified the RBP-gene pairs that showed two-fold expression changes

after knocking out the specific RBP. By integration with eCLIP-Seq results, we identified the genes not only with RBP binding peaks but also showing expression changes as RBP target genes.

Clustering samples based on transcriptional and AS profiles

Median absolute deviation (MAD) was used to select 1,500 most variable genes. Consensus ward linkage hierarchical clustering of the samples and 1,500 genes identified the subtypes with the stability of the clustering increasing from k = 2 to k = 10. For each iteration, we selected 80% of the cancer samples and this process was repeated for 100 times. In addition, we identified the AS events in each cancer and clustered the samples based on PSI profile of these AS events. This process was performed by using the R package-'ConsensusClusterPlus' (Wilkerson and Hayes, 2010).

Survival analysis

The clinical information for 33 types of cancer were downloaded from TCGA website via the TCGAbiolinks R package (Colaprico et al., 2016). The function 'survdiff' in the 'survival' package was used for exploring the difference in survival time among different cancer subtypes.

DATA AND SOFTWARE AVAILABILITY

The identified mutation-AS pairs and detailed information for mutation and AS across 33 types of cancer can be downloaded from Tables S1-S6.

KEY RESOURCES TABLE

Deposited Data		
AS profiles	TCGASpliceSeq	http://bioinformatics.mdanderson.org/TCGASpliceSeq/
Somatic mutations	National Cancer Institute Genomic Data Commons	https://gdc.cancer.gov/
Cancer genes	Cancer Gene Census	http://cancer.sanger.ac.uk/census/
Protein-protein interactions	HI-II-14	http://dx.doi.org/10.1016/j.cell.2014.10.050
RNA-binding proteins	(Sebestyen et al., 2016)	http://genome.cshlp.org/content/early/2016/04/13/gr.199935.115
Cancer hallmarks	(Plaisier et al., 2012)	http://genome.cshlp.org/content/22/11/2302.full
Pathways	Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.jp/kegg/
eCLIP-Seq data	ENCODE	https://www.encodeproject.org/
shRNA-Seq data	ENCODE	https://www.encodeproject.org/

Software and Algorithms		
R	Comprehensive R Archive Network (CRAN)	https://cran.r-project.org/
Consensus clustering	ConsensusClusterPlus	https://www.bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html
Survival analysis	Survival	https://cran.r-project.org/web/packages/survival/index.html
ANNOVAR	ANNOVAR	http://annovar.openbioinformatics.org/
IUPred	IUPred	http://iupred.enzim.hu/
PyMOL	PyMOL v1.8.4.0	https://www.pymol.org/
GSEA	gsea2-3.0_beta_2.jar	http://software.broadinstitute.org/gsea/index.jsp

Figure S1.

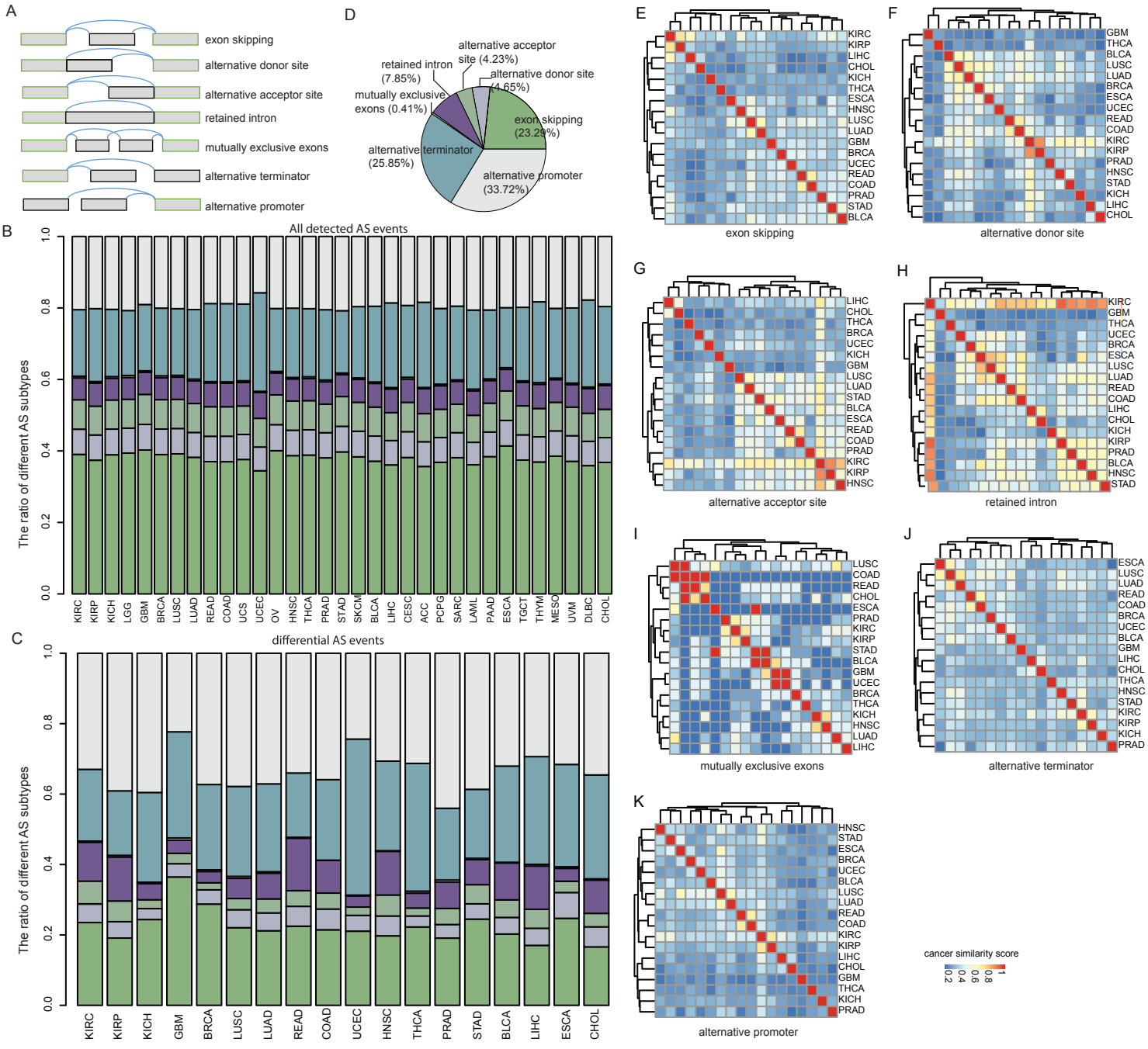

A
exon skipping
alternative donor site
alternative acceptor site
retained intron
mutually exclusive exons
alternative terminator
alternative promoter
D
alternative acceptor site (4.23%)
retained intron (7.85%)
alternative donor site (4.65%)
mutually exclusive exons (0.41%)
exon skipping (23.29%)
alternative terminator (25.85%)
alternative promoter (33.72%)
B
All detected AS events
The ratio of different AS subtypes
KIRC KIRP KICH LGG GBM BRCA LUSC LUAD READ COAD UCS UCEC OV HNSC THCA PRAD STAD SKCM BLCA LIHC CESC ACC PCPG SARC LAML PAAD ESCA TGCT THYM MESO UVM DLBC CHOL
C
differential AS events
The ratio of different AS subtypes
KIRC KIRP KICH GBM BRCA LUSC LUAD READ COAD UCEC HNSC THCA PRAD STAD BLCA LIHC ESCA CHOL
E
exon skipping
F
alternative donor site
G
alternative acceptor site
H
retained intron
I
mutually exclusive exons
J
alternative terminator
K
alternative promoter
cancer similarity score

Figure S1. Differential AS events across cancer types, related to Figure 2.

(A) Seven AS classes analyzed in this study.

(B) The ratio of different AS classes across 33 cancer types.

(C) The ratio of differential AS classes when compared to normal control samples in 18 cancer types.

(D) A pie chart shows the proportion of differential AS classes across 18 cancer types.

(E-K) Hierarchical clustering of cancer types based on the similarity of differential AS patterns for each of the seven AS classes. Cancer similarity is computed as the overlap divided by the minimum number of differential AS events between two cancer types.

Figure S2.

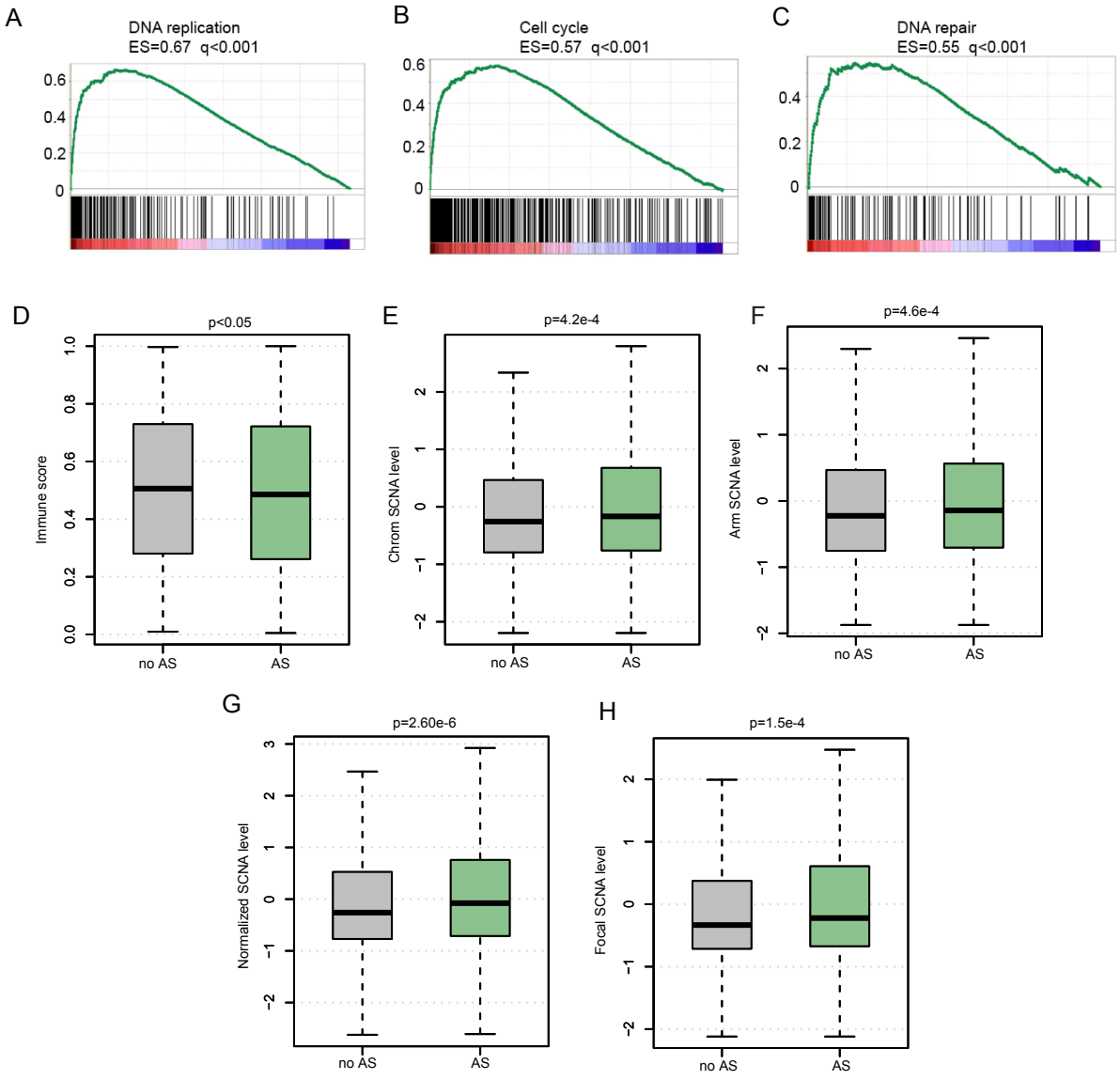

A
DNA replication
ES=0.67 q<0.001
B
Cell cycle
ES=0.57 q<0.001
C
DNA repair
ES=0.55 q<0.001
D
p<0.05
Immune score
no AS
AS
E
p=4.2e-4
Chrom SCNA level
F
p=4.6e-4
Arm SCNA level
G
p=2.60e-6
Normalized SCNA level
H
p=1.5e-4
Focal SCNA level

Figure S2. Enrichment of DNA replication and depletion of immune signature in tumors with AS perturbation, related to Figure 2.

(A-C) RNA sequencing analysis was performed comparing tumor samples with driver AS events versus samples without driver AS events. GSEA plots, enrichment scores (ES), and false discovery rates (FDR; q) are shown for representative gene sets (DNA replication, cell cycle and DNA repair) enriched in tumor samples with differential AS events.

(D) Distribution of immune scores for tumor samples with versus without differential AS events.

(E) Distribution of chromosomal SNCA levels for tumor samples with versus without differential AS events.

(F) Distribution of arm SNCA levels for tumor samples with versus without differential AS events.

(G) Distribution of normalized SCNA levels for tumor samples with versus without differential AS events.

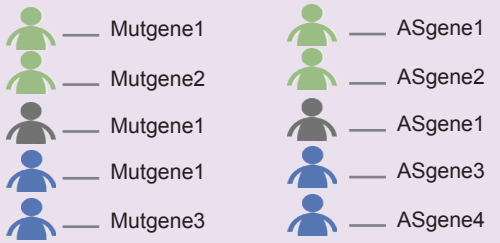
(H) Distribution of focal SCNA levels for tumor samples with versus without differential AS events. P-values (D-H) are computed by Wilcoxon rank-sum test.

Figure S3.

A

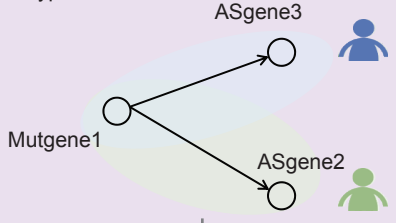
Flowchart

(1) Patient-specific mutations and perturbed alternative splicing events

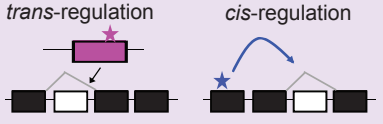


(2) Patient-specific mutation-alternative splicing pairs

(3) Integration of all mutation-AS pairs in a specific cancer type



(4) Driver mutations and alternative splicing events



Datasets and scripts

Download TCGA mutations from GDC (33 maf files)

(1) Construction of mutation and AS matrices

Mutation matrix

$$M(i,j) = \begin{cases} 1, if\ gene\ i\ mutated\ in\ patient\ j \\ 0, if\ gene\ i\ not\ mutated\ in\ patient\ j \end{cases}$$

Identify the patient specific AS perturbations in cancer

$$AS(i,j) = \begin{cases} 1, & if\ PSI(i,j) < Q1 - 1.5IQR \\ 1, & if\ PSI(i,j) > Q3 + 1.5IQR \\ 0, & others \end{cases}$$

$$IQR = Q3 - Q1$$

(Perturbed_AS_personalized.R)

(2) Identify the mutation-AS perturbation in specific cancer sample k

for k in 1:n

gene pair gi-gj:
(a) gi is mutated in patient k;
(b) gj exhibits differential AS in patient k;
(c) gi and gj interact according to known protein interaction and protein-RNA interaction networks

k<=n / *k>n*

(Mutation_AS.R)

(3) Assemble gene pairs across all cancer types

(4) Prioritize driver mutations by Greedy algorithm

Genes are ranked by the number of mis-spliced events they mediate: *G={g1,g2,g3,...,gi,...,gm}*

for i=1:m

Identify the AS perturbations mediated by gene i

Calculate the saturation index (*S*) of perturbed AS events:

$$S = \frac{Mutation\text{-}mediated\ AS}{All\ perturbed\ AS}$$

Saturation index *S*==1 or not

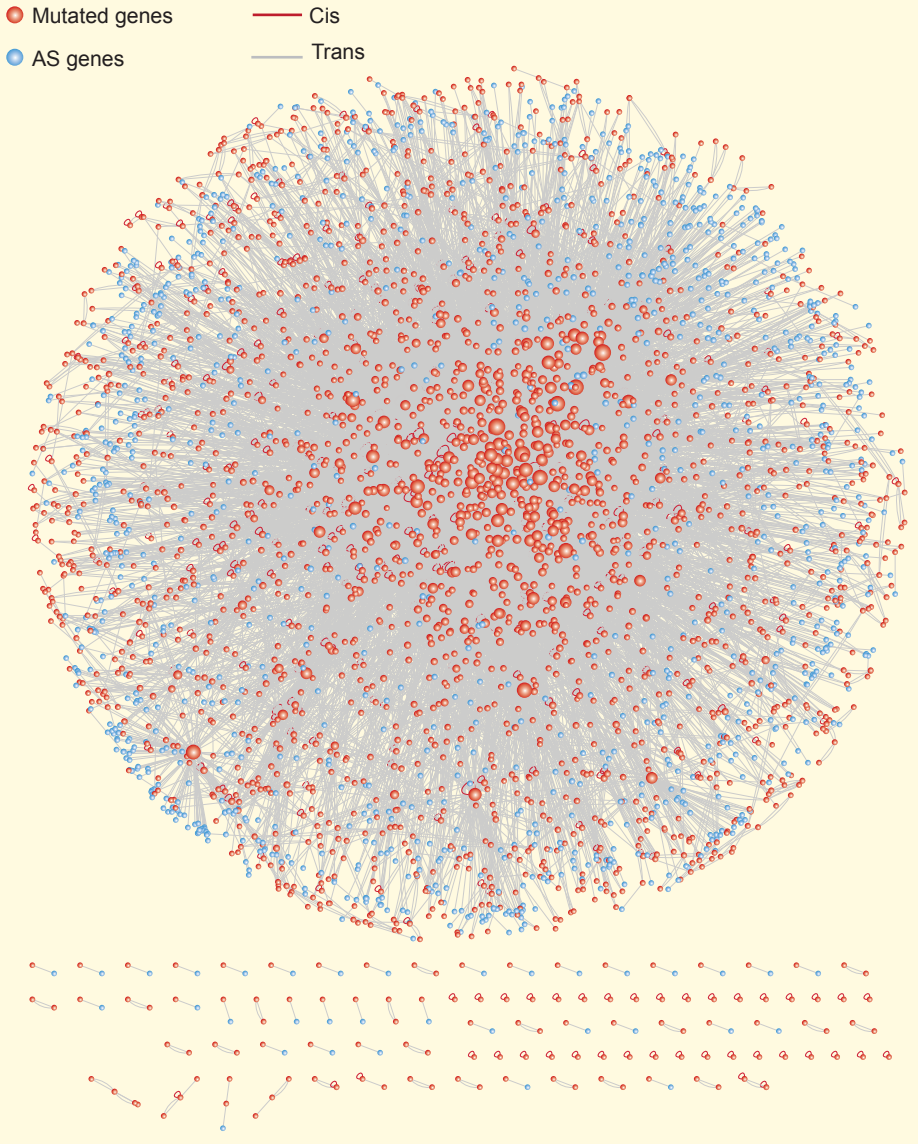
No / Yes

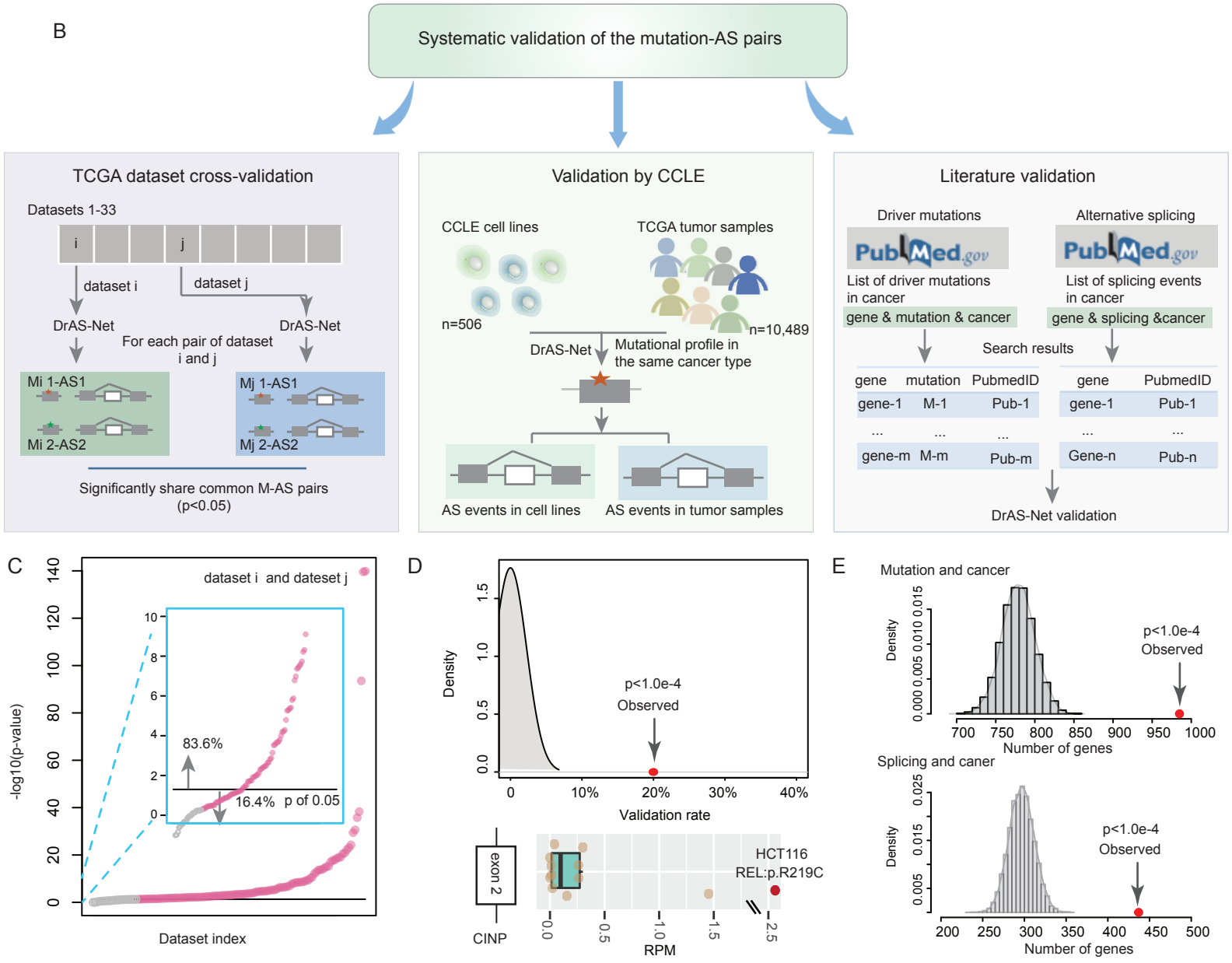
Add another gene

Driver mutations

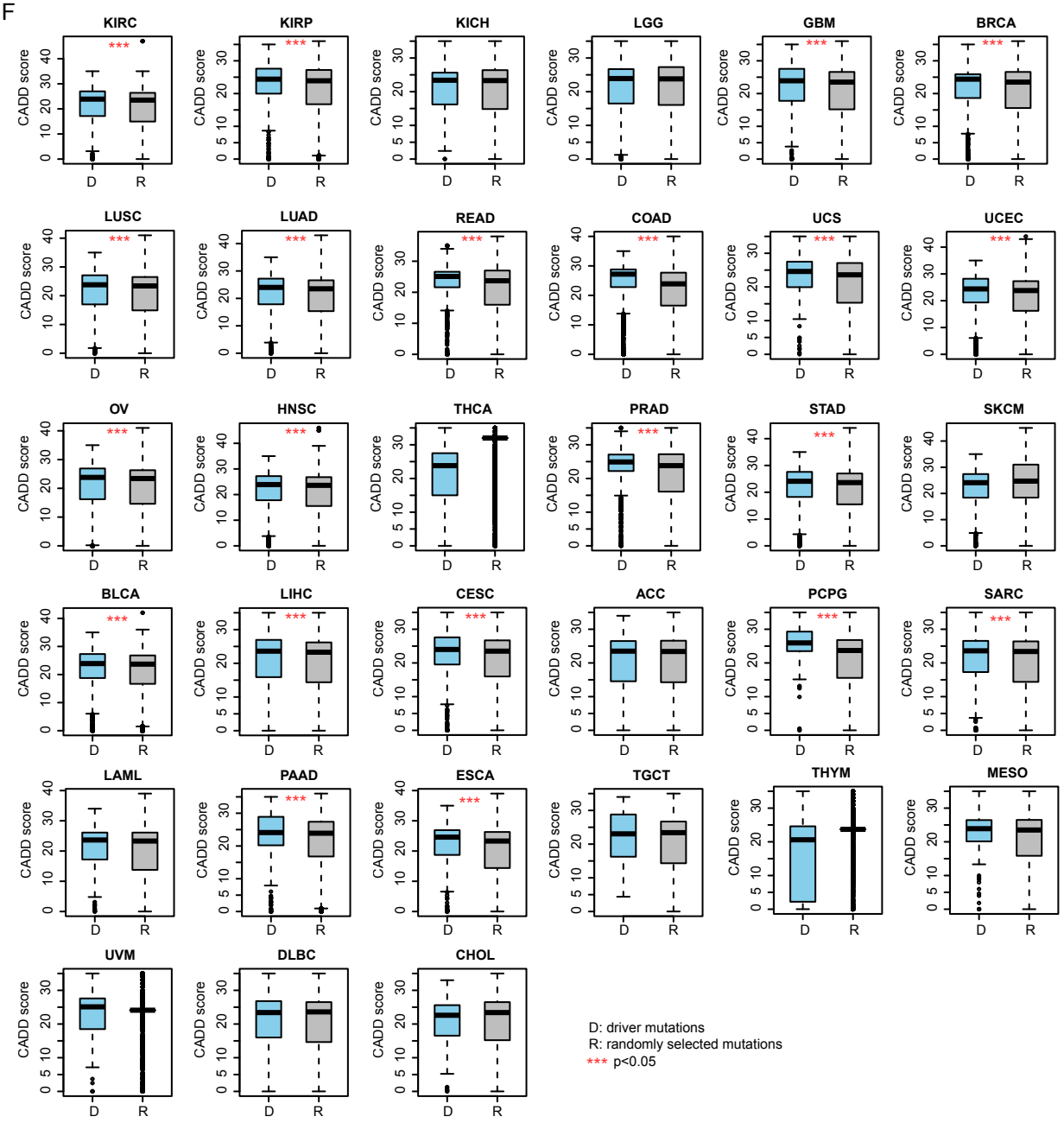
(Trans_pri.R)

DrAS-Net

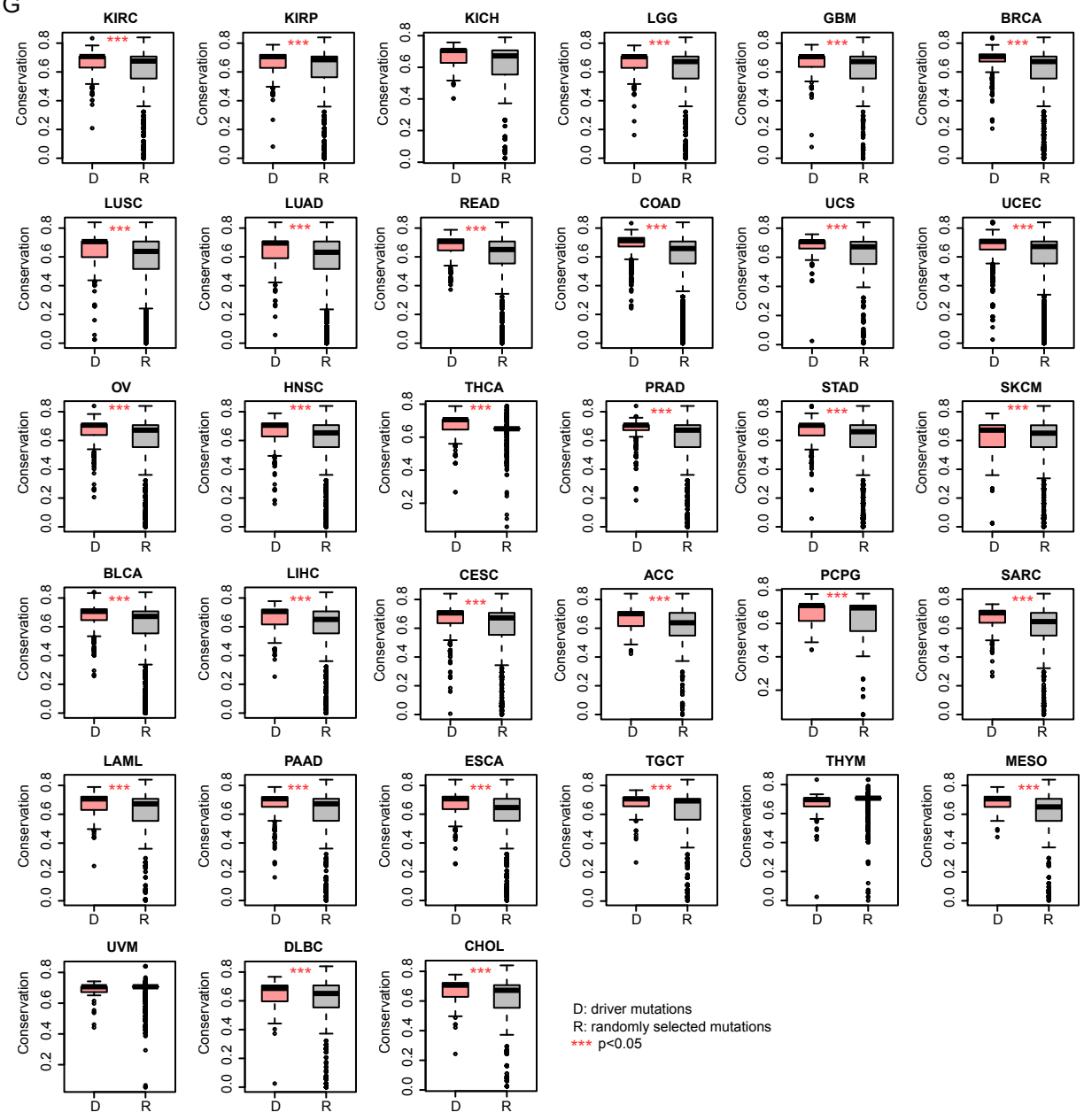



B
Systematic validation of the mutation-AS pairs
TCGA dataset cross-validation
Datasets 1-33
i
j
dataset i
dataset j
DrAS-Net
DrAS-Net
For each pair of dataset i and j
Mi 1-AS1
Mi 2-AS2
Mj 1-AS1
Mj 2-AS2
Significantly share common M-AS pairs (p<0.05)
Validation by CCLE
CCLE cell lines
TCGA tumor samples
n=506
n=10,489
DrAS-Net
Mutational profile in the same cancer type
AS events in cell lines
AS events in tumor samples
Literature validation
Driver mutations
Alternative splicing
PubMed.gov
PubMed.gov
List of driver mutations in cancer
List of splicing events in cancer
gene & mutation & cancer
gene & splicing &cancer
Search results
gene mutation PubmedID
gene-1 M-1 Pub-1
gene-m M-m Pub-m
gene PubmedID
gene-1 Pub-1
Gene-n Pub-n
DrAS-Net validation
C
dataset i and dateset j
-log10(p-value)
83.6%
16.4% p of 0.05
Dataset index
D
Density
p<1.0e-4
Observed
Validation rate
exon 2
HCT116
REL:p.R219C
CINP
RPM
E
Mutation and cancer
Density
p<1.0e-4
Observed
Number of genes
Splicing and caner
Density
p<1.0e-4
Observed
Number of genes

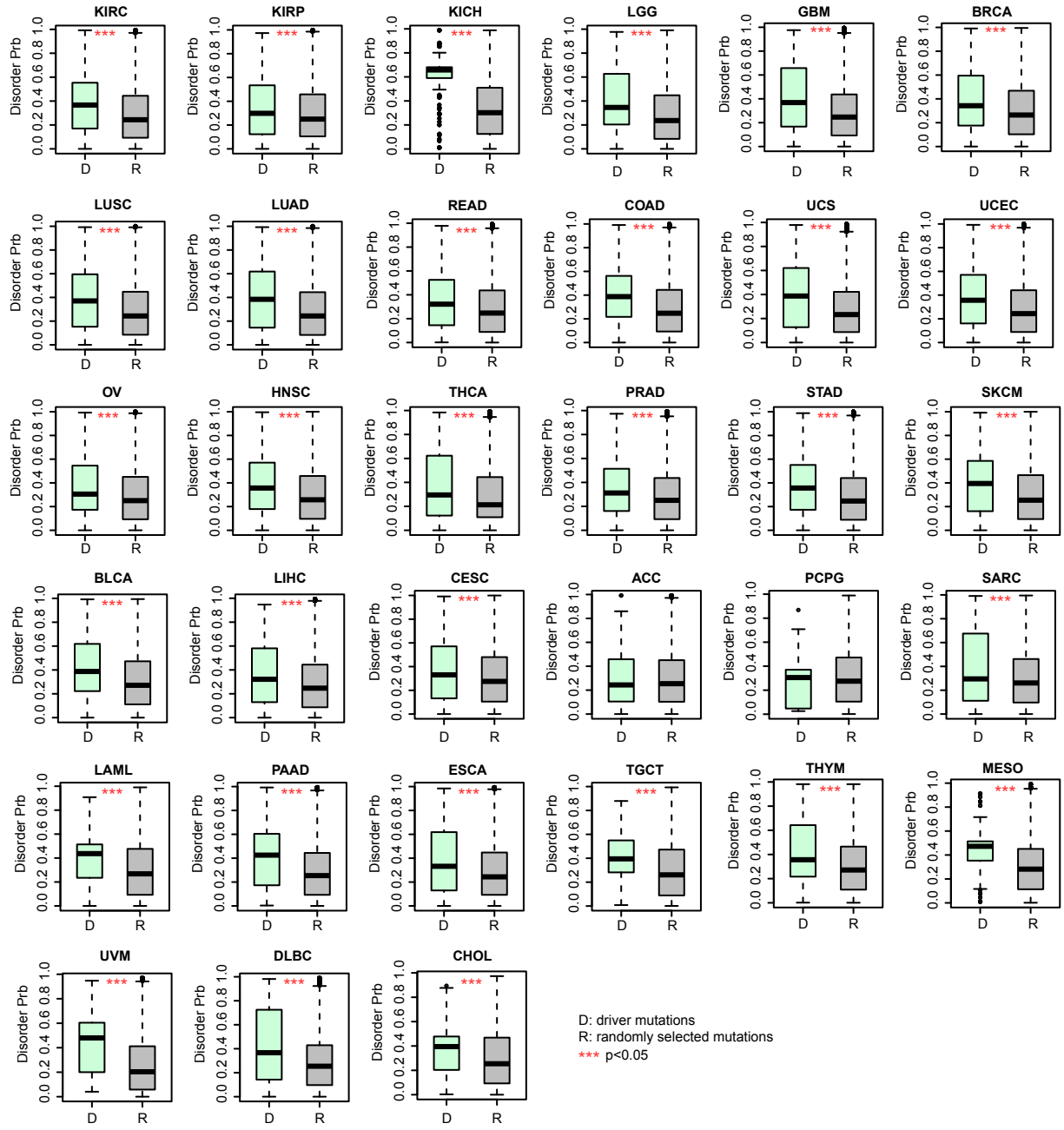
F


KIRC
KIRP
KICH
LGG
GBM
BRCA
LUSC
LUAD
READ
COAD
UCS
UCEC
OV
HNSC
THCA
PRAD
STAD
SKCM
BLCA
LIHC
CESC
ACC
PCPG
SARC
LAML
PAAD
ESCA
TGCT
THYM
MESO
UVM
DLBC
CHOL
CADD score
D
R
***
D: driver mutations
R: randomly selected mutations
*** p<0.05

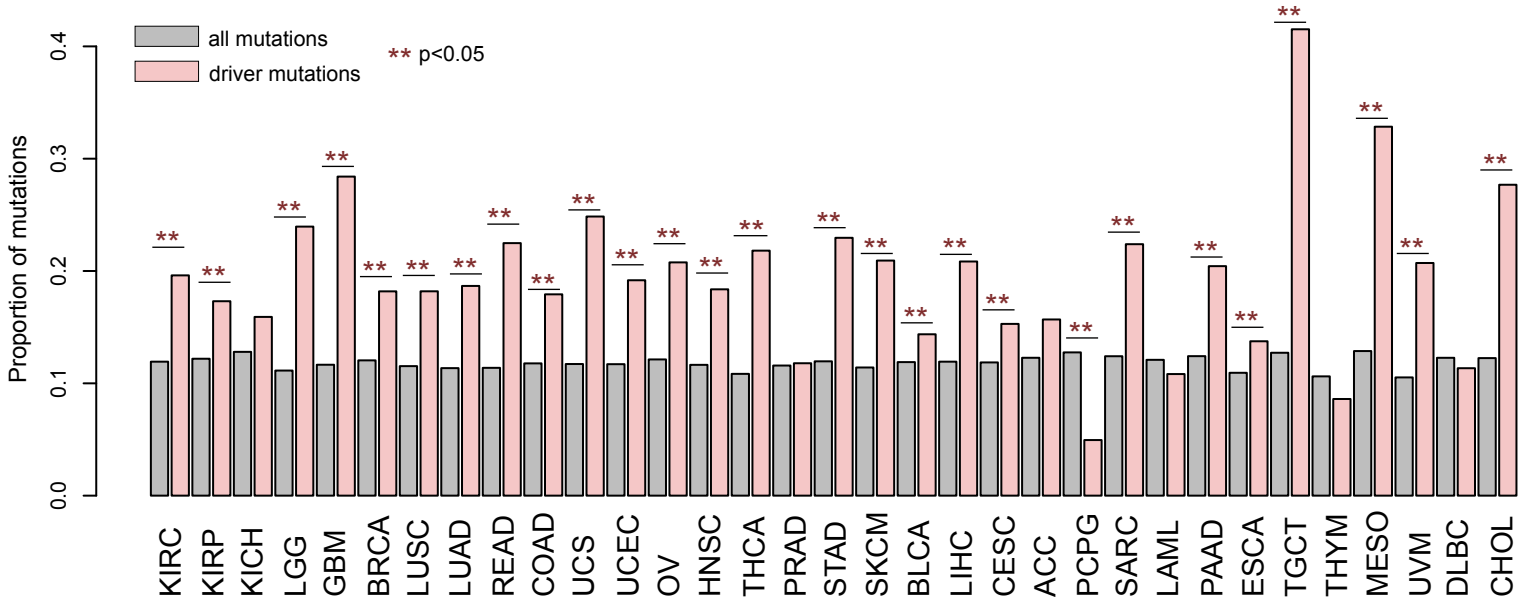
G


KIRC
KIRP
KICH
LGG
GBM
BRCA
LUSC
LUAD
READ
COAD
UCS
UCEC
OV
HNSC
THCA
PRAD
STAD
SKCM
BLCA
LIHC
CESC
ACC
PCPG
SARC
LAML
PAAD
ESCA
TGCT
THYM
MESO
UVM
DLBC
CHOL
Conservation
D
R
***
D: driver mutations
R: randomly selected mutations
*** p<0.05

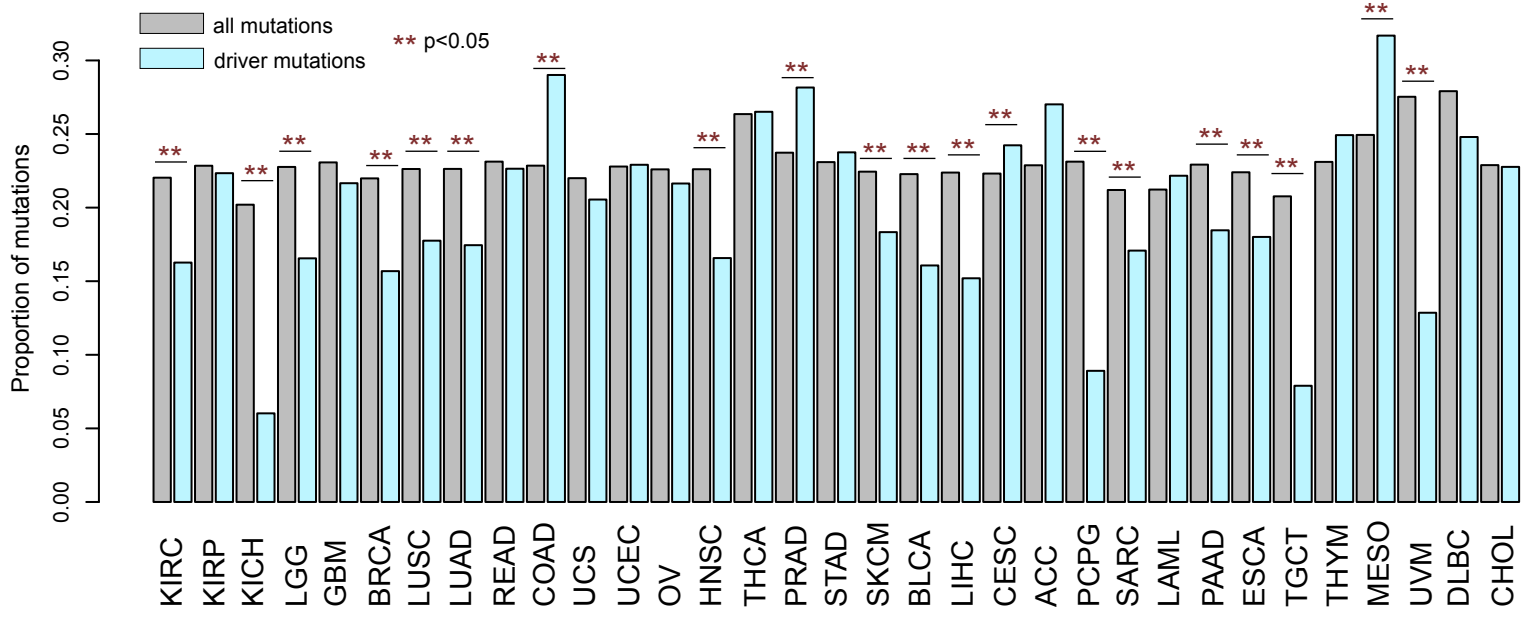
H


KIRC
KIRP
KICH
LGG
GBM
BRCA
LUSC
LUAD
READ
COAD
UCS
UCEC
OV
HNSC
THCA
PRAD
STAD
SKCM
BLCA
LIHC
CESC
ACC
PCPG
SARC
LAML
PAAD
ESCA
TGCT
THYM
MESO
UVM
DLBC
CHOL
Disorder Prb
0.0 0.2 0.4 0.6 0.8 1.0
D
R
***
D: driver mutations
R: randomly selected mutations
*** p<0.05

I ANCHOR analysis on disordered regions



J Protein Pfam domain analysis



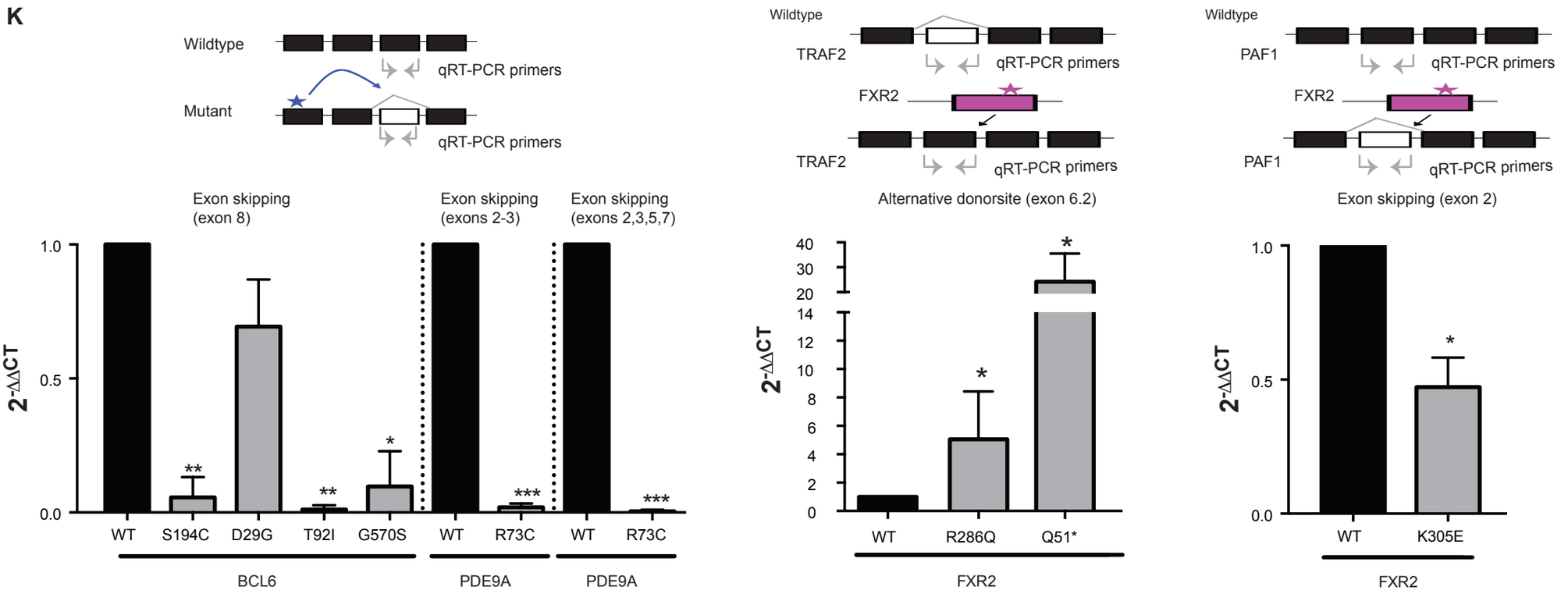

K
Wildtype
qRT-PCR primers
Mutant
qRT-PCR primers
Exon skipping (exon 8)
Exon skipping (exons 2-3)
Exon skipping (exons 2,3,5,7)
2-ΔΔCT
1.0
0.5
0.0
**
**
*
***
***
WT
S194C
D29G
T92I
G570S
WT
R73C
WT
R73C
BCL6
PDE9A
PDE9A
Wildtype
TRAF2
qRT-PCR primers
FXR2
TRAF2
qRT-PCR primers
Alternative donorsite (exon 6.2)
2-ΔΔCT
40
30
20
14
12
10
8
6
4
2
0
*
*
WT
R286Q
Q51*
FXR2
Wildtype
PAF1
qRT-PCR primers
FXR2
PAF1
qRT-PCR primers
Exon skipping (exon 2)
2-ΔΔCT
1.0
0.5
0.0
*
WT
K305E
FXR2

Figure S3. The distribution of functional impact scores of candidate driver mutations and random controls, related to Figure 3.

(A) The work-flow of DrAS-Net. The left part shows the flowchart of the method, the middle panel shows the details of the scripts, and the right panel shows the mutation-AS network across 33 cancer types.

(B) Validation of the inferences from DrAS-Net. The left part is based on 33 dataset cross-validations of TCGA, the middle panel is validation based on independent cell line data, and the right panel shows the validation based on literature.

(C) The distribution of p-values of dataset cross-validations and the proportion of validated dataset pairs.

(D) The validation rates of cell line dataset and random conditions.

(E) The number of genes with literature support for mutated genes and genes with perturbed splicing.

(F) The CADD scores of candidate driver mutations and random controls across 33 cancer types.

(G) The conservation scores of candidate driver mutations and random controls across 33 cancer types.

(H) The disorder probability of candidate driver mutations and random controls across 33 cancer types. Prb, probability.

(I) Proportion of mutations in anchor sites. Pink, driver mutations; Gray, all mutations.

(J) Proportion of mutations in Pfam domains. Light blue, driver mutations; Gray, all mutations.

(K) Experimentally validated mutation-AS pairs. The upper panels show the cartoon of mutation-AS relationships, while the bottom panels show the expression of corresponding exons related to splicing.

D: Driver mutations; R: Randomly selected mutations. P values by one-sided Wilcoxon rank-sum test. *** $p<0.05$.

Figure S4.

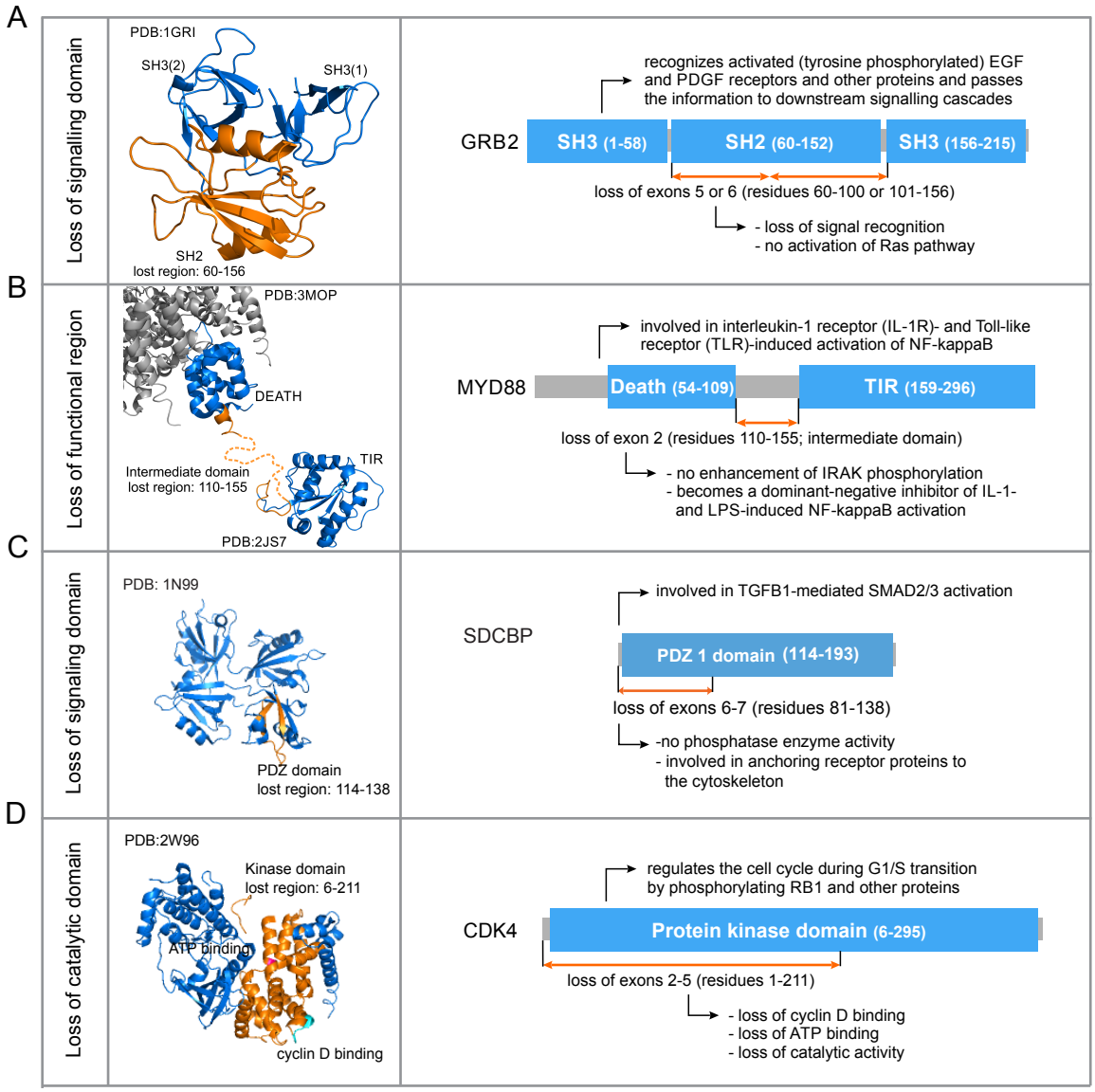
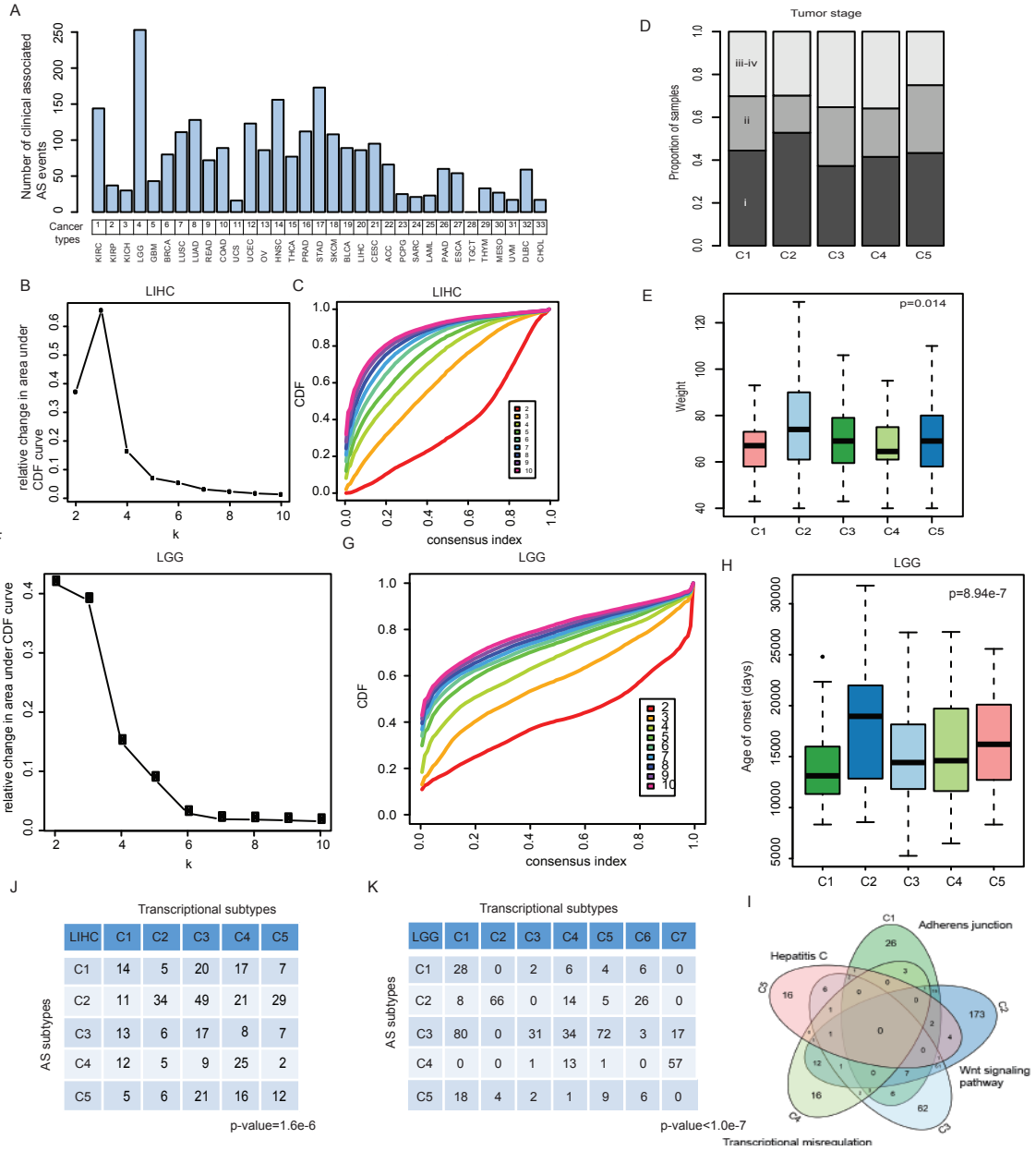

A
Loss of signalling domain
PDB:1GRI
SH3(2)
SH3(1)
SH2
lost region: 60-156
recognizes activated (tyrosine phosphorylated) EGF and PDGF receptors and other proteins and passes the information to downstream signalling cascades
GRB2
SH3 (1-58)
SH2 (60-152)
SH3 (156-215)
loss of exons 5 or 6 (residues 60-100 or 101-156)
- loss of signal recognition
- no activation of Ras pathway
B
Loss of functional region
PDB:3MOP
DEATH
TIR
Intermediate domain
lost region: 110-155
PDB:2JS7
involved in interleukin-1 receptor (IL-1R)- and Toll-like receptor (TLR)-induced activation of NF-kappaB
MYD88
Death (54-109)
TIR (159-296)
loss of exon 2 (residues 110-155; intermediate domain)
- no enhancement of IRAK phosphorylation
- becomes a dominant-negative inhibitor of IL-1- and LPS-induced NF-kappaB activation
C
Loss of signaling domain
PDB: 1N99
PDZ domain
lost region: 114-138
involved in TGFB1-mediated SMAD2/3 activation
SDCBP
PDZ 1 domain (114-193)
loss of exons 6-7 (residues 81-138)
-no phosphatase enzyme activity
- involved in anchoring receptor proteins to the cytoskeleton
D
Loss of catalytic domain
PDB:2W96
Kinase domain
lost region: 6-211
ATP binding
cyclin D binding
regulates the cell cycle during G1/S transition by phosphorylating RB1 and other proteins
CDK4
Protein kinase domain (6-295)
loss of exons 2-5 (residues 1-211)
- loss of cyclin D binding
- loss of ATP binding
- loss of catalytic activity

Figure S4. Functional consequences of losing crucial protein modules due to differential AS events in cancer, related to Figure 4.

The four panels introduce examples for the loss of different types of protein modules –signal transduction domains (A, C), regions with no known folds or interaction motifs (B), and catalytic domains (D) – due to differential AS events observed in samples of different cancer types. The left panels show a corresponding PDB structure for each protein. In the structures, the investigated protein is blue with the lost region highlighted in orange, while interaction partners are marked with gray. In Panel B, the indicated structure is combined from two different PDB entries. The lost region corresponding to the MYD88 intermediate domain is largely missing from both structures and thus it is indicated as a dashed line. In Panel D, ATP binding site is marked in pink, and the cyclin D binding site is marked in cyan. In the right panels, proteins are represented by gray bars with their annotated protein modules from UniProt indicated as blue boxes; For folded domains, the residue boundaries are also shown. Above the proteins a short functional description is provided, while below the proteins the regions corresponding to the differentially spliced exon(s) are marked with orange arrows. The functional consequences of the loss of these regions are also highlighted below.

Figure S5.



LIHC	C1	C2	C3	C4	C5
C1	14	5	20	17	7
C2	11	34	49	21	29
C3	13	6	17	8	7
C4	12	5	9	25	2
C5	5	6	21	16	12

LGG	C1	C2	C3	C4	C5	C6	C7
C1	28	0	2	6	4	6	0
C2	8	66	0	14	5	26	0
C3	80	0	31	34	72	3	17
C4	0	0	1	13	1	0	57
C5	18	4	2	1	9	6	0

Figure S5. Clinical features of AS in representative cancer types, related to Figure 5.

(A) The number of clinical associated AS events in each cancer type.

(B) The Cumulative Distribution Function (CDF) distribution for different number of clusters in LIHC (k=2 to 10).

(C) The relative changes in area under CDF curve with different k in LIHC (k=2 to 10).

(D) Distribution of tumor stage across different cancer subtypes.

(E) Distribution of weight across different cancer subtypes. Statistical difference is calculated by Kruskal-Wallis rank sum test (p=0.014).

(F) The Cumulative Distribution Function (CDF) distribution for different number of clusters in LGG (k=2 to 10).

(G) The relative changes in area under CDF curve with different k in LGG (k=2 to 10).

(H) Distribution of age of onset for different LGG subtypes. Statistical difference is calculated by Kruskal-Wallis rank sum test (p=8.94e-7).

(I) Overlap of mutated genes that mediate AS events in four subtypes. The common top enriched functional terms by genes are marked.

(J) The overlap of samples for different AS subtype and transcriptional subtype in LIHC.

(K) The overlap of samples for different AS subtype and transcriptional subtype in LGG.

References

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D.*, et al.* (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603-607.

Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I.*, et al.* (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic acids research *44*, e71.

Davoli, T., Uno, H., Wooten, E.C., and Elledge, S.J. (2017). Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science *355*.

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433-3434.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics *46*, 310-315.

Latysheva, N.S., Oates, M.E., Maddox, L., Flock, T., Gough, J., Buljan, M., Weatheritt, R.J., and Babu, M.M. (2016). Molecular Principles of Gene Fusion Mediated Rewiring of Protein Interaction Networks in Cancer. Molecular cell *63*, 579-592.

Meszaros, B., Simon, I., and Dosztanyi, Z. (2009). Prediction of protein binding regions in disordered proteins. PLoS computational biology *5*, e1000376.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E.*, et al.* (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics *34*, 267-273.

Plaisier, C.L., Pan, M., and Baliga, N.S. (2012). A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers. Genome research *22*, 2302-2314. Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current protocols in bioinformatics *47*, 11 12 11-34.

Ryan, M., Wong, W.C., Brown, R., Akbani, R., Su, X., Broom, B., Melott, J., and Weinstein, J. (2016). TCGASpliceSeq a compendium of alternative mRNA splicing in cancer. Nucleic acids research *44*, D1018-1022.

Sebestyen, E., Singh, B., Minana, B., Pages, A., Mateo, F., Pujana, M.A., Valcarcel, J., and Eyras, E. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome research *26*, 732-744.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545-15550.

The UniProt, C. (2017). UniProt: the universal protein knowledgebase. Nucleic acids research *45*, D158-D169.

Tompa, P., Davey, N.E., Gibson, T.J., and Babu, M.M. (2014). A million peptide motifs for the molecular biologist. Molecular cell *55*, 161-169.

Watson, I.R., Takahashi, K., Futreal, P.A., and Chin, L. (2013). Emerging patterns of somatic mutations in cancer. Nature reviews. Genetics *14*, 703-718.
Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572-1573.