Molecular Cell, Volume *81*

Supplemental information

A pan-cancer transcriptome analysis of exitron splicing identifies novel cancer driver genes and neoepitopes

Ting-You Wang, Qi Liu, Yanan Ren, Sk. Kayum Alam, Li Wang, Zhu Zhu, Luke H. Hoeppner, Scott M. Dehm, Qi Cao, and Rendong Yang

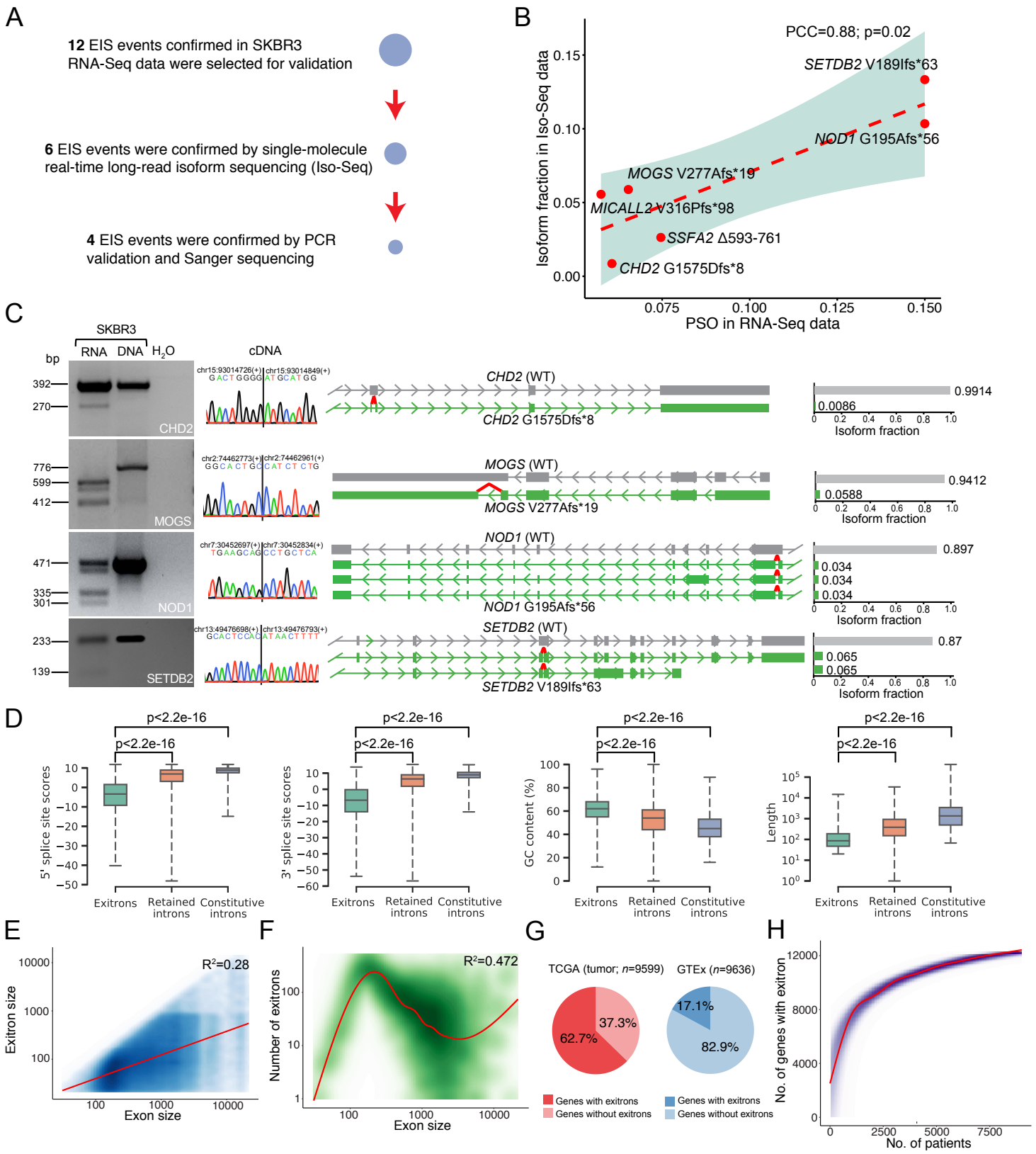
A
12 EIS events confirmed in SKBR3 RNA-Seq data were selected for validation
6 EIS events were confirmed by single-molecule real-time long-read isoform sequencing (Iso-Seq)
4 EIS events were confirmed by PCR validation and Sanger sequencing
B
PCC=0.88; p=0.02
SETDB2 V189Ifs*63
NOD1 G195Afs*56
MOGS V277Afs*19
MICALL2 V316Pfs*98
SSFA2 Δ593-761
CHD2 G1575Dfs*8
Isoform fraction in Iso-Seq data
PSO in RNA-Seq data
C
SKBR3
RNA DNA H2O
bp
cDNA
CHD2
MOGS
NOD1
SETDB2
CHD2 (WT)
CHD2 G1575Dfs*8
MOGS (WT)
MOGS V277Afs*19
NOD1 (WT)
NOD1 G195Afs*56
SETDB2 (WT)
SETDB2 V189Ifs*63
Isoform fraction
D
p<2.2e-16
5' splice site scores
3' splice site scores
GC content (%)
Length
Exitrons
Retained introns
Constitutive introns
E
R²=0.28
Exitron size
Exon size
F
R²=0.472
Number of exitrons
Exon size
G
TCGA (tumor; n=9599)
GTEx (n=9636)
37.3%
62.7%
17.1%
82.9%
Genes with exitrons
Genes without exitrons
H
No. of genes with exitron
No. of patients

Figure S1. Exitron splicing (EIS) events detection in cancer, Related to Figure 1.
(A) Information flow during the validation of EIS events identified in TCGA-BRCA.
(B) The fractions of exitron-spliced isoforms from PacBio long read Iso-Seq data are highly correlated with their PSO values predicted from the RNA-Seq data. Pearson's correlation coefficient (PCC) was used to determine the correlation.
(C) Four EIS events (*CHD2* G1575Dfs*8, *MOGS* V277Afs*19, *NOD1* G195Afs*56 and *SETDB2* V189Ifs*63) were validated by RT-PCR and Sanger sequencing. Isoform fraction of wild-type (grey color) and exitron-spliced transcripts (green color) are counted according to long read Iso-Seq.
(D) Comparisons of 5' and 3' splice site strength (measured using maximum entropy; see Methods), GC content and size distribution of exitrons, retained introns and constitutively spliced introns. All p values were calculated by Mann-Whitney U test.
(E) Exitron size and exitron-containing exon size are correlated (R-Squared=0.28). Exitron size ranges from 21bp to 14,765bp and exon size ranges 31bp to 21,693bp. Both exon size and exitron size are shown on log10 scale. All the points were fitted by a linear regression red line.
(F) Distribution of 129,406 exitrons in TCGA based on their locations in exon of different sizes. Medium-sized exons contain the largest number of exitrons. All the points were fitted by a generalized additive model.
(G) The proportion of genes with and without exitrons in TCGA tumor and GTEx healthy samples.
(H) Saturation analysis of exitron splicing events in TCGA by adding more samples. Each point is a random subset of samples. All the points were fitted by a smoothed red line.

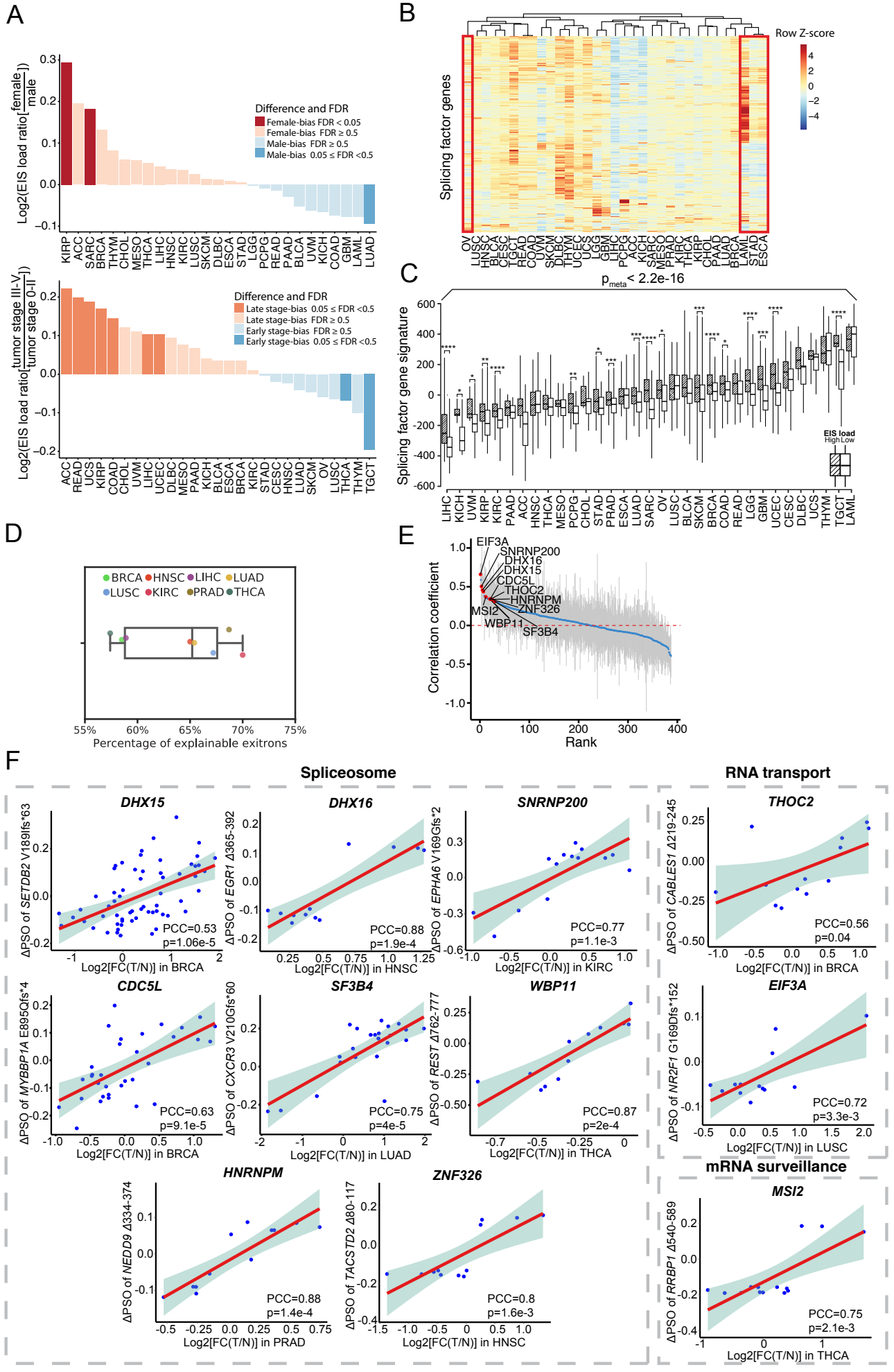
A
Log2(EIS load ratio[female | male])
Difference and FDR
Female-bias FDR < 0.05
Female-bias FDR ≥ 0.5
Male-bias FDR ≥ 0.5
Male-bias 0.05 ≤ FDR <0.5
KIRP ACC SARC BRCA THYM CHOL MESO THCA LIHC HNSC KIRC LUSC SKCM DLBC ESCA STAD LGG PCPG READ PAAD BLCA UVM KICH COAD GBM LAML LUAD
Log2(EIS load ratio[tumor stage III-V | tumor stage 0-II])
Difference and FDR
Late stage-bias 0.05 ≤ FDR <0.5
Late stage-bias FDR ≥ 0.5
Early stage-bias FDR ≥ 0.5
Early stage-bias 0.05 ≤ FDR <0.5
ACC READ UCS KIRP COAD CHOL UVM LIHC UCEC DLBC MESO PAAD KICH BLCA ESCA BRCA KIRC STAD CESC HNSC LUAD SKCM OV LUSC THCA THYM TGCT
B
Row Z-score
Splicing factor genes
OV LUSC HNSC BLCA CESC TGCT READ COAD UVM SKCM UCEC UCS LGG GBM LIHC PCPG ACC KICH SARC MESO PRAD KIRC THCA KIRP CHOL PAAD LUAD BRCA LAML STAD ESCA
C
p_meta < 2.2e-16
Splicing factor gene signature
EIS load
High Low
LIHC KICH UVM KIRP KIRC PAAD ACC HNSC THCA MESO PCPG CHOL STAD PRAD ESCA LUAD SARC OV LUSC BLCA SKCM BRCA COAD READ LGG GBM UCEC CESC DLBC UCS THYM TGCT LAML
D
BRCA HNSC LIHC LUAD
LUSC KIRC PRAD THCA
55% 60% 65% 70% 75%
Percentage of explainable exitrons
E
EIF3A
SNRNP200
DHX16
DHX15
CDC5L
THOC2
HNRNPM
ZNF326
MSI2
WBP11
SF3B4
Correlation coefficient
Rank
F
Spliceosome
RNA transport
mRNA surveillance
DHX15
ΔPSO of SETDB2 V189Ifs*63
Log2[FC(T/N)] in BRCA
PCC=0.53
p=1.06e-5
DHX16
ΔPSO of EGR1 Δ365-392
Log2[FC(T/N)] in HNSC
PCC=0.88
p=1.9e-4
SNRNP200
ΔPSO of EPHA6 V169Gfs*2
Log2[FC(T/N)] in KIRC
PCC=0.77
p=1.1e-3
THOC2
ΔPSO of CABLES1 Δ219-245
Log2[FC(T/N)] in BRCA
PCC=0.56
p=0.04
CDC5L
ΔPSO of MYBBP1A E895Qfs*4
Log2[FC(T/N)] in BRCA
PCC=0.63
p=9.1e-5
SF3B4
ΔPSO of CXCR3 V210GGfs*60
Log2[FC(T/N)] in LUAD
PCC=0.75
p=4e-5
WBP11
ΔPSO of REST Δ762-777
Log2[FC(T/N)] in THCA
PCC=0.87
p=2e-4
EIF3A
ΔPSO of NR2F1 G169Dfs*152
Log2[FC(T/N)] in LUSC
PCC=0.72
p=3.3e-3
HNRNPM
ΔPSO of NEDD9 Δ334-374
Log2[FC(T/N)] in PRAD
PCC=0.88
p=1.4e-4
ZNF326
ΔPSO of TACSTD2 Δ80-117
Log2[FC(T/N)] in HNSC
PCC=0.8
p=1.6e-3
MSI2
ΔPSO of RRBP1 Δ540-589
Log2[FC(T/N)] in THCA
PCC=0.75
p=2.1e-3

Figure S2. EIS dysregulation related to gender, tumor stage and splicing factors, Related to Figure 2.
(A) (top panel) Association between exitron splicing load and gender difference. (bottom panel) Association between exitron splicing load and tumor stage difference.
(B) Cancer types were clustered by the median expression of 404 splicing factor genes.
(C) The comparison of gene expression signature (measured by Z-score) of splicing factor genes on the top (high) vs. bottom (low) quartile patients by exitron splicing (EIS) load across 33 cancer types. p-values were calculated by Mann-Whitney U test with Bonferroni correction. (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$).
(D) The percentage of exitron splicing events in each cancer type that can be explained by a Generalized Additive Model (GAM) of differential expression of splicing factor between tumor and normal samples.
(E) Splicing factor genes ranked by Pearson correlation coefficient of GAM-derived exitron/splicing-factor pairs. Grey bar indicates the 95% confidence interval. Genes of interest are highlighted.
(F) Linear regression indicates a positive correlation of exitron splicing changes (ΔPSO) and differential mRNA expression (Log2(FC)) of corresponding splicing factors involving spliceosome, RNA transport and mRNA surveillance according to KEGG annotation. FC, Fold change; T, tumor; N, normal.

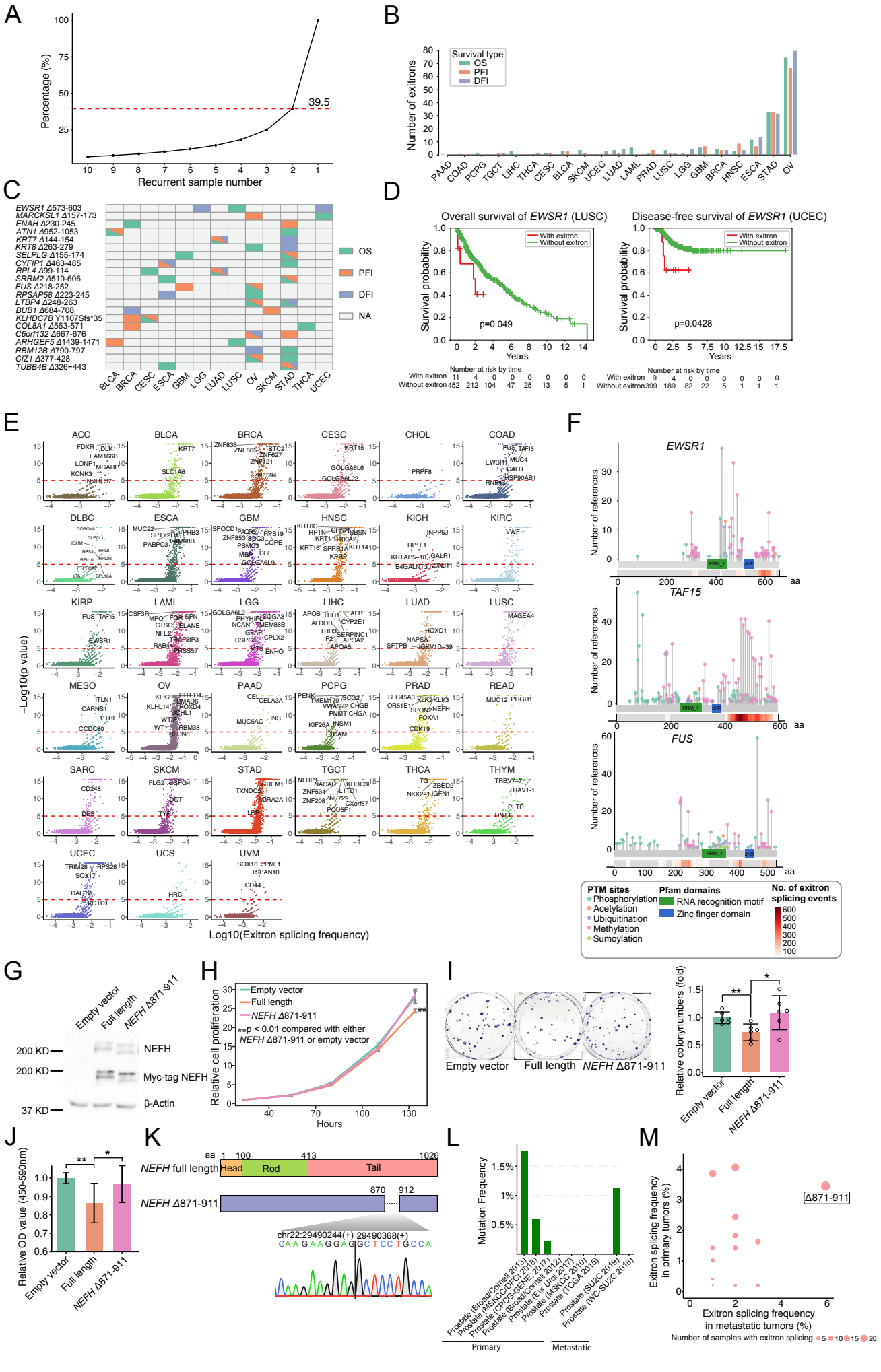
A
Percentage (%)
Recurrent sample number
39.5
B
Number of exitrons
Survival type
OS
PFI
DFI
PAAD COAD PCPG TGCT LIHC THCA CESC BLCA SKCM UCEC LUAD LAML PRAD LUSC LGG GBM BRCA HNSC ESCA STAD OV
C
EWSR1 Δ573-603
MARCKSL1 Δ157-173
ENAH Δ230-245
ATN1 Δ952-1053
KRT7 Δ144-154
KRT8 Δ263-279
SELPLG Δ155-174
CYFIP1 Δ463-485
RPL4 Δ99-114
SRRM2 Δ519-606
FUS Δ218-252
RPSAP58 Δ223-245
LTBP4 Δ248-263
BUB1 Δ684-708
KLHDC7B Y1107Sfs*35
COL8A1 Δ563-571
C6orf132 Δ667-676
ARHGEF5 Δ1439-1471
RBM12B Δ790-797
CIZ1 Δ377-428
TUBB4B Δ326-443
OS
PFI
DFI
NA
BLCA BRCA CESC ESCA GBM LGG LUAD LUSC OV SKCM STAD THCA UCEC
D
Overall survival of EWSR1 (LUSC)
Disease-free survival of EWSR1 (UCEC)
With exitron
Without exitron
Survival probability
p=0.049
p=0.0428
Years
Number at risk by time
With exitron 11 4 0 0 0 0 0 0
Without exitron 452 212 104 47 25 13 5 1
With exitron 9 4 0 0 0 0 0 0
Without exitron 399 189 82 22 5 1 1 1
E
ACC BLCA BRCA CESC CHOL COAD
DLBC ESCA GBM HNSC KICH KIRC
KIRP LAML LGG LIHC LUAD LUSC
MESO OV PAAD PCPG PRAD READ
SARC SKCM STAD TGCT THCA THYM
UCEC UCS UVM
−Log10(p value)
Log10(Exitron splicing frequency)
F
EWSR1
TAF15
FUS
Number of references
aa
PTM sites
Phosphorylation
Acetylation
Ubiquitination
Methylation
Sumoylation
Pfam domains
RNA recognition motif
Zinc finger domain
No. of exitron splicing events
600
500
400
300
200
100
G
Empty vector
Full length
NEFH Δ871-911
200 KD
NEFH
Myc-tag NEFH
37 KD
β-Actin
H
Empty vector
Full length
NEFH Δ871-911
**p < 0.01 compared with either NEFH Δ871-911 or empty vector
Relative cell proliferation
Hours
I
Empty vector
Full length
NEFH Δ871-911
Relative colonynumbers (fold)
J
Relative OD value (450-590nm)
Empty vector
Full length
NEFH Δ871-911
K
NEFH full length
Head
Rod
Tail
NEFH Δ871-911
aa 1 100 413 1026
870 912
chr22:29490244(+) 29490368(+)
CAA GAA GGA GGC T CC T GCC A
L
Mutation Frequency
Prostate (Broad/Cornell 2013)
Prostate (MSKCC/DFCI 2016)
Prostate (SPCG-GENE, 2017)
Prostate (Broad/Cornell 2012)
Prostate (Eur Urol 2017)
Prostate (MSKCC 2010)
Prostate (TCGA 2015)
Prostate (SU2C 2019)
Prostate (WC-SU2C 2018)
Primary
Metastatic
M
Exitron splicing frequency in primary tumors (%)
Exitron splicing frequency in metastatic tumors (%)
Δ871-911
Number of samples with exitron splicing 5 10 15 20

Figure S3. Clinical and functional impacts of tumor-specific exitrons, Related to Figure 3.

(A) The percentage of recurrently spliced TSEs at varied sample numbers.

(B) The distributions of clinically informative TSE splicing events in different cancer types. OS stands for overall survival, PFI stands for progression-free interval, and DFI stands for disease-free interval.

(C) The clinical relevance of fourteen clinically informative TSE splicing events were identified in multiple cancer types. The colored box indicates the association with the different survival types.

(D) Kaplan-Meier curves demonstrate clinical relevance of clinically informative TSE splicing event at *EWSR1* Δ573-603 in LUSC and UCEC. The p value was calculated by a log-rank test.

(E) SEGs identified across TCGA cohorts. For COAD and KIRP, SEGs overlapping with the COSMIC cancer census set are highlighted. For the remaining cancer cohorts, tissue-specific SEGs are highlighted. Coloring is according to cancer type.

(F) Exitron splicing hotspots in FET family proteins are enriched in post-translational modification (PTM) sites. Any position with a PTM site obtains a circle. circles are colored with respect to the corresponding PTM types, and the height of the line depends on the number of references supporting the PTM site at the specified position. The grey bar represents the entire protein with the different amino acid positions (aa). The colored boxes are specific Pfam domains.

(G) Western blotting detected overexpression of Myc-tagged wild-type and exitron-spliced NEFH in PC-3 cells.

(H) Cell proliferation is inhibited by overexpressing wild-type NEFH in PC-3 cells.

(I) Colony formation assays in PC-3 cells with overexpression of wild-type and exitron-spliced NEFH. The number of colonies were quantified and shown at the right panel.

(J) BrdU ELISA assay of PC-3 cells overexpressing wild-type and exitron-spliced NEFH with 24 hrs of BrdU label (n = 10). Y axis, absorbance of 450-590 nm relative to empty vector. There was less incorporation of BrdU in cells expressing wild-type NEFH. All p-values are calculated using unpaired, two-tailed Student's t-test. Error bars indicate ± s.d. *$p < 0.05$, **$p < 0.01$.

(K) Graphical depiction of NEFH protein domains annotated by Uniprot and the location of NEFH Δ871-911 exitron. Electropherogram illustrates the splicing junction of NEFH Δ871-911. aa, amino acid.

(L) Mutation frequency of NEFH among prostate cancer studies from cBioPortal (https://www.cbioportal.org/) and the literature (Quigley et al., 2018 for WC-SU2C).

(M) *NEFH* Δ871-911 is the most frequent TSE splicing commonly present in primary (TCGA-PRAD, *n*=492 samples) and metastatic prostate tumors (WC-SU2C, *n*=101 samples, dbGap: phs001648.v1.p1). Dot color indicates the protein domain (Tail) where TSE splicing events occur.

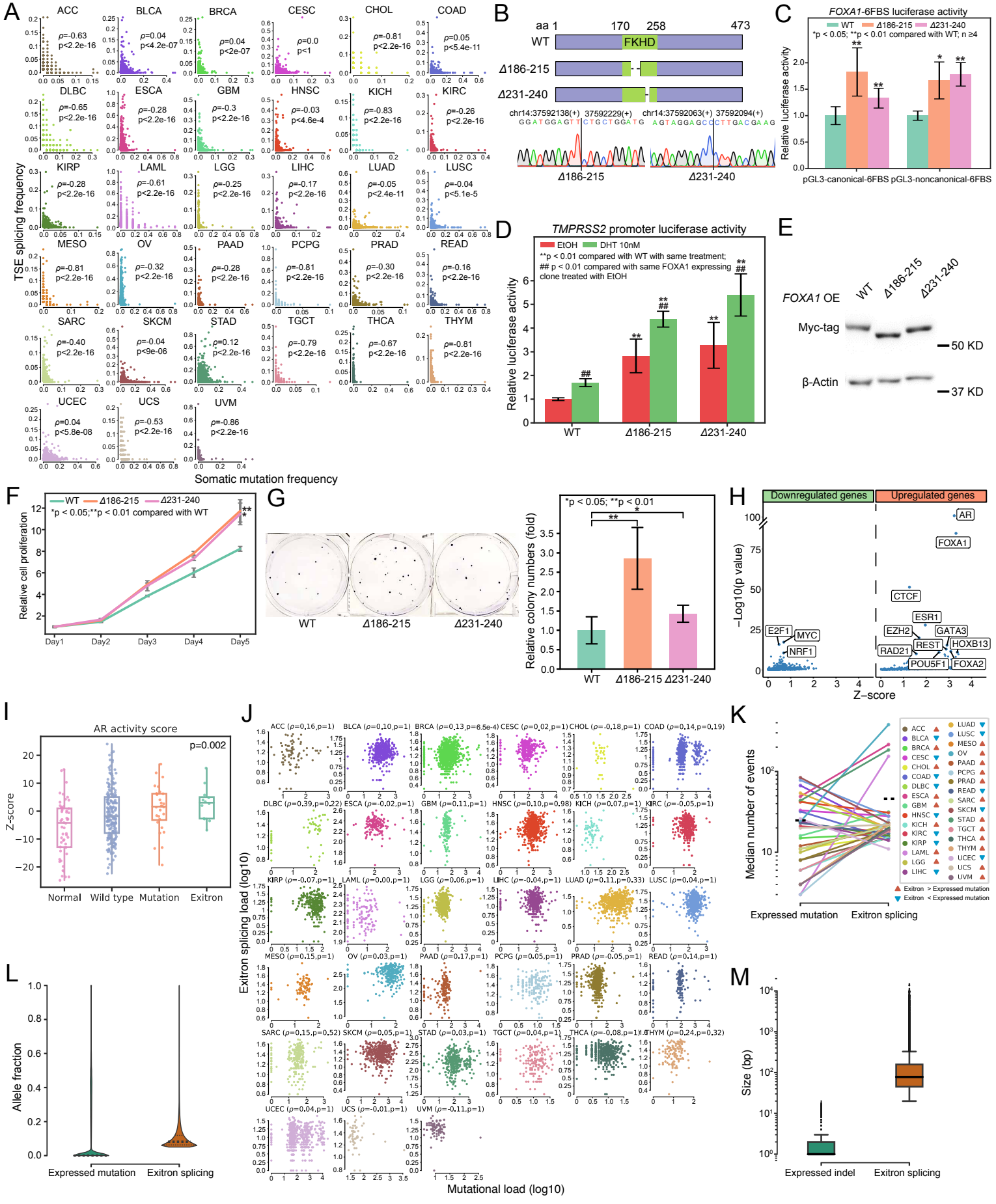
A
ACC ρ=-0.63 p<2.2e-16
BLCA ρ=0.04 p<4.2e-07
BRCA ρ=0.04 p<2e-07
CESC ρ=0.0 p<1
CHOL ρ=-0.81 p<2.2e-16
COAD ρ=0.05 p<5.4e-11
DLBC ρ=-0.65 p<2.2e-16
ESCA ρ=-0.28 p<2.2e-16
GBM ρ=-0.3 p<2.2e-16
HNSC ρ=-0.03 p<4.6e-4
KICH ρ=-0.83 p<2.2e-16
KIRC ρ=-0.26 p<2.2e-16
KIRP ρ=-0.28 p<2.2e-16
LAML ρ=-0.61 p<2.2e-16
LGG ρ=-0.25 p<2.2e-16
LIHC ρ=-0.17 p<2.2e-16
LUAD ρ=-0.05 p<2.4e-11
LUSC ρ=-0.04 p<5.1e-5
MESO ρ=-0.81 p<2.2e-16
OV ρ=-0.32 p<2.2e-16
PAAD ρ=-0.28 p<2.2e-16
PCPG ρ=-0.81 p<2.2e-16
PRAD ρ=-0.30 p<2.2e-16
READ ρ=-0.16 p<2.2e-16
SARC ρ=-0.40 p<2.2e-16
SKCM ρ=-0.04 p<9e-06
STAD ρ=0.12 p<2.2e-16
TGCT ρ=-0.79 p<2.2e-16
THCA ρ=-0.67 p<2.2e-16
THYM ρ=-0.81 p<2.2e-16
UCEC ρ=0.04 p<5.8e-08
UCS ρ=-0.53 p<2.2e-16
UVM ρ=-0.86 p<2.2e-16
TSE splicing frequency
Somatic mutation frequency
B
aa 1
170
258
473
WT
FKHD
Δ186-215
Δ231-240
chr14:37592138(+) 37592229(+)
chr14:37592063(+) 37592094(+)
GGATGGAGTTCTGCTGGATG
AGTAGGAGCCCTTGACGAAG
Δ186-215
Δ231-240
C
FOXA1-6FBS luciferase activity
WT
Δ186-215
Δ231-240
*p < 0.05; **p < 0.01 compared with WT; n ≥4
Relative luciferase activity
pGL3-canonical-6FBS
pGL3-noncanonical-6FBS
D
TMPRSS2 promoter luciferase activity
EtOH
DHT 10nM
**p < 0.01 compared with WT with same treatment;
p < 0.01 compared with same FOXA1 expressing clone treated with EtOH
Relative luciferase activity
WT
Δ186-215
Δ231-240
E
FOXA1 OE
WT
Δ186-215
Δ231-240
Myc-tag
β-Actin
50 KD
37 KD
F
WT
Δ186-215
Δ231-240
*p < 0.05;**p < 0.01 compared with WT
Relative cell proliferation
Day1
Day2
Day3
Day4
Day5
G
WT
Δ186-215
Δ231-240
*p < 0.05; **p < 0.01
Relative colony numbers (fold)
WT
Δ186-215
Δ231-240
H
Downregulated genes
Upregulated genes
−Log10(p value)
Z−score
AR
FOXA1
CTCF
ESR1
EZH2
GATA3
REST
HOXB13
RAD21
POU5F1
FOXA2
E2F1
MYC
NRF1
I
AR activity score
p=0.002
Z-score
Normal
Wild type
Mutation
Exitron
J
Exitron splicing load (log10)
Mutational load (log10)
ACC (ρ=0.16,p=1)
BLCA (ρ=0.10,p=1)
BRCA (ρ=0.13,p=6.5e-4)
CESC (ρ=0.02,p=1)
CHOL (ρ=-0.18,p=1)
COAD (ρ=0.14,p=0.19)
DLBC (ρ=0.39,p=0.22)
ESCA (ρ=-0.02,p=1)
GBM (ρ=0.11,p=1)
HNSC (ρ=0.10,p=0.98)
KICH (ρ=0.07,p=1)
KIRC (ρ=-0.05,p=1)
KIRP (ρ=-0.07,p=1)
LAML (ρ=0.00,p=1)
LGG (ρ=0.06,p=1)
LIHC (ρ=-0.04,p=1)
LUAD (ρ=0.11,p=0.33)
LUSC (ρ=0.04,p=1)
MESO (ρ=0.15,p=1)
OV (ρ=0.03,p=1)
PAAD (ρ=0.17,p=1)
PCPG (ρ=0.05,p=1)
PRAD (ρ=-0.05,p=1)
READ (ρ=0.14,p=1)
SARC (ρ=0.15,p=0.52)
SKCM (ρ=0.05,p=1)
STAD (ρ=0.03,p=1)
TGCT (ρ=0.04,p=1)
THCA (ρ=-0.08,p=1)
THYM (ρ=0.24,p=0.32)
UCEC (ρ=0.04,p=1)
UCS (ρ=-0.01,p=1)
UVM (ρ=-0.11,p=1)
K
Median number of events
Expressed mutation
Exitron splicing
ACC
BLCA
BRCA
CESC
CHOL
COAD
DLBC
ESCA
GBM
HNSC
KICH
KIRC
KIRP
LAML
LGG
LIHC
LUAD
LUSC
MESO
OV
PAAD
PCPG
PRAD
READ
SARC
SKCM
STAD
TGCT
THCA
THYM
UCEC
UCS
UVM
Exitron > Expressed mutation
Exitron < Expressed mutation
L
Allele fraction
Expressed mutation
Exitron splicing
M
Size (bp)
Expressed indel
Exitron splicing

Figure S4. Mutual exclusivity between tumor-specific exitrons and somatic mutations, Related to Figure 4.

(A) The frequency of somatic mutation and exitrons splicing occurred in genes are inversely correlated across TCGA cohorts. Each dot corresponds to a gene. Coloring is according to cancer type. The title of each sub-panel stats the Spearman correlation coefficient. All p values have been corrected for multiple-testing using Bonferroni's method.

(B) Graphical depiction of FOXA1 protein Forkhead DNA-binding domain (FKHD) and locations of two FKHD exitrons. Electropherogram illustrates splicing junctions of FOXA1 exitrons shown in the upper diagram**.**

(C) FOXA1-luciferase reporter assay with results normalized to level of wild-type (WT) FOXA1 activity.

(D) TMPRSS2 promoter luciferase reporter activity with overexpression of WT or various FOXA1 mutants and DHT stimulation.

(E) Representative western blot against C4-2 stable cell lines expressing Myc-tagged WT and exitron-spliced FOXA1.

(F) CellTiter-Glo growth assays indicate that overexpression of exitron-spliced FOXA1 significantly promote growth in prostate cancer cells in standard media conditions.

(G) Colony-formation ability of prostate cancer cell lines overexpressing exitron-spliced FOXA1 is increased relative to cells overexpressing WT FOXA1. Cells were fixed and stained with crystal violet. Assays were done at least thrice and in triplicate wells. The figure is a representative of three experiments with similar results. Quantification was performed by manual counting. All p values are relative to WT, calculated using unpaired, two-tailed Student's t-test. Error bars indicate ± s.d. *$p < 0.05$, **$p < 0.01$.

(H) BART prediction of specific transcription factors mediating observed transcriptional changes. The significant and strong (z-score) mediators of transcriptional responses in patients with FOXA1 exitron splicing events are labelled (BART, Wilcoxon rank-sum test).

(I) Extent of AR pathway activation in FOXA1 exitron spliced, mutated, wild-type tumors and adjacent normal prostate tissues in TCGA study. AR score was calculated using established gene signatures. P value was calculated by Kruskal-Wallis test.

(J) Correlation between the number of tumor-specific exitron splicing events per sample and its corresponding somatic mutational load across TCGA cohorts. Each dot corresponds to a sample. Coloring is according to cancer type. Both exitron splicing load and mutational load are shown on log10 scale. The title of each sub-panel stats the Spearman correlation coefficient. All p values have been corrected for multiple-testing using Bonferroni's method.

(K) The comparison on median number of expressed mutations and tumor-specific exitron splicing events per sample in each TCGA cohort.

(L) The comparison on allele fraction of expressed mutations and exitron splicing events. Allele fractions of expressed mutation and exitron splicing were measured by VAF and PSO respectively.

(M) The comparison on size between expressed indels (inframe and frameshift) and exitron splicing events.

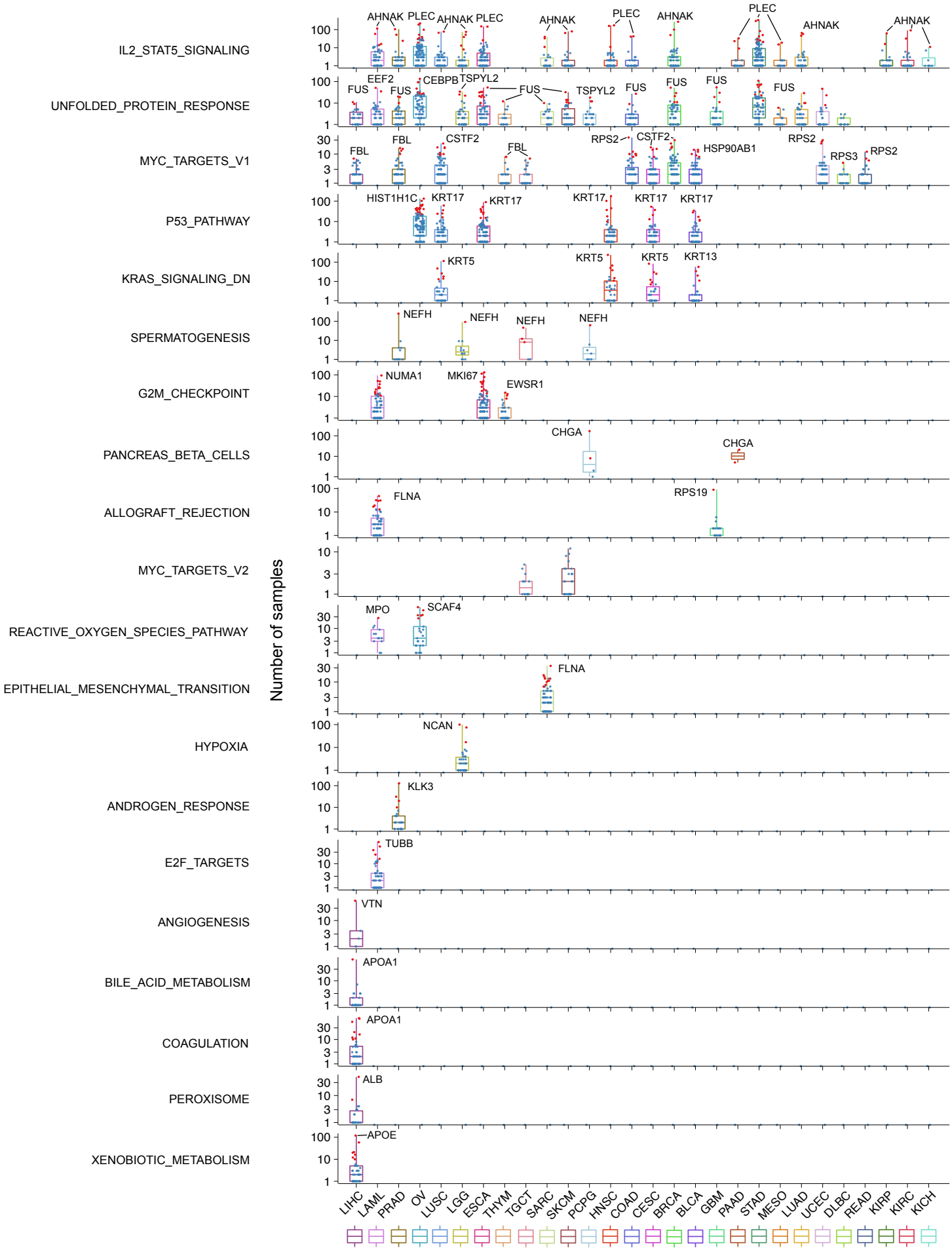

Figure S5. TSE splicing enriched hallmark gene sets identified across TCGA cohorts, Related to Figure 5. SEGs in each tumor cohort and gene set are colored in red. The SEG with the highest number of exitron splicing altered samples in each gene set is highlighted.

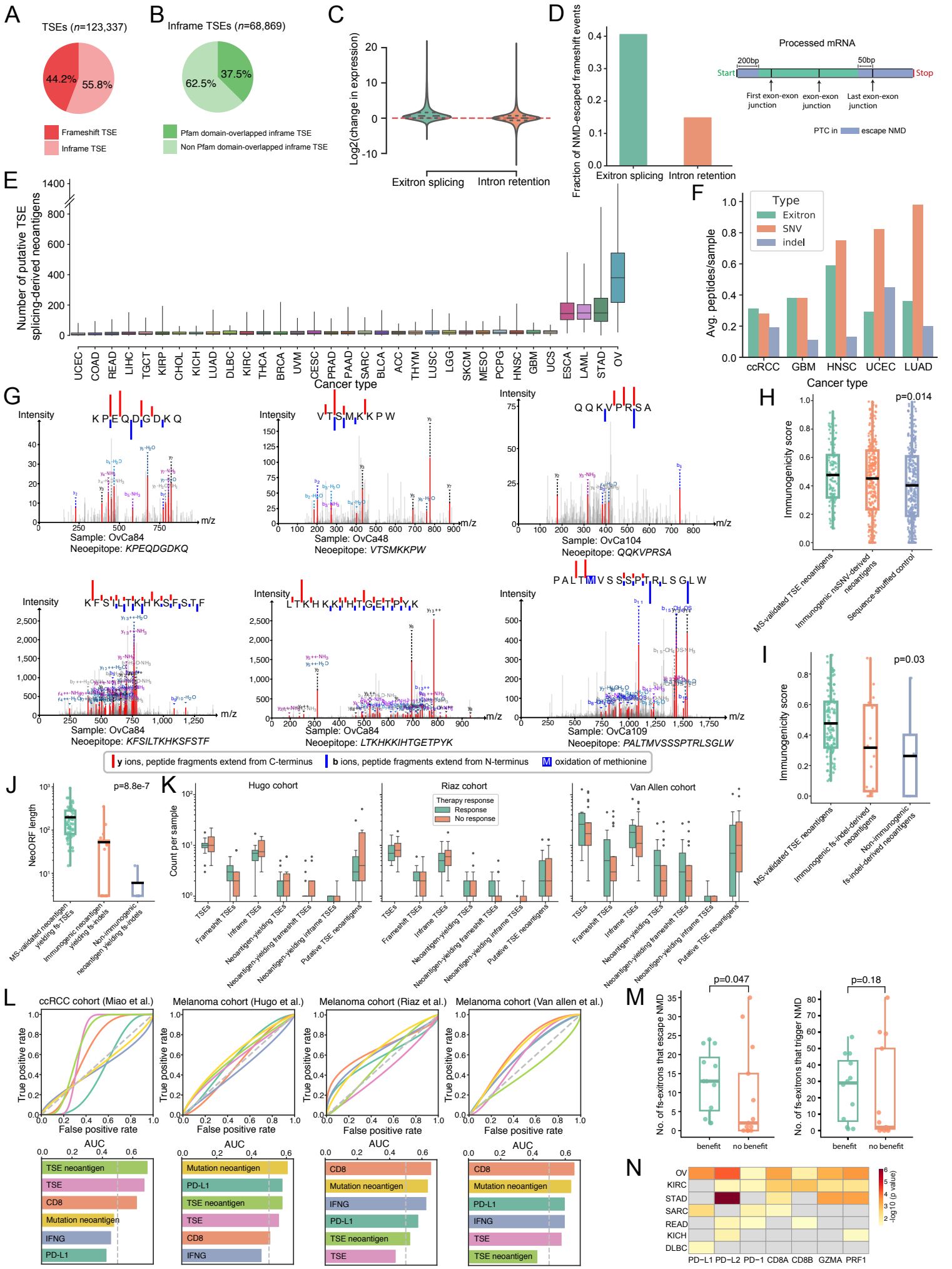
A
TSEs (n=123,337)
44.2%
55.8%
Frameshift TSE
Inframe TSE
B
Inframe TSEs (n=68,869)
62.5%
37.5%
Pfam domain-overlapped inframe TSE
Non Pfam domain-overlapped inframe TSE
C
Log2(change in expression)
Exitron splicing
Intron retention
D
Fraction of NMD-escaped frameshift events
Exitron splicing
Intron retention
Processed mRNA
200bp
50bp
Start
Stop
First exon-exon junction
exon-exon junction
Last exon-exon junction
PTC in
escape NMD
E
Number of putative TSE splicing-derived neoantigens
UCEC COAD READ LIHC TGCT KIRP CHOL KICH LUAD DLBC KIRC THCA BRCA UVM CESC PRAD PAAD SARC BLCA ACC THYM LUSC LGG SKCM MESO PCPG HNSC GBM UCS ESCA LAML STAD OV
Cancer type
F
Type
Exitron
SNV
indel
Avg. peptides/sample
ccRCC GBM HNSC UCEC LUAD
Cancer type
G
Intensity
K P E Q D G D K Q
Sample: OvCa84
Neoepitope: KPEQDGDKQ
V T S M K K P W
Sample: OvCa48
Neoepitope: VTSMKKPW
Q Q K V P R S A
Sample: OvCa104
Neoepitope: QQKVPRSA
K F S I L T K H K K S F S T F
Sample: OvCa84
Neoepitope: KFSILTKHKKSFSTF
L T K H K K I H T G E T P Y K
Sample: OvCa84
Neoepitope: LTKHKKIHTGETPYK
P A L T M V S S S P T R L S G L W
Sample: OvCa109
Neoepitope: PALTMVSSSPTRLSGLW
y ions, peptide fragments extend from C-terminus
b ions, peptide fragments extend from N-terminus
M oxidation of methionine
H
p=0.014
Immunogenicity score
MS-validated TSE neoantigens
Immunogenic nsSNV-derived neoantigens
Sequence-shuffled control
I
p=0.03
Immunogenicity score
MS-validated TSE neoantigens
Immunogenic fs-indel-derived neoantigens
Non-immunogenic fs-indel-derived neoantigens
J
p=8.8e-7
NeoORF length
MS-validated neoantigen yielding fs-TSEs
Immunogenic neoantigen yielding fs-indels
Non-immunogenic neoantigen yielding fs-indels
K
Hugo cohort
Riaz cohort
Van Allen cohort
Therapy response
Response
No response
Count per sample
TSEs
Frameshift TSEs
Inframe TSEs
Neoantigen-yielding TSEs
Neoantigen-yielding frameshift TSEs
Neoantigen-yielding inframe TSEs
Putative TSE neoantigens
L
ccRCC cohort (Miao et al.)
Melanoma cohort (Hugo et al.)
Melanoma cohort (Riaz et al.)
Melanoma cohort (Van allen et al.)
True positive rate
False positive rate
AUC
TSE neoantigen
TSE
CD8
Mutation neoantigen
IFNG
PD-L1
Mutation neoantigen
PD-L1
TSE neoantigen
TSE
CD8
IFNG
CD8
Mutation neoantigen
IFNG
PD-L1
TSE neoantigen
TSE
CD8
Mutation neoantigen
PD-L1
IFNG
TSE
TSE neoantigen
M
p=0.047
No. of fs-exitrons that escape NMD
benefit
no benefit
p=0.18
No. of fs-exitrons that trigger NMD
benefit
no benefit
N
OV
KIRC
STAD
SARC
READ
KICH
DLBC
PD-L1 PD-L2 PD-1 CD8A CD8B GZMA PRF1
-log10 (p value)

Figure S6. Tumor-specific exitrons (TSEs) represent a source of neoepitopes, Related to Figure 6.
(A) The proportion of frameshift and inframe TSEs in TCGA tumors.
(B) The proportion of inframe TSEs overlapping with Pfam domains.
(C) RNA-Seq expression changes in alternatively-spliced versus wild-type samples. The extent of NMD by exitron splicing and intron retention is estimated by comparing the mRNA expression level in samples with the splicing event to the median mRNA expression level of the same gene across all other wild-type tumor samples.
(D) Proportion of NMD-escape events between frameshift exitron splicing events and intron retention. Rules of NMD evasion is determined based on the premature termination codon (PTC) position according to the literature (Lindeboom et al., 2016).
(E) TSE neoantigen counts by cancer type.
(F) Comparison of the contribution of TSE splicing, somatic SNVs and indels to proteomic-confirmed putative neoantigens across five cancer types offered by the CPTAC3 study.
(G) Mass spectra show TSE neoepitopes bound to (left panel) MHC class I molecules and (right panel) MHC class II molecules in Schuster et al. ovarian cancer cohort.
(H) Comparison of the immunogenicity scores of mass spectrometry (MS)-validated TSE neoantigens from CPTAC with nsSNV-derived neoepitopes inducing T-cell responses collected from dbPepNeo database and corresponding sequence-shuffled control.
(I) Comparison of the immunogenicity scores of MS-validated TSE neoantigens from CPTAC with functionally validated immunogenic frameshift (fs) indels-derived neoantigens and non-immunogenic indel-derived neoantigens collected in the literature (Litchfield et al., 2020).
(J) Comparison of neoORF length between MS-validated frameshift TSEs and frameshift indels. P values were calculated by Kruskal-Wallis test. Mean values are highlighted in black.
(K) Association of TSEs, frameshift TSEs, inframe TSEs, neoantigen-yielding TSEs, neoantigen-yielding frameshift TSEs, neoantigen-yielding inframe TSEs and putative TSE neoantigens with response to checkpoint inhibitor therapy in Hugo (n=14 response, n=13 no response), Riaz (n=10 response, n=23 no response) and Van Allen (n=13 response, n=21 no response) melanoma patient cohorts. All comparisons show a non-significant trend ($p > 0.05$, Mann-Whitney rank test). Boxplots show the median, first and third quartiles, whiskers extend to 1.5× the interquartile range, and outliers are plotted individually.
(L) ROC curves for the performance and the area under the ROC curve (AUC) of the TSE neoantigen load, TSE splicing load, mutation neoantigen load, and immune signatures of CD8, PD-L1 and interferon gamma (IFNG) in predicting checkpoint inhibitor response in ccRCC and melanoma cohorts. Grey dash line indicates random predictions.
(M) Efficacy of immunotherapy is predicted by the burden of NMD-escaped frameshift (fs) exitrons (left panel) but not NMD-triggered fs-exitrons (right panel) in clear cell renal cell carcinoma (ccRCC). P values were calculated using Mann-Whitney rank test.
(N) Comparison of gene expression of PD-L1, PD-L2, PD-1, CD8A, CD8B, GZMA and PRF1 between the top and bottom quartile patients based on their TSE splicing-derived neoantigen load across TCGA cancer types. Cancer types with at least one significant p value for the seven gene markers are showed. TSE splicing events with PSO ≥ 0.1 are used for comparison. The p values were calculated by a Mann-Whitney rank test.

Table S1. Cancer abbreviation in TCGA cohort, Related to Figures 1, 2, 3 and 6.

Abbreviation	Study name
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UVM	Uveal Melanoma

Table S4. CPTAC-confirmed neoantigens derived from exitrons, Related to Figure 6.

Neoantigen	Gene	Chr	Start	End	Sample	Cohort
MTRAAWPLM	ARVCF	chr22	19981341	19981603	TCGA-61-1907-01A-01R-1567-13	OV
RAAWPLMTR	ARVCF	chr22	19981341	19981603	TCGA-61-1907-01A-01R-1567-13	OV
HGRAESAPR	DTX2	chr7	76482754	76482816	TCGA-61-1907-01A-01R-1567-13	OV
MAWRMTRAA	ARVCF	chr22	19981341	19981603	TCGA-61-1907-01A-01R-1567-13	OV
RMTRAAWPL	ARVCF	chr22	19981341	19981603	TCGA-61-1907-01A-01R-1567-13	OV
MAVPLLRRM	CTAG2	chrX	154653345	154653496	TCGA-61-1741-01A-02R-1567-13	OV
WMMRKTWMK	CCDC34	chr11	27362975	27363120	TCGA-29-1697-01A-01R-1567-13	OV
SLKVECMPK	CDH11	chr16	64982087	64982256	TCGA-29-1774-01A-01R-1567-13	OV
MTMDMGRFK	KHDRBS1	chr1	32037905	32037990	TCGA-29-1774-01A-01R-1567-13	OV
QSSWPCTFK	ZNF219	chr14	21090795	21090835	TCGA-13-1492-01A-01R-1565-13	OV
FQCAFLHRL	LRRK1	chr15	101065771	101065998	TCGA-24-1435-01A-01R-1566-13	OV
FYFGGPQYL	MYCN	chr2	15942179	15942814	TCGA-24-1435-01A-01R-1566-13	OV
RAAAAAGVR	AEBP2	chr12	19439858	19439919	TCGA-25-1316-01A-01R-1565-13	OV
AAAAGVRRR	AEBP2	chr12	19439858	19439919	TCGA-25-1316-01A-01R-1565-13	OV
AAAAAGVRR	AEBP2	chr12	19439858	19439919	TCGA-25-1316-01A-01R-1565-13	OV
KMRAAAAAG	AEBP2	chr12	19439858	19439919	TCGA-25-1316-01A-01R-1565-13	OV
EMAPSWVLK	MYL6	chr12	56159988	56160073	TCGA-25-1319-01A-01R-1565-13	OV
KEMAPSWVL	MYL6	chr12	56159988	56160073	TCGA-25-1319-01A-01R-1565-13	OV
ERRRSLNSF	NCL	chr2	231455442	231455524	TCGA-25-2399-01A-01R-1569-13	OV
SAFPFPFER	NCL	chr2	231455442	231455524	TCGA-25-2399-01A-01R-1569-13	OV
SFCPSAFPF	NCL	chr2	231455442	231455524	TCGA-25-2399-01A-01R-1569-13	OV
SSRSWMLVR	EMILIN2	chr18	2892048	2892154	TCGA-25-2399-01A-01R-1569-13	OV
MRCWRTTRM	ZBTB46	chr20	63747061	63747212	TCGA-59-2351-01A-01R-1569-13	OV
ATPYSFLSL	VWA5A	chr11	124123723	124123790	TCGA-61-1914-01A-01R-1567-13	OV
LMQATPYSF	VWA5A	chr11	124123723	124123790	TCGA-61-1914-01A-01R-1567-13	OV
MQATPYSFL	VWA5A	chr11	124123723	124123790	TCGA-61-1914-01A-01R-1567-13	OV
IQTPPHDTL	HNRNPD	chr4	82373506	82373588	TCGA-AR-A0TT-01A-31R-A084-07	BRCA
TMWKNHYQK	WWC2	chr4	183261269	183261363	TCGA-A2-A0T7-01A-21R-A084-07	BRCA

Table S5. MS-based imunopeptidomics-confirmed neoantigens derived from exitrons, Related to Figure 6.

Neoantigen	Gene	Chr	Start	End	Sample	MHC type
WGRLLSEY	BTN3A1	chr6	26406024	26406142	OvCa48	classI
VTSMKKPW	BTN3A2	chr6	26368680	26368798	OvCa48	classI
MEMKMRKL	PTMA	chr2	231711925	231711947	OvCa48	classI
KPEQDGDKQ	HLCS	chr21	36936676	36937275	OvCa84	classI
QQKVPRSA	FGD5	chr3	14897954	14898025	OvCa104	classI
TMVSSSPTRLSGLWMT	PRPF8	chr17	1658603	1658682	OvCa58	classII
VARLPLCRREPEP	WDR97	chr8	144109435	144109926	OvCa58	classII
LVEDCLPNGGAQ	WDR97	chr8	144109435	144109926	OvCa58	classII
MKANPALTMVSSSPTRL	PRPF8	chr17	1658603	1658682	OvCa65	classII
KFSILTKHKAFNRS	ZNF208	chr19	21972981	21973953	OvCa70	classII
ILTKHKAYNTFSIL	ZNF208	chr19	21972477	21973953	OvCa70	classII
KFSILTKHKAYNT	ZNF208	chr19	21972477	21973953	OvCa70	classII
LTKHKKIHTGETPYK	ZNF208	chr19	21972057	21972477	OvCa84	classII
DTSSSSTGHATPL	MUC4	chr3	195781609	195783385	OvCa84	classII
KFSILTKHKSFSTF	ZNF208	chr19	21972057	21973953	OvCa84	classII
LHPHPQGAEAAAGLLQRAR	HSPA1B	chr6	31828262	31828846	OvCa99	classII
RGEGSARRQAGQGPDSRPG	HSPA1B	chr6	31828262	31828846	OvCa99	classII
GRGLHPHPQGAEAAAGLL	HSPA1B	chr6	31828262	31828846	OvCa99	classII
LTKHKAYNTFSIL	ZNF208	chr19	21972477	21972981	OvCa104	classII
NPITVRNVGKPFGTAQ	ZNF485	chr10	43616923	43617050	OvCa109	classII
NPALTMVSSSPTRLSGL	PRPF8	chr17	1658603	1658682	OvCa109	classII
ANPALTMVSSSPTRLSGL	PRPF8	chr17	1658603	1658682	OvCa109	classII
PALTMVSSSPTRLSGLW	PRPF8	chr17	1658603	1658682	OvCa109	classII
KPFGTAQALLNIRD	ZNF485	chr10	43616923	43617050	OvCa109	classII
KFSILTKHKAYNTF	ZNF208	chr19	21972477	21973953	OvCa114	classII
KIMKANPALTMV	PRPF8	chr17	1658603	1658682	OvCa114	classII