Supplementary Materials for

A pro-metastatic splicing program regulated by SNRPA1 interactions with structured RNA elements

Lisa Fish, Matvei Khoroshkin, Albertas Navickas, Kristle Garcia, Bruce Culbertson, Benjamin Hänisch, Steven Zhang, Hoang C.B. Nguyen, Larisa M. Soto, Maria Dermit, Faraz K. Mardakheh, Henrik Molina, Claudio Alarcón, Hamed S. Najafabadi, Hani Goodarzi

correspondence to: hani.goodarzi@ucsf.edu

This PDF file includes:

MATERIALS AND METHODS

Cell Culture

All cells were cultured in a 37°C 5% CO2 humidified incubator. The MDA-MB-231 (MDA-parental, ATCC HTB-26) breast cancer cell line, its highly metastatic derivative, MDA-LM2 (*57*), and 293LTV cells (Cell BioLabs LTV-100) were cultured in DMEM medium supplemented with 10% FBS, glucose (4.5 g/L), L-glutamine (4 mM), sodium pyruvate (1 mM), penicillin (100 units/mL), streptomycin (100 μg/mL) and amphotericin (1 μg/mL) (Gibco). The HCC1806-LM2 cell line (an *in vivo* selected highly lung metastatic derivative of the HCC1806 breast cancer line (ATCC CRL-2335)) was cultured in RPMI-1640 medium supplemented with 10% FBS, glucose (2 g/L), L-glutamine (2 mM), sodium pyruvate (1 mM), penicillin (100 units/mL), streptomycin (100 μg/mL) and amphotericin (1 μg/mL) (Gibco). All cell lines were routinely screened for mycoplasma with a PCR-based assay.

shRNA and siRNA-Mediated Gene Knockdown

For stable knockdown of target genes, shRNAs targeting the genes of interest were cloned into the EcoRI and AgeI sites of the pLKO.1 vector (Addgene plasmid #10878), see Table S1 for shRNA sequences. The shRNA constructs were packaged using the ViraSafe lentiviral packaging system (Cell Biolabs, Inc.) using lipofectamine 2000 (Invitrogen) and Opti-MEM (Invitrogen) to transfect the shRNA constructs along with the packaging plasmids into 293LTV cells (Cell Biolabs, Inc.). Virus was harvested 48 hours post-transfection and passed through a 0.45 μm filter. Target cells constitutively expressing luciferase were then transduced for 6-8 hours with the filtered virus in the presence of 8 μg/mL polybrene (Millipore). Transduced cells were selected by treatment with 1.5 μg/mL puromycin.

For transient knockdown of target genes, siRNAs (IDT DNA; see Table S1) were used. Cancer cells were seeded at $1x10^5$ per well and transfected with 100 pmol siRNA using lipofectamine 2000 and Opti-MEM (Invitrogen) per the manufacturer's protocol. Cells were harvested 48-72 hours post-transfection. Knockdown of target genes was assessed by RT-qPCR as described below. For SNRPA1, we also verified the knockdown efficiency by western blot. On average, we observed a log fold-change of -0.4 in knockdown versus control cells. This level of knockdown is comparable with the difference observed by western blot between MDA-LM2 and MDA-parental cells.

Mimetic transfection

MDA-LM2 cells were seeded at $1x10^5$ cells per well in 6-well plates 24 hours before transfection. Each class of S3E variant RNA (as prepared for the Bind-n-Seq experiment, see Table S1 for sequences) was transfected using Lipofectamine 2000 (Invitrogen) and Opti-MEM (Invitrogen) per the manufacturer's protocol. To generate the pooled variant S3E RNA samples, equimolar amounts of the indicated variants were mixed, and 100pmol of each pool was transfected per well. Mock transfections were carried out simultaneously as a control, and all transfections were carried out in duplicate. 48 hours after transfection, Zymo RNA quick prep kit was used to isolate total RNA from the cells, including an on-column DNase treatment. This RNA was processed into RNA-seq libraries using Takara SMARTer stranded total RNA-seq kit v2 – pico input mammalian, and then sequenced on an Illumina HiSeq4000 instrument.

SNRPA1 Overexpression

MDA-parental and HCC1806 cells were engineered to stably express SNRPA1 using lentiviral delivery of the pLX304 vector containing the SNRPA1 ORF. An mCherry ORF-containing pLX304 was used as the control. We verified the overexpression by RT-qPCR and western blot. On average, by western blot we observed a log fold-change of 1.0 in SNRPA1-overexpressing MDA-Parental versus control cells.

Endogenous S3E deletion

PLEC exon31 S3E and ERRFI1 exon3 S3E were deleted in MDA-LM2 cells using the transfection of *Spy*Cas9-crRNA-tracrRNA RNPs. All reagents were from IDT DNA, unless indicated otherwise. Two crRNAs flanking individual S3Es (see Table S1 for sequences) or a non-targeting control crRNA were annealed with tracrRNA separately, in Duplex buffer at 1 μM concentration. The crRNA:tracrRNA duplexes were incubated with Cas9 protein and Cas9 PLUS Reagent (Thermo) in OptiMEM (Gibco) at 60 nM concentration, to form RNPs. The RNPs targeting the same S3E were then mixed in 1:1 molar ratio. The RNP mix was then reverse-transfected into MDA-LM2 cells using CRISPRMAX (Thermo), at 10 nM final concentration. Forty-eight hours post transfection the cells were expanded for genomic DNA and RNA extraction. Quick-DNA Miniprep or Quick-RNA Microprep kits (Zymo) were used for nucleic acid extraction. The edited cells were genotyped by PCR (see Table S1 for primer sequences). PLEC exon31 and ERRFI1 exon3 inclusion was measured by RT-PCR (see Table S1 for primer sequences).

Reporter assays

To systematically address the impact of the S3E structure and sequence to SNRPA1-dependent PLEC exon31 and ERRFI1 exon3 inclusion, two additional minigene reporters were constructed in the pAN30 backbone (*59*). Synthetic DNA (gAN048 and gAN049) pieces, containing truncated PLEC exons 30, 31, 32 and the corresponding introns (preserving the branchpoints) or ERRFI1 exons 2, 3, 4 and the corresponding introns (Twist Biosciences) were Gibson assembled into PacI-digested pAN30, resulting in pAN61 and pAN62, respectively. S3Es were flanked by AsiSI and SbfI sites, and BstXI and MluI sites were inserted into PLEC exon32 and ERRFI1 exon4, proximal to the 5' splice site of these exons. The libraries of S3E variants were cloned into the reporters via AsiSI-SbfI sites and electroporated into MegaX *E. coli* cells (Thermo). Then, the S3E library plasmids were digested with BstXI, the protruding ends were used to anneal a 12 nt random barcode- and MluI site-containing oligonucleotide, and the second strand was synthesized with Klenow fragment, exo- (NEB). The barcoded DNA was then cut with MluI, religated and electroporated into *E. coli* cells.

The barcoded S3E library plasmids were PCR amplified with Illumina flowcell adaptor-containing primers and sequenced as a paired-end run on a MiSeq sequencer (Illumina). This allowed matching each unique barcode, situated in PLEC exon 32 and ERRFI1 exon 4, with the S3E variant in PLEC exon 3 and ERRFI1 intron 3. The barcoded S3E library plasmids were then stably expressed in MDA-LM2 cells by lentiviral transduction followed by puromycin selection, in biological duplicates. 72 hours prior to RNA extraction, the cells were transfected with control or SNRPA1-targeting siRNA, as described above. Biotinylated oligonucleotides complementary to the reporter RNA were annealed with 100 μg of total RNA extracted from barcoded S3E library reporter cells in

1X B&W Buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl), and then pulled-down with streptavidin magnetic beads (Dynabeads Streptavidin C1, Thermo). The purified reporter RNA was then reverse transcribed and PCR amplified with Illumina flowcell adaptor containing primers and sequenced on HiSeq4000 sequencer (UCSF CAT). The resulting reads were used to match the reporter barcodes with the splice junction used.

Metastatic Colonization Assay

Seven- to twelve-week-old age-matched female NOD/SCID gamma mice (NSG, Jackson Labs, 005557) were used for lung colonization assays. For this assay, cancer cells constitutively expressing luciferase were suspended in 100 µL PBS and then injected via tail-vein ($2.5x10^4$ for MDA-LM2, $5x10^4$ for MDA-Parental, $1x10^5$ for HCC1806-LM2). Each cohort contained 4-5 mice, which in NSG background is enough to observe a >2-fold difference with 90% confidence. Mice were randomly assigned into cohorts. Cancer cell growth was monitored *in vivo* at the indicated times by retro-orbital injection of 100 µl of 15 mg/mL luciferin (Perkin Elmer) dissolved in 1X PBS, and then measuring the resulting bioluminescence with an IVIS instrument and Living Image software (Perkin Elmer).

Orthotopic Tumor Growth Assay

Tumor growth assays were performed by injecting cancer cells ($1x10^5$ MDA-LM2 cells or $2x10^5$ HCC1806-LM2 cells) mixed with 25 µL of Matrigel (Corning) into bilateral mammary fat pads of seven- to twelve-week old age-matched female NOD/SCID gamma mice. Tumor volume was assessed weekly by caliper measurements.

Histology

For gross macroscopic metastatic nodule visualization, mice lungs (from each cohort) were extracted at the endpoint of each experiment, and 5 µm thick lung tissue sections were hematoxylin and eosin (H&E) stained. The number of macroscopic nodules was then recorded for each section. An unpaired *t*-test was used to test for significant variations.

Immunohistochemistry

The tissue microarray (TMA) was obtained from the Cooperative Human Tissue Network at the University of Virginia (TMA version: CHTN_BRCaProg2). After deparaffinization by incubation in two baths of xylene for 10 minutes each, TMA was rehydrated by sequential incubation in 100%, 95%, 80% and 60% ethanol for 5 minutes each. The TMA was then rinsed 3X with diH_2O. For antigen retrieval, TMA was boiled in 10 mM Tris, 1 mM EDTA, pH 9.0, and then incubated at room temperature for 35 minutes. The TMA was rinsed 3X with PBS, then incubated in 3% H_2O_2 for 10 minutes. The TMA was rinsed 3X with PBS, and then blocked with 5% milk-PBST at room temperature for 1 hour. The TMA was then incubated with SNRPA1 antibody (Proteintech 17368-1-AP), diluted 1:100 in PBS, in a humidified chamber at 4°C overnight. The TMA was then rinsed 3X with PBST, and then incubated with a biotinylated secondary antibody (Vector Labs BA-1000), diluted in PBST at room temperature for 30 minutes. The TMA was then washed 3X with PBST. Detection was performed using Vectastain ABC HRP Kit (Vector Labs PK-4005) per the manufacturer's instructions. The TMA was then dehydrated by sequential incubation in 60%, 80%, 95%, and 100% ethanol for five minutes each, and

then incubated in two baths of xylene for five minutes each. The slides were air dried and scanned. Staining intensity of each tissue section was assessed by blinded grading of the slides.

Cell Proliferation

At day zero, cancer cells were seeded at $5x10^4$ cells per well. Viable cells were counted by trypsinization and trypan blue staining to determine cell viability using a hemocytometer at day 1, day 3 and day 5. An exponential model was then used to fit a growth rate for each sample ($\ln(N_{t-1}/N_1)= rt$ where t is measured in days). The experiment was performed in biological triplicates, and an unpaired *t*-test was used to test for significant variations.

Cell Invasion

Following morpholino transfection, MDA-LM2 or HCC1806-LM2 cells were starved for 18 hours in DMEM (for MDA-LM2) or RPMI (for HCC1806-LM2) supplemented with 0.2% FBS. 500 µl of the above starving media was then added to the top and bottom of 8 invasion chambers (Corning 354480) sitting in a 24 well plate. $1x10^5$ cells were seeded in each invasion chamber and incubated at 37° C for 24 hours. The bottom and top of each invasion chamber was then washed twice with 500 µl PBS, and a q-tip was used to remove the cells from the top of the chamber during the second wash. Cells were then fixed by adding 300 µl of 4% PFA in PBS to the top and bottom of each chamber, and incubated at 37°C for 15 minutes. The cells were then washed twice with 500 µl PBS. The invasion chamber inserts were cut out using a scalpel and mounted on slides with DAPI-containing mounting media (vector labs H-1500). Slides were imaged at 10X magnification on a fluorescent Nikon Ti Microscope. Five images of different fields were taken from each insert, for a total of 20 images per cell line.

RNA Isolation

Total RNA for RNA-seq and quantitative RT-PCR assays was isolated using the Norgen Biotek total RNA isolation kit with on-column DNase treatment per the manufacturer's protocol.

RT-qPCR

Transcript levels were measured using quantitative RT-PCR by reverse transcribing total RNA to cDNA (SuperScript III, Invitrogen), then using fast SYBR green master mix (Applied Biosystems) or Perfecta SYBR green supermix (QuantaBio) per the manufacturer's instructions. HPRT1 and 18S were used as endogenous controls.

RNA-sequencing

Unless otherwise specified, transcriptomic libraries were prepared using RNA that had been rRNA depleted using Ribo-Zero Gold (Illumina) followed by library preparation with the ScriptSeq-v2 kit (Illumina), and sequenced on an Illumina HiSeq2500 or Illumina NextSeq500 instrument at the Rockefeller genomic resource center. RNA-seq libraries for expression profiling of MDA-parental cells with stable SNRPA1 overexpression were generated using the QuantSeq 3' mRNA-Seq library prep kit fwd (Lexogen) per the manufacturer's protocol, and sequenced on an Illumina HiSeq4000 at UCSF CAT. RNA-seq libraries of siRNA-mediated knockdown of SNRPA1, SNRPB2 and SNRPB in MDA-LM2 cells as well as MDA-LM2 cells transfected with the S3E

variants were prepared using the Takara SMARTer stranded total RNA-seq kit v2 – pico input mammalian, and sequenced on an Illumina HiSeq4000 at UCSF CAT.

For analysis, reads were aligned using STAR (v2) to the human genome (build hg38). Bam files were then merged between replicates and MISO (v0.5) was then used to estimate changes in Ψ (i.e. ΔΨ) for alternatively spliced exons (between samples). pyTEISER was then used to assess whether changes in ΔΨ were associated with S3E occurrence and/or SNRPA1 binding. These associations were reported as mutual information values (MI) in bits, *z*-scores. To examine the reproducibility of the results across splicing analysis tools, we also used rMATS (v4.0.2) to estimate ΔΨ, with replicates provided as separate Bam files. The ΔΨ estimates from MISO were then directly compared to ΔΨ estimates from rMATS.

Bind-N-Seq

The 19 RNA oligonucleotides used in the binding experiment were generated by in vitro transcription. A SNRPA1-bound region in the PLEC transcript was designated as the wild-type (WT) oligonucleotide, and various mutants of this sequence were designed to change the sequence but not the structure (structured variants), the sequence and the structure (unstructured variants), or disrupt the loop region of the wild-type RNA structure (loop variants). The DNA templates used for the in vitro transcription reactions to produce these transcripts were generated by PCR using Phusion hot start II master mix (Thermo), along with synthetic DNA oligonucleotide templates and primers, and included a 5' T7 promoter (IDT) (See Table S1 for sequences). PCR reactions were cleaned up using Zymo DNA Clean and Concentrator spin columns, and the PCR product size was checked using an Agilent Tape Station. The Megascript T7 kit (Ambion) was used per the manufacturer's protocol to perform in vitro transcription reactions. Reactions were incubated at 37°C for 4 hours using 50-60 ng of each DNA template per reaction. The reactions were then treated with 1ul of Turbo DNase at 37°C for 15 minutes. The resulting RNA was cleaned up using Zymo RNA clean and concentrator kit. The RNA was run on a urea-PAGE gel to check size and purity of each RNA oligonucleotide.

RNA oligonucleotides were individually folded before pooling and adding to the binding reaction. 3.34 pmol of RNA oligo in 10.5 µl 20 mM Tris-Cl pH 7.5, 200 mM KCl was heated at 90°C for 2 minutes, then immediately placed on ice. To this, 10.5 µl of 10 mM $MgCl_2$ was added, and the resulting mixture placed at 4°C. The temperature was then increased at a rate of 1°C per min to 30°C, then incubated at 30°C for 10 min. The folded RNAs were then pooled at equimolar concentrations, and an aliquot set aside as the input RNA pool. Binding reactions were performed in duplicate. For each binding reaction, 16.67 pmol recombinant SNRPA1 (Abcam) was mixed with 50 pmol pooled and folded RNA in 1X EMSA buffer (20 mM HEPES pH 7.9, 1 mM EDTA, 125 mM KCl, 5% glycerol, 0.1% Triton-X 100, 1 mM DTT, 0.5 units/µl SuperaseIn (Thermo)), in a final volume of 100 µl. Binding reactions were incubated at 30°C with end-over-end rotation for 30 minutes. Protein-RNA complexes were then isolated by mixing the binding reactions with protein A Dynabeads (Thermo) conjugated to a SNRPA1 antibody (Bethyl A303-948A) and incubating at 30°C with end-over-end rotation for 1 hour. The beads were then washed once with 1 mL 1X EMSA buffer. The immunoprecipitated material was then eluted by adding 10 mM Tris-Cl pH 7.0, 1 mM EDTA, 1% SDS and heating at 70°C for 10 minutes. The RNA from this eluate was then isolated by acid phenol chloroform extraction. Both the SNRPA1-bound RNA fractions as well as the input RNA

pool were processed into RNA-seq libraries using the Takara SMARTer smRNA-Seq kit for Illumina and sequenced on an Illumina NextSeq500 instrument.

Morpholino-Mediated Isoform Modulation

Isoform-specific and standard control morpholinos were ordered from Gene Tools, along with the Endo-Porter transfection reagent. Upon receipt, the morpholinos were dissolved in water to make a stock concentration of 1 mM. The day before transfection, MDA-LM2 or HCC1806- LM2 cells were seeded in 10 cm plates at 1 x 10^6 cells per plate. The following day, the media was replaced with 9 ml complete media, and then 180 µl of morpholino stock (final concentration was 20 µM morpholino) and 30 µl of Endo-Porter were added. The cells were incubated at 37°C for 5 hours, and the media was replaced. 48 hours after transfection, downstream assays were performed, including qPCR and western blotting to check isoform knockdown, *in vitro* and *in vivo* growth, and *in vitro* trans-well invasion.

Western Blotting

Cell lysates were prepared by lysing cells in ice-cold RIPA buffer (25 mM Tris-HCl pH 7.6, 0.15 M NaCl, 1% IGEPAL CA-630, 1% sodium deoxycholate, 0.1% SDS) containing 1X protease inhibitors (Thermo Scientific). Lysate was cleared by centrifugation at 20,000 x g for 15 min at 4°C. Samples were boiled in 1X LDS loading buffer (Invitrogen) and 50 mM DTT. Proteins were separated by SDS-PAGE using 4-12% Bis-Tris NuPAGE gels (SNRPA1) or 3-8% Tris-Acetate NuPAGE gels (PLEC), transferred to nitrocellulose (Millipore), blocked using 5% BSA, and probed using target-specific antibodies. Bound antibodies were detected using infrared dye-conjugated secondary antibodies (Licor) according to the manufacturer's instructions. Antibodies: beta-tubulin (Proteintech 66240-1-Ig), PLEC (Bethyl A304-506A).

S3E pulldown and mass spectrometry

S3E pulldown was performed according to (*60*). RNA baits were prepared by *in vitro* transcription (see Table S1 for sequences), including an S1 aptamer sequence at the 3' end of each transcript, then folded by incubation in 10 mM Tris pH 7.0, 100 mM KCl, 10 mM $MgCl_2$. 25 µg of each RNA bait was then conjugated to streptadivin C1 dynabeads (Invitrogen) by incubating in RNA binding buffer (100 mM NaCl, 50 mM HEPES pH 7.4, 0.5% NP-40, 10 mM $MgCl_2$) and incubated with shaking at 4°C for 30 minutes. Beads were washed 3X with RNA wash buffer (150 or 250 mM NaCl, 50 mM HEPES pH 7.4, 0.5% NP-40, 10 mM $MgCl_2$). The RNA-conjugated beads were then incubated at 4°C for 30 minutes with 400 µg of nuclear extract, 40 units of RNase inhibitor and 20 µg of yeast tRNA (Invitrogen). Nuclear lysate was prepared from MDA-LM2 cells by resuspending cells in cold 50 mM HEPES-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% Triton X-100 and 1X protease inhibitor cocktail (Pierce) and incubating on ice for 10 minutes and then centrifuging at 2000 rpm 4°C for 15 min. The resulting nuclei were then lysed with M-PER buffer (Thermo Scientific), then diluted 10X with 50 mM Tris-HCl pH 7.4, 100 mM NaCl, and supplemented with protease inhibitor cocktail (Pierce). After incubation, RNA was eluted from the beads with buffer containing 16 mM biotin. A fraction of the supernatant was resuspended in 4X LDS buffer for western blot analysis or used for mass spectrometry.

For mass spectrometry analysis, eluted proteins samples were reduced (10 mM DTT, EMD millipore) at 56°C for 45 minutes followed by alkylation in the dark (30 mM iodoacetamide, Sigma). Proteins were digested with 20 ng/µL Endoproteinase LysC (Wako Chemicals) and trypsin (Promega) and halted by acidification. Protein digests was desalted and analyzed by data dependent acquisition using C18 reversed phase nano-LC-MS/MS (Ultimate 3000 coupled to a QExactive Plus, Thermo Scientific) as described previously (*56*). Data was quantified and queried against a Uniprot human database (January 2013) using MaxQuant (version 1.5.0.30) and ProteomeDiscoverer v. 1.4.0.288. Oxidation of methionine and protein N-terminal acetylation was allowed as variable modifications, cysteine carbamidomethyl was set as a fixed modification, and three missed cleavages was allowed. Mass accuracy was set to better than 5 ppm for precursors and False Discovery Rate of 1%.

Absolute quantification of SNRPA1 by mass spectrometry

Intensity-based absolute quantification (iBAQ) of proteins was carried out based on (*61*), with some modifications. Briefly, 400,000 MDA-231-parental cells were lysed in 4% SDS, Tris-HCl pH 7.5, sonicated, and the protein concentration was quantified via BCA assay (Thermo). 17.5 µg of total lysate, corresponding to 40,000 cells, was then spiked in with 5.3 µg of Universal Proteomics Standard Set (UPS2, Sigma-Aldrich), which consists of 48 human proteins formulated into a dynamic range of known amounts spanning six orders of magnitude. An equivalent amount of lysate without the spiked in standard set was taken as control. Both samples were then reduced by the addition of 100 mM DTT and heating for 5 mins at 95˚C, followed by Trypsin digestion and desalting, as previously described (*62*). MS analysis was also performed as previously described (*62*), using a Q-Exactive plus Orbitrap mass spectrometer coupled with a nanoflow ultimate 3000 RSL nano HPLC platform (Thermo Fisher). Each sample was run twice to obtain two technical replicates. Raw MS files were then searched using MaxQuant (version 1.6.3.3) (*63*), with the iBAQ option enabled, and enzyme specificity set to "Trypsin". Default MaxQuant parameters were applied to all other settings. All downstream MS data analysis was performed using Perseus (version 1.6.2.3) (*64*). Briefly, "iBAQ" intensities were log-transformed, median normalized, and the missing values were replaced by imputation using a downshift of 2 and a width of 0.3 standard deviations. The iBAQ values were then averaged between technical replicates and the inverse Log was calculated. The resulting values in the spiked-in averaged runs were then subtracted from the control averaged runs to calculate the corrected iBAQ values for the spiked-in standard proteins. Linear regression was then used to fit the Log of the corrected iBAQ intensities to Log of absolute spiked-in standard protein amounts. The slope and intercept from the linear regression analysis was then used to convert iBAQ intensities to molar amounts for every identified protein in each run. Uncertainty of the fit was calculated by bootstrapping to provide upper and lower ranges. Cellular copy numbers were then calculated by multiplying the estimated molar values by the Avogadro constant, followed by dividing by the number of cells used in the experiment (40,000). Copy numbers for a total of 3,584 proteins were calculated.

Targeted DMS-MaPseq

DMS-MaPseq of the S3Es in ERRFI1 and PLEC was performed as described in (*33*). Briefly, MDA-LM2 cells were incubated in culture with 1.5% DMS (Sigma) at room temperature for 7 minutes, the media was removed, and DMS was quenched with 30%

BME. Total RNA from DMS-treated cells and untreated cells was then isolated using Trizol (Invitrogen). RNA was reverse transcribed using TGIRT-III reverse transcriptase (InGex) and target-specific primers. PCR was then performed to amplify the desired sequences and to add Illumina compatible adapters. The libraries were then sequenced on a MiSeq instrument using MiSeq micro kit v2, 300 cycles (Illumina). See Table S1 for oligo sequences used in library preparation.

Pear (v0.9.6) was used to merge the paired reads into a single combined read. The UMI was then removed from the reads and appended to read names using UMI tools (v1.0). The reads were then reverse complemented (fastx toolkit) and mapped to the amplicon sequences using bwa mem (v0.7). The resulting bam files were then sorted and deduplicated (umi_tools, with method flag set to unique). The alignments were then parsed for mutations (CTK). The mutation frequency at every position was then reported.

CLIP-seq

CLIP-seq for endogenous SNRPA1 in MDA-LM2 cells was performed using irCLIP (*39*), with the following modifications. MDA-LM2 cells were crosslinked with 400 mJ/cm^2 254 nm UV. Cells were lysed in CLIP lysis buffer (1X PBS, 0.1% SDS, 0.5% sodium deoxycholate, 0.5% IGEPAL CA-630) supplemented with 1X protease inhibitors (Thermo Scientific) and SuperaseIN (Thermo Scientific), then treated with DNase I (Promega) for 5 minutes at 37ºC. Lysate was clarified by spinning at 21,000 x g at 4ºC for 15 min. RNA-protein complexes were then immunoprecipitated from the clarified lysate using protein A dynabeads conjugated to anti-SNRPA1 (Bethyl A303-948A) for 2 hours at 4°C. Beads were washed sequentially with high stringency buffer, high salt buffer and low salt buffer. RNA-protein complexes were then nuclease treated on-bead with RNase A (Thermo Scientific), and then ligated to the irCLIP adaptor using T4 RNA ligase (NEB) overnight at 16ºC. RNA-protein complexes were then eluted from beads, resolved on a 4-12% bis-tris NuPAGE gel, transferred to nitrocellulose, then imaged using an Odyssey Fc instrument (Licor). Regions of interest were excised from the membrane and the RNA was isolated by proteinase K digestion followed by pulldown with oligo d(T) magnetic beads (Thermo Scientific). The resulting RNA was then reverse transcribed using superscript IV (Invitrogen) and a barcoded RT primer, purified using MyOne C1 dynabeads (Invitrogen), and then circularized using CircLigase II (Epicentre). Two rounds of PCR were then performed to first amplify the library using adaptor-specific primers and to add sequences compatible with Illumina sequencing instruments. The libraries were then sequenced on an Illumina HiSeq4000 instrument at UCSF CAT.

CLIP-seq for endogenous SF3B1 in MDA-LM2 cells was performed using HITS-CLIP as in (*65*) with the following modifications: MDA-LM2 cells were crosslinked with 400 mJ/cm^2 254 nm UV, and immunoprecipitation was performed with protein A Dynabeads conjugated to SF3B1 antibody (Bethyl A300-997A). After the washes, on-bead dephosphorylation was carried out using T4 PNK (NEB) in PNK dephosphorylation buffer (50 mM Tris-Cl pH 6.8, 5 mM $MgCl_2$, 5 mM DTT). To enable infrared detection of RNA-protein complexes on the membrane the RL3 linker was conjugated to IR800 dye at the 3' end -- the RL3 RNA oligo was synthesized with a 3' azide modification (IDT), and IRDye 800CW-DBCO (Licor) was conjugated to this azide-RL3 oligo via a click reaction by incubating the azide-RL3 with IR Dye 800CW-DBCO in 1X PBS at 25ºC for 4 hours, then purified using SPRI beads. The unlabeled RL3 linker was synthesized with a 3' inverted dT modification (IDT). An Odyssey Fc instrument (Licor)

was used to image the RNA-protein complexes after running on a 4-12% bis-tris NuPAGE gel (Invitrogen) and transfer to Protran BA85 nitrocellulose (Cytiva). After 5' linker ligation, DNase I treatment and reverse transcription, the optimal number of PCR cycles for PCR1 and PCR2 was determined using qPCR by adding SYBR green I and ROX dye (Thermo) to aliquots of the PCR mixture. Subsequently, both rounds of PCR products were size selected and purified using SPRI beads. The final libraries were run on an Agilent TapeStation, and then sequenced on an Illumina HiSeq4000 instrument at UCSF CAT.

The irCLIP data was processed using the CTK package. Briefly, reads were divided based on their internal barcode into two separate replicates. The resulting fastq files were then collapsed and subjected to adapter removal (cutadapt). Barcodes were then stripped and the reads were aligned to the human genome (build hg19) using BWA (bwa aln). PCR duplicates were then removed and mutations were tallied. CTK, Piranha, and clipper (v0.1) were independently used to call peaks. Peaks called by clipper were used for the analyses reported in this paper. SF3B1 HITS-CLIP data were similarly processed and peaks were called using CTK.

RNA EMSA

RNA probes were generated using a MegaScript T7 transcription kit (Invitrogen) to in vitro transcribe PCR-synthesized DNA templates of the S3Es contained in ERRFI1, PLEC, and full-length RNU2, all with a 5' addition of a T7 promoter. IVT reactions were cleaned up by spin column (Zymogen). The resulting RNA was 3' biotinylated by first denaturing 50 pmol of RNA in the presence of 20% DMSO at 85°C for 3 minutes, then immediately placing on ice. Biotinylated cytidine bisphosphate was then ligated to this RNA in a reaction buffer containing 1X RNA ligase buffer (NEB), 1 mM ATP, 1 µl SuperaseIN, 1 nmol pCp biotin (Jena Biosciences), 2 µl T4 RNA ligase (NEB) and 15% PEG 8000 at 16°C overnight. The labeled RNA was then PAGE purified before use in the EMSA assay.

For the binding reactions, the RNA was denatured by heating at 90°C for 2 minutes in 10 mM Tris-HCl pH 7.5, 100 mM KCl, and then immediately placing on ice. $MgCl_2$ was then added to a final concentration of 5 mM, and RNA was heated at a rate of 1 degree per minute to 30°C. For each binding reaction, 20 fmol of each folded RNA was added to a solution containing 20 mM HEPES pH 7.9, 1 mM EDTA, 125 mM KCl, 5% glycerol, 0.1% Triton X-100, 1 mM DTT, 0.1 µl SuperaseIN. Non-specific competitor RNA was also added where indicated, and consisted of 60 nucelotide RNA oligos containing randomized sequences of A and C. To this, 1 pmol of recombinant SNRPA1 protein (Abcam) was added, and the binding reaction was incubated at 30°C for 30 minutes. Glycerol was added to these reactions to a final concentration of 10%, and the reactions were loaded onto a pre-chilled 6% acrylamide gel and run in cold 0.5X TBE. The RNA-protein complexes were transferred to biodyne b nylon (Pierce) in 0.5X TBE and then detected using an HRP-based chemiluminescent nucleic acid detection module kit (Thermo Scientific).

pyTEISER

pyTEISER identifies RNA structural motifs that are informative of changes in gene expression, RNA stability or other quantitative transcriptomic measures. pyTEISER is based on TEISER software and includes a number of critical architectural changes. First,

pyTEISER code base is fully modular: every step of the pipeline is implemented in a separate module which is independent of the other modules. This gives a major speed improvement for research groups that focus on studying a single organism, because it allows the most time-consuming steps to be run just once for a given genome with further refactoring of pre-processed intermediate files. Second, pyTEISER is implemented in Python, which allows easy incorporation in commonly used Python-based scientific pipelines. At the same time, the usage of numba JIT compiler makes pyTEISER efficiency comparable to that of original TEISER which was implemented in C. Third, pyTEISER offers a flexible framework that makes use of RNA secondary structure probing data, such as SHAPE or DMS-seq.

Seed definition: the base unit describing instances of a secondary structure element is called a seed. A seed is an element of length N that has a pre-defined secondary structure (generally, stem-loop) and a degenerate sequence. An example seed would have a secondary structure of '<<<<.....>>>>' and a sequence of 'AAUNNGNGNUNAUU', where N can be any nucleotide. For any given seed, any sequence A of length N can be unambiguously identified as a match, and it can potentially fold into the seed's secondary structure, forming only canonical base pairs or a G-U wobble base pair. If any of these criteria are not met, sequence A is considered a non-match. This definition of seeds is inspired by the concept of context-free grammars, which was implemented in the earlier versions of TEISER. This framework allows for fast scanning of any sequence of length M for matches with a rolling window of length N. In theory, in the worst-case scenario the scanning algorithm takes quadratic time, in practice all the possible analyzed seeds are short (less than 17 nt) long so the algorithm works in $O(C \times M) = O(M)$.

Seed occurrence profile: as an input, pyTEISER uses a set of K sequences together with corresponding transcriptomic measurements. For each of these sequences, we can determine unambiguously if this sequence contains any matches to a given seed. This search can be either performed based solely on primary sequence or it can incorporate secondary structure information, such as RNA SHAPE data, DMS-seq data, or *in silico* RNA folding data. The result of this search is then represented as a binary vector of length K. Each element is set to "1" if the corresponding sequence has a match to the given seed or "0" if it doesn't. The resulting binary vector is called a "seed profile"; it represents which sequences can potentially contain a match to sequence and structural rules described by a given seed.

Structural element profiles: A given structural elements is best viewed as a collection of seeds that together capture the required heterogeneity in its form. This is akin to the relationship of a set of k-mers forming a degenerate sequence motif. In pyTEISER, we use the term "motif" to describe a set of seeds that have similar seed occurrence profiles and likely capture different facets of the same structural element. Formally, a motif is a set of seeds, and its occurrence profile is an overlap of all of the motif's seed occurrence profiles (OR rule).

Generating seeds: pyTEISER generates a comprehensive set of short seeds of stem-loop structures that meet several criteria regarding their stem length, loop length and information content. The information content of the seed X is defined as the negative logarithm of P(X), where P(X) is the probability that a random sequence of the same length matches the seed X. The default boundaries for seed creation are: stem length between 4 and 7 nucleotides, loop length between 4 and 9 nucleotides, number of non-

degenerate nucleotides in seed's sequence between 4 and 6, and information content between 14 and 20 bits. These criteria result in ~70 million seeds.

Sequence scanning: each generated seed is scanned across the provided sequences, resulting in the generation of a seed occurrence profile. Additional criteria can be included to make the seed profiles more specific. First, information provided by *in silico* folding algorithms can be used to exclude the potential seed matches that are unlikely to fold in the predicted way. ViennaRNA software is used to predict folding for the RNA sequence in a window flanking the seed match. Those matches for which the predicted folds differ from seed's secondary structure are excluded. Second, RNA SHAPE data, when available, can be used to guide selection of matches that are specific for a certain biological condition. Nucleotide-based resolution reactivity profiles are included as a pseudo-free energy term which allows higher accuracy of RNA secondary structure predictions (*66*).

Calculation of mutual information values: each seed's occurrence profile is tested to assess whether it is informative of the input transcriptomic measurements. To capture such dependency, we use Mutual Information (MI). Since calculation of MI requires both input vectors to be discrete, we discretize the provided continuous genome-wide measurements (for example, expression values, percent-spliced-in values, etc) in a pre-defined number of equal frequency bins.

Randomization-based statistical testing: pyTEISER runs several non-parametric statistical tests to determine which seeds are significantly informative of expression changes and which ones are not. The tests include (i) a permutation-based estimation of p-value with a null-hypothesis of the expected MI value for a given seed not being different from the expected MI value of a random occurrence profile with the same number of matches and (ii) a permutation-based estimation of the z-score for the MI value for a given seed. All the seeds are sorted by their MI values and a greedy search algorithm identifies the threshold MI value; all the seeds whose MI values exceed the threshold are considered "passed". The threshold is defined as a value of MI at which fewer than 9 out of 10 first seeds below the threshold satisfy the threshold for p-value and/or z-score.

Defining structural elements: the initial random sampling of seeds generated by pyTEISER may be functionally redundant. In other words, several seeds can be very similar to each other and match the same or similar sequences. Such redundant groups of seeds are clustered into structural elements or motifs. Each seed (Q) is compared to a set of seeds previously considered (R) to decide if this seed should form its own motif or if it should be added to one of the existing motifs. For this, we calculate conditional mutual information of occurrence profile of Q with the input data given the occurrence profile of R. We then divide this value by the mutual information of the occurrence profile of Q and the occurrence profile of R. If the resulting value is higher than an indicated threshold (chosen by the user), we conclude that Q and R are two representations of the same underlying element.

Seed optimization: pyTEISER aims to find and describe the structural element(s) that would best explain transcriptomic measurements. Since seeds provide a coarse-grained sampling of the search space, any seed that passes the statistical tests may be suboptimal. Therefore, for each structural element, a representative seed is selected for further optimization in a two-stage algorithm. The first stage involves modification of each

position of the seed. Modifications include replacing the original nucleotide with all possible degenerate nucleotides (15 in total), or changing the base pairing state of the given position. Positions of the seed are modified in a random order to avoid biases. For the seeds resulting from such modification process the occurrence profile is built, and then mutual information of the occurrence profile with the genome-wide measure of interest is calculated. The seed with the best mutual information value is retained. The second stage involves seed elongation. A seed is elongated on both ends using a greedy search to maximize MI values. At each step of elongation all the possible degenerate nucleotides and both base pairing states for the new position are evaluated. Therefore, the optimized seed is guaranteed to have the mutual information value not less than that of the initial seed.

Robustness test: in order to test if the optimized seeds or motifs are robust to changes of the input data, the statistical tests are repeated with down-sampled input data. By default, the tested seed is required to pass p-value and z-score thresholds for at least 6 trials out of 10, where in each trial only 2/3 of the input data is retained.

Calling SSE target exons: For the purposes of this study, we pooled the top performing seeds that represented similar motifs (based on their conditional mutual information ratio) and scanned cassette exons (and 250nt flanking) for matches to this family. The resulting matches were then *in silico* folded using the Vienna package (RNAfold) and sequences whose predicted secondary structure did not match SSE elements were filtered. Subsequently, those sequences with folding energy one-standard below average were selected as final SSE matches.

Implementation and availability: pyTEISER is written in Python 3. The source code and documentation are available at https://github.com/goodarzilab/pyteiser. Precompiled version is available at PyPi at https://pypi.org/project/pyteiser/. The source code and documentation for pyTEISER has been deposited at Zenodo (*56*).

Statistics

The relevant statistical tests for each sup-plot are included in the legends.

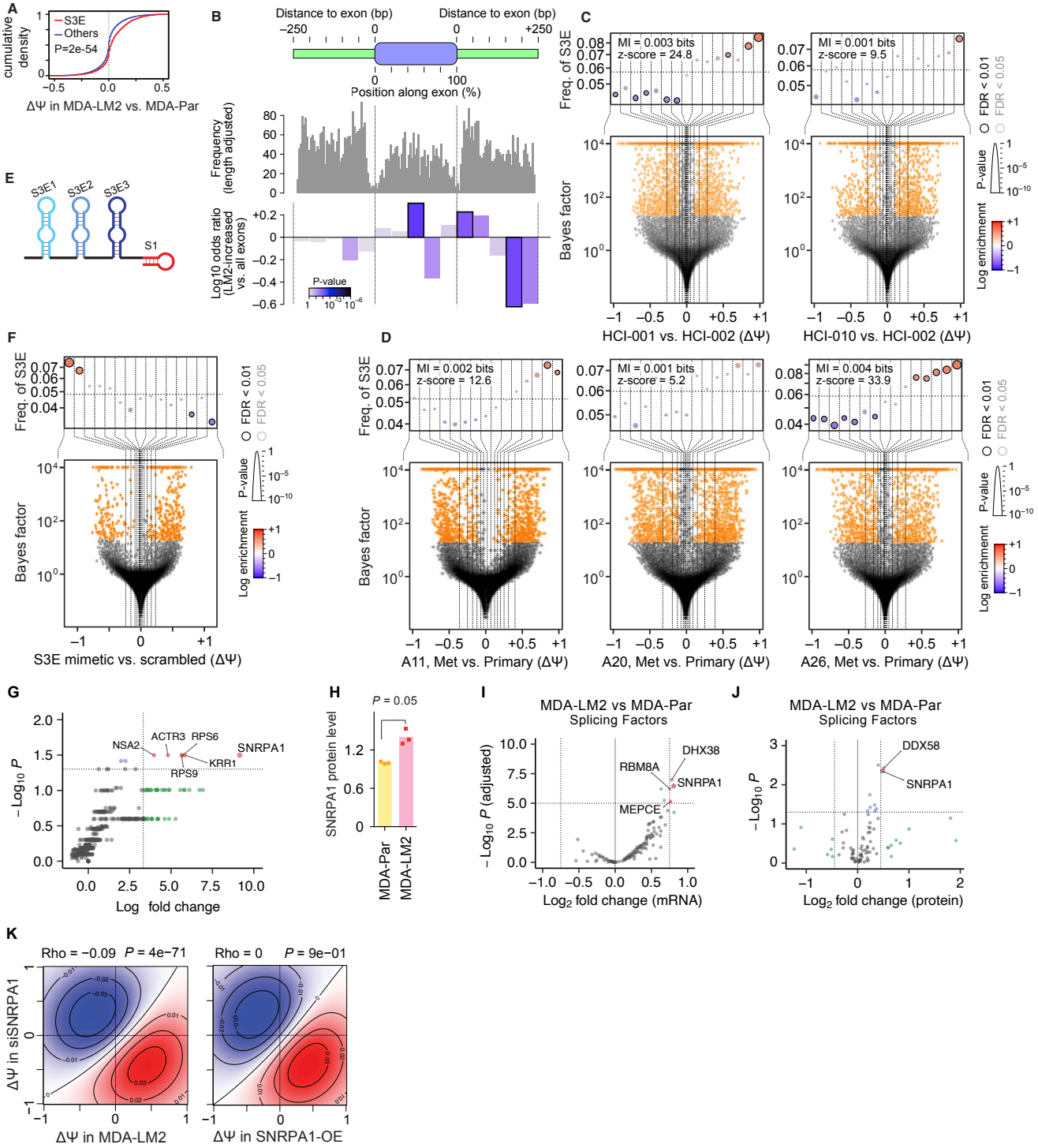

Fig. S1.

SNRPA1 interacts with RNA structural element S3E (SNRPA1-associated structural splicing enhancer) and regulates cassette exon splicing. (**A**) Cumulative distribution plot showing a significant increase in splicing of exons harboring S3Es compared to those without S3Es. P-value calculated using one-tailed Mann-Whitney U-test. (**B**) Top: Schematic representation of S3E-harboring cassette exon and flanking

intronic regions. Middle: Histogram of S3E locations along the cassette exon and flanking introns. Bottom: Relative enrichment of S3Es at each location, comparing S3E instances in exons with increased inclusion in MDA-LM2 vs parental cells to those that occur in non-regulated exons. The color gradient shows the P-value for enrichment or depletion of S3E (Fisher's exact test). Regions that show a significant difference between metastasis-associated exons and other exons are highlighted with a black border (FDR < 0.05). (**C**) Bottom: Volcano plots showing distribution of relative changes in alternative splicing (change in percent spliced in (ΔΨ)) in highly metastatic PDX HCI-001 or HCI-010 compared to poorly metastatic PDX HCI-002. Top: Enrichment of the S3E RNA structural element in exons (and flanking intronic sequences) as a function of ΔΨ between HCI-001/HCI-10 and HCI-002. See Fig. 1A for description of enrichment plots. (**D**) Volcano plots showing differential alternative splicing between metastatic and primary tumors in three triple negative breast cancer patients (bottom), as well as the enrichment of S3E element in exons (top). (**E**) Schematic of synthetic RNA oligo used to co-precipitate S3E-interacting proteins in MDA-LM2 cell lysate, which were then identified by mass spectrometry. (**F**) MDA-LM2 cells were transfected with the S3E3X mimetic depicted in **E**. Bottom: Volcano plot showing distribution of relative changes in alternative splicing (change in percent spliced in (ΔΨ)) in cells transfected with the S3E3X mimetic oligo compared to a control scrambled oligo. Top: Enrichment of the S3E RNA structural element in exons (and flanking intronic sequences) as a function of ΔΨ. (**G**) Volcano plot showing the relative enrichment of S3E3X mimetic oligo interacting proteins in MDA-LM2 cell lysate, as determined by mass spectrometry. P-values calculated using one-tailed Mann-Whitney U-test. (**H**) Bar graphs showing relative SNRPA1 protein level in MDA-parental and MDA-LM2 cells as measured by label free quantitation mass spectrometry, respectively. *n*= 3 biological replicates. P-value calculated using one-tailed Mann-Whitney U-test. (**I-J**) Volcano plots showing the relative splicing factor mRNA (I) and protein (J) abundance in MDA-LM2 cells, compared to MDA-Par, as determined by RNA-seq and label-free quantitation mass spectrometry, respectively. P-values calculated using SESeq2 (I) and t-test (J). **(K)** Two-dimensional heatmaps showing significant anti-correlations of SNRPA1 knockdown-dependent ΔΨ and ΔΨ in MDA-LM2 vs MDA-parental cells and with SNRPA1 overexpression-dependent ΔΨ. The Spearman correlation coefficient and the associated P-value are shown.

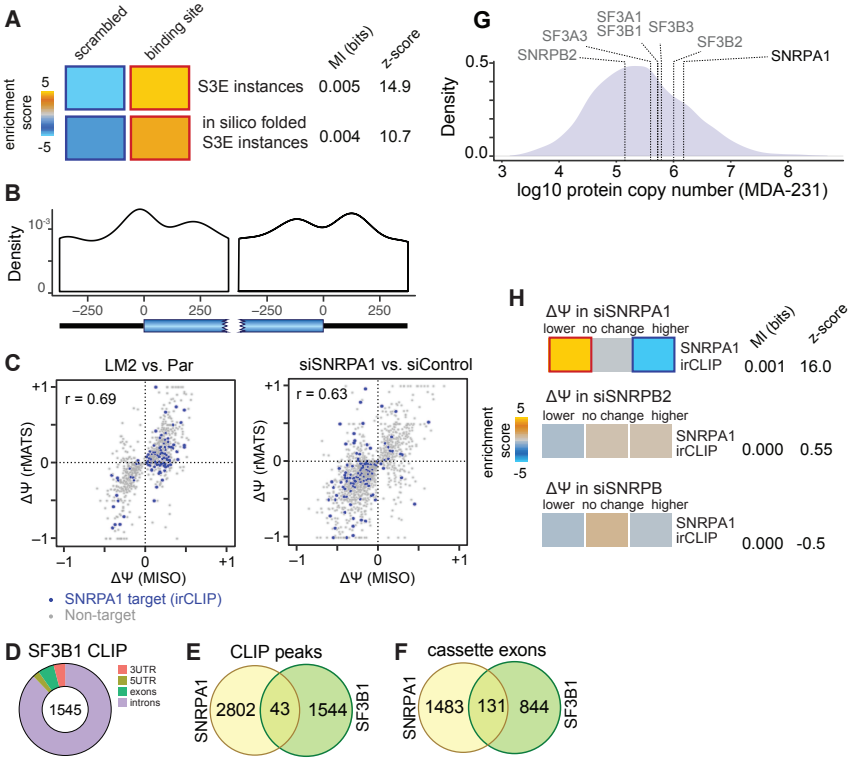

Fig. S2

SNRPA1 controls alternative splicing of a limited set of genes. (**A**) Heatmaps showing enrichment of S3Es in SNRPA1 binding sites identified from CLIP (top), and showing enrichment of matches to S3E among in silico-folded SNRPA1-bound sites identified from CLIP (bottom). Mutual information (MI) value and associated z-score are shown. (**B**) Meta-exon density plot of SNRPA1 irCLIP clusters, showing regions of 500 nt centered around 5' and 3' splice sites, respectively. (**C**) Each scatterplot corresponds to differential splicing in a specific comparison (left: MDA-LM2 cells vs. parental MDA cells; right: SNRPA1 knockdown vs. control cells). Each dot represents one cassette exon, with the x-axis corresponding to splicing changes as quantified by MISO and the y-axis representing splicing changes as quantified by rMATS (using the same RNA-seq data). Blue dots represent cassette exons bound by SNRPA1 (based on irCLIP data). Only exons with MISO Bayes factor larger than 1.0 are shown. (**D**) Chart illustrating the distribution of genomic features among SF3B1 CLIP peaks. (**E**) Venn diagram showing the overlap of SNRPA1 and SF3B1 CLIP peaks. (**F**) Venn diagram showing the overlap of SNRPA1- and SF3B1-bound exons that have a significant increase in cassette exon inclusion in MDA-LM2 cells compared to MDA-Parental cells. (**G**) Density plot showing the distribution of absolute protein quantities in MDA-Parental cells, as determined by intensity-based absolute quantification. U2 snRNP components are indicated. (**H**) Heatmaps showing no significant enrichment of SNRPA1 binding sites in exons (and flanking introns) with decreased ΔΨ in MDA-LM2 cells with siRNA-mediated SNRPB2 (SNRPB'') or SNRPB (SmB/B') knockdown compared to control cells, and a significant enrichment of SNRPA1 binding at these sites in MDA-LM2 cells with SNRPA1 knockdown (using three independently targeting siRNAs) compared to control cells. MI value and associated z-score are shown.

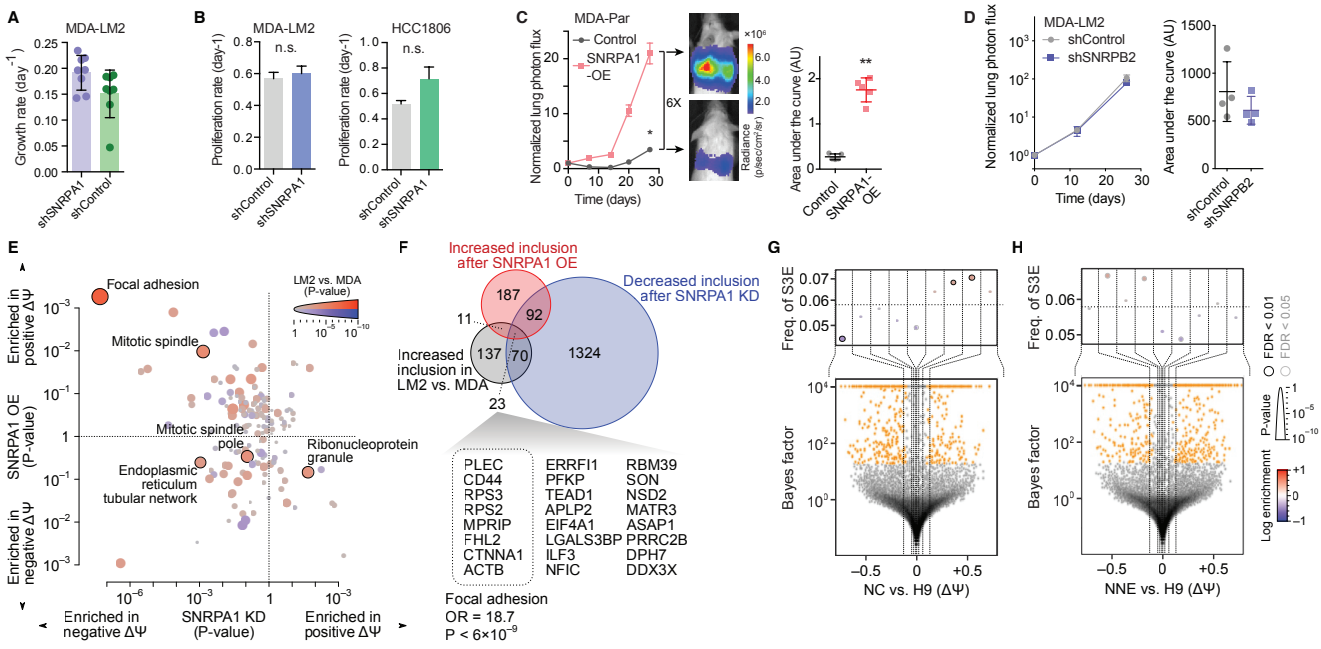

Fig. S3

S3. SNRPA1 promotes metastatic colonization and invasion. (**A**) The orthotopic tumor growth rate of MDA-LM2 cells with shRNA-mediated knockdown of SNRPA1 was measured. $n = 8$ tumors in 4 mice per cohort. (**B**) Comparison of the growth rate of SNRPA1 knockdown and control cells in MDA-LM2 and HCC1806-LM2 backgrounds; bars show mean ± S.E.M; ANOVA used to calculate P; $n = 3$ biological replicates. (**C**) MDA-LM2 cells stably overexpressing SNRPA1 or mCherry (control) were injected via tail vein into NSG mice. Bioluminescence was measured at the indicated times; area under the curve was measured at the final time point. $n = 5$ mice per cohort. **(D)** MDA-LM2 cells stably expressing an shRNA targeting SNRPB2 or a control shRNA were injected via tail vein into NSG mice. Bioluminescence was measured at the indicated times; area under the curve was measured at the final time point. $n = 4$ mice per cohort. Proliferation rate and AU measurements in (A-D) were compared using one-tailed Mann-Whitney U-test. (**E**) Enrichment of GO terms among genes that are differentially spliced as a result of SNRPA1 knockdown (x-axis), as a result of SNRPA1 overexpression (y-axis), or between MDA-LM2 and MDA-Parental cells (dot size and color). Each dot represents one GO cellular component, excluding terms with >500 genes. P-values (axes and dot size/color) are based on Fisher's exact test. GO terms that pass FDR cutoff < 0.05 in the MDA-LM2 vs. MDA-Par comparison are labeled. (**F**) Euler diagram of genes that are differentially spliced in three different contexts. The intersection of all three sets is shown at the bottom. These sets are comprised of genes with exons that have increased inclusion upon SNRPA1 overexpression, have decreased inclusion upon SNRPA1 knockdown, and have increased inclusion in MDA-LM2 vs MDA-parental cells. Genes that are involved in focal adhesion are circled. P-values correspond to Fisher's exact test. (**G**) Enrichment of S3E in differentially spliced cassette exons between neural crest cells (NC) and H9 ESC cells. (**H**) Enrichment of S3E in differentially spliced cassette exons between non-neuronal ectoderm cells (NNE) and H9 ESC cells. Visualizations as in Figure S1C.

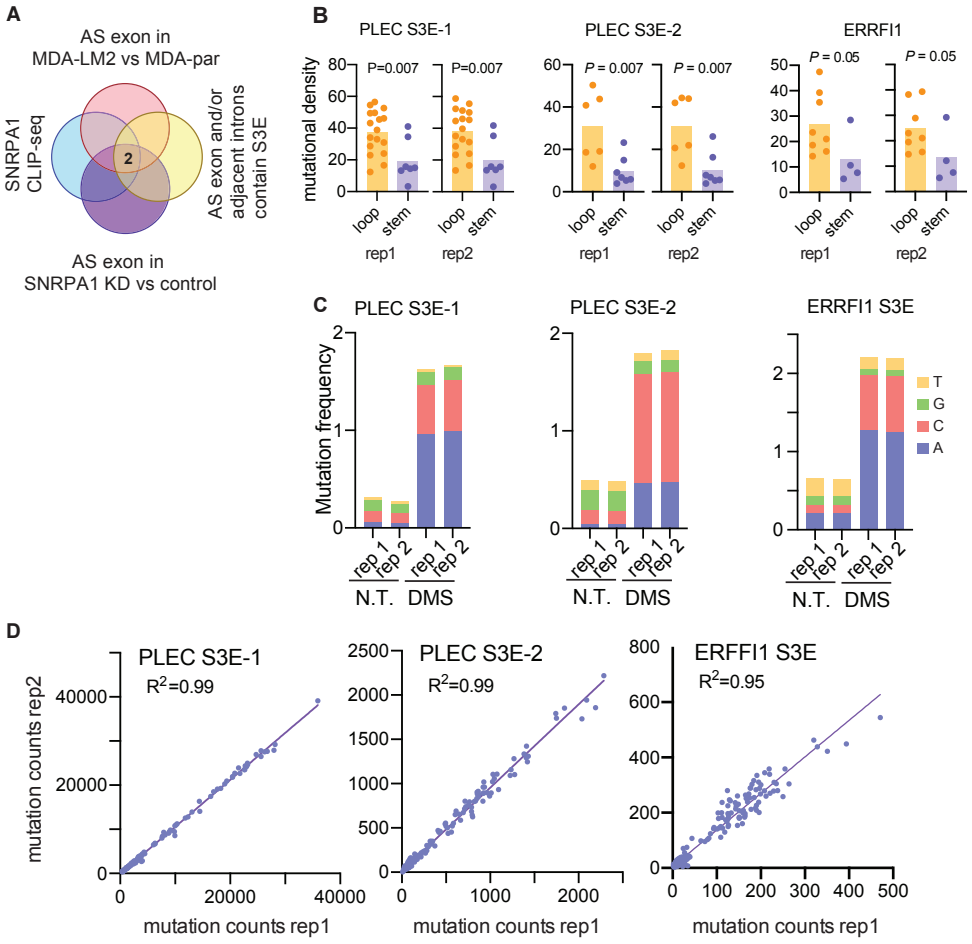

Fig. S4

Targeted DMS-seq of S3E regions is concordant with their *in silico* predicted structures. (**A**) Venn diagram showing the filters used to identify SNRPA1 targets that impact metastasis. Exons that showed more than 10% change in PSI and >5-fold increase in Bayes-factor were selected for this analysis. **(B)** Mutation density in predicted loop and stem regions of PLEC and ERRFI1 S3Es. P-values calculated using one-tailed Mann-Whitney U-test. (**C**) Mutation frequency in DMS-treated and non-treated (N.T.) PLEC and ERRFI1 S3Es, separated by nucleotide. (**D**) Correlation of mutation counts in DMS-seq experimental replicates.

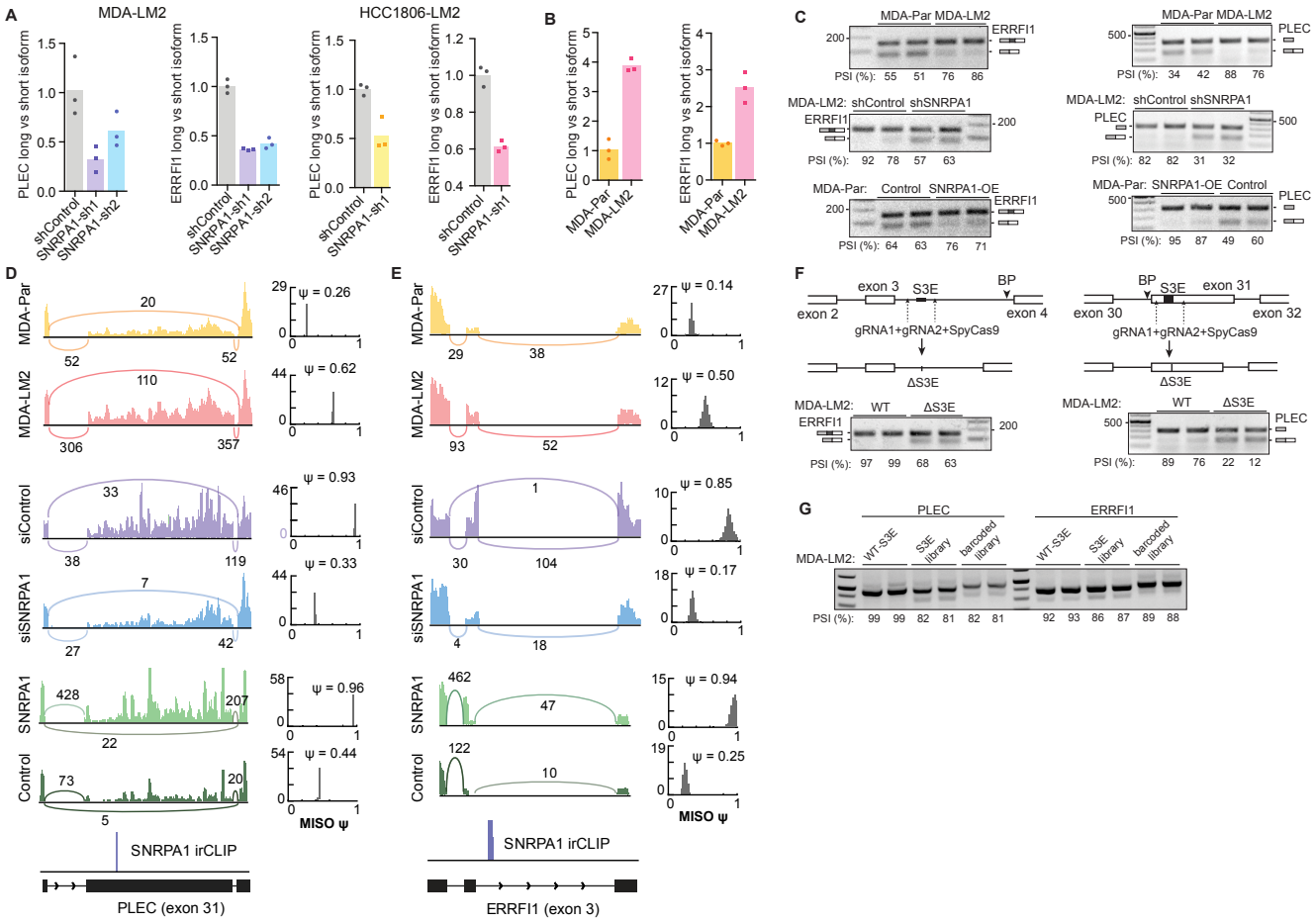

Fig. S5

SNRPA1 modulates alternative splicing of PLEC and ERRFI1. (**A**) Relative levels of PLEC exon 31 and ERRFI1 exon 3 splicing (the ratio of long isoform, i.e. containing the alternative exon, to short isoform) were measured by RT-qPCR in MDA-LM2 and HCC1806-LM2 SNRPA1 knockdown (two independent shRNAs) and control cells; $n = 3$ biological replicates. (**B**) Relative levels of PLEC exon 31 (left) and ERRFI1 exon 3 (right) splicing were measured by RT-qPCR in MDA-parental and MDA-LM2 cells; $n = 3$ biological replicates. **(C)** Percentage spliced in (PSI) of ERRFI1 exon 3 (left) and PLEC exon 31 (right) was determined by RT-PCR and densitometric analysis, in MDA-LM2 and MDA-Parental (top), control and SNRPA1 knockdown MDA-LM2 (middle), and control and SNRPA1 overexpressing MDA-Parental (bottom) cells. Due to PLEC exon 31 length and high GC content, multiplex PCR, amplification of fragments specific for exon 32 (constitutive) and exon 30-32 junction (spliced out), was used. (**D-E**) Sashimi plots (*30*), derived from RNA-seq data, showing relative splicing of PLEC exon 31 (D) and ERRFI1 exon 3 (E), along with CLIP-seq SNRPA1 binding sites from MDA-LM2 cells in MDA-LM2 compared to MDA-parental cells (top); MDA-LM2 SNRPA1 knockdown compared to control cells (middle); and MDA-parental cells overexpressing SNRPA1 compared to those overexpressing mCherry (control) (bottom). For all, MISO Ψ plots show the calculated Ψ of the indicated exons in each cell line. (**F**) Top: Schematic of CRISPR/Cas9-mediated deletion of ERRFI1 exon 3 S3E (left) and PLEC exon 31 S3E (right). Bottom: PSI of ERRFII exon 3 (left) and PLEC exon 31 (right) in MDA-LM2

ΔS3E and control cells, as determined by RT-PCR and densitometric analysis (see (C) for description). (**G**) PSI of reporter cassette exon, comparing PLEC (left) and ERRFI1 (right) wildtype, library variant and barcoded library variant S3Es, determined by reporter-specific RT-PCR and densitometric analysis. See Figure 5B for schematics.

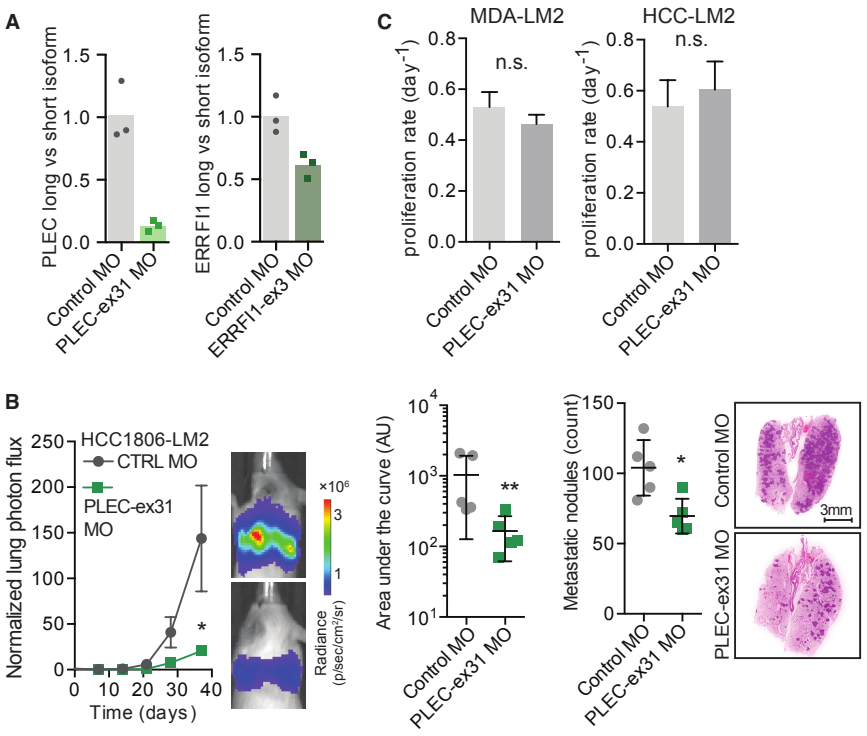

Fig. S6

PLEC isoform switching regulates metastatic capacity of breast cancer cells. (**A**) Levels of PLEC exon 31 (left) and ERRFI1 exon 3 (right) splicing (the ratio of long isoform, i.e. containing the alternative exon, to short isoform) were measured by RT-qPCR in MDA-LM2 cells transfected with exon specific morpholinos (PLEC-ex31 MO, ERRFI1-ex3 MO) or a control MO; $n = 3$ biological replicates. (**B**) HCC1806-LM2 cells transfected with a PLEC-ex31 MO or a control MO were injected via tail vein into NSG mice. Bioluminescence was measured at the indicated times; area under the curve was measured at the final time point. Lungs were stained with H&E and nodules were counted. $n = 5$ mice per cohort. (**C**) Comparison of the proliferation rate of PLEC-ex31 MO and control MO transfected cells in MDA-LM2 and HCC1806-LM2 backgrounds; bars show mean ± S.E.M.; ANOVA used to calculate P; $n = 3$ biological replicates. Proliferation rate, AU and metastatic nodule count measurements in (B-C) were compared using one-tailed Mann-Whitney U-test. *In vivo* bioluminescence measurements in (B) were compared using two-way ANOVA.

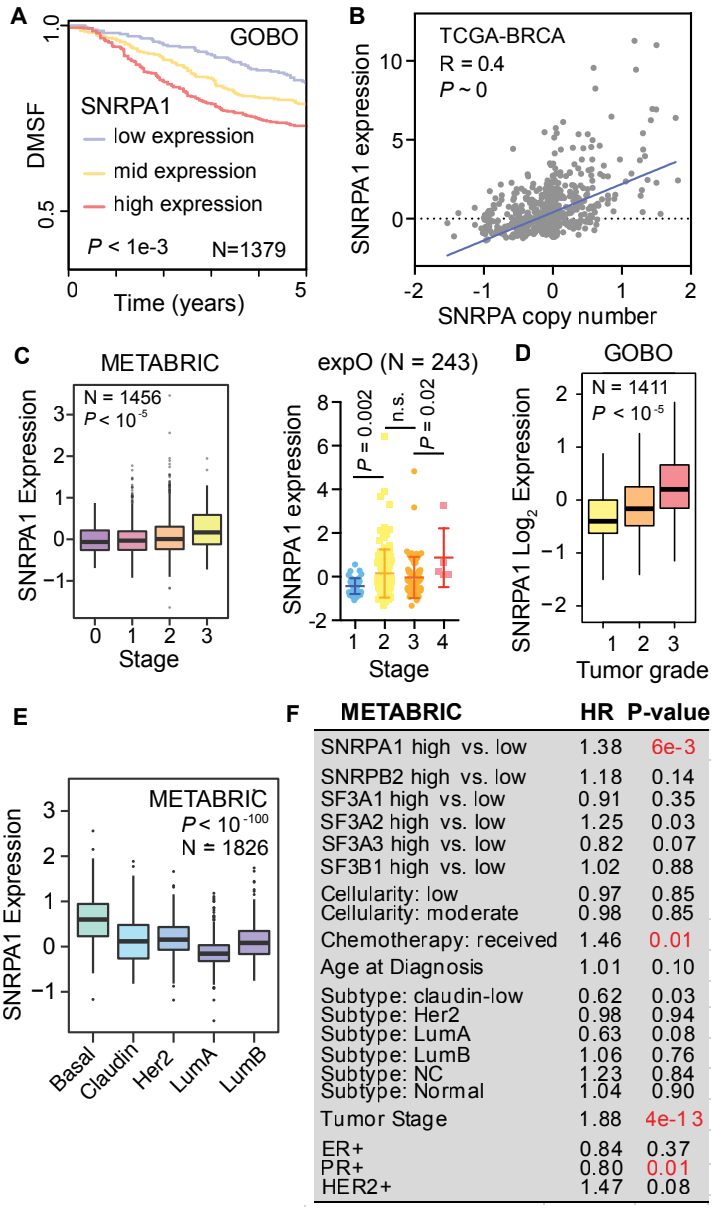

Fig. S7

Clinical associations between SNRPA1 expression and breast cancer progression. (**A**) Survival curve of breast cancer patients stratified by relative SNRPA1 mRNA levels. *n* = 1379. P-value calculated using log-rank test. (**B**) Scatter plot showing a positive correlation between SNRPA1 mRNA levels and SNRPA1 copy number in TCGA-BRCA dataset; *n* = 871. (**C**) SNRPA1 mRNA levels across normal tissue and breast cancer tissue stages I-IV in two independent cohorts; *n* = 1456 (METABRIC) and *n* = 243 (expO; GSE2109). (**D**) SNRPA1 mRNA levels in breast tumors grades 1-3; *n* = 1411. (**E**) SNRPA1 mRNA levels across the indicated breast cancer subtypes. *n* = 1826. (**F**) Hazard ratio and associated *P* values of the listed factors with clinical outcome in METABRIC dataset; n = 1979. SNRPA1 expression levels were compared across groupings using ANOVA. P-value in pairwise comparisons was calculated using one-tailed Mann-Whitney U-test.

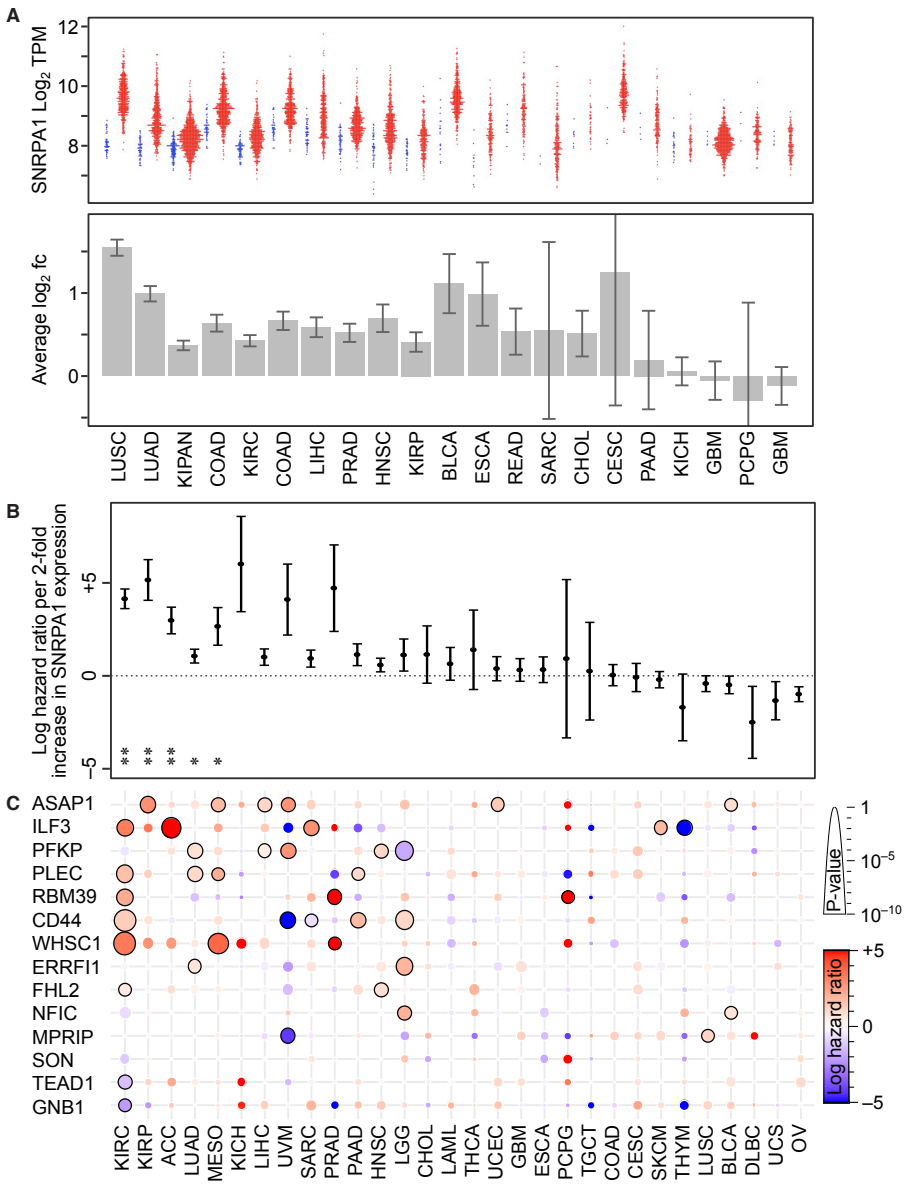

Fig. S8

SNRPA1 expression across non-breast cancer datasets. (**A**) Comparison of SNRPA1 expression between tumor and normal tissues across TCGA cancer types. Top: bee swarm plots showing the distribution of SNRPA1 expression (log_2 of transcripts per million reads, TPM) in normal (blue) and tumor (red) samples. Bottom: Average log_2 fold-change between tumor and normal (error bars represent standard error of mean). (**B**) Association between SNRPA1 expression and overall survival in each TCGA cancer type (Cox regression; error bars represent standard error of mean). Cancer types (x-axis) are in the same order as the next panel. ** FDR < 0.01. * FDR < 0.05. (**C**) Association between expression of SNRPA1 targets and overall survival in each TCGA cancer type (Cox regression) showing relatively little effect on the expression of SNRPA1 targets and survival. The color gradient shows the Cox regression coefficient, whereas the circle size denotes the associated *P*-value. Significant associations (FDR < 0.05) are highlighted with black borders.

Table S1

Nucleic acid sequences used in this study. siRNA, shRNA, gRNA, MO (morpholino oligonucleotide), S3E, oligonucleotide and synthetic DNA sequences used in this study.