SUPPLEMENTARY FIGURES

Figure S1 – related to Figure 1

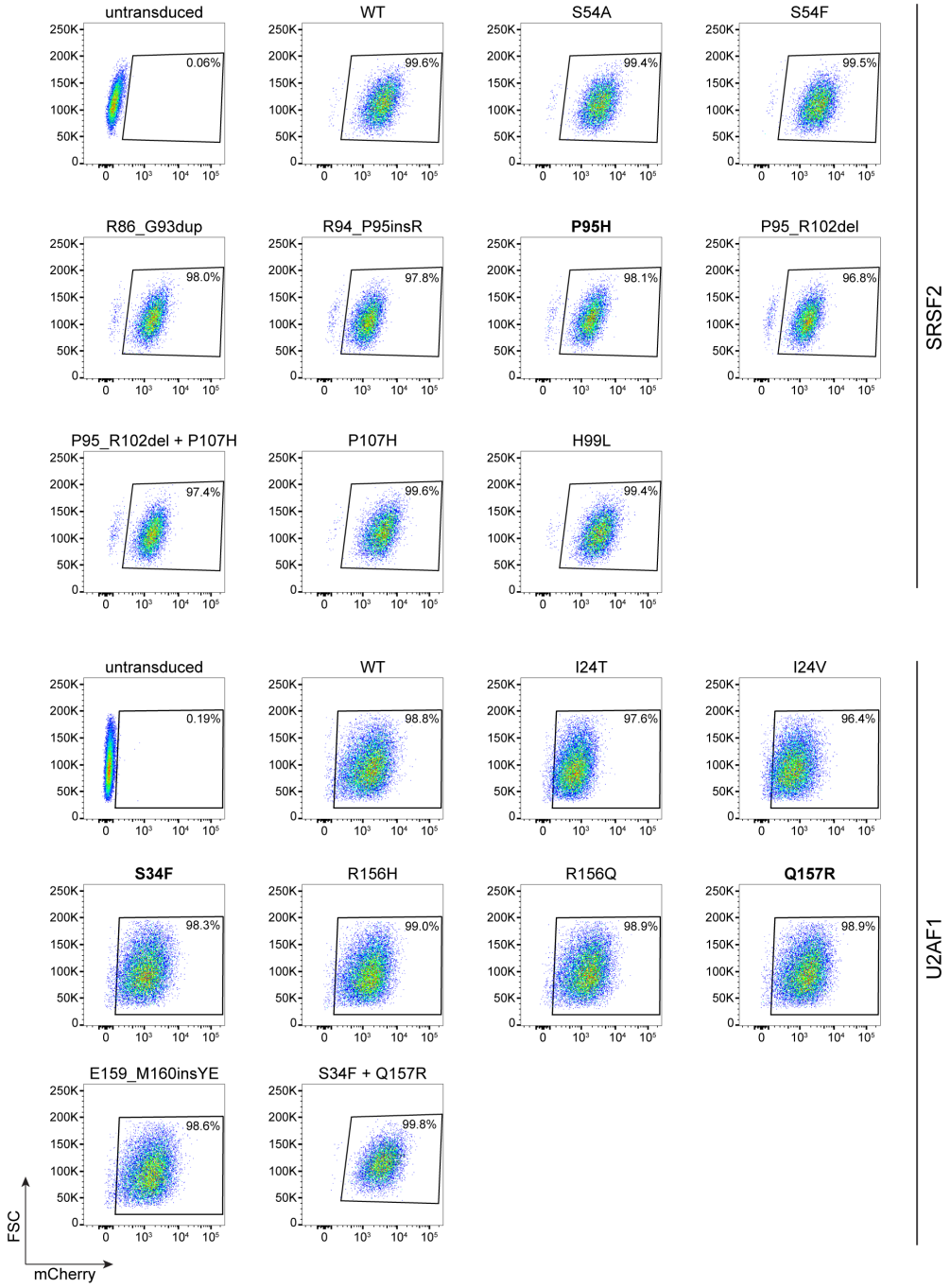

Figure S1 – related to Figure 1. Establishment of K562 cell lines stably expressing transgenic *SRSF2* or *U2AF1*.

Flow cytometry analysis of transgenic *SRSF2*- and *U2AF1*-expressing K562 cell lines. Transgene cassette expresses an mCherry marker. Gates illustrate the populations of transgene-expressing cells that were isolated for further analysis.

Figure S2 – related to Figure 1

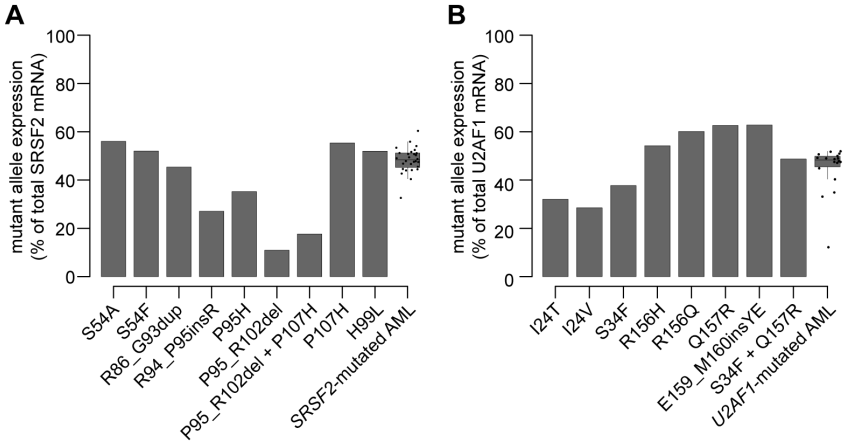

Figure S2 – related to Figure 1. Mutant allele expression in transgenic K562 cell lines. (**A**) Expression of each indicated *SRSF2* allele as a fraction of total SRSF2 mRNA. Box plot indicates mutant *SRSF2* allelic expression in primary AML samples (Lavallée et al, *Nature Genetics*, 2015). Mutant allele expression was computed by RNA-seq for all cases. (**B**) As (**A**), but for *U2AF1* mutations.

Figure S3 – related to Figure 2 and Figure 3

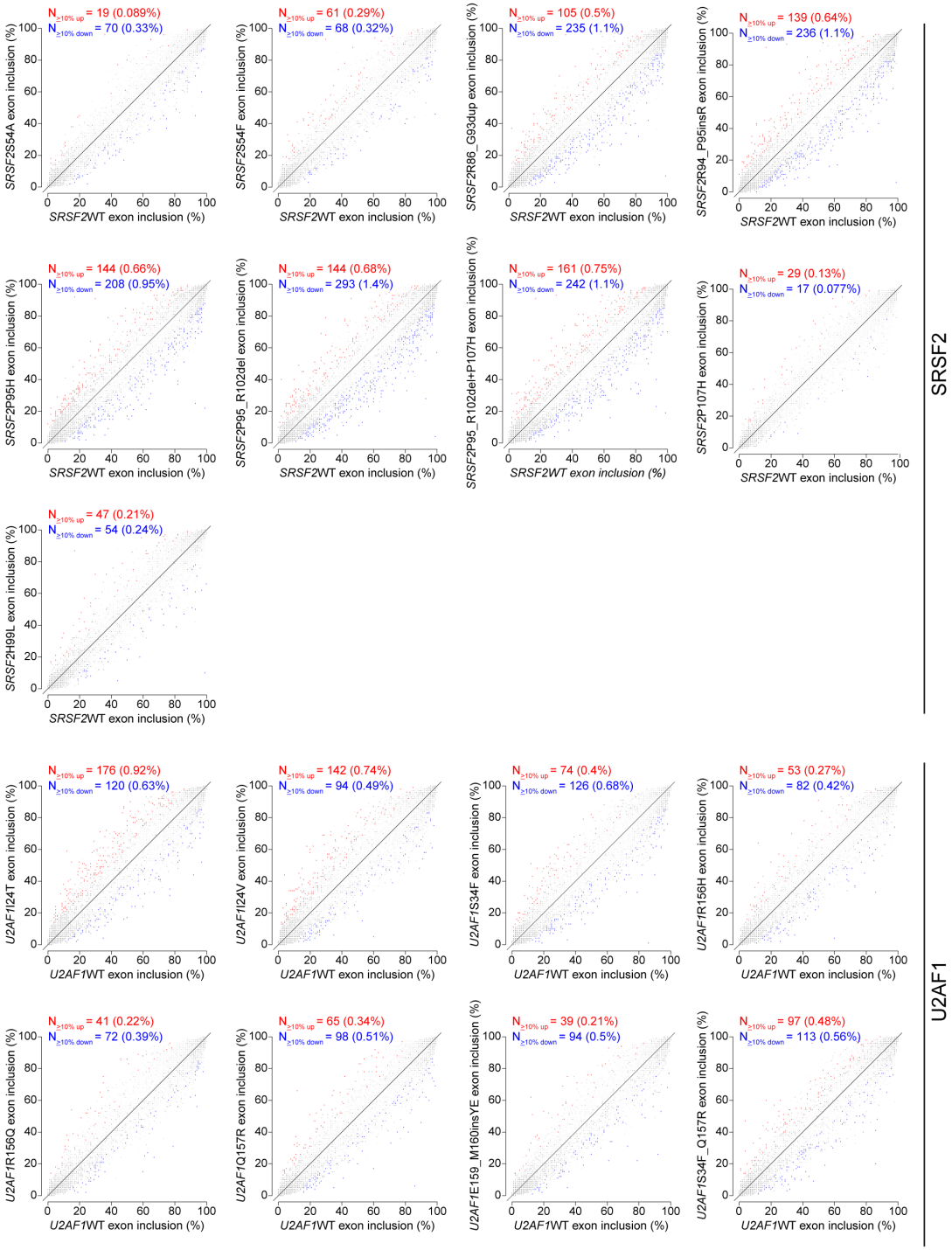

Figure S3 – related to Figures 2 and 3. Cassette exon inclusion in transgenic K562 cell lines. Scatter plots illustrating cassette exon inclusion in K562 cell lines expressing transgenic WT (x axis) or mutant (y axis) alleles of *SRSF2* or *U2AF1*. Red, cassette exons exhibiting significantly increased inclusion in mutant cells; blue, cassette exons exhibiting significantly decreased

inclusion in mutant cells. N, numbers of cassette exons exhibiting significant differential splicing. Percentages are computed with respect to the total numbers of cassette exons exhibiting any evidence of alternative splicing in the analyzed cells. See **Methods** and **Supplementary Methods** for descriptions of identification of significant differential splicing.

Figure S4 – related to Figure 2

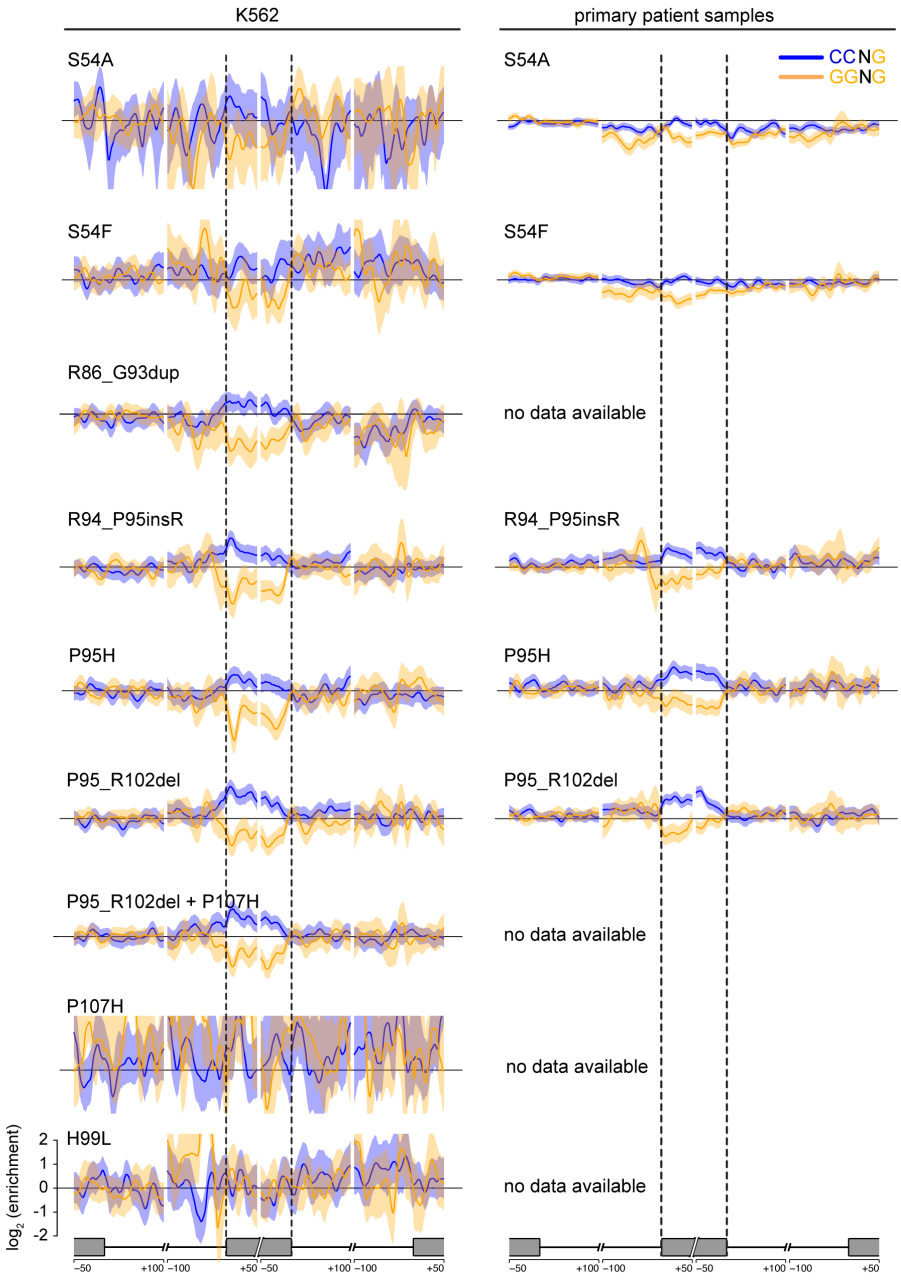

Figure S4 – related to Figure 2. Rare mutations in *SRSF2* alter exonic splicing enhancer (ESE) preference.

Enrichment for CCNG and GGNG motifs within and adjacent to cassette exons that are promoted versus repressed in K562 cells expressing each indicated *SRSF2* mutant allele versus WT-expressing control cells. Samples are identical to those illustrated in **Fig. 2B**. Shading indicates a 95% confidence interval computed as the 2.5th and 97.5th percentiles of enrichment

across all differentially spliced cassette exons for each comparison. Schematic illustrates the genomic loci for which the enrichment analysis was performed; coordinates are defined with respect to the 5' and 3' splice sites, where 0 corresponds to the exon-intron boundaries.

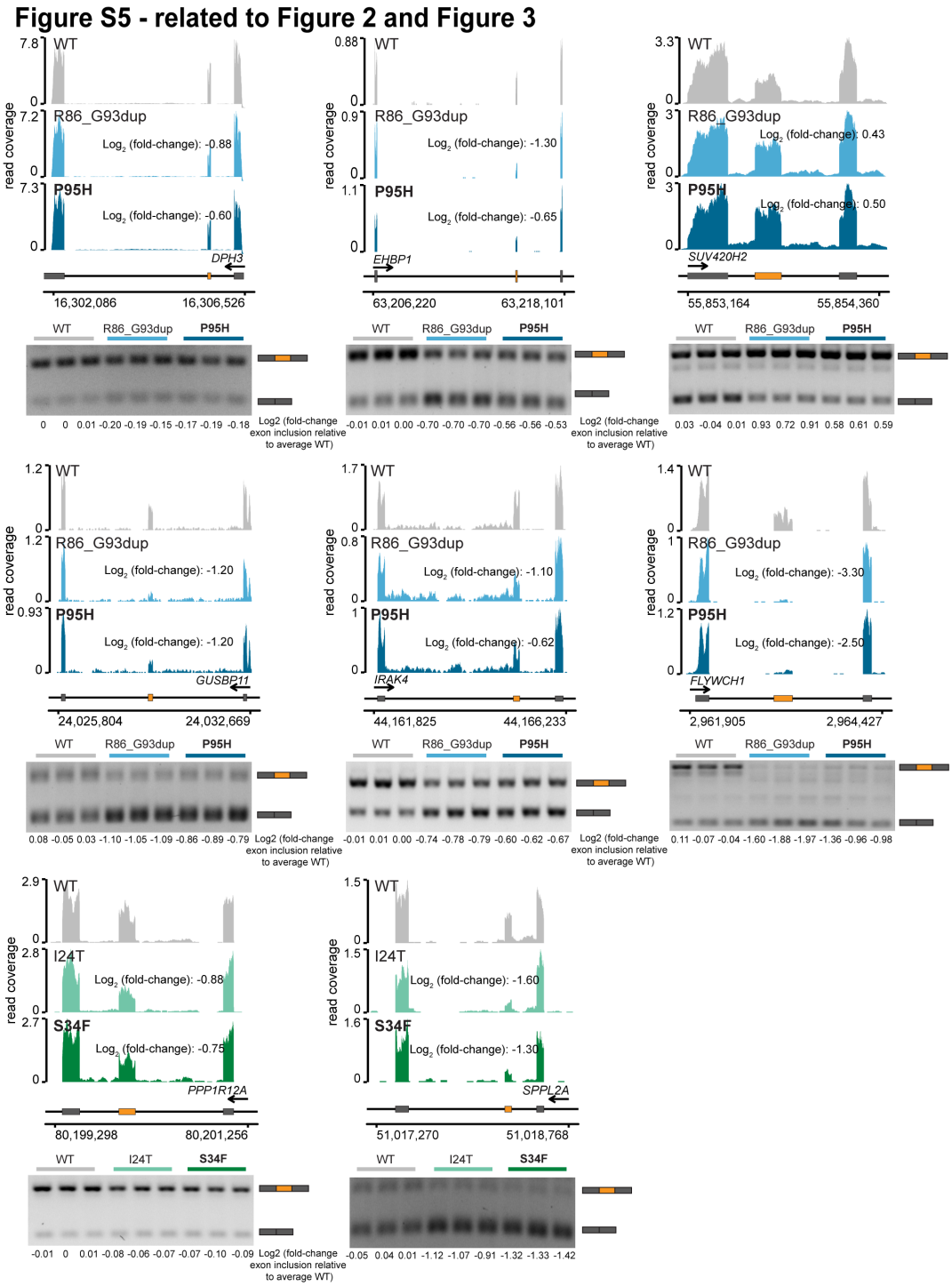

Figure S5 – related to Figures 2 and 3. Rare *SRSF2* and *U2AF1* mutations phenocopy hotspot mutations.

Top, RNA-seq read coverage (top) illustrating increased cassette exon inclusion in the illustrated genes in K562 cells expressing either a hotspot (*SRSF2*P95H or *U2AF1*S34F) or rare

(*SRSF2*R86_G93dup or *U2AF1*I24T) *SRSF2* or *U2AF1* mutation. Log_2 (fold-change) illustrates log_2 (exon inclusion in mutant- versus WT-expressing cells). Bottom, RT-PCR validation of RNA-seq results in technical triplicate.

Figure S6

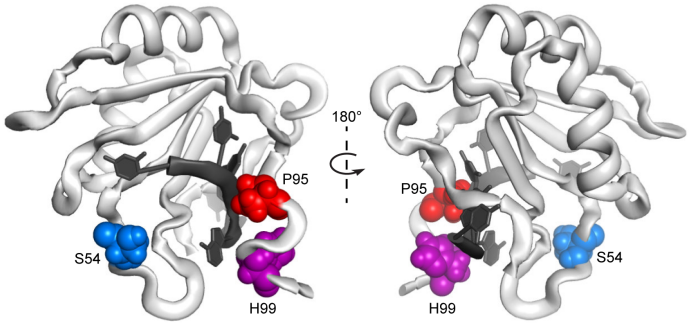

Figure S6. SRSF2 in complex with RNA.

Figure illustrates the solution NMR structure of the SRSF2 RNA recognition motif (PDB ID: 2LEB; Daubner et al, *EMBO Journal*, 2012) in complex with a 5'-UCCAGU-3' RNA substrate (black). Mutations affected the highlighted residues are studied in this manuscript.

SUPPLEMENTARY METHODS

RNA-seq library preparation and sequencing. Total RNA was isolated from K562 cells using the TRIzol reagent. 4 ug of total RNA was then used as the input to make poly(A)-selected, unstranded libraries using a modified protocol of the TruSeq RNA library prep kit v2 (Illumina). After adapter ligation, AMPure XP beads were used to select libraries between 100 and 400 bp. Libraries were amplified using 15 cycles of PCR and DNA fragments of 300 bp were purified after separation on a 2% agarose gel (Qiagen MinElute gel extraction kit). For patient samples, total RNA was isolated using the TRIzol reagent. 500 ng of total RNA was used as the input to make poly(A)-selected, unstranded libraries using the protocol developed for the TruSeq RNA library prep kit v2 (Illumina). Libraries were purified using AMPure XP beads to select for DNA fragments of 300 bp. All purified libraries were sequenced on the Illumina Hi-Seq 2000 with 2x50 bp reads.

Genome annotation and read mapping. Annotations for splicing analysis of cassette exons, competing 5' and 3' splice sites, and retained introns were gathered from MISO v2.0[1]. Splice junctions that were not alternatively spliced in any isoforms from the UCSC knownGene track[2] were defined as constitutive junctions. Gene annotations were defined by merging the UCSC knownGene track with the Ensembl 71 gene annotation[3]. We additionally created an annotation file holding all possible splice junctions obtained by splicing of annotated splice sites as described previously[4]. Reads were mapped to the GRCh37/hg19 human genome assembly using Bowtie v1.0.0[5], RSEM v1.2.4[6], and TopHat v2.1.1[7] as previously described[4].

Isoform expression and differential splicing. Isoform expression levels were estimated as previously described[4]. Unless otherwise specified, a splicing event was classified as differentially spliced if it exhibited a change in isoform ratio of $\geq$10% and a Bayes Factor $\geq$5 with at least 20 informative reads for that event in each sample. Wagenmakers's framework[8] was used to compute Bayes factors associated with differences in isoform ratio between samples.

Cluster analysis. Unsupervised cluster analysis was performed using isoform ratios for cassette exons which had $\geq$100 informative (distinguishing between inclusion and exclusion isoforms)

reads per samples and exhibited an absolute change in isoform ratio ≥10% between any two samples. Ward's method was used for unsupervised clustering following a z-score normalization across samples for each cassette exon.

Sequence logos. For each mutant sample, we identified cassette exons with a minimum 10% change in isoform ratio relative to WT as well as a Bayes factor greater than or equal to 5 in relation to isoform differences between samples. Sequence logos were then created for these cassette exons with the seqLogo package in Bioconductor[9].

Sample similarity. Sample similarity in **Fig. 2C** was calculated using isoform ratios for cassette exons that exhibited differential splicing in association with *SRSF2* mutations in primary AML patient samples with or without *SRSF2* mutations[10]. Isoform ratios (exon inclusion) for these cassette exons were calculated for our transgenic K562 cell lines as well as for individual AML patients. For each AML patient, changes in exon inclusion were defined with respect to the median value for AML samples without *SRSF2* mutations. Pearson correlation coefficients were computed for each possible pairwise comparison of samples and then used to perform unsupervised clustering.

Minigene construction and transfections. An insert containing the RPL21 genomic locus (chr13: 27,828,357-27,829,491 in GRCh37) and a modified H2AFY genomic locus (chr5:134,681,658-134,696,297 in GRCh37) was inserted into the EcoRV site of the pUB6/V5-HisA vector (Invitrogen) by Gibson assembly cloning (NEB). For RPL21, the insert was created by PCR amplifying genomic DNA isolated from K562 cells with primers flanking the genomic range specified above (forward: TTACAGGGGTTTGGGGCAAA, reverse: TGGCAAAGTGAAAAGGGGGT), followed by a nested PCR with primers that exactly match the beginning and end of the specified genomic range (forward: TAATTCGCCAAAATGACGAACACAAAG, reverse: TTAACTTGTTTGTTTACAACAATGCCAAC). The product of this nested PCR was amplified further to create sequence complementary overhangs (forward: tcgagcggccgccactgtgctggatTTAACTTGTTTGTTTACAACAATGCCAAC, reverse:

tccagtgtggtggaattctgcagatTAATTCGCCAAAATGACGAACACAAAG) with a PCR-linearized pUB6 pUB6/V5-HisA vector (forward: ATCCAGCACAGTGGCGGC, reverse: ATCTGCAGAATTCCACCACACTGG). These two PCR products were then combined with a Gibson assembly reaction to create the RPL21 minigene plasmid. Site-directed mutagenesis was used to mutate the native ESE within the cassette exon.

The H2AFY locus was modified during cloning to reduce its size due to constraints of the plasmid. The modified locus consisted of the flanking (upstream and downstream) constitutive and mutually exclusive exons with no modifications as well as the intervening introns, each of which was reduced in size to include the first 100 nucleotides and last 250 nucleotides to preserve sequence elements at the 5' and 3' splice sites that may be important for spliceosome assembly. The sequence for this modified version of H2AFY was synthesized as a gBlock (IDT), with additional 5' and 3' overhangs that have sequence complementarity to the pUB6/V5-HisA vector. This gBlock was then combined with PCR-linearized pUB6 pUB6/V5-HisA vector (forward: ATCCAGCACAGTGGCGGC, reverse: ATCTGCAGAATTCCACCACACTGG) in a Gibson assembly reaction to create the H2AFY minigene plasmid. Site-directed mutagenesis was used to alter nucleotides at the -3 and +1 positions of the 3' splice sites of the mutually exclusive exons as described.

K562 cells were transfected in biological triplicate with the minigene plasmids described above using the Nucleofector II device (Lonza) with the Cell Line Nucleofector Kit V (program T16). RNA was extracted 48 hours post-transfection using the TRIzol reagent and purified using the Qiagen RNeasy Mini kit. RNA was converted to cDNA with Superscript III using a gene-specific primer (ACAACAGATGGCTGGCAACTAGAAG) for pUB6/V5-HisA in order to specifically amplify the RNA transcribed from the minigene.

qRT-PCR. qRT-PCR of the minigene-transfected cDNA was performed in technical triplicate in 5 µL reactions. Each reaction consisted of 1 ng template cDNA, 100 nM primers, and SYBR Green Master mix. The following primers were used: RPL21 inclusion (forward:AAGAGGAGAGGCACCCGATA, reverse:GTACCCATTCCCTTGATGTCTAC), RPL21 exclusion (forward: AGAAAACATGGGAATGGGTACTG, reverse: TGTTTACAACAATGCCAACAGCA), H2AFY upstream inclusion (forward _+1C in exon6_upstream: TGGCCAGAAGCTGAACCTTA, forward_+1T in exon6_upstream:

TGGCCAGAAGTTGAACCTTA, reverse: CAGCGTGTTTCCTAGGTCATC), H2AFY downstream inclusion (forward_+1T in exon6_downstream: CCAGAAGTTGCAAGTTGTACAGG, forward_+1C in exon6_downstream: TGGCCAGAAGCTGCAAGT, reverse: CTCCAGCGTGTTTCCTACTTC).

RT-PCR. Total RNA was isolated from K562 cells using the TRIzol reagent. cDNA was synthesized using SuperScript IV Reverse Transcriptase (ThermoFisher) following the manufacturer's protocol. PCR was performed on synthesized cDNA using primers specific for selected cassette exons (**Table S6**) using Phusion High Fidelity Polymerase (ThermoFisher). Amplicons were quantified with ImageJ (Fiji) following agarose gel electrophoresis.

REFERENCES

1. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*. 2010;7(12):1009–1015.
2. Meyer LR, Zweig AS, Hinrichs AS, et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res*. 2013;41(Database issue):D64–9.
3. Flicek P, Ahmed I, Amode MR, et al. Ensembl 2013. *Nucleic Acids Res*. 2013;41(Database issue):D48–55.
4. Ilagan JO, Ramakrishnan A, Hayes B, et al. U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res*. 2015;25(1):14–26.
5. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.
6. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
7. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105–1111.
8. Wagenmakers E-J, Lodewyckx T, Kuriyal H, Grasman R. Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method. *Cogn Psychol*. 2010;60(3):158–189.
9. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80.
10. Kim E, Ilagan JO, Liang Y, et al. SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. *Cancer Cell*. 2015;27(5):617–630.