Rare and private spliceosomal gene mutations drive partial, complete, and dual phenocopies of hotspot alterations

Joseph Pangallo¹,², Jean-Jacques Kiladjian³, Bruno Cassinat³, Aline Renneville⁴, Justin Taylor⁵, Jacob T. Polaski¹,², Khrystyna North¹,²,⁶, Omar Abdel-Wahab⁵, Robert K. Bradley¹,²,⁶,*

¹Computational Biology Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
²Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
³APHP and Inserm UMRS-1131, IUH, Hôpital Saint-Louis, Paris, France
⁴Hematology Laboratory, Biology and Pathology Center, CHRU of Lille, France; University of Lille Nord de France, Lille, France; UMR-S 1172, Team 3, Cancer Research Institute of Lille, Lille, France
⁵Human Oncology and Pathogenesis Program and Leukemia Service, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
⁶Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

*Correspondence:
Address: 1100 Fairview Ave N, Seattle, WA 98109, USA
Email: rbradley@fredhutch.org
Phone: +1 (206) 667-5662
KEY POINTS

- A subset of patients with hematologic malignancies carry rare spliceosomal gene mutations of unknown disease relevance
- Many rare and even private spliceosomal gene mutations create molecular phenocopies of hotspot mutations and are likely pathogenic
ABSTRACT
Genes encoding the RNA splicing factors *SF3B1*, *SRSF2*, and *U2AF1* are subject to frequent missense mutations in clonal hematopoiesis and diverse neoplastic diseases. Most “spliceosomal” mutations affect specific hotspot residues, resulting in mechanistic changes in exon and splice site recognition that promote disease pathophysiology. However, a subset of patients carry spliceosomal mutations that affect non-hotspot residues, whose potential functional contributions to disease are unstudied. Here, we undertook a systematic characterization of diverse rare and private spliceosomal mutations to infer their likely disease relevance. We utilized isogenic cell lines and primary patient materials to discover that 11 of 14 studied rare and private mutations in *SRSF2* and *U2AF1* induced mechanistic splicing alterations, including partially or completely phenocopying the mechanistic changes induced by hotspot mutations or driving “dual” phenocopies that mimicked two co-occurring hotspot mutations. Our data suggest that many rare and private spliceosomal mutations contribute to disease pathogenesis and illustrate the utility of mechanistic assays to inform precision medicine by inferring the potential disease relevance of newly discovered mutations.
INTRODUCTION

Somatic mutations in genes encoding RNA splicing factors are among the most common genetic changes observed in many hematologic malignancies. Also recurrently observed in solid tumors, albeit at lower frequencies, these “spliceosomal mutations” occur most commonly in SF3B1, SRSF2, and U2AF1 as missense changes at a highly specific set of hotspot residues. Hotspot mutations in SF3B1, SRSF2, and/or U2AF1 are observed in many patients with myelodysplastic syndromes (MDS) and related hematologic diseases and occur at high frequencies of 5-18% in chronic lymphocytic leukemia (CLL), 5-25% of acute myeloid leukemia (AML) in adults, and 14-29% in uveal melanoma.

Consistent with the frequent and recurrent nature of spliceosomal mutations, functional studies indicate that these lesions drive disease. Mutations in SRSF2 and U2AF1 specifically occur at high rates in elderly subjects with clonal hematopoiesis and confer a high risk of transformation to overt myeloid leukemia in this setting. In many cases, concrete links between altered RNA splicing mechanisms, specific target genes, and hallmark disease phenotypes have been identified. For example, SF3B1 mutations alter RNA branchpoint recognition to cause BRD9 mis-splicing and cell transformation; SRSF2 mutations alter exonic splicing enhancer recognition to cause EZH2 mis-splicing and impaired hematopoiesis; U2AF1 mutations alter 3' splice site recognition to cause IRAK4 mis-splicing and aberrant innate immune signaling.

Although the bulk of SF3B1, SRSF2, and U2AF1 mutations affect a small set of hotspot residues, a minority of patients carry non-hotspot mutations, some of which are recurrent despite their relative rarity. The relevance of rare and private (observed in only one patient) spliceosomal lesions to disease is unclear, but they are enriched in hematologic malignancies, preferentially occur as missense changes, and appear in a heterozygous genetic context, like their hotspot counterparts. This situation—where a cancer-relevant gene is subject to hotspot mutations of known significance as well as rare or private mutations of unknown functional consequence—is not unique to splicing factors. Rare and private mutations have frequently been ignored in favor of their more common hotspot counterparts due to the inherent challenges of studying a diverse mutational spectrum. However, advances in mechanistic and functional assays have enabled recent studies to identify pro-tumorigenic roles of rare and even private inherited genetic variants and somatically acquired mutations of previously unknown significance in
BRCA1, EGFR, KRAS, and other cancer-relevant genes\textsuperscript{26-30}. Each of those studies relied on a different approach to classification—e.g., measuring how each rare variant or mutation affected biochemical activity (BRCA1), gene expression profiles (EGFR and others), or tumor outgrowth (KRAS and others)—selected based on known molecular or biological consequences of hotspot mutations.

Here, we conducted a systematic study to infer the likely disease relevance of rare and private mutations in SRSF2 and U2AF1. Our study was motivated in part by a recent report of three CLL patients with novel SF3B1 in-frame deletions whose splicing profiles mimicked those of patients with hotspot SF3B1 mutations\textsuperscript{31}, as well as our recent finding that both rare and common SF3B1 mutations converge on BRD9 mis-splicing across cancer types\textsuperscript{19}. We wondered whether rare and private SRSF2 and U2AF1 mutations might similarly mimic the splicing phenotypes of hotspot mutations, which induce highly specific alterations in exon or 3’ splice site recognition that drive key disease phenotypes\textsuperscript{20-24}. We hypothesized that rare or private SRSF2 and U2AF1 mutations that phenocopied hotspot mechanistic changes were candidate drivers, while mutations that induced no splicing changes were likely passengers. We used this approach in both isogenic cell lines and primary patient materials to infer the likely pathogenicity of non-hotspot SRSF2 and U2AF1 mutations (Fig. 1C).
METHODS

Vector construction and cell line production. An insert containing SRSF2 (or U2AF1) cDNA-FLAG-P2A-mCherry was cloned into the lentiviral vector pRRLSIN.cPPT.PGK-GFP.WPRE (Addgene plasmid # 12252). Mutations in SRSF2 or U2AF1 were then created by site-directed mutagenesis. These plasmids were co-transfected with psPAX2 (Addgene plasmid 12260) and envelope vector pMD2.G (Addgene plasmid 12259) into 293T cells. Lentivirus was collected from the supernatant 48 hour post-transfection. Stable cell lines were made by transducing K562 cells with lentivirus at a MOI of 2.5 (U2AF1) or 5 (SRSF2). Cells were expanded, and mCherry+ cells were collected by FACS. K562 cells were cultured in IMDM supplemented with 10% FBS.

Western blotting. Protein lysates were extracted from K562 cells by resuspension in RIPA buffer. Thirty micrograms of protein were then loaded for SDS-PAGE and transferred to a nitrocellulose membrane. Proteins were probed with the following antibodies: anti-U2AF1 (Bethyl Laboratories, catalog no. A302-080A), anti-SRSF2 (Millipore-Sigma, catalog no. 04-1550), anti-FLAG (Thermo, catalog no. MA1-91878), anti-HistoneH3 (Abcam, catalog no. ab1791).

RNA-seq library preparation and analysis. Total RNA was isolated from K562 cells or patient materials using the TRIzol reagent. 4 ug (K562) or 500 ng (patient materials) of total RNA was used as to make poly(A)-selected, unstranded libraries with the TruSeq RNA library prep kit v2 (Illumina). Purified libraries were sequenced on the Illumina Hi-Seq 2000 with 2x50 bp reads.

Follow RNA-seq read mapping, isoform expression levels were estimated as previously described\textsuperscript{23}. 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 $\geq 5$. Wagenmakers’s framework\textsuperscript{32} was used to compute Bayes factors associated with differences in isoform ratio between samples. A full description of the analysis can be found in Supplementary Methods.

Primary human samples. Studies were approved by the Institutional Review Boards of Memorial Sloan Kettering Cancer Center (MSK; under MSK IRB protocol 06-107) and the Hôpital Saint-Louis and conducted in accordance with the Declaration of Helsinki protocol.
Written informed consent was obtained from all participants. Patient samples were anonymized by the Hematologic Oncology Tissue Bank of MSK and the Hôpital Saint-Louis. Mutational analysis of SRSF2 and U2AF1 was performed on genomic DNA from bone marrow mononuclear cells by targeted sequencing using MSK Heme-PACT assay (for samples from MSK).

**Data availability.** RNA-seq data generated as part of this study were deposited in the Gene Expression Omnibus (accession number GSE135732). Previously published data were downloaded from the Gene Expression Omnibus (GEO) under accession numbers GSE65349, GSE114922, and GSE66917 and GSE67039. TCGA data was downloaded from CGHub. Other data that support this study’s findings are available from the authors upon reasonable request.
RESULTS

Diverse SRSF2 and U2AF1 mutations alter RNA splicing programs

We queried the Catalogue of Somatic Mutations in Cancer (COSMIC) database\textsuperscript{25} to identify all SRSF2 and U2AF1 mutations with confirmed somatic status as of September 17, 2018. We selected eight SRSF2 and six U2AF1 representative non-hotspot mutations for detailed study (Fig. 1B). These mutations exhibited highly variable frequencies (ranging from private to common), represented both missense changes and indels (insertions and deletions), and were present as either single polymorphisms or indels as well as more complex events (involving multiple mutations, such as U2AF1_S34F_Q157R, for which two hotspot mutations co-occurred on the same allele). We systematically determined how each mutation affected RNA splicing mechanisms in both engineered cell lines and primary patient materials, when available, as follows.

We first established cell culture models of each selected SRSF2 and U2AF1 mutation. We modeled each mutation via transgenic expression in K562 cells for two reasons. First, spliceosomal mutations are always co-expressed with a WT allele, which is required for cell survival\textsuperscript{38}. Our lentiviral construct contained a fluorescent marker that permitted titration of transgene expression by flow sorting, which was critical given previous reports that the ratio of mutant to WT protein controls global mis-splicing profiles\textsuperscript{39}. Second, we and others have previously demonstrated that simultaneous expression of a transgenic mutant protein and endogenous WT protein in K562 cells faithfully recapitulates mis-splicing profiles observed in primary patient materials with SRSF2 or U2AF1 mutations\textsuperscript{20,23,40}.

We transduced K562 cells with a lentiviral construct expressing each mutant cDNA (individually) and established stable transgenic cell lines for each selected mutation (Fig. 1C, S1). We additionally established cell lines expressing transgenic wild-type (WT) SRSF2 or U2AF1 as a control for transgene expression, as well as cell lines expressing the hotspot mutations SRSF2_P95H, U2AF1_S34F, and U2AF1_Q157R. We modeled two different U2AF1 hotspot mutations because we previously found that mutations affecting U2AF1’s first versus second zinc finger result in distinct alterations in 3’ splice site recognition\textsuperscript{23}. We confirmed that transgene introduction resulted in relative levels of mutant versus WT SRSF2 and U2AF1 mRNA within physiological ranges observed in patients and that each cell line expressed mutant
protein in the absence of significant perturbations to total (mutant + WT) levels of SRSF2 or U2AF1 relative to untransduced cells (Fig. 1D-E, S2).

We first tested whether expressing rare SRSF2 and U2AF1 mutations altered global splicing programs. We performed high-coverage RNA-seq on each of the 19 distinct cell lines and quantified global isoform expression for ~125,000 alternative splicing events and aberrant retention or splicing of ~160,000 constitutive introns as previously described\(^41\). An unsupervised cluster analysis based on cassette exon inclusion—where we focused on cassette exons because SRSF2 and U2AF1 hotspot mutations primarily affect this category of splicing event\(^20,23\)—revealed allele-specific clustering that was distinct from WT splicing programs in many cases (Fig. 1F-G). This simple analysis suggested that at least some rare mutations influenced splicing programs through as-yet-unknown mechanisms.

**Rare and hotspot SRSF2 mutations converge on altered exonic splicing enhancer preference**

We sought to determine the mechanisms by which rare spliceosomal mutations influenced global splicing programs (Fig. 1F). We first focused on SRSF2 mutations, because all hotspot SRSF2 mutations affect a single residue (P95) and cause identical mechanistic alterations in the RNA splicing process\(^20,21\). Like their hotspot counterparts, rare SRSF2 mutations were associated with a diversity of splicing changes affecting competing splice sites, cassette exons, retained introns, and aberrant splicing or retention of normally constitutive introns, with cassette exons representing the most commonly differentially spliced event. The numbers of significantly differentially spliced events, defined as events with a change in isoform ratio of \(\geq 10\%\) and Bayes factor \(\geq 5\) relative to WT-expressing control cells, varied by an order of magnitude across the different mutations, suggestive of dramatically different functional consequences (Fig. 2A, Table S1).

Hotspot SRSF2 mutations alter SRSF2’s RNA-binding affinity and avidity to induce sequence-specific changes in exonic splicing enhancer (ESE) preference. While WT SRSF2 recognizes a consensus motif SSNG (S = G or C) in pre-mRNA, SRSF2P95H/L/R mutations promote recognition of C-rich variants and repress recognition of G-rich variants\(^20,21,42\). We therefore determined how each rare mutation affected recognition of G- versus C-rich variants of the core SSNG motif. We identified all differentially spliced cassette exons in each cell line (Fig.
S3), identified all occurrences of SSNG motifs in each cassette exon, and computed the enrichment for each SSNG motif variant in cassette exons that were promoted versus repressed in mutant versus WT cells. Six of the eight tested non-hotspot SRSF2 mutations caused significant alterations in C- versus G-rich ESE preference that were restricted to differentially spliced cassette exons, an identical pattern to that observed for the SRSF2P95H hotspot mutation (Fig. 2B, S4). Our approach allowed us to deconvolve complex co-mutation events such as SRSF2P95_R102del+P107H. SRSF2P95_R102del alone phenocopied SRSF2P95 mutations, while SRSF2P107H alone had no effect, suggesting that the first lesion might be pathogenic while the second is functionally silent (Table 1).

We next confirmed our results in the physiological setting of primary patient materials. We searched for non-hotspot SRSF2 mutations in institutional biorepositories as well as published cohorts of patients with AML20,35, CMML20, and MDS34. We identified samples carrying SRSF2S54A/F, SRSF2R94_P95insR, and SRSF2P95_R102del, performed RNA-seq or re-analyzed published data when available, and tested for sequence-specific alterations in ESE preference. In each case, we observed enhanced and spatially restricted recognition of C- versus G-rich SSNG motifs that was consistent with our results from cell culture (Fig. 2B, S4).

Interestingly, although many non-hotspot mutations induced seemingly complete phenocopies of enhanced recognition of C- versus G-rich ESEs, SRSF2S54A/F induced partial phenocopies apparent as decreased recognition of GGNG in the absence of enhanced recognition of CCNG (Table 1, Fig. S4). Unsupervised clustering of K562 cell lines with primary patient samples revealed that global mis-splicing profiles segregated by mechanistic classification, consistent with a central role for altered ESE recognition in driving global mis-splicing programs in cells with rare as well as hotspot SRSF2 mutations (Fig. 2C). We experimentally validated results from RNA-seq by performing RT-PCR on eight distinct mis-splicing events. In each case, the private mutation SRSF2R86_G93dup and the common mutation SRSF2P95H induced concordant mis-splicing in K562 cells (Fig. 2D-E, S5).

We next experimentally confirmed that rare SRSF2 mutations caused aberrant exon recognition in a manner that depended upon altered ESE recognition. As we previously demonstrated that enhanced cassette exon recognition in hotspot mutant cells was due to presence of CCNG motifs20, we here instead tested whether repressed cassette exon recognition was due to presence of GGNG motifs. A cassette exon within RPL21 exhibited significant and
consistent repression in mutant cells and also contained a single GGNG motif, making it an ideal system to test this hypothesis (Fig. 2F). We cloned this cassette exon and flanking introns into a plasmid, introduced a GGTG>CCTG mutation, and expressed both GGTG (native) and CCTG versions of this minigene in K562 cells. We focused on SRSF2R86_G93dup, a private mutation for which we were unable to identify corresponding patient materials but which phenocopied hotspot mutations in cell culture, as well as the rare mutation SRSF2R94_P95insR, for these assays. Cells expressing SRSF2R86_G93dup and SRSF2R94_P95insR both exhibited reduced cassette exon recognition relative to WT cells for the native minigene, as expected, which was abolished by the GGTG>CCTG mutation (Fig. 2G). These results confirmed our genomic inference that rare SRSF2 mutations alter ESE preference and experimentally demonstrate that reduced recognition of G-rich ESEs drives mis-splicing in SRSF2-mutant cells.

**Rare U2AF1 mutations induce both complete and dual phenocopy of altered 3' splice site recognition**

Rare U2AF1 mutations affected a diversity of alternative splicing events as well as a smaller set of normally constitutively spliced introns, with cassette exons exhibiting the most frequent differential splicing (Fig. 3A, Table S3). Unlike SRSF2 hotspot mutations, which induce identical changes in ESE recognition, U2AF1 hotspot mutations give rise to two distinct changes in RNA-binding specificity and 3’ splice site recognition. U2AF1S34F/Y and Q157P/R mutations alter sequence-dependent recognition of the nucleotides preceding and following the AG dinucleotide of the 3' splice site, respectively.23,39,43

We therefore tested how expression of each rare U2AF1 mutant allele altered 3' splice site recognition. We identified cassette exons that were differentially spliced in K562 cells expressing each mutant allele relative to WT cells and computed consensus 3’ splice site sequences that were associated with promoted versus repressed cassette exons (Fig. 3B, S3). Expression of the hotspot mutations U2AF1S34F and Q157R altered recognition of the -3 and +1 sites, as expected. U2AF1R156H phenocopied U2AF1Q157P/R, as did the rare insertion U2AF1E159_M160insYE. The complex co-mutation U2AF1S34F_Q157R drove a “dual” phenocopy, characterized by S34 and Q157 hotspot-like alterations at both the -3 and +1 positions. The rare mutation U2AF1I24T, which affects U2AF1’s first zinc finger like S34F/Y, was also associated with a dual phenocopy that was highly similar to that induced by
U2AF1S34F_Q157R, while U2AF1I24V was similar to U2AF1Q157R (Table 1). To confirm that these 3’ splice site preference alterations were potentially relevant to disease, we extended the above analysis to mutation-matched patient materials. We identified primary patient materials bearing most of the studied rare mutations and compared their transcriptomes to those of WT samples to find similar alterations in consensus 3’ splice sites (Fig. 3B). For U2AF1I24T we only observed alterations at the +1, and not -3, position, rather than the dual phenocopy that was evident in cell culture, potentially due to the relatively low allelic expression of this mutation in the analyzed patient sample (23% vs. 32% allelic expression in the patient samples vs. K562 cells expressing U2AF1I24V). We used RT-PCR to experimentally validate results from RNA-seq, confirming that U2AF1I24T induced similar patterns of mis-splicing as did U2AF1S34F in K562 cells for four distinct splicing events (Fig. 3C-D, S5).

We experimentally confirmed that mis-splicing of exons in cells expressing rare U2AF1 mutations was a direct consequence of altered 3’ splice site recognition. We selected a mutually exclusive exon event within H2AFY for further study, as H2AFY is a robust target of U2AF1S34F/Y in both human patients and murine models whose mis-splicing contributes to impaired hematopoiesis. Like U2AF1S34F, the rare mutations U2AF1I24T/V promoted upstream exon inclusion while repressing downstream exon inclusion (Fig. 3E). We cloned H2AFY’s mutually exclusive exons and flanking introns and exons into a minigene cassette and created mutant versions of the minigene where we mutated the 3’ splice sites of both mutually exclusive exons as follows: (1) swap the nucleotides at the -3 positions, (2) swap the nucleotides at the +1 positions, and (3) swap the nucleotides at both the -3 and +1 positions. We transfected these minigenes into WT and U2AF1I24V cells, where we focused on U2AF1I24V since we were unable to obtain patient samples bearing this lesion for transcriptome analysis, and measured relative levels of upstream versus downstream exon inclusion. These experiments revealed that native C and T at the +1 positions of the upstream and downstream exons were both essential for mutation-dependent splicing, while the nucleotides at the -3 positions could be swapped without consequence (Fig. 3F). These minigene experiments confirm our genomic inference that U2AF1I24V induces H2AFY mis-splicing by altering recognition of the +1 position of the 3’ splice sites of both of H2AFY’s mutually exclusive exons.

Mechanistic classification of mutations explains extent of transcriptome dysregulation
Our analyses of ESE and 3' splice site recognition in SRSF2- and U2AF1-mutant cells and patient materials clearly distinguished between mutations that did or did not alter the normal functions of SRSF2 and U2AF1 (Table 1). Although hotspot SRSF2 and U2AF1 mutations induce distinctive mis-splicing programs that contribute to disease phenotypes, they have also been shown to affect other cellular processes of potential disease relevance including mRNA translation$^{46}$ and R loop formation$^{47,48}$. We reasoned that if a given rare mutation altered a critical cellular process, then that alteration might be reflected in dysregulated gene expression relative to WT cells. This hypothesis is consistent with previous observations that many cancer-causing mutation that act through diverse molecular pathways induce stereotyped and readily detectable alterations in gene expression profiles$^{29}$. We therefore compared the extent of gene expression versus splicing dysregulation to find that hotspot and rare mutations that phenocopied hotspot mutations induced dramatic changes in gene expression, while putative passenger mutations with no apparent effects on ESE or 3' splice site recognition similarly had few effects on global gene expression (Fig. 4A-B, Table S4-5). This analysis supports, although does not prove, our hypothesis that rare SRSF2 or U2AF1 mutations which do not alter ESE or 3' splice site recognition are likely functionally silent passengers.

**Rare SRSF2 and U2AF1 mutations converge on a small set of disease-relevant events**

Although SRSF2 and U2AF1 mutations induce distinct alterations in RNA splicing mechanisms, we wondered whether they might converge on shared downstream targets that contribute to their enrichment in hematologic disease. We speculated that such targets might exhibit concordant differential splicing in association with both hotspot and rare mutations. We therefore identified cassette and mutually exclusive exons within coding genes that were differentially spliced in association with at least three of the five SRSF2P95-like mutations and compared that set to differentially spliced exons found in association with three of the U2AF1S34-like mutations. As expected given SRSF2 and U2AF1 mutations’ distinct mechanistic consequences, as well as these lesions’ preferential enrichment in different disease subtypes$^{1,49}$, the vast majority of differentially spliced exons were SRSF2- or U2AF1-specific. However, three genes were differentially spliced in association with both SRSF2 and U2AF1 mutations (Fig. 4C), of which H2AFY and IRAK4 were particularly notable given their known involvement in hematologic disease. Previous studies demonstrated that U2AF1S34F/Y promote inclusion of the upstream
exon of two mutually exclusive exons within *H2AFY*, which encodes macro-H2A1, thereby perturbing erythroid and granulomonocytic differentiation\textsuperscript{23,44,45}. *U2AF1S34F* similarly promotes inclusion of an *IRAK4* cassette exon to drive the *IRAK4*-long isoform that activates innate immune signaling and is important for leukemic cell function\textsuperscript{24} (Fig. 4D). Our analysis revealed that the rare mutations *U2AF1I24T/V* phenocopied the *H2AFY* and *IRAK4* mis-splicing characteristic of *U2AF1S34F/Y*-mutant cells, and furthermore that both *SRSF2P95* and *SRSF2P95*-like mutations drove *H2AFY* as well as *IRAK4* differential splicing (Fig. 4C, Table S1). Intriguingly, however, *SRSF2* mutations drove *H2AFY* and *IRAK4* mis-splicing that was in direct opposition to that caused by *U2AF1* mutations (Fig. 4E, Table S3). As two of the three coding genes that are shared targets of both hotspot and rare *SRSF2* and *U2AF1* mutations have been previously implicated in the pathology of *U2AF1*-mutant cells, we speculate that differential splicing of *H2AFY* and *IRAK4* may be similarly important for the functional consequences of *SRSF2* mutations.
DISCUSSION

In addition to characterizing the function of rare mutations in SRSF2 and U2AF1, our study illustrates a method for inferring mutational pathogenicity when a biological assay such as tumorigenesis is inaccessible. Although SRSF2 and U2AF1 mutations exhibit the genetic enrichment expected of driver lesions in many dysplastic and neoplastic disorders, they do not confer a growth advantage to cultured transformed cells and are dispensable for the maintenance of at least some xenografts. We therefore took advantage of the stereotyped changes in RNA splicing mechanisms caused by SRSF2 and U2AF1 hotspot mutations—which have been directly linked to disease phenotypes—to classify rare mutations as candidate drivers or passengers. Although unbiased cluster analyses (Fig. 1F,G) separated mutations similarly to subsequent mechanism-based analyses, only mechanism-based approaches can classify pathogenicity with reasonable confidence, given the known role of dysfunctional splicing mechanisms in SRSF2- and U2AF1-mutant hematologic malignancies.

Our mechanistic approach can confidently identify functionally active SRSF2 and U2AF1 mutations that alter ESE or 3’ splice site recognition, but cannot prove that any given mutation is functionally silent. Many cancer driver mutations directly or indirectly dysregulate gene expression, irrespective of the specific mechanisms by which they promote cancer, in a specific manner. Therefore, the concordance between our mechanistic classification of mutations and the extent of transcriptome dysregulation that each induces (Fig. 4A-B) suggests that SRSF2 and U2AF1 mutations that do not detectably alter splicing mechanisms are likely passengers. However, we cannot rule out the possibility that some rare mutations promote disease through mechanisms that are undetectable via transcriptomic analyses. For example, recent studies have reported increased R loop formation in cells expressing SRSF2 and U2AF1 hotspot mutations (although a causative role for R loop formation in dysplastic hematopoiesis or tumorigenesis has not yet been demonstrated). Conversely, although our mechanistic approach accurately tests whether individual rare mutations induce molecular phenocopies of pathogenic hotspot mutations, it only provides a likely estimate (not proof) of pathogenicity. Even variants that we classify as likely pathogenic should be interpreted with care and caution in a clinical setting.

A published structure of SRSF2 offers insight into the potential means by which rare and hotspot mutations cause convergent splicing alterations (Fig. S6). Rare mutations affecting the P95 hotspot presumably induce a similar set of domain movements as those induced by
SRSF2P95H/L/R\textsuperscript{20}, while S54 lies distal to the binding core and so likely affects RNA binding indirectly. H99 interacts with the variable nucleotide in the CCNG motif, potentially explaining why SRSF2H99L did not induce detectable changes in ESE preference.

Our study has several implications for basic and translational studies of spliceosomal mutations. First, since many rare SRSF2 and U2AF1 mutations mechanistically phenocopy the SRSF2P95, U2AF1S34, and U2AF1Q157 hotspot mutations, studying those hotspot mutations will also give insight into the pathology of diverse rarer mutations. Second, because rare and even private SRSF2 and U2AF1 mutations may be pathogenic, non-hotspot mutations should be considered in early detection and monitoring studies\textsuperscript{15} when feasible. Finally, when therapies designed to specifically target cells with spliceosomal mutations enter clinical practice\textsuperscript{38,40,52,53}, patients bearing non-hotspot spliceosomal mutations should be considered as candidates for these therapies. Although performing a mechanistic, whole-transcriptome analysis is not feasible in a clinical setting, continued study of both hotspot and hotspot-phenocopy mutations may reveal specific biomarkers of mutant SRSF2 and U2AF1 activity that can be utilized to rapidly classify novel spliceosomal mutations as drivers or passengers for precision medicine.
ACKNOWLEDGEMENTS
This research was supported in part by the Dept. of Defense Bone Marrow Failure Research Program W81XWH-16-1-0059 (RKB, OA-W), NIH/NIDDK R01 DK103854 (RKB), NIH/NHLBI (R01 HL128239), the Evans MDS Foundation (RKB, OA-W), and the Henry & Marilyn Taub Foundation (OA-W). J.T. is supported by the Conquer Cancer Foundation of the American Society of Clinical Oncology, the American Association for Cancer Research, the American Society of Hematology (ASH), the Robert Wood Johnson Foundation, and the NIH/NCI (K08 CA230319). OA-W is supported by the Pershing Square Sohn Cancer Research Alliance. RKB is a Scholar of The Leukemia & Lymphoma Society (1344-18). The results shown here are in part based upon data generated by the TCGA Research Network: https://cancergenome.nih.gov/.

AUTHOR CONTRIBUTIONS
J-JK, BC, AR, JT, and OA-W provided patient material. JP performed experiments and computational analyses. JTP and KN contributed to data interpretation. JP and RKB wrote the paper.

COMPETING INTERESTS
OA-W has served as a consultant for H3 Biomedicine, Foundation Medicine Inc., Merck, and Janssen, and serves on the Scientific Advisory Board of Envisagenics Inc.; OA-W has received prior research funding from H3 Biomedicine unrelated to the current manuscript.
REFERENCES


<table>
<thead>
<tr>
<th>Mutation</th>
<th>n</th>
<th>Mechanistic classification</th>
<th>Evidence</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>S54A</td>
<td>1</td>
<td>partial phenocopy of P95</td>
<td>cell line + patient</td>
<td>this study</td>
</tr>
<tr>
<td>S54F</td>
<td>1</td>
<td>partial phenocopy of P95</td>
<td>cell line + patient</td>
<td>this study</td>
</tr>
<tr>
<td>R86_G93dup</td>
<td>1</td>
<td>phenocopy of P95</td>
<td>cell line</td>
<td>this study</td>
</tr>
<tr>
<td>R94_P95insR</td>
<td>11</td>
<td>phenocopy of P95</td>
<td>cell line + patient</td>
<td>this study</td>
</tr>
<tr>
<td><strong>P95H</strong></td>
<td>448</td>
<td>hotspot</td>
<td>cell line + patient</td>
<td>(previously studied)</td>
</tr>
<tr>
<td><strong>P95L</strong></td>
<td>280</td>
<td>hotspot</td>
<td>cell line + patient</td>
<td>(previously studied)</td>
</tr>
<tr>
<td><strong>P95R</strong></td>
<td>168</td>
<td>hotspot</td>
<td>cell line + patient</td>
<td>(previously studied)</td>
</tr>
<tr>
<td>P95_R102del</td>
<td>79</td>
<td>phenocopy of P95</td>
<td>cell line + patient</td>
<td>this study</td>
</tr>
<tr>
<td>P05_R102del + P107H</td>
<td>7</td>
<td>phenocopy of P95</td>
<td>cell line</td>
<td>this study</td>
</tr>
<tr>
<td>P107H</td>
<td>7</td>
<td>silent</td>
<td>cell line</td>
<td>this study</td>
</tr>
<tr>
<td>H99L</td>
<td>2</td>
<td>silent</td>
<td>cell line</td>
<td>this study</td>
</tr>
<tr>
<td>I24T</td>
<td>5</td>
<td>dual phenocopy of S34 and Q157 (likely)</td>
<td>cell line + patient</td>
<td>this study</td>
</tr>
<tr>
<td>I24V</td>
<td>1</td>
<td>phenocopy of S34 (likely)</td>
<td>cell line</td>
<td>this study</td>
</tr>
<tr>
<td>S34F</td>
<td>308</td>
<td>hotspot</td>
<td>cell line + patient</td>
<td>(previously studied)</td>
</tr>
<tr>
<td>S34Y</td>
<td>92</td>
<td>hotspot</td>
<td>cell line + patient</td>
<td>(previously studied)</td>
</tr>
<tr>
<td>R156H</td>
<td>30</td>
<td>phenocopy of Q157</td>
<td>cell line + patient</td>
<td>this study</td>
</tr>
<tr>
<td>R156Q</td>
<td>2</td>
<td>silent</td>
<td>cell line</td>
<td>this study</td>
</tr>
<tr>
<td>Q157R</td>
<td>66</td>
<td>hotspot</td>
<td>cell line + patient</td>
<td>(previously studied)</td>
</tr>
<tr>
<td>Q157P</td>
<td>121</td>
<td>hotspot</td>
<td>cell line + patient</td>
<td>(previously studied)</td>
</tr>
<tr>
<td>E159_M160insYE</td>
<td>8</td>
<td>phenocopy of Q157</td>
<td>patient</td>
<td>this study</td>
</tr>
<tr>
<td>S34F + Q157R</td>
<td>1</td>
<td>dual phenocopy of S34 and Q157</td>
<td>cell line</td>
<td>this study</td>
</tr>
</tbody>
</table>

**Table 1. Mechanistic classification of studied mutations.** n, number of times that each mutation has been reported in COSMIC. Mechanistic classification inferred from exonic splicing enhancer preferences and 3' splice site preferences associated with each mutation. The mechanistic consequences of hotspot SRSF2 and U2AF1 mutations were previously studied by several groups²⁰-²³,⁵⁴.
Figure 1. Strategy for mechanistic classification of rare, non-hotspot SRSF2 and U2AF1 mutations.

(A) Hotspot (bold) and select rare and private mutations affecting SRSF2 and U2AF1. RRM, RNA recognition motif; RS, arginine/serine-rich domain; UHM, U2AF homology motif; Zn, zinc finger domain.

(B) Numbers of reported mutations in SRSF2 and U2AF1 in the Catalogue of Somatic Mutations in Cancer (COSMIC) database as of September 17, 2018. SRSF2S54A was identified in a patient sample but is not present in COSMIC.
(C) Schematic of our strategy for transgenically expressing individual mutations in cell culture and performing subsequent transcriptome analyses.

(D) Western blot for FLAG, SRSF2, and Histone H3 (H3) using lysate from untransduced K562 cells or K562 cells that stably expressed FLAG-tagged WT or mutant SRSF2 (mutation indicated above). H3 is a loading control. FLAG and SRSF2 band intensities were quantified using ImageJ and normalized to the respective band intensity for H3.

(E) As (D), but for U2AF1.

(F) Heat map and associated dendrogram representing an unsupervised cluster analysis based on cassette exon inclusion levels computed from the transcriptomes of K562 cells stably expressing the indicated alleles of SRSF2 (left) or U2AF1 (right). Exon inclusion values were z-score normalized.
Figure 2. Rare mutations in SRSF2 alter exonic splicing enhancer (ESE) preference.

(A) Differentially spliced events identified in K562 cells expressing each indicated SRSF2 mutant allele relative to WT-expressing control cells. Percentages represent the distribution of differentially spliced events among the indicated event types for each mutation.
(B) Enrichment for each indicated variant of the SSNG motif within cassette exons that were promoted versus repressed in cells expressing mutant versus WT SRSF2. The enrichment for a given motif was defined as the number of instances in all promoted exons divided by the number of instances in all repressed exons. Error bars represent 95% confidence intervals estimated by bootstrapping. The transcriptomes of patient samples bearing SRSF2S54A (polycythemia + hyperleukocytosis + myelofibrosis) and SRSF2S54F (chronic myelomonocytic leukemia) were sequenced for this study; RNA-seq data from patient samples bearing SRSF2R94_P95insR (AML), SRSF2P95H (CMML), and SRSF2P95_R102del (AML) were previously published²⁰.

(C) Heat map and dendrogram illustrating the global similarity of splicing programs in K562 cells and AML samples expressing the indicated alleles of SRSF2. Dendrogram illustrates the results of an unsupervised clustering based on differential splicing in each indicated sample relative to WT-expressing control cells (K562) or a median computed over all WT samples (AML). AML patient data was previously published²⁰.

(D) Top, RNA-seq read coverage illustrating increased cassette exon inclusion in PRMT2 in K562 cells expressing either a hotspot (P95H) or private (R86_G93dup) SRSF2 mutation. Log₂ (fold-change) illustrates log₂ (exon inclusion in mutant- versus WT-expressing cells). Bottom, RT-PCR validation of RNA-seq results in technical triplicate. Log fold-changes for RT-PCR computed with respect to the mean signal for WT.

(E) As (D), but for a cassette exon in C5orf4 that is repressed by mutant SRSF2.

(F) Relative inclusion of a cassette exon within RPL21 expressed from its endogenous locus in K562 cells expressing mutant versus WT SRSF2. Error bars represent 95% confidence intervals for the relative inclusion ratio, computed by propagating the 95% confidence intervals for the two isoforms to the ratio for mutant versus WT SRSF2 by standard rules for error propagation during division of quantities with individual errors.

(G) As (F), but where the RPL21 cassette exon is expressed from a minigene transfected into K562 cells and contains the indicated ESEs. GGTG is the native sequence; CCTG is a mutated ESE that is predicted to be well-recognized in the presence of mutant SRSF2. Bars represent the mean ± standard deviation, measured by qRT-PCR and computed over three biological replicates.
Figure 3. Rare mutations in U2AF1 alter 3' splice site recognition.

(A) Differentially spliced events identified in K562 cells expressing each indicated U2AF1 mutant allele relative to WT-expressing control cells. Percentages represent the distribution of differentially spliced events among the indicated event types for each mutation.
(B) Sequence logos representing consensus 3’ splice sites of cassette exons that are differentially spliced in K562 cells expressing mutant versus WT U2AF1. Gray boxes highlight sequence preferences at the -3 and +1 positions which are similar to those observed in cells expressing the U2AF1S34F/Y or U2AF1Q157P/R hotspot mutations. RNA-seq data from patient samples bearing U2AF1I24T (adrenocortical carcinoma), U2AF1S34F (AML), U2AF1R156H (MDS), U2AF1Q157R (AML), and U2AF1E159_M160insYE (AML) were previously published34-37.

(C) Top, RNA-seq read coverage illustrating increased cassette exon inclusion in RHBDD2 in K562 cells expressing either a hotspot (S34F) or rare (I24R) U2AF1 mutation. Log2 (fold-change) illustrates log2 (exon inclusion in mutant- versus WT-expressing cells). Bottom, RT-PCR validation of RNA-seq results in technical triplicate. Log fold-changes for RT-PCR computed with respect to the mean signal for WT.

(D) As (C), but for mutually exclusive exons in H2AFY. The upstream (orange) exon is the exon for which inclusion is calculated.

(E) Relative inclusion of the upstream versus downstream exon for two mutually exclusive exons within H2AFY expressed from its endogenous locus in K562 cells expressing mutant versus WT U2AF1 as estimated by RNA-seq. Error bars represent 95% confidence intervals for the relative inclusion ratio, computed by propagating the 95% confidence intervals for the two isoforms to the ratio for mutant versus WT SRSF2 by standard rules for error propagation during division of quantities with individual errors.

(F) As (E), but where the H2AFY mutually exclusive exons are expressed from a minigene transfected into K562 cells and contain 3’ splice sites with the indicated sequences. AG is the AG dinucleotide of the 3’ splice site. Bars represent the mean ratio of inclusion of the upstream:downstream exons ± standard deviation, estimated by qRT-PCR and computed over three biological replicates.
**Figure 4.** Hotspot and rare **SRSF2** and **U2AF1** induce transcriptome dysregulation and converge on **H2AFY** and **IRAK4** mis-splicing.

(A) Scatter plot comparing the numbers of differentially expressed genes (x axis) and differentially spliced cassette exons (y axis) in K562 cells expressing each indicated SRSF2 mutation versus WT-expressing control cells. Differentially expressed genes were defined as those genes with expression \( \geq 1 \) TPM in both samples, \(|\log_2 \text{ (fold-change)}| \geq \log_2 (1.5)\), and Bayes factor \( \geq 10\). See Table 1 for additional information on mechanistic classification of each mutation.

(B) As (A), but for the indicated **U2AF1** mutations.

(C) Venn diagram illustrating the sets of coding genes containing cassette exons and mutually exclusive exons that were differentially spliced in association with both hotspot and rare SRSF2 and/or U2AF1 mutations relative to control WT-expressing cells. Differentially spliced exons were defined as those exhibiting a change in isoform ratio \( \geq 10\% \) and a Bayes factor \( \geq 1\). Diagram restricted to genes containing cassette exons or mutually exclusive exons that were differentially spliced in association with at least three SRSF2P95-like mutations (**SRSF2R86_G93dup**, **SRSF2R94_P95insR**, **SRSF2P95H**, **SRSF2P95_R102del**, and **SRSF2P95_R102del**).
SRSF2P95_R102del + P107H considered) and three U2AF1S34-like mutations (U2AF1I24T, U2AF1I24V, and U2AF1S34F considered).

(D) Inclusion of a cassette exon within IRAK4 in K562 cells expressing each indicated U2AF1 allele. Error bars represent 95% confidence intervals as estimated by MISO55.

(E) As (D), but for cells expressing each indicated SRSF2 allele.