Understanding the mechanistic and functional consequences of splicing factor mutations

Joseph Pangallo

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Robert K. Bradley, Chair

Andrew Hsieh

Arvind Subramaniam

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Understanding the mechanistic and functional consequences of splicing factor mutations

Joseph Pangallo

Chair of the Supervisory Committee:
Full Member Robert K. Bradley
Fred Hutchinson Cancer Research Center

RNA splicing is a highly conserved eukaryotic process by which a precursor mRNA is converted into a mature mRNA. Precise regulation of RNA splicing is essential for proper cell development and maintenance. Disruption of RNA splicing is frequently associated with disease. Recent studies identified mutations in genes encoding RNA splicing factors in clonal hematopoiesis and diverse neoplastic diseases. The majority of spliceosomal mutations affect hotspot residues, while a subset of patients carry mutations in non-hotspot residues. Studies have demonstrated that some hotspot spliceosomal mutations result in splicing changes that promote disease; however, it is still unclear how remaining hotspot and non-hotspot spliceosomal mutations promote disease. To address this, I performed RNA-seq and quantified splicing dysregulation in isogenic cell lines and primary patient materials carrying splicing factor mutations. I identified

11 rare mutations in splicing factors *SRSF2* and *U2AF1* with likely disease pathogenicity, as well as two mis-spliced events with disease relevance driven by a hotspot mutation in *SF3B1*.

1. Rare and private spliceosomal gene mutations drive partial, complete, and dual phenocopies of hotspot alterations (Pangallo et al. Blood 2020). Genes encoding the RNA splicing factors *SRSF2* and *U2AF1* are subject to frequent missense mutations in diverse hematopoietic and other neoplastic diseases. Studies to date have demonstrated that hotspot mutations affecting *SRSF2* and *U2AF1* result in splicing changes that contribute to disease pathology. However, it is unclear how non-hotspot, or rare mutations affect splicing and disease pathology. In order to understand the disease relevance of rare mutations in *SRSF2* and *U2AF1*, we performed RNA-seq on isogenic cell lines and primary patient materials that express several rare *SRSF2* and *U2AF1* mutants. We found that 11 of 14 studied rare mutations induced distinct splicing alterations by phenocopying alterations in exon and splice site recognition that are induced by their hotspot counterparts (**Figure 1**). Our study suggests that many rare and private spliceosomal mutations contribute to disease pathology.

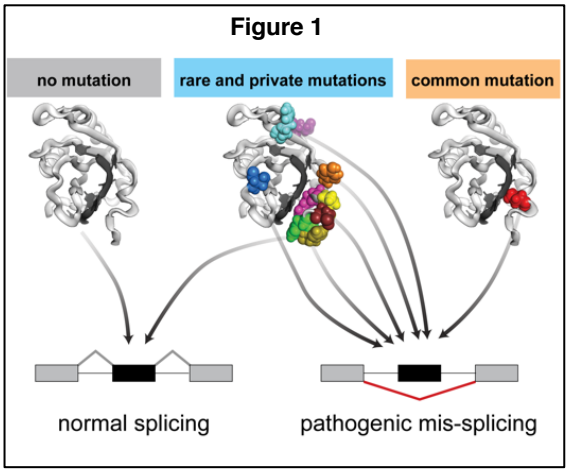

Figure 1. Cartoon illustrating that many rare and private mutations in SRSF2 and *U2AF1* induce mis-splicing that may contribute to disease.

2. Using MDS patient-derived induced pluripotent stem cells to determine the functional consequences of mis-splicing due to SF3B1 mutations. Mutations in RNA splicing factor *SF3B1* are observed in ~80% of patients with MDS-RARS (refractory anemia with ring sideroblasts). The high frequency as well as early occurrence of *SF3B1* mutations during disease development suggest that *SF3B1* mutations contribute to the pathology of MDS-RARS. However, direct links between *SF3B1* mutations and the phenotypes of MDS-RARS such as

impaired erythropoiesis are unclear. In order to study the contribution of *SF3B1* mutations to MDS-RARS, we performed RNA-seq in a SF3B1 mutant (G742D) MDS patient-derived induced pluripotent stem cell (iPSC) line undergoing erythropoiesis. We observed *SF3B1*G742D associated mis-splicing in *ABCB7* and *TMEM14C* (**Figure 2**), two genes whose protein products are involved in erythropoiesis. Our study suggests that mis-splicing of these key genes contribute to the pathology of MDS-RARS. Future studies will directly test the impact of mis-spliced *ABCB7* and *TMEM14C* on MDS-RARS phenotypes.

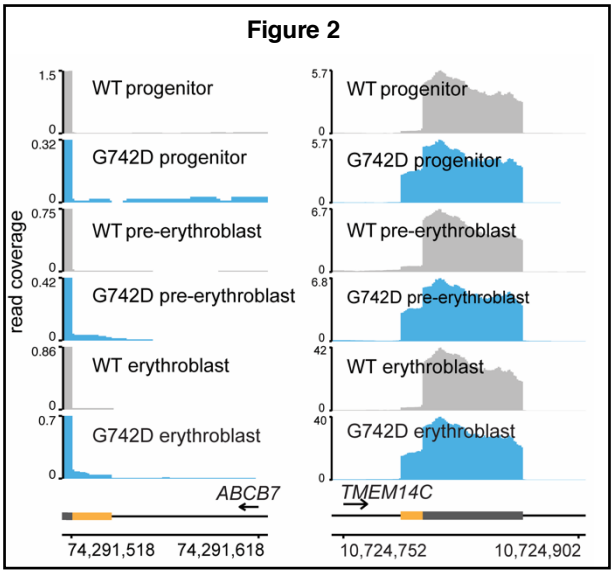

Figure 2. RNA-seq read coverage illustrating differential splicing at a competing 3' splice site in *TMEM14C*.

To sum up, in my thesis research, we found that rare and private mutations in *SRSF2* and *U2AF1* induce hotspot-like splicing alterations. As a result, we conclude that these mutations may contribute to disease pathology and should be studied further in a translational setting. Meanwhile, a common mutation affecting *SF3B1* in MDS patient-derived iPS cells drives mis-splicing of genes involved in erythropoiesis. Future studies will need to test whether these mis-spliced events directly drive an MDS phenotype such as impaired erythroid differentiation.

Together, these studies suggest that continued study of rare and common splicing factor mutations may reveal specific biomarkers that can be utilized to classify spliceosomal mutations as drivers or passengers for precision medicine.

TABLE OF CONTENTS

List of Figures

Chapter 2 Figures:

Chapter 3 Figures:

Acknowledgements

This dissertation would not have been possible without the help and support from many people.

First, I would like to thank Rob Bradley for being an amazing mentor and PI. It was very clear from the first day of my rotation that Rob is a great scientist, and an even better person. He deeply cares about each member of his lab. I really appreciate all the time Rob puts into mentoring. He knew that I was very interested in learning and improving my computational skills, so we discussed projects that would require computational analysis. Rob has helped me grow as an independent scientist, and always challenges me to ask specific questions while still remembering the big picture of the science. I feel very prepared for the next step of my scientific career because of the positive lab environment that Rob created.

I would also like to thank all members of the Bradley Lab, past and present. Each and every one of you has been an absolute joy to work with and has played an essential role in my scientific growth. Everyone who was already part of the lab when I joined – Heidi Dvinge, Janine Ilagan, Qing Feng, Sujatha Jagannathan, Heather Johns – were great scientific role models. I had the privilege of getting to follow their impressive scientific development. I really appreciate how kind, thoughtful, and brilliant they all are. They provided me with a great foundation for my own scientific growth. They also started a daily coffee break tradition, which was a great source for science and life discussions. Everyone who joined the lab after me – Guo-Liang "Chewie" Chew, Dylan Udy, Khrystyna North, Jose Pineda, Jake Polaski, James Thomas, Emma Hoppe, Emma De Neef, Austin Gabel, Andrea Belleville – helped me continue the tradition of a friendly and collaborative lab environment. I look forward to seeing where their

scientific careers take them. Also, I really want to thank them for continuing our daily coffee tradition.

Outside of the Bradley Lab, I would really like to thank my committee of Andrew Hsieh, Arvind Subramaniam, Stephen Tapscott, and Savan Ram. I really appreciate all the time they have taken out of their busy schedule to guide my thesis research. I want to thank them for asking me challenging questions during meetings. Many of those questions forced me to step out of my scientific comfort zone and learn new things. I would also like to thank neighboring labs for help and support at various times throughout graduate school. Thank you to the Computational Biology program administrators, shared resource facilities, the MCB graduate program, and fellow classmates. I was surrounded by wonderful colleagues and friends at both UW and Fred Hutch.

Finally, I want to thank my parents, sister, in-laws, and my wife Taylor. Their love and support have been instrumental in helping me navigate through life and graduate school, and I could never thank them enough.

Chapter 1. Introduction

RNA splicing is a highly conserved eukaryotic process by which a precursor mRNA is converted into a mature mRNA. RNA splicing allows for the production of multiple protein isoforms arising from a single precursor mRNA. As a result, precise regulation of RNA splicing is essential for proper cell development and maintenance[1-4] (e.g, neural development, hematopoiesis).

Disruption of normal RNA splicing is frequently associated with disease. This includes abnormal intron retention associated with cancer transcriptomes[5], point mutations in splice sites causing mis-splicing of the corresponding intron and promotion of disease[6], as well as mutations in genes encoding RNA splicing factors which alter splicing of hundreds to thousands of genes[7-16].

Mutations in genes encoding RNA splicing factors are among the most common genetic alterations observed in hematologic malignancies, and to a lesser extent, solid tumors[7-12]. RNA splicing factor mutations are most commonly observed in *SF3B1*, *SRSF2*, and *U2AF1* at a specific set of hotspot residues[13-14]. Mutations in *SF3B1*, *SRSF2*, and/or *U2AF1* are observed in patients at frequencies of 50-60% in myelodysplastic syndromes (MDS) and related diseases[15], 5-18% in chronic lymphocytic leukemia[11,12,16,17], 5-25% in acute myeloid leukemia (AML) in adults[18], and 14-29% in uveal melanoma[19,20].

The high frequency with which splicing factor mutations occur as well as recent functional studies suggest the mutations drive disease. Numerous examples of altered RNA splicing of specific target genes driving disease phenotypes have been observed. For example, *U2AF1* mutations induce mis-splicing of *IRAK4*, thereby leading to aberrant innate immune signaling[21]. *SRSF2* mutations induce mis-splicing of *EZH2*, which results in impaired hematopoiesis[22]. *SF3B1* mutations induce mis-splicing of *BRD9*, resulting in tumorigenesis[23].

However, due to the diverse mutational landscape as well as the high number of diseases that are associated with splicing factor mutations, many links between altered RNA splicing and disease phenotypes have not been identified.

Although most mutations in *SF3B1*, *SRSF2*, and *U2AF1* affect specific hotspot residues, a small subset of patients carry mutations in non-hotspot residues. The functional consequences of these rare mutations, and in turn, their contribution to disease, are unclear.

In this dissertation, I sought to address two main questions :

1. What is the likelihood that non-hotspot splicing factor mutations drive disease?

In Chapter 2, I describe a study in which we infer the likely disease relevance of rare and private mutations in *SRSF2* and *U2AF1*. This was motivated by recent work that demonstrated three CLL patients with novel *SF3B1* in-frame deletions whose splicing profiles mimicked patients with hotspot mutations. We hypothesized that rare and private mutations in *SRSF2* and *U2AF1* might also mimic splicing profiles of their hotspot counterparts, and are therefore likely candidate driver mutations. To test this, we performed RNA-seq in both isogenic cell lines and primary patient materials. We characterized the effects of 14 rare or private mutations in *SRSF2* and *U2AF1* on their ability to alter exon and splice site recognition, and found that 11 out of 14 mutations mimicked the effects of their hotspot counterparts.

2. How do splicing factor mutations contribute to disease phenotypes?

Mutations in each individual splicing factor are often associated with different subtypes of disease. For example, ~80% of patients with MDS-RARS (refractory anemia with ring

sideroblasts) carry a mutation in *SF3B1*. The specificity of *SF3B1* mutations in MDS-RARS suggests that *SF3B1* mutations induce a specific set of molecular changes that contribute to MDS-RARS phenotypes (e.g., erythropoiesis). In Chapter 3, I describe a study in which we performed RNA-seq in a *SF3B1* mutant iPS cell line model at successive stages of erythropoiesis. We identified two SF3B1 mutant-induced mis-spliced genes (*ABCB7*, *TMEM14C)* that could potentially explain the impaired erythropoiesis phenotype observed in MDS-RARS patients.

References

1. Weyn-Vanhentenryck SM, Feng H, Ustianenko D, et al. Precise temporal regulation of alternative splicing during neural development. *Nat Commun*. 2018;9(1):2189.
2. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol*. 2017;18(7):437-451.
3. Pimentel H, Parra M, Gee S, et al. A dynamic alternative splicing program regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res*. 2014;42(6):4031-4042.
4. Pimentel H, Parra M, Gee SL, Mohandas N, Pachter L, Conboy JG. A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res*. 2016;44(2):838-851.
5. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med*. 2015;7(1):45. Published 2015 May 15.
6. Scotti MM, Swanson MS. RNA mis-splicing in disease. Nat Rev Genet. 2016;17(1):19-32.
7. Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*. 2011;478(7367):64–69.
8. Graubert TA, Shen D, Ding L, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat Genet*. 2011.
9. Papaemmanuil E, Cazzola M, Boultwood J, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med*. 2011;365(15):1384–1395.
10. Visconte V, Makishima H, Jankowska A, et al. SF3B1, a splicing factor is frequently mutated in refractory anemia with ring sideroblasts. *Leukemia*. 2011.

11. Wang L, Lawrence MS, Wan Y, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med*. 2011;365(26):2497–2506.
12. Quesada V, Conde L, Villamor N, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet*. 2011.
13. Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer*. 2016;16(7):413–430.
14. Seiler M, Peng S, Agrawal AA, et al. Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. *Cell Rep*. 2018;23(1):282–296.e4.
15. Kennedy JA, Ebert BL. Clinical Implications of Genetic Mutations in Myelodysplastic Syndrome. J Clin Oncol. 2017;35(9):968-974.
16. Rossi D, Bruscaggin A, Spina V, et al. Mutations of the SF3B1 splicing factor in chronic lymphocytic leukemia: association with progression and fludarabine-refractoriness. *Blood*. 2011;118(26):6904–6908.
17. Ramsay AJ, Rodríguez D, Villamor N, et al. Frequent somatic mutations in components of the RNA processing machinery in chronic lymphocytic leukemia. Leukemia. 2012.
18. Yoshimi A, Lin K-T, Wiseman DH, et al. Coordinated alterations in RNA splicing and epigenetic regulation drive leukemogenesis. *Nature.* In press.
19. Martin M, Maßhöfer L, Temming P, et al. Exome sequencing identifies recurrent somatic mutations in EIF1AX and SF3B1 in uveal melanoma with disomy 3. *Nat Genet*. 2013.
20. Harbour JW, Roberson EDO, Anbunathan H, et al. Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma. *Nat Genet.* 2013.
21. Smith MA, Choudhary GS, Pellagatti A, et al. U2AF1 mutations induce oncogenic IRAK4 isoforms and activate innate immune pathways in myeloid malignancies. *Nat Cell Biol.* 2019;21(5):640-650.
22. 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.
23. Inoue D, Chew GL, Liu B, et al. Spliceosomal disruption of the non-canonical BAF complex in cancer. *Nature.* 2019;574(7778):432-436.

Chapter 2. Rare and private spliceosomal gene mutations drive partial, complete, and dual phenocopies of hotspot alterations

This research was originally published in *Blood*. Pangallo J, Kiladjian J-J, Cassinat B, Renneville A, Taylor J, Polaski JT, North K, Abdel-Wahab O, Bradley RK. Rare and private spliceosomal gene mutations drive partial, complete, and dual phenocopies of hotspot alterations. *Blood*. 2020;135(13):1032-1043.

2.1 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 splicing changes 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 distinct splicing alterations, including partially or completely phenocopying the alterations in exon and splice site recognition 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 molecular assays to inform precision medicine by inferring the potential disease relevance of newly discovered mutations.

2.2 Introduction

Somatic mutations in genes encoding RNA splicing factors are among the most common genetic changes observed in many hematologic malignancies[1-6]. 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[7,8]. 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) [5,6,9,10], 5-25% of acute myeloid leukemia (AML) in adults[11], and 14-29% in uveal melanoma[12,13].

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[14,15]. In many cases, concrete links between altered RNA splicing, 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[16-19]; *SRSF2* mutations alter exonic splicing enhancer recognition to cause *EZH2* mis-splicing and impaired hematopoiesis[20,21]; *U2AF1* mutations alter 3' splice site recognition to cause *IRAK4* mis-splicing and aberrant innate immune signaling[22-24].

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 (**Fig. 1A-B**) [25]. 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 molecular 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[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[31], as well as our recent finding that both rare and common *SF3B1* mutations converge on *BRD9* mis-splicing across cancer types[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[20-24]. We hypothesized that rare or private *SRSF2* and *U2AF1* mutations that phenocopied hotspot-induced changes in splicing were candidate drivers, while mutations that induced few or 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**).

2.3 Results

Diverse *SRSF2* and *U2AF1* mutations alter RNA splicing programs

We queried the Catalogue of Somatic Mutations in Cancer (COSMIC) database[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 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[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[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[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[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.

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

We sought to determine how 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 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 ≥10% and Bayes factor ≥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, *SRSF2*P95H/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 *SRSF2*P95H hotspot mutation (**Fig. 2B, S4**). Our approach allowed us to deconvolve complex co-mutation events such as *SRSF2*P95_R102del+P107H. *SRSF2*P95_R102del alone phenocopied *SRSF2*P95 mutations,

while *SRSF2*P107H 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 AML[20,35], CMML[20], and MDS[34]. We identified samples carrying *SRSF2*S54A/F, *SRSF2*R94_P95insR, and *SRSF2*P95_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, *SRSF2*S54A/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 *SRSF2*R86_G93dup and the common mutation *SRSF2*P95H 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 motifs[20], 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 *SRSF2*R86_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 *SRSF2*R94_P95insR, for these assays. Cells expressing *SRSF2*R86_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. *U2AF1*S34F/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 *U2AF1*S34F and Q157R altered recognition of the -3 and +1 sites, as expected. *U2AF1*R156H phenocopied *U2AF1*Q157P/R, as did the rare insertion *U2AF1*E159_M160insYE. The complex co-mutation *U2AF1*S34F_Q157R drove a "dual" phenocopy, characterized by S34 and Q157 hotspot-like alterations at both the -3 and +1 positions. The rare mutation *U2AF1*I24T, 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 *U2AF1*S34F_Q157R, while *U2AF1*I24V was similar to *U2AF1*Q157R (**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 *U2AF1*I24T 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 *U2AF1*I24T). We used RT-PCR to experimentally validate results from RNA-seq, confirming that *U2AF1*I24T induced similar patterns of mis-splicing as did *U2AF1*S34F 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 *U2AF1*S34F/Y in both human patients and murine models whose mis-splicing contributes to impaired hematopoiesis[23,44,45]. Like *U2AF1*S34F, the rare mutations *U2AF1*I24T/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 *U2AF1*I24V cells, where we focused on *U2AF1*I24V 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 *U2AF1*I24V 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, 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 *SRSF2*P95-like mutations and compared that set to differentially spliced exons found in association with three of the *U2AF1*S34-like mutations. As expected given *SRSF2* and *U2AF1* mutations' distinct consequences for splicing, 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 *U2AF1*S34F/Y promote inclusion of the upstream exon of two mutually exclusive exons within *H2AFY*, which encodes macro-H2A1, thereby perturbing erythroid and granulomonocytic differentiation[23,44,45]. *U2AF1*S34F 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[24] (**Fig. 4D**). Our analysis revealed that the rare mutations *U2AF1*I24T/V phenocopied the *H2AFY* and *IRAK4* mis-splicing characteristic of *U2AF1*S34F/Y-mutant cells, and furthermore that both *SRSF2*P95 and *SRSF2*P95-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.

2.4 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[20,23,39,50]. We therefore took advantage of the stereotyped changes in RNA splicing caused by *SRSF2* and *U2AF1* hotspot mutations—which have been directly linked to disease phenotypes[20-24]—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 the latter can classify pathogenicity with reasonable confidence, given the known role of dysfunctional exon and splice site recognition in *SRSF2*- and *U2AF1*-mutant hematologic malignancies.

Our 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 means by which they promote cancer, in a specific manner[29]. Therefore, the concordance between our 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 exon or splice site recognition are likely passengers. However, we cannot rule out the possibility that some rare mutations promote disease through means 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)[47,48]. Conversely, although our 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[51] 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 *SRSF2*P95H/L/R[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 *SRSF2*H99L 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 generate molecular phenocopies of the *SRSF2*P95, *U2AF1*S34, and *U2AF1*Q157 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[15] when feasible. Finally, when therapies designed to specifically target cells with spliceosomal mutations enter clinical practice[38,40,52,53], patients bearing non-hotspot spliceosomal mutations should be considered as candidates for these therapies. Although performing a 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.

REFERENCES

1. Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*. 2011;478(7367):64–69.
2. Graubert TA, Shen D, Ding L, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat Genet*. 2011.
3. Papaemmanuil E, Cazzola M, Boultwood J, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med*. 2011;365(15):1384–1395.
4. Visconte V, Makishima H, Jankowska A, et al. SF3B1, a splicing factor is frequently mutated in refractory anemia with ring sideroblasts. *Leukemia*. 2011.
5. Wang L, Lawrence MS, Wan Y, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med*. 2011;365(26):2497–2506.
6. Quesada V, Conde L, Villamor N, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet*. 2011.
7. Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer*. 2016;16(7):413–430.
8. Seiler M, Peng S, Agrawal AA, et al. Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. *Cell Rep*. 2018;23(1):282–296.e4.
9. Rossi D, Bruscaggin A, Spina V, et al. Mutations of the SF3B1 splicing factor in chronic lymphocytic leukemia: association with progression and fludarabine-refractoriness. *Blood*. 2011;118(26):6904–6908.
10. Ramsay AJ, Rodríguez D, Villamor N, et al. Frequent somatic mutations in components of the RNA processing machinery in chronic lymphocytic leukemia. *Leukemia*. 2012.
11. Yoshimi A, Lin K-T, Wiseman DH, et al. Coordinated alterations in RNA splicing and epigenetic regulation drive leukemogenesis. *Nature*. In press.
12. Martin M, Maßhöfer L, Temming P, et al. Exome sequencing identifies recurrent somatic mutations in EIF1AX and SF3B1 in uveal melanoma with disomy 3. *Nat Genet*. 2013.
13. Harbour JW, Roberson EDO, Anbunathan H, et al. Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma. *Nat Genet*. 2013.
14. Abelson S, Collord G, Ng SWK, et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature*. 2018;559(7714):400–404.
15. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat. Med.* 2018;24(7):1015–1023.
16. DeBoever C, Ghia EM, Shepard PJ, et al. Transcriptome Sequencing Reveals Potential Mechanism of Cryptic 3' Splice Site Selection in SF3B1-mutated Cancers. *PLoS Comput Biol*. 2015;11(3):e1004105.
17. Darman RB, Seiler M, Agrawal AA, et al. Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. *Cell Rep*. 2015;13(5):1033–1045.
18. Alsafadi S, Houy A, Battistella A, et al. Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nat Commun*. 2016;7:10615.
19. Inoue D, Chew G-L, Liu B, et al. Spliceosomal disruption of the non-canonical BAF complex in cancer. *Nature*. In press.
20. 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.

21. Zhang J, Lieu YK, Ali AM, et al. Disease-associated mutation in SRSF2 misregulates splicing by altering RNA-binding affinities. *Proceedings of the National Academy of Sciences*. 2015.
22. Brooks AN, Choi PS, de Waal L, et al. A Pan-Cancer Analysis of Transcriptome Changes Associated with Somatic Mutations in U2AF1 Reveals Commonly Altered Splicing Events. *PLoS ONE*. 2014;9(1):e87361.
23. Ilagan JO, Ramakrishnan A, Hayes B, et al. U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res*. 2015;25(1):14–26.
24. Smith MA, Choudhary GS, Pellagatti A, et al. U2AF1 mutations induce oncogenic IRAK4 isoforms and activate innate immune pathways in myeloid malignancies. *Nat Cell Biol*. 2019;21(5):640–650.
25. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019;47(D1):D941–D947.
26. Starita LM, Young DL, Islam M, et al. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics*. 2015;200(2):413–422.
27. Starita LM, Islam MM, Banerjee T, et al. A Multiplex Homology-Directed DNA Repair Assay Reveals the Impact of More Than 1,000 BRCA1 Missense Substitution Variants on Protein Function. *Am J Hum Genet*. 2018;103(4):498–508.
28. Findlay GM, Daza RM, Martin B, et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature*. 2018;372(7726):2235–222.
29. Berger AH, Brooks AN, Wu X, et al. High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer Cell*. 2016;30(2):214–228.
30. Kim E, Ilic N, Shrestha Y, et al. Systematic Functional Interrogation of Rare Cancer Variants Identifies Oncogenic Alleles. *Cancer Discov*. 2016;6(7):714–726.
31. Agrawal AA, Seiler M, Brinton LT, et al. Novel SF3B1 in-frame deletions result in aberrant RNA splicing in CLL patients. *Blood Adv*. 2017;1(15):995–1000.
32. 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.
33. Durham BH, Getta B, Dietrich S, et al. Genomic analysis of hairy cell leukemia identifies novel recurrent genetic alterations. *Blood*. 2017;130(14):1644–1648.
34. Pellagatti A, Armstrong RN, Steeples V, et al. Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. *Blood*. 2018;132(12):blood–2018–04–843771–1240.
35. Lavallée V-P, Baccelli I, Krosl J, et al. The transcriptomic landscape and directed chemical interrogation of MLL-rearranged acute myeloid leukemias. *Nat Genet*. 2015.
36. Zheng S, Cherniack AD, Dewal N, et al. Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. *Cancer Cell*. 2016;29(5):723–736.
37. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059–2074.
38. Lee SC-W, Dvinge H, Kim E, et al. Modulation of splicing catalysis for therapeutic targeting of leukemia with mutations in genes encoding spliceosomal proteins. *Nat. Med.* 2016.
39. Fei DL, Motowski H, Chatrikhi R, et al. Wild-Type U2AF1 Antagonizes the Splicing Program Characteristic of U2AF1-Mutant Tumors and Is Required for Cell Survival. *PLoS Genet*. 2016;12(10):e1006384.

40. Shirai CL, White BS, Tripathi M, et al. Mutant U2AF1-expressing cells are sensitive to pharmacological modulation of the spliceosome. *Nat Commun*. 2017;8:14060.
41. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med*. 2015;7(1):45.
42. Liang Y, Tebaldi T, Rejeski K, et al. SRSF2 mutations drive oncogenesis by activating a global program of aberrant alternative splicing in hematopoietic cells. *Leukemia*. 2018;478:64.
43. Okeyo-Owuor T, White BS, Chatrikhi R, et al. U2AF1 Mutations alter sequence specificity of pre-mRNA binding and splicing. *Leukemia*. 2014.
44. Shirai CL, Ley JN, White BS, et al. Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo. *Cancer Cell*. 2015;27(5):631–643.
45. Yip BH, Steeples V, Repapi E, et al. The U2AF1S34F mutation induces lineage-specific splicing alterations in myelodysplastic syndromes. *J Clin Invest*. 2017.
46. Palangat M, Anastasakis DG, Fei DL, et al. The splicing factor U2AF1 contributes to cancer progression through a noncanonical role in translation regulation. *Genes Dev*. 2019;33(9-10):482–497.
47. Chen L, Chen J-Y, Huang Y-J, et al. The Augmented R-Loop Is a Unifying Mechanism for Myelodysplastic Syndromes Induced by High-Risk Splicing Factor Mutations. *Mol Cell*. 2018;69(3):412–425.e6.
48. Nguyen HD, Leong WY, Li W, et al. Spliceosome Mutations Induce R loop-Associated Sensitivity to ATR Inhibition in Myelodysplastic Syndrome. *Cancer Res*. 2018;78(18):canres.3970.2017–5374.
49. Haferlach T, Nagata Y, Grossmann V, et al. Landscape of Genetic Lesions in 944 Patients with Myelodysplastic Syndromes. *Leukemia*. 2013.
50. Fei DL, Zhen T, Durham B, et al. Impaired hematopoiesis and leukemia development in mice with a conditional knock-in allele of a mutant splicing factor gene U2af1. *Proceedings of the National Academy of Sciences*. 2018;46(44):201812669–E10446.
51. Daubner GM, Cléry A, Jayne S, et al. A syn-anti conformational difference allows SRSF2 to recognize guanines and cytosines equally well. *EMBO Journal*. 2011;31(1):162–74.
52. Obeng EA, Chappell RJ, Seiler M, et al. Physiologic Expression of Sf3b1(K700E) Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation. *Cancer Cell*. 2016;30(3):404–417.
53. Seiler M, Yoshimi A, Darman R, et al. H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. *Nat. Med.* 2018.
54. Przychodzen B, Jerez A, Guinta K, et al. Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms. *Blood*. 2013.
55. 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.

2.5 Tables

	Mutation	n	Mechanistic classification	Evidence	Reference
SRSF2	S54A	1	partial phenocopy of P95	cell line + patient	this study
	S54F	1	partial phenocopy of P95	cell line + patient	this study
	R86_G93dup	1	phenocopy of P95	cell line	this study
	R94_P95insR	11	phenocopy of P95	cell line + patient	this study
	P95H	448	hotspot	cell line + patient	(previously studied)
	P95L	280	hotspot	cell line + patient	(previously studied)
	P95R	168	hotspot	cell line + patient	(previously studied)
	P95_R102del	79	phenocopy of P95	cell line + patient	this study
	P05_R102del + P107H	7	phenocopy of P95	cell line	this study
	P107H	7	silent	cell line	this study
	H99L	2	silent	cell line	this study
U2AF1	I24T	5	dual phenocopy of S34 and Q157 (likely)	cell line + patient	this study
	I24V	1	phenocopy of S34 (likely)	cell line	this study
	S34F	308	hotspot	cell line + patient	(previously studied)
	S34Y	92	hotspot	cell line + patient	(previously studied)
	R156H	30	phenocopy of Q157	cell line + patient	this study
	R156Q	2	silent	cell line	this study
	Q157R	66	hotspot	cell line + patient	(previously studied)
	Q157P	121	hotspot	cell line + patient	(previously studied)
	E159_M160insYE	8	phenocopy of Q157	cell line + patient	this study
	S34F + Q157R	1	dual phenocopy of S34 and Q157	cell line	this study

Table 1. Mechanistic classification of studied mutations. n, number of times that each mutation has been reported in COSMIC. Classification inferred from exonic splicing enhancer preferences and 3' splice site preferences associated with each mutation. The consequences of hotspot *SRSF2* and *U2AF1* mutations were previously studied by several groups[20-23,54].

2.6 Figures and legends

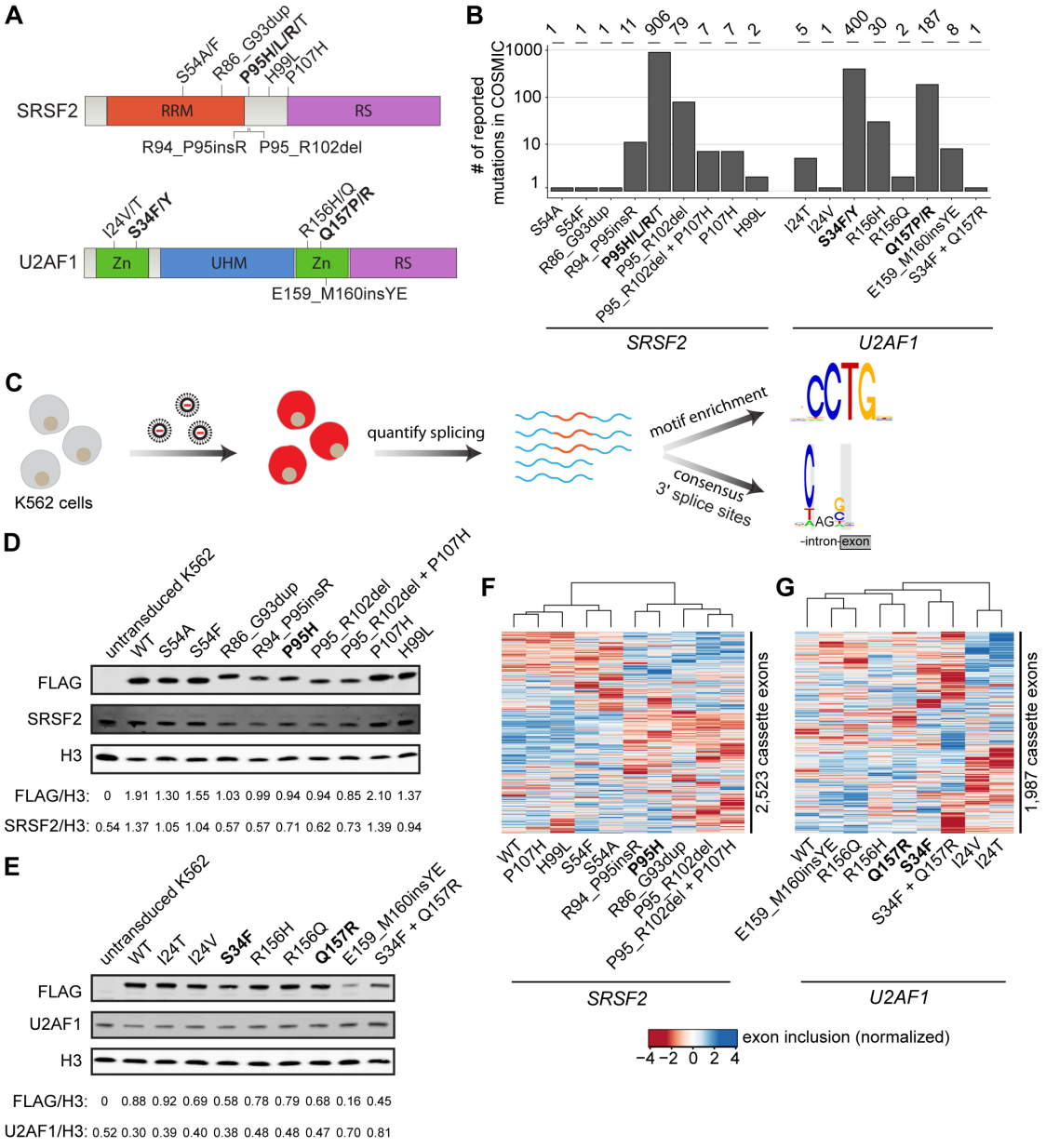

Figure 1. Strategy for classification of rare, non-hotspot *SRSF2* and *U2AF1* mutations.
(**A**) Hotspot (bold) and select rare and private mutations affecting SRSF2 and U2AF1. RRM, RNA recognition motif; RS, arginine/serine-rich domain; UHM, U2AF homology motif; Zn, zinc finger domain.
(**B**) Numbers of reported mutations in *SRSF2* and *U2AF1* in the Catalogue of Somatic Mutations in Cancer (COSMIC) database as of September 17, 2018. *SRSF2*S54A 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.

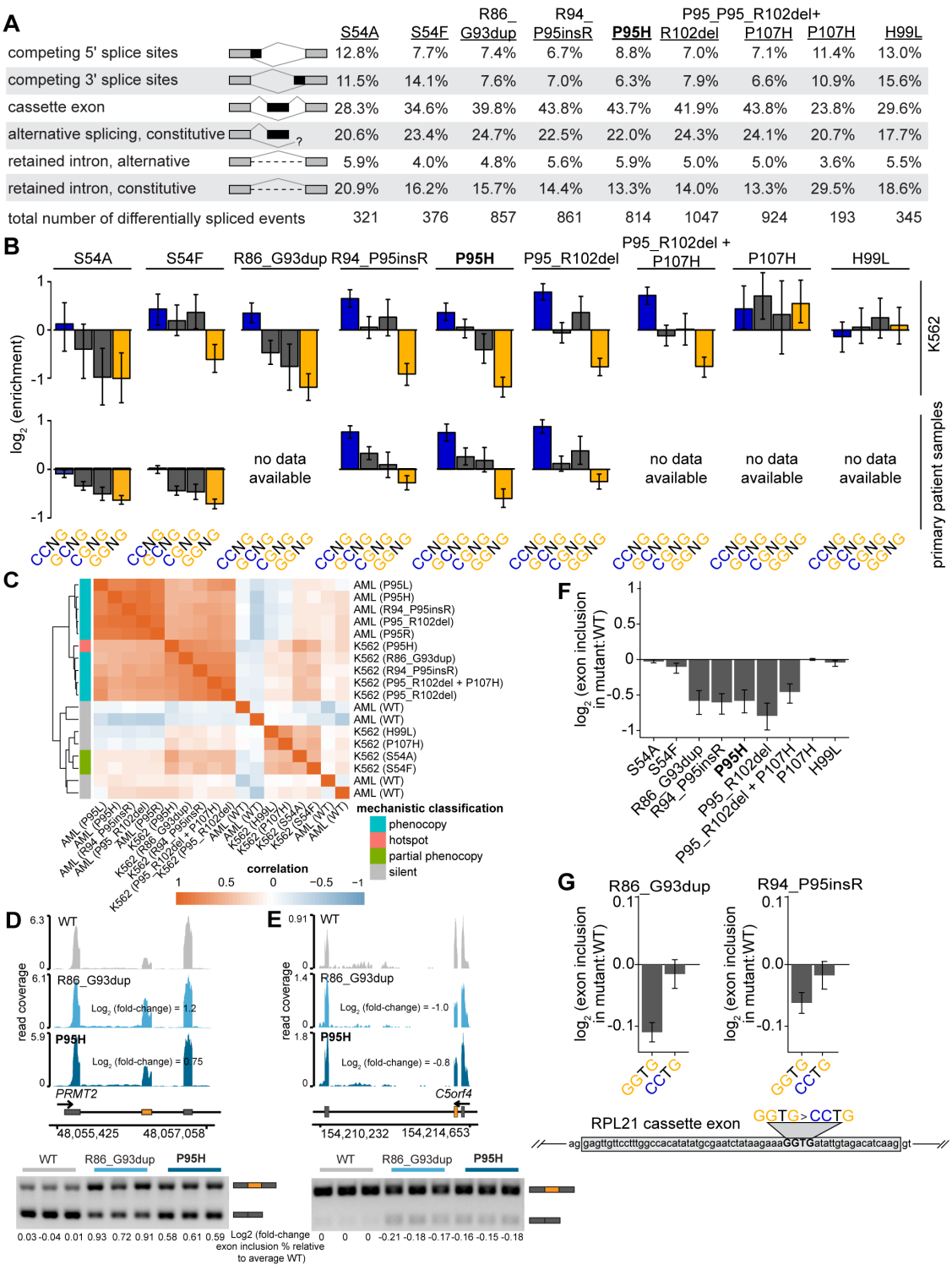

Figure 2. Rare mutations in *SRSF2* alter exonic splicing enhancer (ESE) preference. (**A**) Differentially spliced events identified in K562 cells expressing each indicated *SRSF2* mutant allele relative to WT-expressing control cells. Percentages represent the distribution of differentially spliced 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 *SRSF2*S54A (polycythemia + hyperleukocytosis + myelofibrosis) and *SRSF2*S54F (chronic myelomonocytic leukemia) were sequenced for this study; RNA-seq data from patient samples bearing *SRSF2*R94_P95insR (AML), *SRSF2*P95H (CMML), and *SRSF2*P95_R102del (AML) were previously published[20].
(**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 published[20].
(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.

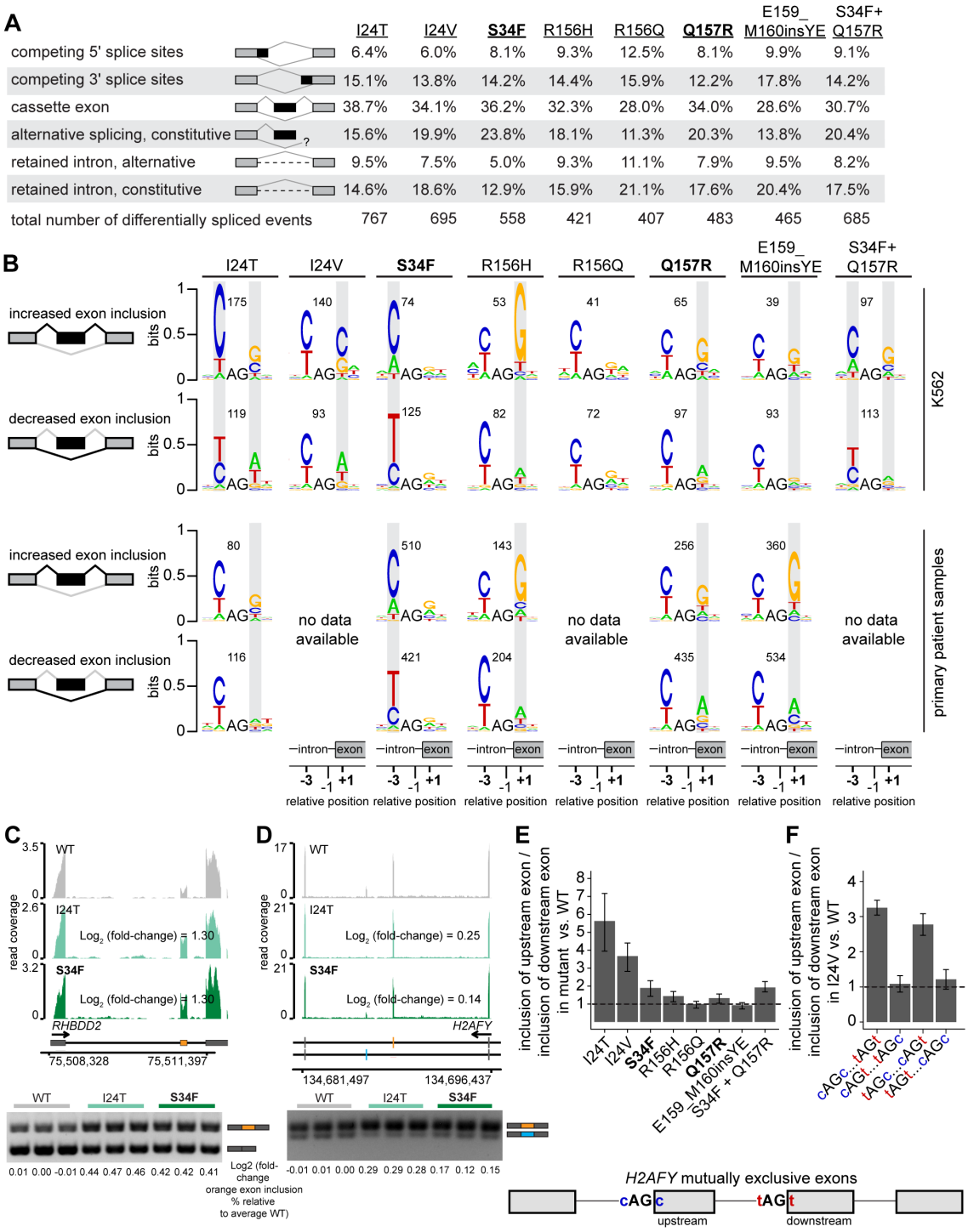

Figure 3. Rare mutations in *U2AF1* alter 3' splice site recognition.

(**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 *U2AF1*S34F/Y or *U2AF1*Q157P/R hotspot mutations. RNA-seq data from patient samples bearing *U2AF1*I24T (adrenocortical carcinoma), *U2AF1*S34F (AML), *U2AF1*R156H (MDS), *U2AF1*Q157R (AML), and *U2AF1*E159_M160insYE (AML) were previously published[34-37].
(C) Top, RNA-seq read coverage illustrating increased cassette exon inclusion in RHBDD2 in K562 cells expressing either a hotspot (S34F) or rare (I24T) 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.

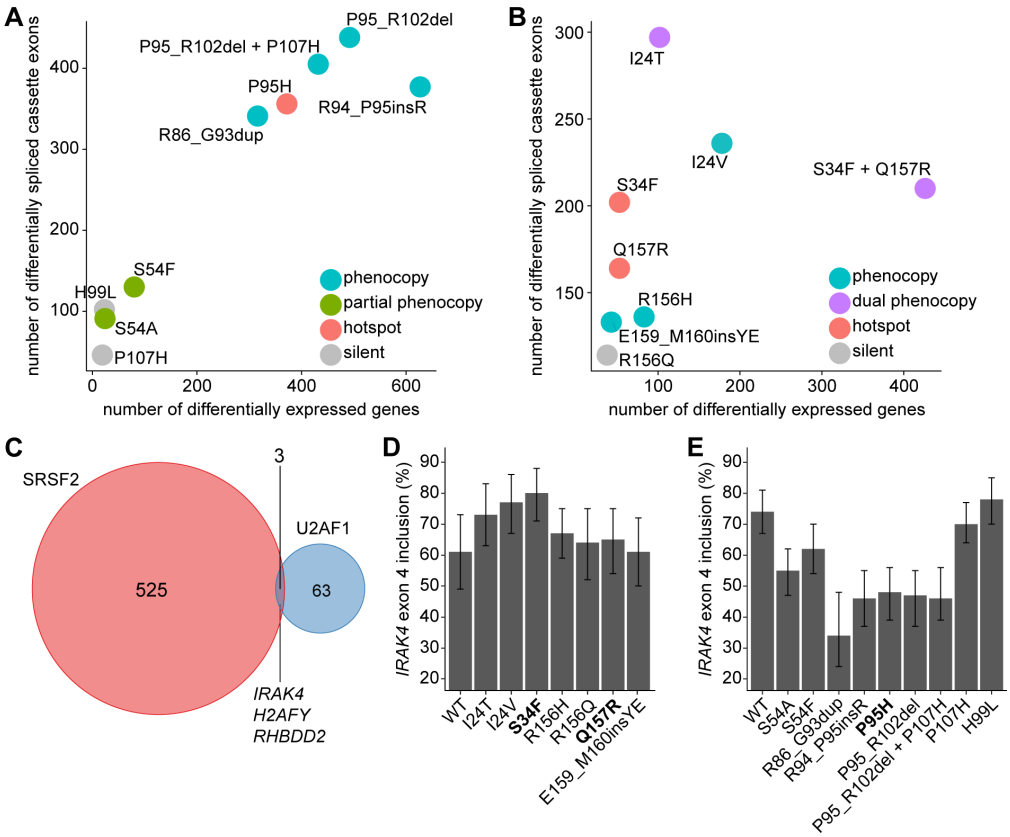

Figure 4. Hotspot and rare *SRSF2* and *U2AF1* induce transcriptome dysregulation and converge on *H2AFY* and *IRAK4* mis-splicing.

(**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 ≥ 1 TPM in both samples, $|\log_2(\text{fold-change})| \geq \log_2(1.5)$, and Bayes factor ≥ 10. See **Table 1** for additional information on 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 ≥ 10% and a Bayes factor ≥ 1. Diagram restricted to genes containing cassette exons or mutually exclusive exons that were differentially spliced in association with at least three *SRSF2*P95-like mutations (*SRSF2*R86_G93dup, *SRSF2*R94_P95insR, *SRSF2*P95H, *SRSF2*P95_R102del, and *SRSF2*P95_R102del + P107H considered) and three *U2AF1*S34-like mutations (*U2AF1*I24T, *U2AF1*I24V, and *U2AF1*S34F 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 MISO[55].

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

2.7 Supplementary figures and legends

Figure S1 – related to Figure 1

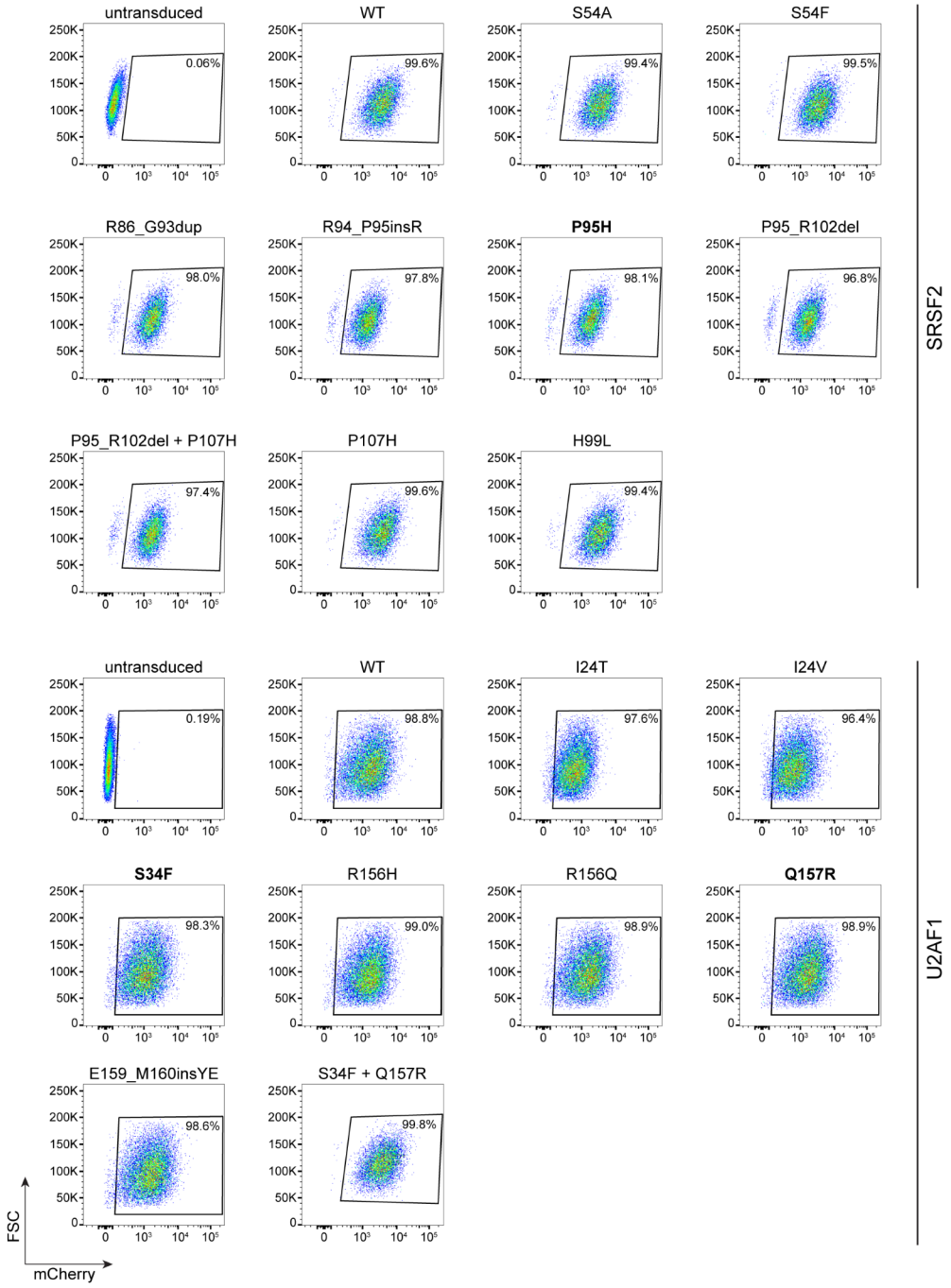

Figure S1 – related to Figure 1. Establishment of K562 cell lines stably expressing transgenic *SRSF2* or *U2AF1*.

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.

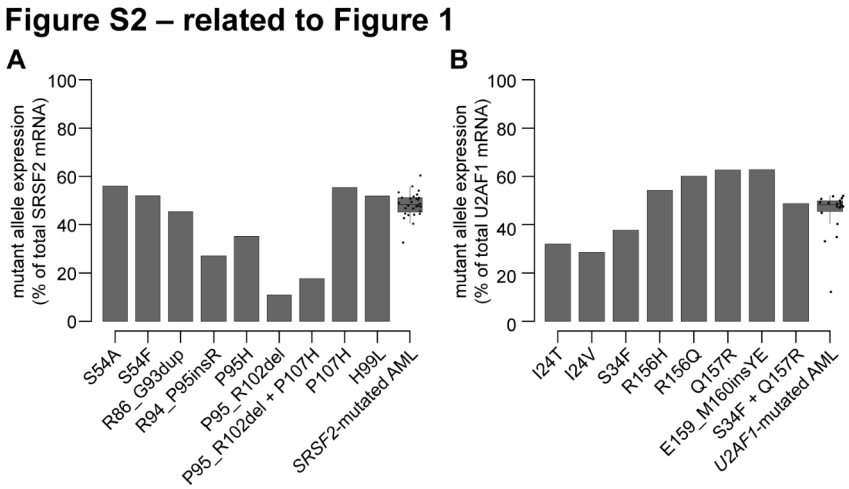

Figure S2 – related to Figure 1. Mutant allele expression in transgenic K562 cell lines.
(**A**) Expression of each indicated *SRSF2* allele as a fraction of total SRSF2 mRNA. Box plot indicates mutant *SRSF2* allelic expression in primary AML samples (Lavallée et al, *Nature Genetics*, 2015). Mutant 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

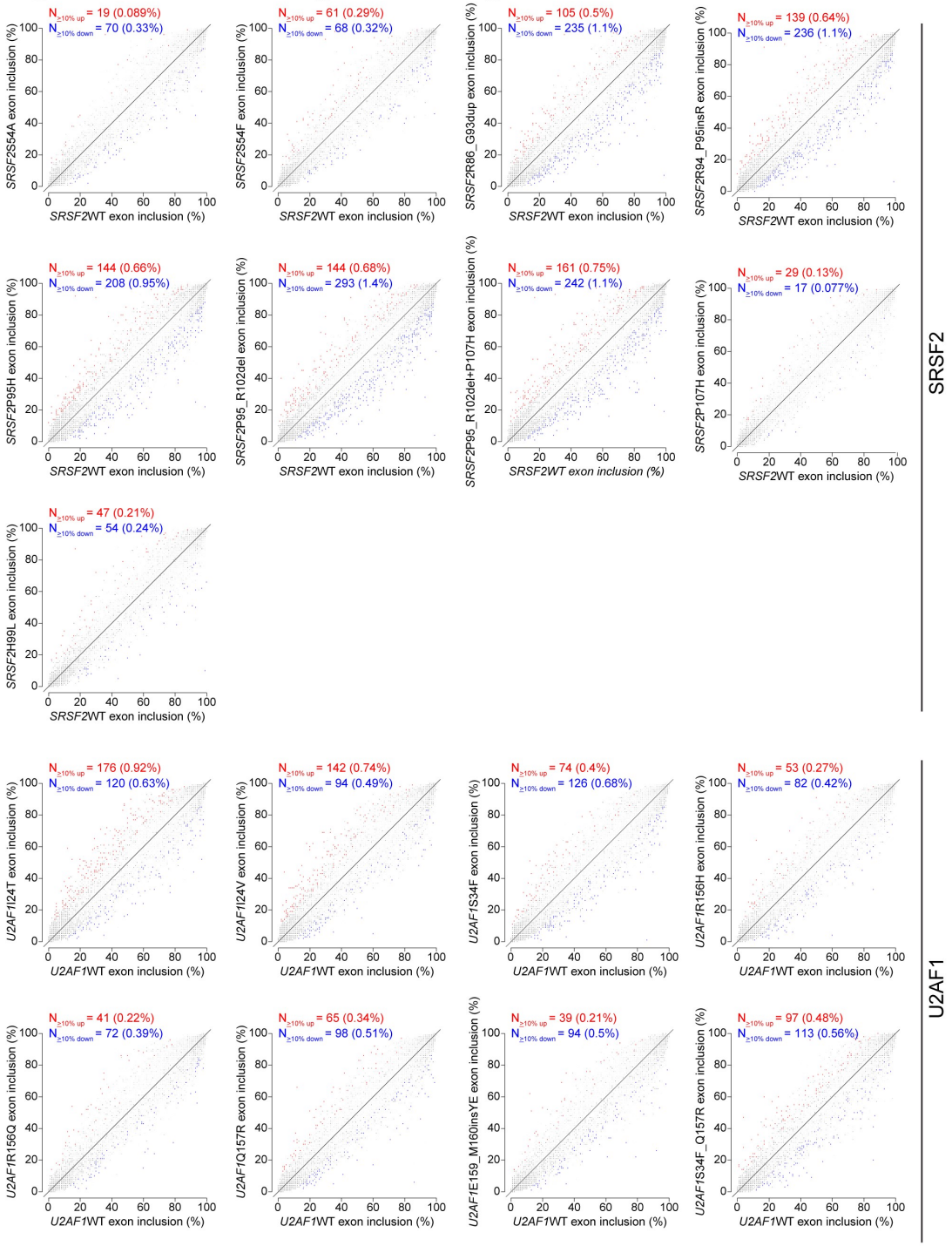

Figure S3 – related to Figures 2 and 3. Cassette exon inclusion in transgenic K562 cell lines. Scatter plots illustrating cassette exon inclusion in K562 cell lines expressing transgenic WT (x axis) or mutant (y axis) alleles of SRSF2 or U2AF1. Red, cassette exons exhibiting significantly increased inclusion 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 **Materials and methods** for descriptions of identification of significant differential splicing.

Figure S4 – related to Figure 2

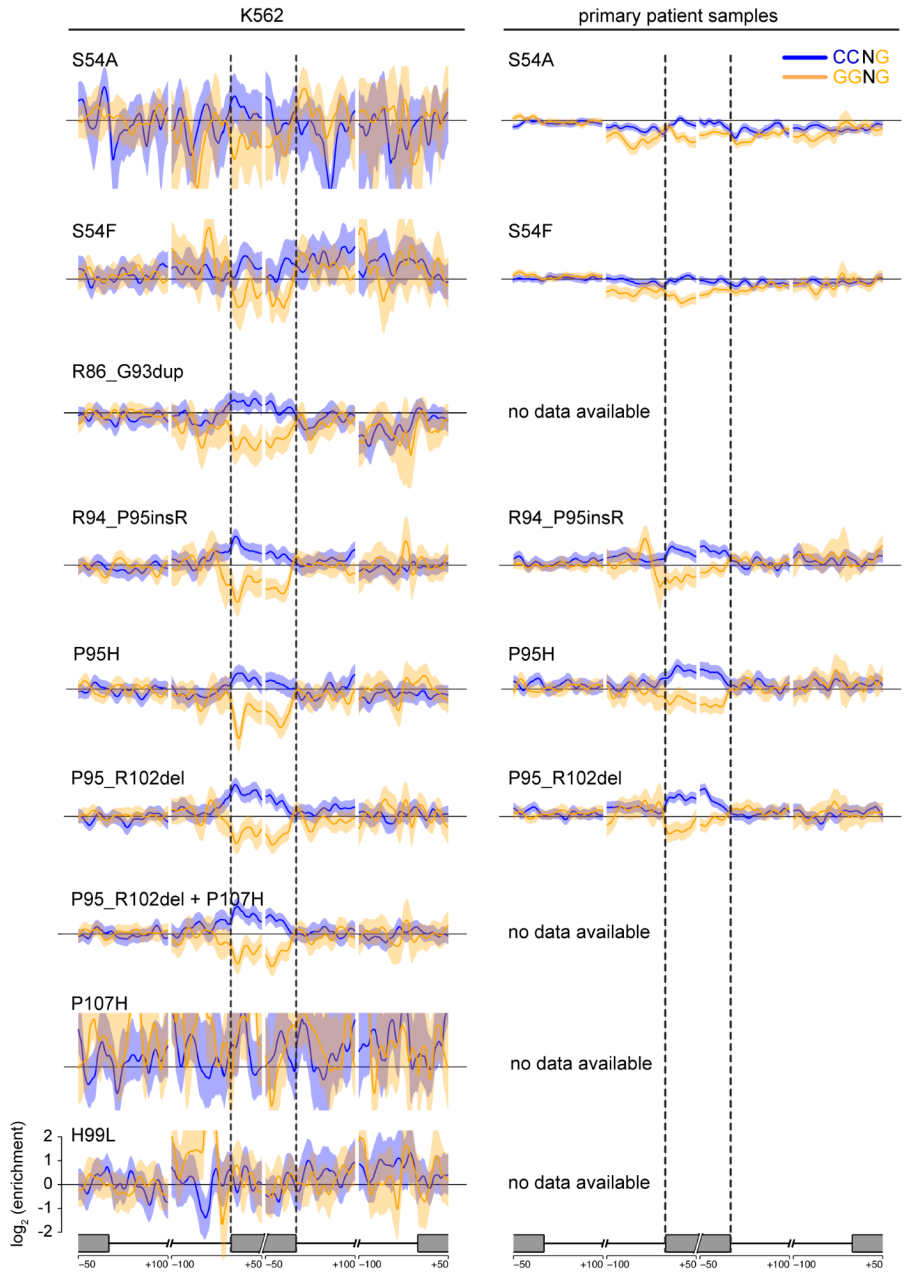

Figure S4 – related to Figure 2. Rare mutations in *SRSF2* alter exonic splicing enhancer (ESE) preference.

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.

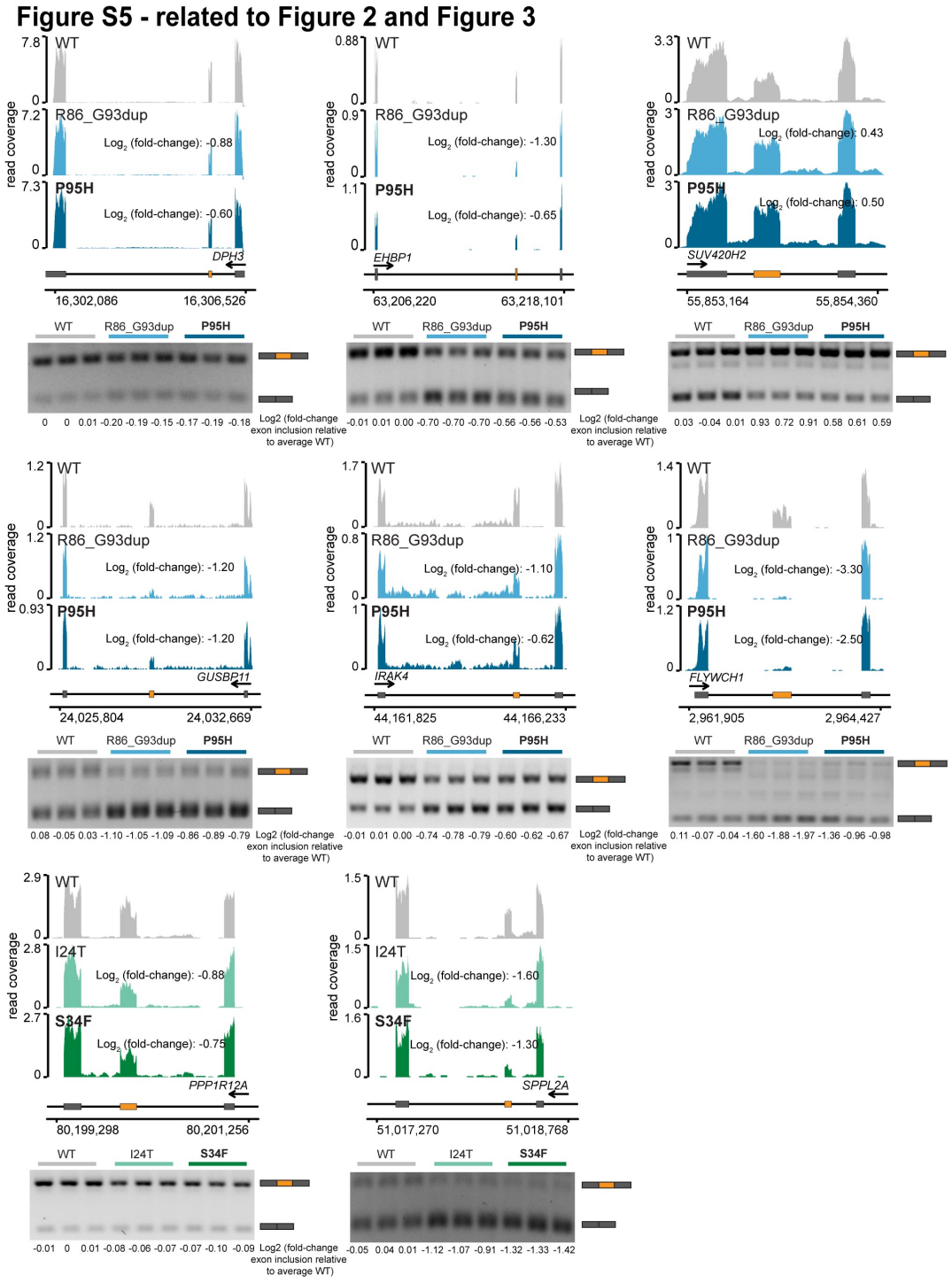

Figure S5 – related to Figure 2 and Figure 3. Rare SRSF2 and U2AF1 mutations phenocopy hotspot mutations.

Top, RNA-seq read coverage (top) illustrating increased cassette exon inclusion in the illustrated genes in K562 cells expressing either a hotspot (SRSF2P95H or U2AF1S34F) or rare (SRSF2R86_G93vdup or U2AF1I24T) SRSF2 or 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.

Figure S6

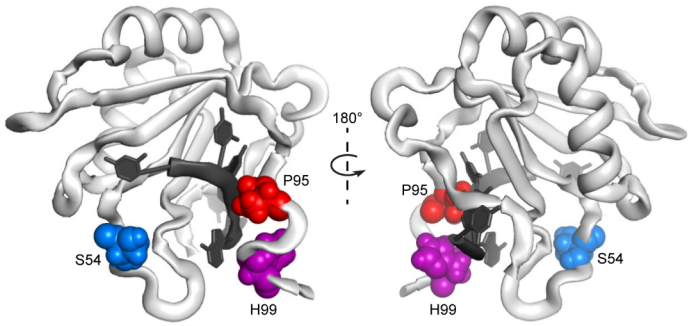

Figure S6. SRSF2 in complex with RNA.
Figure illustrates the solution NMR structure of the SRSF2 RNA recognition motif (PDB ID: 2LEB; Daubner et al, EMBO Journal, 2012) in complex with a 5'-UCCAGU-3' RNA substrate (black). Mutations affecting the highlighted residues are studied in this manuscript.

2.8 Materials and 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[23]. Unless otherwise specified, a splicing event was classified as differentially spliced if

it exhibited a change in isoform ratio of ≥10% and a Bayes Factor ≥5. Wagenmakers's framework[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 **Materials and 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[33] (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[20], GSE114922[34], and GSE66917 and GSE67039[35]. TCGA data was downloaded from CGHub[36,37]. Other data that support this study's findings are available from the authors upon reasonable request.

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 v2.01. Splice junctions that were not alternatively spliced in any isoforms from the UCSC knownGene track2 were defined as constitutive junctions. Gene annotations were defined by merging the UCSC knownGene track with the Ensembl 71 gene annotation3. We additionally created an annotation file holding all possible splice junctions obtained by splicing of annotated splice sites as described previously4. Reads were mapped to the GRCh37/hg19 human genome assembly using Bowtie v1.0.05, RSEM v1.2.46, and TopHat v2.1.17 as previously described4.

Isoform expression and differential splicing. Isoform expression levels were estimated as previously described4. Unless otherwise specified, a splicing event was classified as differentially spliced if it exhibited a change in isoform ratio of >10% and a Bayes Factor >5 with at least 20 informative reads for that event in each sample. Wagenmakers's framework8 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 >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 Bioconductor9.

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 mutations10. 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/V5HisA 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 genespecific 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.

2.9 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/.

2.10 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.

2.11 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.

Chapter 3. Using MDS patient-derived induced pluripotent stem cells to determine the functional consequences of mis-splicing due to SF3B1 mutations.

3.1 Introduction

Mutations in RNA splicing factors are the most common class of mutation in Myelodysplastic Syndromes (MDS)[1-4] and are frequently observed in other hematological malignancies as well as solid tumors[5]. Mutations in genes encoding splicing factors *SF3B1*, *SRSF2*, or *U2AF1* are observed in ~50-60% of MDS patients[6-7]. These mutations occur in a heterozygous manner and are almost always mutually exclusive[5].

Mutations in each individual splicing factor are often associated with different subtypes of disease. For example, ~80% of patients with MDS-RARS (refractory anemia with ring sideroblasts) carry a mutation in *SF3B1*[6]. This subtype is characterized by development of ring sideroblasts and poor erythropoiesis. The high frequency and early occurrence of SF3B1 mutations suggests that mutant SF3B1 contributes to the phenotypes of MDS-RARS[1,5,6].

There are challenges in modeling MDS-RARS. Current SF3B1 mutant mouse models develop anemia, but they do not develop ring sideroblasts[8]. The lack of tools that faithfully model MDS-RARS phenotypes has made it difficult to determine clear links between SF3B1 mutant-induced splicing, specific target genes, and MDS-RARS phenotypes. Motivated by this challenge, a recent study developed a SF3B1 mutant (G742D) MDS patient-derived induced pluripotent stem cell (iPSC) line that develops ring sideroblasts and demonstrates impaired erythropoiesis[9].

As a regulated splicing program is essential for successful erythropoiesis[10,11], we hypothesized that mis-splicing due to *SF3B1* mutations leads to poor erythropoiesis. This study aims to directly link SF3B1 mutant-induced mis-splicing to impaired erythropoiesis by studying splicing dysregulation in an iPS cell line model that recapitulates MDS-RARS phenotypes[9]. Because this cell line also develops ring sideroblasts, we hypothesized that SF3B1-mutant induced splicing is also responsible for this phenotype. To test this theory, we performed RNA-seq in an iPS cell line model at successive stages of erythropoiesis. Our goal was to identify splicing changes associated with mutant SF3B1 that are responsible for the development of ring sideroblasts and impaired erythropoiesis in MDS-RARS patients.

3.2 Results

Induction of erythroid differentiation alters RNA splicing programs.

In order to test the effects of *SF3B1* mutations on erythroid differentiation and development of ring sideroblasts, we utilized iPSC lines that demonstrated both phenotypes. These iPSC lines were created by reprogramming cells obtained from a patient with MDS with refractory anemia (MDS-RA1)[9]. These iPSC lines contained four notable genetic alterations, with the following inferred order of acquisition: Reciprocal translocation between chromosomes 4 and 12 [t(4;12)], *SF3B1*G742D, *EZH2*R685H, del(5q). We also utilized an isogenic iPSC line containing none of these genetic alterations (referred to as WT).

SF3B1 mutations in MDS are associated with impaired erythropoiesis[12]. We therefore induced erythroid differentiation in CD34+ iPSCs to test the effects of *SF3B1* mutations on erythroid differentiation. Cells were collected by FACS sorting at the progenitor ($CD34^{+}$), pre-

erythroblast ($CD34^{-}CD71^{+}$), and early erythroblast ($CD71^{+}CD235a^{+}$, from here on out will be referred to as erythroblast) stages. Because the presence of *SF3B1*G742D impaired further differentiation into mature erythroblasts ($CD71^{lo}CD235a^{+}$)[9], we stopped collection at the early erythroblast stage.

We sought to determine how differentiating iPSC progenitors into erythroblasts influenced global splicing patterns. We performed RNA-seq on cells at the stages of differentiation described above and quantified global splicing programs across differentiation stages **(Fig. 1A)**. Unsupervised clustering analysis demonstrated a differentiation stage-specific splicing program, regardless of *SF3B1* mutation status **(Fig. 1B)**. This suggests that the inability of mutant SF3B1 iPSC progenitors to differentiate into mature erythroblasts is not due to mutant-specific disruption of the global splicing program.

Erythroid differentiation of iPS progenitors induces RNA splicing alterations in hundreds of splicing events

We next sought to characterize the degree of differential splicing in iPS progenitors undergoing erythropoeisis. We performed RNA-seq on samples collected at each stage of erythropoiesis and quantified isoform expression for ~125,000 alternative splicing events as previously described[13]. We performed pairwise comparisons of samples across each successive stage of differentiation (progenitor-vs-preerythroblast, preerythroblast-vs-erythroblast), and quantified the number of differentially spliced events for each comparison **(Fig. 2A,B)**. An event was considered differentially spliced if it exhibited a change in isoform ratio of $\geq$ 10% and a bayes factor ≥ 5 . We observed differential splicing in hundreds of events, which is consistent with other studies that measured and identified changes in splicing isoform ratios during erythropoiesis[10,11].

SF3B1 mutations do not globally alter erythroid-specific splicing programs

As global splicing profiles clustered by differentiation stage, and not *SF3B1* mutation status, we next sought to determine the influence of *SF3B1* mutations on erythroid-specific splicing profiles. We isolated the set of differentially spliced events in WT iPS cells undergoing erythropoiesis. (**Fig. 2A)**. Unsupervised clustering analysis demonstrated that splicing profiles were segregated by differentiation state **(Fig. 3A)**. This suggests that *SF3B1* mutations do not dramatically alter the erythroid splicing program induced in WT iPSC cells, and that impaired downstream differentiation is more likely due to mis-splicing of a small set of events.

Disease relevant mis-spliced events induced by *SF3B1* mutations persist throughout erythroid differentiation of patient derived iPS progenitor cells

We hypothesized that specific SF3B1 mutant-induced mis-spliced events resulted in impaired differentiation of iPS progenitors into mature erythrocytes. As erythropoiesis is a dynamic process, we aimed to provide strong evidence of consistent mis-splicing in a dynamic state as well. We identified 73 differentially spliced events in SF3B1 WT-vs-G742D cells that were common throughout erythropoiesis **(Fig. 4A)**. Events in genes *MAP3K7, BRD9,* and *TMEM14C* were particularly notable due to their known involvement in important cellular processes. A fourth gene, *ABCB7*, was not included in this list of differentially spliced events, but was explored due to previous studies that implicated this gene in Refractory Anemia with Ring Sideroblasts (RARS)[14,15].

Recent work done demonstrated that *SF3B1*K700E induced increased usage of an intron-proximal 3′ splice site in MAP3K7, resulting in hyperactive NF-kB signaling[16]. Our analysis

revealed this identical change for *SF3B1*G742D **(Fig. 4B)**. In addition, we recently showed mis-expression of an exon in *BRD9* due to *SF3B1* mutations, which resulted in decreased *BRD9* protein levels and subsequent cellular transformation[17]. The usage of this exon was phenocopied for *SF3B1*G742D **(Fig. 4C)**. We also observed mutant-induced increased usage of an intron-proximal 3′ splice site in *TMEM14C* **(Fig. 4D)**. Interestingly, this splicing event occurs in the 5′ UTR, and preliminary work done by a collaborating lab indicates that this mutant-induced splicing change impairs translation of this gene. As *TMEM14C* is essential for erythropoiesis[18], this suggests a possible functional role for this mis-spliced event in promoting impaired erythropoiesis. Finally, we see SF3B1 mutant-induced usage of a competing 3′ splice site in *ABCB7* **(Fig. 4E)** with a corresponding decrease in expression of *ABCB7* RNA transcripts **(Fig.4F)**, consistent with other studies[19,20]. Data on corresponding protein levels of ABCB7 in our iPSC lines is currently being investigated. Despite previous studies that link reduced ABCB7 protein levels to Refractory Anemia with Ring Sideroblasts (RARS)[19,21], the direct functional consequence of this mis-spliced event in relation to impaired erythropoiesis and ring sideroblast formation is unclear.

3.3 Discussion

SF3B1 mutations are highly associated with MDS-RARS, but the molecular basis for which these mutations promote disease is not entirely clear. The data presented here support a model in which SF3B1 mutant-induced MDS-RARS hallmark phenotypes of ring siderobasts and impaired erythropoiesis are likely not due to global splicing disruption, but rather mis-splicing of individual genes.

We identified four genes with consistent mis-splicing profiles in *SF3B1*G742D iPSC lines undergoing erythropoiesis, two of which have a prominent role in erythroid differentiation. This includes SF3B1 mutant-associated usage of a competing 3′ splice site in *TMEM14C* that may lead to impaired translation (experiments still in progress). As downregulation of TMEM14C was shown to impair erythropoiesis[18], this finding is likely relevant to MDS-RARS. We also observed usage of a competing 3′ splice site in *ABCB7* promoted by mutant SF3B1. This results in a transcript that is predicted to be degraded by nonsense-mediated decay. Previous studies showed that downregulation of ABCB7 impaired erythroid differentiation[14,15], which suggests a functional role for this mis-spliced event in promoting MDS-RARS.

We cannot rule out the possibility that the identified mis-spliced events are due to other genetic alterations (e.g., t(4;12)] , *EZH2*R685H, del(5q)) in our iPSC lines. We mitigated this potential issue by selecting mis-spliced events common to other *SF3B1* mutant datasets[15,17,22]. However, future experiments will also contain gene-corrected *SF3B1* iPSC clones in order to demonstrate SF3B1-mediated effects.

Our study has implications for clinical research of MDS patients carrying splicing factor mutations. The only curative treatment of MDS is hematopoietic stem cell transplantation, but fewer than 5% of patients are eligible for this treatment[23]. DNA hypomethylating agents (HMAs) are another approved treatment for MDS. However, only 10-15% of patients treated with HMAs experience complete response[24], and clinical studies have been unable to demonstrate consistent predictions of response to HMAs with respect to individual disease-associated mutations[25]. As a result, more targeted treatments for MDS are in development. For example, treatment with splicing inhibitor molecules selectively target cells expressing mutant splicing factors[26,27]. One such drug, H3B-8800, is currently undergoing clinical trial to determine the efficacy of splicing

inhibition in patients carrying a mutation in a RNA splicing factor[27]. Despite the promising results of splicing inhibitor treatment, these drugs inhibit global splicing, and we cannot rule out the potential for unwanted side effects. Our study addresses this concern for side effects by identifying specific mis-spliced events with disease relevance. Future confirmation that these specific splicing events alter disease phenotypes will move us one step closer to designing more precise therapies for MDS patients carrying splicing factor mutations.

REFERENCES

1. Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*. 2011;478(7367):64-69.
2. Graubert TA, Shen D, Ding L, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat Genet*. 2011;44(1):53-57.
3. Foy A, McMullin MF. Somatic *SF3B1* mutations in myelodysplastic syndrome with ring sideroblasts and chronic lymphocytic leukaemia. *J Clin Pathol*. 2019;72(11):778-782.
4. Visconte V, Makishima H, Jankowska A, et al. SF3B1, a splicing factor is frequently mutated in refractory anemia with ring sideroblasts. *Leukemia*. 2012;26(3):542-545.
5. Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. RNA splicing factors as oncoproteins and tumour suppressors. *Nat Rev Cancer*. 2016;16(7):413-430.
6. Kennedy JA, Ebert BL. Clinical Implications of Genetic Mutations in Myelodysplastic Syndrome. *J Clin Oncol*. 2017;35(9):968-974.
7. Haferlach T, Nagata Y, Grossmann V, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia*. 2014;28(2):241-247.
8. Obeng EA, Chappell RJ, Seiler M, et al. Physiologic Expression of Sf3b1(K700E) Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation. *Cancer Cell*. 2016;30(3):404-417.
9. Hsu J, Reilly A, Hayes BJ, et al. Reprogramming identifies functionally distinct stages of clonal evolution in myelodysplastic syndromes. *Blood*. 2019;134(2):186-198.
10. Pimentel H, Parra M, Gee S, et al. A dynamic alternative splicing program regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res*. 2014;42(6):4031-4042.
11. Pimentel H, Parra M, Gee SL, Mohandas N, Pachter L, Conboy JG. A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res*. 2016;44(2):838-851.
12. Malcovati L, Cazzola M. Recent advances in the understanding of myelodysplastic syndromes with ring sideroblasts. *Br J Haematol*. 2016;174(6):847-858.

13. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med*. 2015;7(1):45.
14. Boultwood J, Pellagatti A, Nikpour M, et al. The role of the iron transporter ABCB7 in refractory anemia with ring sideroblasts. *PLoS One*. 2008;3(4):e1970.
15. Nikpour M, Scharenberg C, Liu A, et al. The transporter ABCB7 is a mediator of the phenotype of acquired refractory anemia with ring sideroblasts. *Leukemia*. 2013;27(4):889-896.
16. Lee SC, North K, Kim E, et al. Synthetic Lethal and Convergent Biological Effects of Cancer-Associated Spliceosomal Gene Mutations. *Cancer Cell*. 2018;34(2):225-241.e8.
17. Inoue D, Chew GL, Liu B, et al. Spliceosomal disruption of the non-canonical BAF complex in cancer. *Nature*. 2019;574(7778):432-436.
18. Yien YY, Robledo RF, Schultz IJ, et al. TMEM14C is required for erythroid mitochondrial heme metabolism. *J Clin Invest*. 2014;124(10):4294-4304.
19. Darman RB, Seiler M, Agrawal AA, et al. Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. *Cell Rep*. 2015;13(5):1033-1045.
20. Dolatshad H, Pellagatti A, Liberante FG, et al. Cryptic splicing events in the iron transporter ABCB7 and other key target genes in SF3B1-mutant myelodysplastic syndromes. *Leukemia*. 2016;30(12):2322-2331.
21. Nikpour M, Scharenberg C, Liu A, et al. The transporter ABCB7 is a mediator of the phenotype of acquired refractory anemia with ring sideroblasts. *Leukemia*. 2013;27(4):889-896.
22. Bergot T, Lippert E, Douet-Guilbert N, Commet S, Corcos L, Bernard DG. Human Cancer-Associated Mutations of SF3B1 Lead to a Splicing Modification of Its Own RNA. *Cancers (Basel)*. 2020;12(3):652.
23. Khan C, Pathe N, Fazal S, Lister J, Rossetti JM. Azacitidine in the management of patients with myelodysplastic syndromes. *Ther Adv Hematol*. 2012;3(6):355-373.
24. Bejar R, Lord A, Stevenson K, et al. TET2 mutations predict response to hypomethylating agents in myelodysplastic syndrome patients. *Blood*. 2014;124(17):2705-2712.
25. Sanz GF, Ibañez M, Such E. Do next-generation sequencing results drive diagnostic and therapeutic decisions in MDS? Blood Adv. 2019 Nov;3(21) 3454-3460.
26. Lee SC, Dvinge H, Kim E, et al. Modulation of splicing catalysis for therapeutic targeting of leukemia with mutations in genes encoding spliceosomal proteins [published correction appears in Nat Med. 2016 Jun 7;22(6):692]. *Nat Med*. 2016;22(6):672-678.
27. Seiler M, Yoshimi A, Darman R, et al. H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. *Nat Med*. 2018;24(4):497-504.

3.4 Figures and legends

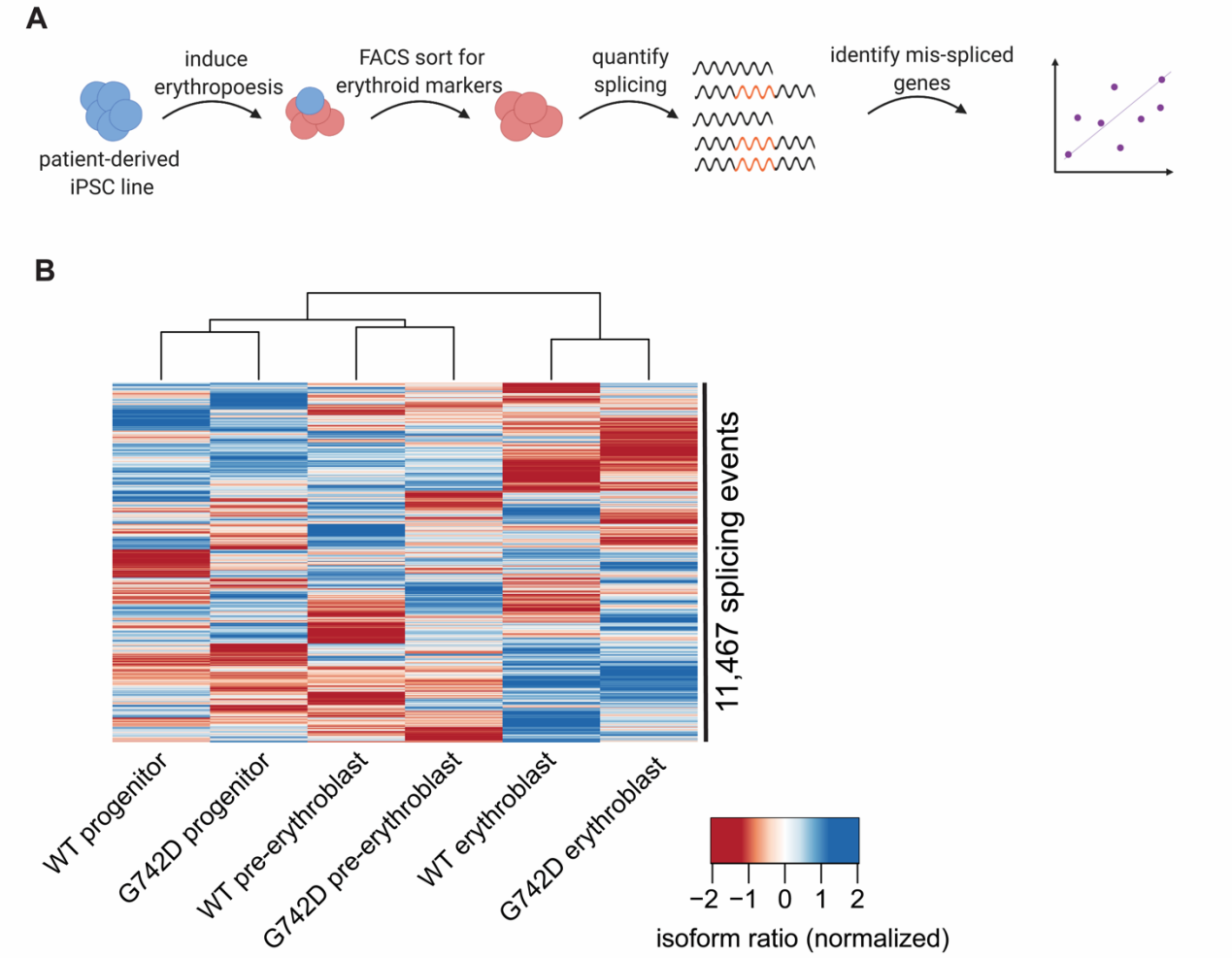

Figure 1. Strategy for studying mis-splicing during erythroid differentiation in a SF3B1 mutant iPSC model.

(A) Schematic of the strategy to study our SF3B1 mutant iPSC model during erythroid differentiation

(B) Heat map and associated dendrogram representing an unsupervised cluster analysis based on splice isoform ratios computed from the transcriptomes of iPS cells induced to undergo erythroid differentiation. Isoform ratios were z-score normalized.

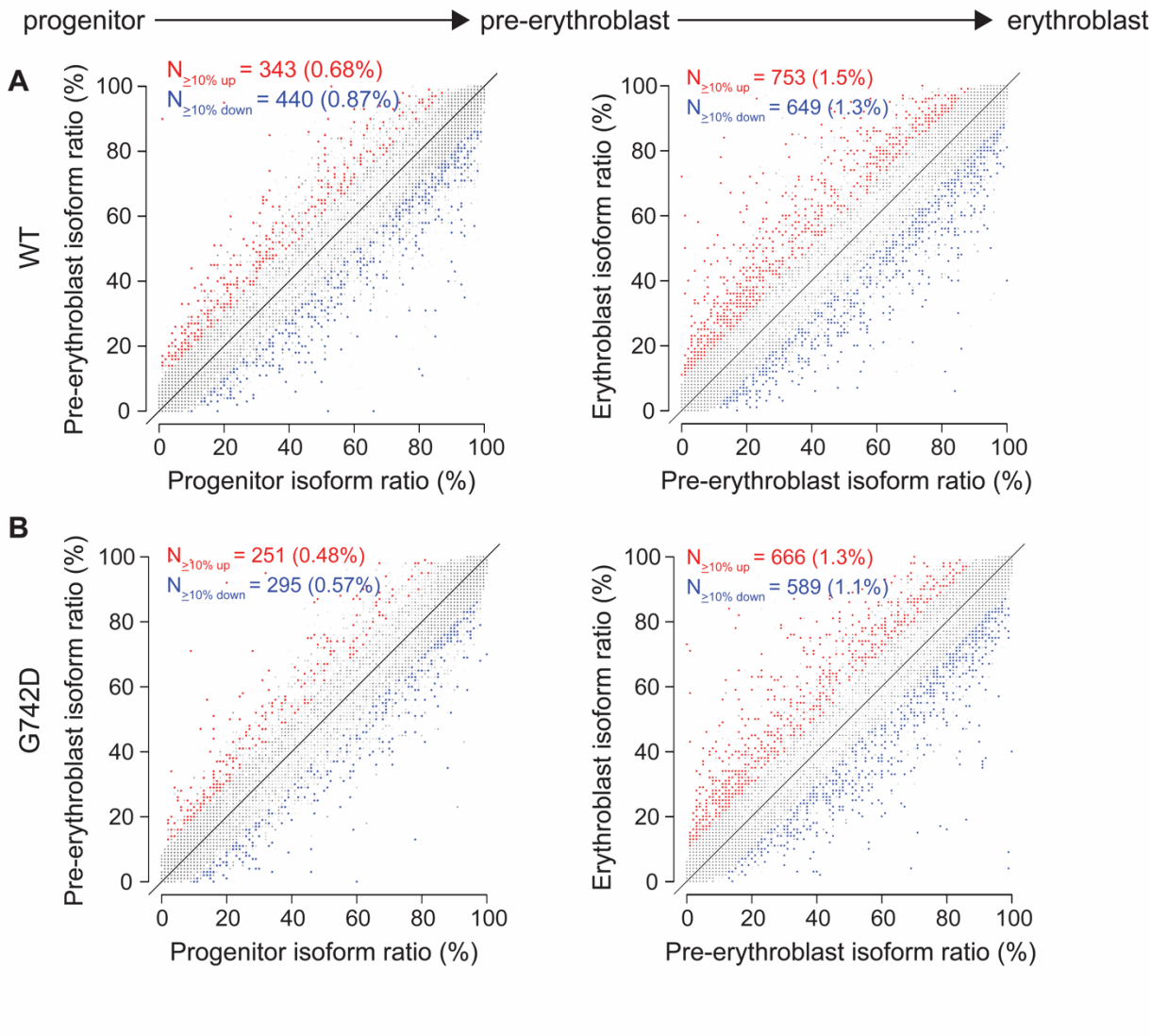

Figure 2. Induction of erythroid differentiation in iPS cells induces changes in hundreds of splicing events.

(A) Scatter plots illustrating splice isoform ratios in an iPSC model induced to undergo erythroid differentiation. Red, splicing events exhibiting significantly increased isoform ratio in the more advanced differentiation stage; blue, splicing events exhibiting significantly decreased isoform ratio in the more advanced differentiation stage. N, numbers of events exhibiting differential splicing. Percentages are computed with respect to the total numbers of splicing events exhibiting any evidence of alternative splicing in the analyzed cells. See Methods for descriptions of identification of significant differential splicing.

(B) As **(A)**, but for *SF3B1*G742D iPS cells.

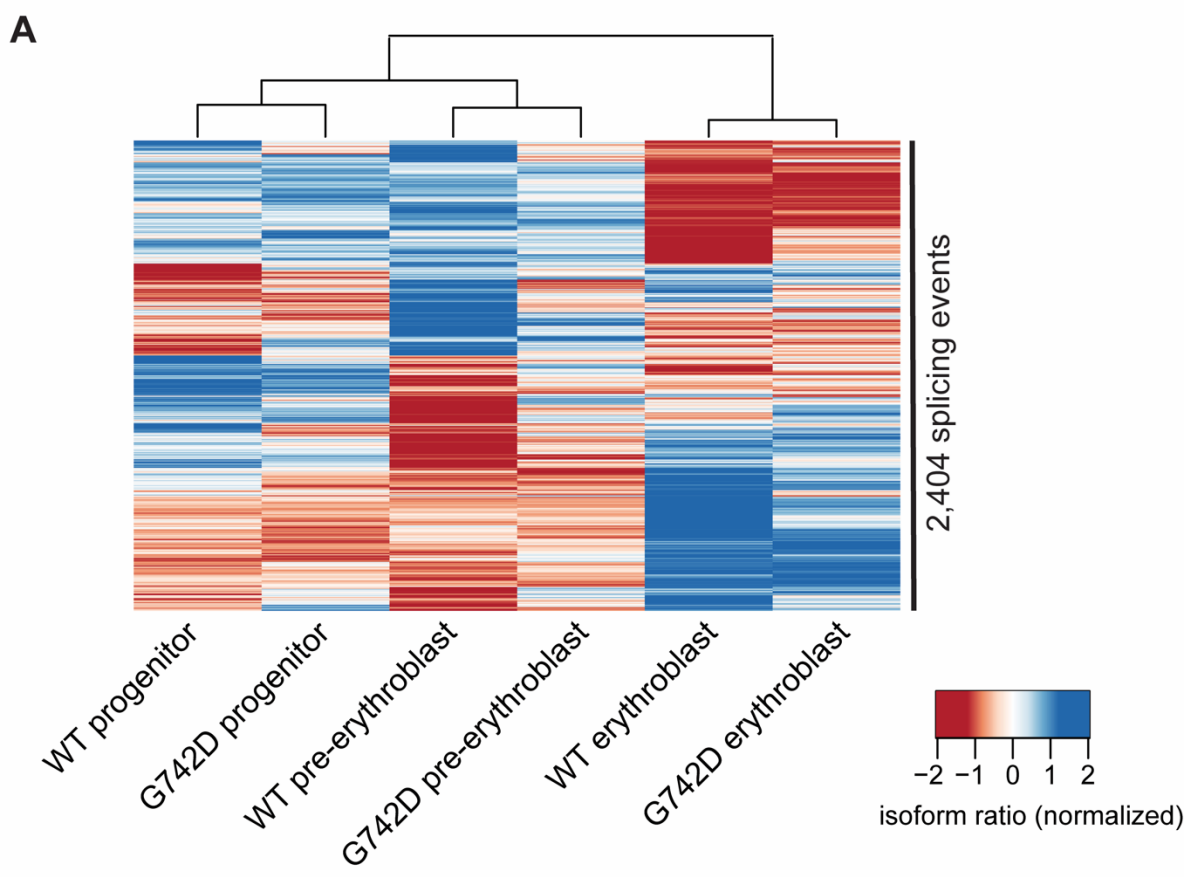

Figure 3. Erythroid differentiation splicing programs cluster by differentiation state, not SF3B1 mutation status.

(A) Heat map and associated dendrogram representing an unsupervised cluster analysis based on splice isoform ratios computed from the transcriptomes of iPS cells induced to undergo erythroid differentiation. Splicing events were restricted to those which were differentially spliced in the WT iPSC model. Isoform ratios were z-score normalized.

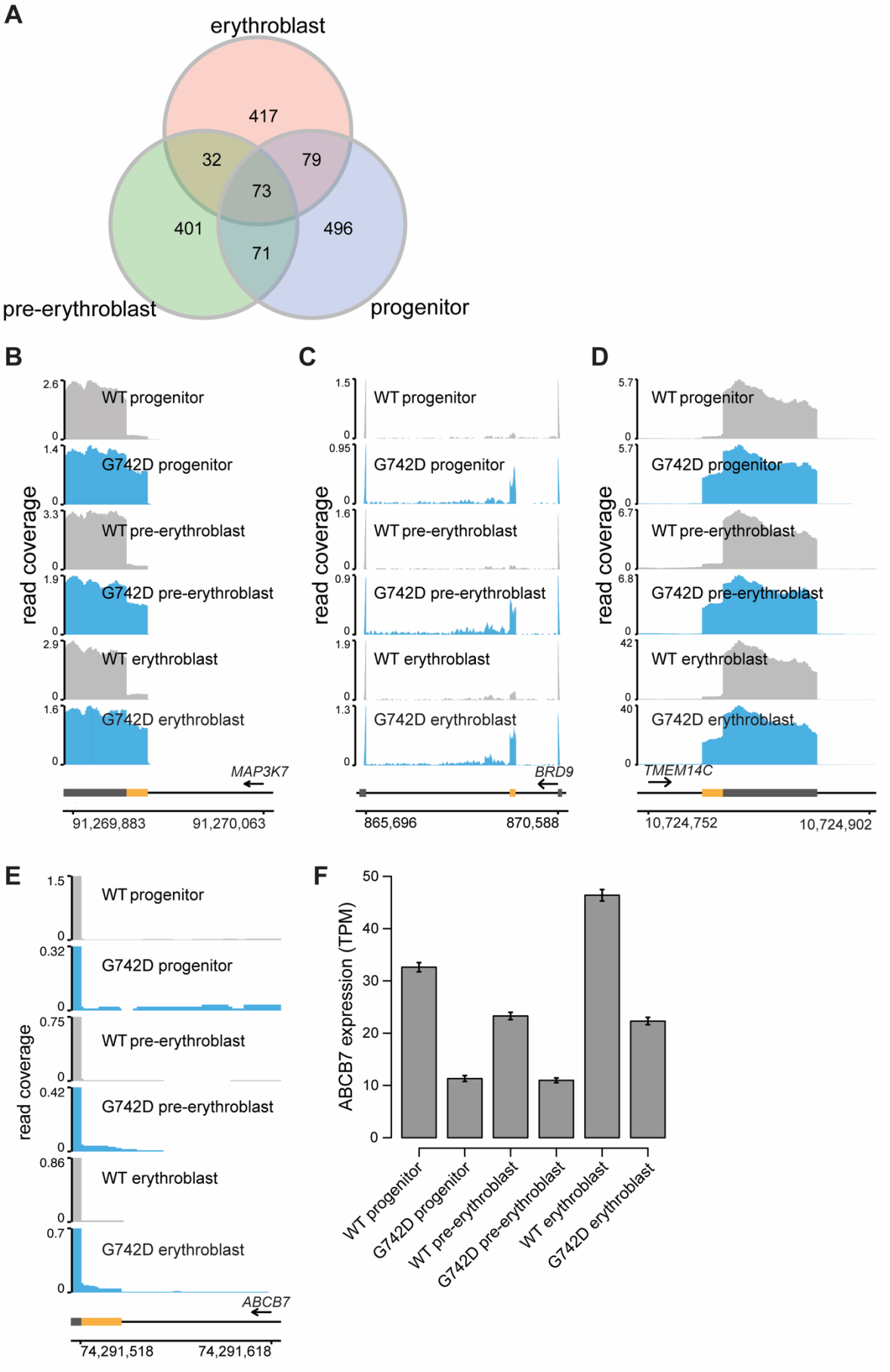
A
erythroblast
417
32
79
73
401
496
71
pre-erythroblast
progenitor
B
C
D
read coverage
WT progenitor
G742D progenitor
WT pre-erythroblast
G742D pre-erythroblast
WT erythroblast
G742D erythroblast
MAP3K7
91,269,883
91,270,063
BRD9
865,696
870,588
TMEM14C
10,724,752
10,724,902
E
F
ABCB7
74,291,518
74,291,618
ABCB7 expression (TPM)
WT progenitor
G742D progenitor
WT pre-erythroblast
G742D pre-erythroblast
WT erythroblast
G742D erythroblast

Figure 4. Figure 4. SF3B1 mutant-induced mis-splicing of genes occurs during erythropoiesis. **(A**) Venn diagram illustrating the sets of splicing events that were differentially spliced at each stage of differentiation in G742D cells relative to WT cells. Differentially spliced events were defined as those exhibiting a change in isoform ratio $\geq$ 10% and a Bayes factor $\geq$ 5.
(B) RNA-seq read coverage illustrating differential splicing at a competing 3' splice site in *MAP3K7*.
(C) RNA-seq read coverage illustrating increased inclusion of a poison exon in *BRD9*.
(D) RNA-seq read coverage illustrating differential splicing at a competing 3' splice site in *TMEM14C*.
(E) RNA-seq read coverage illustrating differential splicing at a competing 3′ splice site in *ABCB7.*
(F) Bar plot illustrating mRNA expression of *ABCB7* in the indicated samples. TPM, transcripts per million (TMM-normalized). Error bars represent 95% confidence intervals.

3.5 Materials and methods

RNA-seq library preparation and analysis. Total RNA was isolated from flow-sorted, MDS patient-derived iPS cells induced to undergo erythroid differentiation. 500 ng 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.

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]. Gene annotations were defined by merging the UCSC knownGene track with the Ensembl 71 gene annotation[2,3]. 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]. A splicing event was classified as differentially spliced if it exhibited a change in isoform ratio of >10% and a Bayes Factor >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 splicing events which had >20 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.

Erythroid differentiation of iPSCs. Erythroid differentiation was induced as previously described[11,12].

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.
11. Lee HY, Gao X, Barrasa MI, et al. PPAR-a and glucocorticoid receptor synergize to promote erythroid progenitor self-renewal. Nature. 2015;522(7557):474-477.
12. Hsu J, Reilly A, Hayes BJ, et al. Reprogramming identifies functionally distinct stages of clonal evolution in myelodysplastic syndromes. Blood. 2019;134(2):186-198.

Chapter 4. Discussion

The goal of my thesis work was to better understand the contribution of splicing factor mutations to disease. We demonstrated that 11 out of 14 studied rare and private mutations in *SRSF2* and *U2AF1* phenocopy splicing alterations seen in their hotspot counterparts, indicating a likely disease relevance for those mutations. In addition, we also identified mis-splicing of *ABCB7* and *TMEM14C* in a mutant SF3B1 iPSC line that may explain the impaired erythropoiesis phenotype frequently associated with MDS patients carrying *SF3B1* mutations.

Multiple lines of evidence support clinical implications of the findings in this dissertation. First, therapies that disrupt the spliceosome are shown to specifically target cells carrying hotspot spliceosomal mutations[1-4]. Because many rare and private *SRSF2* and *U2AF1* mutations phenocopy splicing alterations in hotspot mutations, patients carrying non-hotspot spliceosomal mutations should be considered when these treatments enter clinical practice. Additionally, studies have shown that modulation of mis-spliced events induced by mutant splicing factors can reverse associated disease phenotypes. For example, *U2AF1*S34F-induced mis-splicing of *H2AFY* and *IRAK4* was modulated to reverse the erythroid differentiation defect observed in *U2AF1*S34F cells[5]; reversal of reduced EZH2 expression caused by *SRSF2* mutations rescued the impaired hematopoiesis phenotype observed in *SRSF2* mutant cells[6]; correcting *BRD9* mis-splicing with antisense oligonucleotides suppressed tumor growth in SF3B1 mutant cells[7]. This supports the idea of treating patients bearing splicing factor mutations with molecules that target specific splicing events, which if successful could significantly reduce side effects. Further experiments are needed to confirm if mis-splicing of *ABCB7* and *TMEM14C* in the SF3B1 mutant iPSC lines presented in this dissertation directly lead to the associated

disease phenotype of impaired erythropoiesis. If true, *ABCB7* and/or *TMEM14C* may be potential targets for a therapy in which a specific splicing event is modulated.

Although the results described in this dissertation demonstrate disease relevance of rare, private, and hotspot splicing factor mutations, they do not consider the effects of mutations that co-occur with splicing factor mutations. Consideration of co-occurring mutations is important for several reasons. First, splicing factor mutations frequently co-occur with other oncogenic drivers[8]. In fact, one study describes an AML patient cohort in which 47% of patients with mutant SRSF2 also had a mutation in *IDH2*[9]. Second, mouse models engineered to express a single mutant splicing factor induce splicing alterations observed in corresponding mutant cell lines and primary patient materials but are unable to recapitulate all hallmark disease phenotypes. *SF3B1*$^{K700E/+}$ mice show impaired erythropoiesis, but do not develop ring sideroblasts[3]; *U2AF1*^{S34F} mice develop inconsistent defects in blood cell lineages and *U2AF1*S34F stem cells from these mice do not demonstrate a competitive repopulation advantage[10]; *SRSF2*P95H$^{/+}$ mice develop multilineage dysplasia, but *SRSF2*P95H stem cells from these mice do not have a competitive repopulation advantage[6]. In contrast to mice expressing mutant SRSF2 alone, mice engineered to co-express mutant IDH2 and SRSF2 demonstrated altered splicing defects, which consequently contributed to an enhanced proliferative capacity in vivo[9]. This is not surprising based on studies demonstrating that distinct splicing factor mutations induce distinct splicing alterations[1,3,6,7,10,11], so one can reasonably assume that diseases containing two or more driver mutations would also result in distinct splicing alterations. Several iPSC models that carry mutations co-occurring with splicing factor mutations, including the one described in Chapter 3, have demonstrated cooperation between multiple genetic alterations to induce phenotypic changes observed in patients[13,14]. As a result, future models of the rare mutations in *SRSF2* and

U2AF1 that were described in Chapter 2 should consider that co-expression of another driver mutation will be able to more accurately model the molecular and phenotypic changes observed in patients. In doing so, current and future treatments of diseases associated with splicing factor mutations can be tested for efficacy in a way that is more accurately linked to the genetic makeup of the disease.

Overall, I believe that the key findings described in this dissertation will motivate future studies on rare, private, and hotspot spliceosomal mutations with the hope of developing safe, effective, and more targeted therapies in an advancing era of personalized medicine.

REFERENCES

1. Lee SC-W, Dvinge H, Kim E, et al. Modulation of splicing catalysis for therapeutic targeting of leukemia with mutations in genes encoding spliceosomal proteins. *Nat. Med.* 2016.
2. Shirai CL, White BS, Tripathi M, et al. Mutant U2AF1-expressing cells are sensitive to pharmacological modulation of the spliceosome. Nat Commun. 2017;8:14060.
3. Obeng EA, Chappell RJ, Seiler M, et al. Physiologic Expression of Sf3b1(K700E) Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation. Cancer Cell. 2016;30(3):404–417.
4. Seiler M, Yoshimi A, Darman R, et al. H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. Nat. Med. 2018.
5. Yip BH, Steeples V, Repapi E, et al. The U2AF1S34F mutation induces lineage-specific splicing alterations in myelodysplastic syndromes [published correction appears in J Clin Invest. 2017 Sep 1;127(9):3557]. J Clin Invest. 2017;127(6):2206-2221.
6. 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.
7. Inoue D, Chew GL, Liu B, et al. Spliceosomal disruption of the non-canonical BAF complex in cancer. Nature. 2019;574(7778):432-436
8. Cazzola M, Della Porta MG, Malcovati L. The genetic basis of myelodysplasia and its clinical relevance. Blood. 2013;122(25):4021-4034.
9. Yoshimi A, Lin KT, Wiseman DH, et al. Coordinated alterations in RNA splicing and epigenetic regulation drive leukaemogenesis. Nature. 2019;574(7777):273-277.

10. Shirai CL, Ley JN, White BS, et al. Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo. Cancer Cell. 2015;27(5):631-643.
11. Ilagan JO, Ramakrishnan A, Hayes B, et al. U2AF1 mutations alter splice site recognition in hematological malignancies. Genome Res. 2015;25(1):14-26.
12. Bejar R, Lord A, Stevenson K, et al. TET2 mutations predict response to hypomethylating agents in myelodysplastic syndrome patients. Blood. 2014;124(17):2705-2712.
13. Chang CJ, Kotini AG, Olszewska M, et al. Dissecting the Contributions of Cooperating Gene Mutations to Cancer Phenotypes and Drug Responses with Patient-Derived iPSCs. Stem Cell Reports. 2018;10(5):1610-1624.
14. Hsu J, Reilly A, Hayes BJ, et al. Reprogramming identifies functionally distinct stages of clonal evolution in myelodysplastic syndromes. Blood. 2019;134(2):186-198.