Published in final edited form as:
Leukemia. 2021 April ; 35(4): 1108–1120. doi:10.1038/s41375-020-1002-y.

Complex Landscape of Alternative Splicing in Myeloid Neoplasms

Courtney E. Hershberger[1], **Devlin C. Moyer**[1], **Vera Adema**[2], **Cassandra M. Kerr**[2], **Wencke Walter**[3], **Stephan Hutter**[3], **Manja Meggendorfer**[3], **Constance Baer**[3], **Wolfgang Kern**[3], **Niroshan Nadarajah**[3], **Sven Twardziok**[3], **Mikkael A. Sekeres**[2], **Claudia Haferlach**[3], **Torsten Haferlach**[3,+], **Jaroslaw P. Maciejewski**[2,+], **Richard A. Padgett**[1,*,+]

[1]Cardiovascular and Metabolic Sciences Department, Cleveland Clinic Foundation, Cleveland, Ohio, United States

[2]Translational Hematology and Oncology Research, Cleveland Clinic Foundation, Cleveland, Ohio, United States

[3]Munich Leukemia Laboratory, Munich, Germany

Abstract

Myeloid neoplasms are characterized by frequent mutations in at least seven components of the spliceosome that have distinct roles in the process of pre-mRNA splicing. Hotspot mutations in *SF3B1*, *SRSF2*, *U2AF1* and loss of function mutations in *ZRSR2* have revealed widely different aberrant splicing signatures with little overlap. However, previous studies lacked the power necessary to identify commonly mis-spliced transcripts in heterogeneous patient cohorts. By performing RNA-Seq on bone marrow samples from 1,258 myeloid neoplasm patients and 63 healthy bone marrow donors, we identified transcripts frequently mis-spliced by mutated splicing factors (SF), rare SF mutations with common alternative splicing (AS) signatures, and SF-dependent neojunctions. We characterized 17,300 dysregulated AS events using a pipeline designed to predict the impact of mis-splicing on protein function. Meta-splicing analysis revealed a pattern of reduced levels of retained introns among disease samples that was exacerbated in patients with splicing factor mutations. These introns share characteristics with "detained introns," a class of introns that have been shown to promote differentiation by detaining pro-proliferative

[*]Corresponding author and lead contact: Rick A. Padgett, Lerner Research Institute, 9500 Euclid Avenue, Cleveland, Ohio 44195, Location: NE2-305, padgetr@ccf.org, Phone: (216) 445-2692.
[+]These authors contributed equally to this work
Author Contributions
C.E.H, R.A.P., and J.P.M., and T.H. conceptualized the project. C.E.H, R.A.P., and J.P.M., T.H., and M.A.S. prepared the manuscript, with contributions from all authors. C.M. Kerr assisted in sample preparation. C.E.H, D.C.M., V.A., C.H., N.N., S.T., S.H., M.M., W.W., C.B., W.K., and T.H. performed data curation. C.E.H performed formal analyses and D.C.M contributed to formal analyses for figures 4 and 6.

Competing Interests
T.H., C.H. and W.K. serve as the management board of the MLL Munich Leukemia Laboratory. N.N, S.T., S.H, M.M., W.W., and C.B. are employed by the MLL Munich Leukemia Laboratory. The MLL offers diagnostic services for leukaemias and lymphomas, including cytomorphology, cytochemistry, immunophenotyping, cytogenetics, FISH, and a broad spectrum of molecular assays. In addition, MLL runs several research studies based on a combination of methods for routine use and also including whole genome sequencing and whole transcriptome sequencing.

transcripts in the nucleus. In this study, we have functionally characterized 17,300 targets of mis-splicing by the SF mutations, identifying a common pathway by which AS may promote maintenance of a proliferative state.

Introduction

Myeloid neoplasms, including myelodysplastic syndromes (MDS), acute myeloid leukemia (AML), and myelodysplastic/myeloproliferative neoplasms (CMML, MDS/MPN-RS-T and MDS/MPN-U) are characterized by high somatic mutation rates in components of the spliceosome (1–3). Recurrent hotspot mutations have been observed in the spliceosomal factor genes *SF3B1*, *SRSF2*, *U2AF1*, *PRPF8* and *DDX41*(1–3). Truncating mutations or chromosomal deletions have also been described in *ZRSR2*, *LUC7L2*, *DDX41* and *PRPF8* (2–8). A role for dysregulated alternative splicing (AS) in malignancies had been described before these splicing factor (SF) gene mutations were discovered, but the high frequency of specific SF mutations is unique to myeloid neoplasms among cancers. These seven SFs have been identified in different spliceosomal complexes and interact differently with pre-mRNA (9). Concordantly, patients harboring these mutations experience different clinical phenotypes, responses to therapies and outcomes (4, 6, 10–15). Therefore, it is likely that common mis-splicing events contribute to pathogenesis, yet there are mutation-specific alterations that explain the morphologic and clinical diversity of cases affected by different mutations in and deletions of SFs.

Previous work using engineered cell lines characterized the mechanisms by which alternative splicing (AS) patterns result from missense mutations in *SF3B1*, *SRSF2*, *U2AF1* and *PRPF8*, as well as loss of function mutations in *ZRSR2* (1, 2, 4, 15–25). Genetically engineered murine models harboring mutations in *SF3B1*, *SRSF2* and *U2AF1* mimicked these general splicing patterns, but did not fully recapitulate the AS events observed in human cell lines, likely due to differences in AS between species (8, 12, 20, 23, 25–31). Both murine models and engineered human leukemic cell lines were instrumental in characterizing the molecular mechanisms of these mutated splicing factors but phenotypically, have not fully explained pivotal alterations explaining clonal expansion. Early AS analyses of patient samples confirmed that overall splicing patterns observed in model systems were present in patient bone marrow (2, 4, 6, 16–21, 23, 24, 26–29, 31–35). However, they were unable to provide a definitive list of splicing changes being limited by sample sizes (fewer than 10 patients harboring mutations) and disease heterogeneity. Studies that have included more patients have identified a more comprehensive list of splicing changes (8, 22, 36, 37), but focused on a single disease, missense mutations alone, and did not assess the functional consequences of the identified splicing changes.

Fully understanding alternative splicing patterns in malignancies is essential because AS often confers functional changes on the translated protein, potentially creating oncogenes or inactivating tumor suppressor genes. The aims of characterizing global mis-splicing in cancers are two-fold: to identify isoforms that confer pathogenicity and to identify biomarkers that could direct detection or treatment.

Materials and Methods

Sample collection, processing and sequencing

Bone marrow samples were collected between 09/2005–04/2017 and diagnosed (MDS, AML, MDS/MPN) using WHO 2016 guidelines. All patients gave written informed consent for scientific evaluations and the study was approved by the Internal Review Board and adhered to the tenets of the Declaration of Helsinki. 250ng of total RNA was prepared with the Illumina TruSeq Total Stranded RNA library preparation kit. Paired-end reads of 100bp were sequenced by the NovaSeq 6000 with median depth of 50 million reads per sample. Sequences were aligned using STAR(38) (STAR 2.5.0) to human reference genome hg19. DNA was prepared with Illumina TruSeq PCR free library prep kit and sequenced (paired-end) on NovaSeq 6000 or HiSeqX with median 100x coverage. DNA sequences were aligned to human reference genome hg19 using Isaac3 (Illumina).

Identifying somatic mutations and chromosomal deletions

Somatic mutations were called by Streka2 (Illumina) and filtered using an in-house pipeline (see supplementary methods). Copy Number Variation was derived from whole genome sequencing data using GATK4 (Broad Institute). Deletions of *LUC7L2*, *DDX41*, *PRPF8* and *ZRSR2* were confirmed by karyotype.

Gene Expression and Splice Junction Identification and Quantification

Gene counts were estimated with Cufflinks (v2.2.1) and normalized by applying Trimmed mean of M-values (TMM) normalization method(39). Novel splice junctions were identified by StringTie (40). Novel and annotated splice junctions were quantified by rMATS 4.0.1 (41) and filtered for coverage and variance. Upon grouping by SF mutation or SF expression, rMATS STAT was run to identify dysregulated AS events, p-values were adjusted for multiple hypothesis testing with Bonferroni correction (Qvalues ≤ 0.05 and inclusion level difference ≥ 5% were considered significant).

Hierarchical clustering

Unsupervised hierarchical clustering in figures 2c, supplemental figure 2e, supplemental figure 2f, supplemental figure 6a–h was performed using pheatmap (42).

Functional assessment of AS events

Inclusion coordinates of AS events were intersected with features from the Exon Ontology database(43). AS events that correlated with survival were identified using rMATS SURVIV (44) and p-values were adjusted with Bonferroni correction for multiple hypothesis testing. Gene Set Enrichment Analysis was performed on low-inclusion and high-inclusion groups using GSEA software (45) on the Hallmark Gene Sets. The genes containing excluded introns identified to be significant in 20 or 30 comparisons of SF groups and healthy controls were analyzed by Gene Ontology(46) (GO) for enrichment in a molecular function (FDR ≤0.05)

Identification of Disease Specific Splicing Events (DSSEs)

Splice junctions observed in disease samples with zero reads across junctions in the healthy bone marrow samples were selected. DSSEs were filtered for coverage: 3 reads covering the junction in a single sample, 20 reads covering the junction across the MDS or AML cohorts, and 10 samples with ≥ 10% exclusion and 0 reads across all GTEX tissues. DSSEs were translated, the reading frame confirmed by blastp, and tested for binding to MHC class I using NetMHC (HLA supertypes, 9mer peptide sequences, ≥Rank % 2).

Additional method details are described in the Supplemental Materials, code used to perform analyses is available in our Github repository https://github.com/hershbc/AS_Analysis_myeloid_neoplasms. DNA and RNA sequences are available by request to Torsten Haferlach.

Results and Discussion

Seven splicing factors are frequently mutated in myeloid neoplasms

To identify alternative splicing (AS) patterns in myeloid neoplasms, we performed deep RNA next generation sequencing (NGS) on the bone marrow cells of 1,258 patients with myeloid neoplasms (MDS, AML, CMML, MDS/MPN) and 63 healthy bone marrow controls. In parallel, we performed whole genome sequencing to identify somatic variants (Supplemental Fig. 1a). We identified 728 patients harboring mutations or deletions in splicing factors *DDX41, LUC7L2, PRPF8, SF3B1, SRSF2, U2AF1*, and *ZRSR2* (Fig. 1a). The proportion of SF mutations varied from 38–98% of each disease cohort (Fig. 1b). 1,177 patients (94%) harbored mutations in at least one SF gene and/or 66 non-SF myeloid neoplasm driver genes (Supplemental Fig. 1a). Previously documented hotspot mutations in SF genes (*U2AF1*^{S34}, *U2AF1*^{Q157}, *SRSF2*^{P95}, and *SF3B1*^{K700}, *DDX41*^{R525H}, *PRPF8*^{D1598}) were identified alongside rarely occurring and uncharacterized missense mutations in these genes (Supplemental Fig. 1b). Predicted loss of function (LoF) events (nonsense or frame shift mutations and chromosomal deletions) were also observed in *DDX41, LUC7L2, PRPF8* and *ZRSR2* (Supplemental Fig. 1b). Samples with hotspot mutations in *U2AF1*, *SRSF2*, and *SF3B1* were mutually exclusive (Supplemental Fig. 1c)(1, 3, 7, 47). In contrast, 73/168 samples with LoF lesions in DDX41, ZRSR2, PRPF8 and LUC7L2 had an aberration in at least one other SF, although the variant allele frequency indicated they may not be in present in the same clone. (Supplemental Fig. 1d).

LoF mutations and chromosomal deletions result in decreased expression of SFs

Since LoF of *LUC7L2* and *ZRSR2* has been shown to impact hematopoietic differentiation (4, 48), we measured expression of these genes in all samples. Expression of *LUC7L2* and *ZRSR2* was decreased overall in myeloid neoplasms compared to healthy control samples, and patients harboring LoF lesions had further decreased expression of these genes (*LUC7L2*: fold change(FC)=0.79 Log_2CPM p<0.001, *DDX41*: FC=0.86 Log_2CPM p<0.001, *PRPF8*: FC=0.91 Log_2CPM p<0.001, *ZRSR2* Males: FC =0.85 Log_2CPM p=0.00098) (Fig. 1c). We observed that many SF wild-type (SF WT) samples also exhibited low *LUC7L2* and/or *ZRSR2* levels, suggesting that mRNA levels of these two genes may be targeted by multiple mechanisms. Uncharacterized somatic mis-sense mutations, mutations within introns, or SNPs in LUC7L2 and ZRSR2 may contribute to the downregulation of these genes. Additionally, slow expression may be a secondary effect caused by a dysregulated transcription factor or epigenetic modifier.

Conversely, *DDX41* and *PRPF8* showed elevated gene expression in disease samples compared to healthy controls. Therefore, patients harboring LoF lesions exhibited expression levels of *DDX41* and *PRPF8* that were closer to the range of expression observed in healthy controls. *DDX41* and *PRPF8* are the only two splicing factors that are targeted by both LoF and missense mutations in myeloid neoplasms. Rare germline *DDX41* LoF mutations can predispose patients to develop MDS, often requiring a second hit to cause disease (5). *PRPF8* missense mutations result in change of function, and patients with lower expression are known to exhibit fewer AS changes than those with missense mutations (6).

Identification and quantification of annotated and unannotated splice junctions

The potential for activation of novel splice sites and the limited annotation of existing isoforms led us to identify AS events *de novo* by creating a custom splicing annotation for our cohort. Over 5×10^5 potential splicing events supported by RNA sequencing data were annotated by StringTie and quantified by rMATS. Stringent filtering (see methods) resulted in 170,006 AS events with good sequencing coverage (Supplemental Fig. 2a). For each patient, we calculated percent spliced in (PSI) for alternatively spliced / skipped exons (SE), alternatively spliced / retained introns (RI), and use of alternative 5’ splice sites (A5SS) or alternative 3’ splice sites (A3SS) (Supplemental Fig. 2b). Alternatively spliced exons were the most prevalent AS event observed (Fig. 2a). As expected, analysis of PSI for each AS event revealed that the majority differed across the cohort by only 5–20%, but >1000 AS events had a range greater than 80% (Fig. 2b, Supplemental Fig. 2d). Ninety percent of these AS events had a mean inclusion level between 0–10% or 90–100%. Thus, a single isoform was often dominant among the transcripts that we quantified, and an even balance between inclusion and exclusion of a coding sequence was rare. (Supplemental Fig. 2c). Notably, the alternatively spliced exons we measured were most often included (mean PSI 90–100%), while the retained introns were most often removed (mean PSI 0–10%).

AS signatures cluster based on the presence of SF mutations, SF expression levels, and disease

Prior to dividing the cohorts by mutation and/or disease, we used unsupervised hierarchical clustering to identify groups of patients with common splicing signatures. Clustering of the 20,000 most variable AS events (those with the largest PSI range) of each type (SE, RI, A5SS, A3SS) revealed that SF mutation status alone does not explain AS differences between samples. Additional factors, such as low SF expression and disease, contributed to the splicing signatures of each patient. (Fig. 2c, Supplemental Fig. 2e). To better understand the similarities and differences of splicing patterns in patients that harbor missense mutations within the same splicing factor, while avoiding sub-clustering due to disease, samples in the MDS cohort containing *SF3B1*, *U2AF1*, and *SRSF2* mutations were also independently clustered alongside SF WT samples. Previous studies have suggested that other mutations in *SF3B1* heat repeats 5–9 are functionally similar to the K700E hotspot (16, 22, 32, 37). Indeed, we see that mutations in this region also co-cluster in our analysis

(Supplemental Fig. 2f). Previous work has also suggested that mutations in the two zinc finger domains in *U2AF1* have entirely separate targets (21, 22, 28) or form sub-clusters due to both common and distinct targets (37). Our data showed *U2AF1* hotspot mutations forming sub-clusters. This suggests that patients harboring U2AF1^{S34} or U2AF1^{Q157} share some overlapping and some distinct dysregulated splicing or that the two mutations affect the same exon or intron, but in different ways. As expected, the various *SRSF2* mutations encompassing P95 also generally co-clustered, including the P95-R102 deletion(20, 37).

Upon clustering of AS signatures, the samples harboring LoF in *DDX41, LUC7L2, PRPF8* and *ZRSR2* did not form groups (data not shown). However, when we annotated all patients in the lowest expression decile, irrespective of mutation, we observed distinct groups (Fig. 2c, Supplemental Fig. 2e). Low *LUC7L2* and/or low *ZRSR2* patients co-clustered and segregated distinctly from low *DDX41* and/or low *PRPF8* patients (Fig. 2c, Supplemental Fig. 2e). We observed some overlap between low *LUC7L2* and low *ZRSR2* groups, as well as between low *DDX41* and low *PRPF8* groups (Fig. 2d, Supplemental Fig. 2g). The mRNA levels of *LUC7L2* and *ZRSR2* are correlated, as are the mRNA levels of *DDX41* and *PRPF8* (Fig. 2d). The splicing signatures of samples with low *DDX41* and low *PRPF8* clustered with healthy controls. (Fig. 2c, Supplemental Fig. 2e) This is concordant with our observation that low *DDX41* and low *PRPF8* groups had similar expression levels to that of healthy controls (Fig. 1c). We also examined the distribution of non-SF mutations (Supplemental Fig. 2h) but none appeared to explain any of the additional clustering in AS signatures.

The results of our unsupervised clustering strengthened our confidence in our groupings for downstream AS analyses. We noted that AS events drive the formation of clusters that distinguish disease groups, with the differences between MDS and AML patients being the most dramatic. While we expected missense mutations in and around hotspots to cluster, samples with low expression of *LUC7L2*, *ZRSR2*, *PRPF8* and *DDX41* also exhibited distinct splicing signatures.

17,300 dysregulated AS events between SF groups, disease controls, and healthy controls

We identified distinct dysregulated AS events in molecularly-defined SF groups (stratified by disease type, co-clustering mutations, or SF low expression) that differed from disease controls (SF WT or SF high expression) or from healthy controls. These comparisons resulted in 17,300 AS events that were significantly dysregulated in at least one of 60 analytic comparisons (ΔPSI ≥ 5%, *q*-value ≤ 0.05)(Fig. 3a,b, Supp. Table 1–6).

Identifying AS events for each SF group that significantly differed from both disease and healthy controls revealed considerable overlap of AS patterns (Fig. 3b). High overlap is indicative of SF mutations or downmodulations that produce splicing patterns that are distinct from both disease control and healthy controls. The percentage of dysregulated AS events in the disease control comparisons that are also significant when compared to healthy control is indicated.

Mis-splicing patterns are consistent among diseases with common SF mutations

We then characterized the types of dysregulated splicing for each molecularly defined SF group. The results were concordant with other studies conducted on SF missense mutations in myeloid neoplasm patient samples, on engineered cell lines, and in murine models (Fig. 3c). Although the number of significant AS events varied between disease cohorts, trends such as increased exon exclusion in *SRSF2*^{P95} samples (8, 15, 20, 22, 25, 31, 37) and altered A3SS usage in *SF3B1*^{MT} samples were consistent across diseases (Fig. 3c) (2, 16–19, 22, 26, 27, 32, 33, 35–37). *U2AF1* mutations displayed trends of dysregulated exon inclusion and alternative 3' splice site usage (1, 21–24, 28–30, 34, 37, 49), and *U2AF1*^{S34} was associated with a greater numbers of events than *U2AF1*^{Q157} (Fig. 3c). Differences in the number of significant AS events between disease cohorts with the same SF mutation may be partially related to sample size, underscoring the importance of procuring large cohorts for the identification of significantly dysregulated AS events (Supplemental Fig. 3c).

Expression levels of SFs have broad impact on cassette exon and intron splicing

To understand the impact of low expression of SFs on alternative splicing, we identified dysregulated AS events associated with differences in expression of *DDX41*, *LUC7L2*, *PRPF8* and *ZRSR2*. The expression-based comparisons yielded far greater numbers of significant AS events than the SF mutation cohorts (Fig. 3d). Taken in conjunction with the unsupervised hierarchical clustering data, our results indicate that expression of SFs have a broad impact on AS in hematopoietic progenitor cells.

Notably, *LUC7L2* and *ZRSR2* groups showed similar patterns of intron mis-splicing, with the majority of introns being more frequently excluded (4, 37) (Supplemental Fig 3d). This trend was also observed in our *SF3B1*^{MT} samples and has been highlighted recently (22). Additionally, a re-assessment of AS studies in *SF3B1* models and patients revealed that exclusion of RI is commonly found in RNA-Seq data (18, 19, 26, 27, 37).

Recurrent Targets of Mis-Splicing

To identify the genes whose splicing is most frequently affected by SF mutations and low SF expression, we highlighted genes concordant among the same molecularly defined SF group, regardless of disease status (Extended Data Fig. 4a). Mis-spliced genes specific to a single disease could be involved in individual disease heterogeneity. Although no gene was ubiquitously mis-spliced among the 8 SF groups, the most frequently affected genes included ubiquitination factors (*USP24, MARCH6, USP25, USP15, UBR4*), splicing factors (*RBM39, LUC7L2, SNRPA1, SMU1, LSM14A*), transcription factors (*BTAF1, CHD2, ZEB2*) and cancer-related genes (*RB1CC1, PHIP, SP100, SPAG9, PCM1, BIRC6, PTPRC*) (Supplementary Fig. 4b).

We also identified the most frequently dysregulated AS events (i.e. significant in the majority of comparisons, green column Fig. 3a). The 50 most frequently affected AS events were located in genes for ubiquitination and deubiquitination factors (*UBE2D3*, *UB32L3*, *USP15*, *RNF167*, and *MARCH6*), mitochondrial factors (*MRPL33*, *YME1L1*, and *UQCRC2*), transcription factors (*NCOR1*, *TFDP1*, and *BTAF1*), DNA repair factors (*TDP1* and *SMC5*), genes with a role in oncogenesis (*OCIAD1*, *SMEK1*, and *SPAG9*), and genes implicated in leukemia (*ABI1* and *IKZF1*) (Fig. 4a). Of the 17,300 mis-spliced events, 74% were significant in at least 2 of the 60 comparisons shown in Figure 3a. Therefore, we have made the q-values, mean PSI values, and ΔPSI values of each significant AS event available in easily queriable tables (Supplemental Table 1a–e) to encourage mechanistic studies of dysregulated AS events in myeloid neoplasms.

Predicted impact of dysregulated AS on biological and clinical outcomes

We have provided an unprecedented resource in the identification of highly significant AS events driven by splicing factor mutations that are reproducible across multiple myeloid neoplasm subtypes. However, the question of whether or not the these mis-splicing events results in degradation by NMD or confers altered function remains, and understanding which mis-spliced transcripts are most likely to contribute to the disease phenotype is essential.

To this end, we designed a pipeline to infer the functional impact of AS dysregulation *in silico*. We identified hundreds of AS events that activate cryptic splice sites, shift the reading frame, and correlate with decreased transcript expression (Fig. 4b, Supplemental Table 4). We also identified AS events that are likely translated and encode functional protein-domains (Fig.4b, Supplemental Table 2). Finally, we analyzed associations between AS events and overall survival as well as and Hallmark pathway dysregulation (Fig. 4b,c, Supplemental Fig. 4c,d, Supplemental Table 3). Integration of these AS data, gene expression data, clinical variables and exon ontology annotations provided insights into the functional impact of the most recurrently dysregulated AS events (Fig. 4a) in myeloid neoplasms. For example, *SRSF1* A3SS 4 introduced protein structure and a binding site, potentially altering the function of *SRSF1*. *TRA2A* exon 3b was positively correlated with TRA2A gene expression. Inclusion of TRA2A exon 3b and *SRSF1* A3SS 4 were both inversely correlated with survival in AML and with expression of oxidative phosphorylation targets (Fig. 4b,c, Supplemental Table 2,3). These examples aside, we provide the functional assessment (all analyses in Fig. 4b) for each of the 17,300 significantly dysregulated events identified in our study (Supplemental Tables 2, 3a–e, 4).

Mis-spliced introns distinguish disease from healthy bone marrow

The differences between the AS patterns in myeloid neoplasms compared to healthy bone marrow are poorly understood, therefore we focused on characterizing the AS events that were significant when the SF groups were compared to healthy controls, as opposed to disease controls. The mis-splicing events that deviated most strongly from healthy bone marrow included both SF specific and pan-disease mis-splicing events (Fig 5a). We collected the AS events that most often distinguished myeloid malignancies from healthy bone marrow regardless of splicing factor mutation. Within this group, excluded RIs and A5SSs were overrepresented (Fig 5b). This set of introns displayed unusual features; the majority of these introns were poorly spliced in healthy bone marrow (mean PSI 40–100%), with increased exclusion across multiple SF and SF WT groups (Fig. 5c). The high PSIs of the introns identified in healthy controls were unexpected given the low proportion of introns with a mean PSI >10% (Supplemental Fig. 2c). The introns were also significantly shorter on average (2393 bp *vs*. 6190 bp) and very few intron inclusion levels were correlated with changes in transcript expression (Fig. 5d, Supplemental Fig. 5a,b). RI-containing transcripts were enriched in GO terms such as gene expression, RNA splicing and RNA transport (Fig. 5e).

The high inclusion levels in quiescent cells, lack of correlation with transcript stability, and shorter length suggest that we have identified a group of “detained introns.” Detention of introns is a known mechanism of translational control in quiescent cells, although the specific introns involved are likely to be tissue specific (50–55). Increased IR is essential for differentiation of early progenitor cells into myeloid lineage cells(52, 56, 57). Increased removal of specific retained introns is also observed in self-renewing cells as well as in breast cancer and *SF3B1* mutant MDS (18, 19, 22, 26, 27, 37, 58). We observed this trend in multiple splicing factor groups (Fig 3c,d), but also across all myeloid neoplasm samples within our cohort. This is the first indication that detained introns are excluded in the majority of myeloid neoplasm progenitor cells and suggests that the failure to retain introns may be connected to the failure of differentiation in myeloid lineage cells.

Classification of rare mutations in SFs

Uncharacterized mutations outside of hotspot positions (Supplemental Fig. 1b) were not included in our SF groups, but it is possible that these newly identified putative somatic mutations in the SFs would have a similar impact on AS. Having established AS signatures for each SF molecularly-defined group, we used hierarchical clustering to identify uncharacterized missense mutations in the SFs that phenocopy the AS patterns (Fig. 6a, Supplemental Fig. 6a–h). Previously uncharacterized mutations seen in our cohort, including those within *SF3B1*; K700_V701insE, M784_K785delinsI, W658C and H662Y, cluster with the previously reported heat repeat mutations (Supplemental Fig. 6a). Mutations in *SRSF2*; F57Y and D73G clustered with P95 mutations, while P96L did not (Supplemental Fig. 6b). *U2AF1*^{S34} and *U2AF1*^{Q157} clustered separately and no other *U2AF1* mutations clustered with S34(F,Y) (Supplemental Fig. 6 c,d). However, other *U2AF1* mutations (C154S, Y158_E159dup and R156H) clustered with Q157(P,R) (Supplemental Fig. 6d). In contrast, M1f., M172L and Q157H did not cluster with either S34F or Q157(P,R). The analysis also revealed point mutations that may confer loss of function on *LUC7L2* and *ZRSR2*. Patients with missense mutations in *LUC7L2*; R71H, splice site mutations, and alternate start sites (M1) clustered with low *LUC7L2* expressers (Supplemental Fig. 6e). Patients with mutations in *ZRSR2*; G438_R442dup, C172Y, H330Y, Y271C, H330, and splice site mutations clustered with low *ZRSR2* expressers (Supplemental Fig. 6f). These data contribute to our understanding of functional domains in SFs, and provide opportunities to personalize therapy based on common splicing signatures, rather than specific missense mutations.

Identification of neojunctions

The complex analysis performed also indicated that AS in myeloid neoplasms generates predictable novel coding sequences that could potentially be targeted for diagnostics or therapy. AS can generate novel protein junctions through gain or loss of coding RNA sequences or the introduction of frameshift alterations(59, 60). We classified myeloid disease-specific splicing events (DSSEs) as exon exclusion events that were observed in the disease cohorts but did not occur in the healthy controls. After filtering for reproducibility and testing for binding to class I MHC, we identified 925 DSSEs (Fig. 6b) (61, 62). Twenty DSSEs were identified (≥%5 exon exclusion) in the majority of patients in one or more SF mutation group (Fig. 6c). Several DSSEs were splicing-factor specific and reproducible between MDS and AML cohorts, e.g. ***SF3B1*** (*FMNL1, PILRB, FAM143A*), ***ZRSR2*** (*EIF4G3, ZNF207*), and ***SRSF2*** (*RASGRP2, CD2BP2*). Others were found in 20–60% of MDS and AML patients regardless of SF mutational status (*CPXM1, TPT1, IPO4*). Neojunctions generated through AS can be translated into neoepitopes that have the potential to be targeted through immunotherapy(59, 60). Our study has identified neojunctions that are specific to SF mutations as well as neojunctions that are pan-disease.

In summary, we identified unique AS events based on SF mutations and SF expression as well as shared events between all patients, irrespective of mutations. Therefore, SF mutations and downmodulation exaggerate pathological splicing patterns, in addition to producing individual characteristics. Our data indicate that a collection of ordinarily retained introns are excluded in patients, potentially blocking differentiation and maintaining a myeloid precursor phenotype. Thus, the search for a discrete abnormality or singular event will not be successful. This highlights the continuum of disease; enhanced intron exclusion is observed in all patient samples and genetic aberrations in SFs exacerbate further intron exclusion. Although we no longer expect to find a singular AS event driving pathogenesis, diseased AS signatures provide multiple targets with the potential to be targeted by treatment(59, 60). Both SF-specific and pan-cancer novel AS events that arose in myeloid neoplasms were identified in our work. These events, combined with our classification of uncommon SF mutations, identify specific groups of myeloid neoplasm patients expressing novel coding sequence that can be used for diagnosis or potential targeted therapy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by U.S. NIH grants R01HL123904, R01HL132071 and R35HL135795 and Torsten Haferlach Leukämiediagnostik Stiftung. The authors wish to acknowledge the Vera and Joseph Dresner Foundation as a funding source and thank them for their contributions to this work. CEH was supported by NIH F31HL131140. We thank the staff of MLL for sample preparation and sequencing, RC Dietrich for helpful discussions and comments regarding the paper, DM Rotroff for assistance with clustering analysis, and T Radivoyevitch for advice on statistics.

References

1. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. Nature. 2011;478 (7367):64–9. [PubMed: 21909114]
2. Makishima H, Visconte V, Sakaguchi H, Jankowska AM, Abu Kar S, Jerez A, et al. Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis. Blood. 2012;119(14):3203–10. [PubMed: 22323480]
3. Haferlach T, Nagata Y, Grossmann V, Okuno Y, Bacher U, Nagae G, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. Leukemia. 2014;28(2):241–7. [PubMed: 24220272]

4. Madan V, Kanojia D, Li J, Okamoto R, Sato-Otsubo A, Kohlmann A, et al. Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. Nature communications. 2015;6:6042.

5. Polprasert C, Schulze I, Sekeres MA, Makishima H, Przychodzen B, Hosono N, et al. Inherited and Somatic Defects in DDX41 in Myeloid Neoplasms. Cancer cell. 2015;27(5):658–70. [PubMed: 25920683]

6. Kurtovic-Kozaric A, Przychodzen B, Singh J, Konarska MM, Clemente MJ, Otrock ZK, et al. PRPF8 defects cause missplicing in myeloid malignancies. Leukemia. 2015;29(1):126–36. [PubMed: 24781015]

7. Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. Blood. 2013;122(22):3616–27; quiz 99. [PubMed: 24030381]

8. Yoshimi A, Lin KT, Wiseman DH, Rahman MA, Pastore A, Wang B, et al. Coordinated alterations in RNA splicing and epigenetic regulation drive leukaemogenesis. Nature. 2019.

9. Cvitkovic I, Jurica MS. Spliceosome database: a tool for tracking components of the spliceosome. Nucleic Acids Res. 2013;41(Database issue):D132–41. [PubMed: 23118483]

10. Wang C, Chua K, Seghezzi W, Lees E, Gozani O, Reed R. Phosphorylation of spliceosomal protein SAP 155 coupled with splicing catalysis. Genes Dev. 1998;12(10):1409–14. [PubMed: 9585501]

11. Wu S, Romfo CM, Nilsen TW, Green MR. Functional recognition of the 3' splice site AG by the splicing factor U2AF35. Nature. 1999;402(6763):832–5. [PubMed: 10617206]

12. Shao C, Yang B, Wu T, Huang J, Tang P, Zhou Y, et al. Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome. Nat Struct Mol Biol. 2014;21(11):997–1005. [PubMed: 25326705]

13. Liu HX, Chew SL, Cartegni L, Zhang MQ, Krainer AR. Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. Mol Cell Biol. 2000;20(3):1063–71. [PubMed: 10629063]

14. Daubner GM, Clery A, Jayne S, Stevenin J, Allain FH. A syn-anti conformational difference allows SRSF2 to recognize guanines and cytosines equally well. Embo j. 2012;31(1):162–74. [PubMed: 22002536]

15. Zhang J, Lieu YK, Ali AM, Penson A, Reggio KS, Rabadan R, et al. Disease-associated mutation in SRSF2 misregulates splicing by altering RNA-binding affinities. Proc Natl Acad Sci U S A. 2015;112(34):E4726–34. [PubMed: 26261309]

16. Darman RB, Seiler M, Agrawal AA, Lim KH, Peng S, Aird D, 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):1033–45. [PubMed: 26565915]

17. Dolatshad H, Pellagatti A, Fernandez-Mercado M, Yip BH, Malcovati L, Attwood M, et al. Disruption of SF3B1 results in deregulated expression and splicing of key genes and pathways in myelodysplastic syndrome hematopoietic stem and progenitor cells. Leukemia. 2015;29(5):1092–103. [PubMed: 25428262]

18. Dolatshad H, Pellagatti A, Liberante FG, Llorian M, Repapi E, Steeples V, et al. Cryptic splicing events in the iron transporter ABCB7 and other key target genes in SF3B1-mutant myelodysplastic syndromes. Leukemia. 2016;30(12):2322–31. [PubMed: 27211273]

19. Kesarwani AK, Ramirez O, Gupta AK, Yang X, Murthy T, Minella AC, et al. Cancer-associated SF3B1 mutants recognize otherwise inaccessible cryptic 3' splice sites within RNA secondary structures. Oncogene. 2017;36(8):1123–33. [PubMed: 27524419]

20. Kim E, Ilagan JO, Liang Y, Daubner GM, Lee SC, Ramakrishnan A, et al. SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. Cancer cell. 2015;27(5):617–30. [PubMed: 25965569]

21. Ilagan JO, Ramakrishnan A, Hayes B, Murphy ME, Zebari AS, Bradley P, et al. U2AF1 mutations alter splice site recognition in hematological malignancies. Genome Res. 2015;25(1):14–26. [PubMed: 25267526]

22. Shiozawa Y, Malcovati L, Galli A, Sato-Otsubo A, Kataoka K, Sato Y, et al. Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. Nature communications. 2018;9(1):3649.

23. Smith MA, Choudhary GS, Pellagatti A, Choi K, Bolanos LC, Bhagat TD, et al. U2AF1 mutations induce oncogenic IRAK4 isoforms and activate innate immune pathways in myeloid malignancies. Nat Cell Biol. 2019;21(5):640–50. [PubMed: 31011167]

24. Brooks AN, Choi PS, de Waal L, Sharifnia T, Imielinski M, Saksena G, 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. [PubMed: 24498085]

25. Kon A, Yamazaki S, Nannya Y, Kataoka K, Ota Y, Nakagawa MM, et al. Physiological Srsf2 P95H expression causes impaired hematopoietic stem cell functions and aberrant RNA splicing in mice. Blood. 2018;131(6):621–35. [PubMed: 29146882]

26. Obeng EA, Chappell RJ, Seiler M, Chen MC, Campagna DR, Schmidt PJ, et al. Physiologic Expression of Sf3b1(K700E) Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation. Cancer cell. 2016;30(3):404–17. [PubMed: 27622333]

27. Mupo A, Seiler M, Sathiaseelan V, Pance A, Yang Y, Agrawal AA, et al. Hemopoietic-specific Sf3b1-K700E knock-in mice display the splicing defect seen in human MDS but develop anemia without ring sideroblasts. Leukemia. 2017;31(3):720–7. [PubMed: 27604819]

28. Shirai CL, Ley JN, White BS, Kim S, Tibbitts J, Shao J, et al. Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo. Cancer cell. 2015;27(5):631–43. [PubMed: 25965570]

29. Yip BH, Steeples V, Repapi E, Armstrong RN, Llorian M, Roy S, et al. The U2AF1S34F mutation induces lineage-specific splicing alterations in myelodysplastic syndromes. J Clin Invest. 2017;127(9):3557.

30. Fei DL, Zhen T, Durham B, Ferrarone J, Zhang T, Garrett L, et al. Impaired hematopoiesis and leukemia development in mice with a conditional knock-in allele of a mutant splicing factor gene U2af1. Proc Natl Acad Sci U S A. 2018;115(44):E10437–e46. [PubMed: 30322915]

31. Colla S, Ong DS, Ogoti Y, Marchesini M, Mistry NA, Clise-Dwyer K, et al. Telomere dysfunction drives aberrant hematopoietic differentiation and myelodysplastic syndrome. Cancer cell. 2015;27(5):644–57. [PubMed: 25965571]

32. DeBoever C, Ghia EM, Shepard PJ, Rassenti L, Barrett CL, Jepsen K, et al. Transcriptome sequencing reveals potential mechanism of cryptic 3’ splice site selection in SF3B1-mutated cancers. PLoS Comput Biol. 2015;11(3):e1004105. [PubMed: 25768983]

33. Visconte V, Makishima H, Jankowska A, Szpurka H, Traina F, Jerez A, et al. SF3B1, a splicing factor is frequently mutated in refractory anemia with ring sideroblasts. Leukemia. 2012;26(3):542–5. [PubMed: 21886174]

34. Przychodzen B, Jerez A, Guinta K, Sekeres MA, Padgett R, Maciejewski JP, et al. Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms. Blood. 2013;122(6):999–1006. [PubMed: 23775717]

35. Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N Engl J Med. 2011;365(15):1384–95. [PubMed: 21995386]

36. Conte S, Katayama S, Vesterlund L, Karimi M, Dimitriou M, Jansson M, et al. Aberrant splicing of genes involved in haemoglobin synthesis and impaired terminal erythroid maturation in SF3B1 mutated refractory anaemia with ring sideroblasts. British journal of haematology. 2015;171(4):478–90. [PubMed: 26255870]

37. Pellagatti A, Armstrong RN, Steeples V, Sharma E, Repapi E, Singh S, et al. Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. Blood. 2018;132(12):1225–40. [PubMed: 29930011]

38. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics. 2015;51:11.4.1–9. [PubMed: 26334920]

39. Robinson MD, McCarthy Dj Fau - Smyth GK, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. (1367–4811 (Electronic)).

40. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5. [PubMed: 25690850]

41. Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A. 2014;111(51):E5593–601. [PubMed: 25480548]
42. Kolde R pheatmap: Pretty Heatmaps. 2019.
43. Tranchevent LC, Aube F, Dulaurier L, Benoit-Pilven C, Rey A, Poret A, et al. Identification of protein features encoded by alternative exons using Exon Ontology. Genome Res. 2017;27(6):1087–97. [PubMed: 28420690]
44. Shen S, Wang Y, Wang C, Wu YN, Xing Y. SURVIV for survival analysis of mRNA isoform variation. Nature communications. 2016;7:11548.
45. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50. [PubMed: 16199517]
46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000;25(1):25–9. [PubMed: 10802651]
47. Patnaik MM, Lasho TL, Finke CM, Hanson CA, Hodnefield JM, Knudson RA, et al. Spliceosome mutations involving SRSF2, SF3B1, and U2AF35 in chronic myelomonocytic leukemia: prevalence, clinical correlates, and prognostic relevance. American journal of hematology. 2013;88(3):201–6. [PubMed: 23335386]
48. Kotini AG, Chang CJ, Boussaad I, Delrow JJ, Dolezal EK, Nagulapally AB, et al. Functional analysis of a chromosomal deletion associated with myelodysplastic syndromes using isogenic human induced pluripotent stem cells. (1546–1696 (Electronic)).
49. Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. Nature genetics. 2011;44(1):53–7. [PubMed: 22158538]
50. Yap K, Lim ZQ, Khandelia P, Friedman B, Makeyev EV. Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. Genes Dev. 2012;26(11):1209–23. [PubMed: 22661231]
51. Boothby TC, Zipper RS, van der Weele CM, Wolniak SM. Removal of retained introns regulates translation in the rapidly developing gametophyte of Marsilea vestita. Dev Cell. 2013;24(5):517–29. [PubMed: 23434411]
52. Wong JJ, Ritchie W, Ebner OA, Selbach M, Wong JW, Huang Y, et al. Orchestrated intron retention regulates normal granulocyte differentiation. Cell. 2013;154(3):583–95. [PubMed: 23911323]
53. Boutz PL, Bhutkar A, Sharp PA. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev. 2015;29(1):63–80. [PubMed: 25561496]
54. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, et al. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. 2014;24(11):1774–86. [PubMed: 25258385]
55. 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 research. 2016;44(2):838–51. [PubMed: 26531823]
56. Edwards CR, Ritchie W, Wong JJ, Schmitz U, Middleton R, An X, et al. A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. Blood. 2016;127(17):e24–e34. [PubMed: 26962124]
57. Ni T, Yang W, Han M, Zhang Y, Shen T, Nie H, et al. Global intron retention mediated gene regulation during CD4+ T cell activation. Nucleic acids research. 2016;44(14):6817–29. [PubMed: 27369383]
58. Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. RNA splicing factors as oncoproteins and tumour suppressors. Nat Rev Cancer. 2016;16(7):413–30. [PubMed: 27282250]
59. Shen L, Zhang J, Lee H, Batista MT, Johnston SA. RNA Transcription and Splicing Errors as a Source of Cancer Frameshift Neoantigens for Vaccines. Scientific reports. 2019;9(1):14184. [PubMed: 31578439]

60. Kahles A, Lehmann KV, Toussaint NC, Huser M, Stark SG, Sachsenberg T, et al. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer cell. 2018;34(2):211–24.e6. [PubMed: 30078747]

61. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics. 2016;32(4):511–7. [PubMed: 26515819]

62. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. 2003;12(5):1007–17. [PubMed: 12717023]

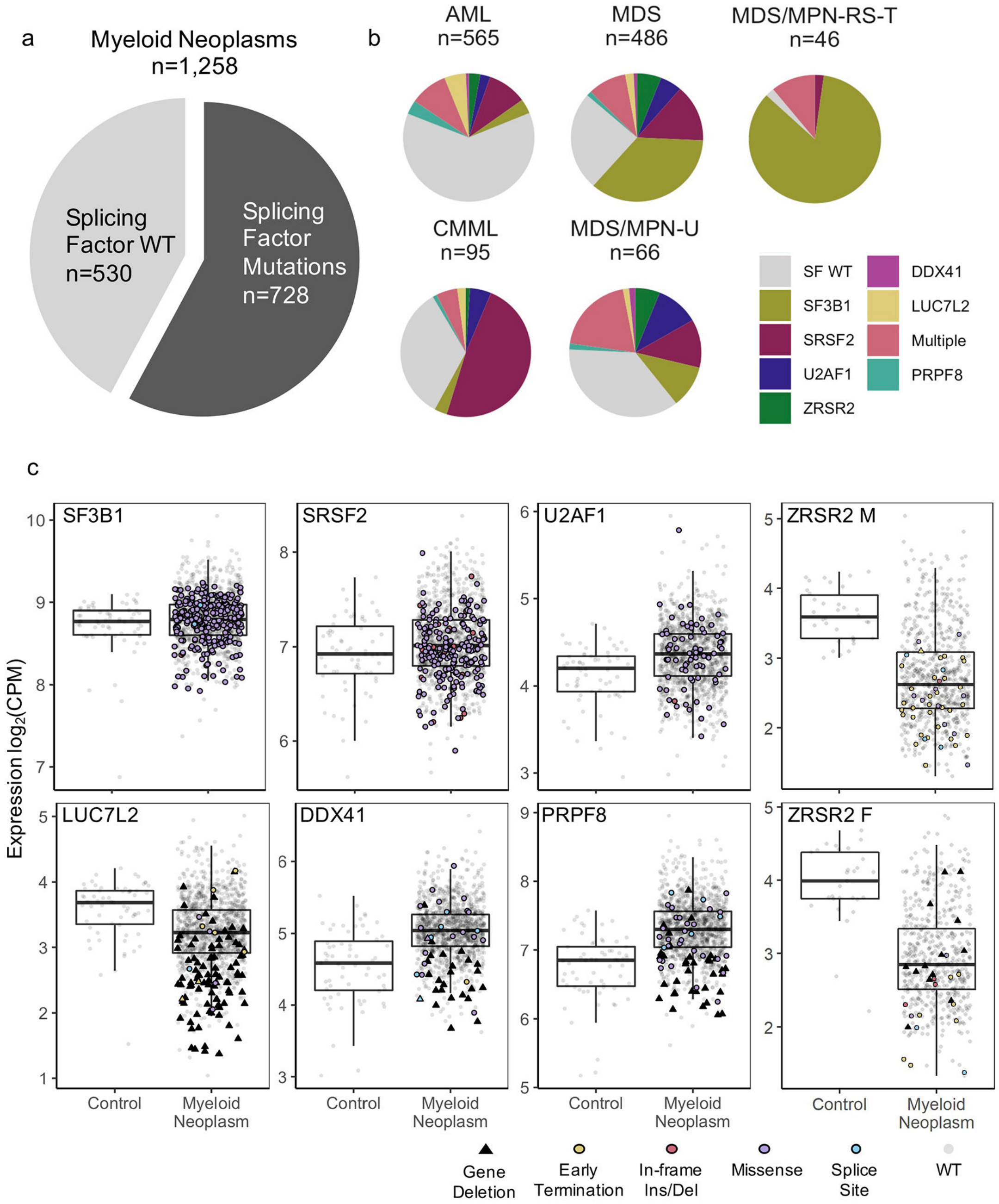


Figure 1: Overview of splicing factor (SF) mutations and expression levels in the cohorts. **(a)** Percentage of patients harboring SF mutations**. (b)** Frequency of SF mutations in the cohort listed by disease from patients with one of five myeloid neoplasms (AML=213/565, MDS=365/486, CMML=63/95, MDS/MPN-U=42/66, MDS/MPN-RS-T=45/46). **(c)** mRNA expression in log2 counts per million reads (CPM) of SFs in healthy BM controls and in myeloid neoplasm patient samples, with and without mutations and/or chromosomal deletions. ZRSR2, an X-linked gene, is grouped by patient sex.

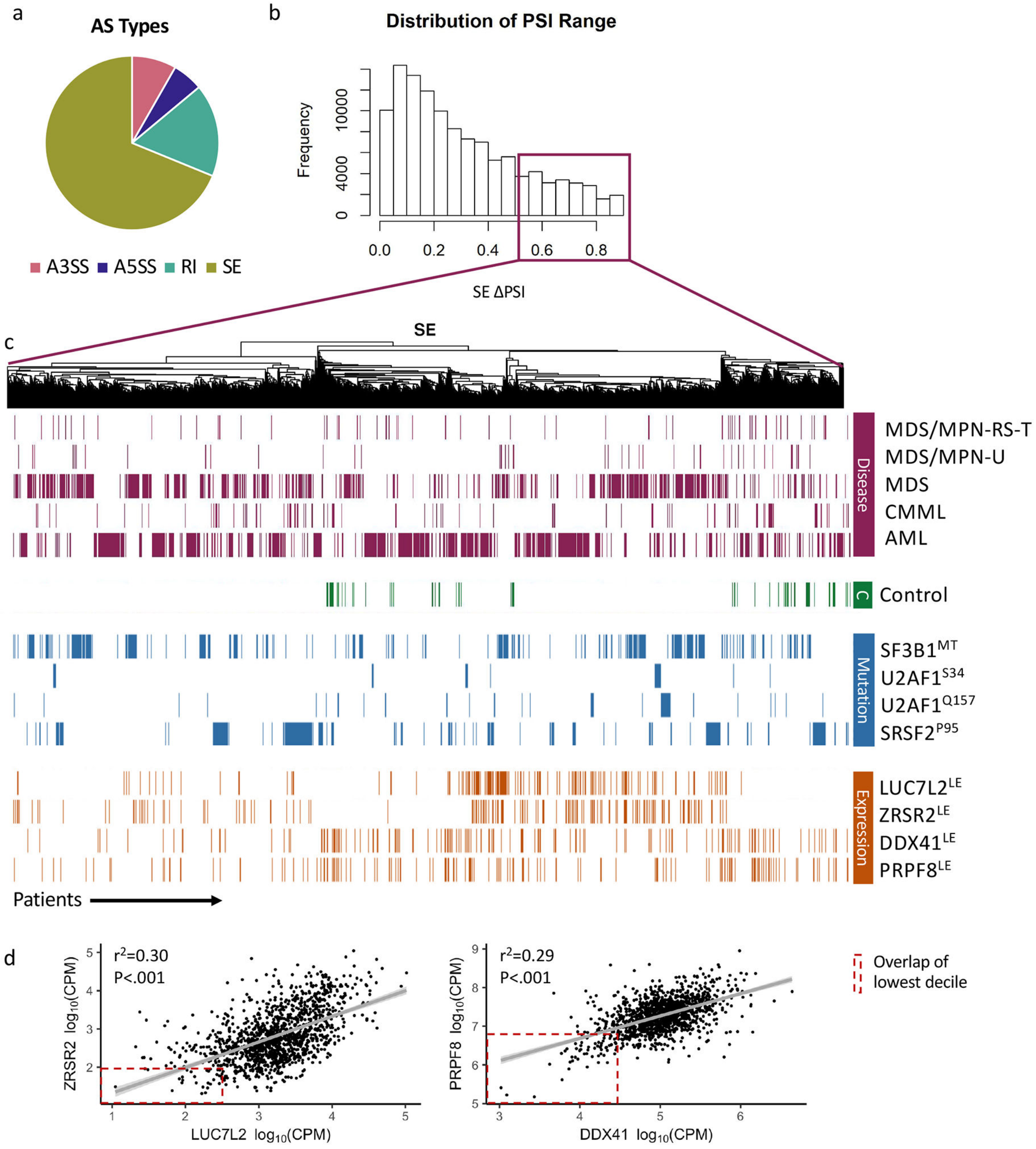
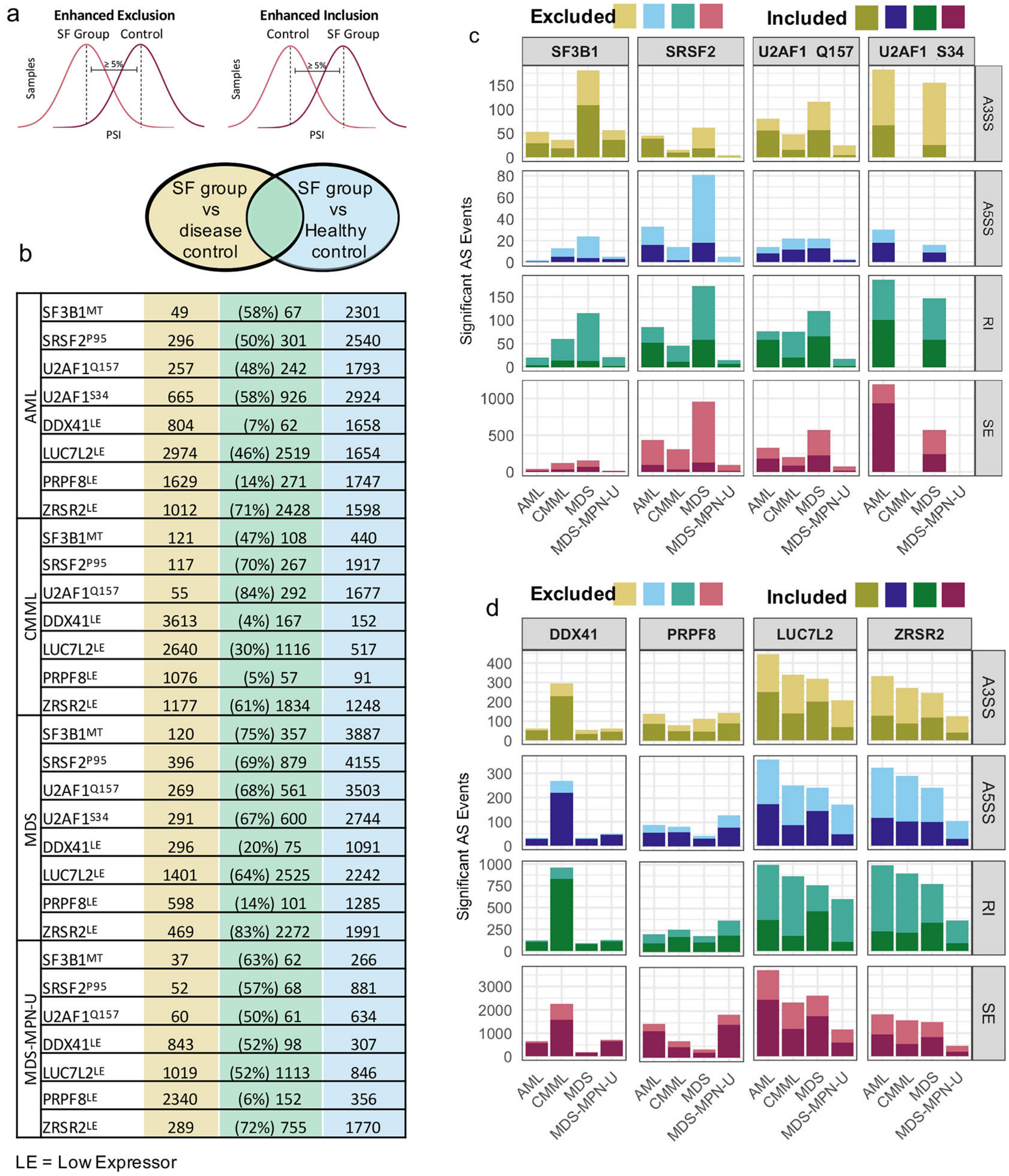


Figure 2:

Overview of alternative splicing (AS) events observed in patient and control samples and quantified by rMATS. **(a)** Distribution of AS types across 170,006 AS events: alternatively spliced / skipped exons (SE), 68.8%; alternatively spliced introns / retained introns (RI), 17.2%; exons with alternative 5' splice sites (A5SS), 5.7%; and exons with alternative 3' splice sites (A3SS), 8.3%. **(b)** Distribution of the percent spliced in (PSI) Range observed across all samples. **(c)** Unsupervised hierarchical clustering of the 20,000 alternative spliced exons with the largest range. Samples are annotated by disease, splicing factor mutation (*SF3B1*, *SRSF2*, *U2AF1*) or splicing factor expression (*DDX41*, *LUC7L2*, *PRPF8*, *ZRSR2*). “LE” indicates that the sample is among the 10% lowest expressers (samples were ranked by

Log_2CPM) in the cohort. **(d)** Correlations of ZRSR2 expression with LUC7L2 expression and DDX41 expression with PRPF8 expression (r^2 and p value generated by linear regression). Overlap of low expressers indicated with red box.



AML	SF3B1^{MT}	49	(58%) 67	2301
	SRSF2^{P95}	296	(50%) 301	2540
	U2AF1^{Q157}	257	(48%) 242	1793
	U2AF1^{S34}	665	(58%) 926	2924
	DDX41^{LE}	804	(7%) 62	1658
	LUC7L2^{LE}	2974	(46%) 2519	1654
	PRPF8^{LE}	1629	(14%) 271	1747
	ZRSR2^{LE}	1012	(71%) 2428	1598
CMML	SF3B1^{MT}	121	(47%) 108	440
	SRSF2^{P95}	117	(70%) 267	1917
	U2AF1^{Q157}	55	(84%) 292	1677
	DDX41^{LE}	3613	(4%) 167	152
	LUC7L2^{LE}	2640	(30%) 1116	517
	PRPF8^{LE}	1076	(5%) 57	91
	ZRSR2^{LE}	1177	(61%) 1834	1248
MDS	SF3B1^{MT}	120	(75%) 357	3887
	SRSF2^{P95}	396	(69%) 879	4155
	U2AF1^{Q157}	269	(68%) 561	3503
	U2AF1^{S34}	291	(67%) 600	2744
	DDX41^{LE}	296	(20%) 75	1091
	LUC7L2^{LE}	1401	(64%) 2525	2242
	PRPF8^{LE}	598	(14%) 101	1285
	ZRSR2^{LE}	469	(83%) 2272	1991
MDS-MPN-U	SF3B1^{MT}	37	(63%) 62	266
	SRSF2^{P95}	52	(57%) 68	881
	U2AF1^{Q157}	60	(50%) 61	634
	DDX41^{LE}	843	(52%) 98	307
	LUC7L2^{LE}	1019	(52%) 1113	846
	PRPF8^{LE}	2340	(6%) 152	356
	ZRSR2^{LE}	289	(72%) 755	1770

LE = Low Expressor

Figure 3: Distribution of dysregulated alternative splicing events by disease and SF mutation or expression.

(a) Schematic showing identification of mis-spliced AS events. **(b)** Overlap of dysregulated splicing events comparing splicing factor group to disease controls and healthy controls. The number of AS events commonly dysregulated are in green. Of all significant AS events in SF group vs disease control, the percentage of those that were also significant in SF group vs healthy control is displayed. **(c)** Number of mis-spliced (ΔPSI ≥ 5%, q-value ≤ 0.05) AS events in samples with SF mutations compared with disease controls. Cohorts are divided by disease and SF mutation status. Lighter color indicates enhanced exclusion of exon or intron; darker color indicates enhanced inclusion of exon or intron. **(d)** Number of mis-spliced AS events in samples with low expression (10% lowest of cohort) of *DDX41*, *LUC7L2*, *PRPF8*, or *ZRSR2* compared to high expressers (10% highest in cohort), legend description as in figure 3c.

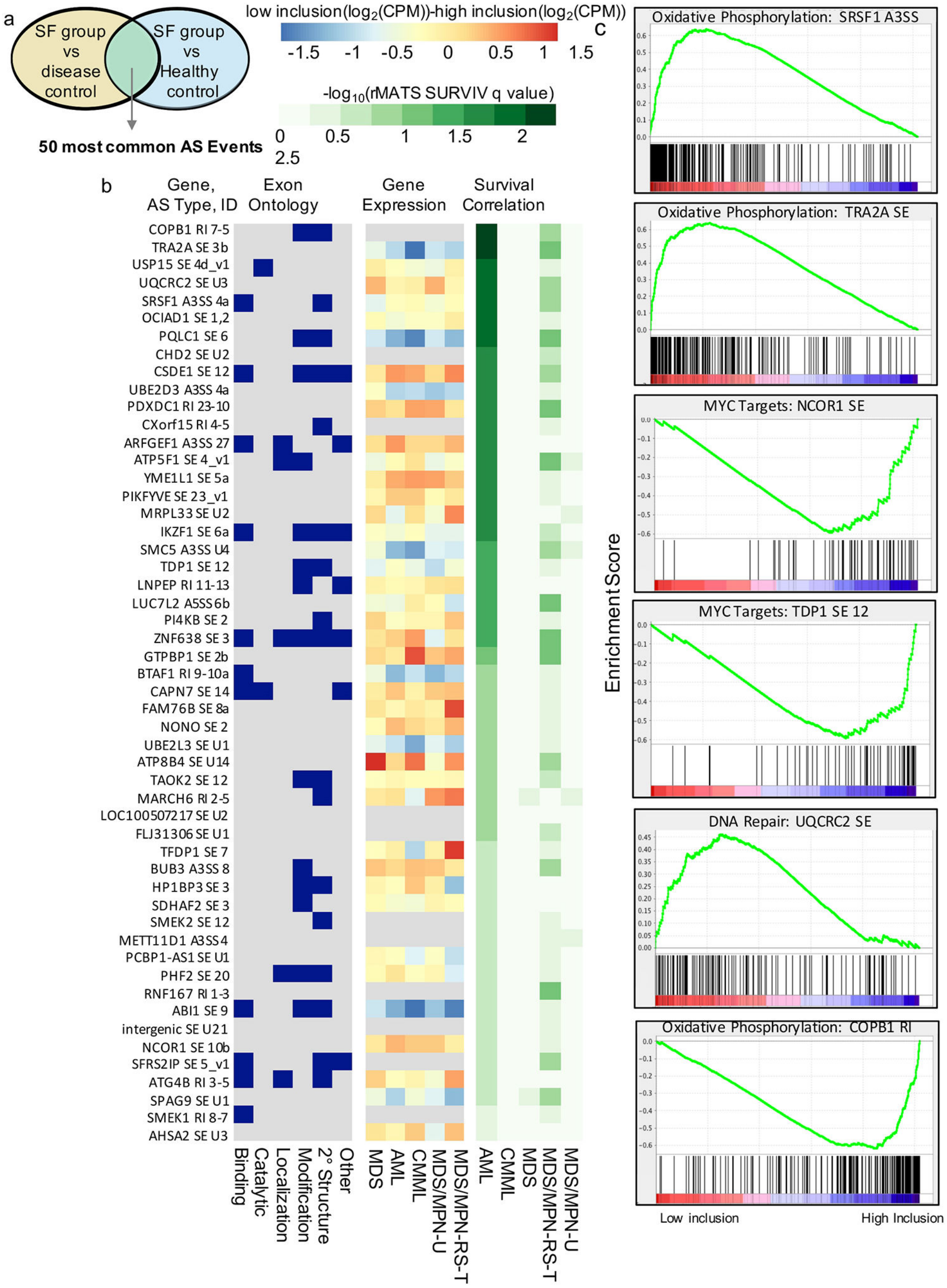


Figure 4: Predicted impact of AS on biological and clinical outcomes.

(a) Selection of most frequently dysregulated AS events from Figure 3c. **(b)** *AS event IDs:* Gene symbols, AS type, and custom exon identifier of the AS event. *Exon Ontology:* Binary chart showing the presence of exon ontology features in dysregulated exons. AS events were classified as containing structural elements, PTMs (post translational modifications), localization signatures, catalytic domains, or binding sites based on intersection with Exon Ontology DB annotation*. Gene Expression:* The cohorts were stratified by the inclusion of each AS event and plotted the expression level difference between the two groups, identifying many AS events whose inclusion correlated with changes in gene expression*.*

Survival Correlation: Correlation of significant AS events and survival in each cohort using RMATS-SURVIV. Results are displayed as –log10 transformed q-values, indicating the false discovery rate (FDR). **(c)** MDS and AML cohorts were combined and stratified by AS event inclusion level. Pathway analysis revealed dysregulated expression of Hallmark Gene sets in low- versus high inclusion groups.

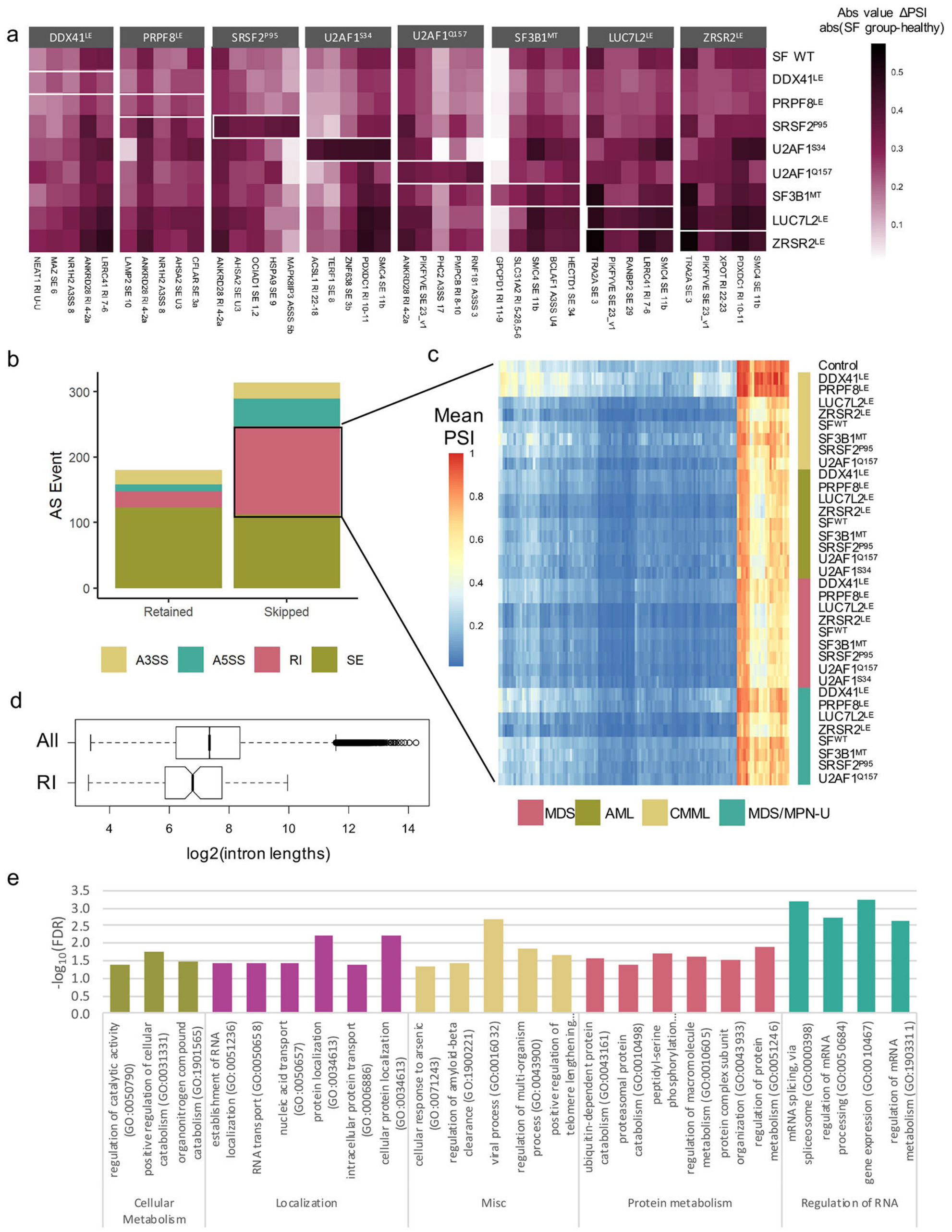


Figure 5: Enhanced exclusion of RIs in SF groups compared to healthy controls.

(a) Top 5 mis-spliced gene in each SF group compared to healthy bone marrow (significant in 4 diseases, largest median ΔPSI). **(b)** Distribution of AS types in 494 AS events found to be significant in >= 20 SF groups compared to healthy controls. **(c)** Mean PSI of each SF group for each excluded RI identified in Fig. 6A. **(d)** Length of all introns and RIs in (b) (T-test, unpaired, two-tailed, unequal variances, P<0.001); retained introns are shorter (2393 bp *vs*. 6190bp). **(e)** Most significant GO terms describing RIs in (b).

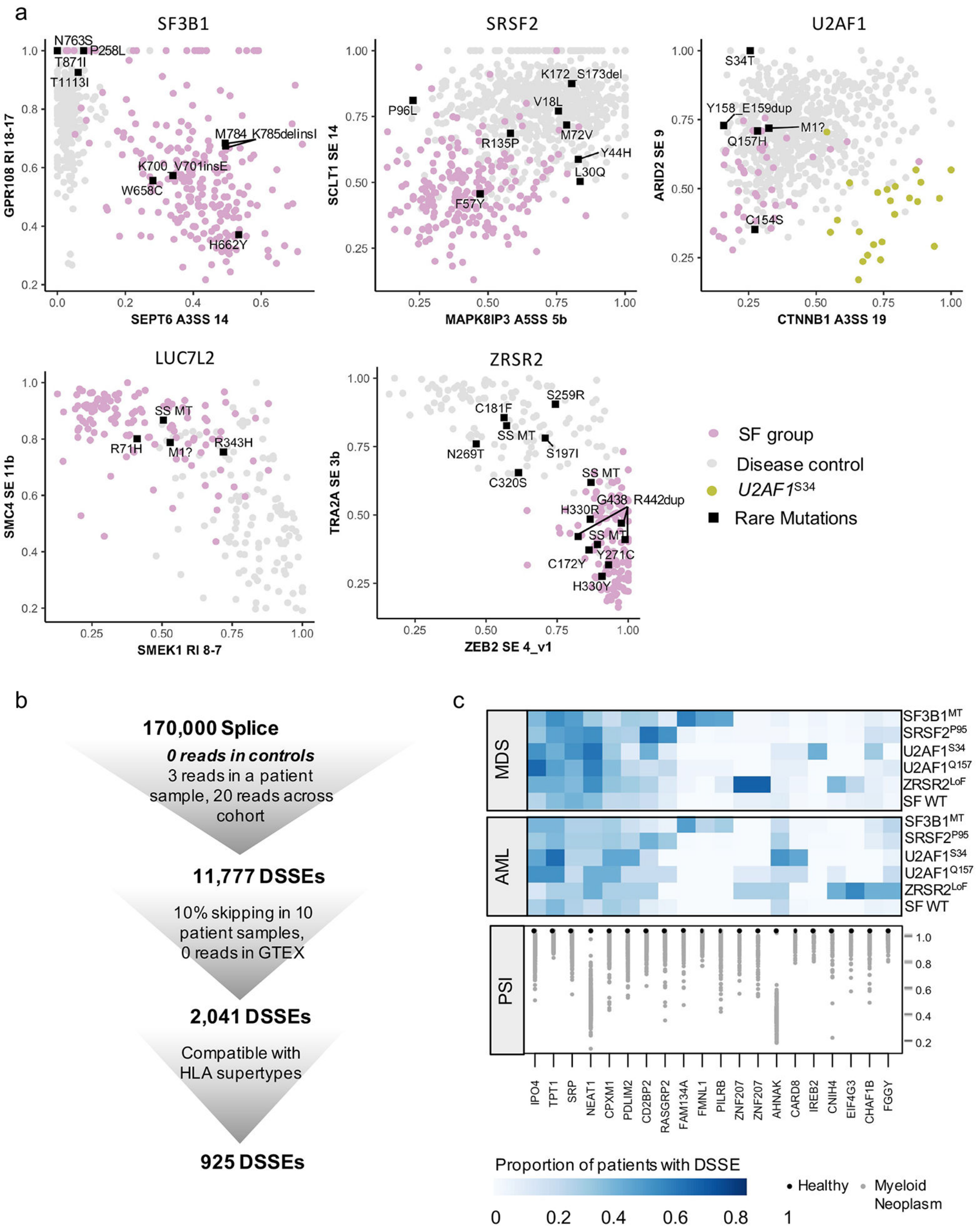


Figure 6: Characterization of rare somatic mutations in SFs and identification of neojunctions.
(a) Plot of PSIs for the two strongest AS events, found in at least 3 disease groups. Rare mutations that have not been recorded by the three databases (COSMIC, ClinVar, Cancer Hotspot) or have not been confirmed to be somatic in COSMIC or pathogenic in ClinVar are displayed as labeled rectangles. **(b)** Schematic of disease-specific splicing events (DSSE): identification and filtering to identify strong and reproducible DSSEs. **(c)** Disease specific splicing events in MDS and AML that are found in ≥50% patients in at least one SF group and never in healthy controls.