U2AF1 mutations alter splice site recognition in hematological malignancies

Aravind Ramakrishnan[1,2†], Janine O. Ilagan[3,4†], Michele E. Murphy[1], Ahmad S. Zebari[3,4], Philip Bradley[3], and Robert K. Bradley[3,4*]

[1]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[2]Division of Medical Oncology, School of Medicine, University of Washington, Seattle, WA, USA

[3]Computational Biology Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[4]Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[†]These authors contributed equally to this work.

[*]Correspondence: rbradley@fhcrc.org

Running title: U2AF1 mutations alter splice site recognition

Keywords: myelodysplastic syndromes, acute myeloid leukemia, RNA splicing, U2AF1, U2AF35

ABSTRACT

Whole-exome sequencing studies have identified common mutations affecting genes encoding components of the RNA splicing machinery in hematological malignancies; however, the molecular consequences of these mutations are unknown. Here, we synthesize patient data, cell culture experiments, and structural modeling to systematically determine how mutations affecting the 3' splice site recognition factor U2AF1 alter its normal role in RNA splicing in myeloid malignancies. In contrast to initial reports that U2AF1 mutations cause loss of function and global splicing failure, we find that these U2AF1 mutations instead cause gain of function to promote or repress specific variants of the consensus 3' splice site. Mutations affecting the first and second zinc fingers give rise to different alterations in splice site preference, which influence the similarity of splicing programs in tumors. These allele-specific effects are consistent with a computationally predicted model of U2AF1 in complex with RNA. We created comprehensive global maps of differential splicing driven by each U2AF1 mutation, and use these maps to identify genes that are consistently affected by U2AF1 mutations both *in vivo* and in cell culture. Many such genes participate in molecular pathways that have been previously implicated in myeloid cancers, including DNA methylation (*DNMT3B*), X inactivation (*H2AFY*), and the DNA damage response (*ATR*). Our findings provide a comprehensive description of the mechanistic consequences of a spliceosomal gene mutation in cancer, and suggest that U2AF1 mutations may contribute to tumorigenesis by driving widespread quantitative changes in splicing that affect diverse cellular pathways.

INTRODUCTION

Myelodysplastic syndromes (MDS) represent a heterogeneous group of blood disorders characterized by dysplastic and ineffective hematopoiesis. Patients frequently suffer from cytopenias, and are at increased risk for disease transformation to acute myeloid leukemia (AML) (Tefferi and Vardiman 2009). The only curative treatment is hematopoietic stem cell transplantation, for which most patients are ineligible due to advanced age at diagnosis. The development of new therapies has been slowed by our incomplete understanding of the molecular mechanisms underlying the disease.

Recent sequencing studies of MDS patient exomes identified common mutations affecting genes encoding components of the RNA splicing machinery, with ~45-85% of patients affected (Yoshida et al. 2011; Papaemmanuil et al. 2011; Visconte et al. 2011; Graubert et al. 2011). Spliceosomal genes are the most common targets of somatic point mutations in MDS, suggesting that dysregulated splicing may constitute a common theme linking the disparate disorders that comprise MDS. Just four genes—*SF3B1*, *SRSF2*, *U2AF1*, and *ZRSR2*—carry the bulk of the mutations, which are mutually exclusive and occur in heterozygous contexts (Yoshida et al. 2011). Targeted sequencing studies identified high-frequency mutations in these genes in other hematological malignancies as well, including chronic myelomonocytic leukemia and AML with myelodysplastic features (Yoshida et al. 2011). Of the four commonly mutated genes, *SF3B1*, *U2AF1*, and *ZRSR2* encode 3' splice site recognition factors (Cvitkovic and Jurica 2012; Shen et al. 2010), suggesting that altered 3' splice site recognition is an important feature of the pathogenesis of MDS and related myeloid neoplasms. However, both the mechanistic consequences and downstream targets of spliceosomal gene mutations are unknown, hindering our understanding of how spliceosomal gene mutations contribute to disease initiation and progression.

U2AF1 (also known as U2AF35) may provide a useful model system for beginning to unravel the molecular consequences of MDS-associated spliceosomal gene mutations. U2AF1 mutations are highly specific—they uniformly affect the S34 and Q157 residues within the first and second CCCH zinc fingers of the protein—making comprehensive studies of all mutant alleles feasible (Figure 1A). Furthermore, U2AF1's biochemical role in binding the AG dinucleotide of the 3' splice site is relatively well-defined (Wu et al. 1999; Zorio and Blumenthal 1999; Merendino et al. 1999). U2AF1 preferentially recognizes the core RNA sequence motif

yAG|r, which matches the genomic consensus 3' splice site and intron|exon boundary that crosslinks with U2AF1 (Wu et al. 1999). Nevertheless, our understanding of U2AF1:RNA interactions is incomplete. U2AF1's U2AF homology motif (UHM) is known to mediate U2AF1:U2AF2 heterodimer formation (Kielkopf et al. 2001); however, both the specific protein domains that give rise to U2AF1's RNA binding specificity and the normal function of U2AF1's zinc fingers are unknown (Figure 1B). Accordingly, the precise mechanistic consequences of U2AF1 mutations are difficult to predict.

Since the initial reports of common U2AF1 mutations in MDS, the molecular and biological consequences of U2AF1 mutations have been controversial. An early study overexpressed mutant U2AF1 in HeLa cells and suggested that U2AF1 mutations cause loss of function, globally dysfunctional splicing marked by frequent inclusion of premature termination codons and intron retention, and cell cycle arrest (Yoshida et al. 2011), while another study expressed the mutant protein in 293T cells and found evidence of increased basal splicing but decreased cassette exon recognition based on minigene assays (Graubert et al. 2011). However, some recent studies have not identified widespread changes in splicing associated with these mutations in patient transcriptomes (Makishima et al. 2012; Lindsley and Ebert 2013b). Resolving this controversy has been hindered by the various systems used in different studies (e.g., ectopic expression in non-hematopoietic cells versus studies of specific genes in patient transcriptomes), as well as the absence of a comprehensive map of genome-wide splicing changes caused by U2AF1 mutations in hematopoietic cells.

Here, we systematically determined how MDS-associated mutations in U2AF1 affect the protein's normal function. U2AF1 mutations cause gain of function, with different U2AF1 mutations giving rise to distinct alterations in the consensus 3' splice site motif yAG|r recognized by U2AF1, both *in vivo* and in cell culture. These alterations in splice site preference result in genome-wide, yet specific, changes in both constitutive and alternative splicing. Our results resolve the controversy surrounding the molecular consequences of U2AF1 mutations, give insight into the normal function of U2AF1's poorly studied zinc finger domains, and provide a comprehensive map of splicing changes induced by each mutant allele that may contribute to disease.

RESULTS

Leukemias with U2AF1 mutations exhibit altered consensus 3' splice sites

Previous biochemical studies showed that U2AF1 recognizes the core sequence motif yAG|r of the 3' splice site (Wu et al. 1999; Zorio and Blumenthal 1999; Merendino et al. 1999). Accordingly, we hypothesized that U2AF1 mutations might result in preferential activation or repression of 3' splice sites in a sequence-specific manner. To test this hypothesis, we compared the global transcriptomes of U2AF1 wild-type (WT) and mutant *de novo* adult AML samples that were sequenced as part of TCGA, The Cancer Genome Atlas (Cancer Genome Atlas Research Network 2013). Seven of the 169 samples with available RNA-seq data carried a U2AF1 mutant allele. For each mutant U2AF1 sample, we enumerated all cassette exons that were differentially spliced between the sample and an average U2AF1 WT sample, requiring a minimum change in isoform ratio of 10% (Bayes factor $\geq$ 2.5). We identified ~300 differentially spliced cassette exons for each U2AF1 mutant sample, corresponding to ~2% of all ~15,000 cassette exons that were alternatively spliced in the AML samples and had sufficient read coverage for statistical analysis (1.9–3.6% of exons were differentially spliced, depending upon the particular sample, with a median of 2.1%). Exons whose inclusion was increased or decreased in U2AF1 mutant samples exhibited different consensus nucleotides at the -3 and +1 positions flanking the AG of the 3' splice site. As these positions correspond to the yAG|r motif bound by U2AF1, this data strongly supports our hypothesis that U2AF1 mutations alter 3' splice site recognition activity in a sequence-specific manner (Figure 1C, S1).

U2AF1-associated alterations in splice site preference are allele-specific

Mutations at different residues of U2AF1 were associated with distinct alterations in the consensus 3' splice site motif yAG|r of differentially spliced exons. S34F and S34Y mutations, affecting the first zinc finger, were associated with nearly identical alterations at the -3 position in all six S34 mutant samples; in contrast, the Q157P mutation in the second zinc finger was associated with alterations at the +1 position. These sequence preferences (C >> T at the -3 position for S34F/Y and G >> A at the +1 position for Q157P) differ markedly from the human 3' splice site consensus. C/T and G/A appear at similar frequencies at the -3 and +1 positions in the human genome, and both minigene and genomic studies of competing 3' splice sites indicate

that C and T are approximately equally effective at the -3 position (Smith et al. 1993; Bradley et al. 2012). Our finding that S34 mutations are associated with alterations in the preferred nucleotide at the -3 position is consistent with a recent study of a subset of the patients analyzed in our manuscript, in which the authors found increased/decreased preference for T at the -3 position for exons with increased exclusion/inclusion in mutant samples (Przychodzen et al. 2013).

Transgenic expression of mutant U2AF1 is sufficient to alter 3' splice site preferences

We next tested whether these alterations in 3' splice site preference are a direct consequence of U2AF1 mutations. We generated K562 erythroleukemic cell lines that stably expressed a single FLAG-tagged U2AF1 allele (WT, S34F, S34Y, or Q157R) at modest levels in the presence of the endogenous protein. This expression strategy, where the transgene was expressed at levels of 1.5-5.7X endogenous U2AF1, is consistent with the co-expression of WT and mutant alleles at approximately equal levels that we observed in AML transcriptomes (Figure 2A). A similar pattern of co-expression of WT and mutant alleles has been previously reported in MDS patients carrying U2AF1 mutations (Graubert et al. 2011). We separately knocked down (KD) endogenous U2AF1 to ~13% of normal protein levels in the absence of transgenic expression to test whether the mutations cause gain or loss of function (Figure 2B). To identify potential changes in 3' splice site preference, we performed deep RNA-seq (~100M 2x49 bp reads per sample) and compared alternative splicing of cassette exons in cells expressing transgenic mutant or WT alleles. This provided sufficient read coverage to test ~20,000 cassette exons that were alternatively spliced in K562 cells for differential splicing.

We found that 1.8–2.6% of alternatively spliced cassette exons were differentially spliced in cells expressing mutant versus WT U2AF1, similar to the frequencies that we observed in AML transcriptomes. These cassette exons that were promoted or repressed by each mutant allele exhibited sequence preferences at the -3 and +1 positions that were highly similar to those observed in AML patient samples (Figure 2C). Mutations affecting identical residues (S34F/Y and Q157P/R) caused similar alterations in 3' splice site preference, while mutations affecting different residues did not, confirming the allele-specific consequences of U2AF1 mutations. In contrast, cassette exons that were differentially spliced following KD of endogenous U2AF1 exhibited no sequence-specific changes in 3' splice site preference (Figure 2D). We therefore

conclude that S34 and Q157 mutations cause gain of function, in contrast to the loss of function suggested by previous studies (Yoshida et al. 2011; Makishima et al. 2012). Gain, rather than loss, of function is consistent with the empirical absence of inactivating (nonsense or frameshift) U2AF1 mutations observed in patients (Yoshida et al. 2011). Taken together, our data demonstrate that U2AF1 mutations cause allele-specific alterations in the preferred 3' splice site motif yAG|r both *in vivo* and in cell culture.

S34 mutations preferentially promote AG-dependent 3' splice sites

U2AF1 mutations alter 3' splice site preferences, yet only specific cassette exons are affected by expression of any mutant allele. In both AML transcriptomes and our K562 system, we consistently observed that approximately 2% of alternatively spliced cassette exons responded to U2AF1 mutations (e.g., exhibited a change in inclusion or exclusion ≥10%) in each sample. Previous biochemical studies have found that only a subset of exons have "AG-dependent" 3' splice sites that require U2AF1 binding for proper splice site recognition (Reed 1989; Wu et al. 1999). We therefore speculated that exons responsive to expression of U2AF1 mutant alleles might be unusually reliant upon U2AF1 recruitment for normal splicing. Supporting this hypothesis, the vast majority of cassette exons promoted by S34F expression decreased in inclusion upon U2AF1 KD in WT cells (Figure 3A). Defining exons that decreased in inclusion following U2AF1 KD as U2AF1-dependent, we observed an approximately six to eight-fold enrichment for overlap between S34F or S34Y-promoted exons and U2AF1-dependent exons (Figure 3B). We observed no similarly strong enrichment for Q157R-promoted exons. Therefore, S34, but not Q157, mutations preferentially promote strongly AG-dependent splice sites.

U2AF1 allele specificity influences tumor similarity

While U2AF1 mutations are frequently treated as functionally equivalent in clinical studies, our RNA-seq analyses indicate that mutations within the first and second zinc fingers give rise to distinct alterations in splice site preference both *in vivo* and in cell culture. We therefore hypothesized that these allele-specific effects might influence tumor similarity. We found that AML transcriptomes with S34F and S34Y mutations exhibited greater overlap in global splicing with each other than with the Q157P sample (Figure 3C). These results were replicated in our K562 system (Figure 3D). We accordingly conclude that the mechanistic observations described

above—that different U2AF1 alleles cause distinct alterations in splice site preference—are directly relevant to global splicing programs in tumors.

U2AF1 mutations cause broad changes in global splicing

We next assembled a comprehensive map of global changes in splicing driven by each U2AF1 mutant allele in our K562 system. We tested ~125,000 annotated alternative splicing events for differential splicing, and furthermore assayed ~160,000 constitutive splice junctions for evidence of novel alternative splicing or intron retention. The resulting catalogs of differentially spliced events (File S1-3) revealed that all major classes of alternative splicing events, including cassette exons, competing splice sites, and retained introns, were affected by U2AF1 mutations. The mutant alleles additionally drove changes in constitutive splicing, including alternative splicing and retention of constitutively spliced introns, although at lower rates (Figure 4A).

We used these global maps of differentially expressed isoforms to test whether U2AF1 mutation-induced alterations in 3' splice site preference influenced splicing events other than cassette exon recognition. With the exception of competing 5' splice sites—which are unlikely to be directly regulated by processes occurring at the 3' splice site—each identified class of splicing events exhibited allele-specific sequence preferences at the -3 and +1 positions consistent with our original findings in patient transcriptomes (Figure 4B). We conclude that U2AF1 mutations directly affect all classes of splicing events that are frequently regulated through 3' splice site recognition, thereby giving rise to the observed widespread changes in global splicing.

U2AF1 mutations do not cause global splicing dysfunction

A previous study reported that the S34F mutation caused broad splicing dysfunction, including overproduction of aberrant mRNAs slated for degradation and widespread intron retention, based on transgenic expression and RNA-seq experiments (Yoshida et al. 2011). These results are surprising, since presumably both tumors and healthy tissues require a competent splicing machinery. In contrast, another study used minigene reporters to find that S34F expression caused more efficient splice site recognition in a basal splicing reporter, but increased exon skipping in an alternative splicing reporter (Graubert et al. 2011). As these previous studies relied on acute expression in HeLa (Yoshida et al. 2011) and 293T cells (Graubert et al. 2011), it is unclear how their results might generalize to hematopoietic cells *in vivo*.

We accordingly tested whether splicing dysfunction occurs globally in AML transcriptomes with or without U2AF1 mutations. Quantifying expression of ~10,000 cassette exons whose inclusion or exclusion was predicted to trigger nonsense-mediated decay (NMD), we found that U2AF1 mutant samples did not exhibit globally increased expression of NMD-inducing isoforms relative to WT samples. Similarly, mutant and WT samples exhibited similar levels of constitutive intron retention and exon skipping (Figure 4C-D, S2-4). These findings replicated in our K562 system, in which no mutant allele gave rise to global increases in expression of NMD substrates, constitutive intron retention, or exon skipping. In contrast, KD of endogenous U2AF1 induced all three hallmarks of aberrant splicing, again consistent with the mutations causing gain, not loss, of function (Figure 4E-H, S5).

The differences between previously published results and our observations are likely due to differing experimental designs. Yoshida et al acutely (rather than stably) expressed the S34F allele at very high levels (50X WT levels) in HeLa cells, whereas we expressed each allele at modest levels (1.5-5.7X WT levels) in hematopoietic cells to mimic the co-expression of WT and mutant alleles that occurs in patients (Figure 2A). Maintaining a balance between WT and mutant allele expression is likely important to prevent broad splicing dysfunction in the face of U2AF1 mutations.

U2AF1 mutations likely modify U2AF1:RNA interactions

As U2AF1 mutations drive alterations in the preferred 3' splice site motif yAG|r—the same motif that is recognized and bound by U2AF1 (Wu et al. 1999)—we next investigated whether U2AF1 mutations might act by modifying U2AF1's RNA binding activity. U2AF1's RNA binding specificity could originate from its U2AF homology motif (UHM) and/or its two CCCH zinc fingers. The UHM mediates U2AF heterodimer formation and binds a consensus 3' splice site sequence with low affinity (Kielkopf et al. 2001), suggesting that the UHM is insufficient to generate U2AF1's sequence specificity. As U2AF1's zinc fingers are independently required for U2AF RNA binding (Webb and Wise 2004), and our data indicates that zinc finger mutations alter splice site preferences, we hypothesized that U2AF1's zinc fingers might directly interact with the 3' splice site.

To evaluate this hypothesis, we started from the experimentally determined structure of the UHM domain (Kielkopf et al. 2001), modeled the conformations of the zinc finger domains

bound to RNA by aligning them to the CCCH zinc finger domains in the TIS11d:RNA complex structure (Hudson et al. 2004), and sampled the conformations of the two short linker regions using fragment assembly techniques (Leaver-Fay et al. 2011). The RNA was built in two segments taken from the TIS11d complex, one anchored in the N-terminal zinc finger and one in the C-terminal finger. We modeled multiple 3' splice site sequences (primarily variants of uuAG|ruu), and explored a range of possible alignments of the 3' splice site within the complex. The final register was selected on the basis of energetic analysis and manual inspection using known features of the specificity pattern of the 3' splice site (in particular, the lack of a significant genomic consensus at the -4 and +3 positions, consistent with the experimental absence of a crosslink between U2AF1 and the -4 position (Wu et al. 1999)).

Based on these simulations, we propose a theoretical model of U2AF1 in complex with RNA wherein the zinc finger domains guide recognition of the yAG|r motif, consistent with the predictions of our mutational data. The model has the following features (Figure 5A, File S4). The first zinc finger contacts the bases immediately preceding the splice site, including the AG dinucleotide (Figure 5B-C), while the second zinc finger binds immediately downstream (Figure 5D). The RNA is kinked at the splice site and bent overall throughout the complex so that both the 5' and 3' ends of the motif are oriented toward the UHM domain and U2AF2 peptide. Contacts compatible with the 3' splice site consensus are observed at the sequence-constrained RNA positions. The mutated positions S34 and Q157 are nearby the bases at which perturbed splice site preferences are observed for their respective mutations. Moreover, the modified preferences can, to some extent, be rationalized by contacts seen in our simulations. S34 forms a hydrogen bond with U(-1), and preference for U at -1 appears to decrease upon mutation; the Q157P mutation would improve electrostatic complementarity with G at +1 by removing a backbone NH group, in agreement with increased G preference in this mutant. Our predicted model, in which the first zinc finger recognizes the AG dinucleotide, is also consistent with our finding that S34, but not Q157, mutations preferentially promote strongly AG-dependent splice sites.

U2AF1 mutations affect diverse cellular pathways

U2AF1 mutations likely contribute to disease initiation and progression through downstream mis-spliced effectors, rather than by altering the sequence specificity of 3' splice site recognition

per se. We accordingly sought to determine how U2AF1-induced splicing changes affected broader biological processes using the comprehensive maps of differential splicing described above.

Gene ontology analysis indicated that genes involved in the cell cycle, DNA repair, chromatin modification, methylation, and RNA processing pathways, among others, are enriched for differential splicing in K562 cells expressing mutant U2AF1. This could be due to high basal rates of alternative splicing within these genes, which frequently are composed of many exons, or instead caused by specific targeting by mutant alleles of U2AF1. Upon correcting for gene-specific variations in the number of possible alternatively spliced isoforms, these pathways were no longer enriched in gene ontology analyses. Therefore, U2AF1 mutations appear to broadly affect global splicing rather than specifically targeting a few biological pathways.

Identifying candidate downstream effectors of cancer-associated mutations is challenging, and prone to high false-positive rates. Candidates identified from tumor transcriptome sequencing are not necessarily direct targets of the mutation of interest, while candidates identified from transgene expression experiments may not prove relevant to *in vivo* disease processes. High levels of inter-tumoral variation in gene expression between even genetically similar tumors, as well as intrinsic biological stochasticity, further complicate the discovery process. A recent study reported that a patient carrying a U2AF1 mutation exhibited intron retention within the *TET2* gene (Makishima et al. 2012); however, we did not observe abnormal splicing of *TET2* in our experiments.

We sought to address the inherent difficulties of identifying candidate downstream effectors of U2AF1 mutations by leveraging our combination of patient data and transgene expression experiments. Our results indicate that S34 and Q157 mutations cause distinct splicing phenotypes, and so should be analyzed separately; accordingly, we considered only S34 mutations here, as only one AML sample carried a Q157 mutation. We identified events that were consistently differentially spliced in U2AF1 S34 mutant (N = 6) and WT (N = 162) samples ($p < 0.01$, Mann-Whitney U test; File S5), and intersected those with events that we previously identified as differentially spliced in our maps of splicing changes associated with S34F and S34Y expression (File S1-3). This approach was very conservative in the sense of requiring differential splicing in AML transcriptomes (to help ensure disease relevance) as well

as in our K562 system following expression of S34F and S34Y (to help restrict to direct targets of U2AF1 mutations).

We identified a total of 67 genes that satisfied our criteria as consistently differentially spliced in association with S34 mutations. The consistently differentially spliced genes that we identified encode proteins that participate in diverse cellular pathways (Table 1, File S6). For example, a cassette exon at the 3' end of *ATR* gene, which encodes a PI3K-related kinase that activates the DNA damage checkpoint, is included at high rates in AML patients carrying S34, but not Q157, mutations. Similar allele-specific exon inclusion occurs in K562 cells expressing transgenic U2AF1 (Figure 6A). This cassette exon alters the C terminus of the ATR protein, is predicted to subject the mRNA to degradation, and is highly conserved, suggesting that it is physiologically relevant (Figure 6B).

Suggestively, many genes identified in our analysis encode proteins that participate in biological pathways implicated in MDS pathogenesis. Genes encoding proteins involved in epigenetic regulation, including DNA methylation and chromatin modification, are the most common mutational targets in MDS after spliceosomal genes (Lindsley and Ebert 2013a). We observed consistent differential splicing in genes encoding the *de novo* methyltransferase DNMT3B (Figure 6C), the Polycomb group protein SCML2, and the histone modifiers EHMT1 and KDM3A, suggestive of a potential connection between U2AF1 mutations and epigenetic effects. Similarly, the gene *H2AFY*, encoding the core histone macro-H2A.1, was consistently differentially spliced in association with U2AF1 mutations (Figure 6D). Macro-H2A.1 is important for X chromosome inactivation (Hernández-Muñoz et al. 2005), and loss of X chromosome inactivation was recently shown to cause a MDS-like disease in mice (Yildirim et al. 2013), suggesting that U2AF1-driven differential splicing of macro-H2A.1 could potentially be relevant to disease processes.

Given our conservative approach, the number of differentially spliced genes that we have identified likely constitutes a lower bound. For example, three distinct events within *ASXL1* were differentially spliced in AML transcriptomes and our K562 system in association with U2AF1 mutations. This is intriguing, as *ASXL1* is a common mutational target in myelodysplastic syndromes and related disorders (Gelsi-Boyer et al. 2009); furthermore, *U2AF1* and *ASXL1* mutations co-occur more frequently than expected by chance (Thol et al. 2012). However, the altered splicing that we observed was always more extreme in one system than in the other

(Figure 6E), and no single *ASXL1* event passed our statistical thresholds in both the AML and K562 data. Therefore, while *ASXL1* is frequently differentially spliced in association with U2AF1 mutations, we cannot confidently identify a specific exon that consistently responds to the mutations. Accordingly, the genes listed in Table 1 probably constitute only a subset of the potentially important downstream targets of U2AF1 mutations.

DISCUSSION

Here, we have provided a comprehensive elucidation of the mechanistic consequences of a spliceosomal gene mutation in cancer, as well as a global map of splicing changes driven by U2AF1 mutations in hematopoietic cells. Our results indicate that U2AF1 mutations cause highly specific alterations in 3' splice site recognition in myeloid neoplasms, likely by modifying U2AF1's RNA binding activity. Taken together with the high frequency of mutations targeting U2AF1 and other 3' splice site recognition factors, our results strongly support the hypothesis that specific alterations in spliceosomal protein activity are important contributors to the molecular pathology of MDS and related hematological disorders.

Intriguingly, we observed consistent differential splicing of multiple genes such as *DNMT3B* that participate in molecular pathways previously implicated in MDS pathogenesis. It is tempting to speculate that differential splicing of a few such genes in well-characterized pathways explain how U2AF1 mutations drive disease. However, we instead hypothesize that spliceosomal mutations contribute to dysplastic hematopoiesis and tumorigenesis by dysregulating a multitude of genes involved in many aspects of cell physiology. This hypothesis is consistent with two notable features of our data. First, the splicing changes are widespread. Hundreds or thousands of exons are differentially spliced in response to U2AF1 mutations, belying the notion that just a few key downstream targets are affected by mis-splicing. Second, the splicing changes are relatively moderate. In both the AML and K562 data, we observed broad quantitative changes in splicing, but almost no "isoform switches," wherein a previously minor isoform becomes the major isoform (the *ATR* gene illustrated in Figure 6 is a notable exception, although the change is smaller in the K562 system). Therefore, we expect that specific targets such as *DNMT3B* may contribute to, but probably do not wholly explain, U2AF1 pathophysiology. As additional data from tumor sequencing and transgene expression experiments become available—for example, as more patients transcriptomes carrying Q157 mutations are sequenced—precisely identifying disease-relevant changes in splicing will become increasingly reliable.

Our understanding of the molecular consequences of U2AF1 mutations will also benefit from further experiments conducted during the differentiation process. Both the AML and K562 data arose from relatively "static" systems, in the sense that the bulk of the assayed cells were not actively undergoing lineage specification. U2AF1 mutations likely cause similar changes in

splice site recognition in both precursor and more differentiated cells, but altered splice site recognition could have additional consequences in specific cell types. A recent study reported that regulated intron retention is important for granulopoiesis (Wong et al. 2013), consistent with the idea that as-yet-unrecognized shifts in RNA processing may occur during hematopoiesis. By disrupting such global processes, altered splice site recognition could contribute to the ineffective hematopoiesis that characterizes MDS.

Relevance to future studies of spliceosomal mutations

Our mechanistic findings of altered splice site recognition contrast with initial reports that U2AF1 mutations result in loss of function and widely aberrant splicing. As described above, the discrepancies between previous transgene expression studies and our own are likely due to the use of non-hematopoietic versus hematopoietic cells, acute versus stable expression, and high versus moderate levels of expression. The concordance that we observe between analyses conducted using *in vivo* (AML patient) and *in vitro* (K562 cells) systems reinforces the importance of mimicking natural expression patterns, to the extent possible, in transgene experiments.

Expressing transgenes at relatively physiological levels is probably also important for phenotypic studies. Yoshida et al observed G2/M cell cycle arrest following expression of the S34F allele in their HeLa system (Yoshida et al. 2011). We performed cell cycle assays using propidium iodide and FACS in our K562 cells, and did not observe signs of aberrant cell cycle progression in cells expressing any mutant allele. We further tested the generality of these results by stably expressing the WT, S34F, S34Y, or Q157R alleles in the TF-1 erythroleukemic cell line and performing cell cycle assays, whereupon we again did not observe G2/M arrest or another obvious cell cycle defect (Figure S6). As with our studies of splicing, this difference in cell cycle phenotype is likely due to the high expression used by Yoshida et al versus the moderate levels that we used. Attempting to mimic normal expression levels in patients will probably prove relevant to studies of mutations in spliceosomal genes other than *U2AF1* as well.

Both mechanistic and phenotypic studies of cancer-associated somatic mutations frequently focus on single mutant alleles, even when multiple distinct mutations affecting that gene occur at high rates. Similarly, distinct mutations affecting the same gene are frequently grouped together in prognostic and other clinical studies, thereby implicitly assuming that

different mutations have similar physiological consequences. Our finding that different U2AF1 mutations are not mechanistically equivalent illustrates the value of comprehensive studies of all high-frequency mutant alleles, when feasible. The distinctiveness of S34 and Q157 mutation-induced alterations in 3' splice site preference suggests that they may constitute clinically relevant disease subtypes, potentially contributing to the heterogeneity of MDS. Mutations affecting other spliceosomal genes may likewise have allele-specific consequences. For example, mutations at codons 625 versus 700 of SF3B1 are most commonly associated with uveal melanoma (Harbour et al. 2013; Martin et al. 2013) versus MDS (Yoshida et al. 2011; Graubert et al. 2011; Papaemmanuil et al. 2011; Visconte et al. 2011) and chronic lymphocytic leukemia (Quesada et al. 2011). Accordingly, stratifying patients by allele may prove fruitful for both mechanistic and clinical studies of spliceosomal gene mutations in diverse disorders.

Our study additionally illustrates how investigating disease-associated somatic mutations can give insight into the normal function of proteins. The domains responsible for U2AF1's RNA binding specificity have remained elusive since U2AF1's biochemical role in AG-dependent splicing was elucidated almost 15 years ago. With a fairly restricted set of assumptions, we computationally predicted a family of models in which the first zinc finger is consistently responsible for recognizing the AG dinucleotide. As a computational prediction, the model must be tested with future experiments. Nonetheless, given the concordance between our theoretical model of U2AF1:RNA interactions and our mutational data, we speculate that this model will provide a useful framework for future biochemical studies of U2AF1 function in both healthy and diseased cells.

METHODS

Vector construction

A plasmid encoding U2AF1 cDNA (NCBI identifier NM_006758) was purchased from Open Biosystems and used as a template to generate constructs encoding U2AF1 + Gly Gly + FLAG, which were then cloned into the BamH1/Xho1 sites of pUB6/V5-His A vector (Invitrogen). Site-directed mutagenesis with the Phusion polymerase was used to generate constructs encoding the S34F, S34Y and Q157R alleles. Several PCR amplifications were then performed to generate bicistronic constructs of the form U2AF1 + Gly Gly + FLAG + T2A + mCherry (T2A is the cleavage sequence EGRGSLLTCGDVEENPGP). These inserts were then cloned into the BamH1/Sal1 sites of the self-inactivating lentiviral vector pRRLSIN.cPPT.PGK-GFP.WPRE (Addgene Plasmid 12252). The resulting plasmids co-express U2AF1 and mCherry under control of the PGK promoter.

Viral infection and cell culture

293T cells were maintained in DMEM supplemented with 10% fetal calf serum (FCS). To generate viral supernatant, lentiviral vectors were co-transfected into 293T cells along with the packaging vector PsPAX2 (Addgene plasmid 12260) and envelope vector pMD2.G (Addgene plasmid 12259) using the calcium phosphate method. Viral supernatants were harvested at 48 and 72 hours after transfection, filtered through a 0.45 um filter and concentrated by centrifugation at 5000g for 24 hours. K562 erythroleukemia cells were grown in RPMI-1640 supplemented with 10% FCS. To generate stable cell lines, one million K562 cells were resuspended in growth media supplemented with 8 ug/mL protamine sulfate and infected with concentrated lentiviral supernatant. Cells were then expanded and transduced cells expressing mCherry were isolated by fluorescence activated cell sorting (FACS) using a Becton Dickinson FACS Aria II equipped with a 561 nm laser. TF-1 cells were grown in HPGM supplemented with 1 ng/mL of GM-CSF, and stable cell lines were generated using a similar procedure as for K562 cells. For RNAi studies, K562 cells were transfected with a control (non-targeting) siRNA (Dharmacon D-001810-03-20) or a siRNA pool against U2AF1 (Dharmacon ON-TARGETplus SMARTpool L-012325-01-0005) using the Nucleofector II device from Lonza with the Cell Line

Nucleofector Kit V (program T16), and RNA and protein were collected 48 hours after transfection.

Cell cycle analysis

Approximately one million K562 or TF-1 cells expressing either WT or mutant U2AF1 were pelleted and resuspended in 500 uL of hypotonic lysis buffer (0.1 % Triton X-100, 1 mg/mL sodium citrate, 50 ug/mL propidium iodide, 500 ng/mL RNase A). Cells were kept in the dark and incubated on ice for about 2 hours and then at room temperature for 15 minutes, after which they were returned to ice for data collection. FACS analysis was performed using a Becton Dickinson LSRII flow cytometer.

mRNA sequencing

Total RNA was obtained by lysing 10 million K562 cells for each sample in Trizol and RNA was extracted using Qiagen RNA easy columns. Using 4 ug of total RNA, we prepared poly(A)-selected, unstranded libraries for Illumina sequencing using a modified version of the TruSeq protocol. After adapter ligation, AMPure XP Beads were used to select 100 – 400 bp DNA fragments by varying bead-to-library volume ratios. 0.5X beads were added to the sample library to select for fragments < 400 bp followed by 1X beads to select for > 100 bp fragments. DNA fragments were amplified using 15 cycles of PCR and separated by 2% agarose gel electrophoresis. DNA fragments (300 bp) were purified using the Qiagen MinElute gel extraction kit. RNA-seq libraries were then sequenced on the Illumina HiSeq 2000 to a depth of approximately 100 million 2x49 bp reads per sample.

Accession numbers

For the AML analysis, BAM files were downloaded from CGHub ("LAML" project) and converted to FASTQ files of unaligned reads for subsequent read mapping. For the HeLa cell analysis, FASTQ files were downloaded from DDBJ series DRA000503, and the reads were trimmed to 50 bp (after removing the first five bp) to restrict to the high-quality portion of the sequencing reads. A similar trimming procedure was performed in the original manuscript (Yoshida et al. 2011).

Genome annotations

MISO v2.0 annotations were used for cassette exon, competing 5' and 3' splice sites, and retained intron events (Katz et al. 2010). Constitutive junctions were defined as splice junctions that were not alternatively spliced in any isoform of the UCSC knownGene track (Meyer et al. 2013). For read mapping purposes, the following specific annotation files were created. A gene annotation file was created by combining isoforms from the MISO v2.0 (Katz et al. 2010), UCSC knownGene (Meyer et al. 2013), and Ensembl 71 (Flicek et al. 2013) annotations, and a splice junction annotation file was created by enumerating all possible combinations of annotated splice sites as previously described (Hubert et al. 2013).

RNA-seq read mapping

Reads were mapped to the UCSC hg19 (NCBI GRCh37) human genome assembly using Bowtie (Langmead et al. 2009), RSEM (Li and Dewey 2011), and TopHat (Trapnell et al. 2009). RSEM v1.2.4 was modified to call Bowtie v1.0.0 with the -v 2 mapping strategy. RSEM was then invoked with the arguments --bowtie-m 100 --bowtie-chunkmbs 500 --calc-ci --output-genome-bam on the gene annotation file. The resulting BAM file was then filtered to remove alignments with mapq scores of 0 and require a minimum splice junction overhang of 6 bp. Unaligned reads were then aligned with TopHat v2.0.8b with the arguments --bowtie1 --read-mismatches 2 --read-edit-dist 2 --no-mixed --no-discordant --min-anchor-length 6 --splice-mismatches 0 --min-intron-length 10 --max-intron-length 1000000 --min-isoform-fraction 0.0 --no-novel-juncs --no-novel-indels --raw-juncs on the splice junction file, with --mate-inner-dist and --mate-std-dev determined by mapping to constitutive coding exons as determined with MISO's exon_utils.py script. The resulting alignments were then filtered as described and merged with RSEM's results to generate a final BAM file.

Isoform expression measurements

MISO (Katz et al. 2010) and v2.0 of its annotations were used to quantify isoform ratios for all cassette exons, competing 5' and 3' splice sites, and retained introns. Alternative splicing of constitutive junctions and retention of constitutive introns was quantified in an unbiased manner as previously described (Hubert et al. 2013). All analyses were restricted to splicing events with at least 20 relevant reads (reads supporting either or both isoforms) that were alternatively

spliced in our data. Events were defined as differentially spliced between two samples if they satisfied the following criteria: (1) at least 20 relevant reads in both samples, (2) a change in isoform ratio of at least 10%, and (3) a Bayes factor greater than or equal to 2.5 (AML data) or 5 (K562 data). A more relaxed Bayes factor was used for the AML data to compensate for the lower read coverage of the dataset, which reduces the power of statistical significance testing. Wagenmakers's framework (Wagenmakers et al. 2010) was used to compute Bayes factors for differences in isoform ratios between samples.

AML WT and mutant comparisons

To identify splicing events that were differentially spliced in each AML sample with a U2AF1 mutation, each U2AF1 mutant sample was compared to an average U2AF1 WT sample. The average U2AF1 WT sample was created by averaging isoform ratios over all 162 U2AF1 WT samples.

Sequence logos

Sequence logos were created with v1.26.0 of the seqLogo package in Bioconductor (Gentleman et al. 2004).

Gene ontology enrichment analysis

Gene ontology analysis was performed with GOseq (Young et al. 2010). To identify enriched pathways, the set of all genes that were differentially spliced in K562 cells expressing a mutant versus WT allele of U2AF1 was used as input to GOseq with the "Hypergeometric" method. This identified the cell cycle, DNA repair, chromatin modification, methylation, and RNA processing pathways as enriched with a maximum false discovery rate of 0.01. However, this analysis did not take into account the varying frequency of alternative splicing in different genes. To take this into account, GOseq with called with a bias correction defined for each gene as Σ (geometric mean of number of relevant reads), where the sum is taken over all splicing events annotated for that gene. This bias correction takes into account the inherent bias for detecting alternative splicing within a gene with many exons, high levels of transcription, etc. After incorporating that bias correction, the previously identified pathways were no longer enriched,

indicating that U2AF1 mutations do not preferentially target specific groups of genes beyond those that are frequently alternatively spliced.

Western blotting

Protein lysates from K562 cells pellets were generated by resuspension in RIPA buffer and protease inhibitor along with sonication. Protein concentrations were determined using the Bradford protein assay. 10 ug of protein was then subjected to SDS-PAGE and subsequently transferred to nitrocellulose membranes. Membranes were blocked with 5% milk in Tris-buffered saline (TBS) for 1 hour at room temperature and then incubated with primary antibody 1:1000 anti-U2AF1 (Bethyl Laboratories), anti-FLAG (Thermo), or anti-α-tubulin (Sigma) for 1 hour at room temperature. Blots were washed with TBS containing 0.005% Tween 20 and then incubated with the appropriate secondary antibody for 1 hour at room temperature.

Protein structure prediction

Models of U2AF1 (residues 9-174) in complex with a RNA fragment extending from the 3' splice site positions -4 to +3 were built by combining template-based modeling, fragment assembly methods, and all-atom refinement. Models were built using the software package Rosetta (Leaver-Fay et al. 2011) with template coordinate data taken from the UHM:ULM complex structure (Kielkopf et al. 2001) (PDB ID 1jmt: residues A/46-143) and the TIS11d:RNA complex structure (Hudson et al. 2004) (PDB ID 1rgo: U2AF1 residues 16-37 mapped to A/195-216; residues 155-174 mapped to A/159-179; RNA positions -4 to -1 mapped to D/1-4; RNA positions +1 to +3 mapped to D/7-9). The remainder of the modeled region (residues 9-15, 38-45, and 144-154) was built using fragment assembly (with templated regions held fixed) in a low-resolution representation (backbone heavy atoms and side chain centroids) and force field. The fragment assembly simulation consisted of 6000 fragment-replacement trials, for which fragments of size 6 (trials 1-3000), 3 (trials 3001-5000), and 1 (trials 5001-6000) were used. The RNA was modeled in two pieces, one anchored in the N-terminal zinc finger and the other in the C-terminal zinc finger, with docking geometries taken from the TIS11d:RNA complex. A pseudo-energy term favoring chain closure was added to the potential function to reward closure of the chain break between the RNA fragments. The fragment assembly simulation was followed by all-atom refinement during which all side chains as well as the non-

templated protein backbone and the RNA were flexible. Roughly 100,000 independent model building simulations were conducted, each with a different random number seed and using a randomly selected member of the 1rgo NMR ensemble as a template. Low-energy final models were clustered to identify frequently sampled conformations (the model depicted in Figure 5A was the center of the largest cluster). We explored a range of possible alignments of the splice site RNA within the complex, with the final model selected on the basis of all-atom energies, RNA chain closure, manual inspection, and known sequence features of the 3' splice site motif.

DATA ACCESS

The reported RNA-seq data from K562 cells will be deposited into the NCBI GEO database upon manuscript acceptance.

ACKNOWLEDGEMENTS

We thank Beverly Torok-Storb for project assistance and advice, and Sue Biggins, Toshi Tsukiyama, and members of the Bradley lab for comments on the manuscript. This research was supported by the Hartwell Innovation Fund (RKB, AR), Damon Runyon Cancer Research Foundation DFS 04-12 (RKB), NIH/NCI P30 CA015704 recruitment support (RKB), Fred Hutchinson Cancer Research Center institutional funds (RKB), NIH/NCI training grant T32 CA009657 (JOI), NIH/NIDDK P30 DK056465 pilot study (JOI), NIH/NHLBI U01 HL099993 (AR), NIH/NIDDK K08 DK082783 (AR), the J.P. McCarthy Foundation (AR), the Storb Foundation (AR), and NIH/NIGMS R01 GM088277 (PB).

AUTHOR CONTRIBUTIONS

AR designed the cell culture and transgene expression strategies. JOI created Illumina libraries and performed biochemistry. AR, JOI, MEM, and ASZ performed experimental work, including cloning, cell culture, and flow cytometry. RKB and PB performed computational analyses, prepared figures, and wrote the manuscript, with contributions from other authors. RKB and AR initiated the study.

DISCLOSURE DECLARATION

The authors declare that no competing interests exist.

FIGURE LEGENDS

Figure 1. U2AF1 mutations are associated with altered 3' splice site preferences in AML transcriptomes.

(A) U2AF1 domain structure (UniProt Consortium 2012; Kielkopf et al. 2001) and common mutations. CCCH, CCCH zinc finger.

(B) Schematic of U2AF1 interaction with the 3' splice site of a cassette exon (black).

(C) Consensus 3' splice sites of cassette exons with increased, unchanged, or decreased inclusion in U2AF1 mutant relative to WT samples. Boxes highlight sequence preferences at the -3 and +1 positions that differ from the normal 3' splice site consensus. Vertical axis, information content in bits. N, number of cassette exons in each category with sufficient read coverage for analysis. Data for all U2AF1 mutant samples shown in Figure S1.

Figure 2. Transgenic expression of mutant U2AF1 alters 3' splice site preferences in U2AF1 WT cells.

(A) Mutant allele expression as a percentage of total U2AF1 mRNA in AML transcriptomes (left), our K562 cells (center), and Yoshida et al's HeLa cells (right). AML patient IDs indicated on horizontal axis. Numbers above bars indicate the ratio of mutant to WT allele expression (c.f., 1.5-5.7X in the K562 cells versus 50.2X in the HeLa cells).

(B) Western blots showing levels of transgenic FLAG-tagged and total U2AF1 in K562 cells stably expressing the indicated alleles (top) and levels of endogenous U2AF1 in K562 cells following transfection with a non-targeting siRNA or a siRNA pool against U2AF1 (bottom).

(C) Consensus 3' splice sites of cassette exons with increased, unchanged, or decreased inclusion in cells expressing transgenic mutant versus WT alleles of U2AF1. Compare to Figure 1C.

(D) Consensus 3' splice sites of cassette exons with increased, unchanged, or decreased inclusion in K562 cells transfected with a siRNA pool against U2AF1 relative to a cells transfected with a non-targeting siRNA.

Figure 3. The consequences of U2AF1 mutations are allele-specific.

(A) Density plot of change in cassette exon inclusion following U2AF1 KD for all alternatively spliced cassette exons (gray) and S34F-promoted cassette exons (red). Most cassette exons

promoted by S34F are repressed by U2AF1 KD, indicating that they are U2AF1-dependent. As a density plot (normalized histogram), the vertical axis is arbitrary and so not shown.
(B) Overlap between cassette exons that are promoted or repressed by mutant allele expression (rows) and U2AF1 KD (columns). Third column indicates the enrichment for U2AF1 dependence, defined as the overlap between exons promoted by mutant allele expression and exons repressed versus promoted by U2AF1 KD. S34, but not Q157, mutations preferentially promote U2AF1-dependent exons (bold).
(C) Overlap between cassette exons promoted (left) and repressed (right) in patients with U2AF1 mutations relative to WT patients. Percentages measure similarity between patients as the fraction of promoted/repressed cassette exons shared only between each pair of patients.
(D) As (C), but for our K562 system.

Figure 4. U2AF1 mutations cause global splicing changes, but not splicing failure.
(A) Splicing events that are differentially spliced in cells expressing transgenic S34F versus WT U2AF1. Pie chart illustrates the distribution of differentially spliced events among different classes of splicing events. Percentages specify the fraction of alternatively spliced events of each class that are affected by S34F expression.
(B) Presence or absence of altered 3' splice site preferences C >> T (S34F/Y) and G >> A (Q157R) at the -3 and +1 positions for different classes of splicing events affected by mutant U2AF1 expression.
(C) Global levels of NMD substrates in AML transcriptomes. U2AF1 mutant transcriptomes (red) do not exhibit higher levels of NMD substrates, even with the conservative comparison to the average WT sample. Distance from the center measures the splicing dissimilarity (restricted to NMD-inducing cassette exons) between each AML transcriptome and the average of all U2AF1 WT samples. Numbers, patient IDs.
(D) Global levels of retained constitutive introns in AML transcriptomes. U2AF1 mutant transcriptomes (red) do not exhibit increased constitutive intron retention.
(E) Levels of NMD-inducing isoforms of cassette exon events in K562 cells expressing transgenic WT or S34F U2AF1. S34F expression does not result in higher levels of NMD substrates. N, number of cassette exons events for which the NMD-inducing isoform is increased (red) or decreased (blue) in S34F-expressing cells. Percentages specify the fraction of NMD-

inducing alternatively spliced cassette exon events that are affected by S34F expression. Events that do not change are rendered transparent.

(F) Levels of properly spliced constitutive introns in K562 cells expressing transgenic WT or S34F U2AF1. S34F expression does not result in increased constitutive intron retention; instead, constitutive splicing is moderately more efficient in cells expressing transgenic S34F versus WT U2AF1.

(G) Levels of NMD-inducing isoforms of cassette exon events in K562 cells following control KD or U2AF1 KD.

(H) Levels of properly spliced constitutive introns in K562 cells following control KD or U2AF1 KD.

Figure 5. Theoretical model of the U2AF1:RNA complex.

(A) Overview, with the zinc finger domains colored cyan, the RNA in salmon, and the UHM beta sheet in blue and alpha helices in red. The frequently mutated positions S34 and Q157 are shown in stick representation. ZF, zinc finger.

(B-D) Interactions with individual bases characteristic of the 3' splice site consensus. Green dotted lines indicate hydrogen bonds and favorable electrostatic interactions; RNA and selected side chains are shown in stick representation.

Figure 6. U2AF1 mutations affect diverse cellular processes.

(A) Inclusion of cassette exon in the *ATR* gene in AML transcriptomes and K562 cells expressing transgenic U2AF1. Error bars, 95% confidence intervals.

(B) Cassette exon at the 3' end of the *ATR* gene. Conservation is phastCons (Siepel et al. 2005) track from UCSC (Meyer et al. 2013). Red stop sign, stop codon.

(C-E) Inclusion of cassette exons in the *DNMT3B*, *H2AFY*, and *ASXL1* genes.

SUPPORTING INFORMATION

Figure S1. U2AF1 mutations are associated with altered 3' splice site preferences in AML transcriptomes.

As Figure 1C, but for all seven samples with U2AF1 mutations (including those shown in Figure 1).

Figure S2. U2AF1 mutations are not associated with increased levels of NMD substrates in AML transcriptomes.

Plot illustrates the relative levels of NMD-inducing isoforms of alternatively spliced cassette exon events in each AML sample, with samples ordered by increasing level of NMD substrates. For each sample, all NMD-inducing isoforms that were increased or decreased ≥10% relative to the median over all samples were identified, and the quantity 100 x (# increased - # decreased) / (# increased + # decreased) was plotted on the vertical axis. Therefore, a positive value indicates a global increase in levels of NMD substrates, and vice versa. Samples with U2AF1 mutations (black) do not exhibit higher levels of NMD substrates than do samples without U2AF1 mutations (gray). Numbers above bars indicate the number of differentially expressed events for each sample.

Figure S3. U2AF1 mutations are not associated with increased retention of constitutive introns in AML transcriptomes.

Plot illustrates the relative levels of properly spliced out constitutive introns in each AML sample, with samples ordered by increasing level of proper splicing (intron removal). For each sample, all constitutive introns with evidence of increases/decreases in splicing ≥10% relative to the median over all samples were identified, and the quantity 100 x (# increased - # decreased) / (# increased + # decreased) was plotted on the vertical axis. Therefore, a positive value indicates a global increase in properly spliced constitutive introns, and vice versa. While retention of constitutive introns is common in AML transcriptomes—most bars are below 0—samples with U2AF1 mutations (black) do not exhibit higher levels of constitutive intron retention than do samples without U2AF1 mutations (gray). Numbers above bars indicate the number of differentially retained constitutive introns for each sample; the plot is restricted to these events

(e.g., the vast majority of constitutive introns are never retained in any sample, and those introns are not analyzed here since the plot is restricted to events that differ between samples).

Figure S4. U2AF1 mutations are not associated with increased exon skipping in AML transcriptomes.

Plot illustrates the relative levels of inclusion of alternatively spliced cassette exons that are NMD-irrelevant in each AML sample, with samples ordered by increasing level of exon inclusion. For each sample, all NMD-irrelevant cassette exons whose inclusion was increased or decreased ≥10% relative to the median over all samples were identified, and the quantity 100 x (# increased - # decreased) / (# increased + # decreased) was plotted on the vertical axis. Therefore, a positive value indicates a global increase in cassette exon inclusion, and vice versa. Samples with U2AF1 mutations (black) do not exhibit higher levels of cassette exon skipping than do samples without U2AF1 mutations (gray). Numbers above bars indicate the number of differentially expressed events for each sample; the plot is restricted to these events. NMD-irrelevant cassette exons are defined as events for which either both or neither of the inclusion and exclusion isoforms are NMD substrates. The plot is restricted to NMD-irrelevant events to distinguish exon inclusion from NMD.

Figure S5. U2AF1 KD, but not U2AF1 mutant allele expression, induces dysfunctional splicing.

(A) Levels of NMD-inducing isoforms of cassette exon events in K562 cells expressing transgenic WT or mutant U2AF1, or transfected with a control siRNA or siRNA pool against U2AF1.

(B) Levels of properly spliced constitutive introns in K562 cells expressing transgenic WT or mutant U2AF1, or transfected with a control siRNA or siRNA pool against U2AF1.

(C) Levels of exon inclusion for NMD-irrelevant cassette exons in K562 cells expressing transgenic WT or mutant U2AF1, or transfected with a control siRNA or siRNA pool against U2AF1.

Figure S6. Expression of U2AF1 mutant alleles is not associated with obvious cell cycle defects.

Cell cycle analyses of (A) K562 cells, or (B) TF-1 cells expressing transgenic WT or mutant U2AF1. Horizontal axis is level of propidium iodide as measured by Peridinin Chlorophyll Protein Complex.

File S1. Differentially spliced events in K562 cells expressing S34F.
Splicing events that are differentially spliced in K562 cells expressing transgenic S34F versus WT U2AF1. Each row of the table corresponds to isoform 1 of a splicing event, where isoform 1 is defined as follows: inclusion isoform for cassette exons ("se"), most intron-proximal isoform for competing 5' and 3' splice sites ("a5ss", "a3ss"), inclusion of upstream exon for mutually exclusive exons ("mxe"), splicing of retained introns annotated as alternative ("ri") or constitutive ("ci"), and canonical splicing of constitutive junction ("cj"). Each row is assigned a unique identifier specifying the event type and coordinates of the upstream junctions for isoforms 1 and 2 of the event; this event identifier format is a modification of the format used by MISO (Katz et al. 2010). The columns of the table are defined as follows: "coords", genomic coordinates containing the event; "spliceSites", dinucleotides at the 5' and 3' splice sites of the upstream junction of isoform 1; "nmdTarget", whether the specified isoform is a predicted NMD target, where a value of "NA" indicates that the event is not NMD relevant (e.g., neither or both isoforms are predicted substrates for NMD); "deltaPsi", change in isoform ratio between the two samples; "bayesFactor", Bayes factor associated with the sample comparison; "gene", gene ID; "geneName", gene name; "geneDescription", gene description. Gene IDs, names, and descriptions are from Ensembl, when available.

File S2. Differentially spliced events in K562 cells expressing S34Y.
As File S1, but for S34Y expression.

File S3. Differentially spliced events in K562 cells expressing Q157R.
As File S1, but for Q157R expression.

File S4. Theoretical model of the U2AF1:RNA complex.
Computationally predicted model encompassing 3' splice site residues -4 to +3 built using fragment assembly. Multi-model PDB file contains the center (model 1) and 19 randomly

selected members (models 2-20) of the largest cluster after structure-based comparison and clustering of all low-energy models.

File S5. Differentially spliced events in AML samples with S34 mutations.
As File S1, but for the AML data, based on comparisons between samples with S34F or S34Y mutations and samples with no U2AF1 mutations. The sample with a Q157 mutation was ignored for this analysis. Here, the "deltaPsi" column specifies the difference in isoform ratio for the medians of the two sample groups, and "bayesFactor" is replaced by "pval", as *p*-values are computed using the Mann-Whitney U test for group comparisons.

File S6. Consistently differentially spliced events in AML samples and K562 cells.
Events that are differentially spliced in AML samples with S34 mutations (File S5) as well as K562 cells expressing S34F or S34Y (File S1-2).

TABLES

name	description
AL589743.1	
AP2M1	adaptor-related protein complex 2, mu 1 subunit
ATAD3B	ATPase family, AAA domain containing 3B
ATF2	activating transcription factor 2
ATG13	autophagy related 13
ATR	ataxia telangiectasia and Rad3 related
BPTF	bromodomain PHD finger transcription factor
BUB3	BUB3 mitotic checkpoint protein
CCDC53	coiled-coil domain containing 53
CHCHD7	coiled-coil-helix-coiled-coil-helix domain containing 7
CHKB	choline kinase beta
CLIP1	CAP-GLY domain containing linker protein 1
CNOT2	CCR4-NOT transcription complex, subunit 2
DIS3	DIS3 mitotic control homolog (*S. cerevisiae*)
DLG1	discs, large homolog 1 (*Drosophila*)
DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta
EHMT1	euchromatic histone-lysine N-methyltransferase 1
FAM60A	family with sequence similarity 60, member A
FCGR2A	Fc fragment of IgG, low affinity IIa, receptor (CD32)
FGFR1OP2	FGFR1 oncogene partner 2
FXR1	fragile X mental retardation, autosomal homolog 1
GNAS	GNAS complex locus
GTF2I	general transcription factor IIi
GUSB	glucuronidase, beta
H2AFY	H2A histone family, member Y
KDM3A	lysine (K)-specific demethylase 3A
LUC7L	LUC7-like (*S. cerevisiae*)
MAP3K3	mitogen-activated protein kinase 3
MEF2C	myocyte enhancer factor 2C
MFF	mitochondrial fission factor
MLLT10	myeloid/lymphoid or mixed-lineage leukemia (*trithorax* homolog, *Drosophila*)
MRRF	mitochondrial ribosome recycling factor
MTA1	metastasis associated 1
MTHFSD	methenyltetrahydrofolate synthetase domain containing
MUM1	melanoma associated antigen (mutated) 1
MYNN	myoneurin
PABPC4	poly(A) binding protein, cytoplasmic 4 (inducible form)
PACRGL	PARK2 co-regulated-like
PICALM	phosphatidylinositol binding clathrin assembly protein
PIWIL4	piwi-like 4 (Drosophila)
PLEKHM2	pleckstrin homology domain containing, family M (with RUN domain) member 2
POLB	polymerase (DNA directed), beta
PPHLN1	periphilin 1
PRRC2C	proline-rich coiled-coil 2C
RAB6A	RAB6A, member RAS oncogene family
RALGDS	ral guanine nucleotide dissociation stimulator
RBM3	RNA binding motif (RNP1, RRM) protein 3
RIPK2	receptor-interacting serine-threonine kinase 2
RNF216	ring finger protein 216
RNF34	ring finger protein 34, E3 ubiquitin protein ligase
RP11-464F9.1	
RWDD1	RWD domain containing 1
SCARB1	scavenger receptor class B, member 1
SCML2	sex comb on midleg-like 2 (*Drosophila*)
SEC31A	SEC31 homolog A (*S. cerevisiae*)
SECISBP2	SECIS binding protein 2
SETD4	SET domain containing 4
SLC30A6	solute carrier family 30 (zinc transporter), member 6
SPAG5	sperm associated antigen 5
SRP19	signal recognition particle 19kDa
SUN2	Sad1 and UNC84 domain containing 2
TAF1D	TATA box binding protein (TBP)-associated factor, RNA polymerase I, D, 41kDa
TBC1D5	TBC1 domain family, member 5
TNRC6A	trinucleotide repeat containing 6A
TPD52L2	tumor protein D52-like 2
USP33	ubiquitin specific peptidase 33
ZFAND1	zinc finger, AN1-type domain 1

Table 1. Genes that are differentially spliced in association with U2AF1 mutations.

Genes that contain events that are differentially spliced in both the AML data (S34 mutant versus WT samples) and K562 data (S34 mutant versus WT sample). Descriptions taken from Ensembl.

REFERENCES

Bradley RK, Merkin J, Lambert NJ, Burge CB. 2012. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biol* **10**: e1001229.

Cancer Genome Atlas Research Network. 2013. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**: 2059–2074.

Cvitkovic I, Jurica MS. 2012. Spliceosome Database: a tool for tracking components of the spliceosome. *Nucleic Acids Res*.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. *Nucleic Acids Res* **41**: D48–55.

Gelsi-Boyer V, Trouplin V, Adélaïde J, Bonansea J, Cervera N, Carbuccia N, Lagarde A, Prébet T, Nezri M, Sainty D, et al. 2009. Mutations of polycomb-associated gene ASXL1 in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *Br J Haematol* **145**: 788–800.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80.

Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, Krysiak K, Harris CC, Koboldt DC, Larson DE, et al. 2011. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat Genet*.

Harbour JW, Roberson EDO, Anbunathan H, Onken MD, Worley LA, Bowcock AM. 2013. Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma. *Nat Genet*.

Hernández-Muñoz I, Lund AH, van der Stoop P, Boutsma E, Muijrers I, Verhoeven E, Nusinow DA, Panning B, Marahrens Y, van Lohuizen M. 2005. Stable X chromosome inactivation involves the PRC1 Polycomb complex and requires histone MACROH2A1 and the CULLIN3/SPOP ubiquitin E3 ligase. *Proc Natl Acad Sci USA* **102**: 7635–7640.

Hubert CG, Bradley RK, Ding Y, Toledo CM, Herman J, Skutt-Kakaria K, Girard EJ, Davison J, Berndt J, Corrin P, et al. 2013. Genome-wide RNAi screens in human brain tumor isolates reveal a novel viability requirement for PHF5A. *Genes Dev* **27**: 1032–1045.

Hudson BP, Martinez-Yamout MA, Dyson HJ, Wright PE. 2004. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nat Struct Mol Biol* **11**: 257–264.

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**: 1009–1015.

Kielkopf CL, Rodionova NA, Green MR, Burley SK. 2001. A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. *Cell* **106**: 595–605.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, et al. 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Meth Enzymol* **487**: 545–574.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

Lindsley RC, Ebert BL. 2013a. Molecular pathophysiology of myelodysplastic syndromes. *Annu Rev Pathol* **8**: 21–47.

Lindsley RC, Ebert BL. 2013b. The biology and clinical impact of genetic lesions in myeloid malignancies. *Blood*.

Makishima H, Visconte V, Sakaguchi H, Jankowska AM, Abu Kar S, Jerez A, Przychodzen B, Bupathi M, Guinta K, Afable MG, et al. 2012. Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis. *Blood* **119**: 3203–3210.

Martin M, Maßhöfer L, Temming P, Rahmann S, Metz C, Bornfeld N, van de Nes J, Klein-Hitpass L, Hinnebusch AG, Horsthemke B, et al. 2013. Exome sequencing identifies recurrent somatic mutations in EIF1AX and SF3B1 in uveal melanoma with disomy 3. *Nat Genet*.

Merendino L, Guth S, Bilbao D, Martínez C, Valcárcel J. 1999. Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG. *Nature* **402**: 838–841.

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, et al. 2013. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* **41**: D64–9.

Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, Pellagatti A, Wainscoat JS, Hellstrom-Lindberg E, Gambacorti-Passerini C, et al. 2011. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med* **365**: 1384–1395.

Przychodzen B, Jerez A, Guinta K, Sekeres MA, Padgett R, Maciejewski JP, Makishima H. 2013. Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms. *Blood*.

Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, Ramsay AJ, Beà S, Pinyol M, Martínez-Trillos A, et al. 2011. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet*.

Reed R. 1989. The organization of 3' splice-site sequences in mammalian introns. *Genes Dev* **3**: 2113–2123.

Shen H, Zheng X, Luecke S, Green MR. 2010. The U2AF35-related protein Urp contacts the 3' splice site to promote U12-type intron splicing and the second step of U2-type intron splicing. *Genes Dev* **24**: 2389–2394.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Smith CW, Chu TT, Nadal-Ginard B. 1993. Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Mol Cell Biol* **13**: 4939–4952.

Tefferi A, Vardiman JW. 2009. Myelodysplastic syndromes. *N Engl J Med* **361**: 1872–1885.

Thol F, Kade S, Schlarmann C, Löffeld P, Morgan M, Krauter J, Wlodarski MW, Kölking B, Wichmann M, Görlich K, et al. 2012. Frequency and prognostic impact of mutations in SRSF2, U2AF1, and ZRSR2 in patients with myelodysplastic syndromes. *Blood* **119**: 3578–3584.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111.

UniProt Consortium. 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **40**: D71–5.

Visconte V, Makishima H, Jankowska A, Szpurka H, Traina F, Jerez A, O'Keefe C, Rogers HJ, Sekeres MA, Maciejewski JP, et al. 2011. SF3B1, a splicing factor is frequently mutated in refractory anemia with ring sideroblasts. *Leukemia*.

Wagenmakers E-J, Lodewyckx T, Kuriyal H, Grasman R. 2010. Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method. *Cogn Psychol* **60**: 158–189.

Webb CJ, Wise JA. 2004. The splicing factor U2AF small subunit is functionally conserved between fission yeast and humans. *Mol Cell Biol* **24**: 4229–4240.

Wong JJ-L, Ritchie W, Ebner OA, Selbach M, Wong JWH, Huang Y, Gao D, Pinello N, Gonzalez M, Baidya K, et al. 2013. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**: 583–595.

Wu S, Romfo CM, Nilsen TW, Green MR. 1999. Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature* **402**: 832–835.

Yildirim E, Kirby JE, Brown DE, Mercier FE, Sadreyev RI, Scadden DT, Lee JT. 2013. Xist RNA is a potent suppressor of hematologic cancer in mice. *Cell* **152**: 727–742.

Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, et al. 2011. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**: 64–69.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**: R14.

Zorio DA, Blumenthal T. 1999. Both subunits of U2AF recognize the 3' splice site in Caenorhabditis elegans. *Nature* **402**: 835–838.

Figure 1

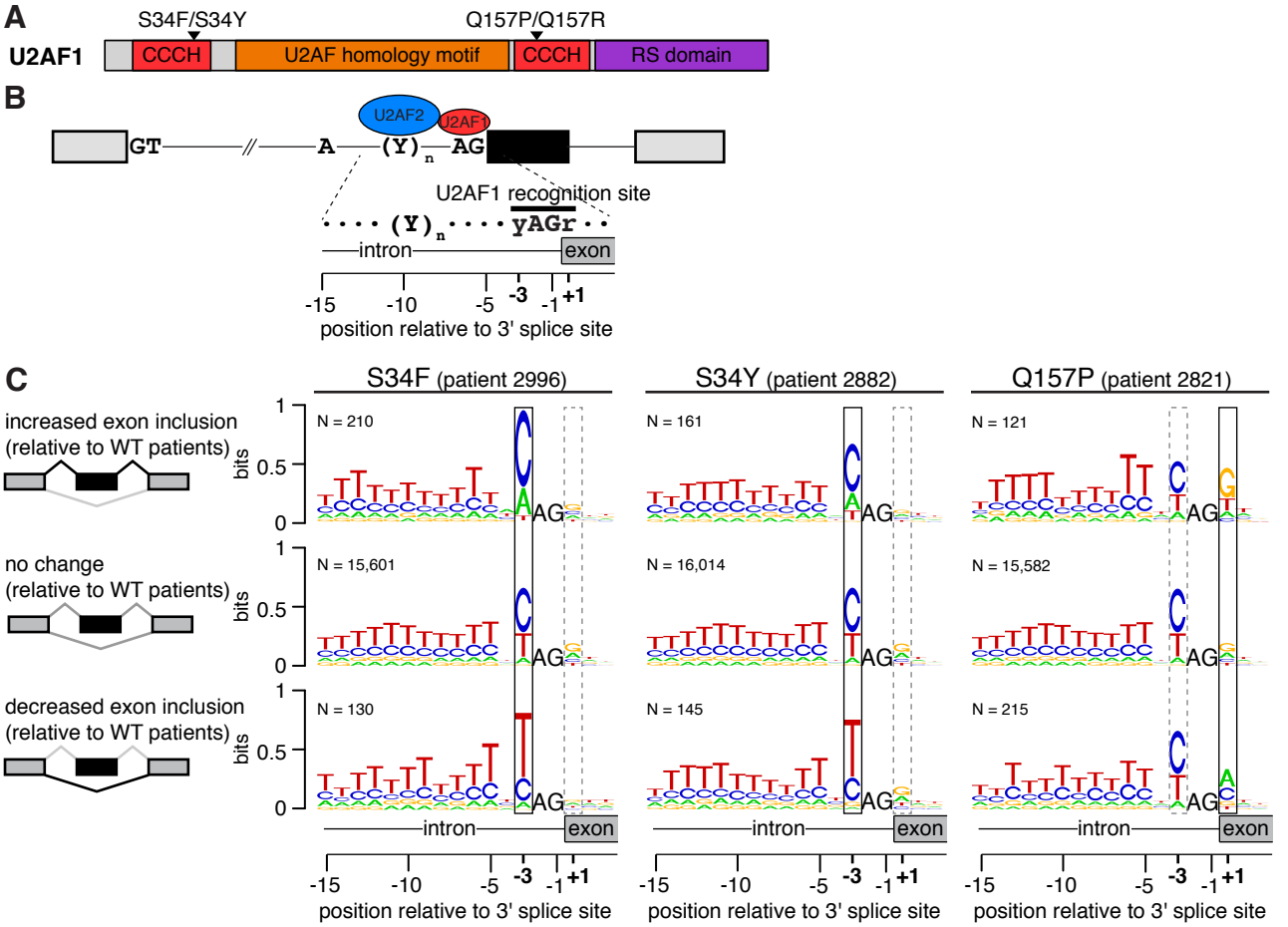

A
S34F/S34Y
Q157P/Q157R
U2AF1
CCCH
U2AF homology motif
CCCH
RS domain
B
U2AF2
U2AF1
GT
A
(Y)n
AG
U2AF1 recognition site
(Y)n
yAGr
intron
exon
-15
-10
-5
-3
-1
+1
position relative to 3' splice site
C
S34F (patient 2996)
S34Y (patient 2882)
Q157P (patient 2821)
increased exon inclusion
(relative to WT patients)
no change
(relative to WT patients)
decreased exon inclusion
(relative to WT patients)
bits
N = 210
N = 161
N = 121
N = 15,601
N = 16,014
N = 15,582
N = 130
N = 145
N = 215

Figure 2

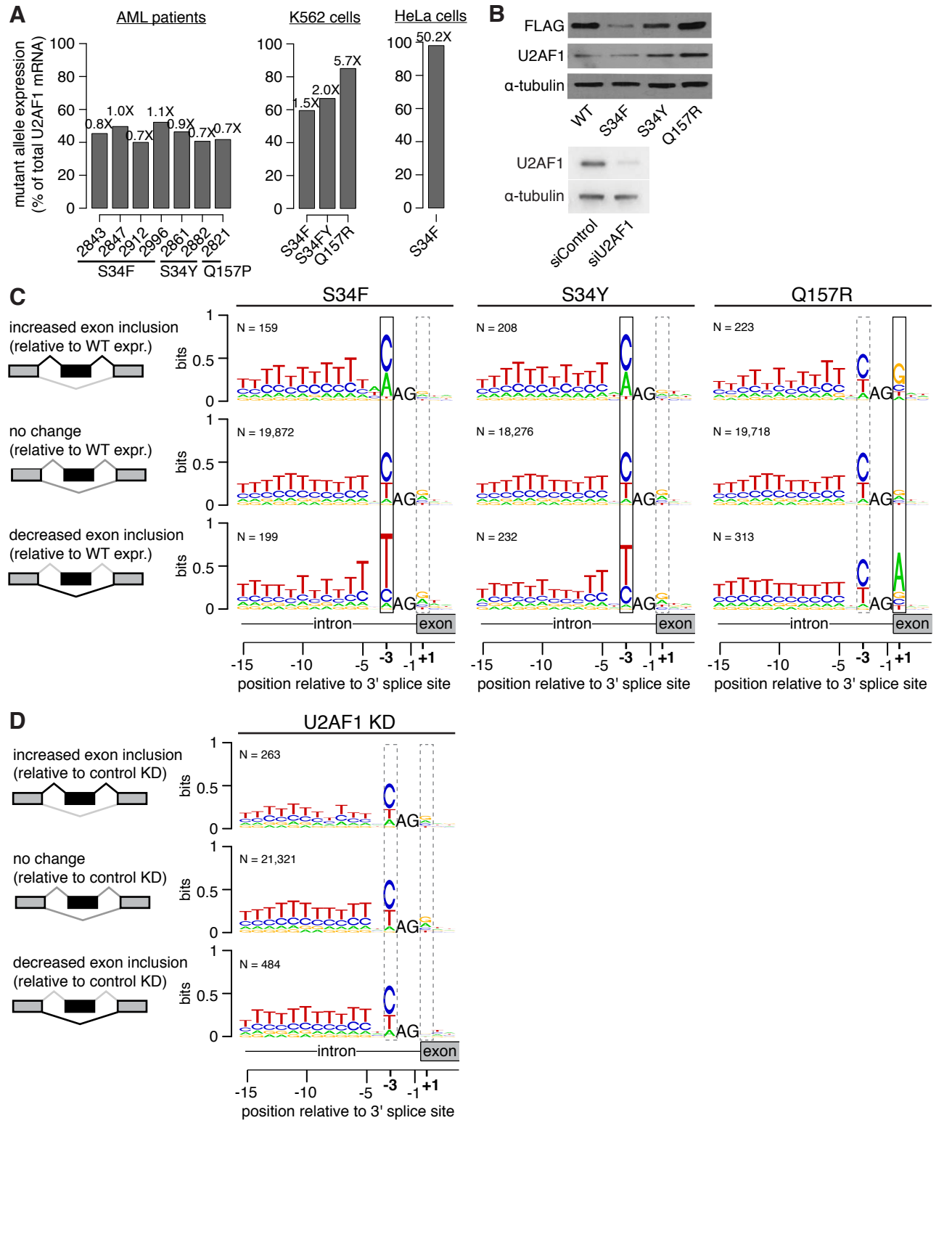
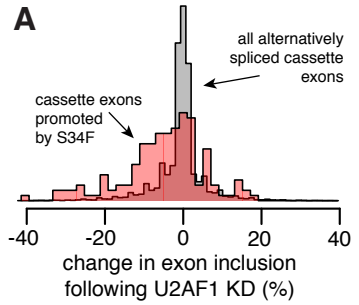
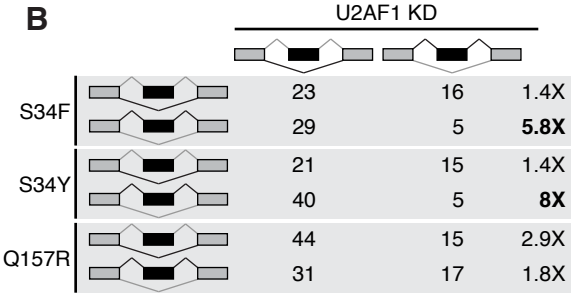
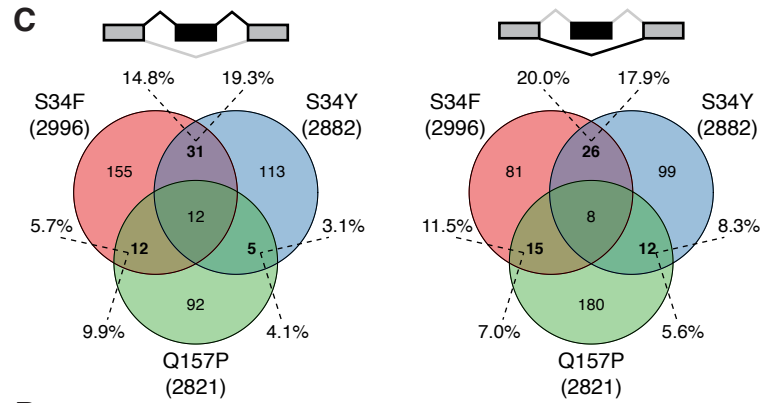

A
AML patients
K562 cells
HeLa cells
mutant allele expression
(% of total U2AF1 mRNA)
0.8X
1.0X
0.7X
1.1X
0.9X
0.7X
0.7X
1.5X
2.0X
5.7X
50.2X
2843
2847
2912
2996
2861
2882
2821
S34F
S34Y
Q157P
S34F
S34FY
Q157R
S34F
B
FLAG
U2AF1
α-tubulin
WT
S34F
S34Y
Q157R
U2AF1
α-tubulin
siControl
siU2AF1
C
S34F
S34Y
Q157R
increased exon inclusion
(relative to WT expr.)
no change
(relative to WT expr.)
decreased exon inclusion
(relative to WT expr.)
bits
N = 159
N = 19,872
N = 199
N = 208
N = 18,276
N = 232
N = 223
N = 19,718
N = 313
intron
exon
position relative to 3' splice site
D
U2AF1 KD
increased exon inclusion
(relative to control KD)
no change
(relative to control KD)
decreased exon inclusion
(relative to control KD)
N = 263
N = 21,321
N = 484

Figure 3


A
all alternatively spliced cassette exons
cassette exons promoted by S34F
-40
-20
0
20
40
change in exon inclusion following U2AF1 KD (%)


B
U2AF1 KD
S34F
23
16
1.4X
29
5
5.8X
S34Y
21
15
1.4X
40
5
8X
Q157R
44
15
2.9X
31
17
1.8X


C
14.8%
19.3%
S34F (2996)
S34Y (2882)
155
31
113
12
5.7%
12
5
3.1%
92
9.9%
4.1%
Q157P (2821)
20.0%
17.9%
S34F (2996)
S34Y (2882)
81
26
99
8
11.5%
15
12
8.3%
180
7.0%
5.6%
Q157P (2821)

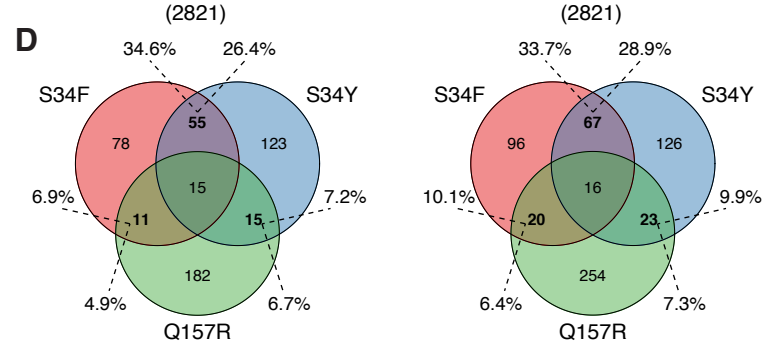
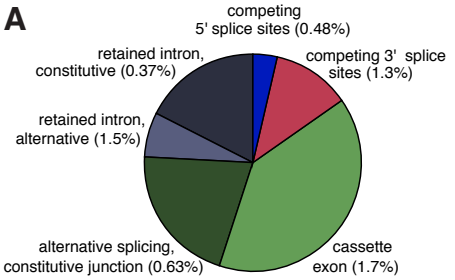
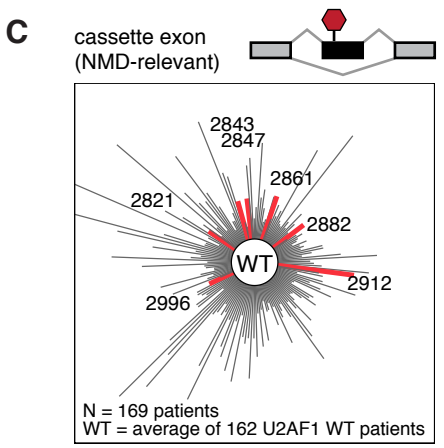

D
34.6%
26.4%
S34F
S34Y
78
55
123
15
6.9%
11
15
7.2%
182
4.9%
6.7%
Q157R
33.7%
28.9%
S34F
S34Y
96
67
126
16
10.1%
20
23
9.9%
254
6.4%
7.3%
Q157R

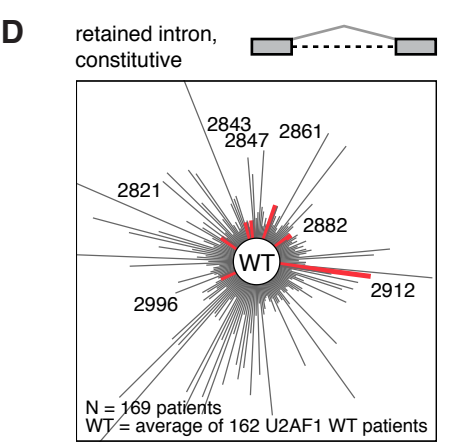
Figure 4

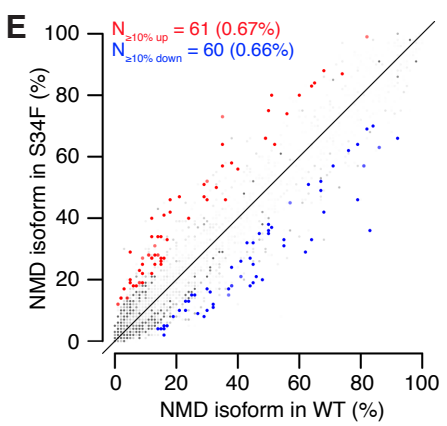


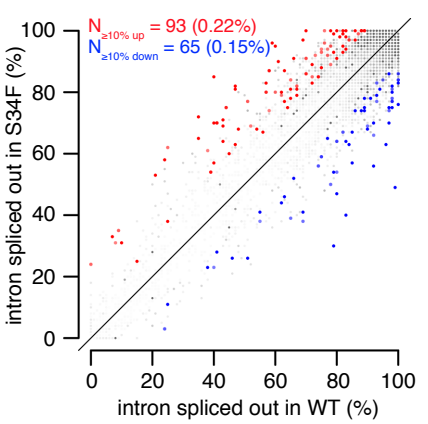
B

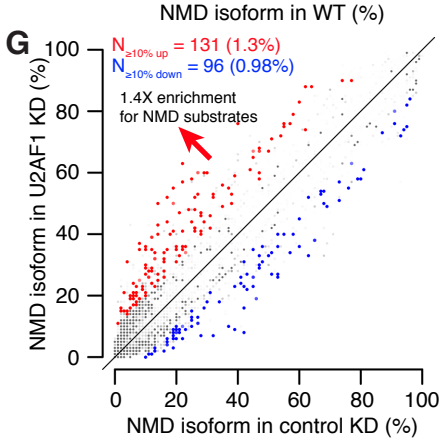
	S34F	S34Y	Q157R
competing 5' splice sites			
competing 3' splice sites	-3	-3	+1
cassette exon	-3	-3	+1
alternative splicing, constitutive junction	-3	-3	+1
retained intron, alternative	-3		+1
retained intron, constitutive	-3		+1











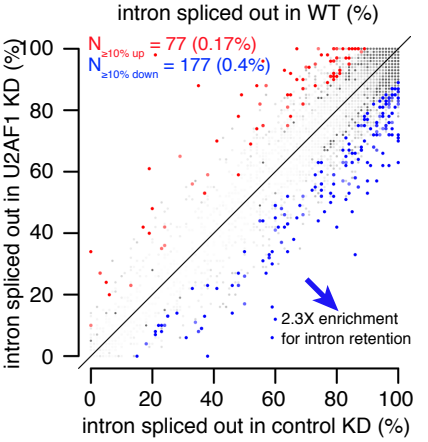


Figure 5

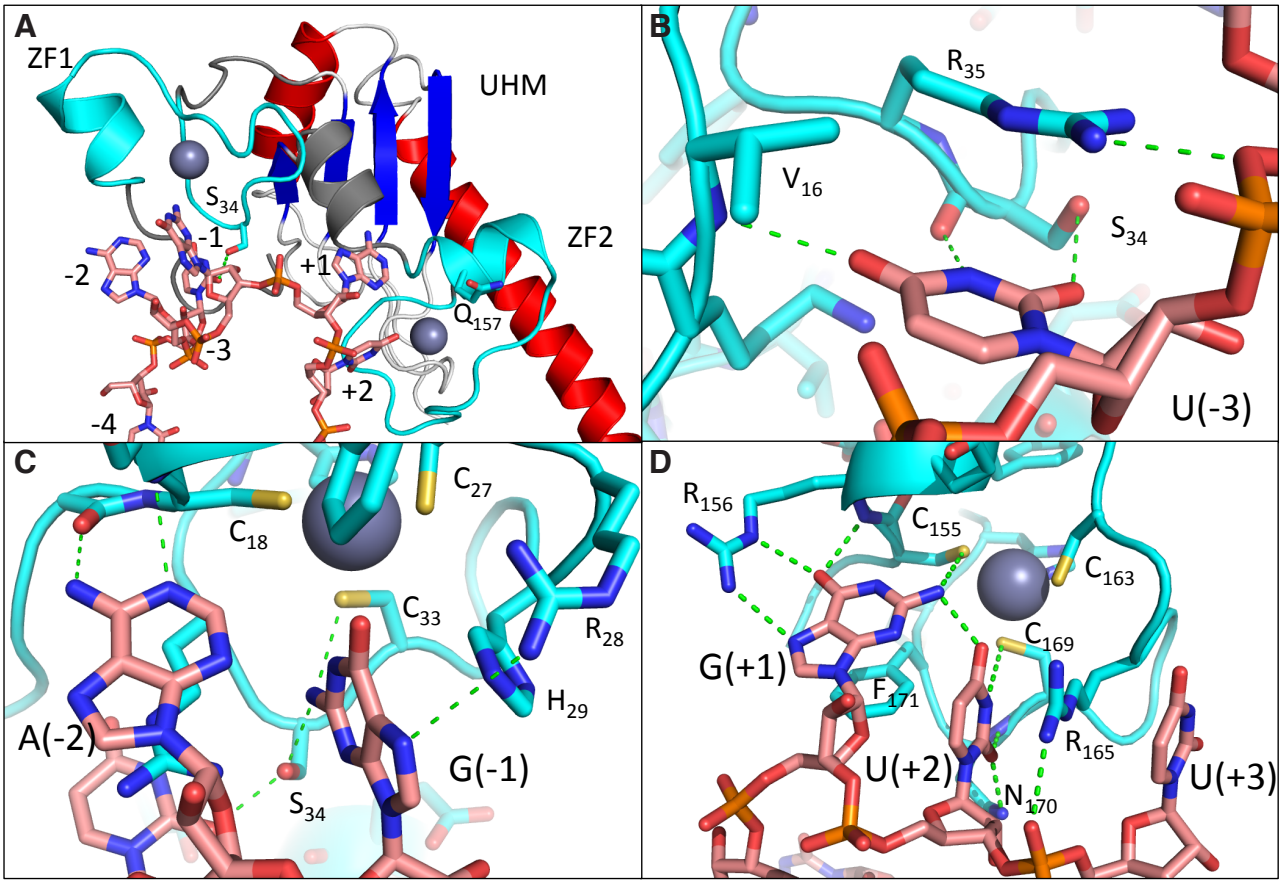

A
ZF1
UHM
ZF2
S34
-1
+1
-2
-3
+2
-4
Q157
B
R35
V16
S34
U(-3)
C
C27
C18
C33
R28
H29
A(-2)
G(-1)
S34
D
R156
C155
C163
C169
G(+1)
F171
R165
U(+2)
N170
U(+3)

Figure 6

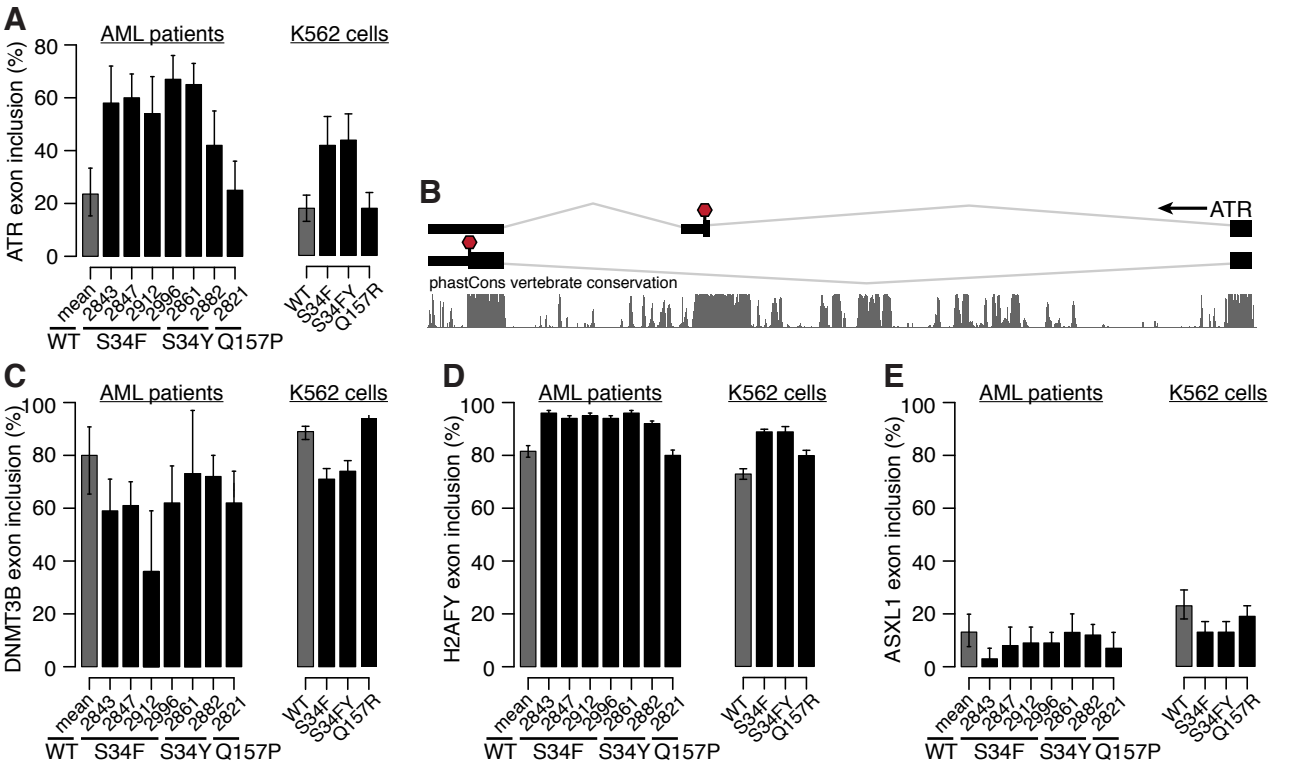

A
AML patients
K562 cells
ATR exon inclusion (%)
0
20
40
60
80
mean
2843
2847
2912
2996
2861
2882
2821
WT
S34F
S34Y
Q157P
WT
S34F
S34FY
Q157R
B
ATR
phastCons vertebrate conservation
C
AML patients
K562 cells
DNMT3B exon inclusion (%)
0
20
40
60
80
100
mean
2843
2847
2912
2996
2861
2882
2821
WT
S34F
S34Y
Q157P
WT
S34F
S34FY
Q157R
D
AML patients
K562 cells
H2AFY exon inclusion (%)
0
20
40
60
80
100
mean
2843
2847
2912
2996
2861
2882
2821
WT
S34F
S34Y
Q157P
WT
S34F
S34FY
Q157R
E
AML patients
K562 cells
ASXL1 exon inclusion (%)
0
20
40
60
80
100
mean
2843
2847
2912
2996
2861
2882
2821
WT
S34F
S34Y
Q157P
WT
S34F
S34FY
Q157R

Figure S1

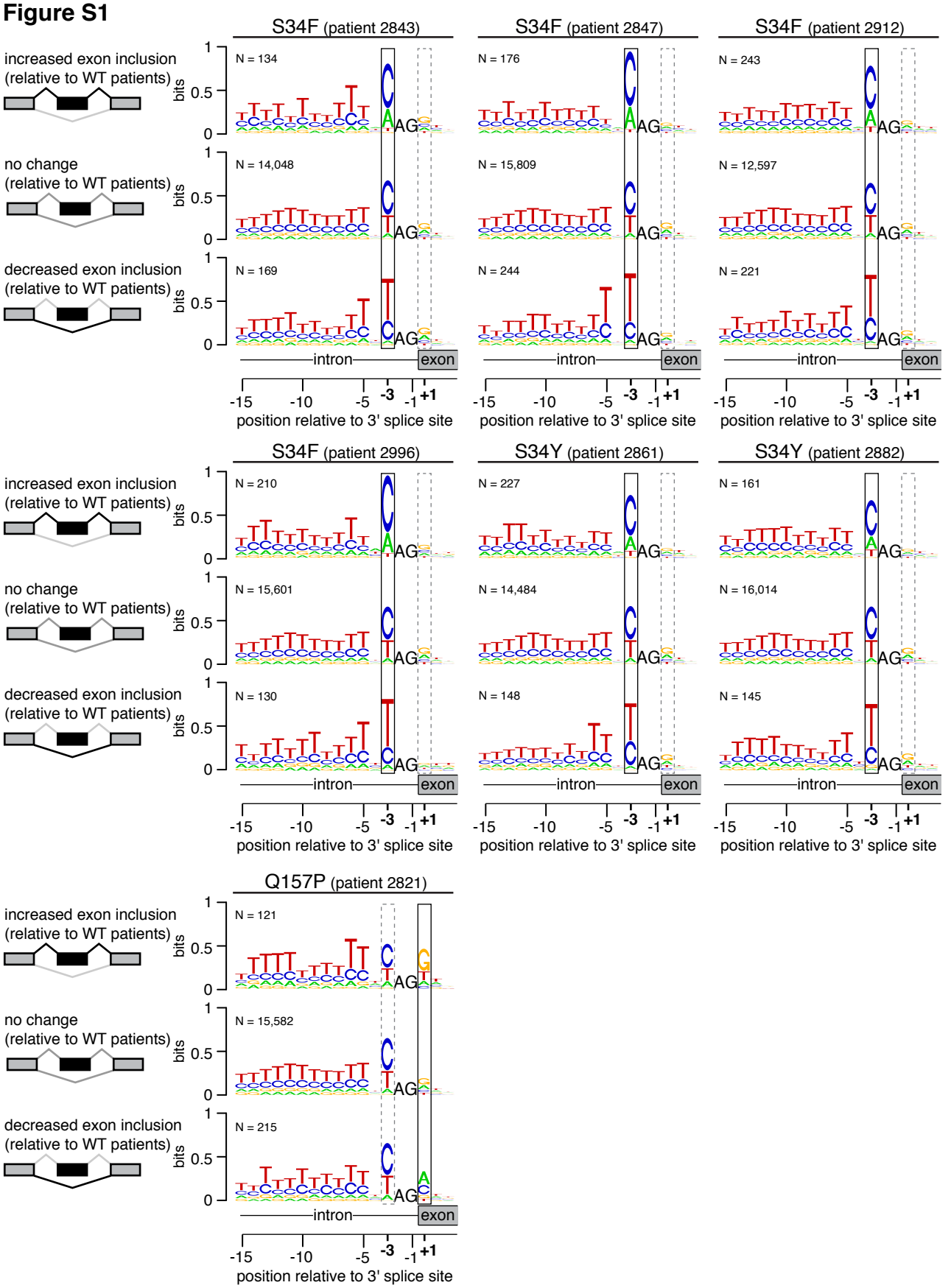

S34F (patient 2843)
S34F (patient 2847)
S34F (patient 2912)
S34F (patient 2996)
S34Y (patient 2861)
S34Y (patient 2882)
Q157P (patient 2821)
increased exon inclusion (relative to WT patients)
no change (relative to WT patients)
decreased exon inclusion (relative to WT patients)
N = 134
N = 14,048
N = 169
N = 176
N = 15,809
N = 244
N = 243
N = 12,597
N = 221
N = 210
N = 15,601
N = 130
N = 227
N = 14,484
N = 148
N = 161
N = 16,014
N = 145
N = 121
N = 15,582
N = 215
bits
intron
exon
-15 -10 -5 -3 -1 +1
position relative to 3' splice site

Figure S2

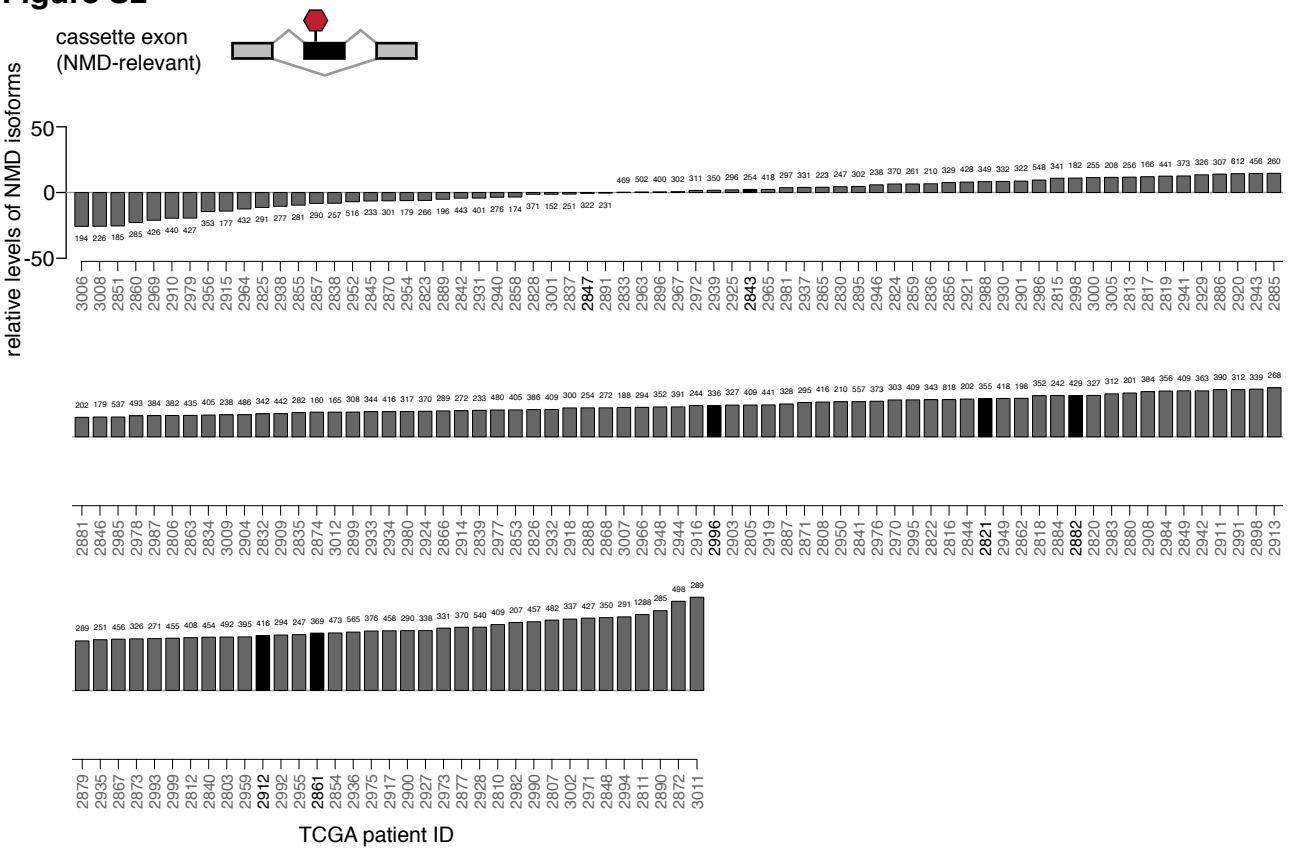

cassette exon
(NMD-relevant)
relative levels of NMD isoforms
50
0
-50
TCGA patient ID

Figure S3

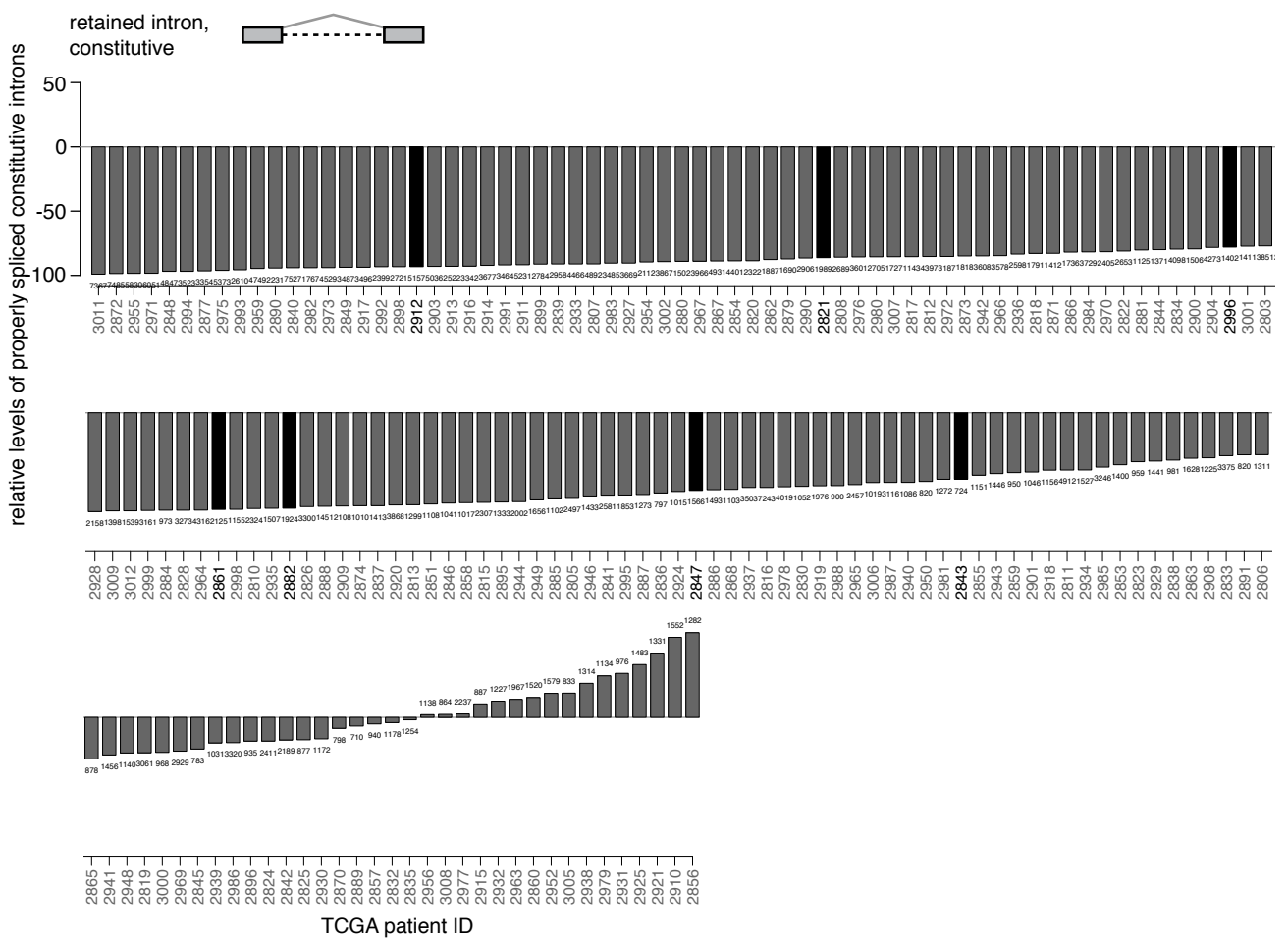

retained intron,
constitutive
relative levels of properly spliced constitutive introns
50
0
-50
-100
TCGA patient ID

Figure S4

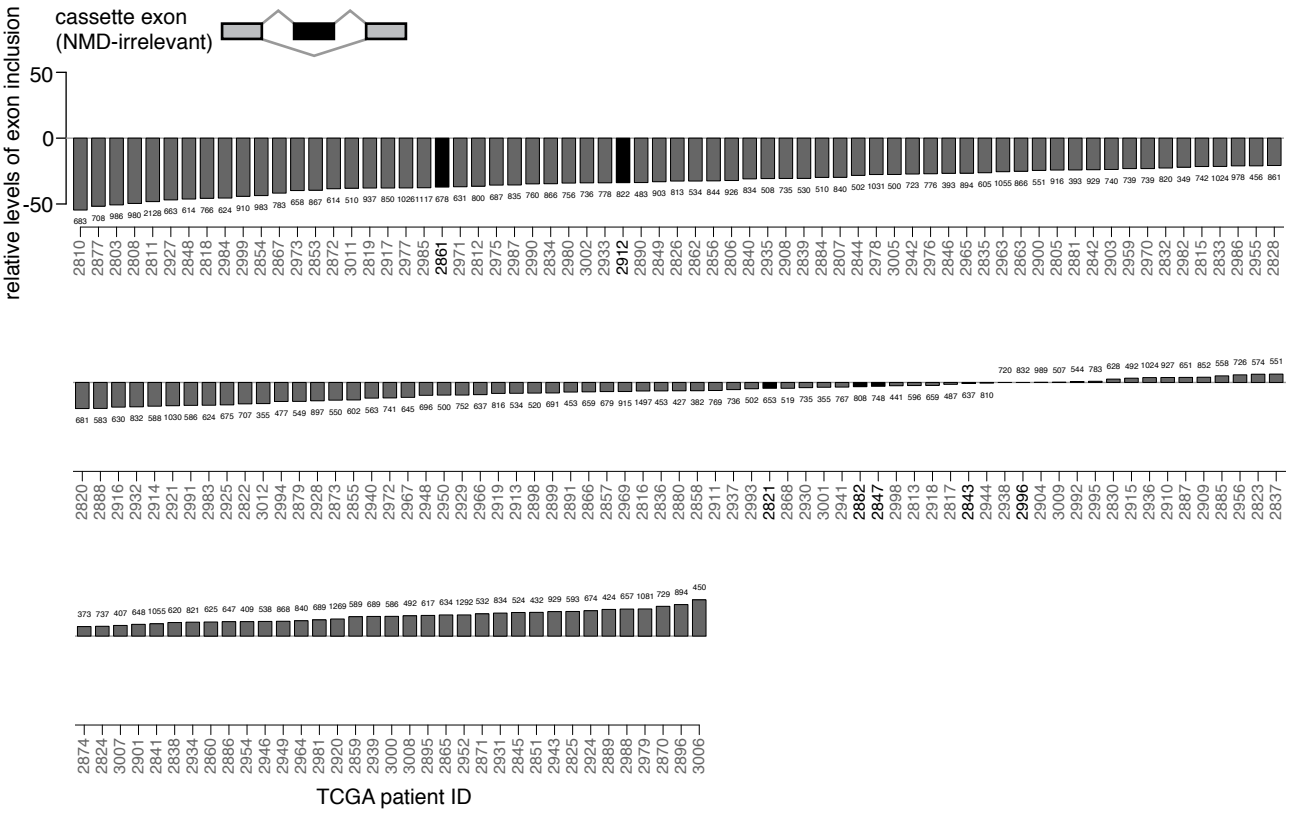

cassette exon
(NMD-irrelevant)
relative levels of exon inclusion
50
0
-50
TCGA patient ID

Figure S5

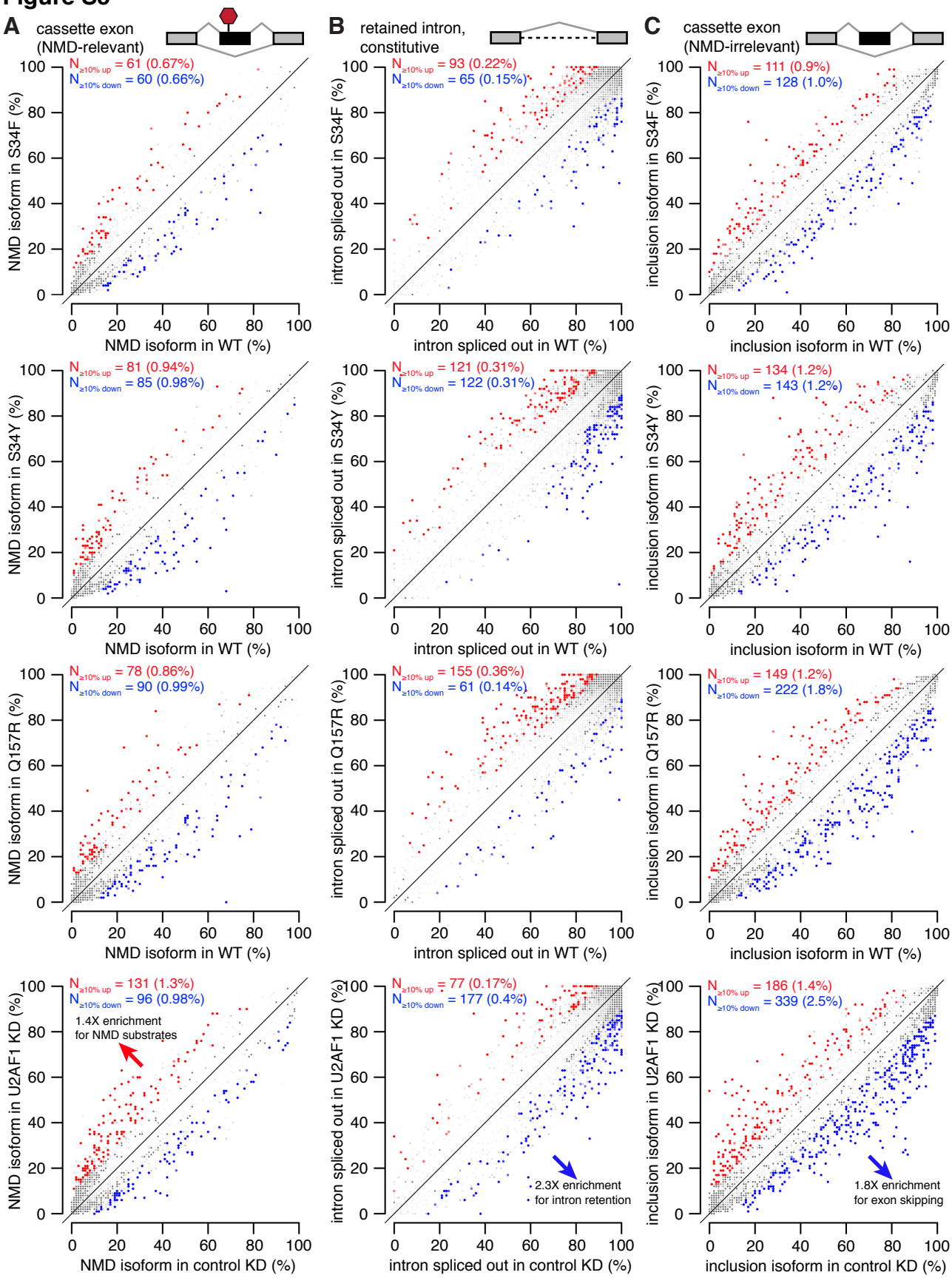
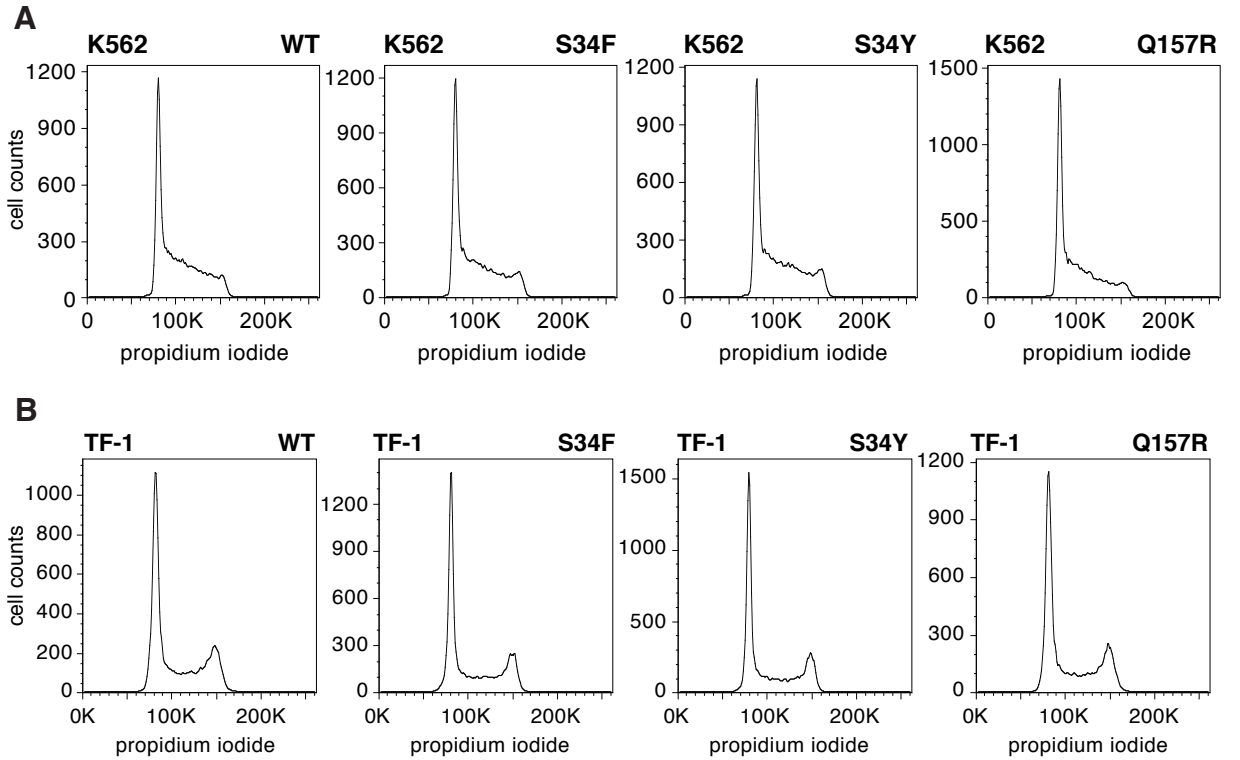

A
cassette exon
(NMD-relevant)
B
retained intron,
constitutive
C
cassette exon
(NMD-irrelevant)
N≥10% up = 61 (0.67%)
N≥10% down = 60 (0.66%)
NMD isoform in S34F (%)
NMD isoform in WT (%)
N≥10% up = 93 (0.22%)
N≥10% down = 65 (0.15%)
intron spliced out in S34F (%)
intron spliced out in WT (%)
N≥10% up = 111 (0.9%)
N≥10% down = 128 (1.0%)
inclusion isoform in S34F (%)
inclusion isoform in WT (%)
N≥10% up = 81 (0.94%)
N≥10% down = 85 (0.98%)
NMD isoform in S34Y (%)
NMD isoform in WT (%)
N≥10% up = 121 (0.31%)
N≥10% down = 122 (0.31%)
intron spliced out in S34Y (%)
intron spliced out in WT (%)
N≥10% up = 134 (1.2%)
N≥10% down = 143 (1.2%)
inclusion isoform in S34Y (%)
inclusion isoform in WT (%)
N≥10% up = 78 (0.86%)
N≥10% down = 90 (0.99%)
NMD isoform in Q157R (%)
NMD isoform in WT (%)
N≥10% up = 155 (0.36%)
N≥10% down = 61 (0.14%)
intron spliced out in Q157R (%)
intron spliced out in WT (%)
N≥10% up = 149 (1.2%)
N≥10% down = 222 (1.8%)
inclusion isoform in Q157R (%)
inclusion isoform in WT (%)
N≥10% up = 131 (1.3%)
N≥10% down = 96 (0.98%)
1.4X enrichment
for NMD substrates
NMD isoform in U2AF1 KD (%)
NMD isoform in control KD (%)
N≥10% up = 77 (0.17%)
N≥10% down = 177 (0.4%)
2.3X enrichment
for intron retention
intron spliced out in U2AF1 KD (%)
intron spliced out in control KD (%)
N≥10% up = 186 (1.4%)
N≥10% down = 339 (2.5%)
1.8X enrichment
for exon skipping
inclusion isoform in U2AF1 KD (%)
inclusion isoform in control KD (%)

Figure S6


A
K562 WT
K562 S34F
K562 S34Y
K562 Q157R
cell counts
propidium iodide
B
TF-1 WT
TF-1 S34F
TF-1 S34Y
TF-1 Q157R
cell counts
propidium iodide