bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

U2AF1 mutations alter splice site recognition in hematological malignancies

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

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

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

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

[4]Division of Medical Oncology, School of Medicine, University of Washington, 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

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ABSTRACT

Whole-exome sequencing studies have identified common mutations affecting genes encoding components of the RNA splicing machinery in hematological malignancies. Here, we sought to determine how mutations affecting the 3' splice site recognition factor U2AF1 alter its normal role in RNA splicing. We find that *U2AF1* mutations influence the similarity of splicing programs in leukemias, but do not give rise to widespread splicing failure. *U2AF1* mutations cause differential splicing of hundreds of genes, affecting biological pathways implicated in myeloid disease such as DNA methylation (*DNMT3B*), X chromosome inactivation (*H2AFY*), the DNA damage response (*ATR*, *FANCA*), and apoptosis (*CASP8*). We show that *U2AF1* mutations alter the preferred 3' splice site motif in patients, in cell culture, and *in vitro*. Mutations affecting the first and second zinc fingers give rise to different alterations in splice site preference and largely distinct downstream splicing programs. These allele-specific effects are consistent with a computationally predicted model of U2AF1 in complex with RNA. Our findings suggest that *U2AF1* mutations contribute to pathogenesis by causing quantitative changes in splicing that affect diverse cellular pathways, and give insight into the normal function of U2AF1's zinc finger domains.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 proteins involved in 3' splice site recognition (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.

U2AF1 (also known as *U2AF35*) may provide a useful model system to dissect 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 (Figure 1B), 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

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. 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 found that overexpression of mutant U2AF1 in HeLa cells resulted in dysfunctional splicing marked by frequent inclusion of premature termination codons and intron retention (Yoshida et al. 2011), while another early study reported increased exon skipping in a minigene assay following mutant U2AF1 expression in 293T cells, as well as increased cryptic splice site usage in the *FMR1* gene in MDS samples (Graubert et al. 2011). *U2AF1* mutations have been suggested to cause both alteration/gain of function (Graubert et al. 2011) and loss of function (Yoshida et al. 2011; Makishima et al. 2012). More recently, two studies analyzed acute myeloid leukemia transcriptomes from The Cancer Genome Atlas (TCGA), and found that exons with increased or decreased inclusion in samples with *U2AF1* mutations exhibited different nucleotides prior to the AG of the 3' splice site (Przychodzen et al. 2013; Brooks et al. 2014), suggesting that *U2AF1* mutations may cause specific alterations to the RNA splicing process.

To determine how *U2AF1* mutations alter RNA splicing in hematopoietic cells, we combined patient data, cell culture experiments, and biochemical studies. We found that *U2AF1* mutations cause splicing alterations in biological pathways previously implicated in myeloid malignancies, including epigenetic regulation and the DNA damage response. *U2AF1* mutations drive differential splicing by altering the preferred 3' splice site motif in an allele-specific manner. Our results identify downstream targets of *U2AF1* mutations that may contribute to pathogenesis, show that different *U2AF1* mutations are not functionally equivalent, and give insight into the normal function of U2AF1's zinc finger domains.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

RESULTS

U2AF1 mutations are associated with distinct splicing programs in AML

We first tested whether *U2AF1* mutations were relevant to splicing programs in leukemias with an unbiased approach. We quantified genome-wide cassette exon splicing in the transcriptomes of 169 *de novo* adult acute myeloid leukemia (AML) samples that were sequenced as part of The Cancer Genome Atlas (Cancer Genome Atlas Research Network 2013) and performed unsupervised cluster analysis. Five of the seven samples carrying a *U2AF1* mutation clustered together (Figure 1C). One of the samples that fell outside of this cluster had a Q157 rather than S34 *U2AF1* mutation, and the other carried a mutation in the RNA splicing gene *KHDRBS3* in addition to a *U2AF1* mutation, potentially contributing to its placement in an outgroup. Both of the outgroup *U2AF1* mutant samples additionally had low mutant allele expression relative to wild-type (WT) allele expression (Figure 1D). These results suggest that *U2AF1* mutations are associated with distinct splicing patterns in patients, and are consistent with a recent report that spliceosomal mutations define a distinct subgroup of myeloid malignancies based on gene expression and DNA methylation patterns (Taskesen et al. 2014).

U2AF1 mutations alter RNA splicing in blood cells

To determine how *U2AF1* mutations affect RNA splicing in an experimentally tractable system, we generated K562 erythroleukemic cell lines that stably expressed a single FLAG-tagged *U2AF1* allele (WT, S34F, S34Y, Q157P, or Q157R) at modest levels in the presence of the endogenous protein (Figure 2A). This expression strategy, where the transgene was expressed at levels of 1.8-4.7X endogenous *U2AF1*, is consistent with the co-expression of WT and mutant alleles at approximately equal levels that we observed in AML transcriptomes (Figure 1D,2B). Similar 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 U2AF1 protein levels in the absence of transgenic expression to test whether the mutations cause gain or loss of function.

To identify mutation-dependent changes in splicing, we performed deep RNA-seq on these K562 cell lines stably expressing each mutant allele (~100M 2x49 bp reads per cell line). This provided sufficient read coverage to measure quantitative inclusion of ~20,000 cassette

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

exons that were alternatively spliced in K562 cells. Unsupervised cluster analysis of global cassette exon inclusion in these cell lines placed S34F/Y and Q157P/R as distinct groups and revealed that mutations within the first and second zinc fingers are associated with largely distinct patterns of exon inclusion (Figure 2C). This is consistent with our cluster analysis of AML transcriptomes, where the one sample with a Q157 mutation was placed as an outgroup to samples with S34 mutations.

We next assembled comprehensive maps of splicing changes driven by *U2AF1* mutations in AML transcriptomes, K562 cells expressing mutant *U2AF1*, and K562 cells following *U2AF1* KD. We tested ~125,000 annotated alternative splicing events for differential splicing, and assayed ~160,000 constitutive splice junctions for evidence of novel alternative splicing or intron retention. We required a minimum change in isoform ratio of 10% to call an event differentially spliced. As our cluster analysis of K562 cells indicated that S34 and Q157 mutations generate distinct splicing patterns, we compared the six S34 AML samples to all *U2AF1* WT AML samples. We separately identified splicing changes caused by both S34F and S34Y or both Q157P and Q157R in K562 cells relative to the WT control cells. The resulting catalogs of differentially spliced events revealed that all major classes of alternative splicing events, including cassette exons, competing splice sites, and retained introns, were affected by *U2AF1* mutations (Figure 2D, File S1-5). Cassette exons constituted the majority of affected splicing events, followed by alternative splicing or intron retention of splice junctions annotated as constitutive in the UCSC genome browser (Meyer et al. 2013).

Thousands of splicing events were affected by each common *U2AF1* mutation, but the fraction of differentially spliced events was relatively low. For example, >400 frame-preserving cassette exons were differentially spliced in association with S34Y vs. WT *U2AF1* expression; however, those >400 cassette exons constituted only ~3.6% of frame-preserving cassette exons that are alternatively spliced in K562 cells (Figure 2E). Expression of any mutant allele caused differential splicing of 2-5% of frame-preserving cassette exons, with a bias towards exon skipping (Figure S1A). We did not observe increased intron retention or expression of isoforms that are predicted substrates for degradation by nonsense-mediated decay (NMD) in association with any *U2AF1* mutation. Instead, constitutive intron removal appeared slightly more efficient in cells expressing mutant versus WT *U2AF1* (Figure 2F,S1B-C). In contrast, we did observe increased expression of NMD substrates and mRNAs with unspliced introns following *U2AF1*

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

KD (Figure S1B-C). Consistent with these findings in K562 cells, AML samples carrying *U2AF1* mutations did not exhibit increased levels of NMD substrates or intron retention (Figure 2G,S2-4). We conclude that S34 and Q157 *U2AF1* mutations cause splicing changes affecting hundreds of exons, but do not give rise to widespread splicing failure.

These results contrast with an early report that the *U2AF1* S34F mutation causes overproduction of mRNAs slated for degradation and genome-wide intron retention (Yoshida et al. 2011). The discrepancy between those results and our observations are likely due to differing experimental designs. This previous study acutely expressed the S34F allele at 50X WT levels in HeLa cells, whereas we stably expressed each allele at 1.8-4.7X WT levels in blood cells (Figure 1D). Maintaining a balance between WT and mutant allele expression—like that observed in AML and MDS patients—may be important to maintain efficient splicing.

U2AF1 mutations cause differential splicing of cancer-relevant genes

We next sought to identify downstream targets of *U2AF1* mutations that might contribute to myeloid pathogenesis. We took a conservative approach of requiring differential splicing in AML transcriptomes as well as in K562 cells to help identify disease-relevant events that are likely direct consequences of *U2AF1* mutations. We intersected differentially spliced events identified in three distinct comparisons: AML S34 vs. WT samples, K562 S34 vs. WT expression, and K562 Q157 vs. WT expression. 16.8% of AML S34-associated differential splicing was phenocopied in K562 S34 cells versus 4.6% for K562 Q157 cells, consistent with allele-specific effects of *U2AF1* mutations (Figure 3A). The relatively low overlap of ~17% between AML and K562 S34-associated differential splicing is due to differences in gene expression patterns between these cell types as well as the modest nature of splicing changes caused by *U2AF1* mutations, such that many changes fall near the border of our statistical thresholds for differential splicing (Figure 2E). This analysis revealed 54 splicing events that were affected by both S34 and Q157 mutations in AML transcriptomes and K562 cells. When we instead intersected genes containing differentially spliced events—not requiring that identical exons or splice sites be affected—we found a substantially increased intersection of 140 genes (Table 1).

Many genes that were differentially spliced in association with *U2AF1* mutations participate in biological pathways previously implicated in myeloid malignancies. For example,

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DNMT3A encodes a *de novo* DNA methyltransferase and is a common mutational target in myelodysplastic syndromes and acute myeloid leukemia (Walter et al. 2011; Ley et al. 2010). Multiple exons of its paralog *DNMT3B*, including an exon encoding part of the methyltransferase domain, are differentially spliced in AML patients carrying *U2AF1* mutations as well as in K562 cells expressing *U2AF1* mutant alleles (Figure 3B-D). Similarly, different exons of *ASXL1* are alternatively spliced in association with S34 mutations in AML transcriptomes and K562 cells, although the same exons are not consistently affected (File S1-5). *ASXL1* is a common mutational target in myelodysplastic syndromes and related disorders (Gelsi-Boyer et al. 2009), and *U2AF1* and *ASXL1* mutations co-occur more frequently than expected by chance (Thol et al. 2012). Other genes participating in epigenetic processes are differentially spliced as well, such as *H2AFY* (Figure 3E). *H2AFY* encodes the core histone macro-H2A.1, which is important for X chromosome inactivation (Hernández-Muñoz et al. 2005). As loss of X chromosome inactivation causes a MDS-like disease in mice (Yildirim et al. 2013), differential splicing of macro-H2A.1 could potentially be relevant to disease processes.

Isoform switches, wherein a previously minor isoform becomes the major isoform, were relatively rare but did occur. 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 association with S34, but not Q157, mutations. This cassette exon alters the C terminus of the ATR protein, may render the mRNA susceptible to nonsense-mediated decay, and is highly conserved (Figure 3F-G). S34 mutations similarly cause an isoform switch from an intron-proximal to an intron-distal 3' splice site of *CASP8* that is predicted to shorten the N terminus of the protein (Figure 3H).

We noticed that splicing changes frequently affected multiple genes relevant to a specific biological process, such as DNA damage (*ATR* and *FANCA*; Figure 3G,I). Consistent with this observation, gene ontology analysis indicated that genes involved in the cell cycle, chromatin modification, DNA methylation, DNA repair, and RNA processing pathways, among others, are enriched for differential splicing in both AML transcriptomes and K562 cells in association with *U2AF1* mutations. This enrichment 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

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ontology analyses. We conclude that *U2AF1* mutations preferentially affect specific biological pathways, but that this enrichment is due to frequent alternative splicing within such genes rather than specific targeting by *U2AF1* mutant protein.

U2AF1 mutations cause allele-specific alterations in the 3' splice site consensus

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 the splicing changes caused by *U2AF1* mutations might be due to preferential activation or repression of 3' splice sites in a sequence-specific manner. To test this hypothesis, we identified consensus 3' splice sites of cassette exons that were promoted or repressed in AML transcriptomes carrying *U2AF1* mutations relative to WT patients. 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%. 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 supports our hypothesis that *U2AF1* mutations alter 3' splice site recognition activity in a sequence-specific manner (Figure 4A).

Mutations affecting 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, while the Q157P mutation in the second zinc finger was associated with alterations at the +1 position (Figure S5). In contrast, cassette exons that were differentially spliced in *U2AF1* WT samples did not exhibit altered consensus sequences at the -3 or +1 positions (Figure S6). These results confirm the findings of two recent studies of this cohort of AML patients—which reported a frequent preference for C instead of T at the -3 position of differentially spliced cassette exons in *U2AF1* mutant samples (Przychodzen et al. 2013; Brooks et al. 2014)—and extend their observations of altered splice site preference to show allele-specific effects of *U2AF1* mutations, which have not been previously identified.

U2AF1 mutation-dependent sequence preferences (C/A >> T at the -3 position for S34F/Y and G >> A at the +1 position for Q157P) differ from the genomic consensus for

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cassette exons. C/T and G/A appear at similar frequencies at the -3 and +1 positions of 3' splice sites of cassette exons (Figure S5,6), and 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). The consensus 3' splice sites associated with promoted/repressed cassette exons in *U2AF1* mutant transcriptomes also differ from U2AF1's known RNA binding specificity. A previous study identified a core tAG|g motif in the majority of RNA sequences bound by U2AF1 in a SELEX experiment (Wu et al. 1999). Comparing that motif with preferences observed in *U2AF1* mutant transcriptomes, we hypothesize that S34 mutations promote unusual recognition of C instead of T at the -3 position, while Q157 mutations reinforce preferential recognition of G instead of A at the +1 position.

We next tested whether these alterations in 3' splice site preference are a direct consequence of *U2AF1* mutations. Comparing K562 cells expressing a *U2AF1* mutant vs. WT allele, we found that 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 2B). 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* did not exhibit sequence-specific changes at the -3 or +1 positions of the 3' splice site (Figure 4C). We therefore conclude that S34 and Q157 mutations cause alteration or gain of function, consistent with the empirical absence of inactivating (nonsense or frameshift) *U2AF1* mutations observed in patients.

U2AF1 mutations preferentially affect U2AF1-dependent 3' splice sites

U2AF1 mutations are associated with altered 3' splice site consensus sequences, yet only a relatively small fraction of cassette exons are affected by expression of *U2AF1* mutant alleles. Previous biochemical studies 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 that are sensitive to *U2AF1* mutations might also rely upon U2AF1 recruitment for normal splicing. We empirically defined U2AF1-dependent exons as those with decreased inclusion following *U2AF1* KD, and computed the overlap between

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

U2AF1-dependent exons and exons that were affected by *U2AF1* mutant allele expression. For every mutant allele, we observed an enrichment for overlap with U2AF1-dependent exons, suggesting that *U2AF1* mutations preferentially affect exons with AG-dependent splice sites (Figure 4D).

U2AF1 mutations alter the preferred 3' splice site motif yAG|r

Our genomics data shows that cassette exons promoted/repressed by *U2AF1* mutations have 3' splice sites differing from the consensus. We therefore tested whether altering the core 3' splice site motif of an exon influenced its recognition in the presence of WT versus mutant *U2AF1* protein. We created a minigene encoding a cassette exon of *ATR*, which responds robustly to S34 mutations in AML transcriptomes and K562 cells, by cloning the 5' end of the *ATR* genomic locus into a plasmid. The minigene exhibited mutation-dependent splicing of the cassette exon, as expected, although splicing was less efficient than from the endogenous locus. We then mutated the -3 position of the cassette exon's 3' splice site to A/C/G/T and measured cassette exon inclusion in WT and S34Y K562 cells. Robust mutation-dependent increases in splicing required the A at the -3 position found in the endogenous locus, consistent with the unusual preference for A observed in our analyses of AML and K562 transcriptomes. We additionally observed a small but reproducible increase for C (Figure 5A). We next performed similar experiments for Q157-dependent splicing changes. We created a minigene encoding a cassette exon of *EPB49* (encoding the erythrocyte membrane protein band 4.9), mutated the +1 position of the 3' splice site to A/C/G/T, and measured cassette exon inclusion in WT and Q157R K562 cells. Cassette exon recognition was suppressed by Q157R expression when the +1 position was an A, consistent with our genomic prediction, and was not affected by Q157R when the +1 position was mutated to another nucleotide. Therefore, for both *ATR* and *EPB49*, robust S34 and Q157-dependent changes in splicing require the endogenous nucleotides at the -3 and +1 positions.

We then tested how *U2AF1* mutations influence constitutive, rather than alternative, splicing in an *in vitro* context. We used the adenovirus major late (AdML) substrate, a standard model of constitutive splicing, and mutated the -3 position of the 3' splice site to C/T. We measured AdML splicing efficiency following *in vitro* transcription and incubation with nuclear extract of K562 cells expressing WT or S34Y *U2AF1*. The AdML substrate exhibited sequence-

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

specific changes in splicing efficiency in association with *U2AF1* mutations. Consistent with our genomic analyses, AdML with C/T at the -3 position was more/less efficiently spliced in S34Y versus WT cells (Figure 5C). Taken together, our data demonstrate that *U2AF1* mutations cause sequence-specific alterations in the preferred 3' splice site motif in patients, in cell culture, and *in vitro*.

U2AF1 mutations may modify U2AF1:RNA interactions

As *U2AF1* mutations alter 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 could potentially modify 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 domain mediates U2AF heterodimer formation and is sufficient to promote splicing of an AG-dependent pre-mRNA substrate (Guth et al. 2001; Kielkopf et al. 2001). However, this domain binds a consensus 3' splice site sequence with low affinity (Kielkopf et al. 2001), suggesting that it may be 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 whether this hypothesis is sterically possible, we started from the experimentally determined structure of the UHM domain in complex with a peptide from U2AF2 (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)).

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 6A, File S6). The first zinc finger contacts the bases immediately preceding the splice site, including the AG dinucleotide (Figure 6B-C), while the second zinc finger binds immediately downstream (Figure 6D). 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.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DISCUSSION

Here, we have described the mechanistic consequences of *U2AF1* mutations in hematopoietic cells, as well as provided a catalog of splicing changes driven by each common *U2AF1* mutation. *U2AF1* mutations cause highly specific alterations in 3' splice site recognition in myeloid neoplasms. Taken together with the high frequency of mutations targeting *U2AF1* and genes encoding other 3' splice site recognition factors, our results support the hypothesis that specific alterations in 3' splice site recognition are important contributors to the molecular pathology of MDS and related hematological disorders.

We observed consistent differential splicing of multiple genes such as *DNMT3B* and *FANCA* that participate in molecular pathways previously implicated in blood disease. 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, hundreds of exons are differentially spliced in response to *U2AF1* mutations. Second, many of the splicing changes are relatively modest. In both the AML and K562 data, we observed relatively few isoform switches, with the *ATR* and *CASP8* examples illustrated in Figure 3 being notable exceptions. Therefore, we expect that specific targets such as *DNMT3B* probably contribute to, but do not wholly explain, *U2AF1* pathophysiology. As additional data from tumor transcriptome sequencing become available—for example, as more patient 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

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

disrupting such global processes, altered splice site recognition could contribute to the ineffective hematopoiesis that characterizes MDS.

Relevance to future studies of spliceosomal mutations

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 studying 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, we speculate that 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. With a fairly restricted set of assumptions, we computationally predicted a family of models in which the first zinc finger of U2AF1 recognizes the AG dinucleotide of the 3' splice site. 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, this model may provide a useful framework for future studies of U2AF1 function in both healthy and diseased cells.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 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, K562 cells were infected with concentrated lentiviral supernatants at a MOI of ~5,in growth media supplemented with 8 ug/mL protamine sulfate . 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. 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.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mRNA sequencing

Total RNA was obtained by lysing 10 million K562 cells for each sample in TRIzol and RNA was extracted using Qiagen RNeasy 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

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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). Because the AML data had approximately two-fold lower read coverage than the K562 data, we reduced the Bayes factor by a factor of two to compensate for the loss in statistical power. Wagenmakers's framework (Wagenmakers et al. 2010) was used to compute Bayes factors for differences in isoform ratios between samples.

Cluster analysis

To perform the cluster analysis of AML transcriptomes (Figure 1C) and K562 cells (Figure 2C), we identified cassette exons that displayed changes in isoform ratios ≥10% across the samples,

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and then further restricted to cassette exons with at least 100 informative reads across all samples. An informative read is defined as a RNA-seq read that supports either isoform, but not both. We created a similarity matrix using the Pearson correlation computed from the z-score normalized cassette exon inclusion values, and clustered the samples using Ward's method.

AML WT and mutant comparisons

To identify splicing events that were differentially spliced in AML S34 samples vs. WT samples (File S1), we used the Mann-Whitney U test and required $p < 0.01$. To identify splicing events that were differentially spliced in each AML sample with a U2AF1 mutation (Figure 4A,S5), 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 was 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.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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), anti-Histone H3 (Abcam), 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.

Minigenes

An insert containing the *ATR* genomic locus (chr3:142168344-142172070) or *EPB49* genomic locus (chr8:21938036-21938724) was cloned into the EcoRV site of pUB6/V5-HisA vector (Invitrogen) by Gibson assembly cloning (NEB). Site-directed mutagenesis was used to generate different nucleotides at the -3 position at 3' splice site. Plasmids containing minigenes were transfected using the Nucleofector II device from Lonza with the Cell Line Nucleofector Kit V (program T16). RNA was collected after 48h. We isolated total RNA from K562 cells using TRIzol and extracted RNA using the Qiagen RNeasy kit. Complementary DNA (cDNA) was generated using 1 ug of total RNA with a primer specific to the minigene transcript immediately upstream of the poly(A) tail (ACAACAGATGGCTGGCAACTAGAAG). Assays were performed in biological triplicate. Triplicates of equal amounts of 6 ng cDNA were used in a 5 uL reaction with 2.5 uL 2x SYBR Green PCR Master Mix (Life Technologies), and 50 nM forward and reverse primers. *ATR* primers: inclusion (forward AACTGGAGAAGTTGTCAATGAAAAG, reverse GGGTCTTGGCTTAATGAGGTC) and exclusion (forward TCAATGAAAAGGCCAAGACC, reverse TCAATAGATAACGGCAGTCCTGT). *EPB49* primers: inclusion (forward GCCTGCAGAACGGAGAGG, reverse CTCAAGCCGCATCCGATCC) and exclusion (forward, GCCTGCAGATCTATCCCTATGAAAT, reverse CTCAAGCCGCATCCGATCC).

***In vitro* splicing**

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A pre-mRNA substrate transcribed from the AdML derivative HMS388 was used in all splicing reactions (Jurica et al. 2002; Reichert et al. 2002). Site-directed mutagenesis was used to generate T or C at the -3 nucleotide of the 3' splice site. We linearized the DNA template using BamH1. T7 run-off transcription was used to generate G(5′)ppp(5′)G-capped radiolabeled pre-mRNA using UTP [α-^{32}P]. K562 nuclear extracts were isolated following a published protocol (Sakaki et al. 2012; Folco et al. 2012) with a minor modification (high salt buffer contains 20 mM HEPES pH7.9, 1.2 M KCl, 1.5 mM MgCl2, 25% glycerol, 0.5 mM DTT, and 0.2 mM PMSF). We incubated 10 nM pre-mRNA substrates in standard splicing conditions: 60 mM potassium glutamate, 3 mM magnesium acetate, 2 mM ATP, 5 mM creatine phosphate, 0.05 mg/mL tRNA, and 40% K562 nuclear extract for 1hr at 30°C. RNA was extracted using phenol/chloroform/isoamyl and precipitated with ethanol. RNA species were separated in a 12% denaturing polyacrylamide gel and visualized using a phosphoimager. For quantification in Figure 5, each species was normalized by subtracting the background and then dividing by the number of uracil nucleotides in that species. The percentage of the second step products was calculated by dividing the second step species (spliced mRNA and lariat intron) by the total of all species in the lane.

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 internally 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-

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 RNA-seq data from K562 cells has been deposited into the NCBI GEO database (accession number GSE58871).

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), Ellison Medical Foundation AG-NS-1030-13 (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

JOI designed the molecular genetics and biochemistry experiments. AR designed the cell culture and *U2AF1* expression strategies. JOI, AR, BH, MEM, and ASZ performed experimental work,

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

including cloning, cell culture, and flow cytometry. RKB and PB performed computational analyses 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.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

FIGURE LEGENDS

Figure 1. *U2AF1* mutations contribute to splicing programs in AML.

(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) Heat map illustrating similarity of alternative splicing programs in AML transcriptomes. Dendrogram is from an unsupervised cluster analysis based on cassette exon inclusion levels. Blue, samples with *U2AF1* mutations.

(D) *U2AF1* mutant allele expression as a percentage of total *U2AF1* mRNA in AML transcriptomes. Numbers above bars indicate the ratio of mutant to WT allele expression.

Figure 2. *U2AF1* mutations alter splicing, but do not cause splicing failure.

(A) Western blots showing levels of FLAG-tagged 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).

(B) *U2AF1* mutant allele expression as a percentage of total *U2AF1* mRNA in K562 cells.

(C) Heat map of K562 cells expressing *U2AF1* mutant alleles. Dendrogram is from an unsupervised cluster analysis based on cassette exon inclusion levels.

(D) *U2AF1* mutation-dependent changes in splicing for AML S34 vs. WT patients, K562 S34F or S34Y vs. WT expression, K562 Q157P or Q157R vs. WT expression, and K562 KD vs. control KD. Percentages indicate the fraction of mutation-dependent splicing changes falling into each category of splicing event.

(E) Levels of cassette exon inclusion in K562 cells expressing WT or S34Y *U2AF1*. N, numbers of alternatively spliced cassette exons with increased/decreased inclusion. Percentages, fraction of alternatively spliced cassette exons that are affected by S34Y expression. Events that do not change significantly are rendered transparent. Plot restricted to cassette exon events that are predicted to not induce nonsense-mediated decay (NMD).

(F) Levels of NMD-inducing isoforms of cassette exon events in K562 cells expressing WT or S34Y *U2AF1*.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(G) Levels of NMD-inducing isoforms of cassette exon events in AML transcriptomes. Distance from the center measures the splicing dissimilarity between each AML transcriptome and the average of all *U2AF1* WT samples, defined as the sum of absolute differences in expression of NMD-inducing isoforms.

Figure 3. *U2AF1* mutations affect genes involved in disease-relevant cellular processes.

(A) Overlap between mutation-dependent differential splicing in AML S34F/Y patients, K562 S34F/Y cells, and K562 Q157P/R cells. Overlap taken at the level of specific events (left) or genes containing differentially spliced events (right). Percentages, the fraction of differentially spliced events (left) or genes containing differentially spliced events (right) in S34F/Y AML transcriptomes that are similarly differentially spliced in K562 cells expressing S34F/Y or Q157P/R U2AF1 alleles.

(B) *DNMT3B* gene structure and protein domains (UniProt Consortium 2012). Upstream 5' UTR not shown. PWWP, Pro-Trp-Trp-Pro domain. ADD, ATRX-DNMT3-DNMT3L domain. Red stop sign, stop codon.

(C-D) Inclusion of *DNMT3B* cassette exons. Error bars, 95% confidence intervals as estimated from read coverage levels by MISO (Katz et al. 2010).

(E) Inclusion of *H2AFY* cassette exon.

(F) Cassette exon at 3' end of *ATR*. Conservation is phastCons (Siepel et al. 2005) track from UCSC (Meyer et al. 2013).

(G) Inclusion of cassette exon in *ATR*.

(H) Usage of intron-proximal 3' splice site of *CASP8*.

(I) Inclusion of cassette exon in *FANCA*.

Figure 4. *U2AF1* mutations alter 3' splice site consensus sequences.

(A) Consensus 3' splice sites of cassette exons with increased or decreased inclusion in *U2AF1* mutant relative to WT AML transcriptomes. 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 with increased or decreased inclusion in each sample. Data for all U2AF1 mutant samples shown in Figure S5.

(B) As (A), but for K562 cells expressing the indicated mutant allele vs. WT.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(C) As (A), but for K562 cells following *U2AF1* KD or control KD.

(D) 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 affected by mutant allele expression and exons repressed versus promoted by *U2AF1* KD.

Figure 5. *U2AF1* mutations cause sequence-dependent changes in 3' splice site recognition.

(A) Schematic of *ATR* minigene (top) and inclusion of *ATR* cassette exon transcribed from minigenes with A/C/G/T at the -3 position of the 3' splice site in K562 cells expressing WT or S34Y *U2AF1* (bottom). Error bars, standard deviation from biological triplicates.

(B) Schematic of *EPB49* minigene (top) and inclusion of *EPB49* cassette exon transcribed from minigenes with A/C/G/T at the +1 position of the 3' splice site in K562 cells expressing WT or Q157R *U2AF1* (bottom).

(C) Schematic of AdML pre-mRNA substrate used for *in vitro* splicing (top) and *in vitro* splicing of AdML substrate incubated with nuclear extract from K562 cells expressing WT or S34Y *U2AF1* (bottom). Percentages are the fraction of second step products (spliced mRNA and lariat intron) relative to all RNA species after 60 minutes of incubation. RNA, input radiolabeled RNA. GG, pre-mRNA with the AG dinucleotide replaced by GG to illustrate the first step product of splicing. Black dot, exonucleolytic “chew back” product of the lariat intermediate.

Figure 6. 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.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SUPPORTING INFORMATION

Figure S1. *U2AF1* mutant allele expression does not cause splicing failure.

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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* mutations are associated with altered 3' splice site consensus sequences in AML transcriptomes.

As Figure 4A, but for all seven AML samples with *U2AF1* mutations.

Figure S6. *U2AF1* WT AML transcriptomes do not exhibit altered 3' splice site consensus sequences.

As Figure 4A, but for seven randomly chosen AML samples without *U2AF1* mutations.

File S1. Differentially spliced events in AML samples with S34 mutations.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Splicing events that are differentially spliced in AML samples with S34F/Y mutations versus WT samples. 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", difference in isoform ratio between the two sample groups; "pval", *p*-value computed with the Mann-Whitney U test for group comparisons; "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 S34F.
As File S1, but for K562 cells expressing S34F versus WT *U2AF1*. Here, the "deltaPsi" column specifies the difference in isoform ratio between the mutant vs. WT cells, and "pval" is replaced by "bayesFactor", the Bayes factor associated with the sample comparison.

File S3. Differentially spliced events in K562 cells expressing S34Y.
As File S2, but for S34Y expression.

File S4. Differentially spliced events in K562 cells expressing Q157P.
As File S2, but for Q157P expression.

File S5. Differentially spliced events in K562 cells expressing Q157R.
As File S2, but for Q157R expression.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

File S6. 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.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TABLES

name	description	name	description
ABI1	abl-interactor 1	*MRPS28*	mitochondrial ribosomal protein S28
AGTPBP1	ATP/GTP binding protein 1	*MTA1*	metastasis associated 1
AKAP9	A kinase (PRKA) anchor protein (yotiao) 9	*MTL5*	metallothionein-like 5, testis-specific (tesmin)
AL589743.1	NA	*MYNN*	myoneurin
ALG2	asparagine-linked glycosylation 2, alpha-1,3-mannosyltransferase homolog (S. cerevisiae)	*N4BP2*	NEDD4 binding protein 2
ANKMY1	ankyrin repeat and MYND domain containing 1	*NCAPG2*	non-SMC condensin II complex, subunit G2
ANKRD36	ankyrin repeat domain 36	*NOM1*	nucleolar protein with MIF4G domain 1
ANKRD42	ankyrin repeat domain 42	*NPIP*	nuclear pore complex interacting protein
ARHGEF11	Rho guanine nucleotide exchange factor (GEF) 11	*NT5C3*	5'-nucleotidase, cytosolic III
ASPM	asp (abnormal spindle) homolog, microcephaly associated (Drosophila)	*ODF2L*	outer dense fiber of sperm tails 2-like
ATAD3B	ATPase family, AAA domain containing 3B	*OSBPL3*	oxysterol binding protein-like 3
ATF2	activating transcription factor 2	*PACRGL*	PARK2 co-regulated-like
ATXN2	ataxin 2	*PAPD7*	PAP associated domain containing 7
B3GALNT2	beta-1,3-N-acetylgalactosaminyltransferase 2	*PCM1*	pericentriolar material 1
BAZ1A	bromodomain adjacent to zinc finger domain, 1A	*PHF7*	PHD finger protein 7
BCCIP	BRCA2 and CDKN1A interacting protein	*PIGG*	phosphatidylinositol glycan anchor biosynthesis, class G
BCOR	BCL6 corepressor	*PILRB*	paired immunoglobin-like type 2 receptor beta
BIRC6	baculoviral IAP repeat containing 6	*PKD1P1*	NPIP-like protein 1
BPTF	bromodomain PHD finger transcription factor	*PKP4*	plakophilin 4
C10orf137	chromosome 10 open reading frame 137	*PLEKHM2*	pleckstrin homology domain containing, family M (with RUN domain) member 2
C17orf61-PLSCR3	Uncharacterized protein	*POLA1*	polymerase (DNA directed), alpha 1, catalytic subunit
C17orf62	chromosome 17 open reading frame 62	*POLD3*	polymerase (DNA-directed), delta 3, accessory subunit
C1orf63	chromosome 1 open reading frame 63	*PRKAR2A*	protein kinase, cAMP-dependent, regulatory, type II, alpha
C22orf39	chromosome 22 open reading frame 39	*PRRC2C*	proline-rich coiled-coil 2C
C9orf142	chromosome 9 open reading frame 142	*PTDSS2*	phosphatidylserine synthase 2
CAPN7	calpain 7	*RABGGTB*	Rab geranylgeranyltransferase, beta subunit
CAPRIN2	caprin family member 2	*RBM12*	RNA binding motif protein 12
CASP8	caspase 8, apoptosis-related cysteine peptidase	*RBM5*	RNA binding motif protein 5
CBWD2	COBW domain containing 2	*RDH13*	retinol dehydrogenase 13 (all-trans/9-cis)
CCDC138	coiled-coil domain containing 138	*REV1*	REV1, polymerase (DNA directed)
CCDC14	coiled-coil domain containing 14	*RHOT1*	ras homolog family member T1
CCP110	centriolar coiled coil protein 110kDa	*RINT1*	RAD50 interactor 1
CD47	CD47 molecule	*RNF216*	ring finger protein 216
CDCA7	cell division cycle associated 7	*RP11-1415C14.4*	NA
CHCHD7	coiled-coil-helix-coiled-coil-helix domain containing 7	*RPRD2*	regulation of nuclear pre-mRNA domain containing 2
CNOT2	CCR4-NOT transcription complex, subunit 2	*RTFDC1*	replication termination factor 2 domain containing 1
COG1	component of oligomeric golgi complex 1	*SAC3D1*	SAC3 domain containing 1
CSNK1E	casein kinase 1, epsilon	*SCLY*	selenocysteine lyase
DCUN1D4	DCN1, defective in cullin neddylation 1, domain containing 4 (S. cerevisiae)	*SEC31B*	SEC31 homolog B (S. cerevisiae)
DDX26B	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 26B	*SETD4*	SET domain containing 4
DHX32	DEAH (Asp-Glu-Ala-His) box polypeptide 32	*SNHG16*	small nucleolar RNA host gene 16 (non-protein coding)
DMTF1	cyclin D binding myb-like transcription factor 1	*SPPL2A*	signal peptide peptidase like 2A

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DNHD1	dynein heavy chain domain 1
DNM1L	dynamin 1-like
DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta
DPP9	dipeptidyl-peptidase 9
DRAM2	DNA-damage regulated autophagy modulator 2
DROSHA	drosha, ribonuclease type III
ENOSF1	enolase superfamily member 1
ENTPD6	ectonucleoside triphosphate diphosphohydrolase 6 (putative)
FAM219B	family with sequence similarity 219, member B
GIT2	G protein-coupled receptor kinase interacting ArfGAP 2
GPCPD1	glycerophosphocholine phosphodiesterase GDE1 homolog (S. cerevisiae)
GTF2I	general transcription factor IIi
HDAC10	histone deacetylase 10
HERC2	HECT and RLD domain containing E3 ubiquitin protein ligase 2
HNRNPH1	heterogeneous nuclear ribonucleoprotein H1 (H)
HPS1	Hermansky-Pudlak syndrome 1
IKBIP	IKBKB interacting protein
KDM4B	lysine (K)-specific demethylase 4B
KDM4C	lysine (K)-specific demethylase 4C
KLC1	Kinesin light chain 1
LTBP3	latent transforming growth factor beta binding protein 3
LUC7L3	LUC7-like 3 (S. cerevisiae)
MAP4K2	mitogen-activated protein kinase kinase kinase kinase 2
MAPK9	mitogen-activated protein kinase 9
MELK	maternal embryonic leucine zipper kinase
METTL22	methyltransferase like 22
MNAT1	menage a trois homolog 1, cyclin H assembly factor (Xenopus laevis)
MPHOSPH9	M-phase phosphoprotein 9
SRRM1	serine/arginine repetitive matrix 1
SRRM2	serine/arginine repetitive matrix 2
SRRT	serrate RNA effector molecule homolog (Arabidopsis)
ST3GAL3	ST3 beta-galactoside alpha-2,3-sialyltransferase 3
STRADA	STE20-related kinase adaptor alpha
TAF1	TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250kDa
TAF1D	TATA box binding protein (TBP)-associated factor, RNA polymerase I, D, 41kDa
TBC1D5	TBC1 domain family, member 5
THAP9-AS1	NA
TMEM116	transmembrane protein 116
TMEM5	transmembrane protein 5
TNRC18	trinucleotide repeat containing 18
TP53BP1	tumor protein p53 binding protein 1
TPP2	tripeptidyl peptidase II
TRMT13	tRNA methyltransferase 13 homolog (S. cerevisiae)
TTN-AS1	NA
TUBGCP4	tubulin, gamma complex associated protein 4
VPS41	vacuolar protein sorting 41 homolog (S. cerevisiae)
VTI1A	vesicle transport through interaction with t-SNAREs 1A
WDR33	WD repeat domain 33
WDR6	WD repeat domain 6
WHSC1	Wolf-Hirschhorn syndrome candidate 1
WRNIP1	Werner helicase interacting protein 1
ZDHHC16	zinc finger, DHHC-type containing 16
ZNF195	zinc finger protein 195
ZNF251	zinc finger protein 251
ZNF514	zinc finger protein 514
ZNF559	zinc finger protein 559

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

Genes that contain events that are differentially spliced in the AML S34 samples (versus WT samples), K562 S34 samples (versus WT), and K562 Q157 samples (versus WT). Descriptions taken from Ensembl.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

Brooks AN, Choi PS, de Waal L, Sharifnia T, Imielinski M, Saksena G, Pedamallu CS, Sivachenko A, Rosenberg M, Chmielecki J, et al. 2014. A Pan-Cancer Analysis of Transcriptome Changes Associated with Somatic Mutations in U2AF1 Reveals Commonly Altered Splicing Events. *PLoS ONE* **9**: e87361.

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.

Folco EG, Lei H, Hsu JL, Reed R. 2012. Small-scale nuclear extracts for functional assays of gene-expression machineries. *J Vis Exp*.

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*.

Guth S, Tange TØ, Kellenberger E, Valcárcel J. 2001. Dual function for U2AF(35) in AG-dependent pre-mRNA splicing. *Mol Cell Biol* **21**: 7673–7681.

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,

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

Jurica MS, Licklider LJ, Gygi SR, Grigorieff N, Moore MJ. 2002. Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. *RNA* **8**: 426–439.

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.

Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, Larson DE, Kandoth C, Payton JE, Baty J, Welch J, et al. 2010. DNMT3A mutations in acute myeloid leukemia. *N Engl J Med* **363**: 2424–2433.

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

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,

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

Reichert VL, Le Hir H, Jurica MS, Moore MJ. 2002. 5' exon interactions within the human spliceosome establish a framework for exon junction complex structure and assembly. *Genes Dev* **16**: 2778–2791.

Sakaki K, Yoshina S, Shen X, Han J, Desantis MR, Xiong M, Mitani S, Kaufman RJ. 2012. RNA surveillance is required for endoplasmic reticulum homeostasis. *Proceedings of the National Academy of Sciences* **109**: 8079–8084.

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.

Taskesen E, Havermans M, van Lom K, Sanders MA, van Norden Y, Bindels E, Hoogenboezem R, Reinders MJT, Figueroa ME, Valk PJM, et al. 2014. Two splice factor mutant leukemia subgroups uncovered at the boundaries of MDS and AML using combined gene expression and DNA-methylation profiling. *Blood*.

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.

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

Walter MJ, Ding L, Shen D, Shao J, Grillot M, McLellan M, Fulton R, Schmidt H, Kalicki-Veizer J, O'Laughlin M, et al. 2011. Recurrent DNMT3A mutations in patients with myelodysplastic syndromes. *Leukemia* **25**: 1153–1158.

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

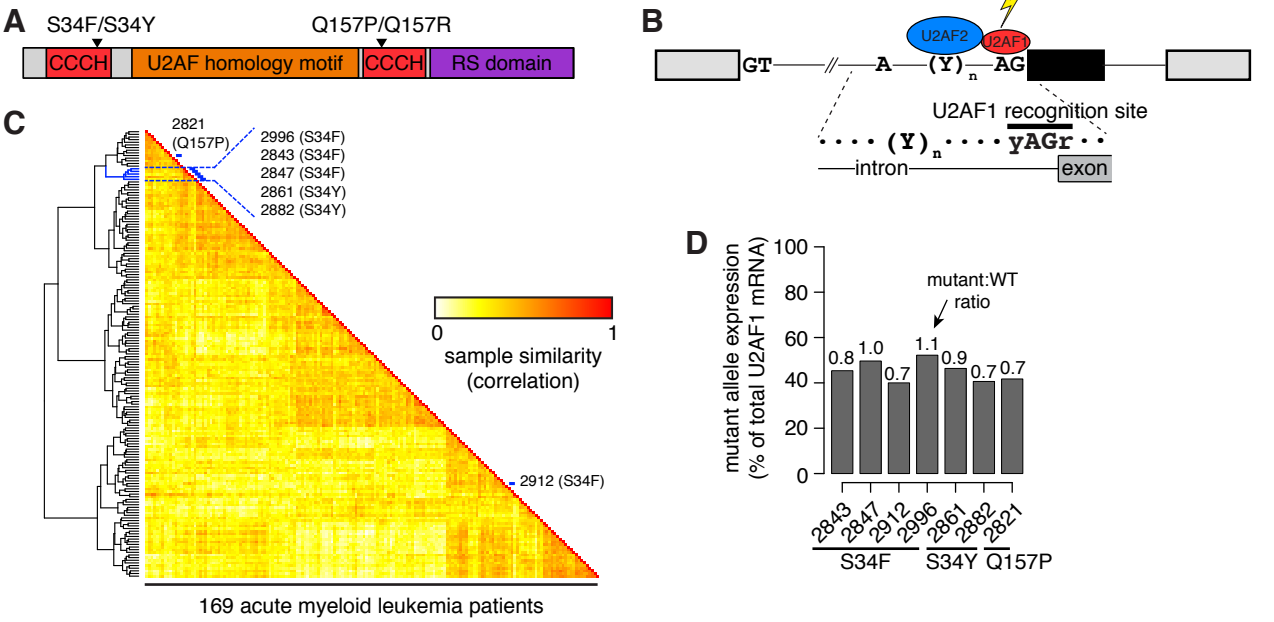
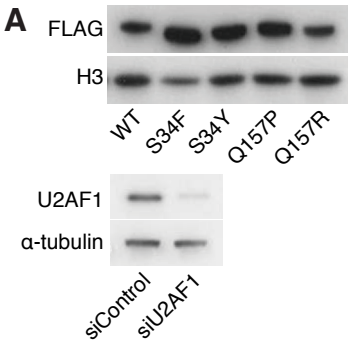
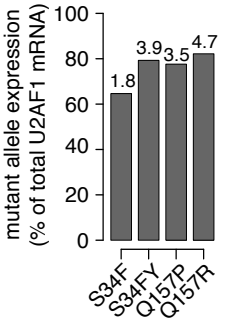
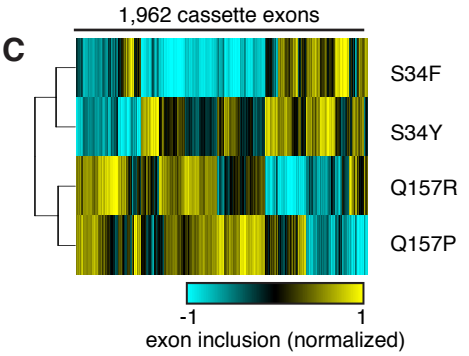

A
S34F/S34Y
Q157P/Q157R
CCCH
U2AF homology motif
CCCH
RS domain
B
U2AF2
U2AF1
GT
A
(Y)n
AG
U2AF1 recognition site
(Y)n
yAGr
intron
exon
C
2821
(Q157P)
2996 (S34F)
2843 (S34F)
2847 (S34F)
2861 (S34Y)
2882 (S34Y)
2912 (S34F)
0
1
sample similarity
(correlation)
169 acute myeloid leukemia patients
D
mutant allele expression
(% of total U2AF1 mRNA)
100
80
60
40
20
0
mutant:WT
ratio
0.8
1.0
0.7
1.1
0.9
0.7
0.7
2843
2847
2912
2996
2861
2882
2821
S34F
S34Y
Q157P

Figure 2

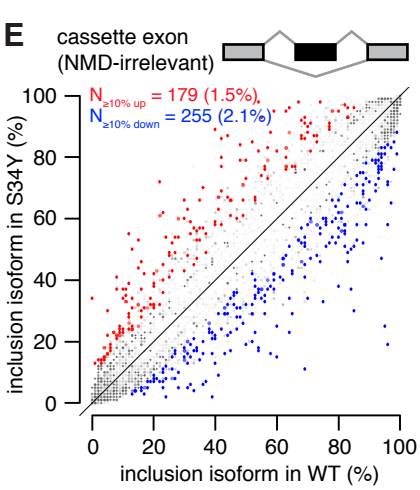


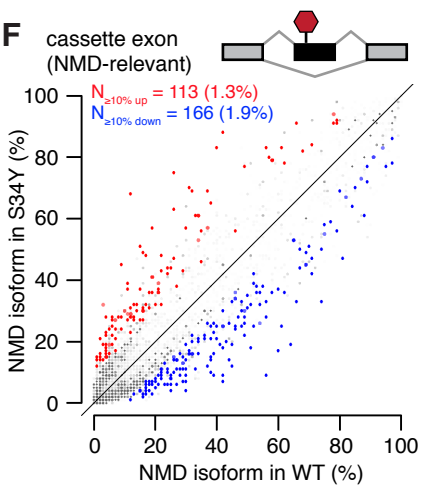




D

	AML patients	K562 cells		
	S34F/Y	S34F/Y	Q157P/R	KD
competing 5' splice sites	2.3%	3.6%	4.5%	5.1%
competing 3' splice sites	9.6%	10.0%	10.0%	10.0%
cassette exon	38.0%	41.0%	42.0%	42.0%
alternative splicing, constitutive	15.0%	19.0%	20.0%	22.0%
retained intron, alternative	4.6%	7.8%	5.7%	6.5%
retained intron, constitutive	30.0%	19.0%	18.0%	14%





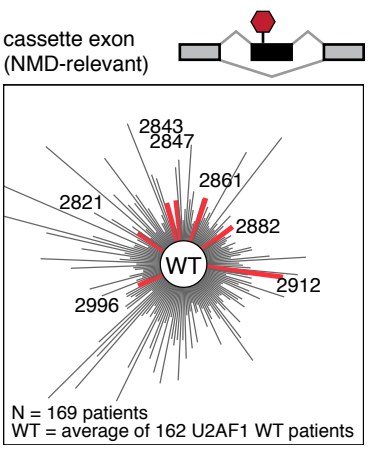


Figure 3

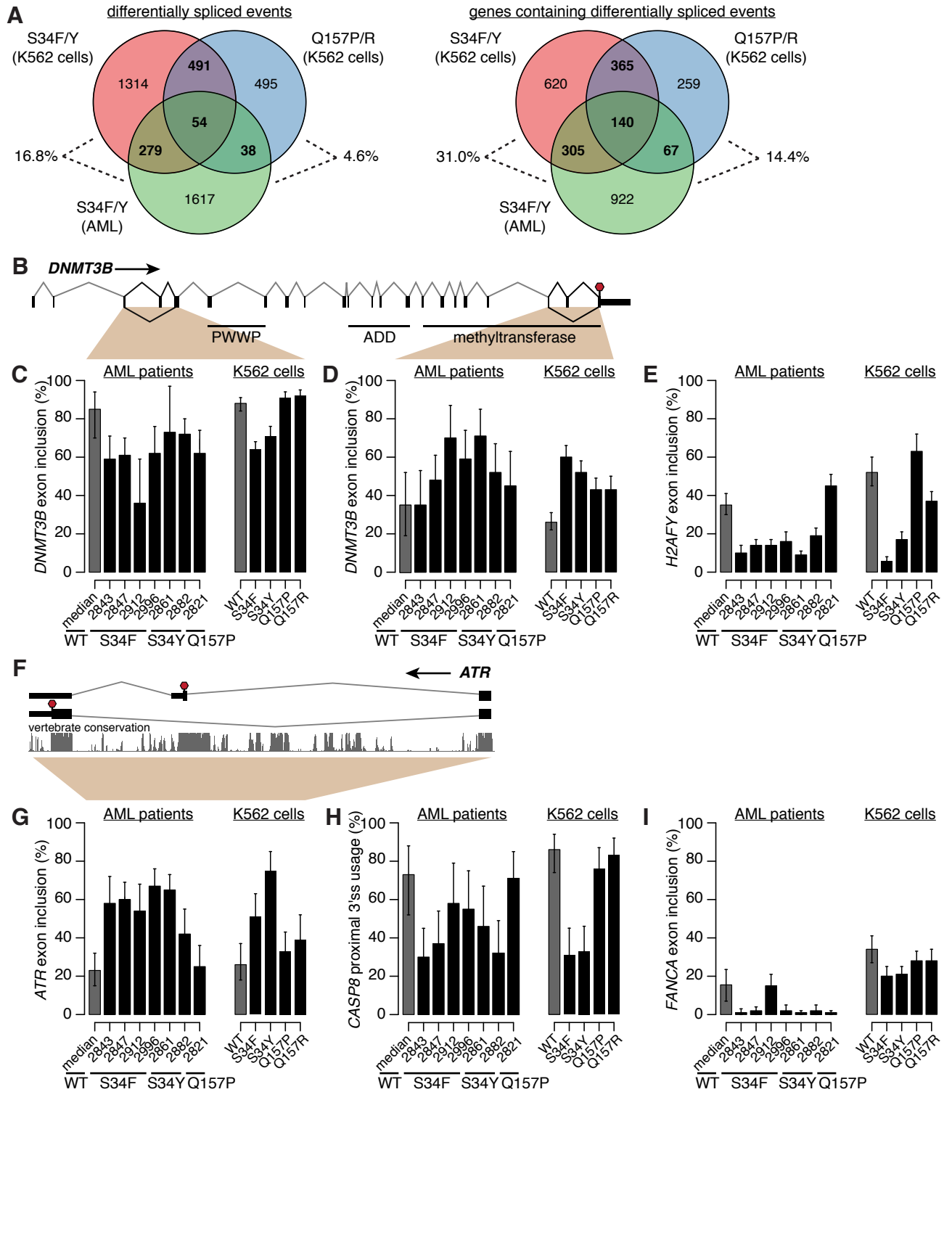
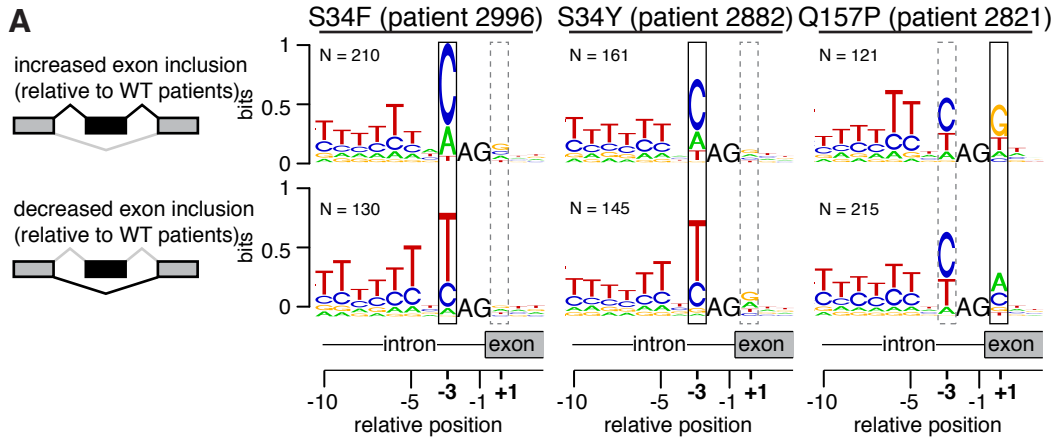
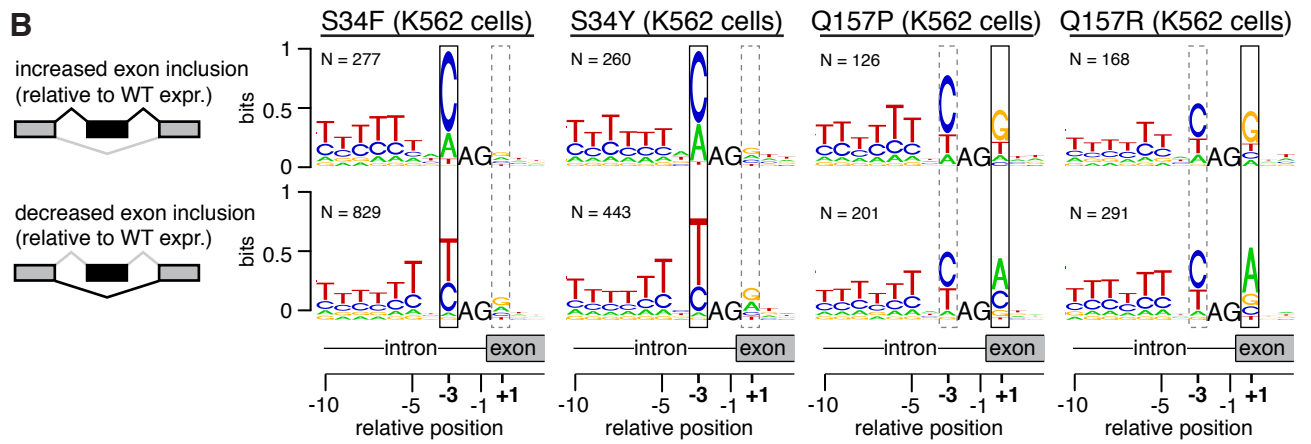
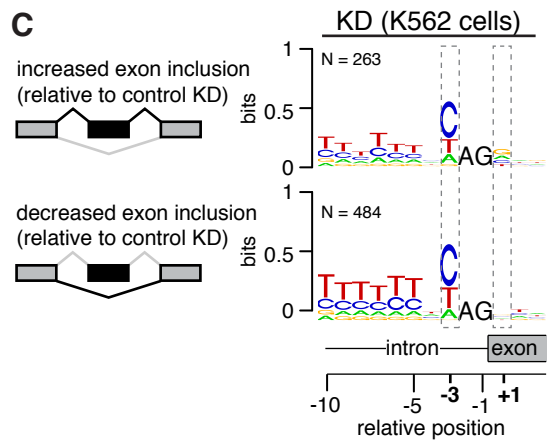

A
differentially spliced events
S34F/Y (K562 cells)
Q157P/R (K562 cells)
S34F/Y (AML)
1314
491
495
54
279
38
1617
16.8%
4.6%
genes containing differentially spliced events
620
365
259
140
305
67
922
31.0%
14.4%
B
DNMT3B
PWWP
ADD
methyltransferase
C
AML patients
K562 cells
DNMT3B exon inclusion (%)
median
2843
2847
2912
2996
2861
2882
2821
WT
S34F
S34Y
Q157P
Q157R
D
DNMT3B exon inclusion (%)
E
H2AFY exon inclusion (%)
F
ATR
vertebrate conservation
G
ATR exon inclusion (%)
H
CASP8 proximal 3'ss usage (%)
I
FANCA exon inclusion (%)

Figure 4



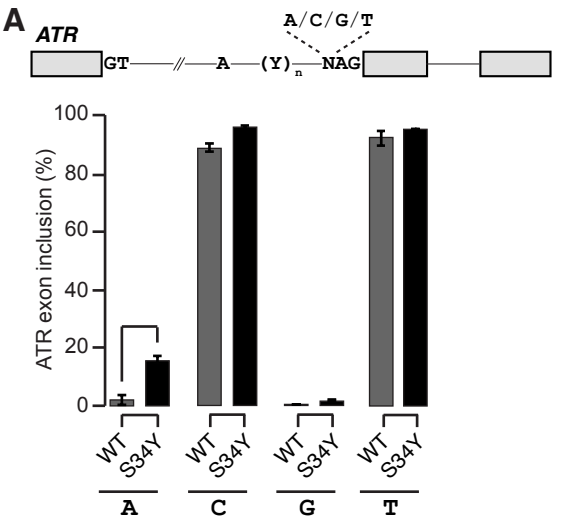


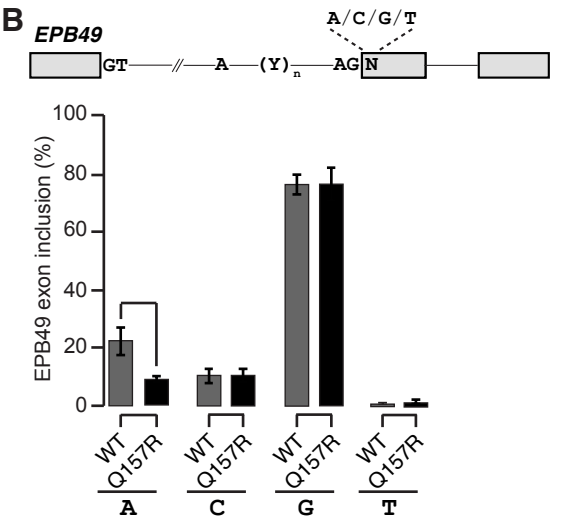


D

		U2AF1 KD		
		U2AF1-dependent	U2AF1-independent	enrichment for U2AF1 dependence
S34F		78	41	1.9X
		25	23	1.1X
S34Y		33	33	1.0X
		39	9	4.3X
Q157P		43	10	4.3X
		12	8	1.5X
Q157R		46	12	3.8X
		13	13	1.0X

Figure 5


A
ATR
A/C/G/T
GT
A
(Y)n
NAG
100
80
60
40
20
0
ATR exon inclusion (%)
WT
S34Y
A
C
G
T


B
EPB49
A/C/G/T
GT
A
(Y)n
AG
N
100
80
60
40
20
0
EPB49 exon inclusion (%)
WT
Q157R
A
C
G
T

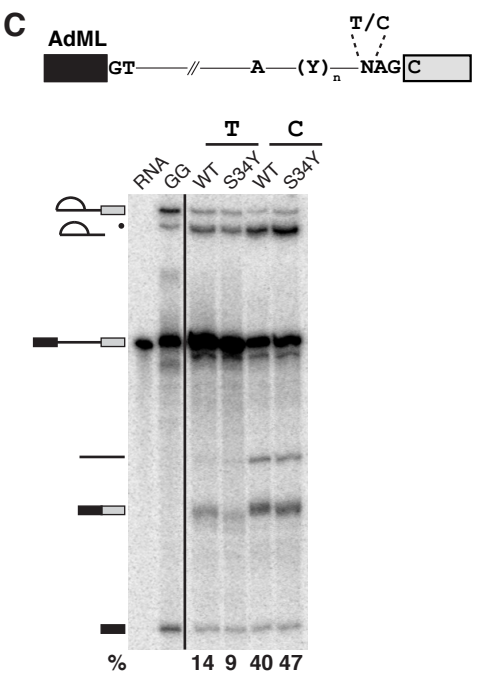
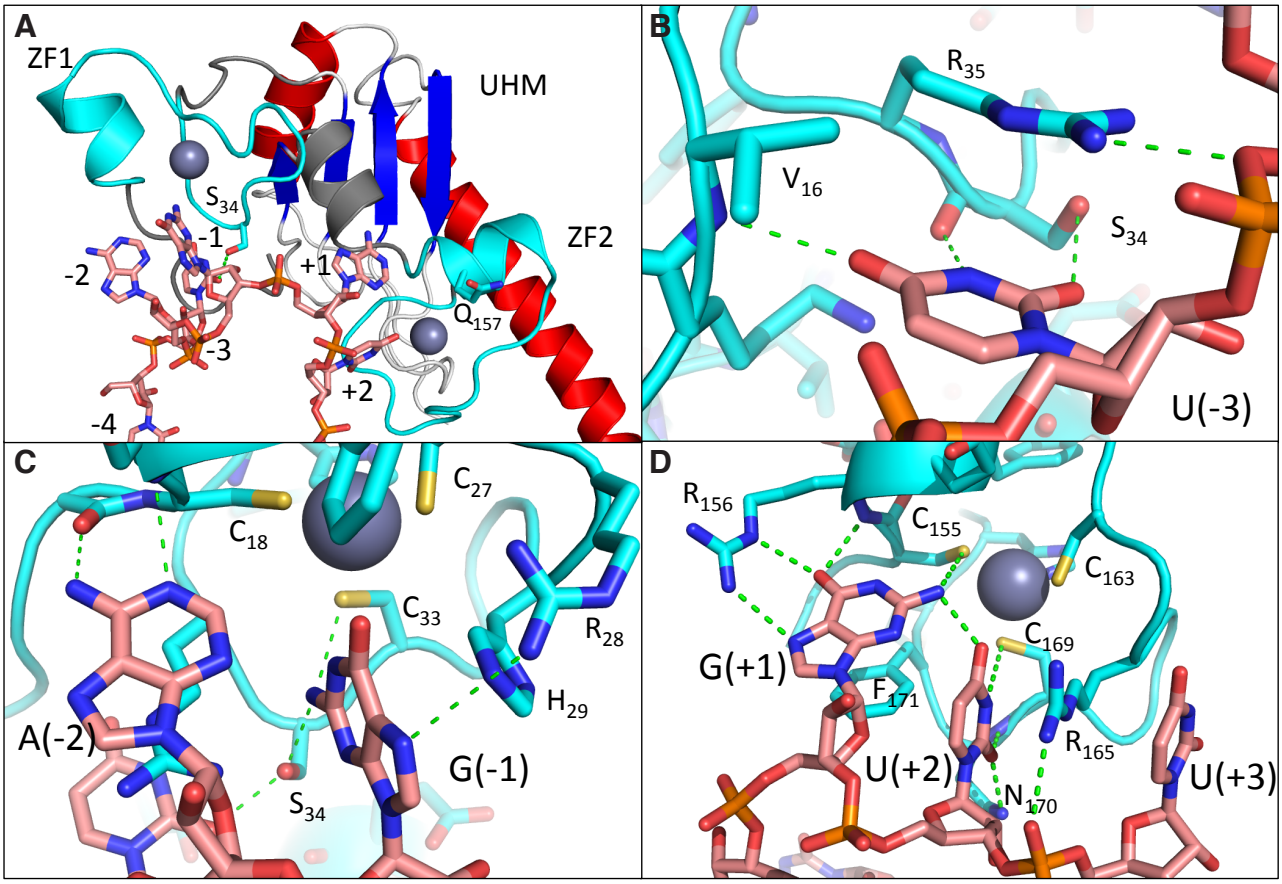

C
AdML
T/C
GT
A
(Y)n
NAG
C
T
C
RNA
GG
WT
S34Y
WT
S34Y
PP
%
14 9 40 47

Figure 6


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)

bioRxiv preprint doi: https://doi.org/10.1101/001107; this version posted June 28, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Figure S1

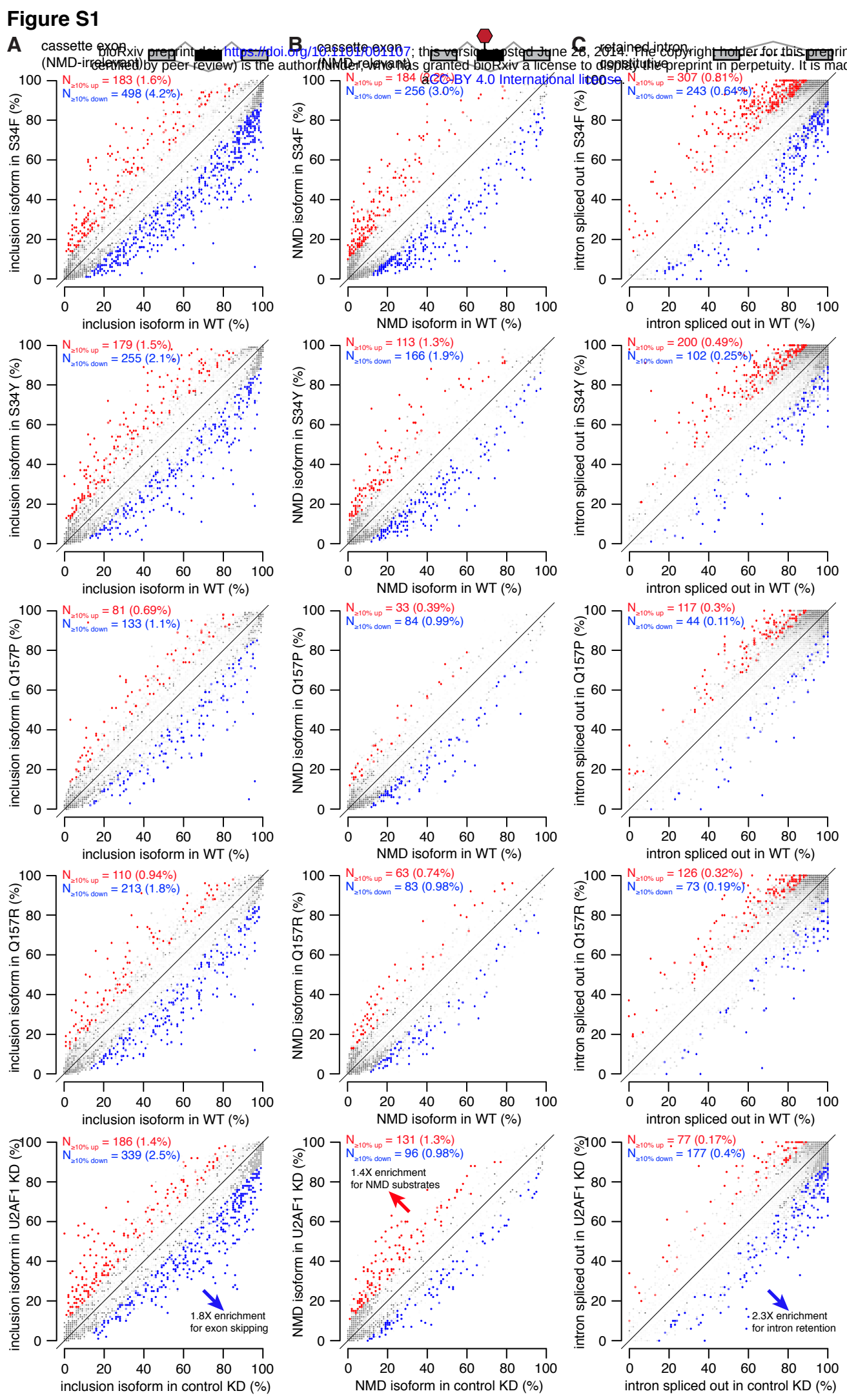


Figure S2

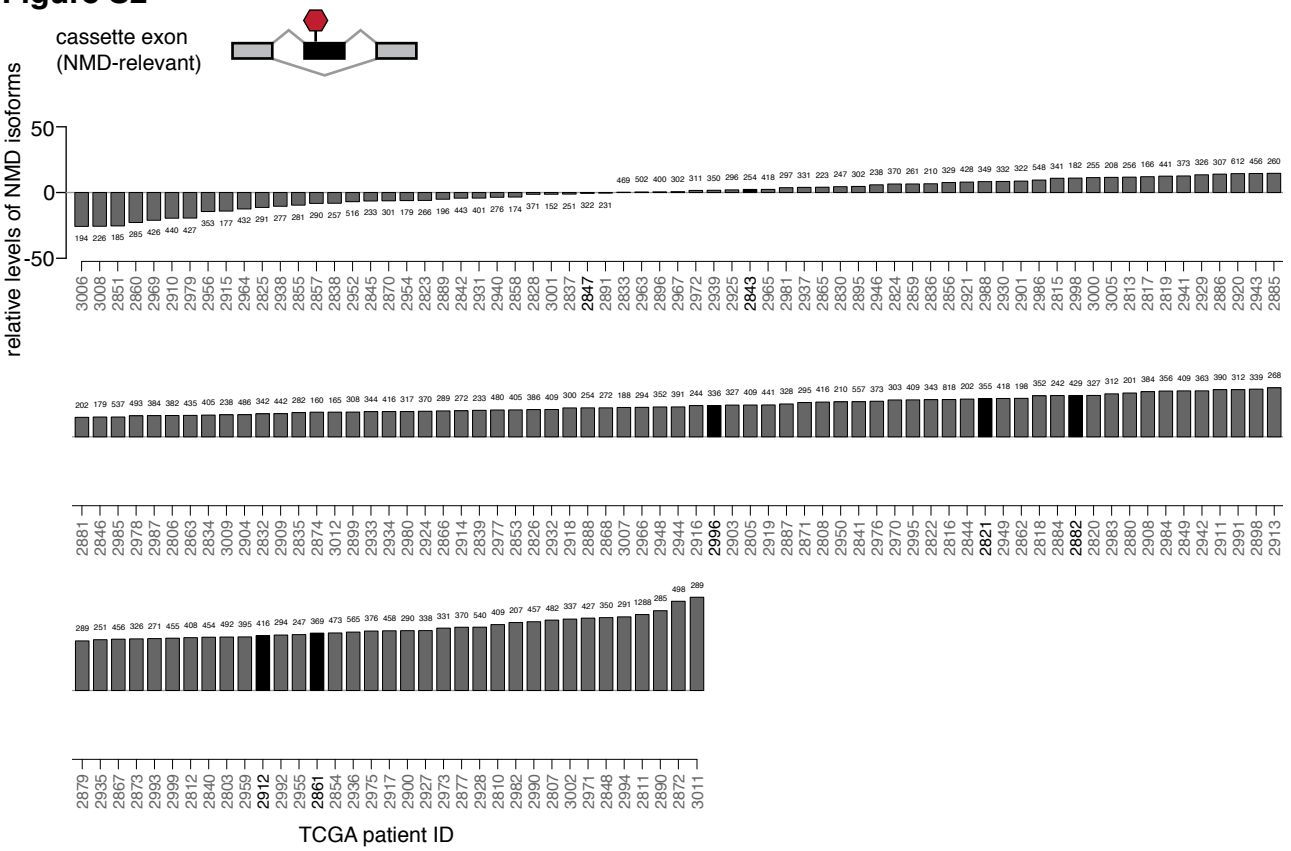

cassette exon
(NMD-relevant)
relative levels of NMD isoforms
50
0
-50
194 226 185 285 426 440 427 353 177 432 291 277 281 290 257 516 233 301 179 266 196 443 401 276 174 371 152 251 322 231 469 502 400 302 311 350 296 254 418 297 331 223 247 302 238 370 261 210 329 428 349 332 322 548 341 182 255 208 256 166 441 373 326 307 612 456 260
3006 3008 2851 2860 2969 2910 2979 2956 2915 2964 2825 2938 2855 2857 2838 2952 2845 2870 2954 2823 2889 2842 2931 2940 2858 2828 3001 2837 2847 2891 2833 2963 2896 2967 2972 2939 2925 2843 2965 2981 2937 2865 2830 2895 2946 2824 2859 2836 2856 2921 2988 2930 2901 2986 2815 2998 3000 3005 2813 2817 2819 2941 2929 2886 2920 2943 2885
202 179 537 493 384 382 435 405 238 486 342 442 282 160 165 308 344 416 317 370 289 272 233 480 405 386 409 300 254 272 188 294 352 391 244 336 327 409 441 328 295 416 210 557 373 303 409 343 818 202 355 418 198 352 242 429 327 312 201 384 356 409 363 390 312 339 268
2881 2846 2985 2978 2987 2806 2863 2834 3009 2904 2832 2909 2835 2874 3012 2899 2933 2934 2980 2924 2866 2914 2839 2977 2853 2826 2932 2918 2888 2868 3007 2966 2948 2944 2916 2996 2903 2805 2919 2887 2871 2808 2950 2841 2976 2970 2995 2822 2816 2844 2821 2949 2862 2818 2884 2882 2820 2983 2880 2908 2984 2849 2942 2911 2991 2898 2913
289 251 456 326 271 455 408 454 492 395 416 294 247 369 473 565 376 458 290 338 331 370 540 409 207 457 482 337 427 350 291 1288 285 498 289
2879 2935 2867 2873 2993 2999 2812 2840 2803 2959 2912 2992 2955 2861 2854 2936 2975 2917 2900 2927 2973 2877 2928 2810 2982 2990 2807 3002 2971 2848 2994 2811 2890 2872 3011
TCGA patient ID

Figure S3

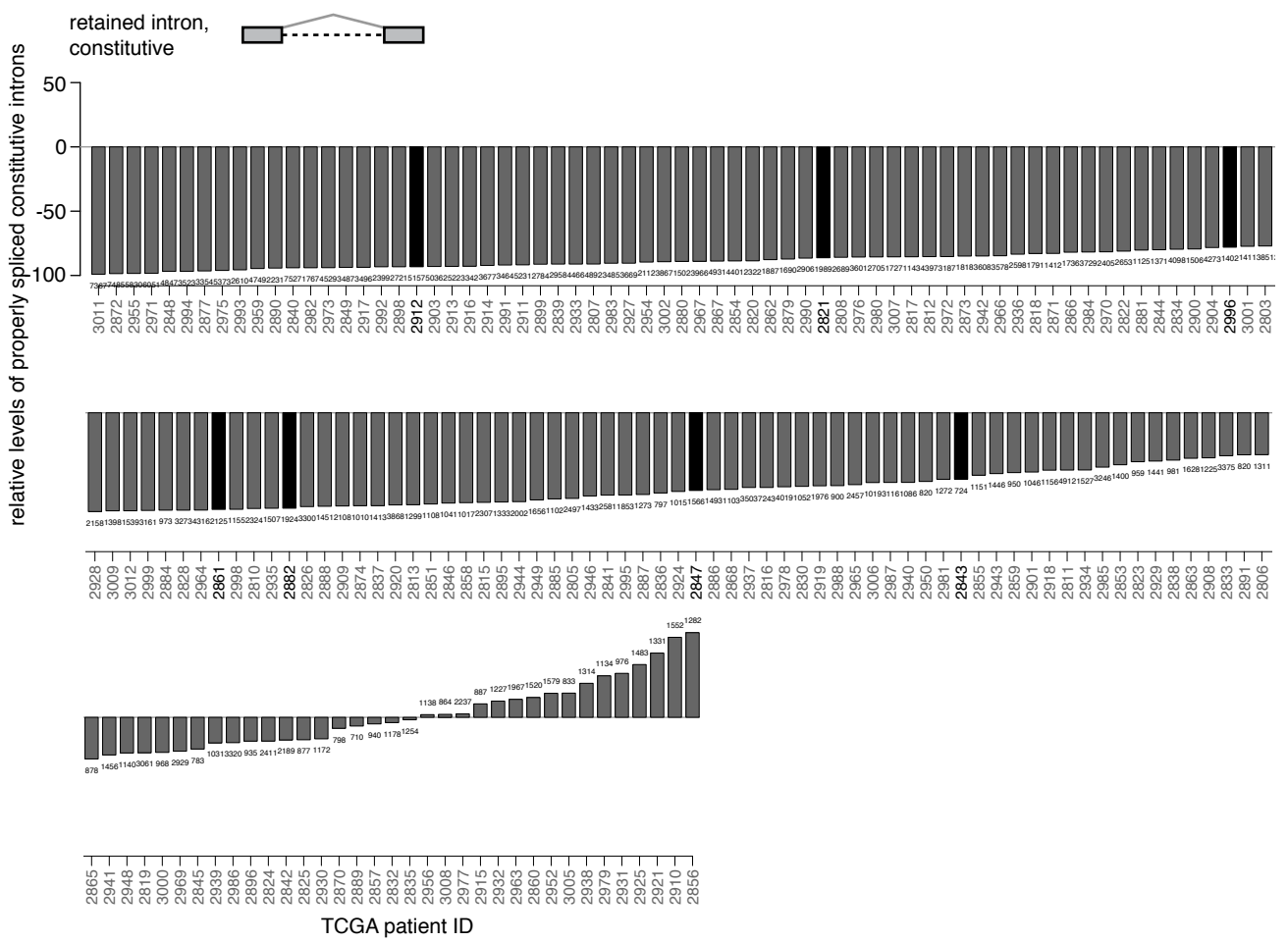

retained intron,
constitutive
relative levels of properly spliced constitutive introns
50
0
-50
-100
TCGA patient ID

Figure S4

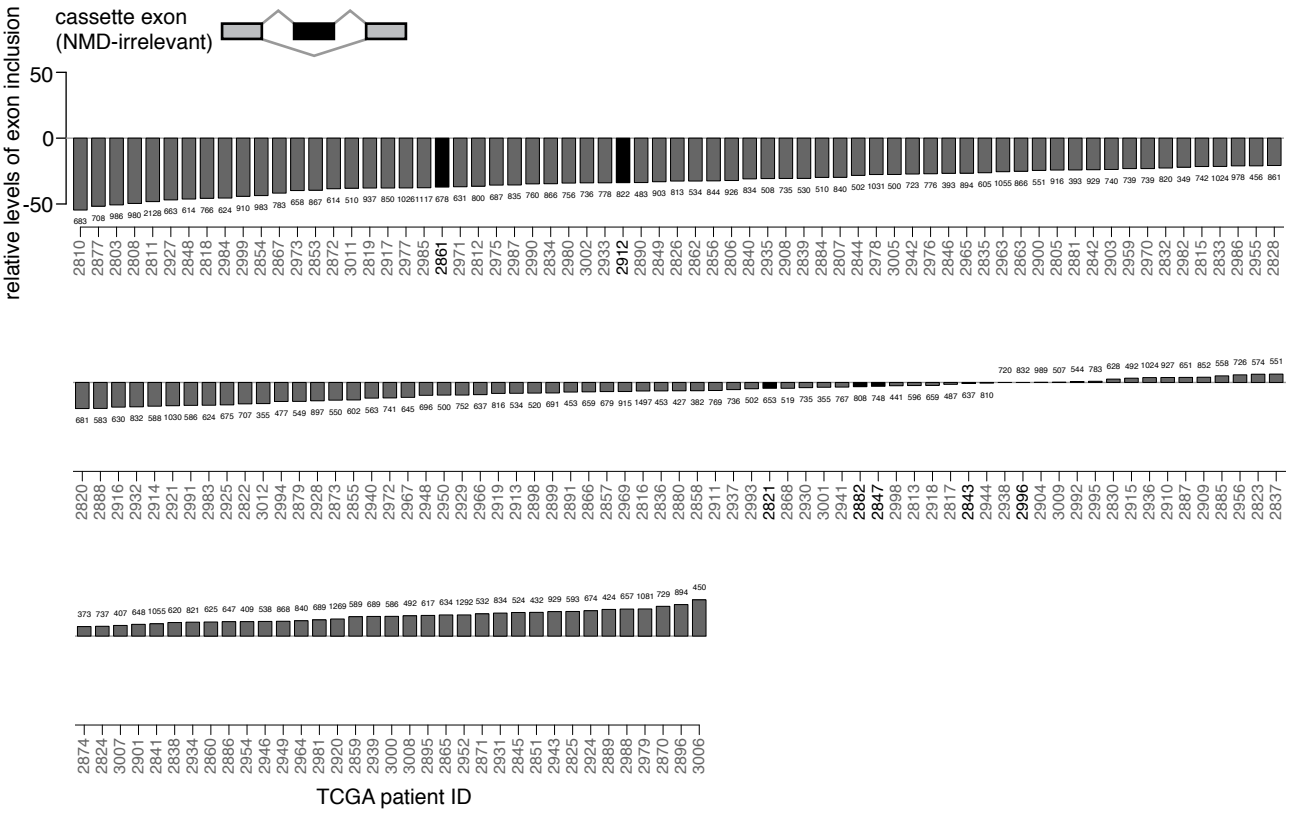

cassette exon
(NMD-irrelevant)
relative levels of exon inclusion
50
0
-50
TCGA patient ID

Figure S5

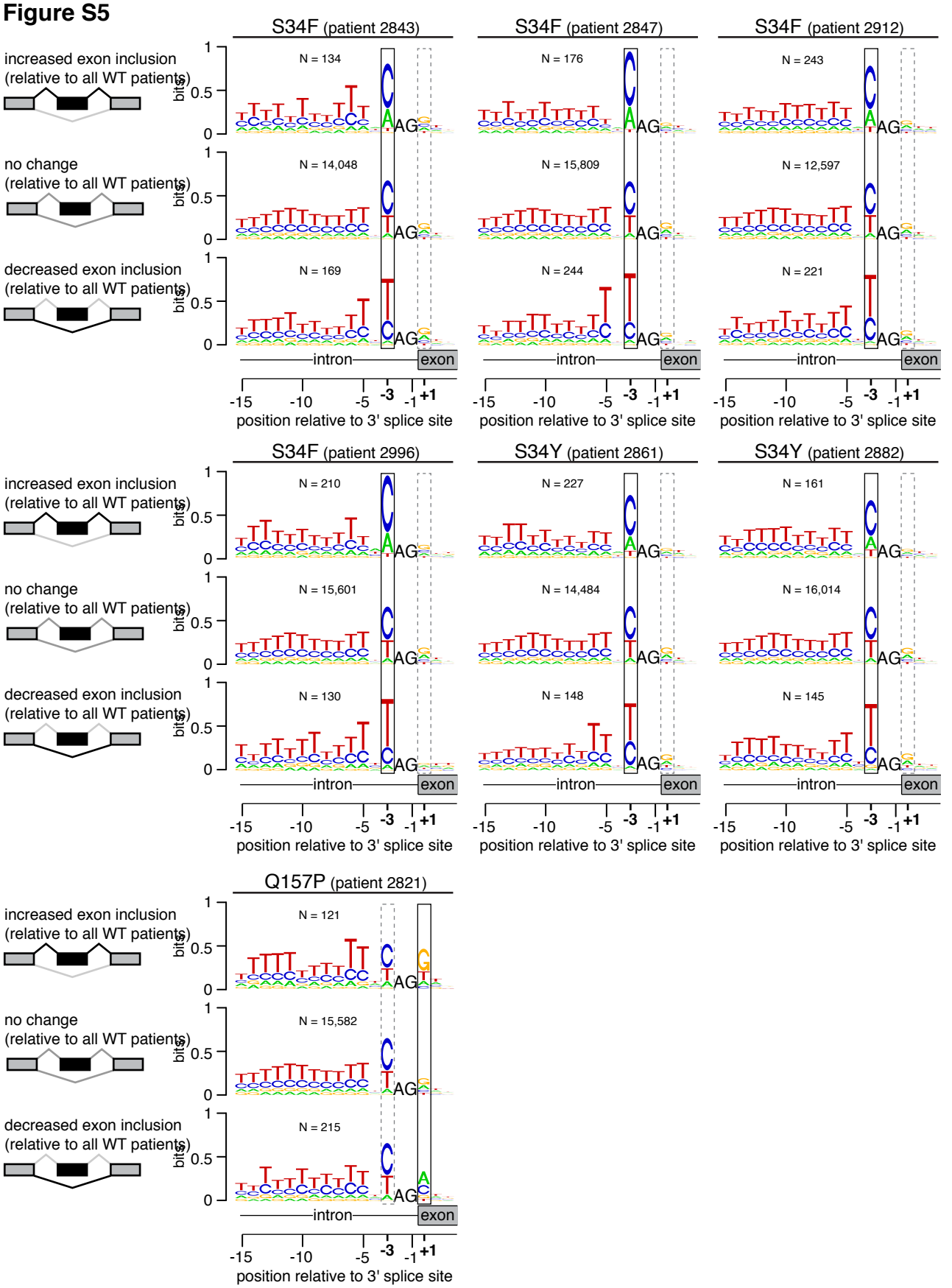
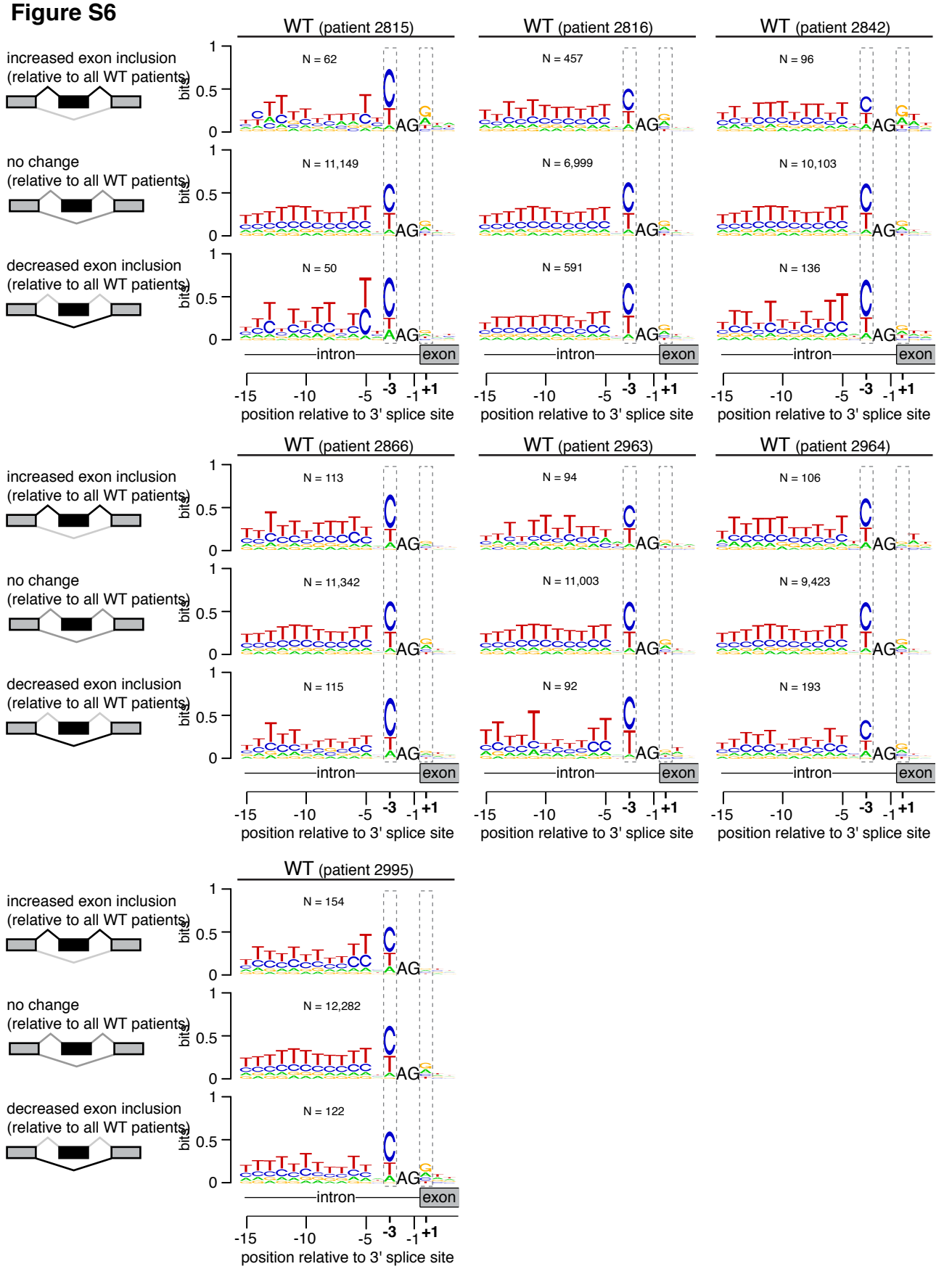

S34F (patient 2843)
S34F (patient 2847)
S34F (patient 2912)
S34F (patient 2996)
S34Y (patient 2861)
S34Y (patient 2882)
Q157P (patient 2821)
increased exon inclusion
(relative to all WT patients)
no change
(relative to all WT patients)
decreased exon inclusion
(relative to all WT patients)
N = 134
N = 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 S6


WT (patient 2815)
WT (patient 2816)
WT (patient 2842)
WT (patient 2866)
WT (patient 2963)
WT (patient 2964)
WT (patient 2995)
increased exon inclusion (relative to all WT patients)
no change (relative to all WT patients)
decreased exon inclusion (relative to all WT patients)
N = 62
N = 457
N = 96
N = 11,149
N = 6,999
N = 10,103
N = 50
N = 591
N = 136
N = 113
N = 94
N = 106
N = 11,342
N = 11,003
N = 9,423
N = 115
N = 92
N = 193
N = 154
N = 12,282
N = 122
bits
intron
exon
-15 -10 -5 -3 -1 +1
position relative to 3' splice site