PATTERNS OF MISSPLICING DUE TO SOMATIC U2AF1 MUTATIONS IN MYELOID NEOPLASMS

Bartłomiej Przychodzen, Andres Jerez, Kathryn Guinta, Mikkael A. Sekeres, Richard Padgett, Jaroslaw P. Maciejewski, Hideki Makishima

1Department of Translational Hematology and Oncology Research, Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH, USA;
2Department of Molecular Genetics, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA.

Key words: Splicing, U2AF1, U2AF35, MDS, AML

Running Title: Missplicing of genes in AML and MDS due to somatic U2AF1 mutations.

Scientific section: MYELOID NEOPLASIA

Corresponding author: Hideki Makishima MD., Ph.D.
Taussig Cancer Institute/R40
9500 Euclid Avenue
Cleveland OH USA 44195
Phone: 216-445-5962
FAX: 216-636-2498
E-mail: makishh@ccf.org
Key Points:

1. Gene specific, splicing patterns of U2AF1 mutants

2. Mechanism of selective splicing pattern of U2AF1 mutants

Abstract

Recently, recurrent mutations of spliceosomal genes were frequently identified in myeloid malignancies, as well as other types of cancers. One of these spliceosomal genes, *U2AF1*, was affected by canonical somatic mutations in aggressive type of myeloid malignancies. We hypothesized that *U2AF1* mutations causes defects of splicing (missplicing) in specific genes and that such misspliced genes might be important in leukemogenesis. We analyzed RNA deep-sequencing to compare splicing patterns of 201,837 exons between the cases with *U2AF1* mutations (N=6) and wild-type (N=14). We identified different alternative splicing patterns in 35 genes comparing cells with mutant and wild-type *U2AF1*. *U2AF1* mutations are associated with abnormal splicing of genes involved in functionally important pathways, such as cell cycle progression and RNA processing. In addition, many of these genes are somatically mutated or deleted in various cancers. Of note is that the alternative splicing patterns associated with *U2AF1* mutations were associated with specific sequence signals at the affected splice sites. These novel observations support the hypothesis that *U2AF1* mutations play a significant role in myeloid leukemogenesis due to selective missplicing of tumor-associated genes.
Introduction

Pre-mRNA splicing is one of the vital physiologic functions in eukaryotic gene expression. Most human genes are spliced in two or more patterns to produce mRNAs that encode protein variants, a process known as physiologic alternative splicing. Alternative exon usage is determined by the selection of splice sites and results in exon skipping or retention. Accordingly, alteration of the exon usage ratio alters the proportion of mRNA isoforms with and without the affected exon. For instance, exon skipping caused by alternative splicing has been found to be altered in various cancers.

Splicing is carried out by a complicated and dynamic molecular machine known as the spliceosome. Errors in splicing can be a result of somatic mutations of spliceosomal components leading to aberrant and potentially pathological mRNA isoform composition.

Recently, somatic mutations of several spliceosomal proteins (U2AF1, SRSF2, SF3B1) have been identified in myeloid malignancies, in particular myelodysplastic syndromes (MDS), MDS/myeloproliferative neoplasms (MDS/MPN), and secondary acute myeloid leukemia (sAML). While these mutations are functionally related through effects on the splicing machinery, the downstream consequences of these mutations may be diverse and involve different oncogenic pathways. These spliceosomal factor mutations are associated with specific pathomorphologic features, clinical phenotypes and coexisting somatic mutational patterns, suggesting that downstream consequences of individual mutations may be distinct. SRSF2 mutations are strongly associated with chronic myelomonocytic leukemia (CMML), mutations in U2AF1 are more common in advanced myelomonocytic leukemias with poor outcome, while mutations in SF3B1 are associated with the presence of ring sideroblasts, conveying a comparatively benign prognosis.

Based on the canonical location within the affected spliceosomal gene, these missense mutations are unlikely to be simply hypomorphic but, rather they appear to result in change of function. Here we explore the effects on patterns of alternative splicing due to mutations in the splicing factor U2AF1.

The emergence of spliceosomal mutations as a novel leukemogenesis mechanism raises several questions: what are the critical downstream target genes affected, what is the molecular context of these genes, and how they are specifically targeted by the mutant U2AF1 protein? In this study, we used next generation genomic platforms to investigate: i) U2AF1 mutant-specific splicing patterns, ii) specific genes affected by missplicing and iii) the coexistence of other molecular defects involving these misspliced genes in cancer.
METHODS

Patients

Tumor DNA was obtained from patients bone marrow. Informed consent for sample collection was obtained according to protocols approved by the Institutional Review Board and in accordance with the Declaration of Helsinki. Diagnoses of MDS, MDS/MPN, MPN, and sAML were confirmed and assigned according to World Health Organization (WHO) classification criteria. The clinical characteristics of patients investigated in this study are presented in Table 1.

DNA Sequencing

Selected exons of the \textit{U2AF1} gene were amplified and subjected to direct genomic sequencing using standard techniques on the ABI 3730xl DNA analyzer (Applied Biosystems) as previously described\textsuperscript{17-19}. Positive mutations were detected by bidirectional sequencing and confirmed using germline DNA obtained from non-clonal CD3+ T cell fraction. Whole exome capture was accomplished based on liquid phase hybridization of sonicated genomic DNA having 150 - 200bp of mean length to the bait cRNA library synthesized on magnetic beads (SureSelect\textsuperscript{®}, Agilent Technology), according to the manufacture’s protocol. SureSelct Human All Exon 50Mb kit was used for targeted, exome capture. The captured targets were subjected to massive sequencing using Illumina HiSeq 2000. Generation of .bam files with its preprocessing and detection of somatic point mutations / insertions and deletions was done as previously described\textsuperscript{7}. Additionally, for detailed analyses, exome sequencing data (N=197) on AML patients obtained through TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) were used.

Whole RNA Sequencing

We have used publically available RNAseq data from TCGA data portal for 97 patients (https://tcga-data.nci.nih.gov/tcga). We selected 6 cases harboring \textit{U2AF1} mutation (c.101C>T, p.S34F, N=4 and c.101C>A, p.S34Y, N=2) for which deep RNAseq\textsuperscript{20} data was available. We also selected 14 cases that were wild-type (WT) for any spliceosomal factor mutation. To further demonstrate specificity of \textit{U2AF1} mutations with respect to other spliceosomal mutations (\textit{SRSF2}, \textit{SF3B1}, \textit{U2AF26}) we selected 7 additional cases with mutations in these other spliceosomal factor genes.

Global Differential Splicing Pattern Analysis

We quantified exon inclusion ratios based on paired-end RNAseq data. SpliceTrap software (http://rulai.cshl.edu/splicetrap/) was used to quantify the frequency of inclusion of each exon \textsuperscript{21} and to extract counts of
paired end reads that span each exon junction in the genome. For this purpose, each exon was tested for inclusion or exclusion with respect to adjacent exons (Supplementary Figure 3). SpliceTrap considers individual exons in whole genome and is not limited by analysis of known repository of transcripts. This unbiased method is a suitable approach for possible novel discovery of unknown/unexpected splicing variants. Each exon was tested with respect to adjacent exons. Within each triplet, each exon was termed as A, B and C, exon B being the one screened for every triplet in the transcriptome. According to this method, we counted reads spanning between exon A/B, B/C and A/C where reads spanning the A/C junction reflect the proportion of mRNA missing exon B. The sum of reads between exons A/B and B/C divided by 2 reflects the proportion of mRNA that contains exon B. In order to estimate the frequency of exon B skipping, we divided the number of reads spanning A/C by half of the sum of the reads spanning A/B and B/C. By following these guidelines we extracted alternative splicing patterns for 20 patients (6 U2AF1 mutant patients and 14 spliceosomal WT patients). Using the frequency of skipped reads to represent the skipping ratio is independent from variation in coverage between different RNAseq samples and is normalization step itself. The unpaired T-test was used in order to assess the difference of exon usage between these two groups. For each exon tested we compared average exon usage between U2AF1 mutants and the WT group, with associated p-values generated. Statistical difference of p<0.0001 and average difference of ±15% in frequency of exon usage was considered as valid for an exon tested. Using this approach, we detected changes in exon skipping (excess of shorter mRNA missing an exon) as well as in exon retention (excess of longer mRNA incorporating an exon) (Supplementary Figure 3).

**Reverse Transcription-PCR Analysis of U2AF1 Mutant and WT patients**

RNA was extracted from bone marrow or peripheral blood mononuclear cells of patients with and without U2AF1 mutations by TRIzol (Invitrogen, Carlsbad, CA, USA). Reverse transcription-polymerase chain reaction (RT-PCR) was performed using a primer pair specific for detecting exon skipping (exon7; CEP164). 100ng of cDNA was amplified in 35 cycle RT-PCR reaction at annealing temperature of 60°C. The status of exon skipping or retention was determined by size differences determined by gel electrophoresis.

**Sequence analysis of regions around 3’ and 5’ splice site.**

Sequence information was extracted from adjacent 3’ and 5’ splice site for all exons that were surveyed for differential exon usage. For the 3’ splice sites, sequence was extracted from 20bp upstream to 3bp downstream of the intron/exon junction. For the 5’ splice sites, sequence was extracted from 3bp upstream to 5bp downstream of the exon/intron.
junction. Exons were divided into 3 groups according to exon usage levels (exon skipping 0-5%, 40-60% and 90-100%). All of the splice site sequences (Human Genome release hg19) were obtained through the Table Browser available at the UCSC genome browser website (http://www.genome.ucsc.edu/). Exon usage levels were obtained from all WT samples used in this study. Additionally we extracted sequence information for the genes that were found to have different exon usage level in U2AF1 mutants. These were divided into two subcategories – exons that were found to have an excess of exon skipping and exons that had an excess of exon retention. The last group of sequences was a randomly selected set of 1000 exons. Sequence logos were generated using the WebLogo22,23 online application (http://weblogo.berkeley.edu). Sequence logos were used as a graphical representation of overrepresentation of certain nucleotides around 5’ and 3’ splice sites. The overall height of the stack represents sequence conservation at a given position and the height of symbols representing nucleotides indicates the frequency of a given nucleotide at given position (see Figure 4 and Supplementary Figure 1). An increased height of any nucleotide is an indication of a higher frequency of the specific nucleotide at that position.

Expression Analysis

Expression array data (Affymetrix Human Genome U133 Plus 2.0 Array) were obtained through the TCGA (The Cancer Genome Atlas) Data Portal (https://tcga-data.nci.nih.gov/tcga). To select low U2AF1 expressor patients, samples that had expression lower than 2 SD from the mean were selected. Expression data for U2AF1 was normally distributed. The statistical difference between normal and low U2AF1 expressors was assessed using unpaired T-test.

Publicly available databases.

The February 2009 human reference sequence (GRCh37) produced by the Genome Reference Consortium was used as reference genome (UCSC genome browser; http://genome.ucsc.edu/cgi-bin/hgGateway). Somatic mutation data was searched by Catalogue of somatic mutations in cancer (COSMIC) database in Welcome Trust Sanger Institution website (http://www.sanger.ac.uk/genetics/CGP/cosmic/). Each potential mutation was compared against databases of known SNPs, including Entrez Gene (http://www.ncbi.nlm.nih.gov/gene) and the Ensemble Genome Browser (http://useast.ensembl.org/index.html).

Cytogenetics and single nucleotide polymorphism array (SNP-A) analyses.

SNP-A assays were processed as previously described.24,25 Affymetrix Human Mapping 250K NSP and Human Genome-Wide SNP Array 6.0 Kit (Affymetrix, Santa Clara, CA) were used. Patients with SNP-A lesions concordant with
metaphase cytogenetics or typical lesions known to be recurrent required no further analysis. Germline changes reported in our internal or publicly-available (Database of Genomic Variants; http://projects.tcag.ca/variation) copy number variation (CNV) databases were considered non-somatic and excluded from further analysis. Results were obtained using CNAG (v3.0)26 (Affymetrix Human Mapping 250k NSP kit) or Genotyping Console (Affymetrix Genome-Wide SNP Array 6.0 kit). All other lesions were additionally confirmed as somatic or germline by analysis of CD3-sorted cells.27

Statistical analysis.

For comparison of the exon usage levels between groups, with and without U2AF1 mutations, statistical analysis was performed using the described workflow. We used 201,837 exons/variables and 20 observations (N=6 U2AF1 mutants and N=14 U2AF1 WT). The probability that a particular score would occur by a chance was assessed using permutation testing and a random model28. We employed randomization based significance testing. This leads to the notion of normalized scores, expressed as the number of standard deviations from the mean of the random distribution. The returned P-values were >=5 standard deviation from expected P-value of a random set (student T-test was used to generate the P-values). That allowed one to reject the null hypothesis of being no difference between the 2 cohorts tested.
Results

Detection of U2AF1 mutations and genotypic associations.

We examined the U2AF1 mutational status of a cohort of patients (N=524) with various hematologic malignancies, including MDS, MDS/MPN, MPN and AML using Sanger sequencing and NGS exome sequencing to identify cases for further analysis. Of these, 46 cases harboring heterozygous somatic mutations in U2AF1 (9%) were found (Figure 1). Overall, U2AF1 mutations were more frequently present in male patients compared to female patients (83% vs 17%, p<.001) and were nearly evenly distributed among patients with AML (7%, in both sAML as well as primary (p)AML), MDS (10%), MPN (8%), MDS/MPN, (14.5%) and MDS with higher risk (12.5%). There were eight distinct missense mutations, including A26V (n=1), R35L (n=1), S34Y (n=4), S34F (n=21), R156Q (n=2), Q157P (n=13), Q157R (n=3) and G213A (n=1) (Supplementary Table 1). The two most frequent mutations, S34F(47%) and Q157P(29%) accounted for more than 75% of all mutations detected. Almost invariably (97.8%) the mutations were localized in one of the 2 zinc finger domains (Figure 1A).

Mutational screening detected concomitant mutations in DNMT3A, RAS family genes (KRAS/NRAS), ASXL1, RUNX1, TET2, CBL and IDH family genes (IDH1/2) in 25%, 25%, 24%, 13%, 11%, 12% and 5% of patients, respectively (Supplementary Figure 2). There was only one case (refractory anemia with ring sideroblasts, trisomy 8) harboring a double U2AF1/SF3B1 spliceosomal mutation. Mutations in a different spliceosomal factor, SRSF2, were mutually exclusive in our cohort. Additionally, 11/40 patients didn’t harbor any additional mutations from the panel of genes tested (Supplementary Figure 2).

Functional importance of U2AF1 mutants

To understand the functional consequences of U2AF1 mutations, we studied mutation-specific exon usage patterns as determined by deep RNA-sequencing of U2AF1 mutant (n=6) and WT (n=14) cases. Using T-test (p<.0001), average and absolute difference in exon usage ratio (>±15%) as criteria, we successfully tested 201,837 exons in 17,097 genes. Using this approach, we found 35 exons in 35 genes with a significantly altered pattern of inclusion or exclusion in U2AF1 mutant cases compared to spliceosomal WT cases (Figure 2). The U2AF1 mutant-specific splicing patterns were categorized into two groups: exon skipping (lower exon usage) or exon retention (higher exon usage) with respect to WT (Supplementary Figure 3). Most (77%) of the significantly altered exons showed more exon skipping in patients carrying U2AF1 mutation. The rest (23%) represented increased exon retention patterns. Among the genes exhibiting differential exon usage patterns, we identified genes involved in different stages of mitosis (CEP164, EHMT1, WAC and ATR). Another distinct group of genes identified consists of genes involved in RNA processing (PTBP1, STRAP, PPWD1, PABPC4 and UPF3B) (Figure 2).

To confirm the aberrant pattern of alternative splicing in U2AF1 mutants, we amplified cDNA containing a specific exon found to be alternatively skipped based on RNAseq reads. As an example, we selected CEP164 exon 7, which was observed to be most frequently skipped in U2AF1 mutants compared to cases with WT U2AF1. Using primers in exon 6 and exon 8, skipping of exon 7 yielded a 220 bp product while inclusion of exon 7 yielded a 298 bp (Figure 3C). As predicted by the RNAseq results, only the exon 7 skipped product was observed in U2AF1 mutant cases while in 6 out of
7 U2AF1 WT cases, both 298bp and 220bp bands were detected, suggesting that exon 7 was partially skipped. These findings were further confirmed by using primers in exon 6 and exon 7, in which amplification products were detected only in the WT U2AF1 cases, but not in the U2AF1 mutant cases (Figure. 3C).

Transcriptional analysis of U2AF1 WT patients with low U2AF1 expression
To better understand the functional role of the U2AF1 mutations, we also compared exon usage levels in patients with low expression of U2AF1. If the U2AF1 mutations were simply hypomorphic, one would expect to find similar missplicing patterns in patients with low levels of U2AF1 expression and those with mutant U2AF1. We focused our analyses on exon usage of genes that were identified to have differential splicing pattern between U2AF1 mutant and WT cases (Figure 3A). Low expressors of U2AF1 had similar (p>0.05) usage levels (<±10%) as that of WT in 19 exons (55%), similar usage to U2AF1 mutant in 1 exon (3%) and intermediate exon usage to WT and mutant in 14 exons (42%) (Figure 3B). This result indicates that low expression of U2AF1 does not create the same aberrant splicing patterns observed in U2AF1 mutants. Additionally all the exons of 35 genes that were differentially spliced in U2AF1 mutant were screened but no significant differences were found.

Comparison of splicing patterns in U2AF1 mutant patients with other spliceosomal factor mutations
While mutations in several different splicing factors have been described in myeloid neoplasms, each factor seems to have a unique distribution and different effects on survival. This suggests that the target genes might be different for each mutant factor. To investigate this idea, we selected cases with somatic mutations in the splicing factors SF3B1, SRSF2 and U2AF26 and compared the splicing pattern of representative genes to our U2AF1 mutant cases. As shown in Figure 3D, each factor mutation was associated with different changes in the splicing patterns of these genes. The dotted lines correspond to the level of exon skipping seen in WT cases.

Sequence signatures of splice sites associated with missplicing in U2AF1 mutants
To explain differences in exon usage between U2AF1 mutants and WT, we analyzed the splice site sequences surrounding all the exons analyzed in this study. Figure 4 shows the sequence patterns flanking the alternative exons for the subset that showed increased exon skipping (top line), the subset that showed increased exon retention (second line) and the flanking exons (bottom line). The splice site sequence patterns generally match the consensus sequences for 3’ and 5’ splice sites with the exception of the -3 nucleotide relative to the 3’ intron/exon junction (boxed position 16 in Figure 4). This position had a higher frequency of thymidine (83%) adjacent to exons that were skipped more frequently in U2AF1 mutants and a very low frequency of thymidine adjacent to exons that were more frequently included in U2AF1 mutants. The consensus of all 3’ splice sites shows a nearly equal probability of a thymidine or a cytosine at this position as is seen in the bottom line of Figure 4. As discussed below, this position is immediately adjacent to the AG dinucleotide that is known to be bound by U2AF1. To determine if this sequence signature was related to the frequency of skipping of the adjacent exon, we bounded all alternative exons by their skipping frequencies and analyzed the splice sites sequences.
(Supplementary Figure 1). All subsets of alternative exons had a very similar pattern of C and T at position 16 that matches the overall consensus sequence.

Mutations and deletions in the misspliced genes by U2AF1 mutants
To further explore the common pathophysiology between exon usage alteration in U2AF1 mutants and other somatic molecular events occurring in myeloid neoplasms, the genes in which excess exon skipping or retention was detected were searched for mutations and deletions. Out of 35 genes misspliced in U2AF1 mutants, deletion of the corresponding locus was observed in 34 genes (97%) and somatic mutations were observed in all 35 (100%) genes (Supplementary Table 2). Remarkably, some of these somatic mutations were located in the exact exons for which exon usage was changed by U2AF1 mutation. For example, two missense and one nonsense mutation were observed in the same RRM domain of PTBP1 which was skipped more frequently in U2AF1 mutant cases (Figure 2). In another example, the CEP164 locus, in which exon 7 was highly skipped, was frequently deleted (chr11q23.3) in myeloid malignancies (Supplementary Table 2) while in a solid tumor cohort, missense/nonsense/frameshift mutations were reported as well (COSMIC database).
Discussion

Recently, frequent recurrent \textit{U2AF1} mutations in myeloid malignancies were reported to result in splicing alteration, which causes exon skipping or less expression due to unspliced pre-mRNA.\textsuperscript{3,4,7} In this study, we identified distinct mutation-specific exon usage patterns as the functional consequences of \textit{U2AF1} mutations. \textit{U2AF1} mutations are associated with abnormal splicing of genes involved in functionally important pathways, including cell cycle progression and RNA processing. Moreover, some of these genes are somatically mutated or deleted in various cancers. Of note is that missplicing patterns associated with \textit{U2AF1} mutations were observed in exons flanked by a characteristic splice site sequence bias. These findings supply novel information on how the recurrent \textit{U2AF1} mutations might participate in the pathophysiology of myeloid malignancies.

In this study, deep RNA-sequencing of \textit{U2AF1} mutant cases showed significant alterations of splicing patterns in multiple genes. Functionally related gene groups were affected by missplicing due to \textit{U2AF1} mutations. For example, genes involved in different stages of mitosis (\textit{CEP164}, \textit{EHMT1}, \textit{WAC} and \textit{ATR}) or in RNA processing (\textit{PTBP1}, \textit{STRAP}, \textit{PPWD1}, \textit{PABPC4} and \textit{UPF3B}) were affected. Another affected gene, \textit{CEP164}, is one of the centrosomal proteins involved in G2/M checkpoint control and nuclear divisions.\textsuperscript{29,30} In various malignancies, the \textit{CEP164} locus is frequently deleted or affected by missense/nonsense/frameshift mutations. Thus, the \textit{CEP164} locus demonstrates 3 different types of loss of function: deletion, mutation and splicing defects due to \textit{U2AF1} somatic mutations. Of note is that \textit{CEP164} and \textit{ATR} proteins interact with each other in the DNA damage signaling cascade\textsuperscript{30} and \textit{ATR} is one of the genes frequently mutated in myeloid malignancies. These findings indicate that molecular events due to splicing defects and somatic mutation/deletion might be leukemogenic events via common gene targets.

One of the misspliced genes found here, \textit{PTBP1}, is known to regulate alternative splicing events through interactions with pyrimidine-rich RNA sequences.\textsuperscript{31} \textit{PTBP1} may also inhibit the binding of U2 snRNP to certain pre-mRNAs indicating that \textit{PTBP1} could be in the same complex as \textit{U2AF1} or competing with it.\textsuperscript{32} Splicing regulatory genes (for example \textit{PTBP1}) misspliced by \textit{U2AF1} mutations might indirectly promote additional splicing defects. Interestingly, the shorter spliced variant of \textit{PTBP1} that is overproduced in \textit{U2AF1} mutant cases is missing of the second quasi-RNA recognition motif (RRM) domain, which is functionally associated with RNA binding.\textsuperscript{33} Moreover, in solid tumors, somatic mutations were reported in this RRM domain.\textsuperscript{34-37} These findings suggest that \textit{U2AF1} mutations might modify the isoforms of other spliceosomal proteins (for example \textit{PTBP1}) by changing splicing pattern or that other spliceosomal genes modified by \textit{U2AF1} mutations could indirectly promote other splicing defects as well as the direct effects of \textit{U2AF1} mutations.

The \textit{U2AF1} protein is part of the heterodimeric U2 Auxiliary Factor (U2AF) along with the \textit{U2AF2} protein. \textit{U2AF2} binds to the polypyrimidine tract upstream of the 3’ splice junction while \textit{U2AF1} binds to the invariant AG dinucleotide at the 3’ splice junction. The binding of the \textit{U2AF} complex to the 3’ splice site is one of the early steps in spliceosome formation. There is evidence that the requirement of \textit{U2AF1} differs among 3’ splice sites suggesting that it can serve a regulatory role in alternative splicing decisions. This theory is supported by the finding that most mutations detected by us and other groups are located in either of the two zinc finger domains which are likely involved in RNA binding.\textsuperscript{3,4,7}
Further support comes from our finding of a unique sequence feature in 3’ splice sites affected by $U2AF1$ mutations. The identity of the nucleotide immediately upstream of the 3’ splice site AG appears to regulate how well the adjacent exon is spliced in the $U2AF1$ mutants. Normally this nucleotide is either a T or C and was not thought to be recognized by U2AF1. It now appears that this nucleotide is recognized differently by the mutant U2AF1 compared to the WT U2AF1. The molecular basis of this altered recognition is under investigation.

Mutations in myeloid neoplasms of each component of the spliceosome are almost always mutually exclusive $^{4,7}$, even if the proteins cooperate with each other in splicing. This implies that defects in these different genes might contribute to modifying spliceosomal function in a unique way and that $U2AF1$ and other spliceosomal mutations cannot occur in a cumulatively synergistic way. Furthermore, the spectrum of mutations in these genes suggests that they are not simply loss of function alleles but, rather, have altered functions. To support this theory, we showed that patients with low expression of $U2AF1$ revealed a splicing pattern similar to that of WT but different from $U2AF1$ mutant. Another explanation of distinct splicing pattern observed in $U2AF1$ mutant is that low expressors of $U2AF1$ might just induce compensatory mechanism on a spliceosomal machinery. Further investigation of this observation is needed to clarify the mechanism.

Clinically, in our cohort, we find that $U2AF1$ mutations are more frequent in more proliferative phenotypes, including MDS/MPN and high risk MDS, which require a new therapeutic strategy. Previous reports also showed that $U2AF1$ mutations are associated with high incidence of leukemic evolution and poor prognosis.$^{3,4,7}$ In younger patients, more intensive chemotherapy or stem cell transplantation will be indicated in cases with $U2AF1$ mutations. In elderly patients, more specific drug therapy should be applied, for example, molecular-targeted-therapy. In this study, we identified down-stream splicing defects in several genes that are functionally important in various cancers. Such molecules could be proposed as novel therapeutic targets in $U2AF1$ mutant cases.

Our RNA sequencing analysis was applied to the most comprehensive splice sites in coding regions, which provided us with completely novel findings associated with prevalent $U2AF1$ mutations. Despite of disability to remove false positive risk thoroughly, genetically reproducible splicing patterns were identified in functionally important genes. Null-model comparisons were more reasonable statistical methodology in this study than multiple testing corrections. Further basic experiments, for example, conditional knock-in mutant animal models will clarify the detail pathophysiological significance of splicing defects in myeloid neoplasms with various types of spliceosome gene mutations.

In summary, our study validates the change of function nature of $U2AF1$ mutations and describes a set of significantly misspliced genes, functionally correlated, and almost invariably affected by a concomitant molecular alteration, establishing a novel mechanism of leukemogenesis of myeloid malignancies.
Acknowledgments

The results published here are in part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigator and institutions that constitute the TCGA research network can be found at http://cancergenome.nih.gov.

This work was supported by National Institutes of Health (Bethesda, MD; NIH) grants RO1-GM104059 (R.A.P.), RO1HL-082983 (J.P.M.), U54 RR019391 (J.P.M.), K24 HL-077522 (J.P.M.), a grant from the AA & MDS International Foundation (Rockville, MD), the Robert Duggan Charitable Fund (Cleveland, OH; J.P.M.), and Scott Hamilton CARES grant (Cleveland, OH; H.M.)

Authorship

B.P. and H.M. designed research, performed research, collected data, performed statistical analysis and wrote the manuscript. K.G. collected data. A.J. and M.A.S. interpreted data, and wrote the manuscript. R.P. designed research, contributed analytical tools, collected data, analyzed and wrote the manuscript. J.P.M. designed research, analyzed and interpreted data, and wrote the manuscript.

Conflict of interest disclosure: The authors declare no competing financial interests.

Corresponding author:

Hideki Makishima.

Taussig Cancer Institute/R40, 9500 Euclid Avenue, Cleveland, OH 44195 e-mail: makishh@ccf.org
References


Figure 1. Distribution and frequency of U2AF1 mutations across gene domains and different hematological malignancies.
A) Three isoforms of U2AF1 are shown with the two zinc finger domains (ZNF) and the RNA recognition motif (RRM) highlighted. Almost all identified U2AF1 missense mutations are located in one of the two ZNF domains. B) Comparison of the frequency of U2AF1 mutations between different hematological malignancies. MDS/MPN and high risk MDS (RAEB1/2) showed the most frequent mutations (14% and 13% respectively) whereas pAML showed the least (6%).
Figure 2. Differences of exon usage frequencies in genes that were identified. Exon skipping frequencies were based on RNAseq data, averaged and presented as bar graphs. Left panel: Bars in dark blue color represent U2AF1 mutants and dark brown bars represent WT. The order of genes was determined using the average difference between U2AF1 mutant and WT exon skipping frequency. Right panel: Detailed frequency of exon skipping of all exons screened for PTBP1 (upper panel) and CEP164 (lower panel). Additional mutational information is depicted for both selected genes. CEP164 lower panel contains additional SNP karyotyping data depicting samples that had the CEP164 locus deleted (highlighted in red).
Figure 3. Transcriptional analysis of patients with splicing factor mutations. A) Comparison of levels of U2AF1 mRNA between U2AF1 mutants, WT cases and WT cases with low expression of U2AF1 (red, blue and green colors, respectively). The mean expression level is indicated by the dotted line. B) Exon skipping levels in three genes comparing U2AF1 mutants, WT and WT with low expression levels (red, blue and green bars, respectively). C) Validation of RNAseq results on exon 7 of the CEP164 gene using an independent set of patients by RT-PCR. D) Comparison of exon skipping levels between patients bearing mutations in different spliceosomal factors: SF3B1, SRSF2, U2AF26 and U2AF1.
Figure 4. Frequencies of nucleotides surrounding 3’ and 5’ splice sites adjacent to exons affected by U2AF1 mutations. Exons that were more skipped or more retained in U2AF1 mutants were combined into two groups and the splice site consensus sequences were derived (top and middle lines). Adjacent splice sites were analyzed as a control set (bottom line). Nucleotide frequencies are represented using WebLogo software. The height of each stack represents the information content of that position in bits. The height of each letter represents the frequency of occurrence of each nucleotide.
Patterns of missplicing due to somatic *U2AF1* mutations in myeloid neoplasms

Bartlomiej Przychodzen, Andres Jerez, Kathryn Guinta, Mikkael A. Sekeres, Richard Padgett, Jaroslaw P. Maciejewski and Hideki Makishima