RECURRENT MUTATIONS IN THE *U2AF1* SPLICING FACTOR IN MYELODYSPLASTIC SYNDROMES

Timothy A. Graubert*[1,3,6], Dong Shen*[4], Li Ding*[2,4], Theresa Okeyo-Owuor[1], Cara L. Lunn[1], Jin Shao[1], Kilannin Krysiak[1], Christopher C. Harris[4], Daniel C. Koboldt[4], David E. Larson[4], Michael D. McLellan[4], David J. Dooling[4], Rachel M. Abbott[4], Robert S. Fulton[4], Heather Schmidt[4], Joelle Kalicki-Veizer[4], Michelle O'Laughlin[4], Marcus Grillot[1], Jack Baty[5], Sharon Heath[1], John L. Frater[6], Talat Nasim[7,8], Daniel C. Link[1,3], Michael H. Tomasson[1,3], Peter Westervelt[1,3], John F. DiPersio[1,3], Elaine R. Mardis[2,3,4], Timothy J. Ley[1,2,3,4], Richard K. Wilson[2,3,4], and Matthew J. Walter [1,2,3]

[1]Department of Internal Medicine, Division of Oncology, Washington University, St. Louis, MO, USA
[2]Department of Genetics, Washington University, St. Louis, MO, USA
[3]Siteman Cancer Center, Washington University, St. Louis, MO, USA
[4]The Genome Institute, Washington University, St. Louis, MO, USA
[5]Division of Biostatistics, Washington University, St. Louis, MO, USA
[6]Department of Pathology and Immunology, Washington University, St. Louis, MO, USA
[7]Department of Medical and Molecular Genetics, King's College, Guy's Hospital, London, UK
[8]National Institute for Health Research (NIHR), Biomedical Research Centre, Guy's and St. Thomas' NHS Foundation Trust and King's College London, UK

*these authors contributed equally.

Corresponding Author:

Matthew J. Walter, MD
Washington University School of Medicine
Division of Oncology, Stem Cell Biology Section
Campus Box 8007
660 South Euclid Avenue
St. Louis, MO 63110 USA

Phone: 314/362-9409
Fax: 314/362-9333
email: mjwalter@dom.wustl.edu

Text Word Count: 1968
Introductory Paragraph Word Count: 147
Figures: 3
Tables: 1
Pages (including figures): 23
References: 47
Format: Letter

INTRODUCTORY PARAGRAPH

Myelodysplastic syndromes (MDS) are hematopoietic stem cell disorders that often progress to chemotherapy-resistant secondary acute myeloid leukemia (sAML). We used whole genome sequencing to perform an unbiased comprehensive screen to discover all the somatic mutations in a sAML sample and genotyped these mutations in the matched MDS sample. Here we show that a missense mutation affecting the serine at codon 34 (S34) in *U2AF1* was recurrently mutated in 13/150 (8.7%) *de novo* MDS patients, with suggestive evidence of an associated increased risk of progression to sAML. U2AF1 is a U2 auxiliary factor protein that recognizes the AG splice acceptor dinucleotide at the 3' end of introns and mutations are located in highly conserved zinc fingers in U2AF1[1,2]. Mutant U2AF1 promotes enhanced splicing and exon skipping in reporter assays *in vitro*. This novel, recurrent mutation in *U2AF1* implicates altered pre-mRNA splicing as a potential mechanism for MDS pathogenesis.

MAIN TEXT

Myelodysplastic syndromes (MDS) are a heterogenous group of hematopoietic stem cell disorders characterized by dysplastic blood cell formation and peripheral blood cytopenias. Up to 30% of patients with MDS will progress to a highly chemotherapy-resistant secondary acute myeloid leukemia (sAML). Whole genome sequencing (WGS) offers an unbiased approach to discover all the genetic mutations present in cancer genomes and has been used to identify novel mutations in *de novo* and therapy-related AML genomes[3-7]. Here we report the results of WGS of an MDS-derived sAML sample and the matched normal (skin) sample. We performed WGS using 100 base pair paired-end reads and obtained 39.1x and 38.2x haploid and 99.3% and 98.9% diploid coverage of the sAML and normal samples, respectively (**Supplementary Table 1**). We divided the genome into non-overlapping tiers, as previously described[4], and validated putative mutations using deep sequencing of captured DNA isolated from the sAML, normal, and MDS samples. We validated 507 somatic single nucleotide variants (SNVs) in the sAML sample, including 30 SNVs in protein coding regions (tier 1 mutations). 505/507 SNVs preexisted in the MDS sample, including 30 tier 1 mutations (**Supplementary Fig. 2, Supplementary Tables 2, 3**). The same codon in *U2AF1* (*U2AF35)* was mutated in 2 additional MDS-derived sAML cases analyzed by whole genome sequencing (data not shown). This was the sole recurrent mutation in these cases. To determine the frequency of this mutation in MDS, we sequenced the entire coding region of *U2AF1*, including 9 exons, in diagnostic bone marrow and paired normal (skin) samples from 150 consecutively accrued *de novo* MDS patients (including the index case) and identified 13 patients (8.7%) with missense mutations affecting the highly conserved serine at amino acid position 34 (S34) in U2AF1 (**Fig**. **1a**). The same nucleotide was mutated in all samples, resulting in either a S34F (n=11) or S34Y (n=2)

substitution (**Supplementary Table 4**). One patient with an S34F mutation (UPN 947519) also had a *U2AF1* Q157R mutation located in the second zinc finger (**Fig**. **1a**). No other somatic SNVs affecting *U2AF1* were detected in these samples. Subsequent analysis focused on the highly recurrent S34 mutations.

U2AF1 is the small (35 kDa) subunit of U2 snRNP auxiliary factor (U2AF) that is involved in pre-mRNA processing (splicing), and it forms a heterodimer with the larger subunit U2AF2 (U2AF65)[2]. U2AF1 binds the 3' AG splice acceptor dinucleotide of the pre-mRNA target intron[2] and U2AF2 binds the adjacent polypyrimidine tract. PCR amplicons spanning the S34 codon were generated from genomic DNA and cDNA templates made from nucleic acids harvested from unpurified MDS bone marrow cells, and subjected to deep sequencing to obtain mutant allele frequencies. Importantly, there was no deletion or uniparental isodisomy (UPD) that spanned the *U2AF1* locus (chromosome 21q22.3) based on SNP arrays and whole genome sequencing data for the index case. Read counts from 11 of the 13 genomic DNA samples (including the sAML sample from the index case and serial MDS samples from two other patients) showed that the S34 mutant allele frequencies were ~40-50%, indicating that the majority of cells in the samples contained a heterozygous mutation, even though the myeloblast counts ranged from 0-21% in the MDS samples (**Fig. 1b**). Similar results were obtained from cDNA deep sequencing (~30-50% mutant allele frequency), indicating that both the S34 mutant and wild-type alleles were expressed in all samples, regardless of the myeloblast count (**Fig. 1c**). In addition, there was no difference in the total levels of *U2AF1* mRNA or the dominant *U2AF1* isoform that was expressed in unfractionated MDS bone marrow samples from patients with and without *U2AF1* mutations (**Supplementary Fig. 4a**). Collectively, these results suggest that *U2AF1* mutations are an early, initiating genetic event in MDS pathogenesis.

Although eight of the mutant samples had myeloblast counts > 5%, patients with *U2AF1* mutations were not restricted to a particular International Prognostic Scoring System (IPSS) category and had a median IPSS score of 1 (range 0-3) (**Supplementary Table 5**). Patients with a del(20q) karyotype were more likely to harbor a *U2AF1* mutation (P=0.01), although the number of patients with mutations and del(20q) is small (n=4) (**Table 1**). No difference in event-free or overall survival was observed in patients with or without *U2AF1* mutations (**Fig. 2a-b**). However, the 2 mutant patients with the longest overall survival had received hematopoietic stem cell transplants (**Supplementary Fig. 1**). Patients with *U2AF1* mutations had an increased probability of progression to sAML (P=0.03) (**Fig. 2c**), an observation that will require confirmation in a larger cohort. This corresponds to a *U2AF1* mutation frequency of 15.2% (7/46 patients) in the subset of MDS patients that progress to sAML vs. 5.8% (6/104 patients) in the subset that did not. Since there was no statistical difference in the myeloblast count or IPSS distribution of patients with or without *U2AF1* mutations (**Table 1**), the mutant genotype does not appear to be a surrogate for these well-established predictors of sAML risk.

Splicing involves cleavage of intronic sequences from pre-mRNA, followed by ligation of the remaining exons together to produce a mature mRNA product[8]. Inclusion and exclusion of different exons or utilization of alternative 3' splice sites during pre-mRNA processing produces multiple protein isoforms that can have different functions within a cell, and alternative splicing can be affected by the levels of U2AF1 in a cell[9-12]. It is unknown which domain of U2AF1 binds the pre-mRNA. Interestingly, the S34 and Q157 residues are located within zinc finger domains (**Fig. 1a**) that may be important for RNA binding activity. Indeed, the U2AF1 zinc fingers are structurally similar to the murine and human ZFP36 family zinc fingers (both

$CX_8CX_5CX_3H$ zinc fingers)[13-15] that bind RNA and the noncanonical RNA recognition motif (RRM; also known as U2AF homology motif, UHM) in U2AF1 only weakly binds RNA[16].

To examine the effects of the S34F mutation on U2AF1 splicing activity, we utilized previously described and validated *in vitro* double-reporter splicing and minigene reporter assays[9,17]. The double-reporter plasmid constitutively expresses β-galactosidase, while luciferase is expressed only if appropriate splicing removes an upstream intron that contains translational stop codons. Transient co-expression of the double-reporter plasmid pTN24 and the S34F mutant *U2AF1* cDNA in 293T cells resulted in a significant increase in splicing (as detected by an increase in the luciferase/β-galactosidase ratio), compared to co-expression of wild-type *U2AF1*, despite similar total U2AF1 protein levels in all samples (**Fig. 3a**, $P<0.001$). The level of splicing was similar in cells depleted of endogenous U2AF1 compared to control cells (**Fig. 3b**, left column), and the increase in splicing observed with the S34F mutant *U2AF1* is independent of endogenous U2AF1 levels (**Fig. 3b**, right column, $P<0.001$ when compared to vector alone). This suggests that splicing activity in this assay is insensitive to U2AF1 levels, and increased splicing mediated by the mutant protein is attributable to a novel gain-of-function activity.

Next, we examined exon skipping using a minigene reporter plasmid (a human gene fragment containing an upstream and downstream exon surrounding an intron-flanked exon). Appropriate splicing produces an mRNA with all 3 exons, while exon skipping fuses the upstream and downstream exons only. We measured the levels of exon skipping using a *GH1* minigene reporter plasmid in cells transiently co-transfected with either wild-type or S34F mutant *U2AF1* cDNA. The proportion of transcripts with a skipped exon (lower PCR band) relative to the appropriately spliced *GH1* minigene (upper PCR band) is increased in cells

expressing the S34F *U2AF1* mutant compared to control or wild-type *U2AF1* expression (**Fig. 3c**, P=0.01). This increase in exon skipping mediated by mutant *U2AF1* remained significant after depletion of endogenous *U2AF1* in 293T cells (**Fig. 3c**, P<0.02). We also observed an increased utilization of alternative 3' cryptic splice sites in the *FMR1* gene in clinical MDS samples with *U2AF1* mutations compared to MDS samples without *U2AF1* mutations (**Supplementary Fig. 3a**). The alternative splicing of *FMR1* was confirmed to be mutant *U2AF1*-dependent using a *FMR1* minigene splicing reporter assay *in vitro* (**Supplementary Fig. 3b**).

Collectively, these results suggest that *U2AF1* S34 mutations may result in subtle increases in splicing efficiency (**Fig. 3a)** -- or possibly altered isoform expression (**Fig. 3c**) -- which could induce gene expression changes. To test whether *U2AF1* mutations alter global gene expression levels, we analyzed mRNA microarray data (Affymetrix U133plus2) obtained from bone marrow CD34+ cells purified from 6 MDS patients with a *U2AF1* mutation, 9 MDS patients without a mutation, and 4 normal donors[18]. The *U2AF1* mutant samples did not segregate together using an unsupervised hierarchical clustering algorithm with all 19 samples, however, the 6 *U2AF1* mutant samples did segregate together when compared to the normal control samples (**Supplementary Fig. 4b**). Next, we identified the genes that were significantly different between control and mutant patients using Significance Analysis of Microarrays (SAM)[19]. SAM identified 401 dysregulated probesets (51 up-regulated and 351 down-regulated) in *U2AF1* mutant versus control samples (FDR<0.005) (**Supplementary Fig. 4b, Supplementary Table 6**). Three of the most enriched functional annotation categories for genes that are down-regulated in *U2AF1* mutant samples are splicing and RNA recognition motif (RRM) genes (enrichment scores 2.5-4.3) (**Supplementary Fig. 4c, Supplementary Table 6**).

These results suggest that a compensatory down-regulation of splicing genes may exist in *U2AF1* mutant samples. These gene categories were not down-regulated in *U2AF1* wild-type MDS patients compared to controls (data not shown), suggesting that down-regulation of splicing and RRM genes is not common to all MDS samples, but instead is associated with *U2AF1* mutations.

Mutations in *U2AF1* represent a novel mechanism that could alter gene expression in MDS and expand the list of commonly mutated genes in MDS that may affect transcription or translation, including *RPS14, TET2, EZH2, ASXL1,* and *DNMT3A*[20-26]. Only two patients in our cohort of 150 *de novo* MDS samples had both a *DNMT3A* and *U2AF1* mutation (**Supplementary Table 7**). Both of these mutations appear to be early genetic events in MDS, given their high mutant allele burdens in patients with early stage disease[26].

U2AF1 is highly conserved (**Fig. 1a**)[1], and homozygous loss is lethal in many organisms[27-29]. We did not observe any nonsense, frameshift or missense mutations affecting the coordinating CCCH residues in the zinc fingers, again suggesting the S34 mutations are not loss-of-function mutations. The corresponding amino acid in the human ZFP36L2 (a ZFP36 family member) protein interacts with RNA through a hydrogen bond, further suggesting that the S34 position may be important for RNA binding[13]. Additionally, interactions between conserved aromatic amino acids in ZFP36 family members and RNA bases stabilize the protein RNA complexes formed[13]. The two amino acid substitutions we identified at S34 add a bulky aromatic ring (phenylalanine or tyrosine) to the zinc finger, which may alter, or even enhance, binding of the zinc finger to RNA. We suggest that the S34F/Y mutations in U2AF1 alter the specificity of U2AF1-dependent splicing. Pre-mRNAs with strong polypyrimidine tracts can splice independently of U2AF1 *in vitro*, whereas weak polypyrimidine tracts are more dependent

on U2AF1 for appropriate splicing[2,11]. Therefore, the pattern of U2AF specificity (determined by both U2AF1 and U2AF2) may be influenced by the nucleotide sequence in pre-mRNAs and may be an important factor in determining which genes are altered in cells expressing mutant U2AF1.

Alternative splicing has been described for a wide range of cancers[30,31], although the underlying mechanisms that influence cancer pathogenesis remain largely unknown. Alterations in the transcriptome mediated by alternative splicing may contribute directly to cancer, or indirectly by engaging some other pathway. The identification of somatic mutations in spliceosome genes in MDS by our group and others[32-34], raises the possibility that mutations in splicing factors, including *U2AF1*, may be responsible for the observed alterations of splicing in cancer. Ultimately, cancer cells may generate genetic diversity in a large number of genes by selecting cells with mutations in U2AF1 or other spliceosome proteins. Identification of key target genes affected by *U2AF1* mutations will be critical for our understanding of how these mutations contribute to MDS pathogenesis.

ACKNOWLEDGEMENTS

This work was supported by NIH grants R01HL082973 (Graubert), RC2HL102927 (Graubert), U54HG003079 (Wilson), P01CA101937 (Ley), and a Howard Hughes Medical Institute Physician-Scientist Early Career Award (Walter). Technical assistance was provided by the Alvin J. Siteman Cancer Center High Speed Cell Sorting Core, the Molecular and Genomic Analysis Core, the Biomedical Informatics Core, and the Tissue Procurement Core which are supported by an NCI Cancer Center Support Grant P30CA91842. Additional technical assistance was provided by Masayo Izumi. We thank Dr. Kinji Ohno (Nagoya University Graduate School of Medicine, Japan) for minigene constructs. We thank Dr. Kathleen Hall (Washington University School of Medicine) for helpful scientific discussions.

AUTHORS CONTRIBUTIONS

Timothy A. Graubert: project leader, study design, execution and analysis, manuscript preparation.

Dong Shen: project leader, sequence analysis.

Li Ding: project leader, supervisor data analysis team.

Theresa Okeyo-Owuor: in vitro splicing assays.

Cara L. Lunn: quantitative reverse transcriptase PCR, in vitro splicing assays.

Jin Shao: microarray data analysis and PCR assays.

Kilannin Krysiak: gene expression analysis.

Christopher C. Harris: sequence analysis.

Dan C. Koboldt: capture validation data analysis.

David E. Larson: mutation analysis and annotation.

Michael D. McLellan: auto-analysis and manual review of validation data.

David J. Dooling: IT and data management, data analysis automation leader.

Rachel M. Abbott: variant validation production.

Robert S. Fulton: variant validation oversight.

Heather Schmidt: manual review of variants.

Joelle Kalicki-Veizer: manual review of variants.

Michelle O'Laughlin: variant validation production.

Marcus Grillot: clinical data management and specimen acquisition.

Jack Baty: statistical analysis of clinical variables and outcomes.

Sharon Heath: clinical data management and specimen acquisition.

John L. Frater: clinical hematopathology review.

Talat Nasim: design of in vitro dual reporter splicing assay.

Daniel C. Link: study design, execution and analysis and manuscript preparation.

Michael H. Tomasson: study design, execution and analysis.

Peter Westervelt: clinical data and specimen acquisition, study design, execution and analysis.

John F. DiPersio study design, execution and analysis and manuscript preparation.

Elaine R. Mardis: project conception, analysis coordination and manuscript preparation.

Timothy J. Ley: project conception, study design, manuscript preparation.

Richard K. Wilson: project conception and oversight, manuscript preparation.

Matthew J. Walter: project leader, study design, analysis coordination and manuscript preparation.

DATA ACCESS SECTION

Sequence and SNPa data have been deposited in dbGAP under accession number phs000159.v3.p2. Gene expression profiling data have been deposited in GEO under accession number GSE30195.

SUPPLEMENTARY INFORMATION

Please see Supplementary Note for Text, Results, Figures (4), Tables (9), and References.

COMPETING INTERESTS STATEMENT

The authors have no competing interest to declare.

FIGURE LEGENDS

Figure 1. ***U2AF1* mutations found in patients with myelodysplastic syndromes (MDS).** (**a**) Missense mutations were detected in codons 34 and 157 of U2AF1. The ZnF1 (zinc finger 1), UHM (U2AF homology motif), ZnF2 (zinc finger 2), and RS (arginine-serine rich) domains are shown. The amino acid sequence of the ZnF1 domain is highly conserved (shaded). The zinc coordinating and mutated residue are shown in blue (asterisks) and red (arrow), respectively. (**b**) Deep sequencing of *U2AF1* using DNA collected from paired normal, MDS, or secondary AML (sAML) samples. Mutant allele frequencies represent the proportion of sequencing reads supporting the mutant allele reads/total reads. Total read counts are shown below (mean 5,651 reads/sample). The mutation is present in the majority of cells (mutant allele frequency 31.4-48.2%) in all cases. (**c**) Deep sequencing of cDNA from MDS or sAML samples. The mutant allele is expressed in all cases. UPN, unique patient number.

Figure 2. Impact of *U2AF1* mutations on clinical outcome. (**a**) Overall and (**b**) Event-free survival are not impacted by *U2AF1* genotype. (**c**) The probability of secondary AML (sAML) progression is increased in patients with *U2AF1* mutations (P=0.03).

Figure 3. *U2AF1* S34F mutation induces splicing alterations. (**a**) Transient coexpression of the pTN24 double-reporter splicing construct with or without a splicing enhancer (Tra2α), splicing inhibitor (hnRNPG), wild-type *U2AF1*, or mutant (S34F) *U2AF1* cDNA in 293T cells. The double-reporter construct constitutively expresses β-galactosidase, while luciferase is expressed only if an upstream intron which contains multiple stop codons is spliced out. In the absence of the Tra2α splicing enhancer or hnRNPG splicing inhibitor, expression of the mutant

U2AF1 increases splicing of the pTN24 construct resulting in an increase in the luciferase/β-galactosidase expression compared to expression of wild-type *U2AF1* ($P<0.001$). Tra2α (positive control) and hnRNPG (negative control) cause increased or decreased splice efficiency, respectively. A representative Western blot of U2AF1 levels in the same cells used for luciferase assays is shown below each combination of plasmids. (**b)** Transient coexpression of the pTN24 double-reporter splicing construct with a control plasmid (vector) or mutant (S34F) *U2AF1* cDNA in the presence of a control siRNA or siRNA targeting the endogenous *U2AF1* in 293T cells. Expression of the S34F mutant *U2AF1* results in an increase in splicing of the pTN24 construct with an increase in the luciferase/β-galactosidase expression compared to a control plasmid ($P<0.001$). The result is independent of endogenous U2AF1 expression levels. A representative Western blot of U2AF1 is shown below each condition. (**c**) The *GH1* minigene was transiently transfected into 293T cells with a control plasmid (pcDNA3.1-YFP), wild-type *U2AF1* cDNA, or mutant (S34F) *U2AF1* cDNA in the presence of a control siRNA or siRNA targeting the endogenous *U2AF1*. RNA was harvested 48 hours later and a reverse transcriptase (RT) reaction was performed to create cDNA. PCR using the indicated primers resulted in a fully spliced 505 base pair amplicon or a 386 base pair amplicon that skips the middle exon shaded in black (exon skipping). A representative PCR gel image is shown and the ratio of the lower band (exon skipping = amplicon b) relative to the fully spliced upper band (amplicon a) is shown above each condition. Expression of the S34F mutant *U2AF1* results in an increase in exon skipping compared to control or wild-type *U2AF1* ($P<0.02$). bp, base pair; WT, wildtype; Mut, mutant; T7, T7 primer. Error bars, s.d.

Table 1. Patient Characteristics

	U2AF1 wild-type	*U2AF1* S34F/Y mutant	p-value[a]
n (%)	137 (91.3)	13 (8.7)	
Age at diagnosis – yr ±SD	61 ±14	59 ±10	0.37[b]
range (median)	20-87 (62)	42-79 (59)	
Median survival	975 days	729 days	0.81[c]
Gender			
Male – no.(%)	82 (59.9)	10 (76.9)	0.37
Blood counts			
WBC (K/mcL)	6 ±11	6 ±11	0.22[b]
ANC (K/mcL)	3 ±7	3 ±6	0.43[b]
Hb (g/dL)	10 ±2	10 ±2	0.94[b]
PLT (K/mcL)	91 ±88	78±66	0.94[b]
Bone marrow blasts (%)	6 ±6	8 ±7	0.41 [b]
FAB subtype – no.(%)			0.44
RA	63 (46.0)	4 (30.8)	
RARS	5 (3.7)	0 (0)	
RAEB	64 (46.7)	8 (61.5)	
RAEB-T	5 (3.7)	1 (7.7)	
Cytogenetics – no.(%)			
normal	66 (48.2)	3 (23.1)	0.14
-Y only	2 (1.5)	0 (0)	1.00
-5, del(5q)	29 (21.2)	1 (7.7)	0.47
-7, del(7q)	16 (11.7)	1 (7.7)	1.00
-17, del(17q)	5 (3.7)	0 (0)	1.00
del(20q)	8 (5.8)	4 (30.8)	***0.01***
+8	16 (11.7)	2 (15.4)	0.66
complex (≥ 3)	35 (25.6)	1 (7.7)	0.19
other	15 (11.0)	3 (23.1)	0.19
not available	4 (2.9)	0 (0)	1.00
IPSS – no.(%)			0.50
low	22 (16.1)	1 (7.7)	
INT-1	54 (39.4)	6 (45.2)	
INT-2	37 (27.0)	2 (15.4)	
high	19 (13.9)	4 (30.8)	
not available	5 (3.7)	0 (0)	

WBC, white cell blood count; ANC, absolute neutrophil count; Hb, hemoglobin; PLT, platelet count; FAB, French American British; RA, refractory anemia; RARS, refractory anemia with ringed sideroblasts; RAEB, refractory anemia with excess blasts; RAEB-T, refractory anemia with excess blasts, in transformation; IPSS, International Prognostic Scoring System.

[a] Fisher's exact, except where indicated; [b] Wilcoxon test; [c] Log-rank

ONLINE METHODS

Flow sorting of bone marrow samples. Bone marrow cells from the secondary AML (sAML) sample, cryopreserved in 10% DMSO, were rapidly thawed at 37°C, washed, and stained with PE-Cy7 conjugated hCD45, clone J.33 (Beckman Coulter), and FITC conjugated anti-hCD34, clone 581 (Beckman Coulter). The blast population (low SSC/CD45 dim) was sorted using a Reflection high speed cell sorter (Sony iCyt) directly into lysis buffer and genomic DNA was prepared by column purification (Qiagen DNeasy).

Whole genome sequence production. Four DNA libraries were generated for paired-end sequencing: two from the tumor sample (flow-sorted sAML myeloblasts), and two from the normal sample (punch biopsy of unaffected skin). Sequence data was generated using both Illumina GAIIx and Illumina Hiseq platforms in 2 x 100 paired-end reads. Reads were aligned individually to NCBI Build 36 of the human reference sequence using BWA 0.5.5 and SAMtools r544. Alignments were merged into a single BAM file and marked for duplicates using Picard 1.17 (http://picard.sourceforge.net). Only non-duplicate reads were used for all downstream analyses.

Somatic mutation detection. Candidate point mutations were predicted using SomaticSniper (D. Larson et al, in press), previously referred to as glfSomatic[4,35]. Putative single nucleotide variants (SNVs) with somatic score of 40 and average mapping quality of 40 were considered high-confidence (HC); all others were deemed low-confidence (LC). Small (<100 bp) insertion/deletion events (indels) were called using a combination of GATK[36], IndelGenotyper, Pindel[37], and a modified version of SAMtools[38]. Both SNVs and indels were annotated using gene structure and conservation information, and classified by tier as previously described[4]. Briefly, tier 1 contains all changes in the amino acid coding regions of annotated exons, consensus splice-site regions, and RNA genes (including microRNA genes). Tier 2 contains changes in highly conserved regions of the genome or regions that have regulatory potential. Tier 3 contains mutations in the nonrepetitive part of the genome that do not meet tier 2 criteria, and tier 4 contains mutations in the remainder of the genome. High confidence tier 2 and tier 3 mutations, and all tier 1 mutations (regardless of confidence) were selected for validation (see below).

To identify somatic DNA copy number changes from whole genome sequencing (WGS) data, reads aligned by BWA[39] were binned into contiguous, non-overlapping 1 kb windows. Copy number for each bin was normalized to the median copy number for each chromosome in tumor and normal separately. A Hidden Markov Model algorithm[40] was used to generate a list of segments with copy number expressed as $\log_2$ (tumor/normal). Copy number changes were also supported by demonstrating loss of heterozygosity (LOH) in the affected regions. In brief, heterozygous SNPs were identified in WGS data from the normal sample (>10 reads of >q10 quality with non-reference allele frequencies of 0.4-0.6). The variant allele frequencies at these positions were then averaged in bins of 20 consecutive SNPs and visualized for the normal and tumor samples separately. Deletions, amplifications, inter-, and intrachromsomal rearrangements were also predicted using the BreakDancer algorithm[41].

Mutation validation. To comprehensively evaluate tier 1-3 predictions, we utilized a custom solid-phase capture platform. We selected all tier 1 SNV predictions (HC and LC) and all HC tier 2-3 SNVs. Tier 1-3 indel predictions were also included. In addition, we used this approach to validate SV predictions (deletions and rearrangements). We identified 8-16 SNPs that were

heterozygous in the normal DNA sample (determined using the WGS and SNPa data) that were located within the affected segments and 8 SNPs from flanking normal regions. The genomic positions of SNVs and indels (with a 200 bp margin) and SVs (with a 400 bp margin) were submitted for probe design. Probes were synthesized on custom HD2.1 long oligonucleotide arrays (Roche NimbleGen). Whole genome amplified DNA (REPLI-g, Qiagen) from the normal (skin), unfractionated MDS, and unfractionated sAML samples was used as bait for capture on the arrays and the recovered DNA (enriched for target sequences) was resequenced on the Illumina GAIIx platform.

At least 10x coverage was obtained for ~87.16% of the target sequence for all samples (**Supplementary Table 1**). Reads were mapped using BWA[39], deduplicated, and merged into BAM files. The reference or somatic status at the nucleotide of interest was then determined for each sample using VarScan2[42] with the following parameters: min-coverage=10, min-var-freq=0.05, somatic-p-value<0.01, validation=1. To validate low-frequency (2-5%) SNVs, we re-ran VarScan with adjusted parameters: min-coverage=100, min-var-freq=0.02, somatic-p-value<0.01, validation=1. In validation mode, VarScan reads data from tumor and normal samples simultaneously, performing pair-wise comparisons at every position covered in both samples. Each position is classified as Reference (wild-type), Germline, LOH, or Somatic, based upon a comparison of the consensus genotypes and supporting read counts (Fisher's Exact test). Positions called Somatic are further subjected to our internally-developed false-positive filter which removes sequencing- and alignment-related artifacts using several criteria (read count, mapping quality, average read position, strand representation, homopolymer-like sequence context, mismatch quality sum difference, trimmed read length, Q2 distance) and were manually reviewed. Chromosome X and Y somatic positions are determined using the false-positive filter and manual review. SIFT and PolyPhen2 computational algorithms were used to predict whether *U2AF1* mutations were damaging, as previously described[43,44].

Sanger sequencing. To screen for recurrence of *U2AF1* mutations, we performed Sanger sequencing using whole genome amplified DNA extracted from unfractionated bone marrow aspirates and paired normal tissue (skin) from 150 individual patients with de novo MDS. PCR amplicons covering all 9 exons and splice sites in *U2AF1* were sequenced using BigDye chemistry and analyzed on an ABI 3730 sequencer (primer sequences in **Supplementary Table 8**). Sequence variants were called by The Genome Institute's mutational profiling pipeline and manually reviewed. Potential somatic mutations (present in the bone marrow sample and not detectable in skin) were confirmed by independent PCR and sequencing.

Deep sequencing of U2AF1 mutations in DNA and cDNA. Unfractionated bone marrow samples from 11 patients with validated *U2AF1* mutations were selected for deep sequencing to estimate clone size. Whole genome amplified DNA from normal (skin), MDS, and sAML samples were amplified by PCR individually using barcoded primers (**Supplementary Table 8**). The products were then pooled and sequenced on the Roche/454 platform. In parallel, RNA was extracted from MDS and sAML samples (Trizol, Invitrogen), converted to cDNA using the Ovation RNA-seq Kit (NuGEN), and amplified with barcoded primers spanning intron/exon boundaries (**Supplementary Table 8**). Reads were aligned to Hs36 using BWA-SW[39]. Following alignment, BAM and pileup files were generated using SAMtools and analyzed by Picard to remove duplicates. Only uniquely mapped bases with >q20 scores were retained. Reads supporting the reference or variant allele were identified by VarScan2.

***SNP array analysis*.** Genomic DNA samples (not subjected to whole genome amplification) from the normal, MDS, and sAML specimens (not flow-sorted) were hybridized to Affymetrix 6.0 SNP arrays (Affymetrix, Inc.). Analysis of copy number alterations and copy neutral loss of heterozygosity was performed using the Partek Genomics Suite (Partek, Inc).

***mRNA expression profiling*.** Total RNA was harvested from unfractionated bone marrow cells (69% myeloblast) from the secondary AML sample from UPN 266395 and hybridized to the Affymetrix Exon 1.0 ST array. Raw data was extracted using the Affymetrix Expression Console (Affymetrix, Inc.) and analyzed in Prism 5.04 (GraphPad Software, Inc.).

Total RNA was harvested from CD34+ purified MDS bone marrow samples (n=15) and control bone marrow (n=4) and hybridized to the Affymetrix U133plus2 array, as previously reported[18]. Supervised hierarchical clustering was performed using Ward's clustering algorithm with a Euclidean distance similarity measure in Spotfire (TIBCO Software Inc). Significance Analysis of Microarrays (SAM), Gene Set Enrichment Analysis (GSEA), and Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 were performed as previously described[19,45,46].

Reverse transcriptase PCR. cDNA was made from RNA using Moloney Murine Leukemia Virus (M-MLV) reverse transcriptase or Superscript III kit (Invitrogen). Quantitative real-time RT-PCR was performed using TaqMan Universal PCR Master Mix (Applied Biosystems) (primer and probe sequences for GAPDH are provided in[47]). The *U2AF1* primer and probe set spans exons 2-3 (Hs00739599_m1, Applied Biosystems). All samples were run in duplicate on a 7300 Real-Time PCR system (Applied Biosystems) and analyzed using the relative standard curve method. Non-quantitative RT-PCR for *U2AF1* mRNA isoform expression was performed and loaded on a 10% polyacrylamide gel (Forward: 5'-GCCTCCATCTTCGGCACCGA-3', Reverse: 5'-GGCATGGCTCAGAATCGCCC-3').

Generation of U2AF1 expression plasmids. RNA from a patient bone marrow biopsy (UPN 571656) was used to generate both *U2AF1* (wildtype) and *U2AF1* (S34F mutant) expression vectors. cDNA was generated from patient bone marrow RNA using Superscript III kit (Invitrogen). Both wild type and mutant *U2AF1* cDNAs were obtained via PCR amplification, cloned with the Topo Cloning kit (Invitrogen), and sequenced for verification. The *U2AF1* cDNAs were then shuttled into the EcoRI site of the pcDNA3.1+ vector (Invitrogen) for transient transfection experiments.

Luciferase-β-galactosidase double-reporter assay. 293T cells were seeded in a 6-well plate (1 x 10^6 per well) and cultured in DMEM (Gibco/Invitrogen) supplemented with 10% FBS and L-glutamine. Following overnight culture, cells were co-transfected with the expression vectors (*U2AF1* wildtype or S34F mutant, or empty vector) and the pTN24 splicing reporter plasmid (containing a constitutively expressed β-galactosidase reporter for transfection normalization and a luciferase reporter that is conditional on removal of a translational stop codon by splicing) with or without splicing modulators hnRNPG and Tra2α[17]. In some experiments, cells were also co-transfected with 30nM *U2AF1*-specific siRNA (5'-CGUAGAAAGUGUUGUAGUUGAUUGA-3'; IDT, Inc.) or 30nM siRNA scramble control (Dharmacon). Cells were harvested 48 hours following transfection, and reporter expression was detected, as previously described[17] using the Dual Light Reporter System (Applied Biosystems) and analyzed by calculating the ratio of luciferase to β-galactosidase signal. Changes in U2AF1 levels were confirmed by Western blot using antibodies specific for U2AF1 (SAS1300700, Sigma-Aldrich) or β-actin (A5441, Sigma-

Aldrich) as a loading control. Three independent experiments were performed and the data was analyzed using a Student's t-test.

Minigene constructs and transfection. 293T cells were cultured and transfected with *U2AF1* expression vectors, as above. Cells were also co-transfected with a *GH1* minigene splicing reporter construct[9]. In other experiments, cells were co-transfected with a *FMR1* minigene splicing reporter construct containing partial sequence from exons 14 and 15 and the complete intronic sequence. Amplification of the *FMR1* DNA fragment (including the intron) was achieved using the *FMR1-201* E15 set of primer sequences previously published[10]. The amplified fragment was cloned into the TopoTA vector (Invitrogen), purified following BamH1 and XhoI digestion, and subsequently cloned into pcDNA3.1 (Invitrogen). Co-transfection of *U2AF1* or control siRNAs was also performed, as above. Cells were harvested 48 hours following transfection, and RNA was extracted using the RNeasy reagent (Qiagen) following the manufacturer's instructions. The RNA was used as a template for cDNA synthesis via RT-PCR with random hexamers and oligo(dT) primers. Changes in minigene splicing were then measured by PCR of the cDNA using a T7 forward primer and gene-specific reverse primers as previously described[9,10] and quantified by densitometry. *U2AF1* knockdown by siRNA and U2AF1 reconstitution were confirmed by Western blot analysis, as above. Three to four independent experiments were performed and the data was analyzed using a Student's t-test.

LITERATURE CITED

1. Webb, C.J. & Wise, J.A. The splicing factor U2AF small subunit is functionally conserved between fission yeast and humans. *Mol Cell Biol* 24, 4229-40 (2004).
2. Wu, S., Romfo, C.M., Nilsen, T.W. & Green, M.R. Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature* 402, 832-5 (1999).
3. Welch, J.S. et al. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA* 305, 1577-84 (2011).
4. Mardis, E.R. et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* 361, 1058-66 (2009).
5. Link, D.C. et al. Identification of a Novel TP53 Cancer Susceptibility Mutation Through Whole-Genome Sequencing of a Patient With Therapy-Related AML. *JAMA* 305, 1568-1576 (2011).
6. Ley, T.J. et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456, 66-72 (2008).
7. Ley, T.J. et al. DNMT3A mutations in acute myeloid leukemia. *N Engl J Med* 363, 2424-33 (2010).
8. Wahl, M.C., Will, C.L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* 136, 701-18 (2009).
9. Fu, Y., Masuda, A., Ito, M., Shinmi, J. & Ohno, K. AG-dependent 3'-splice sites are predisposed to aberrant splicing due to a mutation at the first nucleotide of an exon. *Nucleic Acids Res* 39, 4396-404 (2011).
10. Kralovicova, J. & Vorechovsky, I. Allele-specific recognition of the 3' splice site of INS intron 1. *Hum Genet* 128, 383-400 (2010).

11. Pacheco, T.R., Coelho, M.B., Desterro, J.M., Mollet, I. & Carmo-Fonseca, M. In vivo requirement of the small subunit of U2AF for recognition of a weak 3' splice site. *Mol Cell Biol* 26, 8183-90 (2006).
12. Pacheco, T.R., Moita, L.F., Gomes, A.Q., Hacohen, N. & Carmo-Fonseca, M. RNA interference knockdown of hU2AF35 impairs cell cycle progression and modulates alternative splicing of Cdc25 transcripts. *Mol Biol Cell* 17, 4187-99 (2006).
13. Hudson, B.P., Martinez-Yamout, M.A., Dyson, H.J. & Wright, P.E. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nat Struct Mol Biol* 11, 257-64 (2004).
14. Lai, W.S., Kennington, E.A. & Blackshear, P.J. Interactions of CCCH zinc finger proteins with mRNA: non-binding tristetraprolin mutants exert an inhibitory effect on degradation of AU-rich element-containing mRNAs. *J Biol Chem* 277, 9606-13 (2002).
15. Liang, J., Song, W., Tromp, G., Kolattukudy, P.E. & Fu, M. Genome-wide survey and expression profiling of CCCH-zinc finger family reveals a functional module in macrophage activation. *PLoS One* 3, e2880 (2008).
16. Kielkopf, C.L., Rodionova, N.A., Green, M.R. & Burley, S.K. A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. *Cell* 106, 595-605 (2001).
17. Nasim, M.T. & Eperon, I.C. A double-reporter splicing assay for determining splicing efficiency in mammalian cells. *Nat Protoc* 1, 1022-8 (2006).
18. Graubert, T.A. et al. Integrated genomic analysis implicates haploinsufficiency of multiple chromosome 5q31.2 genes in de novo myelodysplastic syndromes pathogenesis. *PLoS ONE* 4, e4583 (2009).
19. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98, 5116-21 (2001).
20. Delhommeau, F. et al. Mutation in TET2 in myeloid cancers. *N Engl J Med* 360, 2289-301 (2009).
21. Ebert, B.L. et al. Identification of RPS14 as a 5q- syndrome gene by RNA interference screen. *Nature* 451, 335-9 (2008).
22. Langemeijer, S.M. et al. Acquired mutations in TET2 are common in myelodysplastic syndromes. *Nat Genet* 41, 838-42 (2009).
23. Ernst, T. et al. Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nat Genet* 42, 722-6 (2010).
24. Nikoloski, G. et al. Somatic mutations of the histone methyltransferase gene EZH2 in myelodysplastic syndromes. *Nat Genet* 42, 665-7 (2010).
25. Gelsi-Boyer, V. et al. Mutations of polycomb-associated gene ASXL1 in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *Br J Haematol* 145, 788-800 (2009).
26. Walter, M.J. et al. Recurrent DNMT3A mutations in patients with myelodysplastic syndromes. *Leukemia* 25, 1153-8 (2011).
27. Golling, G. et al. Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development. *Nat Genet* 31, 135-40 (2002).
28. Rudner, D.Z., Kanaar, R., Breger, K.S. & Rio, D.C. Mutations in the small subunit of the Drosophila U2AF splicing factor cause lethality and developmental defects. *Proc Natl Acad Sci U S A* 93, 10333-7 (1996).
29. Zorio, D.A. & Blumenthal, T. U2AF35 is encoded by an essential gene clustered in an operon with RRM/cyclophilin in Caenorhabditis elegans. *RNA* 5, 487-94 (1999).

30. Grosso, A.R., Martins, S. & Carmo-Fonseca, M. The emerging role of splicing factors in cancer. *EMBO Rep* 9, 1087-93 (2008).
31. David, C.J. & Manley, J.L. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev* 24, 2343-64 (2010).
32. Visconte, V. et al. SF3B1, a splicing factor is frequently mutated in refractory anemia with ring sideroblasts. *Leukemia,* Sep 2. doi: 10.1038 (2011).
33. Papaemmanuil, E. et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med,* Sept 26. doi: 10.1056 (2011).
34. Yoshida, K. et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* 478, 64-9 (2011).
35. Ding, L. et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* 464, 999-1005 (2010).
36. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297-303 (2010).
37. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2865-71 (2009).
38. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-9 (2009).
39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-60 (2009).
40. Baum, L.E. & Eagon, J.A. An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology. *Bulletin of the American Mathematical Society* 73, 360-363 (1967).
41. Chen, K. et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 6, 677-81 (2009).
42. Koboldt, D.C. et al. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* (2009).
43. Adzhubei, I.A. et al. A method and server for predicting damaging missense mutations. *Nat Methods* 7, 248-9 (2010).
44. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31, 3812-4 (2003).
45. Dennis, G., Jr. et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4, P3 (2003).
46. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102, 15545-50 (2005).
47. Fortier, J.M. et al. POU4F1 is associated with t(8;21) acute myeloid leukemia and contributes directly to its unique transcriptional signature. *Leukemia* 24, 950-7 (2010).

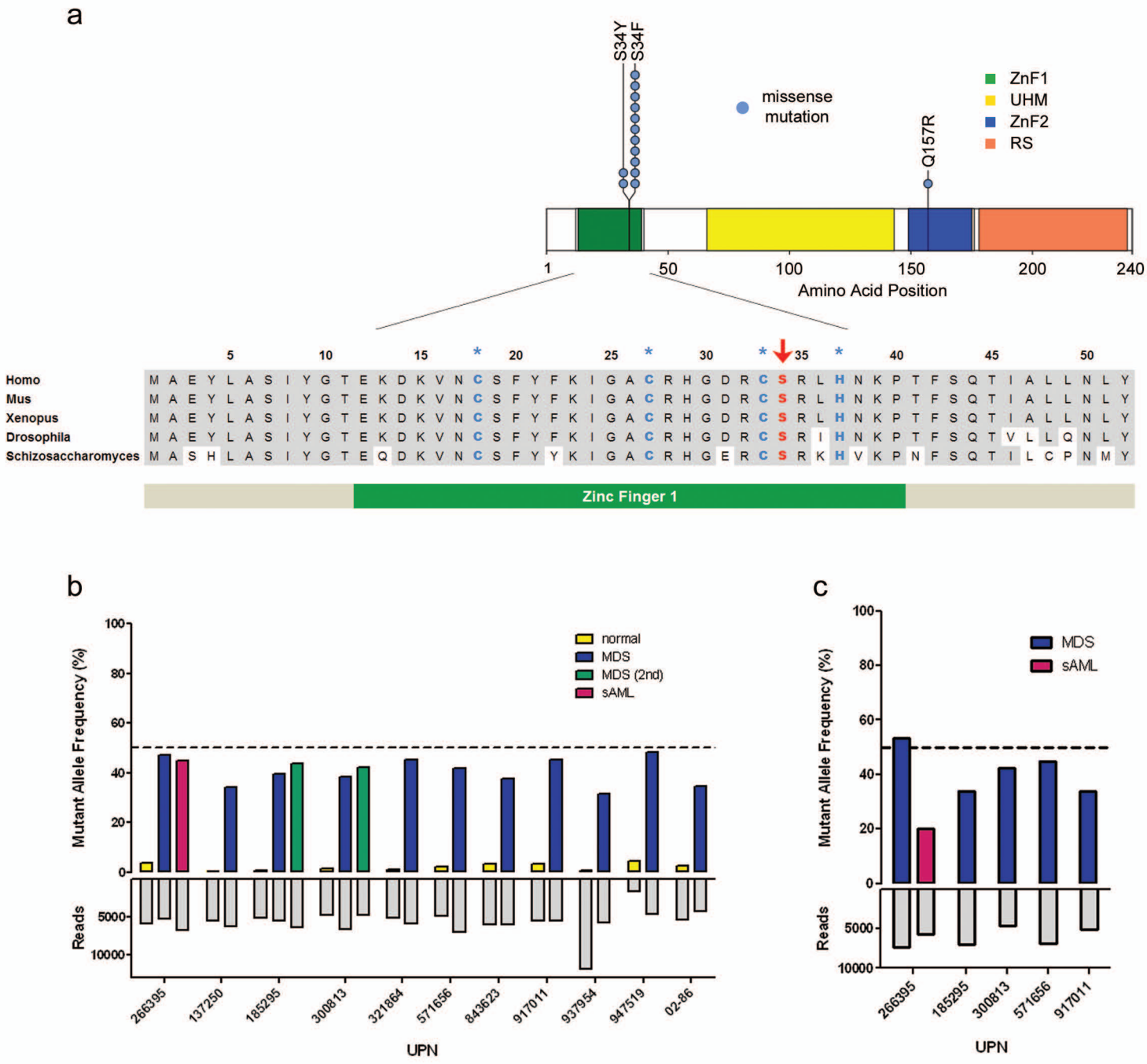
a
S34Y
S34F
missense mutation
Q157R
ZnF1
UHM
ZnF2
RS
1
50
100
150
200
240
Amino Acid Position
5
10
15
20
25
30
35
40
45
50
Homo M A E Y L A S I Y G T E K D K V N C S F Y F K I G A C R H G D R C S R L H N K P T F S Q T I A L L N L Y
Mus M A E Y L A S I Y G T E K D K V N C S F Y F K I G A C R H G D R C S R L H N K P T F S Q T I A L L N L Y
Xenopus M A E Y L A S I Y G T E K D K V N C S F Y F K I G A C R H G D R C S R L H N K P T F S Q T I A L L N L Y
Drosophila M A E Y L A S I Y G T E K D K V N C S F Y F K I G A C R H G D R C S R I H N K P T F S Q T V L L Q N L Y
Schizosaccharomyces M A S H L A S I Y G T E Q D K V N C S F Y Y K I G A C R H G E R C S R K H V K P N F S Q T I L C P N M Y
Zinc Finger 1
b
Mutant Allele Frequency (%)
normal
MDS
MDS (2nd)
sAML
Reads
266395
137250
185295
300813
321864
571656
843623
917011
937954
947519
02-86
UPN
c
Mutant Allele Frequency (%)
MDS
sAML
Reads
266395
185295
300813
571656
917011
UPN

Figure 1

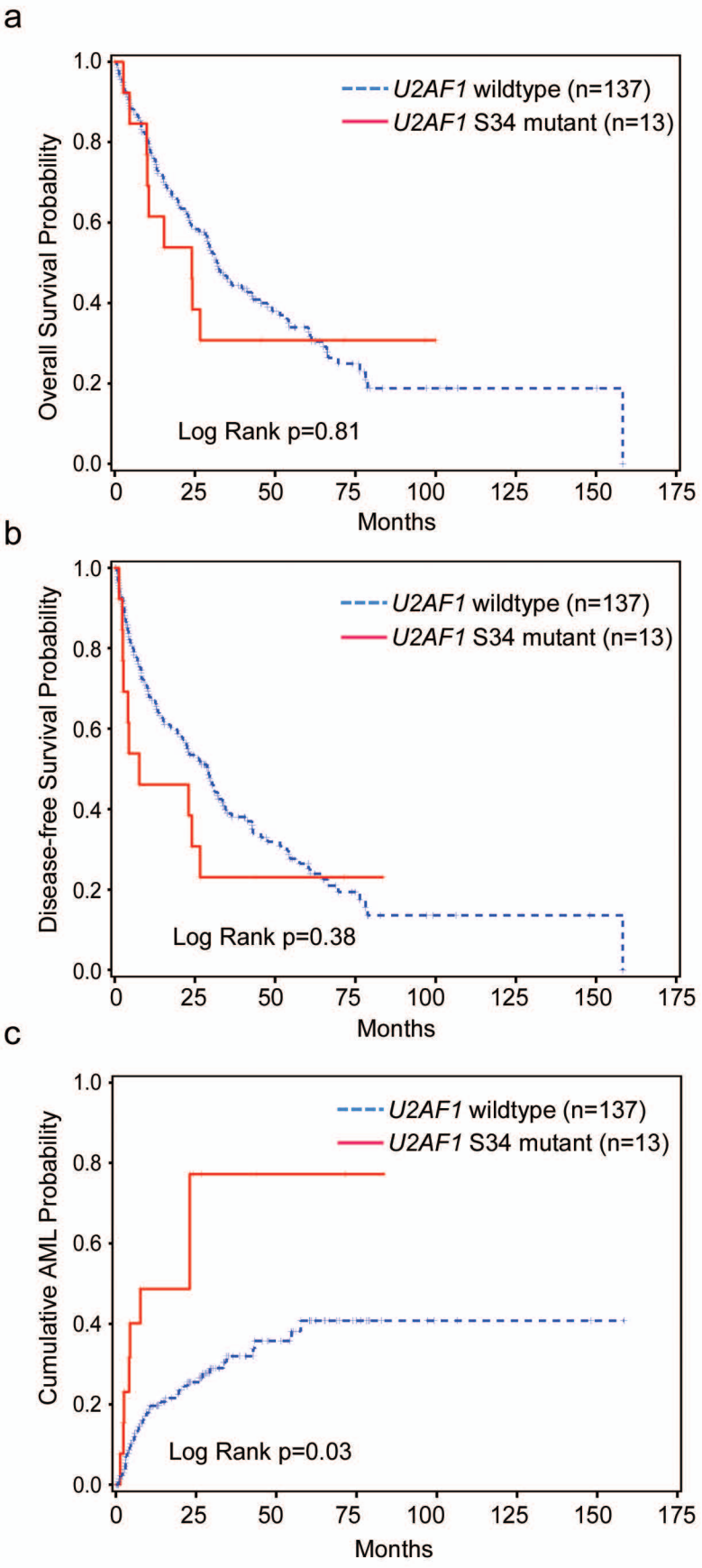
a
1.0
0.8
0.6
0.4
0.2
0.0
Overall Survival Probability
U2AF1 wildtype (n=137)
U2AF1 S34 mutant (n=13)
Log Rank p=0.81
0
25
50
75
100
125
150
175
Months
b
1.0
0.8
0.6
0.4
0.2
0.0
Disease-free Survival Probability
U2AF1 wildtype (n=137)
U2AF1 S34 mutant (n=13)
Log Rank p=0.38
0
25
50
75
100
125
150
175
Months
c
1.0
0.8
0.6
0.4
0.2
0.0
Cumulative AML Probability
U2AF1 wildtype (n=137)
U2AF1 S34 mutant (n=13)
Log Rank p=0.03
0
25
50
75
100
125
150
175
Months

Figure 2

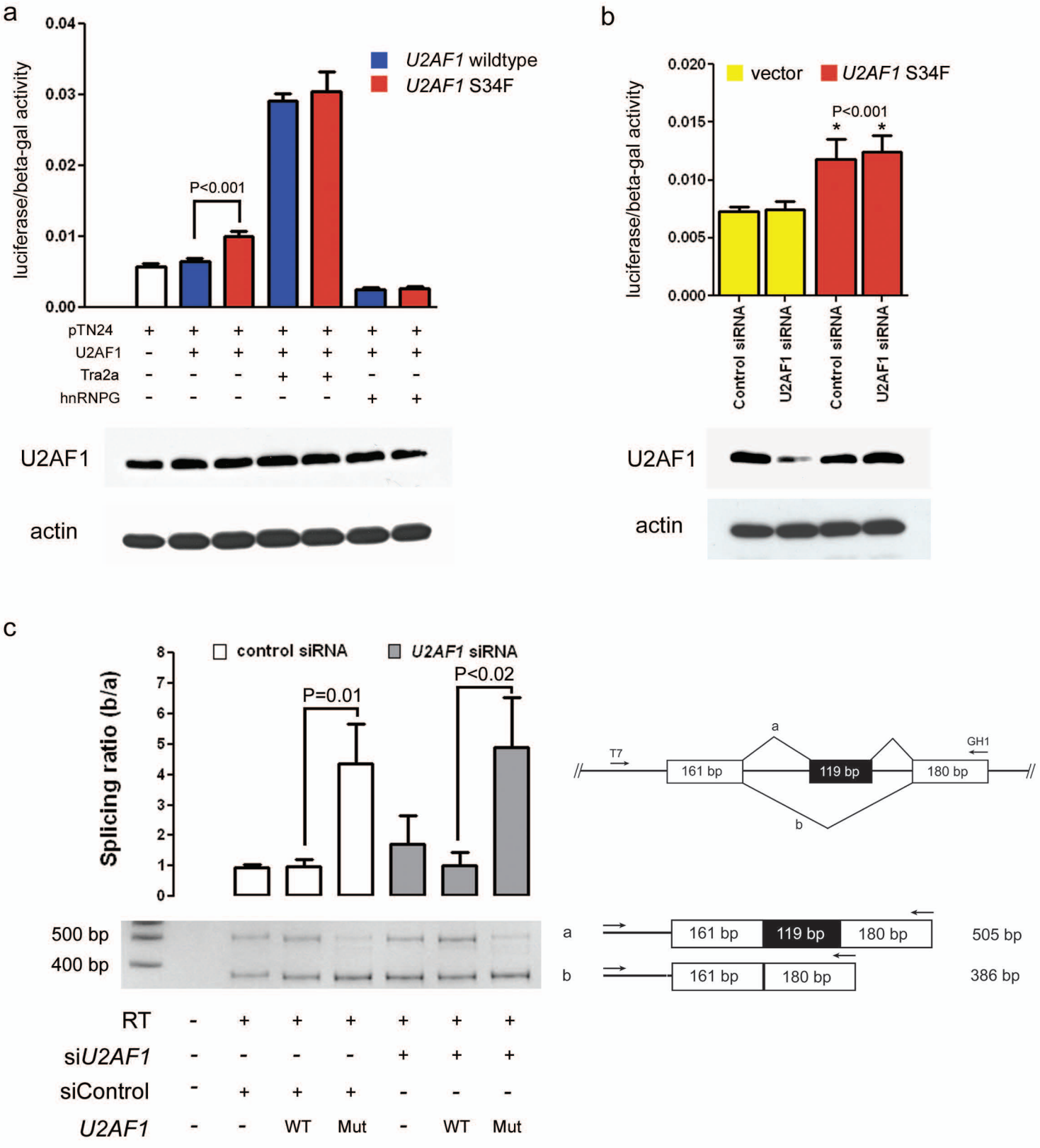
a
luciferase/beta-gal activity
0.04
0.03
0.02
0.01
0.00
U2AF1 wildtype
U2AF1 S34F
P<0.001
pTN24 + + + + + + +
U2AF1 - + + + + + +
Tra2a - - - + + - -
hnRNPG - - - - - + +
U2AF1
actin
b
luciferase/beta-gal activity
0.020
0.015
0.010
0.005
0.000
vector
U2AF1 S34F
P<0.001
* *
Control siRNA
U2AF1 siRNA
Control siRNA
U2AF1 siRNA
U2AF1
actin
c
Splicing ratio (b/a)
control siRNA
U2AF1 siRNA
P=0.01
P<0.02
T7
a
GH1
161 bp
119 bp
180 bp
b
500 bp
400 bp
a
161 bp
119 bp
180 bp
505 bp
b
161 bp
180 bp
386 bp
RT - + + + + + +
siU2AF1 - - - - + + +
siControl - + + + - - -
U2AF1 - - WT Mut - WT Mut

Figure 3