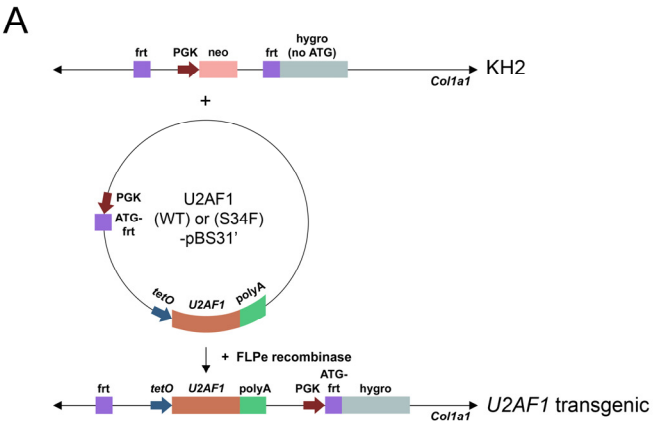
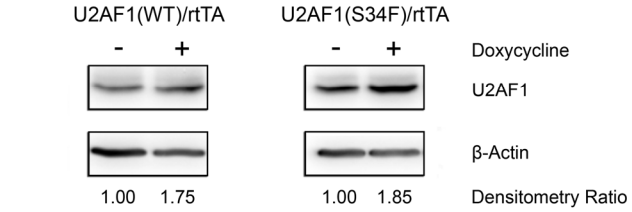
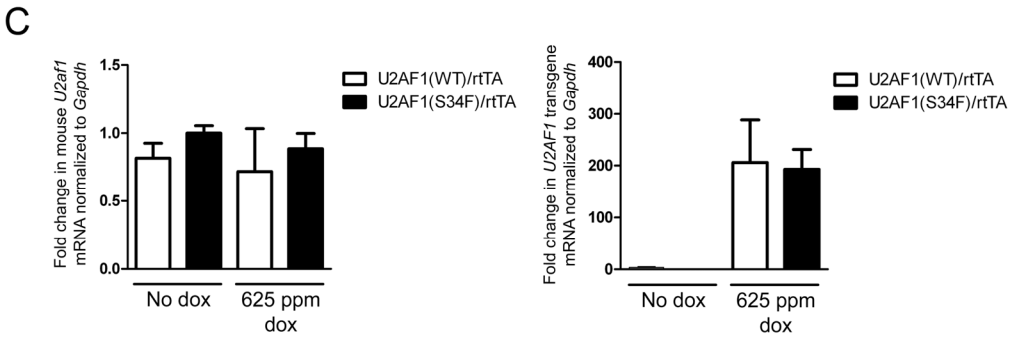
I. SUPPLEMENTAL DATA

Table S1, related to Figure 1. Spliceosome genes. Provided as an Excel file.


A
frt
PGK
neo
frt
hygro
(no ATG)
KH2
Col1a1
+
PGK
ATG-
frt
U2AF1
(WT) or (S34F)
-pBS31'
tetO
U2AF1
polyA
+ FLPe recombinase
frt
tetO
U2AF1
polyA
PGK
ATG-
frt
hygro
Col1a1
U2AF1 transgenic


B
U2AF1(WT)/rtTA
U2AF1(S34F)/rtTA
-
+
-
+
Doxycycline
U2AF1
β-Actin
1.00
1.75
1.00
1.85
Densitometry Ratio


C
Fold change in mouse U2af1 mRNA normalized to Gapdh
1.5
1.0
0.5
0.0
U2AF1(WT)/rtTA
U2AF1(S34F)/rtTA
No dox
625 ppm dox
Fold change in U2AF1 transgene mRNA normalized to Gapdh
400
300
200
100
0
U2AF1(WT)/rtTA
U2AF1(S34F)/rtTA
No dox
625 ppm dox

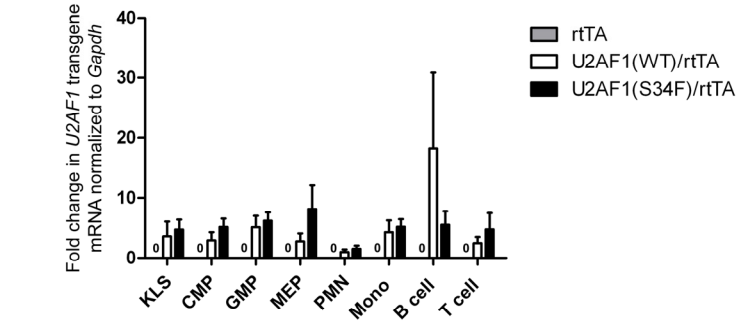
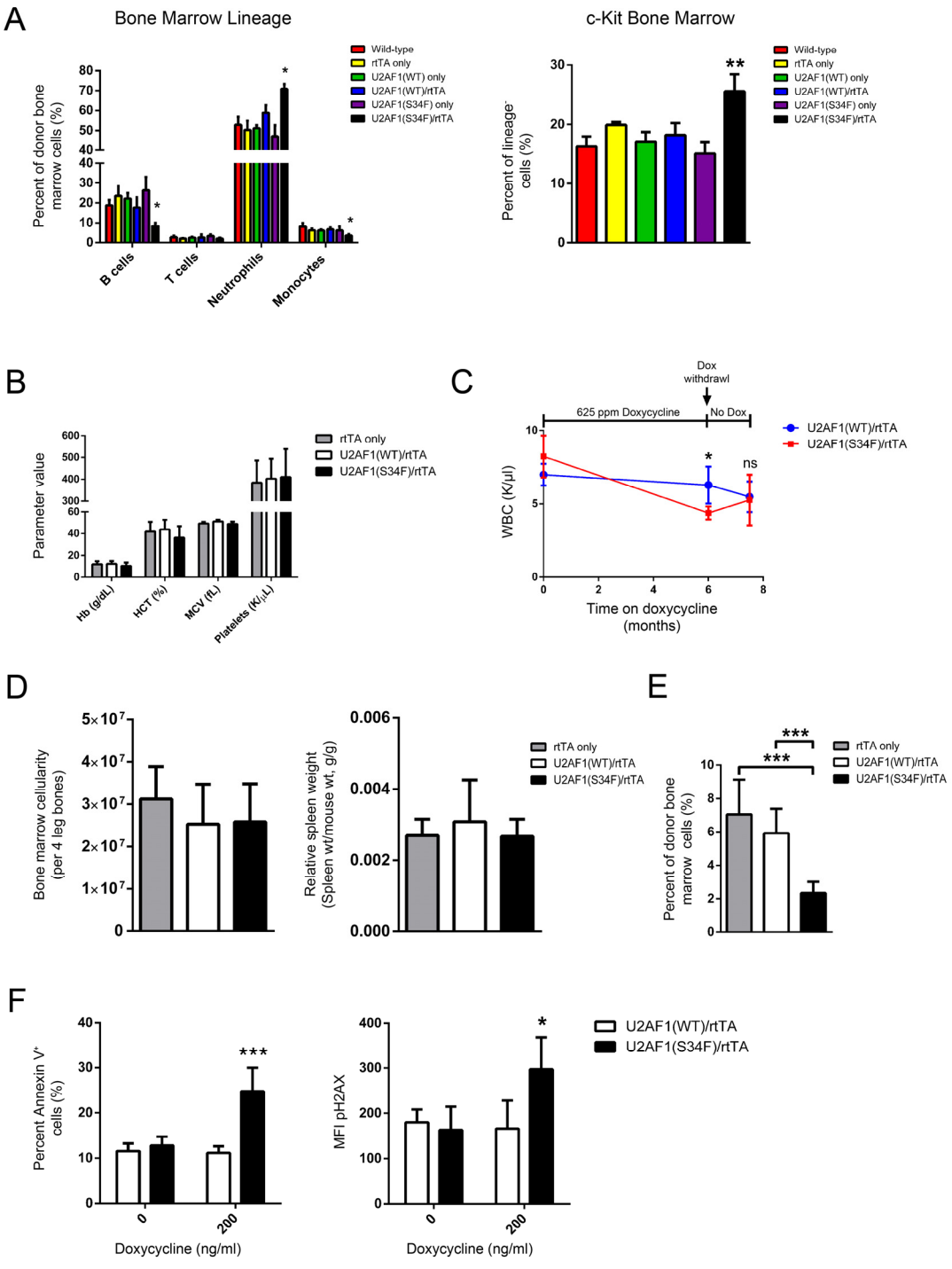

D
Fold change in U2AF1 transgene mRNA normalized to Gapdh
40
30
20
10
0
rtTA
U2AF1(WT)/rtTA
U2AF1(S34F)/rtTA
KLS
CMP
GMP
MEP
PMN
Mono
B cell
T cell

Figure S1, related to Figure 2. Doxycycline-inducible U2AF1(S34F) or U2AF1(WT) transgenic mice. (A) Diagram of FLP/FRT targeting approach for integration of the human U2AF1(WT) or U2AF1(S34F) cDNA into the *Col1a1* locus in mouse KH2 ES cells via the pBS31' plasmid. ES cell clones were selected by hygromycin resistance, which is conferred to cells by the integration of the pBS31' vector; pBS31' integration places a PGK promoter element and ATG start codon (previously lacking) in front of a hygromycin resistance gene to induce its expression. (B) Western blot of doxycycline-induced U2AF1 expression in U2AF1(S34F)/rtTA and U2AF1(WT)/rtTA c-Kit-enriched bone marrow cells following 24 hours in vitro culture with or without 75 ng/ml doxycycline, a concentration in vitro that induces ~2 fold *U2AF1* transgene mRNA expression relative to endogenous *U2af1* by pyrosequencing (data not shown) (n=2 mice pooled per genotype). The U2AF1/β-actin densitometry ratio is normalized to the corresponding "no doxycycline" sample. Results are representative of 2 independent experiments.
(C) Detection of endogenous mouse *U2af1* (left panel) and exogenous *U2AF1* transgene (right panel) expression by quantitative RT-PCR in transgenic mouse bone marrow following doxycycline induction for 5 days in vivo where indicated; results are normalized to *Gapdh*.
(D) Detection of *U2AF1* transgene expression by quantitative RT-PCR in different hematopoietic cell lineages of bone marrow following 5 days of doxycycline induction (n=3 for U2AF1(WT)/rtTA and U2AF1(S34F)/rtTA, n=2 for rtTA only). Results are normalized to *Gapdh*. Cell lineages are defined by cell surface markers used for flow cytometry sorting: donor-derived ($CD45.2^+$, $CD45.1^-$) KLS (stem cell-enriched population; lin^-, $c\text{-}Kit^+$, $Sca\text{-}1^+$), CMP (common myeloid progenitors; lin^-, $c\text{-}Kit^+$, $Sca\text{-}1^-$, $CD34^+$, $Fc\gamma^-$), GMP (granulocyte-monocyte progenitors; lin^-, $c\text{-}Kit^+$, $Sca\text{-}1^-$, $CD34^+$, $Fc\gamma^+$), MEP (megakaryocyte-erythroid progenitors; lin^-, $c\text{-}Kit^+$, $Sca\text{-}1^-$, $CD34^-$, $Fc\gamma^-$), PMN (neutrophils; $CD115^-$, $Gr\text{-}1^+$), monos (monocytes; $CD115^+$, $Gr\text{-}1^+$), B cells ($B220^+$), T cells ($CD3e^+$).


A
Bone Marrow Lineage
Percent of donor bone marrow cells (%)
Wild-type
rtTA only
U2AF1(WT) only
U2AF1(WT)/rtTA
U2AF1(S34F) only
U2AF1(S34F)/rtTA
B cells
T cells
Neutrophils
Monocytes
c-Kit Bone Marrow
Percent of lineage cells (%)
B
Parameter value
Hb (g/dL)
HCT (%)
MCV (fL)
Platelets (K/μL)
C
Dox withdrawl
625 ppm Doxycycline
No Dox
WBC (K/μl)
ns
Time on doxycycline (months)
D
Bone marrow cellularity (per 4 leg bones)
Relative spleen weight (Spleen wt/mouse wt, g/g)
E
Percent of donor bone marrow cells (%)
F
Percent Annexin V+ cells (%)
Doxycycline (ng/ml)
MFI pH2AX

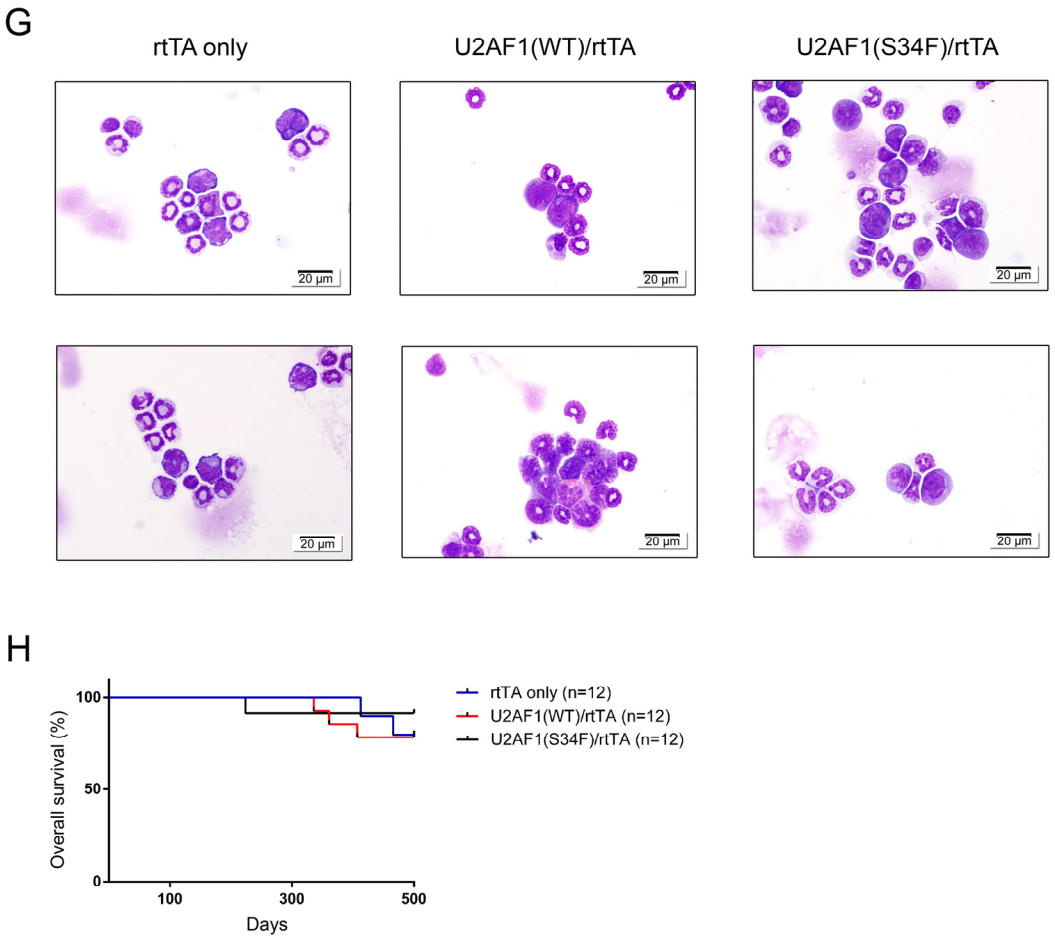


Figure S2, related to Figure 3. Hematopoietic alterations in U2AF1(S34F)/rtTA are mutant- and doxycycline-specific and reversible. (A) Flow cytometry for donor-derived monocytes ($CD115^+$, $Gr\text{-}1^+$), neutrophils ($CD115^-$, $Gr\text{-}1^+$), B cells ($B220^+$), T cells ($CD3e^+$) (left panel) and stem/progenitor-enriched cells (lin^-, $c\text{-}Kit^+$) (right panel) in mice transplanted with pooled bone marrow from one of six possible genotypes produced from U2AF1(S34F)/rtTA and U2AF1(WT)/rtTA transgenic mouse lines (n=3-5). (B) Peripheral blood hemoglobin (Hb), hematocrit (HCT), mean corpuscular volume (MCV), and platelet counts from transplanted mice after 1 month of doxycycline (n=9-11). (C) Reversibility of U2AF1(S34F)-induced leukopenia, measured by peripheral blood WBC counts of transplanted mice before doxycycline, after 6

months of doxycycline, and 6 weeks after doxycycline removal (*p<0.05, ns=not significant, n=4-5). (D) U2AF1(S34F)/rtTA-recipient mice have no differences in bone marrow cellularity, measured by the number of bone marrow cells from 4 leg bones (left panel), and relative spleen weight/mouse weight (right panel) following 1 month of doxycycline (n=9-11). (E) Flow cytometry for donor-derived bone marrow monocytes ($CD115^+$, $Gr\text{-}1^+$) in mice transplanted with transgenic donor marrow following 5 days of doxycycline (n=4-9). (F) U2AF1(S34F)/rtTA mouse bone marrow cells have increased Annexin V^+ (left panel) and intracellular phospho-H2AX (right panel) in vitro compared to U2AF1(WT)/rtTA bone marrow cells cultured for 5 days in media with myeloid cytokines (SCF, FLT3L, IL3, TPO) with increasing amounts of doxycycline (n=4 mice each). (G) Bone marrow cytospins were examined for dysplasia and morphologic abnormalities. Mice of all 3 genotypes had normal bone marrow myeloid cell morphology following 1 year of doxycycline. Representative micrographs per genotype are shown (n=2 each). (H) Kaplan-Meier curve of overall survival of transplanted mice (n=12). Time on doxycycline (in days) is plotted. All data are represented as mean (+/- SD). (*p<0.05, **p<0.01, ***p<0.001)

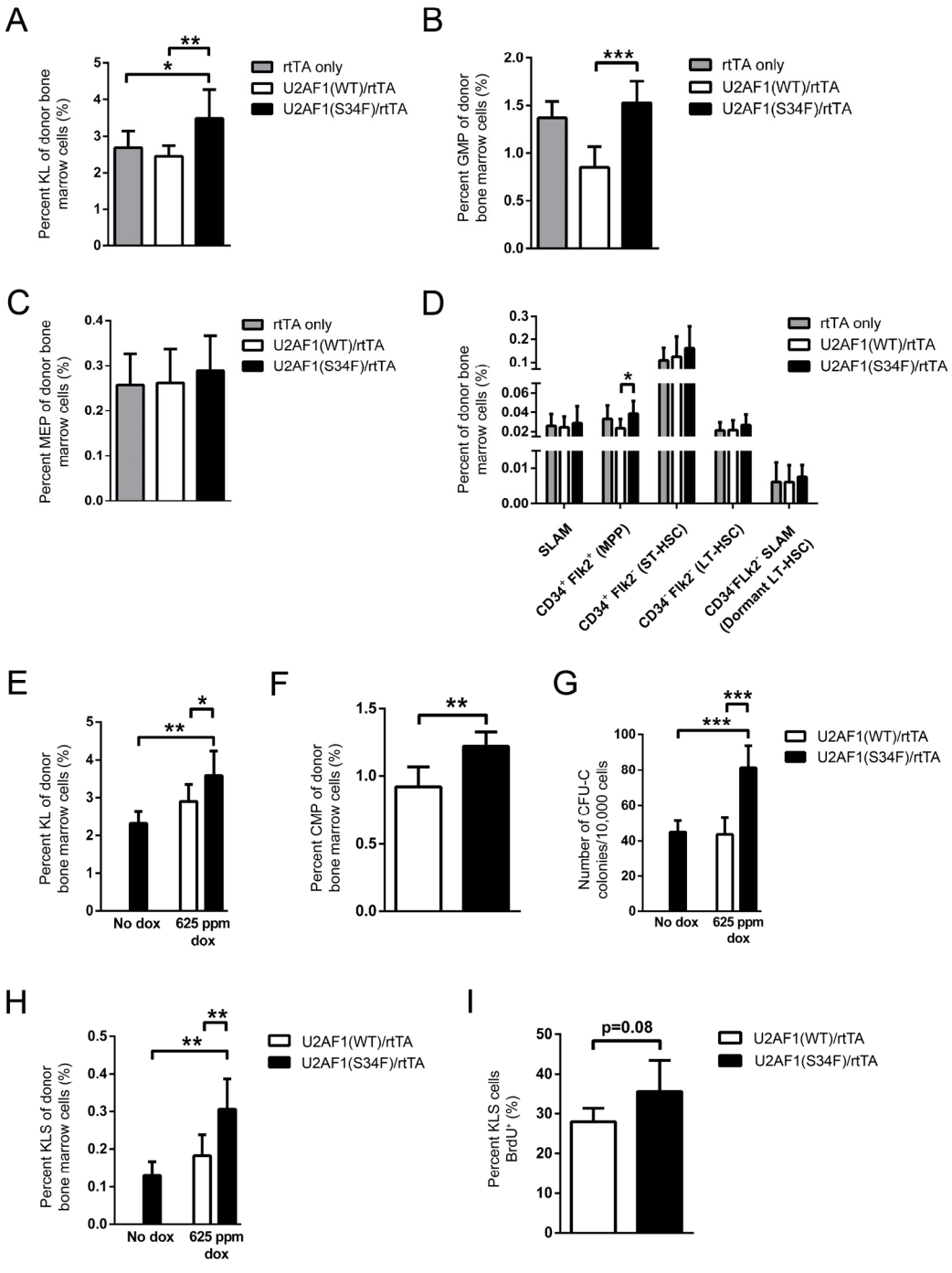

A
Percent KL of donor bone marrow cells (%)
rtTA only
U2AF1(WT)/rtTA
U2AF1(S34F)/rtTA
B
Percent GMP of donor bone marrow cells (%)
C
Percent MEP of donor bone marrow cells (%)
D
Percent of donor bone marrow cells (%)
SLAM
CD34+ Flk2- (MPP)
CD34+ Flk2- (ST-HSC)
CD34- Flk2- (LT-HSC)
CD34-Flk2- SLAM (Dormant LT-HSC)
E
Percent KL of donor bone marrow cells (%)
No dox
625 ppm dox
F
Percent CMP of donor bone marrow cells (%)
G
Number of CFU-C colonies/10,000 cells
H
Percent KLS of donor bone marrow cells (%)
I
p=0.08
Percent KLS cells BrdU+ (%)

Figure S3, related to Figure 4. U2AF1(S34F)-recipient mouse stem and progenitor cell expansion occurs as early as 5 days, but is not associated with specific HSC sub-populations. (A-C) Flow cytometry for donor-derived bone marrow KL (lin^-, $c\text{-}Kit^+$, $Sca\text{-}1^-$) cells (A), granulocyte-monocyte progenitor, GMP (lin^-, $c\text{-}Kit^+$, $Sca\text{-}1^-$, $CD34^+$, $Fc\gamma^+$) cells (B), and megakaryocyte-erythroid progenitor, MEP (lin^-, $c\text{-}Kit^+$, $Sca\text{-}1^-$, $CD34^-$, $Fc\gamma^-$) cells (C), in mice transplanted with transgenic donor marrow following 1 month of doxycycline (n=9-11). (D) Flow cytometry of donor-derived KLS (lin^-, $c\text{-}Kit^+$, $Sca\text{-}1^+$) subpopulations [SLAM ($CD150^+$, $CD48^-$), MPP ($CD34^+$, $Flk2^+$), ST-HSC ($CD34^+$, $Flk2^-$), LT-HSC ($CD34^-$, $Flk2^-$), dormant LT-HSC (SLAM, $CD34^-$, $Flk2^-$)] in the bone marrow of transplanted mice following 1 month of doxycycline (n=9-11). (E-G) U2AF1(S34F)/rtTA-recipient mice have increased bone marrow progenitor cells following 5 days of doxycycline treatment as shown by flow cytometry for donor-derived progenitors KL (E, n=8) and CMP (lin^-, $c\text{-}Kit^+$, $Sca\text{-}1^-$, $CD34^+$, $Fc\gamma^-$) (F, n=4) and by colony forming CFU-C assays (G, n=4-12) in mice transplanted with transgenic donor marrow. (H) U2AF1(S34F)/rtTA-recipient mice have an increase in the stem-cell enriched fraction of bone marrow following 5 days of doxycycline treatment by flow cytometry for donor-derived KLS (lin^-, $c\text{-}Kit^+$, $Sca\text{-}1^+$) cells (n=8). (I) U2AF1(S34F)/rtTA-recipient mice have increased BrdU staining of KLS cells, as measured by intracellular flow cytometry for DNA-incorporated BrdU measured in donor-derived bone marrow KLS cells following 5 days doxycycline (n=5). Data are represented as mean (+/- SD). (*$p<0.05$, **$p<0.01$, ***$p<0.001$)

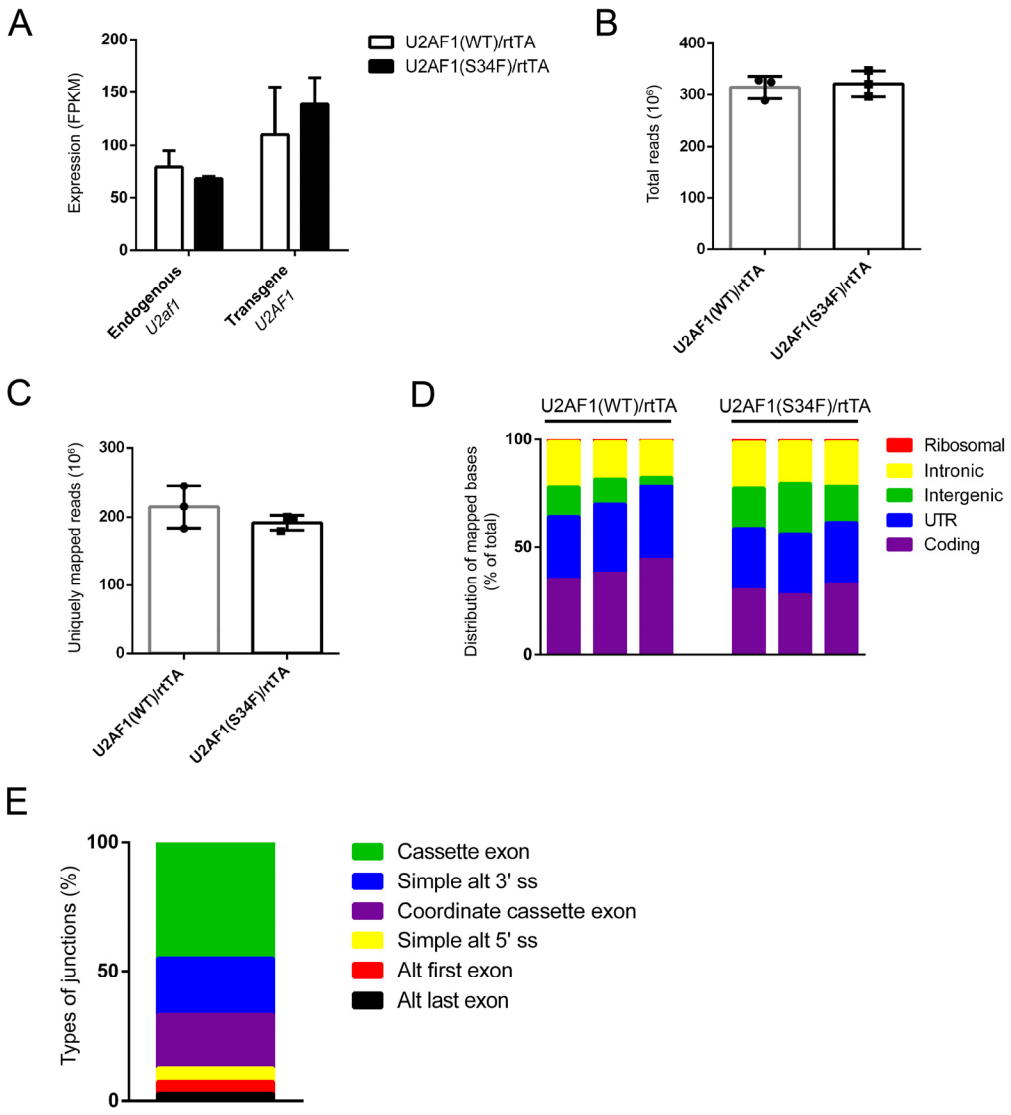


Figure S4, related to Figure 5. RNA-seq of common myeloid progenitors has consistent metrics across genotypes, shows comparative expression of endogenous *U2af1* and transgenic *U2AF1*, and reveals that exon skipping is the predominant mode of altered splicing. (A) FPKM values of *U2AF1* human transgenes [mapped to an artificial chromosome representing the U2AF1(WT)-pBS31' plasmid] and endogenous mouse *U2af1* in U2AF1(S34F)/rtTA and U2AF1(WT)/rtTA bone marrow CMPs. (B, C) Total reads (B) and

uniquely mapped reads (C) from CMP RNA-seq samples. (D) Distribution of mapped bases of U2AF1(S34F)/rtTA and U2AF1(WT)/rtTA CMP samples. Data plotted for each mouse individually. (E) Distribution of junctions identified as dysregulated in U2AF1(S34F)/rtTA samples compared to U2AF1(WT)/rtTA controls by indicated classifications (n=367). UTR, untranslated region; Alt, alternative; ss, splice site.

Table S2, related to Figure 5. Genes differentially expressed (DESeq FDR < 0.1) in common myeloid progenitor cells of U2AF1(S34F)/rtTA-recipient mice			
1110038B12Rik	*Ddx46*	*Ints8*	*Ptrf*
2210020M01Rik	*Dmc1*	*Itga1*	*Pvt1*
2610035D17Rik	*Dock1*	*Itgb2l*	*Rbm34*
4933403G14Rik	*Eda2r*	*Kctd14*	*Rnase6*
5330426P16Rik	*Eef2k*	*Klrb1f*	*Rnd3*
6530418L21Rik	*Eif5b*	*Klrk1*	*Rnf17*
AC027184.1	*Epha2*	*Lgals3*	*S100a4*
Aldh1a1	*Fahd2a*	*Lilrb3*	*Scarna3a*
Anapc2	*Fam113a*	*Lsp1*	*Shisa2*
Api5	*Fes*	*March1*	*Slamf8*
Arl11	*Fkbp1b*	*Mfsd7b*	*Slc16a1*
Batf3	*Gm10925*	*Mrpl9*	*Slc46a3*
BC094916	*Gm6377*	*Ms4a4c*	*Slc9a9*
Bccip	*Gpr44*	*Ms4a6c*	*Smc2*
Camsap2	*Grap*	*mt-Atp6*	*Snora24*
Cct8	*Gsn*	*mt-Atp8*	*Snord118*
Cd226	*Gstm5*	*mt-Cytb*	*Snrpf*
Cd300lg	*H2-Aa*	*mt-Nd1*	*Spice1*
Cd36	*H2-Ab1*	*mt-Nd2*	*Ssb*
Cd47	*H2-DMb1*	*mt-Nd4*	*Sulf2*
Cfh	*H2-Eb1*	*mt-Nd5*	*Susd1*
Ciita	*H2-T23*	*Naaa*	*Tlr3*
Clec9a	*Hck*	*Ncl*	*Tomm70a*
Clk3	*Hist1h2bb*	*Nudcd2*	*Trim21*
Crip1	*Hist1h4c*	*P2ry10*	*Trp53inp1*
Cst3	*Hnf4a*	*Phlda3*	*Txnl4a*
CT025673.1	*Hnrnpa2b1*	*Plbd1*	*Unc13d*
Cxcl16	*Hspd1*	*Plcxd2*	*Zadh2*
Cxcr4	*Id2*	*Plxdc1*	*Zbtb46*
Cyth4	*Ide*	*Ppa2*	*Zcwpw1*
D16H22S680E	*Ifi205*	*Prpsap2*	*Zfp365*
Ddah2	*Ighg1*	*Ptpn3*	*Zmat3*

Table S3, related to Figure 5. Pathways enriched (GOseq FDR < 0.1) in genes differentially expressed in common myeloid progenitor cells of U2AF1(S34F)/rtTA-recipient mice.

Category	Term/Pathway	FDR	Enrichment
Biological Process (GO:BP)			
GO:0033194	response to hydroperoxide	0.063	32.65
GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II	0.014	31.10
GO:0002495	antigen processing and presentation of peptide antigen via MHC class II	0.004	28.64
GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	0.004	25.92
GO:0002478	antigen processing and presentation of exogenous peptide antigen	0.041	21.77
GO:0019884	antigen processing and presentation of exogenous antigen	0.092	16.13
GO:0048002	antigen processing and presentation of peptide antigen	0.008	15.19
GO:0034341	response to interferon-gamma	0.063	11.83
GO:0019882	antigen processing and presentation	0.063	9.07
GO:0002699	positive regulation of immune effector process	0.036	8.37
GO:0002274	myeloid leukocyte activation	0.014	8.06
GO:0002697	regulation of immune effector process	0.041	5.83
GO:0045087	innate immune response	0.001	5.53
GO:0002252	immune effector process	0.036	4.11
GO:0006955	immune response	0.004	3.67
GO:0002684	positive regulation of immune system process	0.068	3.64
GO:0006952	defense response	0.004	3.53
GO:0045321	leukocyte activation	0.041	3.48
GO:0002376	immune system process	0.001	2.82
GO:0042221	response to chemical	0.080	2.08
GO:0006950	response to stress	0.048	2.04
KEGG Pathways (KEGG)			
5310	Asthma	0.001	29.03
5150	Staphylococcus aureus infection	0.000	22.52
4940	Type I diabetes mellitus	0.000	22.52
5330	Allograft rejection	0.001	20.16
5320	Autoimmune thyroid disease	0.001	19.44
5332	Graft-versus-host disease	0.001	19.44
4672	Intestinal immune network for IgA production	0.001	17.56
5416	Viral myocarditis	0.001	12.32
4612	Antigen processing and presentation	0.001	11.87
5140	Leishmaniasis	0.009	9.72
5322	Systemic lupus erythematosus	0.001	9.46
5323	Rheumatoid arthritis	0.012	9.07
4514	Cell adhesion molecules (CAMs)	0.001	8.86
190	Oxidative phosphorylation	0.003	7.38
4145	Phagosome	0.009	6.35
5012	Parkinson's disease	0.013	6.30

5145	Toxoplasmosis	0.072	5.55

Table S4, related to Figure 5. Junctions differentially spliced (DEXSeq FDR < 0.1) in common myeloid progenitor cells of U2AF1(S34F)/rtTA-recipient mice. Provided as an Excel file.

Table S5, related to Figure 6. Junctions differentially spliced (FDR < 0.1) as identified by Fisher's combined probability method in 3 datasets comparing mutant U2AF1 versus wild-type U2AF1: mouse CMP, human AML, and human CD34^{+} cells. Provided as an Excel file.

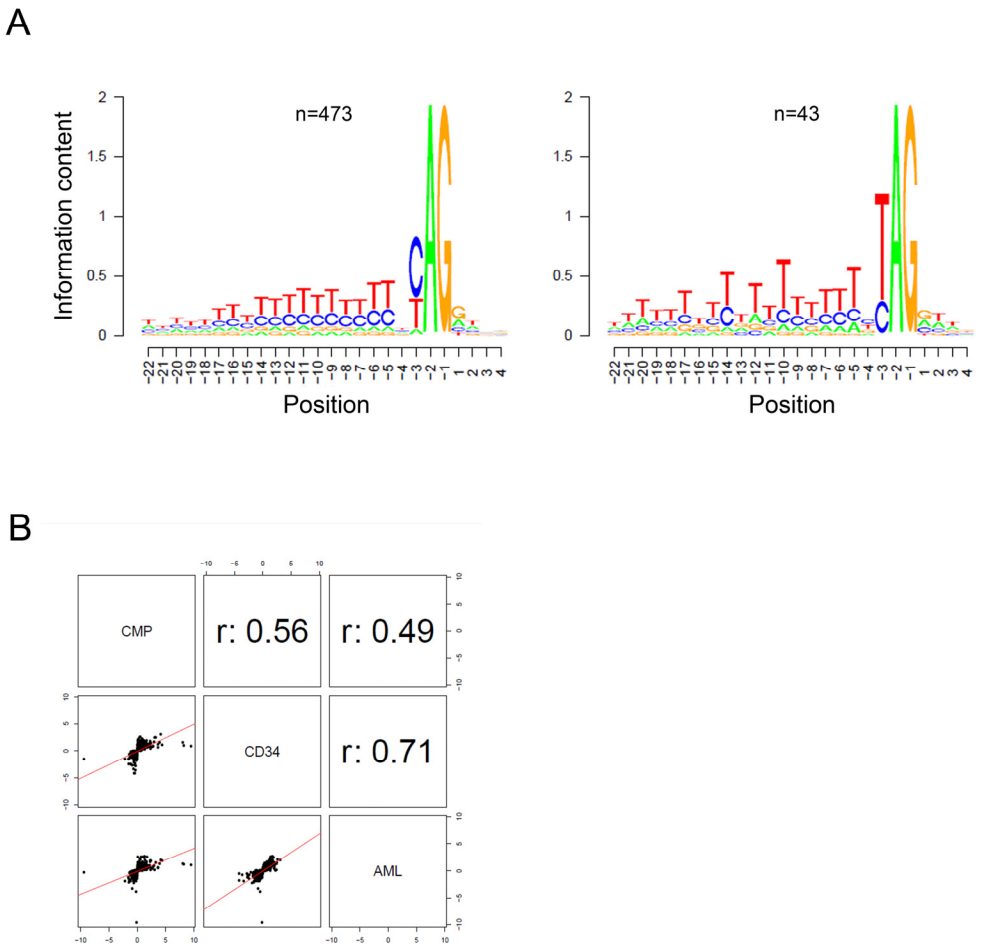


Figure S5, related to Figure 6. Meta-analysis identifies junctions with 3' splice site bias and correlated fold change across mouse CMP, human AML, and human CD34^{+} data sets. (A) Examination of control (left panel) or skipped exon events (right panel) identified by Fisher's combined probability method identifies an increased frequency of uracil (T) in the -3 position relative to the AG dinucleotide at 3' intronic splice sites. (B) Scatter plot of the junctions in the Fisher datasets (FDR<0.1) and the coefficient of correlation (r value).

Table S6, related to Figure 6. Pathways enriched (GOseq FDR < 0.1, Enrichment >2) in differentially spliced genes as identified by Fisher's combined probability method in 3 datasets comparing mutant U2AF1 versus wild-type U2AF1: mouse CMP, human AML, and human CD34$^+$ cells.			
Category	**Term/Pathway**	**FDR**	**Enrichment**
Biological Process (GO:BP)			
GO:0033119	negative regulation of RNA splicing	0.076	7.45
GO:0043484	regulation of RNA splicing	0.035	4.64
GO:0006415	translational termination	0.023	2.82
GO:0044033	multi-organism metabolic process	0.023	2.82
GO:0045047	protein targeting to ER	0.023	2.74
GO:0072599	establishment of protein localization to endoplasmic reticulum	0.023	2.74
GO:0019080	viral gene expression	0.024	2.73
GO:0070972	protein localization to endoplasmic reticulum	0.023	2.67
GO:0006414	translational elongation	0.025	2.63
GO:0006413	translational initiation	0.024	2.62
GO:0019058	viral life cycle	0.023	2.60
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	0.087	2.59
GO:0000398	mRNA splicing, via spliceosome	0.087	2.59
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	0.027	2.58
GO:0006613	cotranslational protein targeting to membrane	0.028	2.55
GO:0000956	nuclear-transcribed mRNA catabolic process	0.040	2.48
GO:0008380	RNA splicing	0.027	2.39
GO:0006402	mRNA catabolic process	0.087	2.34
GO:0006412	translation	0.000	2.33
GO:0006401	RNA catabolic process	0.076	2.32
GO:0006397	mRNA processing	0.028	2.31
GO:0016071	mRNA metabolic process	0.000	2.25
GO:0006396	RNA processing	0.006	2.10
Molecular Function (GO:MF)			
GO:0044822	poly(A) RNA binding	0.000	2.43
GO:0003723	RNA binding	0.000	2.22
GO:0003735	structural constituent of ribosome	0.014	2.17
KEGG Pathways (KEGG)			
3040	Spliceosome	0.012	3.08
3010	Ribosome	0.008	2.69

Table S7, related to Table 1. Recurrently mutated genes (RMG) in MDS and AML. Provided as an Excel file.

II. SUPPLEMENTAL EXPERIMENTAL PROCEDURES

A. Mouse genotyping

Mouse genotypes were determined from genomic tail DNA prepared using NaOH extraction (HotSHOT method) (Truett et al., 2000), and PCR was performed with the following primers (Stratman et al., 2003). For *M2rtTA*, we used the forward primer 5'-CAATCGAGATGCTGGACAGGCATCATAC-3' and the reverse primer 5'-TCTTTTGCTACTTGATGCTCCTGTTCCTCC-3', with a product size of 301 bp. For *U2AF1*(WT) or *U2AF1*(S34F) transgene, we used the forward 5'-TCAGTATGAGATGGGAGAATGCACACGAG-3' and the reverse primer 5'-TTTGCCCCCTCCATATAACATGAATTTTACAA-3', with a product size of 357 bp.

B. Pyrosequencing

To detect both the exogenous human transgene *U2AF1* and the endogenous mouse *U2af1* transcripts expressed in transgenic cells, cDNA was prepared (as described in the Experimental Procedures) followed by PCR using primers that would amplify both species of transcript (forward 5'-TTAGCCAGACCATTGCCCTCTTG-3', reverse 5'-CGGCTGTCCATTAAACCAACGGT-3'; 293 bp). Then to detect the relative ratio of human and mouse cDNA in transgenic mouse bone marrow cells, pyrosequencing for single nucleotide polymorphisms that differ between mouse and human cDNAs was performed on the above-generated PCR products with pyrosequencing primers (5'-CTGGTGGGGAACGTGTA-3', 5'-CGGTTTGCGCTGTGC-3').

C. Flow cytometry

Flow cytometry was performed on bone marrow, peripheral blood, and spleen to determine the cell lineage distribution of donor-derived mature cells using the following cell surface receptors (all antibodies were obtained from eBioscience unless otherwise indicated, clone ID provided if available): CD45.1-eFluor450 (A20), CD45.2-FITC (104, BD Biosciences), CD115-PE (AF8980), Gr-1-APCeFluor780 (RB6-8C5), B220-PerCPCy5.5 (RA3-6B2), and

CD3e-APC (145-2C11). Incubations were performed in FACS buffer following pre-incubation with Fcγ Receptor 16/32 block (93).

Flow cytometry for donor-derived hematopoietic progenitor and stem cells was performed on bone marrow and spleen using the following cell surface markers (all antibodies are from eBioscience, unless indicated, clone ID provided if available): CD45.1-APC (A20), CD45.1-eFluor450 (A20), CD45.2-PE (104), CD45.2-AlexaFluor700 (104), Biotin-conjugated lineage [Gr-1 (RB6-8C5), Cd3e (145-2C11), B220 (RA3-6B2), Ter119 (TER-119), and CD41(eBioMWReg30)], streptavidin secondary-eFluor605NC, c-Kit-APCeFluor780 (ACK2), Sca-1-PerCP-Cy5.5 (D7), CD34-FITC (RAM34), Fcγ-eFluor450 (93), CD150-PE (TC15-12F12.2, Biolegend), Cd48-PE-Cy7 (HM48-1), Flk2-APC (A2F10). All antibodies were incubated in FACS buffer.

D. Quantitative RT-PCR of *U2AF1* transgene and endogenous *U2af1*

To examine the effect of exogenous U2AF1(WT) and U2AF1(S34F) expression on endogenous mouse *U2af1* expression levels, bulk bone marrow was obtained from transgenic mice without doxycycline treatment or following 5 days of transgene induction with doxycycline in vivo. To examine the expression levels of the *U2AF1* transgene in the different hematopoietic lineages of mouse bone marrow cells, donor-derived cells were flow cytometry-sorted from 4-5 recipient mice transplanted with transgenic donor marrow (pooled for each n value) following 5 days of transgene induction with doxycycline. RNA was prepared from cells using Trizol (Ambion) following manufacturer's recommendation. DNA was removed using Turbo DNA-free kit (Ambion), and cDNA was prepared using the Superscript III kit (Invitrogen). Quantitative-RT-PCR was performed for the *U2AF1* transgene (Forward primer 5'-AAAGATCTGGGCGATTCTGA-3', Reverse primer 5'-TATTTGTGAGCCAGGGCATT-3'), endogenous mouse *U2af1* (Forward primer 5'-ACCGGGAGAGATCTGGAC-3', Reverse primer 5'-GAACTGGCTTGTCAACCAAC-3'), and mouse *Gapdh* (Forward primer 5'-TGCACCACCAACTGCTTAG-3', Reverse primer 5'-GGATGCAGGGATGATGTTC-3') using

Sybergreen (Applied Biosystems). Individual cDNA samples were normalized according to their levels of *Gapdh*.

E. Western blot analysis of U2AF1 transgene induction

Soluble whole cell protein lysate was prepared from mouse bone marrow cells using lysis buffer (50 nM HEPES, pH 7.5, 150 mM NaCl, 1 mM EDTA, 10% Glycerol, 0.1% Tween-20 plus protease inhibitors) and sonication, and 35 μg of total protein per sample was resolved via SDS-PAGE electrophoresis. Protein was transferred to PVDF membrane (Millipore), and immunoblotting with anti-U2AF1 (Abcam, ab86305) and anti-β-Actin (Sigma, AC-15) antibodies was performed per the manufacturer's recommendation. Primary antibody detection was performed using horseradish peroxidase-conjugated secondary antibodies (Cell Signaling) per the manufacturer's recommendation. Immunoreactive bands were detected via SuperSignal West Pico (or Femto) Chemiluminescent substrate (ThermoScientific) and imaged using myECL Imager (ThermoScientific). Densitometry was performed using ImageJ.

F. BrdU Incorporation in KLS cells

Transplanted mice were given doxycycline for 5 days to induce transgene expression. Twenty-four and 12 hours prior to sacrifice, mice were given an I.P. injection of BrdU (1 mg/mouse). Bone marrow was harvested and donor-derived KLS (lin^-, c-Kit^+, Sca-1^+) cells were stained for flow cytometry as described in the Experimental Procedures. Following KLS staining, bone marrow cells were processed per manufacturer's recommendation for the FITC BrdU Flow Kit (BD Biosciences). Flow cytometry data were analyzed with FlowJo software.

G. Primer design for RT-PCR of MDS patient samples

PCR primers used for RT-PCR and gel electrophoresis of MDS patient bone marrow samples were designed to span the splice junction such that both the canonical and alternatively spliced isoforms were amplified and resolved. Primer sequences and expected PCR product sizes for each gene are described below. For *H2AFY*, we used the forward primer 5'-CAGTCCTCTCCACCAAGAGCC-3' and the reverse primer 5'-

CACCTTTCTTCTCCAGCGTGTT-3', with product sizes 153 bp and 162 bp. For *BCOR*, we used the forward primer 5'- GACAGCAGCCACACTGAGAC-3' and the reverse primer 5'-TCTTCCGACCAGCTTCTGTT-3', with product sizes 296 bp and 398 bp. For *GNAS*, we used the forward primer 5'- TGCAGAAGGACAAGCAGGTCT-3' and the reverse primer 5'-TGATGTCCTGCACTTTGGTTG-3', with product sizes 153 bp and 201 bp. For *PICALM*, we used the forward primer 5'- GTTGACTTTGAATCTGTGTTTGGAA-3' and the reverse primer 5'-TTCTGAGAGGCCACTGTTGG-3', with product sizes 165 bp and 305 bp. For *KDM6A*, we used the forward primer 5'- ATCAGCCCATGGATGCTTTA-3' and the reverse primer 5'-CCACCACTCCAATTGTCAGA-3', with product sizes 226 bp, 361 bp, 382 bp. For *KMT2D*, we used the forward primer 5'- GTGCCCGATCAGAGCCTAA-3' and the reverse primer 5'-GTGCCCGATCAGAGCCTAA-3', with product sizes 65 bp and 113 bp. For *MED24*, we used the forward primer 5'- CACGGCAAAGCAGAGGAATG-3' and the reverse primer 5'-AATGCTCGATGGCAGTCCAA-3', with product sizes 253 bp and 310 bp.

H. RNA-seq alignment and preprocessing

For the mouse CMP data set, RNA-seq analysis was performed with The Genome Institute's Genome Modeling System within the model group '65d63496836e40e79660e078a80106cf' using the RNA-seq processing-profile '734554c622244c0993999d8b7a08a12f.' Quality of raw RNA sequence data in the form of FastQ data files was assessed using FastQC version 0.10.0 (http://www.bioinformatics.babraham.ac. uk/projects/fastqc/). Paired 2x100bp sequence reads from the stranded cDNA library were aligned to the mouse genome reference sequence (mm9/NCBI build 37). Initial segmented alignments were performed using bowtie version 2.1.0 (Langmead and Salzberg, 2012) followed by spliced alignments with TopHat version 2.0.8 (Kim et al., 2013). TopHat was supplied transcript models in GTF format using the '-g' parameter that represented known and predicted mouse transcripts obtained from Ensembl version 67 (Flicek et al., 2013) and was passed the parameters '--library-type fr-firststrand' indicating the

strandedness of the library. Binary sequence alignment (BAM) files obtained by alignment of RNA-seq reads with TopHat were summarized by use of SAMStat version 1.08 and SAMtools version 0.1.18/r982 (specifically, the idxstats and flagstat utilities) (Li et al., 2009). The quality of alignments was assessed using Picard version 1.85 (specifically, the RnaSeqMetrics utility) (http://picard.sourceforge.net/). Following alignment, expression estimates in the form of Fragments per Kilobase of exon per Million bases mapped (FPKM) were calculated by Cufflinks version 2.1.1 (Trapnell et al., 2010) using default parameters except for '--num-threads 4 --max-bundle-length 10000000 --max-bundle-frags 10000000.' Transcript models were supplied to Cufflinks using the '-g' option and the same GTF described above. Transcripts corresponding to mitochondrial and ribosomal genes, as indicated by Ensembl gene biotypes, were masked during calculation of transcript expression estimates. Exon-exon junction counts were obtained by parsing the 'junctions.bed' file produced by TopHat. This file reports the coordinates of all introns observed by splice-aware alignment of reads to the genome and the number of reads supporting each. Each observed exon-exon junction was annotated with the genomic coordinates of its defining acceptor and donor sites and cross-referenced against the mouse transcripts in Ensembl version 67 to determine the number, if any, of exons skipped. Each junction was then assigned to one of five classes describing its relationship to known Ensembl transcripts: 'DA', 'NDA', 'D', 'A', and 'N'. 'DA' junctions are those where the exon-exon combination corresponds to one or more known Ensembl transcripts. 'NDA' refers to a novel connection of donor and acceptor pair that is individually used in Ensembl transcripts but has not been observed in that combination. 'D' refers to junctions that use a known donor site but a novel acceptor site. 'A' refers to junctions that use a known acceptor site but a novel donor site. 'N' refers to junctions that do not correspond to any Ensembl transcript at donor or acceptor sites. Gene-level read counts were obtained by excluding unmapped reads or secondary read alignments from the BAM files (using samtools view –F 0x0100) and processing the resulting

reads using HTseq (Anders et al., 2014) (version 0.5.4p1; with parameters: --mode intersection-strict --minaqual 1 --stranded reverse --type exon --idattr gene_id).

Alignment of human (AML and CD34) data sets were performed similarly to the above, with only minor changes noted below. Alignment and preprocessing of the CD34 data set has been described previously (Okeyo-Owuor et al., 2014). The CD34 data set in model group '70069' used RNA-seq processing profile '2819506,' while the AML data set in model group '74964' used RNA-seq processing profile '2836606.' Reads were trimmed to remove 'SPIA' adapters (ligated during cDNA synthesis) using the read trimmer 'flexbar' version 229 (https://wiki.gacrc.uga.edu/wiki/FAR) with the following parameters set: '--adapter CTTTGTGTTTGA --adapter-trim-end LEFT --nono-length-dist --threads 4 --adapter-min-overlap 7 --max-uncalled 150 --min-readlength 25.' After trimming, reads were aligned to a modified version of the human genome reference sequence (NCBI build 37 lite) with alternative haplotype sequences omitted. Transcript models supplied to TopHat represented human transcripts from Ensembl version 67. In neither case were the human cDNA libraries stranded, hence TopHat was not passed a library type parameter as above. SAMtools version 0.1.16/r963 was used to summarize BAM files. FPKMs were calculated using Cufflinks version 2.0.2 and parameters '--num-threads 4 --max-bundle-length 10000000.' BAM files were filtered using samtools view –F 0x0104 prior to being passed to HTseq. As these libraries were not stranded, HTseq was passed the parameters --mode intersection-strict --minaqual 1 --stranded no --type exon --idattr gene_id.

I. Junction- and gene-level differential expression analysis

Differential junction- and gene-level expression across WT and S34F mutant U2AF1 samples was inferred using DEXSeq version 1.21.1 (Anders et al., 2012) and DESeq2 version 1.6.1 (Anders and Huber, 2010), respectively, in R. Both use negative binomial generalized linear models (GLMs) to infer statistically significant differences between condition-specific (here, WT and S34F) bin counts (here, corresponding to total reads mapped to a junction or to a gene).

The CMP and AML data sets were derived from unpaired experimental designs. The CMP data set is comprised of three WT U2AF1 and three S34F U2AF1 samples. We analyzed a subset of the AML data set (Cancer Genome Atlas Research, 2013), comprised of six samples harboring U2AF1 S34(F/Y) mutations and 102 control samples that did not harbor copy number or point mutations in *U2AF1* or in any of 222 other spliceosome-associated genes (Table S1). Samples harboring U2AF1 (S34F) mutations have UPNs 245450, 445045, 569053, and 633734 and those harboring U2AF1 (S34Y) mutations have UPNs 400830 and 957664. UPNs 274429 [U2AF1(Q157P)], 415172 [SF3B1(K700E]. Control samples have UPNs: 203, 237, 269, 295, 104851, 113971, 123804, 141273, 142074, 146218, 150288, 150951, 156704, 180168, 186481, 195182, 208027, 225373, 237983, 242129, 254137, 255421, 258135, 273919, 275786, 290344, 296361, 303642, 311636, 319955, 322110, 326772, 327733, 329614, 332131, 345701, 346190, 348685, 375182, 399253, 407992, 410324, 412761, 418499, 427366, 431799, 433325, 452198, 456892, 507696, 509754, 545259, 548327, 553863, 558395, 573988, 578179, 593890, 594368, 595704, 605322, 606061, 617776, 619751, 632729, 635258, 670224, 671473, 690397, 702808, 717456, 723101, 737451, 740266, 766126, 767969, 775109, 804168, 807615, 808642, 816067, 817156, 826984, 831711, 849660, 851929, 861663, 862507, 866660, 868231, 869586, 907786, 914247, 923966, 933124, 974749, 984036, 987523, 989176, 991612, 992966, 997292. This approach and a similar edgeR-based (Robinson et al., 2010) analysis of these data have been described previously (Okeyo-Owuor et al., 2014). Similarly, we have previously (Okeyo-Owuor et al., 2014) described an edgeR-based analysis of the CD34 data set, which was derived from a paired experimental design. This data set is comprised of three WT U2AF1 samples, each of which is paired with one of three S34F U2AF1 samples. Both designs are accommodated by GLMs.

DEXSeq was applied to junctions belonging to genes having at least two exons (as annotated in Ensembl version 67) and which met a minimum read threshold. This threshold required more than five junction reads in at least half of the samples (i.e., three for CD34 and 54

for AML). Given the greater expected heterogeneity of the CMP data, we loosened this requirement and required the minimum number of reads in at least two of the samples. ‘DA’, ‘NDA’, ‘D’, and ‘A’ junctions were analyzed together, with ‘N’ junctions (i.e., in which neither the 3’ nor the 5’ splice site was annotated in Ensembl) excluded from analysis. Both DEXSeq and DESeq2 compare a full model to a reduced model. For the unpaired designs (CMP and AML), we used DEXSeq to detect potential interaction between the ‘condition’ (i.e., WT or MT) and the expression ‘expr’ of a junction in a sample ‘sample’ using a full model that describes (the log of) ‘expr’ as a sum of terms of the (log of the) baseline gene-level expression (across all samples), (the log of) the fraction of reads mapped to the junction, (the log of) the fold change in the overall gene expression in ‘sample,’ and an interaction term between ‘condition’ and the junction. Hence, DEXSeq effectively detects alternative splicing (dysregulated junctions) after controlling for overall changes in gene expression between conditions. DEXSeq implements this by augmenting a data matrix whose rows are junctions and whose columns are samples with additional columns (one per sample) holding the sum of junction reads in that sample across all ‘other’ junctions in the gene. Then, the above full model [Eqn 6 of (Anders et al., 2012)] is fit to each row of the augmented matrix using the formula expr ~ sample + exon + condition:exon. Here, we retain the ‘exon’ terminology to refer to junctions, which may be ‘this’ to describe the original junction counts or ‘other’ to refer to the augmented junction counts within the gene. This is compared to the reduced model expr ~ sample + exon [Eqn 5 of (Anders et al., 2012)]. The DEXSeq manual (http://www.bioconductor.org/packages/release/bioc/vignettes/DEXSeq/inst/doc/DEXSeq.pdf) provides additional detail. For the paired experimental design (i.e., the CD34 data set), we added an additional interaction term between the biological replicate ‘replicate’ and the junction, following the general method of incorporating an additional covariate in Eqn 7 of (Anders et al., 2012). Hence, the full model formula is expr ~ sample + exon + replicate:exon + condition:exon and the reduced formula is expr ~ sample + exon + replicate:exon. Having defined the formulae,

analysis proceeded by invoking estimateSizeFactors, estimateDispersions with formula set to the full model formula, testForDEU with the full and reduced models specified, and DEXSeqResults. The log fold change was obtained by calling results with parameters name = 'exonthis.conditionMT' and independentFiltering = FALSE, rather than estimateExonFoldChanges.

DESeq2 was applied to genes having more than 20 reads in the above minimum number of samples. For unpaired experiments (i.e., CMP and AML), we directly tested for a condition-specific effect on gene expression by instantiating DESeqDataSetFromMatrix with design = ~condition. For the paired experiment (i.e., CD34), we introduced an additional term to absorb replicate-level effects and hence specified design = ~replicate + condition. Analysis was then performed by invoking DESeq and results were obtained by invoking the results function with parameter independentFiltering=FALSE.

J. Principal component analysis

We performed principal component analysis (PCA) of the alternative 3' splicing ratios involving junctions skipping a single exon in the CMP and AML data sets. Such events include the skipping of cassette exons as well as the mutually exclusive skipping of a single exon. We define the alternative 3' splicing ratio (3' SR) for a junction in a sample as the ratio of reads involving that junction to the sum of reads of junctions sharing its 5' splice site. We analogously define the alternative 5' splicing ratio (5' SR). Given the expected involvement of U2AF1 in 3' splice site selection, we anticipate that most direct, robust differences induced by its mutation will involve alternative 3' splice sites. Hence, we focus on PCA of 3' SR and refer to it simply as SR in the main text. Similar results were obtained with PCA of 5' SR (data not shown). SR is similar to percent spliced in (i.e., PSI or Ψ), used to infer dysregulated splicing in methods such as MISO (Katz et al., 2010) and rMATS (Shen et al., 2012; Shen et al., 2014). Some definitions of PSI extend the straightforward notion of SR to reflect, for example, mapping biases due to isoform length (Katz et al., 2010). We avoid these complications (and the assumptions inherent

in inferring transcripts from short sequencing reads) in restricting our principal component (rather than differential) analyses to SR.

SRs were analyzed if the corresponding junction had more than five reads in all samples. A junction's SRs were standardized by centering by the mean of the junction's SRs across samples and dividing by the standard deviation of that junction's SRs across samples. PCA was performed on the standardized SRs using prcomp with default parameters in R. Additional scaling was performed by ggbiplot with default parameters and the resulting, scaled rotated data (returned in the x component from prcomp) was plotted using plot3d.

K. Meta-analysis

To integrate results from the three datasets, we use Fisher's method (Brown, 1975)

$$x_i^{fisher} = -2 \sum_{j=\{CMP,CD34,AML\}} \ln p_i^j$$

to combine p_i^{CMP}, p_i^{CD34}, and p_i^{AML}, the p values for junction (or gene) *i* across the CMP, CD34, and AML data sets, respectively, into a single statistic x_i^{fisher}. A p value may be computed from x_i^{fisher} since it follows a chi-squared distribution with 2 *k* degree of freedom (with *k* = 3, the number of combined p values). Fisher's method will be anti-conservative unless the p values are independent under the null hypothesis (of no dysregulation). Since less than 1% of potential junctions are inferred to be dysregulated, we can assess independence across all junctions without concern that the few violating the null hypothesis will perturb the results. We did so by creating scatterplots and calculating correlations of p values across all pairwise combinations of data sets. We saw no evidence for correlation, with correlation values and slopes of best-fit linear models close to zero and with no apparent visual correlation in the scatterplots.

L. Consensus sequence analysis

Consensus 3' splice site sequences corresponding to skipped exons, alternative 3' splice sites, and control junctions were computed via seqLogo (version 1.32.1). Junctions resulting from a skipped exon ("skipped exon junctions") were defined as those with a single

skipped exon (as annotated in Ensembl 67) and inferred to be differentially expressed with an FDR < 0.1. As in the PCA analysis above, this includes all cassette exon skipping events and some mutually exclusive exon skipping events. The 3' splice site of the skipped exon itself was determined as that belonging to a junction that was not identical to the skipped exon junction, but which shared its 5' splice site and involved no skipped exons. Junctions participating in alternative 3' splice site selection were defined as those (1) having an FDR < 0.1; (2) for which the same 5' splice site was paired with exactly two expressed (i.e., meeting the above minimum read and sample number requirements) 3' splice sites; and (3) for which neither of the junctions involved the skipping of any exon. We refer to the differentially expressed splice site as the alternative 3' splice site and the other of the two splice sites in the pair as the canonical 3' splice site. By restricting to these simple cases (of only two possible alternatives), we could be confident that the alternative splice site was preferentially used relative to the single canonical splice site; this direct contrast would not have been possible with multiple alternative splice sites, all of which may be in competition with each other as well as with the canonical splice site. These splice sites were extended to define the flanking sequences of skipped exons, alternative 3' splice sites, and canonical 3' splice sites. Skipped exon, canonical 3', and alternative 3' splice sites were analyzed separately according to whether they occurred more [$\log_2$(fold change) > 0] or less [$\log_2$(fold change) < 0] in the S34F sample relative to WT. Control flanking regions were defined as those corresponding to junctions that showed no evidence for differential expression—i.e., had a | $\log_2$(fold change) | < 0.001. The human and mouse sequences of these flanking regions were downloaded using getSeq from the R library BSgenome (version 1.34.0), passing it genome identifiers from the subsidiary R libraries BSgenome.Hsapiens.UCSC.hg19 or BSgenome.Mmusculus.UCSC.mm9, respectively. A position frequency matrix (over the entire amino acid and nucleotide alphabet) was created using consensusMastrix. This was transformed to a position weight matrix via makePWM and passed to seqLogo.

M. Gene enrichment and pathway analysis

GOseq (version 1.18.0)(Young et al., 2010) analysis was performed to determine enrichment of genes with dysregulated junctions (DEXSeq FDR < 0.1) within KEGG pathways (Kanehisa and Goto, 2000), Gene Ontology (Ashburner et al., 2000) biological process (BP) and molecular function (MF) terms, oncogenes and tumor suppressor genes annotated in the Cancer Gene Census (Futreal et al., 2004), spliceosome genes, and genes recurrently mutated in AML and MDS. Enrichment for dysregulated genes (DESeq FDR < 0.1) was performed similarly.

The Cancer Gene Census was downloaded on Nov 17, 2014 (http://cancer.sanger.ac.uk/cancergenome/projects/census/). Genes were characterized as oncogenes if they were unambiguously designated as having a dominant molecular genetic function (Dom) in the Cancer Gene Census and were characterized as tumor suppressor genes if they were unambiguously designated as recessive (Rec). As not all entries were associated with Ensembl ids, gene symbols were translated to Ensembl ids using a translation table downloaded from Ensembl. This was obtained by clicking on the Biomart link at http://may2012.archive.ensembl.org/index.html, choosing Ensembl Genes 67 under database, choosing Homo sapiens genes (GRCh37.p7) under dataset, and selecting 'Ensembl Gene ID' and 'Associated Gene Name' exposed by progressively selecting 'Attributes,' then 'Features,' then 'Gene.' The locally curated list of putative spliceosome-associated genes includes 246 unique gene symbols, with a subset having translations to 223 unique Ensembl ids (Table S1). The locally curated list of genes recurrently mutated in AML or MDS contains 284 unique gene symbols, all of which have translations to unique Ensembl ids (Table S7).

Each GOseq analysis first created a probability weighting function for genes in the human or mouse genome, as appropriate, using nullp. GO BP, GO MF, and KEGG analyses were performed by passing the resulting probability weighting function to goseq and specifying test.cats = "GO:BP", "GO:MF", or "KEGG," respectively. p values calculated by GOseq were

corrected for multiple testing using the method of Benjamini and Hochberg (Benjamini and Hochberg, 1995) as implemented in the R function p.adjust. Enrichments were calculated as the ratio of two proportions: the proportion of the number of dysregulated genes in the category (e.g., GO term or pathway) relative to the total number of dysregulated genes and the proportion of the total number of genes in the category relative to the total number of assayed genes (i.e., genes tested for dysregulated junctions via DEXSeq or for overall expression differences via DESeq). The Ensembl ids associated with each GO term were determined by reverse mapping (via goseq:::reversemapping) the list of categories associated with genes returned by getgo. A similar reverse mapping was obtained from KEGG pathways to Ensembl ids by passing the parameter fetch.cats = “KEGG” to getgo. Finally, Ensembl ids were translated to gene symbols using biomart: A biomart object was obtained by calling useMart with parameter biomart = “ensembl” and dataset = “mmusculus_gene_ensembl” or dataset = “hsapiens_gene_ensembl,” as appropriate. Ensembl ids were then translated to gene symbols using getBM with parameters attributes = c(“ensemble_gene_id”, “external_gene_name”), filters = “ensemble_gene_id”, and the biomart object returned by useMart. The curated lists for oncogenes, tumor suppressors, genes recurrently mutated in AML or MDS, and spliceosome genes were tested by specifying the gene2cat parameter of goseq as a list with gene-named entries with value indicating whether or not the respective gene belongs to the curated list.

GOseq infers and corrects for biases in the enriched genes. By default, it attempts to infer a bias associating gene length with dysregulation. However, plots (output by nullp) of gene length versus splicing dysregulation ratio (i.e., the fraction of genes at a given length that had dysregulated junctions) showed no apparent correlation, so that the bias factor was similar across all genes. We therefore tested several other biases of potential correlation in the number and/or expression level (in counts) of junctions tested for dysregulation in a gene. Specifically, we used the sum over tested junctions within a gene of the geometric mean of the respective junction’s counts (Ilagan et al., 2014), the sum over tested junctions within a gene of the

arithmetic mean of the respective junction's counts, and the number of tested junctions within a gene. Each of these was specified using the bias.data of nullp. We found the best correlation between bias and dysregulation ratio using the number of tested junctions bias, whose results we report in the main text. GOseq pathway analysis of gene-level differential expression continued to use the default gene-length bias.

N. Intersection of junctions across data sets

We calculated a p value to assess the significance of observing 17 dysregulated junctions in common and with concordant direction of log fold change across all three data sets—CMP, AML, and CD34—using simulation. 219, 162, and 1,652 junctions were dysregulated (DEXseq FDR < 0.1) independently across the CMP, AML, and CD34 data sets, respectively, and additionally were amongst the 99,178 junctions tested (and, hence, necessarily homologous) across all three data sets. Therefore, for each of n=1,000 iterations and from the set of 99,178 homologous junctions, we randomly sampled 219, 162, and 1,652 junctions and corresponding FDR values from the CMP, AML, and CD34 data sets, respectively. The p value was calculated as the proportion out of n=1,000 iterations the number of iterations having at least 17 concordantly dysregulated (DEXseq FDR < 0.1) across the three data sets.

O. Endogenous *U2af1* and transgenic *U2AF1* expression analysis by RNA-seq

We compared endogenous *U2af1* expression to transgenic (human) *U2AF1* expression through their respective FPKMs. Given the sequence differences between the human and mouse genes, the human transgenes in our RNA-seq experiments did not map to the mouse genome (as described in **RNA-seq alignment and preprocessing** above). Therefore, we modified the above alignment and expression quantitation steps. We created an augmented mouse mm9 genome having an artificial chromosome holding the sequence from the U2AF1(WT)-pBS31' plasmid. The resulting FASTA was passed to TopHat. Since the plasmid encodes the cDNA (i.e., an uninterrupted transcript) for U2AF1, we did not pass TopHat an augmented transcriptome via the '-g' parameter, but continued to supply it with the Ensembl

transcript model as described above. We did, however, partition the plasmid/artificial chromosome into separate regions for expression quantitation based on visually contiguous, non-zero expression (as viewed in IGV) across at least one of the six samples. This led us to define a region from 1132 to 1854 corresponding to human *U2AF1*. We used Cufflinks 2.0.2 to quantitate their expression as described above, except with a GTF file augmented from above to include these new regions. We then used the FPKM values calculated for the *U2AF1* region of this artificial chromosome and for the *U2af1* gene as the expression levels of the human transgenic *U2AF1* and mouse endogenous *U2af1*, respectively.

III. SUPPLEMENTAL REFERENCES

Anders, S., Pyl, P. T., and Huber, W. (2014). HTSeq--A Python Framework to Work with High-throughput Sequencing Data. Bioinformatics. DOI: 10.1093/bioinformatics/btu638

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics *25*, 25-29.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met *57*, 289-300.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. Nature Reviews Cancer *4*, 177-183.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research *28*, 27-30.

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods *7*, 1009-1015.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth *9*, 357-359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Shen, S., Park, J. W., Huang, J., Dittmar, K. A., Lu, Z. X., Zhou, Q., Carstens, R. P., and Xing, Y. (2012). MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. Nucleic Acids Research *40*, e61.

Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., and Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences of the United States of America. DOI: 10.1073/pnas.1419161111

Stratman, J. L., Barnes, W. M., and Simon, T. C. (2003). Universal PCR genotyping assay that achieves single copy sensitivity with any primer pair. Transgenic Research *12*, 521-522.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotech *28*, 511-515.

Truett, G. E., Heeger, P., Mynatt, R. L., Truett, A. A., Walker, J. A., and Warman, M. L. (2000). Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT). BioTechniques *29*, 52, 54.