An important class of intron retention events in human erythroblasts is regulated by cryptic exons proposed to function as splicing decoys

MARILYN PARRA,[1] BEN W. BOOTH,[1] RICHARD WEISZMANN,[1] BRIAN YEE,[2] GENE W. YEO,[2,3] JAMES B. BROWN,[4] SUSAN E. CELNIKER,[1] and JOHN G. CONBOY[1]

[1]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
[2]Department of Cellular and Molecular Medicine, University of California at San Diego, La Jolla, California 92037, USA
[3]Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597
[4]Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

ABSTRACT

During terminal erythropoiesis, the splicing machinery in differentiating erythroblasts executes a robust intron retention (IR) program that impacts expression of hundreds of genes. We studied IR mechanisms in the *SF3B1* splicing factor gene, which expresses ~50% of its transcripts in late erythroblasts as a nuclear isoform that retains intron 4. RNA-seq analysis of nonsense-mediated decay (NMD)-inhibited cells revealed previously undescribed splice junctions, rare or not detected in normal cells, that connect constitutive exons 4 and 5 to highly conserved cryptic cassette exons within the intron. Minigene splicing reporter assays showed that these cassettes promote IR. Genome-wide analysis of splice junction reads demonstrated that cryptic noncoding cassettes are much more common in large (>1 kb) retained introns than they are in small retained introns or in nonretained introns. Functional assays showed that heterologous cassettes can promote retention of intron 4 in the *SF3B1* splicing reporter. Although many of these cryptic exons were spliced inefficiently, they exhibited substantial binding of U2AF1 and U2AF2 adjacent to their splice acceptor sites. We propose that these exons function as decoys that engage the intron-terminal splice sites, thereby blocking cross-intron interactions required for excision. Developmental regulation of decoy function underlies a major component of the erythroblast IR program.

Keywords: SF3B1; alternative splicing; intron retention

INTRODUCTION

Intron retention (IR) is a common variant of alternative splicing in which selected intron(s) are specifically retained in an otherwise spliced and polyadenylated transcript. IR can be developmentally or physiologically regulated as an important component of gene regulation in many cell types (Jacob and Smith 2017). IR transcripts can be stored in the nucleus to be spliced in response to appropriate signals, thus serving as a source of new mRNA (Ninomiya et al. 2011; Boothby et al. 2013; Mauger et al. 2016), or they can represent dead-end RNAs that are degraded in the nucleus (Pendleton et al. 2018). Other IR transcripts are transported to the cytoplasm where they are degraded by nonsense-mediated decay (Wong et al. 2013). Partial diversion of transcriptional output into IR isoforms also functions as a post-transcriptional pathway to modulate expression levels (Wong et al. 2013; Braunschweig et al. 2014; Shalgi et al. 2014; Boutz et al. 2015; Ni et al. 2016). Some IR transcripts can be recruited to ribosomes to function in translation (Li et al. 2016), while others might serve as miRNA sponges in the nucleus (Schmitz et al. 2017).

The diversity of IR programs and functions suggests that a number of regulatory pathways exist, allowing cells to integrate multiple inputs in order to independently regulate different subsets of IR transcripts. Consistent with this notion, coherent subsets of genes can be regulated by IR, especially those encoding RNA binding proteins and spliceosomal factors (Wong et al. 2013; Braunschweig et al. 2014; Shalgi et al. 2014; Boutz et al. 2015; Pimentel et al. 2016). How these subprograms are regulated is not well understood. Recent studies have implicated transcription rate/pausing, specific splicing factors (SRSF4 and HNRNPLL), and DNA and protein

Corresponding author: jgconboy@lbl.gov
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.066951.118.

methylation factors (by MECP2, in granulocytes, and by PRMT5, in glioblastoma) as important effectors of IR (Braunschweig et al. 2014; Cho et al. 2014; Boutz et al. 2015; Braun et al. 2017; Wong et al. 2017). In the brain, signaling-dependent splicing was observed to require NMDA-type glutamate receptor or calmodulin-dependent protein kinase pathways for removal of retained introns upon neuronal activation (Mauger et al. 2016). There are also a few recent examples showing that individual IR events can be regulated by feedback mechanisms to control physiological pathways (Bergeron et al. 2015; Park et al. 2017; Pendleton et al. 2017; Pirnie et al. 2017).

Terminal erythropoiesis is an excellent model system for studies of IR. Primary erythroblasts differentiating in culture carry out a robust IR program (Edwards et al. 2016; Pimentel et al. 2016) impacting the expression of many important erythroid genes. In mature erythroblasts, IR transcripts comprise 25%–50% of steady state RNA for genes encoding essential splicing factors (SF3B1 and others), mitochondrial iron importers required for heme biosynthesis (mitoferrins, encoded by *SLC25A37* and *SLC25A28*), and major cytoskeletal proteins (alpha spectrin, encoded by *SPTA1*). Moreover, the cellular complement of IR events is continuously remodeled in a differentiation stage-dependent manner through the combined effects of differentiation-independent and -dependent IR networks (Pimentel et al. 2016).

In the current study, we focused on regulatory mechanisms for a subset of genes represented by *SF3B1*. SF3B1 is an essential splicing factor that functions in normal 3′ splice site regulation. Mutations in the *SF3B1* gene are found in many MDS (myelodysplastic syndrome) patients, where they induce RNA processing errors due to altered 3′ splice site selection (Obeng et al. 2016) and changes in exon skipping (Jin et al. 2017). We used a combination of comparative genomics, RNA-seq and RT-PCR analysis, and minigene splicing reporters to discover highly conserved decoy exons that strongly influence the level of intron 4 retention. Furthermore, bioinformatics analysis of RNA-seq data demonstrated that decoy exons are a common feature of large (>1 kb) retained introns. We propose that decoy exons interact nonproductively with intron-terminal splice sites to block intron excision, and that this mechanism regulates a critical subset of IR events in differentiating erythroblasts.

RESULTS

SF3B1 retained intron 4 (i4) harbors cryptic exon(s) that are highly conserved

Comparative genomic analysis showed that *SF3B1* i4 is extremely conserved among vertebrate genomes (Fig. 1A). Three 125–200 nt regions are 93%–98% identical from chicken to human, and core areas of two are 79%–94% identical from zebrafish to man (Supplemental Fig. S1). Sequence inspection revealed several pairs of consensus 3′ and 5′ splice sites in these ultra conserved regions, as well as in three additional conserved regions, predicting six short exons of 29–56 nt (Fig. 1A, E4a–E4f). The extraordinary conservation of these exons to fish (E4d and E4e), reptiles (E4b, E4c, and E4f), and mammals (E4a) is shown in Supplemental Figure S1, and suggests that these cryptic exons might have important function(s) in *SF3B1* regulation. Consistent with the idea that the splicing machinery can recognize these exons, the 3′ splice site factors U2AF1 and U2AF2 cross-linked to most of the cryptic exons in duplicate eCLIP experiments performed on K562 erythroleukemia cells (Fig. 1A, lower panels) and to all six exons in HEPG2 cells (not shown). Since all except E4c would introduce premature termination codons (PTCs), we hypothesized a noncoding function for these exons.

In order to study expression of cryptic exons in *SF3B1* intron 4, we performed transcriptome analysis of early erythroblast progenitors (culture day 9; D9) and mature erythroblasts (culture day 16; D16) that had been treated with cycloheximide plus emetine to inhibit nonsense-mediated decay (NMD). This strategy increased the relative abundance of transcripts containing exons with PTC, allowing us to validate expression of two cryptic exons (E4b and E4e) in the RNA-seq profiles (Fig. 1B). When this RNA was examined by RT-PCR under conditions optimized for amplification of small exon inclusion products but not larger intron retention products, inclusion of two cryptic exons was confirmed (Fig. 1C). Furthermore, analysis of individual RNA-seq reads revealed rare splice junctions that validate expression of all six predicted cassettes, and indicate extensive connections between and among these cryptic exons and the flanking constitutive exons 4 and 5 (Fig. 1D). Two important features were noted in this analysis. First, each cryptic exon was represented by at least one unique RNA-seq read that connected it to both upstream and downstream exons, confirming its ability to be recognized and spliced as a discrete exon. Second, splice junctions that link constitutive exons 4 and 5 with intron-internal sites were surprisingly abundant. In the D16 sample, E4–E4b and E4–E4e junctions together represented ∼33% of splice junction reads (261/778) that connect E4 to downstream sequences, while ∼30% of E5 upstream splice junction reads (223/739) involved E4b or E4e. Splice junctions that connect E4b to E4e were also common. At lower frequency, splice junctions connected the other cryptic exons to each other and to the flanking constitutive exons (Fig. 1D).

Together these experiments confirmed that the intron-terminal splice sites of i4 interact with internal splice sites associated with cryptic exons. We hypothesized that the cryptic exons might function as decoys whose splice sites could compete with the cross-intron interactions required for intron excision, and thereby might promote intron retention. We reasoned that many of these interactions might exhibit "leaky" splicing, leaving behind splice junctions that serve as indirect evidence for these interactions. This model

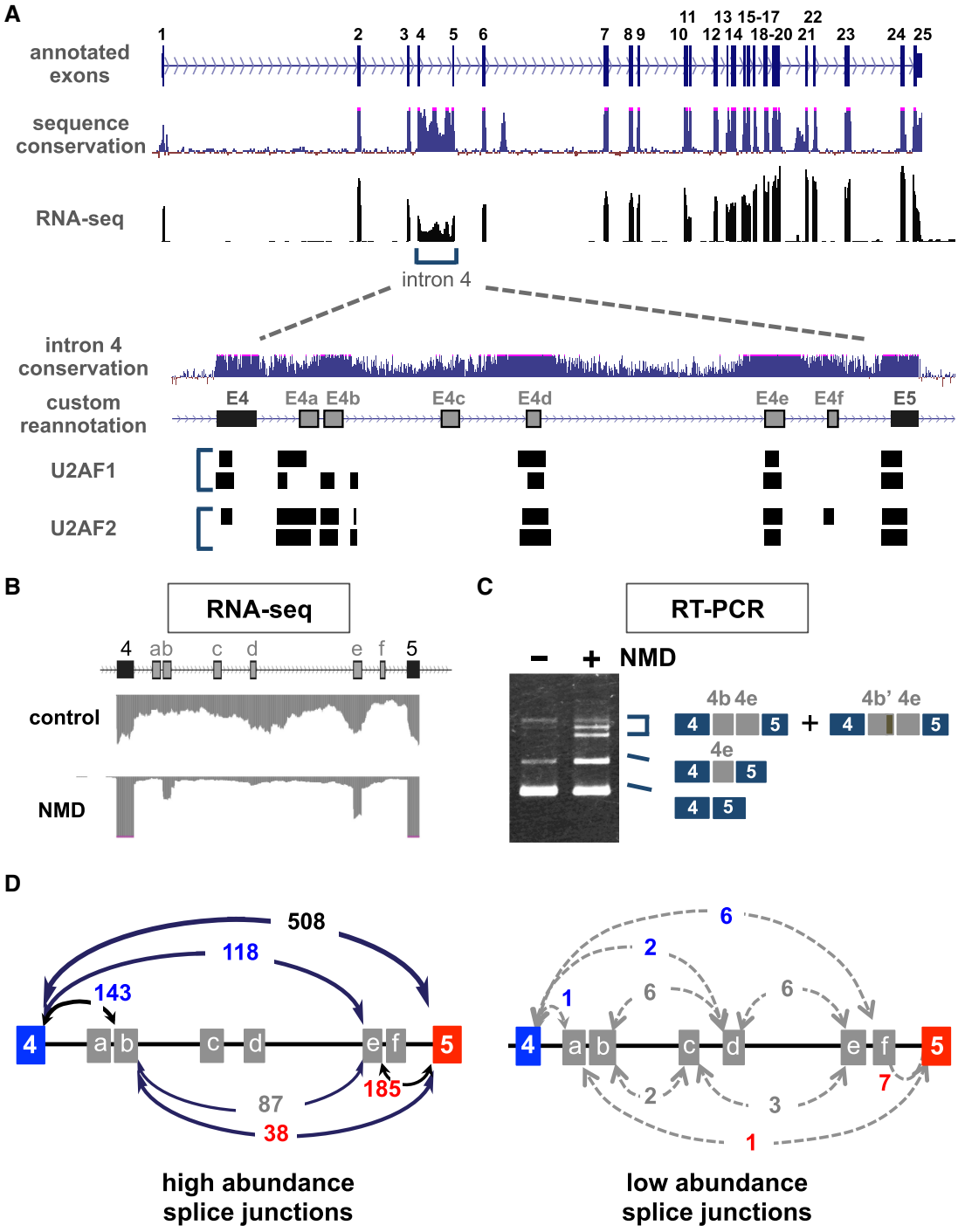

FIGURE 1. Structure and expression of *SF3B1* intron 4. (*A*) *Top* portion, genome browser tracks showing exons, conservation, and RNA-seq coverage of the *SF3B1* gene. *Bottom* portion, closeup of i4 region showing predicted cryptic exons E4a-E4f and binding data for U2AF subunits. (*B*) RNA-seq coverage in control cells versus NMD-inhibited cells. (*C*) RT-PCR analysis of cryptic exon expression. (*D*) Plots showing RNA-seq junctions in NMD-inhibited D16 erythroblasts.

suggested several testable predictions: that deleting decoy exons or mutating their splice sites should reduce intron retention; that known 3′ splice site factors should cross-link to the cryptic junctions; and that analogous decoy exons/splice junctions should be common in retained introns.

Decoy exons promote *SF3B4* i4 retention

We investigated the sequence requirements for intron retention using a series of minigene splicing reporters. The wild-type (WT) construct contains a 4.7 kb fragment of the

SF3B1 gene spanning exons 3–6 and includes full-length intron sequences in this region (Fig. 2A). This reporter was spliced to produce two major products when transfected into K562 cells: a fully spliced RNA containing exons 3–6, and a larger transcript that has excised introns 3 and 5 but specifically retained i4 (Fig. 2B, left lane). In contrast, little intron retention was observed in HEK (human embryonic kidney) cells (Fig. 2B, right lane), demonstrating cell type-specific regulation of i4 retention.

To test whether the cryptic exons in intron 4 function as decoys to promote intron retention, we constructed a series of splicing reporters in which these putative decoys were deleted. Deletion of individual exons, together with short flanking intron sequences as indicated in Supplemental Table 1, resulted in slightly reduced IR compared with the WT construct. Deletion of E4e had the greatest effect, suggesting that it possesses the strongest IR activity (Fig. 2C). Deletion of all six decoys resulted in a smaller intron with minimal IR (Fig. 2D, compare first two lanes), while adding back a single copy of E4e restored substantial retention (lane +E4e). As a control, we showed that loss of IR in the decoy-deficient reporter was not due to the smaller size of the intron, since E4e-containing constructs with similar-sized introns retained significant IR (lanes -E4abcd and –E4abc).

Decoy exon E4e can promote IR at heterologous sites

We next investigated whether E4e could promote IR in different intronic contexts. First, the ability of E4e to function at a heterologous position in i4 was tested. Indeed, the reduction in IR observed when E4e was deleted from the wild-type construct (Fig. 3B, compare WT with Δ4e) was rescued when E4e was substituted at the E4d site (mut21).

We then asked whether IR-promoting elements in i4 could induce IR in a nonretained intron. Intron 5 (i5) in the *SF3B1* gene ordinarily exhibits negligible IR in either the endogenous gene (not shown), or in splicing reporters (Fig. 3B, right gel, lanes WT). In construct mut31, most of i5 (~1.1 kb) was replaced with a similar length of i4 including E4a through E4e, but ~160 nt of the natural i5 sequence was maintained at either end. This arrangement induced substantial retention of i5. When E4e alone was inserted at a random site within i5 (mut24.2), modest IR activity was observed, but it was accompanied by substantial inclusion of E4e. Thus, decoy exons can promote IR in a heterologous intron, independent of specific sequences at the intron-terminal splice sites or flanking constitutive exons. The neighboring sequence environment of a decoy is critical, however, since a "permissive" sequence context in intron 5 apparently enhanced exon E4e inclusion rather than the IR phenotype observed in intron 4. In fact, for this experiment it was necessary to mutate a cryptic 5′ splice site in intron 5 (marked by the "X"), because otherwise the major product was strong inclusion of an aberrant E4e-related exon.

Candidate decoy exons occur in many retained introns

Studies of IR-promoting elements in introns of *OGT* (Park et al. 2017) and *ARGLU1* (Pirnie et al. 2017) previously showed that internal intronic elements, including at least

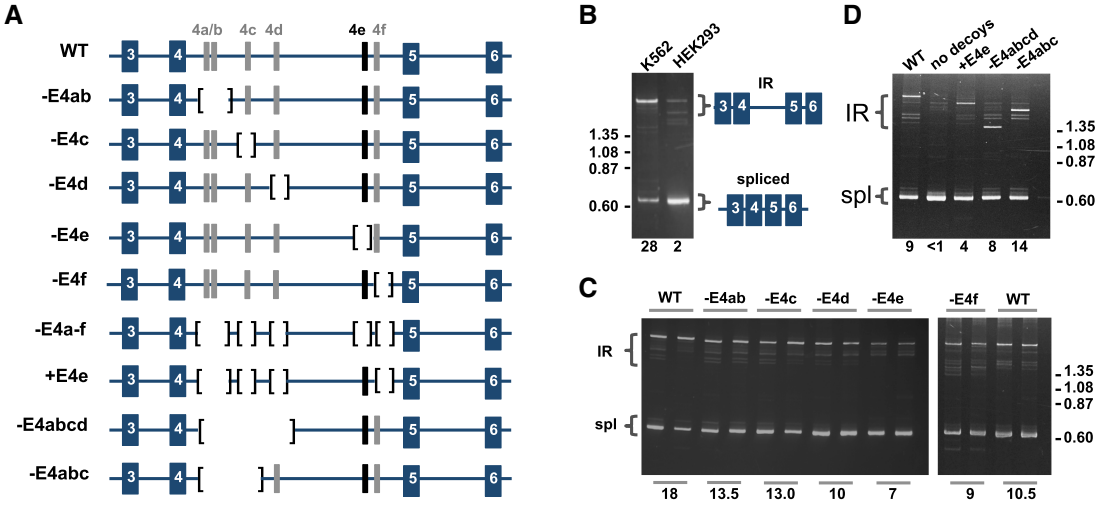

FIGURE 2. Intron retention assays in minigene splicing reporters. (*A*) Structure of minigene splicing reporters spanning the E3–E6 region of *SF3B1*. Shown are the wild-type reporter (WT) and a series of variants with deletions of candidate decoy exons. (*B*) Splicing assay performed with the wild-type reporter in K562 and HEK293 cells. (*C*) RT-PCR assays show that single-decoy deletions have mild effects on IR. Numbers *below* the lanes indicate apparent IR percent as determined by densitometry. Construct –E4f was analyzed (along with the WT control) in a separate experiment. Band intensities were corrected for differences in size of the amplification products, but the absolute value for intron retention is likely underestimated due to amplication bias of the large IR band relative to more efficiently amplified small spliced products. Duplicate lanes represent independent transfections performed in parallel under identical conditions. (*D*) Splicing assays with multidecoy deletions show that even a single decoy (E4e) can promote IR.

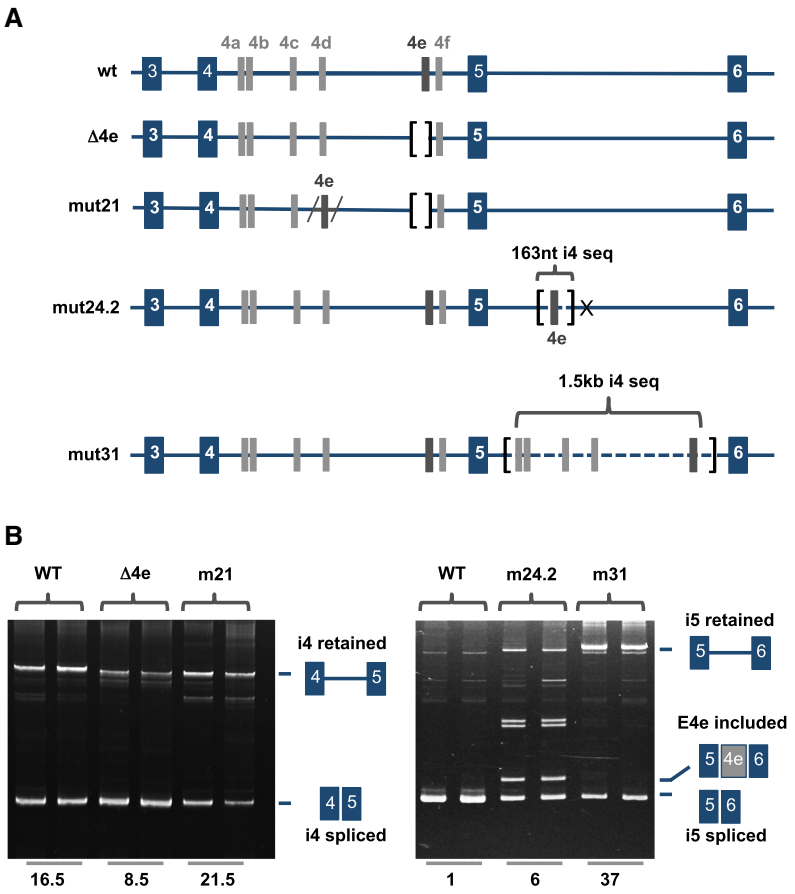

FIGURE 3. Testing IR activity of a decoy exon at heterologous sites. (*A*) Structure of minigene splicing reporters testing function of E4e in a different region of i4 (mut21) or in i5 (mut24.2, mut31). (*B*) RT-PCR analysis of intron retention showing that decoy exons can promote retention at heterologous sites in the same intron or in another intron. Numbers *below* the lanes indicate apparent IR percent as determined by densitometry, corrected for differences in band size. Transfections were performed in parallel under identical conditions, and each experiment was performed at least twice.

one cryptic cassette exon, can promote IR. In the *OGT* gene, an intron splicing silencer (ISS) inhibits intron 4 removal, thus favoring intron retention and contributing to *O*-GlcNAc homeostasis (Park et al. 2017). This *OGT* intron was strongly retained in erythroblasts, and splice junction data revealed a novel ~143–153 nt cassette exon that partially overlaps the reported ISS and possesses alternative 3′ and 5′ splice sites (Supplemental Figs. S2, S3). In *ARGLU1*, a highly conserved cassette exon was previously shown to promote IR in HeLa cells (Pirnie et al. 2017). This exon was expressed in erythroblasts as a 58–200 nt cassette, depending on alternative splice site usage (Supplemental Figs. S2, S3). To begin identifying additional decoys, we manually inspected splice junction reads in highly retained introns reported earlier (Pimentel et al. 2016). Among the subset of introns with good candidate decoys, several mapped to transcripts that encode widely expressed RBPs (DDX39B, SNRNP70, and FUS) or to important erythroid-specific proteins (SPTA1 and KEL). Splice junction reads showed in NMD-inhibited cells showed that these putative decoys could splice to the flanking constitutive exons (Supplemental Fig. S3), and RNA-seq profiles demonstrated that each of the decoys was flanked on both sides by retained intron(s). This genomic configuration is consistent with our earlier finding that PTC-containing alternative exons are often situated between two consecutive retained introns (Pimentel et al. 2016). Together these observations suggested that decoy cassettes might be a common phenomenon among retained introns.

Next, to address the genome-wide potential for decoy exon-mediated IR, we correlated IR and splice junction data for all introns in erythroid-expressed genes. Among 20,534 retained introns identified in the early erythroblast D9 sample, 770 predicted cassettes were identified (Supplemental Table 2). Late erythroblast D16 cells, that have downregulated expression of many genes, exhibited 411 cassettes in 9677 retained introns. Interestingly, the incidence of candidate decoys was strongly dependent on intron length; both D9 and D16 transcriptomes exhibited a much higher frequency of cassettes in longer retained introns (≥1 kb; 12%–17%) than in shorter retained introns (~2%). Consistent with this observation, the median length of retained introns with cassettes was much greater (1176–1280 nt) than that of retained introns lacking cassettes (258–311 nt). These results provided the first clue that distinct IR mechanisms might preferentially regulate introns of different lengths.

However, the analysis above does not take into account that some of the cassettes are likely to represent "conventional" NMD-inducing exons that play no role in IR. We reasoned that the real frequency of IR-promoting cassettes could be estimated by comparing cassette frequency in retained versus nonretained introns, since the latter group by definition lacks IR-promoting cassettes. Table 1 shows that the frequency of cassettes was three- to fourfold higher in the large (>1 kb) retained introns (R-introns) than in the comparable nonretained introns (N-introns), for both D9 and D16 transcriptomes. Based on the larger overall numbers of N-introns, and the greater frequency of cassettes in R-introns, we conclude that most noncoding cassette exons in erythroblasts are located in nonretained introns and are not relevant to IR. However, a majority of cassettes that are located in R-introns likely do act to promote IR.

TABLE 1. Frequency of candidate decoy exons in different intron categories

	R-introns >1 kb			R-introns <1 kb		
Sample	Total	+cass.	% Cassette	Total	+cass.	% Cassette
D9	4038	482	11.9	16,496	288	1.7
D16	1436	238	16.6	8241	173	2.1
	N-introns>1 kb			N-introns<1 kb		
Sample	Total	+cass.	% Cassette	Total	+cass.	% Cassette
D9	82,853	3600	4.3	36,664	179	0.5
D16	78,454	3443	4.4	41,360	244	0.6

These observations support the concept of two mechanistically distinct classes of retained introns: longer introns in which decoy exons often promote IR, and shorter introns that are retained via decoy-independent mechanism(s). Overall, the data indicates that several hundred introns may be regulated by the decoy mechanism in human erythroblasts. The true number could be higher, if some functional decoys were not detected because they are poorly spliced and did not generate splice junction reads.

Heterologous decoy exons can promote intron 4 retention

We proposed that heterologous IR-associated cassette exons could substitute for *SF3B1* E4e to promote retention of i4. To explore this hypothesis, we assayed the function of several exons identified above, i.e., those located in highly retained introns from genes prominently expressed in late stage erythroblasts. Each candidate decoy, together with its natural splice sites and at least 40 nt of flanking intron sequences, was inserted into splicing reporter ΔE4e, which has low intrinsic IR (Fig. 4A). Changes in retention were then assessed in transfected K562 cells (Fig. 4B). As a positive control we first tested the 58nt *ARGLU1* exon, and in parallel tested several additional candidate decoys. In the minigene assay, the *ARGLU1* decoy exhibited strong IR activity as indicated by increased intensity of the IR bands and decreased intensity of the spliced products compared with ΔE4e (Fig. 4B, compare lanes 2 and 3). Even stronger IR activity was associated with the *OGT* cassette (lane 4). *DDX39B* intron 6 was predicted to encode a decoy of 122–369 nt, depending on the use of alternative splice sites; this decoy also exhibited strong IR activity (lane 5). A 60 nt cassette in *SNRNP70* intron 7 showed variable but relatively low activity (lane 6). In contrast, neither an 85 nt predicted cassette in *FUS* intron 7, nor a 33 nt cassette in *KEL* intron 6, had detectable IR activity (lanes 7–8).

Decoy splice sites are critical for IR function

Using the strong *OGT* decoy exon as a model, we tested the prediction that splice sites are essential for the IR-promoting activity of decoy exons. The decoy-deficient *SF3B1* splicing reporter exhibited little or no retention (Fig. 5, lanes 1–2), while insertion of the *OGT* decoy promoted strong IR (lanes 3–4). Mutating GT dinucleotides to CT at both 5′ splice sites of the decoy exon (Supplemental Fig. S2) essentially abrogated IR in favor of completely spliced transcripts (Fig. 5, lanes 5–6). At the 3′ splice site, mutation of the two alternative AG dinucleotides also eliminated retention of the full-length intron (Fig. 5, lanes 7–8), but had a more complex phenotype due to activation of a cryptic 3′ splice site that decreased the size of the decoy exon. Fully spliced (intron excision) transcripts were not observed in this mutant. Instead, (truncated) decoy inclusion products were generated together with partial IR products retaining only the downstream portion of the intron. The observation that splice site mutations almost completely abrogated retention of the full intron strongly supports the decoy exon model.

Binding of U2AF1 and U2AF2 to decoy exons

If decoy exons compete with intron terminal splice sites to enhance IR, the decoy model predicts that decoy splice junctions must be recognized by the splicing machinery. We showed above that U2AF1 and U2AF2, important 3′ splice site factors, bind to the 3′ splice site regions of *SF3B1, OGT, ARGLU1*, and *DDX39B* decoy exons, but little or no binding was associated with candidate decoys that did not promote IR (Fig. 1; Supplemental Fig. S3). Here we tested the association between decoy exons and U2AF binding more globally by examining ENCODE eCLIP data that define U2AF1 and U2AF2 binding sites in the K562 transcriptome. The U2AF binding profiles in Figure 6 indicate that preferential binding in retained introns

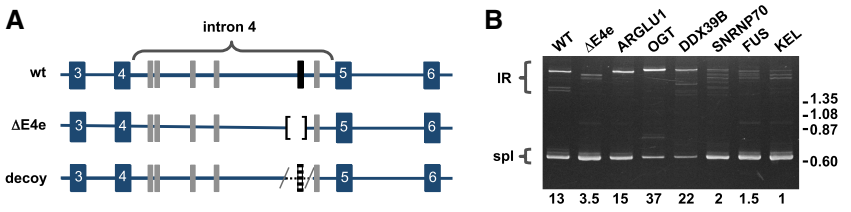

FIGURE 4. Functional testing of heterologous decoy exons in the *SF3B1* splicing reporter. (*A*) Splicing reporters used to assay heterologous decoys. Construct "decoy" had E4e deleted and replaced with candidate decoys from other genes. (*B*) RT-PCR analysis of intron retention activity associated with heterologous decoys, cloned from the indicated genes. IR, intron retention product; spl, spliced product. Numbers *below* the lanes indicate apparent IR percent as determined by densitometry, corrected for differences in band size. Numbers at the *right* of the gel indicate size markers in kb.

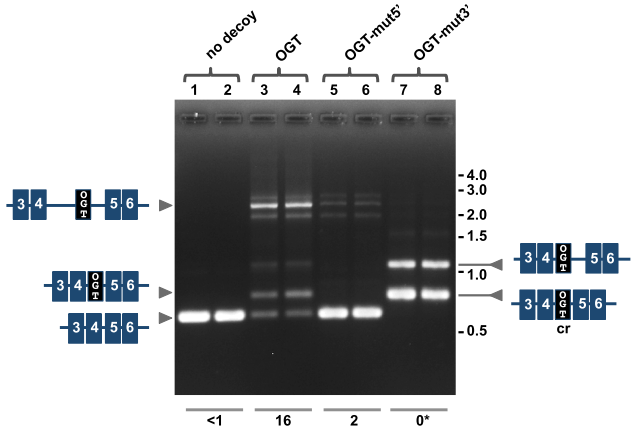

FIGURE 5. Testing importance of decoy splice sites for promotion of IR. Gel shows splicing results from the indicated reporters. Wild-type and mutated splice sites were as follows: 5′ splice site: WT = ATGgtaacgggt; mut5′ = ATGctaacggct; 3′ splice site: WT = tttagaagGTT; mut3′ = tttacaacGTT. Underlined nucleotides were altered from "g" in WT to "c" in mutants. WT, wild type; cr, cryptic splice site. Duplicate lanes represent independent transfections performed in parallel under identical conditions. Numbers *below* gel indicate apparent retention of full length intron. (*) Partial intron retention, of downstream sequences only, was ~33%. Size markers in kb are indicated at the *right* margin of the gel.

occurs primarily at the 3′ splice site region of decoy exons, as well as the 3′ splice site of the downstream constitutive exon. Good binding peaks for both U2AF1 and U2AF2 were observed at the expected location upstream of these exons, similar to binding patterns for a control set of known cassettes expressed in K562 cells (native cassette exons). As shown in the figure, comparable binding profiles were obtained in replicate eCLIP experiments for both U2AF1 and U2AF2. These binding profiles support the decoy model's prediction that decoy exons bind to U2AF splicing factors.

DISCUSSION

Experimental and computational data support the concept that decoy-mediated IR is a novel mechanism for regulating an important component of the erythroid IR program. Experimental analysis of *SF3B1* minigene splicing reporters demonstrated that highly conserved decoy exons are required for optimum intron retention, that decoy exons can induce IR in an ordinarily nonretained intron, and that heterologous decoys from other retained introns can promote *SF3B1* i4 retention. In parallel, computational analysis of RNA-seq data from NMD-inhibited cells suggested that a broader program of decoy exon-regulated IR is executed during terminal erythropoiesis. Our model expands the range of splicing outcomes available to noncoding cassette exons. As shown in Figure 7, skipping of these exons leads to productive splicing and generation of translatable mRNAs; inclusion of decoy exons yields unstable transcripts that are subject to NMD; and nonproductive interaction with intron-terminal splice sites represents a novel IR outcome. Notably, the decoy exon mechanism greatly extends the model reported recently for *ARGLU1*, where it was proposed that unproductive splicing complexes assembled at the alternative exon disfavor intron splicing so as to promote its retention (Pirnie et al. 2017). The model is also consistent with previous reports that RIs are enriched adjacent to alternative exons (Braunschweig et al. 2014; Pimentel et al. 2016). Depending on variables including splice site strength, nearby enhancer and silencer elements, and physiological context, we envision that dual-function cassettes may post-transcriptionally direct transcript outcomes preferentially toward NMD or toward IR. At the extremes, some cryptic cassettes may function solely to induce NMD, while others may function predominantly to promote IR. The latter subset of decoys might be difficult to detect using splice junction criteria, since they would rarely splice to the flanking constitutive exons.

The role of decoy-mediated IR in terminal erythropoiesis remains to be investigated. In human erythroblasts, our data indicate that decoy-regulated RIs comprise several hundred erythroblast retention events, mostly involving >1 kb introns that are mechanistically distinct from the more numerous population of smaller RIs that generally lack decoys. Given our earlier finding that several major spliceosomal factors possess highly retained introns (Pimentel et al. 2016), and new data showing that some of these possess IR-promoting decoys, we suggest this mechanism could modulate changes in splicing capacity of late stage erythroblasts as they reduce gene expression in preparation for enucleation. Another possibility is that regulated IR could contribute to balanced expression of competing or cooperating genes. For example, modulation of IR plays an important role in regulating expression for competing OGT and OGA enzymes to ensure *O*-GlcNAc homeostasis (Park et al. 2017), and the documented decoy exon in *OGT* likely contributes to that control. Speculatively, since the alpha spectrin gene *SPTA1* has a decoy exon in retained intron 20, IR could help ensure balanced expression of alpha and beta spectrin subunits, two structural polypeptides that form long heterodimers and provide mechanical support to the red cell membrane skeleton.

The molecular mechanism by which decoy sites can nonproductively engage intron-terminal splice sites requires further study. Decoy splice site function must be carefully tuned, since strong splice sites would favor decoy exon inclusion over intron retention, while splice site-inactivating mutations would abrogate IR-promoting activity. One feature that may be critical is the presence of multiple competing splice sites at the decoy exon. Multiple 5′ and/or 3′ splice sites are a feature of the decoy exon in *ARGLU1* (Pirnie et al. 2017) as well as *OGT* and *DDX39B* (Supplemental Fig. S2). Moreover, even when explicit splice junction multiplicity is not evident, some decoys possess potentially competing splice site motifs in the proximal introns (results not shown). On the other hand, cassette exons encoded in retained introns from *FUS* and *KEL* did not appear to have alternative splice sites, and did not exhibit IR activity in the splicing reporter assay.

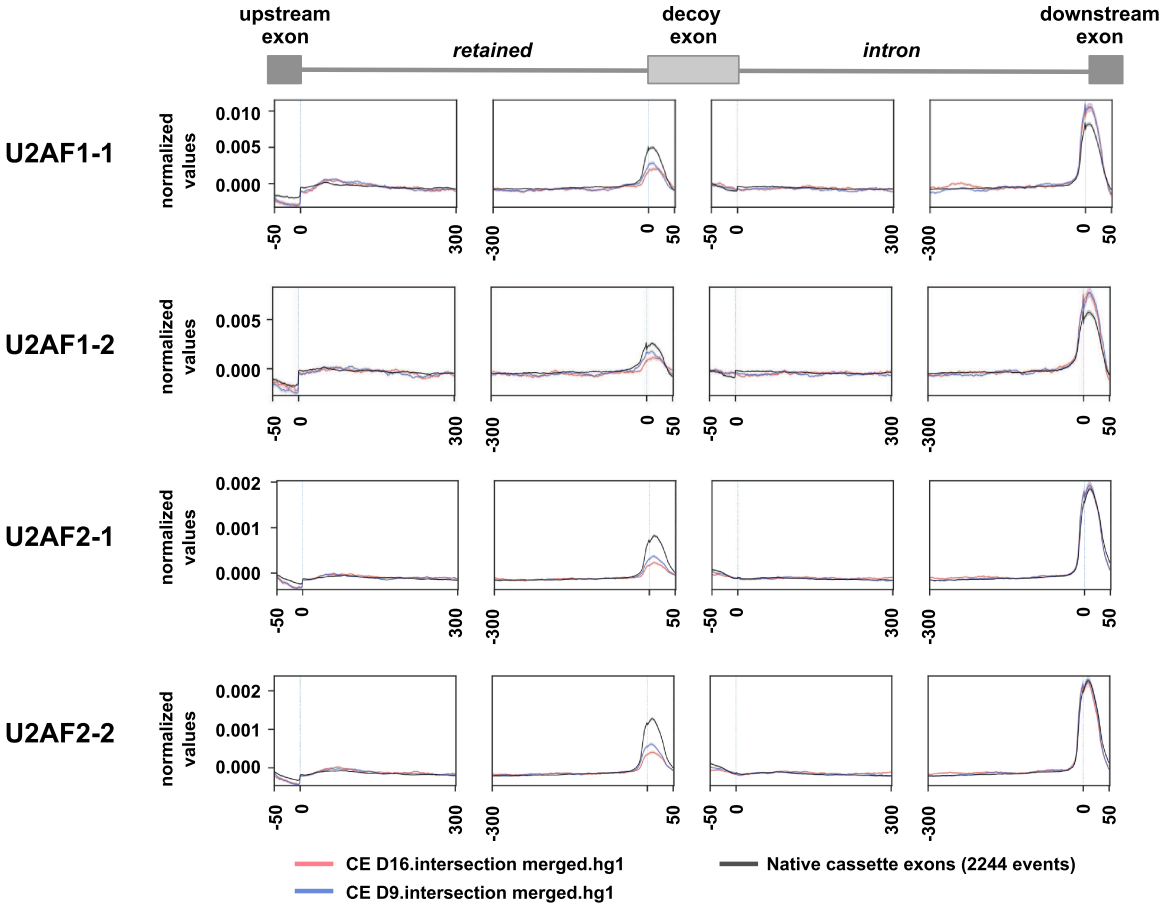

FIGURE 6. U2AF binding profiles in retained introns with candidate decoys. Enhanced CLIP binding profiles for U2AF1 and U2AF2 in K562 cells. Profiles show enriched binding for U2AF factors near candidate decoys was similar to that observed for a control set of cassette exons in the same cells. U2AF1-1 and U2AF1-2 represent replicate eCLIP experiments for U2AF1; U2AF2-1 and U2AF2-2 represent replicate eCLIP experiments for U2AF2.

Competition between splice sites is a fundamental governing principle of alternative splicing. The concept that competing sites can be spliced inefficiently, functioning mainly as decoys to block functional use of other sites, is also well grounded in previous work. An early example was the Drosophila P-element gene, where U1 snRNP binding to an exonic pseudo-5′ splice site was shown to compete with the normal downstream 5′ splice site to promote retention of the adjacent intron in somatic cells (Siebel et al. 1992). Subsequent studies showed that a decoy 3′ acceptor can engage the 5′ splice site of an upstream exon to promote its skipping in caspase-2 (Côté et al. 2001) and other transcripts (Havlioglu et al. 2007). The concept that a noncoding exon might engage splice sites of both flanking exons to induce IR was suggested

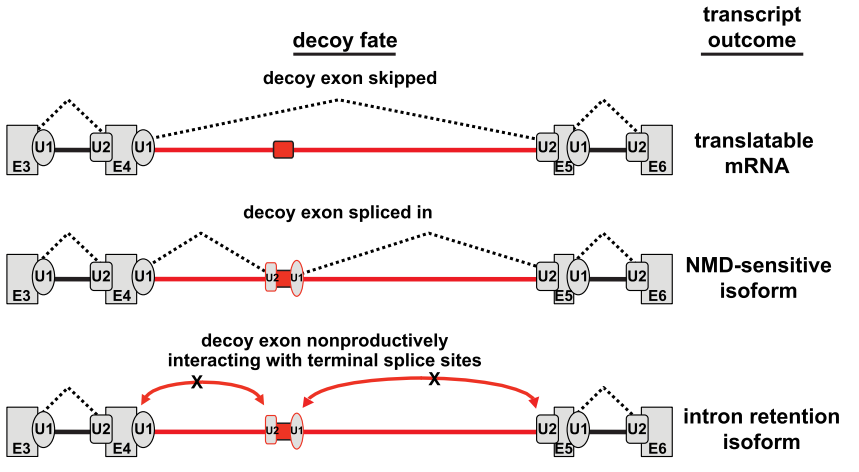

FIGURE 7. Decoy exon splicing model. Behavior of the decoy exon (red) can dictate three distinct fates for the pre-mRNA: Skipping of the exon leads to production of mRNA; inclusion of the decoy generates an NMD-sensitive isoform; and nonproductive interaction with flanking exons yields the intron retention transcript. Dotted lines indicate splicing; red curved lines indicate nonproductive interactions (no splicing). Decoy function could be regulated through the action of nearby enhancer and/or silencer elements (not shown).

recently (Pirnie et al. 2017), and the decoy exon data presented here greatly expands the potential role of decoy splice sites in regulation of splicing outcomes. Interestingly, the appreciation of decoy exons might help to explain recent observations in other systems. Far-distal branchpoints located >100 nt upstream of annotated 3′ splice sites are more common in retained introns than in constitutive introns (Pineda and Bradley 2018); the increased frequency in retained introns could be due in part to branchpoints associated with cryptic decoy exons rather than the annotated constitutive exons. In other studies, U2AF was shown to bind at numerous intronic locations not corresponding to known 3′ splice sites (Shao et al. 2014). Our data suggest that some of these likely correspond to decoy exons that promote intron retention. Finally, it is possible that U1 snRNP binding at some decoy 5′ splice sites might have an additional function in suppressing premature cleavage and polyadenylation at cryptic polyadenylation signals (Kaida et al. 2010).

In conclusion, the decoy model represents a new and distinctive component of the erythroid alternative splicing program. We speculate that developmentally regulated IR events in other cell types may also be regulated by decoy exons, and that physiological control of such programs could be mediated by context-dependent combinations of splicing enhancer and silencer proteins that impact recognition of the decoys. We anticipate that analogous subsets of IR events might be regulated by decoy mechanisms in other developmental or physiological contexts. Possible candidates could include differentiating granulocytes (Wong et al. 2013, 2017), activated T cells (Ni et al. 2016), stimulated neurons (Mauger et al. 2016), cells subjected to proteotoxic stress (Shalgi et al. 2014), proliferating versus differentiated muscle cells (Llorian et al. 2016), differentiating germ cells (Naro et al. 2017), etc. Identifying the RBPs that regulate these decoy-dependent programs will be an important goal of future studies in this area.

MATERIALS AND METHODS

Erythroblast culture

Cells were cultured as described previously (Hu et al. 2013). For NMD experiments, cultures were divided and one half of each culture was incubated with 100 μg/mL emetine for 8 h and 100 μg/mL cycloheximide for 4 h. Experiments were done in biological duplicates for a total of eight samples (two differentiation stages, two conditions plus or minus NMD inhibitors, and two biological replicates.

RNA and RNA library preparation

RNA was isolated from erythroblasts (4×10^6 cells from day 9 and 16×10^6 cells on day 16) according to the manufacturer's instructions (Qiagen). Sequencing libraries were prepared from 500 ng of total RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (catalog number E7490, protocol revision 5.0), NEBNext Ultra Directional RNA Library Preparation Kit for Illumina (catalog number E7420, protocol revision 6.0), and NEBNext Multiplex Oligos for Illumina (catalog number E7600, protocol revision 2.0) with the following modifications: E7420, Section1.2 ("mRNA Isolation, Fragmentation and Priming Total RNA") Step 37, we decreased the incubation time from 15 min to 5 min; E7420, Section 1.3 ("First Strand cDNA Synthesis"), Step 2,we increased the incubation time from 15 min to 50 min; for the size selection we used 40 μL AMPure XP beads for the first bead selection and 20 μL AMPure XP beads for the second bead selection (targeting an insert size of 300–450 bp and a final library size of 400–550 bp); and in E7420, Section 1.9A ("PCR library Enrichment"), Step 2 we used 14 cycles for PCR cycling and dual index primers (i507 and i705–i712). Individual libraries were normalized to 10 nM and eight samples were pooled per lane. Sequencing was performed at UC Berkeley's QB3 Vincent J. Coates Genomics Sequencing Laboratory on an Illumina HiSeq4000 instrument, generating 150 bp paired end reads.

RNA-seq analysis

For each sample we produced 13–60 M total reads and 4–30 M mapped reads. Replicates were merged and aligned to the GRCh38.p8 version of the human genome using TopHat version: 2.1.1 (Trapnell et al. 2009; Kim et al. 2013). BAM files generated were used for the alternative splicing analysis and for the cassette reannotation. All the splicing-junction data formatted as bigBed and bigWig files obtained in this study were uploaded onto the UCSC genome browser; it can be accessed by copying the following link into a web browser: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&hubUrl=https://sina.lbl.gov/seqdata/conboy-ucsc/hub.txt. The FASTQ files were run through the STAR aligner (Dobin et al. 2013) to produce expression scores as input for DEXSeq (Anders et al. 1984; Reyes et al. 2013)

Custom cassette reannotation scripts

A custom reannotation tool was written using the Rust language to reannotate an input annotation to include additionally discovered cassette exons. The source code can be found at GitHub: https://github.com/bdgp/cassette_reannotation. The reannotation tool was given a curated subset of the NCBI RefSeq annotation version GCF_000001405.34 based on GRCh38.p8. This curated subset only included features beginning with NM, NR, and YP. TopHat version: 2.1.1 (Trapnell et al. 2009; Kim et al. 2013) generated alignment files of the sample reads were also passed in to the reannotation tool. The tool finds constitutive splice pairs in the input annotation, then searches for overlapping reads in the input read alignment files that provide evidence for unannotated cassette exons. The tool requires at least two paired-end fragment splices for both the start and the stop of the cassette which also must splice at least one of the annotated flanking exons. The tool also requires contiguous read coverage throughout the discovered cassette. The tool then produces a reannotation that includes all of the features of the input annotation, along with newly created transcript features. The newly created transcript features are based on transcript features existing in the input annotation, but with the discovered cassettes added.

Alternative splicing analysis

The reannotated annotation including discovered cassettes, along with the alignment files, were passed to SplAdder (Kahles et al. 2016) http://github.com/ratschlab/spladder to discover retained introns and alternative exons in the annotation. Some minor bug fixes, disabling of code assertions and parameter tuning were required for SplAdder to perform the analysis. We discovered that the least-stringent default parameter set was still too stringent to detect many retained introns in our sample data, so the individual parameters were made even less stringent. Details of the parameters used in our run of SplAdder can be found in the Supplemental Material.

The retained intron adjusted PSI values returned by SplAdder unfortunately did not account for exons overlapping the retained intron, so a custom tool was written to correct the adjusted PSI values. SplAdder computes the adjusted PSI values using the average per-base coverage over the intron (intron_cov) and the number of reads with splices confirming the intron (intron_conf). The tool written to correct the adjusted PSI values recomputes the average per-base coverage by subtracting out the coverage of exonic regions. The tool uses this corrected intron_cov value along with the intron_conf directly from SplAdder to compute the corrected adjusted PSI value for the retained intron.

Splicing reporter assays

A 4.7 kb region of the human *SF3B1* gene extending from the 3′ end of intron 2 to the 5′ end of intron 6, was amplified using the following primers: F: 5′-tggaattctgcagatAAGGAGGGCTTAGACATCAC AC-3′; R: 5′-gccagtgtgatggatCTATGGCAACCCAAGCAGA-3′. The fragment was cloned into pcDNA3.0 using In-Fusion methods (Gibson 2011) with 15 nt in lower case sequence representing overlap with the ends of EcoRV-linearized vector. The splicing reporter was transfected into K562 using Fugene HD according to the manufacturer's instructions (Promega). RNA was harvested after 48 h and purified according to the manufacturer's instructions (Qiagen), but with the addition of a DNase step to eliminate potential contamination by genomic DNA. RNA was reverse transcribed with Superscript III (Invitrogen) into cDNA using the BGH reverse primer in the vector (5′-tagaaggcacagtcgagg-3′). Spliced products were amplified using a forward primer in exon 3 (5′-catcatctacgagt ttgcttgg-3′) and a reverse primer in the vector (5′-atttaggtgacactatag aatagggc-3′). This strategy amplified minigene-derived transcripts but not endogenous SF3B1 mRNA, as confirmed using RNA from untransfected or empty vector-transfected cells. When assaying IR products, PCR reaction conditions were adjusted to allow for amplification of DNA bands ≥3 kb in length (denaturation at 95°C for 20 sec , annealing at 56°C for 10 sec, extension at 70°C for 2 min 30 sec; 35 cycles) using KOD polymerase in the presence of betaine to enhance amplification. PCR products were analyzed on either 2% agarose gels or 4.5% acrylamide gels. All PCR products discussed in the manuscript were confirmed by DNA sequencing, and all splicing reporter constructs were assayed a minimum of three times. Splicing behavior of test constructs consistently exhibited the same behavior with regard to IR efficiency, relative to control constructs assayed in parallel under identical conditions, despite inevitable variation in baseline intron retention from experiment to experiment.

Enhanced CLIP analysis

Splicing maps were generated using U2AF1 (ENCSR862QCH) and U2AF2 (ENCSR893RAV) eCLIP normalized densities overlapped with selected retained introns containing candidate decoy exons at two time points (D9, D16). These density values were normalized first using an RPM transformation to account for variation in sequencing depth, then normalized again using equivalent densities from a size-matched input sample. To perform this second normalization, both IP and equivalent input signals were transformed into their probability densities to preserve overall shape of binding and to reduce signal dominance from a few events. Input densities for each event were then subtracted from the corresponding IP to remove background signal. The final density value represents the mean of these normalized densities devoid of any value exceeding the 95% median at each position to reduce confounding outlier effects.

In addition to candidate decoy exons, normalized densities were also overlapped with a set of cassette exons, derived from a subset of Gencode (v19) constitutive exons. Within these annotations, we define a "cassette" as any exon found between 10% and 90% spliced in at least 50% of all shRNA knockdown control data (encodeproject.org, all nonspecific target controls, aligned to hg19 with TopHat), filtering any region that is not supported by at least 30 reads. From this set, regions with the highest inclusion average between two replicates were chosen if any regions overlapped to remove any possibility of double counting the eCLIP signal.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work was funded by National Institutes of Health grant 5R01DK108020 (J.G.C.) and by the Director, Office of Science and Office of Biological & Environmental Research of the US Department of Energy (DE-AC02-05CH1123). B.Y. and G.W.Y. are partially supported by the National Institutes of Health under grants HG004659 and HG007005. We acknowledge Brenton Graveley's laboratory for sharing RNA-seq data sets generated within the ENCODE project. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant.

Received April 23, 2018; accepted June 26, 2018.

REFERENCES

Anders RF, Coppel RL, Brown GV, Saint RB, Cowman AF, Lingelbach KR, Mitchell GF, Kemp DJ. 1984. Plasmodium falciparum complementary DNA clones expressed in *Escherichia coli* encode many distinct antigens. *Mol Biol Med* **2:** 177–191.

Bergeron D, Pal G, Beaulieu YB, Chabot B, Bachand F. 2015. Regulated intron retention and nuclear pre-mRNA decay contribute to PABPN1 autoregulation. *Mol Cell Biol* **35:** 2503–2517.

Boothby TC, Zipper RS, van der Weele CM, Wolniak SM. 2013. Removal of retained introns regulates translation in the rapidly developing gametophyte of *Marsilea vestita*. *Dev Cell* **24:** 517–529.

Boutz PL, Bhutkar A, Sharp PA. 2015. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev* **29:** 63–80.

Braun CJ, Stanciu M, Boutz PL, Patterson JC, Calligaris D, Higuchi F, Neupane R, Fenoglio S, Cahill DP, Wakimoto H, et al. 2017. Coordinated splicing of regulatory detained introns within oncogenic transcripts creates an exploitable vulnerability in malignant glioma. *Cancer Cell* **32:** 411–426.

Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. 2014. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* **24:** 1774–1786.

Cho V, Mei Y, Sanny A, Chan S, Enders A, Bertram EM, Tan A, Goodnow CC, Andrews TD. 2014. The RNA-binding protein hnRNPLL induces a T cell alternative splicing program delineated by differential intron retention in polyadenylated RNA. *Genome Biol* **15:** R26.

Côté J, Dupuis S, Jiang Z, Wu JY. 2001. Caspase-2 pre-mRNA alternative splicing: identification of an intronic element containing a decoy 3′ acceptor site. *Proc Natl Acad Sci* **98:** 938–943.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21.

Edwards CR, Ritchie W, Wong JJ, Schmitz U, Middleton R, An X, Mohandas N, Rasko JE, Blobel GA. 2016. A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. *Blood* **127:** e24–e34.

Gibson DG. 2011. Enzymatic assembly of overlapping DNA fragments. *Methods Enzymol* **498:** 349–361.

Havlioglu N, Wang J, Fushimi K, Vibranovski MD, Kan Z, Gish W, Fedorov A, Long M, Wu JY. 2007. An intronic signal for alternative splicing in the human genome. *PLoS One* **2:** e1246.

Hu J, Liu J, Xue F, Halverson G, Reid M, Guo A, Chen L, Raza A, Galili N, Jaffray J, et al. 2013. Isolation and functional characterization of human erythroblasts at distinct stages: implications for understanding of normal and disordered erythropoiesis in vivo. *Blood* **121:** 3246–3253.

Jacob AG, Smith CWJ. 2017. Intron retention as a component of regulated gene expression programs. *Hum Genet* **136:** 1043–1057.

Jin S, Su H, Tran N-T, Song J, Lu SS, Li Y, Huang S, Abdel-Wahab O, Liu Y, Zhao X. 2017. Splicing factor SF3B1K700E mutant dysregulates erythroid differentiation via aberrant alternative splicing of transcription factor TAL1. *PLoS One* **12:** e0175523.

Kahles A, Ong CS, Zhong Y, Rätsch G. 2016. SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics* **32:** 1840–1847.

Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468:** 664–668.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14:** R36.

Li Y, Bor YC, Fitzgerald MP, Lee KS, Rekosh D, Hammarskjold ML. 2016. An NXF1 mRNA with a retained intron is expressed in hippocampal and neocortical neurons and is translated into a protein that functions as an Nxf1 cofactor. *Mol Biol Cell* **27:** 3903–3912.

Llorian M, Gooding C, Bellora N, Hallegger M, Buckroyd A, Wang X, Rajgor D, Kayikci M, Feltham J, Ule J, et al. 2016. The alternative splicing program of differentiated smooth muscle cells involves concerted non-productive splicing of post-transcriptional regulators. *Nucleic Acids Res* **44:** 8933–8950.

Mauger O, Lemoine F, Scheiffele P. 2016. Targeted intron retention and excision for rapid gene regulation in response to neuronal activity. *Neuron* **92:** 1266–1278.

Naro C, Jolly A, Di Persio S, Bielli P, Setterblad N, Alberdi AJ, Vicini E, Geremia R, De la Grange P, Sette C. 2017. An orchestrated intron retention program in meiosis controls timely usage of transcripts during germ cell differentiation. *Dev Cell* **41:** 82–93.

Ni T, Yang W, Han M, Zhang Y, Shen T, Nie H, Zhou Z, Dai Y, Yang Y, Liu P, et al. 2016. Global intron retention mediated gene regulation during $CD4^+$ T cell activation. *Nucleic Acids Res* **44:** 6817–6829.

Ninomiya K, Kataoka N, Hagiwara M. 2011. Stress-responsive maturation of Clk1/4 pre-mRNAs promotes phosphorylation of SR splicing factor. *J Cell Biol* **195:** 27–40.

Obeng EA, Chappell RJ, Seiler M, Chen MC, Campagna DR, Schmidt PJ, Schneider RK, Lord AM, Wang L, Gambe RG, et al. 2016. Physiologic expression of *Sf3b1*^{K700E} causes impaired erythropoiesis, aberrant splicing, and sensitivity to therapeutic spliceosome modulation. *Cancer Cell* **30:** 404–417.

Park SK, Zhou X, Pendleton KE, Hunter OV, Kohler JJ, O'Donnell KA, Conrad NK. 2017. A conserved splicing silencer dynamically regulates O-GlcNAc transferase intron retention and O-GlcNAc homeostasis. *Cell Rep* **20:** 1088–1099.

Pendleton KE, Chen B, Liu K, Hunter OV, Xie Y, Tu BP, Conrad NK. 2017. The U6 snRNA m^6A methyltransferase METTL16 regulates SAM synthetase intron retention. *Cell* **169:** 824–835.e14.

Pendleton KE, Park SK, Hunter OV, Bresson SM, Conrad NK. 2018. Balance between MAT2A intron detention and splicing is determined co-transcriptionally. *RNA* **24:** 778–786.

Pimentel H, Parra M, Gee SL, Mohandas N, Pachter L, Conboy JG. 2016. A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res* **44:** 838–851.

Pineda JMB, Bradley RK. 2018. Most human introns are recognized via multiple and tissue-specific branchpoints. *Gene Dev* **32:** 577–591.

Pirnie SP, Osman A, Zhu Y, Carmichael GG. 2017. An ultraconserved element (UCE) controls homeostatic splicing of *ARGLU1* mRNA. *Nucleic Acids Res* **45:** 3473–3486.

Reyes A, Anders S, Weatheritt RJ, Gibson TJ, Steinmetz LM, Huber W. 2013. Drift and conservation of differential exon usage across tissues in primate species. *Proc Natl Acad Sci* **110:** 15377–15382.

Schmitz U, Pinello N, Jia F, Alasmari S, Ritchie W, Keightley MC, Shini S, Lieschke GJ, Wong JJ, Rasko JEJ. 2017. Intron retention enhances gene regulatory complexity in vertebrates. *Genome Biol* **18:** 216.

Shalgi R, Hurt JA, Lindquist S, Burge CB. 2014. Widespread inhibition of posttranscriptional splicing shapes the cellular transcriptome following heat shock. *Cell Rep* **7:** 1362–1370.

Shao C, Yang B, Wu T, Huang J, Tang P, Zhou Y, Zhou J, Qiu J, Jiang L, Li H, et al. 2014. Mechanisms for U2AF to define 3′ splice sites and regulate alternative splicing in the human genome. *Nat Struct Mol Biol* **21:** 997–1005.

Siebel CW, Fresco LD, Rio DC. 1992. The mechanism of somatic inhibition of *Drosophila* P-element pre-mRNA splicing: multiprotein complexes at an exon pseudo-5′ splice site control U1 snRNP binding. *Genes Dev* **6:** 1386–1401.

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

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

Wong JJ, Gao D, Nguyen TV, Kwok CT, van Geldermalsen M, Middleton R, Pinello N, Thoeng A, Nagarajah R, Holst J, et al. 2017. Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment. *Nat Commun* **8:** 15134.