Introns in UTRs: Why we should stop ignoring them

Alicia A. Bicknell[1)†], *Can Cenik*[1)†], *Hon N. Chua*[2)], *Frederick P. Roth*[2)] *and Melissa J. Moore*[1)3)*]

Although introns in 5′- and 3′-untranslated regions (UTRs) are found in many protein coding genes, rarely are they considered distinctive entities with specific functions. Indeed, mammalian transcripts with 3′-UTR introns are often assumed nonfunctional because they are subject to elimination by nonsense-mediated decay (NMD). Nonetheless, recent findings indicate that 5′- and 3′-UTR intron status is of significant functional consequence for the regulation of mammalian genes. Therefore these features should be ignored no longer.

Keywords:
ALREX; intron; nonsense-mediated decay; 3′-UTR; 5′-UTR

Introduction

A clearly appreciated role for introns in higher organisms is to allow for alternative splicing, which permits a single gene to encode many different proteins. Less widely appreciated, however, is that the presence of an intron and the act of its removal by the spliceosome can influence almost every step in gene expression from transcription and polyadenylation to mRNA export, localization, translation, and decay [1, 2]. These influences modulate both the levels and localization of expressed proteins. While ~90% of human introns occur within protein coding regions (open reading frames; ORFs), many also reside in untranslated regions (UTRs). Approximately 35% of human 5′-UTRs [3], and between ~6% (NCBI's Reference Sequence; RefSeq) and ~16% (Vertebrate Genome Association; Vega) of human 3′-UTRs are annotated as harboring introns. Yet despite their prevalence, introns in UTRs are rarely considered as distinctive entities with specific regulatory functions. Indeed, until very recently the prevailing view of 5′-UTR introns (5UIs) was that they are only special insofar as they are proximal to the 5′ end of the transcript. Further, a common view of 3′-UTR introns (3UIs) is that they are signatures of nonfunctional transcripts arising solely from genomic noise (e.g. pseudogenes, transposons), genetic mutation, or errors in splicing. This view stems from the observation that mammalian mRNAs with an intron excision site >55 nucleotides downstream of a termination codon are subject to degradation by the nonsense-mediated decay (NMD) pathway [4–8]. Reflecting the widespread view that NMD is restricted to mRNAs encoding inappropriately truncated proteins, NCBI's RefSeq database routinely excludes most 3UI-containing transcripts from its annotated coding transcripts [9]. Nonetheless, recent evidence clearly indicates that 5UIs and 3UIs do have important and unique roles in the regulation of gene expression that should not be overlooked. Below, we describe evidence that the presence or absence of a 5UI has significant consequences for both mRNA nuclear export and cytoplasmic mRNA metabolism, and that 3UIs have multiple roles in modulating normal protein expression.

Splicing directs mRNP formation

All introns can influence gene expression regardless of their position relative

DOI 10.1002/bies.201200073

[1)] Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA, USA
[2)] Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada
[3)] Howard Hughes Medical Institute, Worcester, MA, USA

***Corresponding author:**
Melissa J. Moore
E-mail: Melissa.Moore@umassmed.edu

Abbreviations:
EJC, exon junction complex; **mRNP,** mRNA protein particle; **NMD,** nonsense-mediated decay; **ORF,** open reading frame; **TREX,** transcription export; **UTR,** untranslated region; **3UI,** 3′-UTR intron; **5UI,** 5′-UTR intron.

[†] These authors contributed equally to this work.

to the coding region because they alter the protein makeup of the mRNA protein particle (mRNP). One set of splicing-dependent mRNP proteins is the exon junction complex (EJC), deposited by the spliceosome ~24 nts upstream of exon junctions on spliced mammalian mRNAs [10, 11]. This multiprotein complex remains tightly bound to the mRNA until the first round of translation, when EJCs within the coding region are displaced by ribosomes as they translocate across the message (Fig. 1) [12]. Until then, the EJC core serves both as a molecular marker of prior intron position, and as a binding platform for peripheral proteins. These peripheral factors associate transiently with the core and help regulate the subcellular localization, translation, and decay of the transcript [2, 13–15].

Also deposited on mRNAs during transcription and splicing are the transcription export (TREX) complex and SR proteins. In mammals, the TREX complex is recruited primarily to the 5′ end of transcripts through cooperative action of the nuclear cap-binding complex and the spliceosome [16, 17]. Once bound to the mRNA, the TREX complex promotes nuclear export of fully processed transcripts through the nuclear pore by direct interactions between the TREX component, Aly, and the nuclear export factor, TAP-p15 [18, 19]. SR proteins are best known for their roles in exon definition and as alternative splicing regulators. However, they are also subject to splicing-dependent dephosphorylation, which promotes their tight association with the spliced mRNA. As mRNP components, SR proteins can enhance nucleocytoplasmic export, translation, and decay of their bound mRNAs [20–22]. Thus, as elaborated below, one means by which 5UIs and 3UIs influence gene expression is by promoting the loading of mRNP proteins with downstream functional consequences.

5′-UTR introns and an alternative mRNA export pathway

Initial models suggested that 5UIs evolved under nearly neutral genetic selection, implying that they have no specific function [23]. If this were the case, one would expect 5UIs to be equally distributed among transcripts of all functional classes. Recent analyses, however, have revealed that genes having or lacking 5UIs fall into distinct functional classes, at least in the human and rat genomes. Whereas genes with regulatory functions are enriched for 5UIs, genes encoding proteins targeted to the endoplasmic reticulum (ER) or mitochondria are significantly depleted of such introns [3, 24]. When 5UIs are present, they are necessarily the most 5′ proximal introns in a transcript, and 5′ proximal introns have a disproportionate role in regulating transcription, mRNA export, and translation [16, 25–28]. However, 5′ proximity alone cannot explain the functional distribution of transcripts that do or do not contain 5UIs. Importantly, transcripts possessing only coding-region introns, and in which the first intron has the same proximity to the transcription start site as a 5UI, do not display the same functional distribution as 5UI-containing transcripts [3].

The enrichment of 5UIs in regulatory genes could reflect their tendency to have more transcription factor binding sites, which are often located within the first intron [26]. In addition, deposition of splicing-dependent mRNP components as close as possible to the 5′ end of the mRNA could play a positive role in facilitating rapid export and translation of the newly made mRNAs [28].

On the other hand, some transcripts have evolved to exclude introns from their 5′-UTRs because this allows them to use an alternate mode of nuclear export, the ALREX mRNA export pathway. Unlike the canonical TREX-dependent nuclear export pathway, the ALREX pathway does not require splicing [29]. Instead, ALREX facilitates mRNA export via a specific RNA sequence element located within the 5′ end of the ORF [29, 30]. This sequence element is particularly prominent in transcripts encoding ER and mitochondrial-targeted proteins, the same functional class that is depleted of 5UIs. The current model is that when ALREX elements are present, their position relative to the first intron dictates the method of mRNA export. For transcripts lacking a 5UI, if an ALREX sequence is present at the 5′ end of the ORF, it is likely to be upstream of the first intron (which would be in the ORF). Thus, the ALREX pathway is used to export the mRNA from the nucleus. On the other hand, for transcripts containing a 5UI, the first intron is necessarily upstream of the ORF. These mRNAs are exported by the canonical TREX pathway regardless of whether an ALREX element is present in the ORF (Fig. 2). In support of this model, nucleotide sequences near the 5′ end of the ORF strongly

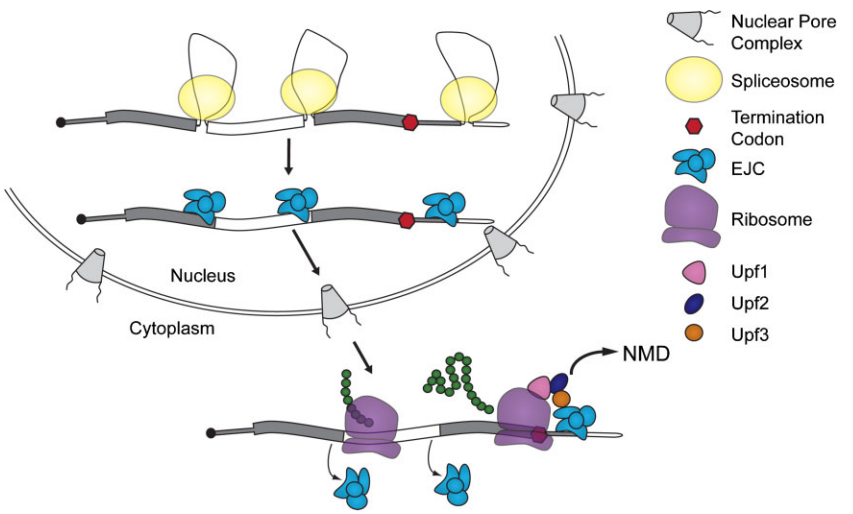


Figure 1. During splicing the EJC is deposited just upstream of splice junctions. Following translation termination, nonsense-mediated decay degrades transcripts harboring an EJC >55 nts downstream of a termination codon. Degradation occurs as a result of interactions between the terminating ribosome, Upf1, Upf2, Upf3, and the EJC.

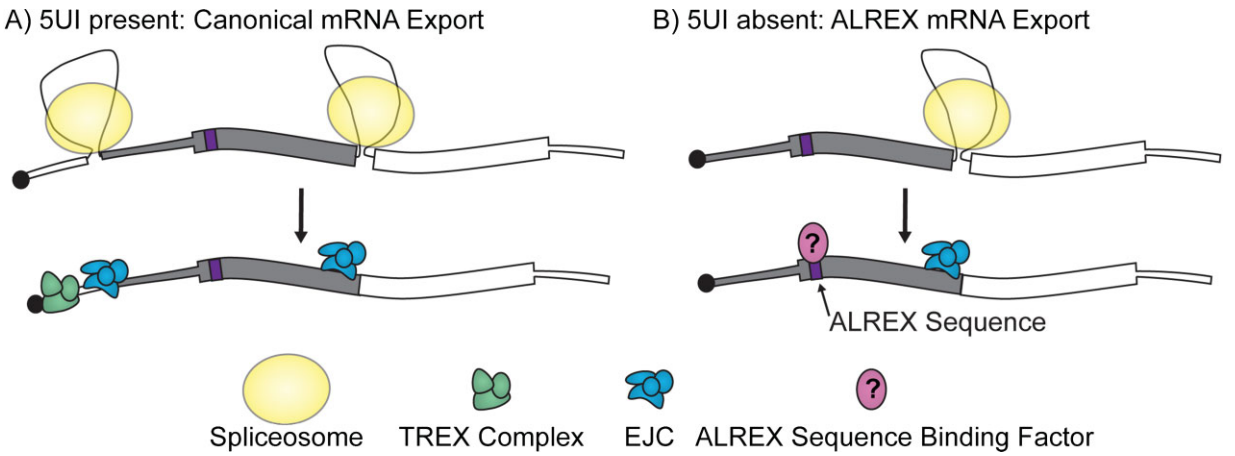


Figure 2. For mRNAs containing an ALREX sequence, position of this sequence relative to the first intron may determine route of mRNA export. **A**: If a 5UI is present, the first intron is upstream of the ALREX sequence, thus canonical TREX-dependent export is used. **B**: If a 5UI is absent, the first intron is downstream of the ALREX sequence, thus the ALREX mRNA export pathway is used.

correlate with 5UI status, and only sequences derived from 5UI-lacking transcripts can support mRNA export in the absence of splicing [30]. Although the ALREX sequence was first identified in ER- and mitochondrial-targeted genes, it can still function as a splicing-independent export element in other sequence contexts [30]. Therefore the ALREX pathway is likely relevant beyond ER and mitochondrial genes, and 5UI status is likely important for regulating export of additional classes of transcripts.

Why do alternate export pathways exist and how do they contribute to overall gene expression? Currently available data suggest that the selection of which nuclear export pathway to use can have downstream functional consequences. For example, a model mRNA exported by the TREX pathway is initially sequestered in stress granules, whereas an almost identical mRNA targeted to the ALREX pathway is not [29]. Thus mRNAs exported by the ALREX pathway may be more readily available for immediate translation under conditions of stress. Another intriguing possibility is that alternative promoter usage leading to inclusion or exclusion of a splicing event upstream of an ALREX element can allow switching between the two export pathways, thereby regulating subsequent mRNA expression. Expressed sequence tags (ESTs), cap analysis of gene expression (CAGE), and RNASeq data indicate that alternate promoter use is widespread in higher eukaryotes; 30–50% of all human and mouse genes have been reported to contain alternate promoters [31, 32]. To a smaller extent, 5UIs can be alternatively retained in the mature mRNA, rather than spliced out [33, 34]. It is currently unknown whether mRNAs containing ALREX elements are enriched for alternate promoter use or alternate intron retention, but this is clearly an interesting avenue of further research.

3′-UTR introns: A whole lot of nonsense?

In mammals, when an mRNA enters the cytoplasm and is translated, the nature of the translation termination event determines whether the transcript will persist and continue to produce protein, or become degraded by the NMD pathway [35]. NMD occurs when Upf1, which is bound to the terminating ribosome, interacts with Upf2, a peripheral EJC protein [36]. EJCs bound within ORFs are removed by the ribosome during translation, but an EJC downstream of the termination codon (i.e. in the 3′-UTR) should persist and stimulate NMD (Fig. 1) [15]. Thus, NMD strictly requires translation of the mRNA target. Current models suggest that the mRNA is translated only once before it is destroyed, producing a single molecule of protein per transcript [37]. Upf1 and Upf2 can also interact on a transcript and stimulate NMD when the transcript has a particularly long 3′-UTR [38]. This additional form of NMD is presumably splicing-independent, and is discussed elsewhere [7, 38].

NMD is perfectly suited to reduce the abundance of several classes of nonfunctional mRNAs. First, mutations or aberrant pre-mRNA splicing frequently introduce premature termination codons (PTCs) upstream of an EJC deposition site. NMD thus prevents production of potentially deleterious truncated proteins [39]. Second, NMD clearly serves to dampen the expression of nonfunctional transcripts arising from pseudogenes, expressed transposons, or integrated retroviruses, which frequently contain termination codons upstream of introns [40]. Finally, during programmed gene rearrangements in the T-cell receptor and immunoglobulin genes, unproductively rearranged alleles generate termination codons upstream of introns, and the resulting mRNAs are degraded by NMD [41].

Because transcripts containing 3UIs can be produced by mistake and because NMD degrades such transcripts, it is often assumed that all 3UI-containing transcripts are nonfunctional. Therefore, in an effort to accurately represent only functional mRNAs in the transcriptome, the NCBI Reference Sequence (RefSeq) database has suppressed the majority of 3UI-containing ''coding'' sequences (accession prefix NM_) and reassigned them ''noncoding'' accession numbers (accession prefix NR_ or XR_). As of April 2011, 846 human transcripts that were considered well-supported by RefSeq had been designated ''noncoding'' solely because they are predicted NMD substrates.

However, the designation of these transcripts as noncoding is not a trivial matter: mRNA prediction algorithms based on ESTs estimate that 35% of all human alternatively spliced isoforms contain 3UIs [42]. Are all of these mRNAs inconsequential for protein production? Or can an mRNA predicted to be destroyed shortly after translation still be considered real and functionally relevant?

Many 3UI-containing transcripts are both functional and protein coding

In addition to facilitating elimination of the nonfunctional mRNAs described above, significant evidence exists that some 3UIs serve to modulate normal gene expression [40, 43–49]. Three classes of 3UI-containing mRNAs are both functional and subject to NMD (Fig. 3). First are mRNAs with short ORFs in the 5′-UTR (upstream ORFs; uORFs). If uORF translation occurs prior to translation of the main ORF, exon junctions within the main ORF are effectively in the 3′-UTR of the uORF, and this can elicit NMD. In this way, translation of uORFs can serve to modulate mRNA levels and thus expression of the main ORF [40]. The second class consists of mRNAs in which a termination codon upstream of an exon junction is "intentionally" introduced by alternative splicing. Such a splicing event can lead to mRNA down-regulation through a process known as alternative-splicing linked to NMD (AS-NMD) [50]. The third class consists of mRNAs that are constitutively spliced within the 3′-UTR [51]. These transcripts are expected to be degraded every time they are translated, unless the NMD pathway is inhibited.

Proteins produced from such 3UI-containing transcripts have been detected in cells. In 2004, Hillman et al. [52] analyzed 1,363 human protein sequences deposited into the Swiss-Prot database and found 107 entries (7.9% of those analyzed) that were derived from transcripts that are apparently subject to NMD. More recently an analysis of mass spectrometry data from the Global Proteome Machine [53] and PeptideAtlas [54] repositories reached the same conclusion: 3UI-containing transcripts can indeed express protein [55]. These results suggest that either the peptides detected in the above proteomics studies are the products of a single round of translation [37], or some endogenous human NMD substrates can undergo multiple rounds of translation prior to being decayed. Consistent with the latter idea, NMD in budding yeast can occur during any round of translation, not just the first [56].

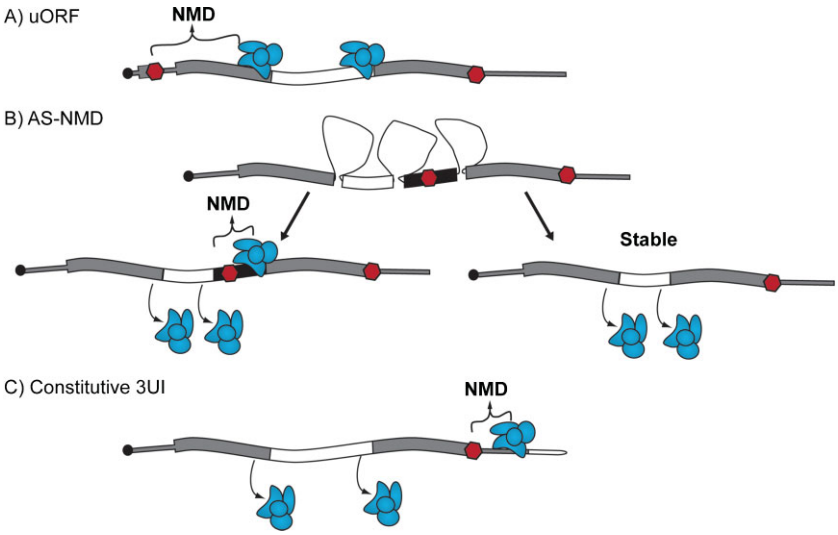


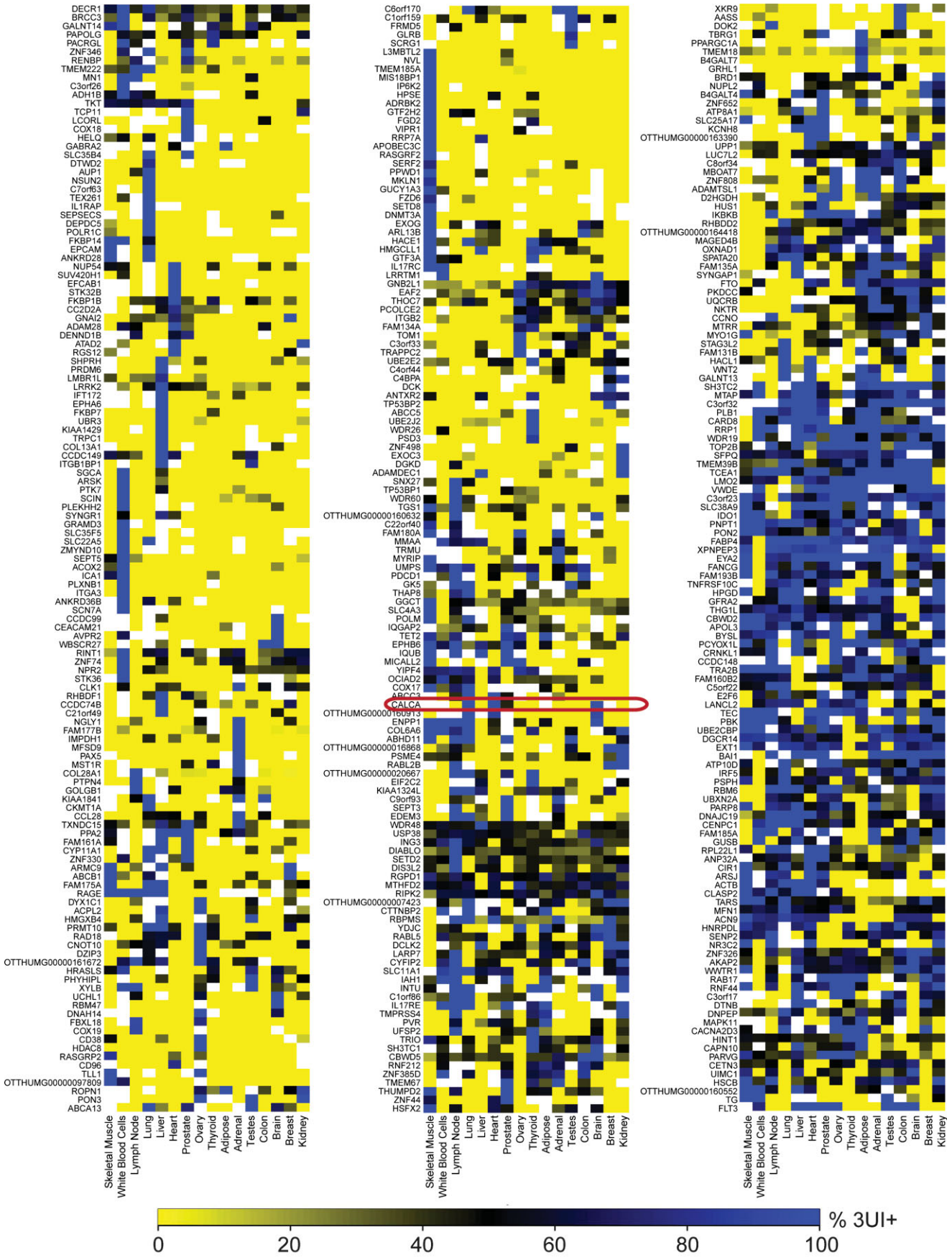
Figure 3. Classes of functional 3UI-containing transcripts. **A**: uORF – translation of the uORF terminates upstream of EJCs in the main ORF, thus eliciting NMD. **B**: AS-NMD – alternative splicing introduces a termination codon upstream of a splice junction. **C**: Constitutive 3UI – all splice forms of the transcript contain introns in the 3′-UTR.

Conservation and tissue-specific expression of 3UI genes

Given that mRNAs containing 3UIs are surprisingly common (an estimated 35% of alternatively spliced isoforms [42] plus those with uORFs and constitutive 3UIs; see above), the challenge is to distinguish functional 3UIs from those representing genomic noise or cellular errors [57]. Two potential indicators of function are conservation and tissue-specific expression. These concepts can be applied to all three classes of 3UI-containing transcripts described above (Fig. 3).

The first class of 3UI-containing transcripts is those with uORFs. Sequence analysis has revealed that approximately 35% of human genes harbor uORFs; of these, 38% are conserved among human, mouse, and rat [58]. Further, ribosome profiling has recently shown that 26% of all translationally active ORFs in mouse embryonic stem cells are actually uORFs [59]. Thus uORFs are prevalent, translated, and often conserved, indicating that many are functional. However, uORFs may have functions besides NMD [60, 61], so the presence of a conserved uORF does not necessarily mean an mRNA is regulated by NMD. For example, it is possible that some ribosomes translating a uORF either fail to recognize the uORF termination codon or reinitiate on the main ORF downstream [62]. Functional studies will be necessary to determine how many mRNAs are regulated via recruitment of NMD factors upon uORF termination.

The second class is AS-NMD transcripts. Among alternative splicing events conserved between human and mouse, approximately 21% introduce termination codons upstream of introns [57, 63]. This estimate, based on traditional transcriptomics, was recently re-examined by an in-depth analysis of the 309 protein coding genes within the ENCODE pilot phase regions


DECR1
BRCC3
GALNT14
PAPOLG
PACRGL
ZNF346
RENBP
TMEM222
MN1
C3orf26
ADH1B
TKT
TCP11
LCORL
COX18
HELQ
GABRA2
SLC35B4
DTWD2
AUP1
NSUN2
C7orf63
TEX261
IL1RAP
SEPSECS
DEPDC5
POLR1C
FKBP14
EPCAM
ANKRD28
NUP54
SUV420H1
EFCAB1
STK32B
FKBP1B
CC2D2A
GNAI2
ADAM28
DENND1B
ATAD2
RGS12
SHPRH
PRDM6
LMBR1L
LRRK2
IFT172
EPHA6
FKBP7
UBR3
KIAA1429
TRPC1
COL13A1
CCDC149
ITGB1BP1
SGCA
ARSK
PTK7
SCIN
PLEKHH2
SYNGR1
GRAMD3
SLC35F5
SLC22A5
ZMYND10
SEPT5
ACOX2
ICA1
PLXNB1
ITGA3
ANKRD36B
SCN7A
CCDC99
CEACAM21
AVPR2
WBSCR27
RINT1
ZNF74
NPR2
STK36
CLK1
RHBDF1
CCDC74B
C21orf49
NGLY1
FAM177B
IMPDH1
MFSD9
PAX5
MST1R
COL28A1
PTPN4
GOLGB1
KIAA1841
CKMT1A
CCL28
TXNDC15
PPA2
FAM161A
CYP11A1
ZNF330
ARMC9
ABCB1
FAM175A
RAGE
DYX1C1
ACPL2
HMGXB4
PRMT10
RAD18
CNOT10
DZIP3
OTTHUMG00000161672
HRASLS
PHYHIPL
XYLB
UCHL1
RBM47
DNAH14
FBXL18
COX19
CD38
HDAC8
RASGRP2
CD96
TLL1
OTTHUMG00000097809
ROPN1
PON3
ABCA13
C6orf170
C1orf159
FRMD5
GLRB
SCRG1
L3MBTL2
NVL
TMEM185A
MIS18BP1
IP6K2
HPSE
ADRBK2
GTF2H2
FGD2
VIPR1
RRP7A
APOBEC3C
RASGRF2
SERF2
PPWD1
MKLN1
GUCY1A3
FZD6
SETD8
DNMT3A
EXOG
ARL13B
HACE1
HMGCL1
GTF3A
IL17RC
LRRTM1
GNB2L1
EAF2
THOC7
PCOLCE2
ITGB2
FAM134A
TOM1
C3orf33
TRAPPC2
UBE2E2
C4orf44
C4BPA
DCK
ANTXR2
TP53BP2
ABCC5
UBE2J2
WDR26
PSD3
ZNF498
EXOC3
DGKD
ADAMDEC1
SNX27
TP53BP1
WDR60
TGS1
OTTHUMG00000160632
C22orf40
FAM180A
MMAA
TRMU
MYRIP
UMPS
PDCD1
GK5
THAP8
GGCT
SLC4A3
POLM
IQGAP2
TET2
EPHB6
IQUB
MICALL2
YIPF4
OCIAD2
COX17
ABCC3
CALCA
OTTHUMG00000160913
ENPP1
COL6A6
ABHD11
OTTHUMG00000016868
PSME4
RABL2B
OTTHUMG00000020667
EIF2C2
KIAA1324L
C9orf93
SEPT3
EDEM3
WDR48
USP38
ING3
DIABLO
SETD2
DIS3L2
RGPD1
MTHFD2
RIPK2
OTTHUMG00000007423
CTTNBP2
RBPMS
YDJC
RABL5
DCLK2
LARP7
CYFIP2
SLC11A1
IAH1
INTU
C1orf86
IL17RE
TMPRSS4
PVR
UFSP2
TRIO
SH3TC1
CBWD5
RNF212
ZNF385D
TMEM67
THUMPD2
ZNF44
HSFX2
XKR9
AASS
DOK2
TBRG1
PPARGC1A
TMEM18
B4GALT7
GRHL1
BRD1
NUPL2
B4GALT4
ZNF652
ATP8A1
SLC25A17
KCNH8
OTTHUMG00000163390
UPP1
LUC7L2
C8orf34
MBOAT7
ZNF808
ADAMTSL1
D2HGDH
HUS1
IKBKB
RHBDD2
OTTHUMG00000164418
MAGED4B
OXNAD1
SPATA20
FAM135A
SYNGAP1
FTO
PKDCC
UQCRB
NKTR
CCNO
MTRR
MYO1G
STAG3L2
FAM131B
HACL1
WNT2
GALNT13
SH3TC2
MTAP
C3orf32
PLB1
CARD8
RRP1
WDR19
TOP2B
SFPQ
TMEM39B
TCEA1
LMO2
VWDE
C3orf23
SLC38A9
IDO1
PNPT1
PCNP2
FABP4
XPNPEP3
EYA2
FANCG
FAM193B
TNFRSF10C
HPGD
GFRA2
THG1L
CBWD2
APOL3
BYSL
PCYOX1L
CRNKL1
CCDC148
TRA2B
FAM160B2
C5orf22
E2F6
LANCL2
TEC
PBK
UBE2CBP
DGCR14
EXT1
BAI1
ATP10D
IRF5
PSPH
RBM6
UBXN2A
PARP8
DNAJC19
CENPC1
FAM185A
GUSB
RPL22L1
ANP32A
CIR1
ARSJ
ACTB
CLASP2
TARS
MFN1
ACN9
HNRPDL
SENP2
NR3C2
ZNF326
AKAP2
WWTR1
RAB17
RNF44
C3orf17
DTNB
DNPEP
MAPK11
CACNA2D3
HINT1
CAPN10
PARVG
CETN3
UIMC1
HSCB
OTTHUMG00000160552
TG
FLT3
Skeletal Muscle
White Blood Cells
Lymph Node
Lung
Liver
Heart
Prostate
Ovary
Thyroid
Adipose
Adrenal
Testes
Colon
Brain
Breast
Kidney
% 3UI+
0
20
40
60
80
100

of the human genome [64]. By including next generation sequencing and RT-PCR data, that study produced an extremely reliable dataset of 162 conserved alternative splicing events. Of these, 27 (17%) introduce 3UIs. Thus, alternative splicing often leads to 3UI-containing transcripts that are conserved and therefore potentially functional.

To address the question of tissue-specificity of 3UI-containing transcripts, Pan et al. [57] undertook a comprehensive analysis of alternatively spliced 3UI-containing isoforms across ten different mouse tissues. Using exon and splice junction arrays to examine inclusion and exclusion of cassette exons, they found little evidence of tissue-specific expression of NMD isoforms. Nonetheless, they also concluded that NMD isoforms are expressed at low levels. Therefore, it is possible that the splicing arrays used in that study, though state-of-the-art at the time, did not provide enough sensitivity to detect expression level differences in rare 3UI-containing transcripts among tissues. In addition, splicing arrays can only measure the specific exon inclusion/exclusion events represented on the array, which are only a small percentage of total alternative splicing events occurring in cells [34]. The 2010 Illumina BodyMap deep sequencing project (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513 accessed on January 11, 2012) provided both enhanced sensitivity and a broader scope of alternate splicing events to address questions of tissue-specific splicing. Our analysis here of these data clearly shows tissue-specific expression of alternatively spliced 3UI-containing transcripts (Fig. 4). For example, CALCA mRNA (Fig. 4; outlined in red) contains six exons and can be alternatively processed to produce distinct mRNAs encoding calcitonin or calcitonin gene-related peptide. The mRNA containing all six exons harbors a stop codon in exon 5, 154 nt upstream of intron 5, and is therefore a potential NMD substrate. This 3UI-containing form is the only CALCA mRNA expressed in the human brain. By contrast, a non-NMD splice-form containing only exons 1–4 predominates in the thyroid. This high degree of variation strongly suggests that there is a functional consequence of including a 3UI in the brain, but not the thyroid. This possibility is consistent with studies showing that the protein encoded by the CALCA gene performs different functions in these two tissues [65].

The third class is constitutively spliced 3UI-containing transcripts. In the Illumina BodyMap deep sequencing data, we have identified 75 human genes for which the only detectable transcript(s) contain 3UIs more than 55 nts downstream of the translation termination codon, and these transcripts are expressed in a tissue-specific manner (Table 1). Previous analyses have identified 152 transcripts containing constitutive 3UIs that are conserved among human, rat, and mouse [51]. Interestingly, constitutive 3UI-containing transcripts that are conserved are particularly enriched in brain, testes, and hematopoietic cells [51]. We do not yet know the complete functional significance of these enrichments, but as described below, NMD is known to be involved in regulating developmental programs in both neurons and hematopoietic cells.

NMD inhibition points to functional 3UI-containing transcripts

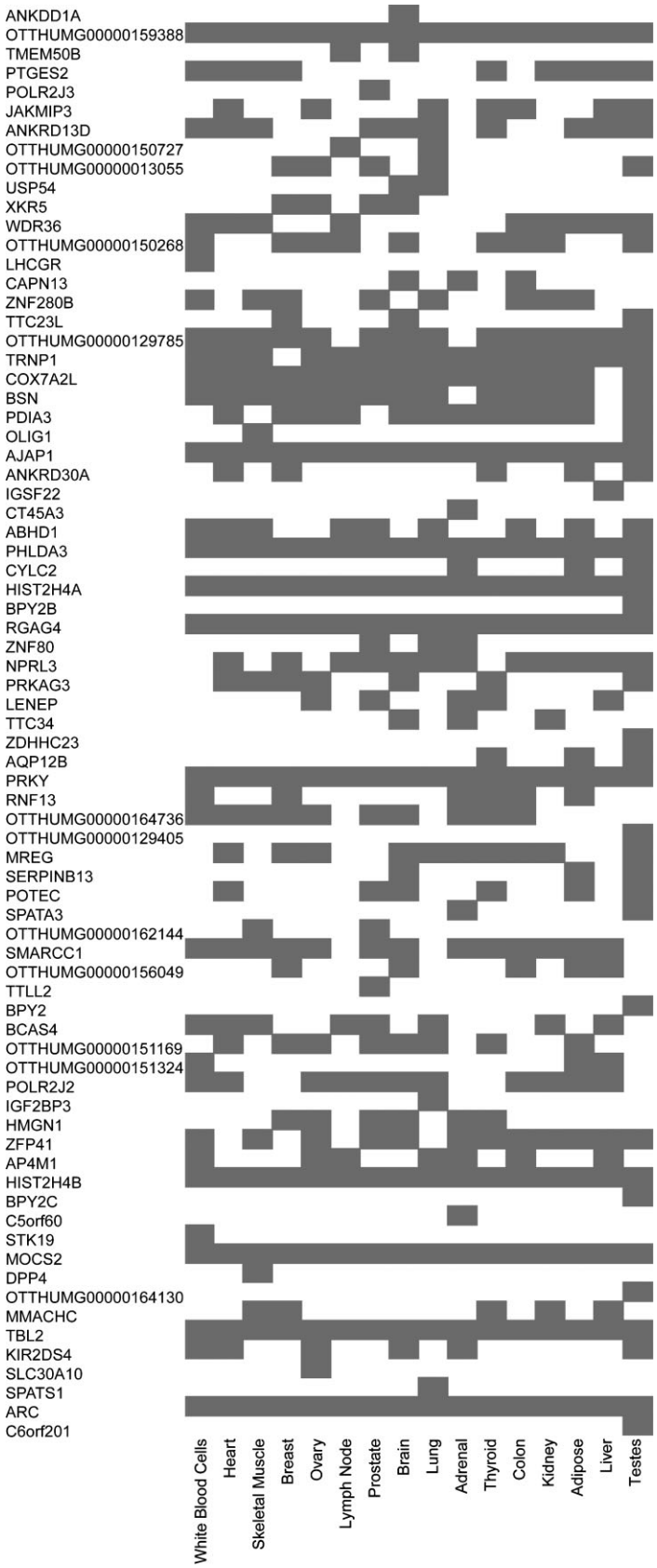
Numerous studies have taken the experimental approach of inhibiting NMD and examining the resulting changes in mRNA levels and splicing patterns [40, 43–49, 57, 66]. By combining these studies, close to 1,000 functional 3UI-containing transcripts have been found to exhibit either increased expression or alternate exon usage upon NMD inhibition. The number and identity of these mRNAs varies considerably by cell type, so the list will likely grow if more cell types are examined. Of course it is still uncertain from those studies which changes in expression and splicing upon NMD inhibition were direct, but NMD inhibition did affect a much higher proportion of 3UI-containing transcripts than non-3UI-containing transcripts [48].

The most dramatic functional enrichment among NMD-affected transcripts is for those encoding RNA-binding proteins [45, 66]. Within this class of genes, exons that introduce stop codons upstream of introns are extremely well-conserved [44, 67]. The introns to either side of these exons are also well-conserved, suggesting the presence of regulatory elements affecting their alternative splicing. This is because many RNA binding proteins use AS-NMD to regulate their own expression homeostatically. That is, as an RNA binding protein increases in abundance, it increasingly binds its own pre-mRNA and facilitates production of a 3UI-containing form subject to NMD, thereby maintaining protein homeostasis [44, 67, 68]. Proteins undergoing this type of regulation include ribosomal proteins [42, 69–71], core spliceosomal proteins [66, 72] and alternative splicing regulators such as hnRNP and SR proteins [45, 67]. In fact, every one of the 11 human SR proteins has a 3UI isoform and regulates its own production by AS-NMD [67].

Also enriched amongst NMD targets are mRNAs involved in development and differentiation [45, 46]. Consistent with this, NMD-deficient mice fail to

Figure 4. Transcripts with 3UIs are expressed in a tissue specific manner. The Illumina BodyMap data was downloaded from Ensembl and aligned to human genome (build hg18) using Tophat v.1.3.1 [83]. These alignments were used to assemble de novo transcriptomes using Cufflinks 1.3.0 [84, 85]. Vega transcript annotations were provided as a reference transcriptome during assembly. Cuffcompare was used to align the Cufflinks output to transcripts in the Vega database. We then selected genes with at least one Vega transcript where the stop codon is more than 55 nts upstream of a splice site. Using the expression level estimates from Cufflinks, we calculated the percentage of total transcripts that contained 3UIs. Only genes where this percentage differs by at least 75% points between at least two tissues are depicted here. When there is no detectable transcript for a given gene in a particular tissue, it is indicated with white color. Genes with detectably expressed transcripts in less than 12 out of 16 tissues are omitted. Red circle highlights the CALCA gene, an example that is discussed in detail in the main body of the article.

Table 1. Expression profiles of mRNAs with constitutive 3UIs from the Illumina BodyMap Project. Grey indicates the tissues where each 3UI-containing mRNA was detected.



complete embryogenesis, with the most prominent defects observed in heart, brain, and hematopoietic development [45, 47, 73, 74]. The neural development program incorporates NMD in at least two important ways, both with widespread effects. First, as neurons differentiate, they down-regulate expression of Ptbp1, an alternative splicing factor. One function of Ptbp1 is to alter splicing of transcripts encoding another splicing factor, Ptbp2, such that a 3UI-containing mRNA isoform is produced. Thus, as neurons differentiate and Ptbp1 levels decrease, the non 3UI-containing form of PTBP2 mRNA is produced. As a result, Ptbp2 protein levels increase, leading to a regulated change in alternative splicing of its many targets, many of which are involved in neuronal differentiation [75]. Some of these Ptbp2 splicing targets are themselves NMD substrates. For example, PSD-95 (also known as DLG4, SAP90), encoding a critical protein in synaptic function, is transcribed throughout neuronal development. However, during early development it is spliced to include a 3UI and the mRNA is degraded. Psd-95 protein is not detected until later in development, when the splicing pattern shifts in favor of the non 3UI-containing form [76].

NMD is additionally linked to the neural differentiation program through miR-128, a microRNA with enriched expression in the brain. Production of miR-128 increases during neural development and one direct target of miR-128 is UPF1 mRNA. Thus both Upf1 levels and NMD efficiency decrease as neurons differentiate. As a result, hundreds of transcripts that would otherwise be subject to NMD are up-regulated, and most of these encode proteins important for neural function [46].

3UI-containing transcripts are also enriched for mRNAs involved in amino acid metabolism, starvation, ER stress, and hypoxia [40, 47, 49]. These enrichments led to the discovery that NMD is inhibited during stress through a mechanism involving phosphorylation of translation initiation factor, eIF2α. As a result, 3UI-containing stress-response transcripts are stabilized, thus promoting stress survival [49].

The examples above highlight several mechanistically distinct ways that 3UIs can be integrated into gene

regulatory pathways. Other possibilities also exist, including regulation of EJC deposition, translational regulation, and regulation of other components of the NMD machinery. All such mechanisms are theoretically subject to cell-type, condition-specific, and/or transcript-dependent control. Therefore, if we are to appreciate the full spectrum of these possibilities, researchers need to become more cognizant of the many protein-coding transcripts that contain 3UIs.

Identifying UTR introns

As detailed above, a wide range of evidence now indicates that whether or not UTR introns are present can significantly affect gene expression. It is therefore important that researchers be able to identify such introns so that they may study their impact on a particular pathway or gene of interest. For this purpose, a wide variety of sources is available, including RefSeq, Vega, AceView, UCSC Known Genes, H-Invitational, and ENCODE [9, 77–81]. Each of these is different in terms of how it approaches the balance between specificity and sensitivity – that is, how it limits its annotations to mRNAs that are functionally relevant (specificity), while at the same time ensuring that as many real mRNAs as possible are represented (sensitivity) [82].

RefSeq achieves high specificity – mRNAs represented in RefSeq have a very high likelihood of being real. Therefore, for 5UI identification, RefSeq is a good starting point. Intronic regions and UTR boundaries are generally well annotated and we have generated a list of all human RefSeq mRNAs with 5UIs, which is publically available [3]. However, RefSeq has relatively lower sensitivity and lacks many real transcripts. In particular, transcripts that are scarce or only expressed in a specific cellular context may not be supported by enough sequence-based evidence to be included in RefSeq. Our analysis thus far has found no specific bias or trend toward inclusion or exclusion of 5UIs in RefSeq. However, in the case of alternate promoter usage, which could lead to inclusion or exclusion of a 5UI, one or more alternate forms could be missing. Therefore, because different groups use different data sources and different methodologies, it is worthwhile to query multiple annotations (listed above) in search of all potential 5′-UTRs for a given transcript.

Because NMD substrates are inherently unstable, sequence evidence for them can be particularly limited. Therefore, sensitivity is even more an issue for 3UI transcripts and multiple data sources should always be used to search for them. In addition to the annotation projects listed above, most primary data from deep sequencing studies are publically available and can be accessed directly. This is an extremely powerful way to look for transcripts that are not yet in the composite databases, but can require more in-depth bioinformatics expertise and effort on the part of the user. Of particular use are sequencing data from experimental conditions that are enriched for 3UI transcripts. Experiments listed above that knockdown or inhibit the NMD machinery are a good source. In addition, the ENCODE project now includes RNAseq data from human nuclear RNA, which is publically available (http://genome.ucsc.edu/ENCODE/). We expect these data to be enriched for 3UI transcripts, as NMD occurs exclusively in the cytoplasm.

In addition to their under-representation due to scarcity, 3UI transcripts are also actively suppressed from some annotation pipelines because they are considered nonfunctional. As discussed above, RefSeq's current policy is to designate most 3UI-containing mRNAs as ''noncoding'' even when a transcript with protein-coding potential is well-supported. The only exceptions in RefSeq are genes for which all available transcripts exhibit 3UIs or those few that have been experimentally verified to produce protein [9]. In genome browsers, ORFs are not displayed for RNAs designated as ''noncoding.'' This may lead researchers into falsely believing that 3UI transcripts contain no ORF and have no protein coding potential. Therefore, when using RefSeq to identify 3UIs, noncoding transcripts must be carefully examined. If these have been designated as noncoding due to predicted NMD targeting, this will be annotated as a ''misc_feature'' on the transcript record.

For identification of 3UIs in human mRNAs, the Vega genome browser, which displays annotations from the Human and Vertebrate Analysis and Annotation (HAVANA) Group at Wellcome Trust Sanger Institute [77] is currently the best place to start. Like RefSeq, the HAVANA group uses a manual curation process, and its annotations therefore have relatively good specificity. However, rather than designating them noncoding, HAVANA includes 3UI-containing transcripts in its coding sequence database, and simply flags them with the NMD biotype.

Conclusions

Introns in both 5′- and 3′-UTRs influence gene expression in ways that are different from introns in coding regions. Within the context of the 5′-UTR, presence or absence of an intron can dictate the mechanism of mRNA export. The export pathway used might depend on alternate promoter usage and could influence gene expression on several levels, including subcellular localization and translational control. In the 3′-UTR, introns can target the mRNA for degradation by NMD. While we have long appreciated the importance of NMD in quality control, only more recently have we begun to understand that NMD also regulates normal gene expression through functional 3UIs. Therefore, it is now time to change the default assumption that 3UI-containing transcripts are non-coding or nonfunctional. Like miRNA binding sites and 5′-UTR structure, introns in UTRs should be regarded as important cis-regulatory elements that modulate multiple levels of gene expression.

References

1. **Le Hir H**, **Nott A**, **Moore MJ.** 2003. How introns influence and enhance eukaryotic gene expression. *Trends Biochem Sci* **28**: 215–20.
2. **Moore MJ**, **Proudfoot NJ.** 2009. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136**: 688–700.
3. **Cenik C**, **Derti A**, **Mellor JC**, **Berriz GF**, et al. 2010. Genome-wide functional analysis of human 5′ untranslated region introns. *Genome Biol* **11**: R29.
4. **Zhang J**, **Sun X**, **Qian Y**, **Maquat LE.** 1998. Intron function in the nonsense-mediated decay of beta-globin mRNA: indications that

pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm. *RNA* **4**: 801–15.
5. **Nagy E**, **Maquat LE.** 1998. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci* **23**: 198–9.
6. **Lejeune F**, **Maquat LE.** 2005. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr Opin Cell Biol* **17**: 309–15.
7. **Nicholson P**, **Yepiskoposyan H**, **Metze S**, **Zamudio Orozco R**, et al. 2010. Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors. *Cell Mol Life Sci* **67**: 677–700.
8. **Chang Y-F**, **Imam JS**, **Wilkinson MF.** 2007. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76**: 51–74.
9. **Pruitt KD**, **Tatusova T**, **Klimke W**, **Maglott DR.** 2009. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37**: D32–6.
10. **Le Hir H**, **Moore MJ**, **Maquat LE.** 2000. Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon-exon junctions. *Genes Dev* **14**: 1098–108.
11. **Le Hir H**, **Izaurralde E**, **Maquat LE**, **Moore MJ.** 2000. The spliceosome deposits multiple proteins 20–24 nucleotides upstream of mRNA exon-exon junctions. *EMBO J* **19**: 6860–9.
12. **Dostie J**, **Dreyfuss G.** 2002. Translation is required to remove Y14 from mRNAs in the cytoplasm. *Curr Biol* **12**: 1060–7.
13. **Tange TÄ**, **Nott A**, **Moore MJ.** 2004. The ever-increasing complexities of the exon junction complex. *Curr Opin Cell Biol* **16**: 279–84.
14. **Bono F**, **Gehring NH.** 2011. Assembly, disassembly and recycling: the dynamics of exon junction complexes. *RNA Biol* **8**: 24–9.
15. **Le Hir H**, **Gatfield D**, **Izaurralde E**, **Moore MJ.** 2001. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J* **20**: 4987–97.
16. **Masuda S**, **Das R**, **Cheng H**, **Hurt E**, et al. 2005. Recruitment of the human TREX complex to mRNA during splicing. *Genes Dev* **19**: 1512–7.
17. **Cheng H**, **Dufu K**, **Lee C-S**, **Hsu JL**, et al. 2006. Human mRNA export machinery recruited to the 5′ end of mRNA. *Cell* **127**: 1389–400.
18. **Strässer K**, **Masuda S**, **Mason P**, **Pfannstiel J**, et al. 2002. TREX is a conserved complex coupling transcription with messenger RNA export. *Nature* **417**: 304–8.
19. **Köhler A**, **Hurt E.** 2007. Exporting RNA from the nucleus to the cytoplasm. *Nat Rev Mol Cell Biol* **8**: 761–73.
20. **Long JC**, **Cáceres JF.** 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417**: 15–27.
21. **Zhong X-Y**, **Wang P**, **Han J**, **Rosenfeld MG**, et al. 2009. SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. *Mol Cell* **35**: 1–10.
22. **Huang Y**, **Steitz JA.** 2005. SRprises along a messenger's journey. *Mol Cell* **17**: 613–5.
23. **Hong X**, **Scofield DG**, **Lynch M.** 2006. Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol* **23**: 2392–404.
24. **Sephton CF**, **Cenik C**, **Kucukural A**, **Dammer EB**, et al. 2011. Identification of neuronal RNA targets of TDP-43-containing ribonucleoprotein complexes. *J Biol Chem* **286**: 1204–15.
25. **Furger A**, **O'Sullivan JM**, **Binnie A**, **Lee BA**, et al. 2002. Promoter proximal splice sites enhance transcription. *Genes Dev* **16**: 2792–9.
26. **Majewski J**, **Ott J.** 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res* **12**: 1827–36.
27. **Matsumoto K**, **Wassarman KM**, **Wolffe AP.** 1998. Nuclear history of a pre-mRNA determines the translational activity of cytoplasmic mRNA. *EMBO J* **17**: 2107–21.
28. **Nott A**, **Meislin SH**, **Moore MJ.** 2003. A quantitative analysis of intron effects on mammalian gene expression. *RNA* **9**: 607–17.
29. **Palazzo AF**, **Springer M**, **Shibata Y**, **Lee C-S**, et al. 2007. The signal sequence coding region promotes nuclear export of mRNA. *PLoS Biol* **5**: e322.
30. **Cenik C**, **Chua HN**, **Zhang H**, **Tarnawsky SP**, et al. 2011. Genome analysis reveals interplay between 5′UTR introns and nuclear mRNA export for secretory and mitochondrial genes. *PLoS Genet* **7**: e1001366.
31. **Carninci P**, **Sandelin A**, **Lenhard B**, **Katayama S**, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**: 626–35.
32. **Davuluri RV**, **Suzuki Y**, **Sugano S**, **Plass C**, et al. 2008. The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet* **24**: 167–77.
33. **Galante PAF**, **Sakabe NJ**, **Kirschbaum-Slager N**, **de Souza SJ.** 2004. Detection and evaluation of intron retention events in the human transcriptome. *RNA* **10**: 757–65.
34. **Wang ET**, **Sandberg R**, **Luo S**, **Khrebtukova I**, et al. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–6.
35. **Mühlemann O**, **Eberle AB**, **Stalder L**, **Zamudio Orozco R.** 2008. Recognition and elimination of nonsense mRNA. *Biochim Biophys Acta* **1779**: 538–49.
36. **Kashima I**, **Yamashita A**, **Izumi N**, **Kataoka N**, et al. 2006. Binding of a novel SMG-1-Upf1-eRF1-eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay. *Genes Dev* **20**: 355–67.
37. **Ishigaki Y**, **Li X**, **Serin G**, **Maquat LE.** 2001. Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20. *Cell* **106**: 607–17.
38. **Singh G**, **Rebbapragada I**, **Lykke-Andersen J.** 2008. A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay. *PLoS Biol* **6**: e111.
39. **Isken O**, **Maquat LE.** 2007. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev* **21**: 1833–56.
40. **Mendell JT**, **Sharifi NA**, **Meyers JL**, **Martinez-Murillo F**, et al. 2004. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat Genet* **36**: 1073–8.
41. **Wang J**, **Vock VM**, **Li S**, **Olivas OR**, et al. 2002. A quality control pathway that down-regulates aberrant T-cell receptor (TCR) transcripts by a mechanism requiring UPF2 and translation. *J Biol Chem* **277**: 18489–93.
42. **Lewis BP**, **Green RE**, **Brenner SE.** 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci USA* **100**: 189–92.
43. **Wittmann J**, **Hol EM**, **Jäck H-M.** 2006. hUPF2 silencing identifies physiologic substrates of mammalian nonsense-mediated mRNA decay. *Mol Cell Biol* **26**: 1272–87.
44. **Ni JZ**, **Grate L**, **Donohue JP**, **Preston C**, et al. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev* **21**: 708–18.
45. **McIlwain DR**, **Pan Q**, **Reilly PT**, **Elia AJ**, et al. 2010. Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay. *Proc Natl Acad Sci USA* **107**: 12186–91.
46. **Bruno IG**, **Karam R**, **Huang L**, **Bhardwaj A**, et al. 2011. Identification of a microRNA that activates gene expression by repressing nonsense-mediated RNA decay. *Mol Cell* **42**: 500–10.
47. **Weischenfeldt J**, **Damgaard I**, **Bryder D**, **Theilgaard-Mönch K**, et al. 2008. NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. *Genes Dev* **22**: 1381–96.
48. **Yepiskoposyan H**, **Aeschimann F**, **Nilsson D**, **Okoniewski M**, et al. 2011. Autoregulation of the nonsense-mediated mRNA decay pathway in human cells. *RNA* **17**: 2108–18.
49. **Wang D**, **Zavadil J**, **Martin L**, **Parisi F**, et al. 2011. Inhibition of nonsense-mediated RNA decay by the tumor microenvironment promotes tumorigenesis. *Mol Cell Biol* **31**: 3670–80.
50. **McGlincy NJ**, **Smith CWJ.** 2008. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem Sci* **33**: 385–93.
51. **Giorgi C**, **Yeo GW**, **Stone ME**, **Katz DB**, et al. 2007. The EJC factor eIF4AIII modulates synaptic strength and neuronal protein expression. *Cell* **130**: 179–91.
52. **Hillman RT**, **Green RE**, **Brenner SE.** 2004. An unappreciated role for RNA surveillance. *Genome Biol* **5**: R8.
53. **Craig R**, **Cortens JP**, **Beavis RC.** 2004. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* **3**: 1234–42.
54. **Desiere F**, **Deutsch EW**, **King NL**, **Nesvizhskii AI**, et al. 2006. The PeptideAtlas project. *Nucleic Acids Res* **34**: D655–8.
55. **Ezkurdia I**, **Del Pozo A**, **Frankish A**, **Rodriguez JM**, et al. 2012. Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Mol Biol Evol* **29**: 2265–83.
56. **Johansson MJO**, **He F**, **Spatrick P**, **Li C**, et al. 2007. Association of yeast Upf1p with direct substrates of the NMD pathway. *Proc Natl Acad Sci USA* **104**: 20872–7.
57. **Pan Q**, **Saltzman AL**, **Kim YK**, **Misquitta C**, et al. 2006. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev* **20**: 153–8.
58. **Iacono M**, **Mignone F**, **Pesole G.** 2005. uAUG and uORFs in human and rodent 5′ untranslated mRNAs. *Gene* **349**: 97–105.

59. **Ingolia NT**, **Lareau LF**, **Weissman JS.** 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**: 789–802.
60. **Vilela C**, **McCarthy JEG.** 2003. Regulation of fungal gene expression via short open reading frames in the mRNA 5′untranslated region. *Mol Microbiol* **49**: 859–67.
61. **Wethmar K**, **Smink JJ**, **Leutz A.** 2010. Upstream open reading frames: molecular switches in (patho)physiology. *BioEssays* **32**: 885–93.
62. **Powell ML**, **Brown TDK**, **Brierley I.** 2008. Translational termination-re-initiation in viral systems. *Biochem Soc Trans* **36**: 717–22.
63. **Baek D**, **Green P.** 2005. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc Natl Acad Sci USA* **102**: 12813–8.
64. **Mudge JM**, **Frankish A**, **Fernandez-Banet J**, **Alioto T**, et al. 2011. The origins, evolution and functional potential of alternative splicing in vertebrates. *Mol Biol Evol* **28**: 2949–59.
65. **Rosenfeld MG**, **Amara SG**, **Evans RM.** 1984. Alternative RNA processing: determining neuronal phenotype. *Science* **225**: 1315–20.
66. **Saltzman AL**, **Pan Q**, **Blencowe BJ.** 2011. Regulation of alternative splicing by the core spliceosomal machinery. *Genes Dev* **25**: 373–84.
67. **Lareau LF**, **Inada M**, **Green RE**, **Wengrod JC**, et al. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446**: 926–9.
68. **Sureau A**, **Gattoni R**, **Dooghe Y**, **Stévenin J**, et al. 2001. SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. *EMBO J* **20**: 1785–96.
69. **Mitrovich QM**, **Anderson P.** 2000. Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in *C. elegans*. *Genes Dev* **14**: 2173–84.
70. **de Lima Morais DA**, **Harrison PM.** 2010. Large-scale evidence for conservation of NMD candidature across mammals. *PLoS ONE* **5**: e11695.
71. **Cuccurese M**, **Russo G**, **Russo A**, **Pietropaolo C.** 2005. Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression. *Nucleic Acids Res* **33**: 5965–77.
72. **Saltzman AL**, **Kim YK**, **Pan Q**, **Fagnani MM**, et al. 2008. Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. *Mol Cell Biol* **28**: 4320–30.
73. **Medghalchi SM**, **Frischmeyer PA**, **Mendell JT**, **Kelly AG**, et al. 2001. Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability. *Hum Mol Genet* **10**: 99–105.
74. **Frischmeyer-Guerrerio PA**, **Montgomery RA**, **Warren DS**, **Cooke SK**, et al. 2011. Perturbation of thymocyte development in nonsense-mediated decay (NMD)-deficient mice. *Proc Natl Acad Sci USA* **108**: 10638–43.
75. **Boutz PL**, **Stoilov P**, **Li Q**, **Lin C-H**, et al. 2007. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev* **21**: 1636–52.
76. **Zheng S**, **Gray EE**, **Chawla G**, **Porse BT**, et al. 2012. PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. *Nat Neurosci* **15**: 381–8, S1.
77. **Wilming LG**, **Gilbert JGR**, **Howe K**, **Trevanion S**, et al. 2008. The vertebrate genome annotation (Vega) database. *Nucleic Acids Res* **36**: D753–60.
78. **Thierry-Mieg D**, **Thierry-Mieg J.** 2006. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* **7**: S12.1–4.
79. **Hsu F**, **Kent WJ**, **Clawson H**, **Kuhn RM**, et al. 2006. The UCSC known genes. *Bioinformatics* **22**: 1036–46.
80. **Genome Information Integration Project And H-Invitational 2**, **Yamasaki C**, **Murakami K**, **Fujii Y**, et al. 2008. The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res* **36**: D793–9.
81. **ENCODE Project Consortium**. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**: e1001046.
82. **Guigó R**, **Flicek P**, **Abril JF**, **Reymond A**, et al. 2006. EGASP: the human ENCODE genome annotation assessment project. *Genome Biol* **7**: S2.1–31.
83. **Trapnell C**, **Pachter L**, **Salzberg SL.** 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–11.
84. **Trapnell C**, **Williams BA**, **Pertea G**, **Mortazavi A**, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–5.
85. **Roberts A**, **Pimentel H**, **Trapnell C**, **Pachter L.** 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**: 2325–9.