Research

An unappreciated role for RNA surveillance

R Tyler Hillman**,†, Richard E Green∗†‡ and Steven E Brenner∗†‡

Addresses: †Department of Bioengineering, University of California, Berkeley, CA 94720-3102, USA. ‡Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720-3102, USA. ∗Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720-3102, USA.

∗ These authors contributed equally to this work.

Correspondence: Steven E Brenner. E-mail: brenner@compbio.berkeley.edu

Published: 2 February 2004

Received: 3 November 2003

Revised: 5 December 2003

Accepted: 2 January 2004

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/2/R8

© 2004 Hillman et al.; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article’s original URL.

Abstract

Background: Nonsense-mediated mRNA decay (NMD) is a eukaryotic mRNA surveillance mechanism that detects and degrades mRNAs with premature termination codons (PTC* mRNAs). In mammals, a termination codon is recognized as premature if it lies more than about 50 nucleotides upstream of the final intron position. More than a third of reliably inferred alternative splicing events in humans have been shown to result in PTC* mRNA isoforms. As the mechanistic details of NMD have only recently been elucidated, we hypothesized that many PTC* isoforms may have been cloned, characterized and deposited in the public databases, even though they would be targeted for degradation in vivo.

Results: We analyzed the human alternative protein isoforms described in the SWISS-PROT database and found that 144 (5.8% of 2,483) isoform sequences amenable to analysis, from 107 (7.9% of 1,363) SWISS-PROT entries, derive from PTC* mRNA.

Conclusions: For several of the PTC* isoforms we identified, existing experimental evidence can be reinterpreted and is consistent with the action of NMD to degrade the transcripts. Several genes with mRNA isoforms that we identified as PTC* - calpain-10, the CDC-like kinases (CLKs) and LARD - show how previous experimental results may be understood in light of NMD.

Background

Alternative pre-mRNA splicing endows genes with the potential to produce a menagerie of protein products. After pre-mRNA is transcribed, a complex system of regulation determines which one of several possible versions of mature mRNA will be produced (reviewed in [1]). Alternative splicing is particularly important in human gene expression, as it affects half or more of human genes [2,3]. The diversity-generating capacity of alternative splicing can be staggering: one notable example, the dscam gene of Drosophila melanogaster, is hypothetically capable of producing 38,016 unique alternative isoforms [4]. However, functional roles for most alternative isoforms remain undiscovered.

It has been known for more than a decade that nonsense and frameshift mutations that induce premature termination codons can destabilize mRNA transcripts in vivo [5,6]. First investigated in yeast and humans, NMD was subsequently observed in a wide range of eukaryotes and is now thought to occur in all eukaryotes [7]. How cells manage to distinguish a
premature termination codon from a normal termination codon has been the subject of intense investigation. Important details have emerged that establish the following mechanistic framework model for NMD in mammals (Figure 1).

During pre-mRNA processing, the spliceosome removes intron sequences. As this occurs, a set of proteins called the exon-junction complex is deposited 20-24 nucleotides upstream of the sites of intron removal [8-11]. The components of this complex serve the dual roles of facilitating export of the mature mRNA to the cytoplasm and remembering the gene structure [12]. According to the current model, as a ribosome traverses the mRNA in its first pioneering round of translation, it displaces all exon-junction complexes in its path until it reaches a stop codon. If the termination codon is on or near the final exon, as is the case for most genes, the ribosome will have displaced all exon-junction complexes. The mRNA will then undergo multiple rounds of translation. (b) If the termination codon is sufficiently far upstream of the final exon, the ribosome will not have displaced all exon-junction complexes. Interactions ensue that result in the degradation of the mRNA by NMD.

Degradation of PTC+ mRNAs is generally thought to occur as a quality-surveillance system - preempting translation of potentially dominant-negative, carboxy-terminal truncated proteins [20]. PTC+ transcripts are aberrantly produced in several ways. The somatic recombination that underlies immune-system diversity frequently generates recombined genes whose transcripts contain a PTC [21]. Inefficient or faulty splicing will often generate a frameshift in the resulting mRNA, inducing a PTC to come into frame. Also, the high processivity of RNA polymerase yields a relatively high error rate, 1 in 10,000 bases [22,23], commonly introducing premature stops. DNA mutations are a source of potentially heritable PTCs. It is estimated that 30% of inherited disorders in humans are caused by a PTC [24]. The numerous diseases whose pathogenesis has been linked to NMD-inducing PTC mutations include aniridia due to the PAX6 gene [25], Duchenne muscular dystrophy due to the dystrophin gene [26], and Marfan syndrome due to the FBN1 gene [27]. In addition to its quality-control role in degrading aberrantly produced PTC+ mRNAs, NMD has also been shown
experimentally to act on a handful of wild-type PTC⁺ mRNAs [28-35]. In Caenorhabditis elegans, for example, expression of the ribosomal proteins L3, L7a, L10a and L12 and the SR proteins SRp20 and SRp30b are regulated posttranscriptionally via the coupling of alternative splicing and NMD [31,32]. In each case, productive isoforms were shown to be produced in vivo, as well as unproductive isoforms with a PTC. Regulated splicing to generate the unproductive isoforms is used as a means of downregulating protein expression, as these mRNA isoforms are degraded by NMD rather than translated to make protein. This system, which we have termed regulated unproductive splicing and translation (RUST), is also used in humans [28-30]. For example, the SR protein SC35 has been shown to autoregulate its own expression using RUST [29]. When levels of SC35 protein are elevated, SC35 binds its own pre-mRNA, inducing the production of PTC⁺ SC35 mRNA. The PTC⁺ SC35 mRNA is destabilized by NMD, resulting in lower levels of SC35 protein. A similar autoregulatory RUST system was also recently discovered to control production of polypyrimidine tract binding protein (PTB) [35].

In a previous study, we found that 35% of human mRNA alternative isoforms reliably inferred from expressed sequence tags (ESTs) are PTC⁺, rendering them apparent targets of NMD (see [36] and a conference report at [37]). Therefore, many wild-type alternative mRNA isoforms may not be translated into functional protein, but instead are targeted for degradation by NMD. The vast majority of PTC⁺ isoforms identified in that study represent previously unrecognized potential targets of NMD. However, EST databases contain expressed sequence for many isoforms that are otherwise uncharacterized. Therefore, it was not obvious how many of the isoforms identified in that study as PTC⁺ were functionally relevant or even previously known. It was also not obvious to what extent those PTC⁺ isoforms represented instances of RUST regulation or simply errors or deregulation in premRNA processing. Regardless, it is clear that NMD has a vital role in regulating mammalian gene expression, as inhibition of NMD is embryonic lethal for mouse [38].

To understand the biological significance of PTC⁺ isoforms and the prevalence of NMD on wild-type transcripts, it is necessary to expand beyond existing isolated RUST examples, while retaining a focus on functionally characterized genes. For this reason, we analyzed the human alternative isoforms described in the SWISS-PROT database. Common routes for gene isoform sequences to be determined and entered into databases include the cloning of intronless mini-genes and the sequencing of unexpected PCR bands. By either method, gene structure cannot be directly observed, and therefore PTCs may be overlooked. Further computational and experimental analyses will also often be oblivious to these features. Because the cloning and characterization of many isoforms predates our current understanding of NMD action, we hypothesized that unrecognized potential targets of NMD may be present even in curated databases like SWISS-PROT. We found that many of these alternative protein isoforms derive from PTC⁺ mRNAs. This is particularly surprising as SWISS-PROT is a heavily curated database of expressed protein sequences. According to the current NMD model, these PTC⁺ mRNAs should be degraded, and therefore the protein isoforms should not be expressed at high abundance. To resolve this apparent conflict, we examined existing experimental evidence and found that, in several cases, results described in the scientific literature are readily explained by NMD action.

Results and discussion

SWISS-PROT protein isoforms from PTC⁺ mRNAs

We examined the human alternative isoforms described in the SWISS-PROT database [39] to determine if any derive from PTC⁺ mRNA (see Materials and methods). For each alternative human protein isoform sequence in SWISS-PROT, we attempted to identify a corresponding cDNA/mRNA sequence in the SWISS-PROT database [39] to determine if any derive from PTC⁺ mRNA (see Materials and methods). For each alternative human protein isoform sequence in SWISS-PROT, we attempted to identify a corresponding cDNA/mRNA sequence. We found that many of these alternative isoforms appear to be expressed at high abundance. To resolve this apparent conflict, we examined existing experimental evidence and found that, in several cases, results described in the scientific literature are readily explained by NMD action.
Assemble alternative human isoform protein sequences

Assembled 4,556 isoforms, representing 1,636 SWISS-PROT genes

Mapped 2,742 isoforms, representing 1,463 genes, to mRNAs and coding sequences

Match SWISS-PROT isoforms to conceptually translated protein sequences in RefSeq or GenPept database (≥99% identity and no gaps)

Genomic sequence

Human genome

144 isoforms (5.8% of 2,483), spread across 107 unique SWISS-PROT entries (7.9% of 1,363), contain premature termination codons (PTCs)

PTC+? If so, NMD candidate

Manual inspection

Figure 2 (see legend on next page)
Calpain-10

Calpain-10 is an ubiquitously expressed protease that is alternatively spliced to produce eight mRNA isoforms [43], found in SWISS-PROT as Q9HC93. The gene for calpain-10 has been intensively studied because a polymorphism in its third intron, UC SNP-43, has been linked to type II diabetes in several populations. Because this polymorphism lies in intronic sequence it does not directly affect the coding potential of any isoform of calpain-10. It was shown that homozgyosity of UC SNP-43 leads to reduced levels of total calpain-10 transcript and is coincident with insulin resistance in skeletal muscle [44]. Previous investigations into how this polymorphism affects transcript abundance have centered on transcriptional regulation [43,45]. In an expression study, Horikawa et al. found four of the eight isoforms to be "less abundant". In our SWISS-PROT survey we found these same four mRNA isoforms to be PTC+, suggesting that NMD may be responsible for this experimental observation (Figure 4). This introduces the possibility that UC SNP-43 may affect the regulation of calpain-10 alternative splicing, favoring production of one or more of the PTC+ isoforms.

CDC-like kinases CLK1, CLK2 and CLK3

CLK1, CLK2 and CLK3 - three members of the CDC-like kinase family (also known as LAMMER kinases and STY kinases; SWISS-PROT entries P49759, P49760, P49761) - were found to have at least one PTC+ splice variant. CLKs are thought to be high-level regulators of alternative splicing, as CLK1 has been shown to activate a set of SR proteins by phosphorylating them [46-48]. The pattern of alternative splicing of each CLK paralog was found to be the same: a full-length isoform and an isoform that skips exon 4 [49]. We found that in each case, skipping exon 4 induces a frameshift that creates a PTC (Figure 5a). The conceptual translations of these PTC+ isoforms, described as 'truncated,' lack most of the coding region, including the kinase domain.

Having observed conservation amongst the human paralogs, we examined the gene structures of the mouse orthologs of each CLK (Figure 5b) to determine if the splicing pattern was shared across species. We identified mouse orthologs through existing RefSeq database annotation. EST evidence of alternative splicing showed that all three mouse CLK orthologs showed the same pattern of alternative splicing, skipping exon 4 to induce a PTC, as seen in the human CLKs. The significance of this evolutionary conservation is underscored by the recent finding that low-abundance alternative exons are 'mostly not conserved' between human and mouse [50]. For the CLK genes, the alternative exons and the introns flanking them are among the most similar regions of these genes (Figure 5b).

We next searched for evidence of more distant conservation of CLK alternative splicing. We identified the single CLK homolog in the sea squirt by using a hidden Markov model of CLKs to search the Ciona intestinalis genome (see Materials and methods). EST evidence clearly indicated that the same alternative splicing pattern seen in human and mouse is also conserved in C. intestinalis. We were not able to observe a set of similar splicing patterns in D. melanogaster (data not shown).

Menegay and co-workers "tested whether expression of CLK1 splice products was subject to regulation by cellular stressors" [48]. They found that "UV exposure or high salt conditions had no effect on the ratio of full-length to truncated splice forms of CLK1. Cycloheximide, however, had a large effect, changing the ratio dramatically in favor of the truncated kinase-less form of mRNA" (see Figure 6). Cycloheximide, a chemical inhibitor of translation, is known to inhibit NMD [51], because NMD depends on translation. Indeed, cycloheximide is now a commonly used reagent for NMD-inhibition experiments (see, for example [28,52,53]). Combined with our finding that the 'truncated' mRNA isoform possesses a PTC, the results of Menegay et al. can be readily explained: the increased abundance of the truncated PTC+ isoform following cycloheximide treatment is likely to be the result of inhibiting NMD, which normally degrades it.

CLK1 has been shown to indirectly affect its own splicing [46]: the presence of high levels of CLK1 protein favors generation of the truncated PTC+ splice variant. However, instead of coding for an inactive, truncated protein isoform, we propose that this PTC+ mRNA isoform may be simply degraded by NMD. Autoregulation of this type would be analogous with that seen for the splicing factors SC35 [29] and PTB [35]. Both

Figure 2 (see previous page)

Many human alternative isoforms in SWISS-PROT derive from PTC+ mRNAs. (a) We analyzed each of the human SWISS-PROT entries containing a VARSPLIC line in its feature table, using this information to assemble protein isoform sequences. Ambiguous VARSPLIC entries led us to discard five entries from our analysis at this point. (b) We next identified cDNA/mRNA sequences corresponding to each protein isoform assembled from SWISS-PROT. BLAST was used to align each protein isoform sequence to translated cDNA/mRNA sequences in GenBank and RefSeq, filtering to ensure only high confidence matches. To obtain the coding sequence of each mRNA/cDNA sequence, we used LocusLink to map each to the correct human genomic contig sequence from the NCBI human genome build 30. We referred to the CDS feature of each GenBank or RefSeq cDNA/mRNA record to identify stop codon locations. (c) We used the SPIDEY mRNA-to-genomic DNA alignment program to determine the gene structure of each mRNA/cDNA isoform sequence. After generating these gene structures, we could determine the PTC+ status on the basis of stop codon location relative to exon-exon junctions. If the termination codon was found to be more than 50 nucleotides upstream of the final intron, the transcript was deemed PTC+ and designated a candidate target of NMD according to the model of mammalian PTC recognition. (d) Each putative PTC+ isoform was manually inspected for errors in gene structure prediction. These errors include false exon predictions due to poly(A) tails and cDNA/mRNA sequence not seen in the corresponding genomic sequence.
SC35 and PTB proteins promote the alternative splicing of PTC+ isoforms of their own mRNAs, which are then degraded by NMD.

LARD/TNFRSF12/DR3/Apo3

Death-domain-containing receptors such as LARD (also known as TNFRSF12, DR3 and Apo3; SWISS-PROT entry Q99831) are known to regulate the balance between lymphocyte proliferation and apoptosis [54]. The term death domain refers to a conserved intracellular region present in receptors such as Fas and tumor necrosis factor receptor 1 (TNFR-1) that is capable of inducing apoptosis when the receptor has bound its ligand (in these cases, Fas ligand and tumor necrosis factor α (TNFα) respectively). The regulation of functional death receptor expression is important in maintaining the balance between lymphocyte proliferation and apoptosis in vivo.

LARD is alternatively spliced to produce 12 isoforms [55]. There is one full-length isoform that encodes a death domain and its expression is pro-apoptotic. Many of the 11 other isoforms, whose functions are unclear, do not encode the death domain. In a study of differential expression of LARD in unstimulated and activated lymphocytes, Screaton and co-workers found that "...there is no change in overall LARD expression in different lymphocyte subsets" [55]. Although total expression levels were unchanged, the pattern of alternative splicing changed dramatically (Figure 7). Unstimulated lymphocytes expressed five 'truncated' isoforms, but very little of the full-length isoform. They found that, "After lymphocyte activation, there is a complete switch in splicing that will expose PHA-blasted [activated] cells to the risk of apoptosis triggered through LARD...The splicing pattern reverses after PHA blasting when isoforms encoding the truncated molecules are much reduced and LARD-1 predominates."
### Table 1

**NMD-candidate SWISS-PROT isoforms**

| Accession number | SWISS-PROT ID | Isoform name | Gene name(s)            | cDNA/mRNA   |
|------------------|---------------|--------------|-------------------------|-------------|
| P78314           | 3BP2_HUMAN    | SHORT        | SH3BP2; 3BP2; RES4-23   | AB000463    |
| P50523           | A1A1_HUMAN    | SHORT        | ATP1A1                  | U16798      |
| Q9N5E7           | ABCD_HUMAN    | 2            | ABCC13                  | AF418600    |
| Q7578            | AD11_HUMAN    | SHORT        | ADAM1I; MDC             | NM_021612   |
| Q9P0K1           | AD22_HUMAN    | 2            | ADAM22; MDC2            | NM_021722   |
| Q9Y6N9           | AI75_HUMAN    | 3            | USH1C; AIE75            | AF039599    |
| Q92667           | AKP1_HUMAN    | 2            | AKAPI1; AKAP149         | NM_139275   |
| P20594           | ANPB_HUMAN    | SHORT        | NPR2; ANPRB             | NM_000907   |
| P18847           | ATF3_HUMAN    | 2            | ATF3                    | NM_004024   |
| Q9H6X2           | ATR1_HUMAN    | MAJOR        | ANTXR1I; ATR; TEM6      | NM_032208   |
| Q9NY97           | B3G7_HUMAN    | 2            | B3GNT1; B3GALT7         | AF288209    |
| Q9HB09           | BC12_HUMAN    | 2            | BCL2L12; BPR            | NM_052842   |
| P13497           | BMPI_HUMAN    | BMP1 6       | BMP1                    | NM_006130   |
|                 |               | BMP1 5       |                         | NM_006131   |
|                 |               | BMP1 4       |                         | NM_006132   |
| Q9H855           | C343_HUMAN    | 4            | CYP3A43                 | AF280111    |
| P01258           | CAL0_HUMAN    | 2            | CALCA; CALCI            | M64486      |
| Q9HC96           | CANA_HUMAN    | B            | CAPN10; KIAA1845        | NM_023084   |
|                 |               | D            |                         | NM_023086   |
|                 |               | E            |                         | NM_023087   |
|                 |               | F            |                         | NM_023088   |
| P28897           | CD38_HUMAN    | 2            | CD38                    | D84277      |
| Q08722           | CD47_HUMAN    | O1 305       | CD47                    | BC037306    |
| Q15519           | CFLA_HUMAN    | 9            | CFLAR; CLARP; MR1T; CASH| AF009617    |
| Q9H2X0           | CHRD_HUMAN    | 3            | CHRD                    | AF209930    |
|                 |               | 4            |                         | AF283325    |
| O43526           | CIQ2_HUMAN    | 3            | KCNQ2                   | NM_004518   |
| Q9NYG8           | CIW4_HUMAN    | 2            | KCN4; TRAAK             | NM_016611   |
| P49759           | CLK1_HUMAN    | SHORT        | CLK1; CLK               | L29222      |
| P49760           | CLK2_HUMAN    | SHORT        | CLK2                    | NM_001291   |
| P49761           | CLK3_HUMAN    | 2            | CLK3                    | NM_001292   |
| Q13286           | CLN3_HUMAN    | 4            | CLN3; BTS               | AF707963    |
| Q99788           | CML1_HUMAN    | MAJOR        | CMKL1; DEZ; CHEM23      | U79526      |
| P27815           | CN4A_HUMAN    | 2            | PDE4A                   | AF69491     |
| Q9H9E3           | COG4_HUMAN    | 2            | COG4                    | AB088369    |
| Q9Y215           | COLQ_HUMAN    | VII          | COLQ                    | NM_080543   |
| Q965M3           | CPXM_HUMAN    | 2            | CPXM                    | BC032692    |
| Q9B7Z0           | CRN1_HUMAN    | 4            | CRNKL1; CRN             | AF318304    |
|                 |               | 5            |                         | AF318305    |
| Q9BUV8           | CT24_HUMAN    | 4            | C20ORF24                | BC044468    |
| P57077           | CU07_HUMAN    | B            | C21ORF7                 | AF269162    |
|                 |               | C            |                         | AF269163    |
| Q9NVD3           | CU18_HUMAN    | B            | C21ORF18                | AF391112    |
| Q92879           | CUG1_HUMAN    | MAJOR        | CUGBP1; BRUNOL2; CUGBP; NAB50 | AF248648 |
| O75675           | DFFB_HUMAN    | BETA         | DFFB; DFF2; DFF40; CAD  | AB028911    |
|                 |               | GAMMA       |                         | AB028912    |
| Accession | Description | DElTA | Major | Minor | SWISS-PROT | Protein Name |
|-----------|-------------|-------|-------|-------|-----------|--------------|
| P25686    | DJB2_HUMAN  | 3     | DNAJB2; HSJ1; HSPF3 | AB028913 |
| Q09013    | DMK_HUMAN   | 1     | DMPK; MDPK | L19268 |
| Q9NYP3    | DONS_HUMAN  | 2     | DONSON; C21orf60 | NM_145794 |
|           |             | 3     |                | NM_145795 |
| Q9NY33    | DPP3_HUMAN  | 2     | DPP3 | NM_130443 |
| O60941    | DTNB_HUMAN  | 3     | DTNB | NM_033147 |
| P29320    | EPA3_HUMAN  | MAJOR | EPHA3; ETK1; ETK; HEK | NM_05233 |
| Q7616     | ERA1_HUMAN  | HERA B | ERA1; HERA | AF082658 |
| Q92731    | ESR2_HUMAN  | 3     | ESR2; NR3A2; ESTRB | BC024181 |
| O00507    | FAFY_HUMAN  | SHORT | USP9Y; USP10; DFFRY | Y13619 |
| P24071    | FCAR_HUMAN  | B DELTA S2 | FCAR; CD89 | NM_133280 |
| P41439    | FOL3_HUMAN  | SHORT | FOLR3 | Z3263 |
| Q95954    | FTCD_HUMAN  | E     | FTCD | AF289024 |
| P59103    | G72_HUMAN   | MAJOR | G72 | AY138546 |
|           |             | 2     |                | NM_172370 |
| Q9UBA6    | G8_HUMAN    | MAJOR | C6orf48; G8 | NM_016947 |
| Q9UBSS    | GBR1_HUMAN  | IE    | GABBR1 | NM_021905 |
| Q9BSJ2    | GCP2_HUMAN  | 2     | TUBGCP2; GCP2 | BC005011 |
| P56159    | GDNR_HUMAN  | 2     | GFRA1; GDNRA; TRNR1; RET1 | NM_145793 |
| O94925    | GLSK_HUMAN  | GAC   | GL5; KIAA0838 | AF15855 |
| Q96958    | HD10_HUMAN  | 4     | HDAC10 | AL022328 |
| Q30201    | HFE_HUMAN   | MAJOR | HFE; HLH | NM_00410 |
| Q9NRM6    | I17S_HUMAN  | 2     | IL1B; IL17B; EVI27 | NM_172345 |
| Q14790    | ICE8_HUMAN  | 7     | CASP8; MCH5 | NM_033537 |
| Q92851    | ICEA_HUMAN  | B     | CASP10; MCH4 | NM_001230 |
|           |             | C     |                | NM_032976 |
| Q92985    | IRF7_HUMAN  | C     | IRF7 | NM_04030 |
| Q01638    | IRL1_HUMAN  | C     | IL1R1; ST2; T1; DER4 | NM_173459 |
| Q14713    | ITP1_HUMAN  | MAJOR | ITGB1BP1; ICAP1 | NM_004763 |
|           |             | 2     |                | NM_022334 |
| Q9HCP0    | KC11_HUMAN  | IS    | CSNK1G1 | NM_022048 |
| P20151    | KKL2_HUMAN  | 3     | KKL2 | AF188745 |
| Q9H2R5    | KLF_HUMAN   | 2     | KKL15 | NM_023006 |
| QW1J2     | LEF1_HUMAN  | B     | LEF1 | AF294627 |
| P19256    | LF3_HUMAN   | SHORT | CD58; LFA3 | X08296 |
| P53667    | LIK1_HUMAN  | 3     | LIMK1; LIMK | NM_016735 |
| Q99698    | LYST_HUMAN  | MAJOR | CHS1; LYST; CHS | NM_000081 |
| P49641    | M2A2_HUMAN  | SHORT | MAN2A2; MANA2X | NM_006122 |
| Q95405    | MADI_HUMAN  | 2     | MADH1P; SARA | NM_007324 |
| P11137    | MAP2_HUMAN  | MAJOR | MAP2 | NM_002374 |
|           |             | MAP2C |                | NM_031845 |
| P27816    | MAP4_HUMAN  | 2     | MAP4 | BC015149 |
| P25912    | MAX_HUMAN   | 3     | MAX | NM_145113 |
| Q15759    | MK11_HUMAN  | BETA 2 | MAPK11; PRKMI1; SAPK2 | NM_002751 |
| O15438    | MRP3_HUMAN  | 3A    | ABC3; CMOAT2; MR3; MPL2 | NM_020037 |

Table 1 (Continued)

NMD-candidate SWISS-PROT isoforms
Table 1 (Continued)

| NMD-candidate SWISS-PROT isoforms |
|-----------------------------------|
| P21757   | MSRE_HUMAN   | II  | MSR1   | NM_002445 |
| Q9H1B4  | NXF5_HUMAN   | MAJOR | NXF5; TAPL1 | NM_032946 |
|         |              | B   |         | NM_033152 |
|         |              | C   |         | NM_033153 |
|         |              | D   |         | NM_033154 |
|         |              | E   |         | NM_033155 |
| Q96QS1  | PHMX_HUMAN   | 5   | PHEMX; TSSC6 | NM_139023 |
|         |              | 4   |         | NM_139024 |
| O14829  | PPE1_HUMAN   | 2   | PPEF1; PPEF; PPP7C | NM_152225 |
| Q9UMR5  | PPT2_HUMAN   | 2   | PPT2 | NM_138934 |
| Q9NQW5  | PRD7_HUMAN   | MAJOR | PRDM7; PFM4 | NM_052996 |
| O14818  | PSA7_HUMAN   | 4   | PSA7 | NM_152255 |
| P55036  | PSD4_HUMAN   | RPN10E | PSMD4; MCB1 | NM_153822 |
| P49768  | PSN1_HUMAN   | I 374 | PSEN1; PSEN1; AD3; PS1 | NM_007319 |
| P23468  | PTGD_HUMAN   | MAJOR | PTPRD | NM_002839 |
| O75771  | R51D_HUMAN   | 2   | RAD51L3; RAD51D | NM_133627 |
| Q93062  | RBO1_HUMAN   | MAJOR | RBPSM | NM_006867 |
| P78653  | RED1_HUMAN   | MAJOR | ADARB1; RED1; DRADA2 | NM_015833 |
| O15126  | SCA1_HUMAN   | 2   | SCAMP1; SCAMP | NM_052822 |
| Q13243  | SFRS5_HUMAN  | SRP40 2 | SFRS5; SRP40; HRS | NM_006925 |
| O60902  | SHX2_HUMAN   | MAJOR | SHOX2; SHO; OG12X | NM_006884 |
| Q13425  | SNP2_HUMAN   | 2   | SNT2B; SNT2B | NM_130845 |
| Q9Y5W8  | SNX1D_HUMAN  | 2   | SNX13; KIAA0713 | NM_015132 |
| O18583  | SON_HUMAN    | E    | SON; NREBP; DBPS; C21ORF50; KIAA1019 | NM_058183 |
| Q15528  | SUR5_HUMAN   | SURF5A | SURF5; SURF-5 | NM_006752 |
| O14763  | T1OB_HUMAN   | MAJOR | TNFRSF10B; DR5; TRAILR2; TRICK; KILLER; ZTNFR9 | NM_003842 |
| Q9BZY9  | TM31_HUMAN   | BETA | TRIM31 | NM_052816 |
| P25445  | TNR6_HUMAN   | 4   | TNFRSF6; APT1; FAS; FAS1 | NM_152873 |
|         |              | 5   |         | NM_152875 |
|         |              | 3   |         | NM_152876 |
|         |              | 2   |         | NM_152877 |
| P00750  | TPA_HUMAN    | SHORT | PLAT | NM_000931 |
| Q93038  | TR12_HUMAN   | 12  | TNFRSF25; TNFRSF12; WSL1; WSL; APO3; DR3; DDR3 | NM_148968 |
|         |              | 4   |         | NM_148969 |
|         |              | 3   |         | NM_148971 |
|         |              | 5   |         | NM_148972 |
|         |              | 6   |         | NM_148973 |
|         |              | 7   |         | NM_148974 |
| Q9BYM8  | U713_HUMAN   | 4   | UBCE7IP3; C20ORF18; XAP4 | NM_031227 |
We found that the five truncated LARD mRNA isoforms shown expressed in unstimulated lymphocytes all have PTCs (isoforms 2, 3, 4, 5 and 6). (Note that SWISS-PROT uses a different numbering scheme in which isoforms 2-6 are known as 12, 3, 5, 6 and 7, respectively.) Only the full-length apoptosis-promoting isoform 1, expressed in activated lymphocytes, is free of a PTC. Although there is presently no evidence of transcript degradation, this precise correlation between PTC-containing isoform expression and lymphocyte activation suggests that the role of alternative splicing in regulating lymphocyte apoptosis may be mediated by NMD.

Conclusions

We found that 144 of the human alternative isoforms described in SWISS-PROT derive from mRNAs that contain PTCs. These mRNAs are apparent targets for NMD, and we expect that most are degraded by this system. In many cases, existing experimental evidence is consistent with this expectation. Because our analysis was restricted to only human entries and many SWISS-PROT records could not be reliably analyzed, it is likely that there remain more unidentified putative NMD targets. We are beginning a collaborative project with SWISS-PROT to identify and suitably annotate these entries. The relevance of this effort is highlighted by the many instances in which existing experimental data can be explained in the light of NMD action.

Materials and methods

SWISS-PROT isoform extraction and assembly

We analyzed each of the 1,641 SWISS-PROT V.41 human entries containing a VARSPLIC line in its feature table [39]. Information contained in each VARSPLIC line was used to assemble protein isoform sequences for 4,556 isoforms from 1,636 unique SWISS-PROT entries. Five entries could not be analyzed due to ambiguous VARSPLIC annotation.

Identification of corresponding cDNA/mRNA sequences

Although SWISS-PROT contains cross-references to cDNA/mRNA sequences for major protein isoforms, cross-references do not exist for many alternative isoforms. To find the cDNA/mRNA sequence corresponding to each SWISS-PROT protein isoform, we used BLAST version 2.2.4 [56] to align each protein isoform sequence to translated cDNA/mRNA sequences from all GenBank [40] and RefSeq cDNA/mRNA sequences in these databases as of 22 March 2003 [41]. In these alignments, we required ≥99% identity over the full length of the SWISS-PROT isoform. In cases of multiple matches, we selected 100% identical matches over 99% identical matches and RefSeq matches over GenBank matches. For SWISS-PROT isoforms matching multiple entries from the same database at the same percent identity, the match associated with the longest cDNA/mRNA sequence was chosen. These rules associated 2,871 alternatively spliced human SWISS-PROT protein isoforms from 1,496 SWISS-PROT entries with a corresponding cDNA/mRNA sequence from either RefSeq or Genbank.

Retrieving coding sequences and genomic loci

We used LocusLink [41] to map each cDNA/mRNA sequence to the correct human genomic contig sequence from the National Center for Biotechnology Information (NCBI) human genome build 30 [57]. The coding sequence (CDS) feature of each GenBank or RefSeq record was used to identify the location of the termination codon. Of the 2,871 alternatively spliced human SWISS-PROT protein isoforms we associated with corresponding cDNA/mRNA sequences, 2,742 had GenBank or RefSeq records that were not
Figure 4
Published expression levels of calpain-10 isoforms are consistent with NMD prediction. (a) A report from Horikawa and co-workers [43] found eight alternative isoforms of calpain-10, of which four are expressed in low abundance. Our analysis found this exact set of four low-abundance isoforms to contain PTCs. (b) Gene structures of alternative mRNA isoforms of calpain-10 show the patterns of alternative splicing and indicate locations of PTCs. Also shown is the position of UCSNP-43, an intronic polymorphism that has been statistically linked to type II diabetes susceptibility in a variety of populations.

Assessing NMD candidacy
The SPIDEY mRNA-to-genomic DNA alignment program [42] was used to determine the location of introns in each cDNA/mRNA alternative isoform sequence. SPIDEY takes as input a cDNA/mRNA sequence and the corresponding genomic sequence, and it generates an alignment that establishes the gene structure. Of the 2,742 alternatively spliced human SWISS-PROT protein isoforms for which both a cDNA/mRNA sequence and stop codon location could be identified, 2,483 resulted in high-confidence SPIDEY alignments, leading us to discard 259 from our analysis. These 2,483 isoform sequences represented 1,363 unique SWISS-PROT entries. We compared the intron positions to the position of the termination codon for each remaining cDNA/mRNA alternative isoform sequence. If the termination codon was found to be more than 50 nucleotides upstream of the final intron, we deemed the transcript to be PTC+ and a polycistronic, allowing us to unambiguously determine the termination codon location for these records. These 2,742 alternative isoforms represented 1,463 unique SWISS-PROT entries.
Splicing to generate a premature termination codon is evolutionarily conserved in CLKs. The CDC-like kinases (CLKs) are splicing regulators that affect splicing decisions through the phosphorylation of SR proteins.

(a) Our screen of SWISS-PROT revealed that human CLK1, CLK2 and CLK3 paralogs all generate PTC+ alternative isoforms. The splicing pattern that generates these isoforms, skipping exon 4, is conserved in each. This splicing pattern causes a frameshift and a PTC. The percent identities from global alignments between corresponding exons and introns are shown in purple.

(b) CLKs were identified in mouse through existing annotation and in the predicted genes of the sea squirt C. intestinalis using an HMM constructed with annotated CLKs from a variety of organisms. An EST analysis revealed that the alternative splicing pattern that generates PTC+ alternative isoforms was conserved in all three sets of orthologs in human and mouse. The same splicing pattern was also found in the only C. intestinalis homolog. A relatively high degree of sequence similarity was found to be present in the introns flanking the alternative exon.

Figure 5
Splicing to generate a premature termination codon is evolutionarily conserved in CLKs. The CDC-like kinases (CLKs) are splicing regulators that affect splicing decisions through the phosphorylation of SR proteins. (a) Our screen of SWISS-PROT revealed that human CLK1, CLK2 and CLK3 paralogs all generate PTC+ alternative isoforms. The splicing pattern that generates these isoforms, skipping exon 4, is conserved in each. This splicing pattern causes a frameshift and a PTC. The percent identities from global alignments between corresponding exons and introns are shown in purple. (b) CLKs were identified in mouse through existing annotation and in the predicted genes of the sea squirt C. intestinalis using an HMM constructed with annotated CLKs from a variety of organisms. An EST analysis revealed that the alternative splicing pattern that generates PTC+ alternative isoforms was conserved in all three sets of orthologs in human and mouse. The same splicing pattern was also found in the only C. intestinalis homolog. A relatively high degree of sequence similarity was found to be present in the introns flanking the alternative exon.
candidate target for NMD according to the model of mammalian PTC recognition [14]. One hundred and seventy-seven alternatively spliced human isoforms from 130 SWISS-PROT entries were identified as possible PTC+ splice variants using these criteria. These predictions required further screening, however, to confirm the veracity of the SPIDEY alignments on which they were based.

We manually reviewed all 177 putatively PTC+ alignments and discarded 33 because of demonstrable errors in the SPIDEY alignments. These errors included a variety of malformed intron predictions and poly(A) tails mistakenly annotated as 3' exons. Isoforms that remained following the application of these manual filters were deemed high-confidence PTC+ mRNAs. This was the case for 5.8% of the isoforms (144 of 2,483) from 7.9% of the unique SWISS-PROT entries studied (107 of 1,363). The SPIDEY alignment for each of these is included in Additional data file 1; the sequence identifiers for each step are included in Additional data file 2.

**CLK analysis**

Human CLK1, CLK2 and CLK3 (SWISS-PROT IDs P49759, P49760, and P49761) were among those SWISS-PROT entries we selected for further examination. They were mapped to RefSeq and GenBank entries, as shown in Figure 2. LocusLink was used to associate each CLK gene sequence to its corresponding genomic contig. For each gene, SPIDEY version 1.35 was run twice, using the vertebrate splice-site setting, to align it with its contig sequence and determine its gene structure. This first SPIDEY alignment was used to define the extent of each gene's locus: the region containing all the coding sequence, introns, and 1,000 nucleotides of flanking sequence on each side. The second SPIDEY alignment was made using just this locus. Custom scripts (available from the authors on request), GFF2APS [58], and manual editing were used to generate the graphical representations of the gene structures shown in Figure 4. Intron and exon sequences were then extracted using the SPIDEY results to delineate exon and intron boundaries. Corresponding exons and introns were globally aligned using ALNIG version 2.0u [59] with default parameters.

Mouse CLKs were identified using RefSeq annotation (NM_009905 - which skipped exon 4 and had a PTC, NM_007712, and NM_007713). Genomic loci sequences were generated and gene structures determined for each mouse CLK gene using SPIDEY, as above. The loci sequences were then used to search the mouse ESTs from dbEST (1 May 2003) [60] using WU-BLAST 2.0u (23 May 2003) [61] with default parameters. Hits with E-values of 10−30 or better were aligned to the locus sequence using SPIDEY. These alignments were examined for evidence of PTC-inducing alternative splicing. The GI numbers for ESTs that exhibited the alternative splicing pattern shown in Figure 3 for each of the mouse CLKs are: CLK1 (full-length): 25185241, 21852543, 12560958, and others; CLK2 (exon 4 skipping): 22822098; CLK3 (exon 4 skipping): 26079129.

The *C. intestinalis* CLK homolog was identified from the database of predicted peptides [62,63] by searching (HMMSEARCH V2.2G) [64] with a HMMER model of known CLKs. This model was generated using HMMBUILD (default parameters) and calibrated using HMMCALIBRATE from a CLUSTAL W V1.83 [65] alignment of the following CLK sequences: NP_004062, NP_003984, NP_003983, NP_031738, BAB33079, NP_031740, NP_065717, AAH43963, NP_599167, NP_031739, NP_477275, EAA12103, NP_741928, BAB67874, and NP_856095. The most significant hit (E-value: 4.4e-243) from *C. intestinalis* was ci010143784. Visual inspection of other, less significant hits revealed that they align with only the kinase domain of the CLK model and none contains the LAMMER motif characteristic of CLKs. A maximum-likelihood tree was generated using PROML V2.3b3 [66] using ci010143784 and the three full-length human CLKs. This tree revealed that the *C. intestinalis* CLK is orthologous to human CLK2. The corre-
sponding cDNA transcript sequence, ci0100143784, was retrieved from the database of predicted transcripts [67].

As above, the locus for this gene was extracted from the genomic contig sequence, Scaffold18, and used to search the database of C. intestinalis ESTs. The following ESTs showed the full-length pattern with no PTC: 24144377, 24820603, 24627564, 24627468, 24866887, and 2482449. The following ESTs showed the alternatively spliced pattern that generates a PTC: 24888181, 24606693, 24823992, and 24893089. The C. intestinalis CLK gene was found to have only 11 exons, whereas human and mouse CLK2 have 13. To determine which exons were homologous, we generated a CLUSTALW multiple-sequence alignment of the known CLK protein sequences listed above and C. intestinalis CLK and used this alignment to identify corresponding regions of DNA sequence. This unambiguously indicated the exon-to-exon alignment shown in Figure 4.

The C. intestinalis CLK gene was found to have only 11 exons, whereas human and mouse CLK2 have 13. To determine which exons were homologous, we generated a CLUSTALW multiple-sequence alignment of the known CLK protein sequences listed above and C. intestinalis CLK and used this alignment to identify corresponding regions of DNA sequence. This unambiguously indicated the exon-to-exon alignment shown in Figure 4.

**Acknowledgements**

We thank Rajiv Bhatnagar, Liana Lareau, Don Rio, Marco Blanchette, Jasper Rine, Karsten Weis and Aaron Garnett for helpful discussions. This work was supported by a Searle Scholarship (01-L-116), NIH grants K22 HG00056 and T32 HG00047, and the Guidant Bioengineering Summer Research Program.

**References**

1. Black DL: *Mechanisms of alternative pre-messenger RNA splicing*, Annu Rev Biochem 2003, 72:291-336.
2. Modrek B, Lee C: *A genomic view of alternative splicing*, Nat Genet 2002, 30:13-19.
3. Kan Z, States D, Gish W: *Selecting for functional alternative splices in ESTs*, Genome Res 2002, 12:1837-1845.
4. Celotto AM, Graveley BR: *Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated*, Genetics 2001, 159:599-608.

**Additional data files**

A table showing the raw SPIDEY output of the 144 NMD-candidate isoform genes (Additional data file 1) and a file containing the identifiers of sequences at each major step of the analysis pipeline (Additional data file 2) are available with the online version of this article.
5. Leeds P, Pelzt SW, Jacobson A, Culbertson MR: The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing premature translocation termination codon. Genes Dev 1991, 5:2303-2314.

6. Kinniburgh AJ, Maq late LE, Schedli T, Rachmielwitz E, Ross J: mRNA-deficient beta o-thalassemia results from a single nucleotide deletion. Nucleic Acids Res 1982, 10:5421-5427.

7. Frischmeyer PA, Dietz HC: Nonsense-mediated mRNA decay in health and disease. Hum Mol Genet 1999, 8:1893-1900.

8. Le Hir H, Izaurralde E, Maq late LE, Moore MJ: The spliceosome deposits multiple 20-24 nucleotides upstream of mRN A exon-exon junctions. EMBO J 2000, 19:6860-6869.

9. Le Hir H, Moore MJ: Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon-exon junctions. Genes Dev 2000, 14:1098-1106.

10. Lykke-Andersen J, Shu MD, Steitz JA: Communication of the position of exon-exon junctions to the mRNA surveillance in a family with Marfan syndrome: a potential mechanism leading to nonsense- mediated decay. Mol Cell Biol 1999, 19:4056-4064.

11. Mitrovich QM, Anderson P: Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans. Genes Dev 2000, 14:2173-2184.

12. Morrison M, Harris KS, Rod B: smg mutants affect the expression of alternatively spliced SR protein mRNAs in Caenorhabditis elegans. Proc Natl Acad Sci USA 1997, 94:9782-9785.

13. Le Hir H, Moore MJ: Nonsense mediated decay downregulates alternatively spliced transcripts bearing nonsense codons. Hum Mol Genet 2003, 12:99-109.

14. Reimund S, Atkinson D, Soret J: Evidence by sequence comparison of a new family of mammalian polypyrimidine tract binding proteins for nonsense-mediated decay. Molecular Cell 2004, 13:91-100.

15. Lewis BP, Green RE, Brenner SE: Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci USA 2003, 100:89-92.

16. Green RE, Lewis BP, Hillman RT, Blan chette M, Lareau LF, Garnett AT, Rio DC, Brenner SE: Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. Bioinformatics 2003, 19(Suppl 1):I118-I121.

17. Medghalchi SM, Frischmeyer PA, Mendell JT, Kelly AG, Lawler AM, Dietz HC: Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for mammary embryonic viability. Hum Mol Genet 2001, 10:99-105.

18. Boeckmann B, Bairach B, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O’Donovan C, Phan I et al.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003, 31:363-370.

19. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: RefSeq and GenBank. Nucleic Acids Res 2003, 31:23-25.

20. Pruitt KD, Maglott DR: RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 2001, 29:137-140.

21. Wheelan SJ, Church DM, Ostell J: Spidey: a tool for mRNA-toprotein-mRNA interactions. Genome Res 2001, 11:152-157.

22. Orgel LE, Benner SA: RNA as the heredity information in the pre-DNA era. J Mol Biol 1999, 289:295-305.

23. Medvedev Y, Cprogrammer of the perfect translation: mRNAs subject to nonsense-mediated mRNA surveillance: the perfect translation. Mol Cell 2003, 1785-1796.

24. Lee SJ, Kim BK, Yoon SB, Kim YW, Kim BJ, Lim HJ, Park SJ: Evidence for the widespread interference of mRNA translation: mRNAs subject to nonsense-mediated mRNA surveillance affect global expression of the yeast and mammalian pathways. Proc Natl Acad Sci USA 2003, 100:106-111.

25. Schibler U, Fratelli M, Cusella-de Angelis I, Rappaport RJ: TIAl or TIAR is required for DT40 cell viability. J Biol Chem 2003, 278:10465-10476.

26. Wilson GM, Sun Y, Sellers J, Lu H, Penkman N, Dillard G, Brewer G: Regulation of AUF1 expression via conserved alternatively spliced elements in the 3′ untranslated region. Mol Cell Biol 1999, 19:4056-4064.

27. Sureau A, Gattoni R, Dooghe Y, Stevenin J, Soret J: SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. EMBO J 2000, 19:1785-1796.

28. Wilson GM, Sun Y, Sellers J, Lu H, Penkman N, Dillard G, Brewer G: Regulation of AUF1 expression via conserved alternatively spliced elements in the 3′ untranslated region. Mol Cell Biol 1999, 19:4056-4064.
53. Lei XD, Chapman B, Hankinson O: Loss of cytoplasmic messenger RNA expression due to nonsense-mediated decay. Mol Pharmacol 2001, 60:388-393.

54. Thome M, Tschopp J: Regulation of lymphocyte proliferation and death by FLIP. Nat Rev Immunol 2001, 1:50-58.

55. Screaton GR, Xu XN, Olsen AL, Cowper AE, Tan R, McMichael AJ, Bell JJ: LARD: a novel lymphoid-specific death domain containing receptor regulated by alternative pre-mRNA splicing. Proc Natl Acad Sci USA 1997, 94:4615-4619.

56. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215:403-410.

57. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al.: Initial sequencing and analysis of the human genome. Nature 2001, 409:860-921.

58. Abril JF, Guigo R: gft2ps: visualizing genomic annotations. Bioinformatics 2000, 16:743-744.

59. Myers EW, Miller W: Approximate matching of regular expressions. Bull Math Biol 1989, 51:5-37.

60. Boguski MS, Lowe TM, Tolstoshev CM: dbEST - database for "expressed sequence tags". Nat Genet 1993, 4:332-333.

61. WU-BLAST archives [http://blast.wustl.edu]

62. Database of predicted C. intestinalis peptides [ftp://ftp.jgi-psf.org/pub/JGI_data/Ciona/v1.0/ciona.prot.fasta.gz]

63. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, De Gregorio A, Gelpke M, Goodstein DM et al.: The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science 2002, 298:2157-2167.

64. HMMER: sequence analysis using profile hidden Markov models [http://hmmer.wustl.edu]

65. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994, 22:4673-4680.

66. Adachi J, Hasegawa M: MOLPHY version 2.3: Programs for Molecular Phylogenetics Based on Maximum Likelihood Tokyo: Institute of Statistical Mathematics; 1996.

67. Database of predicted C. intestinalis transcripts [ftp://ftp.jgi-psf.org/pub/JGI_data/Ciona/v1.0/ciona.mrna.fasta.gz]