714–724 *Nucleic Acids Research, 2005, Vol. 33, No. 2*
doi:10.1093/nar/gki210

The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons

Simon Minovitsky, Sherry L. Gee, Shiruyeh Schokrpur, Inna Dubchak and John G. Conboy*

Life Sciences Division and Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Received November 11, 2004; Revised December 21, 2004; Accepted January 6, 2005

ABSTRACT

Previous studies have identified UGCAUG as an intron splicing enhancer that is frequently located adjacent to tissue-specific alternative exons in the human genome. Here, we show that UGCAUG is phylogenetically and spatially conserved in introns that flank brain-enriched alternative exons from fish to man. Analysis of sequence from the mouse, rat, dog, chicken and pufferfish genomes revealed a strongly statistically significant association of UGCAUG with the proximal intron region downstream of brain-enriched alternative exons. The number, position and sequence context of intronic UGCAUG elements were highly conserved among mammals and in chicken, but more divergent in fish. Control datasets, including constitutive exons and non-tissue-specific alternative exons, exhibited a much lower incidence of closely linked UGCAUG elements. We propose that the high sequence specificity of the UGCAUG element, and its unique association with tissue-specific alternative exons, mark it as a critical component of splicing switch mechanism(s) designed to activate a limited repertoire of splicing events in cell type-specific patterns. We further speculate that highly conserved UGCAUG-binding protein(s) related to the recently described Fox-1 splicing factor play a critical role in mediating this specificity.

INTRODUCTION

Alternative pre-mRNA splicing is a prominent feature of human gene expression, and is often credited with allowing a relatively small number of genes to encode an extremely complex proteome. Particularly intriguing and physiologically relevant is the subset of alternative exons that exhibit developmental- or differentiation-specific expression, i.e. exons whose splicing is switched on and off in a highly regulated manner. Such alternative splicing switches probably play a major role in defining the specialized properties of differentiated mammalian cells by facilitating expression of cell type-specific subsets of the total proteome. Understanding the mechanisms by which splicing switches are regulated is thus critical to a better appreciation of specialization in metazoan organisms.

The molecular switch that mediates inclusion or exclusion of an alternative exon is generally thought to be regulated by the antagonistic activities of splicing factor proteins that interact at positive-acting enhancer elements versus those binding at negative-acting silencer elements in the RNA (1–3). Significant progress in identifying *cis*-regulatory elements in the exons has been made by large-scale screening of RNA sequences to identify those that bind to known splicing factors and/or exhibit functional enhancer or silencer activity (4–11). Increasingly, computational approaches are also contributing to the identification of candidate splicing regulatory elements. Large-scale sequence comparisons have identified several classes of exon splicing enhancers (ESEs) and intron splicing enhancers (ISEs) based on their over-representation in or near exons with weak splice sites (12,13). Computational analysis of sequence motifs important for constitutive splicing has also been reported (14). Finally, as more examples of tissue-specific exons are annotated, another promising strategy for identification of regulatory elements involves computational analysis of specialized datasets of exons exhibiting shared expression patterns (15). Importantly, development of algorithms that predict splicing enhancer elements are improving our ability to recognize new classes of splicing defects that underlie many human diseases (16,17).

Critical splicing regulatory motifs are often located in the introns, and it is interesting to note that such elements play a prominent role in a number of regulated alternative splicing events (2,18–23). Recently, (U)GCAUG elements have been characterized experimentally as intronic splicing regulatory

*To whom correspondence should be addressed. Tel: +1 510 4866973; Fax: +1 510 4866746; Email: JGConboy@lbl.gov

elements (18,24–27) that represent highly specific recognition sequences for novel splicing factors of the Fox-1 family of RNA binding proteins (26). Computational analysis showed further that the UGCAUG hexamer is dramatically over-represented in the proximal downstream intron sequences of many brain-specific exons (15). Together these findings suggested the intriguing hypothesis that Fox proteins may play a critical role in mediating developmental and differentiation-specific splicing switches.

Here, we show that (U)GCAUG association with regulated exons has been highly evolutionarily conserved in the downstream proximal introns near tissue-specific exons of vertebrates from fish to humans. In many cases, number, position and sequence context of (U)GCAUG elements has been strongly conserved, supporting the hypothesis that this element is a critical component of the splicing switch mechanism that mediates tissue-specific splicing events.

MATERIALS AND METHODS

Collection of tissue-specific alternative exon and control exon datasets

'Tissue-specific alternative exon' datasets were related to the human brain-enriched exons studies earlier (15), modified so as to remove very small (<12 nt) and very large exons (>500 nt), and to add a few newly recognized brain-enriched exons. These new genes include exons in EPB41L3 (28), EPB41L1 (28), ANK1 (29), BAIAP (30), DCAMKL1 (31), DNM3, EWSR (32), MYH10 (33) and PTBP2 (34) genes.

For phylogenetic analysis, the orthologous exons were identified in the mouse, rat, dog, chicken and pufferfish genomes using VISTA alignment tools supplemented, where necessary, by manual analysis. Automatic alignment was successful at finding most of the longer alternative exons directly. However, some of the shorter exons lacked sufficient homology to be easily detected at the nucleotide level. In these cases, the flanking constitutive exons were identified, and a conceptual translation of the intervening sequence was successful in identifying the alternative exon.

'Tissue non-specific alternative exons' were derived from the European Bioinformatics Institute database of human alternative exons (http://www.ebi.ac.uk/asd/altextron/index.html), using the so-called cryptic exon database. The first 100 exons in this dataset were extracted and flanking intron sequences retrieved for computational analysis.

'Control exon databases' were generated from randomly selected chromosomal regions by extraction from RefSeq annotation databases to get exon coordinates. Control groups for the mammalian and chicken genomes contained at least several hundred exons. Due to the absence of a VISTA alignment for the dog genome, we utilized the human control dataset for comparative purposes. For Fugu, constitutive exons were extracted from the remainder of the genes in which the brain-enriched exons reside. These control datasets presumably represent predominantly constitutively spliced exons and were used as a reference against which to estimate the statistical significance of sequences over-represented in the alternative exon datasets.

The brain-enriched datasets and the control datasets are available at http://gsd.lbl.gov/splicing/.

Computational analysis

Candidate regulatory elements were predicted computationally by identifying oligonucleotide sequences (or words) that were over- or under-represented in each tissue-specific dataset, relative to the control datasets. The frequencies of each word were calculated at any selected range, relative to the exon boundaries, using the algorithm described previously (15). For each word, a contrast score was calculated as the difference in frequency in the tissue-specific dataset versus the control dataset. The statistical significance of contrast scores was estimated using resampling statistics as described (15) with the following modification. In this paper, the probability refers to the chance that any hexamer might exhibit a given contrast score in a randomly selected subset of the control sample (equal in size to the brain sample).

RT–PCR analysis

Tissue-specific splicing patterns of selected transcripts were characterized using RT–PCR techniques to amplify specified RNA regions from several different mouse tissue sources. RNA was prepared using RNeasy columns according to the manufacturer's instructions (Qiagen, Valencia, CA). One microgram of total RNA was transcribed into cDNA using random hexamer primers in a total volume of 10 μl. Then, 2 μl cDNA was amplified using the following primers: ACVR2, forward, 5′-CACCGAAGCCACCCTATTACAAC-3′; reverse, 5′-CCCCTTGCTTTCACTTCTAACAGC-3′; EPB41L1s, forward, 5′-TGCCTCAGTCAGTGAGAATCACG-3′; reverse, 5′-GGTGCTTCAACAAGACATCCTCTG-3′; EPB41l1b, forward, 5′-GCATCAATGAACTCAAGAGGACCC-3′; reverse, 5′-TTCTTCCTGATTGCCAGACTGC-3′; EPB41L3s, forward, 5′-AAGACAGAAGGAAGAAGGCTGAGG-3′; reverse, 5′-TCCCACTCGTTTGTTAAGGCAG-3′; EPB41L3b, forward, 5′-CCAAGTGTGACTGAGAAAACACAGG-3′; reverse, 5′-CATGCCTCTCCTCTACCAGTATCG-3′; NF1, forward, 5′-GACTGTCTTGTCTCTTGCTTTCGG-3′; reverse, 5′-CACAGCCTTGCACTGCTTTATG -3′; MYH10, 5′-GTCATTCAGTACCTTGCCCACG-3′; reverse, 5′-CGCATTTCCAAAGGATTCCAG-3′; SCN8a, forward, 5′-TATGCGGACAAGGTCTTCACCTAC-3′; reverse, 5′-ATCGGATCTCTGTGTTGTTGCC-3′; PTBP2, forward, 5′-AGGAGACATCCCTACTAGCTGTTCC-3′; reverse, 5′-CATCAGCCATCTGTATCAGAGCAC-3′. Thirty-five cycles of amplification were performed under the following conditions: denaturation for 30 s at 94°C; annealing for 30 s at 55°C; extension for 60 s at 72°C. DNA fragments were analyzed by 5% polyacrylamide gel electrophoresis. The identity of PCR products was confirmed by DNA sequence analysis.

RESULTS

Identification of orthologous tissue-specific exons in several species

A group of 27 tissue-specific, alternatively spliced, cassette exons with predominant expression in the brain was assembled for this study. Table 1 shows the RefSeq names of the human genes in which these exons reside, as well as the size in nucleotides and the hg16/NCBI build 34 coordinates for each exon in the human genome. As described in Methods and also as shown in Table 1, the orthologs for most of these exons were

Table 1. Brain-enriched alternatively spliced exons examined in this study

	Gene name	Exon size (nt)	Human coordinates, July 03	Mouse	Rat	Dog	Chicken	Fugu
1	ACVR2	24	chr2:148879875–148879898(+)	+	+	+	+	+
2	ANK1	24	chr8:41575288–41575311(−)	+	+	+	+	
3	ATP2B4	178	chr1:200879784–200879961(+)	+	+	+	+	
4	BAIAP	84	chr3:65330156–65330239(−)	+	+	+	+	
5	BIN1	93	chr2:127919008–127919100(−)	+		+	+	
6	CACNCB1	63	chr9:136130205–136130267(+)	+	+	+	+	+
7	CLTB	54	chr5:175804403–175804456(−)	+	+	+	+	+
8	DCAMKL1	74	chr13:34160349–34160422(−)	+	+	+	+	+
9	DLG1	36	chr3:198121894–198121929(−)	+	+	+	+	
10	DNM3	30	chr1:169302212–169302241(+)	+	+	+	+	
11	EPB41L1	36	chr20:35498680–35498715(+)	+	+	+	+	
12	EPB41L1	411	chr20:35512839–35513249(+)	+	+	+	+	
13	EPB41L3	36	chr18:5397700–5397735(−)	+	+	+	+	
14	EPB41L3	123	chr18:5388021–5388143(−)	+	+	+	+	
15	EWSR	18	chr22:27994808–27994835(+)	+	+	+	+	
16	FHL1	200	chrX:133997009–133997208(+)	+	+	+		
17	GABRG2	24	chr5:161558631–161560654(+)	+	+	+	+	
18	GRIN1	63	chr9:135399897–135399959(+)	+	+	+	+	+
19	KSR	42	chr17:26073950–26073991(+)	+	+	+	+	+
20	MAG	45	chr19:40495024–40495068(+)	+	+			
21	MYH10	30	chr17:8680527–8680556(−)	+	+	+	+	+
22	MYH10	63	chr17:8634507–8634569(−)	+	+	+	+	+
23	NF1	30	chr17:29675683–29675712(+)	+	+	+		
24	PTPB2	34	chr1:96743776–96743809(+)	+	+	+	+	+
25	PTPRF	27	chr1:43480037–43482062(+)	+	+	+	+	+
26	SCN8a	123	chr12:50460700–50460822(+)	+	+	+	+	+
27	SRC	18	chr20:36700292–36700309(+)	+	+	+	+	+

(+) and (−) indicate the direction of transcription for the selected genes.

identified in several additional vertebrate genomes, including mouse, rat, dog and chicken (indicated by '+'). In addition, the pufferfish orthologs of 11 of these exons were also identified. The DNA sequences for each alternative exon were retrieved, along with 1 kb of upstream and downstream intron, to create five new tissue-specific datasets for the three mammalian species (mouse, rat, dog), one avian species (chicken) and one piscine species (pufferfish). These sequences, together with large datasets of control exon/intron sequences for each species, were used for computational analysis of candidate splicing regulatory elements.

Phylogenetic conservation of intronic UGCAUG sites

In the original dataset of brain-specific exons in the human genome, UGCAUG sites were preferentially localized to the proximal downstream intron (15), with greatest enrichment of sites in the first 100 nt downstream (here defined as the D100 region). We hypothesized that, if UGCAUG is an important splicing regulatory element, it should be evolutionarily conserved near the orthologous alternative exons in other species. This hypothesis was tested by subjecting each of the species-specific datasets to computational analysis with a word search algorithm. Oligonucleotide frequencies were counted and contrast scores computed to quantitate the over-representation of candidate regulatory elements in the tissue-specific exon datasets relative to the control datasets of non-tissue-specific exons (see Methods). A high contrast score therefore suggests that a sequence may be important for regulated alternative splicing but not for constitutive splicing.

Table 2 lists the top ten over-expressed hexamers in each species-specific database, ranked in order by contrast score, for the proximal intron region 0–400 nt downstream of the exons (the D400 region). Similar to our earlier study, UGCAUG was the most over-represented hexamer in the human dataset with a contrast score of 2.29. This was by far the highest contrast score for any unique hexamer in the D400 region, and it was the only statistically significant score ($P < 0.05$), using a conservative measure of the probability that any hexamer would be over-represented to this extent in a dataset of this size. Remarkably, UGCAUG was the most over-represented hexamer observed in the D400 region for all five additional species including mouse, rat, dog, chicken and even pufferfish. In all cases, the contrast scores were statistically significant ($P < 0.05$) and far higher for UGCAUG than for any other hexamer. Moreover, in all of these datasets UGCAUG was either first or second in absolute abundance among all 4096 hexamers within the D400 region (Table 2).

Because the binding specificity of the zebrafish Fox-1 splicing factor was recently reported as the pentamer GCAUG (35), we examined the proportion of total GCAUG motifs in the D400 region that occur in the context of the UGCAUG hexamer. Notably, for all six species the great majority (65–93%) of all GCAUG motifs in the D400 region do reside in a UGCAUG hexamer (Table 3A). AGCAUG consistently ranked second, with 7–20% of the total, while CGCAUG and GGCAUG represented a small minority of cases. Only the UGCAUG hexamer possessed a statistically significant contrast score. This result suggests, at least in the datasets examined here, that UGCAUG plays a functionally dominant role in splicing regulation, and that GCAUG or the related hexamers may be involved in regulating a smaller subset of alternative splicing events. It remains to be determined as to whether the same Fox-1 splicing factor operates at all of these sites via tolerance of sequence variation at position 1,

Table 2. Over-represented hexamers in the proximal downstream intron region D400

Sequence	Brain frequency $\times 10^{-3}$ (rank)	Control frequency $\times 10^{-3}$ (rank)	Contrast score	*P*-value
Human				
UGCAUG	2.53 (1)	0.25 (1310)	2.29	0.031*
UUUUAA	1.72 (2)	0.28 (1103)	1.44	0.240
UUUGUU	1.72 (2)	0.39 (718)	1.33	0.301
UUGUUU	1.63 (5)	0.36 (822)	1.28	0.343
UGUUUU	1.72 (2)	0.45 (555)	1.27	0.347
UGCUUU	1.63 (5)	0.37 (759)	1.26	0.357
UUUUAU	1.45 (7)	0.21 (1563)	1.24	0.375
UCAUUU	1.36 (9)	0.22 (1459)	1.14	0.494
UUUAUU	1.36 (9)	0.24 (1384)	1.12	0.506
GCAUGC	1.27 (11)	0.20 (1597)	1.07	0.580
Mouse				
UGCAUG	2.36 (1)	0.54 (283)	1.81	0.030*
CUGCAU	1.45 (5)	0.38 (836)	1.07	0.391
UUUCUG	1.72 (2)	0.66 (124)	1.07	0.396
AUUUUA	1.36 (7)	0.41 (706)	0.95	0.575
UUUAAA	1.27 (9)	0.42 (627)	0.85	0.803
CAUUUG	1.09 (20)	0.28 (1512)	0.81	0.867
UGAUUU	1.09 (20)	0.35 (1016)	0.74	0.964
ACCAGG	1.18 (13)	0.44 (561)	0.74	0.969
UGCUUU	1.18 (13)	0.45 (528)	0.73	0.974
UUUUAA	1.18 (13)	0.45 (482)	0.71	0.983
Rat				
UGCAUG	3.03 (1)	0.46 (374)	2.56	0.028*
CUGCAU	2.05 (2)	0.30 (1283)	1.75	0.119
UGCUUC	1.46 (4)	0.43 (468)	1.03	0.543
UGUCUG	1.66 (3)	0.67 (53)	0.99	0.594
GCAUGC	1.17 (15)	0.29 (1453)	0.89	0.780
GCAUGU	1.17 (15)	0.36 (791)	0.81	0.908
UCGCUG	0.88 (54)	0.08 (3151)	0.79	0.927
CCCUCU	1.37 (6)	0.58 (112)	0.78	0.941
CCCACC	1.27 (9)	0.50 (264)	0.76	0.958
CUGAUU	1.07 (25)	0.31 (1192)	0.76	0.962
Dog				
UGCAUG	2.73 (1)	0.25 (1310)	2.49	0.024*
UGUUUU	1.76 (2)	0.45 (555)	1.31	0.360
UUUGUU	1.66 (3)	0.39 (718)	1.27	0.391
UUUUUG	1.66 (3)	0.44 (588)	1.22	0.447
UGCUUU	1.56 (5)	0.37 (759)	1.19	0.488
AUUUUU	1.56 (5)	0.40 (685)	1.16	0.521
UCCUUU	1.46 (8)	0.32 (935)	1.14	0.552
UUUAUU	1.37 (9)	0.24 (1384)	1.13	0.565
UUUGCU	1.37 (9)	0.30 (1041)	1.07	0.654
CUGCAU	1.37 (9)	0.30 (1023)	1.06	0.663
Chicken				
UGCAUG	2.12 (2)	0.50 (340)	1.62	0.017*
UGUUUU	2.22 (1)	1.02 (4)	1.19	0.199
GUUUUU	1.69 (4)	0.71 (52)	0.98	0.553
UUUGUU	1.80 (3)	0.91 (13)	0.89	0.778
UUUUUU	1.69 (4)	0.84 (19)	0.85	0.859
CAUGCU	1.27 (14)	0.44 (515)	0.83	0.906
AUGCAU	1.16 (19)	0.36 (890)	0.80	0.941
CUUUUU	1.48 (8)	0.69 (62)	0.79	0.951
UAUAUA	0.95 (53)	0.19 (2299)	0.77	0.977
UGGUUC	0.95 (53)	0.21 (2089)	0.74	0.987
Fugu				
UGCAUG	3.00 (1)	0.61 (227)	2.39	0.010*
UUAAAU	1.62 (6)	0.22 (1635)	1.39	0.789
CCAUUU	1.62 (6)	0.28 (1290)	1.34	0.807
UUUUUC	2.08 (2)	0.77 (76)	1.30	0.884
GGAUGG	1.38 (16)	0.11 (2562)	1.27	0.893
AAAUGG	1.38 (16)	0.11 (2562)	1.27	0.893
UGUGUG	1.85 (4)	0.61 (227)	1.24	0.956
UUGCAU	1.62 (6)	0.39 (757)	1.23	0.956
AUAAUC	1.38 (16)	0.17 (2041)	1.22	0.956
CUUUUU	1.85 (4)	0.66 (163)	1.18	0.964

*Indicates results that are statistically significant (*P*-value < 0.05).

Table 3. Nucleotide preferences flanking GCAUG and UGCAUG

	Human	Mouse	Rat	Dog	Chicken	Fugu
(A) Abundance of hexamers containing GCAUG in the D400 region						
UGCAUG	28 (65%)	28 (74%)	31 (79%)	26 (74%)	20 (67%)	13 (93%)
AGCAUG	8 (19%)	5 (13%)	5 (13%)	5 (14%)	6 (20%)	1 (7%)
CGCAUG	4	2	2	3	2	0
GGCAUG	3	3	1	1	2	0
(B) Abundance of heptamers containing UGCAUG in the D400 region						
UUGCAUG	9	8	6	6	7	6
CUGCAUG	11	11	13	11	6	2
AUGCAUG	5	6	8	4	5	3
GUGCAUG	3	4	4	5	2	2
UGCAUGU	7	10	8	7	5	5
UGCAUGC	7	6	9	6	4	1
UGCAUGA	7	7	8	8	6	2
UGCAUGG	7	6	6	5	5	5

Underlined residues indicate nucleotides flanking the core GCAUG or UGCAUG motif.

or whether distinct Fox-1 related splicing factor(s) bind to hexamers that vary at position 1.

Finally, we asked whether there is any extended sequence preference for UGCAUG hexamers in these brain-enriched datasets. Heptamer sequences containing UGCAUG were therefore analyzed to test for nucleotide preferences flanking the core hexamer. As shown in Table 3B, there is a moderate preference (61–71%) for a pyrimidine nucleotide preceding UGCAUG in all six species (*P*-values: 0.0178 for human, 0.068 for mouse, 0.140 for rat, 0.131 for chicken, 0.0843 for dog and 0.290 for Fugu). Analysis of larger datasets may be necessary to provide a better test of significance. No marked preference for any particular nucleotide immediately following the UGCAUG motif was observed. Together these results suggest that the hexamer UGCAUG is the functional motif shared among many brain-enriched exons.

UGCAUG is a marker of tissue-specific alternative exons

Although this study focused on a group of brain-enriched exons, (U)GCAUG has also been reported as an enhancer of alternative exons with different tissue-specific regulation (24,25,27). We therefore proposed that UGCAUG is a marker specifically of regulated alternative splicing, but not a general marker for all alternative exons. As a test of this hypothesis, we examined the association of UGCAUG with a larger dataset of non-tissue-specific alternative exons. For this purpose, we used a group of 100 alternative exons compiled in the European Bioinformatics Institute database (see Methods). When the word counting algorithm was applied to this dataset, the results revealed both similarities and marked differences compared with the brain-enriched datasets (Table 4). Notably, the top ten contrast scores of the non-tissue-specific dataset were all very U-rich and G-deficient, similar to what was observed in the brain-enriched datasets. However, the non-tissue-specific group differed from the regulated group in one critical respect: there was no significant over-representation of the UGCAUG hexamer in proximal intron sequences of the unregulated group. Among all possible hexamers, UGCAUG ranked as the most over-represented sequence in the D400 region of the human brain-enriched

Table 4. Most over-represented hexamers in the non-tissue-specific dataset, D400 region

Sequence	Frequency in data set $\times 10^{-3}$	Frequency in control $\times 10^{-3}$	Contrast
Human			
AUUUUU	1.60	0.40	1.20
UAUUUU	1.37	0.32	1.05
UUAUUU	1.19	0.26	0.93
UUUUUA	1.22	0.34	0.88
UUUUUC	1.22	0.37	0.85
UUUUUU	1.45	0.60	0.85
CUUUUU	1.24	0.42	0.83
UUCUUU	1.22	0.39	0.83
UUUAUU	1.04	0.24	0.80
UCUUUU	1.09	0.34	0.76

dataset, but was only 184th on the list of top contrast scores for the non-tissue-specific dataset of human alternative exons. This marked difference persisted when the analysis was expanded to include the entire 1 kb flanking intron sequences: almost 90% (24/27) of the human brain-enriched exons possessed UGCAUG within 1 kb upstream or downstream, whereas only 48% (47/99) of the non-tissue-specific exons did so. The latter frequency is not significantly different from random chance.

Distribution of UGCAUG motifs relative to the regulated exons

Our data demonstrates that UGCAUG is frequently localized to the D400 region; however, previous studies have shown that UGCAUG motifs can function at a distance of at least 1 kb from the regulated exon (26). To examine the spatial distribution of UGCAUG motifs in the introns flanking brain-enriched exons, we repeated oligonucleotide counts at 100 nt intervals upstream and downstream of the regulated exons. This analysis demonstrated that the highest frequency of UGCAUG occurred within the D400 region for all six species-specific datasets (Figure 1). A more modest frequency of UGCAUG elements, but still elevated in comparison to the very low incidence of UGCAUG hexamers flanking non-tissue-specific alternative exons (Figure 1, bottom panel), was observed further downstream and in the upstream intron region.

Additional insight into the distribution of conserved UGCAUG hexamers was provided by phylogenetic analysis of individual orthologous exons in the brain-enriched datasets. Figure 2 shows the conserved UGCAUG elements that map near individual tissue-specific exons in the brain datasets. Conserved elements were defined as hexamers that exhibited a similar position and sequence context in two or more species; sporadic elements that occurred in only one species, which were included in the spatial distributions shown in Figure 1, were eliminated from this revised map. Many of these phylogenetically conserved elements were located within 10–400 nt of downstream intron sequence (shaded region); fully 70% (19/27) of these exons possessed at least one conserved UGCAUG in the D400 region. However, several highly conserved UGCAUG elements were located in more distal intronic regions, both upstream and downstream of the regulated exons. In total, about 90% of the alternative exons in the mammalian datasets possessed at least one conserved UGCAUG hexamer within 1 kb. About half of these are spatially conserved in the chicken genome, suggesting that these candidate splicing regulatory elements have been phylogenetically conserved from avians to mammals. However, although UGCAUG elements were strongly associated with tissue-specific exons in the pufferfish genome, their spatial conservation is much reduced.

Conservation of UGCAUG sequence context

We next asked whether spatially conserved UGCAUG elements occur within the context of a larger phylogenetically conserved sequence that might include other important regulatory motifs. For this purpose, we considered separately elements that were very close to the exon, i.e. within 100 nt versus those that mapped >250 nt downstream. The most proximal elements within 100 nt indeed exhibit high conservation of flanking sequence, including the splice site sequences (results not shown), as expected for the proximal intron sequence of orthologous alternative exons (36). Interestingly, distal UGCAUG elements also occur in a highly conserved local sequence context. Figure 3 presents a multiple sequence alignment for six such elements that map 250–600 nt downstream of the regulated exon. Although distal intron sequences are not usually conserved (36), here we observed high conservation of primary nucleotide sequence at least 15 nt on either side of the core UGCAUG motif. Conservation occurred primarily among orthologous members of each dataset, but no obvious sequence similarities were detected between different exons within a dataset. This pattern of phylogenetic conservation suggests that the functional intronic enhancer is somewhat larger than UGCAUG alone. Future studies with larger datasets will be essential to identify additional conserved elements, potentially exhibiting more sequence degeneracy, that are associated with the unique UGCAUG motif.

We considered the possibility that sequence conservation among intronic enhancers could be due to a shared origin of these elements, e.g. by Alu exonization (37). However, the brain-enriched exons in this study occur in non-primate mammals, birds, and even fish, indicating that they pre-date the evolution of Alu elements. In addition, direct sequence analysis revealed no evidence for homology of the brain-enriched exons and their putative intronic enhancers to Alu elements or to other common repeat elements in the human genome (data not shown).

Deficiency of G-triplets in the proximal downstream intron of brain-enriched datasets

It was also of interest to examine the distribution of G triplets, which have been reported to be over-represented in proximal downstream introns (14,38) and to enhance exon inclusion (39–41). Our earlier study reported that, in marked contrast to the over-representation of UGCAUG elements near brain-enriched exons in the human genome, the downstream intron sequence was substantially deficient in G triplets (15). To explore whether deficiency of G triplets adjacent to brain-enriched alternative exons is an evolutionarily conserved feature, we asked which trinucleotides were most under-represented (i.e. exhibited the highest negative contrast scores) in each of the species-specific datasets. Notably, GGG was the only trinucleotide to exhibit a top-ranking negative contrast

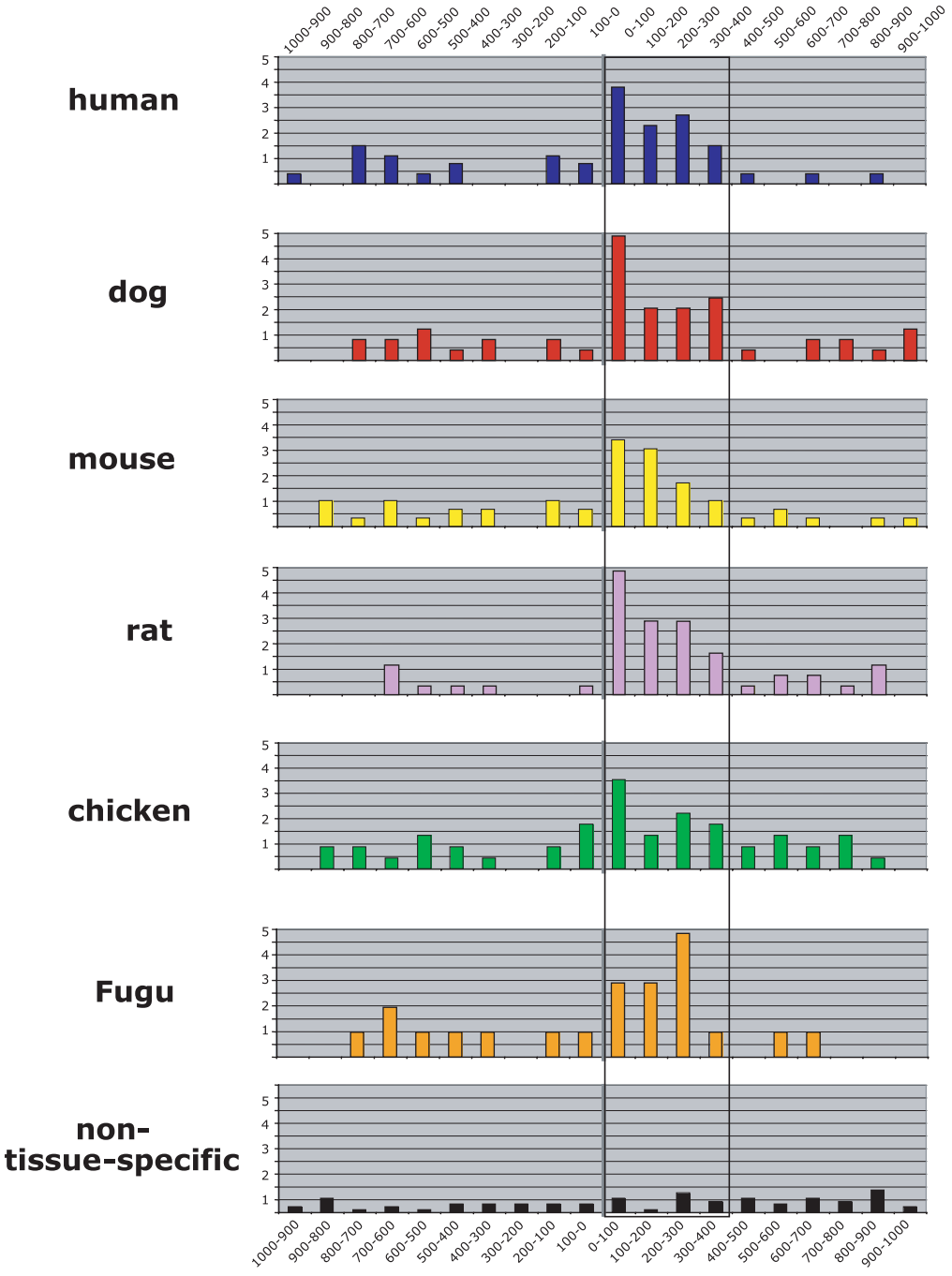

Figure 1. Enrichment of candidate UGCAUG alternative splicing enhancers in the proximal downstream intron region is conserved among vertebrate species. Histograms show the frequency of UGCAUG elements at 100 nt intervals in the introns upstream and downstream of the brain-enriched exons. Top six panels represent the analysis of intron sequences for the vertebrate species indicated at the left. Note the highest abundance of UGCAUG elements is consistently within the downstream ~400 nt (D400) region. The bottom panel shows the much lower incidence of UGCAUG elements that occurs near a group of non-tissue-specific alternative exons. Vertical axis, frequency; horizontal axis, nucleotide range relative to the alternative exon.

score in the D400 region of all the mammalian and avian species (Table 5). Interestingly, although intronic GGG triplets occur less frequently in pufferfish than in mammals (13), they still exhibited a negative contrast score in pufferfish indicating a reduced occurrence in regions flanking regulated alternative exons. G triplets are thus consistently much more abundant in the proximal downstream intron of constitutive exons than in the corresponding intron region of brain-enriched exons. Additional analyses revealed similarly that tetramer and hexamer sequences containing G-triplets were greatly under-represented in the brain-enriched datasets (data not shown). Interestingly, the tissue non-specific dataset also exhibited a significant deficiency of G triplets in the D400 region (Table 5). The latter observation suggests that lack of GGG enhancers in the downstream intron may be a general feature of many alternatively spliced exons, which might then

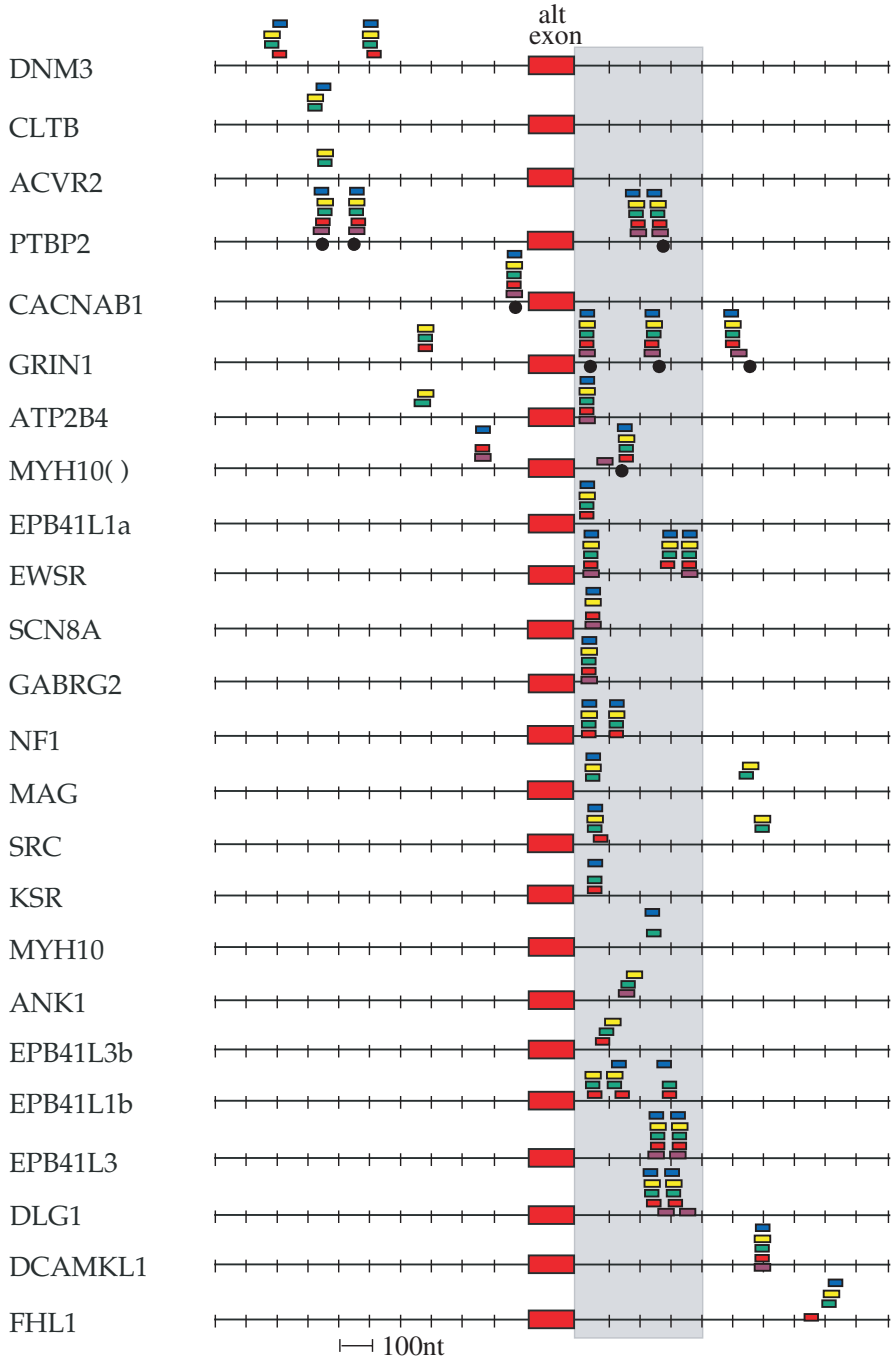

Figure 2. Specific location of UGCAUG hexamers near individual brain-enriched exons is highly conserved evolutionarily. The figure shows all of the brain-enriched exons for which phylogenetically conserved UGCAUG elements were identified in the flanking introns. Location of hexamers is indicated by small rectangles or circles that are color coded to indicate species. Blue, human; yellow, mouse; green, rat; red, dog; brown, chicken; and black circle, Fugu. Regulated brain-enriched exons are at the center; intron sequences to the left represent upstream sequences while those to the right represent downstream sequences.

require an alternative intronic enhancer (such as UGCAUG) to activate splicing in a regulated manner.

Confirmation of tissue-specific splicing in other species

Although alternatively spliced exons are not always conserved between human and mouse (42,43), a critical assumption underlying the analysis of our brain-enriched exon datasets is that the patterns of regulation for tissue-specific alternative exons are indeed phylogenetically conserved. Thus far, most of the published experimental data to support tissue-specific splicing of the exons in our datasets was focused only on the human genes. As an initial test of whether tissue-specific splicing patterns were conserved in other species, we examined the expression of selected exons in several mouse tissues. Figure 4 summarizes the results of an RT–PCR experiment looking at eight such exons. All eight were alternatively spliced, as indicated by the presence of products including

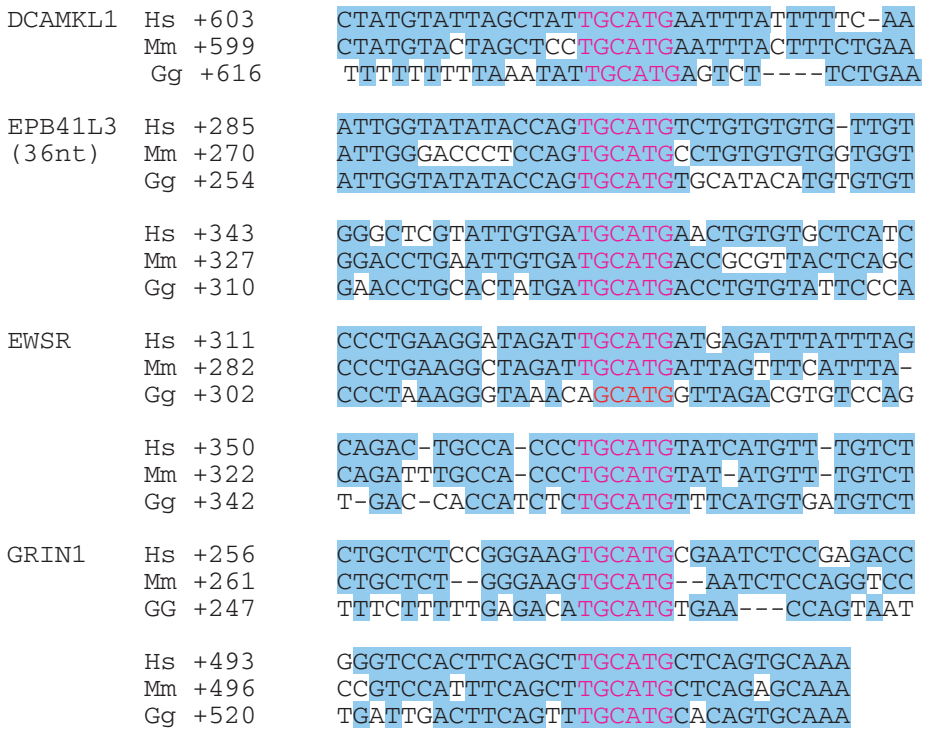

Figure 3. UGCAUG hexamers are often situated within larger regions of conserved intron sequence. Local alignments of nucleotide sequences flanking several conserved UGCAUG elements are shown. Nucleotides shared among at least two species are shaded. Notably, UGCAUG hexamers are located within regions of conservation that may be several hundred nucleotides distal to the regulated exon. Numbers on the left indicate the distance of the UGCAUG element from the exon.

Table 5. Most under-represented trinucleotides, D400 region

Human	Mouse	Rat	Dog	Chicken	Pufferfish	Non-tissue-specific
*GGG −2.19	AGG −1.62	AGG −0.94	GGC −1.90	*GGG −1.10	GAU −1.13	GGC −1.84
AGG −2.00	GGA −1.08	GAA −0.75	*GGG −1.76	GAG −1.09	GGA −0.93	*GGG −1.70
GGC −1.91	CAG −0.84	GGA −0.73	GCC −1.75	AGG −1.03	*GGG −0.68	GCC −1.59
GCC −1.53	*GGG −0.82	AAA −0.47	AGG −1.74	GUG −0.73	UGA −0.67	CCC −1.19
CAG −1.47	GAG −0.79	AAU −0.43	CCC −1.43	AGC −0.70	UGG −0.64	UGG −1.05
CCC −1.20	ACA −0.50	*GGG −0.43	CAG −1.32	GGC −0.65	GCA −0.62	CGG −1.01

*Indicates the position of GGG among under-represented trinucleotides in each dataset.

(filled arrowheads) or excluding (open arrowheads) the designated alternative exon. Moreover, all exhibited tissue-specific expression with modest to substantial inclusion in the brain RNA sample. However, whereas several were quite brain-specific (ACVR2, EPB41L1s, EPB41L3s, EPB41L3b), others exhibited dual specificity for brain and skeletal muscle (NF1, MYH10), or brain and lung (SCN8a), or they exhibited a more complex pattern (PTBP2). These results support the idea that the 'brain-enriched' exons in our databases do exhibit a phylogenetically conserved pattern of activated splicing in the brain. However, they also reveal unexpected complexity in splicing patterns in other tissues, further supporting the notion that UGCAUG elements alone are not sufficient to determine precise tissue specificity.

DISCUSSION

Previous studies have shown that UGCAUG is an intronic splicing regulatory element (18,24,25,27), that it is bound with unusually high specificity by the novel alternative splicing factor, Fox-1 (35), and that it is required for efficient splicing of neural-specific alternative exons (18,44). The demonstration that UGCAUG expression is widely associated with brain-enriched exons (15) and is evolutionarily conserved from fish to man (this paper) further implicates the Fox-UGCAUG system as an important regulator of tissue-specific alternative splicing. Taken together, these biochemical and computational findings suggest a model whereby Fox protein binding to UGCAUG plays a critical role in activating splicing switches for many alternative exons during neural development and differentiation.

As shown in this paper, the intronic UGCAUG motif is phylogenetically and spatially conserved near brain-enriched exons. The majority of these brain-enriched exons possess at least one copy of UGCAUG in the flanking intron, and the remaining few generally have closely related GCAUG pentameric motif(s) in the proximal intron [(15); and from results not shown]. Evolutionary conservation of intronic

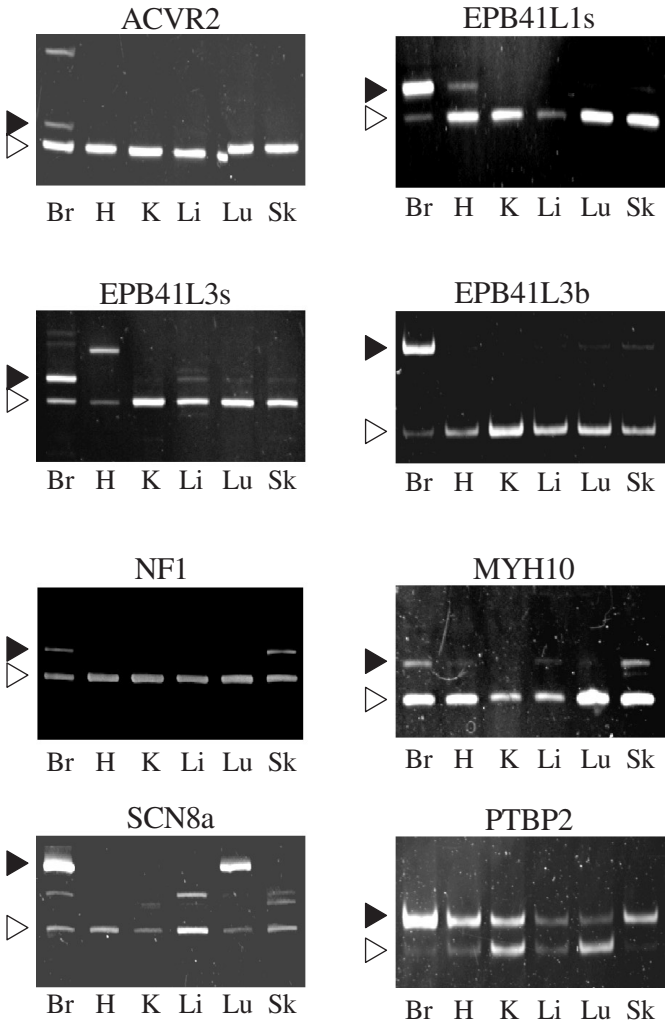

Figure 4. Alternative splicing of brain-enriched alternative exons is conserved in mouse. Alternative splicing of brain-enriched exons was examined by RT–PCR analysis of RNA isolated from six different mouse tissues. The figure shows polyacrylamide gel analyses indicating the amplification of brain-enriched isoforms (filled arrowheads) and non-tissue-specific isoforms (open arrowheads). The identity of the alternative exons is identified above each gel. Tissues used for PCR analysis: Br, brain; H, heart; K, kidney; Li, liver; Lu, lung and Sk, skeletal muscle.

UGCAUG elements has been noted previously (27). However, biochemical studies have demonstrated that in some cases the GCAUG pentamer is functional in splicing assays (25,35). We propose that the presence of linked (U)GCAUG elements may be an essential property for a specialized subset of alternative exons whose splicing is both tissue-specifically regulated and phylogenetically conserved. Consistent with this model, our studies did not identify UGCAUG in association with either constitutive exons or non-tissue-specific alternative exons. We speculate that other experimental approaches such as RESCUE-ISE (13) and the use of support vector machines (45) have not yet identified UGCAUG as a candidate regulatory element only because they have not been applied to the analysis of tissue-specific alternative exons. Almost certainly, these powerful techniques will be invaluable for future studies of splicing regulatory elements.

Alternative exons typically exhibit higher phylogenetic conservation of proximal intron sequences than do constitutive exons (43). Presumably, this conservation reflects a functional role in regulation of alternative splicing; our analyses begin to reveal specific functional motifs characteristic of these conserved introns. Besides the over-representation of UGCAUG elements, introns flanking alternative exons are modestly deficient in purines in general, and highly deficient in G-triplets in particular, compared with the introns flanking control exons. This was a robust finding in all six species-specific brain datasets examined in the study. Since G-triplets have been shown to enhance splicing of neighboring exons (39–41), it is tempting to speculate that the absence of this widely expressed class of intronic enhancers may contribute to the default skipping phenotype of many alternative exons. The ability to efficiently activate splicing of these exons might then require another class of intronic enhancer, e.g. the UGCAUG element or analogous motifs, in order to achieve tissue-specific regulation during development and differentiation. Alternative exons that lack both G-triplets and UGCAUG element(s) might represent a separate class that does not exhibit tissue-specific switching.

We propose that the functional importance of the UGCAUG motif is to serve as a critical *cis*-acting component of alternative splicing switches that trigger many developmental- and differentiation-specific changes in pre-mRNA splicing. Furthermore, we propose that Fox-1 related proteins represent the core of the conserved regulatory machinery that mediates these splicing switches. Direct experimental evidence for Fox-1 splicing regulation via (U)GCAUG motif(s) has been reported (35). Moreover, the RRM domains of human A2BP1/Fox-1 (46) and zebrafish Fox-1 (35) are highly conserved (~91% identity), and comparative genomic analysis indicates that this domain in the mouse, dog, rat and chicken orthologs is identical to human Fox-1 (our unpublished observations). Thus, the UGCAUG-Fox-1 system appears to have been highly conserved through evolution. The unique sequence specificity of this machinery may be an advantageous property for a splicing regulatory network designed to regulate a restricted population of exons in tissue-specific splicing patterns. However, a single isoform of Fox-1 alone clearly cannot account for the diversity of splicing patterns characteristic of complex genes in metazoan organisms. Indeed, the proteome's repertoire of Fox proteins may be fairly complex, involving multiple genes and multiple protein isoforms encoded by each gene. Future studies will be required to assess the relative RNA binding specificities and co-factor binding capabilities for each Fox isoform, in order to elucidate the rules that govern this regulatory system. Such studies may explain why the zebrafish Fox-1 preferentially binds the GCAUG pentamer in SELEX binding assays (35), yet the UGCAUG hexamer exhibits preferential conservation in the genome. Moreover, it seems likely that a variety of splicing co-factors must cooperate with Fox proteins to activate splicing switches for highly selective groups of exons at the appropriate time and place during development. Such co-factors may include members of the SR and hnRNP families, factors related to NOVA-1 (47) or CUG-binding proteins of the CELF family (22), or proteins yet to be characterized. Future experiments will be aimed at identifying additional motifs and co-factors that cooperate with UGCAUG-Fox to

determine precise timing of splicing switches in various cell types.

ACKNOWLEDGEMENTS

The authors thank Thomas Cooper for very helpful comments on the manuscript. This work was supported by National Institutes of Health (NIH) grant HL45182 and by the Director, Office of Biological and Environmental Research, US Department of Energy under contract DE-AC03-76SF00098. Funding to pay the Open Access publication charges for this article was provided by NIH grant HL45182.

REFERENCES

1. Zhu,J., Mayeda,A. and Krainer,A.R. (2001) Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol. Cell*, **8**, 1351–1361.
2. Charlet,B.N., Logan,P., Singh,G. and Cooper,T.A. (2002) Dynamic antagonism between ETR-3 and PTB regulates cell type-specific alternative splicing. *Mol. Cell.*, **9**, 649–658.
3. Rooke,N., Markovtsov,V., Cagavi,E. and Black,D.L. (2003) Roles for SR proteins and hnRNP A1 in the regulation of c-src exon N1. *Mol. Cell. Biol.*, **23**, 1874–1884.
4. Tacke,R. and Manley,J.L. (1995) The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *EMBO J.*, **14**, 3540–3551.
5. Coulter,L.R., Landree,M.A. and Cooper,T.A. (1997) Identification of a new class of exonic splicing enhancers by in vivo selection [Erratum (1997) Mol. Cell. Biol., **17**, 3468.]. *Mol. Cell. Biol.*, **17**, 2143–2150.
6. Liu,H.X., Zhang,M. and Krainer,A.R. (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev.*, **12**, 1998–2012.
7. Schaal,T.D. and Maniatis,T. (1999) Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. *Mol. Cell Biol.*, **19**, 1705–1719.
8. Cavaloc,Y., Bourgeois,C.F., Kister,L. and Stevenin,J. (1999) The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA*, **5**, 468–483.
9. Liu,H.X., Chew,S.L., Cartegni,L., Zhang,M.Q. and Krainer,A.R. (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol. Cell. Biol.*, **20**, 1063–1071.
10. Kim,S., Shi,H., Lee,D.K. and Lis,J.T. (2003) Specific SR protein-dependent splicing substrates identified through genomic SELEX. *Nucleic Acids Res.*, **31**, 1955–1961.
11. Wang,Z., Rolish,M.E., Yeo,G., Tung,V., Mawson,M. and Burge,C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831–845.
12. Fairbrother,W.G., Yeh,R.F., Sharp,P.A. and Burge,C.B. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.
13. Yeo,G., Hoon,S., Venkatesh,B. and Burge,C.B. (2004) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc. Natl Acad. Sci. USA*, **101**, 15700–15705.
14. Zhang,X.H. and Chasin,L.A. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.*, **18**, 1241–1250.
15. Brudno,M., Gelfand,M.S., Spengler,S., Zorn,M., Dubchak,I. and Conboy,J.G. (2001) Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucleic Acids Res.*, **29**, 2338–2348.
16. Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q. and Krainer,A.R. (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.
17. Fairbrother,W.G., Yeo,G.W., Yeh,R., Goldstein,P., Mawson,M., Sharp,P.A. and Burge,C.B. (2004) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.*, **32**, W187–W190.
18. Modafferi,E.F. and Black,D.L. (1997) A complex intronic splicing enhancer from the c-*src* pre-mRNA activates inclusion of a heterologous exon. *Mol. Cell. Biol.*, **17**, 6537–6545.
19. Blanchette,M. and Chabot,B. (1999) Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization. *EMBO J.*, **18**, 1939–1952.
20. Gromak,N., Matlin,A.J., Cooper,T.A. and Smith,C.W. (2003) Antagonistic regulation of alpha-actinin alternative splicing by CELF proteins and polypyrimidine tract binding protein. *RNA*, **9**, 443–456.
21. Forch,P., Puig,O., Martinez,C., Seraphin,B. and Valcarcel,J. (2002) The splicing regulator TIA-1 interacts with U1-C to promote U1 snRNP recruitment to 5′ splice sites. *EMBO J.*, **21**, 6882–6892.
22. Ladd,A.N., Charlet,N. and Cooper,T.A. (2001) The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. *Mol. Cell. Biol.*, **21**, 1285–1296.
23. Ule,J., Jensen,K.B., Ruggiu,M., Mele,A., Ule,A. and Darnell,R.B. (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science*, **302**, 1212–1215.
24. Huh,G.S. and Hynes,R.O. (1994) Regulation of alternative pre-mRNA splicing by a novel repeated hexanucleotide element. *Genes Dev.*, **8**, 1561–1574.
25. Hedjran,F., Yeakley,J.M., Huh,G.S., Hynes,R.O. and Rosenfeld,M.G. (1997) Control of alternative pre-mRNA splicing by distributed pentameric repeats. *Proc. Natl Acad. Sci. USA*, **94**, 12343–12347.
26. Kawamoto,S. (1996) Neuron-specific alternative splicing of nonmuscle myosin II heavy chain-B pre-mRNA requires a cis-acting intron sequence. *J. Biol. Chem.*, **271**, 17613–17616.
27. Lim,L.P. and Sharp,P.A. (1998) Alternative splicing of the fibronectin EIIIB exon depends on specific TGCATG repeats. *Mol. Cell. Biol.*, **18**, 3900–3906.
28. Parra,M., Gee,S., Chan,N., Ryaboy,D., Dubchak,I., Mohandas,N., Gascard,P.D. and Conboy,J.G. (2004) Differential domain evolution and complex RNA processing in a family of paralogous EPB41 (protein 4.1) genes facilitate expression of diverse tissue-specific isoforms. *Genomics*, **84**, 637–646.
29. Gallagher,P.G., Tse,W.T., Scarpa,A.L., Lux,S.E. and Forget,B.G. (1997) Structure and organization of the human ankyrin-1 gene. Basis for complexity of pre-mRNA processing. *J. Biol. Chem.*, **272**, 19220–19228.
30. Laura,R.P., Ross,S., Koeppen,H. and Lasky,L.A. (2002) MAGI-1: a widely expressed, alternatively spliced tight junction protein. *Exp. Cell Res.*, **275**, 155–170.
31. Burgess,H.A. and Reiner,O. (2002) Alternative splice variants of doublecortin-like kinase are differentially expressed and have different kinase activities. *J. Biol. Chem.*, **277**, 17696–17705.
32. Melot,T., Dauphinot,L., Sevenet,N., Radvanyi,F. and Delattre,O. (2001) Characterization of a new brain-specific isoform of the EWS oncoprotein. *Eur. J. Biochem.*, **268**, 3483–3489.
33. Itoh,K. and Adelstein,R.S. (1995) Neuronal cell expression of inserted isoforms of vertebrate nonmuscle myosin heavy chain II-B. *J. Biol. Chem.*, **270**, 14533–14540.
34. Rahman,L., Bliskovski,V., Reinhold,W. and Zajac-Kaye,M. (2002) Alternative splicing of brain-specific PTB defines a tissue-specific isoform pattern that predicts distinct functional roles. *Genomics*, **80**, 245–249.
35. Jin,Y., Suzuki,H., Maegawa,S., Endo,H., Sugano,S., Hashimoto,K., Yasuda,K. and Inoue,K. (2003) A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *EMBO J.*, **22**, 905–912.
36. Sorek,R. and Ast,G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, **13**, 1631–1637.
37. Sorek,R., Lev-Maor,G., Reznik,M., Dagan,T., Belinky,F., Graur,D. and Ast,G. (2004) Minimal conditions for exonization of intronic sequences: 5′ splice site formation in alu exons. *Mol. Cell*, **14**, 221–231.
38. Nussinov,R. (1989) Conserved signals around the 5′ splice sites in eukaryotic nuclear precursor mRNAs: G-runs are frequent in the introns and C in the exons near both 5′ and 3′ splice sites. *J. Biomol. Struct. Dyn.*, **6**, 985–1000.
39. Sirand-Pugnet,P., Durosay,P., Brody,E. and Marie,J. (1995) An intronic (A/U)GGG repeat enhances the splicing of an alternative intron of the chicken beta-tropomyosin pre-mRNA. *Nucleic Acids Res.*, **23**, 3501–3507.
40. McCullough,A.J. and Berget,S.M. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol. Cell. Biol.*, **17**, 4562–4571.

41. McCullough,A.J. and Berget,S.M. (2000) An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5′ splice sites. *Mol. Cell. Biol.*, **20**, 9225–9235.
42. Nurtdinov,R.N., Artamonova,I.I., Mironov,A.A. and Gelfand,M.S. (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum. Mol. Genet.*, **12**, 1313–1320.
43. Sorek,R., Shamir,R. and Ast,G. (2004) How prevalent is functional alternative splicing in the human genome?. *Trends Genet.*, **20**, 68–71.
44. Guo,N. and Kawamoto,S. (2000) An intronic downstream enhancer promotes 3′ splice site usage of a neural cell-specific exon. *J. Biol. Chem.*, **275**, 33641–33649.
45. Zhang,X.H., Heller,K.A., Hefter,I., Leslie,C.S. and Chasin,L.A. (2003) Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res.*, **13**, 2637–2650.
46. Shibata,H., Huynh,D.P. and Pulst,S.M. (2000) A novel protein with RNA-binding motifs interacts with ataxin-2. *Hum. Mol. Genet.*, **9**, 1303–1313.
47. Jensen,K.B., Dredge,B.K., Stefani,G., Zhong,R., Buckanovich,R.J., Okano,H.J., Yang,Y.Y. and Darnell,R.B. (2000) Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*, **25**, 359–371.