Published online 10 July 2007

Nucleic Acids Research, 2007, Vol. 35, No. 14 ***4845–4857***
doi:10.1093/nar/gkm485

A correlation with exon expression approach to identify *cis*-regulatory elements for tissue-specific alternative splicing

Debopriya Das[1], Tyson A. Clark[3], Anthony Schweitzer[3], Miki Yamamoto[1], Henry Marr[1], Josh Arribere[1], Simon Minovitsky[2], Alexander Poliakov[2], Inna Dubchak[2], John E. Blume[3] and John G. Conboy[1,*]

[1]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, [2]Affymetrix, Inc., Santa Clara, CA, 95051 and [3]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA

Received March 12, 2007; Revised June 4, 2007; Accepted June 5, 2007

ABSTRACT

Correlation of motif occurrences with gene expression intensity is an effective strategy for elucidating transcriptional *cis*-regulatory logic. Here we demonstrate that this approach can also identify *cis*-regulatory elements for alternative pre-mRNA splicing. Using data from a human exon microarray, we identified 56 cassette exons that exhibited higher transcript-normalized expression in muscle than in other normal adult tissues. Intron sequences flanking these exons were then analyzed to identify candidate regulatory motifs for muscle-specific alternative splicing. Correlation of motif parameters with gene-normalized *exon* expression levels was examined using linear regression and linear splines on RNA words and degenerate weight matrices, respectively. Our unbiased analysis uncovered multiple candidate regulatory motifs for muscle-specific splicing, many of which are phylogenetically conserved among vertebrate genomes. The most prominent downstream motifs were binding sites for Fox1- and CELF-related splicing factors, and a branchpoint-like element ACUAAC; pyrimidine-rich elements resembling PTB-binding sites were most significant in upstream introns. Intriguingly, our systematic study indicates a paucity of novel muscle-specific elements that are dominant in short proximal intronic regions. We propose that Fox and CELF proteins play major roles in enforcing the muscle-specific alternative splicing program, facilitating expression of unique isoforms of cytoskeletal proteins critical to muscle cell function.

INTRODUCTION

Alternative pre-mRNA splicing is a critical mechanism for regulating gene expression in metazoan organisms, and leads to tremendous protein diversity from a relatively small number of genes. A majority of human genes exhibit some form of alternative splicing. In particular, the human genome encodes a complex alternative splicing program that switches alternative exons on and off according to the needs of individual differentiated cell types. Despite intensive study in recent years, the mechanisms regulating the human alternative splicing program are not yet well understood. The complex decision process, involving which subset of exons on the primary RNA transcript (henceforth, pre-mRNA) will get spliced into the mature mRNA isoform, is mediated by a combination of *cis*-regulatory elements organized across exons and introns (1), quite analogous to the *cis*-regulation of transcription. Global identification of splicing regulatory elements has been difficult and has been primarily restricted to exonic elements (2–6), while limited computational information is available on intronic elements (7–13). However, availability of splicing microarrays (14–16), which can interrogate expression levels of exons genome-wide under any particular biological condition, has opened up new possibilities. In this work, we demonstrate that one can now apply analogous computational approaches used for dissecting transcriptional regulation (17) to decipher the splicing regulatory elements, with genes replaced by exons and promoters by pre-mRNA regions proximal to the splice sites.

A new set of approaches based on correlation with expression has been particularly successful in identifying *cis*-regulatory elements governing transcription (18–21). Here, the premise is that gene expression results from

*To whom correspondence should be addressed. Tel: +(510) 486 6973; Fax: +1 510 486 6746; Email: jgconboy@lbl.gov
Correspondence may also be addressed to Debopriya Das. Tel: +1 510 486 4281; Fax: +1 510 486 6746; Email: ddas@potternexus.lbl.gov

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

integration of multiple signals within the promoter region, as mediated by binding of *trans*-factors to the *cis*-elements. This implies that for an active *cis*-regulatory motif, its parameters [occurrence frequencies and position weight matrix (PWM) scores] must be significantly correlated with the expression levels across genes under any specific biological condition. Multiple studies have established that, using this strategy, one can identify the motifs that are functional under the tested condition. Furthermore, expression data from a single test condition and a reference condition are often sufficient for the analysis. In addition, unlike clustering-based approaches, interacting combinations of motifs can be inferred with high confidence (19,22). Finally, a recent study based on linear splines, which model the sigmoidal nature of transcriptional response, shows that such approaches can accurately identify direct targets of *trans*-factors binding to the active motifs, even when the motifs are very degenerate (22). Target identification in such situations has been quite challenging. Thus, one can delineate the key elements of transcriptional regulatory networks using correlation with expression. This has proven effective in both lower eukaryotes, e.g. yeast (18,19,23), and in mammals (22).

Here we report the first application, to our knowledge, of the correlation with expression approach for identification of *cis*-elements that regulate alternative splicing by integrating pre-mRNA sequence information with the exon microarray data. Specifically, we focused on tissue-specific splicing, as tissue-specific pre-mRNA regions are largely conserved across species (8,24,25), and thus, phylogenetic conservation can be used to evaluate the predictions. We employed an Affymetrix exon microarray (26) to identify 56 muscle-enriched alternative cassette exons, a number of which are predicted to alter the expression of cytoskeletal related genes. We used both linear regression (18) and linear splines (22) to examine whether *cis*-elements in introns adjoining these exons correlate with gene-normalized exon expression in muscle. Multiple motifs that demonstrated statistically significant correlation were also found to be conserved in mouse, chicken and frog. In addition, several of these elements have been previously characterized experimentally as regulators of muscle-specific splicing via binding to members of the Fox (27–33), CELF (34) and PTB (35) families of splicing factors. Taken together, our study shows that correlation with expression is indeed effective in deciphering splicing regulatory elements, and provides the most comprehensive picture yet available of muscle-specific alternative splicing program in humans.

MATERIALS AND METHODS

Identification of muscle-enriched alternative exon and control exon datasets

Total RNA from three biological replicates (three separate individuals) of 16 normal adult human tissues was purchased from BioChain (Hayward, CA, USA). Labeled target was generated from ~200 ng of total RNA and hybridized to a prototype version of the Affymetrix Human Exon Array as described (26). The set of microarrays contain ~1.4 million probesets designed to interrogate, as comprehensively as possible, more than 1 million exon clusters derived from a variety of input sources including annotated genes, cDNA sequences and exon prediction algorithms. Design information and microarray data is available at the GEO database (http://www.ncbi.nlm.nih.gov/geo/; accession number: GSE5791).

Candidate muscle-enriched probesets were identified using the splicing index approach (26,36,37). Exon-level expression was normalized to the expression level of the parent gene by dividing probeset intensities by the median intensity of probesets from exons supported by RefSeq or Ensembl annotations. Exons that exhibited statistically significant differences in inclusion rate were identified using a student's *t*-test on the gene-level normalized probeset intensities (NI). NI values from the three biological replicates of heart and skeletal muscle tissues were compared as a group to the replicates of 14 other non-muscle tissues as a second group. The magnitude of inclusion rate change (splicing index) was estimated by calculating a log ratio (base 2) of the median muscle NI and the median non-muscle NI (26,37). After filtering out non-expressed probesets and genes with low expression, probesets with *t*-test *P*-values <0.001 and splicing index magnitudes of >0.5 were considered candidates for muscle-enriched exons.

Manual filtering of the initial list was performed to select further for high confidence internal cassette exons, by mapping candidate muscle-enriched probeset to their genomic context using the BLAT tool (38) at the UCSC genome browser (http://genome.ucsc.edu). Probesets that overlapped annotated alternative transcriptional starts, alternative polyadenylation sites, or regions with alternative 5′ or 3′ splice sites, were removed from consideration in this study. Exon-level probeset intensities were additionally observed using BLIS (Biotique Systems, Inc. Reno, NV), an integrated genome browser that enables exon expression data from the microarray to be viewed in genomic context. Only probesets that showed clear patterns of muscle enrichment were kept for further analysis. Probesets had to demonstrate higher intensity levels in the muscle tissues and have exon-level data for surrounding probesets consistent with exon skipping in a majority of non-muscle tissues. Probesets were subsequently mapped to the May 2004 human genome (NCBI Build 35) using BLAT (38). Exact exon boundaries were determined by comparison to EST and mRNA sequences requiring consensus splice sites.

For phylogenetic analysis, the orthologous exons were identified in another mammalian genome (mouse; *Mus musculus*), in an avian genome (chicken; *Gallus gallus*) and in an amphibian genome (frog; *Xenopus tropicalis*) using VISTA alignment tools. Automatic alignment was successful at finding most of the longer alternative exons directly, but in a few cases the alignments were adjusted manually. The upstream 200 nt (U200) intronic region was selected as the base 1 to base 200 adjoining the exon in the upstream direction, while downstream 200 nt (D200) intronic region was selected as the base 1 to base 200 downstream of the exon. Alignments of orthologus introns and exons sequences were generated by LAGAN using default parameters (39).

The 'tissue-non-specific alternative' *exon* dataset was derived as described previously (8) from the European Bioinformatics Institute database of human alternative exons (http://www.ebi.ac.uk/asd/altextron/index.html). 'Control exon datasets' 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 were described previously (8). The muscle-enriched datasets and the control datasets is available at: http://vision.lbl.gov/People/ddas/NAR_SPLICE1/

Validation of muscle-enriched expression

A random subset of candidate muscle-enriched exons was selected for validation by RT–PCR, focusing (for ease of amplification) on those $\leqslant 155$ nt in length. RNAs from different human tissues, including heart, skeletal muscle and six non-muscle sources, were purchased from Clontech. One microgram of each RNA source was transcribed into cDNA using random hexamer primers in a total volume of 10 μl. Then, 2 μl cDNA was amplified in a volume of 25 μl, using primers located in the flanking constitutive exons (Supplementary Table 2), for 35 cycles under the following conditions: 30 s at 94°C; 30s at 55°C; 45 s at 72°C. The identity of PCR products was confirmed by DNA sequence analysis.

Correlation with expression

Linear correlation. Counts of hexamers were obtained in a specific pre-mRNA sequence region (upstream or downstream proximal intron). For each region, a linear model was fitted between the logarithm of ratios of gene-normalized exon expression levels and count of each 6-mer word w across a set of exons, $\{n_w^e\}$:

$$\log(\mathrm{NI}_e/\mathrm{NI}_{eC}) = a_w + b_w \cdot n_w^e$$

NI_e is the gene-normalized expression level of exon e in muscle, and C refers to a reference sample. The reference data was taken as the average NI across all tissues. The coefficients a_w and b_w were obtained by a least squares fit. P-values were calculated using an F-test, as described previously (40).

The best fit was obtained for a set of sequences that included the muscle-specific exons (foreground set) and a background set of m sequences ($m = 300$), drawn randomly from a set of manually curated 957 cassette exons across the human genome (11). Since we started with a prioritized set of tissue-enriched sequences, a background set was necessary to model the correct dependence of log ratios on word count. n such random draws were performed ($n = 25$), and a linear fit was obtained for each such draw. A geometric mean of the P-values from all iterations reflects the overall significance of the word.

Linear splines. Linear splines differ from lines by introducing a threshold, called knot, below which the function is constant and linear above it (19,22). A significant difficulty in modeling binding sites via PWMs is that they give rise to a continuous distribution of scores across all possible binding sites. Consequently, a cutoff score needs to be determined to discriminate the true sites from false sites. Such cutoff scores are often based on predetermined background sequences and thresholds, and as a result, are complicated by subjective choices (41). In a linear spline model, the cutoff score corresponds to the knot, and thus, is learnt directly from the input data (22). For each PWM μ of width L, each L-mer in the input sequence was assigned a probability score M:

$$M = [p_1(b_1)p_2(b_2)\ldots p_L(b_L)]^{1/L}$$

where $p_i(b_i)$ is the probability of observing the base b_i at the position i. Thus, the score M always assumes a value between 0 and 1. It is related to binding affinity (42). PWM scores across exons $\{M_e^\mu\}$ for a given motif μ were fitted to the splicing ratios $\{\log(\mathrm{NI}_e/\mathrm{NI}_{eC})\}$ using the following model:

$$\log(\mathrm{NI}_e/\mathrm{NI}_{eC}) = a_\mu + b_\mu \sum_{M_e^\mu > \xi_\mu} \theta\left(M_e^\mu - \xi_\mu, 0\right)$$

where $\theta(x,0)$ is a linear spline: it is x, when $x \geqslant 0$, and zero, otherwise. ξ_μ, termed knot, corresponds to the cutoff score. The coefficients a_μ and b_μ and the location of the knot ξ_μ were determined by a least squares fit. This leads to an unbiased and adaptive determination of the knot ξ_μ for any given PWM. Importantly, in contrast to previous approaches (22), where contribution from only the maximum scoring site was considered, we systematically accounted for contribution from active sites with weaker scores as well. The number of such active sites is adaptively learnt, as displayed in the equation above. Thus, both binding affinity and occurrences of active motifs are accounted for in our approach. The significance of the fit was assessed using an F-test (40). The overall significance of each PWM was enumerated using the same iterative procedure as for the linear regression discussed above.

Over-representation analysis

RNA words. We examined over-representation of candidate oligonucleotide sequences (RNA words) in each tissue-specific dataset, relative to the control datasets, using a hypergeometric distribution. The results were corrected for multiple testing using the false discovery rate (FDR) method (43). The results of this test were generally consistent with the non-parametric approach that we have described previously (7). Furthermore, for each word a contrast score was also calculated as the difference in frequency in the tissue-specific dataset versus the control dataset. Similar results were obtained using two control sets, one composed of predominantly constitutive exons, and the other containing alternative but nontissue-specific exons (8). Like standard motif analyses, repeat elements were not explicitly excluded from this analysis. They are automatically filtered by the correlational analyses. Moreover, only non-overlapping motifs were counted in the word frequency calculations. Manual examination of the sequences revealed no cases of long repeating elements that would influence frequency calculations of the candidate regulatory motifs.

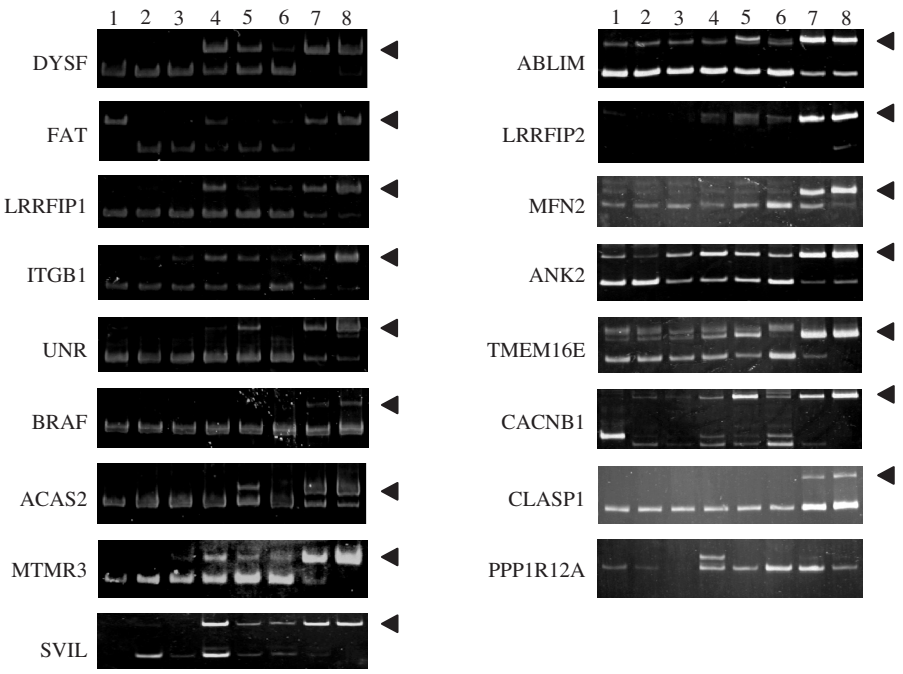

Figure 1. Validation of microarray predictions of muscle-enriched alternative exons. RT-PCR confirmation of muscle-enriched alternative exon expression. Amplifications were performed using primers in the flanking constitutive exons. RNA sources used for amplification by lane number: 1, brain; 2, kidney; 3, liver; 4, stomach; 5, bone marrow; 6, testis; 7, heart; 8, skeletal muscle. Arrowheads indicate positions of the alternative exon inclusion products that are most enriched in heart and skeletal muscle.

PWMs. Over-represented PWMs were obtained using the DME (Discriminating Matrix Enumerator) algorithm (44,45). DME is an enumerative search algorithm that finds the PWMs over-represented in a foreground set relative to a background set. Both intronic regions (upstream and downstream) were searched for over-represented matrices of width 6 nt, using the background sets as above. Default parameter settings were used, except that we varied the average information content of the PWM from 1.0 to 2.0 in steps of 0.1. Fifteen PWMs were obtained for each such setting. Correlation analysis was performed on non-redundant sets of matrices. Matrix similarity was assessed using MatCompare (46).

RESULTS

Identification and characterization of muscle-enriched alternative exons

The human muscle-enriched exon dataset analyzed in this study (Supplementary Table 1) was derived from exon microarray hybridization data using a platform designed to provide a comprehensive genome-wide analysis of annotated and predicted exons (see Methods section). In order to identify motifs that regulate tissue-specific alternative splicing, it is critical to identify a set of alternative exons having similar expression patterns indicative of regulation by a shared splicing program. Therefore, the group of exons studied here was carefully selected by analysis of exon microarray data from a panel of 16 normal adult human tissues. Probesets that exhibited gene-level NI that were significantly higher in heart and skeletal muscle, relative to 14 other tissues, were first identified. For this part of the analysis, we grouped the heart and skeletal muscle exon expression together to enhance the power of the statistical tests. Then a manual filtering process was performed so as to retain only probesets representing cassette exons, and to eliminate probesets corresponding to alternative first and last exons or to exon regions generated from alternative 5′ and 3′ splice sites. The final dataset consisted of 56 muscle-enriched, internal cassette exons. Most of these exons (~80%) are integral multiples of 3 nt in length, with a median length of 84 nt, consistent with the notion that alternative exons are smaller than average constitutive exon length [~145 nt (47,48)]. However, the genes with such alternative exons have a median size of 123 kb, much longer than the average gene length. To explore evolutionary conservation of candidate splicing regulatory elements, we also identified highly conserved orthologs for most of these human muscle-enriched exons in mouse, chicken and frog (Supplementary Table 1). It is important to note that while many of these exons show evidence of alternative splicing in Genbank, most were not previously known to exhibit muscle-enriched splicing and were not identified in the pilot study of muscle-enriched exons by Sugnet *et al*. (11). Therefore, analysis of this dataset should yield novel insights into the vertebrate muscle alternative splicing program, and should provide an opportunity to explore computationally the regulatory motifs that carry out this program.

Muscle-enriched splicing patterns for a random subset of these exons were validated experimentally in the human dataset by RT-PCR (Figure 1). Although splicing patterns were not absolutely muscle specific, in almost every case

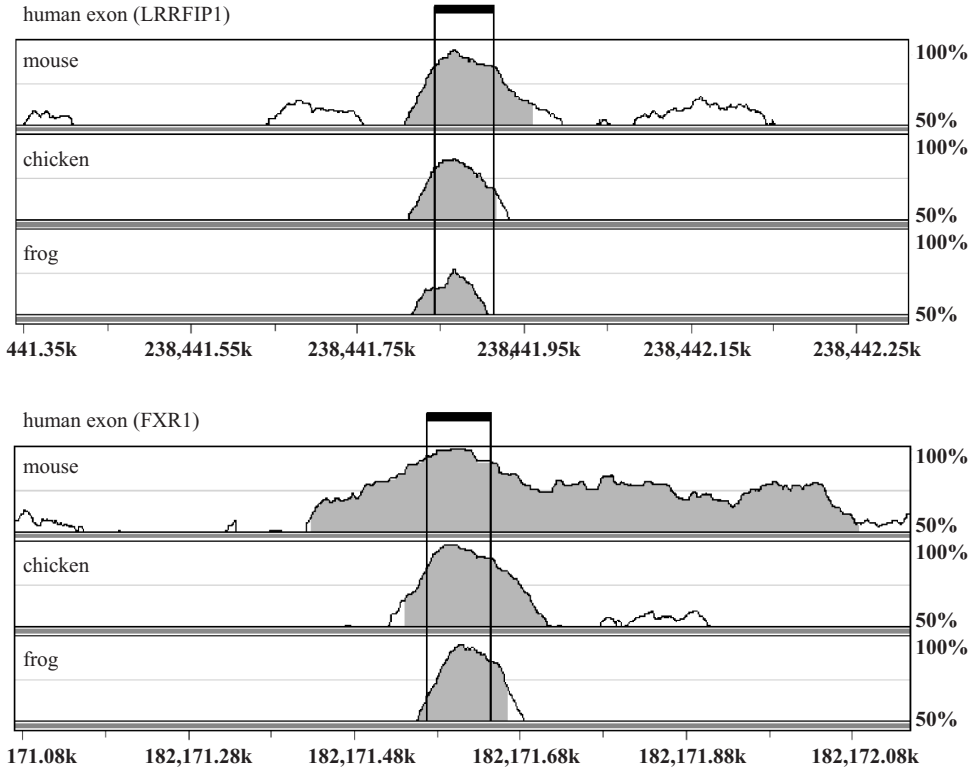

Figure 2. Conservation of intron sequences flanking muscle-enriched exons. Representative VISTA genome alignments of exon and flanking intron sequences from the mouse, chicken and frog genomes with the prototypical muscle-enriched exon from human. Exon boundaries are indicated by vertical lines. Shaded regions indicate sequences that exceed 75% identity, while curves above baseline indicate regions with >50% identity to the human sequence.

the efficiency of exon inclusion was highest in heart and skeletal muscle, confirming the predictions of the exon microarray. Importantly, mRNA and/or EST evidence from the genetic databases (data not shown) demonstrates that the majority of these exons are alternatively spliced in at least one of the other species examined (mouse, chicken or frog), suggesting that the incidence of conserved alternative exons in this specialized dataset is higher than the reported rate for general alternative exons (49). Taken together, these results indicate that the muscle-enriched exons constitute a special class of highly conserved alternative exons.

Intron sequences flanking orthologous alternative exons in the mouse and human genomes tend to be evolutionarily conserved (24), consistent with the observation that *cis*-regulatory elements for tissue-specific alternative splicing are often located in those proximal intron regions. We used VISTA genome alignment tools to compare the proximal intron sequences in this muscle-enriched dataset and extended the evolutionary comparison to include chicken and frog. In the proximal 200-nt upstream (U200) and downstream (D200) introns, mouse sequences were highly similar to their human orthologs (median identity of 61 and 58%, respectively), while chicken and frog introns were much less homologous. The full quantitative data are shown in Supplementary Table 3 and representative alignments of exons with relatively high conservation (FXR1), or lower conservation restricted mainly to the exon (LRRFIP1), are displayed in Figure 2. The reduced overall homology of chicken and frog introns suggests that conserved motifs in these regions are likely conserved specifically for their function as *cis*-regulatory elements for muscle-specific alternative splicing, rather than being passively conserved as part of a larger conserved element.

Frequent occurrence of muscle-enriched exons in genes encoding proteins with functions in cytoskeletal organization

Previous studies have demonstrated that the brain-specific alternative splicing factor, NOVA1, modulates the splicing of many components of the neuronal synapse (50). We hypothesized that the muscle alternative splicing program might similarly coordinate the expression of a particular class of genes that share a common pathway or cellular process. Using the method described previously (22,51), to examine the gene ontology (GO) terms associated with each parent gene for the muscle-enriched exons, we found a strong association with cytoskeleton organization and biogenesis, microtubule stabilization and muscle development (Supplementary Table 4). These associations were statistically significant ($P<0.001$), suggesting that the muscle alternative splicing program is critical for proper expression of the unique cytoskeleton characteristic of vertebrate muscle.

Correlation with exon expression identifies splicing regulatory elements

Alternative splicing regulatory elements responsible for tissue-specific splicing are often located in proximal intron sequences (25). To search for candidate intronic regulatory motifs for the muscle-specific splicing program, we correlated the frequencies of hexamers in specific intronic regions with the logarithm of ratios of gene-normalized exon expression levels in skeletal muscle, across the 56 muscle-enriched exons in the human dataset. The ratio for any exon was enumerated against its average gene-normalized expression level across all the tissues. Thus, it is similar to the splicing index used above. The *cis*-elements exhibiting significant correlation with expression were considered potentially functional in regulating muscle-specific splicing. These were further examined for relative over-representation in introns of muscle-enriched exons, compared to a background set of introns flanking constitutive exons, using a hypergeometric distribution based on word counting in the oligonucleotide sequences. Finally, we examined their spatial conservation through vertebrate evolution (8) by testing whether the motif is over-represented in the other species using exactly the same statistical measures as used for humans.

Here we consider UGCAUG in the downstream 200 nt (D200) of intron sequence as an example. This hexamer represents the binding site for mammalian Fox-1 and Fox-2 splicing factors (31), which have identical RRM domains. UGCAUG has been reported as a common motif in proximal introns adjacent to tissue-specific exons. In a few cases, functional splicing assays have confirmed the importance of this motif in regulation of splicing (27–33). In the large group of muscle-enriched exons studied here, we found a highly significant correlation of UGCAUG frequency with muscle expression ($P = 6.8E{-}05$). The distribution and the linear fit for a single iteration of correlation analysis (see Methods section) are shown in Figure 3A. Similar analysis shows that muscle expression does not correlate with UGCAUG occurrences in the upstream intron, whereas in the downstream intron the magnitude of correlation decreases with distance from the exon as demonstrated by the increasing *P*-values (Figure 3B). These dependencies are further corroborated by strong over-representation of this motif in proximal downstream introns in human and other vertebrates (Figure 3C). Indeed, almost half of the muscle-enriched exons in all four datasets (23/56 in human, 21/54 in mouse, 20/43 in chicken and 19/36 in frog) possessed at least one UGCAUG motif in the first 200 nt of the downstream proximal intron. Together, these results strongly support the hypothesis that UGCAUG is potentially an important regulatory element for muscle-specific alternative splicing, as predicted by the correlation with expression analysis.

Analysis of upstream and downstream intron sequences

We extended the above analysis to identify additional muscle-specific *cis*-elements in upstream and downstream intron sequences. In the downstream 200 nt sequence, we searched all possible hexamers and identified a total of 35 hexamers that were significantly correlated with expression ($P \leqslant 0.05$) and also over-represented in the human dataset ($P \leqslant 0.05$, $q \leqslant 0.2$); nine of these were also over-represented in at least one other species (Table 1A and Supplementary Table 5A). Several of these elements have been previously characterized experimentally as regulators of muscle-specific splicing. These motifs fell into three distinct classes: (i) the Fox1/2-binding motif UGCAUG ($P = 6.8E{-}05$) and two closely related hexamers (GCAUGG, UUUGCA); of note, in all four species the majority (58–76%) of GCAUG motifs in the D200 region occurred in the context of the full UGCAUG hexamer. (ii) UG-rich elements GUGUGU and UGUGUC (correlation *P*-values = 0.032 and 0.005, respectively), that resemble binding sites for the CELF family of splicing factors; and (iii) the novel motif ACUAAC ($P = 0.0006$) and related hexamers CUAACC ($P = 0.004$) and CACUAA ($P = 0.04$). The latter class is similar to the UACUAAC element noted in a recent study of a small group of muscle-specific exons in mouse (11). The distribution of these elements in flanking introns of exons in the human, mouse, chicken and frog datasets is shown in Figure 4A. Importantly, this analysis revealed that UGCAUG was the most over-represented hexamer in all four datasets, and both GUGUGU and ACUAAC were also consistently in the top ~1% of the most over-represented hexamers in these species.

For upstream intron sequences (200 nt), we found a total of 27 hexamers that were significantly correlated with expression and also over-represented in human, of which three were over-represented in at least one other species (Table 1B and Supplementary Table 5B). Many such elements are strongly pyrimidine-rich, characteristic of binding sites for PTB protein, an inhibitor of splicing for many alternative exons (35). In all four species, the muscle-enriched datasets showed strong over-representation of the reported PTB-binding sites, CUCUCU and UCUU, in the proximal upstream intron (Figure 4). CUCUCU was concentrated mainly in the U200 region. UCUU was focused even more tightly in the U100 region (Figure 4B), where it was consistently among the top five over-represented tetramers in all four species. Lesser over-representation of UCUU motifs over a broad area of downstream intron sequences was also noted, perhaps consistent with previous findings that optimal splicing repression by PTB requires binding sites both upstream and downstream of the regulated exon (52,53).

Many of the remaining significant hexamers, both for upstream and downstream introns, have low similarity to the previously discovered elements. Although these may represent novel elements, given that splicing elements are often degenerate, they can also be specific examples of known degenerate motifs. Our analysis using degenerate motifs presented below suggests that the latter possibility is more likely. Finally, for some of the major splicing regulatory elements described above, we observed that the profiles of positional over-representation have been conserved through vertebrate evolution: mouse, chicken and frog. This is displayed in Figure 4 for Fox, CELF and PTB-binding sites. Such strong positional conservation of

Table 1. Significant words identified by linear correlation analysis. Top 10 words in (Panel **A**) downstream and (Panel **B**) upstream introns (length = 200 nt).

Word	Correlation analysis *P*-value	Over-representation analysis		Contrast score	Phylogenetically conserved?	Putative *trans*-factors
		P-value	*q*-value			
Panel A: D200						
UGCAUG	6.8E−05	1.7E−15	6.1E−12	0.0024	Frog, mouse, chicken	Fox-1
ACUAAC	0.0006	2.5E−08	2.3E−05	0.0008	Frog, mouse, chicken	*
GCAUGG	0.0006	3.6E−05	0.004	0.0009	Mouse	Fox-1
CGUGUG	0.0007	0.009	0.12	0.0005		CELF
GCAUGA	0.002	0.002	0.04	0.0006		Fox-1
AGCAUG	0.002	0.0007	0.02	0.0007		Fox-1
UAACCC	0.003	9.6E−05	0.008	0.0006		
CUAACC	0.004	2.9E−05	0.004	0.0007	Frog	*
CACCAA	0.005	0.005	0.08	0.0004		
UGUGUC	0.005	0.006	0.09	0.0007	Chicken	CELF
Panel B: U200						
CCCCUU	0.002	9.8E−05	0.004	0.0009		
UUUCCA	0.002	0.0006	0.02	0.0009		PTB
UCCUCC	0.002	8.3E−05	0.004	0.0007		
UCUCCA	0.002	0.0002	0.007	0.0006		
AUCUCC	0.003	0.02	0.19	0.0002		
CCCCCU	0.003	0.03	0.2	0.0004	Frog	PTB
UCUUUC	0.004	1.1E−07	1.8E−05	0.0020		
CUCCUC	0.006	0.003	0.05	0.0004		
UCAUCU	0.007	0.001	0.02	0.0005		
AAAUCU	0.009	0.003	0.05	0.0005		

Asterisk indicates the previously identified, but as yet uncharacterized, novel element ACUAAC. *q*-value indicates multiple testing correction using the false discovery rate (FDR) method (43). Phylogenetic conservation was assessed by examining relative over-representation of each word in each species and employing the same *P*-value cutoffs as in human ($P \leqslant 0.05$, $q \leqslant 0.2$). Complete list of significant words is shown in Supplementary Table 5.

motifs lends additional support to our findings using correlation with expression.

The Fox-binding site, UGCAUG, was previously shown to be over-represented downstream of brain-enriched alternative exons (7,8), raising the possibility that the brain- and muscle-specific alternative splicing programs might exhibit functional similarities by sharing related components of the splicing machinery. To determine which of the candidate muscle *cis*-regulatory elements might be shared with brain-specific alternative splicing, and which are unique to the muscle program, we compared the frequency of several key *cis*-regulatory motifs in muscle (this study) and brain (8) datasets. Two elements, ACUAAC and UGUGUG, were clearly muscle specific since their frequencies were consistently higher in the D200–D300 region adjacent to muscle-enriched exons compared with the intronic region downstream of brain-enriched exons (Supplementary Figure 1; positive contrast scores). These motifs were also not over-represented in brain relative to control exons (data not shown). In contrast, the motifs UGCAUG and CUCUCU occurred at even higher frequencies in the proximal introns of the brain-enriched dataset than they did in the muscle-enriched dataset (Supplementary Figure 1; negative contrast scores for UGCAUG in the D200–D400 region, and for CUCUCU in the U100 region). Essentially equivalent distribution patterns were observed in the mouse, chicken and frog datasets (data not shown). These results strongly suggest that tissue-specific alternative splicing programs may utilize a combination of unique and shared *cis*-regulatory motifs that will require much additional analysis in the future.

Motifs identified via PWM analysis are consistent with word analyses

Because many splicing factors bind degenerate oligonucleotide sequences in RNA, we extended our analyses to include degenerate motifs through the use of PWMs (3,54). PWMs are probabilistic representations of degenerate binding sites. Over-represented PWMs in introns of 56 muscle-specific exons in the human dataset were obtained using the DME algorithm (44,45). We scanned multiple parameter settings of DME in order to obtain a large number of PWMs and reduce bias from DME. To identify the functional PWMs, we assessed their correlation with muscle expression using linear splines (19,22). Linear splines are among the simplest non-linear variants of linear models. In contrast to many other approaches, they facilitate adaptively learning the cutoff scores of PWMs that discriminate true targets from false targets of *trans*-factors. Previous regression approaches have used either maximum score of the PWM (22) or a global average of PWM scores for all potential binding sites (20) on an input sequence as the predictor variable. However, realistically, a small number of sites, sometimes >1, are bound by the corresponding *trans*-factor. Here we overcame this limitation by including both strength of PWM and the number of putative binding sites in our linear splines approach (see Methods section).

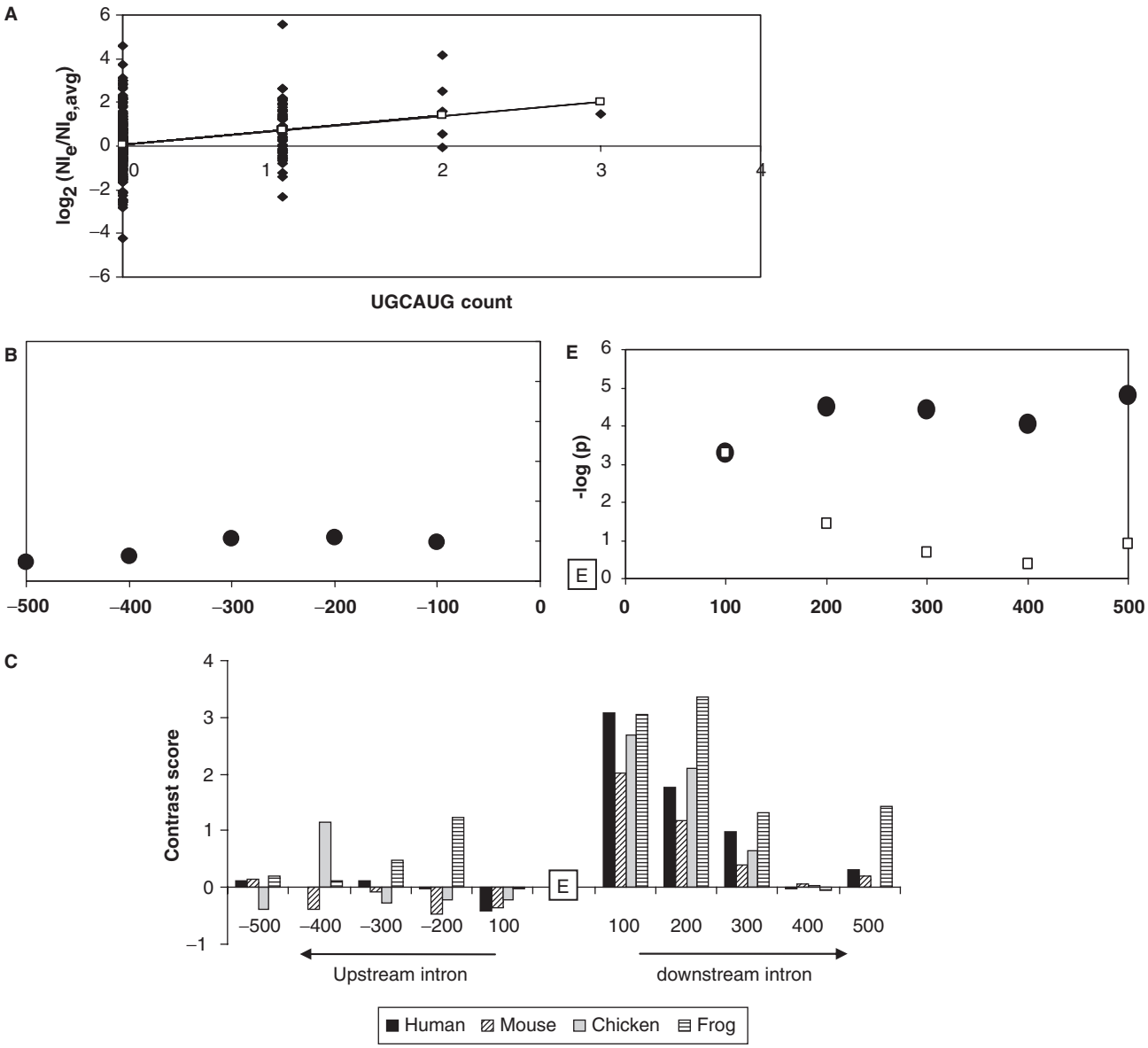

Figure 3. Correlation with exon expression for the UGCAUG regulatory element. (**A**) Linear fit between ratios of gene-normalized exon expression levels and counts of UGCAUG in 200 nt of downstream intronic sequence across 356 exons (56 muscle-specific exons and 300 randomly selected exons) ($P = 4.6E{-}06$). (**B**) Dependence of correlational *P*-values of UGCAUG count with distance. Filled circles indicate correlation with total motif count, while unfilled boxes indicate correlation with bin-wise count (bin size = 100 nt). E.g. for 300 nt, the filled circle reflects the strength of correlation with the count in 1–300 nt of intron, while the unfilled box reflects the correlation with the count in 201–300 nt of intron. Minus sign indicates upstream intron. (**C**) Contrast scores of UGCAUG in upstream and downstream introns of human, mouse, chicken and frog. ('E' in Figure 3B and 3C indicates position of the exon).

Degenerate 6-nt and 4-nt sequences that were over-represented in the proximal downstream intron sequence are shown in Table 2A and B and Supplementary Table 6A and B. Notably, all of the top 10 over-expressed PWM hexamer motifs in the D200 region are consistent with the major over-expressed unique motifs identified above. Among these, the six most statistically significant motifs represent close matches to the Fox-binding site, UGCAUG; two (NHCUAA and HCUAAN) are very similar to the novel ACUAAC element; and the remaining motifs (SUKUGS and CUGYSR) resemble UG-rich-binding site for CELF proteins. Analysis of over-expressed 4-mers in the D50 region revealed that the top-scoring motif is UGCM. While this motif is included in the Fox recognition sequence, other considerations suggest that it would not be sufficient for Fox binding. Instead, UGC likely represents the CUG-rich sequences, characteristic of some CELF-binding sites.

In the U200 region, all of the statistically over-represented motifs were quite pyrimidine-rich relative to the control group (Table 1B and 2C). Further investigation will be required to determine whether these elements are primarily bound by the PTB protein or by additional splicing factor(s). All remaining PWMs that have high significance in upstream and downstream introns exhibit at least partial similarity to the above three elements.

Table 2. Significant position weight matrices (PWMs) identified by correlation analysis using linear splines.

Consensus	*P*-value	Sequence logo
Panel A: D200		
WGCATK	2.3E−07	
WGMHTD	3.4E−07	
GCATRN	8.2E−07	
DWGCAT	3.1E−06	
NWGMWT	3.7E−06	
WGHHTD	7.1E−05	
NHCTAA	1.2E−04	
STKTGS	2.0E−04	
CTGYSR	3.4E−04	
HCTAAN	3.4E−04	
Panel B: D50		
TGCM	4.8E−04	
GCAT	2.8E−03	
CTTG	0.04	

(Continued)

Table 2. Continued.

Consensus	*P*-value	Sequence logo
Panel C: U200		
VYCCHT	9.5E−05	
YMCYYN	1.1E−04	
TYYCCM	5.1E−04	
CCCTNM	8.7E−04	
YMCYYW	9.0E−04	

Panel **A**: Top 10 PWMs of width 6 nt in proximal downstream intron sequences (length = 200 nt).
Panel **B**: All significant PWMs of width 4 nt in proximal downstream intron sequences (length = 50 nt).
Panel **C**: Top 10 PWMs of width 6 nt in upstream intron sequences (length = 200 nt). Complete list of significant PWMs in downstream and upstream 200 nt introns is shown in Supplementary Table 6.

For PWMs with only partial similarity, the similarity is observed either at the 5′ or at the 3′ end of the motif, indicating that the remainder of the motif most probably represents the flanking region. For example, for the PWM RRWGCA, the last four bases match the 5′ end of the UGCAUG, and hence, the first two bases are presumably the flanking region of this putative Fox-binding motif.

Furthermore, in contrast to previous work (22), the new formulation of linear splines used here allowed us to obtain not only the potential target exons, but also the binding sites of the above splicing factors (see Methods section). The results for a representative set of motifs are summarized in the Supplementary Table 7. For the putative Fox-binding motif WGCAUK, we find UGCAUG as the most frequently occurring oligonucleotide sequence, as expected of Fox-binding sites. We have observed similar accuracy in binding site prediction in the context of transcriptional regulation (Das,D., unpublished data). Interestingly, we notice that not all possible combinations of nucleotides of a degenerate PWM are realized in the set of 56 muscle-specific exons. For example, for the candidate CELF-binding motif, CUGYSR, only CUGUGA is predicted as the binding site. These are consistent with the previous observations made in the context of transcriptional regulation (55,56).

DISCUSSION

In this study we have demonstrated that the correlation with expression approach, applied to global exon expression profiles, represents a powerful new tool for identification of *cis*-regulatory motifs for alternative splicing. Using a dataset of high-confidence muscle-enriched alternative exons extracted from human exon microarray data, we correlated motif occurrences in the flanking introns with the splicing index measure of relative muscle enrichment to identify candidate regulatory motifs for the muscle-splicing program. The logic of this strategy is supported by many studies of transcriptional regulation, and a few of splicing regulation (57), showing that functional response often correlates with regulatory motif copy number. The analysis presented here demonstrates that the number of Fox splicing factor binding sites (UGCAUG) correlates strongly with the muscle splicing index (Figure 3A), consistent with previous reports that Fox proteins can regulate various tissue-specific alternative splicing events. The validity of correlation results were further supported by over-representation analysis, by comparative genomics showing that top scoring correlation motifs are phylogenetically conserved among vertebrate genomes, and by previous experimental studies implicating most of the same motifs in regulation of muscle-specific exon(s). Since tissue-specific alternative splicing is rarely an all or nothing phenomenon (e.g. Figure 1), correlation with expression may offer an attractive approach toward understanding complex tissue-specific patterns of alternative splicing. This approach may be particularly effective when PWMs are utilized in the splines-based framework to account simultaneously for both relative affinity and number of motif occurrences, providing insight into both the target exons and binding sites associated with a given motif.

Our immediate goal here in this proof of concept study was to examine whether the correlation with expression method can be used to identify splicing regulatory motifs, and consider muscle-specific alternative splicing program as an example of this application. This analysis strongly implicated several classes of known regulatory factors including Fox (UGCAUG), CELF (GUGUGU and UCUGUG), PTB (CUCUCU and UCUU) and putative KH-type splicing factor (ACUAAC) as important mediators of muscle-enriched splicing. The current study thus confirms and substantially extends earlier reports that these factors can regulate one or a few muscle-enriched exons by providing significant new computational evidence that they correlate with muscle exons in a much larger dataset. Interestingly, there was a notable lack of novel *cis*-elements in the proximal flanking introns that strongly correlate with muscle expression across the entire dataset. This could indicate that much of the fundamental machinery for regulation of generalized muscle-enriched splicing has been identified or, more likely, that additional features need to be incorporated in the algorithms to identify the remaining components. Such features may include wider motifs and motifs located more distally from the regulated exons. It is also possible that there are weaker elements, which may only be revealed when combinatorial interactions among motifs are included in the regression models, or which may be required for spatially or temporally distinct subsets of muscle-enriched exons. To obtain an initial estimate as to which of these factors may be most influential, we extended our study to include PWMs of width 5–7 nt. The results are displayed on our website (http://vision.lbl.gov/People/ddas/NAR_SPLICE1/). We observe that most motifs have similarities to the known motifs as identified above. There is one motif in D200, GGSYVYW, which seems novel. But since it has much higher *P*-value than others ($P = 0.01$), it is not readily clear if it is truly functional. Hence, we suspect that inclusion of combinatorial interactions among motifs may be most effective in revealing the novel motifs. One question that needs to be addressed in future studies, as improved measures of binding specificity become available, is the importance of additional splicing factors such as the muscleblind proteins that are already known to influence the splicing of at least a few muscle-specific alternative exons (51).

A working model that summarizes these findings is presented in Figure 5. Fox, CELF and ACUAAC-binding factors are proposed as positive regulators of muscle-enriched exons via their binding to the downstream proximal intron. The distribution of binding motifs among individual introns suggests that these factors function independently in some cases, and collaboratively in others, to specify muscle-enriched splicing. For Fox proteins an especially widespread role is suggested by the high absolute abundance of UGCAUG-binding motifs: almost half of the muscle-enriched exons in datasets of all four species possess at least one UGCAUG motif in the D200 intronic region, and some of those lacking a proximal UGCAUG have phylogenetically conserved distal UGCAUG motif(s) (data not shown) analogous to the myosin II heavy chain-B neural specific exon (58). It will be interesting in the future to explore how coordination among these and other factors ultimately determines the spatial and temporal details of muscle-enriched splicing events. Based on studies in other systems, PTB is predicted as a negative regulator of splicing, functioning primarily from upstream intronic sites to prevent inappropriate inclusion in non-muscle cell types (35,59,60). Finally, it is important to note that variations of this general model likely pertain to individual exons; in particular, Fox and CELF proteins can also have a negative role in the regulation of exons that are skipped in muscle (27,61–63). Future experimental analysis of these splicing factors, using functional splicing assays and targeted disruption of splicing factor activity *in vivo* (64), will be required to more fully test the predictions of this model.

Some of the *cis*-regulatory elements associated with muscle-enriched alternative exons have previously been observed flanking brain-enriched exons: UGCAUG was the most over-represented motif in proximal downstream intron (7,8,11), and CUCUCU was the second most over-represented motif in the U100 region upstream of brain-enriched exons (7). These observations suggest general roles for Fox- and PTB-related proteins in regulating tissue-specific splicing, at least for muscle and brain, but raise the question as to how tissue specificity is ultimately

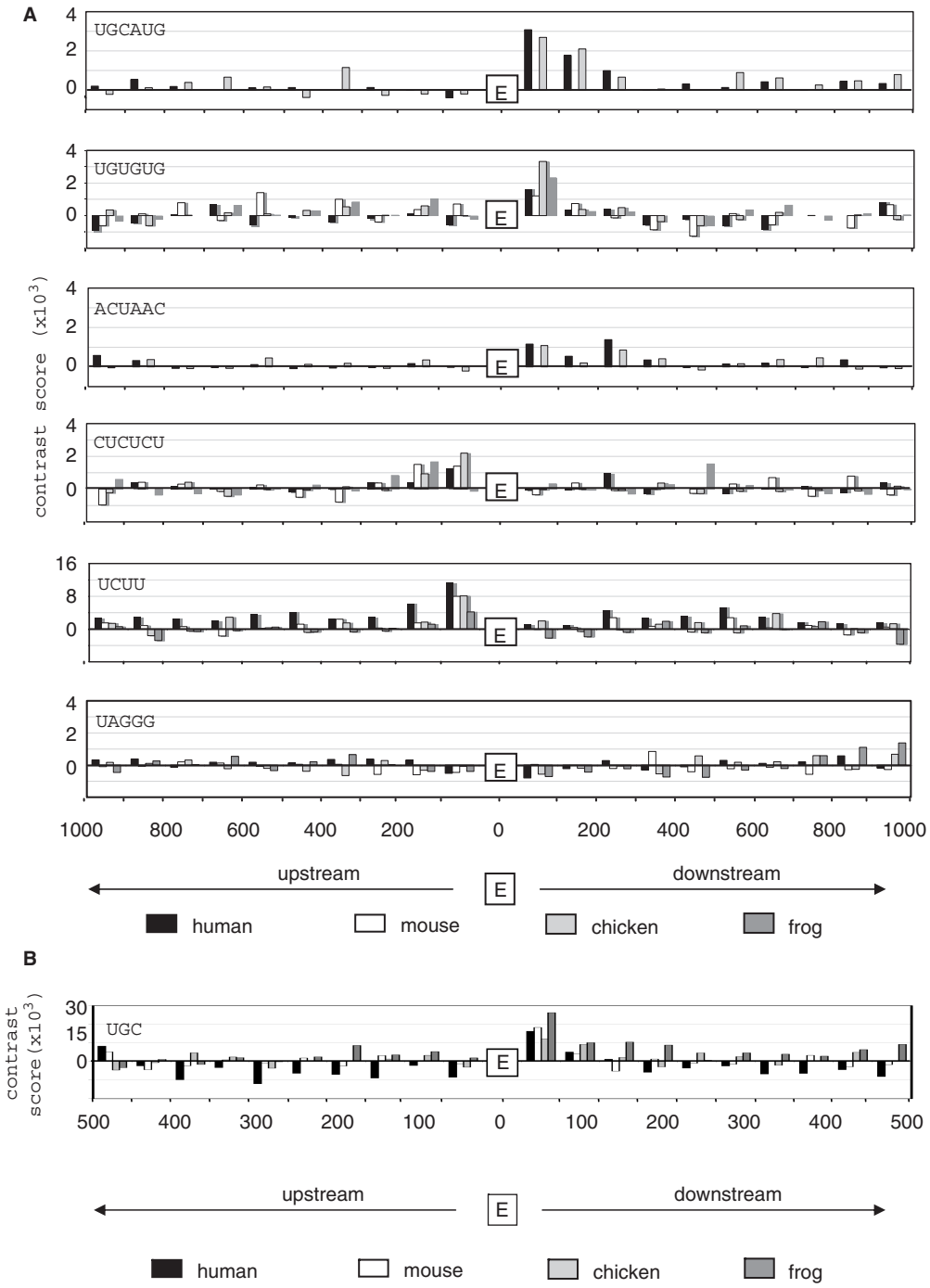

Figure 4. Phylogenetic conservation of regulatory motifs in the proximal intron sequences. Bar graphs show the over-representation of the indicated *cis*-regulatory motifs in the proximal intron sequences for muscle-enriched exons in four vertebrate species. (**A**) Enrichment of selected regulatory motifs in 1-kb flanking intron regions. The highest abundance of UGCAUG, UGUGUG and ACUAAC elements is consistently within the proximal downstream region of ~200 nt, while CUCUCU and UCUU elements were enriched in the proximal upstream intron. The representative hnRNP A1-binding site UAGGG was not over-represented near muscle-enriched exons. (**B**) Analysis of the putative CELF-binding site UGC in 0.5-kb flanking introns. UGC is enriched in the D50 region of all four species. Vertical axis, contrast score, i.e. difference in motif frequency between muscle datasets and control datasets of constitutive exons, occurrences/nt $\times 10^3$; horizontal axis, nt range relative to the alternative exon; E indicates position of the muscle-enriched exon.

determined. Several mechanisms may contribute to determination of temporal and spatial pattern of splicing switches, including tissue-specific differences in transcription and/or alternative splicing of Fox and PTB paralogs (28). Differential expression of additional RNA-binding proteins, such CELF proteins and KH-type ACUAAC-binding proteins in muscle, or NOVA1-related proteins in brain, likely also play a role, as may non-RNA-binding co-factors that preferentially interact with paralogs/isoforms of the primary RNA-binding proteins.

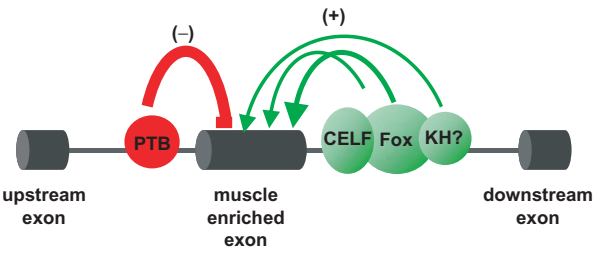

Figure 5. A candidate model showing splicing factors implicated in regulation of conserved muscle-enriched alternative exons. Based on the conserved distribution of splicing factor binding sites across multiple vertebrate orders, the positive correlation with muscle-specific splicing and the high absolute abundance of UGCAUG motifs, Fox proteins are proposed to play a major role in promoting inclusion of muscle-enriched exons. The distribution of binding motifs among individual introns suggests that CELF proteins and by KH-type splicing factor(s) function independently in some cases, and together with Fox proteins in others, to specify muscle-enriched splicing. In contrast, the enrichment of candidate PTB-binding sites in the proximal upstream intron suggests a role in preventing inappropriate inclusion of muscle-specific exons in other cell types.

In summary, normal metazoan development requires not only a transcriptional program, but also an alternative pre-mRNA splicing program to ensure that each gene encodes specific protein isoforms in the appropriate spatial and temporal patterns. Enrichment within the muscle dataset of genes with functions in cytoskeleton organization, microtubule stabilization and muscle development supports the notion that this splicing program is essential for proper expression of the unique muscle cytoskeleton. The exon microarray employed in this study will enhance our ability to track the expression of individual exons during development and differentiation. As we have demonstrated here, this experimental approach is well complemented by the computational approach based on correlation with expression. We anticipate that correlation with exon expression will provide valuable insights into the *cis*-regulation of alternative splicing as additional datasets of tissue-specific exons become available for analysis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Charles Sugnet for use of the dataset of human cassette alternative exons in correlation studies. This work was supported by DE AC03 76SF00098, the National Institutes of Health NIH grant HL45182, National Aeronautics and Space Administration Grant T6275W 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.

Conflict of interest statement. None declared.

REFERENCES

1. Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.
2. 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.
3. 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.
4. 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.
5. Stamm,S., Zhang,M.Q., Marr,T.G. and Helfman,D.M. (1994) A sequence compilation and comparison of exons that are alternatively spliced in neurons. *Nucleic Acids Res.*, **22**, 1515–1526.
6. Zhang,X.H., Kangsamaksin,T., Chao,M.S., Banerjee,J.K. and Chasin,L.A. (2005) Exon inclusion is dependent on predictable exonic splicing enhancers. *Mol. Cell. Biol.*, **25**, 7323–7332.
7. 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.
8. Minovitsky,S., Gee,S.L., Schokrpur,S., Dubchak,I. and Conboy,J.G. (2005) The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons. *Nucleic Acids Res.*, **33**, 714–724.
9. Hui,J., Hung,L.H., Heiner,M., Schreiner,S., Neumuller,N., Reither,G., Haas,S.A. and Bindereif,A. (2005) Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO J.*, **24**, 1988–1998.
10. Miriami,E., Margalit,H. and Sperling,R. (2003) Conserved sequence elements associated with exon skipping. *Nucleic Acids Res.*, **31**, 1974–1983.
11. Sugnet,C.W., Srinivasan,K., Clark,T.A., O'Brien,G., Cline,M.S., Wang,H., Williams,A., Kulp,D., Blume,J.E. *et al.* (2006) Unusual Intron Conservation near Tissue-Regulated Exons Found by Splicing Microarrays. *PLoS Comput. Biol.*, **2**, e4.
12. 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.
13. 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.
14. Clark,T.A., Sugnet,C.W. and Ares,M.Jr. (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, **296**, 907–910.
15. Frey,B.J., Mohammad,N., Morris,Q.D., Zhang,W., Robinson,M.D., Mnaimneh,S., Chang,R., Pan,Q., Sat,E. *et al.* (2005) Genome-wide analysis of mouse transcripts using exon microarrays and factor graphs. *Nat. Genet.*, **37**, 991–996.
16. Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. *et al.* (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
17. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
18. Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.
19. Das,D., Banerjee,N. and Zhang,M.Q. (2004) Interacting models of cooperative gene regulation. *Proc. Natl Acad. Sci. USA*, **101**, 16234–16239.
20. Conlon,E.M., Liu,X.S., Lieb,J.D. and Liu,J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, **100**, 3339–3344.
21. Keles,S., van der Laan,M. and Eisen,M.B. (2002) Identification of regulatory elements using a feature selection method. *Bioinformatics*, **18**, 1167–1175.

22. Das,D., Nahle,Z. and Zhang,M.Q. (2006) Adaptively inferring human transcriptional subnetworks. *Mol. Syst. Biol.*, **2**, 2006–0029.
23. Wang,W., Cherry,J.M., Botstein,D. and Li,H. (2002) A systematic approach to reconstructing transcription networks in Saccharomycescerevisiae. *Proc. Natl Acad. Sci. USA*, **99**, 16893–16898.
24. Sorek,R. and Ast,G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, **13**, 1631–1637.
25. Blencowe,B.J. (2006) Alternative splicing: new insights from global analyses. *Cell*, **126**, 37–47.
26. Clark,T.A., Schweitzer,A.C., Chen,T.X., Staples,M.K., Lu,G., Wang,H., Williams,A. and Blume,J.E. (2007) Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.*, **8**, R64.
27. 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.
28. Nakahata,S. and Kawamoto,S. (2005) Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities. *Nucleic Acids Res.*, **33**, 2078–2089.
29. Underwood,J.G., Boutz,P.L., Dougherty,J.D., Stoilov,P. and Black,D.L. (2005) Homologues of the Caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals. *Mol. Cell. Biol.*, **25**, 10005–10016.
30. Baraniak,A.P., Chen,J.R. and Garcia-Blanco,M.A. (2006) Fox-2 mediates epithelial cell-specific fibroblast growth factor receptor 2 exon choice. *Mol. Cell. Biol.*, **26**, 1209–1222.
31. Ponthier,J.L., Schluepen,C., Chen,W., Lersch,R.A., Gee,S.L., Hou,V.C., Lo,A.J., Short,S.A., Chasis,J.A. *et al.* (2006) Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16. *J. Biol. Chem.*, **281**, 12468–12474.
32. Kabat,J.L., Barberan-Soler,S., McKenna,P., Clawson,H., Farrer,T. and Zahler,A.M. (2006) Intronic alternative splicing regulators identified by comparative genomics in nematodes. *PLoS Comput. Biol.*, **2**, e86.
33. Zhou,H.L., Baraniak,A.P. and Lou,H. (2007) Role for Fox-1/Fox-2 in mediating the neuronal pathway of calcitonin/calcitonin gene-related peptide alternative RNA processing. *Mol. Cell. Biol.*, **27**, 830–841.
34. 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.
35. Spellman,R. and Smith,C.W. (2006) Novel modes of splicing repression by PTB. *Trends Biochem. Sci.*, **31**, 73–76.
36. Gardina,P.J., Clark,T.A., Shimada,B., Staples,M.K., Yang,Q., Veitch,J., Schweitzer,A., Awad,T., Sugnet,C. *et al.* (2006) Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, **7**, 325.
37. Srinivasan,K., Shiue,L., Hayes,J.D., Centers,R., Fitzwater,S., Loewen,R., Edmondson,L.R., Bryant,J., Smith,M. *et al.* (2005) Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods*, **37**, 345–359.
38. Kent,W.J. (2002) BLAT – the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
39. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Program,N.C.S., Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13**, 721–731.
40. Hastie,T., Tibshirani,R. and Friedman,J.H. (2001) *The Elements of Statistical Learning*. Springer Verlag, New York, USA, pp. 46–47.
41. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
42. Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.
43. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
44. Smith,A.D., Sumazin,P., Das,D. and Zhang,M.Q. (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics*, **21** (Suppl. 1), i403–i412.
45. Smith,A.D., Sumazin,P. and Zhang,M.Q. (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc. Natl Acad. Sci. USA*, **102**, 1560–1565.
46. Schones,D.E., Sumazin,P. and Zhang,M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21**, 307–313.
47. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
48. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
49. Nurtdinov,R.N., Artamonova,II, 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.
50. Ule,J., Ule,A., Spencer,J., Williams,A., Hu,J.S., Cline,M., Wang,H., Clark,T., Fraser,C. *et al.* (2005) Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet.*, **37**, 844–852.
51. Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.
52. Amir-Ahmady,B., Boutz,P.L., Markovtsov,V., Phillips,M.L. and Black,D.L. (2005) Exon repression by polypyrimidine tract binding protein. *RNA*, **11**, 699–716.
53. Wagner,E.J., Baraniak,A.P., Sessions,O.M., Mauger,D., Moskowitz,E. and Garcia-Blanco,M.A. (2005) Characterization of the intronic splicing silencers flanking FGFR2 exon IIIb. *J. Biol. Chem.*, **280**, 14017–14027.
54. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
55. Friedman,N., Barash,Y., Elidan,G. and Kaplan,T. (2003) In *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB)*, Berlin, Germany, pp. 28–37.
56. Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.
57. Cooper,T.A. (1998) Muscle-specific splicing of a heterologous exon mediated by a single muscle-specific splicing enhancer from the cardiac troponin T gene. *Mol. Cell. Biol.*, **18**, 4519–4525.
58. Kawamoto,S. (1996) Neuron-specific alternative splicing of non-muscle myosin II heavy chain- B pre-mRNA requires a cis-acting intron sequence. *J. Biol. Chem.*, **271**, 17613–17616.
59. Wagner,E.J. and Garcia-Blanco,M.A. (2001) Polypyrimidine tract binding protein antagonizes exon definition. *Mol. Cell. Biol.*, **21**, 3281–3288.
60. Sharma,S., Falick,A.M. and Black,D.L. (2005) Polypyrimidine tract binding protein blocks the 5′ splice site-dependent assembly of U2AF and the prespliceosomal E complex. *Mol. Cell*, **19**, 485–496.
61. Charlet,B.N., Savkur,R.S., Singh,G., Philips,A.V., Grice,E.A. and Cooper,T.A. (2002) Loss of the muscle-specific chloride channel in type 1 myotonic dystrophy due to misregulated alternative splicing. *Mol. Cell*, **10**, 45–53.
62. Savkur,R.S., Philips,A.V. and Cooper,T.A. (2001) Aberrant regulation of insulin receptor alternative splicing is associated with insulin resistance in myotonic dystrophy. *Nat. Genet.*, **29**, 40–47.
63. Zhang,W., Liu,H., Han,K. and Grabowski,P.J. (2002) Region-specific alternative splicing in the nervous system: implications for regulation by the RNA-binding protein NAPOR. *RNA*, **8**, 671–685.
64. Xu,X. and Fu,X.D. (2005) Conditional knockout mice to study alternative splicing *in vivo*. *Methods*, **37**, 387–392.