Computational Analysis of an Evolutionarily Conserved Vertebrate Muscle Alternative Splicing Program

Debopriya Das*, Tyson A. Clark*, Anthony Schweitzer, Henry Marr, Miki L. Yamamoto, Marilyn K. Parra, Josh Arribere, Simon Minovitsky, Inna Dubchak, John E. Blume, John G. Conboy

* These authors contributed equally to this work

running title: muscle alternative splicing in vertebrate genomes

key words: exon microarray. alternative splicing, Fox1, CELF

Abstract

A novel exon microarray format that probes gene expression with single exon resolution was employed to elucidate critical features of a vertebrate muscle alternative splicing program. A dataset of 56 microarray-defined, muscle-enriched exons and their flanking introns were examined computationally in order to investigate coordination of the muscle splicing program. Candidate intron regulatory motifs were required to meet several stringent criteria: significant over-representation near muscle-enriched exons, correlation with muscle expression, and phylogenetic conservation among genomes of several vertebrate orders. Three classes of regulatory motifs were identified in the proximal downstream intron, within 200nt of the target exons: UGCAUG, a specific binding site for Fox-1 related splicing factors; ACUAAC, a novel branchpoint-like element; and UG- / UGC-rich elements characteristic of binding sites for CELF splicing factors. UGCAUG was remarkably enriched, being present in nearly one-half of all cases. These studies suggest that Fox and CELF splicing factors play a major role in enforcing the muscle-specific alternative splicing program, facilitating expression of a set of unique isoforms of cytoskeletal proteins that are critical to muscle cell differentiation.

Supplementary materials: There are four supplementary tables and one supplementary figure. The tables provide additional detailed information concerning the muscle-enriched datasets, and about over-represented oligonucleotide sequences in the flanking introns. The supplementary figure shows RT-PCR data confirming the muscle-enriched expression of exons predicted from the microarray analysis.

Introduction

Alternative pre-mRNA splicing is a critical mechanism for regulating gene expression in metazoan organisms, and is often cited as a mechanism for generating tremendous protein diversity from a relatively small number of genes. A majority of human genes exhibit some form of alternative splicing, and examples abound for which alternative protein isoforms encoded by a single gene exhibit differences in structure, function, or subcellular localization. Intriguingly, a small but critical proportion of these splicing events is subject to spatial and temporal regulation during differentiation and development. Thus, 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. Biochemical and computational studies have revealed the existence of sequence-specific RNA binding proteins that interact with *cis*-acting regulatory elements in alternative exons, and their flanking intron sequences, to regulate the splicing efficiency of alternative exons. Detailed studies of model pre-mRNAs have established that splicing efficiency is modulated by the activities of stimulatory *trans*-acting factors binding at enhancer motifs in the RNA, and inhibitory factors acting at silencer elements. Several cases in which splicing decisions are determined by the relative concentrations of antagonistic positive and negative factors have been described (e.g., (1-9).

However, the full extent of tissue-specific alternative splicing patterns has not been appreciated on a genome-wide scale, limiting our ability to understand its contribution to cell- type specialization and to understand the rules that govern temporal and spatial patterns of alternative splicing. Biochemical splicing assays, computational analyses of RNA sequences proximal to alternative exons, and CLIP (crosslink and immunoprecipitation) strategies have begun to reveal some of the important elements for tissue-specific splicing. Thus, NOVA1- and Fox1-related splicing factors are required for correct regulation of brain-specific splicing events (10-14), and muscle-specific exons can be regulated by proteins of the CELF family of splicing factors and by muscleblind proteins (3,15-19).

Recent advances in microarray technology now facilitates characterization of gene expression at the level of single exons, and thereby offers an important new tool for broad analysis of alternative splicing programs (11,20-29). In this study, we employed a new Affymetrix exon microarray format to identify 56 muscle-enriched alternative exons, a number of which are predicted to alter the expression of cytoskeletal related genes. Computational analysis of flanking introns revealed a remarkable concentration of binding sites for splicing factors of the Fox1, CELF, and PTB families adjacent to the muscle-enriched exons, a pattern that was conserved in the genomes of several vertebrate orders. Together these observations provide the most comprehensive picture yet available of a muscle-specific

alternative splicing program and suggest that Fox and CELF-related proteins play a critical role in regulating this program.

Results

Identification and characterization of muscle-enriched alternative exons. The prototypical human muscle-enriched exon dataset analyzed in this study (Table I) was derived from genome-wide exon microarray hybridization data. The final dataset consisted of 56 muscle-enriched, internal cassette exons that exhibited an elevated "splicing index" in muscle, i.e., a higher intensity of alternative exon expression, normalized to the parent transcript expression level, in heart and skeletal muscle than in nonmuscle samples. Most of these exons (~80%) are integral multiples of 3nt in length, with a median size of 84nt, consistent with the notion that alternative exons are smaller than average constitutive exons (145nt; ref. 30,31). 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 I.)

Muscle-enriched splicing patterns for a few of these exons were validated experimentally in the human dataset by RT/PCR (Figure 1). Although splicing patterns were not absolutely muscle-specific, in each case the efficiency of exon inclusion was highest in heart and skeletal muscle, thus confirming the predictions of the exon microarray. Importantly, mRNA and/or EST evidence from the genetic databases (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 (32). Together these results indicate that muscle-enriched exons whose alternative splicing exhibits programmed changes during differentiation constitute a special class of highly conserved alternative exons.

Analysis of downstream intron sequences. Alternative splicing regulatory elements responsible for tissue-specific splicing are often located in flanking intron sequences. To search for candidate intronic regulatory motifs for the muscle-specific splicing program, we used a word counting program (13,14) to identify oligonucleotide sequences that are over-represented near muscle-enriched exons, and a hypergeometric analysis to examine motif distribution among the individual exons of each dataset. The significance of candidate regulatory motifs was further investigated by correlation analysis of motif frequency with splicing index. Finally, we examined their spatial and phylogenetic conservation through

vertebrate evolution (14). Motifs meeting these combined criteria were judged to be very strong candidates for regulating muscle-specific alternative splicing events.

Application of these computational strategies to the downstream intron region yielded the results presented in Table I (for the human muscle dataset) and Supplementary Table I (for the mouse, chicken and frog datasets). Table I shows the top 30 motifs in the downstream 200 nt (D200) of intron sequence, ranked in order by the frequency difference (frequency in muscle dataset minus frequency in control dataset). Most of these motifs were highly significant by the hypergeometric measures as well, supporting a potential role in regulation of muscle-specific alternative splicing (Table I).

Correlation of motif frequency with expression data has been used effectively to decipher *cis*-regulatory motif combinations that are functional in regulating gene transcription (33,34). When a similar strategy was applied here to the list of candidate splicing regulatory motifs, nine motifs were found to be positively correlated with the muscle splicing index (correlation p-values < 0.05). These motifs fell into three distinct classes: (1) the Fox1 binding motif UGCAUG (p-value = 6.79×10^{-5}) and five closely related hexamers (UUUGCA, UUGCAU, AUGCAU, GCAUGG, GUGCAU); (2) UG-rich elements GUGUGU and UCUGUG (correlation p-values=0.032 and 0.015, respectively), that resemble binding sites for the CELF family of splicing factors; and (3) the novel motif ACUAAC (p-value=0.0006), which is similar to the UACUAAC element that was noted in a recent study of muscle-specific exons in mouse (28). Each of these classes is analyzed further in the following sections.

Identification of the unique Fox1 splicing factor binding site, UGCAUG, as the strongest candidate regulatory element for muscle-specific alternative splicing.

The most prominent motif by all measures is the Fox1 binding motif, UGCAUG. This hexamer was strongly over-represented in the first 200nt of proximal downstream intron sequence (D200), at a level greater than expected for *any* hexamer, It occurred at 6-10-fold higher frequency than expected by chance, or observed in the comparable D200 region of the control datasets. UGCAUG was also the most over-represented hexamer in the mouse, chicken, and frog datasets (Supplementary Table II), and exhibited the strongest correlation with the muscle splicing index (Table I).

The occurrence of UGCAUG motifs was widespread among muscle-enriched exons: in the D200 region, ~40-50% of exons in the human (23/56), mouse (21/54), chicken (20/43), and frog (19/36) muscle-enriched datasets possessed at least one such motif. Phylogenetic and spatial conservation of UGCAUG elements was further examined by determining abundance as a function of position relative to the regulated alternative exon. Figure 2 shows that in all four species, UGCAUG motifs were specifically concentrated in the D200 region. In contrast, the frequency of UGCAUG elements was substantially

reduced in more distal downstream introns, in the upstream introns, and in all regions of the control dataset. Nontissue-specific alternative exons also lack this concentration of UGCAUG motifs in D200 (14). Interestingly, although UGCAUG motifs in the population of muscle enriched exons were distributed within the broader D200 region, in individual cases UGCAUG motif(s) associated with a particular orthologous exon tended to be highly conserved at specific sites within the D200 across different vertebrate genomes (analogous to (14), Figure 2).

Together these results indicate a major role for Fox1-related splicing factor(s) in the regulation of the muscle alternative splicing program of vertebrates. However, since previous studies have reported GCAUG as the binding site for a Fox1-related zebrafish protein (35), we also examined the distribution of this pentamer in the muscle-enriched datasets. In all four species GCAUG was indeed highly over-represented (data not shown); however, most GCAUG elements in the D200 regions occurred in the context of the full UGCAUG hexamer (human, 28/43= 65%; mouse, 23/40=58%; chicken, 23/31= 74%; frog, 25/33=76%). A similar result was obtained upon analysis of the proximal downstream intron sequences for brain-enriched exons (14). This data suggests that UGCAUG is likely predominant functional regulatory element for muscle-enriched exons, but also allows for the possibility that GCAUG is functional in some cases.

Characterization of the CELF splicing factor binding sites as candidate regulatory elements for muscle-specific alternative splicing. CELF- and muscleblind-related splicing factors have been shown experimentally to play important roles in the alternative splicing of selected muscle-specific exons (3,15-19). We therefore examined our larger muscle-enriched alternative exon dataset for regulatory motifs that might suggest a wider role for CELF proteins in regulation of the muscle alternative splicing program. Such motifs have been reported to include UG-repeat sequences and CUG- or UGC-motifs (3,15).

UGUGUG and GUGUGU were among a group of UG–rich motifs in the top 10 over-represented hexamer motifs of the human muscle dataset, as ranked by frequency difference (Table I and Supplementary Table II). As mentioned above, GUGUGU was also one of the few motifs that was strongly correlated with the muscle splicing index. These findings are in stark contrast to what is observed for brain-enriched alternative exons, where over-representation of UG–rich motifs is not pronounced (14).

Further analysis of UGUGUG was performed by phylogenetic and spatial analysis of its distribution among the introns flanking muscle-enriched exons. Spatially, the highest frequency of UGUGUG motifs was localized in all four species to a narrow region, the first 50nt of downstream intron

(D50), although there was a moderate background frequency in more distal intronic regions (Figure 3). The over-representation in the D50 region was statistically significant when judged by the overall frequency of motifs in the dataset (resampling confidence values <0.005 in human, mouse, and frog; < 0.05 in chicken) or by the proportion of muscle-enriched exons possessing this downstream motif (hypergeometric P values ≤0.01 for all four species), in comparison to the species matched control datasets. The latter measure suggests that the enrichment of UGUGUG was due to a general distribution of the motif among all muscle-enriched exons, rather than from repeated elements in one or a few D50 sequences.

Analysis of trinucleotides in proximal downstream intron sequences revealed that CUG was relatively abundant, but not over-represented relative to the control exon datasets. In contrast, UGC was highly over-represented in the most proximal downstream region of all four vertebrates tested, with a frequency difference ranking first in human, mouse, and frog ($P< 0.0001$ for all three species) and second in chicken ($P<0.004$).

The novel motif ACUAAC is a candidate regulatory element for a subset of muscle enriched exons. The identification of ACUAAC as an over-represented motif that correlates with the muscle splicing index in the human dataset prompted us to further investigate its potential role as a regulator of muscle-specific exons. As shown in Figure 4, ACUAAC was specifically and phylogenetically conserved in the proximal downstream intron region in all four species. Approximately ~ 20% of the exons in each dataset possessed ACUAAC in the D200 region, approximately 4-fold greater than expected by chance. The over-representation of ACUAAC relative to species-matched control exons was highly statistically significant for all four species (hypergeometic p-values: human, $2.5x10^{-08}$; mouse, $2.8\ x10^{-08}$; chicken, $6\ x10^{-04}$; and frog, $1.7\ x10^{-07}$). Moreover, analysis of individual exons revealed high spatial conservation as well: the MTMR3 and RBM9 exons were among nine cases in which the intron locations of ACUAAC elements were conserved in the proximal downstream intron for at least three of the four species. This novel element may therefore be essential for proper regulation of a subset of muscle-enriched exons. This motif is similar to the UACUAAC motif recently reported to be over-represented downstream of a smaller dataset of muscle-enriched exons in the mouse (28).

Analysis of upstream intron sequences: Identification of PTB binding motifs as a candidate regulatory elements for muscle-specific alternative splicing. Computational analysis of upstream intron sequences revealed that over-represented hexamers in the upstream sequence were extremely high in uridine content. However, a similar set of U-rich hexamers was also significantly over-represented among the dataset of tissue-nonspecific alternative exons (14), suggesting that this feature is

characteristic of alternatively spliced exons in general, not a specific feature of muscle regulation. More complex words, that were likely to represent binding sites for splicing regulators, were not consistently found in these datasets.

As an alternative approach for identification of upstream regulatory elements, we performed a phylogenetic analysis of the distribution of motifs for other known splicing regulatory proteins. One of these is PTB, an inhibitor of splicing for many alternative exons (including muscle-specific exons (36-38) via binding to pyrimidine rich sequences such as CUCUCU (39), UCUU (40), and related sequences (41). Putative PTB binding sites were previously reported to be significantly over-represented upstream of brain-enriched alternative exons (13). Here, we chose to examine the distribution of both CUCUCU and UCUU motifs in the muscle enriched datasets, relative to their frequency in the control datasets of constitutive exons, to test for involvement of PTB in regulation of the muscle-enriched exons. In all four species, the muscle-enriched datasets showed strong over-representation of both representative PTB binding sites in the proximal upstream intron (Figure 5). CUCUCU was concentrated mainly in the U200 region. UCUU was focused even more tightly in the U100 region (Figure 5, bottom), where it was consistently among the top five over-represented sequences 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 (40,42,43).

We also probed the role of hnRNP A1, a well known splicing inhibitory protein, by examining the incidence of the prototypical A1 binding site, UAGGG (44), in flanking intron regions. In contrast to the abundance of candidate PTB binding sites, UAGGG was relatively deficient in the proximal intronic regions both upstream and downstream of the muscle-enriched exons (Figure 6). All four datasets exhibited this "A1-deficient" zone near the regulated exons, although the boundaries of the zone varied a bit between species. Together, these results suggested that suggests that PTB may play a widespread role in the negative regulation of muscle-enriched exons in other tissues. While high affinity A1 sites may be less abundant, we cannot rule out involvement of lower affinity of A1 binding sites in the regulation of muscle-specific splicing.

Finally, as a control we examined the distribution of YCAY binding sites for the neural splicing regulator, NOVA1. This element is not expected to play a role in muscle-specific splicing events, and indeed it was much less abundant than the CELF binding motif (results not shown).

Motifs identified via position weight matrix analysis are consistent with word analyses. Because many splicing factors bind RNA degenerate oligonucleotide sequences, we performed additional analyses in

which degeneracy was considered in the motif searches through the use of a position weight matrix (PWM) approach. Over-represented PWMs were obtained using the DME algorithm (63) using multiple parameter setting in order to avoid bias from DME. Functional PWMs were identified by assessing their correlation with muscle expression using linear splines. In contrast to previous approaches (D. Das, Z. Nahle & M.Q. Zhang, Mol. Syst. Biol., 2006, in press), we accounted for both strength of PWM and the number of putative binding sites in this approach. Degenerate 6nt and 4nt sequences that were over-represented in the proximal downstream intron sequence are shown in Table II. 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. Seven of the motifs, including the six most statistically significant sequences, represent close matches to the Fox binding site, UGCAUG; two (NHCUAA and HCUAAN) are very similar to the novel ACUAAC; and the remaining sequence (SUKUGS) resembles UG-rich binding site for CELF proteins. Among the over-expressed 4-mers in the D50 region, the top-scoring motif (UGCM) incorporates the CELF binding motif UGC. Finally, in the U200 region, all of the statistically over-represented hexamers were quite pyrimidine-rich relative to the control group. Further investigation will be required to determine whether PTB is the major splicing regulator that binds to these elements, or whether perhaps additional factor(s) might be involved in regulating muscle exon expression from the position of the upstream intron.

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 (11). 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 GoMiner (45) 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 IV). These correlations were highly statistically significant (p= 0.0001-0.0007), suggesting that the muscle alternative splicing program is critical for proper expression of the unique cytoskeleton characteristic of vertebrate muscle.

Discussion

A central hypothesis of this work is that carefully orchestrated programs of alternative pre-mRNA splicing are essential mediators of gene expression during normal metazoan development. Such programs may regulate defined subsets of alternative exons in highly specific temporal and spatial

patterns so as to modify the structure and function of many important proteins according to the specialized requirements of each differentiated cell type. In this study, a novel genome-wide exon microarray strategy was used to discover a set of highly conserved, muscle-regulated alternative exons. We propose that this unique constellation of muscle-specific splicing events is orchestrated by an alternative splicing program that modulates the expression of many widely expressed genes in a muscle-specific manner. Many of the splicing switches are predicted to alter the expression of cytoskeletal proteins and signaling proteins, functional studies of which may provide fundamental new insights into muscle biology. Thus, alternative splicing switches are likely to be essential for remodeling of muscle cell function during normal differentiation.

It is of great interest to understand the regulatory machinery that coordinates the muscle alternative splicing program. Computational analysis of flanking intron sequences was performed as a first step toward characterizing the *cis*-regulatory motifs that constitute a critical element of this machinery. In addition to simple word counting, we also employed several additional strategies-correlation analysis with muscle expression, phylogenetic analysis of orthologous datasets in other vertebrates, and analysis of more complex motifs using position weight matrices. Together these approaches provided substantial additional support for the generality of the findings for muscle splicing regulation in vertebrates. Most importantly, we identified a specific alternative splicing motif, UGCAUG, the known binding site of Fox1 and Fox2 splicing factors, as a candidate cis-regulatory element for mediating muscle-specific alternative splicing switches. Almost one-half of all muscle-regulated exons contained UGCAUG motif(s) in the proximal intron sequences, and many of the remainders possess such motifs in the more distal intron. This frequency far exceeds the incidence of UGCAUG elements mapping near constitutive exons. Further supporting the functional significance of this motif, UGCAUG over-representation was correlated with the muscle splicing index and was a highly conserved feature in the proximal intron sequences of among vertebrates classes including birds, amphibians, and mammals.

Given the specific (35,46) and high affinity (47) binding of Fox-1 related splicing factors to (U)GCAUG, the Fox proteins are ideally designed to play a broad role in specifying a restricted set of alternative exons to be spliced at the appropriate stage of muscle differentiation. A similar hypothesis was proposed earlier for Fox function in brain-enriched splicing events (12-14) and in smaller sets of muscle-enriched exons (13,28). Therefore, we propose that Fox proteins alone do not determine the tissue-specificity of such splicing switches, but instead they require additional interacting co-factors to accomplish this regulatory feat. Our results further suggest that splicing factors in the CELF family, already well known to regulate several muscle-specific splicing events, likely cooperate with Fox proteins on a broader scale to regulate the muscle splicing program. In a subset of cases, muscle exons

may be regulated by binding of as yet unknown proteins to ACUAAC elements in the downstream proximal intron (28).

A hypothetical model for Fox-mediated alternative splicing switches is shown in Figure 6. In undifferentiated cells and nonmuscle cells, we propose that Fox activity is relatively low in comparison to splicing silencer activity; Fox-regulated exons will thus be predominantly skipped in these cells. Likely silencer proteins include members of the PTB family, since a high focal concentration of PTB binding sites was identified upstream of muscle-enriched exons among the four vertebrate genomes (Figure 4). In contrast, high affinity binding sites for hnRNP A1 (44), another known intronic silencer protein, (e.g., (48)), were relatively deficient in the proximal introns (Figure 5); however, it is possible that lower affinity sites may antagonize Fox and CELF proteins in some genes. Next, the model proposes that differentiation is accompanied by a relative increase in the activity of Fox and/or CELF proteins, leading to greatly increased inclusion of the adjacent alternative exons. This model is consistent with similar proposals for the regulation of individual neural- and muscle-specific exons (1,3) by antagonistic interactions between positive and negative factors, but extends the hypothesis to cover a broad range of exons in the muscle alternative splicing program.

Disruption of normal CELF activity can cause human disease and aberrant splicing of target pre-mRNAs. For example, myotonic dystrophy is a triplet repeat disease in which in which CUG (or alternatively, UGC) repeats alter the splicing program of multiple muscle transcripts (49-56). Analogously, targeted disruption of cardiac CELF activity disrupts alternative splicing causes cardiomyopathy (18). The muscle-enriched exons identified in this study may represent additional targets of CELF activity, and it seems likely that perturbed alternative splicing of these transcripts may contribute to the pathology in these diseases. By analogy, we anticipate that aberrant Fox expression in vivo will also cause disease that includes muscle pathology.

The exon microarray employed in this study has dramatically enhanced our ability to track the expression of individual exons during development and differentiation. This experimental approach complements and extends previous computational strategies for identification of conserved alternative exons (57,58) by confirming and describing the tissue-specificity for alternative exon expression. The high resolution picture of gene expression that is emerging from these studies should have a major impact on our understanding of the biology of the genes, by unmasking many previously unappreciated isoforms of important cellular proteins (59). Moreover, by defining exon datasets with shared expression patterns, these arrays will facilitate computational analysis of candidate regulatory motifs that mediate tissue-specific alternative splicing and give rise to the complex biology. Although the current study only considered internal cassette exons, it should be possible to utilize the same approach

for analysis of tissue-specific alternative first or last exons that must require yet additional regulatory mechanisms. With continued improvements in the technologies for detecting tissue-specific exon expression and the computational approaches for data analysis, we should gain many critical new insights into mechanisms of vertebrate gene expression in health and disease.

Methods

Identification of muscle-enriched alternative exon and control exon data sets. The muscle-enriched alternative exon dataset from humans was identified using an algorithm called the splicing index. Briefly, probeset intensities from individual exons are corrected for transcription rate by dividing the median intensity of probesets from well annotated exons (Ensembl / RefSeq) from the same gene. This gene-level-normalized intensity is then compared between groups of samples using a Student T-Test. For this analysis, we compared three biological replicates each of heart and skeletal muscle to as a group to the three biological replicates of 14 other normal adult human tissues as a second group.

After removing non-expressed genes and probesets that are not detected above background, the T-Test results were further filtered using a 0.5 minimum fold change and remaining probesets are sorted by P-value. In an attempt to generate a list of high-confidence muscle-enriched internal cassette exons the T-Test results were manually filtered by observing the expression data in genomic context. Probesets that mapped to alternative transcriptional starts, alternative 3' ends, alternative poly adenylation sites, and alternative 5' or 3' splice sites were removed. Only probesets that showed clear muscle-enrichment relative to nearby constitutive exons were kept for further analysis. Probesets were then mapped to the May, 2004 human genome using the BLAT tool from the UCSC Genome Browser (60). Exact exon boundaries were determined by comparison to EST and mRNA sequences requiring consensus splice sites. The majority of targeted probesets mapped to previously annotated cassette exons.

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 *tissue-nonspecific alternative exon* dataset was derived as described previously (14) 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 (14).

The muscle-enriched datasets and the control datasets will be available at http://gsd.lbl.gov/splicing/.

Computational analysis. Candidate regulatory elements were predicted computationally by identifying oligonucleotide sequences (words) that were over-represented in each tissue-specific dataset, relative to the control datasets, using the algorithm described previously (13). 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 (13) 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 muscle sample).

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

$$\log(S_e / S_{eC}) = a_w + b_w . n_w^e$$

S_e is the splicing index of exon *e* and *C* refers to a reference condition. The splicing ratio was obtained as the ratio of the splicing index of the exon to its average across all the tissues. The coefficients a_w and b_w were obtained by maximizing the percent reduction in variance. Percent reduction in variance, $\Delta\chi^2$, is defined as (33,34):

$$\Delta\chi^2 = \left[1 - \sum_e (r_e - \bar{r})^2 \Big/ \sum_e (y_e - \bar{y})^2\right] \times 100,$$

where $y_e = \log(S_e / S_{eC})$, $r_e = y_e - y_e^p$ is the residual (*p* indicates the predicted value of *y*), and $\bar{y}$ and $\bar{r}$ are their respective means. p-values were calculated using an F-test (61):

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)}$$

where RSS_1 is the residual sum of squares (RSS) of the final regression model with $p_1 + 1$ terms (p_1 = 1), and RSS_0 is the RSS of the model without a specific motif, which has $p_0 + 1$ terms (p_0 = 0). N is

the number of exons. This statistic has a F distribution with $p_1 - p_0$ numerator degrees of freedom and $N - p_1 - 1$ denominator degrees of freedom.

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 (57). A background set is necessary to model the correct dependence of the log ratios on the word count. 25 such random draws were performed, and a linear fit was obtained for each such case. The significance of each word reported in Table II is the geometric mean of the p-values of all such fits.

Linear Splines. Regulatory motifs for splicing are often degenerate. To account for this degeneracy, we used position weight matrices (PWMs), which are probabilistic representations of binding sites for *trans*-acting factors. PWMs were obtained using the DME algorithm (see below). For each PWMμ 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 (62). For each sequence there is a continuum of scores arising from a large number of putative binding sites. One needs to determine a cut-off score to discriminate the true binding sites from the false sites. Such thresholds can be adaptively determined using linear splines. PWM scores across exons $\{M_e^\mu\}$ for a given motif μ were fitted to the splicing ratios $\{\log(S_e / S_{eC})\}$ using the following model:

$$\log(S_e / S_{eC}) = a_\mu + b_\mu \sum_{M_e^\mu > \xi_\mu} \theta(M_e^\mu - \xi_\mu, 0)$$

where $\theta(x,0)$ is a linear spline: it is x, when $x \geq 0$, and zero, otherwise. ξ_μ, termed knot, corresponds to the cut-off score. The coefficients a_μ and b_μ and the location of the knot ξ_μ were determined so as to maximize $\Delta\chi^2$. It is important to note that $\Delta\chi^2$ depends on the location of ξ_μ. Maximization of $\Delta\chi^2$ leads to an unbiased and adaptive determination of this threshold for any given PWM. The significance of the fit was enumerated using an F-test, as discussed above. Of note, in contrast to previous approaches where contribution from only the maximum scoring site was considered, we systematically accounted for the contribution from active sites with weaker scores as well. The number of such contributing sites is adaptively determined. Thus, both binding affinity and counts of active motifs are accounted for in our approach. p-values for each motif was obtained by fitting splines multiple times to a combined set of foreground and background sequence sets, similar to the case of linear fits.

DME (Discriminating Matrix Enumerator) DME (63,64) is an enumerative search algorithm that finds the PWMs over-represented in a foreground set relative to a background set. Muscle specific exons were used as the foreground set and a set of 1305 randomly chosen control exons from the human genome as background. Both exonic and intronic regions (upstream and downstream) were searched for over-represented matrices of width 6nts. Default parameter settings were used, except we varied the average information content of the PWM from 1.0-2.0 in steps of 0.1. 15 PWMs were obtained for each such setting. Correlation analysis was performed on non-redundant sets of matrices. Matrix similarity was assessed using MatCompare (65).

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.

Figure Legends

Figure 1. Validation of muscle-enriched patterns of alternative exon expression. RT-PCR confirmation of muscle-enriched alternative exon expression. Amplifications were performed usin primers in the flanking constitutive exons. Lanes: 1, brain; 2, kidney; 3, liver; 4, stomach; 5, bone marrow; 6, testis; 7, heart; 8, skeletal muscle. Arrowheads indicates position of the muscle exon inclusion products.

Figure 2. Phylogenetic conservation of Fox1 binding sites in the proximal downstream intron. Histograms show the over-representation of UGCAUG elements in the proximal intron sequences for muscle-enriched exons in four vertebrate species. The highest abundance of UGCAUG elements is consistently within the downstream ~200nt (D200) region. Vertical axis, contrast score, i.e. difference in motif frequency between muscle datasets and control datasets of constitutive exons, occurrences/nt x 10^3; horizontal axis, nt range relative to the alternative exon.

Figure 3. Phylogenetic conservation of predicted CELF binding sites in the proximal downstream

intron. Histograms show the relative over-representation of intronic UGUGUG (upper panel) and UGC (lower panel) elements in the proximal intron sequences for muscle-enriched exons in four vertebrate species. The highest abundance of these candidate CELF binding sites elements is consistently in the proximal downstream intron. The axes are the same as in Fig. 1.

Figure 4. Phylogenetic conservation of ACUAAC elements in the proximal downstream intron. Histograms show the relative over-representation of ACUAAC elements in the proximal intron sequences for muscle-enriched exons in four vertebrate species. The highest abundance of ACUAAC elements is consistently within the downstream proximal intron. The axes are the same as in Fig. 1.

Figure 5. Phylogenetic conservation of putative PTB binding sites in the proximal upstream intron sequences. Histograms show the relative over-representation of CUCUCU (upper panel) and UCUU elements (lower panel) at 100nt intervals upstream and downstream of the muscle-enriched exons, Data for human, mouse, chicken, and frog datasets is included. The highest abundance of CUCUCU elements is consistently within the upstream U200 region, while UCUU is more focused in the U100 region. The axes are the same as in Fig 1.

Figure 6. Relative deficiency of hnRNP A1 binding sites in the proximal intron sequences in phylogenetically conserved. Histograms show the under-representation of UAGGG elements in the proximal intron sequences for muscle-enriched exons in four vertebrate species, relative to constitutive exons. The lowest frequency of UAGGG elements is consistently found within the proximal intron regions upstream and downstream of the regulated exons (boxed). The axes are the same as in Fig. 1.

Figure 7. 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 and the positive correlation with muscle-specific splicing, CELF and Fox proteins are proposed to play major roles in promoting inclusion of muscle-enriched exons. A smaller subset of muscle exons may also be enhanced by ACUAAC-binding protein(s) represented in this model as protein X. In contrast, the enrichment of candidate PTB binding sites in the proximal upstream intron suggests a role in preventing ectopic inclusion of muscle exons in other cell types.

Figures

Figure 1. Muscle-enriched patterns of alternative exon expression

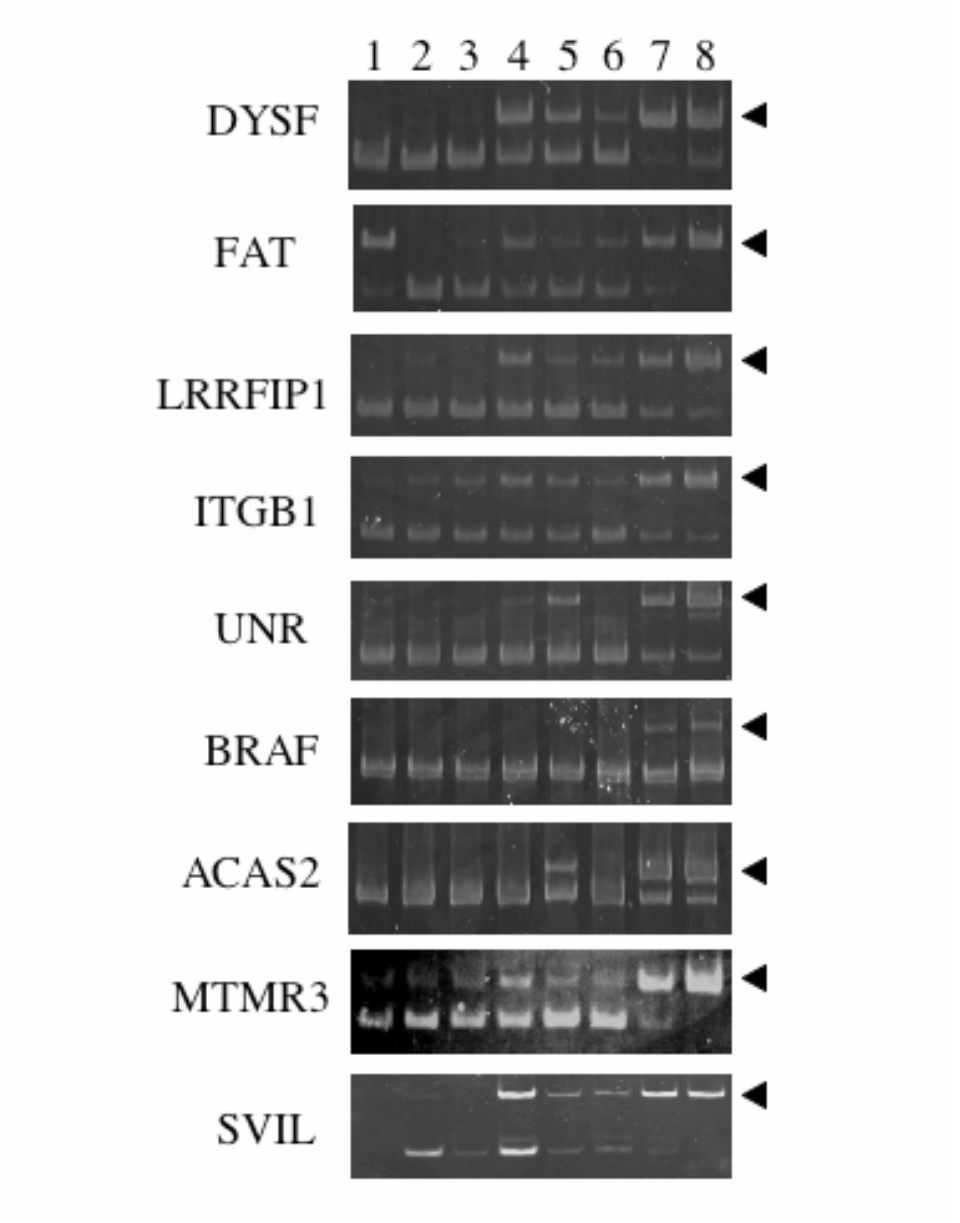

Figure 2. Phylogenetically conserved enrichment of Fox1 binding sites (UGCAUG) in the downstream proximal intron.

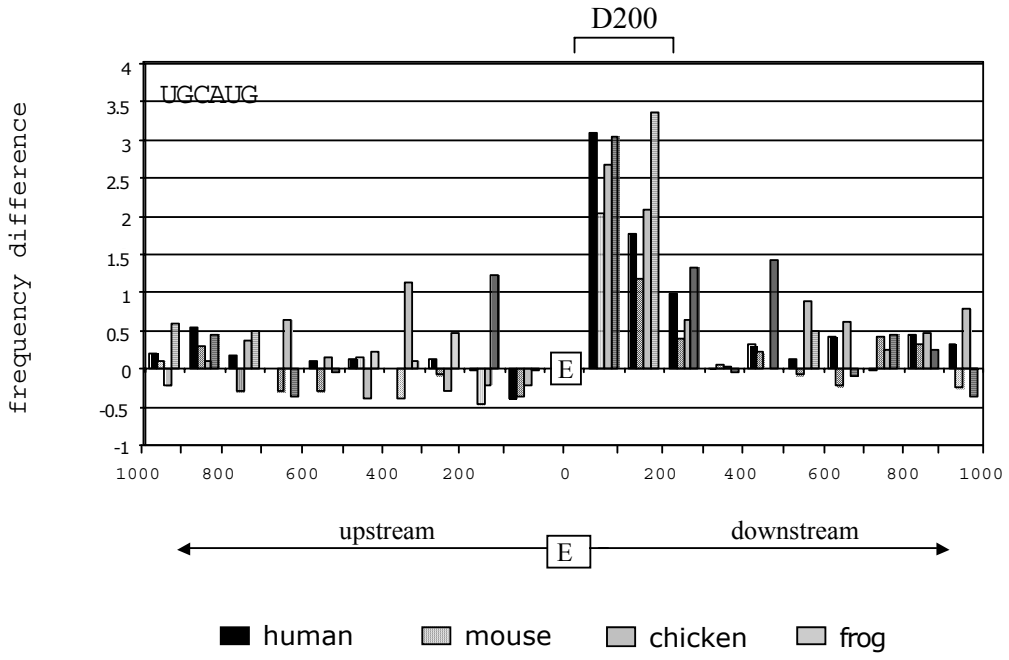

Figure 3. Phylogenetically conserved enrichment of CELF binding sites (UG repeats and UGC) in the downstream intron.

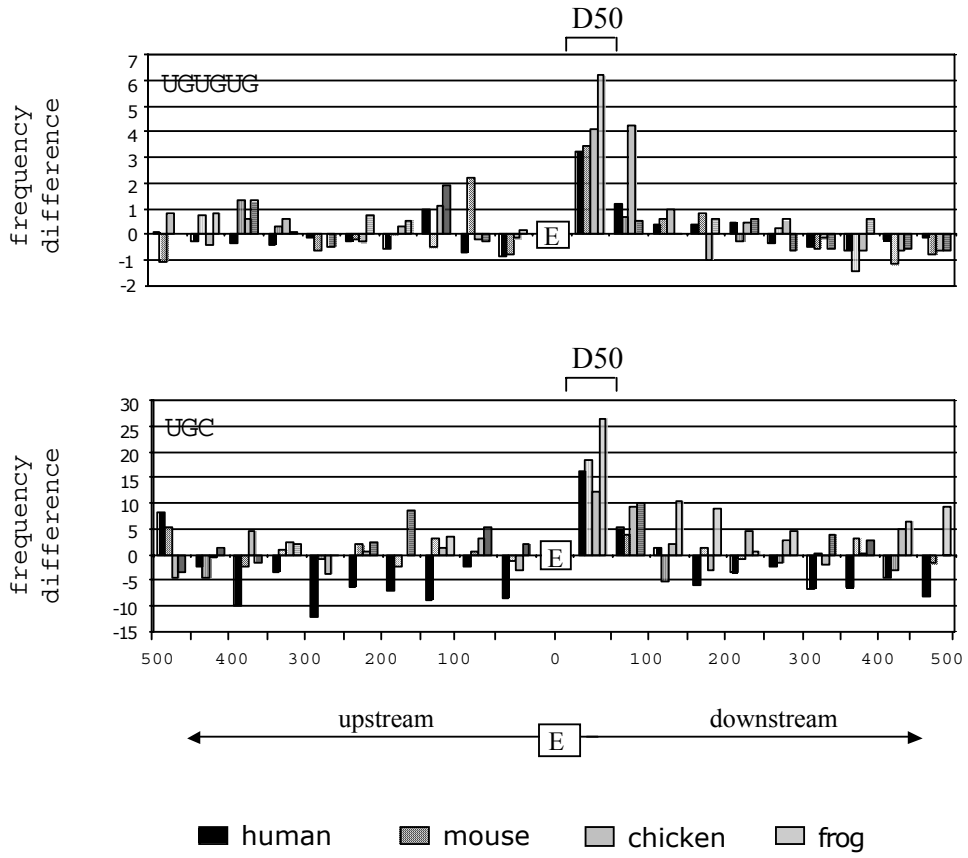

Figure 4. Phylogenetically conserved enrichment of ACUAAC candidate splicing regulatory sites in the downstream intron.

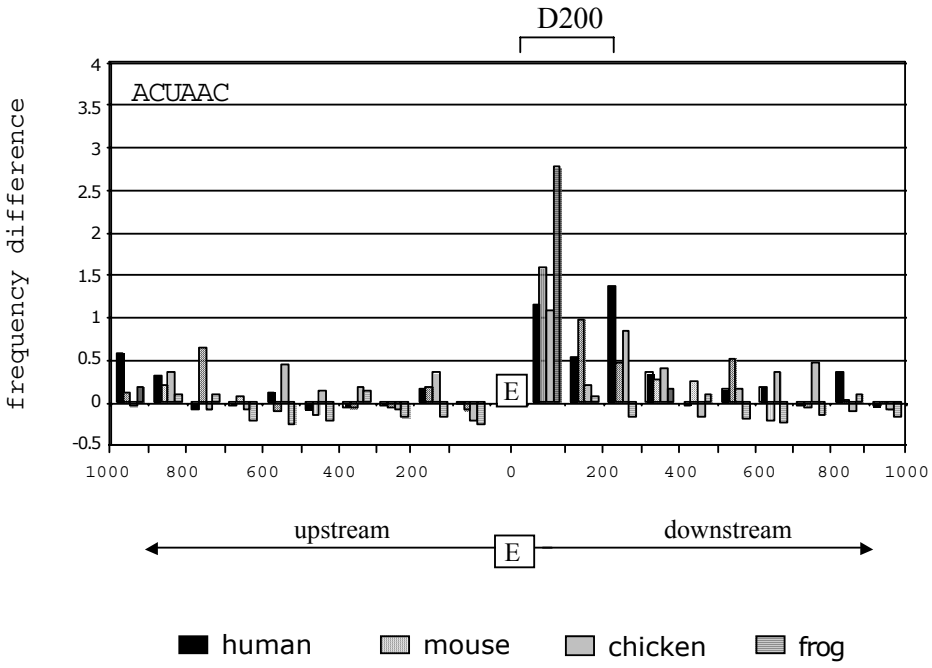

Figure 5. Phylogenetically conserved enrichment of candidate PTB splicing factor binding sites in the proximal upstream intron.

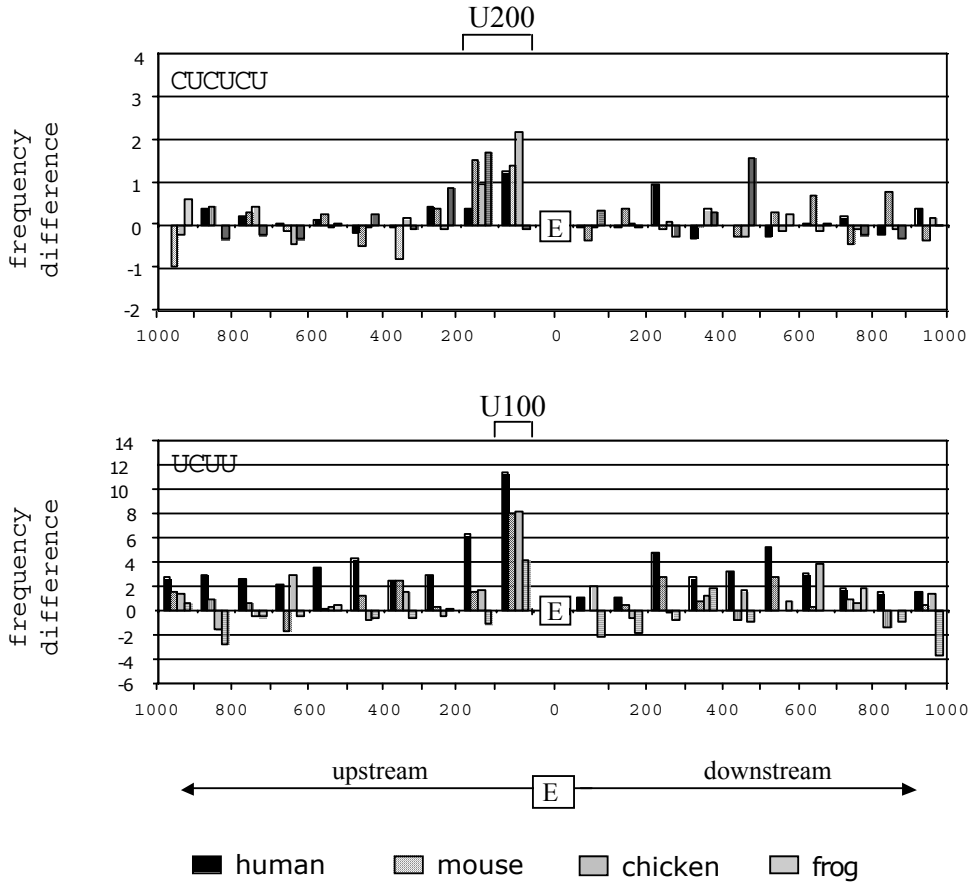

Figure 6. Relative deficiency of high affinity hnRNP A1 binding sites in the proximal intron sequences.

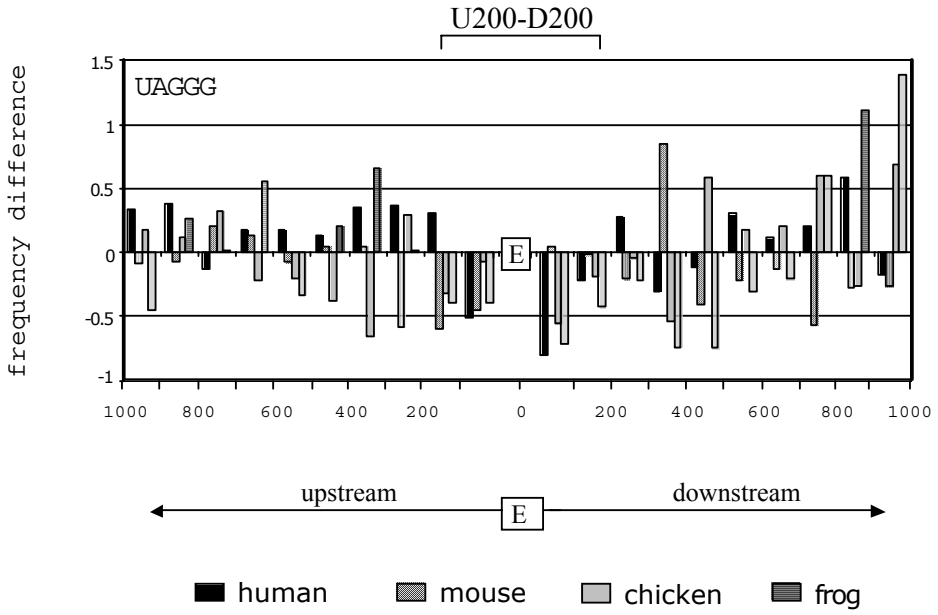

Figure 7. Model for regulation of muscle-enriched exons

Table I. Over-representation and correlation data for candidate hexamer regulatory motifs in the D200 region.

words	muscle freq. (rank)	control freq.	freq. diff.	hypergeo. p-value	hypergeo. q-value	correlation p-value
UGCAUG*	2.72 (1)	0.29	2.44	$2.89x10^{-15}$	$1.05x10^{-11}$	$6.79X10^{-5}$
UUUGCA*	1.56 (4)	0.25	1.31	$8.14X10^{-9}$	$1.49X10^{-5}$	0.0082
UUCUGU	1.56 (4)	0.34	1.21	$2.22X10^{-4}$	$1.17X10^{-2}$	0.248
UGUUUG	1.56 (4)	0.40	1.16	$1.77X10^{-4}$	$1.05X10^{-2}$	0.190
UUGUUU	1.46 (8)	0.37	1.09	$6.14X10^{-4}$	$2.24X10^{-2}$	0.534
GUGUGU*	1.65 (3)	0.58	1.07	$3.57X10^{-4}$	$1.57X10^{-2}$	0.032
UUGCAU*	1.17 (21)	0.11	1.06	$1.21X10^{-7}$	$6.29X10^{-5}$	0.036
UGUGUG	1.85 (2)	0.83	1.02	$2.95X10^{-4}$	$1.38X10^{-2}$	0.121
UUUUCU	1.46 (8)	0.48	0.99	$3.51X10^{-3}$	$6.28X10^{-2}$	0.221
AUGCAU*	1.07 (33)	0.10	0.99	$2.91X10^{-8}$	$2.72X10^{-5}$	0.011
UGUGUU	1.36 (10)	0.39	0.97	$1.36X10^{-4}$	$9.06X10^{-3}$	0.569
AUUUUU	1.26 (16)	0.30	0.96	$3.56X10^{-6}$	$1.10X10^{-3}$	0.468
UGUUUU	1.36 (10)	0.41	0.96	$3.07X10^{-4}$	$1.42X10^{-2}$	0.146
GCUUUU	1.17 (21)	0.21	0.96	$1.33X10^{-4}$	$9.02X10^{-3}$	0.234
UCUGUU	1.26 (16)	0.31	0.96	$1.97X10^{-4}$	$1.12X10^{-2}$	0.453
UUCAAA	1.07 (33)	0.13	0.94	$4.45X10^{-8}$	$3.25X10^{-5}$	0.219
UUAAAU	1.07 (33)	0.13	0.94	$1.21X10^{-7}$	$6.29X10^{-5}$	0.583
CUGCUU	1.36 (10)	0.44	0.92	$4.54X10^{-4}$	$1.86X10^{-2}$	0.199
UUCUAA	1.07 (33)	0.15	0.92	$4.74X10^{-6}$	$1.24X10^{-3}$	0.154
UUUGUU	1.26 (16)	0.35	0.92	$5.31X10^{-2}$	$3.10x10^{-1}$	0.335
GCAUGG*	1.36 (10)	0.47	0.89	$4.47X10^{-5}$	$4.08X10^{-3}$	0.0006
UCUGUG*	1.56 (4)	0.66	0.89	$2.10X10^{-4}$	$1.14X10^{-2}$	0.015
AAUUUU	1.07 (33)	0.19	0.88	$3.66X10^{-4}$	$1.57X10^{-2}$	0.600
GCAUGC	1.07 (33)	0.19	0.88	$1.41X10^{-5}$	$2.44X10^{-3}$	0.088
UGCUUU	1.26 (16)	0.39	0.88	$7.54X10^{-4}$	$2.60X10^{-2}$	0.402
GUGCAU*	1.07 (33)	0.20	0.87	$6.82X10^{-5}$	$5.66X10^{-3}$	0.0068
AGCUUU	1.07 (33)	0.21	0.86	$1.91X10^{-5}$	$2.44X10^{-3}$	0.161
AGAAAU	0.97 (52)	0.14	0.84	$1.59X10^{-5}$	$2.44X10^{-3}$	0.366
UUUUUG	1.17 (21)	0.34	0.83	$1.40X10^{-4}$	$9.16X10^{-3}$	0.479
ACUAAC*	0.88 (69)	0.05	0.82	$2.98X10^{-8}$	$2.72X10^{-5}$	0.0006

* Over-represented motifs with highly significant hypergeometric and correlation p-values.

Table II. Over-represented PWM sequences in the D200, D50, and U200 intron regions.

D200	PWM[1]	P values[2]
1	WGCATK	2.35×10^{-07}
2	WGMHTD	3.39×10^{-07}
3	GCATRN	8.19×10^{-07}
4	DWGCAT	3.06×10^{-06}
5	NWGMWT	3.71×10^{-06}
6	WGHHTD	7.11×10^{-05}
7	NHCTAA	1.24×10^{-04}
8	STKTGS	2.04×10^{-04}
9	CTGYSR	3.39×10^{-04}
10	HCTAAN	3.39×10^{-04}

D50	PWM	P values
1	TGCM	4.76×10^{-04}
2	GCAT	2.81×10^{-03}
3	CTTG	4.49×10^{-02}
4	CCGA	0.64

U200	PWM	P values
1	VYCCHT	9.50×10^{-05}
2	YMCYYN	1.1×10^{-04}
3	TYYCCM	5.06×10^{-04}
4	CCCTNM	8.73×10^{-04}
5	YMCYYW	8.98×10^{-04}
6	TCCMTY	1.11×10^{-03}
7	HTTTCY	1.11×10^{-03}
8	RYYCHY	1.43×10^{-03}
9	YHTTTC	1.45×10^{-03}
10	CYHTYY	2.15×10^{-03}

1. Nucleotide abbreviations: W = A or T; S = G or C; D = A or G or T; V = A or C or G; H = A or C or U; M = A or C; K = G or U; R = A or G; Y = C or U.

2. Shown are the correlation P-values.

References

1. 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.
2. Eperon, I.C., Makarova, O.V., Mayeda, A., Munroe, S.H., Caceres, J.F., Hayward, D.G. and Krainer, A.R. (2000) Selection of alternative 5' splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. *Mol Cell Biol*, **20,** 8303-8318.
3. 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.
4. Guil, S., Gattoni, R., Carrascal, M., Abian, J., Stevenin, J. and Bach-Elias, M. (2003) Roles of hnRNP A1, SR proteins, and p68 helicase in c-H-ras alternative splicing regulation. *Mol Cell Biol*, **23,** 2927-2941.
5. Hanamura, A., Caceres, J.F., Mayeda, A., Franza, B.R., Jr. and Krainer, A.R. (1998) Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. *RNA*, **4,** 430-444.
6. Nasim, M.T., Chernova, T.K., Chowdhury, H.M., Yue, B.G. and Eperon, I.C. (2003) HnRNP G and Tra2beta: opposite effects on splicing matched by antagonism in RNA binding. *Hum Mol Genet*, **12,** 1337-1348.
7. Polydorides, A.D., Okano, H.J., Yang, Y.Y., Stefani, G. and Darnell, R.B. (2000) A brain-enriched polypyrimidine tract-binding protein antagonizes the ability of nova to regulate neuron-specific alternative splicing. *Proc Natl Acad Sci U S A*, **97,** 6350-6355.
8. 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.
9. 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.
10. 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.
11. Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M. *et al.* (2005) Nova regulates brain-specific splicing to shape the synapse. *Nat Genet*, **37,** 844-852.
12. 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.
13. 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.
14. 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.
15. Faustino, N.A. and Cooper, T.A. (2005) Identification of putative new splicing targets for ETR-3 using sequences identified by systematic evolution of ligands by exponential enrichment. *Mol Cell Biol*, **25,** 879-887.
16. Ho, T.H., Charlet, B.N., Poulos, M.G., Singh, G., Swanson, M.S. and Cooper, T.A. (2004) Muscleblind proteins regulate alternative splicing. *Embo J*, **23,** 3103-3112.
17. 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.

18. Ladd, A.N., Taffet, G., Hartley, C., Kearney, D.L. and Cooper, T.A. (2005) Cardiac tissue-specific repression of CELF activity disrupts alternative splicing and causes cardiomyopathy. *Mol Cell Biol*, **25,** 6267-6278.
19. Suzuki, H., Jin, Y., Otani, H., Yasuda, K. and Inoue, K. (2002) Regulation of alternative splicing of alpha-actinin transcript by Bruno-like proteins. *Genes Cells*, **7,** 133-141.
20. 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.
21. Fehlbaum, P., Guihal, C., Bracco, L. and Cochet, O. (2005) A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucleic Acids Res*, **33,** e47.
22. Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R. and Shoemaker, D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302,** 2141-2144.
23. Le, K., Mitsouras, K., Roy, M., Wang, Q., Xu, Q., Nelson, S.F. and Lee, C. (2004) Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res*, **32,** e180.
24. Lee, C. and Roy, M. (2004) Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol*, **5,** 231.
25. Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T.R., Morris, Q.D. *et al.* (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell*, **16,** 929-941.
26. Relogio, A., Ben-Dov, C., Baum, M., Ruggiu, M., Gemund, C., Benes, V., Darnell, R.B. and Valcarcel, J. (2005) Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. *J Biol Chem*, **280,** 4779-4784.
27. Srinivasan, K., Shiue, L., Hayes, J.D., Centers, R., Fitzwater, S., Loewen, R., Edmondson, L.R., Bryant, J., Smith, M., Rommelfanger, C. *et al.* (2005) Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods*, **37,** 345-359.
28. Sugnet, C.W., Srinivasan, K., Clark, T.A., O'Brien, G., Cline, M.S., Wang, H., Williams, A., Kulp, D., Blume, J.E., Haussler, D. *et al.* (2006) Unusual Intron Conservation near Tissue-Regulated Exons Found by Splicing Microarrays. *PLoS Comput Biol*, **2,** e4.
29. Yeakley, J.M., Fan, J.B., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M.S. and Fu, X.D. (2002) Profiling alternative splicing on fiber-optic arrays. *Nat Biotechnol*, **20,** 353-358.
30. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409,** 860-921.
31. 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., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291,** 1304-1351.
32. 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.
33. Bussemaker, H.J., Li, H. and Siggia, E.D. (2001) Regulatory element detection using correlation with expression. *Nat Genet*, **27,** 167-171.
34. Das, D., Banerjee, N. and Zhang, M.Q. (2004) Interacting models of cooperative gene regulation. *Proc Natl Acad Sci U S A*, **101,** 16234-16239.
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. 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.

37. Southby, J., Gooding, C. and Smith, C.W. (1999) Polypyrimidine tract binding protein functions as a repressor to regulate alternative splicing of alpha-actinin mutally exclusive exons. *Mol Cell Biol*, **19,** 2699-2711.
38. Zhang, L., Liu, W. and Grabowski, P.J. (1999) Coordinate repression of a trio of neuron-specific splicing events by the splicing regulator PTB. *RNA*, **5,** 117-130.
39. Chan, R.C. and Black, D.L. (1995) Conserved intron elements repress splicing of a neuron-specific c-src exon in vitro. *Mol Cell Biol*, **15,** 6377-6385.
40. Perez, I., Lin, C.H., McAfee, J.G. and Patton, J.G. (1997) Mutation of PTB binding sites causes misregulation of alternative 3' splice site selection in vivo. *RNA*, **3,** 764-778.
41. Singh, R., Valcarcel, J. and Green, M.R. (1995) Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science*, **268,** 1173-1176.
42. 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.
43. Wagner, E.J. and Garcia-Blanco, M.A. (2002) RNAi-mediated PTB depletion leads to enhanced exon definition. *Mol Cell*, **10,** 943-949.
44. Burd, C.G. and Dreyfuss, G. (1994) RNA binding specificity of hnRNP A1: significance of hnRNP A1 high- affinity binding sites in pre-mRNA splicing. *EMBO J*, **13,** 1197-1204.
45. Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, **4,** R28.
46. 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., Winkelmann, J.C. *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.
47. Auweter, S.D., Fasan, R., Reymond, L., Underwood, J.G., Black, D.L., Pitsch, S. and Allain, F.H. (2006) Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. *EMBO J*, **25,** 163-173.
48. Nasim, F.U., Hutchison, S., Cordeau, M. and Chabot, B. (2002) High-affinity hnRNP A1 binding sites and duplex-forming inverted repeats have similar effects on 5' splice site selection in support of a common looping out and repression mechanism. *Rna*, **8,** 1078-1089.
49. 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.
50. Ho, T.H., Savkur, R.S., Poulos, M.G., Mancini, M.A., Swanson, M.S. and Cooper, T.A. (2005) Colocalization of muscleblind with RNA foci is separable from mis-regulation of alternative splicing in myotonic dystrophy. *J Cell Sci*, **118,** 2923-2933.
51. Jiang, H., Mankodi, A., Swanson, M.S., Moxley, R.T. and Thornton, C.A. (2004) Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons. *Hum Mol Genet*, **13,** 3079-3088.
52. Lu, X., Timchenko, N.A. and Timchenko, L.T. (1999) Cardiac elav-type RNA-binding protein (ETR-3) binds to RNA CUG repeats expanded in myotonic dystrophy. *Hum Mol Genet*, **8,** 53-60.
53. Mankodi, A., Logigian, E., Callahan, L., McClain, C., White, R., Henderson, D., Krym, M. and Thornton, C.A. (2000) Myotonic dystrophy in transgenic mice expressing an expanded CUG repeat. *Science*, **289,** 1769-1773.
54. Mankodi, A., Takahashi, M.P., Jiang, H., Beck, C.L., Bowers, W.J., Moxley, R.T., Cannon, S.C. and Thornton, C.A. (2002) Expanded CUG repeats trigger aberrant splicing of ClC-1 chloride channel pre-mRNA and hyperexcitability of skeletal muscle in myotonic dystrophy. *Mol Cell*, **10,** 35-44.

55. Philips, A.V., Timchenko, L.T. and Cooper, T.A. (1998) Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. *Science*, **280,** 737-741.
56. 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.
57. Sugnet, C.W., Kent, W.J., Ares, M., Jr. and Haussler, D. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac Symp Biocomput***,** 66-77.
58. Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T. and Burge, C.B. (2005) Identification and analysis of alternative splicing events conserved in human and mouse. *Proc Natl Acad Sci U S A*, **102,** 2850-2855.
59. Xu, Q. and Lee, C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res*, **31,** 5635-5643.
60. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res*, **12,** 996-1006.
61. Hastie, T., Tibshirani, R. and Friedman, J.H. (2001) *The Elements of Statistical Learning*. Springer Verlag, New York.
62. 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.
63. 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.
64. 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.
65. Schones, D.E., Sumazin, P. and Zhang, M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21,** 307-313.

Supplementary Table I . Human muscle-enriched exons and their orthologs in mouse, chicken, and frog.

	human exons (56)	representative PSID	Exon Coordinates May04 UCSC	parent gene size	mouse (54)	chicken (43)	frog (36)
1	*MTMR3*	93393158	chr22:28,744,000-28,744,026	145kb	+	+	+
2	*RBAF600*	99925658	chr1:19,221,563-19,221,595	135kb	+	+	+
3	*RSN*	95562371	chr12:121,360,538-121,360,570	151kb	+	+	no
4	*FAT*	98278525	chr4:187,886,671-187,886,706	136kb	+	+	+
5	*INSR*	93783876	chr19:7,101,508-7,101,543	177kb	+	no	no
6	*ACAS2*	93630018	chr20:32,966,682-32,966,720	51kb	+	+	+
7	*TNNT3*	105671755	chr11:1,914,769-1,914,809	19kb	+	+	+
8	*DYSF*	99552338	chr2:71,688,135-71,688,176	233kb	+	+	no
9	*TMEM16E*	96039431	chr11:22,196,368-22,196,409	87kb	+	+	+
10	*RBM9*	103730784	chr22:34,473,160-34,473,202	284kb	+	+	+
11	*MYO9B*	103967726	chr19:17,182,142-17,182,189	112kb	+	+	+
12	*WDFY3*	98362890	chr4:86,005,193-86,005,243	297kb	+	+	no
13	*PRDM2*	100258831	chr1:13,840,919-13,840,974	123kb	+	+	no
14	*ARGBP2*	98280215	chr4:186,931,658-186,931,714	369kb	+	+	+
15	*SORBS1* (60nt)	96137820	chr10:97,072,493-97,072,552	250kb	+	+	+
16	*NBEA*	95353689	chr13:35,118,006-35,118,068	80kb	+	+	+
17	*TEAD3*	97648546	chr6:35,555,815-35,555,877	23kb	+	+	+
18	*SORBS1* (66nt)	96138591	chr10:97,121,731-97,121,796	250kb	+	+	+
19	*MFN2*	100261811	chr1:11,976,846-11,976,913	33kb	+	no	no
20	*LRRFIP2*	98804302	chr3:37,121,950-37,122,018	123kb	+	no	+
21	*LRRFIP1*	99364359	chr2:238,441,843-238,441,914	73kb	+	+	+
22	*SORBS1* (75nt)	96138411	chr10:97,144,748-97,144,822	250kb	+	+	+
23	*Y226*	98619346	chr3:198,906,255-198,906,329	66kb	+	+	no
24	*TPM2*	105745930	chr9:35,674,729-35,674,804	8kb	+	+	+
25	*TPM1*	104276978	chr15:61,143,316-61,143,394	19kb	+	+	+
26	*FXR1*	105005302	chr3:182,171,565-182,171,645	64kb	+	+	+
27	*ITGB1*	104606546	chr10:33,235,997-33,236,077	58kb	+	+	no
28	*ABLIM1*	96108548	chr10:116,235,046-116,235,129	249kb	+	+	no
29	*PPP1R12A*	95444717	chr12:78,673,817-78,673,900	161kb	+	+	no
30	*FBXO31*	94437555	chr16:85,949,518-85,949,604	54kb	+	+	+
31	*ANK2* (92nt)	98509599	chr4:114,656,850-114,656,941	334kb	+	no	+
32	*ANK2* (93nt)	98509569	chr4:114,651,293-114,651,385	334kb	+	+	+
33	*SVIL*	96212858	chr10:29,855,896-29,855,991	278kb	+	+	+
34	*CLASP1*	99203420	chr2:121,865,045-121,865,166	312kb	+	+	+
35	*SORBS1* (102nt)	96138601	chr10:97,121,073-97,121,174	250kb	+	+	+
36	*CAPZB*	99923983	chr1:19,414,542-19,414,674	147kb	+	+	+
37	*PLD1*	98655298	chr3:172,887,177-172,887,290	210kb	+	+	no
38	*BRAF*	97125440	chr7:139,938,650-139,938,806	190kb	+	+	+
39	*CTL2*	103985997	chr19:10,614,573-10,614,697	19kb	+	no	no
40	*SLC25A3*	105519144	chr12:97,491,679-97,491,803	8kb	+	+	+
41	*TPM1*	94896256	chr15:61,123,279-61,123,404	19kb	+	+	+
42	*UNR*	99800031	chr1:114,996,190-114,996,336	33kb	+	no	+
43	*CACNB1*	94197648	chr17:34,595,719-34,595,893	13kb	+	no	no
44	*NEK6*	96612790	chr9:124,134,672-124,134,856	93kb	no	no	no
45	*SMTN*	93390102	chr22:29,821,425-29,821,589	23kb	+	+	no
46	*TBC1D4*	95223108	chr13:74,796,374-74,796,538	197kb	+	+	+
47	*TRIP10*	103991857	chr19:6,697,040-6,697,207	12kb	+	no	no
48	*MAST2*	100199876	chr1:46,095,021-46,095,235	232kb	+	+	+
49	*LDB3*	104624683	chr10:88,436,806-88,437,009	31kb	+	+	+
50	*NACA*	95467208	chr12:55,395,922-55,396,257	14kb	+	no	no
51	*UBE4B*	10026653	chr1:10,100,521-10,100,907	148kb	+	+	+
52	*PRKWNK1*	95711975	chr12:859,000-859,458	155kb	+	+	+
53	*MAPT*	94343664	chr17:41,416,381-41,417,133	131kb	+	no	no
54	*SORBS1* (774nt)	96137735	chr10:97,086,268-97,087,041	250kb	+	+	+
55	*NACA* (1734nt)	95467147	chr12:55,399,777-55,401,510	14kb	no	no	no
56	*TACC2*	96273853	chr10:123,832,152-123,837,464	270kb	+	no	no

Supplementary Table II. Candidate regulatory motifs for muscle-specific alternative splicing in the proximal downstream intron (D200) of several vertebrate species.

dataset	species	muscle freq. x 10^{-3} (rank)	control freq. x 10^{-3}	frequency difference	p values
human muscle	UGCAUG	2.72 (1)	0.29	2.44	0.0004
	UUUGCA	1.56 (4)	0.25	1.31	0.0351
	UUCUGU	1.56 (4)	0.34	1.21	0.0359
	UGUUUG	1.56 (4)	0.40	1.16	0.0362
	UUGUUU	1.46 (8)	0.37	1.09	0.0363
	GUGUGU	1.65 (3)	0.58	1.07	0.0363
	UUGCAU	1.17 (21)	0.11	1.06	0.0363
	UGUGUG	1.85 (2)	0.83	1.02	0.0370
	UUUUCU	1.46 (8)	0.48	0.99	0.0374
	AUGCAU	1.07 (33)	0.10	0.99	0.0374
mouse muscle	UGCAUG	2.20 (2)	0.65	1.55	0.0122
	ACUAAC	1.34 (9)	0.12	1.21	0.0418
	GCAUGG	1.53 (5)	0.37	1.16	0.0418
	UGUGUG	2.29 (1)	1.28	1.01	0.0579
	UACUAA	1.05 (26)	0.05	1.00	0.0812
	UGCUUG	1.34 (9)	0.37	0.97	0.0812
	CAUUAA	1.05 (26)	0.12	0.93	0.0827
	CUGUGU	1.81 (3)	0.89	0.92	0.0827
	UGUGUU	1.53 (5)	0.61	0.91	0.0827
	UUAACU	0.95 (43)	0.11	0.85	0.1137
chicken muscle	UGCAUG	2.76 (1)	0.48	2.28	<0.0001
	UGUGUG	2.52 (2)	0.62	1.90	<0.0001
	GUGUGU	2.16 (3)	0.41	1.75	<0.0001
	CUGCAU	2.16 (3)	0.71	1.45	0.0001
	AGUGUG	1.80 (10)	0.38	1.41	0.0001
	GUGUGC	1.80 (10)	0.39	1.41	0.0001
	UUUUGU	2.16 (3)	0.77	1.39	0.0001
	UUUGUG	1.92 (7)	0.62	1.30	0.0003
	CUGUUU	1.92 (7)	0.66	1.26	0.001
	AAGUGU	1.56 (15)	0.30	1.26	0.001
frog muscle	UGCAUG	3.58 (1)	0.49	3.09	<0.0001
	GUGUGU	2.15 (5)	0.44	1.71	0.0001
	ACUAAC	1.72 (9)	0.20	1.51	0.0005
	UGUGUU	2.29 (2)	0.85	1.44	0.0007
	CUGUUU	2.15 (5)	0.75	1.40	0.0017
	UUUGCA	2.29 (2)	0.92	1.37	0.0017
	GCAUGU	1.72 (9)	0.38	1.34	0.0017
	UGUGUG	1.86 (7)	0.61	1.25	0.0048
	GCAUGC	1.43 (23)	0.18	1.25	0.0048
	UACUAA	1.72 (9)	0.48	1.24	0.0048
human tissue nonspecific	AUUUUU	1.60 (3)	0.30	1.30	<0.0001
	UAUUUU	1.39 (6)	0.20	1.20	<0.0001
	CUUUUU	1.55 (4)	0.37	1.18	<0.0001
	UUUUUU	1.70 (1)	0.53	1.04	<0.0001
	UCUUUU	1.34 (17)	0.31	1.02	<0.0001
	UUUAAU	1.19 (22)	0.17	0.98	0.0001
	AAAGAA	1.13 (11)	0.16	0.95	0.0001
	UUUUUA	1.24 (22)	0.29	0.90	0.0002
	AAAAAU	1.13 (27)	0.23	0.89	0.0001
	UAAUUU	1.08 (11)	0.19	0.89	0.0001

Supplementary Table III. Candidate regulatory motifs in the proximal upstream intron (U200).

dataset	species	muscle freq. x 10^{-3} (rank)	control freq. x 10^{-3}	frequency difference	p values
human muscle	UUUUUU	3.70 (1)	0.38	3.32	<0.0001
	UUCUUU	3.30 (3)	0.47	2.84	<0.0001
	UUUCUU	3.31 (2)	0.5	2.73	<0.0001
	AUUUUU	2.82 (4)	0.35	2.48	<0.0001
	UUUUGU	2.82 (5)	0.41	2.41	<0.0001
	UUUUUC	2.72 (8)	0.40	2.32	<0.0001
	UCUUUU	2.72 (6)	0.43	2.30	<0.0001
	UUUUUG	2.63 (10)	0.36	2.26	<0.0001
	CUUUUU	2.63 (9)	0.42	2.21	<0.0001
	UUUUCU	2.72 (7)	0.66	2.06	<0.0001
mouse muscle	UUUUCU	2.86 (2)	0.88	1.99	<0.0001
	UUUCUU	2.96 (1)	1.02	1.94	<0.0001
	UUUUUU	2.48 (4)	0.56	1.92	<0.0001
	UUCUUU	2.77 (3)	0.98	1.78	<0.0001
	UUUUUC	2.00 (11)	0.42	1.58	<0.0001
	CUCUCU	2.39 (5)	1.00	1.38	0.0001
	CCUCCU	1.91 (16)	0.56	1.35	0.0014
	CUCCUC	1.91 (14)	0.62	1.29	0.0016
	UUUCCU	2.10 (9)	0.83	1.27	0.0033
	UUUUGU	2.00 (12)	0.77	1.23	0.0035
chicken muscle	UUCUUU	3.72 (1)	1.33	2.39	0.0001
	UUUUUC	3.60 (3)	1.28	2.32	0.0002
	UUUCUU	3.60 (2)	1.34	2.26	0.0002
	UUUUCU	3.36 (4)	1.27	2.09	0.0006
	UUUUUU	3.12 (5)	1.39	1.73	0.0029
	CUUUUU	2.76 (6)	1.06	1.70	0.0033
	CUGUUU	2.52 (10)	0.84	1.68	0.0034
	UUUCCU	2.76 (7)	1.10	1.66	0.0035
	UCUUUU	2.76 (8)	1.14	1.62	0.0043
	CUCUCU	2.16 (17)	0.56	1.60	0.0061
frog muscle	UUUUCU	3.87 (1)	1.93	1.93	0.0011
	UUUUUC	3.58 (2)	1.84	1.74	0.0049
	CCCCCC	1.58 (33)	1.28	1.45	0.0248
	UUUCUC	2.15 (12)	0.73	1.42	0.0248
	UUCCCC	1.58 (32)	2.27	1.35	0.0300
	UUCCCU	1.58 (34)	0.35	1.22	0.0539
	UUUCUU	3.01 (4)	1.86	1.15	0.0808
	CUUUUU	3.15 (3)	2.02	1.13	0.0810
	UUUCCC	1.58 (35)	0.47	1.10	0.0834
	CCUUUU	2.00 (17)	0.90	1.10	0.0834
human tissue nonspecific	UUUUUU	2.99 (1)	0.38	2.61	<0.0001
	AAUUUU	1.70 (7)	0.47	1.54	<0.0001
	UUUUUG	1.86 (4)	0.57	1.49	<0.0001
	CUUUUU	1.86 (5)	0.35	1.44	<0.0001
	UUUCUU	1.91 (3)	0.41	1.33	<0.0001
	UUUUCU	1.96 (2)	0.40	1.30	<0.0001
	UUUUGU	1.70 (6)	0.43	1.29	<0.0001
	AUUUUU	1.49 (11)	0.36	1.15	0.0001
	UUUUUC	1.55 (10)	0.42	1.14	0.0001
	UCUUUU	1.55 (9)	0.66	1.12	0.0001

Supplementary Table IV. Top scoring Gene Ontology categories represented in the muscle-enriched dataset

GO_term	Enrichment	Total genes with term	Muscle exon genes with term	p_value	Description
GO:0007010	6.2	302	7	0.00011	cytoskeleton organization and biogenesis
GO:0007026	75.9	7	2	0.00029	microtubule stabilization
GO:0006996	5.2	358	7	0.00032	organelle organization and biogenesis
GO:0007028	5.1	366	7	0.00037	cytoplasm organization and biogenesis
GO:0007517	9.9	107	4	0.00066	muscle development