7916–7926 *Nucleic Acids Research, 2010, Vol. 38, No. 22* *Published online 11 August 2010*
doi:10.1093/nar/gkq705

A network of conserved co-occurring motifs for the regulation of alternative splicing

Mikita Suyama[1,2], Eoghan D. Harrington[1], Svetlana Vinokourova[3], Magnus von Knebel Doeberitz[3], Osamu Ohara[4,5] and Peer Bork[1,6,*]

[1]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany, [2]Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, 606-8501 Kyoto, Japan, [3]Department of Applied Tumor Biology, Institute of Pathology, University of Heidelberg, 69120 Heidelberg, Germany, [4]Department of Human Genome Research, Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, 292-0818 Chiba, [5]RIKEN Research Center for Allergy and Immunology, Suehiro-cho 1-7-22, Tsurumi-ku, Yokohama, 230-0045 Kanagawa, Japan and [6]Max Delbrück Center for Molecular Medicine, 13125 Berlin-Buch, Germany

Received November 23, 2009; Revised July 21, 2010; Accepted July 26, 2010

ABSTRACT

***Cis*-acting short sequence motifs play important roles in alternative splicing. It is now possible to identify such sequence motifs as conserved sequence patterns in genome sequence alignments. Here, we report the systematic search for motifs in the neighboring introns of alternatively spliced exons by using comparative analysis of mammalian genome alignments. We identified 11 conserved sequence motifs that might be involved in the regulation of alternative splicing. These motifs are not only significantly overrepresented near alternatively spliced exons, but they also co-occur with each other, thus, forming a network of *cis*-elements, likely to be the basis for context-dependent regulation. Based on this finding, we applied the motif co-occurrence to predict alternatively skipped exons. We verified exon skipping in 29 cases out of 118 predictions (25%) by EST and mRNA sequences in the databases. For the predictions not verified by the database sequences, we confirmed exon skipping in 10 additional cases by using both RT–PCR experiments and the publicly available RNA-Seq data. These results indicate that even more alternative splicing events will be found with the progress of large-scale and high-throughput analyses for various tissue samples and developmental stages.**

INTRODUCTION

Alternative splicing is a cellular process that provides distinct, context-dependent isoforms of genes to generate proteomic diversity (1), and is regulated in a tissue- and developmental stage-specific manner. The fraction of human genes that undergo alternative splicing had been first estimated as ~40% based on mRNA and EST sequence data (2,3). It has dramatically increased with the progress of microarray and high-throughput sequencing technologies, and the current estimate is as high as 74–95% of multiexon genes in human (4–6).

Deciphering the set of rules that govern the regulation of alternative splicing has been an important step toward understanding the mechanisms of alternative splicing (7–9). Besides the core splice signals in splice site selection, there are two major groups of *cis*-acting regulatory sequences for both constitutive and alternative splicing: exonic and intronic splicing regulators. The accumulation of cDNA and EST sequences, and the determination of the human genomic sequences have made it possible to computationally analyze exonic splicing regulators (10,11). However, until recently, it had been difficult to identify intronic splicing regulators by comparative genomic approaches because of the lack of sequence information of introns in other mammalian species than human.

The progress of genome sequencing projects of mammalian species has changed the situation, and now we can make use of genome sequence alignments to identify not only intronic splicing regulators but also various *cis*-

*To whom correspondence should be addressed. Tel: +49 6221 387 8526; Fax: +49 6221 387 517; Email: bork@embl.de
Present address:
Eoghan Harrington, Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305-5124, USA.

regulatory elements. The underlying idea is that, due to the purifying selection acting on functionally important sites, regulatory elements evolve much more slowly than the rest of non-functional sequences. Thus, highly conserved sequences in the genome alignments can be considered as candidate regulatory elements (12). Analyses in this direction successfully identified many *cis*-elements in promoter and 3′-untranslated regions (13) and insulator sites (14). Although it has been reported that many of the alternative splicing events are species-specific, i.e. they are not well conserved even among mammalian species (15–17), several studies have shown that genome alignments can be used to identify potential intronic splicing regulators (18–21).

In alternative splicing, not only a single motif can regulate spatio-temporal patterns of isoforms, but there must be combinatorial mechanisms to achieve complex regulation. To address this issue, we first identified highly significant potential intronic splicing regulators by using genome sequence alignments of mammalian species. Then we analyzed the relationships among these potential regulatory elements in terms of their co-occurrence in the intronic sequences.

MATERIALS AND METHODS

mRNA sequences and their coordinates on the human genome

To detect alternatively skipped exons, we used high-quality non-redundant mRNAs from human in the Reference Sequence (RefSeq) Database (22) (release 16, March 2006, http://www.ncbi.nlm.nih.gov/RefSeq). There are several categories in the RefSeq mRNAs, and we only used well-annotated mRNA sequences, i.e. 'reviewed', 'provisional', and 'validated'. For our purposes, this high-confident set of exon skipping is more effective than using all possible event on the cost of false positives, e.g. by including EST sets. The final number of the mRNA sequences we obtained is 21 754. These mRNA sequences are mapped on the human chromosomes (hg18, March 2006, http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/chromFa.zip) by using the BLAT program (23) to determine the coordinates on the human genomic sequences. For each gene, we compared the structure and coordinates of the transcripts to identify skipped exons. Here, we considered an exon as skipped if: (i) the exon is present in a transcript but completely absent in another one; and (ii) the 3′-end of the upstream exon and the 5′-end of the downstream exon have exactly the same coordinates. In total, we identified 1736 skipped exons in 1473 genes.

Genome sequence alignment

The genome sequence alignments of vertebrate genomes generated by the TBA program (24) were downloaded from the UCSC Genome Browser Database (25) (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz17way). Although the genome alignment contains 17 species, we focused only on mammalian genomes (12 species; release date and the assembly identifiers are in the parenthesis): human (March 2006, hg18), chimpanzee (November 2003, panTro1), macaque (January 2006, rheMac2), mouse (February 2006, mm8), rat (November 2004, rn4), rabbit (May 2005, oryCun1), dog (May 2005, canFam2), cow (March 2005, bosTau2), armadillo (May 2005, dasNov1), elephant (May 2005, loxAfr1), tenrec (July 2005, echTel1) and opossum (January 2006, monDom4).

Counting conserved pentamers in the flanking introns of the skipped exons

We searched for intronic splicing regulators within introns 1000-bases upstream and downstream of the skipped exon (Figure 1). We adopted this length cutoff according to some related studies (17,26,27). The 60 residues at the upstream of 3′ splice site of the skipped exon and the 5 residues at the downstream of the 5′ splice site of the skipped exon are excluded from the analysis because these regions contains splice site signals, polypyrimidine tract and branch point sequences, which are rather highly conserved and might bias our analysis. We are aware that some motifs might exist in these regions, but they will again come at the cost of false positives. If the flanking intron of the skipped exon is shorter than 1000 bases, then only the region up to the neighboring exon is taken. In such cases, the 60 residues upstream of 3′ splice site of the neighboring exon and the 5 residues downstream of 5′ splice site of the neighboring exon are also excluded from the analysis to ensure that the region does not contain splice site signals, polypyrimidine tract, or branch point sequences.

Then we retrieved the multiple sequence alignments of genomic sequences corresponding to these regions, and searched for strictly conserved pentamers among at least 10 mammalian species. Here, we masked the regions that are highly conserved (>85% identity) for more than 50-residues usually *cis*-elements for alternative splicing

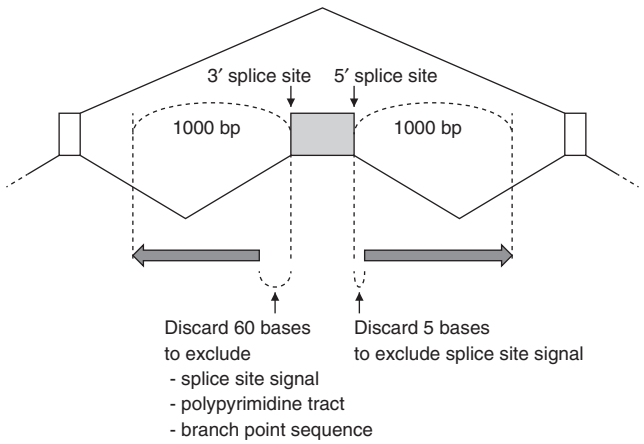

Figure 1. The regions analyzed to search for intronic splicing regulators. The skipped exon is shown in gray box. The regions analyzed are shown in gray arrows. If the flanking intron of the skipped exon is shorter than 1000 bases, then only the region up to the neighboring exon are taken. In such case, the 60 residues upstream of 3′ splice site of the neighboring exon and 5 residues downstream of the 5′ splice site of the neighboring exon are also excluded from the analysis to ensure that the region does not contain splice site signals, polypyrimidine tract, or branch point sequences.

are much shorter and such a long stretch of conserved region might contain unannotated RNA genes embedded within introns or unannotated exons, and counting these conserved regions would bias the analysis. These cutoffs are taken from the average length and sequence identity of orthologous exons between human and mouse (28). In counting conserved pentamers, simple repeats, e.g. runs of Ts (TTTTTTTTTT), are counted only once to prevent over counting.

Statistical significance of conserved pentamers

To assess the statistical significance of the conserved pentamers, first we counted the number of pentamers that are strictly conserved in at least 10 mammalian genomes in the region defined above. We also counted the number of pentamers that are not conserved in the same regions in human. As a control, we counted the numbers of both conserved and unconserved pentamers in the intron between the constitutive exons. For this, we used the same RefSeq gene set (1473 genes), in which we identified exon skipping, and defined an exon as constitutive if there are at least three transcripts for a gene and all the transcripts have exactly the same exon. We set these criteria to minimize the effect of alternative splicing not observed in the current RefSeq database. Then we applied Fisher's exact text to evaluate the specific enrichment of the conserved pentamers in the flanking introns of the skipped exons.

Recent studies have shown that the RNA-Seq method, the transcriptome analysis using next-generation sequencing platforms, reveals more instances of alternative splicing events than the analyses by using mRNA and EST sequences (5,6). This means that some of the constitutive exons that we used as a control set might be alternatively skipped, and might affect our ability to detect motifs. To measure the amount of such cases, we analyzed the exons in our data sets with the RNA-Seq data (5,6) obtained from the GEO database (29). We classified an exon as skipped if there are at least three reads that cover five or more nucleotides of either the upstream and downstream exon. Using the RNA-Seq data from Wang *et al*. (6), only 68 exons out of 3086 exons that we classified as constitutive exons by RefSeq mRNAs are observed to be skipped. We also carried out the same analysis using the RNA-Seq data from Pan *et al*. (5), and detected only 37 exons out of the 3086 constitutive exons as skipped. By merging the results obtained from these two RNA-Seq data, total of 88 (2.9%) constitutive exons are detected as skipped, indicating that the quality of the classification for constitutive exons by using RefSeq mRNAs is high enough to be used as the control.

Since our analysis is based on the conserved pentamers, we may get two or more pentamers from a single highly conserved motif of the length longer than 5 nt. We grouped these pentamers as clustered motifs according to the overlap in position with highly significant probability [$P < 1.0 \times 10^{-12}$ by the cumulative hypergeometric distribution function (30); see below].

Statistical significance of co-occurrence of motifs

The statistical significance of motif co-occurrence was calculated by the cumulative hypergeometric distribution function (30). The probability, P, of obtaining the number of co-occurrence equal to or higher than c purely by chance is calculated by the following equation:

$$P(c) = \sum_{i=c}^{\min(m_1,m_2)} \frac{\binom{m_1}{i}\binom{N-m_1}{m_2-i}}{\binom{N}{m_2}} \quad (1)$$

where m_1 and m_2 are the number of motif1 and motif2, respectively, N the total number of introns examined and i the summation index. For example, if the numbers of motif1 and motif2 are 60 and 80, respectively, and 20 of them co-occur in the total of 800 introns, then the probability to have such co-occurrence purely by chance is calculated as 1.8×10^{-7}.

Prediction of skipped exons in the human gene set

To evaluate whether the motif co-occurrence can be a strong indicator for alternative splicing, we predicted exon skipping in the Ensembl exon set (31) by using the motif co-occurrence and the conservation among the mammalian genomes. For this prediction, we used all Ensembl exons except for those identical with the exons identified in the initial RefSeq analysis. We considered a prediction to be correct if there is at least one for each transcript variant in ESTs from dbEST or mRNAs from GenBank. As a control, all internal exons of the Ensembl gene set are evaluated. The success rates between the predicted set and the control set are evaluated by χ^2 test.

Exon skipping supported by RNA-Seq data

To further confirm our predictions of exon skipping, we used RNA-Seq data (5). From the GEO database at NCBI (29), we downloaded RNA-Seq data (the GEO accession number: GSE13652) for the following six tissues: brain, cerebral cortex, heart, liver, lung and skeletal muscle. For each prediction of exon skipping, we created a 64-nt-long splicing junction sequence, the 5′- and 3′-half of which are from upstream and downstream exons, respectively (5). Next, we mapped the short sequence reads on to the splicing junction sequence by using the bowtie program (32). To eliminate false positive hits, we did not allow any mismatches/indels. We also set a condition that at least five nucleotides of the read should overlap either the upstream or downstream exon.

To see if more tissue data can improve the detection rate of exon skipping events, we applied a re-sampling method to the existing data. Since we have data for six tissue samples, we counted the number of supported exon skipping for all possible combination from one to six samples, and the number of supported exon skipping is plotted along the number of tissue samples we used.

Reverse transcription polymerase chain reaction and amplification of splicing products

Poly (A)+ RNAs from the following 12 human tissues (or brain regions) were purchased from Clontech Laboratories, Inc.(Mountain View, CA): fetal brain, hippocampus and amygdala of adult brain, skeletal muscle, bone marrow, prostate, thymus, pancreas, heart, liver, kidney and testis. We did not use any cancer cell lines, because it is known that there are many cancer-specific alternative splicing events, which might bias the analysis (33). Using the poly (A)+ RNAs, single-stranded cDNAs were prepared as described previously (34). The polymerase chain reaction using the single-stranded cDNA as a template [Reverse transcription polymerase chain reaction (RT–PCR)] was carried out with an LA-Taq polymerase (Takara Bio, Inc., Shiga, JAPAN) in the presence of thermostable Rec A protein as described previously (35). Using thermostable Rec A, we could use the common PCR cycling conditions for all the primer sets given in Supplementary Table S1 as follows: denaturation step, 95°C for 1 min, 30 cycles of PCR consisting of 95°C for 15 s, annealing at 52°C for 30 s and extension at 72°C for 45 s. The final PCR cycle was followed by additional extension reaction at 72°C for 5 min. The PCR products were resolved on 4% NuSeive GTG agarose gel (Takara Bio Inc.) and then detected by fluorescent staining with ethidium bromide. After isolation of the PCR products on agarose gel, DNA sequencing of the PCR products was done on an ABI 3130 (Applied Biosystems, Foster City, CA) using a Big Dye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems).

RESULTS AND DISCUSSION

Identification of intronic splicing regulatory motifs by using genome alignments

To systematically identify and predict sequence motifs that point to the regulation of alternative splicing, we focused on alternatively skipped exons as they are straightforward to measure and are the most frequent type of alternative splicing (36). We compared 21 754 human mRNAs from the well-curated RefSeq database, and identified 1736 skipped exons in 1473 genes (see Supplementary Table S2 for the list of skipped exons).

It has been shown that a certain fraction of alternative splicing is not conserved between human and mouse (16). This means that some alternatively spliced exons that we identified by the comparison of human RefSeq mRNAs might not be alternatively spliced in the other species. However, if the existence of conserved sequences around the exons that are alternatively spliced in human, then this implicates that the conserved sequences is also involved in alternative splicing in other species. Recent studies have also shown that the inter-species comparison by using genome alignments makes it possible to identify many intronic splicing regulators (18–20). Based on these approaches, we retrieved flanking introns for each alternatively skipped exon and constructed alignments of 12 mammalian genomic sequences corresponding to the regions. After filtering out highly conserved 5′ donor and 3′ acceptor sites, polypyrimidine tracts and branch-point sites (Figure 1), we recorded all pentamers that were strictly conserved in at least 10 mammalian genomic sequences. We chose pentamers based on the fact that many splicing factors recognize specific sequences of 4–6 nt in length (37,38). Another reason to use pentamers is that the number of all possible patterns is higher ($4^5 = 1024$) than that for tetramers ($4^4 = 256$), which might help to reduce false counts of patterns. The statistical power comes at the cost of possibly missing some short and degenerated patterns, such as the Nova binding motif (YCAY) (39) or the CU-rich element (40). Longer patterns can be detected though by combining overlapping patterns (see below). To assess whether these pentamers are specifically conserved in the flanking introns of skipped exons, we compared the frequencies of these pentamers with those in the introns flanking constitutive exons as a control and calculated a statistical significance by Fisher's exact test. Then we sorted the pentamers according to the Fisher's *P*-values, and listed 18 pentamers with $P < 1.3 \times 10^{-4}$ (Table 1; see Supplementary Table S3 for the complete list of motifs and the corresponding genes). The cutoff was selected as the highest *P*-value for the pentamer with experimental support (see below). Some of these pentamers are sub-strings of longer motifs; therefore, we clustered these according to their overlaps in the genomic sequence and arrived at a total of 11 distinct motifs.

Although we masked highly conserved regions longer than 50-residues (see 'Materials and Methods' section), there are only 64 intronic regions for the skipped exons out of the total of 1736 skipped exons. As a result, even without the masking, we obtained very similar statistics for the pentamers.

The most significant motif is GCATG (Table1). This pentamer is conserved 157 times in the flanking introns of skipped exons, which corresponds 103 distinct exons from 98 genes. This pentamer, together with the second most significant one, forms a well-characterized hexameric motif, TGCATG, which has been identified as binding site for the Fox-1 protein (37). In some genes this motif is highly conserved among vertebrates (26), and it is also highly frequent in the neighboring introns of skipped exons in nematodes (38). The third and fourth of the pentamers (Table 1) form a hexameric motif, ACTAAC. This motif, which is similar to the branch-point consensus sequence, has been recently identified as specifically enriched in muscle-regulated alternative splicing (27). Among the other significantly overrepresented motifs, some are similar, but not identical, to some previously predicted motifs (Table1). For example, the GTGGTGG G motif is very similar to the purine-rich element that has been shown to be involved in the regulation of alternative splicing in thyroid hormone receptor mRNA (41). This motif is also similar to the G-rich motif, (A/U)GGG, which has been identified as an intronic splicing enhancer (42). Additional novel regulatory motifs are likely to be enriched in our list, but are below our stringent *P*-value cutoff (see Supplementary Table S4 for all the pentamers sorted by the *P*-values).

Table 1. Conserved pentamers with statistical significance

Rank	Conserved pentamer	Clustered motif	*P*-value[a]
1	GCATG	TGCATG	1.0×10^{-29}
2	TGCAT	TGCATG	1.4×10^{-20}
3	ACTAA	ACTAAC	7.5×10^{-12}
4	CTAAC	ACTAAC	9.8×10^{-10}
5	TGCTG	CTGCTGC	8.7×10^{-08}
6	GCTGC	CTGCTGC	3.3×10^{-07}
7	TGCTT	CTTGCTT	7.2×10^{-06}
8	CTTGC	CTTGCTT	8.2×10^{-06}
9	GTGGG	GTGGTGGG	1.1×10^{-05}
10	TTTCT	TTTCT	1.2×10^{-05}
11	AAGAT	AAGAT	3.6×10^{-05}
12	TGGAA	TGGAA	4.2×10^{-05}
13	GCTAA	GCTAA	5.6×10^{-05}
14	CTGCT	CTGCTGC	5.6×10^{-05}
15	AAAGG	AAAGG	6.8×10^{-05}
16	GTGGT	GTGGTGGG	9.8×10^{-05}
17	TCTTG	TCTTG	1.2×10^{-04}
18	GGTGG	GTGGTGGG	1.3×10^{-04}

Pentamers are sorted by *P*-values. The pentamers that cover only a part of known motifs are excluded from the list (see Supplementary Table S4 for a complete list of the pentamers sorted by *P*-value).
[a]*P*-values are calculated by Fisher's exact test.

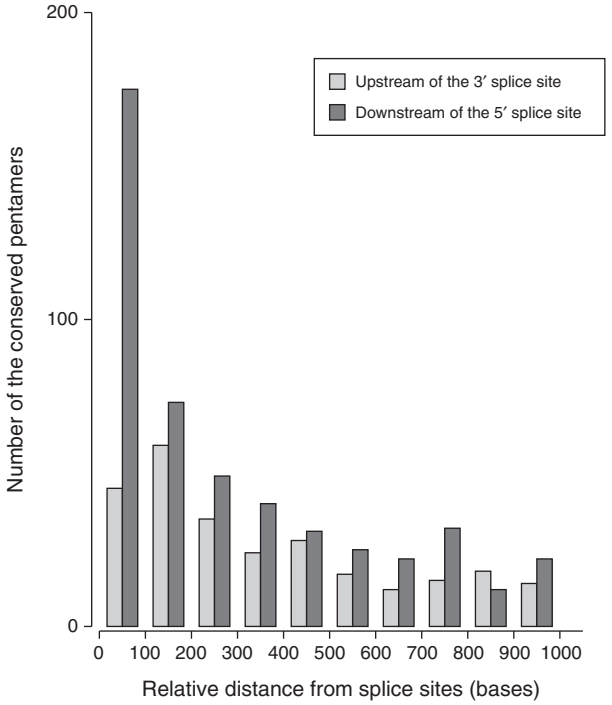

Figure 2. Distribution of the number of the conserved pentamers. The 18 pentamers listed in Table 1 are counted in the flanking introns of alternatively spliced exons. *x*-axis indicates the relative position counting from the corresponding splice site. Upstream of the 3′ splice site and downstream of the 5′ splice site are shown in light and dark gray, respectively.

The length of flanking introns that are considered by different studies varies a lot, e.g. including 100 bases (21) or up to 300 bases (9) from the splice sites. To see the distribution of conserved pentamers around the flanking introns of alternatively skipped exons, we counted the number of 18 pentamers listed in Table 1 (Figure 2). For both upstream of the 3′ splice sites and downstream of the 5′ splice sites, the conserved pentamers are more frequent within 300 bases from the splice sites, but we still see some conserved motifs even up to 1000 bases. Although there must be bona fide motifs in the regions >1 kb apart from the splice sites, the inclusion of those compromises the selectivity of the analysis.

Comparison with the other data sets

We compared the 18 significant pentamers with recently predicted intronic splicing regulators (6,18–20,43) (Table 2). Although all the statistically significant pentamers are also listed in the other data sets at least as a sub-string of the longer patterns, there is no single data set that covers all the 18 pentamers. Four pentamers, GT GGG, AAAGG, GTGGT and GGTGG, which corresponds to two clustered motifs, GTGGTGGG and AAA GG, exist only as a sub-string of the longer patterns in the other data sets. In some data sets, even well-known patterns are missing. For example, the most significant pentamer in our list is missing in three out of four sets of motifs identified by Churbanov *et al.* (20). Another example is ACTAAC, which is missing not only in the three sets of the motifs identified by Churbanov *et al.* (20), but also missing in the list of intron-identity elements identified by Zhang *et al.* (43). Most of the pentamers we identified, especially the ones with lower *P*-values, are also represented in the donor/acceptor intronic flanking region by Voelker *et al.* (18) and the conserved words located in the upstream/downstream of internal exons analyzed by Yeo *et al.* (19).

Analysis of genome alignments with various groups of species

To measure the changes in the power of motif detection with several groups of mammalian species, we analyzed the genome alignments for each group of species. Here, we analyzed the following five groups of species: (i) human and mouse; (ii) Euarchontoglires; (iii) Boreoeutheria; (iv) placental mammals; and (v) mammals (see Table 3 for the species in each group). Since we do not know bona fide *cis*-elements in the genome-wide scale, we focused only on the pentamer, GCATG, which is the part of the well-known *cis*-element for alternative splicing, and counted the number of the conserved pentamers for each groups of species (Table 3). If genome alignments of only human and mouse are analyzed, we obtain more than 300 motifs around skipped exon. But, at the same time, higher number of motifs are found in the introns around constitutive exons, indicating that analyzing only two species, human and mouse, would give a higher ratio of false positive predictions. As a result, we have a higher *P*-value by Fisher's exact test comparing to the sets with more species.

Although the divergence within the group is similar between the set consists of human and mouse (the first

Table 2. Comparison of the 18 pentamers with the patterns in other data sets

	Voelker *et al.* (18)		Yeo *et al.* (19)		Churbanov *et al.* (20)				Wang *et al.* (6)		Zhang *et al.* (43)
	Donor[a] (819)	Acceptor[a] (1007)	Upstream[b] (1069)	Downstream[b] (911)	3′-ISE[c] (814)	3′-ISS[c] (1032)	5′-ISE[c] (478)	5′-ISS[c] (174)	3′SS-intronic[d] (175)	5′SS-intronic[d] (187)	IIE[e] (708)
GCATG	M	M	s	M	–	s	–	–	s	s	s
TGCAT	M	M	M	M	s	–	–	–	s	s	s
ACTAA	M	M	M	s	s	–	–	–	s	s	–
CTAAC	M	M	M	M	M	–	–	–	s	s	–
TGCTG	M	M	s	s	s	s	–	–	s	–	s
GCTGC	M	s	s	s	–	s	s	–	s	–	–
TGCTT	M	M	M	M	–	–	–	–	–	–	s
CTTGC	M	s	M	s	s	s	–	–	s	s	s
GTGGG	s	–	–	s	s	s	s	s	–	–	s
TTTCT	M	M	s	s	s	–	s	–	–	–	s
AAGAT	–	–	–	M	–	s	–	–	s	s	–
TGGAA	s	M	s	s	–	s	–	–	–	–	s
GCTAA	s	M	M	s	s	–	–	–	s	–	–
CTGCT	M	M	M	M	s	–	–	–	–	–	s
AAAGG	s	s	–	–	–	s	–	s	s	–	–
GTGGT	–	–	s	–	–	s	–	–	s	–	s
TCTTG	s	M	M	s	s	–	–	–	s	s	s
GGTGG	s	–	–	s	s	s	s	s	s	–	s

The total number of predicted motifs are indicated in the parenthesis under each data set.
[a]Significantly conserved *n*-mers found in the donor/acceptor intronic flanking region (18).
[b]Conserved words that are located upstream/downstream of internal exons (19).
[c]3′-/5′-splice site-related intronic splicing silencers/enhancers (20).
[d]3′-/5′-splice site intronic motifs (6).
[e]Intron-identity elements (43).
M, exact match; s, the pentamer is included as a sub-string of the longer pattern; –, no match found.

row in Table 3) and Euarchontoglires (the second row in Table 3), number of the GCATG pentamers and the associated *P*-value are very different. This indicates that adding more internal species can drastically improve the power of the *cis*-element detection.

By comparing Euarchontoglires and Boreouetheria, number of the conserved GCATG pentamers around the skipped exons drops from 177 to 158. The difference in the number of motifs might represent the existence of species-specific motifs, i.e. the motifs exist only in Euarchontoglires but do not exist neither in dog nor in cow genomes. Or another explanation for this is that the motifs might not be aligned properly because of the inclusion of distant species.

Motifs are often co-occur in the same flanking introns

When analyzing the genomic positions of these motifs, we found that a significant number of motifs co-occur in the same flanking introns of skipped exons. One such example is found in the downstream of the skipped exon of the nuclear distribution element-like protein (NDEL1; Figure 3A). The two motifs, ACTAAC and TGCATG, are strictly conserved in the 12 mammalian species, and there are no highly conserved segments around this region except for these two motifs. It has been shown that the exon is selectively included in the transcript in melanocyte but not in melanoma (44). This cell line-specific differential expression of the exon suggests a tightly regulated mechanism for the alternative splicing of the exon, and the co-occurring motifs are the most likely candidates to mediate this process. Another example is found in the downstream of the skipped exon of myosin binding protein C (MYBPC1), where TGCATG and TGCTT, are separated by 19–21 residues (Figure 3B). The two examples illustrate that binding of the splicing regulator to the individually well-characterized TGCATG motif appears to be fine-tuned by other factors binding to other neighboring sites. This led to the hypothesis that different motif pair combinations might provide the signature for context-dependent splicing.

By applying a cumulative hypergeometric distribution function [Equation (1)], we obtained statistically significant co-occurring motifs. To determine the cutoff of co-occurrence *P*-value, we compared the distributions of the *P*-values calculated from all possible pairs among the 18 highly significant pentamers and the *P*-values calculated from randomly selected pairs of pentamers (Figure 4). Based on these distributions, we selected the two cutoffs, $P < 1.0 \times 10^{-5}$ and $P < 1.0 \times 10^{-4}$, for statistically significant co-occurrence of motifs. Although we restricted our analysis only to the highly significant motifs listed in Table 1, various combinations of co-occurring motifs are significant, forming a complex network (Figure 5), likely to be the basis for context-dependent regulation. Very recently, Barash *et al.* (9) presented three networks, which consist of motifs and other features, to predict tissue-dependent alternative splicing. Some of the edges in our network are also present in their networks. For example, the edge showing the frequent association of TGCATG and ACTAAC in our

Table 3. GCATG motif in the genome alignments of various groups of species

Species	Conserved GCATG in the introns		*P*-value[a]
	Around skipped exons	Around constitutive exons	
Human + mouse	303	136	1.1×10^{-23}
Euarchontoglires[b]	177	29	1.4×10^{-33}
Boreoeutheria[c]	158	17	6.4×10^{-36}
Placental mammals[d]	140	21	6.8×10^{-28}
Mammals[e]	157	26	1.0×10^{-29}

In the mammals set, we count the motif if it is strictly conserved at least in 10 species.
[a]*P*-values are calculated by Fisher's exact test.
[b]Human, chimpanzee, macaque, mouse, rat, rabbit.
[c]Human, chimpanzee, macaque, mouse, rat, rabbit, dog, cow.
[d]Human, chimpanzee, macaque, mouse, rat, rabbit, dog, cow, armadillo, elephant, tenrec.
[e]Human, chimpanzee, macaque, mouse, rat, rabbit, dog, cow, armadillo, elephant, tenrec, opossum.

network is also present in the network for muscle-specific inclusion of alternative exons (9). Another pair of motifs in our network, TTTCT and TGCATG, can be found in the network for central nervous system-specific inclusion of alternative exons (9). We have these edges (i.e. pairs of motifs) in the same network because we do not take tissue-specificity into account. We consequently analyzed functional GO terms and tissue-specific isoforms in EST and cDNA databases, but did not see any significant association of the co-occurring motifs and functional classes or tissue-specificity (data not shown).

Prediction of alternatively skipped exons by motif co-occurrence

We then tested whether the co-occurrence of the identified motifs in flanking introns of an exon can be applied to predict novel alternatively skipped exons. For this, we extended our analysis to the flanking introns of all annotated exons in the human Ensembl gene set,

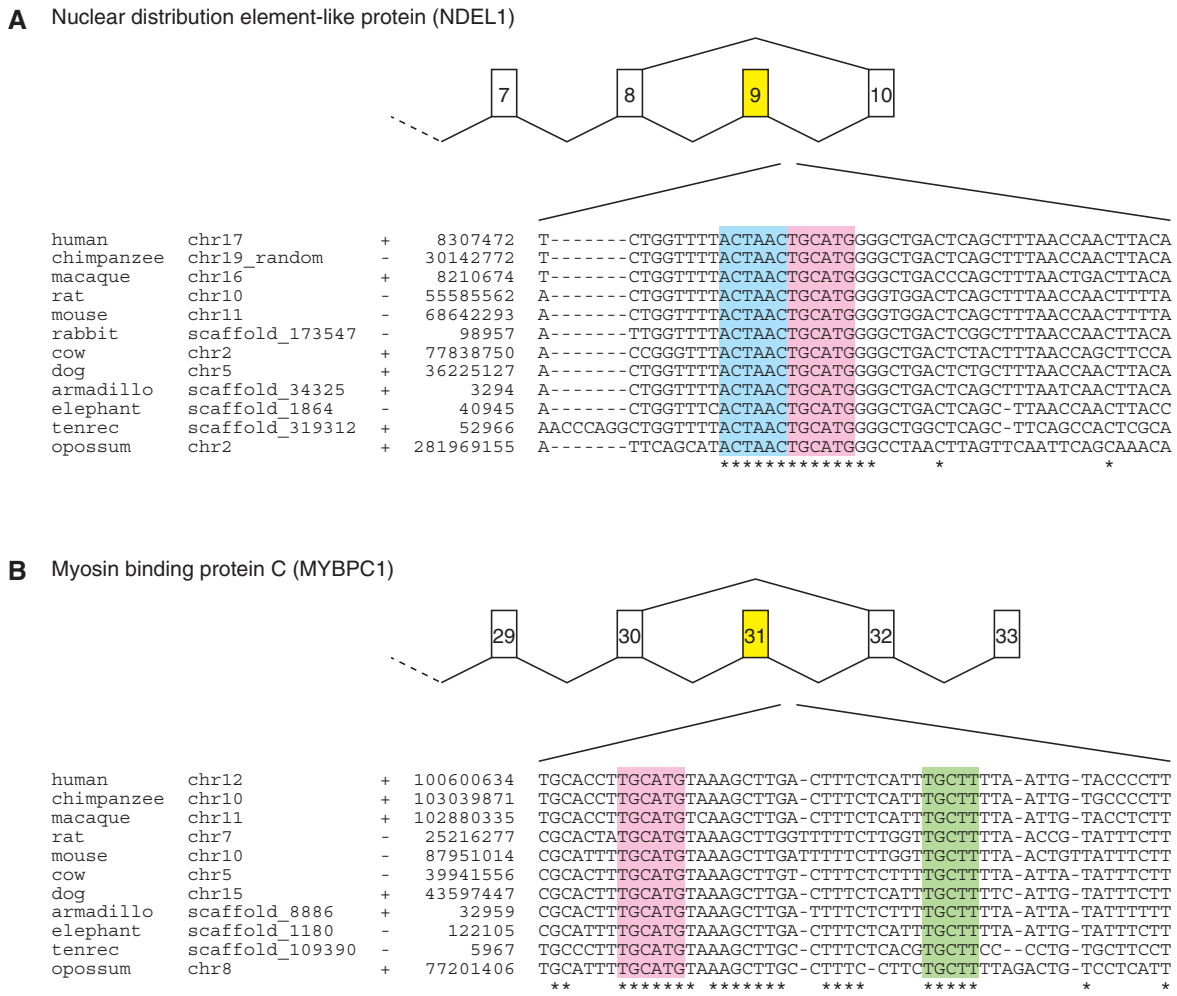

Figure 3. Examples of co-occurring motifs. (**A**) ACTAAC and TGCATG in the flanking intron of the skipped exon of NDEL1 and (**B**) TGCATG and TGCTT in the flanking intron of the skipped exon of MYBPC1. The part of the gene structure around the skipped exon is shown above the alignment with the exon numbers. The format of the alignment: the first column, the name of the species; second column, chromosome or scaffold identifier; the third column, direction of the gene on the chromosome, the fourth column, position on the chromosome. Conserved residues are indicated by asterisks under the alignments. Conserved motifs are indicated by colored boxes.

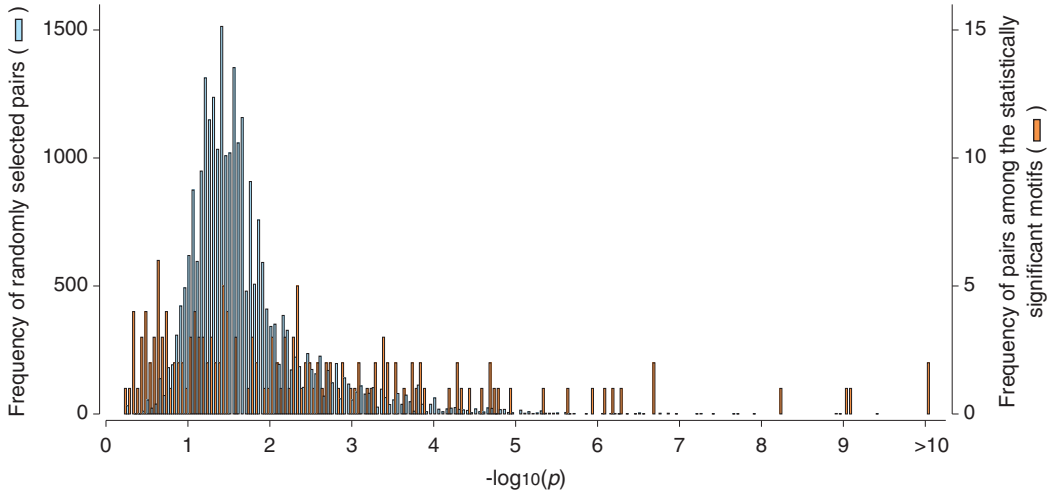

Figure 4. Distribution of the probability calculated by cumulative hypergeometric distribution function. The pairs among the statistically significant motifs are shown in orange and the scale for those is indicated as the vertical axis on the right. The randomly selected pairs are shown in cyan and the scale for those is indicated as the vertical axis on the left.

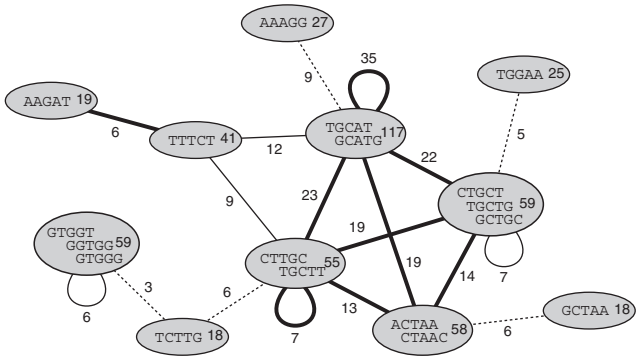

Figure 5. A network of co-occurring motifs. Each node represents a motif. Number of skipped exons with the motif is indicated in the node. If several pentamers consists of a single motif, the pentamers are shown in the node. Each edge represents a pair of co-occurring motifs. Number of skipped exons with a pair of co-occurring motifs is indicated at the edge. Thick and thin lines indicate statistically significant co-occurrence, $P < 1.0 \times 10^{-5}$ and $P < 1.0 \times 10^{-4}$, respectively. The nodes that are not connected with $P < 1.0 \times 10^{-4}$ are linked with dotted lines to the closest motifs. The P-values for these edges are $<1.0 \times 10^{-3}$, except for the one links AAAGG ($P < 2.1 \times 10^{-3}$). The edge connecting the same node indicates the significant co-occurrence of the same motif.

excluding those identified in the initial RefSeq analysis. Using co-occurrence and conservation in mammals of the 11 significant motifs (Table 1), we predicted another 118 alternatively skipped exons in the human Ensembl gene set. Among these 118 exons, 29 (25%) of them have clear indication of exon skipping based on an independent comparison with dbEST and GenBank mRNAs (Supplementary Table S5). This would be a lower estimate of predictive power because of the limited, biased and low-resolution EST coverage in terms of tissues and temporal states. In contrast, of all internal exons in the Ensembl gene set (167 341 in total), only 15 089 (9%) of them seem to be skipped by comparison with dbEST and GenBank mRNAs. This clearly indicates that the number of exon skipping events is significantly higher for the exons with co-occurring motifs in the flanking introns ($P = 3.7 \times 10^{-9}$ by χ^2 text). We further looked for the supporting evidence for the exon skipping in our predictions in recently published RNA-Seq data from six tissue samples (5), and, in addition to the 29 annotated skipping events, RNA-Seq data confirmed 8 additional instances (Supplementary Table S5). To see if more tissue data can improve the detection rate of exon skipping, we applied a re-sampling method to the existing data. The results show that if we have RNA-Seq data for more tissue samples, even more predictions of alternatively skipped exons will be supported although the data are still too sparse to develop a tissue code for exon skipping (Figure 6). For the predictions with no EST and mRNA support, we also randomly selected 10 predicted alternatively skipped exons together with 10 internal exons unlikely to undergo exon skipping and carried out RT–PCR to actively search for the evidence of exon skipping. In 12 tissue libraries, 3 of the 10 exons predicted to be skipped were confirmed to be skipped in certain tissues, whereas no exon skipping was detected for the 10 unpredicted exons as far as we examined (Supplementary Table S6). For example, the skipped form of the third exon of ENST00000256858, as well as the included form, is detected in fetal brain and two regions of adult brain (hippocampus and amygdala) (Figure 7), while it is not skipped in all the other tissues we tested (see Supplementary Figure S1 for all the results of the PCR experiments). This not only confirms the predictive power of motif combinations, but also the low resolution of current transcriptome data with respect to cell type and organism age. For the predictions that the skipped forms have not been confirmed by PCR experiments, we cannot exclude the possibility that the co-occurring motifs might work in a competitive manner. In the future, the addition of other features, such as exon/intron length and RNA secondary structure, should improve the prediction accuracy (9).

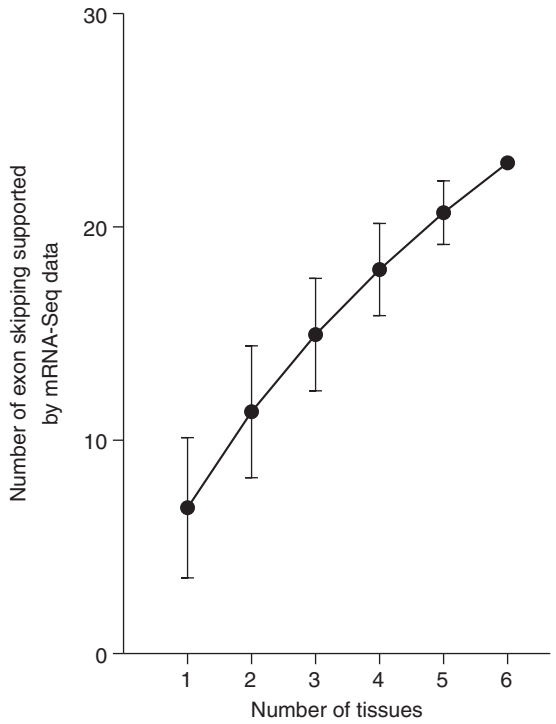

Figure 6. Number of exon skipping supported by RNA-Seq data increase with the number of tissue samples. All possible combinations of tissues are taken into account for each number of tissues. The 118 predictions of exon skipping made for the Ensembl gene set were used.

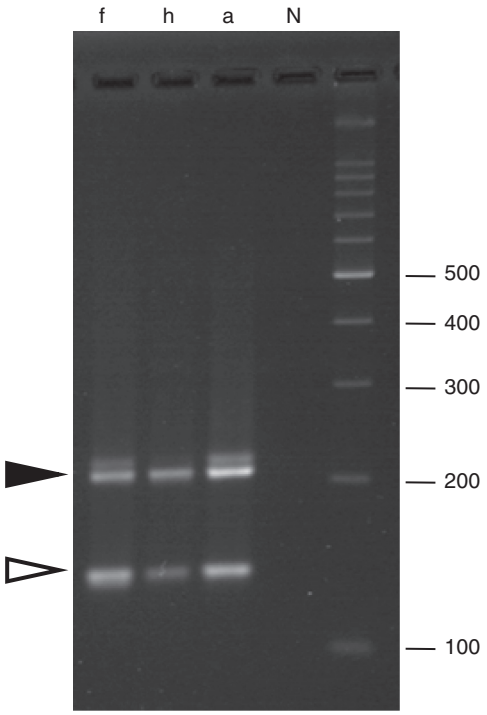

Figure 7. Exon skipping of the third exon of ENST00000256858 confirmed by the PCR experiment. The PCR product lengths for skipped and included forms are 136 bases (open triangle) and 208 bases (closed triangle), respectively (see Supplementary Table S1). F, fetal brain; h, hippocampus; a, amygdala; N, negative control.

In total, 39 exons out of 118 predictions are confirmed to be alternatively spliced (Supplementary Table S5). To see if these 39 cases have some additional features comparing to the rest of the 79 cases, first we calculated the distances of the motifs pairs. The average distances for the confirmed 39 cases and the rest of the 79 cases are 518.1 and 321.3 bases, respectively. Applying Mann–Whitney U-test (45) to these two distributions of the motif distances, we obtained $z' = 1.97$, which corresponds to $P > 0.01$. Then we calculated secondary structure forming potential around the motifs. For each motif, we extracted the sequence from 50 bases upstream to 50 bases downstream of the motif, and applied the RNAfold program (46) to calculated possible folding energy. The average folding energy values for the confirmed 39 cases and the rest of the 79 cases are −19.6 and −17.8 kcal/mol, respectively. The two energy distributions are compared by Mann–Whitney U-test, which gives $z' = 2.56$ ($P > 0.01$). From these analyses, we do not see any statistically significant differences between the experimentally confirmed set and the set yet to be confirmed at least for these two features.

CONCLUSIONS

It has been shown that some distinct splicing factors work together to regulate alternative splicing in some instances (47–49). We show that such co-occurrence of the *cis*-elements is not a special case for a certain pair of motifs but it might be a common principle as almost all motifs identified here form a network that might be the basis for context-specific intronic splicing regulation. The co-occurring elements might act in a cooperative or competing manner to provide complexity in the regulation of alternative splicing, for example, to achieve fine-grained tissue- or developmental stage-specific control. Recently an experimental system has been devised to analyze coordinated regulation of tissue-specific alternative splicing (50), which will allow a more detailed analysis of the regulation provided by the co-occurring elements identified in this study. This kind of combinatorial regulation is analogous to that of transcription, in which combinatorial usage of distinct transcription factors provide a large variety of possibilities for gene expression patterns (51), and this principle might, thus, widespread in genomic regulation. Co-occurrence of *cis*-regulatory elements also suggests that, although one element might be sufficient to explain the regulatory mechanism of alternative splicing under a given condition, other elements might be necessary to fully explain the complex regulation of alternative splicing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Tetsuhiro Moriya for technical assistance in RT–PCR experiments. We also thank the mammalian

genome sequencing consortia for making the genomic sequences freely available before scientific publication.

FUNDING

EU network of excellence EURASNET (to P.B.); Grant-in-Aid for Scientific Research on Priority Areas 'Comparative Genomics' and for Scientific Research (C) from the Ministry of Education, Culture, Sports, Science and Technology of Japan (to M.S.). Funding for open access charge: Grant-in-Aid for Scientific Research (C) from the Ministry of Education, Culture, Sports, Science and Technology of Japan

Conflict of interest statement. None declared.

REFERENCES

1. Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.
2. Brett,D., Hanke,J., Lehmann,G., Haase,S., Delbruck,S., Krueger,S., Reich,J. and Bork,P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, **474**, 83–86.
3. Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.
4. 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.
5. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
6. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
7. Matlin,A.J., Clark,F. and Smith,C.W. (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, **6**, 386–398.
8. Wang,Z. and Burge,C.B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, **14**, 802–813.
9. Barash,Y., Calarco,J.A., Gao,W., Pan,Q., Wang,X., Shai,O., Blencowe,B.J. and Frey,B.J. (2010) Deciphering the splicing code. *Nature*, **465**, 53–59.
10. 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.
11. Zhang,X.H. and Chasin,L.A. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.*, **18**, 1241–1250.
12. Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**, 739–748.
13. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
14. Xie,X., Mikkelsen,T.S., Gnirke,A., Lindblad-Toh,K., Kellis,M. and Lander,E.S. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl Acad. Sci. USA*, **104**, 7145–7150.
15. Modrek,B. and Lee,C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.*, **34**, 177–180.
16. Pan,Q., Bakowski,M.A., Morris,Q., Zhang,W., Frey,B.J., Hughes,T.R. and Blencowe,B.J. (2005) Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet.*, **21**, 73–77.
17. 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. USA*, **102**, 2850–2855.
18. Voelker,R.B. and Berglund,J.A. (2007) A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. *Genome Res.*, **17**, 1023–1033.
19. Yeo,G.W., Van Nostrand,E.L. and Liang,T.Y. (2007) Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet.*, **3**, e85.
20. Churbanov,A., Vorechovsky,I. and Hicks,C. (2009) Computational prediction of splicing regulatory elements shared by Tetrapoda organisms. *BMC Genomics*, **10**, 508.
21. Sorek,R. and Ast,G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, **13**, 1631–1637.
22. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
23. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res*, **12**, 656–664.
24. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
25. Kuhn,R.M., Karolchik,D., Zweig,A.S., Wang,T., Smith,K.E., Rosenbloom,K.R., Rhead,B., Raney,B.J., Pohl,A., Pheasant,M. *et al.* (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.
26. 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.
27. 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.
28. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
29. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.
30. Sudarsanam,P., Pilpel,Y. and Church,G.M. (2002) Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae. *Genome Res.*, **12**, 1723–1731.
31. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
32. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
33. Venables,J.P. (2004) Aberrant and alternative splicing in cancer. *Cancer Res.*, **64**, 7647–7654.
34. Nagase,T., Ishikawa,K., Nakajima,D., Ohira,M., Seki,N., Miyajima,N., Tanaka,A., Kotani,H., Nomura,N. and Ohara,O. (1997) Prediction of the coding sequences of unidentified human genes. VII. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro. *DNA Res.*, **4**, 141–150.
35. Shigemori,Y., Mikawa,T., Shibata,T. and Oishi,M. (2005) Multiplex PCR: use of heat-stable Thermus thermophilus RecA

protein to minimize non-specific PCR products. *Nucleic Acids Res.*, **33**, e126.
36. Clark,F. and Thanaraj,T.A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.*, **11**, 451–464.
37. 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.
38. 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.
39. Buckanovich,R.J. and Darnell,R.B. (1997) The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. *Mol. Cell. Biol.*, **17**, 3194–3201.
40. Chan,R.C. and Black,D.L. (1997) The polypyrimidine tract binding protein binds upstream of neural cell-specific c-src exon N1 to repress the splicing of the intron downstream. *Mol. Cell. Biol.*, **17**, 4667–4676.
41. Hastings,M.L., Wilson,C.M. and Munroe,S.H. (2001) A purine-rich intronic element enhances alternative splicing of thyroid hormone receptor mRNA. *RNA*, **7**, 859–874.
42. Sirand-Pugnet,P., Durosay,P., Brody,E. and Marie,J. (1995) An intronic (A/U)GGG repeat enhances the splicing of an alternative intron of the chicken beta-tropomyosin pre-mRNA. *Nucleic Acids Res.*, **23**, 3501–3507.
43. Zhang,C., Li,W.H., Krainer,A.R. and Zhang,M.Q. (2008) RNA landscape of evolution for optimal exon and intron discrimination. *Proc. Natl Acad. Sci. USA*, **105**, 5797–5802.
44. Watahiki,A., Waki,K., Hayatsu,N., Shiraki,T., Kondo,S., Nakamura,M., Sasaki,D., Arakawa,T., Kawai,J., Harbers,M. *et al.* (2004) Libraries enriched for alternatively spliced exons reveal splicing patterns in melanocytes and melanomas. *Nat. Meth.*, **1**, 233–239.
45. Sokal,R.R. and Rohlf,F.J. (1995) *Biometry*. 3rd edn. W. H. Freeman & Company, New York.
46. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
47. Friedman,B.A., Stadler,M.B., Shomron,N., Ding,Y. and Burge,C.B. (2008) Ab initio identification of functionally interacting pairs of cis-regulatory elements. *Genome Res.*, **18**, 1643–1651.
48. Han,K., Yeo,G., An,P., Burge,C.B. and Grabowski,P.J. (2005) A combinatorial code for splicing silencing: UAGG and GGGG motifs. *PLoS Biol.*, **3**, e158.
49. Smith,C.W. and Valcarcel,J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.*, **25**, 381–388.
50. Kuroyanagi,H., Ohno,G., Mitani,S. and Hagiwara,M. (2007) The Fox-1 family and SUP-12 coordinately regulate tissue-specific alternative splicing in vivo. *Mol. Cell. Biol.*, **27**, 8612–8621.
51. Tjian,R. and Maniatis,T. (1994) Transcriptional activation: a complex puzzle with few easy pieces. *Cell*, **77**, 5–8.