Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2004-059
AIM-2004-022

September 30, 2004

Predictive identification of alternative events conserved in human and mouse

Gene Yeo, Eric Van Nostrand, Dirk Holste, Tomaso Poggio, and Christopher Burge

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Predictive identification of alternative splicing events conserved in human and mouse

Gene Yeo[1,2], Eric Van Nostrand[1], Dirk Holste[1], Tomaso Poggio[2], and Christopher B. Burge[1,3]

[1]Department of Biology and

[2]Department of Brain and Cognitive Sciences, and

Center for Biological and Computational Learning

Massachusetts Institute of Technology

Cambridge, MA 02319

[3] To whom correspondence should be addressed. E-mail: cburge@mit.edu

Abstract

Alternative pre-messenger RNA splicing affects a majority of human genes and plays important roles in development and disease. Alternative splicing (AS) events conserved since the divergence of human and mouse are likely of primary biological importance, but relatively few such events are known. Here we describe sequence features that distinguish exons subject to evolutionarily conserved AS, which we call 'alternative-conserved exons' (ACEs) from other orthologous human/mouse exons, and integrate these features into an exon classification algorithm, ACEScan. Genome-wide analysis of annotated orthologous human-mouse exon pairs identified ~2,000 predicted ACEs. Alternative splicing was verified in both human and mouse tissues using an RT-PCR-sequencing protocol for 21 of 30 (70%) predicted ACEs tested, supporting the validity of a majority of ACEScan predictions. By contrast, AS was observed in mouse tissues for only 2 of 15 (13%) tested exons which had EST or cDNA evidence of AS in human but were not predicted ACEs, and was never observed for eleven negative control exons in human or mouse tissues. Predicted ACEs were much more likely to preserve reading frame, and less likely to disrupt protein domains than other AS events, and were enriched in genes expressed in the brain and in genes involved in transcriptional regulation, RNA processing and development. Our results also imply that the vast majority of AS events represented in the human EST databases are not conserved in mouse, and therefore may represent aberrant, disease- or allele-specific, or highly lineage-restricted splicing events.

Running Title

Conserved Alternative Splicing Events

Introduction

The processing of human primary transcripts to produce the messenger RNAs (mRNAs) that will direct protein synthesis can be constitutive, always yielding the same mRNA, but is often variable, producing multiple alternatively spliced (AS) mRNA products, most commonly by alternative inclusion or exclusion ('skipping') of individual exons (Lopez, 1998; Maniatis and Tasic, 2002; Black and Grabowski, 2003). Alternative pre-mRNA splicing plays a major role in expanding protein diversity and regulating gene expression in higher eukaryotes (Black, 2000; Black, 2003). Regulated AS is crucial in development (Lopez, 1998) and in the nervous system, where specific AS isoforms regulate ion channel function, receptor specificity and neurotransmission (Grabowski and Black, 2001; Black and Grabowski, 2003), and mis-regulation of AS is associated with a number of human diseases (Dredge et al., 2001; Caceres and Kornblihtt, 2002; Faustino and Cooper, 2003).

Expressed sequence tags (ESTs) provide a rich source of information about splicing events occurring in the human and mouse transcriptomes, as they are far greater in abundance and encompass a larger spectrum of cell-type and tissue diversity than complementary DNAs (cDNAs) alone. Considering the set of human ESTs and cDNAs which can be reliably aligned to a human gene locus overlapping a particular exon, this set can be subdivided into those transcripts which include and those which exclude or 'skip' the exon in question. Here, 'skipping' of an exon refers to the situation where a transcript aligns consecutively to an upstream exon and a downstream exon of the gene in question, omitting the given exon. This consideration can be applied to all of the exons in a human gene, and an analogous subdivision can be made of the mouse transcripts that align to exons of the orthologous mouse gene. Each orthologous human/mouse exon pair can then be assigned to one of four categories, $S_{H,m}$, $S_{h,M}$,

$S_{H,M}$, or $S_{h,m}$, depending on whether the exon has observed to be skipped only in human transcripts ($S_{H,m}$), only in mouse ($S_{h,M}$), in both human and mouse ($S_{H,M}$), or whether exon skipping is not observed in either species ($S_{h,m}$).

Using publicly available EST databases totaling over 5 million human and over 3 million mouse ESTs, and databases of ~94,000 and ~91,600 human and mouse cDNAs respectively, thousands of alternative exons can be inferred in each species. However, the overlap between these sets is relatively small, i.e., for only about 240 (~ 1 in 18) of the ~ 4,500 conserved human-mouse exons observed to be skipped in human was transcript evidence found supporting alternative usage (skipping) of the orthologous mouse exon, as discussed below (cf. (Nurtdinov et al., 2003; Sorek and Ast, 2003; Thanaraj et al., 2003)). This observation raises the question of how many of the AS events observable in the human transcriptome are evolutionarily conserved, and therefore presumably contribute to organismal fitness, and how many are aberrant, disease- or allele-specific, or highly lineage-restricted events. Although study of the latter types of events may lead to important insights and applications, a significant fraction of these events may constitute biochemical 'noise' or transient evolutionary fluctuations. On the other hand, conservation of a specific pattern of AS over the ~90 million years since divergence of the mouse and human lineages provides strong evidence of biological function. Therefore, defining the set of AS events conserved between human and mouse is of primary interest in efforts to understand the biological importance of splicing regulation.

Alternative inclusion/exclusion of exons is known to be influenced by a number of factors, such as splice site strength (Smith and Nadal-Ginard, 1989), intron length (Bell et al., 1998), exon length (Black, 1991; Dominski and Kole, 1991) and sometimes pre-mRNA secondary structure (Libri et al., 1991; Estes et al., 1992; Blanchette and Chabot, 1997; Jiang et

al., 2000). Certain *cis*-regulatory elements, including exonic and intronic splicing enhancers (ESEs and ISEs, respectively), as well as exonic and intronic splicing silencers (ESSs and ISSs, respectively) can also control exon skipping by recruiting *trans*-acting splicing factors (Ladd and Cooper, 2002; Black, 2003). Computational studies have identified other sequence features that differ between skipped (also known as 'cassette') exons and constitutive exons in human and mouse genes, including increased conservation in the introns flanking exons skipped in both human and mouse (Sorek and Ast, 2003; Thanaraj et al., 2003; Kaufmann et al., 2004; Sorek et al., 2004a; Sugnet, 2004). These observations motivated us to systematically identify, characterize and integrate sequence features into a classifier that could be used to identify exons subject to evolutionarily conserved exon skipping, here termed 'alternative-conserved exons' (ACEs).

Results and Discussion

Outline of strategy for identification of alternative-conserved exons

Our scheme for identifying ACEs consisted of three phases: 1) 'learning'; 2) prediction; and 3) validation (Fig. 1). In the learning phase, a set of sequence features was identified, including exon and intron length, splice site strength, sequence conservation, and region-specific oligonucleotide composition, which differed between 'training sets' of 241 exons of the class $S_{H,M}$ and ~5,000 exons of the class $S_{h,m}$ defined above (Fig. 2). Additionally, for training purposes, exons of the $S_{h,m}$ class were chosen from genes containing at least one other exon with evidence for AS, as genes with no AS may be under different selective pressures than AS genes (Iida and Akashi, 2000). Next, these features were incorporated into a discriminant classifier, ACEScan, which was used in the prediction phase to predict which of ~96,000 annotated orthologous human/mouse exon pairs not previously known to exhibit conserved AS are in fact ACEs. Finally, in the validation phase, a subset of candidate exons with positive ACEScan scores (designated 'ACEScan-positive' or ACEScan[+] exons for short) was chosen for experimental testing, together with two sets of negative control exons with negative ACEScan scores ('ACEScan-negative' or ACEScan[-] exons): one set with previous transcript evidence for exon skipping in human (S_H category) and one set lacking such evidence (S_h category).

The following features were first incorporated into ACEScan: (i) exon length; (ii) upstream and (iii) downstream intron lengths; (iv) the 5' splice site (5'ss) and (v) 3' splice site (3'ss) scores; (vi) nucleotide percent identity between the orthologous human and mouse exons; and human-mouse intronic sequence conservation (vii) within the last 150 bases upstream and (viii) within the first 150 bases downstream of the exon. In general, exon pairs skipped in both

human and mouse (set $S_{H,M}$) were observed to be shorter than unskipped exon pairs ($S_{h,m}$), were flanked by longer upstream and downstream introns, and possessed significantly weaker splice sites (Fig. 2). Strikingly, exon-pairs in $S_{H,M}$ have significantly higher sequence identity and higher flanking intronic conservation as compared to exon pairs in $S_{h,m}$ (Fig. 2). These observations are similar to and consistent with previous studies (Sorek and Ast, 2003; Kaufmann et al., 2004; Sorek et al., 2004b; Sugnet, 2004). High levels of sequence conservation in the exons and flanking introns is suggestive of conservation of regulatory motifs or RNA structure.

Oligonucleotides useful in discrimination of alternative-conserved exons

Oligonucleotide features designed to score potential *cis*-regulatory elements consisted of the highest-ranking (most biased) over- and under-represented oligonucleotides of length *k* (*k*-mers) in different exon and intron regions. The regions considered were the first and last 100 bases of exons and the proximal 150 bases in the upstream and downstream introns flanking the exon, because of the high levels of sequence conservation in these regions and their proximity to the regulated splice junctions. Counts of conserved oligonucleotides in human-mouse nucleotide alignments of the 150 bases of upstream and downstream intronic sequence and in the entire exon were also scored for enrichment in the set $S_{H,M}$ versus $S_{h,m}$. Inclusion of oligonucleotide counts from unaligned as well as aligned sequences permits scoring of *cis*-regulatory elements that lack strict spatial constraints for function, as well as those that do. Oligonucleotides were ranked by their enrichment as measured by χ^2 value. Several of the motifs identified are similar to motifs known to function in regulation of splicing. For instance, the motifs UGCAU and GCAUG were found to be over-represented in both the upstream and downstream introns flanking exons subjected to conserved skipping (Fig. 2B). Similar sequences, e.g., the hexamer

UGCAUG, are known to be involved in regulation of splicing of the *c-src*, fibronectin, nonmuscle myosin heavy chain and calcitonin genes (Black, 1992; Huh and Hynes, 1994; Modafferi and Black, 1997; Brudno et al., 2001). The UCUCU pentamer, which is similar to sequences involved in splicing repression in the neural-specific N1 exon of the *c-src* transcript (Chan and Black, 1997), was also identified as over-represented in the introns upstream of $S_{H,M}$ exons and in the exons themselves. A number of other U-rich sequences were also over-represented in upstream introns, consistent with previous observations (Shibata et al., 1996; Forch et al., 2000). The sequence UAGGG, which forms a portion of the consensus hnRNP A1 binding site, and can act in negative regulation of splicing (Kashima and Manley, 2003), was also over-represented in $S_{H,M}$ exons relative to unskipped exons. Motifs related to GUAGU, also over-represented in $S_{H,M}$, have been validated as exonic splicing silencers (ESSs) in cultured human cells (Z. Wang and C. B. B., unpublished data). On the other hand, two pentamers which were under-represented in $S_{H,M}$ relative to $S_{h,m}$, CUGGA and AGAAG, resemble consensus ESEs (UGGA and GAGAAG, respectively) identified in previous analyses (Liu et al., 1998; Fairbrother et al., 2002). In fact, more detailed analyses suggest that a significantly higher fraction of oligonucleotides enriched in $S_{h,m}$ matched computationally predicted and experimentally validated ESEs (Fairbrother et al., 2002) as compared to oligonucleotides enriched in $S_{H,M}$ (Table S2). Thus, the oligonucleotide analysis identified a number of known and putative splicing regulatory sequences, and suggests that alternative-conserved exons have a lower density of ESEs and a higher density of ESS sequences relative to constitutive exons. Both of these features would tend to facilitate exclusion by the splicing machinery. Furthermore, the introns flanking ACEs appear to have a higher density of positive- and negative-acting intronic regulatory elements, which are likely to contribute to aspects of regulation. Previously,

candidate ESEs were identified in part based on enrichment in constitutive exons with weak splice sites (Fairbrother et al., 2002). The apparently reduced frequency of ESEs in ACEs, which also have somewhat weak splice sites (Fig. 2), might possibly reflect the differing selective pressures, with constitutive exons presumably being under selection for efficient exon inclusion, while alternative exons are presumably selected for inefficient inclusion under at least some conditions. The apparently increased frequency of ESSs in ACEs relative to constitutive exons is also consistent with this rationale.

Integration and selection of features for accurate exon classification

The task of integrating the general features and oligonucleotide features described above into an algorithm that distinguishes exon pairs in $S_{H,M}$ (positively labeled) from those in $S_{h,m}$ (negatively labeled) was posed as a supervised binary classification problem. Because the potentially large number of features relative to the limited number of positive examples makes this an "ill-posed" problem, a machine-learning algorithm that avoids this "curse of dimensionality" was adapted for our purposes. The regularized least-squares classifier (RLSC), which finds the optimal separating hyperplane in a high-dimensional space that distinguishes two classes of samples, and is similar to the support vector machine algorithm, was utilized (Zhang, 1997; Zhang, 2000; Rifkin et al., 2003; Zhang et al., 2003). As it was not known *a priori* which of the 8,245 features were most important in the classification scheme, models utilizing different combinations of the eight general features and the region-specific oligonucleotide features were compared, and a feature selection protocol was used to reduce the number of parameters and to retain only the most relevant oligonucleotide features.

In order to determine the optimal features and parameters for the classifier, the training data was utilized to generate several models, by varying the choice of general features, the exon or intron regions from which oligonucleotide features were generated, and the number of most discriminative oligonucleotide features included. Cross-validation was used: for each model, 80% of the exon-pairs from $S_{H,M}$ and 80% of the pairs from $S_{h,m}$ were used to train the classifier, which then assigned outputs (predicted classifications) to the remaining 20% of unseen exon pairs. The performance of different models was averaged over 50 iterations of sampling training and test subsets. The model with the best performance utilized 240 oligonucleotides of lengths 4 and 5, and all of the general sequence features (Fig. S1). This model assigned correct labels to ~90 exon pairs for every 100 exon pairs drawn randomly from $S_{H,M}$ and $S_{h,m}$. For an individual exon, the ACEScan score was defined as the mean of the RLSC outputs over 50 random samplings of the training data. The distribution of ACEScan scores for the exon pairs in $S_{H,M}$ ranged from approximately –0.8 to 2.0 (arbitrary units), compared to a range of approximately –1.8 to 0 for most of the exons in $S_{h,m}$ (Fig. 1). At a cutoff score of $\tau = -0.5$, ~85% of exons in $S_{H,M}$ scored above the threshold, and 85% of exons in $S_{h,m}$ scored below τ, demonstrating that ACEScan achieves good (but not perfect) separation of these two classes of exons. Raising the cutoff score to zero, only ~2% of $S_{h,m}$ exons had positive ACEScan scores, compared to ~61% of the exons in $S_{H,M}$, suggesting that ACEScan[+] exon pairs are highly enriched for ACEs.

Experimental validation of conserved AS for 21 of 30 ACEScan[+] exon pairs

A combination of experimental tests and bioinformatic approaches was used to explore the features of ACEScan[+] and ACEScan[-] exon pairs. First, the splicing patterns of a set of 30 arbitrarily chosen ACEScan[+] exons were tested in a battery of human and mouse tissues by

reverse transcriptase PCR (RT-PCR) with primers targeted to flanking exons. ACEScan[+] exons were selected from four intervals: I1 (ACEScan score from 0.0 to 0.5); I2 (0.5 to 1.0); I3 (1.0 to 1.5); and I4 (greater than 1.5), spanning the range of scores of most $S_{H,M}$ exons. Panels of twelve normal human tissues (fetal brain, fetal liver, cerebellum, heart, whole brain, prostate, liver, lung, kidney, bone marrow, skeletal muscle and testis) and eight normal mouse tissues (embryonic mix, whole brain, kidney, skeletal muscle, liver, lung, heart and testis) were assayed. In order to avoid undesired detection of aberrant or disease-specific splicing, tumor or other diseased tissues were not utilized. The products of these 600 RT-PCR reactions (30 exons x 20 tissues) were analyzed by gel electrophoresis, and the identities of PCR products with expected sizes for mRNAs including or excluding the test exon were confirmed by sequencing (Fig. 3A). In all, 4 out of 9, 7 out of 8, 6 out of 8, and 4 out of 5 candidate ACEs in intervals I1, I2, I3 and I4, respectively, were observed to undergo skipping in both human and mouse, while for another 2 exons (both from interval I1) exon skipping was observed only in human tissues (Fig. 3; Table S4). Thus, of 30 predicted ACEs interrogated by RT-PCR, 21 were observed to be skipped in both human and mouse tissue(s), and high rates of validation of AS were seen in all four score intervals. These data support the presence of conserved AS in a majority of ACEScan[+] exons. Although the 30 ACEScan[+] candidates had no previous transcript evidence for skipping, searches of the literature and low-stringency searches of the latest cDNA and EST databases identified evidence for less than half of the AS events observed by RT-PCR, and often single EST hits in one species only (Table S4).

In the examples studied, exon skipping was observed in many different combinations of human and mouse tissues (Table S4), suggesting that many of the features utilized by ACEScan are characteristic of skipped exons generally, regardless of tissue-specificity. Variations in tissue

specificity of AS were observed between human and mouse for some tested exons. However, a general tendency to conserve exon skipping in corresponding tissues was apparent, i.e., 9 of 10 predicted ACEs observed to be skipped in human brain (or cerebellum) were also skipped in mouse brain tissue (Table S4).

Low detection of conserved AS for ACEScan[-] exon pairs

As a negative control, eleven ACEScan[-] exon pairs from the set S_h were chosen from the five score intervals C1 (-0.5 to 0), C2 (-1.0 to –0.5), C3 (-1.5 to –1.0), C4 (-2.0 to –1.5) and C5 (less than –2.0) with at least one pair per interval. Using the same RT-PCR-sequencing assay and the same sets of human and mouse tissues, exon skipping was not observed for any of the eleven negative control exons in any of the 12 human or 8 mouse tissues studied (Table S4). Thus, considering both the human and mouse exons tested, exon skipping was detected for 44 out of 60 ACEScan[+] exons (including 21 ortholog pairs), compared to 0 out of 22 ACEScan[-] exons, a highly significant difference ($P < 0.0001$, Fisher exact test). Of course, for either group of exons failure to detect exon skipping by our RT-PCR assay is not proof that exon skipping does not occur, and some exons not skipped in the tissues studied might be skipped in other untested tissues. However, low-stringency searches of the latest human and mouse EST databases failed to detect any evidence of skipping of the 11 ACEScan[-] exons tested.

As a second type of negative control, an arbitrary set of 15 ACEScan[-] exon pairs were chosen from intervals C2 to C4, with the added requirement that transcript evidence of exon skipping was present for the human member of each exon pair. Using the same RT-PCR-sequencing assay in the same set of 8 mouse tissues as above, exon skipping was detected for only 2 out of the 15 mouse exons tested, suggesting that a substantial majority of these exon

pairs do not undergo evolutionarily conserved AS (i.e. are not ACEs). To explore the potential biological roles of the 13 remaining exons which undergo possible human-specific AS, we examined the tissue sources of the transcripts that showed exon skipping: in 9 of the 13 cases, these transcripts derived exclusively from cancer cell-lines or diseased tissues, suggesting that many of these exons may be skipped primarily in disease states rather than in normal human tissues. The difference in the rate of RT-PCR validation of exon skipping in mouse tissues for the ACEScan[+] exons tested (21 out of 30 = 70%), relative to the ACEScan[-] exons tested (2 out of 26 = ~8%), was also highly significant ($P < 0.002$, Fisher exact test), demonstrating the power of ACEScan to discriminate evolutionarily conserved AS exons from those which are either constitutively spliced or skipped in a species- (or disease-) specific manner. It is also notable that this large difference was observed even though only a few of the ACEScan[+] exons tested had EST or cDNA evidence of skipping, while most (15) of the ACEScan[-] exons had such evidence in human.

Literature-derived AS events contain high frequency of ACEScan[+] exons

The principle that important regulatory elements are usually evolutionarily conserved is well established, and forms the basis of a number of successful comparative genomics approaches for identifying such regulatory regions (e.g., (Loots et al., 2000). To explore the extent to which this principle applies to AS events, and by extension to assess the probable importance of conserved AS events identified by ACEScan, we extracted known exon skipping events from the Manually Annotated Alternatively Spliced Events (MAASE) database (Zheng et al., 2004). This database represents AS events that are described in published work, and therefore are presumably of some importance or interest, where the authors have manually

entered information about the specific nature of the splicing event in the context of aligned transcript and genomic sequences, including literature references. From this database, a total of 29 exon skipping events were identified from mouse which were not in our training set, but for which both the human and mouse orthologous exons were available, making it possible to analyze them using ACEScan. Notably, almost all of the extracted exons had ACEScan scores greater than –0.5 (28 out of 29) and 62% (18 out of 29) were ACEScan[+]. Thus, though small in scale, this analysis of published AS events suggests that a majority of 'interesting' exon skipping events are ACEScan[+] and therefore that most such events are conserved between human and mouse.

Only ~11% of EST/cDNA-derived AS events are likely to be evolutionarily conserved

Of the ~4,300 exon-pairs in with transcript evidence of skipping in human but not mouse (class $S_{H,m}$), only ~7% had positive ACEScan scores (Fig. 1). Together with the observation that ~61% of $S_{H,M}$ exons were ACEScan[+], this low fraction suggests that for only ~11% (= 0.07 / 0.61) of the $S_{H,m}$ exons is AS likely to be conserved in mouse. Thus, a surprising implication of these data is that the vast majority of the AS events inferable from human EST/cDNA-genomic alignments are not evolutionarily conserved in mouse. Instead, most of these events may represent aberrant, disease-specific, or allele-specific splicing, or events whose phylogenetic distribution is highly restricted. At present it is difficult to distinguish among these possibilities, but the transcript analysis described above for ACEScan[-] exons suggests that a significant number may represent disease-associated events.

Functional differences between ACEScan[+] and ACEScan[-] exons

To assess potential functional differences between ACEScan[+] exons, and ACEScan[-] exons that either have or do not have EST or cDNA evidence of exon skipping in human, we analyzed the density of single nucleotide polymorphisms (SNPs), and the frequency of reading frame preservation and protein domain disruption for each of these three classes of exon. Selective pressure on nucleotide sequence was assayed by mapping stringently filtered reference SNPs onto exons that had been scored by ACEScan (Fig. 3B). This analysis found a ~50% higher density of SNPs in in ACEScan[-] S_H exons than in ACEScan[+] exons (this difference is significant at $P < 10^{-5}$ by χ^2 test), suggesting that ACEs are under much more stringent selective pressure to conserve nucleotide sequence than are other exons. Notably, the density of SNPs in ACEScan[-] S_H exons (that have transcript evidence of exon skipping in human) was only ~20% lower than in ACEScan[-] S_h exons, suggesting that ACEScan[-] S_H exons experience selective pressures that are more similar to constitutive exons than they are to ACEs.

Further evidence for the functional roles of many ACEScan[+] S_H exons came from the observation that a far higher fraction of these exons had lengths which were multiples of three (68%, comparable to that seen in the training set of $S_{H,M}$ exons) than was seen for ACEScan[-] S_H exons, for which only ~43% had lengths divisible by three, near background levels for constitutive internal exons (Fig. 3C). This difference is highly significant ($P < 10^{-15}$ by χ^2 test) and implies the existence of strong selective pressure on the alternative protein products derived from alternative splicing of ACEs. Since divisibility of the exon length by 3 was not used in the predictions (only the general size of the exon, with shorter lengths favored over longer lengths), this difference in the frequency of reading frame preservation provides additional evidence that many ACEScan[+] exons are under strong selective pressure related to their role in AS.

The frequency of disruption or removal of a protein domain by AS has been studied by several groups (Kriventseva et al., 2003; Xing et al., 2003; Cline et al., 2004). We found that only ~37% of ACEScan[+] exons overlapped open reading-frame regions encoding Interpro-annotated protein domains by 30 bases (10 codons) or more, a significantly lower fraction than for ACEScan[-] exons studied of either the S_H or S_h classes (Fig. 3D), both of which had similar frequencies of domain disruption, around 50%. (Maximum exon length cutoffs were used to ensure comparable median exon sizes in these three groups, as described in the Figure Legend.) Reducing the minimum overlap to 15 bases did not change the result (data not shown). This finding is generally consistent with the results of Kriventseva and coworkers who observed that protein isoforms arising from AS are more likely to preserve protein domain structure than expected by chance. Taken together, the data shown in Fig. 3 consistently demonstrate that ACEScan[+] exons are under strong selection to conserve function, both at the nucleotide level (Fig. 3B), and at the level of the encoded alternative protein isoform (Figs. 3C, D). In contrast, ACEScan[-] exons show less evidence of selective constraint at the nucleotide level (Fig. 3B) and there is little if any evidence of additional constraints on the protein products derived from exon skipping of ACEScan[-] exons, even when there is transcript evidence that such skipping occurs (Figs. 3C, D).

Sample applications of ACEScan at the gene level

Application of ACEScan to four well-studied genes illustrates some of the strengths and limitations of our approach (Fig. 4). Of the identifiable orthologous human/mouse exon pairs in these genes, eight are known to undergo exon skipping (marked by asterisks), and all of these received positive ACEScan scores, implying that their skipping is likely to be conserved in

mouse. One additional exon (exon 2 of the *GLUR-B* gene) also scored positively; this exon may represent a false positive prediction of the method or an unrecognized AS event. Of these 8 known skipped exons, six preserve reading frame of downstream exons, while skipping of the other two causes a shift in reading frame, and AS events in both of these categories are known to have important functional consequences. For example, electrophysiological studies have shown that skipping of exon 26 of *CACNA1G*, which encodes α_{1g}, a human brain T Ca^{2+} channel α_1 subunit, affects the kinetics of deactivation and recovery from inactivation of the channel (Chemin et al., 2001). This gene has three other cassette (skipped) exons, numbered 14, 34 and 35 (Mittman et al., 1999), the latter two of which were not paired to annotated mouse exons by our automated exon orthology script, and therefore did not receive ACEScan scores. These latter examples remind us that incompleteness of annotation or orthology assignment places certain limits on the sensitivity of comparative methods like ACEScan. Skipping of exons 7 and 8 of the β-amyloid precursor protein (*APP*) gene, implicated in Alzheimer's disease, was detected successfully in a recent large-scale microarray analysis of splicing in human tissues (Johnson et al., 2003). These exons, as well as exon 15 of the *APP* gene received positive ACEScan scores (Fig. 4B); all three of these exons are known to undergo exon skipping (Ponte et al., 1988; Konig et al., 1992). The polyprimidine tract binding protein (*PTB*) gene contains a predicted ACE which is 34 bases in length (Fig. 4C). Skipping of this exon yields a premature stop codon in the downstream exon, generating a substrate for nonsense-mediated mRNA decay (NMD), in an autoregulatory negative feedback loop (Wollerton et al., 2004). This *PTB* exon provides an example of function for one of the ~32% of predicted ACEs which disrupt reading frame. The *GLUR-B* gene, one of the four GluR subunits that assemble to form the AMPA glutamate receptor, contains two well-known skipped exons ('flip' and 'flop', exons 14 and 15), both of

which receive positive ACEScan scores, as well as an exon (number 13, marked with an 'E') that undergoes RNA editing (Sommer et al., 1990; Cha et al., 1994). This edited exon and the downstream intron form an RNA hairpin and are highly conserved in sequence. Despite this high level of exonic and intronic sequence conservation (Cha et al., 1994), this exon receives a negative ACEScan score (Fig. 4D), providing an example of the specificity of our method for AS exons.

Recently, Bejerano et al. reported 111 exonic "ultraconserved" regions (UCRs) with 100% sequence identity between the human and mouse genomes (Bejerano et al., 2004), most of unknown function. Comparing these to our set of predicted ACEs, 33 of the 37 UCRs (~89%) that mapped to internal exons that could be scored by ACEScan received positive ACEScan scores, suggesting that a number of these elements corresponded to ACEs, while the rest may represent other forms of regulatory elements (e.g., RNA editing).

A web server has been set up at (http://genes.mit.edu/acescan) to provide access to ACEScan plots for Ensembl-annotated orthologous human/mouse gene pairs.

Functional characteristics of ACEScan[+] genes.

In order to determine whether genes that contain ACEScan[+] exons, which we refer to as ACEScan[+] genes, are biased towards particular biological activities, we compared these genes to the set of genes not found to contain any ACEScan[+] exons, (ACEScan[-] genes) using Gene Ontology (GO) classifications (http://www.geneontology.org). The results showed that ACE-containing genes are enriched for transcription factor and aminopeptidase activity, and for the actin, RNA and nucleic acid binding GO molecular function categories (Fig. 5A). In terms of GO biological process categories, ACEScan[+] genes were more likely to be involved in

transcriptional regulation, neurogenesis, and development, and less likely to be involved in transport than ACEScan[-] genes. Only slight biases in GO category representation were present in the training set of $S_{H,M}$ genes (Fig. S3). Closer examination of the ACEScan[+] genes that encode RNA binding factors identified ACEScan[+] exons in genes encoding many of the heterogeneous nuclear ribonucleoproteins (hnRNPs), including hnRNPs A1, D, H, I/PTB, K, L, R, M and FUS (Table S3). Some of these correspond to known exon skipping events in these genes, and our RT-PCR-sequencing analysis confirmed skipping for one of the remaining cases (Table S4). Proteins of the hnRNP family are involved with a host of important mRNA-related functions, such as nuclear export, subcellular localization, translation, and stability (Dreyfuss et al., 2002). In addition, some are known to regulate the splicing of their own or other transcripts (Blanchette and Chabot, 1997; Wollerton et al., 2004). It is widely believed that AS is of particular importance in the nervous system (Grabowski and Black, 2001; Black and Grabowski, 2003). To explore the expression patterns of genes containing predicted ACEs, we used microarray data from the Gene Atlas survey of 47 diverse human tissues and cell lines (Su et al., 2002). Overwhelmingly, ACEScan[+] genes were more likely to be differentially expressed in a spectrum of nervous system tissues, including spinal cord, fetal and adult whole brain, and in several brain regions, compared to ACEScan[-] genes (Fig. 5B). Only two cell lines (both ovarian) of the 47 tissues/cell lines studied exhibited similar biases. These results imply an unusually high frequency of conserved AS events in the brain.

In total, 2,041 ACEScan[+] exons were identified in 1,550 genes. Our automated scripts indicate that ~1,750 ACEScan[+] exons do not have previous EST or cDNA evidence for exon skipping, indicating a ~7.3 fold increase in candidate conserved alternative splicing events. Even using low-stringency similarity (BLASTN) searches against the recent update of the dbEST

database (July 22, 2004), we found new EST evidence for exon skipping in both human and mouse in only 51 cases (~2.9%) of the ~1,750 candidates. Initial comparisons to the partially annotated rat genome showed a high correlation between human-mouse and human-rat ACEScan scores, as expected (data not shown).

While this work was in progress, Sorek and coworkers demonstrated that specific features (exon length modulo three, intronic conservation in the intronic regions flanking the exon) can be used in a binary rule-based method for the identification of alternative exons in human (Sorek et al., 2004b). Our computational approach differs in a number of ways. First, ACEScan associates a real-valued score to orthologous human-mouse exon pairs, rather than merely associating a binary label to an exon-pair, which grants much greater flexibility in adjusting the algorithm's sensitivity/specificity. Second, ACEScan does not use the length of the exon modulo three in its predictions. This allows us to assess the degree of selection of ACEs to preserve protein reading frame (Fig. 3C) rather than assuming that reading frame must always be preserved, and it enables ACEScan to identify the subset of ACEs which create mRNAs that encode truncated proteins or which are subject to NMD, an emerging class of regulated AS events (Wollerton et al., 2004). Supporting the validity of this subset of predictions, approximately half of the ACEs validated by our RT-PCR-sequencing protocol had lengths that were not divisible by three (Figure 3A). Third, a much larger set of discriminatory features was utilized in ACEScan, including oligonucleotide features, many of which are likely to represent splicing regulatory elements, and inclusion of these features enhanced the performance of our algorithm. In addition to our higher rate of success of experimental validation of conserved exon-skipping events in both normal human and mouse tissues, our results from testing negative

control exons provides better estimates for the reliability of ACEScan. Lastly, we are the first to identify functional and expression biases in the set of genes containing high-confidence ACEs.

By combining machine-learning techniques with comparative genomic data, we demonstrated a method for efficient large-scale detection of conserved alternative splicing events in mammals. Our method is complementary to recent large-scale microarray-based efforts (Johnson et al., 2003), and identifies close to two thousand previously unrecognized conserved alternative splicing events. These events are likely to play important roles in tuning the functions of a variety of genes, including many that are involved in development and in the nervous system.

Experimental Procedures

Sequence datasets

Chromosome assemblies of the human genome (hg13) and the mouse genome (mm3) were obtained from the UCSC Genome Browser, http://genome.ucsc.edu. Transcript data used included ~94,000 human cDNA and ~91,600 mouse cDNA sequences obtained from GenBank (release 134.0, flatfiles in categories gbpri, gbrod and gbhtc), and ~5×10^6 human expressed sequence tags (ESTs) and ~3.5×10^6 mouse ESTs from dbEST (repository 032703). The GENOA genome annotation script (http://genes.mit.edu/genoa/) was used for spliced alignment of cDNA sequences and ESTs to the human and mouse genomes. GENOA detected matches of significant blocks of identity between a RepeatMasked (A. Smit, P. Green, http://www.repeatmasker.org) cDNA sequence and genomic DNA using BLASTN (Altschul et al., 1990). Matched pairs are then aligned, using the spliced alignment algorithm, MRNAVSGEN (http://genes.mit.edu/genoa/). Subsequently, ESTs were aligned to cDNA-verified genomic regions using SIM4 (Florea et al., 1998). For inclusion in the final GENOA annotation, all ESTs were required to overlap one or more cDNAs, and both the first and the last segments of the spliced alignment were required to exceed 30 nucleotides in length with 90% sequence identity. In addition, the entire EST sequence alignment was required to extend over 90% of the sequence length and have greater than 90% sequence identity.

Overall, GENOA aligned ~86,000 human cDNAs and ~890,000 ESTs, and ~27,000 mouse cDNAs and ~483,000 mouse ESTs. Genes with multiple cDNA alignments were resolved into separate transcript regions containing single genes and candidate regions with alternative exon-intron structures. 5'-terminal and 3'-terminal exons were separated from internal exons and

excluded from further analyses, as they possess different splicing characteristics, as well as sequence composition from internal exons. Exons were categorized as constitutive exons, alternative 3' splice site (3'ss) exons, alternative 5'ss exons, skipped exons, multiply alternatively spliced exons (e.g., exons observed to undergo both exon skipping and alternative 5'ss usage), and exons containing retained introns. Genes with at least one identified alternative splicing (AS) event were categorized as AS genes; all other genes were considered constitutively spliced (CS) genes. An exon was defined as a skipped exon (SE) if it was included in one or more transcripts and excluded at least one other transcript. Specifically, a transcript aligned such that the 3' end of the corresponding upstream exon and the 5' end of the corresponding downstream exon were juxtaposed, was considered as evidence of exon skipping. Human and mouse SEs were identified independently, using transcript data specific to each organism. Orthologous human and mouse genes were obtained from Ensembl version 16 (Hubbard et al., 2002), and exons in each pair of orthologous human and mouse genes were matched by BLASTN.

Exon-intron sequence regions and feature extraction

The following sequence features were extracted for each conserved human-mouse exon pair: exon length, upstream intron length, downstream intron length, 5'ss (donor site) and 3'ss (acceptor) scores, exon conservation (percent identity), upstream and downstream 150-base intron region conservation (CLUSTALW alignment score (Thompson et al., 1994)), and a list of oligonucleotide occurrence counts, described below. Length features were transformed to logarithmic ($\log_{10}$) scale and splice sites were scored using a maximum entropy model(Yeo and Burge, 2004). Exons were divided into four different regions: the last 150 bases of the upstream intron (or the entire intron for introns shorter than 150 bases), the first 150 bases of the

downstream intron (or the entire intron), the first 100 bases of the exon (or the entire exon), and the last 100 bases of the exon (or the entire exon). Occurrence counts for all oligonucleotides of length *k* for *k* ranging from 3 to 6 nucleotides were calculated from the four regions described above. Counts were generated separately from unaligned and ClustalW-aligned regions. In either case, all overlapping *k*-mers contained completely in the given region were counted. *k*-mers that occurred less than twice in the $S_{H,M}$ and $S_{h,m}$ training sets were excluded from further analysis. For training of ACEScan, *k*-mers were ranked by enrichment in $S_{H,M}$ versus $S_{h,m}$ exons/introns, as scored using a χ^2-statistic for a 2 x 2 contingency table, with Yates correction factor. For each region in $S_{H,M}$ and $S_{h,m}$ (rows contingency table), the number of occurrences of each *k*-mer and the number of occurrences of all remaining *k*-mers were determined (table columns). The oligonucleotide features were ranked and the top *N* features were extracted and concatenated into a *(M+N)*-dimensional vectors, where *M* is the number of general sequence features used. The top-ranked oligonucleotide features used by ACEScan included some 5-mers and some 4-mers (Table S1), but no 3-mers or 6-mers.

Classification and sampling

The regularized least-squares classifier (RLSC) was used to learn the features from $S_{H,M}$ and $S_{h,m}$. The RLSC has a quadratic loss function and requires the solution of a single system of linear equations (Rifkin et al., 2003). Due to the unbalanced size of the two sets, i.e. there were about 25 times more exon-pairs in $S_{h,m}$ (negative examples) than in $S_{H,M}$ (positive examples), errors made on the positive examples cost a multiplicative factor of β times greater than the penalty for errors made on the negative examples. The binary-class RLSC classification problem was stated as

$\min_f (1/L) \Sigma_{i=1:L} w_i (y_i - f(x_i))^2 + \lambda \|f\|^2_K$ (1)

where f and $\|f\|^2_K$ are the function and function norm induced in a reproducing kernel Hilbert-space, respectively, L is the size of the training set, λ is the "tradeoff" between generalization and over-fitting and w_i is a misclassification penalty, set to β if sample x_i had a positive label (y_i=1), otherwise set to 1. To address the potential for incorrect labeling of Sh,m exons because of incomplete coverage by transcript data, the misclassification parameter β for positively labeled data was set to 5, higher than the value for negatively labeled data. Assuming a solution f^* of the form

$f^*(\boldsymbol{u}) = \Sigma_{i=1:L} c_i K(\boldsymbol{u}, \boldsymbol{x_i})$ (2)

where $K(\boldsymbol{u},\boldsymbol{v}) = \langle \boldsymbol{u},\boldsymbol{v} \rangle$, c_i are coefficients, and $\underline{K}$ is the L x L kernel matrix satisfying $\underline{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and $\mathbf{W}$ is the diagonal matrix of penalties w_i, the problem was rewritten in matrix notation and the optimal $\mathbf{c}$, defining $\boldsymbol{c} = [c_1 \ldots c_l]^T$ was found, by substituting Eq. (2) into Eq. (1),

$(\underline{K} + \lambda L \mathbf{W}^{-1})\, \mathbf{c} = \mathbf{y}$. (3)

Fixing λ and β and solving for $\mathbf{c}$ using the conjugate gradient method (implemented in Matlab), test examples were assigned an output according to Eq. (2). In order to solve Eq. (3) efficiently, $\underline{K}$ was expressed as $\mathbf{A}\mathbf{A}^T$, where $\mathbf{A}$ was the L x d matrix of training examples with d features. By first computing $\boldsymbol{\alpha} = \mathbf{A}^T\mathbf{c}$, the outputs for unlabeled ('test') examples were obtained by matrix multiplication of $\mathbf{B}$ and $\boldsymbol{\alpha}$, where $\mathbf{B}$ is the n x d matrix of n unlabeled examples.

At each iteration, 80% of the conserved human-mouse exon pairs from $S_{H,M}$ and 80% of the samples from $S_{h,m}$ were used to train the classifier, and prediction outputs were generated for the remaining 20% of $S_{H,M}$ and $S_{h,m}$. Area under the curve (AUC) values were obtained for each iteration, and the average AUC value was used to measure model performance (described below). Empirically, it was found that λ=0.01 and β=5 gave optimal performance. Empirically, it

was also determined that the model labeled ***i*** in Fig. S1 obtained the highest AUC value, at a cutoff of ~ –0.5. The ACEScan score for an exon pair was defined as the mean prediction output over 500 random samples of the training set. Similarly, when ACEScan was utilized to score unseen Ensembl-annotated human-mouse exon pairs (i.e. exon pairs not in the training set), each pair was assigned an ACEScan score calculated as the mean output from 50 random samples of the training data from $S_{H,M}$ and $S_{h,m}$. The approach of taking the average output from many different samplings of the training set corresponds closely to the use of "bagging" in statistical machine learning (Duda et al., 2001). The set of ACEScan[+] exons will be made available on the web at [http://genes.mit.edu/acescan] at the time of publication.

Performance measures

Receiver Operating Curve (ROC) analysis (Swets, 1988) was used to assess the performance of models in binary hypothesis testing. An ROC plot graphically represents the true positive rate (on the *y*-axis) versus the false positive rate (*x*-axis) as a function of the threshold used in prediction, and displays the tradeoff between the sensitivity and the false positive rate (increases in sensitivity are generally accompanied by an increase in false positives). The integrated area under the ROC curve (AUC) was used to measure performance (higher AUC values correspond to improved classification performance).

Gene Ontology analysis

Gene Ontology (GO) identifiers (IDs) for each Ensembl-annotated gene were obtained from EnsMart (release version 19.1). Organizational principles (molecular function, biological process) were obtained from http://www.geneontology.org. For each term (e.g., neurogenesis,

GO ID:0007399), the fraction of genes containing predicted ACEs and not containing predicted ACEs, relative to the genes under the overall principle (e.g. GO ID:0007399 was found under biological process) was compared by a χ^2 test of significance, with Yates sample size correction (Glantz, 1997). Adjusting for multiple hypothesis testing using Bonferroni correction (217 terms were compared with at least 10 genes belonging to the term for molecular function; 187 terms were compared for biological process), enriched terms were identified at a significance cutoff of $P<0.05$.

Gene expression analysis

Affymetrix HG-U95A microarray gene expression from 47 human tissues and cell-lines previously published by Su and colleagues (Su et al., 2002) were obtained from the Gene Expression Atlas (http://expression.gnf.org). Mappings for Affymetrix probe identifiers were obtained from EnsMart (release 19.1). Average difference (AD) values lower than 20 were standardized to 20, as described (Su et al., 2002). Genes expressed in a tissue or cell-line at greater than 2 times the standard deviation above the median expression across tissues or cell-lines were defined as tissue-specifically expressed in that tissue or cell-line. For each tissue, the fraction of genes containing predicted ACEs and not containing predicted ACEs, relative to the set of all tissue-specifically expressed genes was compared using a χ^2 test, with the Yates sample size correction. Adjusting for multiple hypothesis testing using Bonferroni correction, enriched tissues were again identified at a significance cutoff of $P<0.05$.

Single nucleotide polymorphism (SNP) analysis

8,408 high-quality reference SNPs (an 11-mer with 11-mer flanks on both sides) were obtained (Fairbrother et al., 2004) and mapped to exons scored by ACEScan. The SNP density for a set of exons was calculated by dividing the total number of SNPs contained in the exons by the total length of all exons within the set.

Protein domain analysis

Human Ensembl transcripts and Ensembl annotated Pfam protein features (Bateman et al., 2004) were obtained from the EnsMart database (Ensembl version 22.34). The start and end locations of each annotated protein feature with respect to the translated transcript were obtained and compared to the coordinates of the exons in the transcript. A protein feature was considered to overlap an exon if W bases or more of the exon was within the feature. W was adjusted from 5 to 30 bases in steps of 5 bases to test the robustness of the measurement. A χ^2 test was performed to determine if high-scoring ACEScan exons overlapped exons at a lower or higher rate compared to low-scoring ACEScan exons that were designated as skipping (and non-skipping) events using transcript alignments from ESTs.

Experimental validation

The Invitrogen Superscript III First-Strand synthesis system for RT-PCR (Cat. No. 18080-051) was used to generate cDNAs from 3-4 ug of total RNA from human tissues (whole brain, fetal brain, heart, fetal liver, cerebellum, prostate, liver, lung, kidney, skeletal muscle, bone marrow and testis) and mouse tissues (whole brain, testis, liver, lung, skeletal muscle, kidney, heart and a pool from embryonic 5, 11, 15, and 17-day tissues) from Clontech (BD Biosciences) using

oligo(dT) primers. The Invitrogen Taq DNA polymerase kit (Cat. No. 18038-042) was utilized with primers designed using the primer3 program (Rozen and Skaletsky, 2000) targeted to exons flanking candidate ACEs. Forty cycles of PCR using an ABI 9700 thermocycler were conducted at denaturing temperature of 94^{o} C for 30s, annealing at 58 ^{o} C for 30 s, and elongation at 72^{o} C for 30-100 seconds depending on the size of the predicted products. PCR products were resolved on a 2% agarose gel (Merck) at 116 volts in TBE buffer. Bands of the expected size were gel-purified using the QIAquick Gel Extraction Kit (Qiagen Cat. No. 28704) according to the manufacturer's instructions. Each isolated band was amplified by additional rounds of PCR with the same primers before sequencing.

Acknowledgments

The authors thank E. Lander, P. Sharp and Z. Wang for helpful discussions. This work was supported by grants from the National Science Foundation and the National Institutes of Health (C.B.B.), and from the Lee-Kuan-Yew graduate fellowship from Singapore (G.Y.).

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.
Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., *et al.* (2004). The Pfam protein families database. Nucleic Acids Res *32 Database issue*, D138-141.
Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321-1325.
Bell, M. V., Cowper, A. E., Lefranc, M. P., Bell, J. I., and Screaton, G. R. (1998). Influence of intron length on alternative splicing of CD44. Mol Cell Biol *18*, 5930-5941.
Black, D. L. (1991). Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? Genes Dev *5*, 389-402.
Black, D. L. (1992). Activation of c-src neuron-specific splicing by an unusual RNA element in vivo and in vitro. Cell *69*, 795-807.
Black, D. L. (2000). Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell *103*, 367-370.
Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem *72*, 291-336.
Black, D. L., and Grabowski, P. J. (2003). Alternative pre-mRNA splicing and neuronal function. Prog Mol Subcell Biol *31*, 187-216.
Blanchette, M., and Chabot, B. (1997). A highly stable duplex structure sequesters the 5' splice site region of hnRNP A1 alternative exon 7B. Rna *3*, 405-419.
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.
Caceres, J. F., and Kornblihtt, A. R. (2002). Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet *18*, 186-193.
Cha, J. H., Kinsman, S. L., and Johnston, M. V. (1994). RNA editing of a human glutamate receptor subunit. Brain Res Mol Brain Res *22*, 323-328.
Chan, R. C., and Black, D. L. (1997). Conserved intron elements repress splicing of a neuron-specific c-src exon in vitro. Mol Cell Biol *17*, 2970.
Chemin, J., Monteil, A., Bourinet, E., Nargeot, J., and Lory, P. (2001). Alternatively spliced alpha(1G) (Ca(V)3.1) intracellular loops promote specific T-type Ca(2+) channel gating properties. Biophys J *80*, 1238-1250.
Cline, M. S., Shigeta, R., Wheeler, R. L., Siani-Rose, M. A., Kulp, D., and Loraine, A. E. (2004). The effects of alternative splicing on transmembrane proteins in the mouse genome. Pac Symp Biocomput, 17-28.
Dominski, Z., and Kole, R. (1991). Selection of splice sites in pre-mRNAs with short internal exons. Mol Cell Biol *11*, 6075-6083.
Dredge, B. K., Polydorides, A. D., and Darnell, R. B. (2001). The splice of life: alternative splicing and neurological disease. Nat Rev Neurosci *2*, 43-50.
Dreyfuss, G., Kim, V. N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nat Rev Mol Cell Biol *3*, 195-205.

Duda, R. O., Hart, P. E., and D.G., S. (2001). Pattern Classification (New York, John Wiley & Sons).
Estes, P. A., Cooke, N. E., and Liebhaber, S. A. (1992). A native RNA secondary structure controls alternative splice-site selection and generates two human growth hormone isoforms. J Biol Chem *267*, 14902-14908.
Fairbrother, W. G., Holste, D., Burge, C. B., and Sharp, P. A. (2004). Single Nucleotide Polymorphism-Based Validation of Exonic Splicing Enhancers. PLoS Biol *2*, E268.
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.
Faustino, N. A., and Cooper, T. A. (2003). Pre-mRNA splicing and human disease. Genes Dev *17*, 419-437.
Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res *8*, 967-974.
Forch, P., Puig, O., Kedersha, N., Martinez, C., Granneman, S., Seraphin, B., Anderson, P., and Valcarcel, J. (2000). The apoptosis-promoting factor TIA-1 is a regulator of alternative pre-mRNA splicing. Mol Cell *6*, 1089-1098.
Glantz, S. A. (1997). Primer of Biostatistics, Fourth edn (New York, McGraw-Hill).
Grabowski, P. J., and Black, D. L. (2001). Alternative RNA splicing in the nervous system. Prog Neurobiol *65*, 289-308.
Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T.*, et al.* (2002). The Ensembl genome database project. Nucleic Acids Res *30*, 38-41.
Huh, G. S., and Hynes, R. O. (1994). Regulation of alternative pre-mRNA splicing by a novel repeated hexanucleotide element. Genes Dev *8*, 1561-1574.
Iida, K., and Akashi, H. (2000). A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene *261*, 93-105.
Jiang, Z., Cote, J., Kwon, J. M., Goate, A. M., and Wu, J. Y. (2000). Aberrant splicing of tau pre-mRNA caused by intronic mutations associated with the inherited dementia frontotemporal dementia with parkinsonism linked to chromosome 17. Mol Cell Biol *20*, 4036-4048.
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.
Kashima, T., and Manley, J. L. (2003). A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. Nat Genet *34*, 460-463.
Kaufmann, D., Kenner, O., Nurnberg, P., Vogel, W., and Bartelt, B. (2004). In NF1, CFTR, PER3, CARS and SYT7, alternatively included exons show higher conservation of surrounding intron sequences than constitutive exons. Eur J Hum Genet *12*, 139-149.
Konig, G., Monning, U., Czech, C., Prior, R., Banati, R., Schreiter-Gasser, U., Bauer, J., Masters, C. L., and Beyreuther, K. (1992). Identification and differential expression of a novel alternative splice isoform of the beta A4 amyloid precursor protein (APP) mRNA in leukocytes and brain microglial cells. J Biol Chem *267*, 10804-10809.

Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S., and Sunyaev, S. (2003). Increase of functional diversity by alternative splicing. Trends Genet *19*, 124-128.
Ladd, A. N., and Cooper, T. A. (2002). Finding signals that regulate alternative splicing in the post-genomic era. Genome Biol *3*, reviews0008.
Libri, D., Piseri, A., and Fiszman, M. Y. (1991). Tissue-specific splicing in vivo of the beta-tropomyosin gene: dependence on an RNA secondary structure. Science *252*, 1842-1845.
Liu, H. X., Zhang, M., and Krainer, A. R. (1998). Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev *12*, 1998-2012.
Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M., and Frazer, K. A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science *288*, 136-140.
Lopez, A. J. (1998). Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. Annu Rev Genet *32*, 279-305.
Maniatis, T., and Tasic, B. (2002). Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature *418*, 236-243.
Mittman, S., Guo, J., and Agnew, W. S. (1999). Structure and alternative splicing of the gene encoding alpha1G, a human brain T calcium channel alpha1 subunit. Neurosci Lett *274*, 143-146.
Modafferi, E. F., and Black, D. L. (1997). A complex intronic splicing enhancer from the c-src pre-mRNA activates inclusion of a heterologous exon. Mol Cell Biol *17*, 6537-6545.
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.
Ponte, P., Gonzalez-DeWhitt, P., Schilling, J., Miller, J., Hsu, D., Greenberg, B., Davis, K., Wallace, W., Lieberburg, I., and Fuller, F. (1988). A new A4 amyloid mRNA contains a domain homologous to serine proteinase inhibitors. Nature *331*, 525-527.
Rifkin, R., Yeo, G., and Poggio, T. (2003). Regularized Least Squares Classification. In Advances in Learning Theory: Methods, Model and Applications, H. Suykens, Basu, Micchelli, Vandewalle, ed. (Amsterdam, IOS Press).
Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol *132*, 365-386.
Shibata, A., Hattori, M., Suda, H., and Sakaki, Y. (1996). Identification of cis-acting elements involved in an alternative splicing of the amyloid precursor protein (APP) gene. Gene *175*, 203-208.
Smith, C. W., and Nadal-Ginard, B. (1989). Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing. Cell *56*, 749-758.
Sommer, B., Keinanen, K., Verdoorn, T. A., Wisden, W., Burnashev, N., Herb, A., Kohler, M., Takagi, T., Sakmann, B., and Seeburg, P. H. (1990). Flip and flop: a cell-specific functional switch in glutamate-operated channels of the CNS. Science *249*, 1580-1585.

Sorek, R., and Ast, G. (2003). Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res *13*, 1631-1637.
Sorek, R., Shamir, R., and Ast, G. (2004a). How prevalent is functional alternative splicing in the human genome? Trends Genet *20*, 68-71.
Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G., and Shamir, R. (2004b). A non-EST-based method for exon-skipping prediction. Genome Res *14*, 1617-1623.
Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A.*, et al.* (2002). Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A *99*, 4465-4470.
Sugnet, C. W., Kent, W.J., Ares JR, M., Haussler, D. (2004). Transcriptome and genome conservation of alternative splicing events in humans and mice. Paper presented at: Pacific symposium on biocomputing (Hawaii, World Scientific).
Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. Science *240*, 1285-1293.
Thanaraj, T. A., Clark, F., and Muilu, J. (2003). Conservation of human alternative splice events in mouse. Nucleic Acids Res *31*, 2544-2552.
Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-4680.
Wollerton, M. C., Gooding, C., Wagner, E. J., Garcia-Blanco, M. A., and Smith, C. W. (2004). Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. Mol Cell *13*, 91-100.
Xing, Y., Xu, Q., and Lee, C. (2003). Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. FEBS Lett *555*, 572-578.
Yeo, G., and Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol *11*, 377-394.
Zhang, M. Q. (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc Natl Acad Sci U S A *94*, 565-568.
Zhang, M. Q. (2000). Discriminant analysis and its application in DNA sequence motif recognition. Brief Bioinform *1*, 331-342.
Zhang, X. H., Heller, K. A., Hefter, I., Leslie, C. S., and Chasin, L. A. (2003). Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. Genome Res *13*, 2637-2650.
Zheng, C. L., Nair, T. M., Gribskov, M., Kwon, Y. S., Li, H. R., and Fu, X. D. (2004). A database designed to computationally aid an experimental approach to alternative splicing. Pac Symp Biocomput, 78-88.

Figure Legends

Figure 1. Schematic overview of the learning and prediction stages of the ACEScan procedure. Human/mouse orthologous gene pairs were taken from EnsMart (www.ensembl.org/Multi/martview). Reciprocal best BLAST hits were used to identify orthologous human-mouse exons within these orthologous genes. Spliced alignment of ESTs to cDNA-verified regions of assembled human and mouse genomic sequences were used to infer splicing patterns of exons. In "Learning", the set $S_{H,M}$ (capital H, M) consists of 241 exons included in some transcript sequences (including ESTs and/or cDNAs) and excluded in other(s) in both human and mouse; the set $S_{h,m}$ (lower-case h, m) consists of ~5,000 exons from genes exhibiting AS that were included in all overlapping transcript sequences in both human and mouse. Sequence features that differed between these sets were identified as described (supporting online text). Random subsets of $S_{H,M}$ and $S_{h,m}$ were used to train the ACEScan algorithm and cross-validation scores were calculated for the unseen subsets of $S_{H,M}$ and $S_{h,m}$. The cross-validated ACEScan score distributions for $S_{H,M}$ (red) and $S_{h,m}$ (black) are shown. For "Prediction", spliced alignment of transcript sequences were used to assign Ensembl annotated exons from ~10,000 human-mouse orthologous gene pairs (not necessarily alternatively spliced) to one of two sets. The set $S_{H,m}$ (capital H, lower-case m) contains exons included in some human transcripts and excluded in others, but included in all mouse transcripts; the set $S_{h,m}$ contains exons not observed to be excluded by any human or mouse transcript ACEScan score distributions for $S_{h,m}$ (blue) and $S_{H,m}$ (purple) are shown. Exons with ACEScan score > 0 are predicted ACEs.

Figure 2. Sequence features that differ between conserved alternative and constitutive human-mouse exons. **(A)** Features typical of exons of the $S_{H,M}$ (alternatively spliced) and $S_{h,m}$ (constitutive) training sets are depicted. $S_{H,M}$ exons had shorter median exon length (93 bp versus 126 bp, $P<10^{-22}$), longer upstream intron length ($P<0.005$), longer downstream intron length ($P<10^{-5}$), weaker 5' and 3' splice site scores ($P<10^{-5}$ and $P<0.02$, respectively using MaxEntScan scores (http://genes.mit.edu), higher exon sequence conservation (percent identity; $P<10^{-46}$), and higher conservation (CLUSTALW alignment score) in the 150 bp intron regions immediately upstream and downstream of the exon ($P<10^{-63}$ and $P<10^{-66}$, respectively). For each feature, the Kolmogorov-Smirnov (KS) test was used to test the null hypothesis of independent samples drawn from the same underlying population. Length and splice site score values are shown for human exons/introns; mouse values were similar. Average percent identity for alignments of flanking intron regions are shown in a 9-base sliding window for $S_{H,M}$ (red) and $S_{h,m}$ (dashed, black) exons. **(B)** Pentanucleotides utilized by ACEScan. Over- (under-) represented pentamers in exon or 150-base flanking intron regions of $S_{H,M}$ versus $S_{h,m}$ exons are shown in red (black). Pentamer frequencies were analyzed separately for CLUSTALW-aligned regions only ('aligned') or entire region ('unaligned'). Exon 5'-, 3'-ends refer to first/last100 bases of exon.

Figure 3. Validation of ACEScan[+] predictions using experimental and computational approaches. **(A)** Experimental validation via RT-PCR and sequencing of subsets of candidate ACEScan[+] exons and negative control ACEScan[-] exons in panels of normal human and mouse tissues with primers in flanking upstream and downstream

exons (described in Experimental Methods). Graphical representations of splicing patterns (inclusion/exclusion) and the number of exon-pairs observed to be excluded and included are designated in red and black respectively. The three randomly selected subsets tested were: (i) 30 ACEScan[+] exon pairs; and as negative controls, (ii) 15 ACEScan[-] S_H exon pairs (with EST/cDNA evidence for both inclusion and exclusion of the human exon, indicated by horizontal lines representing spliced transcripts); and (iii) 11 ACEScan[-] S_h exon pairs (with no transcript evidence for skipping in either human or mouse). **(B)** SNP density in ACEScan[+], ACEScan[-] S_H, and ACEScan[-] S_h exons. The number of stringently filtered SNPs per 10,000 bases were computed for each exon set. **(C)** Fraction of $S_{H,M}$ exons, ACEScan[+] S_H exons, and ACEScan[-] S_H exons that have lengths that were a multiple of 3, and the background fraction of frame-preserving constitutive exons. **(D)** Analysis of protein domain preservation of ACEScan[+], ACEScan[-] S_H, and ACEScan[-] S_h exons that maintain reading-frame (i.e. length modulo 3). Maximum exon size cutoffs (150, 110 and 108 bp for ACEScan[+], ACEScan[-] S_H, and ACEScan[-] S_h exons respectively) were utilized to avoid exon length biases. The median length of exons in each subset was 84 bases, with no significant difference in the distribution of sizes among the sets (by a Kruskal-Wallis non-parametric test). The minimum number of exonic bases overlapping the protein domain was set to 30 bases.

Figure 4. ACEScan plots for well-known alternatively spliced genes. ACEScan scores for all orthologous human-mouse exons identified by our automated procedure (see legend to Fig. 1) are shown. Known alternative exons are indicated by asterisks; the

known RNA edited exon of GLUR-B is indicated by the letter 'E'. The following known AS exons are illustrated: (**A**) Exons 14 (69 bp) and 26 (54 bp) of the human voltage-dependent T-type calcium channel alpha 1G subunit gene (*CACNA1G*, Ensembl gene identifier ENSG00000006283); (**B**) Exons 7 (168 bp), 8 (57 bp) and 15 (54 bp) of the human amyloid beta protein precursor gene (*APP*, ENSG00000142192); (**C**) Exon 11 (34 bp) of the human polypyrimidine tract-binding protein gene (*PTB*, ENSG00000011304); (**D**) Exons 14 (115 bp) and 15 (249 bp) of the human glutamate receptor, AMPA 2 gene (*GLURB*, ENSG00000120251).

Figure 5. Functional classification and expression patterns of genes with predicted ACEs. (A) Gene Ontology (GO) "Molecular function" and "Biological process" categories which differed significantly ($P < 0.05$) in representation between genes containing predicted ACEs (black bars) and genes not containing predicted ACEs (white bars) are shown. Statistical significance was assessed using χ^2 statistics with Bonferroni correction for multiple hypothesis testing. GO categories are ordered from right to left in order of increasingly significant bias towards genes containing predicted ACEs. Only one category (transport) was significantly biased towards genes without predicted ACEs. (B) Tissues exhibiting significant over- / under-representation of genes containing predicted ACEs ($P < 0.05$) in tissue-specifically expressed genes. Genes differentially expressed (greater than two-fold higher than median value across all tissues) in a HG-U95A microarray study (see main text) were considered tissue-specifically expressed. Tissues are ordered from right to left in order of increasingly significant bias towards genes containing predicted ACEs.

Table 1. ACEScan scores of mouse skipped exons from the MAASE database (Zheng et al., 2004).

No.	MAASE Gene Name	MAASE ID	Ensembl mouse identifier	Exon #	Human exon length (bp)	Mouse exon length (bp)	ACEScan score
1	NOS1 - Nitric-oxide synthase, brain	267255	29361	15	102	102	0.12
2	MBP - Myelin basic protein	292857	41607	9	78	78	-0.41
3	MBP - Myelin basic protein	292857	41607	12	33	33	0.06
4	MBP - Myelin basic protein	292857	41607	13	120	123	-0.39
5	ACSL6 - Long-chain-fatty-acid--CoA ligase 6	325273	20333	13	78	78	0.10
6	ACSL6 - Long-chain-fatty-acid--CoA ligase 6	325273	20333	14	78	78	0.36
7	MCF2L - Guanine nucleotide exchange factor DBS	330913	31442	13	93	93	-0.39
8	MCF2L - Guanine nucleotide exchange factor DBS	330913	31442	14	74	75	-0.14
9	SPTAN1 - Spectrin alpha chain, brain	340288	57738	11	60	60	0.72
10	SPTAN1 - Spectrin alpha chain, brain	340288	57738	44	18	16	0.46
11	TEC - Tyrosine-protein kinase Tec	272000	29217	11	66	66	-0.61
12	TNNT2 - Troponin T, cardiac muscle isoforms	273074	26414	7	30	37	0.11
13	KCNMA1 - Calcium-activated potassium channel alpha subunit 1	275823	63142	27	174	174	0.63
14	KCNMA1 - Calcium-activated potassium channel alpha subunit 1	275823	63142	33	81	81	0.39
15	KCNMA1 - Calcium-activated potassium channel alpha subunit 1	275823	63142	39	29	29	1.45
16	CAST - Calpain inhibitor	287853	21585	9	66	57	-0.18
17	CAST - Calpain inhibitor	287853	21585	34	45	36	-0.33
18	DTNA - Dystrobrevin alpha	315011	24302	25	78	78	1.16
19	DTNA - Dystrobrevin alpha	315011	24302	33	21	21	1.58
20	DTNA - Dystrobrevin alpha	315011	24302	46	93	93	-0.38
21	MBNL2 - Muscleblind-like 2 Isoform 1	315631	22139	9	54	54	2.08
22	TNNT3 - Troponin T, fast skeletal muscle isoforms	329096	61723	8	18	18	-0.1
23	TNNT3 - Troponin T, fast skeletal muscle isoforms	329096	61723	9	21	15	0.22
24	TNNT3 - Troponin T, fast skeletal muscle isoforms	329096	61723	24	41	41	0.54
25	TNNT3 - Troponin T, fast skeletal muscle isoforms	329096	61723	25	41	41	-0.42
26	TPM3 - Tropomyosin alpha 3 chain	272578	27940	9	76	76	0.38
27	TPM3 - Tropomyosin alpha 3 chain	272578	27940	10	76	76	0.16
28	TPM3 - Tropomyosin alpha 3 chain	272578	27940	14	79	83	-0.05
29	PHKA1 - Phosphorylase B kinase alpha regulatory chain, skeletal muscle	292503	34055	22	177	177	0.39

Table 1. MAASE id, gene name, and exon sizes for the identified mouse and orthologous human exons. Exon pairs were then scored by ACEScan.

FIGURE 1.

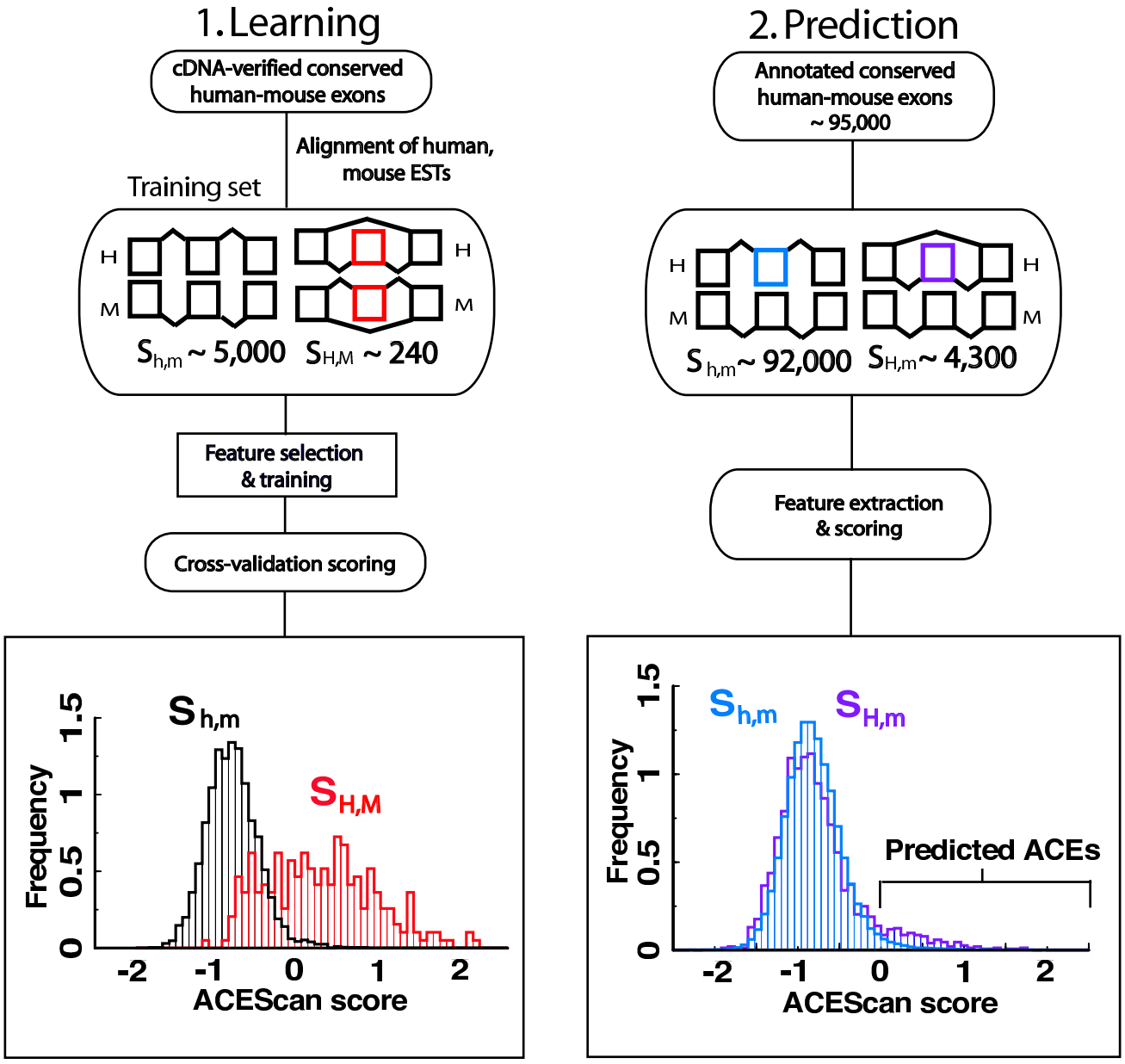
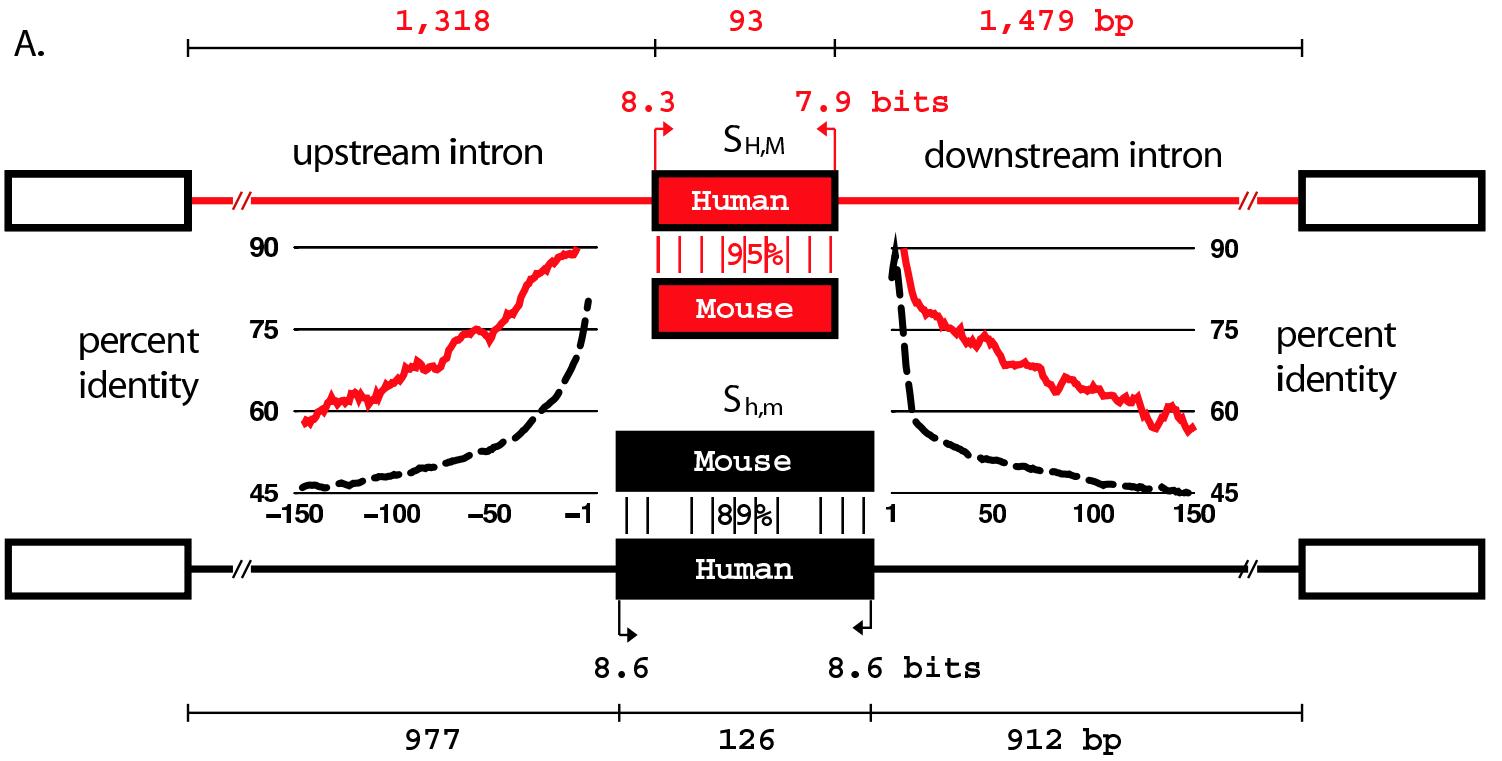

1. Learning
cDNA-verified conserved human-mouse exons
Alignment of human, mouse ESTs
Training set
H
M
H
M
S h,m ~ 5,000
S H,M ~ 240
Feature selection & training
Cross-validation scoring
S h,m
S H,M
Frequency
0
0.5
1
1.5
-2
-1
0
1
2
ACEScan score
2. Prediction
Annotated conserved human-mouse exons ~ 95,000
H
M
H
M
S h,m ~ 92,000
S H,m ~ 4,300
Feature extraction & scoring
S h,m
S H,m
Predicted ACEs
Frequency
0
0.5
1
1.5
-2
-1
0
1
2
ACEScan score

FIGURE 2.



B.

Upstream intronic region

Unaligned	Aligned
GCCGC	UGCAU
UUUCC	GCAUG
UCUCU	CGGGG
UGUCU	ACACU
UUGUC	UUAAC
CUCUU	ACUAC
CUUUC	CCUAU
UCUUU	UCCAU
CGCCG	CAUAU
UCCUU	CUAUU
CCGCU	GUGAG
UGCAU	
CCUUU	
UUCUC	
UUUGC	
UUUUU	
GUUUC	
GGACA	
UGGAG	
GGCUG	
GACAG	

Exon region

Exon 5' end	Aligned		Exon 3' end
CCUCC	UGUAG	UAGGG	CCUCC
CUCCC	GUAGU	CCUCG	CCCUC
CCCUC	ACUAG	CGGCG	CAAUC
CAAUC	UAGAA	UACGA	CUCCC
UCCCU	CCUCC	UCUCU	UCCCU
CCCCC	CGAAG	AGUAG	GUCCC
ACAUA	CCCGC	CCGCG	UCCUU
CGGCG	CUACG	AGUUA	UCCCC
CCCGC	UAACG	UUAAU	GAAAG
UAACC	UAAAC	UAGGA	CCCGC
UCUCU	UCUAG	ACUGA	CGAAG
CAGGG	UAGUG	UCCUU	AAAAG
CGUGG	UAACC	GAAAG	UGAAA
	CGCGG	CUCGC	UGGCC
	CGCCC	UAAUU	CAUCA
	UAGGC	CAAUC	
	CACGA	CCGCU	
	CGUAG	CCCCC	
	CCCUA	GGCGG	
	GUCGU	CUAGC	
	UUACG	CUCUC	
	AUCAA	ACCUG	
	UGGAG	CUGGA	
	CAUCA	GCUGG	
		CUGGC	

Downstream intronic region

Unaligned	Aligned
GCAUG	UGCAU
UGCAU	GCAUG
CAUGC	UCGCA
AUGCA	UGUGC
CUAAC	CAUUG
GUUUG	UGCCG
ACUAA	UCGCG
CACUA	CUAAC
CACUU	GUUUG
CAAAU	GUGAG
CAUGG	CAGGG
UAAGU	GGACA
UAAGA	AUGAU
UGAGU	GACAG
GUAAG	
GUGAG	

FIGURE 3.

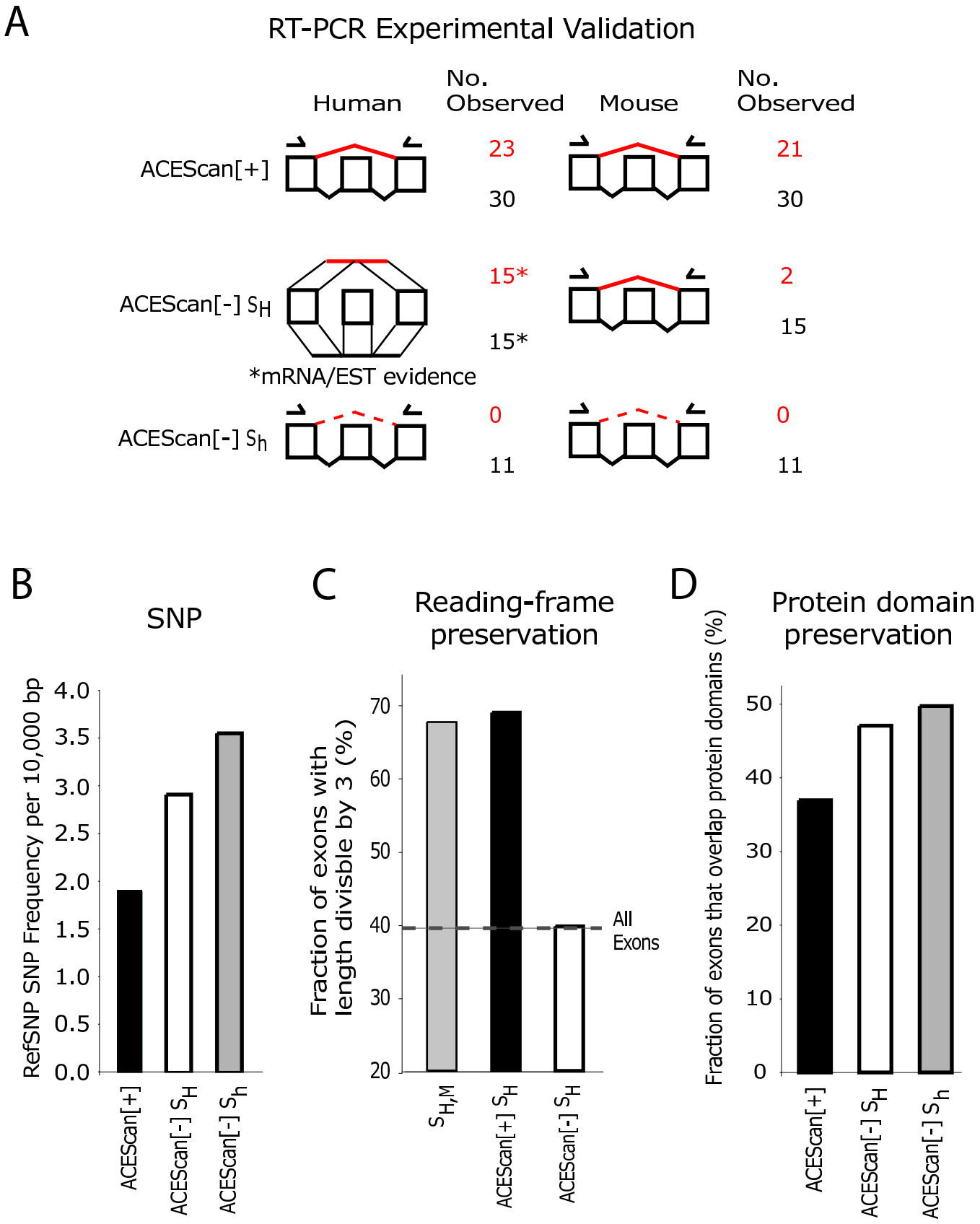

A
RT-PCR Experimental Validation
Human
No. Observed
Mouse
No. Observed
ACEScan[+]
23
30
21
30
ACEScan[-] S_H
15*
15*
2
15
*mRNA/EST evidence
ACEScan[-] S_h
0
11
0
11
B
SNP
RefSNP SNP Frequency per 10,000 bp
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
ACEScan[+]
ACEScan[-] S_H
ACEScan[-] S_h
C
Reading-frame preservation
Fraction of exons with length divisble by 3 (%)
70
60
50
40
30
20
All Exons
$S_{H,M}$
ACEScan[+] S_H
ACEScan[-] S_H
D
Protein domain preservation
Fraction of exons that overlap protein domains (%)
50
40
30
20
10
0
ACEScan[+]
ACEScan[-] S_H
ACEScan[-] S_h

FIGURE 4.

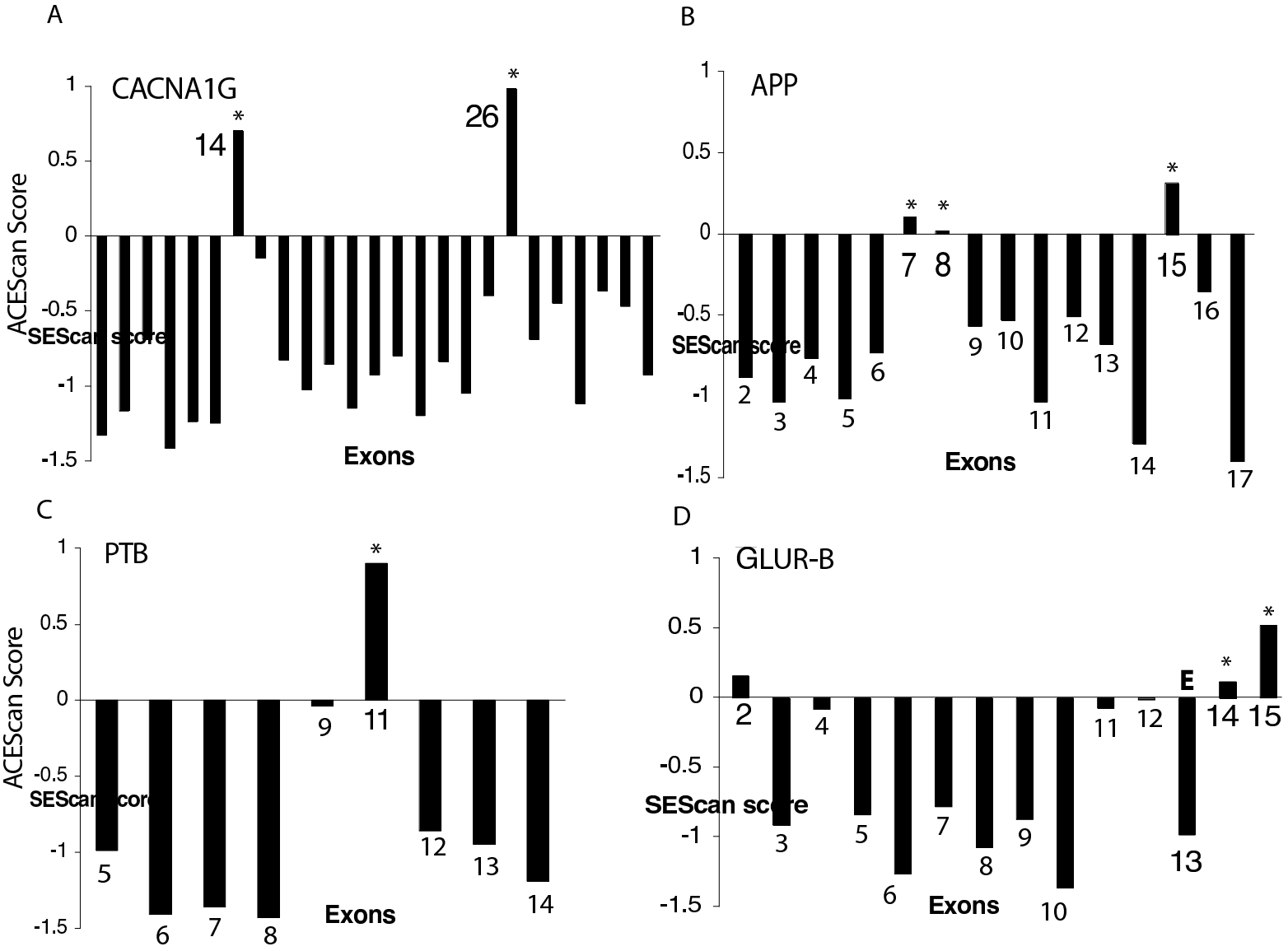
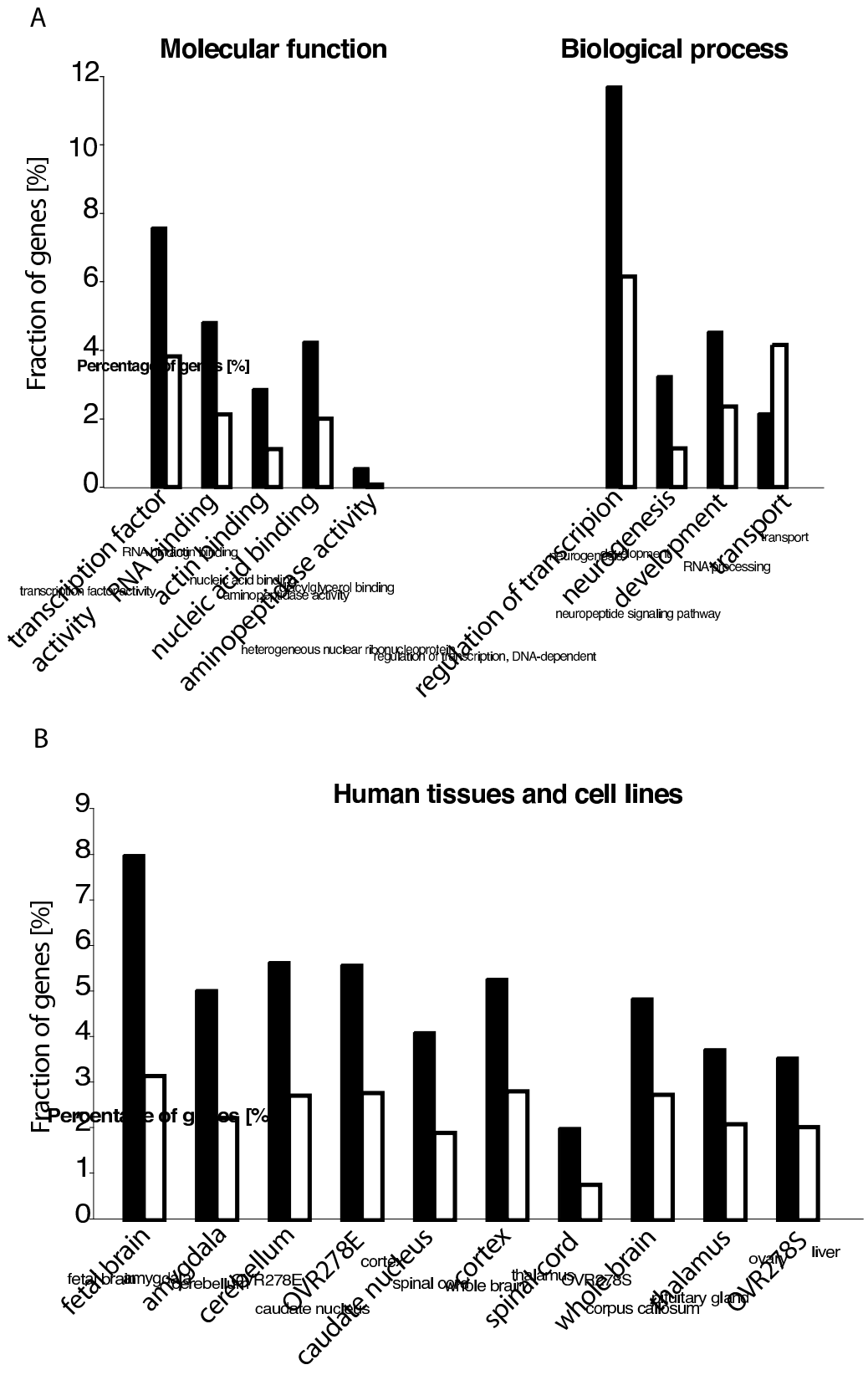

A
CACNA1G
ACEScan Score
SEScan score
Exons
14
26
B
APP
SEScan score
Exons
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
C
PTB
ACEScan Score
SEScan score
Exons
5
6
7
8
9
11
12
13
14
D
GLUR-B
SEScan score
Exons
2
3
4
5
6
7
8
9
10
11
12
13
14
15

FIGURE 5.


A
Molecular function
Biological process
Fraction of genes [%]
12
10
8
6
4
2
0
transcription factor activity
RNA binding
actin binding
nucleic acid binding
aminopeptidase activity
regulation of transcription
neurogenesis
development
RNA processing
transport
neuropeptide signaling pathway
heterogeneous nuclear ribonucleoprotein
regulation of transcription, DNA-dependent
B
Human tissues and cell lines
Fraction of genes [%]
9
8
7
6
5
4
3
2
1
0
fetal brain
amygdala
cerebellum
OVR278E
caudate nucleus
cortex
spinal cord
whole brain
thalamus
OVR278S
liver

Legends to Supporting Figures

Fig. S1. Performances of different models differing in the choice of features and the number of oligonucleotide features. A. The average area under the curve (AUC) values obtained from cross-validation with different models using different numbers of top ranking oligonucleotide features are shown. B. Different models (a to i) are shown in the table, utilizing (denoted with a tick mark, otherwise a cross) different combinations of features, such as core features (5'ss, 3'ss scores, exon lengths, upstream and downstream flanking intron lengths), upstream and downstream intron alignment scores, exon alignment or similarity scores; and oligonucleotide features from aligned regions (exon alignment features, upstream and downstream intron alignment features), as well as unaligned regions (exon features, flanking intron features).

Fig. S2. Fraction of tetramers and pentamers differentially represented as calculated by the χ^2 statistic with Yates correction, at different cutoffs (3.841, 6.635, 10.828) in $S_{H,M}$ versus $S_{h,m}$ in the different regions (aligned exon region, aligned upstream and downstream intronic regions, unaligned 5' and 3' ends of exons, and unaligned upstream and downstream intron) relative to all enumerable tetramers and pentamers. This illustrates that aligned exonic regions contain the most numbers of different tetramer and pentamer features between $S_{H,M}$ and $S_{h,m}$.

Fig. S3. The fraction of genes containing putative ACEs from $S_{H,M}$ used in training, the fraction of genes containing predicted ACEs, and the fraction of genes from all Ensembl-annotated genes annotated in the various Gene Ontology (GO) terms.

Table S1.

Rank	Upstream intron	Rank	Exon 5' end	Rank	Exon 3' end	Rank	Downstream intron	Rank	Aligned upstream intron	Rank	Aligned exon	Rank	Aligned downstream intron
12	tttc	11	cctcc	15	tccc	27	tgcat	10	tgcat	21	tgtag	2	gcatg
16	cttt	23	tccc	25	cctcc	80	gcatg	20	cgct	22	gtagt	6	tgcat
19	gccgc	30	cctc	39	ccctc	109	tcgca	28	gcat	26	actag	7	catg
47	tttt	46	ctccc	41	caatc	111	tccg	34	ccgc	31	tagaa	13	gcat
50	tttcc	61	ccctc	51	ctccc	136	cttt	37	gcatg	33	cctcc	17	catgc
63	ttgc	94	caatc	56	cctc	144	cgca	38	ctat	35	cgaag	18	tgca
65	tctct	119	tccct	67	ccct	165	tgtgc	43	ttgc	45	cccgc	24	ttgc
73	tctc	125	ctcc	74	aaag	168	cattg	48	cgtc	53	ctacg	29	acta
75	cgct	135	ccccc	96	tccct	171	tgccg	49	gtcg	55	taacg	58	atgca
78	ccgc	167	acata	110	gtccc	185	tcgcg	57	cgggg	64	taaac	60	ctaa
84	tccg	180	cggcg	113	tcctt	201	ccgc	59	acta	68	tctag	76	ctaac
97	tgtct	196	aaag	118	aggg	213	ctaac	72	caat	69	tagtg	77	gtttg
99	ttgtc	206	aggg	140	ctcc	217	gtttg	81	cggg	79	taacc	91	gccg
102	ctctt	207	cccc	152	tcccc	219	gacag	82	acgt	89	cgcgg	93	actaa
117	ctttc	215	cccgc	199	gaaa	184	ggaca	83	cacg	95	cgccc	122	ccaa
124	gccg	220	taacc	203	gaaag	194	atgat	86	atcc	101	taggc	123	cacta
126	tcttt	222	tctct	211	cccgc	198	agag	87	acact	103	cacga	139	cactt
132	tctt	223	caggg	225	tagg	141	aggg	90	ttaac	106	cgtag	166	cggt
133	cgccg	212	tgat	234	cgaag	98	caggg	100	actac	128	cccta	175	caat
134	tcctt	230	cgtgg	236	aaaag	85	gagg	105	cctat	129	gtcgt	176	caaat
146	tcct	162	gacc	238	tgaaa	88	cagg	107	gccg	130	ttacg	205	catgg
148	ccgct	114	ctgg	172	catca	42	tgag	115	caac	137	taggg	221	atgc
149	tgcat			195	gtac	32	gtgag	120	cgtt	138	cctcg	228	cact
151	ttcc			155	tggcc			127	cgca	142	cggcg	150	agag
159	ccttt			70	ctgg			145	tccat	143	tacga	92	taagt
169	ctct							154	catat	153	tctct	40	taaga
182	ttctc							160	tgcg	158	agtag	36	gagt
187	ctat							161	ggcg	163	ccgcg	14	tgagt
200	tttg							178	cccg	164	agtta	9	taag
204	tttgc							186	ccga	170	ttaat	8	gtaa
224	ttttt							188	cgat	173	tagga	5	gtga
232	gtttc							190	catc	174	actga	4	tgag
208	ggac							214	ctatt	177	tccttt	3	gtaag
210	gagg							226	cgga	183	gaaag	1	gtgag
227	aggc							240	tccg	189	ctcgc		
229	gacag							237	cctg	191	taatt		
179	ggctg							239	ccag	193	caatc		
181	cagg							192	aaaa	197	ccgct		
156	tggag							147	cagg	209	ccccc		
121	ggaca							157	gagg	216	ggcgg		
131	gagc							116	gtgag	218	ctagc		
62	tggg							104	ctgg	231	ctctc		
44	ggag							66	tgag	233	ccgaa		
										235	atcaa		
										202	tggag		
										108	catca		
										112	acctg		
										71	ctgga		
										52	gctgg		
										54	ctggc		

Table S1. The highest ranked 240 oligonucleotide sequences utilized in ACEScan. These oligonucleotide features are ranked from most significantly differentially represented between $S_{H,M}$ and $S_{h,m}$ (1) to least significantly differentially represented (240) by the χ^2 statistic with Yates correction. Within each column, features are ordered by the level of enrichment in $S_{H,M}$ versus $S_{h,m}$, and represented according to pertinent sequence regions. Features in black (red) represent oligonucleotides enriched in $S_{H,M}$ ($S_{h,m}$).

Table S2.

Cutoff for enrichment	**3.84**	**5.02**	**6.64**
Fraction of oligonucleotides enriched in $S_{H,M}$ exons that overlap RESCUE-annotated ESE hexamers	0.36	0.39	0.39
Fraction of oligonucleotides enriched in $S_{h,m}$ exons that overlap RESCUE-annotated ESE hexamers	0.88	0.90	0.83
p value	0	1.0E-13	1.0E-6

Table S2. Overlap of over and under-represented oligonucleotides in human $S_{H,M}$ versus $S_{h,m}$, and RESCUE-annotated ESEs. Oligonucleotides (4,5-mers) that were enriched in $S_{H,M}$ or $S_{h,m}$ at different cutoffs (3.84, 5.02, 6.64, corresponding to χ^2 values for different *p* value cutoffs) were considered to overlap a RESCUE-annotated ESE if a subsequence of length 4 or 5 bases of an oligonucleotide was an exact match to a continuous subsequence constructed from a RESCUE-annotated ESE hexamer (Fairbrother et al., 2002). RESCUE-annotated ESEs are under-represented in the set of *k*-mers that are enriched in $S_{H,M}$, as opposed to the set that are enriched in $S_{h,m}$. The statistical significance was assessed, by using the χ^2 test with Yates correction term.

Table S3.

Gene Name	Human Ensembl ID	Human Exon No.	Exon length (bp)	Mouse Ensembl ID	Mouse Exon No.	Exon length (bp)	ACE-Scan Score	Evidence
HnRNP A1	135486	7	156	036021	3	159	0.81	Ref. (Blanchette and Chabot, 1997)
HnRNP D0	138668	2	57	000568	2	57	0.21	Ref. (Wagner et al., 1998)
HnRNP H	169045	5	139	007850	5	139	0.78	Table S4
HnRNP I/PTB	011304	11	34	006498	8	34	0.90	Ref. (Wollerton et al., 2004)
HnRNP K	165119	16	170	021546	14	170	0.33	-
HnRNP K	165119	6	44	021546	4	44	1.39	EST BI115223
HnRNP K	165119	8	72	021546	6	72	1.02	EST AK096385
HnRNP L	104824	6	73	015165	6	73	0.16	-
HnRNP M	099783	3	53	002291	3	53	1.00	4 Human, 3 Mouse ESTs
HnRNP R	125944	2	108	028666	3	119	0.28	-
HnRNP FUS	089280	6	35	030795	6	35	0.29	-

Table S3. Heterogeneous nuclear ribonucleoproteins with predicted ACEs. The first column contains the Ensembl-identifier (last 6 digits) of the human and the orthologous mouse genes, the second column contains the exon number of the predicted exon in the longest transcript of the Ensembl-annotated genes, the third column contains the exon length, and the fourth column contains the corresponding ACEScan score and the last column contains literature or transcript-based evidence for exon skipping.

Table S4.

ID	Human Ensembl ID	Human Exon	Human Exon Length (bp)	Mouse Ensembl ID	Mouse Exon	*Gene Name*	SEScan Score	Skipping in human tissues (RT-PCR)	Skipping in mouse tissues (RT-PCR)	Other evidence
S1	091129	24	36	020598	24	*NRCAM*	1.73	Ce	Br, Te, Emb	Ref. (Wang et al., 1998)
S2	078328	5	93	008658	5	*A2BP1*	1.67	-	Sk, Lu, Li, Te, Emb	Ref. (Shibata et al., 2000)
S3	83312	20	76	009470	19	*TNPO1*	1.61	Br, Te, He, Ce, Sk, Bm, FBr	Br	
S4	079819	12	63	019978	12	*EPB41L2*	1.59	Br, Pr, Li, Lu, Ki, Sk, Bm, Te	Br, Ki, Li, Sk, Te, Emb, He, Lu	
S5	137764	17	27	032364	16	*MAPKK 5*	1.48	-	Ki, Li	Human NM_002757.2 skips 17 & 18
S6	137764	18	33	032364	17	*MAPKK 5*	1.38	-	Ki, Li	Human NM_002757.2 skips 17 & 18
S7	175388	4	33	048320	4	*Q8N787*	1.26	Pr, Li, Lu, Ki, Sk, Bm, Te	Ki, Sk, Lu, Li, Te, Emb, He	
S8	156113	9	92	021780	2	*KCNMA1*	1.22	Te	Li, Lu	
S9	182872	4	74	031060	5	*RBM 10*	1.05	Li, Lu, Te	Br, Ki, Sk, Lu, Li, Te, Emb, He	
S10	144331	4	163	027016	3	*ZNF533*	1.03	Br, Pr, Te, He, Li, Ce, Sk, Bm, FLi, Ki, FBr	Br, Li	
S11	130558	4	220	026833	4	*OLFM1*	0.99	FBr	Br, Ki	
S12	155970	9	39	039478	8	NM_181723	0.95	Ce, FBr, Br	Br	
S13	172660	8	35	020680	8	*TAF15*	0.92	Br, Pr	Br, Ki, Sk, Lu	
S14	112062	9	80	024004	9	*MAPK14*	0.80	Pr, Li, Lu, Ki, Sk, Bm, Te	Sk, Li	
S15	169045	5	139	007850	5	*HNRNPH*	0.78	Br, Pr, Li, Lu, Ki, Sk, Bm, Te	Br, Li, Te	
S16	079819	13	147	019978	13	*EPB41L2*	0.68	Br, Pr, Li, Lu, Ki, Sk, Bm, Te	Br, Ki, Li, Sk, Te, Emb, He, Lu	
S17	114098	11	83	032468	13	NM_014154	0.67	Li, Lu, Ki	Sk, Lu	
S18	169057	2	125	031393	2	*MECP2*	0.36	-	Br, Li, Te, Ki, Lu, Bm, He	Ref. (Kriaucionis and Bird, 2004)
S19	149970	9	147	025658	8	NM_014927	0.21	Br, Pr, Li, Lu, Ki, Sk, Bm, Te	Br, Ki, Te	
S20	122367	4	368	021798	5	*LDB3*	0.14	Te	Br, Ki, Li, Te, Emb, He	
S21	101977	5	117	031139	3	*MCF2*	0.10	Br	Br, Ki	
S22	060237	12	279	045962	Matched sequence	*PRKWNK1*	0.26	Ki, Sk, Bm	Not Skipped	Mouse annotation unclear

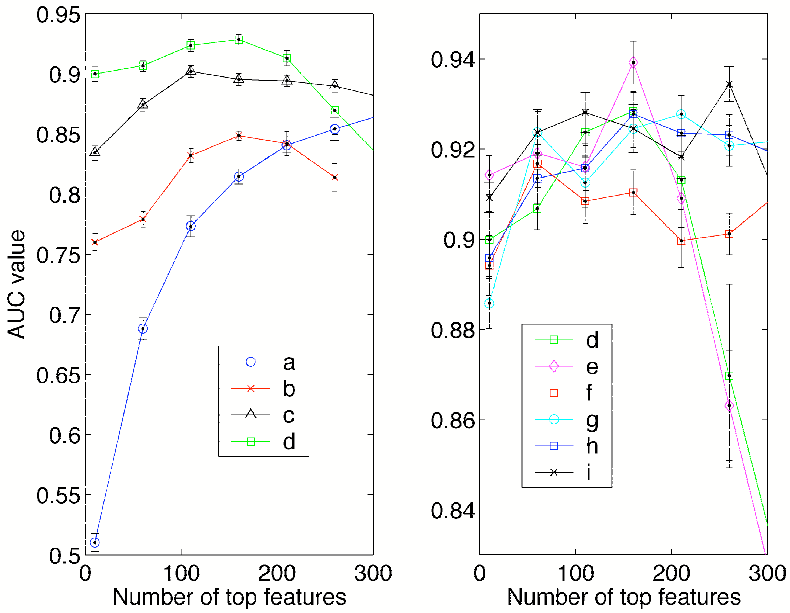
					in intron				
S23	136531	3	92	026992	6	*Scn3a*	0.10	Pr, Te, Ki, FBr	Not Skipped	
S24	153944	2	41	034017	2	*MSI2*	1.75	Not Skipped	Not Skipped	
S25	121964	5	92	036890	9	NM_024659	1.30	Not Skipped	Not Skipped	
S26	155966	19	53	031189	5	*FMR2*	1.06	Not Skipped	Not Skipped	
S27	182197	3	108	038616	4	*EXT1*	0.69	Not Skipped	Not Skipped	mRNA AK130054 (Lu): Exons 2-6 skipped
S28	079739	7	116	025791	8	*PGM1*	0.36	Not Skipped	Not Skipped	
S29	164692	33	108	029661	37	*COL1A2*	0.14	Not Skipped	Not Skipped	
S30	117676	11	89	003644	11	*RPS6KA1*	0.14	Not Skipped	Not Skipped	
C1	69188	22	202	041592	16	*SDK2*	-0.27	Not Skipped	Not Skipped	
C2	73670	22	120	020926	22	*ADAM11*	-0.81	Not Skipped	Not Skipped	
C3	183773	6	97	022763	6	*NM_144704*	-1.41	Not Skipped	Not Skipped	
C4	143761	3	111	020440	3	*ARF1*	-1.62	Not Skipped	Not Skipped	
C5	146904	4	403	029859	5	*EPHA1*	-2.01	Not Skipped	Not Skipped	
C6	140859	6	112	031788	6	*KIFC3*	-0.99	Not Skipped	Not Skipped	
C7	018236	13	176	000107	14	*CNTN1*	-0.71	Not Skipped	Not Skipped	
C8	008405	12	104	020038	12	*CRY1*	-1.15	Not Skipped	Not Skipped	
C9	178035	8	91	006666	8	*IMPDH2*	-1.15	Not Skipped	Not Skipped	
C10	164070	7	245	025757	7	*OS94_HUMAN*	-1.66	Not Skipped	Not Skipped	
C11	130812	5	271	038742	6	*ANGPTL6*	-1.55	Not Skipped	Not Skipped	
E1	166164	10	108	031660	10	*BRD7*	-0.98	EST/mRNA evidence	Not Skipped	BX377621 (placenta cot 25-normalized) BI860548 (mammary adenocarcinoma cell line) AV747130 (Adult Pituitary)
E2	132849	32	84	028562	22	*NM_005799*	-0.66	EST/mRNA evidence	Not Skipped	mRNA AJ224748 (HeLa) Exon is not seen in any other mRNAs (5) or ESTs (2)
E3	100395	3	134	022394	3	*L3MBTL2*	-0.86	EST/mRNA evidence	Not Skipped	BU171558 (melanotic melanoma) Exon is included in all other mRNAs (8) and ESTs (>40)
E4	107897	3	121	026781	2	*ACBD5*	-0.73	EST/mRNA evidence	Br, Ki, Emb	ESTs (neuroblastoma, BM011542,

										BM011474); mRNAs (BC025309, neuroblastoma; AB082527, brain)
E5	120992	6	74	025903	6	*LYPLA1*	-0.65	EST/mRNA evidence	Not Skipped	BQ434220 (embryonic carcinoma), H04075 (placenta at birth), BE246387 (leukopheresis)
E6	076513	13	161	041870	13	*ANKRD13*	-1.24	EST/mRNA evidence	Not Skipped	AK095130 (substantia nigra)
E7	103876	12	102	030630	12	*FAH*	-1.26	EST/mRNA evidence	Not Skipped	S63549 (tyrosinemia patients), Exon included in 30 ESTs
E8	120137	3	177	018846	4	*PANK3*	-1.03	EST/mRNA evidence	Not Skipped	U46305 (Pancreatic cancer)
E9	124198	35	131	027682	35	*ARFGEF2*	-1.04	EST/mRNA evidence	Not Skipped	BU173319 (Retinablastoma)
E10	106443	5	160	029629	6	*PHF14*	-1.05	EST/mRNA evidence	Not Skipped	BF816107 (adenocarcinoma)
E11	170248	10	178	032504	10	*PDCD6IP*	-1.56	EST/mRNA evidence	Not Skipped	BQ961575 (leiomyosarcoma)
E12	107937	3	104	021149	3	*GTPBP4*	-1.58	EST/mRNA evidence	Not Skipped	AK097093 (spleen, skips exon but retains downstream intron)
E13	134899	12	145	026048	12	*ERCC5*	-1.59	EST/mRNA evidence	Not Skipped	AA191090 (hNT neuroteratocarcinoma neurons)
E14	150753	3	165	022234	3	*CCT5*	-1.83	EST/mRNA evidence	Not Skipped	AU143554 (retinoblastoma)
E15	132170	6	451	000440	7	*PPARG*	-1.54	EST/mRNA evidence	Br, Ki, Li, Emb, Te	BI524664 (NIH_MGC_122), CA426975 (subchondral bone)

Table S4. Candidate ACEs verified by RT-PCR experiments and subsequent sequencing, or by literature and transcript-based evidence. Samples S1 to S30 are candidate ACEs, samples C1 to C5 are low scoring ACEs (scored <0), and samples E1 to E15 are low scoring ACEs with human EST/mRNA evidence of skipping. (Abbreviations for tissues are as follows. Ce: cerebellum, Br: whole brain, FBr: fetal brain, Sk: skeletal muscle, Li: liver, Lu: lung, He: heart, Bm: bone marrow, Te: testis, Pr: prostate, Ki: kidney, FLi: fetal liver, Emb: embryonic mix).

Fig. S1.

A.



B.

Model index	A	b	c	d	e	f	g	h	i
Core features	x	√	√	√	√	√	√	√	√
Upstream intron alignment score	x	x	x	√	√	√	x	√	√
Downstream intron alignment score	x	x	x	√	√	√	√	x	√
Exon alignment score	x	x	√	x	√	x	√	√	√
Upstream intron alignment features	x	x	x	x	x	√	x	√	√
Downstream intron alignment features	x	x	x	x	x	√	√	x	√
Exon alignment features	x	x	√	x	x	x	√	√	√
Flanking intron features	√	√	x	√	√	x	x	x	√
Exon features	√	√	x	√	√	x	x	x	√

Fig. S2.

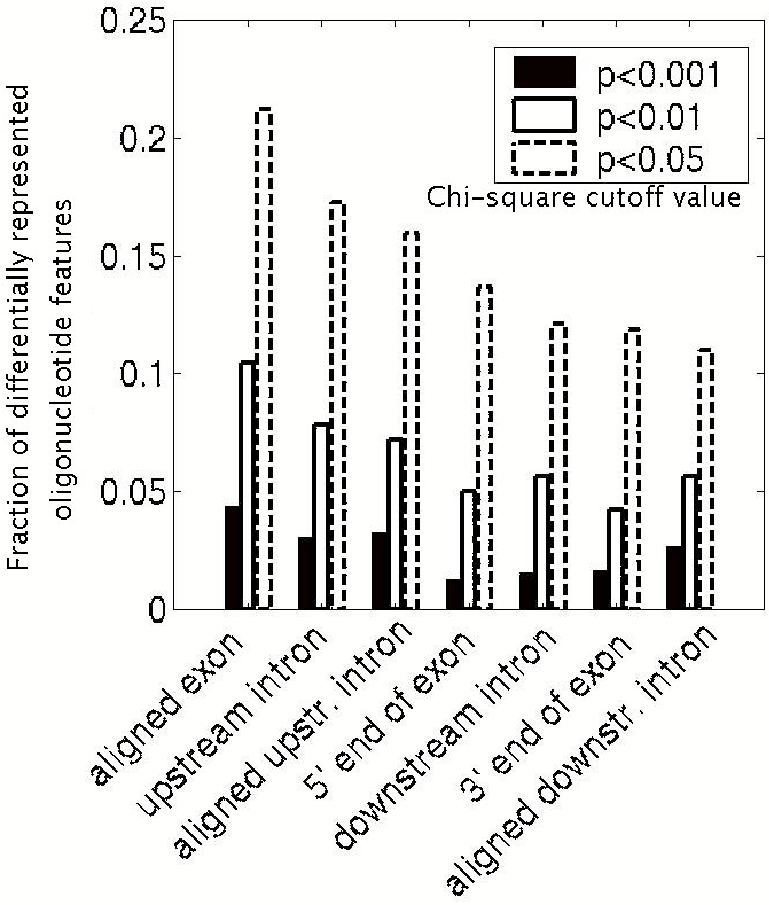
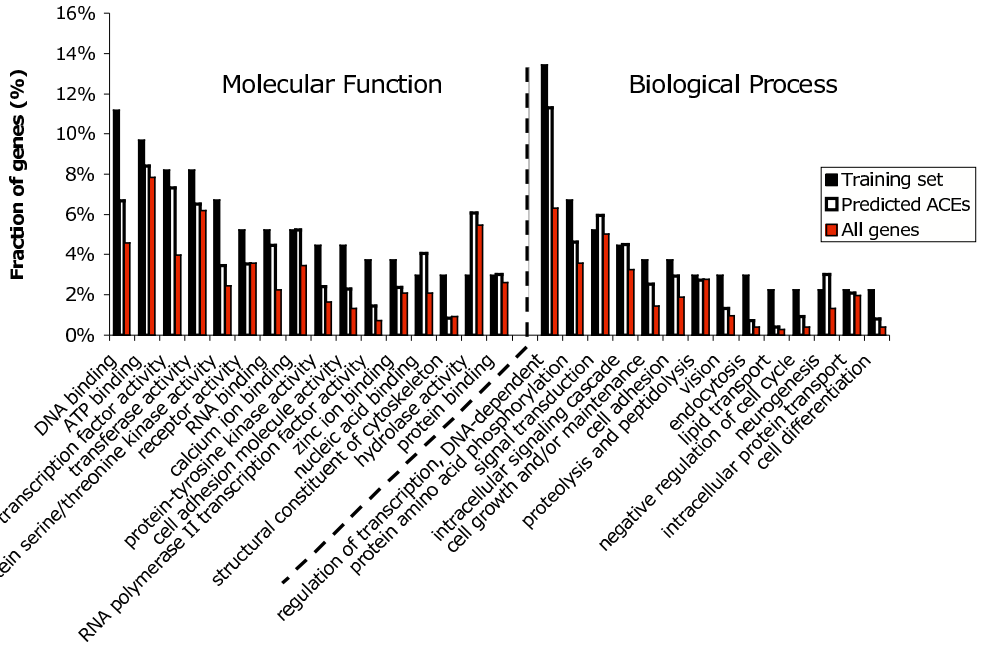

0.25
0.2
0.15
0.1
0.05
0
Fraction of differentially represented oligonucleotide features
p<0.001
p<0.01
p<0.05
Chi-square cutoff value
aligned exon
upstream intron
aligned upstr. intron
5' end of exon
downstream intron
3' end of exon
aligned downstr. intron

Fig S3.


Fraction of genes (%)
16%
14%
12%
10%
8%
6%
4%
2%
0%
Molecular Function
Biological Process
Training set
Predicted ACEs
All genes
DNA binding
ATP binding
transcription factor activity
transferase activity
protein serine/threonine kinase activity
receptor activity
RNA binding
calcium ion binding
protein-tyrosine kinase activity
cell adhesion molecule activity
RNA polymerase II transcription factor activity
zinc ion binding
nucleic acid binding
structural constituent of cytoskeleton
hydrolase activity
protein binding
regulation of transcription, DNA-dependent
protein amino acid phosphorylation
intracellular signaling cascade
signal transduction
cell growth and/or maintenance
cell adhesion
proteolysis and peptidolysis
vision
endocytosis
lipid transport
negative regulation of cell cycle
neurogenesis
intracellular protein transport
cell differentiation

References to Supplementary

Blanchette, M., and Chabot, B. (1997). A highly stable duplex structure sequesters the 5' splice site region of hnRNP A1 alternative exon 7B. Rna *3*, 405-419.
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.
Kriaucionis, S., and Bird, A. (2004). The major form of MeCP2 has a novel N-terminus generated by alternative splicing. Nucleic Acids Res *32*, 1818-1823.
Shibata, H., Huynh, D. P., and Pulst, S. M. (2000). A novel protein with RNA-binding motifs interacts with ataxin-2. Hum Mol Genet *9*, 1303-1313.
Wagner, B. J., DeMaria, C. T., Sun, Y., Wilson, G. M., and Brewer, G. (1998). Structure and genomic organization of the human AUF1 gene: alternative pre-mRNA splicing generates four protein isoforms. Genomics *48*, 195-202.
Wang, B., Williams, H., Du, J. S., Terrett, J., and Kenwrick, S. (1998). Alternative Splicing of Human NrCAM in Neural and Nonneural Tissues. Mol Cell Neurosci *10*, 287-295.
Wollerton, M. C., Gooding, C., Wagner, E. J., Garcia-Blanco, M. A., and Smith, C. W. (2004). Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. Mol Cell *13*, 91-100.