ARTICLES

Deciphering the splicing code

Yoseph Barash[1,2]*, John A. Calarco[2]*, Weijun Gao[1], Qun Pan[2], Xinchen Wang[1,2], Ofer Shai[1], Benjamin J. Blencowe[2] & Brendan J. Frey[1,2,3]

Alternative splicing has a crucial role in the generation of biological complexity, and its misregulation is often involved in human disease. Here we describe the assembly of a 'splicing code', which uses combinations of hundreds of RNA features to predict tissue-dependent changes in alternative splicing for thousands of exons. The code determines new classes of splicing patterns, identifies distinct regulatory programs in different tissues, and identifies mutation-verified regulatory sequences. Widespread regulatory strategies are revealed, including the use of unexpectedly large combinations of features, the establishment of low exon inclusion levels that are overcome by features in specific tissues, the appearance of features deeper into introns than previously appreciated, and the modulation of splice variant levels by transcript structure characteristics. The code detected a class of exons whose inclusion silences expression in adult tissues by activating nonsense-mediated messenger RNA decay, but whose exclusion promotes expression during embryogenesis. The code facilitates the discovery and detailed characterization of regulated alternative splicing events on a genome-wide scale.

Transcripts from approximately 95% of multi-exon human genes are spliced in more than one way, and in most cases the resulting transcripts are variably expressed between different cell and tissue types[1,2]. This process of alternative splicing shapes how genetic information controls numerous critical cellular processes, and it is estimated that 15% to 50% of human disease mutations affect splice site selection[3].

Tissue-dependent splicing is regulated by *trans*-acting factors, *cis*-acting RNA sequence motifs, and other RNA features, such as exon length and secondary structure. For nearly two decades, researchers have sought to define a regulatory splicing code in the form of a set of RNA features that can account for abundances of spliced isoforms[4–8]. Through detailed investigation of a small number of examples of regulated splicing[9], it is clear that a splicing code must account for various features that act together to control splicing. Furthermore, a code should enable the reliable prediction of the regulatory properties of previously uncharacterized exons and the effects of mutations within regulatory elements.

Here we describe a method for inferring a splicing regulatory code that addresses these challenges (Fig. 1a). We evaluate the code using a variety of criteria, describe and verify predictions made by the code, and demonstrate the usefulness of the code in scientific exploration.

Isoform quantification and RNA features

Our method takes as an input a collection of exons and surrounding intron sequences and data profiling how those exons are spliced in different tissues. The method assembles a code that can predict how a transcript will be spliced in different tissues.

We used data profiling 3,665 cassette-type alternative exons across 27 diverse mouse tissues, including whole-embryo stages and a variety of adult tissues[10]. For each exon and each tissue, this data set provides a percentage inclusion value, which is an estimate of the fraction of transcripts that include the exon[11]. Tissues were grouped to form four tissue types: central nervous system (CNS) tissues, muscle tissues, digestive system tissues, and whole embryos, with embryonic stem cells added to the latter group (Supplementary Information 1 and Fig. 1). To correct for tissue-independent baseline exon inclusion levels and measurement noise, we converted the percentage inclusion value for each data point (exon and tissue type) to three probabilities, q^{inc}, q^{exc} and q^{nc}, of increased exon inclusion ('inc'), increased exon exclusion, that is, skipping ('exc'), and no change ('nc'). q denotes the set of three probabilities and we refer to it as a 'splicing pattern'. Of all exons exhibiting a high probability of increased inclusion ($q^{\mathrm{inc}} \geq 0.99$) or exclusion ($q^{\mathrm{exc}} \geq 0.99$), 51%, 23%, 32% and 25% showed changes in CNS, muscle, embryonic and digestive tissues, respectively, whereas 24% were regulated in more than one tissue type.

Code assembly requires a set of relevant features derived from exonic and intronic sequences. We constructed a compendium of 1,014 features of four types: known motifs, new motifs, short motifs and features describing transcript structure (see later and Supplementary Information 2). Motif features correspond to consensus sequences, sequence clusters, or position-specific score matrices, and may be associated with specific RNA regions (Fig. 1a).

A literature survey yielded 171 'known motifs' found near tissue-regulated exons[10,12–14] or associated with splicing factor binding preferences, including [U]GCAUG (bound by A2bp1 and A2bp2, referred to here as Fox proteins[15,16]), YCAY clusters (bound by Nova1 and Nova2, referred to as Nova[17,18]), CU-rich sequences (bound by Ptbp1 and its neuronal variant Ptbp2, referred to as PTB and nPTB[19,20]), YGCUYK-like CUG- and UG-rich sequences (bound by Mbnl, Cugbp and related CELF-like factors[21]), ACUAAY (bound by Quaking-like Qk and Star[22] factors, later referred to as Qkl), U-rich sequences (bound by Tia1 and Tial1, later referred to as Tia1/Tiar), and elements associated with serine/arginine-rich (SR) and heterogenous nuclear ribonucleoprotein factors. As interspecies conservation of intronic sequences is associated with alternative splicing[1,12,23], we included interval-averaged conservation levels as features and also region-specific motif scores weighted by conservation. Note that although a motif may be known, its regulatory activity in the context of other features is usually not.

The compendium includes 326 'new motifs' that have weak or no known evidence for roles in tissue-dependent splicing, including 12 clusters of validated or putative exonic and intronic splicing enhancers

[1]Biomedical Engineering, Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto M5S 3G4, Canada. [2]Banting and Best Department of Medical Research and Department of Molecular Genetics, Donnelly Centre, University of Toronto, 160 College Street, Toronto M5S 3E1, Canada. [3]Microsoft Research, 7 J. J. Thomson Avenue, Cambridge CB3 0FB, UK.
*These authors contributed equally to this work.

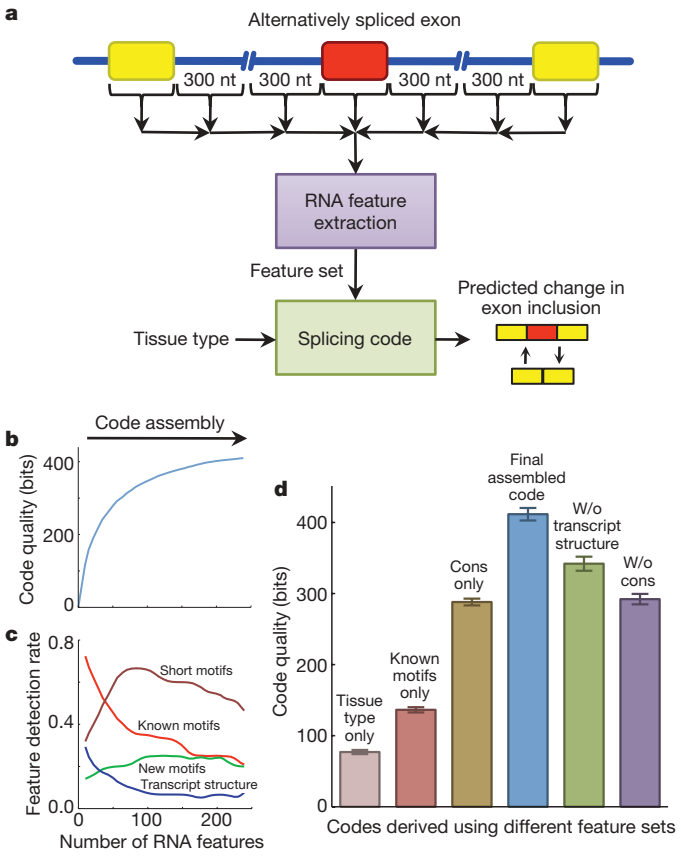

Figure 1 | Assembling the splicing code. a, The code extracts hundreds of RNA features (known/new/short motifs and transcript structure features) from any exon of interest (red), its neighbouring exons (yellow) and intervening introns (blue). It then predicts whether or not the exon is alternatively spliced, and if so, whether the exon's inclusion level will increase or decrease in a given tissue, relative to others. **b**, **c**, Code assembly proceeds by recursively adding features to maximize an information measure of code quality (**b**), and different feature types are preferred at different stages of assembly (**c**). **d**, The final assembled code achieves higher code quality than simpler codes derived using previously reported features and feature subsets. Cons, conservation; w/o, without. Error bars represent 1 s.d.

(ESEs and ISEs) and silencers (ESSs and ISSs), which are 6–8 nucleotides long and were identified without regard to possible tissue-dependent roles[24–26], and 314 5–7-nucleotide-long motifs that are conserved in intronic sequences neighbouring alternative exons[27]. There are also 460 region-specific counts of 1–3-nucleotide 'short motifs', because such features were previously associated with alternative splicing[28]. We included 57 'transcript structure' features implicated in determining spliced transcript levels, such as exon/intron lengths, regional probabilities of secondary structures[29], and whether exon inclusion/exclusion introduces a premature termination codon (PTC).

In addition to the feature compendium, we constructed a set of ~1,800 'unbiased motifs' by performing a *de novo* search[10] for each tissue type and direction of splicing change (Supplementary Information 3). Later, we report results obtained with and without using these features.

Assembling a high-information code

Our method seeks a code that is able to predict the splicing patterns of all exons as accurately as possible, based solely on the tissue type and proximal RNA features. The putative features for a particular exon are appended to make a feature vector r, and the corresponding prediction in tissue type c is denoted $p(c,r)$. Like q, $p(c,r)$ consists of probabilities of increased inclusion or exclusion, or no change. The code is combinatorial and accounts for how features cooperate or compete in a given tissue type, by specifying a subset of important features, thresholds on feature values and softmax parameters[30] relating active feature combinations to the prediction $p(c,r)$ (Supplementary Information 4).

We use a measure of 'code quality' that is based on information theory[31] (see Methods). It can be viewed as the amount of information about genome-wide tissue-dependent splicing accounted for by the code. A code quality of zero indicates that the predictions are no better than guessing, whereas a higher code quality indicates improved prediction capability.

To assemble a code, our method recursively selects features from the compendium, while optimizing their thresholds and softmax parameters to maximize code quality (Supplementary Information 5). The code quality increased monotonically during assembly, but diminished gains were observed after 200 features were included (Fig. 1b, c, based on fivefold cross-validation). The final assembled code contained ~200 features. When a code was assembled using the compendium plus the unbiased motifs, the increase in code quality did not exceed 1 s.d. in error (data not shown), but, interestingly, some of the unbiased motifs that did not correspond to any compendium features were selected and subsequently experimentally verified (see later).

To quantify the contributions of its different components, we compared our final assembled code to partial codes whose only inputs were the tissue type, previously described motifs, conservation levels, or the compendium with transcript structure features or conservation levels removed (Fig. 1d).

Predicting alternative splicing

On the task of distinguishing alternatively spliced exons from constitutively spliced exons, our method achieves a true positive rate of more than 60% at a false positive rate of 1% (Supplementary Information 6). To address the more difficult challenge of predicting tissue-dependent regulation, we applied the code to various sets of unique test exons (exons not similar to those used during code assembly) and verified the predictions using microarray data, PCR with reverse transcription (RT–PCR) and focused studies (see later and Supplementary Information 5).

We first asked whether the theoretical ranking of the different codes shown in Fig. 1d corresponds well to their relative abilities to predict microarray-assessed tissue-dependent regulation (see Methods). Indeed, the final assembled code achieved significantly higher accuracy than the partial codes (Fig. 2a). For exons in genes with median expression in the top 20th percentile, at a false positive rate of 1%, a true positive rate of 21% was achieved, and this rose to 51% for a false positive rate of 10%.

We next asked how well the splicing code predicts significant differences in the percentage exon inclusion between pairs of tissues, for cases where the predicted difference is large (Fig. 2b and Supplementary Fig. 12). For microarray data, the splicing code correctly predicted the direction of change (positive or negative) in 82.4% of cases ($P < 1 \times 10^{-30}$, Binomial test; see Methods). For RT–PCR evaluation, 14 exons that the splicing code predicted would exhibit significant tissue-dependent splicing were profiled in 14 diverse tissues. The splicing code correctly predicted the direction of change in 93.3% of cases ($P < 1 \times 10^{-10}$, Binomial test). A scatterplot comparing predictions and measurements (Fig. 2c) illustrates that the code is able to predict an exon's direction of regulation better than its percentage inclusion level. Figure 2d shows RT–PCR data and predictions for four representative exons.

To assess whether the code recapitulates results from experimental studies of individual exons and tissue-specific splicing factors, we surveyed 97 CNS- and/or muscle-regulated exons targeted by Nova, Fox, PTB, nPTB and/or unknown factors[18,19,32–39]. For each test exon, we extracted its features, applied the code and examined whether or not it correctly predicts splicing patterns in CNS or muscle tissues (Supplementary Table 3). The code's predictions were correct for 74% of the combined set of 97 exons ($P < 1 \times 10^{-41}$, Bernoulli test), 65%

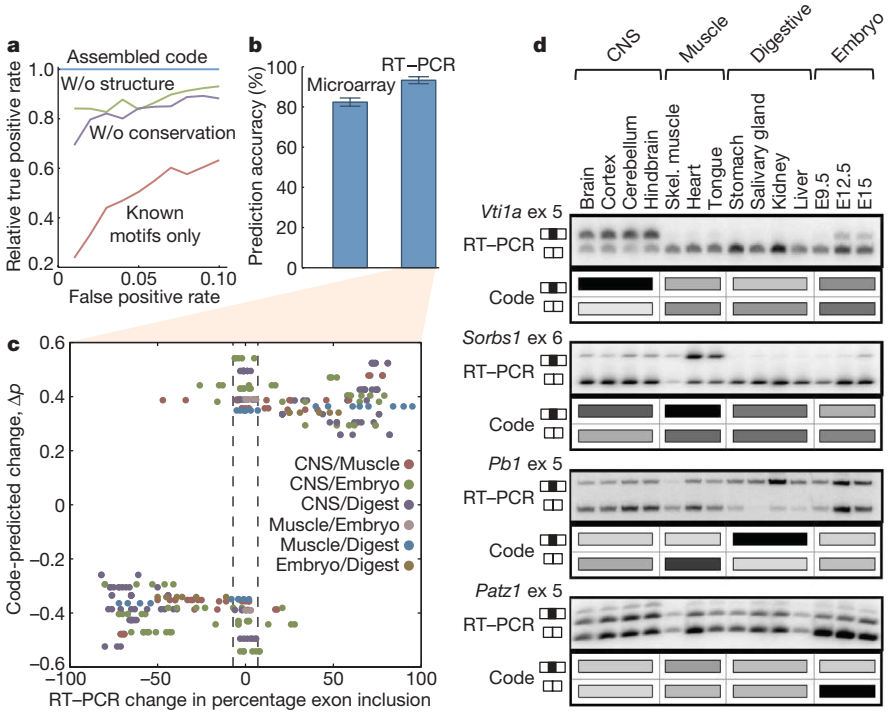

Figure 2 | Predicting tissue-regulated alternative splicing. **a**, Classification rates for the final assembled code and simpler codes, assessed using microarray data ($n = 28{,}920$). **b**, Accuracy of the code in predicting microarray- and RT–PCR-measured changes in exon inclusion levels between pairs of tissues ($n = 346$ and $n = 208$). Error bars represent 1 s.d. **c**, For each exon and pair of tissues, the RT–PCR-measured change in the percentage inclusion is plotted against the code-predicted change in the probability of exon inclusion. Dashed lines indicate RT–PCR differences exceeding 1 s.d. in measurement error. **d**, RT–PCR data for four exons, plus code predictions indicating relative increases (dark shading) or decreases (light shading) in the exon inclusion level.

of the Nova targets ($P < 1 \times 10^{-20}$), 95% of the Fox targets ($P < 1 \times 10^{-15}$) and 91% of the PTB/nPTB targets ($P < 1 \times 10^{-8}$).

To our knowledge, this is the first time tissue-dependent splicing changes have been predicted from sequence information alone and the prediction accuracy has been quantitatively evaluated.

Interpretation of the splicing code

Figure 3a shows components of the code that have strongest regulatory evidence (also see Supplementary Information 7–9). The consensus sequences for motif features are accompanied (in parentheses) by names of potential binding proteins, but it should be kept in mind that a different or unknown factor could bind instead. The direction of a feature's regulatory activity (increased inclusion or exclusion) is indicated by colour (red or blue), and if a feature has an effect in both directions (for example, because it works in combination with another factor) both colours are shown. Short motifs are not included, but are shown in Supplementary Fig. 9.

The complexity of the code is reflected by the number of tissue-specific features per exon, the median of which varies from 12 (CNS) to 19 (embryo) when excluding short motifs (Supplementary Fig. 10). The code reveals tissue-specific combinations of features that are potentially synergistic (the number of features must exceed a threshold for regulation) or antagonistic (the direction of the regulatory effects of two features is opposite). Other features are associated with several tissues or are predicted to act in a tissue-independent manner. Many aspects of the code compare well with known results, whereas others are new, and others challenge known results, as explained later.

Position- and tissue-specific effects of elements that match the known binding motifs for Fox, Nova, Mbnl, Cugbp, Tia1/Tiar, PTB, nPTB and Qkl proteins are mostly consistent with previous results, but interesting differences arise. The code predicts regulatory elements that are deeper into introns than previously appreciated; PTB/nPTB-like CU-rich elements were found to often reside 250 to 300 nucleotides upstream of CNS-regulated exons, which is considerably farther than previously reported[10,14] (Fig. 4 and later). Mbnl sites were found mostly in upstream introns of exons upregulated in CNS, but were also found in the downstream introns of muscle-regulated exons. Consistent with previous computational analyses[17] and *in vivo* cross-linking and immunoprecipitation (CLIP) assays[18], in adult CNS tissues, Nova elements in upstream introns and alternative exons were associated with exon exclusion, whereas elements in the 5′ region of downstream introns were primarily associated with exon inclusion. However, contrasting effects were observed in embryonic tissues (Fig. 3a), where Nova elements in the 5′ region of downstream introns were primarily associated with exon exclusion. Although the position of the first AG upstream of the alternative exon was previously associated with alternative splicing[40], our code associates it with tissue-specific (predominantly CNS) regulation. The motif ACUAAC was previously associated with Qkl factors and reported as enriched downstream of exons upregulated in muscle[1,13]. Our code identifies this feature, but also predicts that its presence in the upstream intron regulates CNS-specific splicing, which is consistent with studies implicating its *trans*-factor as a regulator in human brain tissue[41].

To determine interactions between regulatory features, we identified pairs of features that are unexpectedly frequent in the code and generated feature interaction networks (Fig. 3b–d and Supplementary Information 8). Although some combinations arise primarily from feature similarity (for example, Mbnl and Cugbp binding sites), others correspond to bona fide mechanisms, some of which have been verified. Figure 3b shows that nPTB, Mbnl and Cugbp binding sites jointly occur in the 3′ region of introns upstream of exons upregulated in CNS tissues; later, this interaction is examined using mutated minigene reporters. The combination of nPTB binding sites in upstream introns, nPTB binding sites in downstream introns, and short alternative exons shows the general utility of a previously proposed mechanism in which PTB facilitates RNA looping resulting in exon exclusion[42]. Our code indicates that this mechanism may be disabled in CNS tissues, causing increased exon inclusion.

The code shows that combinations of several features act more frequently than previously appreciated. In a few isolated cases, short exons, weak splice sites and low ESE counts were previously found to result in an 'exclusion by default' mechanism that can be reversed by other features[9]. As seen in Fig. 3, short exons, weak 3′ splice sites and low ESE counts are frequently associated with CNS-specific exon inclusion. Over six times more CNS-regulated exons have low values (lowest 20th percentile) for these features than non-CNS-regulated exons ($P < 1 \times 10^{-8}$, Binomial test). The code also reveals elements near flanking exons (for example, ESEs and ESSs) that participate in regulation.

Transcript structure features were found to have strong effects in the code, with interesting and potentially important biological implications.

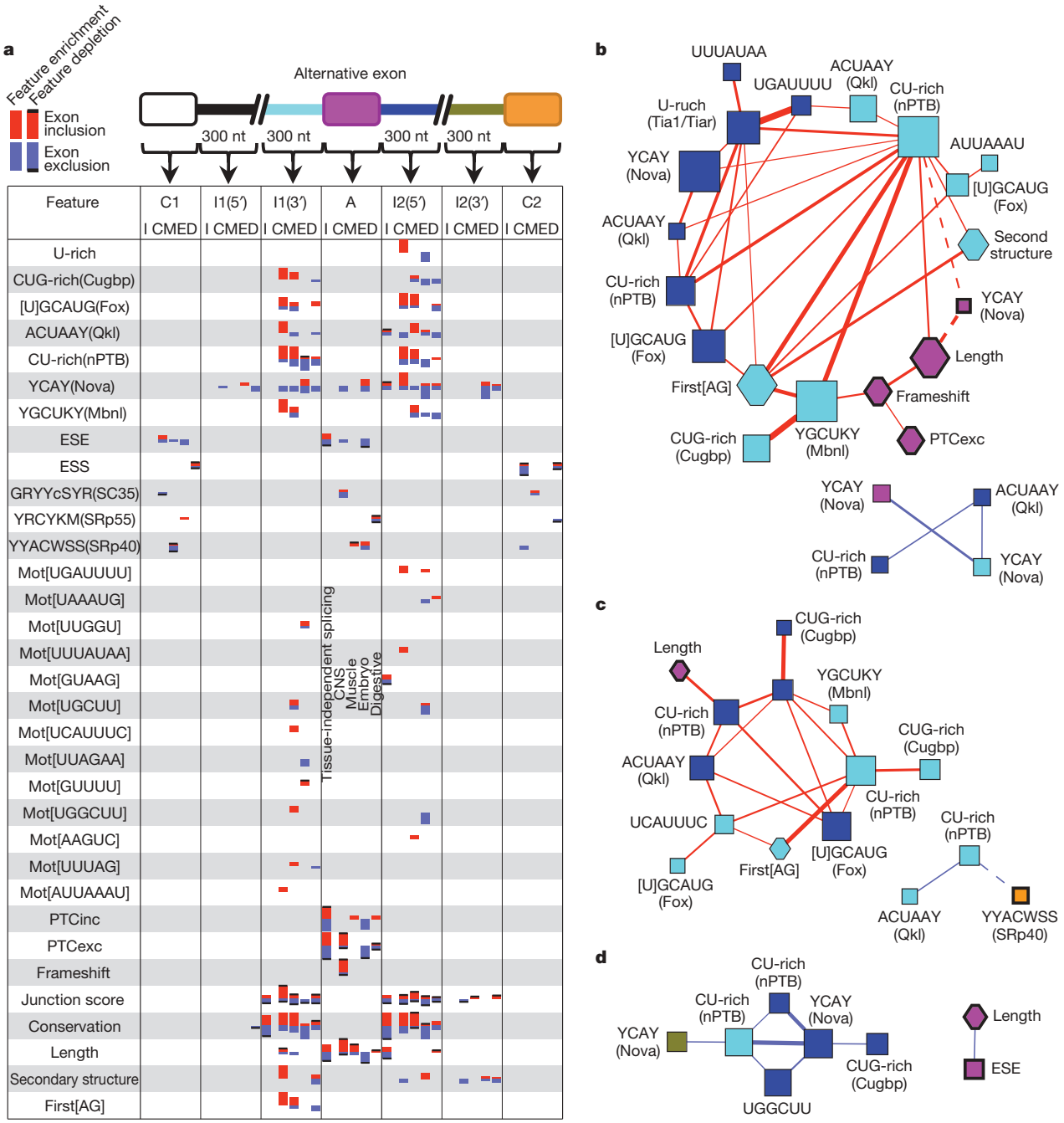

Figure 3 | Graphical depiction of the splicing code. **a**, The region-specific activity of each feature in increased exon inclusion (red bar) or exclusion (blue bar) is shown for CNS (C), muscle (M), embryo (E) and digestive (D) tissues, plus a tissue-independent mixture (I). A bar with/without a black hat indicates activity due to feature depletion/enrichment. Bar size conveys enrichment *P*-value; $P < 0.005$ in all cases. Potential feature binding proteins are shown in parentheses. **b**–**d**, Unexpectedly frequent feature pairs were identified and used to generate feature interaction networks for CNS (**b**), muscle (**c**) and embryonic (**d**) tissues. Node size and colour indicate the feature's *P*-value and region (see colour key in **a**). Red/blue edges correspond to increased inclusion/exclusion and edge thickness conveys interaction *P*-value (false discovery rate-corrected Fisher test); $P < 0.05$ in all cases. A thick/thin node boundary indicates activity due to feature depletion/enrichment.

For example, the code predicts a class of exons that insert a PTC after inclusion in transcripts and that are skipped in embryonic tissues but included in adult tissues. Later, we describe experiments indicating that these exons have an important role in the regulation of developmental stage-specific gene expression.

Predicting regulatory feature maps

By flagging regulatory elements in RNA sequences surrounding an alternative exon, the splicing code yields a visual feature map that partially accounts for how the exon is regulated. Predicted feature maps were initially evaluated by their overlap with 376 nucleotides of RNA sequence analysed by mutagenesis in more than 60 splicing reporter constructs from *Agrn*[33], *Src*[19,43], *Casp2* (ref. 35) and the *Slo* K^+ STREX exon[44]. Our feature maps (Supplementary Figs 2–7 and Supplementary Information 10) achieve an overlap of 90% with a statistical significance of $P < 0.002$ (empirical, using maps from unrelated exons). In contrast, feature maps constructed using only known motifs achieve an overlap of 38% ($P = 0.004$) and maps derived solely from conservation information[27] achieve poor specificity ($P = 0.27$).

Code-generated feature maps can be used to guide focused mechanistic studies. We examined exon 16 of the *Daam1* gene, which our

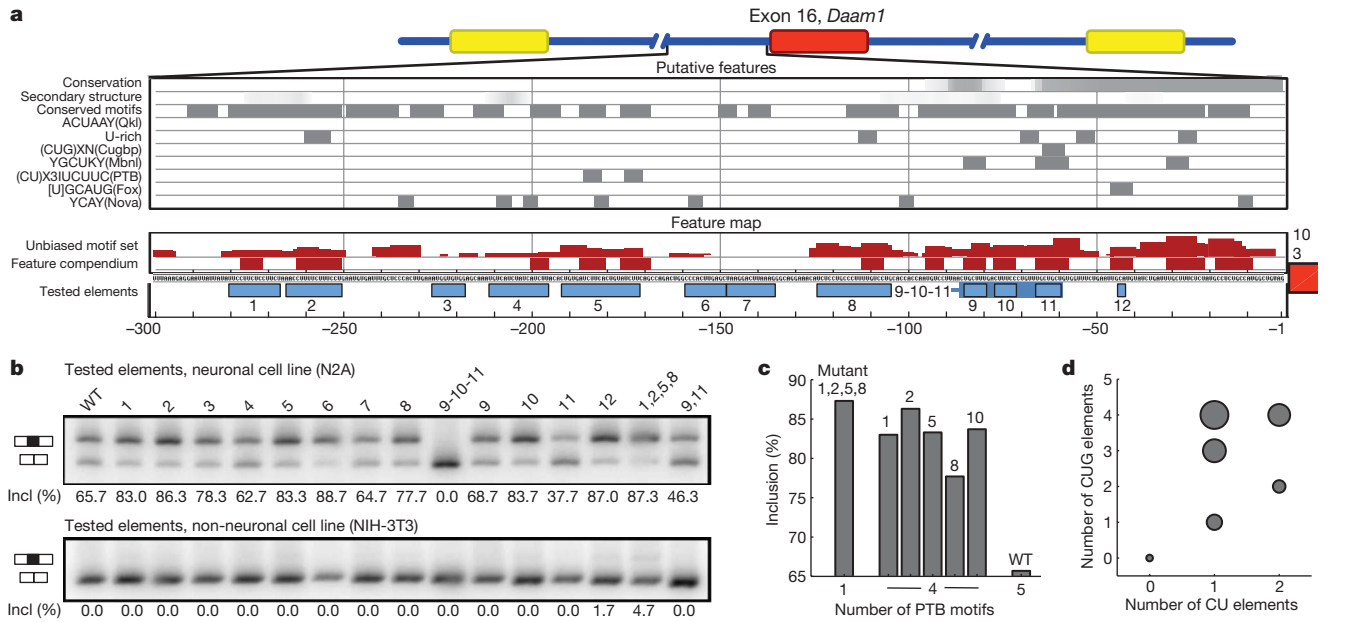

Figure 4 | Validation of a regulatory feature map. Regulatory elements in the intron upstream of exon 16 in *Daam1* predicted to be associated with CNS-specific increased exon inclusion. **a**, Putative features (grey blocks), along with code-selected features from the compendium and the unbiased motif set (red blocks). Twelve segments were selected for testing (blue blocks), including one not overlapping with predictions (7), and 15 minigene reporters with single- or combined-segment substitutions were constructed and transfected into neuroblastoma (N2A) and epithelial (NIH-3T3) cells. **b**, RT–PCR results for the wild type and 15 mutants. **c**, Mutations of several nPTB-like elements support code-predicted synergistic interactions. **d**, Mutations of several CU and CUG elements between 55 and 90 nucleotides support code-predicted antagonistic interactions. Symbol size indicates the percentage exon inclusion (0–83.7%).

code ranked among the top 0.3% for CNS-specific increased inclusion. It was recently shown[39] that this exon is specifically included in CNS tissues, but the precise locations of elements that mediate neural-specific splicing of the exon were not determined. In our code, most features were found within 300 nucleotides of upstream intron sequence and the corresponding feature map (red blocks in Fig. 4a) yields the following new predictions: an unusually high density of regulatory elements (only 7.8% of other exons in our data set had as high a density); novel motifs GGAGC (215–219 nucleotides) and CUGGC (159–163 nucleotides); three well-separated regions (72–78, 106–124 and 267–280 nucleotides) that resemble nPTB binding elements, but do not score well using known motif definitions for this protein[19,42,45]; and predicted inactivity of several features derived only from conservation[27] (137–142, 146–150 and 284–290 nucleotides).

We tested the code-generated feature map by performing extensive mutagenesis using a *Daam1* exon 16 precursor mRNA reporter[39] that faithfully recapitulates endogenous splicing patterns. Fifteen mutant minigene reporters were constructed by replacing segments comprising a total of 150 nucleotides of intron sequence with random sequences pre-filtered to avoid introducing regulatory features (Fig. 4a, blue blocks, Supplementary Information 12 and Supplementary Table 5). Semi-quantitative RT–PCR assays were used to estimate the percentage inclusion of exon 16 in transcripts from the wild-type and mutant reporters, after their transfection into mouse neuroblastoma (N2A) cells and non-neural fibroblast (NIH-3T3) and myoblast (C2C12) cell lines (Fig. 4b and data not shown). Of the 15 mutants, 14 contain segments overlapping with predicted elements and indeed the percentage inclusion in N2A cells for all of these mutants was significantly different from wild type ($P < 0.005$, normal test, s.d. of 0.8% estimated from three transfections; Supplementary Information 12). Mutant 7 was designed to test a region predicted to have no regulatory role and, indeed, its percentage inclusion is within 1 s.d. of the wild-type value. Although the code could not confidently predict the direction of percentage inclusion change, it predicted that all mutants would have higher exon inclusion in CNS tissues relative to other tissues, and this prediction was confirmed for 14 out of 15 mutants by comparison of the results from the N2A and NIH3T3 cell lines (Fig. 4b and Supplementary Information 12).

Disruption of the two new motifs in mutants 3 and 6 resulted in changes in inclusion levels ($P < 1 \times 10^{-100}$, normal test). Mutants 1, 8 and 10 disrupted the three new CU-rich nPTB-like motifs and showed increased exon inclusion ($P < 1 \times 10^{-100}$). *In vivo* mouse whole brain CLIP and high-throughput sequencing indicates that nPTB indeed binds sequences overlapping these CU-rich elements (D. Licalatosi and R. Darnell, personal communication). Mutants 2 and 5 disrupted elements that correspond to previously defined nPTB motifs[19,42,45]. As expected, these mutants showed increased exon inclusion ($P < 1 \times 10^{-100}$), but to a lesser extent than for the new CU-rich elements. The code used conservation to predict and reject functional elements. Mutant 10 disrupted an element that the code identified using conservation plus CU-richness (72–78 nucleotides), and confirmed that this element is functional (see earlier). In contrast, mutant 7 disrupted a region that overlapped two conserved elements, but which the code predicted would not be functional. Indeed, no significant change in the percentage inclusion was observed.

According to the code, the number of nPTB motifs must exceed a threshold before CNS regulation occurs. To investigate interactions between nPTB motifs (disrupted by mutants 1, 2, 5, 8 and 10), in Fig. 4c we plot the percentage inclusion for these mutants, a combination mutant (segments 1, 2, 5 and 8), and wild type. Although the presence of four nPTB motifs slightly suppresses inclusion compared to only one, greater suppression occurs with five nPTB motifs, indicating a synergistic interaction. The code also predicted that CUG-rich Cugbp/Mbnl-like elements close to the nPTB element in mutant 10 would enhance inclusion. To explore combinations of these elements, we counted the number of nPTB-like CU elements and the number of Cugbp/Mbnl-like CUG elements from 55 to 90 nucleotides, and repeated this procedure for mutants 9, 10 and 11, a combination of mutants 9 and 11 (labelled 9,11), and a mutant (labelled 9-10-11) that disrupted all elements in this region, including the CUG motif between regions 10 and 11. Consistent with the code, these elements were found to interact antagonistically (Fig. 4d).

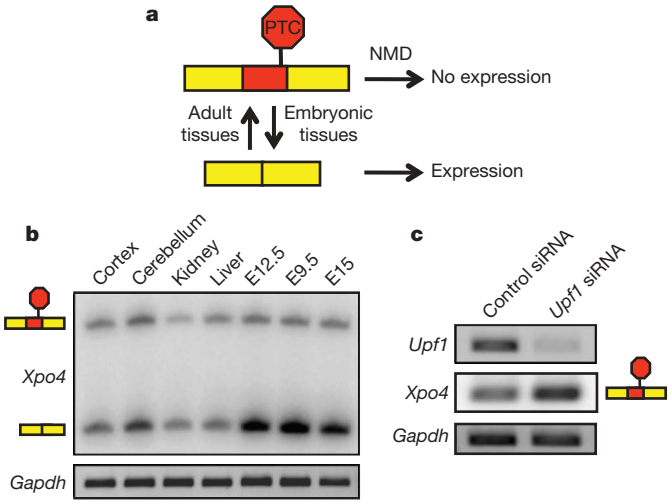

Figure 5 | The code predicts a mechanism for developmental regulation. **a**, The code identified a class of PTC-introducing exons predicted to activate NMD when included in adult tissues, but to allow mRNA expression when skipped in embryonic tissues. **b**, RT–PCR data monitoring splicing and mRNA expression levels of transcripts from *Xpo4*, which contains a code-predicted PTC-introducing exon, in four adult tissues (cortex, cerebellum, kidney and liver) and three embryonic samples (embryonic day (E)9.5, E12.5 and E15). **c**, RT–PCR data monitoring mRNA levels of the NMD factor Upf1 and the PTC-containing Xpo4 isoform in neuroblastoma (N2A) cells transfected with control siRNAs or Upf1 siRNAs. The Xpo4 PTC-containing isoform was selectively amplified using an exon-specific primer. *Gapdh* mRNA levels represent a loading control.

Alternative splicing-controlled gene expression

The code revealed a mechanism underlying the regulation of specific genes during development, whereby a class of alternative exons that introduce a PTC and activate nonsense-mediated mRNA decay (NMD) are included in adult tissues to suppress mRNA expression, but are skipped in embryonic tissues to activate mRNA expression (Fig. 5a).

Supporting this predicted mechanism, microarray data indicated that 30 out of 38 genes with exons in the above class have higher ($P < 0.05$, *t*-test) mRNA expression levels in embryonic tissues compared to adult tissues. Several high-scoring predictions were confirmed by RT–PCR assays as having low or non-detectable levels of PTC-containing isoforms in the profiled tissues, and relatively abundant levels of the exon-skipped isoforms in embryonic tissues (Fig. 5b and Supplementary Fig. 17a). To confirm the role of NMD in mRNA regulation, a short interfering RNA (siRNA) pool capable of knocking down the essential NMD factor Upf1 was transfected into N2A cells and changes in splice isoform levels were monitored by RT–PCR assays. Knockdown of Upf1 resulted in increased levels of the PTC-containing isoforms in five out of six examples with detectable expression of these isoforms in N2A cells (Fig. 5c and Supplementary Fig. 17b).

Genes containing the developmentally regulated PTC-introducing exons identified by the splicing code include those with previously described roles in development and disease. Exportin 4 (Xpo4, also known as Exp4) is a particularly interesting example (Fig. 5b, c). It is a nuclear export receptor for the translation initiation factor eIF5A[46], and a nuclear import receptor for SRY-related HMG-box (Sox) family transcription factors[47], which have key roles in regulating embryonic development, and are required for the maintenance of stem cell pluripotency[48]. Notably, eIF5A is amplified in certain cancers and a recent oncogenomics-based RNA interference screen further identified human XPO4 as being required for the proliferation of XPO4-deficient tumours[49]. These findings support the conclusion that Xpo4 expression must be tightly controlled such that it is active during embryogenesis but downregulated in adult tissues, to avoid possible deleterious consequences including oncogenesis.

Discussion

The method we used to infer a splicing code produced a testable map for how RNA features work together to regulate tissue-dependent alternative splicing. The utility of the code is supported by evaluation of its ability to predict splicing patterns for previously unanalysed exons in major tissue types, including CNS, muscle, embryo and digestive tissues; recapitulation of results from previous studies of muscle- and brain-dependent splicing including targets of Nova, Fox and PTB/nPTB; evaluation of RNA segments predicted to have regulatory function; an automatically generated, interpretable, graphical depiction of the code; and discovery of a class of exons whose alternative splicing regulates gene expression differently in adult and embryonic tissues, by introducing PTCs. Unlike high-throughput sequencing and microarray profiling, our code can successfully predict tissue-regulation of exons independently of transcript expression levels (Supplementary Information 11).

To facilitate future research, we developed a web tool (accessible at http://genes.toronto.edu/wasp), which can be used to explore new regulatory elements and how these elements work in combination to shape the transcriptional landscape. The tool can scan previously uncharacterized exons, predict tissue-dependent splicing patterns, and produce downloadable exploratory feature maps linked to the UCSC genome browser. Users can also download data sets comprising the feature vectors and prediction targets described earlier. As an example of the tool's utility, we wanted to explore exons that might be involved in human neurological disorders, so we used the code to predict previously uncharacterized CNS-regulated exons in widely expressed genes associated with Parkinson's disease, Alzheimer's disease, and several other disorders (Supplementary Information 11, Supplementary Table 4 and Supplementary Fig. 14). In many cases, the newly identified CNS-regulated exons are predicted to affect critical protein domains and one of the exons overlaps patient genomic deletions linked to neurological disease.

A unique aspect of our approach is that it searches for a regulatory code that maximizes a quantifiable measure of code quality, so as to jointly account for many features and produce a predictive splicing code. Interesting future directions include incorporating *in vivo* CLIP data[18], high-throughput *in vitro* protein-RNA binding data[50], further splicing profiling (for example, RNA-Seq) data, and different learning algorithms.

It is apparent from examining the splicing code deciphered in the present study that large numbers of sequence features are generally required to achieve tissue-regulated splicing. We anticipate that the splicing code described here will be useful in future studies directed at understanding the mechanisms by which these elements and *trans*-acting factors combine to regulate tissue-dependent splicing regulation, and how these mechanisms go awry in human diseases.

METHODS SUMMARY

Code assembly: A splicing code is inferred such that the prediction $p(c_i,r_i)$, based on the RNA feature vector r_i and tissue type c_i (where $i = 1,\ldots,14{,}460$ indexes exon-tissue type data points), is as close to the measured splicing pattern q_i as possible, for all data points. To achieve this, we introduce an information theoretic[31] measure of 'code quality':

$$\text{Code quality} = \sum_{i \in \text{data points}} \sum_{s \in \{\text{inc, exc, nc}\}} q_i^s \log\left(\frac{p^s(c_i,r_i)}{\bar{q}^s}\right)$$

Where $\bar{q}^s$ is the average probability of increased inclusion ($s = \text{inc}$), increased exclusion ($s = \text{exc}$) or no change ($s = \text{nc}$), taken over all exons and tissue types. The code quality can also be viewed as the data set log-likelihood, up to an additive constant. In Fig. 2a, thresholds $q^{\text{inc}} \geq 0.99$ and $q^{\text{exc}} \geq 0.99$ were applied to obtain 28,920 binary indicators (3.4% positive) and matching code prediction probabilities were obtained using fivefold cross validation. In Fig. 2b, c, cases in which the microarray- or RT–PCR-measured tissue-difference in the percentage exon inclusion exceeded one s.d. in expected error[11] (5%) were selected, and microarray cases were further screened so that transcripts in both tissues were among the top 20% in expression. For every test exon and pair of tissues c and c', the difference in predictions for the two tissues, $\Delta p = p(c,r) - p(c',r)$, was computed and high confidence cases ($|\Delta p| > 0.5$) were used for testing.

Computational techniques, splicing reporter constructs, cell transfections and RT–PCR assays are described in Supplementary Information.

Received 9 December 2009; accepted 9 March 2010.

1. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).
2. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genet.* **40,** 1413–1415 (2008).
3. Wang, G.-S. & Cooper, T. A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Rev. Genet.* **8,** 749–761 (2007).
4. Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14,** 802–813 (2008).
5. Hartmann, B. & Valcarcel, J. Decrypting the genome's alternative messages. *Curr. Opin. Cell Biol.* **21,** 377–386 (2009).
6. Hallegger, M., Llorian, M. & Smith, C. W. Alternative splicing: global insights. *FEBS J.* **277,** 856–866 (2010).
7. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463,** 457–463 (2010).
8. Blencowe, B. J. Alternative splicing: new insights from global analyses. *Cell* **126,** 37–47 (2006).
9. Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72,** 291–336 (2003).
10. Fagnani, M. *et al.* Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biol.* **8,** R108 (2007).
11. Shai, O., Morris, Q. D., Blencowe, B. J. & Frey, B. J. Inferring global levels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics* **22,** 606–613 (2006).
12. Sugnet, C. W. *et al.* Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput. Biol.* **2,** e4 (2006).
13. Das, D. *et al.* A correlation with exon expression approach to identify *cis*-regulatory elements for tissue-specific alternative splicing. *Nucleic Acids Res.* **35,** 4845–4857 (2007).
14. Castle, J. C. *et al.* Expression of 24,426 human alternative splicing events and predicted *cis* regulation in 48 tissues and cell lines. *Nature Genet.* **40,** 1416–1425 (2008).
15. Minovitsky, S., Gee, S. L., Schokrpur, S., Dubchak, I. & Conboy, J. G. The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons. *Nucleic Acids Res.* **33,** 714–724 (2005).
16. Kawamoto, S. Neuron-specific alternative splicing of nonmuscle myosin II heavy chain-B pre-mRNA requires a cis-acting intron sequence. *J. Biol. Chem.* **271,** 17613–17616 (1996).
17. Ule, J. *et al.* An RNA map predicting Nova-dependent splicing regulation. *Nature* **444,** 580–586 (2006).
18. Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456,** 464–469 (2008).
19. Chan, R. C. & Black, D. L. The polypyrimidine tract binding protein binds upstream of neural cell-specific c-src exon N1 to repress the splicing of the intron downstream. *Mol. Cell. Biol.* **17,** 4667–4676 (1997).
20. Ashiya, M. & Grabowski, P. J. A neuron-specific splicing switch mediated by an array of pre-mRNA repressor sites: evidence of a regulatory role for the polypyrimidine tract binding protein and a brain-specific PTB counterpart. *RNA* **3,** 996–1015 (1997).
21. Faustino, N. A. & Cooper, T. A. Identification of putative new splicing targets for ETR-3 using sequences identified by systematic evolution of ligands by exponential enrichment. *Mol. Cell. Biol.* **25,** 879–887 (2005).
22. Galarneau, A. & Richard, S. Target RNA motif and target mRNAs of the Quaking STAR protein. *Nature Struct. Mol. Biol.* **12,** 691–698 (2005).
23. Sorek, R. & Ast, G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13,** 1631–1637 (2003).
24. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297,** 1007–1013 (2002).
25. Zhang, X. H. & Chasin, L. A. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.* **18,** 1241–1250 (2004).
26. Stadler, M. B. *et al.* Inference of splicing regulatory activities by sequence neighborhood analysis. *PLoS Genet.* **2,** e191 (2006).
27. Yeo, G. W., Nostrand, E. L. & Liang, T. Y. Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet.* **3,** e85 (2007).
28. Xiao, X., Wang, Z., Jang, M. & Burge, C. B. Coevolutionary networks of splicing *cis*-regulatory elements. *Proc. Natl Acad. Sci. USA* **104,** 18583–18588 (2007).
29. Shepard, P. J. & Hertel, K. J. Conserved RNA secondary structures promote alternative splicing. *RNA* **14,** 1463–1469 (2008).
30. Bishop, C. M. *Pattern Recognition and Machine Learning*. (Springer, 2006).
31. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27,** 379–423 (1948).
32. Wollerton, M. C., Gooding, C., Wagner, E. J., Garcia-Blanco, M. A. & Smith, C. W. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol. Cell* **13,** 91–100 (2004).
33. Wei, N., Lin, C. Q., Modafferi, E. F., Gomes, W. A. & Black, D. L. A unique intronic splicing enhancer controls the inclusion of the agrin Y exon. *RNA* **3,** 1275–1288 (1997).
34. Lim, L. P. & Sharp, P. A. Alternative splicing of the fibronectin EIIIB exon depends on specific TGCATG repeats. *Mol. Cell. Biol.* **18,** 3900–3906 (1998).
35. Côté, J., Dupuis, S., Jiang, Z. & Wu, J. Y. Caspase-2 pre-mRNA alternative splicing: identification of an intronic element containing a decoy 3′ acceptor site. *Proc. Natl Acad. Sci. USA* **98,** 938–943 (2001).
36. Hayakawa, M. *et al.* Muscle-specific exonic splicing silencer for exon exclusion in human ATP synthase γ-subunit pre-mRNA. *J. Biol. Chem.* **277,** 6974–6984 (2002).
37. Jin, Y. *et al.* A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *EMBO J.* **22,** 905–912 (2003).
38. Zhang, C. *et al.* Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes Dev.* **22,** 2550–2563 (2008).
39. Calarco, J. A. *et al.* Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell* **138,** 898–910 (2009).
40. Gooding, C. *et al.* A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. *Genome Biol.* **7,** R1 (2006).
41. Wu, J. I., Reed, R. B., Grabowski, P. J. & Artzt, K. Function of quaking in myelination: regulation of alternative splicing. *Proc. Natl Acad. Sci. USA* **99,** 4233–4238 (2002).
42. Oberstrass, F. C. *et al.* Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science* **309,** 2054–2057 (2005).
43. Markovtsov, V. *et al.* Cooperative assembly of an hnRNP complex induced by a tissue-specific homolog of polypyrimidine tract binding protein. *Mol. Cell. Biol.* **20,** 7463–7479 (2000).
44. Xie, J., Jan, C., Stoilov, P., Park, J. & Black, D. L. A consensus CaMK IV-responsive RNA sequence mediates regulation of alternative exons in neurons. *RNA* **11,** 1825–1834 (2005).
45. Pérez, I., Lin, C. H., McAfee, J. G. & Patton, J. G. Mutation of PTB binding sites causes misregulation of alternative 3′ splice site selection *in vivo*. *RNA* **3,** 764–778 (1997).
46. Lipowsky, G. *et al.* Exportin 4: a mediator of a novel nuclear export pathway in higher eukaryotes. *EMBO J.* **19,** 4362–4371 (2000).
47. Gontan, C. *et al.* Exportin 4 mediates a novel nuclear import pathway for Sox family transcription factors. *J. Cell Biol.* **185,** 27–34 (2009).
48. Lefebvre, V., Dumitriu, B., Penzo-Méndez, A., Han, Y. & Pallavi, B. Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. *Int. J. Biochem. Cell Biol.* **39,** 2195–2214 (2007).
49. Zender, L. *et al.* An oncogenomics-based *in vivo* RNAi screen identifies tumor suppressors in liver cancer. *Cell* **135,** 852–864 (2008).
50. Ray, D. *et al.* RNACompete: Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnol.* **27,** 667–670 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements. We thank D. L. Black, D. Botstein, S. E. Brenner, C. B. Burge, B. Chabot, R. Durbin, X.-D. Fu, B. R. Graveley, T. R. Hughes, N. Jojic, C. W. J. Smith, S. Tavazoie and members of our various laboratories for discussions or comments on the manuscript; D. D. Licatalosi and R. B. Darnell for communicating unpublished results; and S. Chaudhry for RT–PCR work. B.J.F. also thanks C. M. Bishop and D. J. C. MacKay for hosting him during his sabbatical in Cambridge. This research was funded by a grant from Genome Canada through the OGI to B.J.B., B.J.F. and others; an NSERC/CFI/OIT CRC grant to B.J.F.; CIHR grants to B.J.F. and B.J.B.; an NCIC grant to B.J.B.; and NSERC EWR Steacie and Canadian Institute for Advanced Research Fellowships to B.J.F.

Author Contributions. Y.B. and B.J.F. developed the predictive framework and code assembly algorithms, analysed validation rates, and with B.J.B. and J.A.C. extracted predictions for regulatory mechanisms. Y.B., B.J.B. and B.J.F. produced the feature compendium. J.A.C. performed wet laboratory experiments. Q.P. generated exon and intron datasets. W.G. and Y.B. developed the web tool with input from the other authors. X.W. analysed exons from neurological disorder-associated genes. O.S. estimated the percentage inclusion values. B.J.F., B.J.B. and Y.B. designed the study and wrote the manuscript with input from the other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.J.F. (frey@psi.toronto.edu) or B.J.B. (b.blencowe@utoronto.ca).