A Biophysical Model for Identifying Splicing Regulatory Elements and Their Interactions

Ji Wen[1], Zhibin Chen[2], Xiaodong Cai[1]*

[1] Department of Electrical and Computer Engineering, University of Miami, Coral Gables, Florida, United States of America, [2] Department of Microbiology and Immunology, University of Miami, Miami, Florida, United States of America

Abstract

Alternative splicing (AS) of precursor mRNA (pre-mRNA) is a crucial step in the expression of most eukaryotic genes. Splicing factors (SFs) play an important role in AS regulation by binding to the *cis*-regulatory elements on the pre-mRNA. Although many splicing factors (SFs) and their binding sites have been identified, their combinatorial regulatory effects remain to be elucidated. In this paper, we derive a biophysical model for AS regulation that integrates combinatorial signals of cis-acting splicing regulatory elements (SREs) and their interactions. We also develop a systematic framework for model inference. Applying the biophysical model to a human RNA-Seq data set, we demonstrate that our model can explain 49.1%–66.5% variance of the data, which is comparable to the best result achieved by biophysical models for transcription. In total, we identified 119 SRE pairs between different regions of cassette exons that may regulate exon or intron definition in splicing, and 77 SRE pairs from the same region that may arise from a long motif or two different SREs bound by different SFs. Particularly, putative binding sites of polypyrimidine tract-binding protein (PTB), heterogeneous nuclear ribonucleoprotein (hnRNP) F/H and E/K are identified as interacting SRE pairs, and have been shown to be consistent with the interaction models proposed in previous experimental results. These results show that our biophysical model and inference method provide a means of quantitative modeling of splicing regulation and is a useful tool for identifying SREs and their interactions. The software package for model inference is available under an open source license.

Citation: Wen J, Chen Z, Cai X (2013) A Biophysical Model for Identifying Splicing Regulatory Elements and Their Interactions. PLoS ONE 8(1): e54885. doi:10.1371/journal.pone.0054885

Editor: Xiaoning Qian, University of South Florida, United States of America

Received October 18, 2012; **Accepted** December 17, 2012; **Published** January 30, 2013

Funding: This work was supported by National Science Foundation under grant CCF-0746882 to XC. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: x.cai@miami.edu

Introduction

A key step in eukaryotic gene expression is to remove introns from precursor messenger RNA (pre-mRNA) so that exons can be joined together to form the mature mRNA. By including different exons in the mRNA, alternative splicing (AS) can generate different isoforms from a single gene to increase proteomic diversity. Recent studies have found that ~95% of human genes undergo AS [1,2]. The importance of AS is highlighted recently by the findings that AS related mutations can cause many human diseases including cancer [3,4].

AS can be regulated via several mechanisms, often in a tissue-specific manner [5,6]. One mechanism is that recognition of splice sites by the spliceosome is influenced by a class of RNA binding proteins named splicing factors (SFs) that can bind to the *cis*-acting splicing regulatory elements (SREs) on the pre-mRNA. These SREs are categorized as exonic splicing enhancers (ESEs) and silencers (ESSs), and intronic splicing enhancers (ISEs) and silencers (ISSs) based on their locations and effects on splicing [7]. Recently, several experimental [8–10].and computational methods have been employed to identify SREs. The computational methods can be largely categorized into three approaches. The first enrichment-based approach is to identify SREs as short nucleotide sequences (typically hexamers or octamers) that are statistically enriched in a carefully selected set of introns and exons against a background or negative dataset [11–13]. The second conservation-based approach utilizes comparative genomic methods to identify evolutionarily conserved motifs in introns and exons, which can also be combined with the enrichment-based approach to identify SREs [14–17]. The third regression-based approach exploits both sequence information and expression levels of different isoforms in a unified framework [18,19]. Comparing with the other two approaches, the regression-based approach offers flexibility of identifying combinatorial regulatory effects of multiple SREs. However, the current regression methods for AS were not developed systematically from a theoretical base, which may limit their performance.

Multiple SREs could act cooperatively to promote or repress splicing by regulating exon or intron definition [20,21]. However, given the huge number of possible pairs of SREs in different regions, experimental approaches for identifying SRE pairs is expensive and time-consuming if feasible. For this reason, computational methods become an important means of selecting candidate SRE pairs in a systematic and high-throughput manner. Several recent computational works have studied cooperative SRE pairs in AS regulation. Ke and Chasin [22].searched for frequently co-occurred SRE pairs from two ends of exons that mediate exon definition. Friedman *et al.* [23] identified cooperative SRE pairs from two ends of human and mouse introns possibly mediating intron definition. Suyama *et al.* [24] analyzed conserved pentamers that often co-occur in the same region of upstream or downstream introns, which may arise from cooperative binding of different SFs or actually from a single long motif. These different types of SRE pairs from different regions reveal that cooperative interaction between SREs may be a common mechanism in AS regulation. However, these SRE pair detection methods did not incorporate expression data into the analysis, and like the enrichment-based approach to the detection of single SRE, they could not exploit sequence and expression data in a systematic way, which may limit their detection power.

In this paper, we first derive a novel biophysical regression model for the regulation of AS, which can capture both main effects of individual SREs and combinatorial effects of multiple SREs. We then develop a systematic framework to infer the regression model, which in turn identifies both single SREs and different types of cooperative SRE pairs. The key feature of our model inference framework is that we employ the shrinkage technique [25] to identify a small number of SREs and SRE pairs from a huge number of all possible SREs and their pairs. Our numerical results show that our regression model can explain a significant portion of the variance in the data comparable to the best result for transcription achieved by a nonlinear biophysical model [26]. Using an RNA-Seq data set [1], we identify 619 SREs and 196 SRE pairs, some of which are verified with previous experimental results. For example, binding sites of PTB in upstream and downstream of introns of alternatively spliced exons (ASEs), binding sites of hnRNP F/H in two ends of downstream intron, and binding sites of hnRNP E/K in the same region are identified as interacting SRE pairs, which are consistent with existing experimental evidences.

Results

Biophysical Model for AS Regulation

The basal machinery of splicing is known as the spliceosome, a large multicomponent ribonucleoprotein complex having U1, U2, U4, U5 and U6 snRNPs as its main building blocks [27]. Splicing begins with a multi-step process of spliceosome assembly around the splice sites and the branch point. SFs bound to nearby SREs can influence spliceosome assembly by facilitating or inhibiting the subunits of spliceosome to recognize the splice sites [7,28,29]. They can also regulate splicing through other mechanisms such as regulation of the transition from exon definition to intron definition [5,30]. Moreover, multiple SFs and the spliceosome can interact cooperatively or antagonistically to affect the splicing process.

We model the spliceosome assembling process with a chemical reaction:

$$S + RNA \rightleftharpoons RNA \cdot S, \tag{1}$$

where S denotes the spliceosome. This simplification is similar to the one used in the derivation of a biophysical model [26,31,32] for transcription where assembly of the RNA polymerase (RNAP) complex is simplified to one reaction. If an SF binding to an SRE interacts with the spliceosome, it increases or decreases the equilibrium constant of the above reaction, which in turn enhances or inhibits splicing. Let us consider a gene with one ASE and let I_1 (I_2) be the isoform that includes (excludes) the ASE. Let E_{I_1} and E_{I_2} be the expression levels of I_1 and I_2, respectively. We assume that the pre-mRNA of the gene with assembled spliceosome produces I_1, whereas the pre-mRNA without assembled spliceosome produces I_2. Therefore, the probability of producing I_1 is equal to the probability that the pre-mRNA is bound by the spliceosome. The first probability can also be expressed as $E_{I_1}/(E_{I_1}+E_{I_2})$, while the second probability can be derived from the biophysical chemical reactions modeling spliceosome assembly and binding of SREs to the pre-mRNA as detailed in Materials and Methods. Based on this observation, we derive the following regression model in Materials and Methods to capture the regulatory effects of SREs on the splicing of an ASE:

$$y = \beta_0 + \sum_{i \in \mathcal{M}} x_i \beta_i + \sum_{(i,j) \in \mathcal{I}} x_i x_j \beta_{ij} + \epsilon, \tag{2}$$

where $y = \log(f^{t_1}) - \frac{1}{N-1} \sum_{t=1, t \neq t_1}^{N} \log(f^t)$, with $f^{t_1} = E_{I_1}^{t_1} / E_{I_2}^{t_1}$ for tissue t_1 and t being the index of N tissues; $\mathcal{M}$ and $\mathcal{I}$ are the set of potential SREs and the set of potential SRE pairs, respectively; x_i is a binary variable to indicate presence ($x_i = 1$) or absence ($x_i = 0$) of the ith SRE; β_i reflects the contribution of the ith SRE to the splicing of the ASE, and β_{ij} indicates the cooperative contribution of the ith SRE and the jth SRE; ε is the measurement error, which is modeled as a Gaussian random variable with zero mean. Note that the ith and jth SREs in the third term at right hand side of (2) can be from the same region or two different regions as described in Figure 1A. The first term $\log(f^{t_1})$ in y is the logarithm of the ratio between the expression levels of two isoforms including or excluding the ASEs; it is determined by the basal splicing level and the regulatory effects of SFs binding to the SREs. The second term in y is added to remove the basal splicing level determined by the spliceosome alone as described in Materials and Methods. Thus, given the splicing profile y of a set of ASEs and a set of candidate SREs, we can identify SREs or SRE pairs that have regulatory effect by finding β_i or β_{ij} that are not equal to zero with certain statistical significance. The method for model inference to determine β_i or β_{ij} is presented in next section. The condition to determine if an SRE is a enhancer or silencer based on the sign of β_i and the expression level of the SF binding to the SRE is given in Proposition 1 in Materials and Methods. Whether an interacting pair of two SREs is an enhancer or silencer, however, is difficult to be inferred from the sign of β_{ij}.

Note that model (2) is linear with respect to unknown parameters β_i and β_{ij}. The linearity facilitates model inference even when the number of parameters is very large. As shown in Materials and Methods, linear model (2) is directly derived from biophysical principles. In contrast, the biophysical model for gene transcription [26,31–33] is a nonlinear model with respect to unknown parameters to be inferred. While it is possible to infer parameters of the nonlinear model when the number of parameters is small [26], model inference becomes difficult when the number of parameters is relatively large. The linear regression models for transcription [34–37] or for splicing [18,19] are an approximation of the nonlinear biophysical models, as shown in [36]. Although these linear models enable efficient model inference, their performance is limited by the approximation inherent to the models. A nonlinear model based on regression spline [38] was also developed to approximate the nonlinear biophysical model. Comparing with linear regression, regression spline reduces approximation error, but increases the complexity of model inference.

Framework for Model Inference

We applied the biophysical model to identify SREs and SRE pairs involved in alternative splicing. As described in Materials and Methods, We first determined a set of ASEs from the UCSC KnownGene table [39], then extracted all the hexamers in five

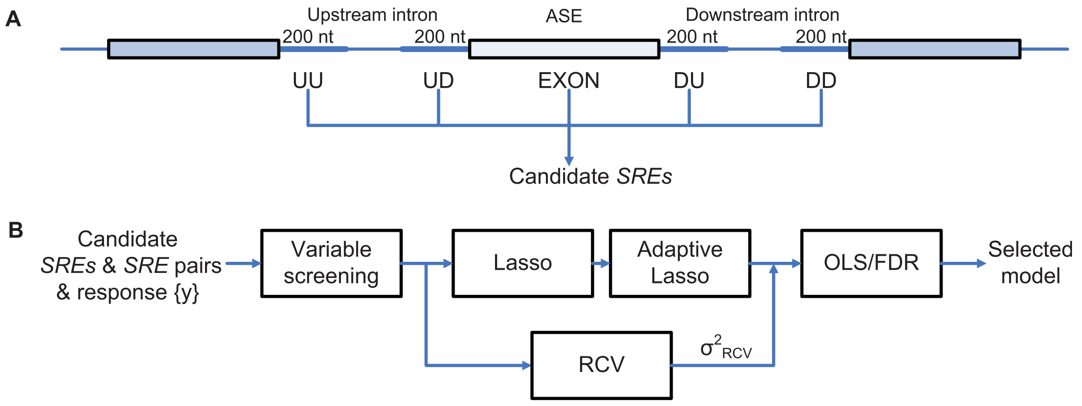

Figure 1. Five regions around ASEs used to extract SREs and the model inference framework. (**A**) All hexamers in the five regions around ASEs are considered as candidate SREs. UU (upstream/upstream) stands for the 5′ end of the upstream intron. UD (upstream/downstream) denotes the 3′ end of the upstream intron. DU and DD are defined in a similar way. EXON stands for the ASE region. (**B**) The inference framework for detecting active SREs and SRE pairs. RCV: refitted cross-validation; OLS: ordinary least squares regression.
doi:10.1371/journal.pone.0054885.g001

regions around ASEs as candidate SREs (Figure 1A), and obtained y for this set of ASEs from the RNA-Seq data. With this data set, model inference was carried out using four components described in Figure 1B.

Since we considered all 6-mers and their interactions, the regression model (2) contained $5\times 4^6=20480$ variables for main effects of possible SREs and $>10^8$ (20480·20479/2) variables for interactions of possible SRE pairs. The huge number of variables not only requires huge computation for model inference, but also may yield a large number of false SREs. To overcome these problems, we developed the four-component framework in Figure 1B to select reliable SREs without overfitting the model. In the first variable screening component, we used the sure independence screening method [40] described in Materials and Methods to remove 6-mers that have very small correlation with the response variable y.

In the second component, we adopted the Lasso [41] and the adaptive Lasso [42] to perform penalized multiple regression to select variables. Descriptions of the Lasso, the adaptive Lasso and the cross-validation procedure for determining the parameter λ of the Lasso and the adaptive Lasso are given in Materials and Methods. The Lasso is known to shrink many variables, with no or small correlation with the response variable, to zero [41], and thus yields a sparse model that only contains a small number of variables. Using both the Lasso and the adaptive Lasso was to ensure more reliable variable selection as suggested in [43]. This component produced a sequence of models for different values of λ. The minimum variance of the residual error at $\lambda=\lambda_{min}$ determined by cross-validation was denoted as σ^2_{min}.

Although the Lasso and the adaptive Lasso only retained a small number of variables in the model, we wanted to ensure that the overfitting problem did not occur. To this end, we added the third component named refitted cross-validation (RCV) [44] to the inference procedure. RCV reliably estimates the variance σ^2_{RCV} of the residual error in a linear model of ultrahigh dimension. We then compared σ^2_{min}, with σ^2_{RCV}. If $\sigma^2_{min}>\sigma^2_{RCV}$, we selected variables that gave σ^2_{min}; otherwise, we identified the value of λ that yielded a residual variance equal to σ^2_{RCV} and selected variables with this λ.

The adaptive Lasso together with RCV selected a set of SREs and SRE pairs, but it did not give p-value for each variable in the model. In the last component, we used ordinary least squares (OLS) method to refit the model with the variables selected by the adaptive Lasso. We then used the p-values provided by OLS to select variables at a false discovery rate (FDR) ≤ 0.01 [45]. This final set of variables were identified as SREs and SRE pairs. A software package implementing the inference framework is freely available under an open source license.

Performance of the Model and the Regression Framework

As described in Materials and Methods, we selected a set of ASEs for each tissue from the KnownGene table and calculated the inclusion ratio $E_{I_1}/(E_{I_1}+E_{I_2})$ from the RNA-Seq data [1] for each ASE, which was then used to calculate the response y in model (2). We applied our biophysical model and inference framework to this data set to identify SREs and SRE pairs. The number of ASEs used in model inference, the number of SREs and SRE pairs in the final model and the percentage variance explained by the final model are given in Table 1. In each tissue, our biophysical model explained 49.1%–66.5% of the variance in the data (see R^2 in Table 1), which was comparable to that achieved by the best model for transcription reported in [26]. More specifically, the linear model for gene transcription in [34] explained 9.6% of the variance on average, while the regression spline model for gene expression that incorporated interaction terms explained 13.9% to 32.9% of the variance [38]. Most recent work by Gertz *et al.* [26] fitted a nonlinear biophysical model to the expression data of synthetic genes. Their models explained 44–59% of the variance in gene expression. Thus, considering the non-synthetic genes we used, our model has captured a large fraction of the variance in the splicing response.

Overall, 619 different SREs and 196 SRE pairs were detected from different tissues. Specifically, Table S1 contains all the SREs and SRE pairs identified from different tissues, regression coefficient and p-value for each SRE or SRE pair. It also includes some SFs that have experimental evidence (SELEX [46–65], RNAcompete [66] and other experiments [67–75]) to bind to the identified SREs. These SREs and SRE pairs consist of 854 different hexamers. Many SREs are very similar to each other, which may arise from an SRE longer than 6 nucleotides (nts) or from a degenerate motif. Note that although they were detected in different tissues, the SRE and SRE pairs *per se* do not give rise to tissue-specific isoforms, because they are always there in pre-mRNA sequences in different tissues. It is the tissue-specific

Table 1. Summary of the final selected models in each tissue.

Tissue	No. of ASEs	No. of SREs	No. of SRE pairs	R^2 (%)
Adipose	1345	60	30	52.9
Brain	1411	65	19	49.1
Breast	1399	82	24	58.3
Colon	1236	64	13	49.4
Heart	1302	84	21	60.5
Liver	995	76	19	66.5
Lymph node	1405	77	24	55.7
Skeletal muscle	1174	85	18	63.2
Testes	1601	79	28	51.0

The second column is the number of ASEs used in model inference. The third and fourth columns list the number of SREs and SRE pairs in the final model. The last column gives the percentage of the variance explained (R^2) by the final model.

doi:10.1371/journal.pone.0054885.t001

expression of SFs that regulates tissue-specific splicing through SREs.

We compared our 854 SREs with the results of Castle *et al.* [12], who performed a systematic screening of all 4 to 7-mers for cis-regulatory motifs enriched near ASEs using microarray. We only kept 5 to 7-mers with a p-value smaller than 10^{-3} (Bonferroni-corrected p-values as described in [12]) for the comparison. In total, 137 non-redundant 5 to 7-mers were left (91 5-mers, 28 6-mers and 18 7-mers). We considered it as a match if a 7-mer contains one of our SREs, a 5-mer is part of our SREs, or a 6-mer exactly matches one of our SREs. This yielded 103 k-mers, $k=5,6,7$, that could find at least one match in our SREs. In order to evaluate the significance of the overlap, we generated a list of 6-mers from the 137 k-mers of Castle *et al.* with the following procedure. For each 5-mer, 8 different 6-mers containing the 5-mer were obtained by padding a nucleotide to the beginning or the end of the 5-mer. For each 7-mer, 2 different 6-mers were obtained by extracting the first or the last 6 nts. In total, 639 different 6-mers were extracted from Castle's k-mers, $k=5,6,7$. A significant number of 6-mer (180) were found in both our 854 SREs and the 639 6-mers obtained from 137 k-mers of Castle *et al.* (p-value $= 9.37\mathrm{e}{-7}$ from Fisher's exact test).

We also compared our 854 SREs with the binding sites of 25 SFs experimentally identified with SELEX [46–65] or RNAcompete [66]. Each of [46–65] attempted to determine the binding sites of 1 to 3 SFs using SELEX, and [46–65] reported SELEX results for a total of 25 SFs. For each of 25 SFs, we obtained a set of RNA sequences selected with SELEX from one of [46–65]. If there are more than one SELEX results for an SF, we used the most recent SELEX result. We then extracted the consensus sequences embedded in this set of RNA sequences as the binding sites of the SF. If a consensus binding site is 4 or 5 nt long, one more nucleotide was also extracted from each side of the original selected sequence to obtain hexamers. If a consensus binding site is longer than 6 nts, all the hexamers included in the consensus were extracted. For RNAcompete, the 7-mers listed in Figure 2 of [66] were used as consensus binding sites, and two hexamers were taken from each 7-mer. In total, 709 different hexamers were obtained from the consensus binding sites. A significant number (175) of hexamers were found in both our 854 SREs and the 709 hexamers obtained from SELEX or RNAcompete (p-value = 0.004 from Fisher's exact test). In Table S1, we gave the SF that was identified in [46–66] to bind to one of the 175 hexamers. Although the overlap between our predicted SREs and the binding sites of SFs determined with SELEX and RNAcompete is statistically significant, a relative large number (679) of our predicted SREs are not included in these experimental results, which implies that our result contains some novel SREs.

Experimental Evidence of SREs

Several well-defined SREs involved in tissue-specific AS have been detected in our work. We will take SREs bound by Fox-1 protein, polypyrimidine tract binding protein (PTB), quaking protein (QKI), muscleblind-like protein (MBNL) as examples (listed in Table 2) to illustrate how to understand the results and compare them with the available experimental evidences.

Fox protein recognizes [U]GCAUG as its SRE and it has been shown to be one of the most conserved regulators of tissue-specific AS in metazoans. Fox-1 is exclusively expressed in brain, heart, and skeletal muscle as reported in [67], which is consistent with the RNA-Seq data we used as shown in Figure 2. Its paralog Fox-2 has relatively low expression level in all tissues. In our results, we detected two SREs containing UGCAUG as summarized in Table 2. These 2 SREs were detected in heart or muscle, which is consistent with the tissues where Fox-1 is expressed. Note that the second SRE UGCAUG(UU)-GCAUGU(UU) detected in upstream region in heart is an interaction term in the regression model. Since our model includes interaction between SREs in the same region or from different regions, it is possible that an SRE longer than 6 nts is detected as an interaction term. This interaction term actually arises from a 7-mer SRE UGCAUGU(UU) [68] (Among the 28 UGCAUG(UU)-GCAUGU(UU) pairs used for inference in heart, 27 are derived from UGCAUGU(UU)). Using the inference method for regulatory effects described in Materials and Methods, we found that the SRE that we identified from muscle are enhancers in the downstream region (DU) (Table 2), which is consistent with the computational analysis [1] and the experimental evidence [67,76].

Another well-defined SF is PTB which binds to UC-rich SREs and has high binding affinity to UCUU and/or UCUCU [69,77]. Two SREs we identified are possible binding sites of PTB (Table 2). They are found in the upstream region in adipose and lymph node. Both SREs are detected from the tissues where PTB are over-expressed as shown in Figure 2. Using Proposition 1 in Materials and Methods, These two SREs were determined to be a silencer in the upstream of ASEs. This result coincides with position-dependent alternative splicing activity of the PTB as identified using microarrays [78].

One of QKI's binding site ACUAAY [59,70] was detected in the upstream intron of the ASE in muscle (Table 2). The SRE ACUAAC was previously reported as an over-represented motif in the downstream region of ASEs in muscle [18] and was also predicted as an upstream intronic SRE specific to the central nervous system [79]. Our result indicates that this upstream SRE detected in muscle is a silencer, which is consistent with the experimental result [80].

The MBNL family protein can bind to the motif YGCU(U/G)Y [71]. Two sequences of this motif can be found in our SREs (Table 2). However, the function of this motif is complicated. First, both MBNL and CELF family proteins can bind to similar motifs [81,82], although different protein isoforms may have different binding specificities [83]. Second, MBNL can either promote or repress splicing of specific ASEs on different pre-mRNAs by antagonizing the activity of CELF proteins [71]. Thus, although our result indicates that this SRE is related to AS regulation in muscle and lymph node, it is difficult to interpret its regulatory effect based on current result.

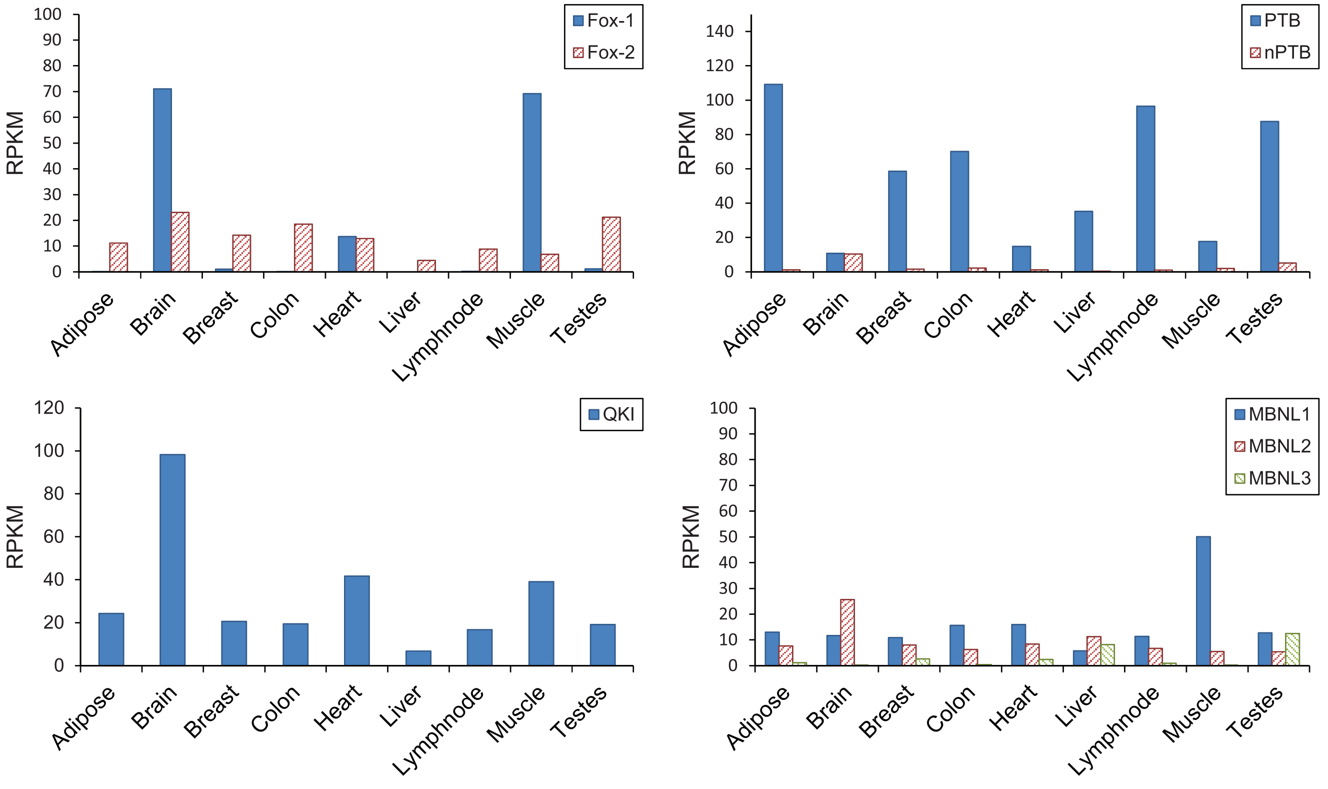

Figure 2. Expression level of several splicing factors in 9 tissues. The expression levels are calculated from the RNA-Seq data [1] as reads per kilobases per million mapped reads (RPKM).
doi:10.1371/journal.pone.0054885.g002

Experimental Evidence of SRE Pairs

We also identified 196 different cooperative SRE pairs (Table S1). Thirty-nine percent (77 SRE pairs) of them are interactions of SREs in the same region. As discussed in the previous section, a part of this kind of interaction may come from a single SRE longer than 6 nts. These SRE pairs are actually not related to interactions, although they also have biological meanings, since a longer conserved SRE may have stronger regulatory effect than a shorter SRE. We have marked all 23 such SRE pairs in Table S1. Among the remaining interactions between different regions (119 SRE pairs), the most frequent interactions are SREs in region pairs UD-DU, UU-UD and UD-DD. Region pair UD-DU may reflect the effect in the exon definition stage of spliceosome assembly, and region pair UU-UD may reflect the effect in the intron definition stage. A summary of interaction between SRE pairs in different regions is given in Figure 3. Several detected SRE pairs are listed in Table 3.

Table 2. Selected examples of the detected SREs.

Putative SF	SRE(region)	P-value	# Occ.	Tissue	Effect
Fox-1	UGCAUG(DU)	7.02E−13	98	muscle	enhancer
Fox-1	UGCAUG(UU)/ GCAUGU(UU)	1.34E−3	28	heart	unknown
PTB	CUCUCU(UD)	9.76E−6	140	lymph node	silencer
PTB	UUCUCU(UD)	9.33E−4	201	adipose	silencer
QKI	ACUAAC(UD)	3.36E−8	39	muscle	silencer
MBNL	UGCUGC(UU)	1.22E−6	110	lymph node	unknown
MBNL	UGCUGC(EXON)	2.61E−3	65	muscle	unknown

# Occ. is abbreviation for number of occurrences.
doi:10.1371/journal.pone.0054885.t002

Several previously identified SREs were detected in our interaction results. They either cooperate with their own or other different SREs in a different region or in the same region. Two SRE pairs bound by PTB [69,77] were identified in our result. All the four SREs located at the upstream region of the 3′ splice site which can also be a part of an extended polypyrimidine tract, although the 3′ splice site consensus of 15 nts containing the classical polypyrimidine tract [7,11] has been removed in our analysis. One SRE pair UUCUCU-UCCUCU was identified in the upstream intron in brain. The other interaction detected involves PTB's binding site UCUUCC in the UD region and CCUUCU in the DD region (Table 3). The downstream CCUUCU resembles the PTB binding sites and might also be bound by PTB. In a cooperative model for the regulatory mechanism of PTB, it was proposed that a single PTB or a PTB dimer could loop out the branch point upstream of 3′ splice site or the ASE by binding to several pyrimidine tracts upstream and downstream to repress splicing [84]. A single PTB-binding element had weak silencing activity, while multiple PTB-binding sites at both upstream and downstream of the ASE or of a branch point could have strong inhibitory effect [85]. This model and the experimental results in [84,85] are consistent with the interaction pairs we identified.

hnRNP F/H can bind to GGGA and G-rich SREs [72,73]. The proposed mechanism underlying the splicing regulation of hnRNP

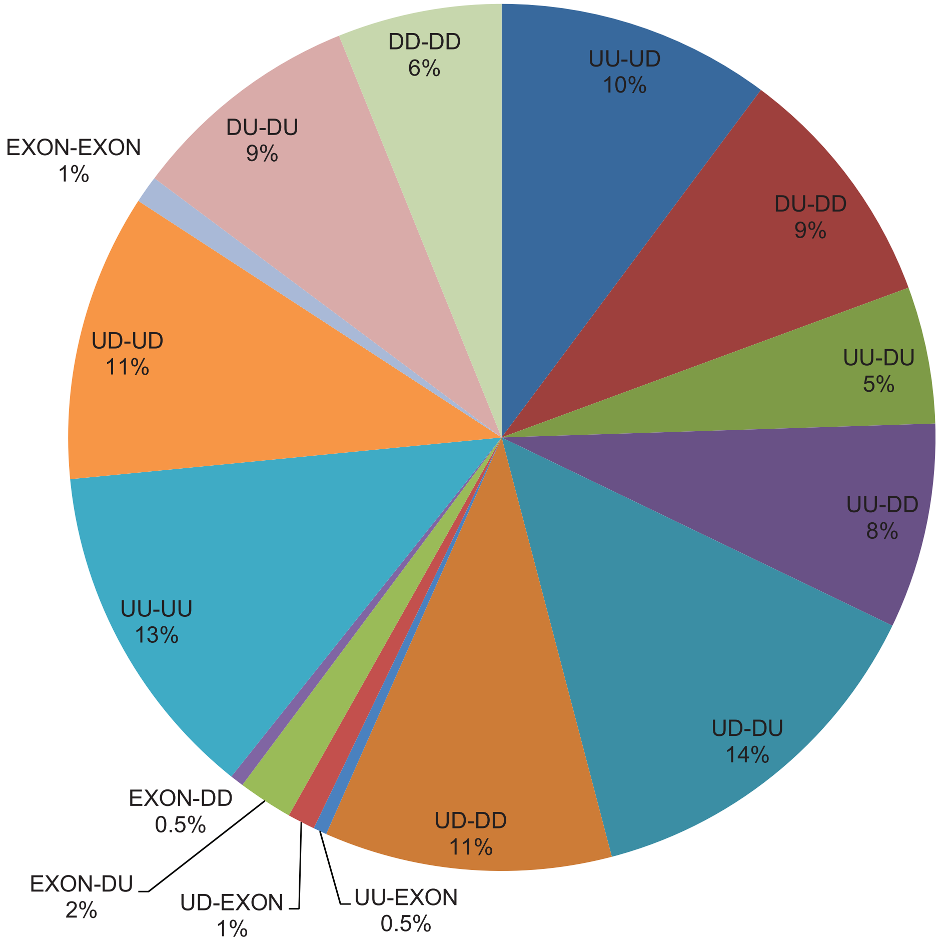

Figure 3. Percentage of different types of SRE pairs. Two hundred and forty-one SRE pairs are detected in different or same regions. This figure shows breakdown of the SRE pairs in different regions.
doi:10.1371/journal.pone.0054885.g003

F/H involves an interaction between proteins bound at both ends of an intron that loops out the intron and brings distantly separated exons into closer proximity [86]. Consistent with this model, one pair of SREs GGGGCA(DU)/UGGGGA(DD) (Table 3) putatively bound by the hnRNP F/H was detected in two ends of the downstream intron.

Both hnRNP E and hnRNP K proteins contain three copies of the KH domain arranged in a very similar manner, and they are the major poly-C binding proteins in mammalian cells [55,74]. We extracted all the SRE pairs in our result that contain at least 4 Cs. Two such SRE pairs (Table 3) were found to resemble the binding motif of the hnRNP E/K [55]. One SRE pair is due to a longer motif ACCCCUC at downstream intron. The other SRE pair was detected as interaction in the same region. This result may reflect the fact that the three KH domains bind RNA synergistically while a single KH domain appears to have very low level of RNA binding activity [87].

An interesting outcome of this work was the identification of many AU-rich elements in SREs and SRE pairs. The AU-rich elements have been identified previously by comparative analysis as a large class of conserved mammalian ISEs [17]. In plant splicing system, the AU-rich sequences in upstream and downstream introns appear to be involved in early intron recognition and stabilization of spliceosomal complex [88]; and multiple short AU-rich SREs could cooperatively modulate splice site usage [89]. However, the interaction between AU-rich SREs has not been reported in mammals. In our result, The SRE pairs involved in two AU-rich SREs from different regions are listed in Table 3. Among these SRE pairs, 7 pairs were detected in upstream and downstream introns, and 4 were detected at two ends of the same intron. These AU-rich SREs resemble the binding sites of TIA-1/TIAR, Hu protein and Sam68 [49,90]. Hu proteins can bind to both upstream and downstream SREs. It has been speculated that binding of Hu protein to multiple sites both upstream and downstream of an ASE could loop out the ASE and as a result block exon definition to repress splicing [90]. Moreover, Hu proteins can promote exon inclusion by binding to AU-rich sequences that are conserved at both upstream and downstream of ASE [91]. The intricate experimental results stress the importance of further mutation analysis to study the cooperative interactions of AU-rich binding proteins.

We also identified many interactions between AU-rich elements and binding sites of other known SFs. For example, We found one interactions between upstream AU-rich SRE and a downstream SRE resembling QKI's binding site. Another example is the interactions between upstream AU-rich SRE and a CU-rich SRE resembling polypyrimidine tract (Table 3). These results indicate that AU-rich SREs and cooperative pairs may play an important regulatory role in mammalian AS and worth further experimental investigations. In summary, various cooperative mechanisms could

Table 3. Selected examples of the detected cooperative SRE pairs.

Putative SF_1	SRE_1 (region)	Putative SF_2	SRE_2 (region)	P-value	Tissue
PTB	UUCUCU(UD)	PTB	UCCUCU(UD)	2.02E−4	brain
PTB	UCUUCC(UD)	PTB	CCUUCU(DD)	1.02E−6	brain
hnRNP F/H	GGGGCA(DU)	hnRNP F/H	UGGGGA(DD)	1.31E−3	liver
hnRNP E/K	CCCCAG(UU)	hnRNP E/K	CCGCCC(UU)	8.02E−8	lymph node
hnRNP E/K	ACCCCU(DU)	hnRNP E/K	CCCCUC(DU)	8.70E−4	adipose
T/H/S	AUUUAU(UU)	T/H/S	UAAAUG(UD)	2.33E−10	brain
T/H/S	AUUUAC(DU)	T/H/S	AAUAAA(DD)	3.04E−9	lymph node
T/H/S	AAAUUU(UD)	T/H/S	UUUUUU(DU)	3.71E−7	lymph node
T/H/S	AAUAUG(UU)	T/H/S	AAAAAU(UD)	3.75E−5	heart
T/H/S	AUUUAG(UD)	T/H/S	UUAUAU(DU)	1.02E−4	testes
T/H/S	UUUAAU(UD)	T/H/S	UUUUAU(DD)	1.45E−4	adipose
T/H/S	AAAUUC(UU)	T/H/S	AAAUUU(DU)	6.08E−4	lymph node
T/H/S	UUUUAA(UD)	T/H/S	CAUUAU(DU)	8.93E−4	adipose
T/H/S	UUUAAU(UU)	T/H/S	AAAAUA(UD)	1.09E−3	heart
T/H/S	GUUUUA(UD)	T/H/S	UUUUAU(DD)	1.90E−3	adipose
T/H/S	AAUAUU(UD)	T/H/S	AUUUUG(DD)	3.06E−3	breast
T/H/S	UAUUUA(UD)	QKI	AACUAA(DU)	2.47E−6	liver
T/H/S	AUUUAA(UU)	PTB	CUUUUC(UD)	6.12E−4	heart

T/H/S represent TIA-1/TIAR, Hu family or Sam68 proteins.
doi:10.1371/journal.pone.0054885.t003

be detected in our result as an interaction, which implies the important role of interaction in splicing regulation.

Discussion

Our regression model for splicing regulation is derived from the same biophysical principle regarding protein and nucleic acids interaction as the one used to derive the model for transcription [26,31–33]. However, our model uses the ratio of the expression levels of two isoforms or equivalently the ratio of binding and unbinding probabilities of the spliceosome to the mRNA as the response variable, whereas the model for transcription uses the gene expression level or equivalently the binding probability of the RNAP to the promoter as the response variable. For this reason, our model for splicing is linear with respect to unknown parameters to be inferred, whereas the model for transcription is nonlinear. While the nonlinear model for transcription with relatively small number of unknown parameters can be directly inferred [26], a linear approximation [34–37] and an nonlinear approximation using splines [18,38] have been proposed to facilitate model inference. It was shown that the spline approximation [37,38] can offer significantly better performance than the linear approximation in terms of the variance explained by the model. The linear regression model [19] and the spline regression model [18] for splicing regulation use the expression level of an isoform as the response variable. Therefore these two models are also approximate models. In contrast, our regression model is directly derived from the biophysical principle and the linearity of our model with respect to unknown parameters enables efficient model inference even when the number of unknown parameters is very large. This explains why the variance explained by our model is comparable to the best result achieved by the nonlinear biophysical model for transcription [26]. Our model may be improved to explain more variance, for example, by including interactions involved more than two SREs, by accounting for the number of occurrences of each SRE, and by including SREs of length not necessarily equal to six nucleotides, although this can increase the complexity of model inference. On the other hand, if we are interested in the identification of SREs interacting with the binding sites of a specific SF, we can design well-controlled experiments which measure the splicing profile of all the genes before and after knockdown of a specific SF, and then apply our model to identify SRE pairs related to the SF.

To the best of our knowledge, this is the first time that the regression-based approach is employed to systematically identify cooperative SRE pairs. Moreover, our regression framework can identify SRE pairs without being affected by GC content. Current methods for identifying cooperative SRE pairs in splicing regulation use hypergeometric test to find the co-occurrence of SRE pairs over-presented in different regions [22–24]. Since they only use sequence information, some sequence features that are not related to splicing regulation may increase the FDR. For example, without correction of GC content, the majority of the motif pairs detected in [22,23] with high p-values share similar GC contents, being either GC-rich or AU-rich. For this reason, they corrected the GC content by grouping the ASEs with similar GC content [22,23]. However, If AU-rich or GC-rich SRE pairs indeed have regulatory effects, correction for GC content might introduce bias and under- or over-estimate the statistical significance of the SRE pairs. In fact, it has been shown that the AU-rich motif is conserved in mammalian introns [17] and could be bound by TIA-1/TIAR, Hu protein or Sam68 [49,90]. In our work, since our model uses the isoform expression information in addition to the sequence information, it can automatically handle the GC content problem. If there is no correlation between the splicing response and the GC- or AU-rich SRE pairs, these pairs will not be identified since they are just sequence features irrelevant to splicing regulation. On the other hand, if they do have regulatory effect, our method will identify them as SRE pairs based on their correlation with the response.

Since we want to detect all possible SREs and SRE pairs, our linear regression model contains a very large number of candidate SREs and their pairs as variables. The challenging problem in model inference is to select correct variables without overfitting the data. We employed four techniques to tackle this problem. First, variable screening is used to exclude SREs and SRE pairs that are present in less than 1% of the samples or have no or very small correlation with the response variable. Second, regularized inference methods, the Lasso and the adaptive Lasso, were employed in conjunction with cross-validation to select a small number of SREs and SRE pairs. Third, RCV is used to estimate the residual variance which was further used to prevent the possible overfitting problem. Finally, the FDR was calculated to retain only the most statistically significant SREs and SRE pairs in the final model. Overall, these steps combine the state-of-the-art techniques and form an effective framework to reduce the FDR and prevent model overfitting, without compromising the power of detection.

The SREs and SRE pairs we identified have a significant overlap with a set of SREs identified with experiments. The regulatory effects of several well-defined SREs were correctly inferred from our model. For several different interaction patterns proposed based on the experimental results, our model successfully identified them as SRE pairs in the same regions as the proposed ones and provided more insight into their interactions. Note that we can identify SRE pairs at two ends of intron [23], at two ends of exon [22], and in the same region [24] in one framework, and capture the combinatorial regulatory effects of multiple SREs more faithfully. We also report AU-rich SRE pairs as a putative interaction pattern that is important and prevalent in human splicing regulation. In summary, our biophysical regression model provides a useful platform for discovering splicing regulators and unraveling splicing regulatory mechanisms.

Materials and Methods

The Biophysical Model

Suppose a gene contains an ASE, and it can generate an isoform I_1 with the ASE or another isoform I_2 without the ASE. Since splicing is coupled with transcription and the product emerging from this coupled process is either I_1 or I_2, we can consider the splicing of each pre-mRNA independently. We model the multi-step assembly of spliceosome S to the splice sites of the ASE on each individual pre-mRNA as a single chemical reaction:

$$S + \underset{\substack{\downarrow \\ I_2}}{RNA} \rightleftharpoons \underset{\substack{\downarrow \\ I_1}}{RNA} \cdot S, \tag{3}$$

where RNA denotes the state in which S is not fully assembled, and RNA represents the state in which S is fully assembled around the ASE. Then I_1 is produced from the RNA state and I_2 is produced from the RNA state. The probability of having a spliceosome assembled around the ASE is equal to the probability of the RNA state in reaction (3), which can be expressed as $P_s = \frac{[RNA \cdot S]}{[RNA] + [RNA \cdot S]}$, where [RNA·S], [RNA.and [S.stand for the concentrations of RNA·S, RNA and S, respectively. Since the equilibrium constant of reaction (3) is given by $k_s = \frac{[RNA \cdot S]}{[RNA][S]}$, we can write P_s as follows:

$$P_s = \frac{k_s[S]}{1 + k_s[S]} = \frac{q_s}{1 + q_s}, \tag{4}$$

where $q_s = k_s[S]$. Similar to the gene expression model [36,38], the dynamic changes of the concentration of I_1 denoted as E_{I_1} can be written as:

$$\frac{dE_{I_1}}{dt} = k_g P_s - k_d E_{I_1}, \tag{5}$$

where k_g and k_d are synthesis and degradation rates, respectively. In the steady state where $dE_{I_1}/dt = 0$, we have $E_{I_1} = \frac{k_g}{k_d} P_s = \alpha P_s$, where $\alpha = k_g/k_d$. Likewise, concentration of the second isoform $E_{I_2} = \gamma(1 - P_s)$, where γ is another constant. Thus, the ratio of E_{I_1} and E_{I_2} can be written as:

$$\frac{E_{I_1}}{E_{I_2}} = \frac{\alpha}{\gamma} \frac{P_s}{1 - P_s} = c \cdot q_s, \tag{6}$$

where constant $c = \alpha/\gamma$. Now we consider an SF that can bind to an SRE around the ASE and influence the assembly of the spliceosome. The pre-mRNA can have four possible states: 1) bound by both S and SF (RNA), 2) bound by S only (RNA), 3) bound by SF only (RNA), and 4) bound by neither of them (RNA). Then the probability P_s of having a spliceosome assembled around the ASE is equal to the probability of states 1) and 2). Following [26,31,32], we can write P_s as $P_s = (z_1 + z_2)/(z_1 + z_2 + z_3 + z_4)$, where z_i is the Boltzmann weight for state i. Let w be the cooperative factor reflecting the interaction between SF and S, $q_{sf} = k_{sf}[SF]$ where $k_{sf} = \frac{[RNA \cdot SF]}{[RNA][SF]}$, then it is not difficult to find that $z_1 = w q_s q_{sf}$, $z_2 = q_s$, $z_3 = q_{sf}$ and $z_4 = 1$, which yields:

$$P_s = \frac{q_s + w q_s q_{sf}}{1 + q_{sf} + q_s + w q_s q_{sf}}. \tag{7}$$

If $w = 1$, SF and S bind to the transcript independently and P_s in (7) is simplified to that in (4). If $w > 1$, the binding of SF to SRE increases the probability of spliceosome assembly, which implies that the SF is an enhancer. If $w < 1$, binding of the SF has a negative effect on spliceosome assembly and the SF is a repressor. The ratio of the expression levels of I_1 and I_2 can be written as:

$$\frac{E_{I_1}}{E_{I_2}} = \frac{\alpha}{\gamma} \frac{P_s}{1 - P_s} = c \cdot q_s \frac{1 + w q_{sf}}{1 + q_{sf}}. \tag{8}$$

If two SFs can cooperatively bind to their SREs around the ASE and interact with the spliceosome, it is not difficult to derive the following ratio:

$$\frac{E_{I_1}}{E_{I_2}} = c \cdot q_s \frac{1 + w_1 q_{sf_1} + w_2 q_{sf_2} + w_{12} w_1 w_2 q_{sf_1} q_{sf_2}}{1 + q_{sf_1} + q_{sf_2} + w_{12} q_{sf_1} q_{sf_2}}, \tag{9}$$

where $q_{sf_1} = k_{sf_1}[SF_1]$ with $k_{sf_1} = \frac{[RNA \cdot SF_1]}{[RNA][SF_1]}$, q_{sf_2} and k_{sf_2} are defined similarly for SF_2, $w_i, i = 1,2$, is the cooperativity factor between SF_i and S, and w_{12} is the cooperativity factor between two SFs. If $w_{12} = 1$, there is no cooperative interaction between two SFs, and they enhance or repress spliceosome assembly independently. In this case, (9) can be simplified as

$$\frac{E_{I_1}}{E_{I_2}} = c \cdot q_s \frac{1 + w_1 q_{sf_1}}{1 + q_{sf_1}} \frac{1 + w_2 q_{sf_2}}{1 + q_{sf_2}}. \tag{10}$$

If $w_{12} \neq 1$, we can express (9) as:

$$\frac{E_{I_1}}{E_{I_2}} = c \cdot q_s \frac{1 + w_1 q_{sf_1}}{1 + q_{sf_1}} \frac{1 + w_2 q_{sf_2}}{1 + q_{sf_2}} \phi, \tag{11}$$

where $\phi = (\frac{1 + w_1 q_{sf_1} + w_2 q_{sf_2} + w_{12} w_1 w_2 q_{sf_1} q_{sf_2}}{1 + q_{sf_1} + q_{sf_2} + w_{12} q_{sf_1} q_{sf_2}}) / (\frac{1 + w_1 q_{sf_1}}{1 + q_{sf_1}} \frac{1 + w_2 q_{sf_2}}{1 + q_{sf_2}})$. If we define $b_1 = \log(\frac{1 + w_1 q_{sf_1}}{1 + q_{sf_1}})$, $b_2 = \log(\frac{1 + w_2 q_{sf_2}}{1 + q_{sf_2}})$ and $b_{12} = \log(\phi)$, we can write (11) as:

$$\log(\frac{E_{I_1}}{E_{I_2}}) = \log(c \cdot q_s) + b_1 + b_2 + b_{12}. \tag{12}$$

The first term reflects the basal level of splicing determined by the spliceosome alone, the second and third terms are the effects of interactions between the spliceosome and each individual SF, while the last term is the effect of the interaction between two SFs.

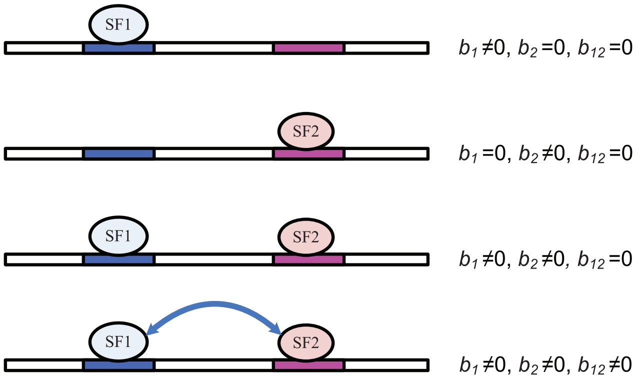
SF1
$b_1 \neq 0$, $b_2 = 0$, $b_{12} = 0$
SF2
$b_1 = 0$, $b_2 \neq 0$, $b_{12} = 0$
SF1
SF2
$b_1 \neq 0$, $b_2 \neq 0$, $b_{12} = 0$
SF1
SF2
$b_1 \neq 0$, $b_2 \neq 0$, $b_{12} \neq 0$

Figure 4. Illustration of equation (12). Two possible SREs are considered in this example, one for SF1, and the other one for SF2. Four different conditions are shown. Binding of either SF can affect the probability of spliceosome assembly. The arrow connecting two SFs indicates the interaction between two SFs. The contribution to spliceosome assembly from SFs is represented by b_1, b_2 and b_{12}.
doi:10.1371/journal.pone.0054885.g004

This can be seen from the fact that $b_i = 0$ if $w_i = 1$ and $b_{12} = 0$ if $w_{12} = 1$. In other words, if the ith SRE affects splicing, then $b_i \neq 0$; otherwise $b_i = 0$; Similarly, if two SREs interact with each other and affect splicing jointly, then $b_{12} \neq 0$; otherwise $b_{12} = 0$. Four different conditions are depicted in Figure 4.

Note that $b_i, i = 1,2$, is determined by k_{sf_i}, $[SF_i]$ and w_i. Since an SF can bind to the same set of SREs around different ASEs, we assume that k_{sf_i} and w_i are the same for different ASEs. Therefore, if we consider a set of ASEs in the same tissue or under the same condition, where $[SF_i]$ is fixed, b_i is identical for these ASEs. Similarly, we assume that b_{12} is a constant for different ASEs in the same tissue. On the other hand, since different exons may have strong or weak splice sites and different genes may have different degradation rates, the first term $\log(c \cdot q_s)$ in (12) may be different for different exons even in the same tissue or under the same condition. Since our goal is to infer b_1, b_2 and b_{12} from data of multiple ASEs in the same tissue, we need to remove the exon-specific effects from the model.

If expression levels of isoforms in two tissues t_1 and t_2 are available, the first term in (12) is identical in these two tissues, and can be removed by forming the following model:

$$\log\left(\frac{E_{I_1}^{t_1}}{E_{I_2}^{t_1}}\right) - \log\left(\frac{E_{I_1}^{t_2}}{E_{I_2}^{t_2}}\right) = (b_1^{t_1} - b_1^{t_2}) + (b_2^{t_1} - b_2^{t_2}) + (b_{12}^{t_1} - b_{12}^{t_2}), \quad (13)$$

The data of tissue t_2 can be regarded as a reference. Subtraction of the reference data from the data of tissue t_1 removes the exon-specific effects. When data of multiple tissues are available, we can arbitrarily choose a tissue as the reference. However, since the expression level of each isoform is estimated from the noisy measurements, a better reference can be obtained by averaging the data of multiple tissues, which is similar to the strategy used in [18]. Specifically, suppose we have a set of data $E_{I_1}^t, E_{I_2}^t$ for tissue $t = 1, \ldots, T$, we can remove the first term in (12) by forming the following model:

$$\begin{aligned}
&\log\left(\frac{E_{I_1}^{t_1}}{E_{I_2}^{t_1}}\right) - \frac{1}{T-1}\sum_{\substack{t=1 \\ t \neq t_1}}^{T} \log\left(\frac{E_{I_1}^{t}}{E_{I_2}^{t}}\right) \\
&= \Big(b_1^{t_1} - \frac{1}{T-1}\sum_{\substack{t=1 \\ t \neq t_1}}^{T} b_1^t\Big) + \Big(b_2^{t_1} - \frac{1}{T-1}\sum_{\substack{t=1 \\ t \neq t_1}}^{T} b_2^t\Big) \qquad (14) \\
&+ \Big(b_{12}^{t_1} - \frac{1}{T-1}\sum_{\substack{t=1 \\ t \neq t_1}}^{T} b_{12}^t\Big).
\end{aligned}$$

If we define $y = \log\left(\frac{E_{I_1}^{t_1}}{E_{I_2}^{t_1}}\right) - \frac{1}{T-1}\sum_{\substack{t=1 \\ t \neq t_1}}^{T} \log\left(\frac{E_{I_1}^{t}}{E_{I_2}^{t}}\right)$, $\beta_1 = b_1^{t_1} - \frac{1}{T-1}\sum_{\substack{t=1 \\ t \neq t_1}}^{T} b_1^t$, $\beta_2 = b_2^{t_1} - \frac{1}{T-1}\sum_{\substack{t=1 \\ t \neq t_1}}^{T} b_2^t$, and $\beta_{12} = b_{12}^{t_1} - \frac{1}{T-1}\sum_{\substack{t=1 \\ t \neq t_1}}^{T} b_{12}^t$, then (14) can be simplified as:

$$y = \beta_1 + \beta_2 + \beta_{12}. \quad (15)$$

So far, we have assumed that y can be measured without any error. If the measurement error is taken into account, equation (15) becomes:

$$y = \beta_1 + \beta_2 + \beta_{12} + \epsilon, \quad (16)$$

where ε is the measurement error modeled as a Gaussian random variable with zero mean. Model (16) is derived under the assumption that two SREs are present to regulate splicing. Since we do not know which SRE or SRE pair contributes to splicing, we can include all potential SREs and their pairwise interaction in the model by adding a parameter to the model for each potential SRE and SRE pairs. Moreover, when we use this model to identify SREs and their interactions, we need to apply it to a set of ASEs. However, different ASEs may have different SREs. To overcome this problem, we include all possible SREs (typically hexamers) and their pairwise interactions in the model, but multiply β_i by a binary variable $x_i \in \{0,1\}$ that indicates if the corresponding SRE is present in the ith ASE, and similarly multiply β_{ij} by $x_i x_j$. This gives rise to the model in (2). We can also include interactions involving more than two SREs, but this will dramatically increase the number of unknowns that is already very large, which will make model inference extremely difficult if not impossible. For this reason, we only include pairwise interactions in our model.

Inference of Regulatory Effects

Our model inference framework will infer β_i and β_{ij} in regression model (2). If β_i or β_{ij} is not equal to zero with certain statistical significance, then we determine that the ith element or the pair involving the ith and jth elements has regulatory effect. However, it is unclear whether it is an enhancer or silencer, since it is the sign of $w_i - 1$ or $w_{ij} - 1$, not the sign of β_i or β_{ij} that determines an enhancing or inhibitory effect. We will next show that we can infer the regulatory effect of the ith SRE from β_i and $[SF_i]$ that is the concentration of the SF that can bind to the SRE. The enhancing or inhibitory effect of an SRE pair however is difficult to infer.

Let us first consider the situation where only one reference sample is used as in (13). In this case, $\beta_i = b_i^{t_1} - b_i^{t_2} = f([SF_i^{t_1}]; w_i) - f([SF_i^{t_2}]; w_i)$, where $f([SF_i^t]; w_i) = \log(\dfrac{1 + w_i k_{sf_i}[SF_i^t]}{1 + k_{sf_i}[SF_i^t]})$. Then it can be easily shown that the sign of β_i is the same as the sign of $(w_i - 1)([SF_i^{t_1}] - [SF_i^{t_2}])$. Therefore, if $([SF_i^{t_1}] - [SF_i^{t_2}])\beta_i > 0$, then $w_i > 1$ and SRE i is an enhancer; otherwise, $w_i < 1$ and SRE i is a silencer.

For model (2) or (15) where multiple tissues are used as the reference, the following proposition can be used to infer the regulatory effect of an SRE:

Proposition 1 *Given the condition*

$$[SF_i^{t_1}] > \frac{1}{T-1} \sum_{t \neq t_1} [SF_i^t], \tag{17}$$

we have $w_i > 1$ *if* $\beta_i > 0$, *or* $w_i < 1$ *if* $\beta_i < 0$. *Given the condition*

$$[SF_i^{t_1}] < \left[\frac{1}{T-1} \sum_{t \neq t_1} \frac{1}{[SF_i^t]} \right]^{-1}, \tag{18}$$

we have $w_i > 1$ if $\beta_i < 0$, or $w_i < 1$ if $\beta_i > 0$.

Before we proceed to prove Proposition 1, we make the following comments. The sign of β_i alone can not determine if the SRE is an enhancer ($w_i > 1$) or silencer ($w_i < 1$). It should be used together with condition (17) and (18) to infer the regulatory effect of the SRE. It is difficult to determine the regulatory effect of an SRE interacting pair bound by two SFs.

Proof: For simplicity, we will omit subscript i throughout the proof. Define $b = f(q^t) = \log(\dfrac{1 + wq^t}{1 + q^t})$, where $q^t = k_{sf}[SF^t]$ as defined earlier. Then $\dfrac{df(q^t)}{dq^t} = \dfrac{(w-1)}{(1 + wq^t)(1 + q^t)}$. If $w > 1, f(q^t)$ is a monotonically increasing function; otherwise, $f(q^t)$ is a monotonically decreasing function (Figure 5).

Define q_w such that $f(q_w) = \dfrac{1}{T-1} \sum_{t \neq t_1} f(q^t)$, which gives:

$$q_w = \frac{1 - \prod_{t \neq t_1} (\frac{1 + wq^t}{1 + q^t})^{\frac{1}{T-1}}}{\prod_{t \neq t_1} (\frac{1 + wq^t}{1 + q^t})^{\frac{1}{T-1}} - w}, \tag{19}$$

where $0 < w < 1$ or $w > 1$. From (19), we obtain the following result:

$$q_0 = \lim_{w \to 0+} q_w = \prod_{t \neq t_1} (1 + q^t)^{\frac{1}{T-1}} - 1 < \frac{1}{T-1} \sum_{t \neq t_1} q^t,$$

$$q_\infty = \lim_{w \to \infty} q_w = \frac{1}{\prod_{t \neq t_1} (\frac{1 + q^t}{q^t})^{\frac{1}{T-1}} - 1} > \frac{1}{\frac{1}{T-1} \sum_{t \neq t_1} (\frac{1 + q^t}{q^t}) - 1} = \frac{T-1}{\sum_{t \neq t_1} \frac{1}{q^t}},$$

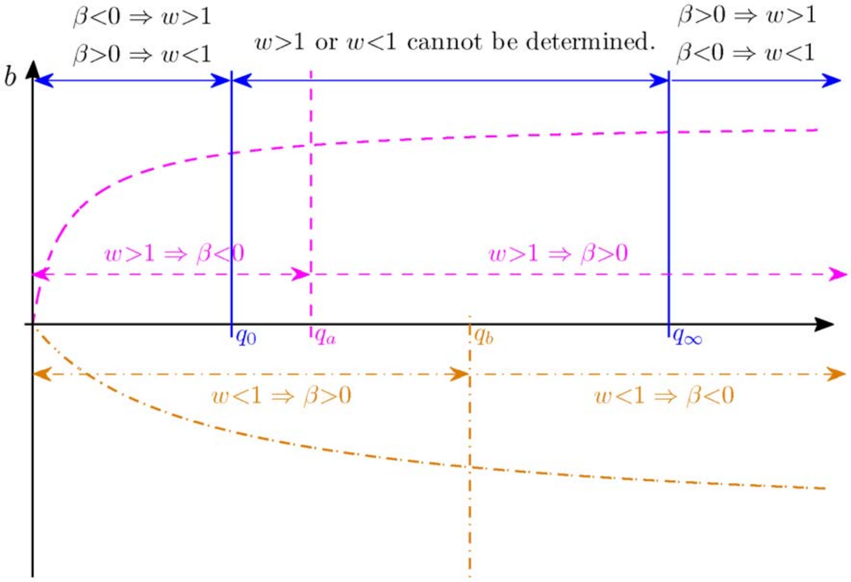
$\beta<0 \Rightarrow w>1$
$\beta>0 \Rightarrow w<1$
$w>1$ or $w<1$ cannot be determined.
$\beta>0 \Rightarrow w>1$
$\beta<0 \Rightarrow w<1$
b
$w>1 \Rightarrow \beta<0$
$w>1 \Rightarrow \beta>0$
q_0
q_a
q_b
q_∞
$w<1 \Rightarrow \beta>0$
$w<1 \Rightarrow \beta<0$

Figure 5. Illustration of Proposition 1. The dashed curve is $b = f(q) = \log(\dfrac{1 + wq}{1 + q})$ for $w > 1$ and the dashed lines with arrows define the region where $\beta < 0$ or $\beta > 0$. The dash-dot curve is $b = f(q) = \log(\dfrac{1 + wq}{1 + q})$ for $w < 1$ and the dash-dot lines with arrows define the region where $\beta > 0$ or $\beta < 0$. The solid lines define the decision region of w described in Proposition 1.
doi:10.1371/journal.pone.0054885.g005

Define $q_a = q_w$ for $w > 1$ and $q_b = q_w$ for $0 < w < 1$. In Lemma 1 given later at the end of this section, we prove that q_w is a monotonically decreasing function of w. Therefore, we have the following inequalities:

$$\frac{T-1}{\sum\limits_{t \neq t_1} \frac{1}{q^t}} < q_a < q_b < \frac{1}{T-1} \sum_{t \neq t_1} q^t. \tag{20}$$

From (14) and (15), we have $\beta = f(q^{t_1}) - \frac{1}{T-1} \sum\limits_{t \neq t_1} f(q^t) = f(q^{t_1}) - f(q_w)$. When $w > 1$, due to the fact that $f(q^t)$ is an increasing function, we have $\beta > 0$ if $q^{t_1} > q^a$ or $\beta < 0$ if $q^{t_1} < q^a$. Similarly, when $w < 1$, we have $\beta < 0$ if $q^{t_1} > q^b$ or $\beta > 0$ if $q^{t_1} < q^b$. This is illustrated in Figure 5. Now suppose that $[SF^{t_1}] > \frac{1}{T-1} \sum\limits_{t \neq t_1} [SF^t]$, since $q^t = k_{sf}[SF^t]$, we have $q^{t_1} > \frac{1}{T-1} \sum\limits_{t \neq t_1} q^t$. Using (20), we have $q^{t_1} > q_b > q_a$. Therefore, we can infer that $w > 1$ if $\beta > 0$, or $w < 1$ if $\beta < 0$ as illustrated in Figure 5. Similarly, if $[SF^{t_1}] < [\frac{1}{T-1} \sum\limits_{t \neq t_1} \frac{1}{[SF^t]}]^{-1}$, we have $q^{t_1} < \frac{T-1}{\sum\limits_{t \neq t_1} \frac{1}{q^t}}$, and thus, $q^{t_1} < q_a < q_b$. We can infer that $w > 1$ if $\beta < 0$, or $w < 1$ if $\beta > 0$, again as illustrated in Figure 5. Note that q_a and q_b are determined by unknown parameter w and they can be anywhere in between q_∞ and q_0. Therefore, if $[\frac{1}{T-1} \sum\limits_{t \neq t_1} \frac{1}{[SF^t]}]^{-1} < [SF^{t_1}] < \frac{1}{T-1} \sum\limits_{t \neq t_1} [SF^t]$, we can not determine whether $w > 1$ or $w < 1$.

Lemma 1

$$h(x) = \frac{1 - \prod\limits_{i=1}^{N} (\frac{1+xq_i}{1+q_i})^{\frac{1}{N}}}{\prod\limits_{t=1}^{N} (\frac{1+xq_i}{1+q_i})^{\frac{1}{N}} - x}. \tag{21}$$

is a monotonically decreasing function for $x \in (0,1) \bigcup (1,\infty)$, *when* q_i, $i = 1,\ldots,N$ *are positive.*

Proof: Define $g(x) = \Pi\,(\frac{1+xq_i}{1+q_i})^{\frac{1}{N}}$. Then $\frac{dh(x)}{dx} = \frac{g'(x)(x-1) - g(x) + 1}{[g(x) - x]^2}$, where $g'(x) = \frac{dg(x)}{dx}$. Let us define the numerator of $\frac{dh(x)}{dx}$ as $J(x)$, since the denominator is positive, we next prove that $J(x) \leq 0$, which implies that $\frac{dh(x)}{dx} \leq 0$ and therefore $h(x)$ is a decreasing function.

$$J(x) = g'(x)(x-1) - g(x) + 1$$

$$= \frac{1}{N} \sum_{i=1}^{N} \frac{q_i}{1+xq_i} g(x)(x-1) - g(x) + 1$$

$$= \left[\frac{1}{N} \sum_{i=1}^{N} \frac{q_i}{1+xq_i}(x-1) - 1 + \frac{1}{g(x)}\right] g(x)$$

$$= \left[\frac{1}{N} \sum_{i=1}^{N} \frac{q_i}{1+xq_i}(x-1) - 1 + \prod_{i=1}^{N} (\frac{1+q_i}{1+xq_i})^{\frac{1}{N}}\right] g(x)$$

$$\leq \left[\frac{1}{N} \sum_{i=1}^{N} \frac{1+xq_i}{1+xq_i} - 1\right] g(x)$$

$$= 0,$$

where the inequality is due to the facts that $g(x) > 0$ and that $\prod\limits_{i=1}^{N} (\frac{1+q_i}{1+xq_i})^{\frac{1}{N}} \leq \frac{1}{N} \sum\limits_{i=1}^{N} (\frac{1+q_i}{1+xq_i})$, since $x > 0$ and $q_i > 0, i = 1,\ldots,N$. Thus, $h(x)$ is monotonically decreasing in (0,1) and $(1,\infty)$. Moreover, since we have:

$$\lim_{x \to 1^-} h(x) = \lim_{x \to 1^+} h(x) = \frac{\sum\limits_{i=1}^{N} \frac{q_i}{1+q_i}}{N - \sum\limits_{i=1}^{N} \frac{q_i}{1+q_i}},$$

$h(x)$ is a monotonically decreasing function for $x \in (0,1) \bigcup (1,\infty)$, and $\frac{dh(x)}{dx} = 0$ if and only if $q_1 = q_2 = \ldots = q_N$.

ASE Selection

The KnownGene table of human February 2009 assembly (hg19) was downloaded from the University of California Santa Cruz (UCSC) genome database [39]. We chose UCSC Known Genes as the reference gene annotation, since they contain a comprehensive gene set that is constructed mostly from experimental data in Genbank and Uniprot [92]. For each gene, the KnownGene table gives all known isoforms of its mRNA transcripts. An exon was selected as an ASE in our dataset if the following five criteria were satisfied: 1) at least one isoform includes the exon, 2) at least one isoform does not include the exon, 3) the upstream $5'$ splice site is the same in all isoforms, and similarly the downstream $3'$ splice site is the same in all isoforms, as illustrated in Figure 6A, 4) the upstream $3'$ splice site is the same in all isoforms with the ASE, and similarly the downstream $5'$ splice site is the same in all isoforms with the ASE, as illustrated in Figure 6A, and 5) both the upstream and downstream introns are of ≥ 400 nts. With these strict criteria, the five regions of the ASE shown in Figure 1A are defined without ambiguity. Note that isoforms in Figure 6B do not satisfy criterion 3), and thus ASEs with isoforms in Figure 6A and 6B are not included in our data set. Similarly, isoforms in Figure 6C do not satisfy criterion 4), and ASEs with isoform in Figure 6A and 6C are not included in our data set. To ensure a reliable estimate of the expression ratio, we only kept ASEs with gene expression level greater than 3 RPKM (reads per kilobases per million mapped reads). This gave a set of ASEs for each tissue. The number of ASEs for each tissue is given in Table 1 and more detailed description of the ASEs including their genomic coordinates are given in Table S2. Most ASEs were used for model inference in almost all tissues. Some ASEs were not used in a specific tissue because they did not pass the minimum expression requirement in that tissue. Note that although almost the same set of ASEs were used, the splicing response variable y of the same

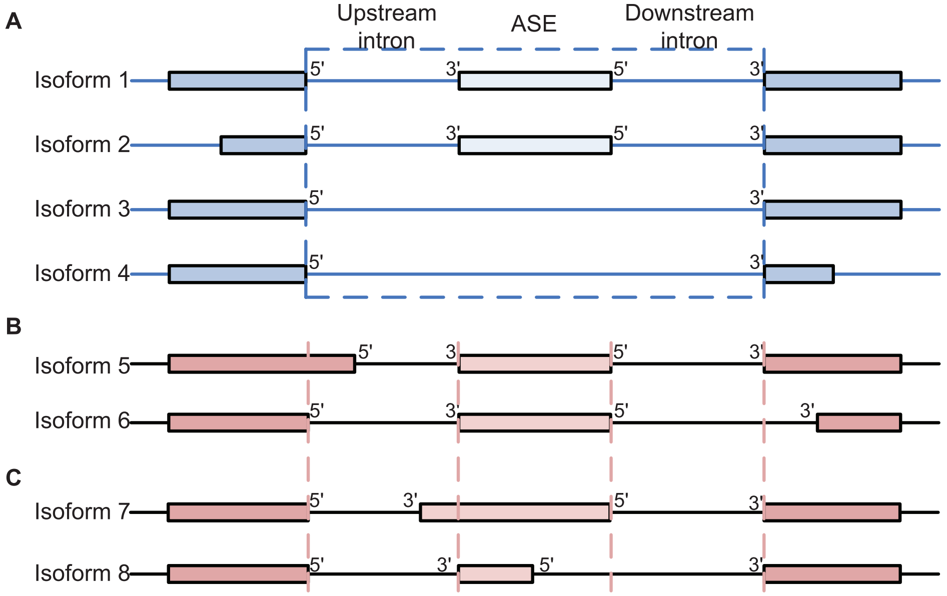

Figure 6. Illustration of ASE selection criteria 3 and 4. (**A**) ASEs that satisfy both criteria 3 and 4. (**B**) ASEs that do not satisfy criterion 3. ASEs in isoforms in both (A) and (B) are not included in our ASE data set. (**C**) ASEs that do not satisfy criterion 4. ASEs in isoforms in both (A) and (C) are not included in our ASE data set.
doi:10.1371/journal.pone.0054885.g006

ASE was different in different tissues, and thus the data used for model inference were in fact different for different tissues.

RNA-Seq Data

The data set in [1] includes RNA-Seq reads from 9 tissues: adipose, whole brain, breast, colon, heart, liver, lymph node, skeletal muscle and testes, as well as several cerebellar cortex samples and cell lines. We only used RNA-Seq data of 9 tissues, which contains over 200 million reads of 32 nts, to detect SREs and cooperative SRE pairs.

Estimation of Expression Level and Inclusion Ratio

We started by mapping the RNA-Seq reads against an expanded human genome (hg19) downloaded from the UCSC genome database, allowing up to two mismatches, using Bowtie (version 0.12.7) [93]. The expanded human genome consists of the UCSC hg19 whole genome reference sequence and the 56 nt long splice-crossing sequences for each exon junction documented in the UCSC KnownGene table. Reads that could be mapped to multiple loci of the genome were excluded, and 140 million uniquely mapped reads were kept for the following analysis.

We next calculated the expression level of each isoform including or excluding a selected ASE in 9 tissues using the algorithm of Jiang *et al.* [94]. Since we only kept uniquely mapped reads, for an exon of length l, we used an effective exon length $l-r-m$, where r is the read length and m is the number of multi-mappable positions of the exon. To find out m, we re-mapped all possible 32-nt subsequences of candidate ASEs and splice junctions against the same expanded genome described above using Bowtie [93]. Moreover, to minimize the effect of non-uniformity of read distribution [95], we only used three exons, including the ASE itself, the adjacent upstream and downstream exons to estimate the expression level of each isoform.

After the expression level of each isoform of the selected gene was calculated, the inclusion ratio (IR) of an ASE in a specific tissue was calculated as the ratio of the expression level of the isoforms with the ASE to the total expression level of all isoforms of the gene, *i.e.*, $IR = \frac{E_{I_1}}{E_{I_1}+E_{I_2}}$, where E_{I_1} is the total expression level of isoforms including the ASE and E_{I_2} is the total expression level of isoforms excluding the ASE.

RNA Sequence Elements

For each tissue and each ASE, we extracted all hexamers in five regions around the ASE, including the 200 nts intronic region adjacent to the upstream 5′ splice site (UU in Figure 1A), the 200 nts intronic region adjacent to the upstream 3′ splice site (UD in Figure 1A), the ASE region (EXON in Figure 1A), the 200 nts intronic region adjacent to the downstream 5′ splice site (DU in Figure 1A) and the 200 nts intronic region adjacent to the downstream 3′ splice site (DD in Figure 1A). The EXON region is the ASE itself if the ASE is less than 200 nts; otherwise, it is the combination of the first and last 100 nts of the ASE. Since the 5′ and 3′ splice sites have the consensus sequences MAG/GURAGU and Y_{10}NCAG/G [11,96], respectively, we excluded the sequences in the window from -3 to 6 around the 5′ splice site and in the window from -14 to 1 around the 3′ splice site in our analysis.

Variable Screening

Before applying the Lasso, we used a strategy similar to the sure independence screening [40] to reduce the dimensionality of the feature space, thereby improving variable selection in terms of both speed and accuracy. Specifically, for the ith hexamer, $i \in (1,\ldots,4^6)$, we used the following simple linear regression to test the correlation between its presence in one of the five regions of ASEs with the response variable:

$$y_e = \beta_0 + x_{ei}\beta_i + \epsilon_e, \tag{22}$$

where $x_{ei}, e=1,\ldots,n$ is a binary variable to indicate if the ith hexamer is present ($x_{ei}=1$) or absent ($x_{ei}=0$) in one of the five regions of the eth ASE, and y_e is the splicing response of the eth ASE as defined earlier, and $\epsilon_e, e=1,\ldots,n$ are independent and identically distributed normal random variables. In some samples, we have inclusion ratio $IR_e=1$ or $IR_e=0$, usually due to the low read abundance of the minor isoform. For these samples, we set $\log(\frac{E_{e,I_1}}{E_{e,I_2}})=10$ for $IR_e=1$ or set $\log(\frac{E_{e,I_1}}{E_{e,I_2}})=-10$ for $IR_e=0$, which is equivalent to $IR_e\approx 0.9999$ or $IR_e\approx 1e^{-5}$. Hexamers having a significant correlation with a p-value < 0.05 were kept in set $\mathcal{M}$ for further analysis with the Lasso and the adaptive Lasso.

In the next step, for each pair of the retained hexamers, their interaction was tested using the model:

$$y_e=\beta_0+x_{ei}\beta_i+x_{ej}\beta_j+x_{ei}x_{ej}\beta_{ij}+\epsilon_e. \tag{23}$$

Interaction terms with a p-value < 0.05 were also kept in set $\mathcal{I}$ for further analysis. To reduce the possible false positive effects, we also required that the co-occurrence frequency of the two hexamers in an interaction pair was significant (p-value<0.05 from a hypergeometric test based on the null hypothesis that the presence of the first hexamer is independent of the presence of the second hexamer) in the five regions of the selected ASEs defined earlier, and that any hexamer or hexamer-pair must be present in at least 1% of the ASEs.

The Lasso and the Adaptive Lasso

We define $\mathbf{y}=(y_1,y_2,...,y_e,...,y_n)^T$ and $\mathbf{x}_i=(x_{1i},x_{2i},...,x_{ei},...,x_{ni})^T$, where y_e and x_{ei} are defined earlier. We also define $\mathbf{x}_i.*\mathbf{x}_j$ as element-wise multiplication of two vectors. The Lasso procedure was performed by solving the following problem [41]:

$$\begin{aligned}\{\hat\beta_0,\hat\beta_i,\hat\beta_{ij}\}=\underset{\beta_0,\beta_i,\beta_{ij}}{\operatorname{argmax}}\{\|\mathbf{y}-\beta_0-\sum_{i\in\mathcal{M}}\mathbf{x}_i\beta_i-\sum_{(i,j)\in\mathcal{I}}\mathbf{x}_i.*\mathbf{x}_j\beta_{ij}\|^2\\ +\lambda(\sum_{i\in\mathcal{M}}|\beta_i|+\sum_{(i,j)\in\mathcal{I}}|\beta_{ij}|)\}.\end{aligned}\tag{24}$$

The optimal value of parameter λ was obtained using 100-fold cross-validation based on the mean squared prediction error. Then we chose $\hat w_i=1/|\hat\beta_i|$ and $\hat w_{ij}=1/|\hat\beta_{ij}|$ and solved the following adaptive Lasso problem [42]:

$$\begin{aligned}\{\hat\beta'_0,\hat\beta'_i,\hat\beta'_{ij}\}=\underset{\beta'_0,\beta'_i,\beta'_{ij}}{\operatorname{argmax}}\{\|\mathbf{y}-\beta'_0-\sum_{i\in\mathcal{M}}\mathbf{x}_i\beta'_i-\sum_{(i,j)\in\mathcal{I}}\mathbf{x}_i.*\mathbf{x}_j\beta'_{ij}\|^2\\ +\lambda(\sum_{i\in\mathcal{M}}\hat w_j|\beta'_i|+\sum_{(i,j)\in\mathcal{I}}\hat w_{ij}|\beta'_{ij}|)\}.\end{aligned}\tag{25}$$

The optimal value of λ was also obtained using 100-fold cross-validation. We solved these problems using the coordinate descent algorithm of Friedman et al. [97] implemented in the 'glmnet' package.

Refitted Cross-validation

RCV is a technique to estimate residual variance in linear regression models of ultrahigh dimension [44]. In our case, the n samples were randomly split into two even datasets. We applied the Lasso to the first dataset to select a set $\mathcal{V}$ of variables from the variables in $\mathcal{M}$ and $\mathcal{I}$ resulted from the variable screening procedure. We then again used the Lasso to refit the model with the variable set $\mathcal{V}$ to the second dataset. The refitting process selected a set $\mathcal{V}'$ of variables from $\mathcal{V}$. Finally, the variance of the residual error $\hat\sigma_1^2$ is estimated from the second dataset with variables in $\mathcal{V}'$ using the OLS method. We reversed the role of the two datasets, and obtained another estimate of the variance of the residual error, $\hat\sigma_2^2$. The final estimate is then defined as $\hat\sigma^2=(\hat\sigma_1^2+\hat\sigma_2^2)/2$. We repeat this process 100 times by randomly splitting the dataset, and the average variance σ^2_{RCV} of the 100 estimates was the final estimate of the residual variance.

Final Model Selection and Correction for Multiple Testing

The adaptive Lasso procedure produced a sequence of models for different values of λ. For each λ, we extracted the variables of the model and estimated the residual variance σ^2 with these selected variables using the OLS method. For the sequence of models, we selected the one having the smallest variance that was larger than σ^2_{RCV} as the optimal model.

Although the adaptive Lasso selected a set of variable in the final model, it did not give p-value for each variable. We then used the OLS method to refit the model and calculated the p-value for each variable. Based on these p-values, we chose the variables at an FDR ≤ 0.01 [45]. The final model contained these variables, and the percentage of the variance explained (R^2) by this model was calculated as:

$$R^2=1-\sum_{e=1}^{n}(y_e-\hat y_e)^2/\sum_{e=1}^{n}(y_e-\bar y)^2, \tag{26}$$

where $\hat y_e$ is the predicted value of y_e from the final model, and $\bar y$ is the sample mean of y_e. Note that R^2 was also used in [26,34,38] as the figure of merit for performance evaluation. The software package for model inference framework is available under an open source license.

Supporting Information

Table S1 The SREs and SRE pairs identified in 9 tissues.
(XLS)

Table S2 The ASEs used in the analysis.
(XLS)

Dataset S1 The R script implementing the model inference framework.
(GZ)

Author Contributions

Conceived and designed the experiments: XC JW ZC. Performed the experiments: JW. Analyzed the data: JW XC ZC. Wrote the paper: JW XC ZC.

References

1. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470–476.
2. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40: 1413–1415.
3. Cooper TA, Wan L, Dreyfuss G (2009) RNA and disease. Cell 136: 777–793.

4. David CJ, Manley JL (2010) Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev 24: 2343–2364.
5. Chen M, Manley JL (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol 10: 741–754.
6. Licatalosi DD, Darnell RB (2010) RNA processing and its regulation: global insights into biological networks. Nat Rev Genet 11: 75–87.
7. Wang Z, Burge CB (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA 14: 802–813.
8. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, et al. (2003) CLIP identifies Nova-regulated RNA networks in the brain. Science 302: 1212–1215.
9. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, et al. (2004) Systematic identification and analysis of exonic splicing silencers. Cell 119: 831–845.
10. Djordjevic M (2007) SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways. Biomol Eng 24: 179–189.
11. Chasin LA (2007) Searching for splicing motifs. Adv Exp Med Biol 623: 85–106.
12. Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, et al. (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet 40: 1416–1425.
13. Wen J, Chiba A, Cai X (2010) Computational identification of tissue-specific alternative splicing elements in mouse genes from RNA-Seq. Nucleic Acids Res 38: 7895–7907.
14. Sugnet CW, Srinivasan K, Clark TA, O'Brien G, Cline MS, et al. (2006) Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLOS Comput Biol 2: e4.
15. Kabat JL, Barberan-Soler S, McKenna P, Clawson H, Farrer T, et al. (2006) Intronic alternative splicing regulators identified by comparative genomics in nematodes. PLOS Comput Biol 2: e86.
16. Yeo GW, Van Nostrand EL, Liang TY (2007) Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. PLOS Genet 3: e85.
17. Voelker RB, Berglund JA (2007) A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. Genome Res 17: 1023–1033.
18. Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, et al. (2007) A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. Nucleic Acids Res 35: 4845–4857.
19. Wang X, Wang K, Radovich M, Wang Y, Wang G, et al. (2009) Genome-wide prediction of cis-acting RNA elements regulating tissue-specific pre-mRNA alternative splicing. BMC Genomics (Suppl 1): 54.
20. Witten JT, Ule J (2011) Understanding splicing regulation through RNA splicing maps. Trends Genet 27: 89–97.
21. Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, et al. (2012) Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Reports 1: 167–178.
22. Ke S, Chasin L (2010) Intronic motif pairs cooperate across exons to promote pre-mRNA splicing. Genome Biol 11: R84.
23. Friedman BA, Stadler MB, Shomron N, Ding Y, Burge CB (2008) Ab initio identification of functionally interacting pairs of cis-regulatory elements. Genome Res 18: 1643–1651.
24. Suyama M, Harrington ED, Vinokourova S, von Knebel Doeberitz M, Ohara O, et al. (2010) A network of conserved co-occurring motifs for the regulation of alternative splicing. Nucleic Acids Res 38: 7916–7926.
25. Tibshirani R (1997) The lasso method for variable selection in the cox model. Stat Med 16: 385–395.
26. Gertz J, Siggia ED, Cohen BA (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature 457: 215–218.
27. Wahl MC, Will CL, Lührmann R (2009) The spliceosome: design principles of a dynamic RNP machine. Cell 136: 701–718.
28. Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: Towards a cellular code. Nat Rev Mol Cell Biol 6: 386–398.
29. Izquierdo JM, Majos N, Bonnal S, Martinez C, Castelo R, et al. (2005) Regulation of fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. Mol Cell 19: 475–484.
30. Sharma S, Kohlstaedt LA, Damianov A, Rio DC, Black DL (2008) Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. Nat Struct Mol Biol 15: 183–191.
31. Shea MA, Ackers GK (1985) The or control system of bacteriophage lambda: A physical-chemical model for gene regulation. Journal of Molecular Biology 181: 211–230.
32. Buchler NE, Gerland U, Hwa T (2003) On schemes of combinatorial transcription logic. Proc Natl Acad Sci U S A 100: 5136–5141.
33. Djordjevic M, Sengupta AM, Shraiman BI (2003) A biophysical approach to transcription factor binding site discovery. Genome Res 13: 2381–2390.
34. Bussemaker HJ, Li H, Siggia ED (2001) Regulatory element detection using correlation with expression. Nat Genet 27: 167–174.
35. Conlon EM, Liu XS, Lieb JD, Liu JS (2003) Integrating regulatory motif discovery and genome-wide expression analysis. Proc Natl Acad Sci U S A 100: 3339–3344.
36. Das D, Zhang MQ (2007) Predictive models of gene regulation: application of regression methods to microarray data. Methods Mol Biol 377: 95–110.
37. Das D, Pellegrini M, Gray JW (2009) A primer on regression methods for decoding cis-regulatory logic. PLOS Comput Biol 5: e1000269.
38. Das D, Banerjee N, Zhang MQ (2004) Interacting models of cooperative gene regulation. Proc Natl Acad Sci U S A 101: 16234–16239.
39. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The UCSC genome browser database: update 2011. Nucleic Acids Res 39: D876–D882.
40. Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc B 70: 849–911.
41. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58: 267–288.
42. Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101: 1418–1429.
43. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition. Springer Series in Statistics. Heidelberg: Springer.
44. Fan J, Guo S, Hao N (2012) Variance estimation using refitted cross-validation in ultrahigh dimensional regression. J R Stat Soc B 74: 37–65.
45. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57: 289–300.
46. Görlach M, Burd CG, Dreyfuss G (1994) The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein c proteins. J Biol Chem 269: 23074–23078.
47. Burd CG, Dreyfuss G (1994) RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-m rna splicing. EMBO J 13: 1197–1204.
48. Dember LM, Kim ND, Liu KQ, Anderson P (1996) Individual RNA recognition motifs of TIA-1 and TIAR have different RNA binding specificities. J Biol Chem 271: 2783–2788.
49. Lin Q, Taylor SJ, Shalloway D (1997) Specificity and determinants of Sam68 RNA binding. J Biol Chem 272: 27274–27280.
50. Buckanovich RJ, Darnell RB (1997) The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. Mol Cell Biol 17: 3194–3201.
51. Liu HX, Zhang M, Krainer AR (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev 12: 1998–2012.
52. Tacke R, Tohyama M, Ogawa S, Manley JL (1998) Human Tra2 proteins are sequence-specific activators of pre-mrna splicing. Cell 93: 139–148.
53. Cavaloc Y, Bourgeois CF, Kister L, Stévenin J (1999) The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. RNA 5: 468–483.
54. Liu HX, Chew SL, Cartegni L, Zhang MQ, Krainer AR (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. Mol Cell Biol 20: 1063–1071.
55. Thisted T, Lyakhov DL, Liebhaber SA (2001) Optimized RNA targets of two closely related triple KH domain proteins, heterogeneous nuclear ribonucleoprotein K and αCP-2KL, suggest distinct modes of RNA recognition. J Biol Chem 276: 17484–17496.
56. Lerga A, Hallier M, Delva L, Orvain C, Gallais I, et al. (2001) Identification of an RNA binding specificity for the potential splicing factor TLS. J Biol Chem 276: 6807–6816.
57. Peng R, Dye BT, Pérez I, Barnard DC, Thompson AB, et al. (2002) PSF and p54nrb bind a conserved stem in U5 snRNA. RNA 8: 1334–1347.
58. Faustino NA, Cooper TA (2005) Identification of putative new splicing targets for ETR-3 using sequences identified by systematic evolution of ligands by exponential enrichment. Mol Cell Biol 25: 879–887.
59. Galarneau A, Richard S (2005) Target RNA motif and target mRNAs of the quaking STAR protein. Nat Struct Mol Biol 12: 691–698.
60. Hori T, Taguchi Y, Uesugi S, Kurihara Y (2005) The RNA ligands for mouse proline-rich RNA-binding protein (mouse prrp) contain two consensus sequences in separate loop structure. Nucleic Acids Res 33: 190–200.
61. Ponthier JL, Schluepen C, Chen W, Lersch RA, Gee SL, et al. (2006) Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16. J Biol Chem 281: 12468–12474.
62. Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, et al. (2006) An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum Mol Genet 15: 2490–2508.
63. Paradis C, Cloutier P, Shkreta L, Toutant J, Klarskov K, et al. (2007) hnRNP I/PTB can antagonize the splicing repressor activity of SRp30c. RNA 13: 1287–1300.
64. Reid DC, Chang BL, Gunderson SI, Alpert L, Thompson WA, et al. (2009) Next-generation selex identifies sequence and structural determinants of splicing factor binding in human pre-mrna sequence. RNA 15: 2385–2397.
65. Goers ES, Purcell J, Voelker RB, Gates DP, Berglund JA (2010) MBNL1 binds GC motifs embedded in pyrimidines to regulate alternative splicing. Nucleic Acids Res 38: 2467–2484.
66. Ray D, Kazan H, Chan E, Castillo LP, Chaudhry S, et al. (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol 27: 667–670.
67. Kuroyanagi H (2009) Fox-1 family of RNA-binding proteins. Cell Mol Life Sci 66: 3895–3907.
68. Auweter SD, Fasan R, Reymond L, Underwood JG, Black DL, et al. (2006) Molecular basis of rna recognition by the human alternative splicing factor fox-1. EMBO J 25: 163–173.

69. Sauliere J, Sureau A, Expert-Bezancon A, Marie J (2006) The polypyrimidine tract binding protein (PTB) represses splicing of exon 6B from the β-tropomyosin Pre-mRNA by directly interfering with the binding of the U2AF65 subunit. Mol Cell Biol 26: 8755–8769.
70. Zearfoss NR, Clingman CC, Farley BM, McCoig LM, Ryder SP (2011) Quaking regulates hnrnpa1 expression through its 3′ utr in oligodendrocyte precursor cells. PLOS Genet 7: e1001269.
71. Ho TH, Charlet-B N, Poulos MG, Singh G, Swanson MS, et al. (2004) Muscleblind proteins regulate alternative splicing. EMBO J 23: 3103–3112.
72. Caputi M, Zahler AM (2001) Determination of the rna binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. J Biol Chem 276: 43850–43859.
73. Fisette JF, Toutant J, Dugr-Brisson S, Desgroseillers L, Chabot B (2010) hnRNP A1 and hnRNP H can collaborate to modulate 5′ splice site selection. RNA 16: 228–238.
74. Chaudhury A, Chander P, Howe PH (2010) Heterogeneous nuclear ribonucleoproteins (hnrnps) in cellular processes: Focus on hnrnp e1's multifunctional regulatory roles. RNA 16: 1449–1462.
75. Sanford JR, Wang X, Mort M, VanDuyn N, Cooper DN, et al. (2009) Splicing factor sfrs1 recognizes a functionally diverse landscape of rna transcripts. Genome Res 19: 381–394.
76. Zhang C, Zhang Z, Castle J, Sun S, Johnson J, et al. (2008) Defining the regulatory network of the tissue-specific splicing factors fox-1 and fox-2. Genes Dev 22: 2550–2563.
77. Perez I, Lin CH, McAfee JG, Patton JG (1997) Mutation of ptb binding sites causes misregulation of alternative 3′ splice site selection in vivo. RNA 3: 764–778.
78. Llorian M, Schwartz S, Clark TA, Hollander D, Tan LY, et al. (2010) Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. Nat Struct Mol Biol 17: 1114–1123.
79. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, et al. (2010) Deciphering the splicing code. Nature 465: 53–59.
80. Wu JI, Reed RB, Grabowski PJ, Artzt K (2002) Function of quaking in myelination: Regulation of alternative splicing. Proc Natl Acad Sci U S A 99: 4233–4238.
81. Timchenko LT, Miller JW, Timchenko NA, DeVore DR, Datar KV, et al. (1996) Identification of a (CUG)n triplet repeat rna-binding protein and its expression in myotonic dystrophy. Nucleic Acids Res 24: 4407–4414.
82. Miller JW, Urbinati CR, Teng-Umnuay P, Stenberg MG, Byrne BJ, et al. (2000) Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. EMBO J 19: 4439–4448.
83. Kino Y, Mori D, Oma Y, Takeshita Y, Sasagawa N, et al. (2004) Muscleblind protein, MBNL1/EXP, binds specifically to CHHG repeats. Hum Mol Genet 13: 495–507.
84. Oberstrass FC, Auweter SD, Erat M, Hargous Y, Henning A, et al. (2005) Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science 309: 2054–2057.
85. Amir-ahmady B, Boutz PL, Markovtsov V, Phillips ML, Black DL (2005) Exon repression by polypyrimidine tract binding protein. RNA 11: 699–716.
86. Martinez-Contreras R, Fisette JF, Nasim FuH, Madden R, Cordeau M, et al. (2006) Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. PLOS Biol 4: e21.
87. Paziewska A, Wyrwicz LS, Bujnicki JM, Bomsztyk K, Ostrowski J (2004) Cooperative binding of the hnRNP K three KH domains to mRNA targets. FEBS Lett 577: 134–140.
88. Brown JWS, Simpson CG (1998) Splice site selection in plant pre-mRNA splicing. Ann Rev Plant Physiol and Plant Mol Biol 49: 77–95.
89. Merritt H, McCullough AJ, Schuler MA (1997) Internal AU-rich elements modulate activity of two competing 3′ splice sites in plant nuclei. Plant J 12: 937–943.
90. Zhu H, Hinman MN, Hasman RA, Mehta P, Lou H (2008) Regulation of neuron-specific alternative splicing of neurofibromatosis type 1 pre-mRNA. Mol Cell Biol 28: 1240–1251.
91. Wang H, Molfenter J, Zhu H, Lou H (2010) Promotion of exon 6 inclusion in HuD pre-mRNA by Hu protein family members. Nucleic Acids Res 38: 3760–3770.
92. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, et al. (2006) The UCSC known genes. Bioinformatics 22: 1036–1046.
93. Langmead B, Trapnell C, Pop M, Salzberg S (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.
94. Jiang H, Wong WH (2009) Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 25: 1026–1032.
95. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.
96. Sun H, Chasin LA (2000) Multiple splicing defects in an intronic false exon. Mol Cell Biol 20: 6414–6425.
97. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Software 33: 1–22.