

# Prediction of Mutant mRNA Splice Isoforms by Information Theory-Based Exon Definition

Eliseos J. Mucaki,<sup>1</sup> Ben C. Shirley,<sup>2</sup> and Peter K. Rogan<sup>1,2\*</sup>

<sup>1</sup>Department of Biochemistry, Western University, London, Ontario, Canada; <sup>2</sup>Department of Computer Science, Western University, London, Ontario, Canada

Communicated by Michael Dean

Received 3 October 2012; accepted revised manuscript 4 January 2013.

Published online 24 January 2013 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22277

**ABSTRACT:** Mutations that affect mRNA splicing often produce multiple mRNA isoforms, resulting in complex molecular phenotypes. Definition of an exon and its inclusion in mature mRNA relies on joint recognition of both acceptor and donor splice sites. This study predicts cryptic and exon-skipping isoforms in mRNA produced by splicing mutations from the combined information contents ( $R_i$ , which measures binding-site strength, in bits) and distribution of the splice sites defining these exons. The total information content of an exon ( $R_{i,\text{total}}$ ) is the sum of the  $R_i$  values of its acceptor and donor splice sites, adjusted for the self-information of the distance separating these sites, that is, the gap surprisal. Differences between total information contents of an exon ( $\Delta R_{i,\text{total}}$ ) are predictive of the relative abundance of these exons in distinct processed mRNAs. Constraints on splice site and exon selection are used to eliminate nonconforming and poorly expressed isoforms. Molecular phenotypes are computed by the Automated Splice Site and Exon Definition Analysis (<http://splice.uwo.ca>) server. Predictions of splicing mutations were highly concordant (85.2%;  $n = 61$ ) with published expression data. *In silico* exon definition analysis will contribute to streamlining assessment of abnormal and normal splice isoforms resulting from mutations. Hum Mutat 00:1–9, 2013. © 2013 Wiley Periodicals, Inc.

**KEY WORDS:** exon definition; mRNA; cryptic splicing; gap surprisal; information theory

## Background

mRNA processing mutations, which are responsible for a wide range of human diseases [Divina et al., 2009], alter the abundance

and/or structures of mature transcripts. These mutations often occur proximate to exon/intron boundaries, but are frequently found at other sequence locations within introns or exons. Mutations that abolish or weaken recognition of natural splice acceptor or donor sites often produce transcripts lacking corresponding exons or activate adjacent cryptic splice sites of the same phase. Alternatively, mutations activate cryptic splice sites whose strength exceeds existing natural sites elsewhere in the unspliced transcript. The resultant molecular phenotypes may include isoforms with altered exon length and, in some instances, reduced or leaky expression of normal isoforms. We propose an approach based on information theory to predict the structures and approximate abundance of the output molecules generated directly or indirectly by splicing mutations.

Berget's exon definition model [Berget, 1995] provides a mechanism for recognizing multiple small exons against a background of considerably larger intronic sequences. Accurate exon recognition can be complicated by pseudoexonic structures present in introns that mimic natural exon structures [Ibrahim et al., 2005]. To discriminate between these structures, accurate spliceosomal recognition relies on relatively high affinities of the recognition sequences in natural exons and the presence of other splicing regulatory elements. Exons and adjacent introns also contain splicing enhancer (ESE, ISE) and silencer (ESS, ISS) sequences close to or overlapping constitutive splice sites, which may assist or suppress exon recognition through interactions with additional proteins [Berget, 1995; Graveley and Maniatis, 1998]. Recognition of an exon may therefore depend, to some degree, on the combined effects of each of these proteins [Goren et al., 2010]; however, the factors that recognize the acceptor and donor splice sites are often sufficient [Hwang and Cohen, 1997].

Information theory can be used to measure the conservation of nucleotide sequences bound by individual proteins or protein complexes. In splicing, information theory-based models of donor and acceptor splice sites reveal which nucleotides are permissible at both highly conserved and variable positions in individual sites [Robberson et al., 1990; Schneider, 1997]. These sequences are recognized prior to intron excision, and this recognition is related to the strength of the spliceosome–splice site interaction [Berget, 1995]. The strengths of spliceosome–splice site interactions are related to the corresponding individual information content,  $R_i$ , of the RNA sequence [Rogan et al., 1998]. The automated splice site analysis (ASSA) server was developed to determine changes in  $R_i$  values that result from mutations at individual splice sites, branch point sequences, and/or splicing regulatory sequences [Nalla and Rogan, 2005]. We propose that wild-type and mutant exons can be defined by the cumulative  $R_i$  values of each of these distinct binding sites contributing to exon recognition ( $R_{i,\text{total}}$ ), based on the fact

Additional Supporting Information may be found in the online version of this article.

\*Correspondence to: Peter K. Rogan, Department of Biochemistry, Schulich School of Medicine and Dentistry, Western University, London, ON N6A 2C1, Canada. E-mail: [progan@uwo.ca](mailto:progan@uwo.ca)

Contract grant sponsors: Natural Sciences and Engineering Research Council of Canada (371758-2009); Canadian Breast Cancer Foundation; Canada Foundation For Innovation; Canada Research Chairs; Compute Canada; Western University; and Cytonomix Inc.

that information is additive for independent sources of uncertainty [Jaynes, 1957].

In contrast with splice sites across an intron, cognate pairs of donor and acceptor splice sites from the same exon tend to be separated by a narrow range of distances in the unspliced transcript [the most common internal exon size is 96 nucleotides (nt)]. Single-exon recognition tends to be constrained by preferred distances between the U2 and U1 spliceosomal binding sites across the same exon [Hwang and Cohen, 1997]. We previously presented a model to define exon sequences which incorporates the information contents of both splice sites and preferences for frequent exon lengths of all natural exons [Rogan, 2009]. A general approach was used that minimized entropy of a pair of binding sites separated by a variable-length interstitial sequence. Given a set of exons flanked on either side by 100 nt intron sequences, the most accurate model (99% correctly detected exon boundaries) was derived by bootstrapping sets of 4,000 sequences with left (acceptor) and right (donor) sites of 31 nt (9.7 bits) and 15 nt (8.1 bits) in length. In the present study, we ensure that pairs of splice sites of opposite polarity are derived from the same exon by incorporating the surprisal function ([Tribus, 1961]; also termed self-information by Shannon [Cover and Thomas, 2006]), which corrects for both frequent and uncommon or rare intersite distances that are unlikely to form an exon. This is based on the observation that long internal exons are recognized inefficiently [Robberson et al., 1990], though they do occur (1,115 known internal exons >1,000 nt; [Bolisetty and Beemon, 2012]). The total exon information content ( $R_{i,\text{total}}$ ) is significantly reduced by this gap surprisal value, if either the predicted exon length is suboptimal or splice site pairs are derived from different exons, but is nearly unchanged for common exon lengths.

Here, we analyze splicing mutations according to changes in total exon information. Because the Automatic Splice Site and Exon Definition Analysis (ASSEDA) server predicts changes in expression relative to wild-type levels, it is assumed that the gene is expressed in the tissue being assayed, and that all splicing regulatory factors required for its expression are present in the relevant cell type in which the mutation is analyzed. Multiple splice isoforms may be produced from activated cryptic splice sites of the same polarity as the mutated splice site. The exons with highest information contents have the highest abundance, analogous to previous analyses of individual splicing mutations [Rogan et al., 1998]. The  $R_{i,\text{total}}$  values for different exons in normal and mutant sequences are directly compared to estimate their relative inclusion or exclusion in mature mRNA. Information theory-based exon definition models generate testable predictions of splice isoforms and can reveal splice isoforms that have not been previously described.

## Materials and Methods

### Exon Information Content

We derive the information content of a spliced exon from the cumulative contributions of the nucleic-acid-binding sites recognized by the spliceosomal machinery and the distribution distances separating binding sites within the same exon. Given a set  $S$  of  $n$  different binding sites in an exon, each of which is recognized by  $m$  different proteins, then  $S = \{x_n, \text{ where } 1 \leq n \leq m\}$ . The total information content,  $I_s$ , of all sites in  $S$  is

$$I_s = \sum_{n=1}^m R_i(x_n) \text{ bits} \quad (1)$$

The information content of each site,  $R_i(x_n)$  (measured in bits) is derived from a weight matrix ( $R_{iw}$ ) representing the sequence conservation of each nucleotide in that sequence. The derivation has been presented previously [Rogan et al., 1998; Schneider, 1997].

The information contents of each set of binding sites are modified to account for the probability that these sites occur within the same exon. This requires a gap surprisal term that depends on the transcriptome-wide distribution of the lengths separating them. The gap surprisal is applied to a set of sites within the same exon. Each combination of different binding proteins ( $x_1, x_2, \dots$ ) is described by a distinct distribution. The number of different unordered pairs of binding sites, given  $n$  different sites, corresponds to  $\binom{n}{2}$  different gap surprisal terms. The gap surprisal for two binding sites ( $x_p$  and  $x_q$ ), separated by  $L$  nucleotides  $g(L_{pq})$ , is

$$g(L_{pq}) = -\log_2 [P(L_{pq})] \text{ bits} \quad (2)$$

where  $L_{pq}$  is the distance between  $x_p$  and  $x_q$  sites. We calculate  $P(L_{pq})$  from experimentally validated intersite distances from human genes. Equation (4) signifies that the greater the distance between two sites, the larger the gap surprisal (greater penalty) will be, resulting in a biological reduction of larger than consensus exon length occurrence.

Denoting  $G(L_s)$ , the total gap surprisal of  $\binom{n}{2}$  different pairs of sites in set  $S$ ,

$$G(L_s) = \sum_{1 \leq p \leq n} \sum_{p < q \leq n} g(L_{pq}) \quad (3)$$

The total information content ( $R_{i,\text{total}}$ ) is defined by combining Equations (1) and (3),

$$R_{i,\text{total}} = \sum_{n=1}^m R_i(x_n) + \sum_{1 \leq p \leq n} \sum_{p < q \leq n} g(L_{pq}) \quad (4)$$

To calculate the  $R_{i,\text{total}}$  of an internal exon, we consider the simplest case with a constitutive set of donor and acceptor splice sites ( $n = 2$ ). We define  $x_1$  as the acceptor and  $x_2$  to be the donor site.  $x_n$  has been extended to incorporate other types of binding sites, including splicing regulatory factors, SF2/ASF (*SRSF1*) and SC35 (*SRSF2*), that modify exon recognition. These factors act to enhance splicing when the recognition sites are located within exons (ESE) and repress splicing (ISS) if occurring in the intron adjacent to constitutive splice sites [Lim et al., 2011]. The sign of this term in  $R_{i,\text{total}}$  is positive if the binding site is exonic and negative if it is intronic. The pairwise distribution of functional binding sites in the transcriptome is required to determine  $g(L_{pq})$ . For the first and last exons of a gene,  $R_{i,\text{total}}$  is the sum of the  $R_i$  value of the single splice site in that exon adjusted for  $g(L)$ , where  $L$  is exon length based on length distributions for the corresponding terminal exons. The sign of the  $g(L_{pq})$  term is negative for exonic locations (ESE) and reversed for intronic sites (ISS). We calculate and compare  $R_{i,\text{total}}$  values for the strengths of the constitutive splice sites in an exon before and after a mutation (details are provided in Supp. Methods). Isoforms with either different donor or acceptor sites may be predicted for each mutation. Because the lengths of these isoforms may vary considerably from each other, an analysis of compound mutations at different gene locations has been disabled in molecular phenotypic analysis. The exon definition algorithm requires at least one natural site from an exon to be contained in the predicted isoforms; thus, cryptic or pseudoexons activated by intronic mutations are not reported. Nevertheless, an individual information analysis with the ASSEDA server can detect changes at such sites as well as detect preexisting, neighboring cryptic splice sites of opposite polarity, which together could be recognized as novel exons.

## Populating the Annotation Database

The ASSEDA server is based on human genome reference sequence hg19 (GRCh37), GenBank and RefSeq cDNA accessions (downloaded from genome.ucsc.edu, July 2011), and SNP (dbSNP 135) tables. Genome-wide information weight matrices for automatically curated acceptor ( $n = 108,079$ ) and donor ( $n = 111,772$ ) splice sites (acceptor\_genome and donor\_genome, respectively, described in [Rogan et al., 2003]) were used in the  $R_{i,\text{total}}$  calculation. The reference sequence was scanned with these matrices to determine the  $R_i$ s of known natural splice sites and used to populate a MySQL database table (ALL\_RL, modified from the *all\_mRNA.txt* and the *refSeqAli.txt* from the UCSC genome browser).

The frequencies of different exon lengths occurring in the RefSeq database were determined for the gap surprisal calculation. The distribution of internal exon lengths better fit both lognormal and negative binomial distributions (Relative Standard Error [RSE] = 0.0006), compared with a Gaussian (RSE = 0.003). The lengths of first (best RSE = 0.006) and last (best RSE = 0.004) exons fit these distributions less well because long exons are more frequent. Gap surprisal values were normalized for all of these exon classes separately, by offsetting these values for predicted exon length relative to most frequent exon length in each type, which was reassigned  $G(L_s) = 0$  bits. Separate distributions were compiled, respectively, for first, internal, and last exons and stored in separate database tables. The start and end positions of the first and last exons were relaxed to include any coordinate within a 200 nt window once to avoid duplication of exons in the gap surprisal calculation (this accounts for variation in the methods used to generate the cDNAs that are mapped onto the genomic sequence).

## Incorporating Models of Splicing Regulatory Sequences into $R_{i,\text{total}}$

The impact of mutations in ISS or ESEs at SF2/ASF or SC35 binding sites on constitutive splicing can be predicted by selecting the option to incorporate this term into the  $R_{i,\text{total}}$  computation (on the Options page). Information weight matrices,  $R_i(b,l)$ , for SF2/ASF, SC35, SRP40 (*SRSF5*), and SRP55 (*SRSF6*) were derived from previously published data [Liu et al., 1998, 2000; Smith et al., 2006] and supplemented by experimentally validated binding sites curated from subsequent publications (sequence logos and weight matrices are available in Supp. Table S1). After scanning the reference genome and locating all predicted binding sites with the SF2/ASF and SC35  $R_i(b,l)$  matrices, their distributions,  $g(L_{pq})$ , were determined separately for intronic and exonic binding sites in closest proximity to adjacent constitutive splice sites. In computing  $R_{i,\text{total}}$ , the strongest preexisting splicing regulatory site affected by the mutation (with the highest initial  $R_i$  value) is selected by the server, unless the final  $R_i$  value of a second site surpasses that of the preexisting site upon introduction of the mutation (then the second site is reported). The gap surprisal table that is applied is based on which splicing regulatory protein is selected, and the location of the site.

## Description of Server

The ASSEDA server subsumes ASSA's capability to analyze changes in  $R_i$ , and additionally predicts molecular phenotypes based on changes in  $R_{i,\text{total}}$ . ASSEDA and ASSA use the same interface to

input sequence variants: Human Genome Organization (HUGO) approved gene symbols, Human Genome Variation Society (HGVS) mutation nomenclature, dbSNP identifiers, sequence window range around the mutation coordinate, and selected weight matrices as input (Fig. 1A; [Nalla and Rogan, 2005]). Mutation syntaxes are then translated into equivalent Delila instructions [Schneider et al., 1984]. The ASSEDA server contains a new option that allows analysis of either information of individual splice sites, the molecular phenotype consisting of all predicted transcript isoforms based on exon information, or both (for system architecture and program flow diagrams, see Supp. Figs. S1 and S2). Upon submission of a mutation, a set of GenBank accession identifiers (ID) corresponding to mRNAs associated with the submitted gene is suggested. These IDs now include mRNAs in the NCBI Reference Gene Sequence database (<http://www.ncbi.nlm.nih.gov/RefSeq/>; RefSeq). The IDs are differentiated according to GenBank accessions (in green) and RefSeq IDs (in blue). The longest mRNA accession number is selected by default, and the genomic structure of each RefSeq accession is hyperlinked to the selected ID.

The window range is a primary determinant of the number of potential isoforms reported because larger windows capture additional potential cryptic splice sites. The feasibility of exon formation is assessed by their  $R_{i,\text{total}}$  values, and by using rule-based filters to ensure that only likely isoforms are reported. These eliminate cryptic exons with misordered splice sites, overlapping donor and acceptor sites, internal exons less than 30 nt in length [Dominiski and Kole, 1991] predicted splice isoforms with <1% of exon inclusion relative to the mutated, natural exon strength ( $\Delta R_{i,\text{total}}$  between two isoforms <6.65 bits). The server highlights isoforms with negligible expression when their  $R_{i,\text{total}}$  values are at least 1 bit below that of the  $R_{i,\text{total}}$  of the mutated exon. Tabular results can be sorted by column and is paginated, which is particularly helpful for mutations in which numerous cryptic exons are predicted. All rows with potentially expressed isoforms are uncolored, but the wild-type exon is indicated in red. Splice isoforms that either cannot be expressed or minor forms (<5% of the major expressed form) that would not be detectable experimentally are, by default, filtered out. Without filtering, rows containing nonfunctional or minimally expressed predicted isoforms are highlighted in distinct colors: (1) exons with misordered splice sites (light blue), (2) potential cryptic exons with lower  $R_{i,\text{total}}$  values than normal or mutated exon ( $\leq 1\%$  predicted expression; pink), (3) isoforms with both incorrect splice site order and have low  $R_{i,\text{total}}$  values (green). The minimum reportable  $R_{i,\text{total}}$  value may also be selected using horizontal sliding scale bar, which filters out potential exons below this threshold.

The server draws a set of box glyphs (Fig. 2A) depicting a set of exon structures and lengths of potential isoforms that are most likely to form exons. The index of each isoform and its  $R_{i,\text{total}}$  value are also indicated next to each structure as well as the approximate chromosome coordinates of the normal and cryptic exons. Mutations at natural splice sites, which decrease  $R_{i,\text{total}} \geq 7$  bits (or abolish the site), are depicted as exon-skipped isoforms; however, this threshold is adjustable (Options menu).

The server also generates separate custom tracks of each isoform and uploads them to the UCSC genome browser, where they are displayed in the context of the exon containing the mutation as an embedded frame within ASSEDA. Each isoform is represented by a different spectrally encoded track, with highest to lowest energy wavelengths ordered in correspondence with their respective  $R_{i,\text{total}}$  values.

**A.**

Mutation entry by gene name

Designated gene name:

Mutation/variant:

Window range:  (Should be <=5,000)  
(Default is set to twice the length of the longest exon)

Analyze following sites:  
(Use Ctrl to select multiple options)

acceptor  
 donor  
 branch\_point  
 SC35  
 SRp40  
 SF2\_ASF

Point Mutation (default)

Molecular Phenotype Prediction by Exon Definition

**B.**

Isoform number	Exon definition model						Donor			Acceptor					$\Delta R_{i,total}$
	Donor coordinate	Position relative to natural Site	Closest natural site	Acceptor coordinate	Position relative to natural Site	Closest natural site	Initial (R <sub>i</sub> )	Final (R <sub>i</sub> )	$\Delta R_i$	Initial (R <sub>i</sub> )	Final (R <sub>i</sub> )	$\Delta R_i$	Initial R <sub>i,total</sub>	Final R <sub>i,total</sub>	
1	41208981	87	41209068	41209153	0	41209153	2.3	2.3	0.0	6.0	6.0	0.0	6.9	6.9	0.0
2	41209140	-72	41209068	41209202	49	41209153	1.9	1.9	0.0	3.5	3.5	0.0	3.9	3.9	0.0
3	41209125	-57	41209068	41209202	49	41209153	0.9	0.9	0.0	3.5	3.5	0.0	3.5	3.5	0.0
4	41208935	133	41209068	41209153	0	41209153	0.2	0.2	0.0	6.0	6.0	0.0	3.2	3.2	0.0
15	41208935	133	41209068	41209178	25	41209153	0.2	0.2	0.0	2.1	2.1	0.0	-1.1	-1.1	0.0
16	41208935	133	41209068	41209175	22	41209153	0.2	0.2	0.0	1.0	1.0	0.0	-2.1	-2.1	0.0
WT	41209068	0	41209068	41209153	0	41209153	6.9	6.9	0.0	6.0	6.0	0.0	6.9	6.9	0.0

**C.**

Isoform number	Exon definition model						Donor			Acceptor					$\Delta R_{i,total}$
	Donor coordinate	Position relative to natural Site	Closest natural site	Acceptor coordinate	Position relative to natural Site	Closest natural site	Initial (R <sub>i</sub> )	Final (R <sub>i</sub> )	$\Delta R_i$	Initial (R <sub>i</sub> )	Final (R <sub>i</sub> )	$\Delta R_i$	Initial R <sub>i,total</sub>	Final R <sub>i,total</sub>	
1	41208981	87	41209068	41209153	0	41209153	2.3	2.3	0.0	6.0	6.0	0.0	6.9	6.9	0.0
4	41208935	133	41209068	41209153	0	41209153	0.2	0.2	0.0	6.0	6.0	0.0	3.2	3.2	0.0
WT	41209068	0	41209068	41209153	0	41209153	6.9	6.9	0.0	6.0	6.0	0.0	6.9	6.9	0.0

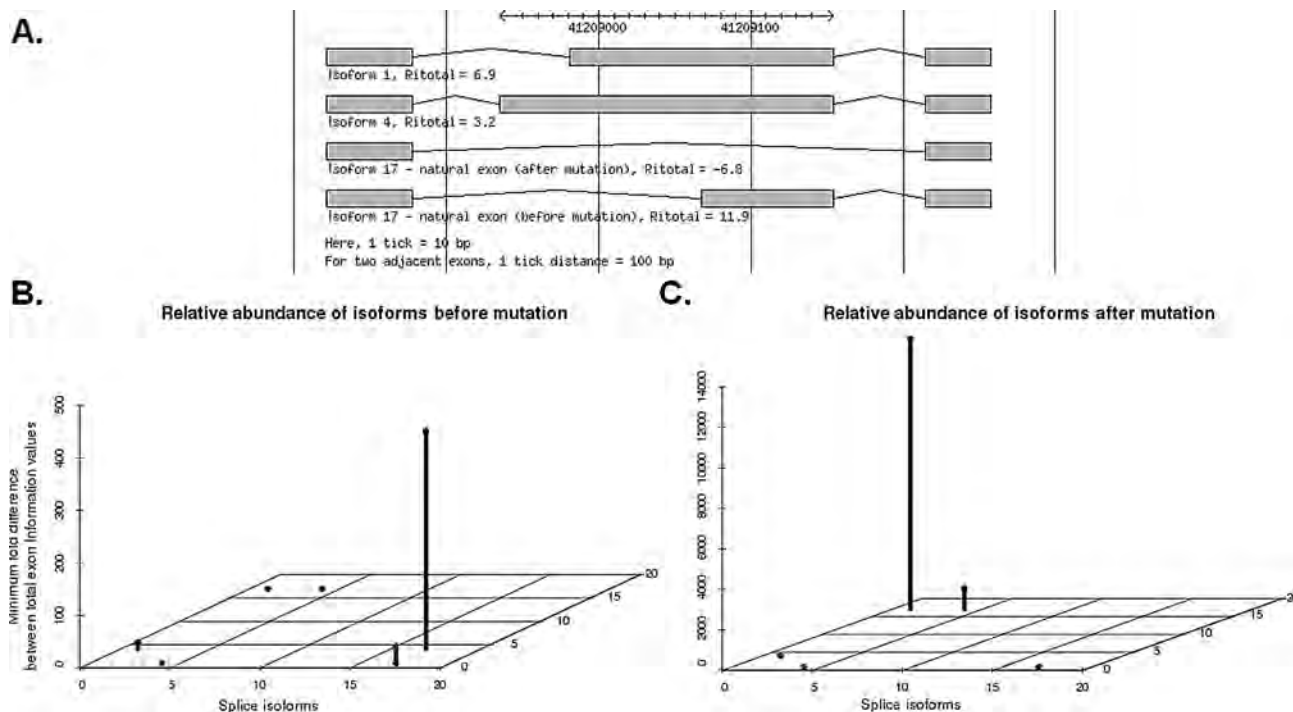
Wild type exon (WT)	Prospective isoforms (PI)	Isoforms not conforming natural exon defn (NC)
Negligibly expressed isoforms (NE)	Non-conforming natural exon defn isoforms which are negligibly expressed (NC-NE)	

**Figure 1.** Server input and results for *BRCA1* mutation, chr17:g.41209068G>A (NM\_007294.3:c.5277+1G>A). **A:** User input. The window size of 200 nt increases the number of potential cryptic isoforms reported beyond the default length. **B:** Resulting table after applying splicing mechanism and exon abundance filters (isoforms 5–14 are not presented because of space limitations). The column headings show key binding site locations, initial and final values and changes in  $R_i$ , as well as changes in  $R_{i,total}$ . The natural or mutated exon is listed in table row 17 (WT in legend below). Cells 1 and 4 (PI) indicate predicted cryptic isoforms with  $R_{i,total}$  values comparable or exceeding the strength of the natural exon ( $R_{i,total}$  final). Splice isoforms with  $R_{i,total} \leq 1$  bit (> twofold lower abundance; NE in legend) of the mutated natural exon are minimally expressed and filtered out. Rows 2 and 3 indicate predicted exons with misordered splice sites (NC), and rows 15 and 16 show exons, which also would be minimally expressed (NC-NE); C) Only three of 35 potential isoforms are reported for the input mutation after filtering on these criteria.

### Relative Abundance of Predicted Splice Isoforms

The server also displays pairwise differences in relative abundance for all predicted isoforms. The relative abundance or fold change in binding affinity of a single binding site is  $\leq 2^{\Delta R_i}$ , where  $\Delta R_i$  is the difference between the respective  $R_i$ s of wild type and mutant type of the site [Schneider, 1997]. We extend the idea of relative abundance of single binding site to multiple binding sites by comparing their  $R_{i,total}$  values. Suppose  $n$  and  $m$  are two alternative splice isoforms sharing at least one common splice site and their respective total

information contents are  $R_{i,total(n)}$  and  $R_{i,total(m)}$ . If  $R_{i,total(n)} > R_{i,total(m)}$ , then the relative abundance of  $n$  over  $m$  will be  $\leq 2^{\Delta R_{i,total(nm)}}$ , where  $\Delta R_{i,total(nm)} = R_{i,total(n)} - R_{i,total(m)}$ . Relative transcript abundance is displayed as a multidimensional graph (with *scatterplot3d*, an R package for visualization of three-dimensional multivariate data). The graph shows predicted pairwise differences in exon abundance ( $z$ -axis) of the  $x$ -axis isoform relative to the one on the  $y$ -axis, both before (left graph) and after mutation (right graph). The isoform designations correspond to those shown in the other molecular phenotype tabs.



**Figure 2.** Structure and relative abundance of predicted isoforms. Isoforms are depicted graphically according to their exon structures, relative abundance, and custom browser tracks in separate tabs. Isoform numbers in the figure refer to designations in Figure 1C. **A:** The scale above shows the genome coordinates of each of the isoforms. All of the prospective isoforms (sorted by  $R_{i,\text{total}}$ ) are scaled according to their genomic coordinates (above glyphs). The exon-skipping splice form is displayed for mutations wherein resulting  $R_{i,\text{total}} < 0$  bits. **(B and C)** Plots indicating predicted pairwise ( $x,y$  axes) relative minimum fold differences in abundance ( $z$ -axis) of each isoform both before and after changes in  $R_{i,\text{total}}$  due to the mutation. Results are depicted for *BRCA1*, chr17:g.41209068G>A. **B:** The natural wild-type exon (isoform 17) has the highest level of expression. After the mutation **(C)**, isoform 1, which activates a downstream cryptic splice site, is expected to be the dominant splice form. Note that the scale of the  $z$ -axis will change between the panels, depending on the range of  $\Delta R_{i,\text{total}}$  values resulting from the mutation.

## Results

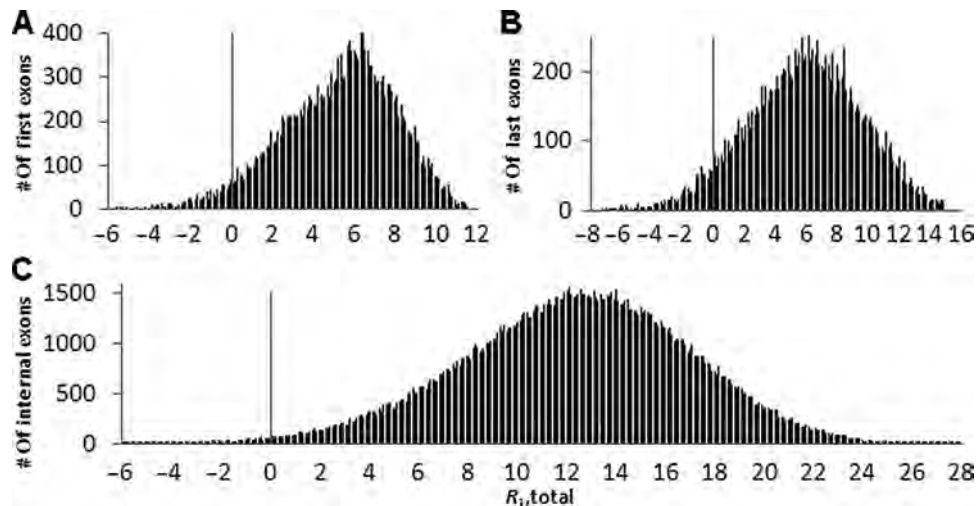
### Exon Definition by Information Analysis of Functional Exons

Gap surprisal values of all exon lengths were determined from their respective frequencies in the exome of all RefSeq genes. The gap surprisal penalty was then normalized so that the most common internal exon length (96 nt;  $n = 172,250$ ) was zero bits, by subtracting a constant value of 6.59 bits (its  $\log_2$  frequency). Less frequent exon lengths were scaled to this value by subtracting this constant from their respective gap surprisal values. First and terminal exons are, respectively, missing either a donor or an acceptor splice site, and exhibit a broader range of exon lengths. Separate gap surprisal distributions were computed for these exons. The most frequent first and last exons were, respectively, 158 nt ( $n = 23,471$ ) and 232 nt ( $n = 21,261$ ) in length, corresponding to gap surprisals of 7.8 and 9.4 bits, respectively.  $R_{i,\text{total}}$  values were  $>0$  bits for 98.9% of internal exons, 95.3% of first exons, and 93.1% of last exons (Fig. 3). The distribution of  $R_{i,\text{total}}$  values for each of these exon classes had an approximately Gaussian distribution. Although inclusion of the gap surprisal term resulted in fewer false-positive splice isoforms [Dominski and Kole, 1992; Robberson et al., 1990], a slightly higher proportion of first and last exons had negative  $R_{i,\text{total}}$  values. Because most of these splice sites in these exons exhibited positive  $R_i$  values (72% of first, 87% last exons), the negative  $R_{i,\text{total}}$  values may be the result of other unknown factors contributing to recognition of these exons not accounted for, or to suboptimal gap surprisal functions.

### Interpretation of Splicing Mutations by Exon Definition Analysis

The previous ASSA has been upgraded to the ASSEDA server (<http://splice.uwo.ca>). In addition to individual splice site analysis, ASSEDA computes changes in total exon information resulting from variants that differ from the most current version of the genome reference sequence. Mutations are still inputted through a Web interface [Nalla and Rogan, 2005] with a predefined sequence window of sufficient length to encompass the spectrum of potential splice sites activated by a mutation. The  $R_{i,\text{total}}$  values of prospective exons within this window are computed with the Molecular Phenotype option on the main page of the server.

The server accepts and reports genomic DNA coordinate notations (as well as the Intervening Sequence [IVS] notations) supported by the most recently proposed Locus Reference Genomic variant format [Dagleish et al., 2010], which is consistent with HGVS recommendations. A typical molecular phenotypic prediction is indicated in Figure 1 (*BRCA1* IVS20+1G>A or HGVS designation chr17: g.41209068C>T; Supp. Table S2, Mutation #4). The tabular results indicate genomic coordinates of donor and acceptor sites, their relative distance from the closest natural site, and the change in  $R_i$  for these sites. Each row indicates  $R_{i,\text{total}}$  both before and after mutation for a different set of exon boundaries corresponding to a distinct predicted isoform. Predicted isoforms are sorted according to these values, whose fold differences in binding affinity are  $\leq 2^{\Delta R_{i,\text{total}}}$  [Schneider, 1997].



**Figure 3.** Distribution of the  $R_{i,\text{total}}$  of annotated exons. Histogram of  $R_{i,\text{total}}$  values for exons in the RefSeq database are illustrated for first (a), last (b), and internal exons (c). Nearly all internal exons exhibit total information contents exceeding zero bits (98.9%). The gap surprisal functions for first and last exons are not optimized for single splice site exons (4.7% and 7.0%, respectively, have  $R_{i,\text{total}}$  values below zero bits). The majority of false-negative internal exons contain one or both splice sites that are either weak or are not recognized by either the U1 or U2 spliceosomes.

Initially, 20 potential isoforms are found for this mutation, of which those with the highest  $R_{i,\text{total}}$  values and the affected natural exon are indicated (Fig. 1B). Based on the mechanism of exon recognition and the  $\Delta R_{i,\text{total}}$  values, only a subset of these indexed isoforms is likely to be expressed. Splice site polarity is specified such that a functional acceptor splice site cannot occur downstream of a natural donor splice site to define an exon, and vice versa [Berget, 1995]. The server eliminates exons with misordered splice sites, removing many false-positive splice isoforms that do not conform to the natural mRNA splicing mechanisms. Pairs of splice donor and acceptor sites that either overlap each other are also not considered as potential exons [Nalla and Rogan, 2005; Robberson et al., 1990]. Predicted low abundance natural and cryptic isoforms with undetectable expression (Figs. 1B and C) are also filtered out.

The structures and lengths of each potential isoform (natural, cryptic, skipped) are also displayed in a separate tab (Fig. 2A). The central exon affected by the mutation is drawn to scale; however, flanking intron sequences are condensed for the presentation. In the example given above, the exon 20 donor site in chr17: g.41209068C>T ( $R_{i,\text{total}}$  11.9  $\rightarrow$  -6.6 bits) is inactivated and a corresponding isoform with exon skipping is shown. The relative abundance ( $z$ -axis) of different pairs of indexed isoforms ( $X$  and  $Y$ ) before (Fig. 2B) and after (Fig. 2C) mutation also predicts a number of cryptic isoforms. Isoform 1 uses a preexisting donor 87 nt downstream that is at least 13,307-fold (i.e.,  $\leq 2^{13.7 \text{ bits}}$ ) more abundant than the mutated exon, but would not normally be detected because it is 32-fold ( $\leq 2^{5.0}$ ) less abundant than the normal exon. mRNA analyses have shown that this mutation results in both cryptic and skipped splice forms [Sanz et al., 2010]; however, isoform 4, which contains 133 of intronic sequence (Figs. 1C and 2A), was not detected.

The molecular phenotype analysis panel also displays wild-type and predicted mutant isoforms as separate BEDGRAPH custom tracks on the UCSC genome browser. The combination of these predictions with other browser tracks (i.e., sequenced mRNAs, ESTs, and known SNPs within a gene) can distinguish mutant from naturally occurring alternative splice forms or previously described variants that may be associated with a suspected mutation.

## Validation

To assess whether the proposed model of exon definition produced results consistent with observed mutant spliced products, we evaluated a series of reported splicing mutations for which end-point (Supp. Table S2) and quantitative (Supp. Table S3) expression studies had been performed. Mutant isoforms and relative abundance were predicted for splicing mutations in the nephropathic cystinosis gene product (*CTNS*), cardiac myosin binding protein C (*MYBPC3*), fumarylacetoacetate hydrolase (*FAH*), the Ellis van Creveld syndrome gene product (*EVC*), thalassemia (*HBB*), coagulation factor XII (*F12*), adenosine deaminase (*ADA*), cyclin-dependent kinase inhibitor 2A (*CDKN2A*), iduronate 2-sulfatase (*IDS*), phenylalanine hydroxylase (*PAH*), paired-like homeodomain 2 (*PITX2*), phosphomannomutase 2 (*PMM2*), and early-onset breast cancer (*BRCA1* and *BRCA2*).

A detailed analysis of one of these mutations illustrates the importance of the gap surprisal function and of post-hoc filtering out misordered and weak exons. A splicing mutation, *CDKN2A*: IVS2+1G>T (g.21970900G>T; Supp. Table S2, #21), abolishes a natural donor site, and numerous potential cryptic donor sites are unmasked ( $n = 61$  potential exons). After filtering nonconforming or negligibly expressed isoforms, eight prospective exons of sizes 133, 234, 166, 680, 435, 973, 742, 515, and 308 nt remain. Application of the gap surprisal term reduces the number of exon combinations. Exons of length 308, 435, 515, 680, 742, and 973 nt are much less common, and the gap surprisal penalties are correspondingly larger (5.9, 7.1, 8.3, 8.6, 8.6, and 9.9 bits, respectively), significantly lower their  $R_{i,\text{total}}$  values. The three highest predicted 133, 166, and 234 contain correctly ordered splice sites (with penalties of 0.8, 1.4, and 3.0 bits, respectively), two of which have previously been reported [Rutter et al., 2003]. Similarly, *PMM2*: IVS3-1G>C (Supp. Table S2 #41) mutation predicts 263 possible exon structures, 35 isoforms when the gap surprisal penalty is applied, and only four after filtering. Clearly, one of the strengths of the present approach is that it effectively eliminates improbable or poorly expressed isoforms.

The accuracy of information-based exon definition prediction was compared with expression data for these mutations (Supp.

Table S2). There is generally good concordance with documented splice isoforms. All mutations either weakened or inactivated natural splice sites or activated cryptic splice sites as expected. Among the mutations tested, 36 of 41 mutations were completely consistent with published results, including cryptic isoforms. In some instances, a reported cryptic exon was predicted, but was not determined to be the most abundant splice isoform.

Information analysis correctly predicted several types of splicing abnormalities in different genes. There were 31 mutations that resulted in formation of one or more cryptic exons (Supp. Table S2). Exons using these cryptic splice sites were predicted for 28 of the 31 mutations, 20 of which had the highest  $R_{i,\text{total}}$  values. “The cryptic splice isoforms for the remaining eight mutations were predicted to be among the highest in abundance, save one (Supp. Table S2 #10).” Complete intron retention was reported for one mutation (#40), whereas nine mutations were found to result in exon skipping only (#1, 7, 8, 11, 14, 23, 26, 37, and 41). Previously, we have shown that large changes in  $\Delta R_i$  can result in exon skipping as well as leaky splicing [Rogan et al., 1998]. All of these mutations decreased  $R_{i,\text{total}}$  of the natural exon, although in one case, the extent was marginally below significance (#14; 0.8 bits). Exon skipping was reported for mutations #7, 8, 23, and 24 rather than reduced levels of exon inclusion suggested by the exon definition analysis. These mutations reduced the predicted exon abundance by ninefold to 23-fold relative to the normally spliced product. This level of expression is close to the detection limit of a minor cryptic splice isoform for most analytic methods [Rogan et al., 1998], and may explain why only exon skipping was documented for these mutations [Claes et al., 2002, 2003; Macias-Vidal et al., 2009; Tompson et al., 2007]. Additionally, the discrepancy could simply be because of the limitations of the type of in vitro analyses performed.

Exon definition analysis of the remaining mutations showed partial discordance to published mRNA evidence. In three cases, the reported cryptic site used had an  $R_i < 0$  bits (#10, 15, 32). Mutation #27,  $R_{i,\text{total}}$  of the natural and the proven activated cryptic site does not quite reach the threshold for a functional site defined by information theory. In the final case (#22), the creation of a cryptic donor is predicted (2.7 bits), but the resultant 425-nt exon is not observed ( $R_{i,\text{total}} < 0$ ).

In simple molecular recognition systems, information theory-based methods predict binding-site affinity changes, as  $R_i$  is a direct related to binding affinity. However, while spliceosomal binding affinity is crucial to constitutive splice site recognition [Berget, 1995], other factors such as interactions with regulatory factors can influence splicing outcomes [De Conti et al., 2012]. The present model attempts to account for the effects of these factors on exon inclusion by incorporating the contributions of multiple binding sites. Predicted isoforms were compared with detected mRNA species for eight splicing mutations that have been assessed with quantitative methods (predominantly quantitative RT-PCR; Supp. Table S3). ASSEDA correctly predicted a decrease in wild-type splicing in most instances. Cryptic exons detected (Supp. Table S3: #2, 4, 5, 8) were also correctly predicted, and were ranked highest based on  $R_{i,\text{total}}$  values (using a window size 200 nt) in all but one case (in which the cryptic isoform ranked second; Supp. Table S3: #8). Decreases in the predicted strength of the natural exon were accurately predicted for six of eight mutations (Supp. Table S3: #1–4, 7, 8). The mutation c.653A>G (Supp. Table S3: #5) was found to abolish normal splicing of exon 6 of *PCCB*, although the exon was predicted to have some residual splicing (82.3% decrease in binding efficiency). Conversely, g.6622214G>C (Supp. Table S3: #6) in *NSUN2* was predicted to abolish normal splicing (99.97% decrease in binding efficiency), but a low level of the splice form was still detected (5% of control

expression). When both cryptic isoforms and exon skipping occur (Supp. Table S3: #2 and 8), the ratio between the two splice forms does not seem to relate directly to predicted strength changes. In both cases, splicing of the normal exon is abolished and a cryptic isoform (where  $R_{i,\text{total}} < \text{initial } R_{i,\text{total}}$ ) appears, but they differ in the ratio of exon skipping to cryptic exon expression. This may be associated with the instability of mRNAs with frameshifted cryptic isoforms (Supp. Table S3: #4).

## Impact of ESE/ISS Elements

Elements recognized by splicing regulatory proteins, SF2/ASF (*SRSF1*), SC35 (*SRSF2*), SRp40 (*SRSF5*), SRp55 (*SRSF6*), and hnRNP-H (*HNRNPH1*), can now be analyzed with ASSEDA; however, these matrices are based on many fewer sites (usually <50), and the  $R_i$  values may not be as accurate as constitutive splice sites, especially at the low end of the distribution. The server computes  $R_i$  values of any of these individual sites and can incorporate mutations at either SF2/ASF or SC35 sites into the  $R_{i,\text{total}}$  computation. Because a mutation can affect multiple predicted sites, the site with the highest  $R_i$  value altered by the mutation is analyzed, unless a second cryptic site is strengthened resulting in final  $R_i$  is exceeding that of the original binding site.

A second gap surprisal function, based on the distances between known natural constitutive sites and the closest predicted splicing regulatory site of the same type, was also applied in the  $R_{i,\text{total}}$  calculation. Exon (ESE) and intron (ISS) have independent gap surprisal distributions (Supp. Fig. S3). The ubiquity of these splicing regulatory sequences suggested that their predicted distributions would be biased toward shorter intersite distances; however, there were distinct preferences for certain distances. 17.2% of all exonic SF2/ASF sites were separated by 4 nt from a natural splice site ( $n = 562,786$ ; comparatively, all other distances between zero and 10 nt range from 1.5% to 4.4% in frequency). The most common intronic SF2/ASF sites were 1, 3, and 5 nt from the natural site (9.3%, 7.1%, and 10.5%, respectively;  $n = 562,788$ ). The most common SC35 site intersite exonic distances were 0, 4, and 7 nt (9.5%, 6.5%, and 6.6%, respectively) and intronic distances were spaced 1 and 2 nt from the splice site (9.9% and 9.5%). In all cases, frequency decreased with increased intersite distance. The distribution of predicted SRp40 distances showed no distance bias; there was a gradual inverse relationship between frequency and distance from the natural site (maximum frequency was <0.1% of the sites).

To assess the effect of including SC35 and SF2/ASF sites in the exon definition model, we evaluated 12 reported mutations/variants in either SF2/ASF or SC35 sites that were reported to affect splicing at adjacent splice sites (Supp. Table S4). Eight of 12 predictions of ASSEDA were concordant with the published results (Supp. Table S4: mutations #1–4, 6, 9, and 11 are predicted to weaken splicing and lead to exon skipping; #10 strengthens an intronic SF2/ASF site and activates a cryptic donor). A single-nucleotide difference between *SMN1* and *SMN2* (c.840C>T) is known to alter an SF2/ASF exonic site, resulting in skipping of exon 7 in *SMN2* [Cartegni and Krainer, 2002]. The SF2/ASF variant in *SMN2* reduces  $\Delta R_{i,\text{total}}$  of exon 7 in *SMN2* by 5.7 bits relative in *SMN1*, corresponding to a 52-fold difference in exon recognition, consistent with skipping of this exon in *SMN2* (Supp. Table S4: #1).

Observed alterations of splicing regulatory sites were not predicted in three cases (Supp. Table S4: #5, 7 and 12). For two other mutations, SF2/ASF or SC35 sites were affected (Supp. Table S4: #3 and 8), but interpretation was confounded by concomitant changes in the opposite direction to SRp55 splicing regulatory sequences (the effect of mutation #3 is predicted with ASSEDA's SRp55 model).

## Discussion

We have designed and implemented a novel approach to predict the molecular phenotype of a splicing mutation, producing a probable set of splicing isoforms expressed in mutation carriers. The system is based on information theory-based methods that accurately quantify binding-site affinity [Rogan et al., 1998; Schneider, 1997]. Nonexpressed or very-low-expression exons are filtered out by correcting for suboptimal exon lengths and eliminating incorrectly ordered splice sites. The use of gap surprisal to correct for distances between required sequence features has been previously validated for other types of binding sites [Shultzaberger et al., 2001]. Exon information ( $R_{i,\text{total}}$ ) is computed by the ASSEDA server. The backend databases consist of current human genome sequences (hg19/GRCh37) and gene annotations of exon coordinates, gene names, and mRNA accession numbers, including NCBI Refseq entries (and dbSNP 135). Mutation entry currently supports c. and g. notation, as well as the deprecated IVS-based mutation description.

We implement a simple model for exon definition based on constitutive splice sites, although the theory for extensible framework for incorporation of multiple splice site recognition sequences is derived. Exon-definition-based predictions were compared with known splicing mutations with published mRNA studies, and these predictions were found to be highly concordant (Supp. Table S2). These mutations were sourced from our previous publications so that information theory-based modeling of individual splice sites could be compared with exon definition [Mucaki et al., 2011; Rogan et al., 1998].

The exon definition models imply that rare exons (regardless of length) will have large gap surprisal penalties. This is supported by the fact that, for exons exceeding a few hundred nucleotides in length, the gap surprisal penalty increases with length, until it becomes asymptotic for exon lengths that occur once in the genome. The significant gap surprisal penalties for long exons raise the question as to how well the model performs at the extreme lengths to correctly distinguish natural from decoy exons. The model fails if the contributions of the gap surprisal term exceed the  $R_i$  values of both natural splice sites. In fact, this is generally not the case.

To assess the ability of the server to predict naturally occurring large exons, eight large internal exons in genes *BRCA1-ex11*, *BRCA2-ex11*, *TTN-ex253*, *JARID2-ex7*, *KLHL31-ex2*, *C6orf142-ex4 (MLIP)*, *VCAN-ex8*, and *C17orf53-ex3* were evaluated using ASSEDA (Supp. Table S5). Despite the large-gap (>10 bit) surprisal penalties, the  $R_{i,\text{total}}$  values for each of these exon was still exceeded zero bits. This can be attributed to their strong donor and acceptor sites, which appear to be essential for large exon recognition ([Bolisetty and Beemon, 2012]; the exception being the donor site of *BRCA1* exon 11 [2.9 bits]). These predicted shorter splice forms are present in *BRCA1* mRNA; however, they do not encode full-length protein. For example, the highest ranked prospective isoform for *BRCA1-ex11* was a 118-nt-long alternate splice form (NM\_007298.3). These large exons were not ranked first, as the  $R_{i,\text{total}}$  of smaller exons (<250 nt) tended to have higher overall  $R_{i,\text{total}}$ s (lower gap surprisal penalty). Larger exons tend to have a higher ratio of enhancers to repressors compared with smaller exons [Bolisetty and Beemon, 2012]. This suggests that the gap surprisal function will need to be refined or contributions of other splicing regulatory proteins will need to be incorporated into  $R_{i,\text{total}}$  to correct the ranking of splice isoforms from long exons.

Although the model we have implemented does predict the preponderance of mutant splice isoforms, it has obvious limitations. For example, misordered splice sites excluded in this model can sometimes be activated through the formation of novel cryptic ex-

ons when proximate, preexisting sites of opposite polarity occur within adjacent introns. The server does not currently handle these situations, which are thought to be uncommon and because of the requirement to include at least one functional site of opposite polarity in the mutated exon.

By default, the current models do not take into account other types of splicing-related binding sites that influence splice site selection, including SR proteins that contribute to exon recognition. This is expected to result in discordant interpretation (i.e., Supp. Table S2 #27) or inaccuracies in predicting the abundance predicted splice isoforms because  $R_{i,\text{total}}$  values will be underestimated. Changes in strengths of SF2/ASF and SC35 binding sites can now be included in the  $R_{i,\text{total}}$  calculation, which can improve its accuracy. However, results might be skewed when the altered splicing regulatory sites do not contribute either positively or negatively to normal splicing. There are many more predicted SF2/ASF and SC35 sites in an exon than there is evidence for each having a specific role in exon definition. Furthermore, only a single splicing regulatory site is currently used in the  $R_{i,\text{total}}$  calculation, and this assumption may have to be revisited if the molecular phenotype is the result of multiple synergistic or antagonistic binding events. For example, there are two variants in Supp. Table S4 (#3, 8), which alter either SF2/ASF or SC35 binding sites, but simultaneously change the predicted  $R_i$  value of a proximate SRp55 site. Conceptually, the ASSEDA server can be modified to handle such situations. Nevertheless, predicting changes in molecular phenotype resulting from mutations at multiple, overlapping ESE/ISS sequences will require simultaneous validation of these effects on all of these sequence elements.

In some cases, the published alternate splice form is detected, but is not the strongest (highest  $R_{i,\text{total}}$ ) cryptic exon predicted (Supp. Table S2, #2, 10, 13, 21, 25, 27, 34, and 38). Many of these published mutation phenotypes are based on a gel-based RT-PCR analysis, wherein some cryptic splice isoforms may not have been detected. The gap surprisal term may also be inaccurate, as there are a higher percentage of false-negative natural exons with  $R_{i,\text{total}} < 0$  bits. We suggest using the top five splice forms, including cryptic isoforms with the highest  $R_{i,\text{total}}$  values to perform experimental validation of these predictions.

The gap surprisal contributions for all exon types (first, last, and internal exons) are more variable for larger exons (Supp. Fig. S4A, C, and D). Although stochasticity is clearly discernible because of low counts of longer exons, some of this variability can be attributed to genomic selection for these exon lengths. Certain exon lengths may be overrepresented because of paralogous gene duplications, which tend to inflate their frequency relative to other exons of similar lengths, leading to significantly lower gap surprisal values. Additionally, exons that retain an integral number of codons are more frequent than similarly sized exons that change frame (+1 or -1 length), consistent with previous studies of cassette exons in alternatively spliced genes [Clark and Thanaraj, 2002; Stamm et al., 2006]. This triplet periodicity is well defined for exons ranging in size from 9 to 200 nt (Supp. Fig. S4B), but nevertheless is still apparent in exons beyond 500 nt in length. For these reasons, we avoided fitting gap surprisal values to an algebraic function or smoothing of these distributions.

The development of an exon-definition-based mutation analysis was motivated by the desire to generate predictions that could be directly compared with laboratory expression data. In some instances, these predictions have included strong cryptic exons that have not been previously detected, possibly because the laboratory studies did not directly anticipate the corresponding splice isoforms. The level of concordance we report for previously validated splicing mutations justifies a prospective study of natural and mutant isoforms



predicted by the server, in which all predicted cryptic splice isoforms are tested, and if possible, quantified. It should be feasible to implement algorithms to automate design of isoform specific sequence primers for quantitative expression analysis. This feature will close the circle between bioinformatic methods that predict potential splicing mutations in large-scale genomic DNA sequence studies and the design of experiments to validate the expression patterns with mRNA from the same individual.

## Acknowledgments

We acknowledge other Rogan Laboratory members for valuable comments, and software development and testing. We recognize David Wiseman, Jeff Shantz, and Bruce Richards (Department of Computer Science, Western University) for their assistance with systems administration and maintenance.

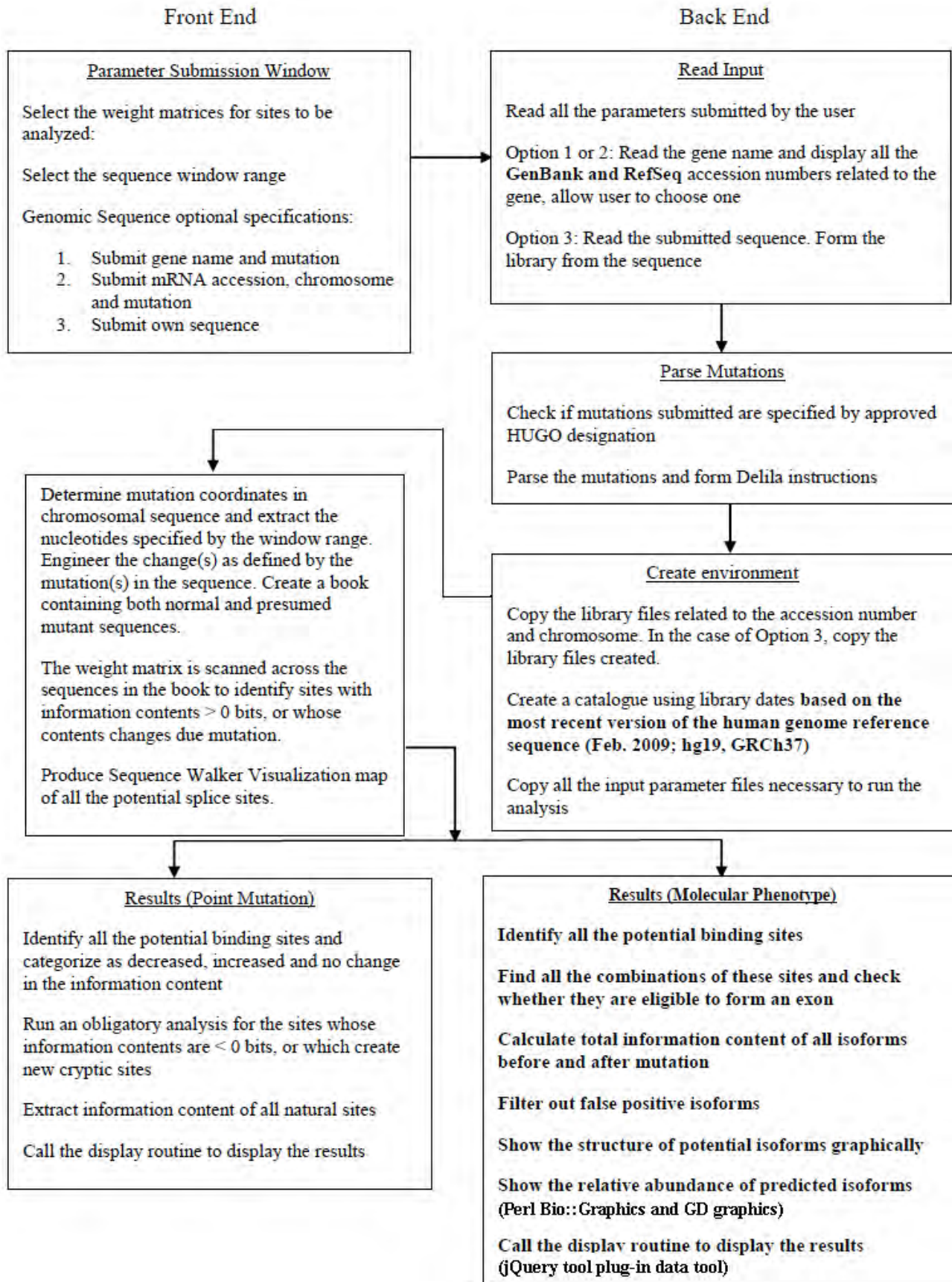
*Authors' contributions:* P.K.R. derived and developed the methods, and the gap surprisal distributions. E.J.M. and B.C.S. implemented these methods, which involved modifying previous software, creation of new modules, and updating databases. E.J.M. analyzed experimental data for validation. E.J.M. and P.K.R. wrote the manuscript, which has been approved by all of the authors.

*Disclosure statement:* P.K.R. is the inventor of US Patent 5,867,402 and founder of Cytognomix, which is developing software based on this technology for complete genome or exome splicing mutation analysis.

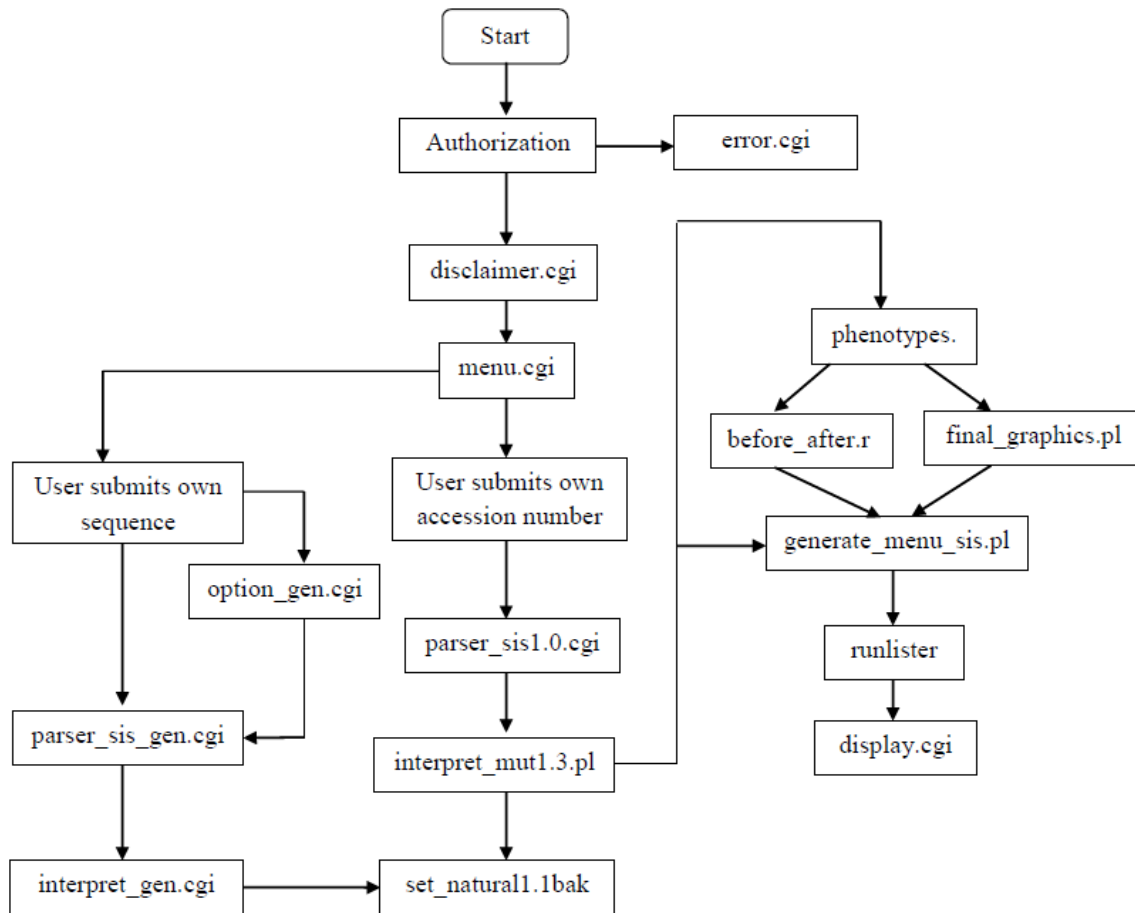
## References

- Berget SM. 1995. Exon recognition in vertebrate splicing. *J Biol Chem* 270:2411–2414.
- Bolisetty MT, Beemon KL. 2012. Splicing of internal large exons is defined by novel cis-acting sequence elements. *Nucleic Acids Res* 40(18):9244–9254.
- Cartegni L, Krainer AR. 2002. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat Genet* 30:377–384.
- Claes K, Poppe B, Machackova E, Coene I, Foretova L, De Paepe A, Messiaen L. 2003. Differentiating pathogenic mutations from polymorphic alterations in the splice sites of BRCA1 and BRCA2. *Genes Chromosomes Cancer* 37:314–320.
- Claes K, Vandesompele J, Poppe B, Dahan K, Coene I, De Paepe A, Messiaen L. 2002. Pathological splice mutations outside the invariant AG/GT splice sites of BRCA1 exon 5 increase alternative transcript levels in the 5' end of the BRCA1 gene. *Oncogene* 21:4171–4175.
- Clark F, Thanaraj TA. 2002. Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum Mol Genet* 11: 451–464.
- Cover TM, Thomas JA. 2006. *Elements of information theory*. Hoboken, NJ: Wiley-Interscience, p. 748.
- Dagleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, Chen Y, McLaren WM, Larsson P, Vaughan BW, Beroud C, Dobson G, et al. 2010. Locus Reference genomic sequences: an improved basis for describing human DNA variants. *Genome Med* 2:24.
- De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* 4(1):49–60.
- Divina P, Kvitkovicova A, Buratti E, Vorechovsky I. 2009. Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping. *Eur J Hum Genet* 17:759–765.
- Dominski Z, Kole R. 1991. Selection of splice sites in pre-mRNAs with short internal exons. *Mol Cell Biol* 11(12):6075–6083.
- Dominski Z, Kole R. 1992. Cooperation of pre-mRNA sequence elements in splice site selection. *Mol Cell Biol* 12:2108–2114.
- Goren A, Kim E, Amit M, Vaknin K, Kfir N, Ram O, Ast G. 2010. Overlapping splicing regulatory motifs—combinatorial effects on splicing. *Nucleic Acids Res* 38:3318–3327.
- Graveley BR, Maniatis T. 1998. Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing. *Mol Cell* 1:765–771.
- Hwang DY, Cohen JB. 1997. U1 small nuclear RNA-promoted exon selection requires a minimal distance between the position of U1 binding and the 3' splice site across the exon. *Mol Cell Biol* 17:7099–7107.
- Ibrahim EC, Schaal TD, Hertel KJ, Reed R, Maniatis T. 2005. Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proc Natl Acad Sci USA* 102:5002–5007.
- Jaynes E. 1957. Information theory and statistical mechanics. *Phys Rev* 106:620–630.
- Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. 2011. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci USA* 108(27):11093–11098.
- 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.
- Liu HX, Zhang M, Krainer AR. 1998. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* 12:1998–2012.
- Macias-Vidal J, Rodes M, Hernandez-Perez JM, Vilaseca MA, Coll MJ. 2009. Analysis of the CTNS gene in 32 cystinosis patients from Spain *Clin Genet* 76:486–489.
- Mucaki EJ, Ainsworth P, Rogan PK. 2011. Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants. *Hum Mutat* 32(7):735–742. doi: 10.1002/humu.21513. Epub 2011 May 5.
- Nalla VK, Rogan PK. 2005. Automated splicing mutation analysis by information theory. *Hum Mutat* 25:334–342.
- Robberson BL, Cote GJ, Berget SM. 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol Cell Biol* 10:84–94.
- Rogan PK, Faux BM, Schneider TD. 1998. Information analysis of human splice site mutations. *Hum Mutat* 12:153–171.
- Rogan PK, Svojanovsky SR, Leeder JS. 2003. Information theory-based analysis of CYP219, CYP2D6 and CYP3A5 splicing mutations. *Pharmacogenetics* 13:207–218.
- Rogan PK. 2009. Ab initio exon definition using an information theory-based approach. CISS 2009, 43rd Annual Conference on Information Sciences and Systems p. 847–852. doi: 10.1109/CISS.2009.5054835.
- Rutter JL, Goldstein AM, Davila MR, Tucker MA, Struwing JP. 2003. CDKN2A point mutations D153spl(c.457G>T) and IVS2+1G>T result in aberrant splice products affecting both p16INK4a and p14ARF. *Oncogene* 22:4444–4448.
- Sanz DJ, Acedo A, Infante M, Duran M, Perez-Cabornero L, Esteban-Cardena E, Lastra E, Pagani F, Miner C, Velasco EA. 2010. A high proportion of DNA variants of BRCA1 and BRCA2 is associated with aberrant splicing in breast/ovarian cancer patients. *Clin Cancer Res* 16:1957–1967.
- Schneider TD, Stormo GD, Yarus MA, Gold L. 1984. Delila system tools. *Nucleic Acids Res* 12:129–140.
- Schneider TD. 1997. Information content of individual genetic sequences. *J Theor Biol* 189:427–441.
- Shultzaberger RK, Bucheimer RE, Rudd KE, Schneider TD. 2001. Anatomy of Escherichia coli ribosome binding sites. *J Mol Biol* 313:215–228.
- Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, Krainer AR. 2006. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum Mol Genet* 15(16):2490–508.
- Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, Tang Y, Barbosa-Morais NL, Thanaraj TA. 2006. ASD: a bioinformatics resource on alternative splicing. *Nucl Acids Res* 34(suppl 1):D46–D55.
- Tompson SW, Ruiz-Perez VL, Blair HJ, Barton S, Navarro V, Robson JL, Wright MJ, Goodship JA. 2007. Sequencing EVC and EVC2 identifies mutations in two-thirds of Ellis-van Creveld syndrome patients. *Hum Genet* 120:663–670.
- Tribus M. 1961. *Thermodynamics and thermodynamics: an introduction to energy, information and states of matter, with engineering applications*. Princeton, NJ: Van Nostrand, p. 649.

## Supplementary Figure S1 – Architecture of the ASSEDA Server



**Supplementary Figure S2 –Flow Chart of the ASSEDA Server.** The program flow chart of the server, with brief descriptions of the programs listed.



**Short Description of all Programs:**

*disclaimer.cgi*: Displays the disclaimer to the user, once they enter the login information.  
*menu.cgi*: Displays the menu file given the script.

*parser\_sis1.0.cgi*: The whole center of all the procedures. Calls all the programs and executes the mutational analysis. Once results are achieved it calls *display.cgi* to display the results menu.  
*parser\_gen.cgi*: It does the same job as *parser\_sis1.0.cgi*, but only in case when user submits his own sequence.

*interpret\_mut1.3.pl*, *interpret\_gen.pl*: Reads the mutation entered, draws out exon information from the database, forms the delila instructions and forms the inst file.

*set\_natural1.1.bak*: This bash script sets the ground for the delila and scan, by forming the directories according to the Ri(b,l) matrix and copying appropriate files to appropriate directories. Once the ground work is done, delila is executed and in case of no errors, runs scan and forms the data file. It calls *real1.1.pl* program to analyze and categorize results into decreased, increased and equal parts.

*phenotypes.pl*: This calculates the total information content of binding sites and finds out the prospective isoforms.

*before\_after.r*: This is an R file which generates the graphical comparison of the abundance of prospective isoforms.

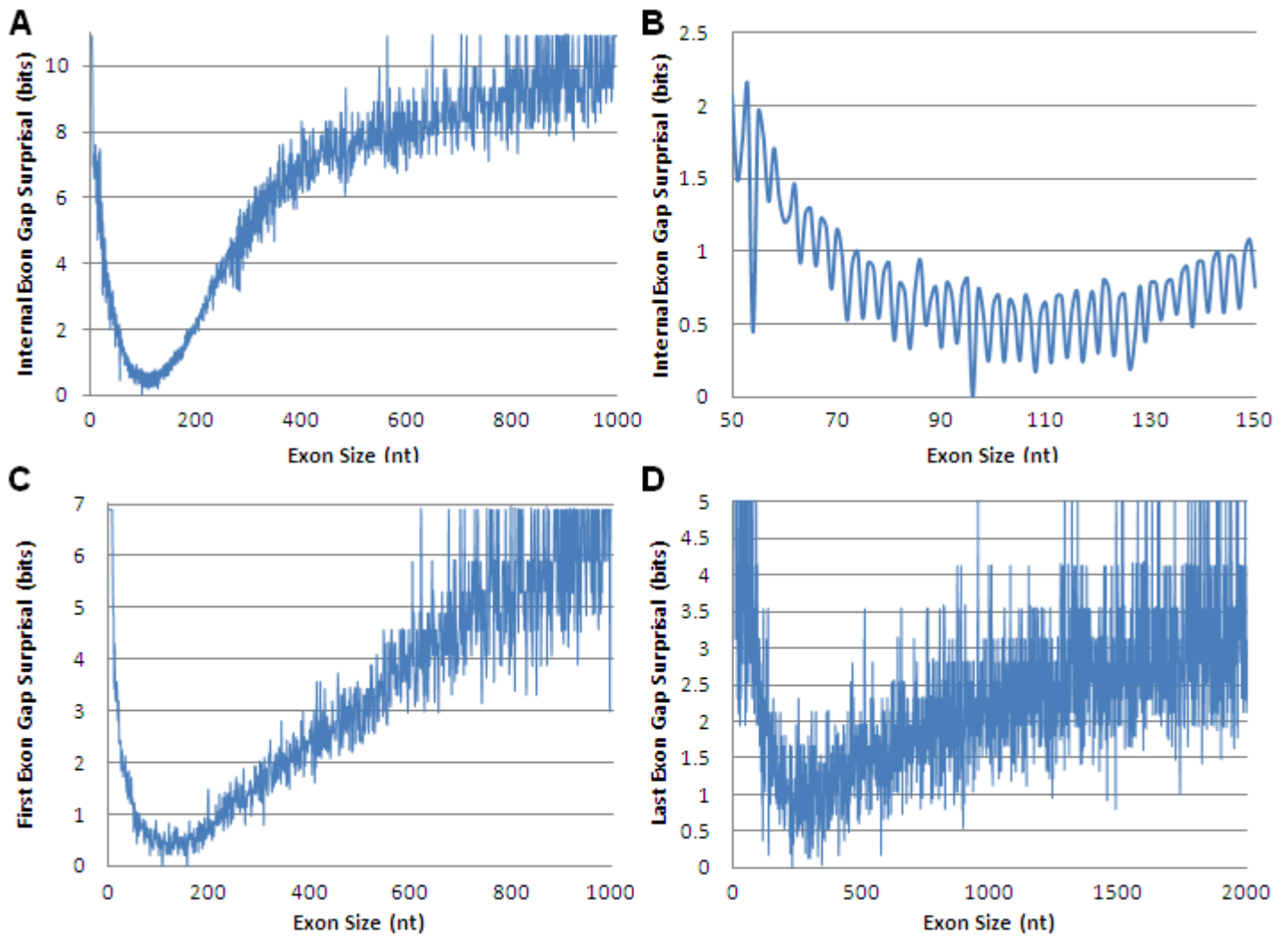
*final\_graphics.pl*: This generates the graphical structures of prospective isoforms.

*generate\_menu\_sis.pl*: This generates the results menu page, dynamically according number of ribl matrices the user selected to analyze.

*runlister*: This runs lister program and generates map pdf file.

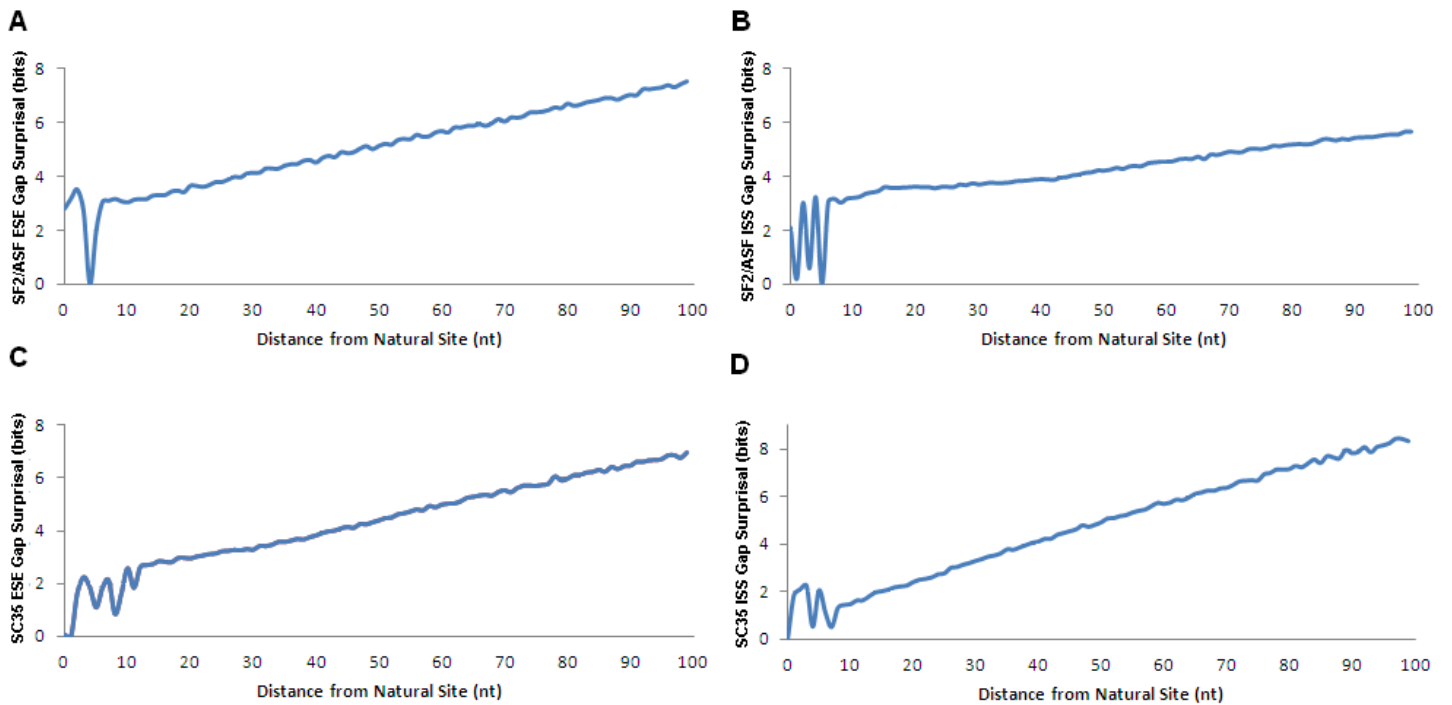
*display.cgi*: It displays the various result pages to the user.

*real1.1.cgi*: Scans through the data file generated by the scan, analyses the results and categorizes the results according to no change, decrease and increase in information content. The results are displayed in html pages.



### Supplementary Figure S3 – Gap Surprisal Distributions for constitutive splice sites of all human exons

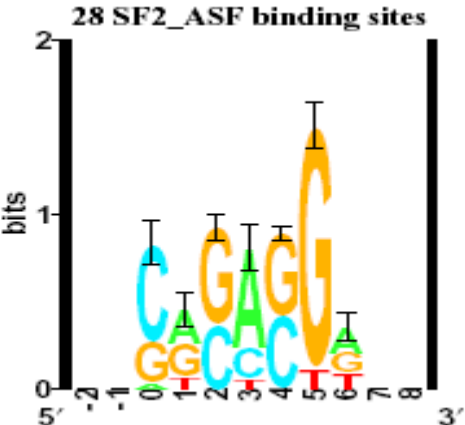
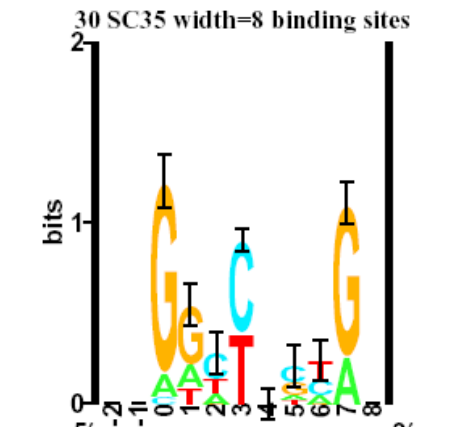
The gap surprisal distribution is computed from the length and frequency of all exons in the genome (see Supplemental methods S1). The length is based on the set of distances between the constitutive donor to acceptor. The results are truncated in the Figure to indicate distributions for exons  $\leq 2000$ nt in length. The gap surprisals are separated by category of exon: internal (panel A), first (panel C) and last (panel D) exons of genes. To illustrate the apparent triplet periodicity of the gap surprisal function associated with open reading frames in exons of common length (50-150 nt), we include panel B. Exons were extracted from the RefSeq database at the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/RefSeq/>).



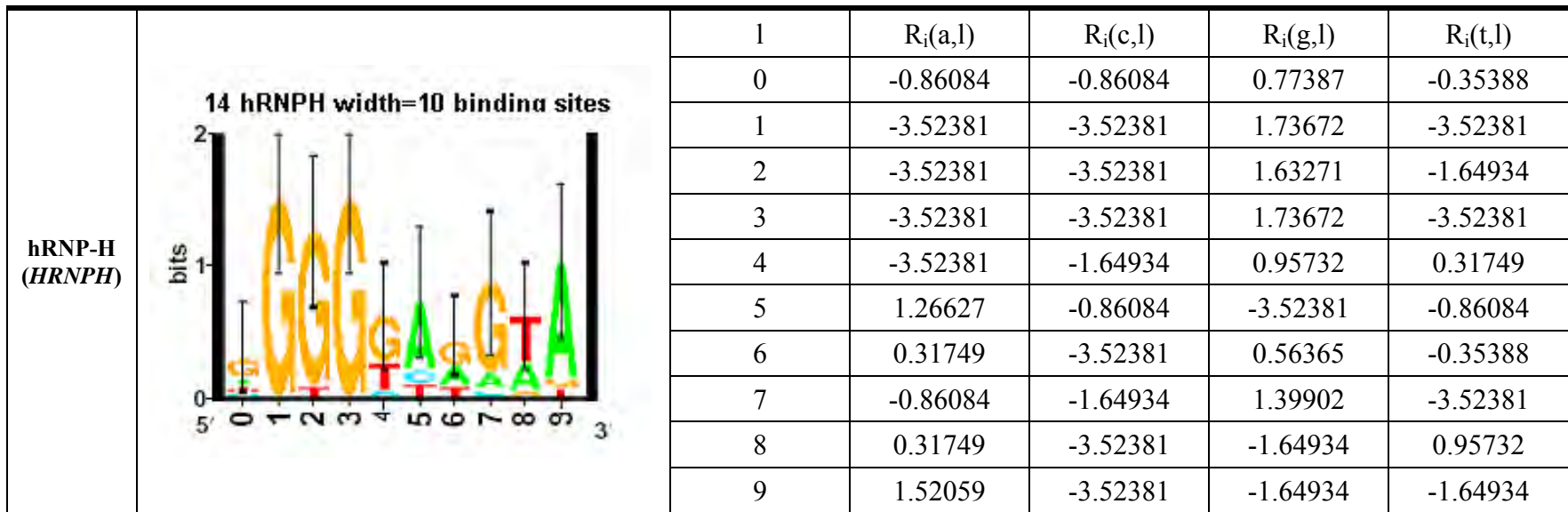
### Supplementary Figure S4 – Gap Surprisal Distributions for SF2/ASF (*SRSF1*) and SC35 (*SRSF2*) sites adjacent to constitutive splice sites in introns and exons

Gap surprisal function distributions were derived for splicing regulatory sequences from the inter-site distance (nt) between all predicted sites of one type (either SC35 or SF2/ASF site) to the nearest constitutive splice site (either donor or acceptor). These distributions are computed separately for intron and exon locations of splicing regulatory sequences. The gap surprisal term and the  $R_i$  value of the corresponding site are added to the other elements of  $R_{i,\text{total}}$ . The contributions of these terms (ie. their signs) are assigned based on whether a binding site is treated as an ISS ( $R_i < 0$ ;  $g(L_{pq}) > 0$ ) or as an ESE ( $R_i > 0$ ;  $g(L_{pq}) < 0$ ). The gap surprisal distributions are displayed for SF2/ASF exonic (A); SF2/ASF intronic (B); SC35 exonic (C); SC35 intronic (D). The windows are truncated at exons  $\leq 100$ nt in the images, however the software computation spans all possible inter-site lengths. A constant value is added to the computed gap surprisal to normalize the values so that the most common intersite distances are not penalized. For SF2/ASF, the most frequent exonic location was at position +4 relative to the splice site (normalization constant: 2.54 bits) and intron location was at position -2 (normalization constant: 3.25 bits). For SC35, the highest frequency exonic location was at position +1 (normalization constant: 3.40 bits) and intronic location was at position -1 (normalization constant: 3.33 bits).

**Supplementary Table S1 - Sequence Logo and Weight Matrix of Splicing Regulatory Sequence Binding Sites**

Enhancer Type	Sequence Logo	Position Weight Matrix*				
ASF/SF2 ( <i>SRSF1</i> )	 <p>28 SF2_ASF binding sites</p>	1	$R_i(a,l)$	$R_i(c,l)$	$R_i(g,l)$	$R_i(t,l)$
		0	-2.965775	1.282153	0.034225	-4.906891
		1	0.619188	-4.906891	0.619188	-0.965775
		2	-4.906891	0.493657	1.121688	-4.906891
		3	1.356153	-0.380812	-4.906891	-1.965775
		4	-4.906891	0.734665	0.941116	-4.906891
		5	-4.906891	-4.906891	1.734665	-1.965775
		6	0.619188	-4.906891	0.204150	-0.158420
SC35 ( <i>SRSF2</i> )	 <p>30 SC35 width=8 binding sites</p>		$R_i(a,l)$	$R_i(c,l)$	$R_i(g,l)$	$R_i(t,l)$
		0	-2.036550	-3.036550	1.663889	-3.036550
		1	-0.229200	-5.000000	1.133374	-0.714620
		2	-0.451590	0.770804	-3.036550	0.133374
		3	-5.000000	1.050912	-5.000000	0.663889
		4	-0.451590	0.285377	0.133374	-0.714620
		5	-1.451590	0.770804	0.422881	-2.036550
		6	-0.451590	0.133374	-3.036550	0.770804
7	-0.03655	-5.000000	1.422881	-5.000000		

SRp40 (SRSF5)	<p>29 SRp40 width=6 binding sites</p>	1	$R_i(a,l)$	$R_i(c,l)$	$R_i(g,l)$	$R_i(t,l)$
		0	1.356397	-4.95420	-1.38057	-0.380570
		1	-2.965530	1.789357	-2.96553	-4.954200
		2	0.619432	-1.96553	0.034469	-0.158180
		3	-2.965530	-1.96553	1.678325	-2.965530
		4	-4.954200	0.619432	1.121932	-4.954200
		5	0.290497	-0.387570	-0.387570	-0.124540
SRp55 (SRSF6)	<p>34 SRp55 width=7 binding sites</p>		$R_i(a,l)$	$R_i(c,l)$	$R_i(g,l)$	$R_i(t,l)$
		0	-0.747840	0.333690	0.944840	-4.58914
		1	0.054720	-1.926170	-2.714670	1.243750
		2	-0.108010	-0.501680	1.102020	-4.589140
		3	-2.714670	1.43323	-2.714670	-0.291460
		4	-1.419210	-2.71467	1.702420	-4.589140
		5	-4.589140	-4.58914	1.372790	0.200940
6	1.102020	0.054720	-0.747840	-4.589140		



\*Information-based position weight matrices were generated using SELEX (Liu et al., 1998) sequences, as well as the sequences of other sites confirmed in published binding studies. Left: sequence logo with error bars indicating 1 standard deviation. Right :information weight matrix ( $R_i[b,l]$ ).



**Supplementary Table S2 -Analysis of published mRNA splice-altering mutations by information theory-based exon definition analysis**

ID	Publications	Gene	Mutation	$R_{i,total(natural)} \rightarrow$	Predicted Result of ASSEDA Server (200nt window)	Published Interpretation	Concordance with predictions
				$R_{i,total(mutant)}$			
				$R_{i,total(strongest cryptic sites [if applicable])}$			
1	Santisteban et al., 1993 <sup>1</sup>	<i>ADA</i>	NM_000022.2:c.478+6T>A g.43254204T>A	12.4 -> 10.9	Natural exon is weakened	Exon 5 skipping	concordant
2	Santisteban et al., 1993 <sup>1</sup>	<i>ADA</i>	NM_000022.2:c.975+1G>A g.43249658G>A	21.0 -> 2.3	Natural donor is abolished, strongest cryptic exons uses pre-existing donors 8, 48, 10, 94, 4 and 43 nt downstream	4 nt intron retention of Exon 10 (reported splice form ranked 5 <sup>th</sup> by exon definition; $R_{i,total}$ 13.7 bits)	concordant
				13.7-14.9			
3	Sanz et al., 2010 <sup>2</sup>	<i>BRCA1</i> <sup>a</sup>	NM_007300.3:c.212+1G>A g.41258472G>A	15.2 -> -3.4	Natural donor is abolished, strongest cryptic exon uses pre-existing donor 22nt upstream	Exon 5 skipping, and a 22nt exonic deletion splice form	concordant
				14.1			
4	Sanz et al., 2010 <sup>2</sup>	<i>BRCA1</i> <sup>a</sup>	NM_007300.3:c.5340+1G>A g.41209068G>A	11.9 -> -6.8	Natural donor is abolished, and the strongest cryptic exon uses pre-existing donor 87nt downstream.	Exon 21 skipping, and 87nt intron retention splice form	concordant
				6.9			
5	Chen et al., 2006 <sup>3</sup>	<i>BRCA1</i> <sup>a</sup>	NM_007300.3:c.441+1G>A g.41256138G>A	13.5 -> -5.1	Natural donor is abolished, and the strongest cryptic exon uses pre-existing donor 62 nt upstream.	62 nt deletion of Exon 7	concordant
				15.2			
6	Sanz et al., 2010 <sup>2</sup>	<i>BRCA1</i> <sup>a</sup>	NM_007300.3:c.5216-1G>A g.41215391G>A	16.6 -> 5.7	Natural acceptor is abolished, and the strongest cryptic exon uses an acceptor created 1 nt downstream.	1 nt deletion of Exon 19	concordant
				-0.5 -> 14.2			

ID	Publications	Gene	Mutation	$R_{i,total(natural)} \rightarrow$	Predicted Result of ASSEDA Server (200nt window)	Published Interpretation	Concordance with predictions
				$R_{i,total(mutant)}$			
				$R_{i,total(strongest cryptic sites [if applicable])}$			
7	Claes et al., 2002 <sup>4</sup>	<i>BRCA1</i> <sup>a</sup>	NM_007300.3:c.135-6T>G g.41258556T>G	15.2 -> 12.0	Natural acceptor is weakened	Exon 5 skipping	concordant
8	Claes et al., 2003 <sup>5</sup>	<i>BRCA1</i> <sup>a</sup>	NM_007300.3:c.134+3A>C g.41267740A>C	10.9 -> 6.4	Natural donor is weakened, and the strongest cryptic exon uses a pre-existing donor 103 nt downstream	Exon 3 skipping	concordant
				8.2			
9	Caux-Moncoutier et al., 2009 <sup>6</sup>	<i>BRCA1</i> <sup>a</sup>	NM_007300.3:c.213-2A>G g.41256975A>G	8.4 -> -6.4	Natural acceptor is abolished, and the top cryptic exon uses a pre-existing acceptor 59 nt upstream	58 nt intron retention of Exon 5	concordant
				13.8			
10	Gutierrez-Enriquez et al., 2009 <sup>7</sup>	<i>BRCA1</i> <sup>a</sup>	NM_007300.3:c.302-1G>A g.41256279G>A	13.5 -> 2.6	Natural acceptor is abolished, cryptic site activated 31 nt upstream	10 nt deletion of Exon 7 (reported splice form ranked 27 <sup>th</sup> by exon definition; $R_{i,total} - 0.7$ bits) <sup>c</sup>	discordant
				1.9			
11	Campos et al., 2003 <sup>8</sup>	<i>BRCA1</i> <sup>a</sup>	NM_007300.3:c.5256+5G>A g.41215345G>A	16.6 -> 13.4	Natural donor is weakened	Exon 18 skipping	concordant
12	Claes et al., 2002 <sup>4</sup>	<i>BRCA1</i> <sup>a</sup>	NM_007300.3:c.212+3A>G g.41258470A>G	15.2 -> 14.4	Natural donor is weakened by < 1 bit, and the top cryptic exon uses a pre-existing donor 22 nt upstream	Exon 5 skipping, and a 22 nt exonic deletion splice form	concordant <sup>d</sup>
				14.1			
13	Chen et al., 2006 <sup>3</sup>	<i>BRCA1</i> <sup>a</sup>	NM_007300.3:c.5049+6T>G g.41222939T>G	8.4 -> 7.1	Natural donor is weakened, and the strongest cryptic exons uses pre-existing donor sites 95 and 93 nt within exon 16	65 nt intron retention of Exon 17 (reported splice form ranked 6 <sup>th</sup> by exon definition; $R_{i,total} - 6.4$ bits)	concordant
				10.3-10.5			

ID	Publications	Gene	Mutation	$R_{i,total(natural)} \rightarrow$	Predicted Result of ASSEDA Server (200nt window)	Published Interpretation	Concordance with predictions
				$R_{i,total(mutant)}$			
				$R_{i,total(strongest cryptic sites [if applicable])}$			
14	Caux-Moncoutier et al., 2009 <sup>6</sup>	<i>BRCA2</i>	NM_000059.3:c.631+3A>G g.32900753A>G	17.4 -> 16.6	Natural donor is weakened by < 1 bit	Exon 18 skipping	concordant <sup>d</sup>
15	Chen et al., 2006 <sup>3</sup>	<i>BRCA2</i>	NM_000059.3:c.8633-2A>G g.32950805A>G	18.6 -> 3.8	Natural acceptor is abolished, and the top cryptic exons uses acceptors 51, 33 nt upstream	Exon 21 skipping, and a 43 nt exonic deletion splice form	discordant
				14.4-15.0			
16	Caux-Moncoutier et al., 2009 <sup>6</sup>	<i>BRCA2</i>	NM_000059.3:c.7977-7C>G g.32937309C>G	13.5 -> 11.5	Natural acceptor is weakened, and the top cryptic exon uses acceptor created by mutation 6 nt upstream	Exon 18 skipping, and a 6 nt intron retention splice form	concordant
				-1.7 -> 9.9			
17	Claes et al., 2003 <sup>5</sup>	<i>BRCA2</i>	NM_000059.3:c.9118-2A>G g.32954142A>G	19.4 -> 4.6	Natural acceptor is weakened, and the strongest cryptic exon uses pre-existing and slightly strengthened acceptor 7 nt downstream	7 nt deletion of Exon 24	concordant
				14.2 -> 14.5			
18	Sanz et al., 2010 <sup>2</sup>	<i>BRCA2</i>	NM_000059.3:c.8395A>G (p.D2723G) g.32937507A>G	13.5	Natural exon is unaffected, but the strongest cryptic exon uses created cryptic donor 164 nt upstream	163 nt deletion of Exon 18	concordant
				-0.6 -> 18.0			
19	Sanz et al., 2010 <sup>2</sup>	<i>BRCA2</i>	NM_000059.3:c.8262G>T (p.D2679Y) g.32937374G>T	13.5	Natural exon is unaffected, but the strongest cryptic exon uses created cryptic donor 298 nt upstream <sup>e</sup>	298 nt deletion of Exon 18	concordant
				-1.9 -> 16.7			
20	Sanz et al., 2010 <sup>2</sup>	<i>BRCA2</i>	NM_000059.3:c.694A>G (p.D156G) g.32900279A>G	11.7	Natural exon is unaffected, but the strongest cryptic exon uses created cryptic donor 9 nt upstream	9 nt deletion of Exon 5	concordant
				-4.6 -> 14.0			

ID	Publications	Gene	Mutation	$R_{i,total(natural)} \rightarrow$	Predicted Result of ASSEDA Server (200nt window)	Published Interpretation	Concordance with predictions
				$R_{i,total(mutant)}$			
				$R_{i,total(strongest cryptic sites [if applicable])}$			
21	Rutter et al., 2003 <sup>9</sup>	CDKN2A	NM_000077.4:c.457+1G>T g.21970900G>T	13.0 -> -5.7	Natural donor is abolished, and the strongest cryptic exons uses pre-existing donor sites found 74 and 175 nt upstream	Exon 2 skipping and a 74 nt exonic deletion splice form (reported splice form ranked 2 <sup>nd</sup> by exon definition; $R_{i,total}$ 8.0 bits)	concordant
				8.0-9.1			
22	Harland et al., 2001 <sup>10</sup>	CDKN2A	NM_000077.4:c.458-105A>G g.21968346A>G	-	Cryptic donor 227 nt downstream of exon is activated by mutation ( $R_i$ -16.0 -> 2.7 bits), but an exon using this site is not predicted	227 nt intron retention of Exon 2	discordant
23	Macias-Vidal et al., 2009 <sup>11</sup>	CTNS	NM_004937.2:c.61+5G>A g.3543566G>A	19.7 -> 16.6	Natural donor is weakened	Exon 3 skipping	concordant
24	Tompson et al., 2007 <sup>12</sup>	EVC	NM_153717.2:c.1886+5G>T g.5795449G>T	9.4 -> 6.0	Natural donor is weakened, and the top cryptic exon uses pre-existing donor site 115 nt downstream	Exon 13 skipping, and a 115 nt intron retention splice form	concordant
				5.8			
25	Arranz et al., 2002 <sup>13</sup>	FAH	NM_000137.1:c.554-1G>T g.80460605G>T	20.3 -> 12.6	Natural acceptor is weakened, and the strongest cryptic exon uses pre-existing acceptor 12 nt upstream	5 nt deletion of Exon 7 due to strengthening of a cryptic site (reported splice form ranked 4 <sup>th</sup> by exon definition; $R_{i,total}$ 8.0 -> 11.1 bits)	concordant
				16.6			
26	Arranz et al., 2002 <sup>13</sup>	FAH	NM_000137.1:c.707-1G>T g.80465355G>T	16.5 -> 8.7	Natural acceptor is weakened, and the strongest cryptic exon uses pre-existing acceptor 70 nt upstream	Exon 9 skipping	concordant
				9.1			

ID	Publications	Gene	Mutation	$R_{i,total(natural)} \rightarrow$	Predicted Result of ASSEDA Server (200nt window)	Published Interpretation	Concordance with predictions
				$R_{i,total(mutant)}$			
				$R_{i,total(strongest cryptic sites [if applicable])}$			
27	Schloesser et al., 1995 <sup>14</sup>	F12	NM_000505.3:c.1681-1G>A g.176829461G>A	-0.1 $\rightarrow$ -11.0	Natural acceptor is weakened, and the strongest cryptic exon uses pre-existing acceptor 176 nt downstream	1 nt deletion of Exon 14 (reported splice form ranked 4 <sup>th</sup> by exon definition; $R_{i,total}$ -15.7 $\rightarrow$ -1.0 bits)	discordant
				1.8			
28	Lapoumeroulie et al., 1987 <sup>15</sup>	HBB	NM_000518.4:c.92+5C>T g.5248155G>A	4.1 $\rightarrow$ 0.9	Natural donor is weakened, and the strongest cryptic exons uses pre-existing donors 16, 38 nt upstream	16 and 38 nt deletion splice forms of Exon 1	concordant
				5.4-5.5			
29	Treisman et al., 1983 <sup>16</sup>	HBB	NM_000518.4:c.92+5G>C g.5248155G>C	4.1 $\rightarrow$ 0.2	Natural donor is weakened, and the strongest cryptic exons uses pre-existing donors 16, 38 nt upstream	16 and 38 nt deletion splice forms of Exon 1	concordant
				5.4-5.5			
30	Treisman et al., 1983 <sup>16</sup>	HBB	NM_000518.4:c.92+6T>C g.5248154T>C	4.1 $\rightarrow$ 2.4	Natural donor is weakened, and the strongest cryptic exons uses pre-existing donors 16, 38 nt upstream	16 and 38 nt deletion splice forms of Exon 1	concordant
				5.4-5.5			
31	Vidaud et al., 1989 <sup>17</sup>	HBB	NM_000518.4:c.92+1G>A g.5248159G>A	4.1 $\rightarrow$ -14.5	Natural donor site is abolished, and the strongest cryptic exon uses pre-existing donors 16, 38 nt upstream	16 and 38 nt deletion splice forms of Exon 1	concordant
				5.4-5.5			
32	Vidaud et al., 1989 <sup>17</sup>	HBB	NM_000518.4:c.93-2A>G g.5248031A>G	12.1 $\rightarrow$ -2.6	Natural acceptor is abolished, and the strongest cryptic exon uses an acceptor created 1 nt upstream	15 nt deletion of Exon 2	discordant
				-5.9 $\rightarrow$ 5.0			

ID	Publications	Gene	Mutation	$R_{i,total(natural)} \rightarrow$	Predicted Result of ASSEDA Server (200nt window)	Published Interpretation	Concordance with predictions
				$R_{i,total(mutant)}$			
				$R_{i,total(strongest cryptic sites [if applicable])}$			
33	Atweh et al., 1985 <sup>18</sup>	<i>HBB</i>	NM_000518.4:c.316-2A>G g.5246958A>G	8.9 $\rightarrow$ -5.8	Natural acceptor is abolished, and the strongest cryptic exon uses pre-existing acceptor 271 nt upstream <sup>e</sup>	~270 nt deletion of Exon 3	concordant
				9.1			
34	Bunge et al., 1993 <sup>19</sup>	<i>IDS</i>	NM_000202.5:c.1007-1G>C g.148568630G>C	6.5 $\rightarrow$ -5.2	Natural acceptor is abolished, and the top cryptic exon with pre-existing acceptor 82 nt upstream	12 nt deletion of Exon 8 (reported splice form ranked 5 <sup>th</sup> by exon definition; $R_{i,total}$ 2.0 $\rightarrow$ 3.4 bits)	concordant
				9.1			
35	Bunge et al., 1993 <sup>19</sup>	<i>IDS</i>	NM_000202.5:c.880-2A>G g.148571973A>G	18.8 $\rightarrow$ 4.1	Natural acceptor is weakened, and the strongest cryptic exons uses pre-existing acceptor sites activated 6, 7 nt upstream, and 51nt downstream	51 nt deletion of Exon 7	concordant
				9.9-10.1			
36	Bunge et al., 1993 <sup>19</sup>	<i>IDS</i>	NM_000202.5:c.419-2A>G g.148582570A>G	21.4 $\rightarrow$ 6.7	Natural acceptor is weakened, and the strongest cryptic exon uses strengthened acceptor 1nt upstream	1 nt intron inclusion of Exon 4	concordant
				13.7 $\rightarrow$ 24.6			
37	Erdmann et al., 2001 <sup>20</sup>	<i>MYBPC3<sup>b</sup></i>	NM_000256.3:c.772+1G>A g.47369974G>A	8.4 $\rightarrow$ -10.3	Natural donor is abolished, and the top cryptic exons uses pre-existing donors 14, 169nt downstream and 52, 57 nt upstream	Exon 7 skipping	concordant
				2.6-3.6			

ID	Publications	Gene	Mutation	$R_{i,total(natural)} \rightarrow$	Predicted Result of ASSEDA Server (200nt window)	Published Interpretation	Concordance with predictions
				$R_{i,total(mutant)}$			
				$R_{i,total(strongest cryptic sites [if applicable])}$			
38	Erdmann et al., 2001 <sup>20</sup>	MYBPC3 <sup>b</sup>	NM_000256.3:c.1928-2A>G g.47361343A>G	9.6 -> -5.1	Natural acceptor is weakened, and the strongest cryptic exon uses pre-existing acceptor 30 and 60 nt downstream, and 22 nt upstream	11 nt deletion of Exon 21 (reported splice form ranked 4 <sup>th</sup> by exon definition; $R_{i,total}$ 1.5 bits) <sup>c</sup>	concordant
				2.2-2.4			
39	Dworniczak et al., 1991 <sup>21</sup>	PAH	NM_000277.1:c.1066-11G>A g.103237568G>A	11.6 -> 11.3	Natural exon change is insignificant, but the strongest cryptic exon uses acceptor created 9 nt upstream by the mutation	9 nt intron inclusion of Exon 12	concordant
				-4.6 -> 10.2			
40	Maciolek et al., 2006 <sup>22</sup>	PITX2	NM_153427.1:c.252+5G>C g.111542315G>C	7.9 -> 4.0	Natural donor is weakened	Complete Intron Retention	concordant
41	Vega et al., 2009 <sup>23</sup>	PMM2	NM_000303.2:c.448-1G>C g.8905494G>C	13.8 -> 2.1	Natural acceptor is abolished, and the cryptic exon using a cryptic site 4nt downstream is strengthened	Exon 3 skipping	concordant
				6.5 -> 8.9			

Published mutations known to affect mRNA splicing in various genes were analyzed using information theory based exon definition analysis. Mutations are given in both HGVS g. and c. format (c. format is mRNA dependent; position 1 is the A of the start codon). The  $\Delta R_{i,total}$  values of mutations of the natural exon resulting from that mutation (as well as potential cryptic exons) are shown in the adjacent column. Interpretations of mutant exons predicted by ASSEDA relative to the published results are also reported. ND = No data, *BRCA1*<sup>a</sup> All mutations for BRCA1 were adjusted by 1 having designation beyond exon 4, when IVS notation is used MYBPC3<sup>b</sup> All IVS mutations for MYBPC3 were adjusted by 1 when IVS notation is used. <sup>c</sup> Must allow negative  $R_i$  values in advanced settings for server to report cryptic exon. <sup>d</sup> These mutations cause an information decrease of just under 1 bit. We call these concordant because they do show a decrease as expected, and any activated cryptic sites detected and closely related in  $R_{i,total}$ . <sup>e</sup> Must expand window range to 500 nt for server to report this cryptic exon.

## References for mutations in Supp. Table S2

- <sup>1</sup> Santisteban I, Arredondo-Vega FX, Kelly S, Mary A, Fischer A, Hummell DS, Lawton A, Sorensen RU, Stiehm ER, Uribe L. 1993. Novel splicing, missense, and deletion mutations in seven adenosine deaminase-deficient patients with late/delayed onset of combined immunodeficiency disease. Contribution of genotype to phenotype. *J Clin Invest* 92:2291-2302.
- <sup>2</sup> Sanz DJ, Acedo A, Infante M, Duran M, Perez-Cabornero L, Esteban-Cardenosa E, Lastra E, Pagani F, Miner C, Velasco EA. 2010. A high proportion of DNA variants of BRCA1 and BRCA2 is associated with aberrant splicing in breast/ovarian cancer patients. *Clin Cancer Res* 16:1957-67.
- <sup>3</sup> Chen X, Truong TT, Weaver J, Bove BA, Cattie K, Armstrong BA, Daly MB, Godwin AK. 2006. Intronic alterations in BRCA1 and BRCA2: effect on mRNA splicing fidelity and expression. *Hum Mutat* 27:427-435.
- <sup>4</sup> Claes K, Vandesompele J, Poppe B, Dahan K, Coene I, De Paepe A, Messiaen L. 2002. Pathological splice mutations outside the invariant AG/GT splice sites of BRCA1 exon 5 increase alternative transcript levels in the 5' end of the BRCA1 gene. *Oncogene* 21:4171-4175.
- <sup>5</sup> Claes K, Poppe B, Machackova E, Coene I, Foretova L, De Paepe A, and Messiaen L. 2003. Differentiating pathogenic mutations from polymorphic alterations in the splice sites of BRCA1 and BRCA2. *Genes Chromosomes Cancer* 37:314-320.
- <sup>6</sup> Caux-Moncoutier V, Pages-Berhouet S, Michaux D, Asselain B, Castera L, De Pauw A, Buecher B, Gauthier-Villars M, Stoppa-Lyonnet D, Houdayer C. 2009. Impact of BRCA1 and BRCA2 variants on splicing: clues from an allelic imbalance study. *Eur J Hum Genet* 17:1471-1480.
- <sup>7</sup> Gutierrez-Enriquez S, Coderch V, Masas M, Balmana J, Diez O. 2009. The variants BRCA1 IVS6-1G>A and BRCA2 IVS15+1G>A lead to aberrant splicing of the transcripts. *Breast Cancer Res Treat* 117:461-465.
- <sup>8</sup> Campos B, Diez O, Domenech M, Baena M, Balmana J, Sanz J, Ramirez A, Alonso C, Baiget M. 2003. RNA analysis of eight BRCA1 and BRCA2 unclassified variants identified in breast/ovarian cancer families from Spain. *Hum Mutat* 22:337.
- <sup>9</sup> Rutter JL, Goldstein AM, Davila MR, Tucker MA, Struewing JP. 2003. CDKN2A point mutations D153spl(c.457G>T) and IVS2+1G>T result in aberrant splice products affecting both p16INK4a and p14ARF. *Oncogene* 22:4444-8.
- <sup>10</sup> Harland M, Mistry S, Bishop DT, Bishop JAN. 2001. A deep intronic mutation in CDKN2A is associated with disease in a subset of melanoma pedigrees. *Hum Mol Genet* 23:2679-2686.
- <sup>11</sup> Macias-Vidal J, Rodes M, Hernandez-Perez JM, Vilaseca MA, Coll MJ. 2009. Analysis of the CTNS gene in 32 cystinosis patients from Spain. *Clin Genet* 76:486-489.
- <sup>12</sup> Tompson SW, Ruiz-Perez VL, Blair HJ, Barton S, Navarro V, Robson JL, Wright MJ, Goodship JA. 2007. Sequencing EVC and EVC2 identifies mutations in two-thirds of Ellis-van Creveld syndrome patients. *Hum Genet* 120:663-670.
- <sup>13</sup> Arranz JA, Pinol F, Kozak L, Perez-Cerda C, Cormand B, Ugarte M, Riudor E. 2002. Splicing mutations, mainly IVS6-1(G>T), account for 70% of fumarylacetoacetate hydrolase (FAH) gene alterations, including 7 novel mutations, in a survey of 29 tyrosinemia type I patients. *Hum Mutat* 20:180-188.



- <sup>14</sup> Schloesser M, Hofferbert S, Bartz U, Lutze G, Lammle B, Engel W. 1995. The novel acceptor splice site mutation 11396(G-->A) in the factor XII gene causes a truncated transcript in cross-reacting material negative patients. *Hum Mol Genet* 4:1235-1237.
- <sup>15</sup> Lapoumeroulie C, Acuto S, Rouabhi F, Labie D, Krishnamoorthy R, Bank A. 1987. Expression of a beta thalassemia gene with abnormal splicing. *Nucleic Acids Res* 15:8195-8204.
- <sup>16</sup> Treisman R, Orkin SH, Maniatis T. 1983. Specific transcription and RNA splicing defects in five cloned beta-thalassaemia genes. *Nature* 302: 591-596.
- <sup>17</sup> Vidaud M, Gattoni R, Stevenin J, Vidaud D, Amselem S, Chibani J, Rosa J, Goossens M. 1989. A 5' splice-region G----C mutation in exon 1 of the human beta-globin gene inhibits pre-mRNA splicing: a mechanism for beta+-thalassemia. *Proc Natl Acad Sci U S A* 86:1041-1045.
- <sup>18</sup> Atweh GF, Anagnou NP, Shearin J, Forget BG, Kaufman RE. 1985. Beta-thalassemia resulting from a single nucleotide substitution in an acceptor splice site. *Nucleic Acids Res* 13:777-790.
- <sup>19</sup> Bunge S, Steglich C, Zuther C, Beck M, Morris CP, Schwinger E, Schinzel A, Hopwood JJ, Gal A. 1993. Iduronate-2-sulfatase gene mutations in 16 patients with mucopolysaccharidosis type II (Hunter syndrome). *Hum Mol Genet* 2:1871-1875.
- <sup>20</sup> Erdmann J, Raible J, Maki-Abadi J, Hummel M, Hammann J, Wollnik B, Frantz E, Fleck E, Hetzer R, Regitz-Zagrosek V. 2001. Spectrum of clinical phenotypes and gene variants in cardiac myosin-binding protein C mutation carriers with hypertrophic cardiomyopathy. *J Am Coll Cardiol* 38:322-330.
- <sup>21</sup> Dworniczak B, Aulehla-Scholz C, Kalaydjieva L, Bartholome K, Grudde K, Horst J. 1991. Aberrant splicing of phenylalanine hydroxylase mRNA: the major cause for phenylketonuria in parts of southern Europe. *Genomics* 11:242-246.
- <sup>22</sup> Maciolek NL, Alward WL, Murray JC, Semina EV, McNally MT. 2006. Analysis of RNA splicing defects in PITX2 mutants supports a gene dosage model of Axenfeld-Rieger syndrome. *BMC Med Genet* 7:59.
- <sup>23</sup> Vega AI, Pérez-Cerdá C, Desviat LR, Matthijs G, Ugarte M, Pérez B. 2009. Functional analysis of three splicing mutations identified in the PMM2 gene: toward a new therapy for congenital disorder of glycosylation type Ia. *Hum Mutat* 30:795-803.

**Supplementary Table S3 - Validation of information theory based exon definition analysis- of mRNA splice-altering mutations by qRT-PCR**

ID	Publications	Gene	Mutation	$R_{i,total(natural)} \rightarrow$	Predicted Result of ASSEDA Server (200nt window)	Published Interpretation	Concordance with predictions (quantification)
				$R_{i,total(mutant)}$			
				$R_{i,total(strongest cryptic sites [if applicable])}$			
1	Sankaran et al., 2012 <sup>1</sup>	<i>GATA1</i>	NM_002049.3:c.220G>C g.48649736G>C	11.9 -> 7.0	Affected donor has 3.4% of its original strength	3-5% exon retention in homozygotes	concordant
				7.3 (151nt upstr.)			
2	Cavallari et al., 2012 <sup>2</sup>	<i>F7</i>	NM_019616.2:c.615+1G>T g.113771190G>T	18.4 -> -0.3	Affected donor should be inactive ; should increase skipping and/or use of cryptic donor 30 nt downstream of wildtype	Wildtype splicing not detected. Cryptic site is used, but exon skipping occurs 5 fold more often.	concordant <sup>a</sup>
				14.3 (30nt downstr.)			
3	Clavero et al., 2004 <sup>3</sup>	<i>PCCA</i>	NM_000282.3:c.1899+3del4 g.101101562del4	17.8 -> 13.7	Due to the 4.1 bit decrease, splice donor has 5.8% of its original strength ; potential exon skipping	1-6.3% exon retention in homozygotes	concordant
4	Clavero et al., 2004 <sup>3</sup>	<i>PCCB</i>	NM_000532.4:c.183+2T>C g.135969402T>C	4.2 -> -2.9	Affected donor should be inactive ; should increase use of cryptic donor 20 nt downstream of wildtype exon 1	Abundance of cryptic site splice form is 9-18% compared to control wildtype splicing; decrease likely due to NMD	concordant <sup>b</sup>
				4.0 (cryptic 20 nt downstream)			

5	Clavero et al., 2004 <sup>3</sup>	<i>PCCB</i>	NM_000532.4:c.653A>G g.136002788A>G	18.9 -> 16.4	Mutation weakens wildtype splice form to 17.7% of original strength; 6nt cryptic site strengthened to near equivalent strength	Wildtype splicing not detected; cryptic splice site use detected but not quantified	<b>discordant for quantitation; concordant for mRNA structure</b>
				13.6 -> 16.8 (cryptic 6nt upstream.)			
6	Martinez et al., 2012 <sup>4</sup>	<i>NSUN2</i>	NM_017755.5:c.538-1C>G g.6622214G>C	12.6 -> 0.9	Predict abolishment of splice acceptor ; should increase skipping and/or use of cryptic donor 7 nt downstream of wildtype	95% reduction of wildtype splicing; 70-80% reduction in total NSUN2 mRNA; cryptic site use not detected	<b>discordant for quantitation; concordant for mRNA structure</b>
				9.8 -> 11.5 (7nt downstream)			
7	Chouery et al., 2008 <sup>5</sup>	<i>TREM2</i>	NM_018965.2:c.40+3delAGG g.41130776delCCT	6.7 -> 2.4	Affected donor has 5.1% of its original strength	Detect “over” 2-fold drop of wildtype splicing in heterozygote ; no cryptic site use detected	concordant
				5.0 (227 nt intronic)			
8	Tanner et al., 2009 <sup>6</sup>	<i>RHO</i>	NM_000539.3:c.936G>A g.129251615G>A	2.6 -> -0.4	Normal splicing is weakened 8-fold; potential skipping and cryptic site use ; 4 nt upstream splice site detected ( -1.8 -> 0.8; ranked 2 <sup>nd</sup> )	5 fold increase in skipping ; 4nt upstream cryptic site is used (not directly quantified; visually equivalent to skipping)	concordant <sup>a</sup>
				3.1 (13nt downstr.)			

Mutations which were annotated with quantifiable methods were directly compared with ASSEDA results to assess accuracy of predicted binding affinity changes. While mRNA structure predictions were concordant, predicted levels of wildtype expression for mutations #5 and 6 were not accurate (predicted to be abolished but remained active and vis versa). Mutations are given in both HGVS g. and c. format (c. format is mRNA dependent; position 1 is the A of the start codon). <sup>a</sup>Relative abundance of cryptic isoform vs. exon skipping events cannot be inferred from these results. <sup>b</sup>Reduced levels of cryptic splice form may be due to activation of nonsense mediated decay, since codon phase is shifted in the cryptic exon.

References for mutations in Supp. Table S3

- <sup>1</sup> Sankaran VG, Ghazvinian R, Do R, Thiru P, Vergilio JA, Beggs AH, Sieff CA, Orkin SH, Nathan DG, Lander ES, Gazda HT. 2012. Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. *J Clin Invest.* 122(7):2439-43.
- <sup>2</sup> Cavallari N, Balestra D, Branchini A, Maestri I, Chuamsunrit A, Sasanakul W, Mariani G, Pagani F, Bernardi F, Pinotti M. 2012. Activation of a cryptic splice site in a potentially lethal coagulation defect accounts for a functional protein variant. *Biochim Biophys Acta.* 1822(7):1109-13.
- <sup>3</sup> Clavero S, Pérez B, Rincón A, Ugarte M, Desviat LR. 2004. Qualitative and quantitative analysis of the effect of splicing mutations in propionic acidemia underlying non-severe phenotypes. *Hum Genet.* 115(3):239-47.
- <sup>4</sup> Martinez FJ, Lee JH, Lee JE, Blanco S, Nickerson E, Gabriel S, Frye M, Al-Gazali L, Gleeson JG. 2012. Whole exome sequencing identifies a splicing mutation in NSUN2 as a cause of a Dubowitz-like syndrome. *J Med Genet.* 49(6):380-5.
- <sup>5</sup> Chouery E, Delague V, Bergougnoux A, Koussa S, Serre JL, Mégarbané A. 2008. Mutations in TREM2 lead to pure early-onset dementia without bone cysts. *Hum Mutat.* 29(9):E194-204.
- <sup>6</sup> Tanner G, Glaus E, Barthelmes D, Ader M, Fleischhauer J, Pagani F, Berger W, Neidhardt J. 2009. Therapeutic strategy to rescue mutation-induced exon skipping in rhodopsin by adaptation of U1 snRNA. *Hum Mutat.* 30(2):255-63.

**Supplementary Table S4 -Analysis of published regulatory ESE/ISS mutations altering mRNA splicing by exon definition analysis**

ID	Publications	Gene	Mutation	$R_{i,total} \text{ (natural)} \rightarrow$	Predicted Result of ASSEDA Server (200nt window)	Published Interpretation	Concordance with predictions
				$R_{i,total} \text{ (mutant)}$			
<b>SF2/ASF (<i>SRSF1</i>)</b>							
1	Miyajima et al., 2002 <sup>1</sup>	<i>SMN1</i> / <i>SMN2</i>	NM_000344.3:c.840C>T g.70247773C>T	16.4 -> 10.7 ( <i>SMN1</i> -> <i>SMN2</i> )	Exon 7 of SMN1 is predicted to be 5.7 bits stronger than SMN2 due to abolished SF2/ASF 6nt upstream of natural acceptor	Single nucleotide difference leads to increased exon skipping in SMN2 compared to SMN1; due to SF2/ASF site	concordant
2	Heintz et al., 2012 <sup>2</sup>	<i>PAH</i>	NM_000277.1:c.1144T>C g.103237478T>C	9.7 -> 5.4	ESE 61nt from natural donor is abolished, decreasing $R_{i,total}$ of natural exon by 4.3 bits (SRp40 site also weakened).	Mutation affects SF2/ASF, SRp20 and SRp40 sites leading to exon skipping.	concordant
				3.3 -> 1.6 (SRp40)			
3	Sun et al., 2009 <sup>3</sup>	<i>RPTOR</i>	NM_020761.2:c.1518A>G g.78854223A>G	6.0 -> 11.6	SF2/ASF is strengthened, increasing natural exon strength by 5.6 bits. SRp55 site is weakened by the same mutation.	G-allele weakens an SRp55 site while strengthening an SF2/ASF site. Increased exon skipping for G-allele is consistent with the possibility that SF2/ASF site may be redundant.	concordant <sup>a</sup>
				6.0 -> 3.2 (SRp55)			
4	Fukao et al., 2010 <sup>4</sup>	<i>ACAT1</i>	NM_000019.3:c.951C>T g.108014720C>T	11.2 -> 6.0	ESE is abolished, decreasing natural site strength by 5.2 bits.	Mutation increases exon 10 skipping by mini-gene analysis.	concordant

5	Gonçalves et al., 2009 <sup>5</sup>	<i>APC</i>	NM_000038.5:c.1918C>G g.112170822C>G	10.9 -> 10.9	No ESE (where $R_i > 0$ bits) is predicted to be altered by mutation with available weight matrices.	SF2/ASF site required for exon 14 inclusion. Mutation weakens ESE and leads to exon 14 skipping.	discordant
6	Burgess et al., 2009 <sup>6</sup>	<i>BEST1</i>	NM_004183.3:c.704T>C g.61724926T>C	18.6 -> 14.3	SF2/ASF site is weakened, leading to a 4.3 bit decrease in natural exon $R_{i,total}$ .	Mutation weakens splicing in ESE-dependant splice assay. Increased exon skipping.	concordant
7	Burgess et al., 2009 <sup>6</sup>	<i>BEST1</i>	NM_004183.3:c.707G>A g.61724929G>A	18.5 -> 18.5	No ESE (where $R_i > 0$ bits) is predicted to be altered by mutation with available weight matrices.	Mutation weakens splicing in ESE-dependant splice assay. Increased exon 6 skipping.	discordant
<b>SC35 (<i>SRSF2</i>)</b>							
8	Jensen et al., 2010 <sup>7</sup>	<i>MOG</i>	NM_002433.4:c.520A>G g.29634012A>G	4.7 -> 2.5	Weak SC35 site is abolished. No other changes in enhancer / repressor elements are predicted.	Mutation leads to increased exon 3 inclusion, suggesting created SRp55 site is active	concordant for SC35 (SRp55 discordant)
9	Tran et al., 2006 <sup>8</sup>	<i>DMD</i>	NM_004012.3:c.1405delTTCA g.32366534delTTCA	14.5 -> 8.9	SC35 is abolished, leading to a 6.1 bit decrease in natural exon $R_{i,total}$	Hybrid minigene assay shows complete exon 38 skipping in presence of mutation	concordant

10	Gabut et al., 2005 <sup>9</sup>	<i>PDHAI</i>	NM_000284.3:c.759+26G>A g.19373648G>A	10.0 (natural $R_{i,total}$ )	Mutation strengthens SC35 downstream of exon 7, activating cryptic donor, increasing cryptic exon splice form $R_{i,total}$ by 2.2 bits	Intronic mutation activates use of cryptic donor by strengthening of pre-existing SC35 site	concordant
				11.3 -> 13.5 (cryptic $R_{i,total}$ )			
11	Colapietro et al, 2003 <sup>10</sup>	<i>NFI</i>	NM_000267.2:c.945GC>AA g.29527496GC>AA	5.6 -> 3.1 ( $R_{i,total}$ [SC35 activated])	Mutation decreases the strength of pre-existing SF2/ASF and 2 sites, leading to a lowered $R_{i,total}$ (3.8 and 2.5 bits, respectively).	Exonic mutation leads to increased exon 7 skipping.	concordant
				4.0 -> 0.2 ( $R_{i,total}$ [SF2/ASF activated])			
12	Raponi et al., 2011 <sup>11</sup>	<i>BRCA1</i>	NM_007300.2:c.231G>T g.41256955C>A	8.4 -> 8.4	Weak SRp40 site is predicted to be abolished by the mutation ; two weakened SRSF1 sites have $R_i < 0$ bits and are filtered out	Increased exon 6 skipping ; Association with proteins 9G8, SC35, SF2/ASF, Tra2, and hnRNP A1 is altered <sup>b</sup>	discordant
				0.2 -> -4.4 (SRp40 site)			

Published mutations known to affect mRNA splicing by altering either SF2/ASF or SC35 splice enhancer elements were analyzed using information theory based exon definition analysis, with the appropriate ESE/ISS advanced option activated (must specify splice enhancer type to test). The  $\Delta R_{i,total}$  values of mutations of the natural exon resulting from that mutation (as well as potential cryptic exons) are shown in the adjacent column. Interpretations of mutant exons predicted by ASSEDA relative to the published results are also reported. Mutations are given in both HGVS g. and c. format (c. format is mRNA dependent; position 1 is the A of the start codon). <sup>a</sup>Mutation causes conflicting changes to multiple ESE sites. Splicing effect must be determined by experimentation. <sup>b</sup>Multiple SR proteins appear to be involved in the splicing of the exon the relative contributions of each as a result of mutation cannot be differentiated by this analysis.

#### References for mutations in Supp. Table S4

- <sup>1</sup> Miyajima H, Miyaso H, Okumura M, Kurisu J, Imaizumi K. 2002. Identification of a cis-acting element for the regulation of SMN exon 7 splicing. *J Biol Chem.* 277(26):23271-7.
- <sup>2</sup> Heintz C, Dobrowolski SF, Andersen HS, Demirkol M, Blau N, Andresen BS. 2012. Splicing of phenylalanine hydroxylase (PAH) exon 11 is vulnerable: molecular pathology of mutations in PAH exon 11. *Mol Genet Metab.* 106(4):403-11.
- <sup>3</sup> Sun C, Southard C, Di Rienzo A. 2009. Characterization of a novel splicing variant in the RAPTOR gene. *Mutat Res.* 9;662(1-2):88-92.
- <sup>4</sup> Fukao T, Horikawa R, Naiki Y, Tanaka T, Takayanagi M, Yamaguchi S, Kondo N. 2010. A novel mutation (c.951C>T) in an exonic splicing enhancer results in exon 10 skipping in the human mitochondrial acetoacetyl-CoA thiolase gene. *Mol Genet Metab.* 100(4):339-44.
- <sup>5</sup> Gonçalves V, Theisen P, Antunes O, Medeira A, Ramos JS, Jordan P, Isidro G. 2009. A missense mutation in the APC tumor suppressor gene disrupts an ASF/SF2 splicing enhancer motif and causes pathogenic skipping of exon 14. *Mutat Res.* 662(1-2):33-6.
- <sup>6</sup> Burgess R, MacLaren RE, Davidson AE, Urquhart JE, Holder GE, Robson AG, Moore AT, Keefe RO, Black GC, Manson FD. 2009. ADVIRC is caused by distinct mutations in BEST1 that alter pre-mRNA splicing. *J Med Genet.* 46(9):620-5.
- <sup>7</sup> Jensen CJ, Stankovich J, Butzkueven H, Oldfield BJ, Rubio JP. 2010. Common variation in the MOG gene influences transcript splicing in humans. *J Neuroimmunol.* 229(1-2):225-31.
- <sup>8</sup> Tran VK, Takeshima Y, Zhang Z, Yagi M, Nishiyama A, Habara Y, Matsuo M. 2006. Splicing analysis disclosed a determinant single nucleotide for exon skipping caused by a novel intraexonic four-nucleotide deletion in the dystrophin gene. *J Med Genet.* 43(12):924-30.
- <sup>9</sup> Gabut M, Miné M, Marsac C, Brivet M, Tazi J, Soret J. 2005. The SR protein SC35 is responsible for aberrant splicing of the E1alpha pyruvate dehydrogenase mRNA in a case of mental retardation with lactic acidosis. *Mol Cell Biol.* 25(8):3286-94.
- <sup>10</sup> Colapietro P, Gervasini C, Natacci F, Rossi L, Riva P, Larizza L. 2003. NF1 exon 7 skipping and sequence alterations in exonic splice enhancers (ESEs) in a neurofibromatosis 1 patient. *Hum Genet.* 113(6):551-4.
- <sup>11</sup> Raponi M, Kralovicova J, Copson E, Divina P, Eccles D, Johnson P, Baralle D, Vorechovsky I. 2011. Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6. *Hum Mutat.* 32(4):436-44.



**Supplementary Table S5 – Analysis of Normally Spliced Large (> 1000 nt) Exons**

Gene	Exon	Size (nt)	Donor $R_i$ (bits)	Acceptor $R_i$ (bits)	Coordinates of exon	$R_{i,total}$ (bits)	Highest Ranked Splice Form ( $R_{i,total}/size/alternate site$ )
<i>BRCA1</i>	11	3426	2.9	9.4	17:41243451-41246877	1.4	11.4 bits / 118 nt / donor
<i>BRCA2</i>	11	4932	7.7	9.3	13:32910401-32915333	6.1	11.0 bits / 169 nt / donor
<i>TTN<sup>a</sup></i>	253	2967	5.7	10.2	2:179453266-179456230	5.0	14.6 bits / 91 nt / donor
<i>JARID2</i>	7	1039	7.0	8.6	6:15496362-15497401	4.7	11.9 bits / 76 nt / donor
<i>KLHL31</i>	2	1205	5.2	15.0	6:53518898-53520103	9.3	13.8 bits / 224 nt / donor
<i>MLIP</i>	4	1572	5.2	11.4	6: 54001512-54003084	5.7	12.5 bits / 155 nt / donor
<i>C17orf53</i>	3	1170	11.0	8.4	17:42225225-42226395	8.5	12.1 bits / 239 nt / acceptor
<i>VCAN</i>	8	5262	9.9	9.3	5:82832825-82838087	8.3	12.0 bits / 1750 nt / acceptor

Large exons (> 1000 nt) were analyzed using ASSEDA. All were found to have positive  $R_{i,total}$  values due to moderate to strong natural site strengths. The right-most column lists the highest ranked prospective isoform predicted by ASSEDA, which are much smaller (< 250 nt) and thus have a lower gap surprisal penalty. As each of these large exon sizes only occur in one exon in the transcriptome, each splice form have the same maximum gap surprisal penalty of 10.9 bits. <sup>a</sup>Representative exon (1 of 5 possible).

## Supplementary Methods 1 – Gap Surprisal Description

Gap Surprisal is the penalty given as per length of the exon. To correctly define the gap surprisal for a combination of splice sites, a table was constructed which relates the gap surprisal to the length of the exon. The whole genome was scanned and the frequencies of different lengths of exons occurring in the genome and their respective probability of occurrence were calculated.

According to Tribus (1961), the amount of self-information contained in a probabilistic event depends only on the probability of that event: the smaller its probability, the larger the self-information associated with receiving the information that the event indeed occurred. The self-information or surprisal  $I(\omega_n)$  associated with outcome  $\omega_n$  with probability  $P(\omega_n)$  is:

$$I(\omega_n) = \log(1/P(\omega_n)) = -\log(P(\omega_n))$$

Here, the base of the logarithm is not specified: if using base 2, the unit of  $I(\omega_n)$  is in bits. The above definition is used to deduce gap surprisal function. The self-information or gap surprisal,  $g(L_n)$ , of observing a pair donor and acceptor site separated by  $L$  nucleotides is  $-\log_2(P(L_n))$  bits. The self-information or gap surprisal,  $g(L_n)$ , of observing a pair donor and acceptor site separated by  $L$  nucleotides is  $-\log_2(P(L_n))$  bits. The gap surprisal is defined as follows

$$\text{Gap Surprisal} = \text{Log}_2 (1/\text{probability of occurrence the exon length}).$$

This function signifies that the greater the distance between the donor and acceptor sites, the larger the gap surprisal (greater penalty) will be, resulting in a biological reduction of larger than consensus exon length occurrence. The gap Surprisal values for different exon lengths were calculated using the above formula. The gap surprisal value for exon lengths that were not Supplementary Figure 3 shows the distribution of gap surprisal for internal exons.

The most frequent length was assigned a gap surprisal of zero, based on the fact that splice sites separated by this distance have a highest likelihood of forming an exon. This length was 96 nucleotides [1901 occurrences among total 172250 occurrences]. The frequency for this particular length 96 was:  $1901/172250 = 0.011036$ . The gap surprisal for the most common, ie. preferred, constitutive exon length is 6.59 bits. To normalize all other gap surprisal terms for all other exon lengths to this value and eliminate the gap surprisal penalty for exons of 96 nucleotides, all of the penalties for all exon lengths were corrected by subtracting 6.59 bits from their respective gap surprisal values.

Total information content of either the acceptor or donor or both was found to be less than zero bits (most of these represent initial and terminal exons, as expected, since these do not contain both donor and acceptor splice sites). To successfully recognize the initial and terminal exons, a separate exon definition distribution was defined for these.

### **Gap Surprisals of First and Last Exons**

As the exon definition hypothesis cannot be applied for first exon since no acceptor site is defined; and for last exon since no donor site is defined, different gap surprisals were defined for selection of these exons. Separate gap surprisal tables were constructed for these exons by scanning refseq and identifying the frequencies of different lengths of first and last exons. It was observed that most frequent length of the first exon was 46 nucleotides and that of last exon was 24 nucleotides. Hence the minimum gap surprisal (0 bits) was assigned to length of 158 for the first exon and a length of 232 for the last exon.