IntSplice: Prediction of the splicing consequences of intronic single nucleotide variations in the human genome

Akihide Shibata, Tatsuya Okuno, Mohammad Alinoor Rahman, Yoshiteru Azuma, Jun-ichi Takeda, Akio Masuda, Kinji Ohno*

Division of Neurogenetics, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, Nagoya 466-8550, Japan

*Address correspondence to:
Kinji Ohno. Division of Neurogenetics, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, 65 Tsurumai, Showa-ku, Nagoya 466-8550, Japan.
E-mail: ohnok@med.nagoya-u.ac.jp

Running Title: IntSplice: Splicing prediction of intronic SNVs

Keywords: Splice acceptor site, aberrant splicing, single nucleotide variations, intronic mutations.

Abbreviations: SNV, single nucleotide variation; Int-SNV, intronic SNV at intronic positions from −50 to −3; Int-α, intronic position -α; Ex+α, exonic position +α; PPT, polypyrimidine tract; BPS, branch point sequence; ss, splice site; *TR*, transcription ratio.

ABSTRACT

Precise spatiotemporal regulation of splicing is mediated by splicing *cis*-elements on pre-mRNA. Single nucleotide variations (SNVs) affecting intronic *cis*-elements possibly compromise splicing, but no efficient tool has been available to identify them. Following an effect-size analysis of each intronic nucleotide on annotated alternative splicing, we extracted 105 parameters that could affect the strength of the splicing signals. However, we could not generate reliable support vector regression models to predict the percent-splice-in (PSI) scores for normal human tissues. Next, we generated support vector machine (SVM) models using 110 parameters to directly differentiate pathogenic SNVs in the human gene mutation database (HGMD) and normal SNVs in the dbSNP database, and we obtained models with a sensitivity of 0.800 ± 0.041 (mean and SD) and a specificity of 0.849 ± 0.021. Our IntSplice models were more discriminating than SVM models that we generated with Shapiro-Senapathy score and MaxEntScan::score3ss. We applied IntSplice to a naturally occurring and nine artificial intronic mutations in *RAPSN* causing congenital myasthenic syndrome. IntSplice correctly predicted the splicing consequences for nine of the ten mutants. We created a web service program, IntSplice (http://www.med.nagoya-u.ac.jp/neurogenetics/IntSplice) to predict splicing-affecting SNVs at intronic positions from -50 to -3.

INTRODUCTION

Higher eukaryotes have evolved by acquiring tissue-specific and developmental stage-specific regulation of alternative splicing of pre-mRNA rather than by acquiring novel genes.[1] Precisely regulated splicing process takes place in the spliceosome, which comprises five small nuclear ribonucleoproteins (U1, U2, U4, U5, and U6 snRNPs) and a large number of non-snRNP proteins.[2] In the first step of the assembly of the spliceosome, U1 snRNP, SF1, U2AF65, and U2AF35 bind to the splicing *cis*-elements at the 5’ splice site (ss), the branch point sequence (BPS), the polypyrimidine tract (PPT), and the 3’ ss, respectively.[3, 4] Single nucleotide variations (SNVs) disrupting these essential *cis*-elements lead to aberrant splicing and cause human diseases. At least 10% of inherited human diseases are caused by mutations affecting the essential splicing *cis*-elements at the 5’ and 3’ ss’s.[5] In addition, intronic and exonic splicing *cis*-elements also confer precise spatiotemporal regulation of constitutive and alternative splicing, which are also frequently disrupted in human diseases.[6] Development of high-throughput sequencing technologies has enabled us to obtain a large number of SNVs from a significant number of individuals. Prediction of the splicing consequences of intronic SNVs, however, remains difficult due to the lack of efficient prediction tools.

Exonic SNVs often disrupt or *de novo* generate exonic splicing enhancers (ESEs) and silencers (ESSs). Several ESE/ESS search tools are available online: ESE finder 3.0,[7] ESRsearch,[8] FAS-ESS,[9] PESXs,[10, 11] RESCUE-ESE,[12] Human Splicing Finder,[13] SpliceAid,[14] SpliceAid2,[15] CRYP-SKIP,[16] Spliceman,[17] and RegRNA 2.0.[18] These tools can be used to predict splicing consequences of exonic SNVs. In contrast to a variety of available tools for inspecting exonic SNVs, only two tools are available to our knowledge to score the 3’ ss. The Shapiro-Senapathy score is calculated using the position-specific scoring matrix (PSSM), representing the frequency of each nucleotide from intronic position −14 (Int-14) to exonic position +1 (Ex+1),[19] which has long been used to predict the splicing effects of SNVs. The MaxEntScan::score3ss scores the 3’ ss from Int-20 to Ex+3.[20] Shapiro-Senapathy score and MaxEntScan, however, were not specifically designed to predict the splicing consequences of intronic SNVs.

We have previously reported that the consensus sequence of human BPS is yUnAy, where “y” represents pyrimidines and “n” represents any nucleotides.[21] Similarly, extensive analyses of human branch points using RNA-seq show that the consensus BPS sequence is “UnAy”.[22-24] The highly degenerative BPS motif, however, prevented us from developing a model to predict the position and the splicing effect of BPS. We also reported that a mutation at the first nucleotide of an exon causes aberrant splicing at the AG-dependent, 3’ ss, where a short

PPT cannot confer sufficient binding affinity for U2AF65 and additional binding of U2AF35 to the 3' ss is required.[25] Here, we present a support vector machine (SVM) model, IntSplice, to predict aberrant splicing due to intronic SNVs (Int-SNVs) at positions from Int-50 to Int-3 (Int-50:Int-3).

MATERIALS AND METHODS

Ethics statement

Studies on a patient with congenital myasthenic syndrome were approved by the ethical review committees of the Mayo Clinic and the Nagoya University Graduate School of Medicine. The studies were performed after an appropriate informed written consent was obtained.

Databases

Sequence motifs of the splicing *trans*-factors were obtained from the SpliceAid database.[14] Exonic and intronic positions of these sequence motifs were not taken into account, because (i) Int-SNVs should not change any exonic motifs; (ii) we did not look into exonic nucleotides when we made our models; and (iii) exonic and intronic positions were not always available in the SpliceAid database. RNA-seq data on the brain, cerebral cortex, heart, liver, skeletal muscle, and lungs in normal humans were obtained from the GEO database (the accession number, GSE13652) in an SRA format.[26] The number of individuals and their demographic features for each tissue were not available for GSE13652.[26] RNA-seq data on the breasts, lymph nodes, testes, adipose tissue, colon, skeletal muscle, liver, and brain in normal humans were similarly obtained with the GEO accession number GSE12946 in an SRA format.[27] For GSE12946, each tissue sample was obtained from a single unrelated individual.[27] The SRA files were converted to fastq files using an SRA toolkit (http://eutils.ncbi.nih.gov/Traces/sra/sra.cgi?view=software). Disease-causing mutations located at positions Int-50:Int-3 were obtained from the Human Gene Mutation Database (HGMD) Professional (Biobase). Some intronic mutations in the HGMD might not be splicing mutations and might affect a transcription enhancer/silencer, a pre-miRNA sequence, or a yet uncharacterized *cis*-element, but the functional consequences of intronic mutations were not always deeply dissected in original papers. We therefore included all intronic mutations at positions Int-50:Int-3, without filtering out non-splicing mutations. Normal SNVs were obtained from dbSNP134. SNVs included in the HGMD were excluded from our analysis. We also excluded SNVs with a global minor allelic frequency (GMAF) of less than 0.01.

Support vector regression (SVR) and support vector machine (SVM) modeling

The RNA-seq data were mapped to the human genome GRCh37/hg19 with ENSEMBL release 64 using TopHat mapper with its default parameters.[28] Splicing efficiency of each individual exon (percent-spliced-in score, PSI) was calculated with the MISO software.[29] For each RNA-seq dataset of the 14 human tissues, we randomly divided 3’ ss’s into five groups. Four groups were arbitrarily chosen to generate an SVR model with the nu-SVR functionality of LIBSVM version 3.17[30] to predict PSIs using 105 parameters. We then tested the validity of the generated SVR model using the remaining fifth group. We made five SVR models by changing the training and validation groups. We generated 100 different combinations of five groups for each RNA-seq dataset and ran the SVR modeling 500 times.

SVM models to distinguish between pathogenic and normal Int-SNVs were generated with 110 parameters using the C-SVC functionality of LIBSVM.[30] A total of 500 different SVM models were generated for 100 different datasets of 1,162 pathogenic and 1,162 normal Int-SNVs. Normal Int-SNVs in each dataset were randomly selected from 16,741 normal SNVs. For SVM modeling, we compared four kernels of “linear”, “polynomial”, “radial basis function”, and “sigmoid”.

For both the SVR and the SVM models, scores of each parameter were normalized using the SVM-scale functionality of LIBSVM,[30] so that each parameter was equally weighted. Perl scripts were run on the RPIMERGY CX400 UNIX server (Fujitsu).

A patient with congenital myasthenic syndrome

The patient, now 29 years old, was hypomotile in utero. After birth, he was floppy, had a poor cry, needed ventilatory support, and had arm and leg contractures. He improved gradually and walked at the age of 14 months. He showed a decremental EMG response in several muscles. His weakness was improved with a cholinesterase inhibitor, pyridostigmine. Sanger sequencing of genomic DNA revealed a homozygous T-to-A substitution at intron 5 (c.913-5T>A) of the *RAPSN* gene. No muscle specimen was available from the patient.

Minigene constructs

To construct the human *RAPSN* minigene, we amplified a genomic segment spanning exons 5 to 7 of *RAPSN* by PCR with KOD Plus DNA polymerase (Toyobo) using genomic DNA isolated from HeLa cells. The 5’ ends of the forward and reverse primers carried the BamHI and

XhoI sites, respectively. The amplified fragment was cloned into the BamHI and XhoI sites of the pcDNA3.1(+) vector (Invitrogen) to generate the pcDNA-*RAPSN* minigene. The naturally occurring (patient) and artificial mutations were engineered into the pcDNA-*RAPSN* construct using the QuikChange Site-Directed Mutagenesis Kit (Stratagene). The presence of artifacts was excluded by sequencing the entire inserts.

Cell culture, transfection, and RT-PCR for splicing analysis

HeLa cells were cultured in DMEM (Sigma-Aldrich) with 10% fetal bovine serum (FBS, Sigma-Aldrich). The cells were plated 24 hrs before transfection in six-well culture plates (1.5×10^5 cells/well), and transfected using the FuGENE 6 transfection reagent (Roche) according to the manufacturer’s instructions. Total RNA was extracted 40 hrs following transfection using the TRIzol reagent (Invitrogen), followed by DNase I treatment. The cDNA was synthesized with an oligo-dT primer using the ReverTra Ace reverse transcriptase (Toyobo). PCR-amplification was performed using the GoTaq DNA polymerase (Promega) with the following primer pair: 5’-ATCATGACCGAGATCGGAAAC-3’ on exon 5 and 5’-GTGGAACCTCACAACGTGC-3’ on exon 7.

MS2-affinity purification of a spliceosomal complex

To synthesize an RNA substrate for the MS2-affinity purification of a spliceosomal complex, we first amplified a genomic segment spanning *RAPSN* exons 5 and 6 from wild-type and mutant pcDNA-*RAPSN* minigenes, and then cloned them into the BamHI and XhoI sites of pcDNA3.1(+) to generate the pcDNA-*RAPSN*-E5-E6 minigenes. A segment spanning three copies of the MS2-binding sites was PCR-amplified from pSP64-MS2 that we previously reported,[31] and was introduced downstream of exon 6 of the pcDNA-*RAPSN*-E5-E6 minigenes using the megaprimer method.[32] The generated pcDNA-*RAPSN*-E5-E6-MS2 minigenes were used as templates to synthesize RNA-substrates using the RiboMAX System (Promega).

An RNA probe (1 pmol) was incubated with 20-fold molar excess of the MS2-MBP fusion protein.[33] Fifty microliters of HeLa nuclear extract (CilBiotech) was preincubated with 10 µl (bead volume) of amylose resin (New England Biolabs) overnight at 4°C. The purified HeLa nuclear extract was then incubated at 37°C for 30 min, with a mixture of the RNA probe and the MS2-MBP fusion protein at final concentrations of 60 mM KCl and 25% HeLa nuclear extract. Ten microliters (bead volume) of amylose resin was added and mixed on a rotary shaker at 4°C for 30 min. After washing four times with washing buffer (20 mM HEPES at pH 8.0, 150 mM KCl, and 0.05% Triton X-100), the resin-bound molecules were eluted with 10 mM maltose solution. The purified proteins were subjected to SDS-PAGE and immunoblot analyses to detect the binding of U2AF65, U2AF35, and U1 snRNP (U1-70K), respectively. The antibodies used were U2AF65 (MC3, sc-53942, Santa Cruz Biotechnology), U2AF35 (N-16, sc-19961, Santa Cruz Biotechnology), and U1-70K (H111, kindly provided by Dr. Akila Mayeda at the Fujita Health University).

RESULTS

Estimation of the effects of individual intronic nucleotides on splicing annotated in the ENSEMBL release 64 database

A diagram showing the flow of our analyses in this communication is shown in Supplementary Figure 1. We first inspected the alternative splicing events annotated in the ENSEMBL release 64 on the GRCh37/hg19 human genome. We restricted our analysis to introns with “AG” dinucleotides at the 3’ end, and not “AC”. We estimated the splicing efficiency of the 3’ ss by defining a new parameter, the transcription ratio (*TR*). When a gene gives rise to *m* different transcripts at a specific position, and *n* transcripts are spliced at the 3’ ss according to ENSEMBL release 64, we defined *TR* for that specific 3’ ss as *n*/*m*. An example of *TRs* is shown in Supplementary Figure 2. Assuming that the 3’ ss with a high *TR* carries a strong splicing signal, we plotted the average *TR* against individual nucleotides from positions Int-50 to Ex+5 at the 3’ ss. The plot revealed that nucleotides at positions Int-13:Int-5, Int-3, Ex+1, and Ex+2 were critical determinants of *TR* (Figure 1).

Prediction of PSIs of 14 tissue-specific RNA-seq data using SVR modeling

The PSIs of individual 3’ ss’s in the RNA-seq data of 14 normal human tissues in GSE13652[26] and GSE12946[27] were calculated with MISO.[29] We first tried to predict a PSI using the primary nucleotide sequence with a linear regression model and with an SVR model. If we can efficiently predict the PSI of a given 3’ ss, we should be able to make a model to identify an intronic splicing mutation. The primary nucleotide sequence alone at positions Int-50:Ex+5, however, was not sufficient to predict the PSI of a given 3’ ss (data not shown). We then extracted 105 parameters that possibly dictate the strength of the splicing signals (Supplementary Table 1). The 105 parameters included individual nucleotides at positions Int-3 and Ex+1 according to Figure 1, the sequence motifs of all the splicing *trans*-factors in the SpliceAid database,[14] the position weight matrix of the human branch point sequence[21] (Supplementary Table 2), variable definitions of PPT, ∆G of a predicted secondary RNA structure based on the mfold program,[34] *etc*. We included ∆G of mfold, because the secondary

RNA structure is a critical determinant of the splicing consequences.[35-37] The RNA-seq data of the 14 tissues, however, generated SVR models with correlation coefficients (*R*) ranging from 0.239 to 0.274 (mean and SD, 0.253 ± 0.011) (Supplementary Figure 3). These SVR models thus failed to predict the PSIs with enough accuracy to estimate the splicing strength of a given 3' ss, and their sole application could not predict the splicing consequence of a given Int-SNV.

Differentiation of pathogenic and normal Int-SNVs using SVM modeling

Next, we tried to differentiate pathogenic SNVs registered in the HGMD and normal SNVs in the dbSNP database at positions Int-50:Int-3. In addition to 14 PSIs calculated with the 14 SVR models stated above, we used the 105 parameters once again to make a prediction model. Among the 119 parameters, however, we excluded 10 parameters that represented the nucleotides at positions Int-2:Ex+5 and at the 5' ss, where no SNV should exist in the current analysis. We also added a parameter indicating whether an "AG" dinucleotide is generated *de novo* by Int-SNVs. We thus used a total of 110 parameters (Supplementary Table 1) to make SVM models. The HGMD included 1,162 pathogenic SNVs at positions Int-50:Int-3, whereas the dbSNP database included 16,741 normal SNVs at positions Int-50:Int-3 with a global minor allelic frequency of greater than 0.01. To match the numbers of SNVs in HGMD and the dbSNP database, we randomly chose 1,162 SNVs from the 16,741 SNVs in the dbSNP database. A dataset of 2,324 pathogenic and normal SNVs was divided into five groups. The datasets of 2,324 SNVs were generated 100 times in order to validate the models repeatedly. For each dataset, four groups were employed to generate an SVM model (IntSplice) with LIBSVM[30] using the 110 parameters to predict whether an SNV belongs to the HGMD or the dbSNP database. We then tested the validity of the SVM model generated using the remaining fifth group, and calculated the sensitivity and the specificity of each model. A total of 500 different SVM models were generated with four kernels of "linear", "polynomial", "radial basis function", and "sigmoid", respectively (Table 1). The sensitivity ranged from 0.710 to 0.769, and the specificity ranged from 0.896 to 0.936. Among the four kernels, the radial basis function generated the most efficient SVM models.

The three best parameters in SVM modeling were the MaxEnt score at Int-20:Int-3 (coefficient = -12.7), the Shapiro-Senapathy score at Int-50:Int-3 (coefficient = -11.2), and the ratio of A/G's at Int-20:Int-8 (coefficient = 10.8) (Supplementary Table 1). As the MaxEnt score and the Shapiro-Senapathy score are comprehensive parameters to dictate the strength of splicing signals, these two parameters were better than individual parameters. Among the individual parameters, the coefficient of the ratio of A/G's at Int-20:Int-8 to SVM modeling was as high as those of the comprehensive parameters. A similar and partly overlapping parameter was the number of G's at Int-12:Int-3 (coefficient = 7.83). Contribution of these individual parameters in SVM modeling suggests that the presence of purines at PPT has a marked negative effect on splicing.

Inclusion of SVR-based prediction of PSI may bias the SVM models in favor of 14 tissues that were included in the SVR modeling. We thus made 500 SVM models without SVR-based prediction of PSI derived from 14 RNA-seq data, and compared the sensitivity and specificity to those with SVR-based prediction of PSI. We found that the sensitivities and specificities of the two SVM models with and without PSI parameters were essentially the same (Supplementary Table 3). As the sum of specificity and sensitivity at Int-50:Int-3 became marginally low by exclusion of PSI parameters, we included PSI parameters in the following analyses.

We also made SVM models with 1,162 pathogenic Int-SNVs and 16,741 normal Int-SNVs at positions Int-50:Int-3 using the radial basis function (unmatched models). We generated 500 different datasets by randomly selecting four-fifth of pathogenic/normal Int-SNVs as a training dataset and the remaining one-fifth of pathogenic/normal Int-SNVs as a validation dataset. SVM models with unmatched datasets had a sensitivity of 0.762 ± 0.030 (mean and SD) and a specificity of 0.905 ± 0.024 (mean and SD). As shown in Table 1, SVM models with matched datasets had a sensitivity of 0.899 ± 0.022 (mean and SD) and a specificity of 0.772 ± 0.027 (mean and SD). Although the sums of sensitivity and specificity were similar between the two datasets (1.671 for unmatched datasets and 1.667 for matched datasets), sensitivity was higher with the matched datasets and specificity was higher with the unmatched datasets. With unmatched datasets, the number of normal SNVs was 14 times (=16,741/1,162) higher than that of pathogenic SNVs. SVM models with unmatched datasets were thus in favor of predicting that Int-SNVs were negative, and specificity became high (0.905) at the cost of low sensitivity (0.762). We supposed that both unmatched and matched models could be used for different purposes. However, in order to detect pathogenic Int-SNVs identified in human diseases, we hoped to keep the sensitivity high as much as possible, and we used SVM models with matched datasets in the following analyses.

Comparison of IntSplice with SVM models generated based on the Shapiro-Senapathy score and MaxEntScan::score3ss

Although Shapiro-Senapathy score[19] and MaxEntScan::score3ss[20] are not designed to predict aberrant splicing due to Int-SNVs, we exploited these scores to predict the splicing consequences of Int-SNVs by setting an automatic cutoff value with SVM. For each of the 100 datasets comprising the 2,324 Int-SNVs at positions Int-50:Int-3 that we used for the IntSplice modeling, we analyzed all the 2,324 Int-SNVs at positions Int-50:Int-3 with Shapiro-Senapathy score and 2,064 Int-SNVs at positions Int-20:Int-3 with MaxEntScan. Shapiro-Senapathy score was originally designed to score the 3' ss up to Int-14, and the scoring matrix was based on the nucleotide sequences available in the year 1987.[19] We thus made a new scoring matrix covering up to Int-50 by analyzing ENSEMBL release 64 (Supplementary Table 2). MaxEntScan was designed to score 3' ss up to Int-20, and was unable to score Int-SNVs at positions Int-50:Int-21.[20] We randomly divided the datasets comprised of 2,324 and 2,064 Int-SNVs into five groups. We made 500 SVM models using either the Shapiro-Senapathy score or the MaxEntScan with the four kernels of “linear”, “polynomial”, “radial basis function”, and “sigmoid” (Table 1), as we did with IntSplice. Again, the radial basis function generated the most efficient models with both the Shapiro-Senapathy score and MaxEntScan. The plots of the sensitivity and the specificity of the radial basis function models generated by IntSplice, Shapiro-Senapathy score, and MaxEntScan, respectively, revealed that the sum of the sensitivity and the specificity of IntSplice was higher than those of the Shapiro-Senapathy score and MaxEntScan for the Int-SNVs at positions Int-50:Int-3 (Figure 2A, Table 1), Int-20:Int-3 (Figure 2B, Table 1), and Int-50:Int-21 (Table 1).

IntSplice: a web service program to predict the pathogenic and normal Int-SNVs using SVM modeling

The aforementioned analyses of the validation datasets indicate that SVM modeling with the radial basis function was able to distinguish between pathogenic and normal Int-SNVs with a sensitivity of 0.772 ± 0.027 (mean and SD) and a specificity of 0.101 ± 0.022 (Table 1). Thus, we generated a global SVM model by including 2,324 SNVs and made a web service program, IntSplice, at http://www.med.nagoya-u.ac.jp/neurogenetics/IntSplice. This program accepts a file in a VCF format with multiple SNVs and predicts whether each SNV affects splicing or not. A given SNV is mapped to all the annotated coding transcripts in ENSEMBL 64, and the program analyzes all the transcripts. If an SNV affects splicing of one or more transcript(s), our program predicts that the SNV is pathogenic and shows the affected ENST transcript numbers. Representative results are shown in Figure 3.

Application of the IntSplice program to intronic mutations of *RAPSN*

We applied our IntSplice program to the naturally occurring and artificial mutations in

RAPSN encoding rapsyn, which makes a scaffold for the muscle nicotinic acetylcholine receptor at the neuromuscular junction.[38, 39] A homozygous *RAPSN* c.913-5T>A mutation was identified in a patient with congenital myasthenic syndrome. Introduction of a minigene spanning *RAPSN* exons 5 to 7 into HeLa cells showed that the *RAPSN* c.913-5T>A mutation caused partial skipping of exon 5 (Figure 4A), and compromised binding of U2AF65 (Figure 4B). In order to investigate which pyrimidine nucleotide in the PPT is essential for splicing of *RAPSN* exon 5, we first substituted “T” for “A” at position Int-9 to make a complete stretch of ten pyrimidines at positions Int-12:Int-3 (“Opt” in Figure 4C). We then serially introduced a mutant “A” from positions Int-11 to Int-3. Introduction of the nine artificial mutants into HeLa cells showed that three mutants at positions Int-6, Int-5, and Int-3 led to skipping of exon 5 (Figure 4C). We also found that the binding of U2AF65 to the Int-6 and Int-5 mutants was compromised, but not that to the Int-3 mutant (Figure 4D). In contrast, the binding of U2AF35 or U1-70K was not affected in any mutant.

The IntSplice, the MaxEntScan-based model, and the Shapiro-Senapathy score-based model correctly predicted aberrant splicing in the patient’s mutation, *RAPSN* c.913-5T>A (Figure 4A). Next, we made the “Opt” construct as a normal reference sequence, and applied these three models to the nine artificial mutants (Figure 4C). The IntSplice, the MaxEntScan-based model, and the Shapiro-Senapathy score-based model erroneously predicted the splicing consequences in one, two, and five mutants, respectively (asterisks in Figure 4C).

DISCUSSION

In an effort to make a model to predict splicing consequences of Int-SNVs, we first analyzed the position-specific effects of the intronic nucleotides on splicing (Figure 1), and extracted parameters that possibly affect the splicing strength (Supplementary Table 1). We calculated the PSIs of 14 RNA-seq data of normal human tissues, and then tried to predict PSIs using SVR models with the 105 extracted parameters. However, the correlation coefficients between the calculated and predicted PSIs were less than 0.3 (Supplementary Figure 3). Next, we generated SVM models to directly differentiate pathogenic SNVs in the HGMD and normal SNVs in the dbSNP database, with 1-specificity (a false positive rate) of ~0.10 and a sensitivity (a true positive rate) of ~0.77, and named it IntSplice. Inefficient prediction with the RNA-seq-based SVR models suggests that prediction of PSI scores is much more difficult than differentiation between normal and pathogenic Int-SNVs. Although SVM models to differentiate normal and pathogenic Int-SNVs with SVM were better than SVR models to predict PSIs, accurate prediction of splicing consequences of Int-SNVs was not available even with SVM modeling. This was likely due to inadequacy of the training datasets and also to lack of parameters that were essential for splicing regulation in living cells. First, our training dataset was comprised of pathogenic Int-SNVs causing Mendelian disorders (HGMD) and normal Int-SNVs in dbSNP with a minor allelic frequency > 0.01. Neither HGMD nor dbSNP database was comprehensive, and the effects of Int-SNVs that were not present in HGMD or dbSNP could not be estimated. Second, among various parameters that enable precise spatiotemporal regulation of splicing *in vivo,* the following parameters could not be taken into account in our SVM modeling. i) Splicing is coupled to transcription, which is regulated by RNA polymerase II, other transcription factors, and chromatin structure.[40] ii) Splicing *cis*-elements that are functional in specific tissue(s) at specific developmental stage(s) have not been fully characterized.[41] iii) The exact mechanisms underlying recognition of degenerative *cis*-elements by a specific RNA-biding protein remain to be elucidated.[42] iv) RNA editing plays a pivotal role in spicing, but RNA editing has not been comprehensively characterized.[43] v) Spatiotemporal regulations of expression and activation of splicing *trans*-factors (RNA-binding proteins) have not been extensively identified.[41]

We compared the prediction efficiency of IntSplice with those of Shapiro-Senapathy score- and MaxEntScan-based SVM models that we generated by applying the same training and validation datasets that were used for IntSplice. Although the sensitivity as well as the sum of the sensitivity and the specificity of IntSplice were better than those of Shapiro-Senapathy score-based and MaxEntScan-based models, the specificity of IntSplice was not as good as that of MaxEntScan-based model for Int-SNVs at positions Int-20:Int-3 (Table 1). A high specificity of MaxEntScan was indeed observed in *RAPSN* mutants. In contrast to IntSplice and Shapiro-Senapathy score-baesd model, MaxEntScan-based model erroneously predicted that the Int-6 and Int-5 mutants were normally spliced, although these caused exon skipping (Figure 4C). MaxEntScan-based model may make the specificity high at the cost of lowering the sensitivity. As we have incorporated both the Shapiro-Senapathy score and MaxEntScan scores in our 110 parameters, we expected that the sensitivity and the specificity of IntSplice were superior to those of Shapiro-Senapathy score-based and MaxEntScan-based models. The better specificity of MaxEntScan compared to IntSplice is possibly accounted for due to the difference in the positions of the Int-SNVs used to produce their respective models: IntSplice was trained with Int-SNVs up to position Int-50, whereas MaxEntScan covered up to position Int-20. Alternatively, the MaxEntScan scores were underestimated among the 110 parameters in the SVM modeling of IntSplice for the sake of an improved sensitivity. Another possibility is that the higher specificity of MaxEntScan was lowered by the lower specificities of the other parameters used in the IntSplice modeling. We hope that our web service program, IntSplice, will reveal yet unidentified splicing mutations at positions Int-50:Int-3, and unveil aberrant splicing in human diseases.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

We are grateful to Dr. Duygu Selcen and Dr. Andrew G. Engel for providing the unpublished c.913-5T>A mutation in the *RAPSN* gene in a patient with congenital myasthenic syndrome. The studies were supported by Grants-in-Aid from the MEXT and MHLW of Japan.

References

1. Black, D.L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72**, 291-336 (2003).
2. Jurica, M.S. & Moore, M.J. Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell.* **12**, 5-14 (2003).
3. Reed, R. Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr. Opin. Genet. Dev.* **6**, 215-220 (1996).
4. Gooding, C., Edge, C., Lorenz, M., Coelho, M.B., Winters, M., Kaminski, C.F. et al. MBNL1 and PTB cooperate to repress splicing of Tpm1 exon 3. *Nucleic Acids Res.* **41**, 4765-4782 (2013).
5. Krawczak, M., Thomas, N.S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J. et al. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum. Mutat.* **28**, 150-158 (2007).
6. Sterne-Weiler, T., Howard, J., Mort, M., Cooper, D.N. & Sanford, J.R. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.* **21**, 1563-1571 (2011).
7. Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q. & Krainer, A.R. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* **31**, 3568-3571 (2003).
8. Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I. et al. Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. *Mol. Cell.* **22**, 769-781 (2006).
9. Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M. & Burge, C.B. Systematic identification and analysis of exonic splicing silencers. *Cell.* **119**, 831-845 (2004).
10. Zhang, Z. & Krainer, A.R. Involvement of SR proteins in mRNA surveillance. *Mol. Cell.* **16**, 597-607 (2004).
11. Zhang, X.H., Kangsamaksin, T., Chao, M.S., Banerjee, J.K. & Chasin, L.A. Exon inclusion is dependent on predictable exonic splicing enhancers. *Mol. Cell. Biol.* **25**, 7323-7332 (2005).
12. Fairbrother, W.G., Yeh, R.F., Sharp, P.A. & Burge, C.B. Predictive identification of exonic splicing enhancers in human genes. *Science.* **297**, 1007-1013 (2002).
13. Desmet, F.O., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M. & Beroud, C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals.

Nucleic Acids Res. **37**, e67 (2009).

14. Piva, F., Giulietti, M., Nocchi, L. & Principato, G. SpliceAid: a database of experimental RNA target motifs bound by splicing proteins in humans. *Bioinformatics.* **25**, 1211-1213 (2009).
15. Piva, F., Giulietti, M., Burini, A.B. & Principato, G. SpliceAid 2: a database of human splicing factors expression data and RNA target motifs. *Hum. Mutat.* **33**, 81-85 (2012).
16. Divina, P., Kvitkovicova, A., Buratti, E. & Vorechovsky, I. Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping. *Eur. J. Hum. Genet.* **17**, 759-765 (2009).
17. Lim, K.H., Ferraris, L., Filloux, M.E., Raphael, B.J. & Fairbrother, W.G. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 11093-11098 (2011).
18. Chang, T.H., Huang, H.Y., Hsu, J.B., Weng, S.L., Horng, J.T. & Huang, H.D. An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. *BMC Bioinformatics.* **14 Suppl 2**, S4 (2013).
19. Shapiro, M.B. & Senapathy, P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* **15**, 7155-7174 (1987).
20. Yeo, G. & Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377-394 (2004).
21. Gao, K., Masuda, A., Matsuura, T. & Ohno, K. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res.* **36**, 2257-2267 (2008).
22. Corvelo, A., Hallegger, M., Smith, C.W. & Eyras, E. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol.* **6**, e1001016 (2010).
23. Taggart, A.J., DeSimone, A.M., Shih, J.S., Filloux, M.E. & Fairbrother, W.G. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat. Struct. Mol. Biol.* **19**, 719-721 (2012).
24. Bitton, D.A., Rallis, C., Jeffares, D.C., Smith, G.C., Chen, Y.Y., Codlin, S. et al. LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. *Genome Res.* **24**, 1169-1179 (2014).
25. Fu, Y., Masuda, A., Ito, M., Shinmi, J. & Ohno, K. AG-dependent 3'-splice sites are predisposed to aberrant splicing due to a mutation at the first nucleotide of an exon. *Nucleic Acids Res.* **39**, 4396-4404 (2011).

26. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature.* **456**, 470-476 (2008).

27. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413-1415 (2008).

28. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* **25**, 1105-1111 (2009).

29. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods.* **7**, 1009-1015 (2010).

30. Chang, C.C. & Lin, C.J. LIBSVM: A Library for Support Vector Machines. *ACM T Intel Syst Tec.* **2**, Article 27 (2011).

31. Rahman, M.A., Masuda, A., Ohe, K., Ito, M., Hutchinson, D.O., Mayeda, A. et al. HnRNP L and hnRNP LL antagonistically modulate PTB-mediated splicing suppression of CHRNA1 pre-mRNA. *Sci Rep.* **3**, 2931 (2013).

32. Ohno, K., Anlar, B., Ozdirim, E., Brengman, J.M., DeBleecker, J.L. & Engel, A.G. Myasthenic syndromes in Turkish kinships due to mutations in the acetylcholine receptor. *Ann. Neurol.* **44**, 234-241 (1998).

33. Das, R., Zhou, Z. & Reed, R. Functional association of U2 snRNP with the ATP-independent spliceosomal complex E. *Mol. Cell.* **5**, 779-787 (2000).

34. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406-3415 (2003).

35. Gahura, O., Hammann, C., Valentova, A., Puta, F. & Folk, P. Secondary structure is required for 3' splice site recognition in yeast. *Nucleic Acids Res.* **39**, 9759-9767 (2011).

36. Plass, M., Codony-Servat, C., Ferreira, P.G., Vilardell, J. & Eyras, E. RNA secondary structure mediates alternative 3'ss selection in Saccharomyces cerevisiae. *RNA.* **18**, 1103-1115 (2012).

37. Pervouchine, D.D., Khrameeva, E.E., Pichugina, M.Y., Nikolaienko, O.V., Gelfand, M.S., Rubtsov, P.M. et al. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. *RNA.* **18**, 1-15 (2012).

38. Ohno, K., Engel, A.G., Shen, X.M., Selcen, D., Brengman, J., Harper, C.M. et al. Rapsyn mutations in humans cause endplate acetylcholine-receptor deficiency and myasthenic syndrome. *Am J Hum Genet.* **70**, 875-885 (2002).

39. Milone, M., Shen, X.M., Selcen, D., Ohno, K., Brengman, J., Iannaccone, S.T. et al. Myasthenic syndrome due to defects in rapsyn: Clinical and molecular findings in 39 patients. *Neurology.* **73**, 228-235 (2009).

40. Kornblihtt, A.R., Schor, I.E., Allo, M., Dujardin, G., Petrillo, E. & Munoz, M.J. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol.* **14**, 153-165 (2013).

41. Giulietti, M., Piva, F., D'Antonio, M., D'Onorio De Meo, P., Paoletti, D., Castrignano, T. et al. SpliceAid-F: a database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res.* **41**, D125-131 (2013).

42. Rahman, M.A., Nasrin, F., Masuda, A. & Ohno, K. Decoding abnormal splicing code in human diseases. *J Invest Genomics* **2(1)**, 00016 (2015).

43. Rieder, L.E. & Reenan, R.A. The intricate relationship between RNA structure, editing, and splicing. *Semin. Cell Dev. Biol.* **23**, 281-288 (2012).

Figure legends

Figure 1. Annotation-based analysis of the effects of intronic nucleotides on splicing. (a) The effect of each intronic nucleotide at positions Int-50:Int-3 and Ex+1:Ex+5 on the average *TR* (see Supplementary Figure 2) according to the ENSEMBL annotation 64. For example, G at position Int-3 is frequently observed in alternatively spliced 3’ ss, yielding a markedly reduced *TR*. **(b)** Schematic of the consensus nucleotide compositions of the BPS (arrow) and PPT.[21]

Figure 2. Sensitivities and specificities of IntSplice, the Shapiro-Senapathy score-based model, and the MaxEntScan-based model. (a) An SVM model generated by four-fifths of the 2324 normal and pathogenic Int-SNVs in the HGMD and dbSNP databases is applied to the remaining one-fifth of the Int-SNVs. The models are generated five times for 100 different datasets. Bars indicate mean and SD. As MaxEntScan is unable to score positions Int-50:Int-21, the MaxEntScan-based models in this region are not indicated. **(b)** IntSplice, the Shapiro-Senapathy score-based model, and the MaxEntScan-based model are generated with 2064 normal and pathogenic Int-SNVs at positions Int-20:Int-3. Mean and SD of the sensitivity and the specificity of 500 SVM models are plotted. Oblique lines indicate where the sums of the sensitivity and the specificity are identical. Note that the oblique lines are not ROC curves, and are auxiliary lines for comparing the sensitivity and the specificity of three models.

Figure 3. Representative results of the IntSplice web service program (http://www.med.nagoya-u.ac.jp/neurogenetics/IntSplice). Predicted results are shown in the “RESULT” column. The rightmost “NOTE” column indicates which exon in which ENSEMBL transcript is predicted to lead to abnormal or normal splicing. The information from the columns “CHROM” to “FILTER” is included in the submitted VCF file, and is not edited by IntSplice. For example, a G-to-A transition at position 73,550,880 of chromosome 10, which is registered in HGMD, is predicted to cause aberrant splicing.

Figure 4. Characterization of the *RAPSN* c.913-5T>A mutation identified in a patient with congenital myasthenic syndrome and of nine artificial mutations. (a) Schematic of pcDNA-*RAPSN* minigene harboring wild-type (wt) and mutant c.913-5T>A (mut) sequences. RT-PCR of the minigenes introduced into HeLa cells are shown. “A”, aberrant splicing; “-“, normal splicing. IntSplice, the Shapiro-Senapathy score-based model, and the MaxEntScan-based model correctly predict aberrant splicing. **(b)** Schematic of MS2-attached

wild-type (wt) and mutant (mut) RNA substrates that originated from pcDNA-*RAPSN*-E5-E6-MS2 minigene. An RNA-affinity-purified spliceosomal complex is immunoblotted with the indicated antibodies. U2AF65 and U2AF35 bind to the PPT and the 3' ss, respectively. U1-70K is a component of U1 snRNP that binds to the 5' ss. A β-globin-MS2 pre-mRNA substrate is employed as a control.[31] NE, nuclear extract (5%). **(c)** "T" (double underlined) is substituted for wild-type "A" at position Int-9 to make an optimized PPT (Opt) carrying an uninterrupted stretch of ten pyrimidines. A mutant "A" nucleotide (shown in red) is serially introduced at positions Int-11:Int-3. RT-PCR of the minigenes introduced into HeLa cells are shown. The splicing consequences predicted by IntSplice, the Shapiro-Senapathy score-based model, and the MaxEntScan-based model are indicated. Incorrectly predicted consequences are marked by an asterisk. "A", aberrant splicing; "-", normal splicing. **(d)** A spliceosome complex is purified and immunoblotted as in **(b)**.

Table 1. Comparison of the SVM kernels

Positions	Tool	SVM kernel	Specificity	Sensitivity
Int-50 to Int-3	PSSM	Linear	0.890 ± 0.020	0.560 ± 0.029
↓	↓	Polynomial	0.971 ± 0.010	0.394 ± 0.028
↓	↓	Radial basis function	0.889 ± 0.020[1]	0.561 ± 0.029[1]
↓	↓	Sigmoid	0.700 ± 0.139	0.587 ± 0.056
↓	IntSplice	Linear	0.934 ± 0.018	0.715 ± 0.029
↓	↓	Polynomial	0.896 ± 0.022	0.769 ± 0.028
↓	↓	Radial basis function	0.899 ± 0.022[1]	0.772 ± 0.027[1]
↓	↓	Sigmoid	0.936 ± 0.019	0.710 ± 0.030
Int-20 to Int-3	PSSM	Linear	0.704 ± 0.052	0.623 ± 0.033
↓	↓	Polynomial	0.833 ± 0.041	0.544 ± 0.031
↓	↓	Radial basis function	0.831 ± 0.044[1]	0.545 ± 0.032[1]
↓	↓	Sigmoid	0.703 ± 0.052	0.624 ± 0.032
↓	IntSplice	Linear	0.909 ± 0.045	0.756 ± 0.034
↓	↓	Polynomial	0.841 ± 0.052	0.808 ± 0.028
↓	↓	Radial basis function	0.821 ± 0.055[1]	0.817 ± 0.030[1]
↓	↓	Sigmoid	0.910 ± 0.054	0.751 ± 0.034
↓	MaxEntScan	Linear	0.937 ± 0.017	0.663 ± 0.031
↓	↓	Polynomial	0.539 ± 0.497	0.471 ± 0.492
↓	↓	Radial basis function	0.924 ± 0.019[1]	0.687 ± 0.033[1]
↓	↓	Sigmoid	0.567 ± 0.174	0.551 ± 0.170
Int-50 to Int-21	PSSM	Linear	0.989 ± 0.008	0.056 ± 0.039
↓	↓	Polynomial	0.998 ± 0.003	0.009 ± 0.018
↓	↓	Radial basis function	0.998 ± 0.004[1]	0.010 ± 0.019[1]
↓	↓	Sigmoid	0.989 ± 0.008	0.056 ± 0.039
↓	IntSplice	Linear	0.950 ± 0.018	0.347 ± 0.086
↓	↓	Polynomial	0.951 ± 0.018	0.343 ± 0.083
↓	↓	Radial basis function	0.949 ± 0.018[1]	0.352 ± 0.086[1]
↓	↓	Sigmoid	0.951 ± 0.018	0.343 ± 0.084

Mean and SD are indicated. MaxEntScan can be applied to SNVs at positions Int-20 to Int-3. [1]SVM modeling with the radial basis function leads to the most discriminating models on average. The sensitivity and the specificity of the radial basis function-based SVM models at positions Int-50:Int-3 and Int-20:Int-3 are plotted in Figure 2.

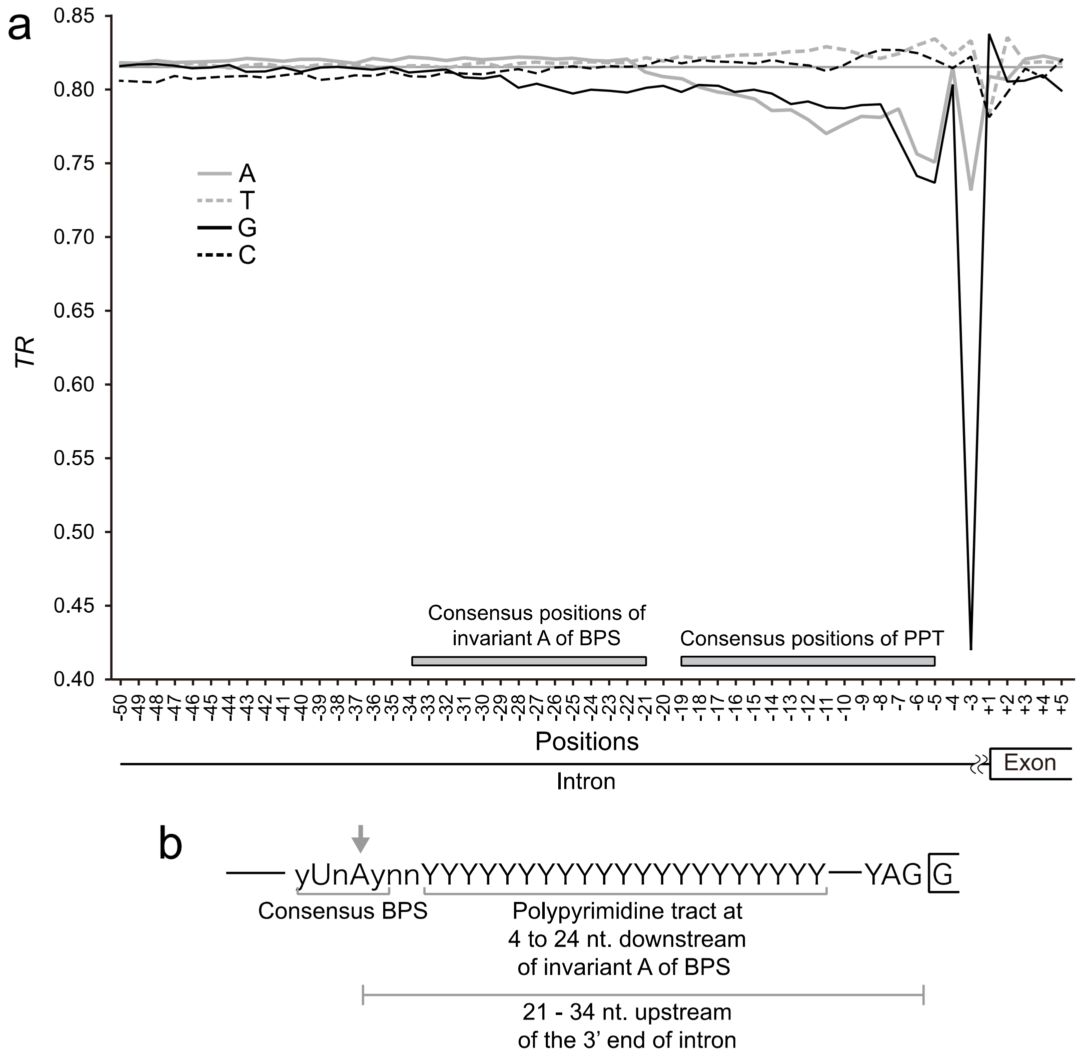
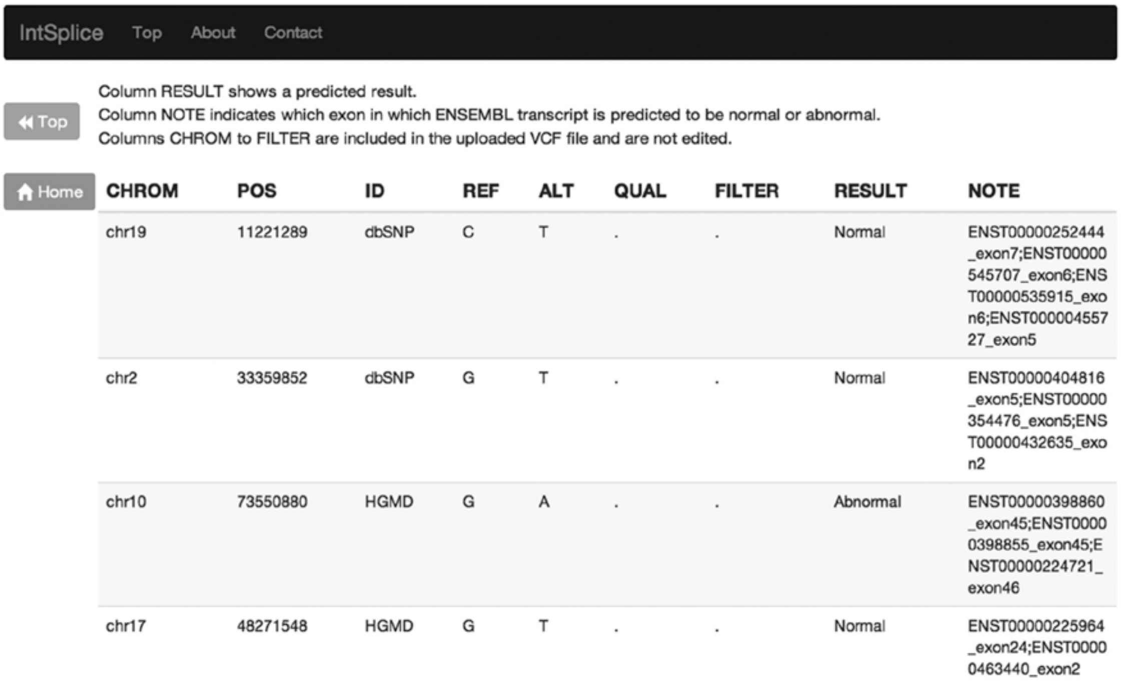


Figure 1. Annotation-based analysis of the effects of intronic nucleotides on splicing. (a) The effect of each intronic nucleotide at positions Int-50:Int-3 and Ex+1:Ex+5 on the average *TR* (see Supplementary Figure 2) according to the ENSEMBL annotation 64. For example, G at position Int-3 is frequently observed in alternatively spliced 3' ss, yielding a markedly reduced *TR*. **(b)** Schematic of the consensus nucleotide compositions of the BPS (arrow) and PPT.[21]

Figure 2. Sensitivities and specificities of IntSplice, the Shapiro-Senapathy score-based model, and the MaxEntScan-based model. (a) An SVM model generated by four-fifths of the 2324 normal and pathogenic Int-SNVs in the HGMD and dbSNP databases is applied to the remaining one-fifth of the Int-SNVs. The models are generated five times for 100 different datasets. Bars indicate mean and SD. As MaxEntScan is unable to score positions Int-50:Int-21, the MaxEntScan-based models in this region are not indicated. **(b)** IntSplice, the Shapiro-Senapathy score-based model, and the MaxEntScan-based model are generated with 2064 normal and pathogenic Int-SNVs at positions Int-20:Int-3. Mean and SD of the sensitivity and the specificity of 500 SVM models are plotted. Oblique lines indicate where the sums of the sensitivity and the specificity are identical. Note that the oblique lines are not ROC curves, and are auxiliary lines for comparing the sensitivity and the specificity of three models.



Column RESULT shows a predicted result.
Column NOTE indicates which exon in which ENSEMBL transcript is predicted to be normal or abnormal.
Columns CHROM to FILTER are included in the uploaded VCF file and are not edited.

CHROM	POS	ID	REF	ALT	QUAL	FILTER	RESULT	NOTE
chr19	11221289	dbSNP	C	T	.	.	Normal	ENST00000252444_exon7;ENST00000545707_exon6;ENST00000535915_exon6;ENST00000455727_exon5
chr2	33359852	dbSNP	G	T	.	.	Normal	ENST00000404816_exon5;ENST00000354476_exon5;ENST00000432635_exon2
chr10	73550880	HGMD	G	A	.	.	Abnormal	ENST00000398860_exon45;ENST00000398855_exon45;ENST00000224721_exon46
chr17	48271548	HGMD	G	T	.	.	Normal	ENST00000225964_exon24;ENST00000463440_exon2

Figure 3. Representative results of the IntSplice web service program (http://www.med.nagoya-u.ac.jp/neurogenetics/IntSplice). Predicted results are shown in the "RESULT" column. The rightmost "NOTE" column indicates which exon in which ENSEMBL transcript is predicted to lead to abnormal or normal splicing. The information from the columns "CHROM" to "FILTER" is included in the submitted VCF file, and is not edited by IntSplice. For example, a G-to-A transition at position 73,550,880 of chromosome 10, which is registered in HGMD, is predicted to cause aberrant splicing.

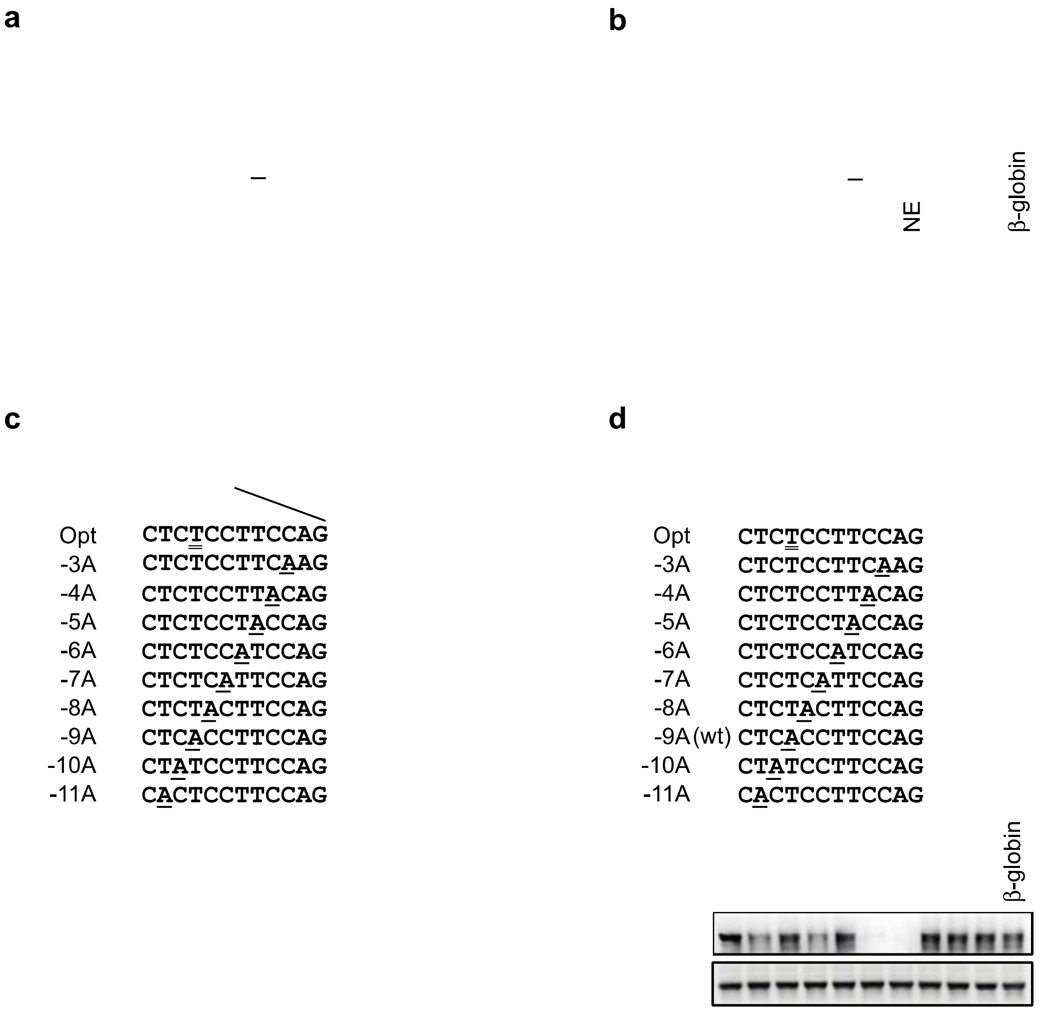
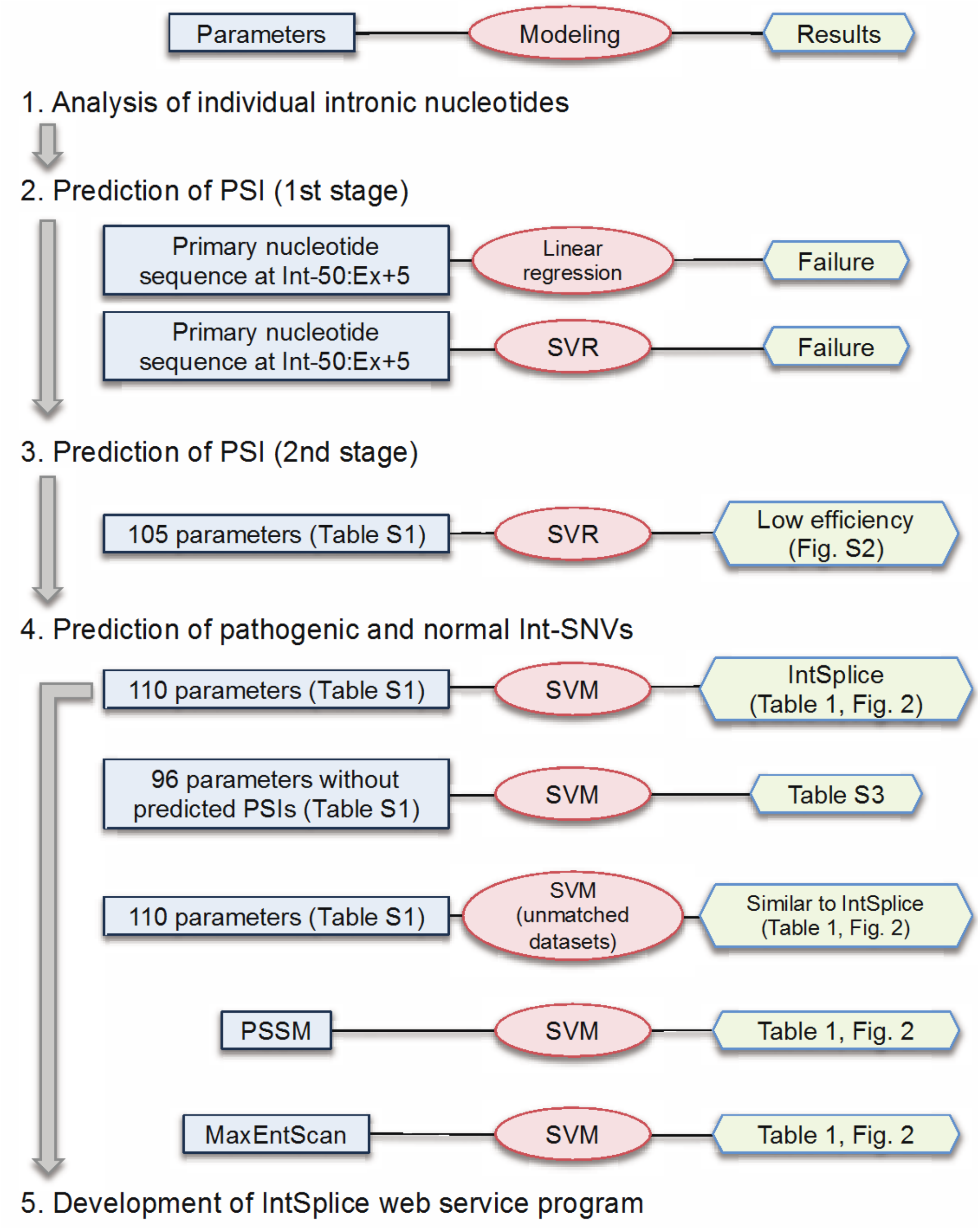
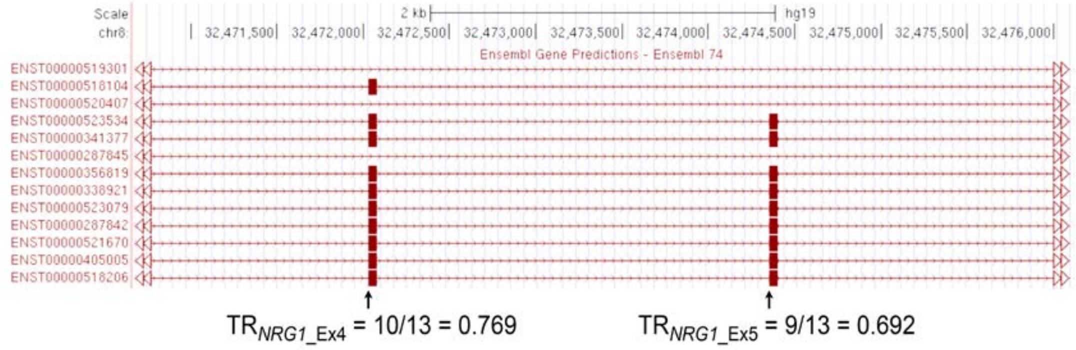


Figure 4. Characterization of the *RAPSN* c.913-5T>A mutation identified in a patient with congenital myasthenic syndrome and of nine artificial mutations. (a) Schematic of pcDNA-*RAPSN* minigene harboring wild-type (wt) and mutant c.913-5T>A (mut) sequences. RT-PCR of the minigenes introduced into HeLa cells are shown. "A", aberrant splicing; "-", normal splicing. IntSplice, the Shapiro-Senapathy score-based model, and the MaxEntScan-based model correctly predict aberrant splicing. **(b)** Schematic of MS2-attached wild-type (wt) and mutant (mut) RNA substrates that originated from pcDNA-*RAPSN*-E5-E6-MS2 minigene. An RNA-affinity-purified spliceosomal complex is immunoblotted with the indicated antibodies. U2AF65 and U2AF35 bind to the PPT and the 3' ss, respectively. U1-70K is a component of U1 snRNP that binds to the 5' ss. A β-globin-MS2 pre-mRNA substrate is employed as a control.[31] NE, nuclear extract (5%). **(c)** "T" (double underlined) is substituted for wild-type "A" at position Int-9 to make an optimized PPT (Opt) carrying an uninterrupted stretch of ten pyrimidines. A mutant "A" nucleotide (shown in red) is serially introduced at positions Int-11:Int-3. RT-PCR of the minigenes introduced into HeLa cells are shown. The splicing consequences predicted by IntSplice, the Shapiro-Senapathy score-based model, and the MaxEntScan-based model are indicated. Incorrectly predicted consequences are marked by an asterisk. "A", aberrant splicing; "-", normal splicing. **(d)** A spliceosome complex is purified and immunoblotted as in **(b)**.

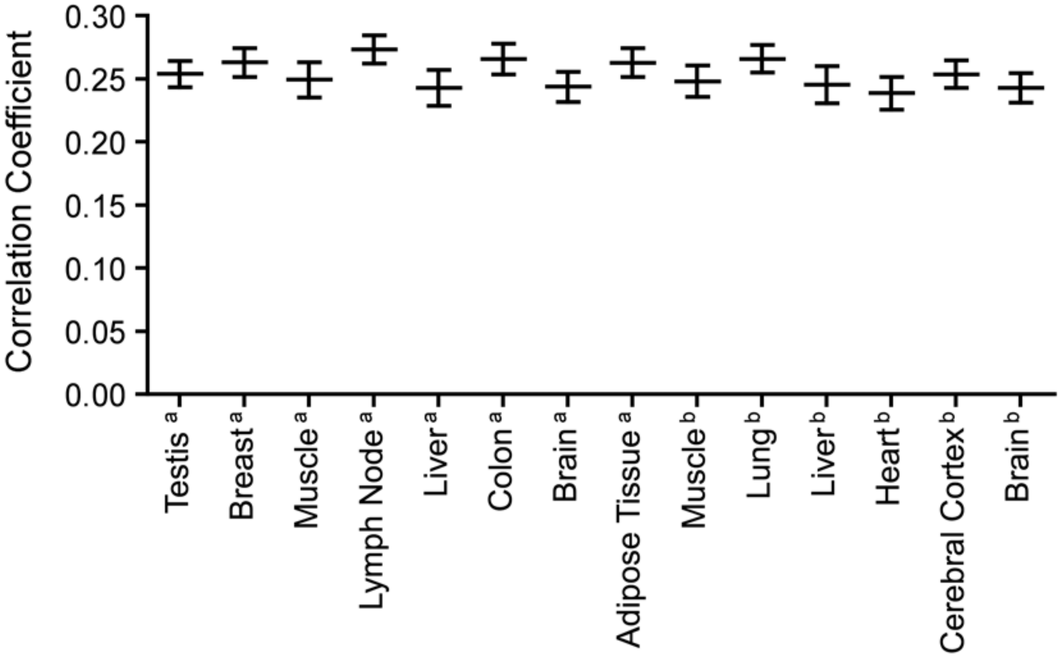
Supplementary Figure S1. A diagram showing the flow of analyses in this communication.


Parameters
Modeling
Results
1. Analysis of individual intronic nucleotides
2. Prediction of PSI (1st stage)
Primary nucleotide sequence at Int-50:Ex+5
Linear regression
Failure
Primary nucleotide sequence at Int-50:Ex+5
SVR
Failure
3. Prediction of PSI (2nd stage)
105 parameters (Table S1)
SVR
Low efficiency (Fig. S2)
4. Prediction of pathogenic and normal Int-SNVs
110 parameters (Table S1)
SVM
IntSplice (Table 1, Fig. 2)
96 parameters without predicted PSIs (Table S1)
SVM
Table S3
110 parameters (Table S1)
SVM (unmatched datasets)
Similar to IntSplice (Table 1, Fig. 2)
PSSM
SVM
Table 1, Fig. 2
MaxEntScan
SVM
Table 1, Fig. 2
5. Development of IntSplice web service program

Supplementary Figure S2. Examples of TR values of exons 4 and 5 of the *NRG1* gene on chromosome 8. TR values are calculated according the ENSEMBL annotation.



Supplementary Figure S3. Correlation coefficients of the SVR models applied to the validation dataset of each tissue-specific RNA-seq data. Each set of RNA-seq data is randomly divided into five groups, and each group is used as a validation dataset. A correlation coefficient between the MISO-generated actual PSIs and the SVR-predicted PSIs is calculated for each validation dataset. Five hundred different SVR models are generated from 100 different combinations of the five groups for each RNA-seq dataset, and the mean and SD are plotted. [a,b]The RNA-seq data are obtained from the GEO database with the accession numbers of GSE13652[a] and GSE12946[b], respectively.



Supplementary Table S1. Parameters to generate SVR and SVM models

Parameter	3’/5’	Position[a]	SVR[b]	SVM[c]
Best-BPS[d]				
Number of nucleotides from best-BPS to Int-3	3'	-50 to -3	0.692	-
Number of G's from best-BPS to Int-3	3'	-50 to -3	-1.730	6.022
Best-BPS at Int-50:-3	3'	-50 to -3	-1.230	-7.720
PPT				
Longest stretch of C/T's with 0 break at Int-50:-3	3'	-50 to -3	2.705	-1.990
Longest stretch of C/T's with 1 break at Int-50:-3	3'	-50 to -3	-0.640	1.317
Longest stretch of C/T's with 2 breaks at Int-50:-3	3'	-50 to -3	0.944	-1.510
Longest stretch of C/T's with 3 breaks at Int-50:-3	3'	-50 to -3	-2.690	0.872
Longest stretch of C/T's with 4 breaks at Int-50:-3	3'	-50 to -3	-0.900	-0.600
Length of a T stretch allowing 0 break	3'	-50 to -3	1.603	-1.160
Length of a T stretch allowing 1 break	3'	-50 to -3	-2.490	0.919
Length of a T stretch allowing 2 breaks	3'	-50 to -3	1.366	-2.460
Length of a T stretch allowing 3 breaks	3'	-50 to -3	-1.260	-1.850
Length of a T stretch allowing 4 breaks	3'	-50 to -3	-2.810	1.025
Best-BPS-PPT[e]				
PWM score of BPS of best-BPS-PPT	3'	-50 to -3	2.199	-0.710
Invariant A of best-BPS-PPT is at Int-34:-21	3'	-50 to -3	-0.280	-8.030
Ratio of C/T's in PPT of best-BPS-PPT	3'	-50 to -3	5.131	-2.800
Ratio of T's in PPT of best-BPS-PPT	3'	-50 to -3	0.454	-7.260
Ratio of G's in PPT of best-BPS-PPT	3'	-50 to -3	4.206	7.562
Length of PPT of best-BPS-PPT	3'	-50 to -3	0.304	-6.310
Individual nucleotides				
A at Int-3	3'	-3	-0.580	2.494
C at Int-3	3'	-3	-3.320	-4.340
G at Int-3	3'	-3	0.421	0.846
T at Int-3	3'	-3	3.495	1.001
A at Ex+1	3'	+1	0.350	-
C at Ex+1	3'	+1	-0.580	-
G at Ex+1	3'	+1	0.603	-
T at Ex+1	3'	+1	-0.360	-

A/G’s at Int-7 to Int-5 (Int-7:-5)	3'	-7 to -5	0.813	-1.190
Ratio of A/G's at Int-20:-8	3'	-20 to -8	1.330	10.800
Number of G's at Int-12:-3	3'	-12 to -3	0.760	7.834
Other parameters				
SD_score	5'	-3 to +6	3.889	-
SD_this_intron	5'	-3 to +6	1.717	-
Exon length	-	-	22.350	-
1/(exon length)	-	-	-12.900	-
Presence of 'GGG' at Int-12:-3	3'	-12 to -3	-1.010	0.431
MaxEnt score of Int-20:+3[1]	3'	-20 to +3	5.336	-12.700
MaxEnt score of Int-3:+6[1]	5'	-3 to +6	9.065	-
MaxEnt score of Int-3:+6[1]	5'	-3 to +6	-4.890	-
Shapiro-Senapathy score[2]	3'	-50 to -3	-5.660	-11.200
Generation of *de novo* AG by an SNV	3'	-50 to -3	-	6.814
∆G of predicted secondary structure by mfold[3]	3'	-50 to +5	-3.900	-1.070
SpliceAid scores of RNA-binding protein[f]				
9G8	3'	-50 to +5	1.887	2.805
CUG-BP1	3'	-50 to +5	-0.490	-0.300
DAZAP1	3'	-50 to +5	-0.150	1.209
ESRP1	3'	-50 to +5	0.914	-1.000
ETR-3	3'	-50 to +5	5.352	0.977
FMRP	3'	-50 to +5	-0.340	0.582
Fox-1	3'	-50 to +5	1.692	2.556
Fox-2	3'	-50 to +5	1.575	1.082
HTra2alpha	3'	-50 to +5	0.936	-0.940
HTra2beta1	3'	-50 to +5	-1.000	-1.080
HuB	3'	-50 to +5	-3.030	4.727
HuC	3'	-50 to +5	0.080	0.101
HuD	3'	-50 to +5	-0.760	-1.760
HuR	3'	-50 to +5	-10.600	1.199
KSRP	3'	-50 to +5	0.278	2.042
MBNL1	3'	-50 to +5	3.243	-0.090
Nova-1	3'	-50 to +5	-4.280	-0.160
Nova-2	3'	-50 to +5	-2.700	0.173

PSF	3'	-50 to +5	-0.240	0.348
RBM25	3'	-50 to +5	1.172	0.996
RBM4	3'	-50 to +5	1.278	-0.600
RBM5	3'	-50 to +5	0.924	2.504
SC35	3'	-50 to +5	0.023	-3.970
SF1	3'	-50 to +5	-1.340	1.926
SF2/ASF	3'	-50 to +5	-0.180	-0.930
SLM-1	3'	-50 to +5	-0.080	-0.100
SLM-2	3'	-50 to +5	-0.400	0.292
SRp20	3'	-50 to +5	-3.250	3.508
SRp30c	3'	-50 to +5	-0.060	-0.920
SRp38	3'	-50 to +5	1.793	1.734
SRp40	3'	-50 to +5	-1.350	4.633
SRp54	3'	-50 to +5	-0.950	-0.350
SRp55	3'	-50 to +5	1.420	2.564
SRp75	3'	-50 to +5	-2.100	0.064
Sam68	3'	-50 to +5	-3.050	-2.000
TDP43	3'	-50 to +5	-0.470	-2.590
TIA-1	3'	-50 to +5	-4.970	-5.690
TIAL1	3'	-50 to +5	1.941	-7.110
YB-1	3'	-50 to +5	0.624	-2.720
ZRANB2	3'	-50 to +5	2.611	-0.490
hnRNP A0	3'	-50 to +5	-0.780	3.318
hnRNP A1	3'	-50 to +5	7.275	2.996
hnRNP A2/B1	3'	-50 to +5	-0.700	1.194
hnRNP C	3'	-50 to +5	-1.810	0.374
hnRNP C1	3'	-50 to +5	-0.660	-1.010
hnRNP C2	3'	-50 to +5	0.120	0.424
hnRNP D	3'	-50 to +5	-0.480	0.015
hnRNP D0	3'	-50 to +5	1.064	-1.000
hnRNP DL	3'	-50 to +5	0.851	1.025
hnRNP E1	3'	-50 to +5	1.266	1.196
hnRNP E2	3'	-50 to +5	1.700	3.600
hnRNP F	3'	-50 to +5	1.509	-3.920

hnRNP H1	3'	-50 to +5	-0.620	-0.020
hnRNP H2	3'	-50 to +5	-0.620	-0.020
hnRNP H3	3'	-50 to +5	-0.310	0.579
hnRNP I (PTB)	3'	-50 to +5	-0.010	3.175
hnRNP J	3'	-50 to +5	-0.550	2.858
hnRNP K	3'	-50 to +5	0.893	3.552
hnRNP L[g]	3'	-50 to +5	0.000	-
hnRNP LL[g]	3'	-50 to +5	0.000	-
hnRNP M	3'	-50 to +5	1.634	-0.860
hnRNP P (TLS)	3'	-50 to +5	-0.550	1.363
hnRNP Q	3'	-50 to +5	-0.280	0.717
hnRNP U	3'	-50 to +5	-0.020	1.446
nPTB	3'	-50 to +5	0.845	2.641
SVR-based prediction of PSI				
Predicted PSI (skeletal muscle RNA-seq)[4]	n.a.	n.a.	-	-2.980
Predicted PSI (lung RNA-seq)[4]	n.a.	n.a.	-	-3.580
Predicted PSI (liver RNA-seq)[4]	n.a.	n.a.	-	-4.070
Predicted PSI (heart RNA-seq)[4]	n.a.	n.a.	-	-3.260
Predicted PSI (cerebral cortex RNA-seq)[4]	n.a.	n.a.	-	-6.550
Predicted PSI (brain RNA-seq)[4]	n.a.	n.a.	-	-8.170
Predicted PSI (testis RNA-seq)[5]	n.a.	n.a.	-	-4.490
Predicted PSI (breast RNA-seq)[5]	n.a.	n.a.	-	3.559
Predicted PSI (skeletal muscle RNA-seq)[5]	n.a.	n.a.	-	-3.270
Predicted PSI (lymph node RNA-seq)[5]	n.a.	n.a.	-	-3.060
Predicted PSI (liver RNA-seq)[5]	n.a.	n.a.	-	-4.540
Predicted PSI (colon RNA-seq)[5]	n.a.	n.a.	-	-7.280
Predicted PSI (brain RNA-seq)[5]	n.a.	n.a.	-	-5.900
Predicted PSI (adipose RNA-seq)[5]	n.a.	n.a.	-	-2.230

[a] Postions to which the indicated parameter is applied.

[b] “SVR” indicates a mean of coefficients for each parameter generated by 500 repetitions of SVR modeling using colon RNA-seq to predict PSIs. “-“ indicates that the parameter was not used for SVR modeling.

[c] "SVM" indicates a mean of coefficients for each parameter generated by SVM modeling (radial

basis function) using 100 different datasets to predict HGMD and dbSNP. “-“ indicates that the parameter was not used for SVM modeling.

[d] Best-BPS, BPS with the highest PWM (position weight matrix) according to Gao *et al*.[6]

[e] Best-BPS-PPT, the best pair of BPS and PPT according to the following algorithm. First, ‘nYnAn’ motif is looked for with an invariant ‘A’ at Int-50:Int-3 and set to be BPS_i ($i \subseteq \mathrm{N}$). Second, the ratio of T/C’s at positions +4 to +24 from the invariant ‘A’ of BPS_i is calculated, while BPS_i downstream of Int-9 is excluded because the length of putative PPT_i becomes less than 7 nucleotides. This gives rise to multiple candidate BPS_i-PPT_i pairs. The sum of the PWM of BPS_i and the T/C ratio in PPT_i is then calculated and a pair with the best sum score is selected.

[f] The exact motif for an RNA-binding protein is searched for at Int-50:Ex+5 and scored according to SpliceAid. The sum of SpliceAid scores is used as a parameter for each RNA-binding protein after the sum is normalized using the SVM-scale functionality of LIBSVM[7].

[g] Only hnRNPs L and LL give a primal variable of 0.000 and do not contribute to SVR modeling. This is likely because 166 binding sites for hnRNPs L and LL in SpliceAid are all long motifs comprised of nine or more nucleotides.

Methods to calculate individual parameters are available upon request.

Supplementary Table S2. Nucleotide frequencies at positions Int-50:Ex+5 of major introns annotated in ENSEMBL 64 to calculate Shapiro-Senapathy scores

Position	A	T	G	C
Int-50	0.252	0.300	0.238	0.210
Int-49	0.250	0.301	0.234	0.215
Int-48	0.249	0.301	0.235	0.214
Int-47	0.250	0.304	0.230	0.216
Int-46	0.249	0.303	0.230	0.217
Int-45	0.250	0.306	0.229	0.216
Int-44	0.250	0.304	0.227	0.220
Int-43	0.248	0.307	0.224	0.221
Int-42	0.249	0.304	0.227	0.220
Int-41	0.250	0.306	0.223	0.221
Int-40	0.249	0.307	0.223	0.221
Int-39	0.249	0.310	0.218	0.223
Int-38	0.247	0.309	0.218	0.226
Int-37	0.249	0.309	0.214	0.228
Int-36	0.248	0.308	0.212	0.232
Int-35	0.249	0.311	0.206	0.234
Int-34	0.249	0.314	0.201	0.236
Int-33	0.249	0.319	0.198	0.235
Int-32	0.249	0.321	0.196	0.234
Int-31	0.251	0.322	0.191	0.236
Int-30	0.250	0.328	0.186	0.236
Int-29	0.251	0.332	0.180	0.237
Int-28	0.252	0.336	0.175	0.237
Int-27	0.252	0.338	0.170	0.240
Int-26	0.250	0.346	0.164	0.240
Int-25	0.249	0.350	0.161	0.240
Int-24	0.243	0.355	0.156	0.245
Int-23	0.238	0.359	0.155	0.247
Int-22	0.229	0.372	0.153	0.246
Int-21	0.217	0.381	0.152	0.250

Int-20	0.205	0.397	0.148	0.251
Int-19	0.185	0.409	0.148	0.258
Int-18	0.169	0.426	0.143	0.262
Int-17	0.153	0.444	0.138	0.265
Int-16	0.141	0.454	0.137	0.268
Int-15	0.131	0.464	0.129	0.277
Int-14	0.122	0.482	0.123	0.272
Int-13	0.111	0.499	0.119	0.271
Int-12	0.103	0.514	0.109	0.274
Int-11	0.097	0.542	0.104	0.256
Int-10	0.097	0.515	0.111	0.277
Int-9	0.109	0.488	0.118	0.286
Int-8	0.118	0.466	0.106	0.309
Int-7	0.123	0.459	0.098	0.321
Int-6	0.099	0.502	0.071	0.328
Int-5	0.102	0.538	0.073	0.287
Int-4	0.245	0.280	0.205	0.270
Int-3	0.068	0.281	0.013	0.638
Int-2	1.000	0.000	0.000	0.000
Int-1	0.000	0.000	1.000	0.000
Ex+1	0.265	0.118	0.469	0.148
Ex+2	0.252	0.359	0.195	0.194
Ex+3	0.261	0.274	0.236	0.230
Ex+4	0.248	0.264	0.224	0.264
Ex+5	0.267	0.276	0.212	0.245

Table S3. Comparison of SVM models with or without SVR-based prediction of PSI of 14 tissues

Positions	SVR-based prediction of PSI	Specificity	Sensitivity
Int-50 to Int-3	with	0.899 ± 0.022	0.772 ± 0.027
↓	without	0.905 ± 0.024	0.762 ± 0.030
↓	difference	0.006	-0.010
Int-20 to Int-3	with	0.821 ± 0.055	0.817 ± 0.030
↓	without	0.829 ± 0.060	0.814 ± 0.030
↓	difference	0.008	-0.004
Int-50 to Int-21	with	0.949 ± 0.018	0.352 ± 0.086
↓	without	0.949 ± 0.018	0.348 ± 0.087
↓	difference	0.000	-0.004

Mean and SD are indicated. The SVM kernel was the radial basis function. SVM models with SVR-based prediction of PSI are identical to those shown in Table 1.

Supplementary references

1. Yeo G., Burge C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377-394 (2004).
2. Shapiro M. B., Senapathy P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* **15**, 7155-7174 (1987).
3. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406-3415 (2003).
4. Pan Q., Shai O., Lee L. J., Frey B. J., Blencowe B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413-1415 (2008).
5. Wang E. T., Sandberg R., Luo S., Khrebtukova I., Zhang L., Mayr C., et al. Alternative isoform regulation in human tissue transcriptomes. *Nature.* **456**, 470-476 (2008).
6. Gao K., Masuda A., Matsuura T., Ohno K. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res.* **36**, 2257-2267 (2008).
7. Chang C. C., Lin C. J. LIBSVM: A Library for Support Vector Machines. *ACM T Intel Syst Tec.* **2**, Article 27 (2011).