genes

MDPI

Article

FexSplice: A LightGBM-Based Model for Predicting the Splicing Effect of a Single Nucleotide Variant Affecting the First Nucleotide G of an Exon

Atefeh Joudaki [1], Jun-ichi Takeda [1], Akio Masuda [1], Rikumo Ode [2], Koichi Fujiwara [2] and Kinji Ohno [1,*]

[1] Division of Neurogenetics, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, 65 Tsurumai, Showa-ku, Nagoya 466-8550, Japan; atefeh.joudaki@gmail.com (A.J.); jtakeda@med.nagoya-u.ac.jp (J.-i.T.); amasuda@med.nagoya-u.ac.jp (A.M.)

[2] Department of Materials Science and Engineering, Nagoya University Graduate School of Engineering, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan; r1kum0.0d3@gmail.com (R.O.); fujiwara.koichi@hps.material.nagoya-u.ac.jp (K.F.)

* Correspondence: ohnok@med.nagoya-u.ac.jp

Abstract: Single nucleotide variants (SNVs) affecting the first nucleotide G of an exon (Fex-SNVs) identified in various diseases are mostly recognized as missense or nonsense variants. Their effect on pre-mRNA splicing has been seldom analyzed, and no curated database is available. We previously reported that Fex-SNVs affect splicing when the length of the polypyrimidine tract is short or degenerate. However, we cannot readily predict the splicing effects of Fex-SNVs. We here scrutinized the available literature and identified 106 splicing-affecting Fex-SNVs based on experimental evidence. We similarly identified 106 neutral Fex-SNVs in the dbSNP database with a global minor allele frequency (MAF) of more than 0.01 and less than 0.50. We extracted 115 features representing the strength of splicing *cis*-elements and developed machine-learning models with support vector machine, random forest, and gradient boosting to discriminate splicing-affecting and neutral Fex-SNVs. Gradient boosting-based LightGBM outperformed the other two models, and the length and nucleotide compositions of the polypyrimidine tract played critical roles in the discrimination. Recursive feature elimination showed that the LightGBM model using 15 features achieved the best performance with an accuracy of 0.80 ± 0.12 (mean and SD), a Matthews Correlation Coefficient (MCC) of 0.57 ± 0.15, an area under the curve of the receiver operating characteristics curve (AUROC) of 0.86 ± 0.08, and an area under the curve of the precision–recall curve (AUPRC) of 0.87 ± 0.09 using a 10-fold cross-validation. We developed a web service program, named FexSplice that accepts a genomic coordinate either on GRCh37/hg19 or GRCh38/hg38 and returns a predicted probability of aberrant splicing of A, C, and T variants.

Keywords: first nucleotide of an exon; splicing-affecting variants; LightGBM model; FexSplice web service program

Citation: Joudaki, A.; Takeda, J.-i.; Masuda, A.; Ode, R.; Fujiwara, K.; Ohno, K. FexSplice: A LightGBM-Based Model for Predicting the Splicing Effect of a Single Nucleotide Variant Affecting the First Nucleotide G of an Exon. *Genes* **2023**, *14*, 1765. https://doi.org/10.3390/genes14091765

Academic Editors: Carlo Maria Di Liegro and Gabriella Schiera

Received: 2 August 2023
Revised: 30 August 2023
Accepted: 4 September 2023
Published: 6 September 2023

1. Introduction

Pre-mRNA splicing is a fundamental process in eukaryotic gene expression that involves the precise removal of introns and the joining of exons to generate mature mRNA. Splicing is mediated by the spliceosome, a dynamic and highly regulated macromolecular complex consisting of target pre-mRNA, small nuclear ribonucleoproteins (snRNPs), and numerous other proteins. The spliceosome catalyzes splicing in two steps. In the first step, the spliceosome is assembled on pre-mRNA, where the intron/exon and exon/intron boundaries comprised specific *cis*-elements including the 5′ splice site (ss), 3′ ss, polypyrimidine tract (PPT), and branch point sequence (BPS) are recognized by *trans*-acting RNA-binding proteins (RBPs) such as the snRNPs, heterogeneous nuclear ribonucleoproteins (hnRNPs), and serine arginine-rich splicing factors (SRSFs) [1,2]. Single nucleotide variations (SNVs) that disrupt *cis*-acting splicing elements and compromise catalytic functions

of *trans*-acting RBPs impair finely tuned alternative and constitutive splicing events [3]. Disruptions in splicing have been implicated in a wide range of diseases including cancer, neurodegenerative disorders, and Mendelian disorders such as congenital myasthenic syndromes [4].

The BPS and PPT are first recognized by SF1 and U2AF65, respectively [5]. Introns with a long PPT do not require the binding of U2AF35 to the intron–exon boundary because U2AF65 is able to bind to PPT strongly, which is called an AG-independent 3′ ss (Figure 1). Conversely, introns with a short or degenerate PPT require the binding of U2AF35 to the intron–exon boundary to reinforce the binding of U2AF65 to PPT, which is called an AG-dependent 3′ ss. We previously reported that SNVs affecting the first nucleotide G of an exon (Fex-SNVs) cause aberrant splicing at the AG-dependent 3′ ss's but not at the AG-independent 3′ ss's [6], which has also been proven at the structural level by others [7]. Serial mutagenesis to gradually increase the length of PPT revealed that a stretch of pyrimidines in PPT needs to be 10 to 15 nucleotides or more to make the 3′ ss insensitive to a Fex-SNV [6]. When the first nucleotide of an exon is not G in the reference sequence, binding of U2AF35 to the intron–exon boundary is predicted to be weak, and such 3′ ss's are mostly AG-independent [8].

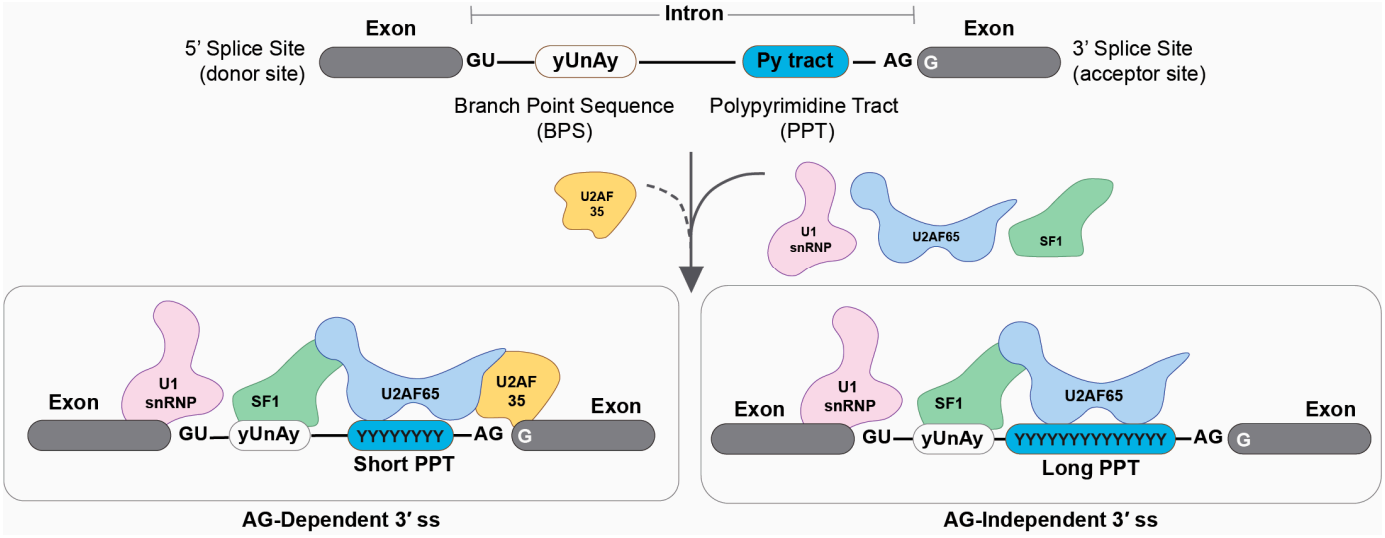


Figure 1. AG-dependent and AG-independent 3′ splice sites (ss's). Introns with a short or degenerate PPT require both U2AF65 and U2AF35 for the recognition of the 3′ ss, which is called the AG-dependent 3′ ss. Introns with a long stretch of PPT strongly bind to U2AF65 and do not require binding of U2AF35, which is called the AG-independent 3′ ss. The 3′ ss's without a G at the first nucleotide of an exon in the reference sequence are mostly AG-independent.

Although the AG-dependence of the 3′ ss predicts the splicing effects of Fex-SNVs, there is no dependable rule to determine the AG-dependence of the 3′ ss. Several prediction tools such as SpliceAI [9] and Collapsed Isoform SpliceAI (CI-SpliceAI) [10] have been developed to predict the splicing consequences of SNVs. However, these tools were not optimized for predicting the splicing effects of Fex-SNVs. We previously developed web service programs of support vector machine (SVM)-based IntSplice (https://www.med.nagoya-u.ac.jp/neurogenetics/IntSplice_v1.0/) (accessed on 1 August 2023) [11] and gradient boosting-based IntSplice2 (https://www.med.nagoya-u.ac.jp/neurogenetics/IntSplice2/) (accessed on 1 August 2023) [12], both of which predict the splicing effects of intronic SNVs at positions −50 to −3, but do not cover Fex-SNVs. To address this challenge, we first curated a dependable dataset that comprised Fex-SNVs and their splicing effects by scrutinizing available articles, and developed a machine-learning model, FexSplice, using Light Gradient Boosting Machine (LightGBM) [13] dedicated to predicting the splicing

effects of Fex-SNVs. We hope that FexSplice sheds light on frequently underestimated splicing-affecting Fex-SNVs.

2. Materials and Methods

2.1. Fex-SNV Dataset

We scrutinized Fex-SNVs in the Human Gene Mutation Database (HGMD) Professional released in April 2020 [14], the ClinVar released on 15 March 2021 [15], and PubMed including a recently published article on splicing variants [16]. We only collected Fex-SNVs with G as the first nucleotide of an exon in the reference sequence.

For HGMD Pro, we chose disease-associated SNVs in the mutation categories of DM (disease-causing mutation) and SM (splicing mutation). For ClinVar 2021 [15], we chose disease-associated SNVs with CLNSIG = pathogenic. We thus identified 801 Fex-SNVs according to the transcript annotations of Ensembl release 101 [17]. We first eliminated Fex-SNVs in the first and last exons because these exons had no upstream and downstream sequences, respectively, and some features could not be extracted from these exons. The predicted amino acid substitutions of Fex-SNVs were annotated in HGMD Pro, ClinVar, and the literature, but their effects on pre-mRNA splicing, if any, remained mostly unannotated except for the literature. We thus scrutinized the experimental details of available articles to accurately annotate Fex-SNVs. A Fex-SNV was recognized as splicing-affecting when aberrant splicing was demonstrated using RT-PCR of either the patient sample or a minigene construct. If RefSeq [18] shows two or more splicing isoforms at a Fex-SNV, the Fex-SNV was included when authors addressed which splicing isoform was affected by the Fex-SNV. In contrast, when authors did not address the splicing isoforms, the Fex-SNV was excluded from our dataset. These filtrations reduced the number of Fex-SNVs to 106 splicing-affecting and 5 neutral Fex-SNVs in HGMD Pro, ClinVar, and the literature (Supplementary Table S1a).

For additional neutral Fex-SNVs, we extracted 1005 Fex-SNVs from dbSNP (build 151) on GRCh37/hg19 [19]. The 1005 neutral Fex-SNVs were first filtered by a global minor allelic frequency (MAF) greater than 0.01 and less than 0.5, which produced 156 neutral Fex-SNVs. MAF > 0.5 indicates that the reference nucleotide is minor. To match the numbers of splicing-affecting and neutral Fex-SNVs, we randomly selected 101 out of 156 neutral Fex-SNVs. In the selection, we attempted to exclude Fex-SNVs with similar flanking sequences or neutral Fex-SNVs identified in the course of disease analysis. By adding 5 neutral Fex-SNVs in HGMD Pro, ClinVar, and the literature stated above, we obtained 106 neutral Fex-SNVs (Supplementary Table S1b).

2.2. Extraction of Features

We first extracted 115 features dictating the strength of splicing *cis*-elements, most of which were used to predict the splicing effects of intronic SNVs (IntSplice [11] and IntSplice2 [12]) (Supplementary Table S2). The 115 features included the followings. First, the best BPS was searched for between Int^{-50} to Int^{-3} using the yUnAy motif [20]. The position weight matrix score as well as the conserved branch point "A" nucleotide were evaluated. Second, the length of PPT as well as the ratios of T, G, purines (A/G), and pyrimidines (C/T) in PPT were evaluated. As GGG trinucleotides are frequently recognized by splicing-suppressing hnRNP H and hnRNP K [21,22], the presence of GGG in PPT was evaluated. Third, we previously observed that nucleotides at Int^{-7}, Int^{-6}, Int^{-5}, and Int^{-3}, as well as Ex^{+2} and Ex^{+3}, play critical roles in splicing [11]. We included these nucleotides in our features. Fourth, SD-Score at the 5′ ss [23], MaxEntScan scores at the 3′ and 5′ ss's [24], and Shapiro Senapathy scores [25] at the 3′ and 5′ ss's were included as integrated measures to evaluate the strength of constitutive splicing *cis*-elements. Fifth, RBPs exert essential roles in both alternative and constitutive splicing events [26,27]. In our previous machine-learning model, IntSplice [12], to predict the splicing effects of intronic SNVs, we showed that the inclusion of RBP-biding sites markedly improved the performance. We thus included the sum scores of SpliceAid2 [28] of 71 RBPs in our features. As we could not

predict which specific feature best dictated the strength of splicing signals, we admitted multicollinearity of features. Spearman's rank correlation coefficients of all available pairs of 115 features are indicated in Supplementary Figure S1.

2.3. Machine-Learning Models

We generated machine-learning models with SVM (LinearSVC) [29], random forest (RandomForest) [30], and gradient boosting (LightGBM) [13]. For each model, we optimized hyperparameters using grid search. Feature importance was obtained from each modeling tools with default settings. We also eliminated features one by one using a method of meta-transformer for selecting features based on importance weights [31] by leave-one-out cross-validation (LOOCV). The performance of each model was evaluated by the area under the receiver operating characteristic curve (AUROC), the area under the precision recall curve (AUPRC), and seven statistical measures recommended by the Human Mutation Guidelines (see a legend of Table 1 for details) [32,33]. As we included all the identified splicing-affecting Fex-SNVs in our dataset, we did not create a separate test dataset. Instead, we employed leave-one-out or 10-fold cross-validation.

3. Results

3.1. Generation of Models with LinearSVC, Random Forest, and LightGBM

In this study, we generated machine-learning models to predict whether a Fex-SNV affecting the G nucleotide at the first nucleotide of an exon affects splicing or not. We first created a curated dataset of Fex-SNVs that comprised 106 splicing-affecting and 106 neutral Fex-SNVs (Supplementary Table S1). For each Fex-SNV, we extracted 115 features that dictated the strength of splicing *cis*-elements (Supplementary Table S2). We then generated three machine-learning models: LinearSVC [29], RandomForest [30], and LightGBM [13]. Each model was evaluated by AUROC and AUPRC (Figure 2), as well as seven statistical measures (accuracy, precision, recall/sensitivity, specificity, F1 score, NPV, and MCC) using 10-fold cross-validation (Table 1). LightGBM produced the highest AUROC and the highest scores in six out of the seven statistical measures except for specificity. The importance of 115 features by LightGBM were inspected using 10-fold cross-validation (Figure 3) and will be discussed in detail in the Discussion section.

We next eliminated features one-by-one from the three models using LOOCV (Supplementary Figure S2). Neither LinearSVC nor RandomForest reasonably improved the balanced accuracy by eliminating features. In contrast, the balanced accuracy was maximized at 15 features with LightGBM. Elimination of features from 115 to 15 increased the AUROC of LightGBM model from 0.84 ± 0.08 (mean and SD) to 0.86 ± 0.08 (Figure 2E,G and Table 1). Similarly, elimination of features increased in all the seven statistical measures of LightGBM model by approximately 2% (Table 1).

As expected, the feature importance values of the 15-feature-based LightGBM model using 10-fold cross-validation (Supplementary Figure S3) were similar to those of the 115-feature-based model using 10-fold cross-validation (Figure 3). We herein refer to the 15-feature-based LightGBM model as FexSplice.

3.2. Comparison of FexSplice with SpliceAI and CI-SpliceAI

SpliceAI [9] predicts the positions of ss's using the residual neural networks (ResNet) trained with a 10 Kbp segment annotated in the GTEx database. CI-SpliceAI [10] is based on the SpliceAI and retrained using a collapsed isoform set representative of all manually annotated constitutive and alternative splice sites in GENCODE. SpliceAI [9] and CI-SpliceAI [10] are also able to predict the splicing effects of Fex-SNVs. We calculated the AUROC, the AUPRC, and seven statistical measures of SpliceAI and CI-SpliceAI with our dataset (Supplementary Table S3). FexSplice was trained with our dataset, whereas SpliceAI and CI-SpliceAI were not. Thus, statistical measures of SpliceAI and CI-SpliceAI cannot be unbiasedly compared with those of FexSplice. Nevertheless, precision and specificity were better in SpliceAI and CI-SpliceAI compared to those in FexSplice. This was at the cost of a

much lower recall value of 0.22 in both SpliceAI and CI-SpliceAI compared to 0.78 ± 0.13 (mean and SD) in FexSplice. As SpliceAI and CI-SpliceAI were developed to identify ss's in a large number of candidates in the whole genome, they were likely to be designed to reduce false positives. This may account for high precision and specificity values with low recall values in SpliceAI and CI-SpliceAI.

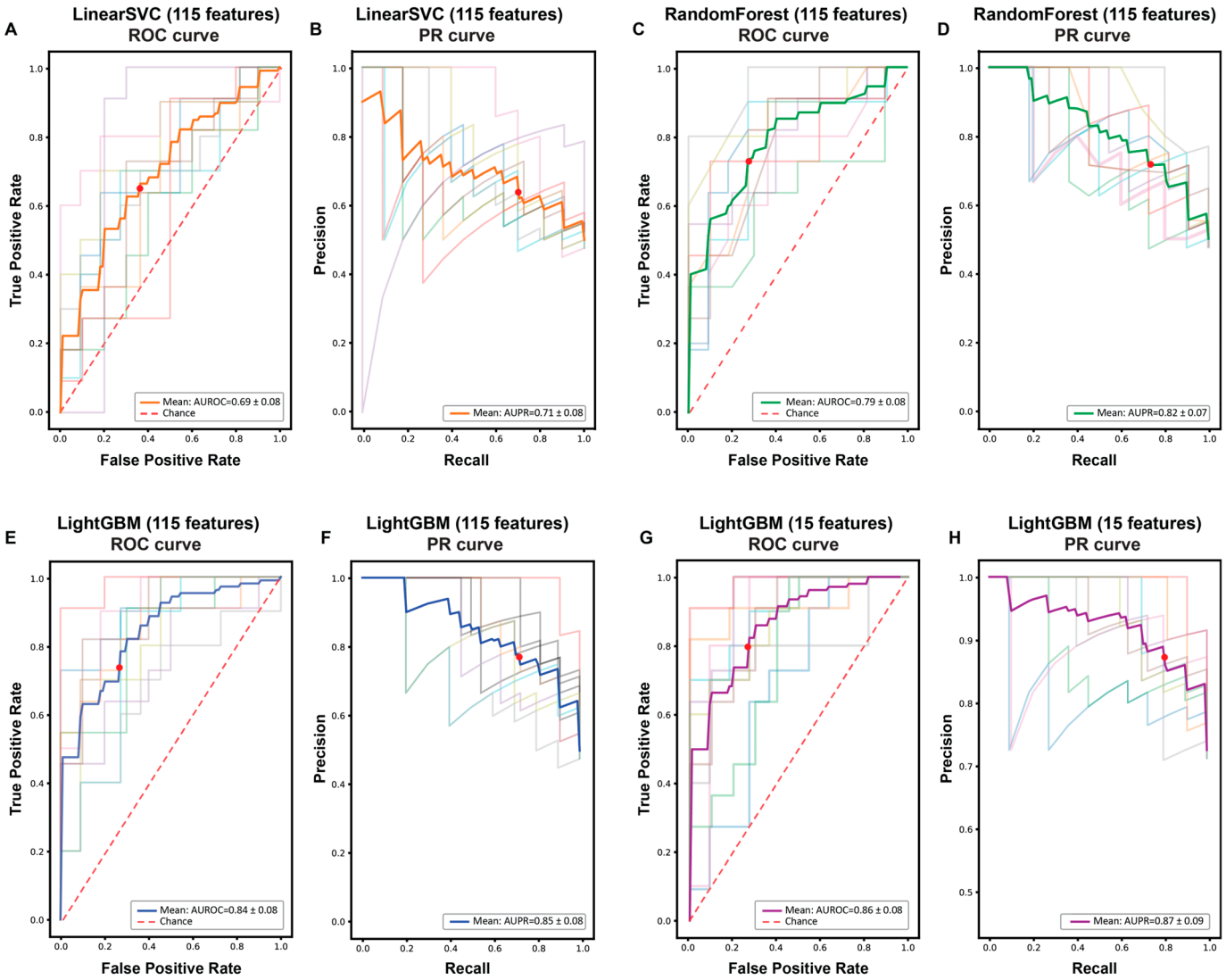


Figure 2. Receiver operating characteristics (ROC) (**A**,**C**,**E**,**G**) and precision recall (PR) (**B**,**D**,**F**,**H**) curves of LinearSVC (**A**,**B**), RandomForest (**C**,**D**), and LightGBM (**E**,**F**) models with 115 features, as well as a LightGBM (**G**,**H**) model with 15 features, all using 10-fold cross-validation. Thin lines represent each of the 10-fold validations, and thick lines represent 10-fold cross-validation. Red dots indicate where the threshold of pathogenic probability is set to 0.5.

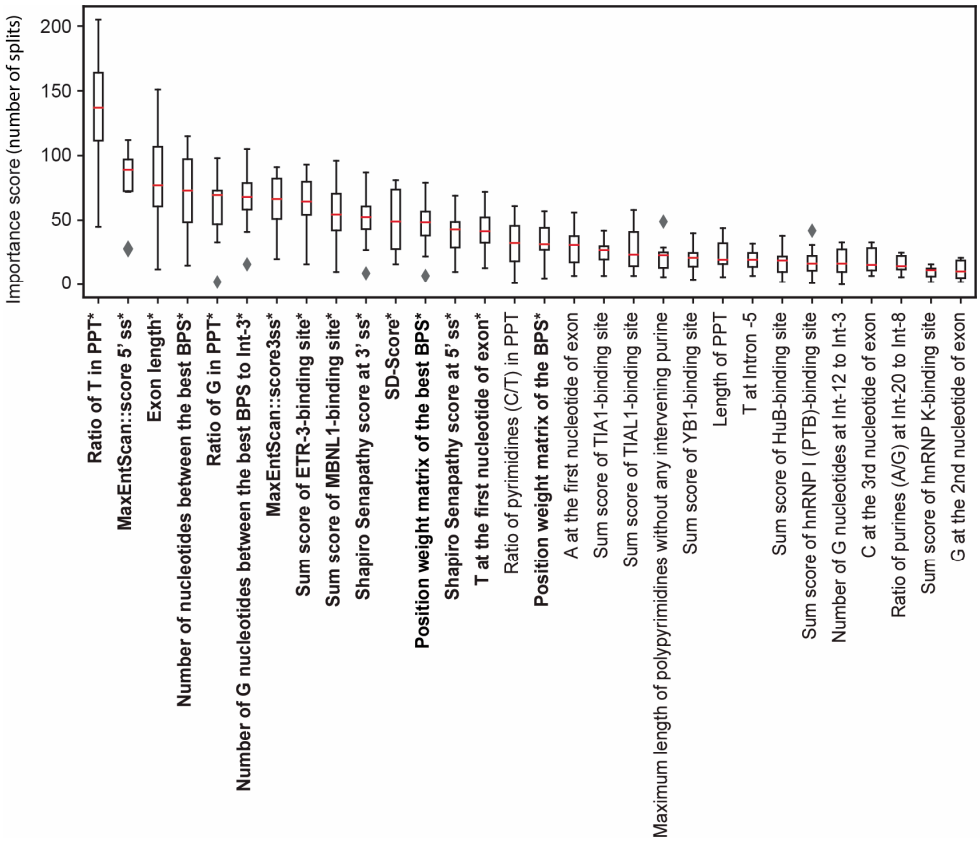


Figure 3. The top 30 features, ranked by their importance, are displayed along with associated median and interquartile range values. This ranking is derived from the feature importance analysis using 10-fold cross-validation of a LightGBM model trained with 115 features. Bold letters with an asterisk indicate 15 features that maximized the AUROC in recursive feature elimination (Supplementary Figure S2), which were used to generate FexSplice.

Table 1. Comparison of nine statistical measures using 10-fold cross-validation of LinearSVC, RandomForest, and LightGBM models with 115 features, as well as a LightGBM model with 15 features.

Model	LinearSVC (115)	Random Forest (115)	LightGBM (115)	LightGBM (15)
Accuracy [1]	0.64 ± 0.10	0.71 ± 0.07	0.75 ± 0.09	0.77 ± 0.07
Precision [2]	0.64 ± 0.09	0.71 ± 0.07	0.77 ± 0.11	0.80 ± 0.12
Recall [3]	0.65 ± 0.15	0.73 ± 0.12	0.74 ± 0.14	0.78 ± 0.13
Specificity [4]	0.63 ± 0.11	0.70 ± 0.08	0.78 ± 0.13	0.77 ± 0.15
F1 score [5]	0.64 ± 0.11	0.71 ± 0.08	0.75 ± 0.10	0.77 ± 0.07
NPV [6]	0.65 ± 0.11	0.73 ± 0.11	0.76 ± 0.11	0.79 ± 0.11
MCC [7]	0.29 ± 0.19	0.43 ± 0.15	0.52 ± 0.18	0.57 ± 0.15
AUROC	0.69 ± 0.08	0.79 ± 0.08	0.84 ± 0.08	0.86 ± 0.08
AUPRC	0.71 ± 0.08	0.82 ± 0.07	0.85 ± 0.08	0.87 ± 0.09

The number of features is indicated in parentheses. Mean and SD are indicated. [1] Accuracy, overall correctness of the classifier: Accuracy = (TP + TN)/(TP + TN + FP + FN); [2] Precision (positive predictive value), correctness of positive predictions: Precision = TP/(TP + FP); [3] Recall (sensitivity or true positive rate), classifier's ability to identify positive instances: Recall = TP/(TP + FN); [4] Specificity (true negative rate), classifier's ability to identify negative instances: Specificity = TN/(TN + FP); [5] F1 Score, balanced metric considering false positives and negatives: F1 Score = 2 * (Precision * Recall)/(Precision + Recall); [6] NPV (negative predictive value), correctness of negative predictions: NPV = TN/(TN + FN); [7] MCC (Matthews correlation coefficient), balanced measure considering all values in the confusion matrix: MCC = (TP * TN − FP * FN)/((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN))1/2. TP (true positive) and FN (false negative) are the numbers of splicing-affecting Fex-SNVs that were predicted to be splicing-affecting and neutral, respectively. FP (false positive) and TN (true negative) are the numbers of neutral Fex-SNVs that were predicted to be splicing-affecting and neutral, respectively.

3.3. Web Service of FexSplice

We developed a web service program, FexSplice, (https://www.med.nagoya-u.ac.jp/neurogenetics/FexSplice) (accessed on 1 August 2023) (Figure 4). The FexSplice web service accepts a genomic coordinate in either GRCh37/hg19 or GRCh38/hg38 and maps it to all the annotated coding transcripts in Ensembl release 101. FexSplice analyzes all the transcripts and generates three possible Fex-SNVs at the given coordinate. LightGBM automatically generates a probability score for each Fex-SNV with 0.5 being the threshold. The default threshold of 0.5 by LightGBM was used in FexSplice. Fex-SNVs with a probability less than 0.5 are predicted to be splicing-insensitive, while those with a probability of 0.5 or more are predicted to be splicing-affecting. When two or more transcripts exist at Fex-SNV, FexSplice predicts the effects of splicing for all the relevant transcripts. Pre-processed genome-wide FexSplice dataset was generated on GRCh37/hg19, and was converted to the GRCh38/hg38 version using LiftOver [34], both of which are downloadable from the FexSplice web site.

Predicted pathogenicity in shown in "Prediction" along with "Pathogenic Probability". Pathogenic Probability ≥ 0.5 is predicted to be "Pathogenic".

Prediction	Pathogenic Probability	Genomic Mutation	Strand	Gene name, ID and exon No. based on Ensembl release 101
Abnormal	0.999	g.57554424C>A	-	FECH::ENSG00000066926::ENST00000262093::exon9::ENSE00003632198IIFECH::ENSG00000066926::ENST00000382873::exon9::ENSE00003632198IIFECH::ENSG00000066926::ENST00000585494::exon9::ENSE00003502974IIFECH::ENSG00000066926::ENST00000591977::exon3::ENSE00003632198
Abnormal	0.998	g.57554424C>G	-	FECH::ENSG00000066926::ENST00000262093::exon9::ENSE00003632198IIFECH::ENSG00000066926::ENST00000382873::exon9::ENSE00003632198IIFECH::ENSG00000066926::ENST00000585494::exon9::ENSE00003502974IIFECH::ENSG00000066926::ENST00000591977::exon3::ENSE00003632198
Abnormal	0.995	g.57554424C>T	-	FECH::ENSG00000066926::ENST00000262093::exon9::ENSE00003632198IIFECH::ENSG00000066926::ENST00000382873::exon9::ENSE00003632198IIFECH::ENSG00000066926::ENST00000585494::exon9::ENSE00003502974IIFECH::ENSG00000066926::ENST00000591977::exon3::ENSE00003632198

Input queries: hg38, 18, 57554424

Figure 4. An example output of the FexSplice web service (https://www.med.nagoya-u.ac.jp/neurogenetics/FexSplice, accessed on 1 August 2023). G.57554424C>A on chromosome 18 (GRCh38/hg38) in *FECH* was previously reported to cause aberrant splicing [6]. The chromosome number and genomic coordinate were entered into the FexSplice web service. Predicted pathogenicity (abnormal in red letters and normal in black letters) and its probability were returned for three possible Fex-SNVs. Pre-processed genome-wide FexSplice datasets on GRCh37/hg19 and GRCh38/hg38 are also available. For g.57554424C>A, SpliceAI predicted a moderate effect on acceptor loss (Δ score = 0.45) and CI-SpliceAI predicted a minor effect on acceptor loss (Δ score = 0.24).

4. Discussion

Our study aimed to develop a model to predict the splicing effect of Fex-SNVs. We scrutinized available articles and curated a dataset that comprised 106 splicing-affecting and 106 neutral Fex-SNVs (Supplementary Table S1). For each Fex-SNV, 115 features dictating the strength of splicing signals were extracted (Supplementary Table S2). Evaluation of the discrimination models by LinearSVC, RandomForest, and LightGBM using 10-fold cross-validation showed that LightGBM produced the highest AUROC, the highest AUPRC, and the highest scores in six out of the seven statistical measures (Table 1). Elimination of the least important feature one-by-one using cross-validation showed that the performance of LightGBM models became the best with 15 features (Supplementary Figure S2).

We evaluated the importance of 115 features (Figure 3) and 15 features (Supplementary Figure S2) both using 10-fold cross-validation and found that highly ranked features were similar between the two models. As our features had multicollinearity (Supplementary Figure S1), high feature importance did not exclusively represent essential features. Nevertheless, the following features were critical. First, among the 115 features (Figure 3), the ratio of T nucleotides in PPT was ranked first and its importance was markedly higher than the other features. The preference of T over C in PPT was previously reported [35,36]. Similarly, the ratio of G nucleotides in PPT was ranked fifth. A more deleterious effect of G than A in PPT on binding to U2AF65 was also previously reported [37]. Additionally,

three other features for PPT and four features for BPS are included in the top 30 features. The importance of PPT in the discrimination models is in accordance with the notion that the AG-dependent 3′ ss's are vulnerable to Fex-SNV. Second, MaxEntScan::5′ss [24], SD-score [23], and Shapiro Senapathy score at 5′ ss [25], all of which represented the splicing signals at the 5′ ss, were ranked second, eleventh, and twelfth, respectively. Unexpectedly, MaxEntScan::5′ss had a higher importance than MaxEntScan::3′ss, which was ranked seventh. The importance of the splicing signals at the 5′ ss is likely to support the exon-recognition model, in which an exon not an intron is recognized as a single unit in pre-mRNA splicing [38]. Third, eight of the top 30 features were for the presence of RBP-binding sites. RBPs exert essential roles in both alternative and constitutive splicing events [26,27]. As indicated in Section 2.2, we previously showed that the inclusion of RBP-biding sites markedly improved the performance of IntSplice, a tool to predict the splicing effects of intronic SNVs [12]. Among the eight RBPs, ETR-3 (CELF2) and MBNL1 were ranked eighth and tenth, respectively. Abnormal downregulation of MBNL and upregulation ETR-3 are hallmarks of myotonic dystrophy, and their effects on pre-mRNA splicing have been extensively studied [39]. However, myotonic dystrophy was not included in either the title or the abstract of any article showing splicing-affecting Fex-SNVs (Supplementary Table S1a). In addition, ETR-3-binding sequences according to SpliceAid2 were observed in 18 out of 106 splicing-affecting and 21 out of 106 neutral Fex-SNVs (p-value = 0.72 by Fisher's exact test). Similarly, MBNL1-binding sequences were observed in 21 out of 106 splicing-affecting and 27 out of 106 neutral Fex-SNVs (p-value = 0.41). Thus, the high feature importance values of ETR-3 and MBNL1 were unlikely to be accounted for by reporting bias of splicing-affecting Fex-SNVs. Although the binding of hnRNP A1 was not included in the top 30 features, hnRNP A1 directly binds to the 3′ ss of *SMN2* exon 7 and suppresses its splicing [40]. However, RBPs are unlikely to bind to the 3′ ss where core spliceosomal components assemble. Thus, the presence of binding sites for RBPs is likely to represent that the splicing signals on and around the exon are weak and that the binding of RBP(s) is required for the exon recognition. Fourth, exonic features such as the exon length and the first-to-third exonic nucleotides played essential roles. We unexpectedly observed that out of the 12 exonic and 12 intronic nucleotides in the 115 features (Supplementary Table S2), four exonic nucleotides (T at Ex^{+1}, A at Ex^{+1}, C at Ex^{+3}, and G at Ex^{+2}) were included in the top 30 features, whereas only one intronic nucleotide (T at Int^{-5}) was included. Aberrant splicing due to T at Ex^{+1} rather than A at Ex^{+1} was previously reported [41]. Crystal structure of U2AF1 (U2AF35) bound to the 3′ ss showed that a nucleotide at Ex^{+2} was not strictly recognized by U2AF1 and a nucleotide at Ex^{+3} was not bound by U2AF1 [7]. Nevertheless, C at Ex^{+3} and G at Ex^{+2} were included in the top 30 features. We previously showed that G at Int^{-3} was markedly detrimental for pre-mRNA splicing, and A at Int^{-3} followed [11]. However, neither nucleotide was included in the top 30 features, which was likely to be masked by multicollinearity of 115 features.

Comparison of FexSplice with SpliceAI and CI-SpliceAI showed that FexSplice outperformed the others in seven out of the nine statistical measures, although FexSplice should be biased by overfitting to our dataset compared to the others. To fairly compare the performance of different tools, models should be generated by an identical training dataset and evaluated by an identical testing dataset, as we previously performed for InMeRF, a tool for predicting the pathogenicity of missense SNVs [42]. We, however, did not recapitulate the generation of models with SpliceAI and CI-SpliceAI. We suppose that the splicing effects of Fex-SNVs have been underestimated in identifying pathogenic variants in human diseases. We hope that FexSplice will help disclose yet unidentified splicing effects of Fex-SNVs, and also understand the physiological mechanisms of the recognition of the 3′ ss's.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes14091765/s1, Figure S1: Heatmap of Spearman's correlation coefficients of 115 feature values to indicate multicollinearity of features; Figure S2: Feature elimination of LinearSVC, RandomForst, and LightGBM models. Figure S3: Feature importance of 15 feature-based

LightGBM models using 10-fold cross-validation; Table S1a: 106 splicing-affecting Fex-SNVs; Table S1b: 106 neutral Fex-SNVs; Table S2. 115 features to dictate the strength of splicing *cis*-elements; Table S3. Comparison of nine statistical measures of FexSplice, SpliceAI, and CI-SpliceAI.

Author Contributions: Conceptualization, K.O.; methodology, A.J., J.-i.T., A.M., R.O. and K.F.; writing—original draft preparation, A.J. and J.-i.T.; writing—review and editing, K.F. and K.O.; supervision, K.O. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by Grants-in-Aid from the Japan Agency for Medical Research and Development (JP22ek0109488 to K.O.), the Japan Society for the Promotion of Science (JP23K18273 to K.O., JP23H02794 to K.O., JP21H02476 to A.M., and JP22K19269 to A.M.), the Ministry of Health, Labour and Welfare of Japan (23FC1014 to K.O.), and the National Center of Neurology and Psychiatry (5–6 to K.O.). A.J. receives scholarship from the THERS Interdisciplinary Frontier Next Generation Researcher Project (JST SPRING, Grant Number JPMJSP2125).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request to the corresponding author.

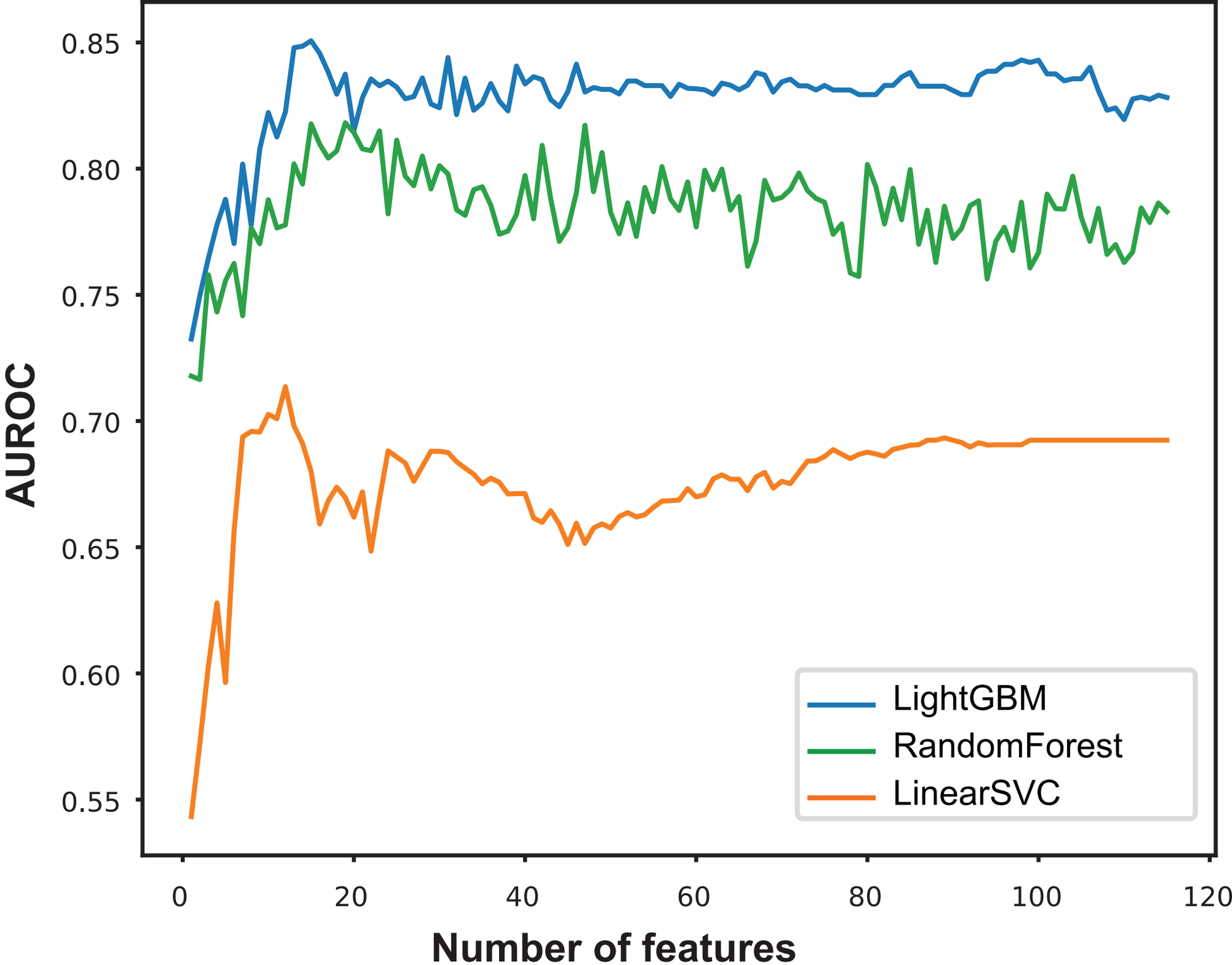
Conflicts of Interest: The authors declare no conflict of interest.

References

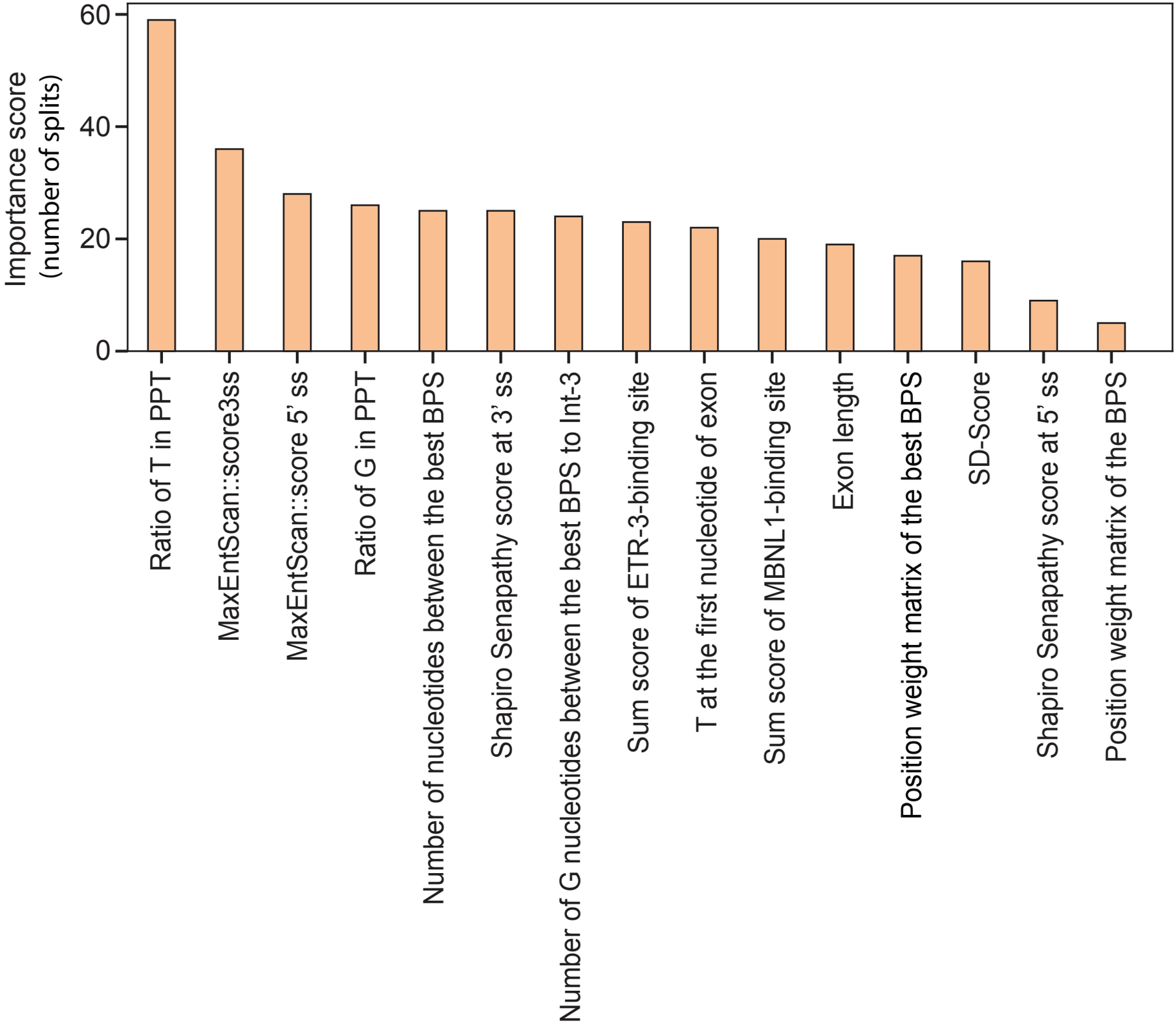
1. Maroney, P.A.; Yu, Y.T.; Jankowska, M.; Nilsen, T.W. Direct Analysis of Nematode Cis- and Trans-Spliceosomes: A Functional Role for U5 SnRNA in Spliced Leader Addition Trans-Splicing and the Identification of Novel Sm SnRNPs. *RNA* **1996**, *2*, 735–745.
2. Ule, J.; Stefani, G.; Mele, A.; Ruggiu, M.; Wang, X.; Taneri, B.; Gaasterland, T.; Blencowe, B.J.; Darnell, R.B. An RNA Map Predicting Nova-Dependent Splicing Regulation. *Nature* **2006**, *444*, 580–586. [CrossRef]
3. Singh, G.; Cooper, T.A. Minigene Reporter for Identification and Analysis of Cis Elements and Trans Factors Affecting Pre-MRNA Splicing. *BioTechniques* **2006**, *41*, 177–181. [CrossRef]
4. Ohno, K.; Ohkawara, B.; Shen, X.-M.; Selcen, D.; Engel, A.G. Clinical and Pathologic Features of Congenital Myasthenic Syndromes Caused by 35 Genes-A Comprehensive Review. *Int. J. Mol. Sci.* **2023**, *24*, 3730. [CrossRef] [PubMed]
5. Guth, S.; Martínez, C.; Gaur, R.K.; Valcárcel, J. Evidence for Substrate-Specific Requirement of the Splicing Factor U2AF(35) and for Its Function after Polypyrimidine Tract Recognition by U2AF(65). *Mol. Cell Biol.* **1999**, *19*, 8263–8271. [CrossRef] [PubMed]
6. 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.* **2011**, *39*, 4396–4404. [CrossRef]
7. Yoshida, H.; Park, S.-Y.; Sakashita, G.; Nariai, Y.; Kuwasako, K.; Muto, Y.; Urano, T.; Obayashi, E. Elucidation of the Aberrant 3′ Splice Site Selection by Cancer-Associated Mutations on the U2AF1. *Nat. Commun.* **2020**, *11*, 4744. [CrossRef]
8. Ohno, K.; Takeda, J.-I.; Masuda, A. Rules and Tools to Predict the Splicing Effects of Exonic and Intronic Mutations. *Wiley Interdiscip. Rev. RNA* **2018**, *9*, e1451. [CrossRef]
9. Jaganathan, K.; Kyriazopoulou Panagiotopoulou, S.; McRae, J.F.; Darbandi, S.F.; Knowles, D.; Li, Y.I.; Kosmicki, J.A.; Arbelaez, J.; Cui, W.; Schwartz, G.B.; et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **2019**, *176*, 535–548.e24. [CrossRef] [PubMed]
10. Strauch, Y.; Lord, J.; Niranjan, M.; Baralle, D. CI-SpliceAI-Improving Machine Learning Predictions of Disease Causing Splicing Variants Using Curated Alternative Splice Sites. *PLoS ONE* **2022**, *17*, e0269159. [CrossRef]
11. Shibata, A.; Okuno, T.; Rahman, M.A.; Azuma, Y.; Takeda, J.-I.; Masuda, A.; Selcen, D.; Engel, A.G.; Ohno, K. IntSplice: Prediction of the Splicing Consequences of Intronic Single-Nucleotide Variations in the Human Genome. *J. Hum. Genet.* **2016**, *61*, 633–640. [CrossRef] [PubMed]
12. Takeda, J.-I.; Fukami, S.; Tamura, A.; Shibata, A.; Ohno, K. IntSplice2: Prediction of the Splicing Effects of Intronic Single-Nucleotide Variants Using LightGBM Modeling. *Front. Genet.* **2021**, *12*, 701076. [CrossRef]
13. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.* **2017**, 3149–3157.
14. Stenson, P.D.; Mort, M.; Ball, E.V.; Evans, K.; Hayden, M.; Heywood, S.; Hussain, M.; Phillips, A.D.; Cooper, D.N. The Human Gene Mutation Database: Towards a Comprehensive Repository of Inherited Mutation Data for Medical Research, Genetic Diagnosis and next-Generation Sequencing Studies. *Hum. Genet.* **2017**, *136*, 665–677. [CrossRef] [PubMed]
15. Landrum, M.J.; Lee, J.M.; Benson, M.; Brown, G.R.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Jang, W.; et al. ClinVar: Improving Access to Variant Interpretations and Supporting Evidence. *Nucleic Acids Res.* **2018**, *46*, D1062–D1067. [CrossRef]
16. Li, K.; Luo, T.; Zhu, Y.; Huang, Y.; Wang, A.; Zhang, D.; Dong, L.; Wang, Y.; Wang, R.; Tang, D.; et al. Performance Evaluation of Differential Splicing Analysis Methods and Splicing Analytics Platform Construction. *Nucleic Acids Res.* **2022**, *50*, 9115–9126. [CrossRef]

17. Howe, K.L.; Achuthan, P.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; Armean, I.M.; Azov, A.G.; Bennett, R.; Bhai, J.; et al. Ensembl 2021. *Nucleic Acids Res.* **2021**, *49*, D884–D891. [CrossRef] [PubMed]
18. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. *Nucleic Acids Res.* **2016**, *44*, D733–D745. [CrossRef] [PubMed]
19. Sherry, S.T.; Ward, M.H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E.M.; Sirotkin, K. DbSNP: The NCBI Database of Genetic Variation. *Nucleic Acids Res.* **2001**, *29*, 308–311. [CrossRef]
20. Gao, K.; Masuda, A.; Matsuura, T.; Ohno, K. Human Branch Point Consensus Sequence Is YUnAy. *Nucleic Acids Res.* **2008**, *36*, 2257–2267. [CrossRef]
21. Huelga, S.C.; Vu, A.Q.; Arnold, J.D.; Liang, T.Y.; Liu, P.P.; Yan, B.Y.; Donohue, J.P.; Shiue, L.; Hoon, S.; Brenner, S.; et al. Integrative Genome-Wide Analysis Reveals Cooperative Regulation of Alternative Splicing by HnRNP Proteins. *Cell Rep.* **2012**, *1*, 167–178. [CrossRef]
22. Oberg, D.; Fay, J.; Lambkin, H.; Schwartz, S. A Downstream Polyadenylation Element in Human Papillomavirus Type 16 L2 Encodes Multiple GGG Motifs and Interacts with HnRNP H. *J. Virol.* **2005**, *79*, 9254–9269. [CrossRef]
23. Sahashi, K.; Masuda, A.; Matsuura, T.; Shinmi, J.; Zhang, Z.; Takeshima, Y.; Matsuo, M.; Sobue, G.; Ohno, K. In Vitro and in Silico Analysis Reveals an Efficient Algorithm to Predict the Splicing Consequences of Mutations at the 5′ Splice Sites. *Nucleic Acids Res.* **2007**, *35*, 5995–6003. [CrossRef] [PubMed]
24. Yeo, G.; Burge, C.B. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J. Comput. Biol.* **2004**, *11*, 377–394. [CrossRef] [PubMed]
25. Shapiro, M.B.; Senapathy, P. RNA Splice Junctions of Different Classes of Eukaryotes: Sequence Statistics and Functional Implications in Gene Expression. *Nucleic Acids Res.* **1987**, *15*, 7155–7174. [CrossRef]
26. Van Nostrand, E.L.; Freese, P.; Pratt, G.A.; Wang, X.; Wei, X.; Xiao, R.; Blue, S.M.; Chen, J.Y.; Cody, N.A.L.; Dominguez, D.; et al. A Large-Scale Binding and Functional Map of Human RNA-Binding Proteins. *Nature* **2020**, *583*, 711–719. [CrossRef] [PubMed]
27. Kawachi, T.; Masuda, A.; Yamashita, Y.; Takeda, J.I.; Ohkawara, B.; Ito, M.; Ohno, K. Regulated Splicing of Large Exons Is Linked to Phase-Separation of Vertebrate Transcription Factors. *EMBO J.* **2021**, *40*, e107485. [CrossRef]
28. 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.* **2012**, *33*, 81–85. [CrossRef]
29. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
30. Breiman, L. *Random Forests*; Springer Science and Business Media LLC: Berlin, Germany, 2001. [CrossRef]
31. Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; et al. API Design for Machine Learning Software: Experiences from the Scikit-Learn Project 2013. *arXiv* **2013**, arXiv:1309.0238.
32. Vihinen, M. Guidelines for Reporting and Using Prediction Tools for Genetic Variation Analysis. *Hum. Mutat.* **2013**, *34*, 275–282. [CrossRef]
33. Grimm, D.G.; Azencott, C.-A.; Aicheler, F.; Gieraths, U.; MacArthur, D.G.; Samocha, K.E.; Cooper, D.N.; Stenson, P.D.; Daly, M.J.; Smoller, J.W.; et al. The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Hum. Mutat.* **2015**, *36*, 513–523. [CrossRef]
34. Hinrichs, A.S.; Karolchik, D.; Baertsch, R.; Barber, G.P.; Bejerano, G.; Clawson, H.; Diekhans, M.; Furey, T.S.; Harte, R.A.; Hsu, F.; et al. The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Res.* **2006**, *34*, D590–D598. [CrossRef]
35. Coolidge, C.J.; Seely, R.J.; Patton, J.G. Functional Analysis of the Polypyrimidine Tract in Pre-MRNA Splicing. *Nucleic Acids Res.* **1997**, *25*, 888–896. [CrossRef] [PubMed]
36. Mullen, M.P.; Smith, C.W.; Patton, J.G.; Nadal-Ginard, B. Alpha-Tropomyosin Mutually Exclusive Exon Selection: Competition between Branchpoint/Polypyrimidine Tracts Determines Default Exon Choice. *Genes Dev.* **1991**, *5*, 642–655. [CrossRef] [PubMed]
37. Singh, R.; Banerjee, H.; Green, M.R. Differential Recognition of the Polypyrimidine-Tract by the General Splicing Factor U2AF65 and the Splicing Repressor Sex-Lethal. *RNA* **2000**, *6*, 901–911. [CrossRef]
38. Tilgner, H.; Nikolaou, C.; Althammer, S.; Sammeth, M.; Beato, M.; Valcárcel, J.; Guigó, R. Nucleosome Positioning as a Determinant of Exon Recognition. *Nat. Struct. Mol. Biol.* **2009**, *16*, 996–1001. [CrossRef]
39. Lopez-Martinez, A.; Soblechero-Martin, P.; de-la-Puente-Ovejero, L.; Nogales-Gadea, G.; Arechavala-Gomeza, V. An Overview of Alternative Splicing Defects Implicated in Myotonic Dystrophy Type I. *Genes* **2020**, *11*, 1109. [CrossRef]
40. Doktor, T.K.; Schroeder, L.D.; Vested, A.; Palmfeldt, J.; Andersen, H.S.; Gregersen, N.; Andresen, B.S. SMN2 Exon 7 Splicing Is Inhibited by Binding of HnRNP A1 to a Common ESS Motif That Spans the 3′ Splice Site. *Hum. Mutat.* **2011**, *32*, 220–230. [CrossRef]
41. Grodecká, L.; Lockerová, P.; Ravčuková, B.; Buratti, E.; Baralle, F.E.; Dušek, L.; Freiberger, T. Exon First Nucleotide Mutations in Splicing: Evaluation of in Silico Prediction Tools. *PLoS ONE* **2014**, *9*, e89570. [CrossRef] [PubMed]
42. Takeda, J.-I.; Nanatsue, K.; Yamagishi, R.; Ito, M.; Haga, N.; Hirata, H.; Ogi, T.; Ohno, K. InMeRF: Prediction of Pathogenicity of Missense Variants by Individual Modeling for Each Amino Acid Substitution. *NAR Genom. Bioinform.* **2020**, *2*, lqaa038. [CrossRef] [PubMed]

Supplementary Figure S1. Heatmap of Spearman's correlation coefficients of 115 feature values to indicate multiple collinearity of features.



Supplementary Figure S2. Feature elimination of LinearSVC, RandomForst, and LightGBM models. Features were eliminated one-by-one by LOOCV using SelectFromModel. Area-under-the-receiver-operating-characteristics-curve (AUROC) is plotted.



Supplementary Figure S3. Feature importance of 15 feature-based LightGBM models by 10-fold cross-validation. The feature importance is represented by the number of times each feature is used to split the data across all trees in the ensemble. Features with higher importance values have a more significant impact on the model.

Supplementary Table S1b. 106 neutral Fex-SNVs

Chrom	GRCh37/hg19	Strand	Ref	Alt	Global MAF	dbSNP	Reference
chr1	9777122	+	G	A	0.032	rs28730668	[1]
chr1	10190773	+	G	A	0.040	rs17034499	
chr1	23997401	+	G	A	0.010	rs116436791	
chr1	26526375	+	G	A	0.007	rs144835190	
chr1	26664968	-	C	T	0.286	rs11247919	
chr1	27428621	-	C	T	0.008	rs139314838	
chr1	36945681	-	C	A	0.011	rs3917923	
chr1	44679335	+	G	A	0.008	rs145692550	
chr1	107946262	+	G	A	0.418	rs12126267	
chr1	192606720	+	G	T	0.020	rs78203847	
chr1	207512742	+	G	C			[1]
chr2	2309927	-	C	A	0.184	rs7580422	
chr2	81008127	-	C	A	0.012	rs116579007	
chr2	186903159	-	C	A	0.054	rs10196464	
chr2	225706613	-	C	T	0.023	rs144638149	
chr3	8719039	-	C	G	0.052	rs113882955	
chr3	109047928	-	C	T	0.156	rs1163439	
chr3	121155122	-	C	T	0.007	rs41540016	
chr3	132416206	-	C	T	0.016	rs77533254	
chr3	155523512	-	C	T	0.053	rs73003554	
chr3	197701913	+	G	T	0.210	rs7373165	
chr4	22348446	-	C	A	0.013	rs10029265	
chr4	38570708	+	G	A	0.216	rs7654396	
chr4	68534392	-	C	T	0.155	rs10010188	
chr4	113468564	-	C	T	0.019	rs78336146	
chr4	121652945	-	C	A	0.028	rs76650962	
chr4	159158677	+	G	A	0.078	rs12504074	
chr4	166606936	+	G	A	0.071	rs80303176	
chr4	185300283	-	C	G	0.009	rs116945722	
chr5	5059677	+	G	A	0.068	rs13436478	
chr5	68398889	+	G	T	0.030	rs28450427	
chr5	72121562	+	G	T	0.013	rs10075937	
chr5	149826526	-	C	T	0.268	rs4841	
chr6	29975095	+	G	C	0.050	rs6940082	
chr6	30384283	+	G	A	0.006	rs11756914	
chr6	32370993	-	C	T	0.027	rs115653647	
chr6	32634447	-	C	G	0.066	rs3210146	
chr6	33048466	+	G	T	0.271	rs1126511	
chr6	49580247	-	C	T	0.070	rs16879498	[1]
chr6	65612113	-	C	T	0.026	rs77020971	
chr6	109621434	+	G	T	0.034	rs78351760	
chr6	139197609	+	G	A	0.276	rs1529151	
chr7	5880392	+	G	C	0.158	rs2240405	

chr7	48643310	+	G	A	0.229	rs7803560	
chr7	72178756	-	C	A	0.126	rs146095374	
chr7	102723513	+	G	A	0.017	rs28482805	
chr7	141540868	-	C	A	0.015	rs73525255	
chr8	42849577	+	G	A	0.005	rs191545829	
chr9	4833154	+	G	A	0.016	rs34215378	
chr9	21813765	+	G	A	0.047	rs10965146	
chr9	34257979	-	C	T	0.048	rs76547823	
chr9	138440554	+	G	C	0.174	rs55695858	
chr9	139110654	-	C	T	0.153	rs12684650	
chr10	15153685	+	G	A	0.006	rs112358622	
chr10	43191254	-	C	G	0.225	rs7088389	
chr10	98006805	-	C	T	0.037	rs11540858	
chr11	2190925	-	C	T	0.016	rs76240471	
chr11	16824599	-	C	T	0.006	rs34556458	
chr11	17548878	-	C	T	0.013	rs55843567	
chr11	31703335	+	G	C	0.020	rs141543586	
chr11	34916657	-	C	T	0.115	rs61734605	
chr11	82985806	-	C	A	0.013	rs115690042	
chr12	6836837	+	G	T	0.063	rs66951403	
chr12	7030660	+	G	A	0.030	rs115894573	
chr12	108152052	+	G	A	0.006	rs375180436	
chr12	121176083	+	G	A	0.182	rs1799958	
chr12	133272470	+	G	T	0.057	rs74727297	
chr13	32954144	+	G	A			[1]
chr13	43463378	-	C	T	0.028	rs79111014	
chr13	52080679	+	G	A	0.462	rs9526773	
chr13	95859035	-	C	A	0.151	rs2274407	
chr15	21137651	-	C	T	0.028	rs925312	
chr15	52510884	-	C	G	0.021	rs73404874	
chr15	75654366	-	C	G	0.045	rs3803464	
chr15	94841430	+	G	A	0.175	rs79161311	
chr16	840531	+	G	A	0.008	rs114122804	
chr16	11594825	-	C	T	0.020	rs71383272	
chr16	19710824	+	G	A	0.008	rs34342607	
chr16	22865985	+	G	A	0.009	rs142571042	
chr17	11835331	+	G	A	0.210	rs17612861	
chr17	25999628	-	C	T	0.249	rs7220339	
chr17	41290674	+	G	C	0.353	rs2292595	
chr17	58263012	-	C	T	0.007	rs77024136	
chr17	61198167	-	C	T	0.455	rs28534579	
chr17	72548117	-	C	T	0.031	rs1699569	
chr17	73235099	-	C	T	0.007	rs76982026	
chr17	73588058	+	G	T	0.237	rs736522	
chr17	79784391	-	C	T	0.200	rs140438341	
chr18	1626240	+	G	T	0.013	rs116114268	

chr18	25727748	-	C	T	0.009	rs17495042	
chr18	74599977	-	C	T	0.154	rs17060015	
chr18	77704680	-	C	T	0.019	rs111363895	
chr19	373531	-	C	T	0.037	rs73489977	
chr19	14527217	-	C	G	0.014	rs185725199	
chr19	33444707	-	C	T	0.398	rs10411735	
chr19	45448695	+	G	A	0.033	rs10425530	
chr19	48235322	+	G	A	0.008	rs368031838	
chr19	51411751	-	C	T	0.019	rs34626614	
chr19	54556228	-	C	G	0.162	rs77931596	
chr19	55349031	+	G	A	0.124	rs58731871	
chr20	47273726	-	C	T	0.160	rs55904123	
chr20	47795781	-	C	T	0.015	rs148170165	
chr22	17588617	+	G	C			[1]
chr22	20961109	-	C	T	0.203	rs13340098	
chr22	20976088	-	C	A	0.022	rs78003513	
chrX	77286898	+	G	A			[1]

Reference

1 Grodecká, L.; Lockerová, P.; Ravčuková, B.; Buratti, E.; Baralle, F.E.; Dušek, L.; Freiberger, T. E

xon first nucleotide mutations in splicing: evaluation of in silico prediction tools. *PLoS One* **2014**, 9, e89570,

Supplementary Table S2. 115 features to dictate the strength of splicing *cis*-elements

Features	3'/Ex/5'	Position[a]
Best-BPS[b]		
Number of nucleotides between the best BPS to Int-3	3'	Int-50 toInt-3
Number of G nucleotides between the best BPS to Int-3	3'	Int-50 to Int-3
Position weight matrix of the best BPS	3'	Int-50 to Int-3
PPT		
Maximum length of polypyrimidines without any intervening purine	3'	Int-50 to Int-3
Best-BPS-PPT[c]		
Position weight matrix of the BPS	3'	Int-50 to Int-3
Branch point is A at the best BPS	3'	Int-50 to Int-3
Ratio of pyrimidines (C/T) in PPT	3'	Int-50 to Int-3
Ratio of T in PPT	3'	Int-50 to Int-3
Ratio of G in PPT	3'	Int-50 to Int-3
Length of PPT	3'	Int-50 to Int-3
Individual nucleotides		
A at Intron -6	3'	Int-6
C at Intron -6	3'	Int-6
G at Intron -6	3'	Int-6
T at Intron -6	3'	Int-6
A at Intron -5	3'	Int-5
C at Intron -5	3'	Int-5
G at Intron -5	3'	Int-5
T at Intron -5	3'	Int-5
A at Intron -3	3'	Int-3
C at Intron -3	3'	Int-3
G at Intron -3	3'	Int-3
T at Intron -3	3'	Int-3
A at the first nucleotide of exon	Ex	Ex+1
C at the first nucleotide of exon	Ex	Ex+1
G at the first nucleotide of exon	Ex	Ex+1
T at the first nucleotide of exon	Ex	Ex+1
A at the 2nd nucleotide of exon	Ex	Ex+2
C at the 2nd nucleotide of exon	Ex	Ex+2
G at the 2nd nucleotide of exon	Ex	Ex+2
T at the 2nd nucleotide of exon	Ex	Ex+2
A at the 3rd nucleotide of exon	Ex	Ex+2
C at the 3rd nucleotide of exon	Ex	Ex+2
G at the 3rd nucleotide of exon	Ex	Ex+2
T at the 3rd nucleotide of exon	Ex	Ex+2
Presence of A or G at Int-7, Int-6, or Int-5	3'	Int-7 to Int-5
Ratio of purines (A/G) at Int-20 to Int-8	3'	Int-20 to Int-8
Number of G nucleotides at Int-12 to Int-3	3'	Int-12 to Int-3
Number of GGG trinucleotides at Int-12 to Int-3	3'	Int-12 to Int-3
Other parameters		
SD-Score	Ex/5'	Ex-3 to Int+6
Exon length	Ex	Ex
MaxEntScan::score3ss	3'/Ex	Int-20 to Ex+3
MaxEntScan::score5ss	Ex/5'	Ex-3 to Int+6
Shapiro Senapathy score at 3' ss	3'/Ex	Int-14 to Ex+1
Shapiro Senapathy score at 5' ss	Ex/5'	Ex-2 to Int+6
SpliceAid2 scores of RNA-bindingprotein[d]		
Sum score of 9G8-binding site	3'/Ex/5'	Int-50 to Int+50

Sum score of CUG-BP1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of DAZAP1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of ESRP1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of ESRP2-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of ETR-3-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of FMRP-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of Fox1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of Fox2-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of HTra2alpha-binding site	3'/Ex/5	Int-50 to Int+50
Sum score of HTra2beta1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of HuB-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of HuC-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of HuD-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of HuR-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of KSRP-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of MBNL1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of Nova1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of Nova2-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of PSF-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of QKI-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of RBM25-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of RBM4-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of RBM5-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SAP155-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SC35-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SF1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SF2/ASF-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SLM1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SLM2-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SRm160-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SRp20-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SRp30c-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SRp38-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SRp40-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SRp54-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SRp55-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of SRp75-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of Sam68-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of TDP43-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of TIA1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of TIAL1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of YB1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of ZRANB2-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP A0-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP A1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP A2/B1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP A3-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP C1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP C2-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP C-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP D0-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP D-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP DL-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP E1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP E2-binding site	3'/Ex/5'	Int-50 to Int+50

Sum score of hnRNP F-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP G-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP H1-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP H2-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP H3-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP I (PTB)-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP J-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP K-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP L-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP LL-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP M-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP P (TLS)-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP Q-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of hnRNP U-binding site	3'/Ex/5'	Int-50 to Int+50
Sum score of nPTB-binding site	3'/Ex/5'	Int-50 to Int+50

[a]The feature was applied to the indicated position. Int+N and Int-N represent the number of intronic nucleotides from the 5' and 3' ss, respectively. Similarly, Ex+N and Ex-N represent the number of exonic nucleotides from the 3' and 5' ss, respectively.

[b]Best-BPS was determined in each intron using the position weight matrix (PWM) of our previous report on human consensus BPS [1]. For example, when a candidate BPS is "CTGAT", the sum of nucleotide probabilities at the five positions becomes 0.470 + 0.746 + 0.177 + 0.923 + 0.420 = 2.736. In a meantime, the best BPS is "CTCAT" with the sum of nucleotide probabilities of 3.007, whereas the worst BPS is "AGATG" with the sum of nucleotide probabilities of 0.360. PWM scores of the best and worst BPS are set to 1.000 and 0.000, respectively. Thus, the PWM score of "CTGAT" becomes 0.897.

Nucleotide probability at each position in human BPS [1]

Consensus	y	T	n	A	y
Position	-3	-2	-1	0	1
A	0.083	0.066	0.166	0.923	0.182
C	0.470	0.160	0.448	0.033	0.331
G	0.127	0.028	0.177	0.017	0.066
T	0.320	0.746	0.210	0.028	0.420

[c]Best-BPS-PPT, the best pair of BPS and PPT was determined according to the following algorithm. First, 'nYnAn' motif was looked for with an invariant 'A' at Int-50:Int-3 and set to be BPS_i. BPS_i located downstream of Int-9 was excluded because the length of PPT became less than 7 nucleotides. Second, the ratio of T/C's at positions +4 to +24 (PPT_i) from the invariant 'A' of BPS_i was calculated. This gave rise to multiple candidate BPS_i-PPT_i pairs at a single intron-exon boundary. The sum of the PWM of BPS_i and the T/C ratio in PPT_i was then calculated and a pair with the best sum score was selected.

[d]The exact motif for an RNA-binding protein was searched for at Int-50:Ex:Int+50 and scored according to SpliceAid 2 [2]. The sum of SpliceAid 2 scores was used as a feature for each RNA-binding protein.

References

1. Gao, K.; Masuda, A.; Matsuura, T.; Ohno, K. Human Branch Point Consensus Sequence Is YUnAy. *Nucleic Acids Res* **2008**, *36*, 2257–2267, doi:10.1093/nar/gkn073.
2. 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* **2012**, *33*, 81–85, doi:10.1002/humu.21609.

Supplementary Table S3. Comparison of 9 statistical measures of FexSplice, SpliceAI, and CI-SpliceAI

Model	FexSplice[b] (15-feature based LightGBM)	SpliceAI[c]	CI-SpliceAI[c]
Accuracy[a]	0.77 ± 0.07	0.54	0.58
Precision[a]	0.80 ± 0.12	0.85	0.82
Recall[a]	0.78 ± 0.13	0.22	0.22
Specificity[a]	0.77 ± 0.15	0.95	0.95
F1 score[a]	0.77 ± 0.07	0.35	0.34
NPV[a]	0.79 ± 0.11	0.49	0.55
MCC[a]	0.57 ± 0.15	0.24	0.25
AUROC	0.86 ± 0.08	0.69	0.65
AUPRC	0.87 ± 0.09	0.75	0.68

[a]Refer to Table 1 for the definitions of statistical measures.
[b]Statistical measures for FexSplice were calculated by 10-fold cross-validation (mean and SD) of our dataset, and are identical to those shown in Table 1.
[c]Statistical measures for SpliceAI and CI-SpliceAI are calculated by our dataset without cross-validation.
Note that FexSplice was trained with our dataset, whereas the others were not.