Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.

Benchmarking deep learning splice prediction tools using functional splice assays

Tabea Riepe[1], Mubeen Khan[2], Susanne Roosing[3], Frans Cremers[4], and Peter 't Hoen[5]

[1]Radboud University Nijmegen Centre for Molecular and Biomolecular Informatics
[2]Radboud Universiteit Donders Institute for Brain Cognition and Behaviour
[3]Radboud University Medical Center
[4]Radboud University Nijmegen Medical Centre
[5]Radboud Universiteit Centre for Molecular and Biomolecular Informatics

September 22, 2020

Abstract

Hereditary disorders are frequently caused by genetic variants that affect pre-mRNA splicing. Whilst genetic variants in the canonical splice motifs are almost always disrupting splicing, the pathogenicity of variants in the non-canonical splice sites (NCSS) and deep intronic (DI) regions are difficult to predict. Multiple splice prediction tools have been developed for this purpose, with the latest tools employing deep learning algorithms. We benchmarked established and deep learning splice prediction tools on gold standard sets of variants in the ABCA4 and MYBPC3 genes associated with Stargardt disease (STGD1) and cardiomyopathy, respectively, with functional assessment in midigene and minigene splice assays. The best performing splice prediction tool for both NCSS and DI variants in ABCA4 was SpliceAI, whilst SpliceSiteFinder-like performed best for NCSS variants in MYBPC3. Overall, the performance in a real time clinical setting is much more modest than reported by the developers of the tools.

Keywords

RNA splicing, variant effect prediction, deep learning, inherited retinal diseases, *ABCA4* , *MYBPC3*

Introduction

An estimated fifty percent of pathogenic variants result in aberrant splicing (López-Bigas, Audit, Ouzounis, Parra, & Guigó, 2005; Pan, Shai, Lee, Frey, & Blencowe, 2008). Genetic variants may affect all sequence elements required for correct splicing including the three core elements that are recognized by the spliceosome: the canonical 5' splice donor site (SDS), the canonical 3' splice acceptor site (SAS) and the branchpoint. Both the SDS and SAS contain conserved dinucleotides. At the SDS the most common encountered dinucleotide is a GT and at the SAS invariably an AG. Alternative dinucleotides for the SDS are known, of which GC with a frequency of 1% is the most common one (Sheth et al., 2006). In contrast with the SDS and SAS, the branchpoint motif is less conserved (Will & Lührmann, 2011). It contains the consensus sequence yUnAy with a conserved uracil (U) and adenine (A) and less conserved pyrimidines (y) (Gao, Masuda, Matsuura, & Ohno, 2008; Rogan, Caminsky, & Mucaki, 2014). The branchpoint is located between 9 and 400 nucleotides (nt) upstream of the SAS (Abramowicz & Gos, 2018). The non-canonical sequences around the canonical splice sites are part of the splice site consensus and therefore also conserved. The non-canonical sequences at the SAS are located from 14 to 3 nt upstream and 2 nt downstream, i.e., in the exon. For the SDS, these are the last two nt of the exon and positions 3 to 6 downstream. In addition to the three main core elements, other *cis* -acting elements such as intronic and exonic splicing enhancers and silencers are involved in splicing (Albert et al., 2018; Glisovic, Bachorik, Yong, & Dreyfuss, 2008).

Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.

Variants in the SDS, SAS, branchpoint and enhancer and silencer motifs can alter splicing (Ohno, Takeda, & Masuda, 2018; Wimmer et al., 2007). Those affecting canonical sequences are considered to have a major effect, where the relevant exon is skipped and even skipping of neighboring exons can be observed. In the presence of alternative splice sites in or outside of the exon, partial exon skipping or exon elongation also have been observed (Fadaie et al., 2019; Fang et al., 2001; Khan, Cornelis, Pozo-Valero, et al., 2020; Labonne et al., 2016; Ramalho et al., 2003; Sangermano et al., 2018; Symoens et al., 2011). Variants in the non-canonical splicing motifs are referred to as non-canonical splice site (NCSS) variants. These may affect splicing by weakening the existing splice site (Bradley et al., 2005; Shaw et al., 2003). On the contrary, deep-intronic (DI) variants can create or strengthen cryptic splice sites (Fadaie et al., 2019; Khan, Cornelis, Pozo-Valero, et al., 2020; Sangermano et al., 2018; Sobczyńska-Tomaszewska et al., 2013; Hanzhen Sun & Chasin, 2000). In general, DI variants will result in pseudo-exon inclusion into the mRNA, when an appropriate naturally existing SAS or SDS is present (Dhir & Buratti, 2010; Romano, Buratti, & Baralle, 2013).

To determine the impact of a putative pathogenic variant or variant of unknown significance (VUS) on splicing, *in silico* splice prediction tools may be employed. The available tools make use of three different algorithms: motif-based algorithms, machine learning algorithms and deep learning algorithms. The novel deep learning tools show promising improvements in the field of *in silico* splice prediction (Cheng et al., 2019; Naito, 2019; Zuallaert et al., 2019), as they do not rely on preselected features. As such, they may capture more complex information such as the distance between different sequence motifs, structural motifs, and non-linear relationships. They may also capture the joint effects of the SDS and SAS, explaining splice site interdependence (Hefferon, Broackes-Carter, Harris, & Cutting, 2002; Khan, Cornelis, Sangermano, et al., 2020; Ohno et al., 2018). Most*in silico* splice prediction tools are trained and evaluated on RNA-seq data, achieving high scores for accuracy and precision that often cannot be reproduced in diagnostics. The reported area under the precision recall curve for SpliceAI for instance is 0.98 (Jaganathan et al., 2019). SpliceAI demonstrated lower performance in small clinical real time test sets (Ellingford et al., 2019; Wai et al., 2020).

Currently, there is no study comparing different deep learning splice prediction tools on a clinically relevant set of variants. In the past, non-deep learning tools have been compared to each other (Jian, Boerwinkle, & Liu, 2014; Moles-Fernández et al., 2018). More recently, one deep learning tool has been compared to non-deep learning tools, in which case the deep learning tool has shown to be more accurate in its predictions and to perform better (Ellingford et al., 2019; Jaganathan et al., 2019; Jian et al., 2014; Ohno et al., 2018). In this study, we compared the motif-based algorithm SpliceSiteFinder-like (Shapiro & Senapathy, 1987), the interaction-based algorithm MaxEntScan (Yeo & Burge, 2004), the machine-learning tools CADD (Rentzsch, Witten, Cooper, Shendure, & Kircher, 2019), GeneSplicer (Pertea, 2001), NNSPLICE (Reese, Eeckman, Kulp, & Haussler, 1997), S-CAP (Jagadeesh et al., 2019) and SPIDEX (Xiong et al., 2015) and the deep learning tools DSSP (Naito, 2019), MMSplice (Cheng et al., 2019), MTSplice (Cheng, Çelik, Kundaje, & Gagneur, 2020), SpliceAI (Jaganathan et al., 2019) and SpliceRover (Zuallaert et al., 2018). A motivation for this selection is given in the Methods section. The comparison was done on two of the largest, high confidence sets of variants that are rare, potentially clinically relevant and for which the effect of splicing has been functionally assessed using mini or midigene assays.

The variants are located in genes coding for ATP binding cassette subfamily A member 4 (*ABCA4*) and Myosin binding protein C (*MYBPC3)* . ABCA4 is a flippase that effectively transports the inactive ligand of rhodopsin and the (color) opsins to the photoreceptor cell cytoplasm. The ligand is then transported to the retinal pigment epithelium where it is converted back to the active ligand and re-united with the opsins. (Molday, Rabin, & Molday, 2000; H. Sun & Nathans, 1997). Biallelic pathogenic variants in *ABCA4* cause Stargardt disease (STGD1), which displays a spectrum of retinal phenotypes encompassing early-onset, classical and late-onset STGD1 depending on the severity of the two alleles (Allikmets et al., 1997; Cremers, Lee, Collin, & Allikmets, 2020; Cremers et al., 1998; Maugeri et al., 2000).*MYBPC3* is involved in muscle contraction in heart muscle cells, and defects are associated with cardiomyopathy (Marston et al., 2009; Van Dijk et al., 2009).

Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.

Methods

Datasets

Seventy-one *ABCA4* NCSS, 81 *ABCA4* DI and 61 *MYBPC3*NCSS variants with functional validation were taken from Khan *et al.* (*ABCA4*) and Ito *et al.* (*MYBPC3*) (Figure 1a, Supplementary Table 1) (Ito et al., 2017; Khan, Cornelis, Pozo-Valero, et al., 2020). The selection criterion for functional validation of the*ABCA4* variants was a 2% difference in splice score for at least two of the Alamut programs (SpliceSiteFinder-like, MaxEntScan, NNSPLICE, GeneSplicer and HSF) and a relative strength of at least 75% for novel splice sites. To assess pathogenicity of putative causative *ABCA4*variants, splice assays were performed using midigenes as previously described (Fadaie et al., 2019; Khan, Cornelis, Pozo-Valero, et al., 2020; Sangermano et al., 2018). In short, midigenes contain multiple exons and introns to create a natural genomic context for testing the effect of variants on splicing. A construct containing the wildtype is then compared to a construct including the mutant variant which is introduced by site-directed mutagenesis. After independent transfection into HEK293T cells, *ABCA4* transcripts were amplified using RT-PCR and separated on a 2% agarose gel to determine the percentage of mutant RNA in comparison with the control line. *ABCA4* variants with more than 20% mutant RNA were classified as splice altering (Sangermano et al., 2018). *MYBPC3* variants with a smaller MaxEntScan score than the reference nucleotide were assessed in minigenes that contained a CMV promoter and a 500 bp oligonucleotide with the relevant intron flanked by exon fragments (Ito et al., 2017). Computational quantification of qPCR transcripts with a significant difference ($p < 0.001$, two-sided Fisher's exact test) between wildtype and mutant transcript was performed, and variants with a significant difference were classified as splice altering. Both the *ABCA4* and *MYBPC3*datasets were aligned to the human genome reference GRCh37/hg19 assembly.

In silico Splice Prediction Tools

In silico splice prediction tools were selected based on the following criteria. First, the tool is freely available. Second, the tool can be applied to a variant in either variant or sequence format. Third, the tool either uses deep learning or is widely applied in routine diagnostics. An overview of all *in silico* prediction tools and their characteristics is provided in Table 1. Delta scores according to formula (1) were calculated for tools that provided a separate score for wild type and variant sequences. The absolute value of the score was used for tools that returned negative values to only compare the magnitude of splice change.

$$Delta\ score = \left| \frac{WT\ score - variant\ score}{\text{Maximum score of the tool}} \right| \qquad (1)$$

The commonly applied tools GeneSplicer, MaxEntScan, NNSPLICE and SpliceSiteFinder-like were accessed from Alamut Visual Software version 2.13 (SOPHiA GENETICS, Lausanne, Switzerland). Missing values likely do not result in a change compared to wildtype and are unlikely to affect splicing. Therefore, they were replaced with zero. When multiple splice sites close to the variant were scored, the score for the canonical splice site was chosen for NCSS variants and the score for the novel created/strengthened splice site was chosen for DI variants.

The other tools were accessed separately from either a website, available scripts or files with precomputed scores. Tools accessed via their website were CADD v1.6 and SpliceRover. For CADD, a VCF file with the variants was uploaded to the website, and raw scores were obtained. SpliceRover required a FASTA sequence with a minimal length of 400 nt. Thus, we included 410 nt long sequences around the variant of interest as input. For 11 variants, which provided an error message, we used a different input length to obtain a score (*ABCA4* : 1000 nt for c.769-605T>C, c.769-1778T>C, c.302+628C>T and 750 for c.769-788A>T;*MYBPC3* : 1000 for c.3815-10T>G, c.2906-12C>T, c.1928-11G>A, c.1625-8C>G, c.1227-9C>A, c.1091-8G>A and 750 for c.906-8T>C). Python scripts were available for DSSP, SpliceAI, MMSplice and MTSplice. Input sequences for DSSP inquired input sequences of 140 nt with the SAS dinucleotide at positions 69 and 70 or the SDS dinucleotide at positions 71 and 72. Donor and acceptor sequences were processed with separate python scripts available on the DSSP GitHub (https://github.com/DSSP-github/DSSP). SpliceAI

Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.

was applied to a variant call format (VCF) file. MMSplice v2.7 and MTSplice were also applied to VCF files but returned multiple scores for most variants. The absolute delta logit PSI scores for the longest transcript and the exon closest to the variant was chosen as primary score. Both tools were included into the same script and the parameter tissue_specificity determined which tool was applied. If tissue_specificity was set to true, MTSplice was chosen, otherwise MMSplice was run on the VCF file. A file with precomputed scores was available for both SPIDEX v1.0 and S-CAP v1.0. The data and all analysis scripts can be found at*https://github.com/cmbi/Benchmarking_splice_prediction_tools*.

MMSplice, MTSplice, Spidex and S-CAP could not calculate a score for more than half of the *ABCA4* DI variants, and we therefore excluded these tools completely for the analysis of DI variants. MMSplice and MTSplice only consider variants located within 300 nt of the SDS or SAS. SPIDEX and S-CAP scores were retrieved from files with precomputed values, which did not include DI variants.

Classification metrics and Receiver Operator Curve

The accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and Matthews correlation coefficient (MCC) were calculated for each dataset. The formulas of the applied classification metrics are provided in Supplementary Table 2. In addition to the standard statistical measures, MCC was used. The MCC is best suited for unbalanced datasets while other metrics are influenced by the size of the positive and negative group. A consensus of the Alamut tools (GeneSplicer, MaxEntScan, NNSPLICE and SpliceSiteFinder-like) is frequently considered in diagnostics. Therefore, an Alamut consensus with 3/4 tools was included in the assessment. Sklearn 0.19.2 for python was used to calculate the area under the curve (AUC) and the optimal cutoff to separate the true positives and true negatives for each prediction tool.

Results

Variants

Seventy-one *ABCA4* NCSS variants, 81 *ABCA4* DI variants and 61 *MYBPC3* NCSS variants were evaluated with a selection of splice prediction tools (Figure 1a, Table 1). The number of variants that alter splicing and variants that have no effect on splicing is provided in Figure 1b. Ninety percent (64 out of 71) of *ABCA4* NCSS variants altered splicing, whilst 74% (60 out of 81) of *ABCA4* DI variants had no effect on splicing. *MYBPC3* NCSS variants showed a more even distribution with 56% (34 out of 61) splice altering variants. For all three datasets, more variants were located near the SDS than the SAS (Figure 1c). Figures 1d and e show the distribution of variants around the SAS and SDS respectively for all NCSS variants. On the donor site most splice altering variants are located at the last exonic position, and on the acceptor site most splice altering variants are located at position -3.

Receiver Operator Curve (ROC) and Area under the curve (AUC)

The ROC curves with AUCs of the five best performing tools for each dataset are provided in Figure 2a-c. For *ABCA4* NCSS variants those tools were SpliceAI, DSSP, S-CAP, Spidex and MaxEntScan. The best performing tools for the *ABCA4* DI variants were SpliceAI, SpliceRover, MaxEntScan, NNSplice and Alamut 3/4. For *MYBPC3* NCSS variants, SpliceSiteFinder-like, Alamut 3/4 MaxEntScan, NNSplice and GeneSplicer achieved the highest AUC. ROC curves and AUCs including all tools are provided in Supplementary Figure 1 and Supplementary Table 3.

Figure 3 groups the different tools for each dataset into categories based on their AUC value. The AUC of MMSplice was always lower than 0.59, which makes it one of the worst performing tools. In addition, CADD showed a low performance with an AUC of 0.64 or lower. SpliceRover, GeneSplicer, SPIDEX and S-CAP performed on average with an AUC between 0.60 and 0.79. Other tools with consistent, average performance and an AUC between 0.7 and 0.81 were the Alamut 3/4 consensus, MaxEntScan and NNSPLICE. SpliceSiteFinder-like, MTSplice and DSSP showed variable performance with a high AUC (0.70-0.79) for one dataset and a low AUC (0.5-0.59) for another dataset. SpliceAI was the best performing tool for *ABCA4* variants with an AUC higher than 0.8, but achieved an AUC of only 0.72 for *MYBPC3* variants.

Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.

Comparison of thresholds

The ROC was used to determine the best threshold for each dataset to classify the variants as splice altering or non-splice altering. Table 2 shows the comparison of the thresholds identified with the ROC curve with the predefined threshold for the different tools suggested by the developers. The best threshold to maximize the number of true positives and true negatives depended highly on the dataset. For MaxEntScan, SpliceSiteFinder-like and NNSplice the best thresholds were higher than the predefined threshold, whereas the best thresholds for SPIDEX and SpliceAI were lower. DSSP thresholds observed for NCSS variants were lower than the predefined threshold whilst the best threshold for DI variants was higher. The thresholds for S-CAP and CADD were difficult to compare to the predefined thresholds as these utilize a threshold depending on the location of the variant. For CADD, the threshold for *ABCA4* DI variants was lower than the suggested threshold. Two tools, SpliceAI and SpliceSiteFinder-like, showed thresholds close to the predefined threshold.

Performance assessment of the splice prediction tools

The accuracy, sensitivity, specificity, PPV, NPV and MCC for each dataset, as defined in Supplementary Table 2, are provided in Tables 3-5. For the *ABCA4* NCSS variants the PPV was above 90% for all tools and the NPV was below 30% for all tools. This can be explained by the imbalance in the variants as the majority of the variants in this dataset affected splicing. Spidex, which was found in the list of the five tools with the highest AUC value, had a low accuracy, sensitivity and NPV compared to other tools. Moreover, it predicted 44 false negatives. The number of false negatives for the other tools is on average 23. The highest MCC, a measure that is optimal for unbalanced test data sets, was found for SpliceAI and DSSP. For *ABCA4* DI variants the tools with the lowest performance assessment corresponded to the tools with the lowest AUC (SpliceSiteFinder-like, DSSP and CADD), and SpliceAI showed the best performance assessment. For the *MYBPC3* NCSS dataset, the tools demonstrating a low performance were also the ones with a low AUC. DSSP showed a low performance on specificity with 46% and SPIDEX on accuracy (49%) and especially sensitivity (18%). Additionally, DSSP predicted 20 false positives (9 on average) and SPIDEX 28 false negatives (13 on average). The tool with the highest MCC was the Alamut 3/4 consensus approach followed by SpliceSiteFinder-like and GeneSplicer.

Discussion

Increasing use of whole genome and whole exome sequencing in routine diagnostics requires *in silico* splice prediction tools to select likely pathogenic variants for further testing. To date, there are studies evaluating single splice prediction tools, but none comparing multiple deep learning tools. This study benchmarked selected established and deep learning *in silico* splice prediction tools based on multiple classification metrics on two of the largest sets of variants for which the effect of splicing is functionally assessed using mini- or midigene assays. The data showed that SpliceAI, the Alamut 3/4 consensus approach, NNSPLICE and MaxEntScan perform well on all datasets. Additionally, this study demonstrated that the choice of the best splice prediction tool may depend on the gene of interest and the type of splice altering variants.

We included NCSS and DI variants in the *ABCA4* gene and NCSS variants in the *MYBPC3* gene. There was no single best performing splice prediction tool for these different datasets. This may be explained in several ways. *ABCA4* and *MYBPC3* are expressed in a tissue specific manner, with high expression in the retina and heart muscle, respectively. The representation of splice patterns in these tissues in the data used for training of the different deep learning algorithms may affect its performance. None of the tools included retina tissue in its training data, as far as we can judge. Moreover, most splice prediction tools focus on the area around the canonical splice sites and were not trained on DI variants, which explains their lower performance on the DI dataset. Another reason for differences in performance may lie in the selection criteria used to functionally assess the *ABCA4* and *MYBPC3* variants. *MYBPC3* variants were selected for functional validation based on MaxEntScan scores, and *ABCA4* variants were selected when they showed a difference in splice score for at least two of the Alamut programs (including Human Splicing Finder) and/or a delta score of at least 2%. This may lead to a positive bias in the performance assessment for the tools

Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.

that were used to select the variants, but we find the opposite, where Alamut 3/4 performs best on the *MYBPC3*data, and MaxEntScan performs relatively well on the *ABCA4*dataset. Yet another source of difference in performance can be found in the functional assays used for their evaluation; *ABCA4* variants were tested in midigenes and *MYBPC3* variants in minigenes. In most cases, minigenes and midigenes result in the same transcripts but when the flanking exons of the minigene vector are stronger than the ones in the gene of interest, they can cause artefacts.

The performance measures of splice prediction tools need to be carefully chosen, in particular when there is an imbalance in the number of splice altering and non-splice altering variants. In the *ABCA4* NCSS dataset most variants affected splicing, while most *ABCA4* DI variants had no effect on splicing. The *MYBPC3* dataset contained about the same number of splice and non-splice altering variants. Imbalance in the dataset influences most classification metrics. If the positive (splice altering) and negative class (non-splice altering) are interchanged during calculation of the metric the metric changes. The only metric not influenced by class imbalance is MCC, and we regard this as the preferred measure in the current setting. One example to demonstrate this are the metrics for Spidex calculated for *ABCA4*NCSS variants with mainly splice altering variants. Spidex showed a specificity of 71%, which is similar to the other tools, but an MCC of only 0.02.

Our results are consistent with previous studies that included smaller number of splice prediction tools. Wai et al. compared Alamut, Human Splicing Finder and SpliceAI on 257 VUSs (NCSS and DI) from blood RNA samples showing that SpliceAI outperformed the other tools with an AUC of 0.951 (Wai et al., 2020). A second study by Ellingford et al. compared SpliceAI, SPIDEX, S-CAP, CADD and MaxEntScan first in a real time assessment of 21 variants and then in variant prioritization of nearly 3000 variants (Ellingford et al., 2019). The real-time assessment showed that SpliceAI and MaxEntScan achieved a good performance. In the variant prioritization of the large cohort only SpliceAI, Spidex and CADD are compared. Here SpliceAI showed the highest AUC (0.96). Our AUC values for SpliceAI were 0.80 (*ABCA4* NCSS), 0.95 (*ABCA4*DI) and 0.72 (*MYBPC3* NCSS). Especially the AUCs of the NCSS datasets are lower than the AUC found in the two other studies. There can be multiple explanations for this. First, our datasets are smaller, making the right prediction for each individual variant more important. Second, we used variants located in only one gene, whereas the above-mentioned studies used variants in a variety of genes. This could indicate that for genes with tissue-specific expression the available splice prediction tools are not specialized enough, for reasons explained above. Third, we evaluated tools based on functional assessment with midi- or minigenes assays, which currently represent the best medium-throughput tools. Still, also this experimental set-up has limitations, since the splice assays were performed in human kidney cells. This means that tissue specific splicing events may be missed. For *ABCA4* it is known that variants can lead to tissue specific pseudo-exon inclusion (Albert et al., 2018). Another limitation is that the percentage mutant RNA of the *ABCA4* variants is determined based on RT-PCR products visualized on agarose gels. RT-PCR has a bias towards smaller segments and this can lead to incorrect classification of the variants. A better alternative would be to use RNA-sequencing.

A general observation made in our benchmark study is that the prediction of the *in silico* tools on a set of clinically relevant variants varies considerably from the performance described in the original paper. SpliceAI, for instance, achieves an area under the precision recall curve (PR-AUC) of 0.98 on RNA-seq data (Jaganathan et al., 2019). For our datasets the PR-AUC is 0.94 for *ABCA4* NCSS variants, 0.91 for *ABCA4* DI variants and 0.75 for *MYBPC3* NCSS variants. The higher performance observed by the authors can be explained by the use of an RNA-seq dataset. Using big RNA datasets to evaluate the performance of a novel algorithm will artificially increase its performance, because naturally occurring high frequency variants have a different effect on splice sites than rare variants affecting splicing. Moreover, circularity, i.e. incomplete independence of the variants used for training and testing, may result in overestimation of the performance of the model very similar properties were already (Grimm et al., 2015). This is why it is important to use a truly independent set of clinically relevant variants to evaluate the performance of the splice prediction tools. Additionally, it is important to use the right evaluation metrics to compare different algorithms. As shown for the*ABCA4* variants imbalance in the dataset influences the classification metrics and therefore also the comparison. The precision recall curve uses the PPV and sensitivity to calculate the AUC. Imbalance in the

Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.

dataset has an influence on both metrics, which makes it difficult to compare highly imbalanced datasets based on the PR-AUC.

To conclude, there are a variety of different splice prediction tools available. It is not easy to choose which tool to use, because different tools may perform better in different contexts. The best performing tools make use of different algorithms, deep learning (SpliceAI), machine-learning (NNSPLICE) and interactions (MaxEntScan). Deep learning has the possibility to improve splice prediction, but is not a guarantee for success: Out of the five deep learning tools, only SpliceAI performed better than the more established tools.

Web Resources

https://github.com/DSSP-github/DSSP

https://github.com/cmbi/Benchmarking_splice_prediction_tool

Acknowledgements

The work of M.K. was supported by RetinaUK, grant no. GR591 (to FPMC).

Data Availability Statement

Variants used for study are publicly available online at LOVD (https://www.lovd.nl/), ClinVar (www.ncbi.nlm.nih.gov/clinvar/) and ExAC (exac.broadinstitute.org/).

References

Abramowicz, A., & Gos, M. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. *Journal of Applied Genetics* , Vol. 59, pp. 253–268. https://doi.org/10.1007/s13353-018-0444-7

Albert, S., Garanto, A., Sangermano, R., Khan, M., Bax, N. M., Hoyng, C. B., ... Cremers, F. P. M. (2018). Identification and Rescue of Splice Defects Caused by Two Neighboring Deep-Intronic ABCA4 Mutations Underlying Stargardt Disease. *American Journal of Human Genetics* ,*102* (4), 517–527. https://doi.org/10.1016/j.ajhg.2018.02.008

Allikmets, R., Singh, N., Sun, H., Shroyer, N. F., Hutchinson, A., Chidambaram, A., ... Lupski, J. R. (1997). A photoreceptor cell-specific ATP-binding transporter gene (ABCR) is mutated in recessive Stargardt macular dystrophy. *Nature Genetics* ,*15* (3), 236–246. https://doi.org/10.1038/ng0397-236

Bradley, K. J., Cavaco, B. M., Bowl, M. R., Harding, B., Young, A., & Thakker, R. V. (2005). Utilisation of a cryptic non-canonical donor splice site of the gene encoding PARAFIBROMIN is associated with familial isolated primary hyperparathyroidism. *Journal of Medical Genetics* , *42* (8). https://doi.org/10.1136/jmg.2005.032201

Cheng, J., Çelik, M. H., Kundaje, A., & Gagneur, J. (2020). MTSplice predicts effects of genetic variants on tissue-specific splicing.*BioRxiv* , 2020.06.07.138453. https://doi.org/10.1101/2020.06.07.138453

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Çelik, M. H., Fairbrother, W. G., Avsec, Ž., & Gagneur, J. (2019). MMSplice: Modular modeling improves the predictions of genetic variant effects on splicing.*Genome Biology* , *20* (1). https://doi.org/10.1186/s13059-019-1653-z

Cremers, F. P. M., Lee, W., Collin, R. W. J., & Allikmets, R. (2020). Clinical spectrum, genetic complexity and therapeutic approaches for retinal disease caused by ABCA4 mutations. *Progress in Retinal and Eye Research* , 100861. https://doi.org/10.1016/j.preteyeres.2020.100861

Cremers, F. P. M., Van De Pol, D. J. R., Van Driel, M., Den Hollander, A. I., Van Haren, F. J. J., Knoers, N. V. A. M., ... Hoyng, C. B. (1998). Autosomal recessive retinitis pigmentosa and cone-rod dystrophy caused by splice site mutations in the Stargardt's disease gene ABCR.*Human Molecular Genetics* , *7* (3), 355–362. https://doi.org/10.1093/hmg/7.3.355

Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.

Dhir, A., & Buratti, E. (2010, February). Alternative splicing: Role of pseudoexons in human disease and potential therapeutic strategies: Minireview. *FEBS Journal* , Vol. 277, pp. 841–855. https://doi.org/10.1111/j.1742-4658.2009.07520.x

Ellingford, J. M., Thomas, H. B., Rowlands, C., Arno, G., Beaman, G., Gomes-Silva, B., ... Black, G. C. (2019). Functional and in-silico interrogation of rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. *BioRxiv* , 781088. https://doi.org/10.1101/781088

Fadaie, Z., Khan, M., Del Pozo-Valero, M., Cornelis, S. S., Ayuso, C., Cremers, F. P. M., ... The, A. (2019). Identification of splice defects due to noncanonical splice site or deep-intronic variants in ABCA4. *Human Mutation* , *40* (12), 2365–2376. https://doi.org/10.1002/humu.23890

Fang, L. J., Simard, M. J., Vidaud, D., Assouline, B., Lemieux, B., Vidaud, M., ... Thirion, J. P. (2001). A novel mutation in the neurofibromatosis type 1 (NF1) gene promotes skipping of two exons by preventing exon definition. *Journal of Molecular Biology* ,*307* (5), 1261–1270. https://doi.org/10.1006/jmbi.2001.4561

Gao, K., Masuda, A., Matsuura, T., & Ohno, K. (2008). Human branch point consensus sequence is yUnAy. *Nucleic Acids Research* ,*36* (7), 2257–2267. https://doi.org/10.1093/nar/gkn073

Glisovic, T., Bachorik, J. L., Yong, J., & Dreyfuss, G. (2008, June 18). RNA-binding proteins and post-transcriptional gene regulation.*FEBS Letters* , Vol. 582, pp. 1977–1986. https://doi.org/10.1016/j.febslet.2008.03.004

Grimm, D. G., Azencott, C. A., Aicheler, F., Gieraths, U., Macarthur, D. G., Samocha, K. E., ... Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation* , *36* (5), 513–523. https://doi.org/10.1002/humu.22768

Hefferon, T. W., Broackes-Carter, F. C., Harris, A., & Cutting, G. R. (2002). Atypical 5' splice sites cause CFTR exon 9 to be vulnerable to skipping. *American Journal of Human Genetics* , *71* (2), 294–303. https://doi.org/10.1086/341664

Ito, K., Patel, P. N., Gorham, J. M., McDonough, B., DePalma, S. R., Adler, E. E., ... Seidman, J. G. (2017). Identification of pathogenic gene mutations in LMNA and MYBPC3 that alter RNA splicing.*Proceedings of the National Academy of Sciences of the United States of America* , *114* (29), 7689–7694. https://doi.org/10.1073/pnas.1707741114

Jagadeesh, K. A., Paggi, J. M., Ye, J. S., Stenson, P. D., Cooper, D. N., Bernstein, J. A., & Bejerano, G. (2019). S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing.*Nature Genetics* , *51* (4), 755–763. https://doi.org/10.1038/s41588-019-0348-4

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., ... Farh, K. K. H. (2019). Predicting Splicing from Primary Sequence with Deep Learning.*Cell* , *176* (3), 535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015

Jian, X., Boerwinkle, E., & Liu, X. (2014). In silico tools for splicing defect prediction: A survey from the viewpoint of end users.*Genetics in Medicine* , Vol. 16, pp. 497–503. https://doi.org/10.1038/gim.2013.176

Khan, M., Cornelis, S. S., Pozo-Valero, M. Del, Whelan, L., Runhart, E. H., Mishra, K., ... Cremers, F. P. M. (2020). Resolving the dark matter of ABCA4 for 1054 Stargardt disease probands through integrated genomics and transcriptomics. *Genetics in Medicine* , *0* (0). https://doi.org/10.1038/s41436-020-0787-4

Khan, M., Cornelis, S. S., Sangermano, R., Post, I. J. M., Groesbeek, A. J., Amsu, J., ... Cremers, F. P. M. (2020). In or out? New insights on exon recognition through splice-site interdependency.*International Journal of Molecular Sciences* , *21* (7), 2300. https://doi.org/10.3390/ijms21072300

Labonne, J. D. J., Chung, M. J., Jones, J. R., Anand, P., Wenzel, W., Iacoboni, D., ... Kim, H. G. (2016). Concomitant partial exon skipping by a unique missense mutation of RPS6KA3 causes Coffin-Lowry syndrome. *Gene* , *575* (1), 42–47. https://doi.org/10.1016/j.gene.2015.08.032

Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G., & Guigó, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? *FEBS Letters* , *579* (9), 1900–1903. https://doi.org/10.1016/j.febslet.2005.02.047

Marston, S., Copeland, O., Jacques, A., Livesey, K., Tsang, V., McKenna, W. J., ... Watkins, H. (2009). Evidence from human myectomy samples that MYBPC3 mutations cause hypertrophic cardiomyopathy through haploinsufficiency*. *Circulation Research* , *105* (3), 219–222. https://doi.org/10.1161/CIRCRESAHA.109.202440

Maugeri, A., Klevering, B. J., Rohrschneider, K., Blankenagel, A., Brunner, H. G., Deutman, A. F., ... Cremers, F. P. M. (2000). Mutations in the ABCA4 (ABCR) gene are the major cause of autosomal recessive cone-rod dystrophy. *American Journal of Human Genetics* ,*67* (4), 960–966. https://doi.org/10.1086/303079

Molday, L. L., Rabin, A. R., & Molday, R. S. (2000). ABCR expression in foveal cone photoreceptors and its role in Stargardt macular dystrophy.*Nature Genetics* , *25* (3), 257–258. https://doi.org/10.1038/77004

Moles-Fernández, A., Duran-Lozano, L., Montalban, G., Bonache, S., López-Perolio, I., Menéndez, M., ... Gutiérrez-Enríquez, S. (2018). Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations?*Frontiers in Genetics* , *9* (SEP), 366. https://doi.org/10.3389/fgene.2018.00366

Naito, T. (2019). Predicting the impact of single nucleotide variants on splicing via sequence-based deep neural networks and genomic features.*Human Mutation* , *40* (9), 1261–1269. https://doi.org/10.1002/humu.23794

Ohno, K., Takeda, J. I., & Masuda, A. (2018, January 1). Rules and tools to predict the splicing effects of exonic and intronic mutations. *Wiley Interdisciplinary Reviews: RNA* , Vol. 9, p. e1451. https://doi.org/10.1002/wrna.1451

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* ,*40* (12), 1413–1415. https://doi.org/10.1038/ng.259

Pertea, M. (2001). GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Research* , *29* (5), 1185–1190. https://doi.org/10.1093/nar/29.5.1185

Ramalho, A. S., Beck, S., Penque, D., Gonska, T., Seydewitz, H. H., Mall, M., & Amaral, M. D. (2003). Transcript analysis of the cystic fibrosis splicing mutation 1525-1G>A shows use of multiple alternative splicing sites and suggests a putative role of exonic splicing enhancers. *Journal of Medical Genetics* , *40* (7), 88. https://doi.org/10.1136/jmg.40.7.e88

Reese, M. G., Eeckman, F. H., Kulp, D., & Haussler, D. (1997). Improved Splice Site Detection in Genie. *Journal of Computational Biology* ,*4* (3), 311–323.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* , *47* (D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Rogan, P. K., Caminsky, N., & Mucaki, E. J. (2014, November 18). Interpretation of mRNA splicing mutations in genetic disease: Review of the literature and guidelines for information-theoretical analysis.*F1000Research* , Vol. 3. https://doi.org/10.12688/f1000research.5654.1

Romano, M., Buratti, E., & Baralle, D. (2013). Role of pseudoexons and pseudointrons in human cancer. *International Journal of Cell Biology* . https://doi.org/10.1155/2013/810572

Sangermano, R., Khan, M., Cornelis, S. S., Richelle, V., Albert, S., Garanto, A., ... Cremers, F. P. M. (2018). ABCA4 midigenes reveal the full splice spectrum of all reported noncanonical splice site variants in Stargardt disease. *Genome Research* , *28* (1), 100–110. https://doi.org/10.1101/gr.226621.117

Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.

Shapiro, M. B., & Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: Sequence statistics and functional implications in gene expression. *Nucleic Acids Research* ,*15* (17), 7155–7174. https://doi.org/10.1093/nar/15.17.7155

Shaw, M. A., Brunetti-Pierri, N., Kádasi, L., Kovácová, V., Van Maldergem, L., De Brasi, D., ... Gécz, J. (2003). Identification of three novel SEDL mutations, including mutation in the rare, non-canonical splice site of exon 4. *Clinical Genetics* ,*64* (3), 235–242. https://doi.org/10.1034/j.1399-0004.2003.00132.x

Sheth, N., Roca, X., Hastings, M. L., Roeder, T., Krainer, A. R., & Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Research* , *34* (14), 3955–3967. https://doi.org/10.1093/nar/gkl556

Sobczyńska-Tomaszewska, A., Ołtarzewski, M., Czerska, K., Wertheim-Tysarowska, K., Sands, D., Walkowiak, J., ... Mazurczak, T. (2013). Newborn screening for cystic fibrosis: Polish 4 years' experience with CFTR sequencing strategy. *European Journal of Human Genetics* , *21* (4), 391–396. https://doi.org/10.1038/ejhg.2012.180

Sun, H., & Nathans, J. (1997). Stargardt's ABCR is localized to the disc membrane of retinal rod outer segments. *Nature Genetics* , Vol. 17, pp. 15–16. https://doi.org/10.1038/ng0997-15

Sun, Hanzhen, & Chasin, L. A. (2000). Multiple Splicing Defects in an Intronic False Exon. *Molecular and Cellular Biology* ,*20* (17), 6414–6425. https://doi.org/10.1128/mcb.20.17.6414-6425.2000

Symoens, S., Malfait, F., Vlummens, P., Hermanns-Lê, T., Syx, D., & de Paepe, A. (2011). A novel splice variant in the N-propeptide of COL5A1 causes an EDS phenotype with severe kyphoscoliosis and eye involvement.*PLoS ONE* , *6* (5). https://doi.org/10.1371/journal.pone.0020121

Van Dijk, S. J., Dooijes, D., Remedios, C. Dos, Michels, M., Lamers, J. M. J., Winegrad, S., ... Van Velden, J. Der. (2009). Cardiac myosin-binding protein C mutations and hypertrophic ardiomyopathy haploinsufficiency, deranged phosphorylation, and cardiomyocyte dysfunction. *Circulation* , *119* (11), 1473–1483. https://doi.org/10.1161/CIRCULATIONAHA.108.838672

Wai, H. A., Lord, J., Lyon, M., Gunning, A., Kelly, H., Cibin, P., ... Archer, L. (2020). Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genetics in Medicine* , *22* (6), 1005–1014. https://doi.org/10.1038/s41436-020-0766-9

Will, C. L., & Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology* , *3* (7), 1–2. https://doi.org/10.1101/cshperspect.a003707

Wimmer, K., Roca, X., Beiglböck, H., Callens, T., Etzler, J., Rao, A. R., ... Messiaen, L. (2007). Extensive in silico analysis of NF1 splicing defects uncovers determinants for splicing outcome upon 5' splice-site disruption. *Human Mutation* , *28* (6), 599–612. https://doi.org/10.1002/humu.20493

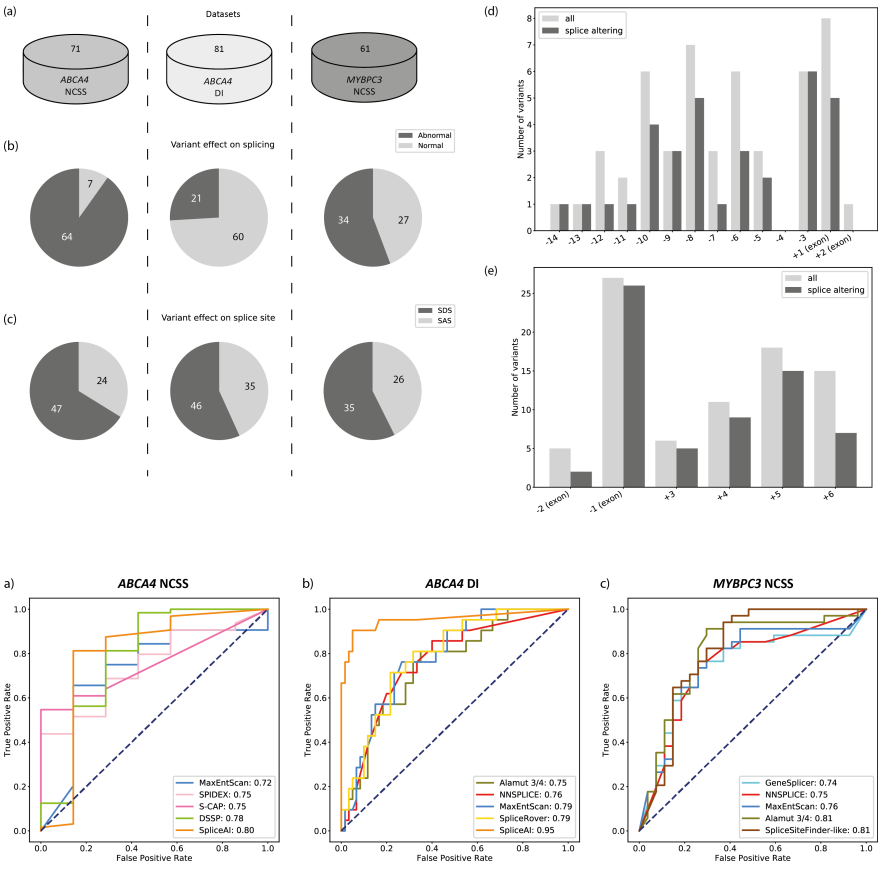
Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., ... Frey, B. J. (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science* , *347* (6218). https://doi.org/10.1126/science.1254806

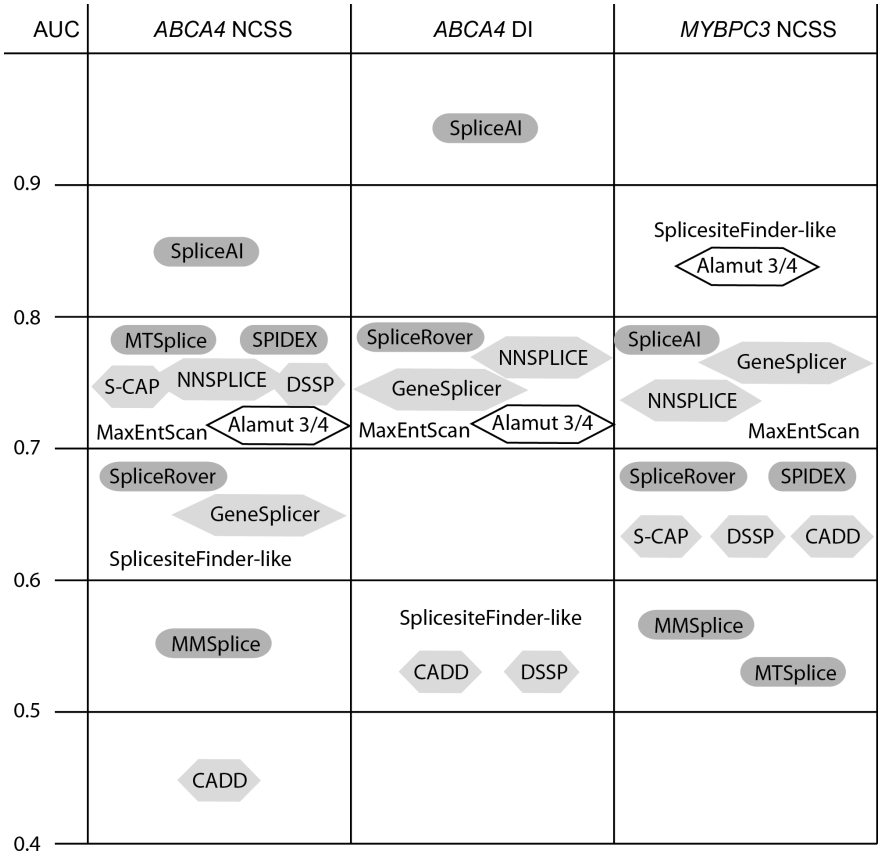
Yeo, G., & Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology* , *11* (2–3), 377–394. https://doi.org/10.1089/1066527041410418

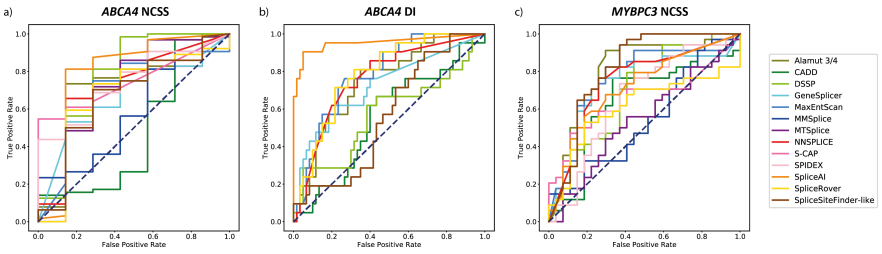
Zuallaert, J., Godin, F., Kim, M., Soete, A., Saeys, Y., & De Neve, W. (2018). Splicerover: Interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics* , *34* (24), 4180–4188. https://doi.org/10.1093/bioinformatics/bty497

Zuallaert, J., Godin, F., Kim, M., Soete, A., Saeys, Y., De Neve, W., ... Xing, Y. (2019). Deep splicing code: Classifying alternative splicing events using deep learning. *Journal of Computational Biology* , *34* (8), 330–333. https://doi.org/10.3390/genes10080587

Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.







Hosted file

`Table_1.pdf` available at https://authorea.com/users/360858/articles/482415-benchmarking-deep-learning-splice-prediction-tools-using-functional-splice-assays

Hosted file

`Table_2.pdf` available at https://authorea.com/users/360858/articles/482415-benchmarking-deep-learning-splice-prediction-tools-using-functional-splice-assays

Hosted file

`Table_3.pdf` available at https://authorea.com/users/360858/articles/482415-benchmarking-deep-learning-splice-prediction-tools-using-functional-splice-assays

Hosted file

Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.

Table_4.pdf available at https://authorea.com/users/360858/articles/482415-benchmarking-deep-learning-splice-prediction-tools-using-functional-splice-assays

Hosted file

Table_5.pdf available at https://authorea.com/users/360858/articles/482415-benchmarking-deep-learning-splice-prediction-tools-using-functional-splice-assays

Posted on Authorea 22 Sep 2020 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.160081230.07101269 | This a preprint and has not been peer reviewed. Data may be preliminary.