Published in final edited form as:

Hum Mutat. 2019 September ; 40(9): 1243–1251. doi:10.1002/humu.23788.

CAGI5 splicing challenge: Improved exon skipping and intron retention predictions with MMSplice

Jun Cheng^{1,2,*}, Muhammed Hasan Çelik¹, Yen Duong Nguyen¹, Žiga Avsec^{1,2}, Julien Gagneur^{1,2,*}

¹Department of Informatics, Technical University of Munich, Boltzmannstraße 3, 85748 Garching, Germany

²Graduate School of Quantitative Biosciences (QBM), Ludwig-Maximilians-Universität München, Germany

Abstract

Pathogenic genetic variants are often primarily affecting splicing. However, it remains difficult to quantitatively predict whether and how genetic variants affect splicing. In 2018, the fifth edition of the Critical Assessment of Genome Interpretation (CAGI 5) proposed two splicing prediction challenges based on experimental perturbation assays: Vex-seq, assessing exon skipping, and MaPSy assessing splicing efficiency. We developed a modular modeling framework, MMSplice, which performed among the best on both challenges. Here we provide insights into the modeling assumptions of MMSplice and its individual modules. We furthermore illustrate how MMSplice can be applied in practice to individual genome interpretation, using the MMSplice VEP plugin and the Kipoi variant interpretation plugin, which are directly applicable to VCF files.

Keywords

Splicing; Variant effect; CAGI; Variant interpretation

1 Introduction

RNA splicing is the process that removes intronic sequence from precursor RNAs to form mature RNAs. Alternative splicing happens when exons are concatenated in alternative combinations (Alberts et al., 2008). Alternative splicing has been shown to be important for tissue development (Baralle & Giudice, 2017). The most common type of alternative splicing in human is exon skipping (Y. Wang et al., 2015). Skipping of an exon can be quantified with the percent-spliced in (PSI, Ψ), which is defined as the fraction of transcripts that includes the exon (Goldstein et al., 2016). Another frequently used splicing metric is the splicing efficiency, which we here define as the fraction of spliced transcripts among spliced and unspliced transcripts (Braberg et al., 2013; Soemedi et al., 2017; Wilhelm et al., 2008). With RNA-seq data, splicing efficiency can be defined for every splice site by considering

*Correspondence: chengju@in.tum.de, gagneur@in.tum.de.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this article

the fraction of reads spanning exon-exon junction reads with reads spanning exon-intron boundaries. Splicing efficiency captures intron retention. Unlike Ψ that is only relevant for splice sites involved in alternative splicing, splicing efficiency is relevant for all splice sites.

Splicing defect is one of the most frequent cause of Mendelian disorders (Li et al., 2016; López-Bigas, Audit, Ouzounis, Parra, & Guigó, 2005). Moreover, thousands of splicing QTLs have been identified and linked to common diseases (Consortium et al., 2015; Li et al., 2016).

Genetic variants affect splicing in two common ways. They can change alternative splicing, in particular exon skipping. They can also change splicing efficiency. Various methods have been developed to predict variant effect on splicing. Early methods focused on scoring the effects of splice regulatory elements, such as splice sites (Soemedi et al., 2017), exon splice enhancers (ESE) and silencers (ESS), intronic splicing enhancers (ISE) and silencers (ISS) (W. G. Fairbrother, 2002; William G. Fairbrother et al., 2004; Z. Wang, Xiao, Van Nostrand, & Burge, 2006; X. H. F. Zhang & Chasin, 2004; X. Zhang & Kangsamaksin, 2005), and branchpoint (Bretschneider, Gandhi, Deshwar, Zuberi, & Frey, 2018; Paggi & Bejerano, 2018). The potential impact of variants using these methods can be assessed with the difference of scores between the reference and the alternative sequence. Other methods focus on predicting Ψ directly. One of the early successful Ψ prediction method was developed by Barash et al (Barash et al., 2010) using mouse transcriptome data. The model learned a “splicing code” from variations of Ψ across exons and across tissues. Although the model was trained only with the reference genome and not with genetic variants, it could predict effect of variants on splicing. A similar model, SPANR, was later on developed for the human genome (Xiong et al., 2015). SPANR was successful in predicting pathogenic variants for several diseases. Even though the approach of learning the splicing code from reference sequence was successful, the model may suffer from evolutionary confounding and fail to learn causal features. To address this issue, large-scale perturbation assays such as massively parallel reporter assay (MPRA) and saturation mutagenesis screens have been developed [18,19,20,21]. In particular, Rosenberg et al (Rosenberg, Patwardhan, Shendure, & Seelig, 2015) probed millions of exonic and intronic random sequences to test their impact on splicing. Their model, HAL, improved upon the state-of-the-art performance at predicting variant effect on exon skipping and alternative donor usage.

Perturbation data are ideal to benchmark computational methods for their predictive power on causal effects. CAGI5 has two splicing challenges with data from such assays: The Vex-seq (Adamson, Zhan, & Graveley, 2018) challenge and the MaPSy (Soemedi et al., 2017) challenge. The tasks of the two challenges were related yet distinct. The Vex-seq experiment assayed 2,059 natural genetic variants, including exonic and intronic single-nucleotide variants (SNVs) and insertion/deletions (indels). The measured quantity was Ψ . The MaPSy experiment measured the impact of 5,761 exonic disease causing missense mutations on splicing. The assay was both performed *in vivo* and *in vitro*. Around 10% of the mutations significantly altered splicing both *in vivo* and *in vitro*. Such variants were defined as exonic splicing mutations (ESM). The measured quantity was splicing efficiency (Methods). Although the two challenges have different measured quantities, we assumed that variant disrupting splicing could affect both Ψ and splicing efficiency. Therefore, we applied a

modular modeling approach, MMSplice (Cheng et al., 2019), where the modules score different gene regions and are shared across challenges. The predictors proposed for each challenge differ only in how they combine the scores of the individual modules.

We have described MMSplice and the modular modeling strategy previously (Cheng et al., 2019). In this CAGI special issue, we focus on the application of MMSplice to the CAGI 5 challenges. In particular, we provide insights into modeling assumptions and about the module architecture. We also emphasize model and variant interpretation, as these are relevant for downstream human genetics application.

2 Methods

2.1 Modular modeling approach for the Vex-seq and MaPSy challenges

The Vex-seq data covered variants from both exons and introns. We noticed that the training data from CAGI for both challenges were limited with 957 training data points for Vex-seq and 4,964 for MaPSy. It is probably difficult to train a model capturing much of the splicing regulatory elements directly from these data. We therefore used complementary data from different sources that are richer (Cheng et al., 2019). We used the GENCODE 24 annotation to train a module to score donor sites and similarly a module to score acceptor sites. In total, 524,569 training data and 131,143 evaluation data were used to train the donor module while 566,822 training data and 141,706 evaluation data were used to train the acceptor module. The modules were trained by training classifiers to distinguish annotated splice sites versus random sequence around the selected splice sites (with some bias to sequence with splice dinucleotide (Cheng et al., 2019)). We further used data from a MPRA that probed the effect of 2 million random sequences on splicing (Rosenberg et al., 2015). The MPRA data had exonic and intronic random sequence, from which we trained modules to score exon and modules to score intron (Cheng et al., 2019). In total we trained six modules: donor, acceptor, 5' exon, 3' exon, 5' intron and 3' intron. Detailed descriptions of all modules and their training methods are given in Cheng et al (Cheng et al., 2019). To score variants for their effect on Ψ (Ψ) and splicing efficiency, we trained separate linear models from modular predictions from common set of modules (Fig. 1). Our modules collectively consider sequence of the whole exon and 100 nt flanking intron from both sides, and therefore score variants in this range.

2.2 Vex-seq challenge

2.2.1 Data processing—The Vex-seq data tested 2,059 variants from the Exome Aggregation Consortium (ExAC) for their effect on Ψ (Adamson et al., 2018). For each variant-exon pair, Ψ for the reference sequence and for the alternative sequence was measured with RNA-Seq with a minigene reporter. The assessed variants included single-nucleotide variants (SNVs) as well as short indels from both exonic and intronic regions. The Vex-seq CAGI challenge provided 957 variants from chromosome 1 to chromosome 8 for training. For each variant, the tested exon coordinates and the associated reference Ψ and Ψ were provided. The test data consisted of 1,054 variants from chromosome 9 to 22 and chromosome X. Reference Ψ values for the exons with reference sequences were provided. The predictors had to predict Ψ for each variant.

2.2.2 Vex-seq model—To predict Ψ for each variant-exon pair from Vex-seq, five modules were applied to the reference sequence and alternative sequence separately. These were the donor module, the acceptor module, the 5' exon module, the 5' intron module, and the 3' intron module. A score difference (ΔS_{Score}) between the reference sequence and the alternative sequence for each module was calculated. A linear model was trained with Vex-seq training data to predict the log odds ratio of Ψ ($\text{logit}(\Psi)$) from the five ΔS_{Score} and using interaction terms between scores of overlapping regions. Denoting the logistic function logit , the model reads:

$$\begin{aligned}\Delta\text{logit}(\Psi) &= \text{logit}(\Psi_{\text{alt}}) - \text{logit}(\Psi_{\text{ref}}) \\ &= \beta_0 + \beta_1 \Delta S_{3' \text{ intron}} + \beta_2 \Delta S_{\text{acceptor}} \\ &\quad + \beta_3 \Delta S_{\text{exon}} + \beta_4 \Delta S_{\text{donor}} + \beta_5 \Delta S_{5' \text{ intron}} \\ &\quad + \beta_6 \mathbb{I} \text{ (Exon overlap splice site modules)} \Delta S_{\text{exon}} \\ &\quad + \beta_7 \mathbb{I} \text{ (5' intron overlap donor module)} \Delta S_{5' \text{ intron}} \\ &\quad + \beta_8 \mathbb{I} \text{ (3' intron overlap acceptor module)} + \epsilon\end{aligned}\quad (1)$$

The difference of Ψ in the natural scale, Ψ , was predicted using the reference value Ψ_{ref} and the predicted log odds ratios (Cheng et al., 2019).

$$\Psi_{\text{alt}} = \text{sigmoid}(\text{logit}(\Psi_{\text{ref}}) + \Delta\text{logit}(\Psi)) \quad (2)$$

where

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

$$\text{logit}(x) = \log \frac{x}{1 - x} \quad (4)$$

2.3 MaPSy challenge

2.3.1 Data processing—The MaPSy experiment tested 5,761 disease-causing exonic variants from Human Gene Mutation Database (HGMD) for their impact on RNA splicing efficiencies (Eq. 5) both *in vivo* and *in vitro* (Soemedi et al., 2017), quantified as:

$$\text{Splicing efficiency} = \log_2 \left(\frac{m_0/m_i}{w_0/w_i} \right), \quad (5)$$

where m_0 is the mutant spliced RNA read count, m_i is the mutant input (unspliced) RNA read count, w_0 is the wild-type spliced RNA read count, w_i is the wild-type input RNA read count (Cheng et al., 2019; Soemedi et al., 2017). Transcripts with skipped exons or mis-splicing were ignored.

2.3.2 MaPSy model

In vivo: The *in vivo* experiment of MaPSy used a three-exon construct with the test exon in the middle. As all variants are exonic, we used three modules that overlap exons: the donor

module, the acceptor module and the 5' exon module. A linear model was trained with the Scores of these three modules to predict splicing efficiency (Eq. 6).

$$\begin{aligned}\Delta\text{splice_efficiency_in_vivo} &= \beta_0 + \beta_1\Delta S_{\text{acceptor}} + \beta_2\Delta S_{\text{exon}} + \beta_3\Delta S_{\text{donor}} \\ &+ \beta_4 \text{ (Exon overlap splice site modules)}\Delta S_{\text{exon}} + \epsilon\end{aligned}\quad (6)$$

In vitro: The *in vitro* experiment of MaPSy used a two-exon construct with the test exon being the second exon in the transcript. Therefore, the test exons did not have donor sites. Consequently, we applied two modules: the acceptor module and the 5' exon module. A linear model was trained similarly as for the *in vivo* model (Eq. 7).

$$\begin{aligned}\Delta\text{splice_efficiency_in_vitro} &= \beta_0 + \beta_1\Delta S_{\text{acceptor}} + \beta_2\Delta S_{\text{exon}} \\ &+ \beta_3 \text{ (Exon overlap acceptor module)}\Delta S_{\text{exon}} + \epsilon\end{aligned}\quad (7)$$

Exon splicing mutation classification: To classify exon splicing mutations (ESMs), we trained a logistic regression model with the predicted *in vitro* splicing efficiency change and 8 other features:

- MMSplice 5' exon module score for the wild-type sequence
- MMSplice donor module score for the wild-type sequence
- MMSplice acceptor module score for the wild-type sequence
- Experiment exon length, which is the exon length in the experimental construct and may differ from the annotated genomic exon
- Log-transformed wildtype *in vitro* input
- Log-transformed mutant *in vitro* input
- Target exon phastcons conservation score
- Target exon flanking intron length in ensemble 75 annotation

Features were selected with 3-fold cross-validation. Besides the above features, scores from CADD, SIFT, phastCons, LoFtool and GC content change were also initially considered but not selected because they did not improve the prediction performance.

2.4 VEP plugin

We have developed an Ensembl VEP (McLaren et al., 2016) plugin which integrates functionalities of our algorithm to VEP. The VEP plugin allows direct analysis of VCF file using the VEP database and services with a common API to existing VEP plugins or pipelines. The plugin is written in Perl based on 'BaseVepPlugin' interface recommended by Ensembl. During the analysis, the plugin executes the following steps: It matches corresponding exons for each variant, obtains reference and alternative sequences using VEP

APIs, sends those sequences to MMSplice python package with standard input and fetches associated scores from the standard output. We found this to be the simplest way for the plugin to communicate with the MMSplice python package. Moreover, a Docker container that contains all the dependencies including VEP is provided to facilitate installation and usage of the plugin at https://github.com/gagneurlab/MMSplice/tree/master/VEP_plugin.

2.5 Evaluation

For all regression tasks, we chose the Pearson correlation (R) as the primary evaluation metric. However, as the Pearson correlation is invariant to affine transformations, we also report the root-mean-square errors (RMSE), which measures the deviation between predicted values and measured ones. For all classification tasks, we report the precision-recall curve and the area under the curve (auPR) for the cases where there is a strong class imbalance. For the cases where the classes are balanced, we chose to use Receiver operating characteristic (ROC) curve and report area under the ROC curve (auROC).

3 Results

3.1 Training performance of modular models

The donor and acceptor modules were trained by classifying annotated splice sites versus random sequence selected around annotated splice sites of the GENCODE 24 genome annotation (Cheng et al., 2019). Both modules were able to distinguish annotated splice sites with high accuracy on the validation data set (auROC=0.98 for both donor and acceptor modules, Fig. 2A,B).

We evaluated our exon modules and intron modules on predicting Ψ_5 and Ψ_3 measured by the MPRA experiment (Cheng et al., 2019; Soemedi et al., 2017). Our 3' exon module and 5' intron module predicted Ψ_3 for the A5SS library with correlation of 0.77 and 0.31 respectively (Fig. 2C,D). Note that all the predictions were done with a single module ignoring all other information. This approach is not comparable to the Rosenberg et al (Rosenberg et al., 2015) approach which used complete sequence information.

3.2 Vex-seq data does not support additive variant effects on the natural scale

The Vex-seq challenge requested to predict Ψ . However, Ψ is bounded to [0,1]. This constrains the predictions. For instance, for a reference Ψ close to 1, Ψ cannot be largely positive. The CAGI 5 organizers therefore also provided the reference Ψ level. MMSplice models additive effects in the log odds scale ($\text{logit}(\Psi)$, Eq. 2, Methods). Application of the logistic function ensures the predictions of the alternative Ψ to be bounded to the [0,1] interval. An alternative approach would have been to model additive effects in the natural scale and to cap all predictions to the [0,1] interval.

We investigated whether the Vex-seq data would support the additive natural scale model. To this end, we looked first at all Vex-seq data for which i) the reference Ψ level was lower than 0.5 and ii) Ψ was positive. If the effects of variants were additive in the natural scale and independent of the reference Ψ level, then we would expect larger deviations for the constructs with Ψ_{ref} close to 0 as they can increase by as much as 1, compare to constructs

with Ψ_{ref} close to 0.5, which are bounded to increase by not more than 0.5. In fact, we observed the opposite trend as Ψ values for variants with Ψ_{ref} close to 0 were significantly smaller ($P = 2.2 \times 10^{-8}$, Fig. 3A). The same was also observed for the Vex-seq data for which i) the reference Ψ level was larger than 0.5 and ii) Ψ was negative (Fig. 3B). Hence, the effects of variants appeared to be larger for Ψ_{ref} close to 0.5 than for Ψ_{ref} close to 0 or 1. This observation further motivated modeling Ψ as a result of the logistic function, which has smallest gradient around 0 and 1 and largest gradient at 0.5.

3.3 Vex-seq challenge: predicting variant effect on exon skipping level

We trained a linear model from the modular predictions to predict Ψ from the 957 training variants provided by Vex-seq challenge. As the Vex-seq variants originated from both intron and exon, five potentially overlapped modules were used: 3' intron, acceptor, 5' exon, donor, and 5' intron. We used 5' exon module instead of 3' exon module because it performed better on the Vex-seq training data (McLaren et al., 2016). In total, 9 parameters were trained (Methods).

On the Vex-seq training data, MMSplice was able to score all variants including indels. When separating the variants into 3' intron, exon and 5' intron, MMSplice had a good performance in all three regions (3' intron: $R=0.78$, $\text{RMSE}=0.09$; Exon: $R=0.61$, $\text{RMSE}=0.11$; 5' intron: $R=0.72$, $\text{RMSE}=0.11$, Fig. 4, Supp. Table S1).

On the unseen test data, MMSplice had similar performance compared to the training data ($R=0.68$, $\text{RMSE}=0.1$) (Cheng et al., 2019), indicating that we did not overfit the training data. Moreover, we outperformed the state-of-the-art methods SPANR ($R=0.26$, $\text{RMSE}=0.14$) (Xiong et al., 2015) and HAL ($R=0.44$, $\text{RMSE}=0.28$) (Rosenberg et al., 2015), showing that the modular approach that leveraging on complimentary data was effective (Cheng et al., 2019). With this model, we were ranked the first on the Vex-seq challenge.

3.4 MaPSy challenge

Encouraged by the results on Vex-seq data, we trained linear models similarly on the MaPSy challenge training data for *in vivo* and *in vitro* separately with log-allelic ratio as response variable (Methods). We first focused on training MMSplice for predicting the log-allelic ratio (splicing efficiency change, Methods). On the training data, MMSplice accurately predicted variant effects on splicing efficiency both *in vivo* ($R=0.59$, $\text{RMSE}=1.02$) and *in vitro* ($R=0.56$, $\text{RMSE}=0.04$) (Fig. 5A,B) (Supp. Table S2). On the unseen test data, MMSplice was still accurate (Cheng et al., 2019). Our log-allelic ratio prediction was the most accurate one in the MaPSy challenge.

We then trained a classifier to classify exon splicing mutations (ESMs) (Methods). On the training data, the classifier had auPR 0.3 (Fig. 5C). On the unseen test variants, the classifier had auPR 0.19 (Fig. 5C) (Supp. Table S2).

3.5 Variant interpretation

To support the interpretation of the predictions made by MMSplice, we followed the *in silico* mutagenesis approach. *In silico* mutagenesis computes predictions for every possible single

nucleotide variant for a given input sequence, and display the predictions in a heatmap called mutation map. The mutation map allows assessing the relative importance of variants compared to other possible variants in the vicinity. The MMSplice implementation followed the Kipoi API (version 0.65), a programmatic standard for predictive models in genomics (Avsec et al., 2018). In particular, it is compatible with the Kipoi variant effect prediction plugin allowing the generation of mutation maps. As an illustrative example we considered the variant (rs746677712, Fig. 6A). This variant lies 5 nucleotides inside the intron near the donor site of exon 5 of the gene *FCGR2B*. MMSplice predicts this variant to increase the skipping of this exon compared to reference sequence with an odds ratio of 0.14 (log odds ratio = -1.99). The mutation maps also shows that, for the considered sequence, only single nucleotide variants on the canonical 5' dinucleotide GT or the last two bases of the exon AG can lead to effects on exon skipping of similar amplitude (Fig. 6A). Similarly, the variant rs773534127 close to the acceptor of exon 5 was predicted to strongly decrease exon inclusion level with an odds ratio of 0.20 (log odds ratio=-1.59), nearly as strong as the predicted effect of single nucleotide variants on the canonical dinucleotide AG (Fig. 6B). Mutation maps are also useful to identify possible splicing regulatory elements as consecutive nucleotides that are predicted to have a strong impact on splicing when mutated. One illustrative example is provided with the mutation map around the variant rs751723286 (Fig. 6C). This single nucleotide variant affects the motif TAGGG, which is the binding site of Heterogeneous Nuclear Ribonucleoprotein A1 (HNRNPA1), an import splicing regulatory RNA binding protein (Burd & Dreyfuss, 1994). The mutation map shows that every mutation on this motif is predicted to increase exon inclusion level, consistent with that HNRNPA1 prevent exon inclusion (Mayeda & Krainer, 1992).

4 Conclusion

We have participated in two CAGI 5 splicing prediction challenges, Vex-Seq and MaPSy, with a single modeling framework MMSplice which ranked among the best on both challenges. The reasons for the success of MMSplice are multiple. First, we have trained the model mostly on richer complementary functional genomics data with about three orders of magnitude more data points than the CAGI challenge data. We have used the CAGI data only to fit a very few number of parameters for each model. Second, we have worked on the log-odds scale rather than on the natural scale. This was not only justified by mathematical convenience but also by the Vex-seq data which showed that the higher impacts of variants were found for intermediate levels of splicing. Third, we made use of an existing high-throughput perturbation assay (Rosenberg et al., 2015) to fit the model. Since the number of publications with massively parallel reporter assays keeps increasing, such datasets will play a major role in building predictive models for genomics in the future as they allow capturing causal effects. Fourth, we have used a modular approach so that we could re-use elements of the model for one challenge in the other challenge

Depending on the assay and the genomic region of the variants, the correlation between MMSplice predicted changes and the measured changes varied between 0.56 (*in vitro* MaPSy) and 0.78 (3' intronic variants for Vex-seq). The effects of many assayed variants are small, and therefore cannot be precisely predicted because their estimates are likely dominated by noise. Moreover, MMSplice might be improved by overcoming following

limitations: First, Ψ is also be affected by transcript stability (half-life time) of different isoforms. Second, the effect of certain splicing motifs dependent on the position with respect to splice sites [CITE], which we did not model. Third, our exon and intron modules were learned from alternative 5' and 3' splicing events instead of exon skipping directly. Hence, they may not capture fully the biology of exon skipping. Fourth, as splicing is tissue-specific, a model that is specific for the target tissue or cell type might perform better.

Our current model will likely, and hopefully, be soon superseded by future models integrating more data. However, we hope that some of the principles identified here will be useful. In particular, we believe that if models would adopt a modular structure and satisfy some reasonable degree of compatibility, the community could more efficiently leverage models from each other. We provide MMSplice and all the individual modules in the model repository Kipoi, which could be helpful to this end.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650. J.C. was supported by the Competence Network for Technical, Scientific High Performance Computing in Bavaria KONWIHR. Z.A. and J.C. were supported by a Deutsche Forschungsgemeinschaft fellowship through the Graduate School of Quantitative Biosciences Munich.

References

- Adamson SI, Zhan L, & Graveley BR (2018). Vex-seq: High-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biology*, 19(1), 1–12. 10.1186/s13059-018-1437-x [PubMed: 29301551]
- Alberts B, Johnson A, Lewis J, Raff M, Roberts K, & And Walter P (2008). *Molecular Biology of the Cell. Amino Acids* (Vol. 54). 10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C
- Avsec Ž, Kreuzhuber R, Israeli J, Xu N, Cheng J, Shrikumar A, ... Gagneur J (2018). Kipoi: accelerating the community exchange and reuse of predictive models for genomics. *BioRxiv*, 1–31. 10.1101/375345
- Baralle FE, & Giudice J (2017). Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, 18(7), 437–451. 10.1038/nrm.2017.27 [PubMed: 28488700]
- Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, ... Frey BJ (2010). Deciphering the splicing code. *Nature*, 465(7294), 53–59. 10.1038/nature09000 [PubMed: 20445623]
- Braberg H, Jin H, Moehle EA, Chan YA, Wang S, Shales M, ... Krogan NJ (2013). From Structure to Systems: High-Resolution, Quantitative Genetic Analysis of RNA Polymerase II. *Cell*, 154(4), 775–788. 10.1016/j.cell.2013.07.033 [PubMed: 23932120]
- Bretschneider H, Gandhi S, Deshwar AG, Zuberi K, & Frey BJ (2018). COSSMO: predicting competitive alternative splice site selection using deep learning. *Bioinformatics* (Oxford, England), 34(13), i429–i437. 10.1093/bioinformatics/bty244
- Burd CG, & Dreyfuss G (1994). RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing. *The EMBO Journal*, 13(5), 1197–204. [PubMed: 7510636]

- Cheng J, Nguyen TYD, Cygan KJ, Çelik MH, Fairbrother WG, Avsec Ž, & Gagneur J (2019). MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biology*, 20(1), 48. 10.1186/s13059-019-1653-z [PubMed: 30823901]
- Consortium TG, Ardlie K, Deluca DS, Segre AV, Sullivan TJ, Young TR, ... Lockhart NC (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235), 648–660. 10.1126/science.1262110 [PubMed: 25954001]
- Fairbrother WG (2002). Predictive Identification of Exonic Splicing Enhancers in Human Genes. *Science*, 297(5583), 1007–1013. 10.1126/science.1073774 [PubMed: 12114529]
- Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, & Burge CB (2004). RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Research*, 32(Web Server issue), W187–90. 10.1093/nar/gkh393 [PubMed: 15215377]
- Goldstein LD, Cao Y, Pau G, Lawrence M, Wu TD, Seshagiri S, & Gentleman R (2016). Prediction and quantification of splice events from RNA-seq data. *PLoS ONE*, 11(5), 1–18. 10.1371/journal.pone.0156132
- Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, ... Pritchard JK (2016). RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285), 600–604. 10.1126/science.aad9417 [PubMed: 27126046]
- 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. 10.1016/j.febslet.2005.02.047 [PubMed: 15792793]
- Mayeda A, & Krainer AR (1992). Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2. *Cell*, 68(2), 365–75. 10.1016/0092-8674(92)90477-T [PubMed: 1531115]
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, ... Cunningham F (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1). 10.1186/s13059-016-0974-4
- Paggi JM, & Bejerano G (2018). A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *RNA*, 24(12), 1647–1658. 10.1261/rna.066290.118 [PubMed: 30224349]
- Rosenberg AB, Patwardhan RP, Shendure J, & Seelig G (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*, 163(3), 698–711. 10.1016/j.cell.2015.09.054 [PubMed: 26496609]
- Soemedi R, Cygan KJ, Rhine CL, Wang J, Bulacan C, Yang J, ... Fairbrother WG (2017). Pathogenic variants that alter protein code often disrupt splicing. *Nature Genetics*, 49(6), 848–855. 10.1038/ng.3837 [PubMed: 28416821]
- Wang Y, Liu J, Huang BO, Xu Y-M, Li J, Huang L-F, ... Wang X-Z (2015). Mechanism of alternative splicing and its regulation. *Biomedical Reports*, 3(2), 152–158. 10.3892/br.2014.407 [PubMed: 25798239]
- Wang Z, Xiao X, Van Nostrand E, & Burge CB (2006). General and Specific Functions of Exonic Splicing Silencers in Splicing Control. *Molecular Cell*, 23(1), 61–70. 10.1016/j.molcel.2006.05.018 [PubMed: 16797197]
- Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, ... Bähler J (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199), 1239–1243. 10.1038/nature07002 [PubMed: 18488015]
- Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, ... Frey BJ (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218), 1254806–1254806. 10.1126/science.1254806
- Zhang XHF, & Chasin LA (2004). Computational definition of sequence motifs governing constitutive exon splicing. *Genes and Development*, 18(11), 1241–1250. 10.1101/gad.1195304 [PubMed: 15145827]
- Zhang X, & Kangsamaksin T (2005). Exon inclusion is dependent on predictable exonic splicing enhancers. *Molecular and Cellular Biology*, 25(16), 7323–7332. 10.1128/mcb.25.16.7323-7332.2005 [PubMed: 16055740]

Figure 1.
MMSplice model for Vex-seq and MaPSy challenges.

Figure 2.

Performance of individual MMSplice modules. (A) ROC curve for the donor module on the evaluation data. (B) ROC curve for the acceptor module on the evaluation data. (C) Predicted (y-axis) versus measured (x-axis) Ψ_3 on the evaluation data from the splicing MPRA (Rosenberg et al., 2015) with the 3' exon module. (D) Predicted (y-axis) versus measured (x-axis) Ψ_3 on the evaluation data from the splicing MPRA (Rosenberg et al., 2015) with the 5' intron module.

Figure 3.

Difference of Ψ depends on reference Ψ . (A) Boxplot Ψ change (y-axis) on different bins of reference Ψ level (x-axis) for variants with reference Ψ smaller than 0.5 and Ψ positive. (B) Boxplot Ψ change (y-axis) on different bins of reference Ψ level (x-axis) for variants with reference Ψ greater than 0.5 and Ψ negative. P-values were calculated by the Mann–Whitney U test.

Figure 4. Evaluation of MMSplice predicting Ψ on Vex-seq training data for variants in 3' intron (left), exon (middle) and right (5' intron). Predicted Ψ (x-axis) versus measured Ψ (y-axis). The dotted line marks the $y=x$ diagonal.

Figure 5. Evaluation MMSplice on MaPSy. Scatter plot of predicted splicing efficiency change versus measured splicing efficiency change for *in vivo* (A) and *in vitro* (B) training data. (C) Precision-Recall curve of classifying exon splicing mutations (ESMs) on MaPSy test data.

Figure 6.

In-silico mutagenesis analysis of example MMSplice predictions. Red color indicates variant increase Ψ while blue color indicates variant decrease Ψ . Alphabet letter height indicates effect magnitude. Gray bars are gene structure schema, thick ones are exons while thin ones are introns. (A) G to C mutation close to the donor site. (B) C to A mutation close to the acceptor site. (C) Exonic G to A mutation.