Received: 28 January 2019 | Revised: 20 June 2019 | Accepted: 10 July 2019

DOI: 10.1002/humu.23869

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

SPECIAL ARTICLE

Assessing predictions of the impact of variants on splicing in CAGI5

Stephen M. Mount[1] | Žiga Avsec[2] | Liran Carmel[3] | Rita Casadio[4] | Muhammed Hasan Çelik[2] | Ken Chen[5] | Jun Cheng[2] | Noa E. Cohen[3,6] | William G. Fairbrother[7] | Tzila Fenesh[8] | Julien Gagneur[2] | Valer Gotea[9] | Tamar Holzer[8] | Chiao-Feng Lin[10] | Pier Luigi Martelli[4] | Tatsuhiko Naito[11] | Thi Yen Duong Nguyen[2] | Castrense Savojardo[4] | Ron Unger[8] | Robert Wang[12,13] | Yuedong Yang[5] | Huiying Zhao[14]

[1]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland

[2]Department of Informatics, Technical University of Munich, Garching, Germany

[3]Department of Genetics, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

[4]Department of Pharmacy and Biotechnology, Biocomputing Group, University of Bologna, Bologna, Italy

[5]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

[6]The integrated program for Computer Science and Computational Biology, School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel

[7]Department of Molecular Biology, Cell Biology, and Biochemistry, Center For Computational Biology, Brown University, Providence, Rhode Island

[8]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel

[9]National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH), Bethesda, Maryland

[10]Translational Informatics, DNAnexus, Mountain View, California

[11]Department of Neurology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[12]Department of Bioengineering, University of California, Berkeley, California

[13]Department of Plant and Molecular Biology, University of California, Berkeley, California

[14]Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China

Correspondence
Stephen M. Mount, Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, College Park, MD 20742.
Email: smount@umd.edu

Funding information
National Human Genome Research Institute, Grant/Award Numbers: U41 HG007346, R13 HG006650; National Science Foundation, Division of BiologicalInfrastructure, Grant/Award Number: ABI 1564785

Abstract

Precision medicine and sequence-based clinical diagnostics seek to predict disease risk or to identify causative variants from sequencing data. The Critical Assessment of Genome Interpretation (CAGI) is a community experiment consisting of genotype-phenotype prediction challenges; participants build models, undergo assessment, and share key findings. In the past, few CAGI challenges have addressed the impact of sequence variants on splicing. In CAGI5, two challenges (Vex-seq and MaPSY) involved prediction of the effect of variants, primarily single-nucleotide changes, on splicing. Although there are significant differences between these two challenges, both involved prediction of results from high-throughput exon inclusion assays. Here, we discuss the methods used to predict the impact of these variants on splicing, their performance, strengths, and weaknesses, and prospects for predicting the impact of sequence variation on splicing and disease phenotypes.

KEYWORDS

CAGI experiment, machine learning, mutation, splicing, variant interpretation

1 | INTRODUCTION

A significant fraction of deleterious mutations in protein-coding genes act through an effect on splicing. Many of these mutations directly impact canonical splice-site dinucleotides (GT at the 5′ splice site and AG at the 3′ splice site) and are routinely scored as splicing mutations. Mutations within the splice-site consensus region (up to three exon nucleotides, and several intron nucleotides, including the canonical dinucleotides) can also have a large effect on splicing. Estimates of the fraction of deleterious mutations in this class range from 9% (Human Gene Mutation Database, Stenson et al., 2017) to 22% (MacArthur et al., 2012). Such mutations are reliably assessed by methods such as MaxEnt (Yeo & Burge, 2004) or information theory (Rogan, Svojanovsky, & Leeder, 2003), and are often included in assessment of potential impact. An even larger set of variants act through auxiliary splicing signals in the exon or flanking introns (Cheung et al., 2019; Ladd & Cooper, 2002; Li et al., 2016,), and as many as 37% of GWAS variants are potentially associated with an effect on splicing (Pal, Yu, Mount, & Moult, 2015). Although it has been known for some time that mutations in auxiliary splicing signals can affect splicing (e.g., Cartegni, Chew, & Krainer, 2002, Yang, Swaminathan, Martin, & Sharan, 2003), they are often ignored when assessing the potential impact of mutations, and the impact of exonic mutations resulting in monogenic disease through an effect on splicing is often attributed to their dual identity as missense mutations (e.g. Zucker et al., 2011).

Recent work with high-throughput assays has identified variants affecting splicing and has established the importance of auxiliary splicing signals. Soemedi et al. (2017) observe that 10% of disease-causing exonic mutations alter splicing in a massively parallel splicing assay (MaPSy). An independent high-throughput assessment of 27,733 human variants (Cheung et al., 2019) from ExAC (Lek et al., 2016) found that 3.8% of rare variants had a high impact on splicing. Eighty-three percent of these lie outside of the splice-site dinucleotides and 62% lie outside of the splice-site consensus region (Cheung et al., 2019), and therefore, in auxiliary splicing signals. Rosenberg, Patwardhan, Shendure, and Seelig (2015) measured the splicing of over 2 million synthetic minigenes incorporating random sequence in splice-site competition assays, and found that the majority of hexamer sequence motifs act as auxiliary splicing signals. Such high-throughput assays demonstrate that many rare variants have the potential to impact splicing. These same assays also provide data sets that facilitate the training of computational tools to predict the impact of variants on splicing.

Mutations in auxiliary splicing signals can alter splicing and therefore, cause disease. Indeed, mutations in auxiliary splicing signals were among the very first mutations with described effects on splicing (Orkin et al., 1982; Mount & Steitz, 1983). However, the identification of mutations in auxiliary splicing signals is not standard because of various factors that make variants affecting auxiliary splicing signals harder to predict relative to splice-site dinucleotides and their surrounding bases, which are by definition fixed in position, and which have greater information content. Two CAGI5 challenges (Vex-seq and MaPSy) directly assess the ability to predict the impact of sequence variants on splicing.

1.1 | Description of the Vex-Seq challenge

The Vex-Seq (variant exon sequencing) method (Adamson, Zhan, & Graveley, 2018) is an in vivo splicing assay that exploits a barcode to track specific variants in a high-throughput assay of exon inclusion (see Figure 1), using a unique molecular identifier to track PCR duplicates. Two cell lines (K562 and HepG2) were assayed for each construct. PSI ("percent spliced in" or Ψ; 100 times the ratio between inclusion reads and the sum of inclusion reads and exclusion reads) was calculated from RNA-seq data for each allele, and delta-Ψ (ΔΨ; the difference between Ψ for the variant and Ψ for the reference) was reported for a training set. Predictors were asked to predict delta-Ψ for the test set of 1,098 variants. Variants were located throughout the affected exon, in its splice sites, and up to 100 nt into flanking introns.

1.2 | Description of the MaPSy challenge

The massively parallel-splicing assay (MaPSy, Soemedi et al., 2017) approach was used to screen 797 reported exonic disease mutations. MaPSy involves a comparison of the ratio of mutant and wild-type (wt) alleles before ("input") and after ("output") a high-throughput splicing assay (see Figure 1). Assays were both "in vivo", via transfection of HEK cells in culture, or "in vitro", in a nuclear extract, using a mini-gene system. The challenge is to predict the degree to which a given variant causes changes in splicing. For the purposes of this challenge, variants were categorized as exonic splicing mutations (ESMs) if they both changed the allelic ratio by 1.5-fold or more and passed a two-sided Fisher's exact test (FET) with a false discovery rate (FDR) of 5% both in vivo and in vitro. Thus, the MaPSy challenge, unlike Vex-seq, was formulated as classification task. Predictors were asked to provide a probability between 0 and 1 that a particular variant was classified as an ESM.

1.3 | A comparison of these challenges

Although these two challenges are essentially similar in that they measure the effect of sequence variants on exon inclusion in an MaPSy, there are important differences.

(a) The MaPSy challenge is inherently categorical in that predictors were asked to say whether or not a particular variant was an exonic splicing mutation. In contrast, the Vex-seq challenge asked for a quantitative measure (ΔΨ) of the impact of the variant on splicing.

(b) The MaPSy challenge is limited to exonic variants, because the variant itself is used to distinguish variant from reference sequences in both input and output. In contrast, because the

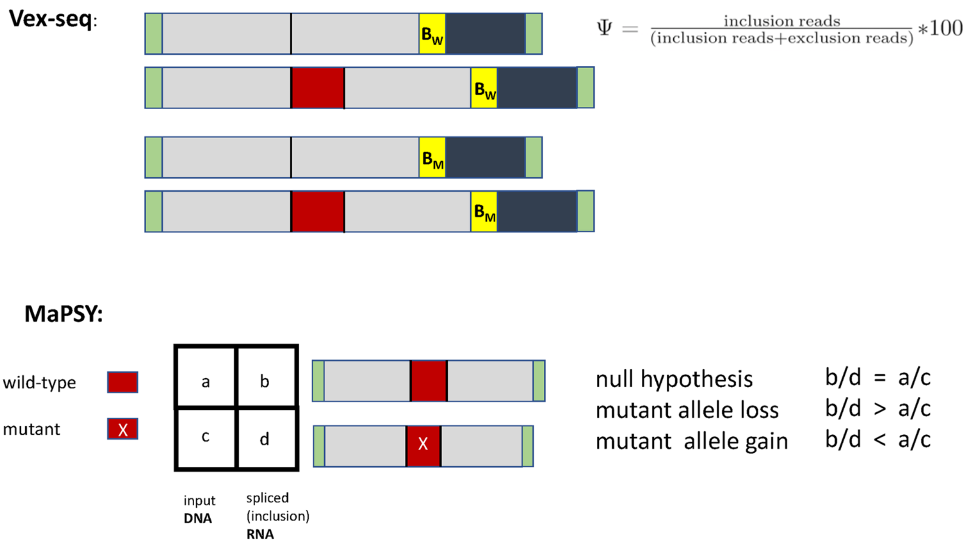
Vex-seq:
B_W
B_W
B_M
B_M
$\Psi = \frac{\text{inclusion reads}}{\text{(inclusion reads+exclusion reads)}} *100$
MaPSY:
wild-type
mutant
X
a
b
c
d
X
input DNA
spliced (inclusion) RNA
null hypothesis b/d = a/c
mutant allele loss b/d > a/c
mutant allele gain b/d < a/c

FIGURE 1 Schematic comparison of the MaPSy and Vex-seq methods. Both methods measure the efficiency of exon inclusion using high-throughput sequencing of PCR products generated by flanking primers (light green) in shared flanking exons. In the case of the Vex-seq challenge (top), inclusion of exons from wild-type and mutant clones are distinguished by barcodes (B_W and B_M, respectively); this allows the efficiency of exon inclusion to be measured for mutations that reside in intron sequence not present in the final PCR product. In the case of the MaPSy challenge (bottom), effects of a mutant on splicing are inferred by comparing the ratio of sequencing counts derived from wild-type (b) and mutant (d) included exons. This ratio (b/d) is compared to the ratio observed for input DNA (a/c). For the in vivo assay, the sequence pairs were incorporated into three-exon minigenes and transfected into HEK293 cells. For the in vitro assay, the library was incorporated into two-exon constructs and incubated in HeLa nuclear extract so that splicing could occur. MaPSy, massively parallel-splicing assay; PCR, polymerase chain reaction; Vex-seq, variant exon sequencing

Vex-seq challenge exploits barcodes in the downstream exon to tag products, intronic variants can be assayed and were included.

(c) The Vex-seq challenge compares the observed support for exon inclusion isoforms with the support for exon-skipping isoforms and ignores all other isoforms. In contrast, the MaPSy challenge evaluates splicing efficiency based solely on the number of correctly spliced (inclusion isoform) reads; all other outcomes (exon skipping, no splicing, RNA degradation) are treated equally because they do not result in reads that are counted.

(d) The MaPSy challenge is designed to measure damaging mutations that cause skipping of a constitutive exon whereas the Vex-seq challenge is designed to measure the rate of inclusion of an alternatively spliced exon.

2 | METHODS

2.1 | Vex-seq data

Training data and answers were provided as $\Delta\Psi$ and mean Ψ of the HepG2 cell assay.

The evaluation metrics below were chosen based on exploratory analysis in which all predictions were compared, to one another and to the "answers," using heatmaps (not shown).

2.2 | Classification of Vex-seq variants

To see whether specific prediction models performed better on variants affecting particular locations relative to exon boundaries, the 1,098 variants were divided into three categories (intron, exon, or splice-site variants; splice sites included three exon and eight intron nucleotides immediately adjacent to the splice site). Specific variants (222 of 1,098) were designated "Hard" if the root-mean square deviation (RMSD) for all predictions of the effect of that variant exceeded 13.4, the standard deviation of true $\Delta\Psi$ values for all variants. Finally, variants were categorized as positive (71 variants), neutral (947 variants), or negative (80 variants) according to whether their observed $\Delta\Psi$ value differed from the mean value of −1.84 by more than a standard deviation.

2.3 | MaPSy data

MaPSy data provided consists of read counts from input DNA and from correctly spliced cDNA.

2.4 | MaPSy predictions

Predictors were asked to provide predictions of the count of reads for in vivo wt spliced, in vivo mutant spliced, in vitro wt spliced, and in vitro mutant; predictions of the ratios observed, the standard deviation for this ratio (as a measure of uncertainty of the prediction), ESM (the probability that this variant is an ESM), and the standard deviation of the ESM.

2.5 | MaPSy exploratory data analysis

Count values for the "answers" were compared using two two-by-two tables of values (wt spliced, mutant spliced; wt input and mutant input), one table for the in vivo experiment and one table for the in vitro experiment. Each was used to calculate *p* values using one-sided FETs. All predictions (the probability, between 0 and 1, that a particular variant was classified as an ESM) were then compared to these four *p* values, and to six composite values: (a) the minimum of the two in vitro *p* values, (b) the minimum of the two in vivo *p* values, (c) the maximum of the mutant *p* values, (d) the maximum of the wt *p* values, (e) the maximum of the two minima (a and b above) and the minimum of the two maxima (c and d above). In addition, all predictions were compared, to one another and to the "answers," using heatmaps (not shown). The evaluations below were chosen based on this exploratory analysis.

2.6 | MaPSy evaluations

Predictions (ESM probabilities) were evaluated using several distinct "answers."

"Official": The answers as provided by the data provider, which "passed the 1.5-fold change and a two-sided FET adjusted with 5% FDR both in vitro and in vivo." Forty-four of 796 variants were considered "official" ESMs.

"Consistent": Cases where the in vivo and in vitro assays indicate differential directional effects on splicing were removed from the official set of ESMs. Twent-six variants were considered "consistent" ESMs. Because the number of official ESMs that are not consistent is surprisingly large (13 in the training set), predictors were instructed that "those should not be regarded as ESMs and be ignored."

"Mutant only": Only cases in the consistent set in which the mutant resulted in reduced splicing both in vivo and in vitro (19 variants) were considered consistent exonic splicing mutations with the expected effect on splicing.

In addition, FET (a one-sided test of the hypothesis that the mutant affects splicing) was used without a fold-change criterion to generate four additional sets of ESMs from the counts.

"FET-vivo-1e−02"—ESM if FET for the mutant case in vivo is less than 10^{-2}

"FET-vitro-1e−02"—ESM if FET for the mutant case in vitro is less than 5×10^{-6}

"FET-vivo-5e−06"—ESM if FET for the mutant case in vivo is less than 10^{-2}

"FET-vitro-5e−06"—ESM if FET for the mutant case in vitro is less than 5×10^{-6}

Predictions were evaluated relative to these seven sets of ESMs by three measures: overall agreement, area under a receiver operating characteristic (ROC) curve, and odds ratio. Overall agreement was calculated as (1−class) × (1−prediction) + (class) × (prediction). Area under ROC curve was calculated using the R package ROCR. The odds ratio was calculated as described below.

2.7 | Odds ratio

An odds ratio was calculated as (*a*/*c*)/(*b*/*d*), where *a* is the number of correctly predicted ESMs; *b* is the number of variants predicted to be ESMs but were not ESMs; *c* is the count of ESMs that were not predicted; and *d* are cases that were neither predicted nor ESMs. These values were calculated as ***a*:** (ESM) × (Pred); ***b*:** (1−ESM) × (Pred); ***c*:** (ESM) × (1−pred); and ***d*:** (1−ESM) × (1−pred), where ESM is the status of the variant (1 or 0) and Pred is the prediction probability.

2.8 | Data

Data were made available by the data providers Brent Graveley (Vex-seq) and Will Fairbrother (MaPSy) subject to the CAGI data use agreement (see https://genomeinterpretation.org/data-use-agreement).

3 | RESULTS

3.1 | VexSeq predictions

The predictors are listed in Table 1a, which briefly summarizes the methods used by each team for the Vex-seq challenge. Because most predictors relied heavily on subsidiary methods and data sources for features, subsidiary features are listed separately in Table 1b.,

3.2 | Performance of VexSeq predictions

Predictions were evaluated by correlations with the $\Delta\Psi$ values provided (Tables 2a–c) and by calculation of RMSD (Table 2d). Because the $\Delta\Psi$ values submitted by team 3 were generally small, they were multiplied by 100 before calculation of RMSD (correcting for assumed confusion between proportion and percentage).

Remarkably, predictions from team 3 performed best by all four measures, in every classification of variants (all variants, intronic variants, exonic variants, splice-site variants, and hard cases), and when predictions were scored categorically (as positive, neutral or negative). The classification of variants revealed some interesting features of the predictions. Groups 3 and 6 stand out as far superior to the others with regard to intron variants in particular. Group 2 performed second-best overall and was particularly strong on variants classified as hard. As expected, splice-site variants generally had a much larger effect ($\Delta\Psi$ RMSD of about 25, vs. about 10 for exon and intron variants). Interestingly, some predictions (e.g., team 6) appeared to perform better by correlation measures than by RMSD, suggesting that whereas the direction of the effect was correctly assessed, the degree was not. Conversely, team 2 performed relatively better on splice sites by the measure of RMSD.

3.3 | MaPSy predictions

The predictors are listed in Table 3, which briefly summarizes the methods each used for the MaPSy challenge. Because most predictors relied heavily on subsidiary methods for features, all subsidiary features, and citations for secondary methods and data sources, are listed separately (Table 1b).

TABLE 1a Summary of prediction methods on the Vex-seq challenge

Group code	Summary of Vex-seq method. Citations for subsidiary methods and data sources are listed below the table
1 JiLin	Machine learning (SVM) on base substitution by position. 1.2 differed from 1.1 in that the sequence was not included, only variant position relative to exon boundaries.
2 Sun Yat-sen	Machine learning (SVM) on secondary features from the DDIG-SN and SilVA, and from the CADD server. ANNOVAR was used to extract ΔΨ values from SPANR. Additional features from a variety of sources. A greedy feature selection strategy was used to find the most effective features.
3 Munich	A novel, compositional, model (MMSplice) was introduced in which five distinct models were used for intron adjacent to the 3′ splice site, the 3′ splice site, the exon, the 5′ splice site, and intron adjacent to the 5′ splice site. These were separately scored and used to generate a composite exon-skipping score. The splice-site models were obtained using convolutional neural network models. The exon and intron models were based on data from MPRA. Distinct models were used for intron adjacent to the 3′ splice site and adjacent to the 5′ splice site. Vex-seq training data were used for module assembly. 3.2 was generated from the sole submission (designated 3.1) by multiplying all values by 100.
4 Berkeley	A random forest regression model was trained using a large number of features, including total allele frequency from the gnomAD database of allele frequencies, MaxEntScan, and many binding site models for specific known splicing regulators.
5 Bar Ilan	Training was done on features listed by Soemedi et al. (2017), plus additional features (change in strength of ESE and ESS features, and output from the Ex-skip web site.) Specific submissions varied as follows: 5.1: "optimal mode"—Using Decision Tree. 5.2: "equal distribution mode"—The classification was adjusted to reflect the class frequencies in the test set. 5.3: Random Forest was used in the optimal mode. 5.4: Linear regression of features was used to directly predict ΔΨ values. 5.5: Only the class assignment is provided with no numeric scores.
6 NHGRI	Spliceport and SPANR values were used. A weighted linear model was used to predict logit-transformed, thresholded Ψ values as a function of SplicePort acceptor and donor scores, as well as logit-transformed, thresholded values of Ψ provided by the SPANR tool. 5.2 used a multiple linear regression model of experimental ΔΨ values as a function of SPANR predictions of ΔΨ and of changes in SplicePort scores induced by the presence of a variant.

Note: Full group names: HILab-JLU-001, JiLin University, China; biomed-ai-th2, Sun Yat-sen University, China; delta_PSI, Technical University of Munich, I12, Germany, DG; berkeley_bioe_26419652, Univ. of California, Berkeley, USA; Biu, Bar Ilan Univ. in collaboration with the Hebrew University of Jerusalem, Israel; NHGRI_Elnitski, National Human Genome Research Institute, USA.
Abbreviatons: MPRA, massively parallel reporter assay; SVM, support vector machine.

TABLE 1b Subsidiary methods and data sets

Several of the predictors incorporated the results of published methods and data sources (referred to here as subsidiary methods).

ANNOVAR is a software tool that functionally annotates genetic variants based on a large set of subsidiary tools and databases (Wang, Li, & Hakonarson, 2010 and annovar.openbioinformatics.org/).

CADD (Combined Annotation-Dependent Depletion) is a method for integrating many diverse annotations into a single measure (C score) for each variant. (see Kircher et al., 2014; Rentzsch, Witten, Cooper, Shendure, & Kircher, 2019, and cadd.gs.washington.edu/).

DDIN-SN is a support vector machine (SVM) model to discriminate disease-causing synonymous variants trained and evaluated on nearly 900 disease-causing variants, incorporating features from the SPIDEX database (Livingstone et al., 2017 and sparks-lab.org/ddig).

DSSP (Deep Splice-Site Prediction system) is a deep neural network-based model that calculates 5′ and 3′ splice-site probability, respectively from a 140-length base sequence, in which the middle nucleotides represent the consensus sequence. It was trained with nearly 3,000 true and 30,000 false splice sites (Naito, 2018 and omictools.com/dssp-2-tool).

EX-SKIP is simple utility that compares the ESE/ESS profile of a wt and a mutated allele to quickly determine which exonic variant has the highest chance to skip this exon (Raponi et al., 2011 and ex-skip.img.cas.cz/).

gnomAD (Genome Aggregation Database) aggregates exome and genome sequencing data from a wide variety of large scale sequencing projects, currently 125,748 exome sequences and 15,708 whole-genome sequences from unrelated individuals (gnomad.broadinstitute.org/ and Lek et al., 2016).

MaxEntScan (Yeo & Burge, 2004) provides MaxEnt scores based on maximum entropy models that may include dependencies between nonadjacent as well as adjacent positions. Specifically, MaxEntScan provides a score for 5′ splice sites (donor sites) based on nine positions (−3 through +6) and 3′ splice sites (acceptor sites) based on 23 positions (−20 through +3; genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html).

MPRA (massively parallel reporter assay) describes the splicing patterns of over two million synthetic minigenes, including degenerate subsequences totaling over 100,000,000 bases of variation (Rosenberg et al., 2015).

SilVA (the Silent Variant Analyzer) is a tool for the automated harmfulness prediction of synonymous (silent) mutations within the human genome based on a number of features, including conservation, codon usage, splice sites, splicing enhancers and suppressors, and messenger RNA-folding free energy (Buske, Manickaraj, Mital, Ray, & Brudno, 2013 and compbio.cs.toronto.edu/silva/).

SPANR (splicing-based analysis of variants) is described by Xiong et al. (2015) as a computational tool that estimates Ψ based on 1,393 sequence features and observed values of Ψ for human exons and ΔΨ values for 650,000 variants (tools.genes.toronto.edu/).

SPIDEX is a precomputed index of SPANR scores for the human genome.

Spliceport provides a score for the strength of splice sites based on a feature-generation algorithm for classification of GT and AG dinucleotides as splice sites or not. (Dogan, Islamaj, Getoor, Wilbur, & Mount, 2007).

TABLE 2 Summary of Vex-seq prediction performance

	1.1	1.2	2.1	2.2	2.3	3.1	4.1	5.1	5.2	5.3	5.4	5.5	6.1	6.2	Ave
(a) Pearson correlation coefficients per group															
All	0.29	0.13	0.57	0.53	0.55	**0.67**	0.43	0.07	0.10	0.49	0.33	0.07	0.50	0.45	–
Exon	0.12	0.16	0.26	0.12	0.19	**0.41**	0.02	−0.15	−0.13	0.10	0.06	−0.14	0.26	0.26	–
Intron	0.01	0.03	0.11	0.00	0.07	**0.33**	0.11	0.02	−0.01	0.07	0.08	0.04	0.20	0.16	–
SS	0.12	0.16	0.26	0.12	0.19	**0.41**	0.02	−0.15	−0.13	0.10	0.06	−0.14	0.26	0.26	–
Hard	0.42	0.19	0.61	0.59	0.62	**0.72**	0.51	0.09	0.10	0.55	0.42	0.10	0.51	0.44	–
Categ.	0.11	0.09	0.24	0.18	0.22	**0.32**	0.12	−0.02	0.06	0.17	0.10	0.01	0.27	0.25	–
(b) Pearson correlation coefficients—direction (sign) only															
All	0.07	0.10	0.10	0.04	0.09	**0.25**	0.25	0.03	−0.08	−0.06	0.06	0.06	−0.03	0.24	–
Exon	0.06	0.17	0.12	0.04	0.07	0.26	**0.28**	0.03	−0.13	−0.10	0.05	0.03	−0.12	0.22	–
Intron	0.03	0.02	0.01	−0.06	0.05	**0.18**	0.17	−0.03	−0.04	−0.03	0.01	0.03	0.03	**0.18**	–
SS	0.40	−0.09	0.79	0.80	0.82	**0.86**	0.60	0.30	0.29	0.73	0.60	0.32	0.61	0.53	–
Hard	0.12	0.12	0.37	0.25	0.23	**0.48**	0.15	−0.09	−0.08	0.12	0.14	0.02	0.37	0.35	–
(c) Spearman correlation coefficients per group															
All	0.15	0.15	0.20	0.12	0.18	**0.38**	0.07	−0.10	−0.08	0.14	0.12	0.00	0.32	0.29	–
Exon	0.15	0.20	0.16	0.08	0.13	**0.41**	0.00	−0.19	−0.17	0.08	0.05	−0.13	0.28	0.26	–
Intron	0.09	0.07	0.06	−0.04	0.05	**0.24**	0.01	−0.07	−0.08	0.02	0.06	0.03	0.22	0.17	–
SS	0.15	0.20	0.16	0.08	0.13	**0.41**	0.00	−0.19	−0.17	0.08	0.05	−0.13	0.28	0.26	–
Hard	0.24	0.18	0.45	0.30	0.44	**0.60**	0.31	−0.05	−0.03	0.31	0.23	0.05	0.49	0.46	–
(d) RMSD variation per group															
All	13.0	13.5	11.3	11.5	11.4	**9.9**	12.3	20.3	20.1	11.9	13.6	13.5	11.7	12.0	13.4
Exon	12.6	12.5	10.9	11.1	10.8	**10.1**	11.8	20.6	20.1	11.0	13.3	12.5	11.3	11.3	12.4
Intron	10.7	9.8	10.1	10.2	10.4	**9.3**	10.2	15.6	16.3	9.8	11.4	9.6	9.4	9.5	9.7
SS	26.0	30.5	17.4	17.3	16.2	**14.5**	22.4	30.3	32.6	22.1	24.0	30.6	22.9	24.6	30.1
Hard	25.8	28.3	22.1	22.4	21.5	**19.1**	23.8	40.5	40.2	24.1	25.7	28.5	24.1	25.0	28.1

Abbreviations: RMSD, root-mean square deviation; SS, splice site

The MaPSy prediction challenge consisted of assigning probabilities that a particular mutation is an exonic splicing mutation (ESM). In other words, it was set up as a classification task, and the “answers” classified every variant with a 1 or a 0. However, this classification is based on four count values from the in vivo assay and four count values from the in vitro assay (see Figure 1). Thus, it is possible to rank the confidence with which each variant is assigned its status, and to compare predictions to p values calculated directly from the counts, and to composite values based on those p values. On the basis of this initial exploratory data analysis, several distinct “answers” were used, in addition to the “official” answer, for evaluation.

3.4 | MaPSy predictions were evaluated using several distinct “answers”

“Official”— The answers as provided by the data provider, which “passed the 1.5-fold change and a two-sided FET adjusted with 5% FDR both in vitro and in vivo.” Forty-four of 796 variants were considered “official” ESMs.

“Consistent”— Cases where the in vivo and in vitro assays indicate differential directional effects on splicing were removed from the official set of ESMs. Twenty-six variants were considered “consistent” ESMs.

“Mutant only”— Only cases in the consistent set in which the mutant resulted in reduced splicing both *in vivo* and *in vitro* (19 variants) were considered consistent exonic splicing mutations with the expected effect on splicing.

In addition, FET (a one-sided test of the hypothesis that the mutant affects splicing) was used without a fold-change criterion to generate four additional sets of ESMs from the counts.

“FET-vivo-1e−02”— ESM if FET for the mutant case in vivo is less than 10^{-2}

“FET-vitro-1e−02”— ESM if FET for the mutant case in vitro is less than 5×10^{-6}

“FET-vivo-5e−06”— ESM if FET for the mutant case in vivo is less than 10^{-2}

“FET-vitro-5e−06”— ESM if FET for the mutant case in vitro is less than 5×10^{-6}

Predictions were evaluated relative to these seven sets of ESMs by two measures: overall agreement (calculated as [1–class] × [1–prediction] + [class] × [prediction]; Table 4a), and area under an ROC

TABLE 3 Summary of MaPSy method

Group code	Summary of MaPSy method
1. Bar Ilan	(Group 5 of the Vex-seq challenge). Training (WEKA) on literature features listed by Soemedi et al. (2017), plus additional features (change in strength of ESE and ESS features, and output from the Ex-skip web site), plus about 20 features that were extracted from a model trained on a previous database of ESM that they compiled. Specific submissions varied as follows: 1.1: Logistic regression on literature features, adding the results of the prediction based on the database as an additional feature; 1.2: Logistic regression on Literature features only; 1.3: Logistic regression on all features compiled from both literature and database; 1.4: Decision tree on literature features, adding the results of database as an additional feature; 1.5: Decision tree using literature only; 1.6: Decision tree on all features; 1.7 Majority vote on submissions 1–3.
2. Bologna	A Random Forest Classifier (RFC) was trained on a subset of features extracted from Soemedi et al. (2017) and including mutation-, exon- and gene-level descriptors, scaled between 0 and 1, positively correlated with ESM. A convolutional neural network (CNN) was applied directly to the 170-mer sequences. The final ESM output is the average of the RFC and CNN outputs.
3. Tokyo	CNNs were used to learn sequence information. Random forest, XGBoost, and logistic or linear regression were trained on a large number of features, including MaxEntScan, ESRseq, and DSSP. Then, stacked generalization with logistic regression was used to combine these.
4. Munich	(Group 3 of the Vex-seq challenge). A novel, compositional model was introduced in which the two splice sites, the intron. and the exon, were separately scored, followed by a composite splicing efficiency score. The splice-site models were obtained using convolutional neural network models. The exon and intron models were based based on data from MPRA. Distinct models were used for intron adjacent to the 3′ splice site and adjacent to the 5′ splice site. MaPSy training data were used for module assembly.
5. Toronto	Features were MaxEntScan, a neural network trained to recognize exon boundaries from sequence, hexamer scores from MPRA and ESS and ESE motifs. Submissions 5.1 and 5.2 differed in the model architectures.

Note: Full group names: Biu: Bar Ilan Univ., in collaboration with the Hebrew University of Jerusalem; BolognaBiocomputing, University of Bologna, Italy; TN, The University of Tokyo, Japan; delta_PSI (ΔΨ), Technical University of Munich, I12, Germany, DG; University of Toronto, Canada.
Abbreviations: DSSP, Deep Splice-Site Prediction system; ESE, exonic splicing enhancer; ESM, exonic splicing mutation; ESS, exonic splicing silencer; MaPSY, massively parallel-splicing assay; MPRA, massively parallel reporter assay.

TABLE 4 Evaluation of MaPSy predictions

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	2	3	4	5.1	5.2
(a) Overall agreement between different answers and prediction (see text)												
Official	0.702	0.820	0.708	0.773	0.707	0.817	0.743	0.715	0.653	**0.916**	0.914	0.902
Consistent	0.718	0.836	0.722	0.786	0.720	0.832	0.759	0.732	0.666	**0.932**	0.931	0.918
Mutant only	0.724	0.846	0.731	0.782	0.714	0.830	0.767	0.742	0.664	0.943	**0.944**	0.931
FET-vivo-1e−02	0.674	0.774	0.681	0.703	0.651	0.755	0.710	0.677	0.632	0.856	**0.859**	0.851
FET-vitro-1e−02	0.654	0.705	0.641	0.644	0.612	0.674	0.667	0.626	0.624	0.745	**0.749**	0.745
FET-vivo-5e−06	0.697	0.812	0.705	0.748	0.683	0.792	0.738	0.704	0.645	0.903	**0.906**	0.895
FET-vitro-5e−06	0.704	0.796	0.700	0.730	0.677	0.763	0.733	0.693	0.652	0.867	**0.871**	0.862
(b) Area under the curve (AUC) for ROC curves (see text)												
Official	0.466	0.489	0.470	0.531	0.550	0.613	0.466	0.334	0.594	**0.692**	0.638	0.618
Consistent	0.657	0.654	0.589	0.547	0.610	0.681	0.631	0.447	**0.835**	0.794	0.756	0.698
Mutant only	0.699	0.700	0.648	0.438	0.518	0.664	0.703	0.482	**0.807**	0.762	0.726	0.701
FETvivo1e02	0.484	0.487	0.493	0.439	0.457	0.535	0.488	0.379	0.532	0.659	0.647	**0.664**
FETvitro1e02	0.465	0.486	0.486	0.469	0.467	0.539	0.474	0.316	0.531	**0.700**	0.647	0.676
FETvivo5e06	0.628	0.616	0.583	0.485	0.502	0.517	0.614	0.484	0.675	**0.684**	0.673	0.654
FETvitro5e06	0.655	0.633	0.595	0.488	0.517	0.502	0.642	0.458	0.691	**0.753**	0.699	0.692
(c) Odds ratio (see text)												
Official	1.00	1.02	0.84	1.48	1.50	**2.15**	0.94	0.50	1.52	**2.15**	1.85	1.31
Consistent	1.71	1.51	1.18	2.00	2.36	3.47	1.44	0.81	2.75	2.87	2.55	1.80
Mutant only	2.12	1.83	1.65	0.62	1.13	2.01	1.83	0.95	2.75	2.04	**3.06**	2.23

Note: Bold values indicate the maximum score for each answer.
Abbreviations: FET, Fisher's exact test; ROC, receiver operating characteristic.

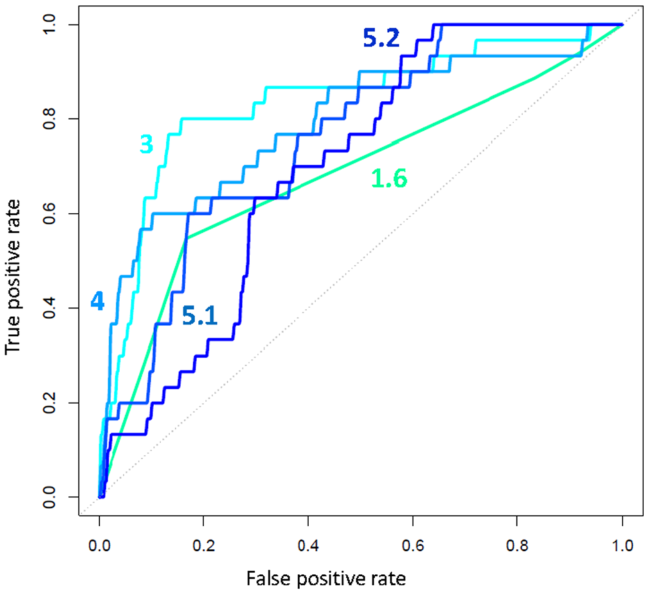

FIGURE 2 Receiver operating characteristic curves for MaPSy predictors on the "consistent" set of ESMs. Each point on the curve reflects a different threshold value for the ESM probability, from which true positive rate (proportion of actual ESMs whose assigned probability is above that threshold) and false positive rate (proportion of actual ESMs whose assigned probability is above that threshold) are calculated. The area under these curves is a measure of the quality of the prediction; the five predictions with the greatest area under these curves (Table 4c) are shown

curve (Table 4b and Figure 2). Finally, an odds ratio was calculated to measure the association between predictions and the "official," "consistent," and "mutant only" "answers" (Table 4c).

All predictors, without exception, and by all five measures, did better on the consistent ESMs than on the official ESMs, and the rank order of prediction scores was similar (Table 4). This is in line with the instruction in the challenge that inconsistent results in the training set should be ignored. Thus, the "consistent" set should be regarded as definitive. For this set, groups 3 and 4 performed best (depending on the measure), with group 5 as a close third.

When ESMs were limited to consistent cases where the mutation reduces inclusion (as would be expected for a true exonic splicing mutation in a constitutive exon), groups 3–5 again performed best (which varied by the measure used). When an odds ratio was calculated (Table 4c), group 1.6 (Bar Ilan, decision tree) performed best.

4 | DISCUSSION

4.1 | Decomposition based on variant location likely improves predictions

In the Vex-seq challenge, predictions from group 3 ("ΔΨ", Munich) performed best by all measures and in every classification of variants. In the MaPSy challenge, the same group (4 in this case), performed best on the "official" ESMs by all measures, and best on "consistent" ESMs by the "overall agreement" criterion. While other aspects of their model could account for this superior performance, one unique feature of their model (MMsplice; Cheng et al., 2019) is the decomposition of sequence surrounding alternatively spliced exons into five distinct regions (upstream intron, acceptor site, exon, donor site, and downstream intron), each of which was evaluated by a distinct neural network. Significantly, this division mirrors the known biochemical mechanisms of exon definition, in which distinct factors bind to messenger RNA precursors to initiate splicing (Black, 2003; Fu & Ares, 2014; Lee & Rio, 2015). Another important point, which was made by Rosenberg et al. (2015), is that the MPRA data on which MMsplice was trained was based on an alternative splice-site assay (local competition), but applies well to this exon inclusion assay, "suggesting a universal mechanism."

4.2 | PSI versus logit PSI

Although the Vex-seq challenge explicitly asked for predictions of ΔΨ, this measure may not directly reflect the strength of splicing signals, which are likely to range from well below what leads to complete exclusion to well above what is sufficient to promote full inclusion. For example, a change of Ψ from 99 to 98 (ΔΨ = −1) is a two-fold increase in the amount of skipping and is likely to reflect a much bigger change in the strength of splicing signals than a change in Ψ from 51 to 49 (ΔΨ = −2). Several of the groups, including the Munich group, used logit(Ψ) rather than Ψ in their analysis for the Vex-seq challenge and then converted the difference back to Ψ to calculate ΔΨ. It is likely that logit(Ψ) is a better space in which to evaluate the effect of splicing variants.

4.3 | Performance based on location of variants

The classification of variants according to their location (intron, exon, and splice site) revealed significant differences between predictors. The core splice-site consensus regions have been well-studied for many years, and our ability to evaluate the impact of mutations within these sequences is considered quite high. In particular, MaxEntScan (Yeo & Burge, 2004) has been the standard. It is, therefore, surprising that performance on this subset of the Vex-seq challenge was not better, and that groups varied as much as they do in their performance. However, when only the direction of the effect of each variant was considered (Table 2b), the correlation between predictions and observations was notably higher for splice sites (0.86) than overall (0.25).

On the subset of the Vex-seq challenge represented by intronic mutations, two groups (3 and 6) stood out. In fact, the other groups did less well than a baseline predictor that simply assigned the mean ΔΨ to every variant. This difference in performance likely reflects the fact that most of the subsidiary features used by most groups are exonic features.

4.4 | Prospects

One goal of the CAGI challenges is to address the goals of precision medicine by providing reliable estimates of the probability that a given variant is pathogenic. However, reliable computational predictions of the results of an in vivo splicing assay cannot directly imply clinical significance, because the splicing assay itself does not perfectly predict clinical results; Soemedi et al. (2017) report an 81% concordance rate with splicing in patient tissue. Furthermore, the disease impact of splicing variants is dependent on the relationship between the degree of loss of function in a gene and the disease phenotype; this varies between genes, but can be estimated from other partial loss-of-function alleles. Despite these problems, we can provide a measure of the association between prediction and the results of a high-throughput assay in the form of an odds ratio, which can be used as one step in a chain of Bayesian inference of the probability of clinical significance (Tavtigian et al., 2018). Table 4c presents the odds ratio associated with each of the MaPSy predictions.

A related concern is that these assays do not fully capture the context within which variants occur, in that sequences outside of the region tested may influence splicing. A large scale analysis of ENCODE data (Kim et al., 2017) revealed a class of splicing events that these authors referred to as "local slowpokes," whose splicing is dependent upon an enhancing effect of neighboring splicing events typically not included in splicing reporter assays. A recent study of splicing prediction from primary sequence with deep learning (Jaganathan, Panagiotopoulou, & McRae, 2019) compared the performance when windows of different size were used, finding that longer windows (up to 10,000 nucleotides) improved performance. These studies together illustrate the limitations of reporter assays with limited sequence context.

Although estimation of the clinical impact of specific variants affecting splicing may not be mature, the prediction tools used here are all very good at identifying which mutations in auxiliary splicing signals identified by clinical exome or whole-genome sequencing are potentially deleterious. Furthermore, both high-throughput splicing assays and prediction methods are undergoing rapid developments (e.g., Jaganathan et al., 2019). Thus, our ability to identify potential causative variants from sequencing data is currently very good, and these methods should be more widely used.

ACKNOWLEDGMENTS

This study was partially supported by NSF grant ABI 1564785 to SMM. The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650. We thank CAGI organizers and data providers for making this challenge happen, especially Steven Brenner, John Moult, Lipika Ray, and Brent Graveley.

ORCID

Stephen M. Mount http://orcid.org/0000-0003-2748-8205
Žiga Avsec http://orcid.org/0000-0002-7790-8936
Rita Casadio http://orcid.org/0000-0002-7462-7039
Muhammed Hasan Çelik http://orcid.org/0000-0001-7185-3711
Ken Chen http://orcid.org/0000-0001-5701-1438
Jun Cheng http://orcid.org/0000-0001-5573-9791
Julien Gagneur http://orcid.org/0000-0002-8924-8365
Valer Gotea http://orcid.org/0000-0001-7857-3309
Pier Luigi Martelli http://orcid.org/0000-0002-0274-5669
Tatsuhiko Naito http://orcid.org/0000-0002-2779-4600
Castrense Savojardo http://orcid.org/0000-0002-7359-0633
Ron Unger http://orcid.org/0000-0003-4153-3922
Robert Wang http://orcid.org/0000-0003-2614-5956

REFERENCES

Adamson, S., Zhan, L., & Graveley, B. (2018). Vex-Seq: High-Throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biology*, *19*(1), 71. https://doi.org/10.1186/s13059-018-1437-x

Black, D. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry*, *72*(1), 291–336.

Buske, O., Manickaraj, A., Mital, S., Ray, P., & Brudno, M. (2013). Identification of Deleterious Synonymous Variants In Human Genomes. *Bioinformatics*, *29*(15), 1843–1850.

Cartegni, L., Chew, S., & Krainer, A. (2002). Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nature Reviews Genetics*, *3*(4), 285–298.

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*, 48.

Cheung, R., Insigne, K., Yao, D., Burghard, C., Wang, J., Hsiao, Y.-H., … Kosuri, S. (2019). A multiplexed assay for exon recognition reveals that an unappreciated fraction of rare genetic variants cause large-effect splicing disruptions. *Molecular Cell*, *73*(1), 183–194. https://doi.org/10.1016/j.molcel.2018.10.037

Dogan, R. I., Getoor, L., Wilbur, W. J., & Mount, S. M. (2007). SplicePort—An interactive splice-site analysis tool. *Nucleic Acids Research*, *35*, 35–W291. https://doi.org/10.1093/nar/gkm407

Fu, X.-D., & Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, *15*(10), 689–701.

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., … Farh, K. K. (2019). Predicting splicing from primary sequence with deep learning. *Cell*, *176*(3), 535–548.

Kim, S. W., Taggart, A. J., Heintzelman, C., Cygan, K. J., Hull, C. G., Wang, J., … Fairbrother, W. G. (2017). Widespread intra-dependencies in the removal of introns from human transcripts. *Nucleic Acids Research*, *45*(16), 9503–9513.

Kircher, M., Witten, D., Jain, P., O'Roak, B., Cooper, G., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315.

Ladd, A., & Cooper, T. (2002). Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biology*, *3*(11), reviews0008.

Lee, Y., & Rio, D. (2015). Mechanisms and regulation of alternative pre-mRNA splicing. *Annual Review of Biochemistry*, *84*, 291–323.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … MacArthur, D. G. Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291.

Li, Y., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., … Pritchard, J. K. (2016). RNA splicing is a primary link between genetic variation and disease. *Science*, *352*(6285), 600–604.

Livingstone, M., Folkman, L., Yang, Y., Zhang, P., Mort, M., Cooper, D., … Zhou, Y. (2017). Investigating DNA-, RNA-, and protein-based features as a means to discriminate pathogenic synonymous variants. *Human Mutation*, *38*(10), 1336–1347.

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., … Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, *335*(6070), 823–828.

Mount, S. M., & Steitz, J. (1983). Lessons from Mutant Globins. *Nature*, *303*(5916), 380–381. https://doi.org/10.1038/303380a0

Naito, T. (2018). Human Splice-Site Prediction with Deep Neural Networks. *Journal of Computational Biology*, *25*(8), 954–961.

Orkin, S., Kazazian, H. H., Jr., Antonarakist, S. E., Ostrert, H., Goff, S. C., & Sexton, J. P. (1982). Abnormal RNA processing due to the exon mutation of pE-globin gene. *Nature*, *300*, 768–769.

Pal, L., Yu, C.-H., Mount, S., & Moult, J. (2015). Insights from GWAS: Emerging landscape of mechanisms underlying complex trait disease. *BMC Genomics*, *16*(Suppl 8), S4.

Raponi, M., Kralovicova, J., Copson, E., Divina, P., Eccles, D., Johnson, P., … Vorechovsky, I. (2011). Prediction of single-nucleotide substitutions that result in exon skipping: Identification of a splicing silencer in BRCA1 exon 6. *Human Mutation*, *32*(4), 436–444.

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

Rogan, P., Svojanovsky, S., & Leeder, S. (2003). Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. *Pharmacogenetics*, *13*(4), 207–218.

Rosenberg, A., Patwardhan, R., Shendure, J., & Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*, *163*(3), 698–711.

Soemedi, R., Cygan, K., Rhine, C., Wang, J., Bulacan, C., Yang, J., … Fairbrother, W. J (2017). Pathogenic variants that alter protein code often disrupt splicing. *Nature Genetics*, *49*, 848–855.

Stenson, P., Mort, M., Ball, E., Evans, K., Hayden, M., Heywood, S., … Cooper, D. (2017). The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, *136*(6), 665–677.

Tavtigian, S. V., Greenblatt, M. S., Harrison, S. M., Nussbaum, R. L., Prabhu, S. A., Boucher, K. M., … ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI) (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med*, *20*(9), 1054–1060. https://doi.org/10.1038/gim.2017.210

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164.

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K., … Frey, B. J. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, *347*(6218), 1254806.

Yang, Y., Swaminathan, S., Martin, B., & Sharan, S. (2003). Aberrant splicing induced by missense mutations in BRCA1: Clues from a humanized mouse model. *Human Molecular Genetics*, *12*(17), 2121–2131.

Yeo, G., & Burge, C. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology*, *11*(2-3), 377–394.

Zucker, M., Rosenberg, N., Peretz, H., Green, D., Bauduer, F., Zivelin, A., & Seligsohn, U. (2011). Point mutations regarded as missense mutations cause splicing defects in the factor XI gene. *Journal of Thrombosis and Haemostasis*, *9*(10), 1977–1984.

How to cite this article: Mount SM, Avsec Ž, Carmel L, et al. Assessing predictions of the impact of variants on splicing in CAGI5. *Human Mutation*. 2019;40:1215–1224. https://doi.org/10.1002/humu.23869