Published online 22 August 2022

Nucleic Acids Research, 2022, Vol. 50, No. 16 **9115–9126**
https://doi.org/10.1093/nar/gkac686

Performance evaluation of differential splicing analysis methods and splicing analytics platform construction

Kuokuo Li[1,4,5,†], Tengfei Luo[2,†], Yan Zhu[2], Yuanfeng Huang[3,6], An Wang[1,4,5], Di Zhang[1,4,5], Lijie Dong[2], Yujian Wang[3,6], Rui Wang[2], Dongdong Tang[1,4,5], Zhen Yu[1,4,5], Qunshan Shen[1,4,5], Mingrong Lv[1,4,5], Zhengbao Ling[2], Zhenghuan Fang[2], Jing Yuan[1,4,5], Bin Li[3,6], Kun Xia[2,7], Xiaojin He[1,4,5,8,*], Jinchen Li[2,3,6,*] and Guihu Zhao[3,6,*]

[1]Department of Obstetrics and Gynecology, The First Affiliated Hospital of Anhui Medical University, Hefei 230022, China, [2]Centre for Medical Genetics & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan, China, [3]Bioinformatics Center & National Clinical Research Centre for Geriatric Disorders & Department of Geriatrics, Xiangya Hospital, Central South University, Changsha, Hunan, China, [4]NHC Key Laboratory of Study on Abnormal Gametes and Reproductive Tract (Anhui Medical University), No 81 Meishan Road, Hefei 230032, Anhui, China, [5]Key Laboratory of Population Health Across Life Cycle (Anhui Medical University), Ministry of Education of the People's Republic of China, No 81 Meishan Road, Hefei 230032, Anhui, China, [6]Department of Neurology, Xiangya Hospital, Central South University, Changsha, Hunan 410008, China, [7]Hengyang Medical School, University of South China, Hengyang, Hunan, China and [8]Anhui Provincial Human Sperm Bank, The First Affiliated Hospital of Anhui Medical University, Hefei 230022, China

Received April 09, 2022; Revised July 01, 2022; Editorial Decision July 23, 2022; Accepted August 01, 2022

ABSTRACT

A proportion of previously defined benign variants or variants of uncertain significance in humans, which are challenging to identify, may induce an abnormal splicing process. An increasing number of methods have been developed to predict splicing variants, but their performance has not been completely evaluated using independent benchmarks. Here, we manually sourced ∼50 000 positive/negative splicing variants from > 8000 studies and selected the independent splicing variants to evaluate the performance of prediction methods. These methods showed different performances in recognizing splicing variants in donor and acceptor regions, reminiscent of different weight coefficient applications to predict novel splicing variants. Of these methods, 66.67% exhibited higher specificities than sensitivities, suggesting that more moderate cut-off values are necessary to distinguish splicing variants. Moreover, the high correlation and consistent prediction ratio validated the feasibility of integration of the splicing prediction method in identifying splicing variants. We developed a splicing analytics platform called SPCards, which curates splicing variants from publications and predicts splicing scores of variants in genomes. SPCards also offers variant-level and gene-level annotation information, including allele frequency, non-synonymous prediction and comprehensive functional information. SPCards is suitable for high-throughput genetic identification of splicing variants, particularly those located in non-canonical splicing regions.

INTRODUCTION

Pre-mRNA splicing, a process in which introns of nascent pre-mRNA are removed, followed by exon ligation, plays an indispensable role in maintaining protein diversity and complicated biological functions in humans (1). Over 90% of human multi-exon genes undergo alternative splicing, generating > 10 mature mRNAs per gene, which are involved in tissue- or cell-specific biological processes (2,3). The *cis*-acting regulatory elements within pre-mRNAs, including the 5′ and 3′ splice sites, branch site, polypyrimidine tract and splicing enhancer or silencer elements (4), interact with *trans*-acting factor-related proteins and regulate alternative

*To whom correspondence should be addressed. Tel: +86 731 8975 2406; Fax: +86 731 8432 7332; Email: lijinchen@csu.edu.cn
Correspondence may also be addressed to Xiaojin He. Tel: +86 731 8975 2406; Fax: +86 731 8432 7332; Email: hxj0117@126.com
Correspondence may also be addressed to Guihu Zhao. Tel: +86 731 8975 2406; Fax: +86 731 8432 7332; Email: ghzhao@csu.edu.cn
†The authors wish it to be known that in their opinion the first two authors should be regarded as Joint First Authors.

splicing processes (5–9). Disruption of these *cis*-acting regulatory elements frequently results in abnormal splicing processes and diseases.

With the development of molecular genetics, particularly whole-exome sequencing (WES) and whole-genome sequencing (WGS), > 100 000 pathogenic variants have been detected in patients in recent decades (10,11). Approximately 15–60% of rare pathogenic variants are located in *cis*-acting regulatory elements (11–14). These variants can affect alternative splicing processes by disrupting splicing sites, generating novel cryptic splice sites and disrupting splice-site usage (15). In addition, splicing quantitative trait loci (sQTLs) have been used to explore the splicing mechanisms underlying human diseases (16). Although next-generation sequencing has accelerated our understanding of splicing variants, the detailed splicing mechanisms remain unknown. Jagadeesh *et al.* found that > 500 potential splicing-related variants were classified as variants of uncertain clinical significance in typical patients (10). However, a proportion of variants of uncertain clinical significance were validated to affect the splicing process and improve prenatal diagnosis and preimplantation genetic testing (17). Therefore, there is an urgent need to further decipher splicing mechanisms and develop high-performance predictive methods to screen potential pathogenic splicing variants in humans.

Previous studies have reported on the development of numerous splicing variant prediction methods, such as MMsplice (18) and CADD-Splice (19) (Supplementary Table S1), based on functionally validated splicing variants, but their performance has not been completely evaluated using independent benchmarks. Moreover, the majority of splicing variants are reported in thousands of studies, making it challenging for general bioinformatic scientists, geneticists and biologists to obtain first-hand information regarding certain splicing variants and genes of interest. Therefore, there is a need to develop an integrated splice variant database. Several studies have attempted to integrate splice variant datasets, such as DBASS (20), MutSpliceDB (21) and SQUIRLS (22). Moreover, although pathogenic splicing prediction scores have been developed, most of these are distributed in different databases or web servers, making it time-consuming to retrieve the variants of interest. Similar to the developed dbNSFP database that provides nonsynonymous deleterious variant prediction (23), integrated splicing prediction scores are necessary to facilitate splicing investigations.

To address this need, we developed a splicing analytics platform called SPCards, curating 21 800 positive and 27 090 negative splicing variants. Furthermore, we integrated splicing variant-predicted scores into SPCards and compared the performance of different methods using benchmark datasets of the integrated splicing variants. Moreover, we integrated other genomic data sources to provide comprehensive variant-level and gene-level annotation, including (i) disease- and phenotype-related information, (ii) allele frequencies in different populations, (iii) gene-level information and (iv) drug–gene interactions. SPCards provides a convenient interface for users to search for specific variants of interest and analyze next-generation sequencing data to screen potential splicing variants using integrated spliced predicted scores.

MATERIALS AND METHODS

Data collection and manual curation

We collected splicing variants from scientific publications in PubMed using the search strategy '(mutation [Title/Abstract] OR variant [Title/Abstract]) AND (splicing [Title/Abstract] OR splice [Title/Abstract])'. We manually screened splicing variant-related publications based on the abstracts downloaded from PubMed and then curated splicing variants based on the full article or supplementary materials. Three databases, DBASS (20), MutSpliceDB (21) and SQUIRLS (22), focus on splicing variant collections in previous studies. Two databases, Gene4Denovo and ClinVar, focus on *de novo* variant and clinically related variant collections, respectively. The Gene4Denovo and ClinVar databases also contain a large number of splicing variants. To obtain a comprehensive list of splicing variants, we integrated splicing variants of five databases into SPCards to make up for the missing splicing variants in the PubMed-based collection. Only the canonical splicing variants of Gene4Denovo were integrated into SPCards. The variants in ClinVar that satisfied the thresholds 'pathogenic' or 'likely pathogenic', 'multiple submitters' and 'no conflicts', and that were annotated as 'splice donor' or 'splice acceptor', were integrated into SPCards.

The collected information for each splicing variant contained the chromosomal location, start position, end position, reference sequence, alternative sequence, distance of variant to splice junctions, whether the variant is located in the canonical splicing region or coding region, whether the variant was located in the acceptor or donor region, the detailed abnormal splicing description, phenotype, validation method and PubMed identifier. In the case of unavailable information, the term 'NA' was used. The complementary DNA (cDNA) information of splicing variants was translated into genomic DNA (gDNA) positions using VarCards (24), the UCSC Genome Browser database (25) or TransVar (26). Overlapping splicing variants (redundant splicing variants) between different publications were removed. All splice variants were checked by experienced scientists.

Splicing variant annotation and splicing prediction algorithm evaluation

We used ANNOVAR to perform a comprehensive annotation of the integrated splicing variants. To uncover their detailed impact, splicing variants were mapped into different transcripts using RefSeq databases. We classified the splicing regions ultimately into six types based on: (i) whether the variant was located in a splicing donor or acceptor region and (ii) the potential confidence in variants in specific regions impacting the splicing process. For the first two types, strong splicing regions were defined as canonical splicing regions, including the donor canonical splicing region (+1, +2) and the acceptor canonical splicing

region (−1, −2). For the second two types, the moderate splicing regions were defined as donor splicing consensus regions (−3 to +8, except +1 and +2) and acceptor splicing regions including splicing consensus regions, potential polypyrimidine tract and a potential branch point (−50 to +2, except −1 and −2). For the third two types, the mild splicing regions were defined as other splicing regions near the donor or acceptor. In addition, we annotated splicing variants using the scores of pathogenic splicing prediction methods: CADD-splice (19), SpliceAI (27), dpsi_max_tissue (28), dpsi_zscore (28), dbscSNV_ADA (29), dbscSNV_RF (29), MMSplice (18), regsnp (30), MaxEntScan (31), GeneSplicer (32), ESRseq (33), Spliceogen (34), SQUIRLS (22), KipoiSplice (35), SSF (36), SPiCE (37) and Synvep (38) (Supplementary Table S1). Synvep was used to predict pathogenic synonymous single nucleotide variants (SNVs) and to identify splice-disrupting variants (38). We integrated the Synvep prediction score into SPCards.

As in our previous study (39), we evaluated the performance of pathogenic splicing prediction methods based on the following nine criteria: (i) positive predictive value (PPV); (ii) negative predictive value (NPV); (iii) false-negative rate (FNR); (iv) sensitivity (TPR, true-positive rate); (v) false-positive rate (FPR); (vi) specificity (TNR, true-negative rate); (vii) accuracy; (viii) Mathew correlation coefficient (MCC); and (ix) area under the curve (AUC). We used 'pROC' packages to evaluate the performance of the pathogenic splicing prediction methods and generated the suggested threshold based on the curated splicing variants. If a variant lacked the prediction score of a specific method, we omitted the variant during the performance evaluation of that method.

Integrated variant-level and gene-level sources

We integrated variant-level and gene-level sources related to splicing variants as in our previous Gene4Denovo study (40). The integrated variant-level sources included the allele frequency in different populations, splicing variant predictive pathogenic scores and 47 predictive pathogenic scores. The allele frequency databases included gnomAD (41), ExAC (42), ESP6500 (43), 1000 Genomes Project (44), Kaviar (45) and HRC (46). For the 18 integrated predicted pathogenicity scores, 10 scores were generated, namely those of MMSplice, MaxEntScan, GeneSplicer, ESRseq, Spliceogen, SQUIRLS, KipoiSplice4, SPiCE, SPiCE_MES and SPiCE_SSF that were based on a local computer, and others were downloaded from corresponding websites (Supplementary Table S1). We only provided prediction score for SNVs. In addition to splicing variant-predicted scores, we also provided the links to web-based splicing prediction servers including NetGene2 (47), CRYPSKIP (48), EX-SKIP (49), ESEfinder (50) and SplicePort (51). Moreover, to provide comprehensive variant information, we integrated 47 predictive pathogenic scores (http://www.genemed.tech/spcards/analysis). We downloaded 46 of the 47 predictive pathogenic scores from dbNSFP v4 (23) and generated ReVe scores based on our previous study (39).

We integrated gene-level sources related to splicing variants from NCBI Gene (52), Gene Ontology (53) and InBio Map protein–protein interaction (54). Gene tolerance scores were retrieved from RVIS (55), LoFmethod (56), GDI (57), Episcore (58), Aggarwala (59), pLI (42) and HIPred (60). In addition, we collected information about gene-related diseases or phenotypes from OMIM, ClinVar, HPO and MGI. Gene-related expression information was collected from BrainSpan, GTEx and the Human Protein Atlas, and information on drug–gene interactions was collected from DGIdb.

Platform construction and interface

The online SPCards (http://www.genemed.tech/spcards) was developed using JavaScript, PHP (Hypertext Preprocessor) and Python on a Linux platform on a Nginx web server. A front-and-back separation model was used. The front-end was based on vue and used the UI MethodKit element, which supports all modern browsers across platforms, including Microsoft Edge, Safari, FireFox and Google Chrome. The back-end was based on Laravel, a PHP web framework. SPCards was developed and supported by versatile browsing and searching functionalities. All of the data are stored in the MySQL database. Users can freely access the genetic data through this web interface. The web interface of SPCards contains search, browse and download modules.

RESULTS

Dataset of positive and negative splicing variants

Based on the splicing variant-related keyword search strategy in PubMed, we primarily found 41 555 splicing-related studies (Supplementary Figure S1). We further curated each study based on the abstract or full articles and integrated splicing variant-related databases. There are 48 890 splicing-related variants in SPCards, comprising 21 800 positive and 27 090 negative variants (Figure 1). A total of 21 800 positive splicing variants were identified in 3345 genes from 8025 studies, comprising 21 244 SNVs and 556 insertion/deletion variants (Table 1). The deepest splicing variant, IVS1 + 36947C > T, created a pseudoexon in the intron. There were 14 335 canonical splicing variants, made up of 8168 and 6167 variants located in the splicing donor and acceptor regions, respectively. The remaining 7465 variants were located in the non-canonical splice region. We classified non-canonical splicing variants into four types: donor splicing consensus region (−3 to +8, except +1 and +2); other splicing regions near the donor; acceptor splicing region (−50 to +2, except −1 and −2) including the splicing consensus region, potential polypyrimidine tract and potential branch point; and other splicing regions near the acceptor. There were 2676, 1619, 2003 and 1167 non-canonical splice variants mapped to the donor splicing consensus region, other splicing regions near the donor, acceptor splicing region and other splicing regions near the acceptor, respectively. Of the non-canonical splice variants, 93.62% (6989/7465) were validated using assays including reverse transcription–polymerase chain reaction (RT–PCR), minigene assay, RNA-seq and multiplex functional assay of splicing using Sort-seq (MFASS). Furthermore, we curated 27 090 assay-validated negative splicing variants, most of

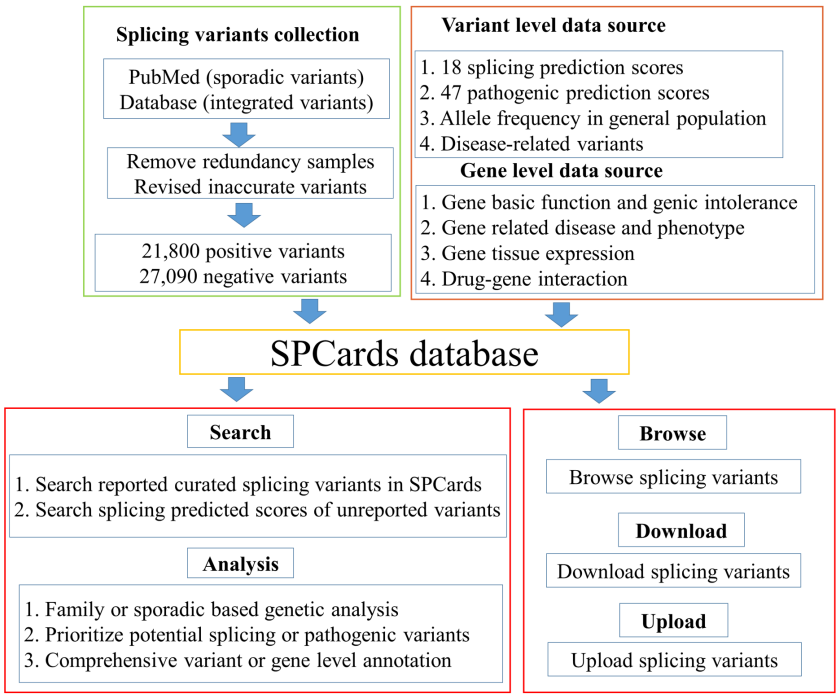

Figure 1. The workflow of SPCards.

Table 1. Summary of positive splicing variants in SPCards

			Donor			Acceptor		
Method	N-gene	N-variant	CSV	−3 to +8 except CSV	Other splicing variants	CSV	−50 to +2 except CSV	Other splicing variants
RT–PCR	1084	3669	1080	852	322	750	431	234
Minigene	364	1142	232	327	121	166	186	110
RNA-seq	176	2080	96	101	760	78	496	549
MFASS	479	1051	164	111	224	85	273	194
Experiment	1081	7736	3319	1024	144	2719	490	40
In silico analysis	2063	6122	3277	261	48	2369	127	40
Total	3345	21 800	8168	2676	1619	6167	2003	1167

RT–PCR, reverse transcription–polymerase chain reaction; MFASS, multiplex functional assay of splicing using Sort-seq; Experiment, the variants were validated by experimental evidence including minigene assay, site-directed mutagenesis or patient-derived RNA sample analysis in the SQUIRLS database; CSV, canonical splicing variants. As the validation method of variants in ClinVar was not available, we classified these variants as *in silico* analysis. −3 to +8 except CSV, the donor splicing consensus region except canonical splicing variants; −50 to +2 except CSV, the region including the donor splicing consensus region, polypyrimidine tract and branch point except canonical splicing variants.

which were validated using MFASS, in 1700 genes (Supplementary Table S2).

Benchmark selection for performance evaluation

To compare the performance of different pathogenic splicing prediction methods, we used integrated positive splicing variants as benchmark datasets. As the splicing variants in SPCards overlapped with the training dataset of splicing prediction methods, we excluded a large proportion of splicing variants from the benchmark to perform an unbiased evaluation of the splicing prediction methods (Supplementary Table S3). First, we removed the *in silico* analysis variants. Second, we only retained splicing variants published from 2019 to 2021, which did not contain the training datasets of developed splicing prediction methods, such as the latest methods CADD-splice and SQUIRLS. A total of 3403 validated positive splicing variants were used for the performance analysis.

Moreover, we source-validated negative splicing variants as benign benchmark datasets for performance evaluation. The number of negative splicing variants (27 090) was significantly larger than that of positive splicing variants (3403). To remove the potential bias of performance evaluation among the different methods, we selected the 3403 negative splicing variants as benign benchmarks (Supplementary Table S4). First, we selected a more rigorous threshold (delta percent spliced in < 0.01) for identifying splicing variants using the MFASS method compared with that of the primary study (delta percent spliced in < 0.5). Second, we used the downsampling method, which randomly selected 3403 variants from the retained 4669 negative splicing variants, comprising 4262 and 407 variants validated using MFASS and other traditional methods, respectively.

Table 2. Performance evaluation based on the SPCards splicing data

Methods	Positive variant (%)	Negative variant (%)	PPV	NPV	Specificity	FPR	Sensitivity	FNR	Accuracy	MCC	AUC
CADD-splice	2998 (88.10)	3378 (99.27)	0.63	**0.70**	0.65	0.35	0.68	0.32	0.66	0.33	0.73
dbscSNV_ADA	1492 (43.84)	446 (13.11)	**0.95**	0.66	0.83	0.17	**0.87**	**0.13**	**0.86**	**0.65**	**0.92**
dbscSNV_RF	1492 (43.84)	446 (13.11)	0.94	0.58	0.83	0.17	0.82	0.18	0.82	**0.59**	**0.89**
dpsi_max_tissue	2900 (85.22)	3360 (98.74)	0.84	0.63	**0.94**	**0.06**	0.37	0.63	0.68	0.38	0.75
dpsi_zscore	2900 (85.22)	3360 (98.74)	0.72	0.67	0.82	0.18	0.53	0.47	0.68	0.37	0.75
ESRseq	2995 (88.01)	3377 (99.24)	0.52	0.56	0.67	0.33	0.41	0.59	0.55	0.08	0.54
GeneSplicer	995 (29.24)	855 (25.12)	0.74	0.61	0.76	0.24	0.58	0.42	0.66	0.35	0.72
KipoiSplice4	2739 (80.49)	3253 (95.59)	0.89	**0.71**	**0.94**	**0.06**	0.53	0.47	0.76	0.53	0.72
MaxEntScan	2965 (87.13)	3377 (99.24)	0.57	0.68	0.52	0.48	0.72	0.28	0.62	0.25	0.63
MMsplice	2897 (85.13)	3377 (99.24)	**0.97**	0.63	**0.99**	**0.01**	0.32	0.68	0.68	0.43	0.71
regsnp	1773 (52.10)	1557 (45.75)	0.89	**0.71**	0.90	0.10	0.67	0.33	0.78	0.58	0.84
SPiCE	1622 (47.66)	459 (13.49)	0.92	0.68	0.70	0.30	**0.91**	**0.09**	**0.86**	**0.60**	**0.90**
SPiCE_MES	1622 (47.66)	460 (13.51)	0.94	0.60	0.80	0.20	**0.85**	**0.15**	**0.84**	**0.59**	**0.89**
SPiCE_SSF	1622 (47.66)	460 (13.51)	0.94	0.53	0.82	0.18	0.79	0.21	0.80	0.53	0.88
SpliceAI	1680 (49.37)	159 (4.67)	**0.98**	0.23	0.84	0.16	0.73	0.27	0.74	0.34	0.83
Spliceogen	2995 (88.01)	3377 (99.24)	0.83	0.66	0.91	**0.09**	0.48	0.52	0.71	0.44	0.72
Squirl	2995 (88.01)	3377 (99.24)	0.74	**0.70**	0.81	0.19	0.61	0.39	0.71	0.43	0.78
Synvep	373 (10.96)	585 (17.19)	0.45	0.67	0.56	0.44	0.56	0.44	0.56	0.12	0.59

SpliceAI, SpliceAI score > 0.1 was integrated into SPCards. The number of true-positive variants and false-positive variants in benchmark data was 3403, respectively. PPV, positive predictive value; NPV, negative predictive value; FNR, false-negative rate; Sensitivity, true-positive rate; FPR, false-positive rate; Specificity, true-negative rate; MCC, Mathew correlation coefficient; AUC, area under the curve. Values in bold are the top performances of predicted methods.

Performance of splicing methods based on AUC

The prediction methods, including CADDsplice, ESRseq, Spliceogen, Squirl, MaxEntScan, dpsi_max_tissue, dpsi_zscore and MMsplice, were available for positive splicing variant-predicted values > 85% (Table 2; Supplementary Table S3). SpliceAI is a genomic prediction method. We integrated highly confident variants with SpliceAI > 0.1 into SPCards and only detected scores in 49.37% of positive and 4.67% of negative splicing variants. Using SpliceAI > 0.1, a significantly higher number of positive variants than negative variants was detected [Fisher's exact test, P = 3.37E-262, odds ratio (OR) = 10.56, confidence interval (CI) = 8.91–12.59], indicating the ability of SpliceAI to detect splicing variants. In addition, the available splicing prediction values for KipoiSplice4 were > 80%.

We evaluated the performances of these methods by measuring the AUC scores and found that dbscSNV_ADA (AUC = 0.9154) and SPiCE (AUC = 0.9046) exhibited AUC scores > 0.9 and showed the best performance compared with other methods, followed by SPiCE_MES (AUC = 0.8887), dbscSNV_RF (AUC = 0.887), SPiCE_SSF (AUC = 0.8813) and regsnp (AUC = 0.8368) (Figure 2A; Supplementary Table S5). These six methods focused on splicing region hotspots, such as the splicing consensus region (Supplementary Table S3). The best performing global genome prediction methods were SpliceAI (AUC = 0.8332) and Squirls (AUC = 0.7819) (Figure 2A). The four prediction methods GeneSplicer (AUC = 0.7178), MaxEntScan (AUC = 0.6276), Synvep (AUC = 0.5905) and ESRseq (AUC = 0.5436) exhibited the lowest power to distinguish splicing variants (Figure 2A).

For the donor splicing consensus region (donor −3 to +8), Kipoisplice (AUC = 0.934), regsnp (AUC = 0.9293), dbscSNV_ADA (AUC = 0.9107), Spliceogen (AUC = 0.9096) and SPiCE (AUC = 0.9059) exhibited AUC scores > 0.9 (Figure 2B; Supplementary Table S5). Spliceogen was the best global genome prediction method, followed by MMsplice (AUC = 0.8854). For the acceptor splicing region (acceptor −50 to +2), only dbscSNV_ADA (AUC = 0.9071) exhibited AUC scores > 0.9, followed by SPiCE (AUC = 0.8724), dbscSNV_RF (AUC = 0.8721) and regsnp (AUC = 0.8584) (Figure 2C; Supplementary Table S5). CADD-splice (AUC = 0.8571) and SpliceAI (AUC = 0.8278) exhibited a high performance in global genome prediction.

The splicing prediction methods exhibited a better performance in the canonical splicing region (+1 and +2 of the splicing donor site and −1 and −2 of the splicing acceptor site) than other regions. As a larger proportion of benchmark variants were located in the canonical splicing region, the AUC of most prediction methods was > 0.7. To evaluate the performance of these methods in non-canonical splicing regions, we removed the canonical splicing variants. The predicted methods also exhibited better performance in the donor region (−3 to +8), except for canonical splicing (Supplementary Figure S2A), and the acceptor region (−50 to +2), except for canonical splicing (Supplementary Figure S2C; Supplementary Table S5). However, these methods showed a significantly reduced performance in other regions near both the donor and acceptor sites (Supplementary Figure S2B, Supplementary Figure S2D; Supplementary Table S5). Despite the weak power of the predicted methods for detecting splicing variants with a long distance to the splicing site, four methods, i.e. SpliceAI (0.7539, 0.6853), dpsi_max_tissue (0.6014, 0.631), dpsi_zscore (0.5885, 0.6295) and Squirls (0.5814, 0.6955), exhibited better AUC performance in the donor and acceptor regions, respectively (Supplementary Figure S2B, Supplementary Figure S2D; Supplementary Table S5).

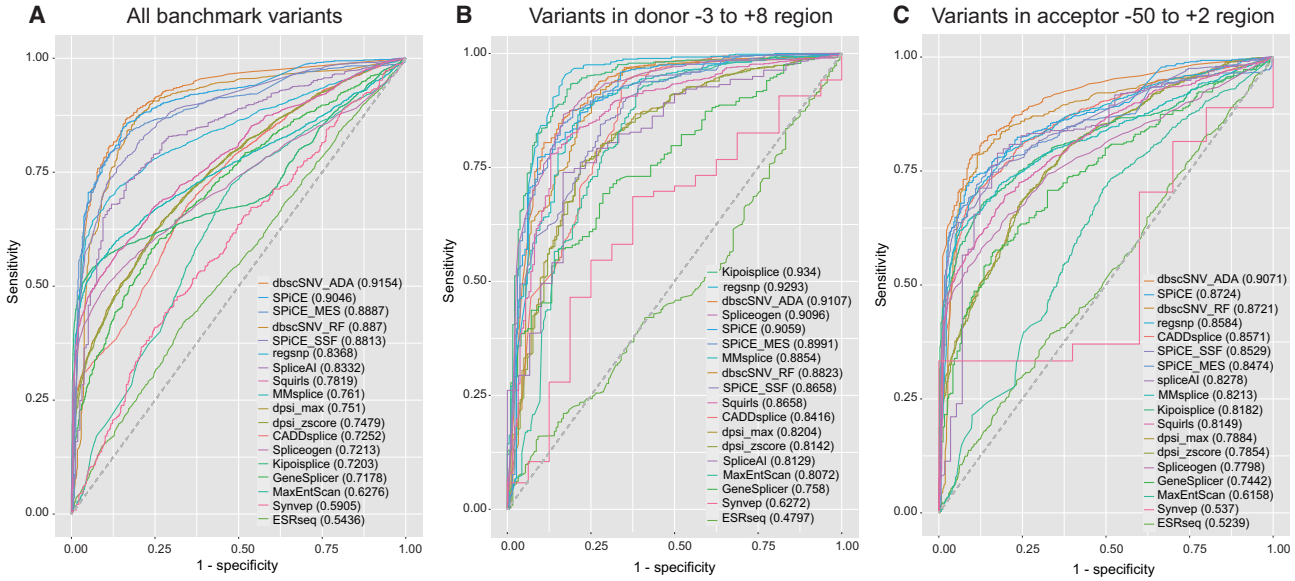

Figure 2. Performance of splicing prediction methods within three regions based on benchmark data. We used integrated functionally validated splicing variants reported from 2019 to 2021 in SPCards. (**A**) All splicing variants. (**B**) Donor −3 to +8, donor splicing consensus region, the variant three bases upstream and eight bases downstream of the donor site. (**C**) Acceptor −50 to +2, the variants 50 bases upstream and two bases downstream of the acceptor site including the acceptor splicing consensus region, potential polypyrimidine tract and potential branch point.

Performance of splicing methods based on the prediction threshold

The prediction threshold was used to further evaluate the performance of the splicing prediction methods. Eleven of the methods had suggestive prediction thresholds and the primary thresholds were used to perform the analysis (Supplementary Table S6). For the other seven methods, we used pROC to generate the best threshold with high sensitivity and specificity, based on the benchmark dataset (Supplementary Table S6). The evaluated performance of the methods is summarized in Table 2. The PPVs ranged from 0.45 to 0.98, and were > 0.9 in the seven methods (SpliceAI, MMsplice, dbscSNV_ADA, SPiCE_MES, dbscSNV_RF, SPiCE_SSF and SPiCE). The NPVs of the methods were generally lower than those of the PPVs, ranging from 0.23 to 0.71, except for CADD-splice, MaxEntScan, ESRseq and Synvep. Only four of the methods (regsnp, KipoiSplice4, Squirl and CADD-splice) had an NPV > 0.7. Moreover, the specificities (1–FPR) and sensitivities (1–FNR) were in the range of 0.52–0.99 and 0.32–0.91, respectively. The specificities of 66.67% of the methods were higher than their sensitivities, particularly for MMsplice and dpsi_max_tissue. This suggests that the benign variants were actually pathogenic and that more moderate cutoff values need to be used to distinguish splicing variants. The accuracies of the five methods, dbscSNV_ADA (0.86), SPiCE (0.86), SPiCE_MES (0.84), dbscSNV_RF (0.82) and SPiCE_SSF (0.8), were > 0.8, and exhibited the highest performances compared with the other methods. Furthermore, dbscSNV_ADA (0.65), SPiCE (0.60), SPiCE_MES (0.59) and dbscSNV_RF (0.59) exhibited the highest MCC scores.

Although we selected a matched number of positive and negative splicing variants to remove bias during the performance evaluation, the available scores for each method were also imbalanced between the positive and negative splicing variants (Table 2). The numbers of positive and negative variants were significantly different for six methods. To test whether this imbalance influenced the performance of the splicing prediction methods, we used the downsampling method to select matched variant numbers among the positive or negative variant datasets. We found that this imbalance only influenced the PPV and NPV of six significantly available score biased methods, i.e. dbscSNV_ADA, dbscSNV_RF, SPiCE, SPiCE_MES, SPiCE_SSF and SpliceAI (Supplementary Table S7). The performance of other computational methods, including specificity, FPR, sensitivity, FNR, accuracy and MCC, was almost consistent with that of the primary evaluation (Table 2; Supplementary Table S7). This indicated that the robustness of the evaluation was not impacted by biased numbers of positive and negative variants.

Owing to the large number of genetic studies focused on cancer-related disorders, we tested whether the performance of the splicing prediction methods was different between cancer- and non-cancer-related genes. We sourced 1172 cancer-related genes in the COSMIC (61,62) and OncoKB (63) databases (Supplementary Table S8) and found 12.61% (858/6806) of cancer-related variants in benchmark. An evaluation of the splicing prediction methods for cancer- and non-cancer-related genes revealed that the performance was slightly higher for cancer-related genes than for non-cancer-related genes. However, we did not find any significant difference between the two types of genes (Supplementary Table S9). This indicated that splicing prediction methods are unbiased for different kind of genes.

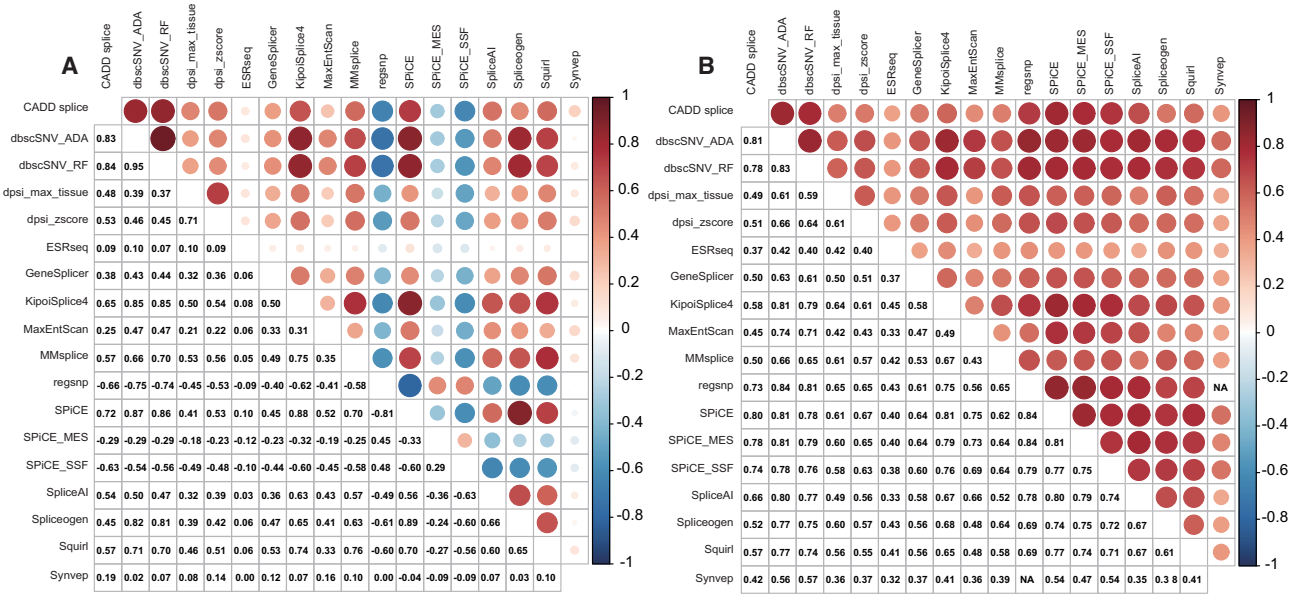

Figure 3. Correlation and consistent prediction ratio among splicing prediction methods. We retained only variants that had prediction scores in both methods for the correlation and consistent prediction ratio analysis. (**A**) Pearson's correlation coefficients (*R*). (**B**) Consistent prediction ratio of binary predictions between pairs of splicing methods. The threshold of the splicing prediction methods is shown in Supplementary Table S6.

Correlation and consistent prediction ratio of computational methods

For the prediction methods, we calculated the Pearson's correlation coefficient between any two methods based on benchmark variants. The scores of regsnp, SPiCE_MES and SPiCE_SSF were negatively correlated with those of the other methods (Figure 3A). To perform a correlation evaluation, we selected the absolute Pearson's correlation coefficient for further analysis. Of the average correlations 55.56% were strong (> 0.5), and SPiCE (0.61), dbscSNV_ADA (0.59), dbscSNV_RF (0.59) and KipoiSplice4 (0.59) exhibited the highest Pearson's correlation coefficients with other methods, followed by the global genome prediction method Squirl (0.55) (Figure 3A; Supplementary Table S10). Furthermore, 33.33% of the correlations were medium (0.3–0.5). The synvep (0.132) and ESRseq (0.126) scores were weakly correlated with those of the other methods (Supplementary Table S10).

We then compared the consistent prediction ratios of the binary predictions with the prediction scores based on benchmark variants (Figure 3B). The average agreement ratios of 66.67% of methods were > 0.6; this was particularly true for dbscSNV_ADA (0.73), SPiCE (0.72), SPiCE_MES (0.71), dbscSNV_RF (0.71) and regsnp (0.71). However, both synvep (0.43) and ESRseq (0.40) exhibited relatively lower consistent prediction ratios compared with those of the other methods (Supplementary Table S11). To provide a more detailed performance evaluation, we examined the prediction methods for positive and negative variants (Supplementary Figure S3). For the positive splicing variants, the prediction ratios of SPiCE (0.76), dbscSNV_ADA (0.75), SPiCE_MES (0.73) and dbscSNV_RF (0.72) were > 0.7 and exhibited higher power in identifying pathogenic variants (Supplementary Table S12). Moreover, the predicted power of the negative splicing variants exhibited slight differences compared with that of the positive splicing variants. MMsplice (0.79), regsnp (0.77), spliceogen (0.74), KipoiSplice4 (0.73) and dpsi_max_tissue (0.73) showed higher consistent prediction ratios. SPiCE (0.56), dbscSNV_ADA (0.65), SPiCE_MES (0.62) and dbscSNV_RF (0.65) exhibited lower consistent prediction ratios for negative splicing variants (Supplementary Table S13).

We further analyzed the consistent percentage of predictions for the prediction methods in the different splicing regions. The methods exhibited a higher consistent percentage of predictions (> 0.5) in the donor splicing consensus region [93.20% (726/779) for −3 to +8] and acceptor splicing region [48.10% (468/973) for −50 to +2] for positive splicing variants (Supplementary Table S14). When canonical variants were removed, the predicted methods exhibited a higher consistent percentage of predictions (> 0.5) in the donor regions [84.28% (134/159) for −3 to +8 except canonical splicing region] but not acceptor regions [30.58% (222/726) for −50 to +2 except canonical splicing region] for positive splicing variants (Supplementary Table S14). We found higher consistent percentage of predictions for negative splicing variants in donor regions [84.28%, (134/159) for −3 to +8 except canonical splicing region] and acceptor regions [96.69%, (700/724) for −3 to +8 except canonical splicing region] (Supplementary Table S15). For the other splicing regions, the splicing prediction methods exhibited significant differences in the predictions of positive and negative splicing variants, and most variants in these region were defined as benign, indicating the incomplete performance of the predicted methods for deep intronic sequence variation (Supplementary Tables S14 and S15).

Table 3. Integrated data sources in SPCards

Category	Data source
Variation-level	
Allele frequency	gnomAD, ExAC, ESP6500, 1000 Genomes Project, Kaviar, HRC
Splicing prediction	CADDsplice, SpliceAI, dpsi_max_tissue, dpsi_zscore, dbscSNV_ADA, dbscSNV_RF, MaxEntScan, GeneSplicer, ESRseq, Spliceogen, Squirl, regsnp, MMsplice, KipoiSplice4, Synvep, SPiCE_SSF, SPiCE_MES, SPiCE
Non-synonymous prediction	ReVe, SIFT, SIFT4G, Polyphen2_HDIV, Polyphen2_HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, VEST4, MetaSVM, MetaLR, MetaRNN, M-CAP, REVEL, MutPred, MVP, MPC, PrimateAI, DEOGEN2, BayesDel_addAF, BayesDel_noAF, ClinPred, LIST-S2, Aloft, CADD_coding, DANN, fathmm-MKL_coding_pred, fathmm-XF_coding, Eigen-raw_coding, Eigen-PC-raw_coding, GenoCanyon_score, integrated_fitCons, GM12878_fitCons, H1-hESC_fitCons, HUVEC_fitCons, LINSIGHT, GERP++_RS, phyloP100way_vertebrate, phyloP30way_mammalian, phyloP17way_primate, phastCons100way_vertebrate, phastCons30way_mammalian, phastCons17way_primate, SiPhy_29way_logOdds, bStatistic_converted
Disease-related	Gene4Denovo, ClinVar, InterVar, ICGC, COSMIC, NCI
Gene-level	
Basic information	UniProtKB, UniProt, Gene Ontology, InterPro, InBio Map, BioSystems
Genic intolerance	RVIS, LoFtool, GDI, Episcore, heptanucleotide context intolerance score, pLI
Disease-related	OMIM, MGI, HPO
Gene expression	BrainSpan, GTEx, The Human Protein Atlas
Target drug	DGIdb

gnomAD, genome aggregation database; EXAC, The Exome Aggregation Consortium; ESP6500, NHLBI GO Exome Sequencing Project; Kaviar, Kaviar Genomic Variant Database; HRC, haplotype reference consortium; RVIS, Residual Variation Intolerance Score; GDI, Human Gene Damage Index; pLI, the probability of being loss-of-function intolerant; OMIM, online Mendelian inheritance in man; MGI, mouse genome informatics; ICGC, International Cancer Genome Consortium; COSMIC, catalogue of somatic mutations in cancer; NCI, NCI-60 Human Tumor Cell Lines Screen; HPO, human phenotype ontology; GTEx, Genotype-Tissue Expression; DGIdb, The Drug Gene Interaction Database.

Integrated variant-level and gene-level sources in SPCards

To accelerate the interpretation of potential splicing variants, we integrated the predicted splicing variant scores into SPCards. Moreover, we integrated other genomic data sources to provide comprehensive variant-level and gene-level annotation information, including (i) allele frequency in gnomAD (41), ExAC (42), ESP6500 (43), 1000 Genomes Project (44), Kaviar (45) and HRC (46); (ii) 47 non-synonymous prediction methods; (iii) disease-related variants in OMIM, MGI, HPO, Gene4Denovo, ClinVar, InterVar, ICGC and NCI; (iv) gene basic information in UniProtKB, UniProt, Gene Ontology, InterPro, InBio Map and BioSystems; (v) genic intolerance in RVIS, LoFmethod, GDI, Episcore, heptanucleotide context intolerance score and pLI; (vi) disease-related genes in OMIM, MGI and HPO; (vii) gene expression in BrainSpan, GTEx and The Human Protein Atlas; and (viii) drug–gene interactions in DGIdb (Table 3). Furthermore, SPCards provides a convenient interface for users to freely analyze their next-generation sequencing data to screen for potential splicing variants using integrated splicing prediction methods (http://www.genemed.tech/spcards).

Search and analysis sections in SPCards

SPCards provides a batch search function that recognizes gene symbols, genomic regions, transcripts, variants, coordinates and distances to splicing junctions (http://www.genemed.tech/spcards/search). SPCards also provides users with a function to perform comprehensive analyses of trio- and non-trio-based genetic data (http://genemed.tech/spcards/analysis). Trio-based analysis can be used to identify *de novo* mutations, homozygous variants, compound heterozygous variants and X-linked variants. Non-trio-based analysis can be used to identify co-segregated rare deleterious variants. Users can define the quality of the variant, the percentage of positive splicing prediction using 18 methods, the result of 47 pathogenic prediction methods, clinically related information and allele frequency in the general population. The results are sent to users via email. Moreover, users can browse the splicing variants in SPCards (http://www.genemed.tech/spcards/browse).

DISCUSSION

With the rapid development of next-generation sequencing technology resulting in an explosion of genetic data, an increasing number of human disease-associated splicing variants have been identified. Approximately 15–60% of rare pathogenic variants are located in *cis*-acting regulatory elements (11–14), which might affect the alternative splicing process (15). However, it is challenging to find potential splicing variants owing to obscure mechanisms and limited accurate prediction methods. Here, we curated almost 50 000 positive and negative variants using 18 prediction methods to provide comprehensive knowledge of splicing variants in SPCards. We also compared the performance of different scores using benchmark datasets of integrated splicing variants.

There were 21 800 positive splicing variants, of which 71.92% (15 678/21 800) were validated using various methods. RT–PCR and RNA-seq were the most commonly used validation methods, followed by minigene assay (Table 1). Previous studies have reported inconsistent results using minigene assay and RT–PCR, which might suggest tissue-

specific RNA processing (64). In addition, the use of multiple validation methods is necessary for mutual authentication to eliminate the potential limitations of a specific method.

In this study, we analyzed the performance of prediction methods using nine criteria based on integrated validated negative and positive splicing variants. To remove potential bias during performance evaluation, we used three strategies. First, we removed the *in silico* analysis variants. Second, we excluded variants published before 2019 which contained the training datasets of the 18 developed splicing prediction methods. Third, as in a previous performance analysis (39,65), we selected matched numbers of positive and negative splicing variants. Although the matched available positive and negative splicing variants of specific methods were also used to explore performance, the only slight change indicated the insusceptibility of these datasets.

According to the benchmark splicing variants, most methods exhibited a high AUC, indicating better and more robust performance in an independently validated dataset. We further classified the splice variants into six types. We observed convergence and divergence of the predicted method performance between the donor and acceptor regions. First, the predicted methods performed better for variants in the donor than in the acceptor region, which might be due to the relatively close distance of the donor region (−3 to +8) compared with that of the acceptor splicing region (−50 to +2). Second, for the donor consistent region (−3 to +8) and acceptor splicing region (−50 to +2), dbscSNV_ADA and SPiCE exhibited the best performance, but CADD-splice also showed a better performance in the acceptor splicing region. Third, when the canonical splicing variants were removed, although dbscSNV_ADA and SPiCE also exhibited a better performance, the other predicted methods, including regsnp, KipoiSplice4, MMsplice and Spliceogen, showed a significantly increased performance. Fourth, for the other regions, the predicted methods that cover a relatively large region of the genome or the whole genome, including SpliceAI, dpsi_max_tissue, dpsi_zscore and Squirl, showed the best performance. These results indicate that the weight coefficient can be applied to predict variants in different regions.

We found that 66.67% of the methods exhibited higher specificity than sensitivity, particularly MMsplice and dpsi_max_tissue. This suggests that the benign variants were actually pathogenic and that clinicians and geneticists should reanalyze the negative genetic data of diseases with high heritability. Furthermore, we provided a suggestive threshold for each predicted method based on the splicing variant benchmark dataset (Supplementary Table S6). The thresholds of dpsi_max_tissue and MMsplice were significantly affected by the benchmark, while another independent benchmark was necessary to further validate these suggestive thresholds.

For the correlation and agreement ratio analyses, the splicing prediction methods exhibited a close relationship with each other. Of the average correlations, 55.56% were strong (> 0.5), including those among the SPiCE (0.61), dbscSNV_ADA (0.59), dbscSNV_RF (0.59) and KipoiSplice4 (0.59) scores. However, the scores of Synvep and ESRseq were poorly correlated with those of other splicing prediction methods, which may indicate different features in the training model. A previous study also showed a poor correlation of the Synvep score with other synonymous SNV scores, such as those of CADD (38). The average agreement ratios of 66.67% of the methods were > 0.6, particularly those of dbscSNV_ADA (0.73), SPiCE (0.72), SPiCE_MES (0.71), dbscSNV_RF (0.71) and regsnp (0.71). All high-performance prediction methods only focused on the splicing variant hotspot region, which contains a large number of training datasets. However, decoding variants in other regions, although challenging, is urgently needed. For example, deep intronic sequence variation results in pseudogenes (66).

The comprehensive database was significantly helpful in understanding pathogenic variants. dbNSFP (23) provides > 40 types of non-synonymous and splice site SNV predictors, which are popular in the field of bioinformatics and genetics and used to identify pathogenic variants (67). We developed a comprehensive splicing platform that offers more prediction methods than Alamut, dbNSFP (23) and VannoPortal (68). A recent study reported the development of ASpedia, which focuses on human alternative splicing (69). ASpedia identified alternative splicing events rather than splicing genetic variants of specific genes at the mRNA level by comparing the different known isoforms in Ensembl and RefSeq. The difference is that SPCards catalogued all the identified splicing genetic variants associated with human diseases at the DNA level from thousands of published studies. In addition, ASpedia requires alternative splicing events as input to retrieve overlapping functional information around alternative splicing regions, such as potential microRNA-binding sites, repeat sequences and protein domains. SPCards can evaluate whether a genomic variant causes an abnormal splicing process by using different integrated splicing prediction scores. Therefore, the two platforms do not compete with each other, instead they complement each other.

This study has some limitations. First, owing to the lack of clearly suggested thresholds of some prediction methods, we generated the cut-off value with the best balance of sensitivity and specificity using the pROC package based on benchmark splicing variants, which might result in bias of performance evaluation. Second, we integrated splicing variants based on keywords in the title and abstract, which might have missed many variants only present in full articles of a large cohort of genetic studies that did not focus on splicing variants. Third, despite curating the most comprehensive validated splicing variant dataset, we cannot remove all the bias among splicing prediction methods. Independent splicing variants are necessary to further validate the performance of these methods.

To date, SPCards is one of the largest and most comprehensive platforms enabling researchers without prior knowledge of bioinformatics to search for and analyze splicing variants of genes of interest using multiple relevant parameters for future functional research. This platform is suitable for high-throughput genetic identification of splicing variants, particularly those located in non-canonical splicing regions.

DATA AVAILABILITY

Splicing variants and splicing predicted scores are available at http://www.genemed.tech/spcards.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the members of the Department of Obstetrics and Gynecology, the First Affiliated Hospital of Anhui Medical University, NHC Key Laboratory of Study on Abnormal Gametes and Reproductive Tract and Center for Medical Genetics, Central South University for valuable discussions regarding this work. We are grateful for resources from the High Performance Computing Center of Central South University.

FUNDING

This work was supported by the National Key R&D Program of China [2021YFC2502100, 2021YFC2700901]; National Natural Science Foundation of China [320705918 to J.C.L., 82101944 to K.K.L.]; Natural Science Foundation of Hunan Province for outstanding Young Scholar [2020JJ3059]; Hunan Youth Science and Technology Innovation Talent Project [2020RC3060 to J.C.L.]; Natural Science Foundation for Young Scientists of Hunan Province, China [2019JJ50974 to G.H.Z.]; and Natural Science Project of University in Anhui Province [KJ2020A0204 to K.K.L.]. Funding for open access charge: National Natural Science Foundation of China.

Conflict of interest statement. None declared.

REFERENCES

1. Scotti,M.M. and Swanson,M.S. (2016) RNA mis-splicing in disease. *Nat. Rev. Genet.*, **17**, 19–32.
2. Park,E., Pan,Z., Zhang,Z., Lin,L. and Xing,Y. (2018) The expanding landscape of alternative splicing variation in human populations. *Am. J. Hum. Genet.*, **102**, 11–26.
3. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
4. Gao,Y., Lin,K.T., Jiang,T., Yang,Y., Rahman,M.A., Gong,S., Bai,J., Wang,L., Sun,J., Sheng,L. *et al.* (2022) Systematic characterization of short intronic splicing-regulatory elements in SMN2 pre-mRNA. *Nucleic Acids Res.*, **50**, 731–749.
5. Sanders,S.J., Schwartz,G.B. and Farh,K.K. (2020) Clinical impact of splicing in neurodevelopmental disorders. *Genome Med.*, **12**, 36.
6. Xu,M., Bai,X., Ai,B., Zhang,G., Song,C., Zhao,J., Wang,Y., Wei,L., Qian,F., Li,Y. *et al.* (2022) TF-Marker: a comprehensive manually curated database for transcription factors and related markers in specific cell and tissue types in human. *Nucleic Acids Res.*, **50**, D402–D412.
7. Jiang,Y., Qian,F., Bai,X., Liu,Y., Wang,Q., Ai,B., Han,X., Shi,S., Zhang,J., Li,X. *et al.* (2019) SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res.*, **47**, D235–D243.
8. Zhang,Y., Song,C., Zhang,Y., Wang,Y., Feng,C., Chen,J., Wei,L., Pan,Q., Shang,D., Zhu,Y. *et al.* (2022) TcoFBase: a comprehensive database for decoding the regulatory transcription co-factors in human and mouse. *Nucleic Acids Res.*, **50**, D391–D401.
9. Chen,J., Zhang,J., Gao,Y., Li,Y., Feng,C., Song,C., Ning,Z., Zhou,X., Zhao,J., Feng,M. *et al.* (2021) LncSEA: a platform for long non-coding RNA related sets and enrichment analysis. *Nucleic Acids Res.*, **49**, D969–D980.
10. Jagadeesh,K.A., Paggi,J.M., Ye,J.S., Stenson,P.D., Cooper,D.N., Bernstein,J.A. and Bejerano,G. (2019) S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nat. Genet.*, **51**, 755–763.
11. Stenson,P.D., Mort,M., Ball,E.V., Shaw,K., Phillips,A. and Cooper,D.N. (2014) The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **133**, 1–9.
12. Wang,G.S. and Cooper,T.A. (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, **8**, 749–761.
13. Krawczak,M., Reiss,J. and Cooper,D.N. (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.*, **90**, 41–54.
14. Lim,K.H., Ferraris,L., Filloux,M.E., Raphael,B.J. and Fairbrother,W.G. (2011) Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl Acad. Sci. U.S.A.*, **108**, 11093–11098.
15. Soemedi,R., Cygan,K.J., Rhine,C.L., Wang,J., Bulacan,C., Yang,J., Bayrak-Toydemir,P., McDonald,J. and Fairbrother,W.G. (2017) Pathogenic variants that alter protein code often disrupt splicing. *Nat. Genet.*, **49**, 848–855.
16. Walker,R.L., Ramaswami,G., Hartl,C., Mancuso,N., Gandal,M.J., de la Torre-Ubieta,L., Pasaniuc,B., Stein,J.L. and Geschwind,D.H. (2019) Genetic control of expression and splicing in developing human brain informs disease mechanisms. *Cell*, **179**, 750–771.
17. He,W.B., Xiao,W.J., Dai,C.L., Wang,Y.R., Li,X.R., Gong,F., Meng,L.L., Tan,C., Zeng,S.C., Lu,G.X. *et al.* (2022) RNA splicing analysis contributes to reclassifying variants of uncertain significance and improves the diagnosis of monogenic disorders. *J. Med. Genet.*, https://doi.org/10.1136/jmedgenet-2021-108013.
18. Cheng,J., Nguyen,T.Y.D., Cygan,K.J., Celik,M.H., Fairbrother,W.G., Avsec,Z. and Gagneur,J. (2019) MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.*, **20**, 48.
19. Rentzsch,P., Schubach,M., Shendure,J. and Kircher,M. (2021) CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.*, **13**, 31.
20. Buratti,E., Chivers,M., Hwang,G. and Vorechovsky,I. (2011) DBASS3 and DBASS5: databases of aberrant 3′- and 5′-splice sites. *Nucleic Acids Res.*, **39**, D86–D91.
21. Palmisano,A., Vural,S., Zhao,Y. and Sonkin,D. (2021) MutSpliceDB: a database of splice sites variants with RNA-seq based evidence on effects on splicing. *Hum. Mutat.*, **42**, 342–345.
22. Danis,D., Jacobsen,J.O.B., Carmody,L.C., Gargano,M.A., McMurry,J.A., Hegde,A., Haendel,M.A., Valentini,G., Smedley,D. and Robinson,P.N. (2021) Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *Am. J. Hum. Genet.*, **108**, 1564–1577.
23. Liu,X., Li,C., Mou,C., Dong,Y. and Tu,Y. (2020) dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.*, **12**, 103.
24. Li,J., Shi,L., Zhang,K., Zhang,Y., Hu,S., Zhao,T., Teng,H., Li,X., Jiang,Y., Ji,L. *et al.* (2018) VarCards: an integrated genetic and clinical database for coding variants in the human genome. *Nucleic Acids Res.*, **46**, D1039–D1048.
25. Haeussler,M., Zweig,A.S., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Hinrichs,A.S., Gonzalez,J.N. *et al.* (2019) The UCSC genome browser database: 2019 update. *Nucleic Acids Res.*, **47**, D853–D858.
26. Zhou,W., Chen,T., Chong,Z., Rohrdanz,M.A., Melott,J.M., Wakefield,C., Zeng,J., Weinstein,J.N., Meric-Bernstam,F., Mills,G.B. *et al.* (2015) TransVar: a multilevel variant annotator for precision genomics. *Nat. Methods*, **12**, 1002–1003.
27. Jaganathan,K., Kyriazopoulou Panagiotopoulou,S., McRae,J.F., Darbandi,S.F., Knowles,D., Li,Y.I., Kosmicki,J.A., Arbelaez,J., Cui,W., Schwartz,G.B. *et al.* (2019) Predicting splicing from primary sequence with deep learning. *Cell*, **176**, 535–548.

28. Jian,X., Boerwinkle,E. and Liu,X. (2014) In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.*, **42**, 13534–13544.
29. Xiong,H.Y., Alipanahi,B., Lee,L.J., Bretschneider,H., Merico,D., Yuen,R.K., Hua,Y., Gueroussov,S., Najafabadi,H.S., Hughes,T.R. *et al.* (2015) RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, **347**, 1254806.
30. Lin,H., Hargreaves,K.A., Li,R., Reiter,J.L., Wang,Y., Mort,M., Cooper,D.N., Zhou,Y., Zhang,C., Eadon,M.T. *et al.* (2019) RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants. *Genome Biol.*, **20**, 254.
31. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.
32. Pertea,M., Lin,X. and Salzberg,S.L. (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.*, **29**, 1185–1190.
33. Ke,S., Shang,S., Kalachikov,S.M., Morozova,I., Yu,L., Russo,J.J., Ju,J. and Chasin,L.A. (2011) Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.*, **21**, 1360–1374.
34. Monger,S., Troup,M., Ip,E., Dunwoodie,S.L. and Giannoulatou,E. (2019) Spliceogen: an integrative, scalable tool for the discovery of splice-altering variants. *Bioinformatics*, **35**, 4405–4407.
35. Avsec,Z., Kreuzhuber,R., Israeli,J., Xu,N., Cheng,J., Shrikumar,A., Banerjee,A., Kim,D.S., Beier,T., Urban,L. *et al.* (2019) The kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.*, **37**, 592–600.
36. Shapiro,M.B. and Senapathy,P. (1987) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.*, **15**, 7155–7174.
37. Leman,R., Gaildrat,P., Le Gac,G., Ka,C., Fichou,Y., Audrezet,M.P., Caux-Moncoutier,V., Caputo,S.M., Boutry-Kryza,N., Leone,M. *et al.* (2018) Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort. *Nucleic Acids Res.*, **46**, 7913–7923.
38. Zeng,Z., Aptekmann,A.A. and Bromberg,Y. (2021) Decoding the effects of synonymous variants. *Nucleic Acids Res.*, **49**, 12673–12691.
39. Li,J., Zhao,T., Zhang,Y., Zhang,K., Shi,L., Chen,Y., Wang,X. and Sun,Z. (2018) Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res.*, **46**, 7793–7804.
40. Zhao,G., Li,K., Li,B., Wang,Z., Fang,Z., Wang,X., Zhang,Y., Luo,T., Zhou,Q., Wang,L. *et al.* (2020) Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. *Nucleic Acids Res.*, **48**, D913–D926.
41. Karczewski,K.J., Francioli,L.C., Tiao,G., Cummings,B.B., Alfoldi,J., Wang,Q., Collins,R.L., Laricchia,K.M., Ganna,A., Birnbaum,D.P. *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.
42. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
43. Fu,W., O'Connor,T.D., Jun,G., Kang,H.M., Abecasis,G., Leal,S.M., Gabriel,S., Rieder,M.J., Altshuler,D., Shendure,J. *et al.* (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216–220.
44. 1000 Genomes Project Consortium, Auton,C., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
45. Glusman,G., Caballero,J., Mauldin,D.E., Hood,L. and Roach,J.C. (2011) Kaviar: an accessible system for testing SNV novelty. *Bioinformatics*, **27**, 3216–3217.
46. McCarthy,S., Das,S., Kretzschmar,W., Delaneau,O., Wood,A.R., Teumer,A., Kang,H.M., Fuchsberger,C., Danecek,P., Sharp,K. *et al.* (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, **48**, 1279–1283.
47. Hebsgaard,S.M., Korning,P.G., Tolstrup,N., Engelbrecht,J., Rouze,P. and Brunak,S. (1996) Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.*, **24**, 3439–3452.
48. Divina,P., Kvitkovicova,A., Buratti,E. and Vorechovsky,I. (2009) Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping. *Eur. J. Hum. Genet.*, **17**, 759–765.
49. Raponi,M., Kralovicova,J., Copson,E., Divina,P., Eccles,D., Johnson,P., Baralle,D. and Vorechovsky,I. (2011) Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6. *Hum. Mutat.*, **32**, 436–444.
50. Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q. and Krainer,A.R. (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.
51. Dogan,R.I., Getoor,L., Wilbur,W.J. and Mount,S.M. (2007) SplicePort—an interactive splice-site analysis tool. *Nucleic Acids Res.*, **35**, W285–W291.
52. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.
53. The Gene Ontology,C. (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331–D338.
54. Li,T., Wernersson,R., Hansen,R.B., Horn,H., Mercer,J., Slodkowicz,G., Workman,C.T., Rigina,O., Rapacki,K., Staerfeldt,H.H. *et al.* (2017) A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Methods*, **14**, 61–64.
55. Petrovski,S., Gussow,A.B., Wang,Q., Halvorsen,M., Han,Y., Weir,W.H., Allen,A.S. and Goldstein,D.B. (2015) The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genet.*, **11**, e1005492.
56. Fadista,J., Oskolkov,N., Hansson,O. and Groop,L. (2017) LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics*, **33**, 471–474.
57. Itan,Y., Shang,L., Boisson,B., Patin,E., Bolze,A., Moncada-Velez,M., Scott,E., Ciancanelli,M.J., Lafaille,F.G., Markle,J.G. *et al.* (2015) The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc. Natl Acad. Sci. U.S.A.*, **112**, 13615–13620.
58. Han,X., Chen,S., Flynn,E., Wu,S., Wintner,D. and Shen,Y. (2018) Distinct epigenomic patterns are associated with haploinsufficiency and predict risk genes of developmental disorders. *Nat. Commun.*, **9**, 2138.
59. Aggarwala,V. and Voight,B.F. (2016) An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.*, **48**, 349–355.
60. Shihab,H.A., Rogers,M.F., Campbell,C. and Gaunt,T.R. (2017) HIPred: an integrative approach to predicting haploinsufficient genes. *Bioinformatics*, **33**, 1751–1757.
61. Tate,J.G., Bamford,S., Jubb,H.C., Sondka,Z., Beare,D.M., Bindal,N., Boutselakis,H., Cole,C.G., Creatore,C., Dawson,E. *et al.* (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **47**, D941–D947.
62. Sondka,Z., Bamford,S., Cole,C.G., Ward,S.A., Dunham,I. and Forbes,S.A. (2018) The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*, **18**, 696–705.
63. Chakravarty,D., Gao,J., Phillips,S.M., Kundra,R., Zhang,H., Wang,J., Rudolph,J.E., Yaeger,R., Soumerai,T., Nissan,M.H. *et al.* (2017) OncoKB: a precision oncology knowledge base. *JCO Precis Oncol.*, **2017**, PO.17.00011.
64. Vettore,S., De Rocco,D., Gerber,B., Scandellari,R., Bianco,A.M., Balduini,C.L., Pecci,A., Fabris,F. and Savoia,A. (2010) A G to C transversion at the last nucleotide of exon 25 of the MYH9 gene results in a missense mutation rather than in a splicing defect. *Eur. J. Med. Genet.*, **53**, 256–260.
65. Zhang,S., He,Y., Liu,H., Zhai,H., Huang,D., Yi,X., Dong,X., Wang,Z., Zhao,K., Zhou,Y. *et al.* (2019) regBase: whole genome base-wise aggregation and functional prediction for human non-coding regulatory variants. *Nucleic Acids Res.*, **47**, e134.
66. Vaz-Drago,R., Custodio,N. and Carmo-Fonseca,M. (2017) Deep intronic mutations and human disease. *Hum. Genet.*, **136**, 1093–1111.
67. Li,K., Wang,G., Lv,M., Wang,J., Gao,Y., Tang,F., Xu,C., Yang,W., Yu,H., Shao,Z. *et al.* (2022) Bi-allelic variants in DNAH10 cause

asthenoteratozoospermia and male infertility. *J. Assist. Reprod. Genet.*, **39**, 251–259.
68. Huang,D., Zhou,Y., Yi,X., Fan,X., Wang,J., Yao,H., Sham,P.C., Hao,J., Chen,K. and Li,M.J. (2022) VannoPortal: multiscale functional annotation of human genetic variants for interrogating molecular mechanism of traits and diseases. *Nucleic Acids Res.*, **50**, D1408–D1416.
69. Hyung,D., Kim,J., Cho,S.Y. and Park,C. (2018) ASpedia: a comprehensive encyclopedia of human alternative splicing. *Nucleic Acids Res.*, **46**, D58–D63.