PDIVAS: Pathogenicity predictor for Deep-Intronic Variants causing Aberrant Splicing

Ryo Kurosawa[1*], Kei Iida[2,3], Masahiko Ajiro[4,5], Tomonari Awaya[1,6], Mamiko Yamada[7], Kenjiro Kosaki[7] and Masatoshi Hagiwara[1*]

Abstract

Background Deep-intronic variants that alter RNA splicing were ineffectively evaluated in the search for the cause of genetic diseases. Determination of such pathogenic variants from a vast number of deep-intronic variants (approximately 1,500,000 variants per individual) represents a technical challenge to researchers. Thus, we developed a Pathogenicity predictor for Deep-Intronic Variants causing Aberrant Splicing (PDIVAS) to easily detect pathogenic deep-intronic variants.

Results PDIVAS was trained on an ensemble machine-learning algorithm to classify pathogenic and benign variants in a curated dataset. The dataset consists of manually curated pathogenic splice-altering variants (SAVs) and commonly observed benign variants within deep introns. Splicing features and a splicing constraint metric were used to maximize the predictive sensitivity and specificity, respectively. PDIVAS showed an average precision of 0.92 and a maximum MCC of 0.88 in classifying these variants, which were the best of the previous predictors. When PDIVAS was applied to genome sequencing analysis on a threshold with 95% sensitivity for reported pathogenic SAVs, an average of 27 pathogenic candidates were extracted per individual. Furthermore, the causative variants in simulated patient genomes were more efficiently prioritized than the previous predictors.

Conclusion Incorporating PDIVAS into variant interpretation pipelines will enable efficient detection of disease-causing deep-intronic SAVs and contribute to improving the diagnostic yield. PDIVAS is publicly available at https://github.com/shiro-kur/PDIVAS.

Keywords Pathogenicity prediction, RNA splicing, Deep intron, Non-coding region, Genomics, Machine learning, Variant interpretation
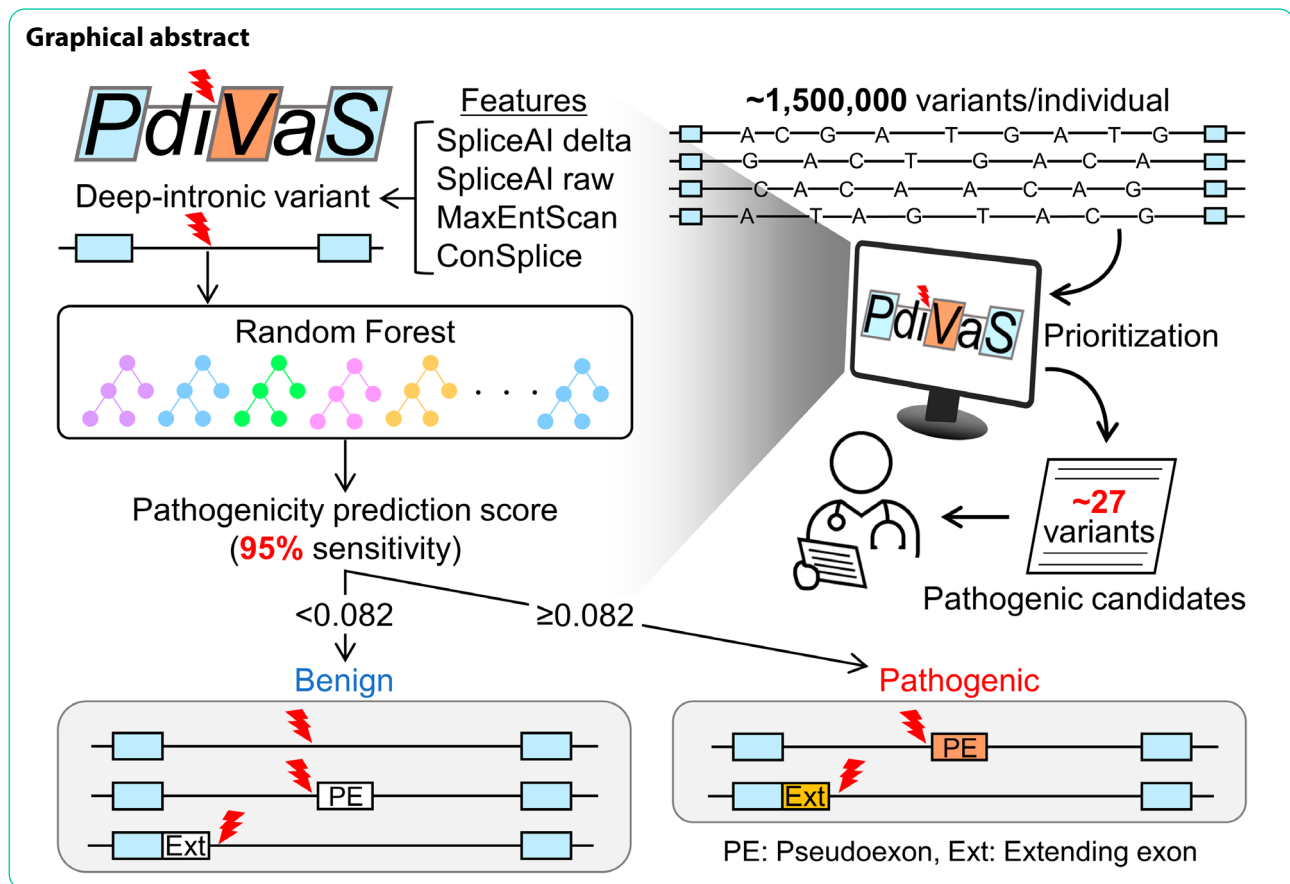
*Correspondence:
Ryo Kurosawa
kurosawa.ryo.43r@st.kyoto-u.ac.jp
Masatoshi Hagiwara
hagiwara.masatoshi.8c@kyoto-u.ac.jp
Full list of author information is available at the end of the article

Graphical abstract

Background

Causative variants of rare genetic diseases remain to be determined in 50–75% of patients, regardless of the technical progress of exome sequencing and genome sequencing [1–4]. One of the major obstacles in the diagnostic process is the technical difficulty of evaluating genetic variants in deep-intronic regions. Intronic variants with splice alterations have been reported as the causative variants of dystrophinopathy, neurofibromatosis type I, and inherited retinal diseases [5–10] because these splice-altering variants (SAVs) create pathogenic pseudoexons or extend existing exons, by affecting recognition by splicing factors (e.g., small nuclear ribonucleoprotein and RNA-binding proteins) (Fig. 1a). The resulting splicing alterations subsequently lead to mRNA destabilization by nonsense-mediated decay (NMD) or functional defects in encoded proteins. Previously, most of the pathogenic deep-intronic SAVs were discovered through the "RNA-based" diagnosis using RT-PCR or RNA sequencing on patient-derived cells or tissues [11–13]. However, causative genes are often expressed in patients' specific tissues, such as the brain and heart, which are rarely available to clinicians.

In contrast to the RNA samples, the genetic variants of almost all genes can be identified through genome sequencing analysis on blood or skin samples, which are readily accessible. However, distinguishing pathogenic deep-intronic variants from the vast number of benign variants is challenging owing to the presence of over 1,800,000 intronic variants in an individual [14]. Therefore, the variants must be filtered through a computational process with pathogenicity predictors to enable clinicians to manually evaluate the candidate variants.

This study presents a novel pathogenicity predictor for deep-intronic variant prioritization, called the Pathogenicity predictor for Deep-Intronic Variant causing Aberrant Splicing (PDIVAS). We demonstrate that PDIVAS offers a clinically applicable strategy for deep-intronic variant prioritization in the diagnosis of rare genetic diseases.

Methods

Variant annotation

The effect of variants on genes was annotated by the Ensembl variant effect predictor (VEP) (version 105, GRCh37) by referring to one transcript per genome region selected by the "--pick_allele_gene" option. The transcript annotation was based on GENCODE V196 [15, 16]. We selected deep-intronic variants according to HGVS nomenclature that describes the relative
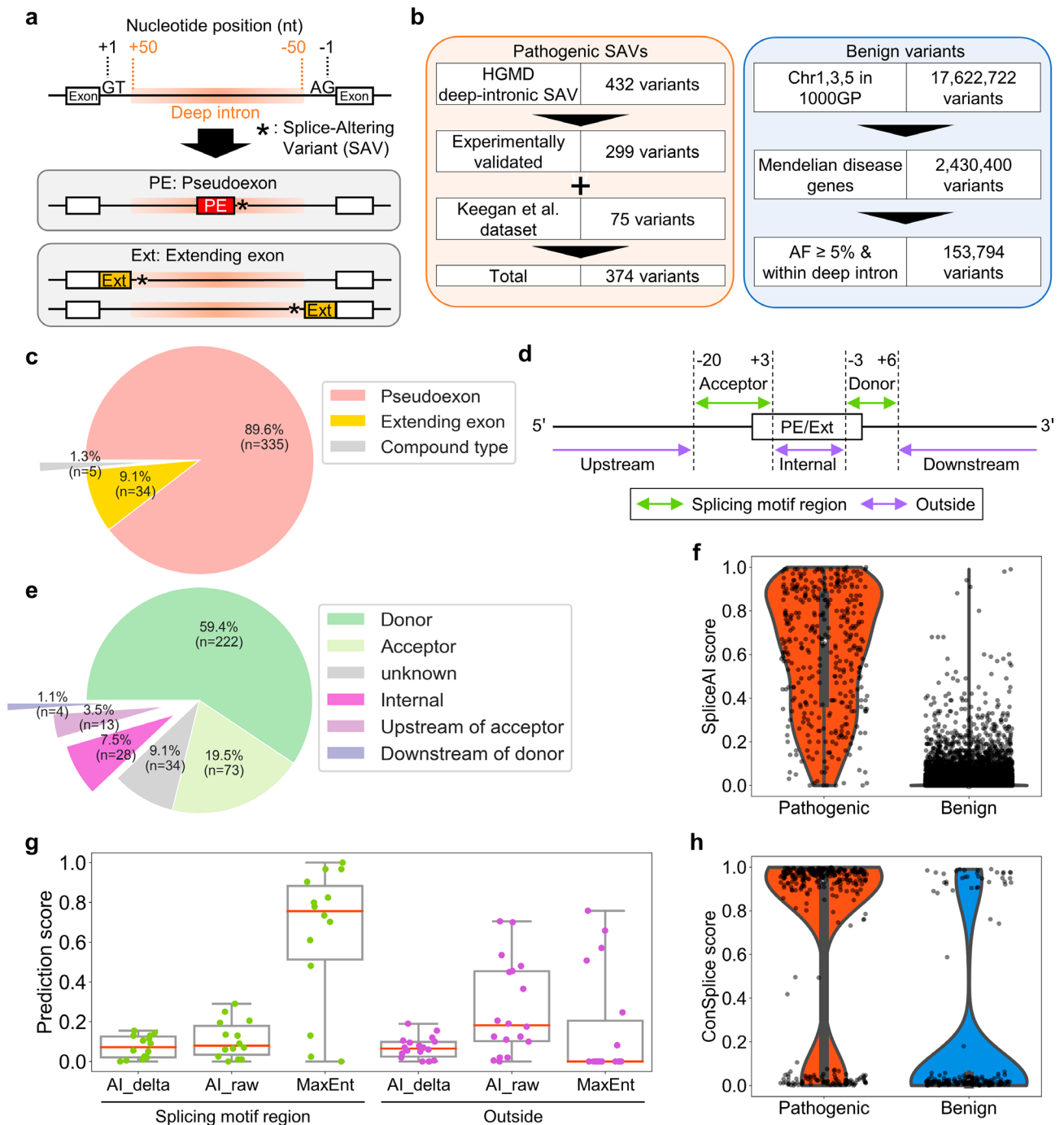
Fig. 1 Prospective features to classify pathogenic and benign variants in a curated dataset. **a** Schematic representation of aberrant splicing induced by deep-intronic splice-altering variants (SAVs). **b** The curated dataset consistes of pathogenic deep-intronic SAVs from the Human Gene Mutation Database (HGMD) and pseudo exon dataset by Keegan et al.[7] and benign deep-intronic variants from the 1000 Genomes Project (1000GP). **c** Splice-type classification of pathogenic SAVs in the curated dataset. Compound type refers to the SAVs that cause both a pseudoexon and an extending exon in the intron region. **d** Definition of region names based on the relative position to the splice sites of pseudoexon (PE) or extending exon (Ext). **e** Classification of pathogenic SAVs on their relative positions to the nearest splice sites of the pseudoexon or extending exons. **f** Violin plot, box plot, and strip plot depicting maximums of SpliceAI donor/acceptor delta gain scores for variants in the curated dataset. **g** Violin plot, box plot, and strip plot depicting prediction scores by mean values of SpliceAI donor/acceptor delta gain scores, mean values of SpliceAI donor/acceptor raw gain scores, and MaxEntScan scores for pathogenic SAVs with ≤0.2 maximums of SpliceAI donor/acceptor delta gain scores (n=44). **h** Violin plot and strip plot depicting ConSplice scores (human splicing constraint metrics) for each variant in the curated dataset with >0.2 maximums of SpliceAI donor/acceptor delta gain scores (n=190). AF, Allele frequency

position of the variant within the gene [17]. The annotation of allele frequency in the gnomAD genome sequencing database (r.2.0.1) was conducted through the VEP "--custom" option referring to the tabix-indexed VCF file downloaded from the Ensembl FTP site [18].

The features from SpliceAI were obtained from the output-customized SpliceAI whose original version was 1.3.1. By default, SpliceAI only outputs four delta scores of acceptor gain/loss and donor gain/loss, which are calculated by subtracting the splice site score on the reference sequence from the alternative sequence with a variant of interest [19]. Our customized SpliceAI script outputs the raw scores for the alternative sequence before subtraction, as well. Furthermore, while default SpliceAI converts the delta scores (acceptor/donor gain) to zero when the predicted site matches the nearest annotated splice site on the gene (with the "-m" mask option), we extended the mask function to all of the referred annotated splice sites on the gene to reduce the number of false positives. The SpliceAI delta_gain_max and mean scores were calculated as the maximum and mean values of acceptor and donor gain delta scores, respectively. The SpliceAI raw_gain_mean score was calculated as the mean value of the raw scores of acceptors and donors. Finally, SpliceAI predictions were performed with the distance option of "-d 300".

The ConSplice feature was obtained from the score-precomputed bed file of the best_splicing_constraint_model provided by the authors [20]. MaxEntScan prediction of the variant's effect on splicing was performed using the plugin module of VEP [21–23]. The feature-extraction algorithm from MaxEntScan is described in Supplementary Figure S1, which corresponds to the interpretation algorithm of Shamsani et al. [23].

Predictors to be compared to PDIVAS were SpliceAI, Pangolin [24], ConSpliceML [20], MaxEntScan, and CADD-Splice [25], which were selected based on the following criteria: (1) the program or the precomputed score file is freely available, (2) the program can assess deep-intronic variants, and (3) the program is operated in a Linux environment and can be applied to large-scale variant analysis. SQUIRLS and SPiP also matched these criteria, but their developers recognized their lower performance on deep-intronic SAVs because of the limited number of training datasets [26, 27]. Therefore, we did not include them in this comparison. Pangolin (v1.0.1) was installed and run with the mask option (--mask True) and distance option (-d 300). To annotate ConSpliceML, we downloaded the score-precomputed VCF file provided by the authors and converted the genome version from GRCh38 to GRCh37 using Picard LiftoverVcf (v.2.27.1). The annotation of ConSpliceML was conducted through the VEP "--custom" option, referring to the precomputed VCF file on the GRCh37 version.

Annotations in CADD-Splice (v1.6) were conducted on plugin modules of the Variant Effect Predictor (VEP). BCFtools, bash scripts, cyvcf2, and Python scripts were used to process the VCF files used in this research [28, 29].

Curated dataset of deep-intronic variants

In this study, we defined "deep intron" as a genomic region ≥ 50 bp away from a splice site of the nearest exon [19]. We tuned our novel predictive model focusing on variants located in the genes responsible for Mendelian diseases (hereinafter called Mendelian disease genes). The model's primary objective is to detect novel causative variants in such known disease-causing genes rather than find unknown disease-causing genes. The Mendelian disease gene lists were collected from the Online Mendelian Inheritance in Man (OMIM) and the Clinical Genomic Database (CGD) [30, 31]. In the OMIM gene list, there were numerous genes whose causative relationships with the registered phenotype were unclear, as well as those that only contributed to the susceptibility to multifactorial diseases (e.g., diabetes and asthma). To only focus on the genes whose contribution to the phenotype is clear in a way of Mendelian inheritance, we extracted genes with annotations of the mode of inheritance (autosomal dominant/recessive, X-linked, Y-linked), and the phenotype mapping key of the molecular basis of the diseases is known. Then, the OMIM and CGD gene lists were combined and non-coding genes were filtered out following the GENCODE V19 annotation, resulting in a final list of Mendelian disease genes (https://github.com/shiro-kur/PDIVAS).

For the benign dataset, the variant lists from the 1000 Genomes Project (Phase 3, GRCh37) were downloaded from the UCSC FTP site. Variants on chromosomes 1, 3, and 5 whose sequences were not used to train SpliceAI were extracted to reduce data size. Additionally, copy number variants and variants in the multi-nucleotide variant format were removed because they could not be evaluated with many of the predictors used in this research. We annotated the remaining variants with VEP and extracted only deep intronic variants with ≥ 5% allele frequency in the 1000 Genomes Project and gnomAD population. Subsequently, Pathogenic deep-intronic SAVs were collected from Human Gene Mutation Database Professional online (HGMD) 2020 and the Keegan et al. dataset [7, 32]. From HGMD, we extracted deep-intronic single nucleotide variants (SNVs) labeled with "disease-causing" and "splicing". We annotated the variants with VEP and variants with < 1% allele frequency in the gnomAD population were extracted. The extracted variants were checked to determine if splice alterations were experimentally validated by reading all of the original reports. The variants whose splice alterations

were validated using RT-PCR or RNA sequencing for patient RNA or minigene/midigene-expressed RNA were included in the final variant dataset. Other variants whose aberrant splicing was only predicted or not specified were filtered out. Keegan et al. [7] also cataloged deep-intronic SAVs with experimental validation in a process independent of HGMD and our reinspection. From the Keegan et al. dataset, SNVs and short deletions (~54nt) were obtained and annotated with VEP to extract only deep-intronic variants, those with <1% allele frequency in the gnomAD population, and located on the Mendelian disease genes. Non-overlapped SAVs were added to the reinforced HGMD dataset to complete the pathogenic dataset. Through the curation process, the relative variant positions to the nearest splice sites in the pseudoexons or extending exons were also checked and classified as "splicing motif region" or "outside" as described in Fig. 1d. The splicing motif region (region of the splice donor and acceptor site motif) conforms to the region defined in MaxEntScan [21].

Model training and testing

For PDIVAS construction, a random forest model provided by scikit-learn (version 1.0.2) was used and trained on the curated dataset. We randomly split the curated dataset into 70% for model training and the remaining 30% for testing the trained model. Furthermore, within the training dataset, five-fold cross-validation was performed. In this process, the training dataset was equally split into five sub-datasets. Four of the five sub-datasets were used for parameter tuning of random forest, and the remaining sub-dataset was used to evaluate the trained model. This process was iteratively repeated for all five combinations of the training and test sub-datasets. The best combination of hyperparameters was determined using the five-fold cross-validation when the average precision metrics were maximized. The average precision was computed with scikit-learn (version 1.0.2) as the weighted mean of the precision attained at each threshold, divided by the difference in recall from the previous threshold. Average precision is more accurate than the area under the curve because it does not require empirical curve construction [33]. The tuned hyperparameters in this model were: (1) the number of decision trees in the forest and the maximum depth of each tree; (2) the ratio of samples extracted from the whole training dataset to train each tree; and (3) the number of features to be used in each tree. Using these optimal hyperparameters, the five-fold cross-validation was conducted again to compare the classification performance of the intermediate models of PDIVAS with SpliceAI and Pangolin and evaluate the stability of the PDIVAS performance. Finally, the random forest model was trained on the entire training dataset using the optimized hyperparameters. The

final PDIVAS model was evaluated on the test dataset with average precisions and maximums of the Matthews correlation coefficient (MCC). For maximum MCC, we calculated individual MCC on every threshold in increments of 0.001 score, and the maximum of the values was extracted from each predictor. We employed these metrics for the performance comparison because they were stable even when the classes in a dataset were highly imbalanced. The predictive performance of PDIVAS was compared to the maximum of SpliceAI acceptor/gain scores, Pangolin gain score, ConSpliceML, MaxEntScan, and CADD-Splice Phred scores. ConSpliceML and CADD-Splice could not score some of the variants because their score-precomputed files did not cover all variant types, such as genomic insertion and deletion. When we calculated the statistical measures, we only referred to the variants with scores.

Variant prioritization for the 1000 genomes project

We referred to the same variant lists of the 1000 Genomes Project under the sub-heading "Curated dataset of deep-intronic variants", above. Variants on chrY were filtered out because the ConSplice score for the PDIVAS score was not available there. Additionally, copy number variants and variants in the multi-nucleotide variant format were removed because they could not be evaluated with numerous predictors used in this study. From the remaining variants, deep-intronic variants of 20,731 protein-coding genes defined on GENCODE V19 were extracted [16]. Subsequently, variants on Mendelian disease genes and variants with <1% allele frequency in both the 1000 Genomes Project and the gnomAD populations were extracted as candidates for pathogenic variants. Finally, the rare variants were scored by PDIVAS, SpliceAI, and Pangolin and filtered by thresholds based on the sensitivity for pathogenic SAVs.

Speed and memory benchmarking

The benchmarking comparing PDIVAS, SpliceAI, and Pangolin pipelines was performed on Intel Xeon Gold 6154 CPUs @ 3.0 GHz (no GPUs). We prepared five VCF files, each containing 10,000 randomly selected deep-intronic variants from the 1000 Genomes Project. Each computation of VEP, SpliceAI, Pangolin, output-customized SpliceAI, and Pangolin was performed on the 10 threads, and its wall time (i.e., the total time to complete the variant annotation) and maximum RAM usage were recorded. The mean wall time and mean RAM usage for five VCF files were used as representative values.

Results

We developed PDIVAS, a novel machine-learning framework to evaluate the pathogenicity of deep-intronic variants. PDIVAS was constructed and optimized through

(1) setting up a labeled dataset of known pathogenic and benign deep-intronic variants, (2) selecting a set of features to classify pathogenic and benign variants, and (3) training the random forest model with selected features. Finally, we evaluated the performance of PDIVAS versus previous predictors through the classification of pathogenic and benign deep-intronic variants, genome sequencing analysis in control individuals, and simulation analysis of patient genome sequences.

Curating a dataset consisting of pathogenic and benign deep-intronic variants

To train the PDIVAS model on truly pathogenic deep-intronic variants, we surveyed the original reports of the variants and collected only those pathogenic variants with experimentally validated splice alterations. As data sources of pathogenic deep-intronic SAVs, we referred to HGMD and a pathogenic pseudoexon dataset constructed by Keegan et al. [7, 32]. From HGMD, we extracted 432 variants located within deep introns out of 26,610 SNVs annotated as both "disease-causing" and "splicing mutation" (Fig. 1b). In this study, we defined deep introns as intronic regions ≥50 bases from the nearest annotated splice sites (Fig. 1a). Reinspection of all the originally reported papers revealed that some lacked experimental validation of their occurrence of aberrant splicing by either splicing minigene/midigene assays or patient-derived RNA sample analysis. To construct a high-quality training dataset, we used only pathogenic SAVs with such experimental evidence. Furthermore, we focused on Mendelian disease genes (n=4,429) (Supplementary Methods). This filtering also clarifies the causality of the variants and the patient phenotypes. We then extracted variants with <1% allele frequency in the gnomAD population. These procedures yielded 290 variants curated as pathogenic SAVs from HGMD (Fig. 1b).

The SAVs curated by Keegan et al. (n=359) were already checked with experimental validation. We also extracted those on Mendelian disease genes with <1% allele frequency in the gnomAD population. The obtained variants were added exclusively to the variant list obtained from HGMD. Our combined approach yielded a final list of 374 pathogenic deep-intronic SAVs located in 180 genes. Among them, 335 SAVs (89.6%) were reported to cause pseudoexons and 34 SAVs (9.1%) caused extending exons, while the remaining 5 SAVs (1.3%) were reported to cause both pseudoexons and extending exons (compound type) (Fig. 1c). The majority of the variants caused aberrant splicing by creating novel splice donors (59.4%) or splice acceptors (19.5%) (Fig. 1d and e). The other variants were outside the splicing motif regions and might influence splicing enhancers or silencers, leading to aberrant splicing.

Benign deep-intronic variants were also collected from genome sequencing data of the 1000 Genomes Project comprised of 2,504 control individuals. We extracted 17,622,722 variants from chromosomes 1, 3, and 5 to reduce the data size (Fig. 1b). After extracting variants on Mendelian disease genes, variants with ≥5% allele frequency in the 1000 Genomes Project and the gnomAD population were collected to assure the benign nature of assorted variants, and the resulting list consisted of 153,794 benign deep-intronic variants. The entire dataset consisting of these pathogenic and benign variants will be called "the curated dataset" hereafter.

Seeking effective features through characterization of the curated dataset

We first reviewed the characteristics of the curated dataset using SpliceAI, a state-of-the-art splicing predictor constructed on a deep neural network. We scored the variants in the curated dataset with SpliceAI and evaluated the results with a 0.2 threshold, a high-sensitivity threshold provided by the developers (Fig. 1f) [19]. As a result, 338 pathogenic deep-intronic SAVs (90% of 374 SAVs) were above the threshold. This indicates that there was still a need for improvement in sensitivity for clinical use by decreasing the number of false negatives. In contrast, 190 benign variants exceeded the threshold. Decreasing the number of these false positives would enable more efficient prioritization of candidate causative variants.

We looked for additional features to support these mispredictions by SpliceAI. For false negatives, we assumed that the result could be effectively complemented by dividing the causes by splice type (SAVs either in splicing motif regions or on the outside (Fig. 1d)). Fourteen SAVs (4.7%) were false negatives in the splicing motif regions (n=295, in total) (Fig. 1g). The false negatives were believed to be supported by MaxEntScan because it specializes in evaluating the splicing motif regions, and its superiority to other splice site motif predictors was previously demonstrated [27]. As expected, MaxEntScan predicted these pathogenic SAVs with higher scores than SpliceAI (0.76 vs. 0.07 *median values*). On the other hand, of the 45 SAVs outside the splicing motif region, 18 SAVs (40%) were false positives. The tendency toward lower sensitivity for SAVs outside of splice sites than in the splicing motif region is consistent with a previous report (Supplementary Figure S2a) [34]. Although the previous report recommended the use of ESRseq scores [35] to support the evaluation of such SAVs affecting splicing enhancers and silencers, the combination of ESRseq scores and SpliceAI resulted in lower predictive specificity while the sensitivity was improved. In this study, we used the SpliceAI raw score as an alternative. The SpliceAI raw score is the splice site score for

the alternative sequence containing the SAV before the score for the reference sequence is subtracted. The score is called the delta score when the reference sequence is subtracted, which is the default output of SpliceAI. We observed that splice site scores on reference sequences were predicted with higher scores for SAVs outside the splicing motif regions than for those within, which might be due to the presence of pre-existing splicing motifs in these regions (Supplementary Figure S2c). Consistently, for the SAVs outside of splicing motif regions with ≤0.2 SpliceAI delta scores, the SpliceAI raw scores were higher than the delta scores (0.18 vs. 0.06 *median values*). This tendency was not as strong for the SAVs in the splicing motif regions (0.08 vs. 0.07 *median values*) (Fig. 1g). This implied that the SpliceAI raw score would improve the predictive sensitivity for SAVs outside the splice sites.

Subsequently, we considered the false-positive results using SpliceAI. We hypothesized that certain predicted variants could cause aberrant splicing, although they had non-deleterious effects on physiological function. However, SpliceAI does not evaluate the deleterious effect of the splicing event because SpliceAI is not trained in pathogenic splicing events. Therefore, a more specific prediction was achieved by combining a deleterious prediction with a human splicing constraint metric from ConSplice, which models mutational constraints on splice-altering variants within the human population. The effectiveness of this approach in predicting deleterious splicing was demonstrated in the ConSpliceML predictor [20]. Likewise, we employed ConSplice for more specialized usage for deep-intronic SAVs. Of the 190 benign variants with ≥0.2 SpliceAI delta score, ConSplice evaluated 153 (81%) variants as being less constrained and having less deleterious effects (<0.2 ConSplice) (Fig. 1h). These findings suggest that the combinatorial use of the SpliceAI delta score, the SpliceAI raw score, MaxEntScan, and ConSplice would be a better pathogenicity predictor than the sole use of the SpliceAI delta score.

PDIVAS performs best in predicting pathogenic deep-intronic SAVs

To verify the advantages of using SpliceAI delta scores, SpliceAI raw scores, MaxEntScan, and ConSplice together, we combined these features into one pathogenicity predictor called PDIVAS. PDIVAS is modeled on a random forest classifier where multiple decision trees are built in parallel and each decision tree defines an input deep-intronic variant as pathogenic or benign in a binary manner (one or zero, respectively), referring to these features (Fig. 2a). The PDIVAS calculated the final prediction score as the fraction of trees that classified the variants as pathogenic. Therefore, a higher PDIVAS score indicates a more likely pathogenic splice alteration predicted in the deep-intronic variant. The score

range is between one and zero. To train the random forest model, we randomly split the curated dataset into a training dataset and a test dataset. The training dataset represented 70% (261 pathogenic and 107,655 benign variants) of the entire dataset and was used to tune the parameters of the random forest model. The independent test dataset contained the remaining 30% with 113 pathogenic and 46,139 benign variants and was used as a hold-out test dataset to evaluate the final random forest model. The predictive accuracy of the final trained model (PDIVAS) was compared with that of SpliceAI, Pangolin, ConSpliceML, MaxEntScan, and CADD-Splice. Pangolin is a model based on SpliceAI architecture. Its training dataset was augmented by incorporating splice site data detected from four mammalian species into the human dataset. ConSpliceML is a random forest model in which ConSplice and the state-of-the-art splicing predictors of SpliceAI and SQUIRLS were used as features to classify pathogenic SAV and benign variants. CADD-Splice is an L2-regularized logistic regression model that incorporates various features of conservation scores, transcription factor binding, DNase I hypersensitivity regions, and splicing features driven by the neural network-based splicing predictor of SpliceAI and MMSplice [36].

To compare the predictive accuracy, pathogenic and benign variants were scored with those predictors (Fig. 2b). Subsequently, the precision, recall, and MCC of these predictors at all thresholds were calculated and described as curves (Fig. 2c). A performance comparison of the predictors was performed on the average precision and maximum MCC (Fig. 2d). PDIVAS achieved the highest performance scores (average precision of 0.92 and maximum MCC of 0.88) of the six predictors. Further, we verified the stable competitiveness of PDIVAS on different training datasets, using five-fold cross-validation on the training dataset. Regardless of the composition of the dataset, PDIVAS was consistently superior to SpliceAI and Pangolin in terms of average precision (0.90 vs. 0.84, 0.77 *median values*) and maximum MCC (0.85 vs. 0.81, 0.76 *median values*) (Supplementary Figure S3). These results indicate that PDIVAS outperforms the other state-of-the-art predictors, in predicting pathogenic deep-intronic SAVs.

Threshold settings for clinical use

PDIVAS thresholds were set based on the sensitivity for pathogenic SAVs in the PDIVAS test dataset (Table 1). Considering regular clinical cases where sensitivity is required rather than specificity, the highest sensitivity (95%) threshold of 0.082   is recommendable. However, Table 1 can also be used as a guide to customize the usage of PDIVAS, depending on research objectives and circumstances. For the later performance comparison between predictors, each threshold was also set on
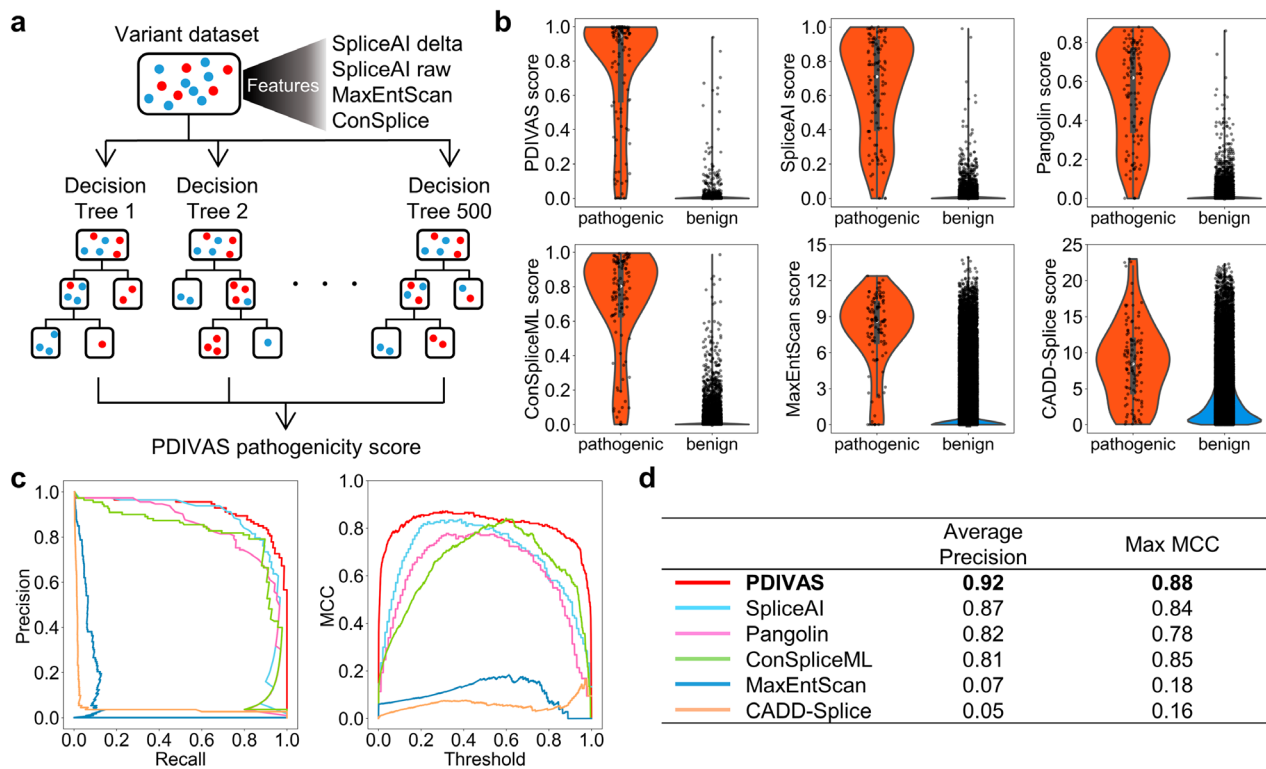
Fig. 2 PDIVAS outperforms the existing predictors in classifying pathogenic and benign variants. **a** A graphic of random forest model used in PDIVAS. The model comprises 500 decision trees to classify pathogenic and benign variants (red and blue plots, respectively), combining multiple features from splicing predictors (output-customized SpliceAI and MaxEntScan) and a human splicing constraint metric of ConSplice. **b** Violin plot, box plot, and strip plot indicating the scores of PDIVAS and five previous   predictors for the test dataset (pathogenic, n = 113; benign, n = 46,139). SpliceAI score is represented as the maximum SpliceAI donor/acceptor delta gain scores. Pangolin score is represented as the Pangolin splice gain score. **c** PR and MCC curve analysis for the PDIVAS and five published predictors. Precision, recall, and MCC are calculated at every threshold of their predictors and are depicted as these curves. The MCC values of Pangolin, MaxEntScan, and CADD-Splice are computed on prediction scores scaled from zero to one for the entire curated dataset (min-max normalization). **d** Comparative evaluation of predictive accuracy on average precision and maximum MCC. PR curve, precision-recall curve; MCC, Matthews correlation coefficients

Table 1 Clinically relevant threshold settings on the predictive sensitivity

Sensitivity (%)	PDIVAS threshold	SpliceAI threshold	Pangolin threshold
60	≥ 0.899	≥ 0.56	≥ 0.48
65	≥ 0.832	≥ 0.50	≥ 0.42
70	≥ 0.763	≥ 0.44	≥ 0.39
75	≥ 0.575	≥ 0.38	≥ 0.34
80	≥ 0.501	≥ 0.33	≥ 0.28
85	≥ 0.340	≥ 0.25	≥ 0.23
90	≥ 0.151	≥ 0.20	≥ 0.16
95	≥ 0.082	≥ 0.11	≥ 0.09

PDIVAS threshold is set on the test dataset while others' thresholds are set on the entire dataset. SpliceAI score is represented as the maximum SpliceAI donor/acceptor delta gain scores. Pangolin score is represented as the Pangolin splice gain score

SpliceAI and Pangolin based on all pathogenic SAVs in the curated dataset.

Features in PDIVAS synergistically work to improve predictive accuracy

In this section, we investigate how the five features in PDIVAS were used to improve the predictive accuracy. Firstly, we evaluated the contribution of these features to the extent to which they increased the purity of pathogenic and benign variants in each node of the decision trees (Fig. 3a). This evaluation showed that the most important contributed feature of the five was SpliceAI_delta_gain_mean, the mean of the splice acceptor and donor scores. This result is reasonable because the most frequent splice type caused by deep-intronic SAV is a pseudoexon, where both the splice acceptor and donor are newly recognized by spliceosome machinery and both gain scores should be considered (Fig. 1c) (Supplementary Figure S4). Meanwhile, extending exons are caused by the creation of only one of the splice sites, and consistently, one of the SpliceAI gain scores is nearly zero (Supplementary Figure S4). Therefore, the SpliceAI_delta_gain_mean might underestimate extending exons. Extending exons were believed to be more appropriately
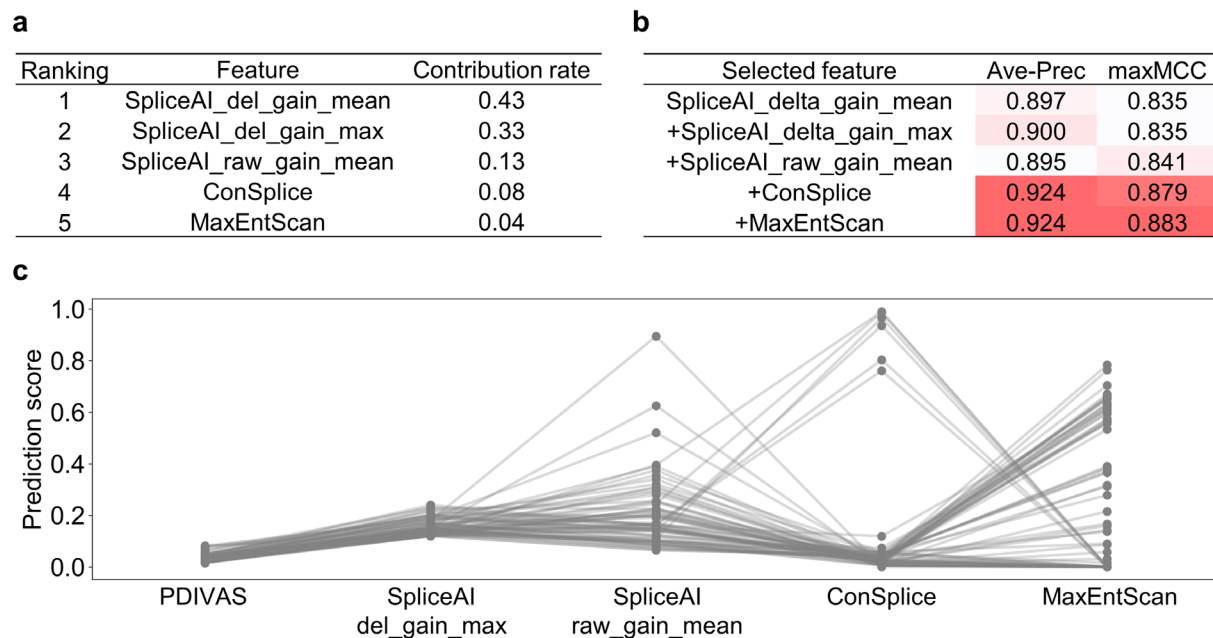
a

Ranking	Feature	Contribution rate
1	SpliceAI_del_gain_mean	0.43
2	SpliceAI_del_gain_max	0.33
3	SpliceAI_raw_gain_mean	0.13
4	ConSplice	0.08
5	MaxEntScan	0.04

b

Selected feature	Ave-Prec	maxMCC
SpliceAI_delta_gain_mean	0.897	0.835
+SpliceAI_delta_gain_max	0.900	0.835
+SpliceAI_raw_gain_mean	0.895	0.841
+ConSplice	0.924	0.879
+MaxEntScan	0.924	0.883

c

Fig. 3 Revealing feature contributions on PDIVAS. **a** The feature contribution rates in the PDIVAS model are computed as the mean of the accumulation of the impurity decrease within each decision tree. SpliceAI_del_gain_mean indicates the mean values of SpliceAI donor/acceptor delta gain scores. SpliceAI_del_gain_max indicates the maximum of them. SpliceAI_raw_gain_max indicates the maximum of SpliceAI donor/acceptor raw gain scores. **b** The predictive accuracy is evaluated when the random forest model is trained on increased features one by one, starting with only one feature of SpliceAI_del_gain_mean. For each feature subgroup, the same datasets are used to train and evaluate each model. This shows the combinational effect of respective features in improving the average precision (Ave-Prec) or a maximum MCC (maxMCC). **c** Comparison of feature scores of benign variants that are predicted as pathogenic by SpliceAI (maximums of SpliceAI donor/acceptor delta gain scores), but correctly predicted as benign by PDIVAS. Thresholds are set to 95% sensitivity for reported pathogenic SAVs (Table 1)

evaluated by the second-most contributed feature, SpliceAI_delta_gain_max only referring to the higher score. The other features SpliceAI_raw_gain_mean, ConSplice, and MaxEntScan also contributed to classifying variants in each node of the decision trees at 13%, 8%, and 4%, respectively.

Further, we demonstrated the contributions of these features using the feature-cumulative training method (Fig. 3b). This involved training a random forest model with progressively increasing numbers of features, beginning with SpliceAI_delta_gain_mean alone, and evaluating the predictive accuracy of each trained model on the same training/test dataset pair. This analysis showed that each feature of SpliceAI_raw_gain_mean, ConSplice, and MaxEntScan improves either the accuracy metrics of average precision or maximum MCC and has a synergetic effect on PDIVAS.

Finally, the feature contributions were studied within the prediction results of benign deep-intronic variants in the test data (n=46,139). When SpliceAI predicted them on the threshold with 95% sensitivity (Table 1), 129 benign variants were predicted as positives. Using PDIVAS on the threshold with 95% sensitivity, 73 variants (57%) of the 129 SpliceAI-predicted were correctly evaluated as negatives (Fig. 3c). Overviewing the feature distributions of the 73 variants suggested that PDIVAS

removed 67 of them because their genomic regions were in low constraint (<0.2 ConSplice) and the predicted SAVs could be tolerated. The remaining six variants had higher ConSplice scores, but the scores of MaxEntScan were zero and the splice alterations of themselves were not expected to occur. These results demonstrate that PDIVAS succeeded in improving the predictive specificity while retaining its high sensitivity for pathogenic deep-intronic SAVs by utilizing these five combined features.

PDIVAS minimizes candidate variants in genome sequencing samples at clinically relevant thresholds

In the clinical genome sequencing analysis, the disease-causing variants must be prioritized from the vast number of deep-intronic variants detected in genome sequencing samples. In this study, we analyzed genome sequencing samples of control individuals as a substitute for undiagnosed patients because patients with Mendelian diseases theoretically differ from other individuals by only one to two pathogenic variants. Control individuals have variants without splice alteration, benign SAVs, and pathogenic SAVs observed in carriers. Ideally, this analysis should extract only those pathogenic variants. Our analysis focused on the 2,504 control individuals from the 1000 Genomes Project [14]. As the genome

samples were collected from various populations, we would obtain more accurate statistics than the analysis on the limited number of patient genome samples, which are rarely publicly accessible. First, by extracting variants within deep introns in protein-coding genes from each genome sequencing sample, 1,570,571 deep-intronic variants per individual were obtained on a mean value (Fig. 4a and Supplementary Figure S5a). Assuming the actual genetic diagnosis process, we further extracted variants of 4,429 genes for Mendelian diseases and those with <1% allele frequency in both 1000 Genomes Project and gnomAD populations [18]. This process retained 14,872 variants per individual with possible pathogenicity at a mean value (Supplementary Figure S5b and S5c). Finally, the effects of their variants were predicted using SpliceAI, Pangolin, and PDIVAS. In this part, we chose only the two predictors to compare with because of their higher performance in the analysis of PR and MCC curves and their wider availability for the prediction of genomic insertions and deletions, as well as SNVs. As a result, PDIVAS predicted approximately 3.0-26.8 variant candidates (*mean values*) per individual on the thresholds with 70–95% sensitivity while SpliceAI predicted approximately 6.3–67.9 variant candidates, and Pangolin approximately 11.9-136.6 variant candidates (Fig. 4b and Supplementary Figure S5d). Even on the most sensitive threshold of 95%, PDIVAS predicted approximately 26.8 candidates, which is approximately 0.4 times and 0.2 times less than SpliceAI and Pangolin, respectively. These results demonstrate that PDIVAS narrows down the number of candidate pathogenic variants to the smallest number and would enable the most efficient clinical genetic diagnosis.

PDIVAS performs best in prioritizing causative variants in genome sequencing samples

We here demonstrated the clinical utility of PDIVAS through the simulation analysis of genetically undiagnosed patients through traditional variant interpretation focusing only on the protein-coding region and exon-intron boundaries. We virtually created patient genome sequencing samples by adding one pathogenic deep-intronic SAV (n=113) in the test dataset to one of the control genome sequencing samples (n=2,504) in the 1000 Genomes Project (Fig. 5a). Using this approach, 282,952 simulated patient genome sequencing samples with various genome backgrounds were obtained. This method was originally developed by Danis et al. [26]. Following the procedures in Fig. 4a, rare deep-intronic variants were extracted and their pathogenicity was predicted using PDIVAS, SpliceAI, and Pangolin for comparison. The predicted variants were ranked based on their scores, with the assumption that a higher-performance pathogenicity predictor could evaluate pathogenic SAVs with higher scores. Conversely, benign variants with non-deleterious splicing alterations or without splicing alterations were predicted to have lower scores. As a result, PDIVAS predicted causative SAVs to be significantly closer to the first rank than SpliceAI and Pangolin (Supplementary Figure S6). Furthermore, when clinicians manually evaluated up to the 5-40th ranks from the prediction results, more numbers of causative SAVs would be detected through PDIVAS, and the undiagnosed rates of the patients were retained to be 0.3-6.5% lower than SpliceAI and 4.9-16.3% lower than Pangolin (Fig. 5b). Even when only the top five were evaluated, 81.7% of the cases would be diagnosed. These results demonstrate that by introducing the diagnostic approach with PDIVAS, clinicians will be able to rapidly find the causative variants because PDIVAS allows them to evaluate a smaller
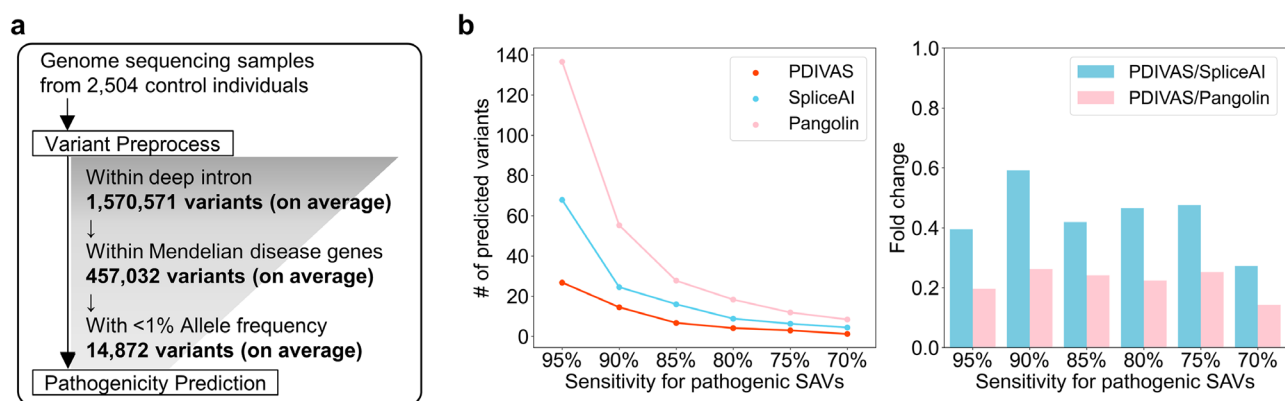

Fig. 4 PDIVAS minimizes candidates of causative deep-intronic variants from genome sequencing samples. **a** Workflow of variant filtering for candidates of pathogenic deep-intronic SAVs in 2,504 genome sequencing samples in the 1000 Genomes Project. Each number is shown as the mean value of 2,504 samples. **b** The median numbers of deep-intronic rare variants predicted at each threshold with sensitivity for pathogenic deep-intronic SAVs in the test dataset (Table 1). To compare the numbers between PDIVAS, SpliceAI, and Pangolin, fold changes are calculated. For SpliceAI prediction, the maximums of SpliceAI donor/acceptor delta gain scores are used. For Pangolin prediction, gain scores are used
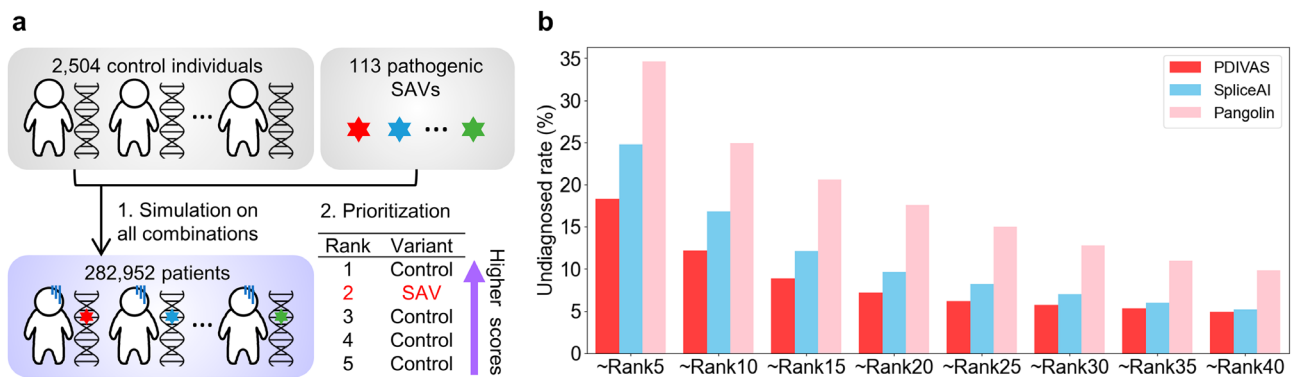
Fig. 5 PDIVAS prioritizes causative variants within simulated patient genome sequences. **a** Diagram indicating the method for simulating patient genome sequences and ranking their predicted variants. **b** Undiagnosed rates are calculated as the percentage of simulated patients whose causative variants are not prioritized within the designated ranks, compared to the total number of simulated patients
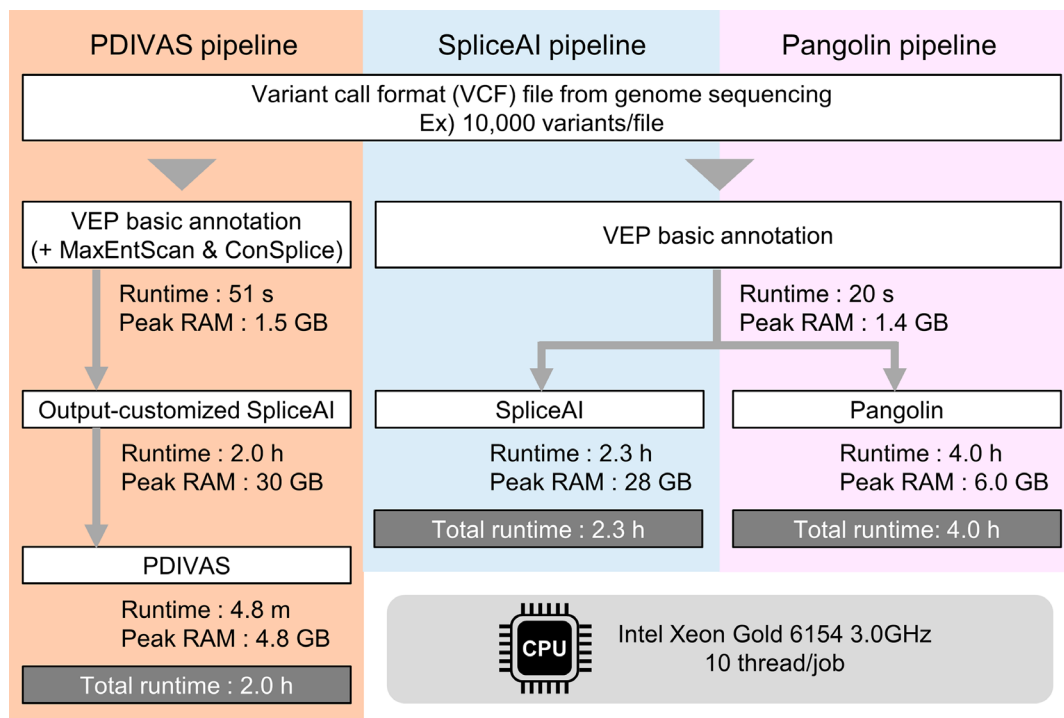

Fig. 6 Variant annotation speed and RAM benchmarks on variants from genome sequencing. Five VCF files were created by randomly extracting 10,000 intronic variants from the 1000 Genomes Project. These files were analyzed on the three pipelines, and the calculation time and RAM usage were recorded on each procedure. The average time and peak RAM were used as representative values. The VEP basic annotation includes the annotation of allele frequencies of the variants and the genes on which the variant is located. Raw scores were output in addition to the default delta scores via output-customized SpliceAI

number of candidates and identify the causative variants with minimal effort.

Accessibility and computing speed of PDIVAS

The command-line interface of PDIVAS is written in Python 3. The application can work to annotate variants in Variant Call Format (VCF) files annotated by VEP and output-customed SpliceAI, already. The output-customized SpliceAI is used to compute SpliceAI raw scores in addition to delta scores (Figs. 1g and 2a). Within the

PDIVAS algorithm, deep-intronic variants with all five features were found and the PDIVAS scores were calculated only for them. We compared the run times of PDIVAS, SpliceAI, and PDIVAS pipelines under 10 threads of CPUs. The average total run time was 2.0, 2.3, and 4.0 h, respectively (Fig. 6). Although the PDIVAS pipeline has one additional procedure compared to SpliceAI and Pangolin pipelines, the computation time on PDIVAS itself was much shorter than SpliceAI and Pangolin. Therefore, we verify that the additional procedure does not affect

Kurosawa *et al. BMC Genomics*   (2023) 24:601

Page 12 of 14

the analysis efficiency. However, we needed more RAMs for the output-customized SpliceAI than Pangolin (30 GB vs. 6.0 GB).

In addition, pre-calculated scores are also available in VCF files for deep-intronic rare SNVs and short indels within Mendelian disease genes. Variant annotation based on the pre-calculated files can be performed on basic variant annotation tools such as VEP, AnnoVar, snpEff, BCFtools, and Vcfanno [29, 37–39]. The source code and pre-calculated scores of PDIVAS were available at https://github.com/shiro-kur/PDIVAS.

Discussion

In this study, we have presented PDIVAS, an efficient algorithm to prioritize deep-intronic variants causing aberrant splicing in genome sequencing data. PDIVAS reaches an average precision of 0.92 and a maximum MCC of 0.88, which are the highest among the 6 existing predictors. When applying PDIVAS to patient genome sequence analysis, the highest sensitivity threshold of 0.082 (Table 1) is recommended for a higher diagnostic rate. Even with the high sensitivity, the number of pathogenic candidate variants is reduced to about 27 per individual, which is small enough for clinicians to manually evaluate their relevance to patient symptoms. Moreover, it ranks most of the causative variants within the top 5 in simulated patient genomes. These results indicate that incorporating PDIVAS into routine analysis pipelines would improve the efficiency of variant interpretation and contribute to increasing the detection rate of causative variants.

There are three major technical characteristics of PDIVAS. First, the PDIVAS training dataset includes only truly pathogenic SAVs that were checked by us and the curators of HGMD. With only the HGMD curation, the dataset also included variants that are expected to cause splicing, but the events were not experimentally validated although they were given the labels of "Splice" variants. Some actually might not cause aberrant splicing and are not the cause of the diseases. Therefore, we again checked all original reports of pathogenic deep-intronic SAV candidates and extracted experimentally validated ones. Further, we augmented the dataset with that curated by Keegan et al. independently of HGMD, where the experimental validation of splice alterations was also strictly checked.

Second, PDIVAS is trained not to classify SAVs and non-SAVs, but to distinguish pathogenic SAVs and benign variants. As some common variants and homozygous variants in general populations are also observed to induce exonization within deep introns, some SAVs do not cause deleterious effects on physiological functions [40]. This indicates that the task of classifying SAVs and non-SAVs is not enough for pathogenicity prediction.

Therefore, we incorporated the deleterious effect prediction by ConSplice, as well as splicing features. As shown in Fig. 3c, during PDIVAS prediction, ConSplice helps remove benign SAVs from the SplieAI-predicted SAVs. This prediction of deleterious effect is a technical advantage over SpliceAI and Pangolin, which are not trained on pathogenic splicing events.

Third, PDIVAS is modeled for the specific use of pathogenicity prediction on deep-intronic variants. We divided deep-intronic SAVs into those within the splicing motif regions and those outside of them and observed their SpliceAI-score distributions. Through detailed observation, we reached the conceptualization of SpliceAI raw scores and MaxEntScan for better sensitivity. To implement their combination, we modeled the random forest optimized to evaluate the pathogenicity of deep-intronic variants. These features were not incorporated into the previous predictor of ConSpliceML.

Determining causative variants is beneficial for patients because it could change their clinical management [41, 42]. The pathogenic pseudoexons created by deep-intronic SAVs can be pharmacologically targeted, as we previously demonstrated in two pathological models: the IVS4+866 C>T causative variant of *NEMO* of anhidrotic ectodermal dysplasia with immunodeficiency, and the c.3849+10kbC>T of *CFTR* of cystic fibrosis [43, 44]. Pathogenic pseudoexons often harbor suboptimal splicing sites and are prone to be regulated through alternative splicing factors, and serine/threonine-rich splicing factors (SRSFs) are identified as exon recognition facilitators in the above cases. Thus, the small-molecule inhibitor of CDC-like kinase (CLK) that activates SRSFs through phosphorylation of the RS domain [45], leads to the inhibition of pathogenic pseudoexon recognition and recovery from disease-associated phenotypes in cellular models [43, 44]. The pathogenicity interpretation of the deep-intronic variants with PDIVAS will shed light on previously overlooked pathogenic deep-intronic SAVs in rare genetic diseases, hopefully improving diagnostic yield and the possibility of clinical management.

Conclusions

Here, we developed a Pathogenicity predictor for Deep-Intronic Variants causing Aberrant Splicing (PDIVAS). The PDIVAS was trained to differentiate between pathogenic and benign deep-intronic variants. The predictive accuracy of PDIVAS was optimized not only by predicting splicing alterations with multiple splicing predictors but also by evaluating the deleterious effect of the predicted splice event with a human splicing constraint metric. By implementing PDIVAS into variant interpretation pipelines, a small number of pathogenic variant candidates were extracted. Particularly, the threshold of 0.082 is recommendable for the users to maximize the

diagnostic rates since the threshold has 95% sensitivity for previously reported pathogenic variants. Efficient variant interpretation by PDIVAS would contribute to resolving numerous genetically undiagnosed cases whose deep-intronic causative variants were previously overlooked. The source code to run PDIVAS and the precomputed PDIVAS scores for all rare deep-intronic SNVs, short insertion, and deletion within Mendelian disease genes are now available at https://github.com/shiro-kur/PDIVAS.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09645-2.

Additional File 1: Supplementary Figure S1-6.

Data Availability
The PDIVAS source code, command-line interface, and predictions for all rare deep-intronic SNVs, short insertion, and short deletion within genes of Mendelian disease are available at https://github.com/shiro-kur/PDIVAS. ConSplice scores and precomputed scores of ConSpliceML [20] are available at https://home.chpc.utah.edu/~u1138933/ConSplice/. Precomputed scores of CADD-Splice [25] are available at https://krishna.gs.washington.edu/download/CADD/v1.6/GRCh37/. The pathogenic splice-altering variants from HGMD [32] were downloaded from the HGMD website http://www.hgmd.cf.ac.uk/ under a HGMD commercial license. Due to the HGMD commercial licensing, we are not allowed to share these variants publicly. 1000 Genomes Project variants [14] are publicly available at http://hgdownload.cse.ucsc.edu/gbdb/hg19/1000Genomes/phase3/. gnomAD [18] variants are publicly available at https://ftp.ensembl.org/pub/data_files/homo_sapiens/GRCh37/variation_genotype/gnomad.genomes.r2.0.1.sites.noVEP.vcf.gz. Gene list from OMIM [30] is available to users from academic institutions and non-profit organizations at https://www.omim.org/downloads. Gene lists from CGD [31] are publicly available at https://research.nhgri.nih.gov/CGD/download/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
¹Department of Anatomy and Developmental Biology, Graduate School of Medicine, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan
²Faculty of Science and Engineering, Kindai University, 3-4-1 Kowakae, Higashi-osaka, Osaka 577-8502, Japan
³Medical Research Support Center, Graduate School of Medicine, Kyoto University, Yoshida- Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan
⁴Division of Cancer RNA Research, National Cancer Center Research Institute, Tokyo 104- 0045, Japan
⁵Department of Drug Discovery Medicine, Graduate School of Medicine, Kyoto University, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan
⁶Laboratory of Tumor Microenvironment and Immunity, Graduate School of Medicine, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan
⁷Center for Medical Genetics, Keio University School of Medicine, Tokyo 160-8582, Japan

References
1. Ankala A, da Silva C, Gualandi F, Ferlini A, Bean LJ, Collins C, Tanner AK, Hegde MR. A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield. Ann Neurol. 2015;77(2):206–14.
2. Taylor JC, Martin HC, Lise S, Broxholme J, Cazier JB, Rimmer A, Kanapin A, Lunter G, Fiddy S, Allan C, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet. 2015;47(7):717–26.
3. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T, et al. The genetic basis of mendelian phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet. 2015;97(2):199–215.
4. Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, Ward P, Braxton A, Wang M, Buhay C, et al. Molecular findings among patients referred for clinical whole-exome sequencing. JAMA. 2014;312(18):1870–9.
5. Vaz-Drago R, Custódio N, Carmo-Fonseca M. Deep intronic mutations and human disease. Hum Genet. 2017;136(9):1093–111.
6. Petersen USS, Doktor TK, Andresen BS. Pseudoexon activation in disease by non-splice site deep intronic sequence variation - wild type pseudoexons constitute high-risk sites in the human genome. Hum Mutat. 2022;43(2):103–27.
7. Keegan NP, Wilton SD, Fletcher S. Analysis of pathogenic Pseudoexons reveals novel mechanisms driving cryptic splicing. Front Genet. 2021;12:806946.
8. Okubo M, Noguchi S, Awaya T, Hosokawa M, Tsukui N, Ogawa M, Hayashi S, Komaki H, Mori-Yoshimura M, Oya Y et al. RNA-seq analysis, targeted long-read sequencing and in silico prediction to unravel pathogenic intronic events and complicated splicing abnormalities in dystrophinopathy. Hum Genet 2022.
9. Pros E, Gomez C, Martin T, Fabregas P, Serra E, Lazaro C. Nature and mRNA effect of 282 different NF1 point mutations: focus on splicing alterations. Hum Mutat. 2008;29(9):E173–193.
10. Sangermano R, Garanto A, Khan M, Runhart EH, Bauwens M, Bax NM, van den Born LI, Khan MI, Cornelis SS, Verheij J, et al. Deep-intronic ABCA4 variants explain missing heritability in Stargardt disease and allow correction of splice defects by antisense oligonucleotides. Genet Med. 2019;21(8):1751–60.
11. Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A, Haack TB, Graf E, Schwarzmayr T, Terrile C, et al. Genetic diagnosis of mendelian disorders via RNA sequencing. Nat Commun. 2017;8:15824.
12. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, Bolduc V, Waddell LB, Sandaradura SA, O'Grady GL et al. Improving genetic diagnosis in mendelian disease with transcriptome sequencing. Sci Transl Med 2017, 9(386).
13. Murdock DR, Dai H, Burrage LC, Rosenfeld JA, Ketkar S, Muller MF, Yepez VA, Gagneur J, Liu P, Chen S et al. Transcriptome-directed analysis for mendelian disease diagnosis overcomes limitations of conventional genomic testing. J Clin Invest 2021, 131(1).
14. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.

15. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. Genome Biol. 2016;17(1):122.
16. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766–73.
17. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE. HGVS recommendations for the description of sequence variants: 2016 update. Hum Mutat. 2016;37(6):564–9.
18. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
19. Jaganathan K, Panagiotopoulou SK, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535–48. e524.
20. Cormier MJ, Pedersen BS, Bayrak-Toydemir P, Quinlan AR. Combining genetic constraint with predictions of alternative splicing to prioritize deleterious splicing in rare disease studies. BMC Bioinformatics. 2022;23(1):482.
21. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004;11(2–3):377–94.
22. Vallee MP, Di Sera TL, Nix DA, Paquette AM, Parsons MT, Bell R, Hoffman A, Hogervorst FB, Goldgar DE, Spurdle AB, et al. Adding in Silico Assessment of potential splice aberration to the Integrated evaluation of BRCA gene unclassified variants. Hum Mutat. 2016;37(7):627–39.
23. Shamsani J, Kazakoff SH, Armean IM, McLaren W, Parsons MT, Thompson BA, O'Mara TA, Hunt SE, Waddell N, Spurdle AB. A plugin for the Ensembl variant effect predictor that uses MaxEntScan to predict variant spliceogenicity. Bioinformatics. 2019;35(13):2315–7.
24. Zeng T, Li YI. Predicting RNA splicing from DNA sequence using pangolin. Genome Biol. 2022;23(1):103.
25. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021;13(1):31.
26. Danis D, Jacobsen JOB, Carmody LC, Gargano MA, McMurry JA, Hegde A, Haendel MA, Valentini G, Smedley D, Robinson PN. Interpretable prioritization of splice variants in diagnostic next-generation sequencing. Am J Hum Genet. 2021;108(9):1564–77.
27. Leman R, Parfait B, Vidaud D, Girodon E, Pacot L, Le Gac G, Ka C, Ferec C, Fichou Y, Quesnelle C et al. SPiP: Splicing Prediction Pipeline, a machine learning tool for massive detection of exonic and intronic variant effects on mRNA splicing. Hum Mutat 2022.
28. Pedersen BS, Quinlan AR. cyvcf2: fast, flexible variant analysis with Python. Bioinformatics. 2017;33(12):1867–9.
29. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM et al. Twelve years of SAMtools and BCFtools. Gigascience 2021, 10(2).
30. Amberger JS, Hamosh A. Searching online mendelian inheritance in man (OMIM): a knowledgebase of human genes and genetic phenotypes. Curr Protoc Bioinformatics. 2017;58(1 2 1–1 2):12.
31. Solomon BD, Nguyen AD, Bear KA, Wolfsberg TG. Clinical genomic database. Proc Natl Acad Sci U S A. 2013;110(24):9851–5.
32. Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. The human gene mutation database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. Curr Protoc Bioinformatics 2012, Chap. 1:Unit1 13.
33. Zhou QM, Zhe L, Brooke RJ, Hudson MM, Yuan Y. A relationship between the incremental values of area under the ROC curve and of area under the precision-recall curve. Diagn Progn Res. 2021;5(1):13.
34. Moles-Fernández A, Domènech-Vivó J, Tenés A, Balmaña J, Diez O, Gutiérrez-Enríquez S. Role of Splicing Regulatory Elements and in Silico Tools usage in the identification of Deep Intronic Splicing Variants in Hereditary Breast/Ovarian Cancer genes. Cancers. 2021;13(13):3341.
35. Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. 2011;21(8):1360–74.
36. Cheng J, Nguyen TYD, Cygan KJ, Celik MH, Fairbrother WG, Avsec Z, Gagneur J. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. Genome Biol. 2019;20(1):48.
37. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
38. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92.
39. Pedersen BS, Layer RM, Quinlan AR. Vcfanno: fast, flexible annotation of genetic variants. Genome Biol. 2016;17(1):118.
40. Sakaguchi N, Suyama M. In silico identification of pseudo-exon activation events in personal genome and transcriptome data. RNA Biol. 2021;18(3):382–90.
41. Group NIS, Krantz ID, Medne L, Weatherly JM, Wild KT, Biswas S, Devkota B, Hartman T, Brunelli L, Fishler KP, et al. Effect of whole-genome sequencing on the Clinical Management of acutely ill infants with suspected genetic disease: a Randomized Clinical Trial. JAMA Pediatr. 2021;175(12):1218–26.
42. Stark Z, Tan TY, Chong B, Brett GR, Yap P, Walsh M, Yeung A, Peters H, Mordaunt D, Cowie S, et al. A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. Genet Med. 2016;18(11):1090–6.
43. Boisson B, Honda Y, Ajiro M, Bustamante J, Bendavid M, Gennery AR, Kawasaki Y, Ichishima J, Osawa M, Nihira H, et al. Rescue of recurrent deep intronic mutation underlying cell type-dependent quantitative NEMO deficiency. J Clin Invest. 2019;129(2):583–97.
44. Shibata S, Ajiro M, Hagiwara M. Mechanism-based Personalized Medicine for cystic fibrosis by suppressing Pseudo exon inclusion. Cell Chem Biol. 2020;27(12):1472–1482e1476.
45. Zhou Z, Fu XD. Regulation of splicing by SR proteins and SR protein-specific kinases. Chromosoma. 2013;122(3):191–207.

Publisher's Note