Genetics in Medicine Open (2024) 2, 101850

ELSEVIER

Genetics in Medicine OPEN
An Official Journal of the ACMG

www.journals.elsevier.com/genetics-in-medicine-open

ARTICLE

Combined bioinformatic and splicing analysis of likely benign intronic and synonymous variants reveals evidence for pathogenicity

Owen R. Hirschi[1,2], Stephanie A. Felker[3], Surya P. Rednam[1,2], Kelly L. Vallance[4], D. Williams Parsons[1,2], Angshumoy Roy[1], Gregory M. Cooper[3], Sharon E. Plon[1,2,*]

[1]Baylor College of Medicine, Houston, TX; [2]Texas Children's Cancer Center, Texas Children's Hospital, Houston, TX; [3]HudsonAlpha Institute for Biotechnology, Huntsville, AL; [4]Cook Children's Medical Center, Fort Worth, TX

ARTICLE INFO

Article history:
Received 16 November 2023
Received in revised form
7 May 2024
Accepted 8 May 2024
Available online 15 May 2024

Keywords:
Intronic
RNA
Splicing
Synonymous
Variant Classification

ABSTRACT

Purpose: Clinical variant analysis pipelines likely have poor sensitivity to the effects on splicing from variants beyond 10 to 20 bases of exon-intron boundaries. Here, we demonstrate the value of SpliceAI to inform curation of rare variants previously classified as benign/likely benign (B/LB) under current guidelines.

Methods: Exome sequencing data from 576 pediatric cancer patients enrolled in the Texas KidsCanSeq study were filtered for intronic or synonymous variants absent from population databases, predicted to alter splicing via SpliceAI (>0.20), and scored >10 by combined annotation-dependent depletion. Rare synonymous or intronic B/LB variants in 61 genes submitted to ClinVar were also evaluated and RNA further assessed in monocyte-derived messenger RNA and/or an in vitro splice reporter assay in HEK-293T cells.

Results: SpliceAI-supplemented analysis of the KidsCanSeq cohort revealed a *DICER1* intronic variant that resulted in missplicing in RNA from a proband with a personal and family history of pleuropulmonary blastoma but negative clinical exome and panel reports. Analysis of 34,188 B/LB ClinVar variants yielded 18 variants predicted to cause disrupted reading frames. Assessment of 8 variants (*DICER1* $n = 4$, *CDH1* $n = 2$, *PALB2* $n = 2$) by in vitro splicing assay demonstrated abnormal splice products (mean 66%; range 6% to 100%). When available, phenotypic information from submitting laboratories demonstrated *DICER1*-associated tumors in 2 families (1 variant) and breast cancer in 3 families (2 *PALB2* variants).

Conclusion: Incorporation of SpliceAI in variant curation pipelines may improve classification of B/LB intronic and synonymous variants and highlight putative pathogenic variants for functional assays and RNA analysis, thereby increasing diagnostic yield for rare diseases.

The Article Publishing Charge (APC) for this article was paid by Sharon E. Plon.

Owen R. Hirschi and Stephanie A. Felker contributed equally to this article.

[*]Correspondence and requests for materials should be addressed to Sharon E. Plon, Feigin Tower, Suite 1200.20, 1102 Bates Avenue, Houston, TX 77030. *Email address:* splon@bcm.edu

doi: https://doi.org/10.1016/j.gimo.2024.101850

Introduction

Identifying variants in clinically relevant genes that affect splicing is challenging yet essential for accurate genetic diagnosis for many diseases. Current variant classification guidelines from the American College of Medical Genetics and Genomics/Association for Molecular Pathology only assert that variants within 2 base pairs of a splice junction in a gene of interest are pathogenic (P) candidates if loss of function of that gene is a known disease mechanism.[1] Intronic variants farther away from these splice sites, however, may also affect splicing and be overlooked by these criteria. Such variants can result in the loss of canonical donor, acceptor, and branch point sites required for messenger RNA (mRNA) splicing and instead result in the gain of noncanonical splice sites leading to alternative products that alter gene function and potentially lead to loss of function. Variants predicted to be synonymous with regard to the protein sequence may also disrupt splicing; yet, these variants are similarly deprioritized in clinical pipelines and may be penalized by commonly used predictive metrics, such as combined annotation-dependent depletion (CADD), which incorporate coding consequence in its prediction of deleteriousness.[2]

Intronic or synonymous variants affecting splicing can be rescued in genomic analysis with the use of bioinformatic tools, such as SpliceAI, a deep neural network predictor of splice site activity in the pre-mRNA sequence.[3] The tool uses genomic variation as input and outputs the position of potential splice acceptor or donor site loss or gain; it also provides a "delta score," which is the maximum probability of splicing events affected by the variant within a user-determined window flanking the variant. SpliceAI is an improvement upon previous approaches, such as MaxEntScan, which detects splicing variants only if they disrupt canonical splicing motifs within 9 base pairs (bp) on the donor (5′) splice site, and 23 bp of the acceptor (3′) splice site and thus unable to analyze variants deeper into introns and exons.[4]

Here, we demonstrate the added value of SpliceAI to identify rare variants previously classified as likely benign (LB) in clinically significant cancer predisposition genes *DICER1* (HGNC:17098), *CDH1* (HGNC:1748), and *PALB2* (HGNC:26144). The identified variants were experimentally shown to result in aberrant splicing products leading to premature stop codons (PTC).

Materials and Methods

Texas KidsCanSeq (KCS) cohort

The Texas KCS study is a Clinical Sequencing Evidence-Generating (CSER) Consortium[5] study that recruited pediatric cancer probands under 18 years of age. The study was approved by Baylor College of Medicine institutional review board, which served as the central institutional review board for all 6 participating sites. Probands and participating parents submitted blood or saliva samples for parallel clinical germline hereditary cancer panel and exome sequencing,[6] with results reported back to the medical record. Consent included permission to perform subsequent research analyses with data shared with the CSER consortium.[7]

Cohort variant analysis and variant filtration

The germline exome variant call files (VCFs) from the Texas KCS pediatric cancer probands were analyzed for variants within a list of 181 cancer predisposition genes (Supplemental Table 1). Variants were filtered to retain those with a total read depth of greater than 10 reads and with FILTER = "PASS" by the xAtlas variant caller.[8] The resulting proband VCFs were then merged using bcftools (v1.13),[9] and the variants were annotated via SpliceAI (v1.3.1) (masked scoring option) to determine variant effects on splice site acceptors or donors within 50 bp of the variant.[3] We filtered for those with delta SpliceAI scores of over 0.2, which is the lowest threshold to predict variants affecting splicing.[3] The resulting variants were then annotated using Ensembl Variant Effect Predictor (v102).[10] We prioritized variants with a Genome Aggregation Database (gnomAD) v3.1.1[11] allele count of less than 20, a CADD (v1.6) score over,[10,12] SpliceAI gains or losses scored over 0.2, and without an existing P or likely pathogenic (LP) classification in ClinVar. Variants were then curated for those in genes associated with the respective proband cancer phenotype.

RNA Analysis

Viably frozen peripheral blood mononuclear cells (PBMCs) from cohort probands were analyzed. PBMCs were thawed in a mixture of Roswell Park Memorial Institute media and 10% fetal bovine serum (VWR) then allowed to incubate for 48 hours. Cells were incubated with or without 100 μg/ml emetine (Sigma-Aldrich), which inhibits translation and thereby prevents transcript degradation via nonsense mediated decay (NMD), for 6 hours. Total cellular RNA was extracted using Qiagen RNAeasy Micro kit (Qiagen). A 2-step reverse transcriptase polymerase chain reaction (RT-PCR) analysis was performed: random hexamer primed Superscript first-strand synthesis system (Invitrogen) using 30 ng of total RNA followed by PCR using 2 oligonucleotides (5′-TGACTTGCTATGTCGCCTTG-3′ and 5′-GGTCAGTTGCAGTTTCAGCA-3′) specific to *DICER1* (NM_177438.3) exon 5 and 6. Products were gel purified using QIAquick Gel and PCR cleanup kit (Qiagen) and validated via Sanger sequencing (Eurofins Genomics).

Analysis of variants submitted to the ClinVar database

Variants submitted to the ClinVar database (hosted by the National Center for Biotechnology Information)[13] within

DICER1 were acquired (October 2022). Intronic and synonymous variants with no predicted amino acid or termination change were annotated using SpliceAI (v1.3.1) and Variant Effect Predictor (v102) as detailed in Materials and Methods—Cohort Variant Analysis and Variant Filtration (see above). Variants were prioritized for further analysis using the following criteria: (1) designated benign (B) or LB in ClinVar, (2) absent from gnomAD v3.1.1,[11] and (3) both SpliceAI gains and losses scored over 0.2 and CADD (v1.6) score over.[10,12]

Variants submitted to ClinVar in *DICER1* and 60 other genes with Clinical Genome Resource (ClinGen) Variant Curation Expert Panel (VCEP)-approved rules were downloaded (April 2023). B/LB variants in all 61 genes were filtered and annotated using the same methodology as above. This methodology was similarly applied to variants of uncertain significance (VUS) occurring in *DICER1*. For each variant tested via the in vitro splicing assay, the ClinVar submitters were contacted for phenotype information, when available.

In vitro splicing assay

A splice reporter assay based on the vector system (pDESTSplice; AddGene #32484) was used following the protocol by Kishore, et al.[14] Desired sequences were either synthesized via gBlocks (IDT) or extracted from human genomic DNA (Promega). The gBlock or extracted DNA harbored the exon-intron-exon junctions of interest along with 215 base pairs of intronic sequence, when applicable, flanking both exons and attB1 sites, created via primers for extracted DNA (Supplemental Table 2). Three different versions of each gBlocks were made: a reference version matching GRCh38, 1 with the allele of interest, and 1 with a common nonreference allele with an allele count in gnomAD greater than 2 near to the variants of interest[11] (Supplemental Table 2). Each gBlock was then cloned into pDONR221 using Gateway cloning following the manufacturer's protocol (Invitrogen), verified by Sanger sequencing, and recombined into pDESTSplice. Extracted DNA was cloned into pDONR221 using Gateway cloning following the manufacturer's protocol and the sequence was then verified. For regions not amenable to gBlocks, Phusion site-directed mutagenesis was performed following the manufacturer's protocol (Thermo Scientific) using 5′-phosphorylated-primers (Supplemental Table 2) to generate the allele of interest or the common gnomAD allele in the pDONR221 vector containing reference sequence followed by sequence verification and recombination into pDESTSplice. Reporter clones were then isolated from bacteria using a QIAprep spin miniprep kit (Qiagen). Transfection of 800 ng of each plasmid into 100k HEK-293T cells was done in triplicate using Lipofectamine 3000, following manufacturer's protocol for 24-well plates (ThermoFisher). After 24-hour incubation, total cellular RNA was extracted using Qiagen RNAeasy Micro kit. RT-PCR using 300 ng of total cellular RNA was performed in 2 steps: random hexamer primed Superscript first-strand synthesis system (Invitrogen) followed by PCR using 2 oligonucleotides for rat insulin exons in pDESTSplice (5′-CCTGCTCATCCTCTGGGAGC-3′ and 5′- AGGTCTGAAGGTCACGGGCC-3′). Products were gel purified using QIAquick Gel and PCR cleanup kit (Qiagen) and analyzed on agarose gel electrophoresis and analyzed via Sanger sequencing (Azenta Life Sciences). Intensity of gel bands was quantified using ImageJ 1.53t[15] and normalized to a background control. Graphs, simple linear regression, and statistics were generated using GraphPad Prism (v10).

Results

Texas KCS analysis

As a part of the CSER Consortium, the Texas KCS study recruited pediatric cancer probands at 6 clinical sites across Texas between 2018 to 2021 with solid tumors, lymphomas, or histiocytic disorders. Clinical germline exome and targeted panel sequencing was performed on 576 probands and reported in the medical record. Further analysis of the exome VCF files as described in Materials and Methods resulted in 30 B/LB/VUS variants in cancer predisposition genes of interest with a SpliceAI gain or loss scored over 0.2, absence in gnomAD v3.1.1, and CADD > 10. In total, 3 heterozygous variants in genes consistent with the patient's tumor phenotype were identified. The first 2 variants were in *SUFU* (NM_016169.4:c.177C>T p.(Arg393Trp)) in a medulloblastoma proband and in *RB1* (NM_000321.3:c.1960+1G>A) in a retinoblastoma proband. The former was reported as a VUS and later determined to be generally inconsistent with proband's cancer subtype, and the latter variant had been previously reported as P on the clinical exome and panel reports. Therefore, neither of these variants were selected for further functional study.

The third variant, NM_177438.3:c.574-26A>G in *DICER1* (termed variant A) was from a patient with pleuropulmonary blastoma (PPB). The proband was diagnosed with PPB at 15 months and was expected to have *DICER1*-related tumor predisposition syndrome, given the parents also reported a family history of PPB in a paternal cousin from another country with unclear work-up. However, no *DICER1* variants (P or VUS) were on the clinical germline exome and germline panel reports. Variant A is intronic and 26 bp from the canonical splice acceptor at the 3′ end of the fifth intron of *DICER1*, outside the range of intronic variants typically evaluated by clinical platforms. Subsequent analysis of genomic DNA from either saliva (parents) or blood (proband) confirmed the presence of the intronic variant in the proband DNA and paternal transmission (Figure 1A). The splicing event predicted by SpliceAI is a loss of the canonical splice acceptor site and the creation of a splice

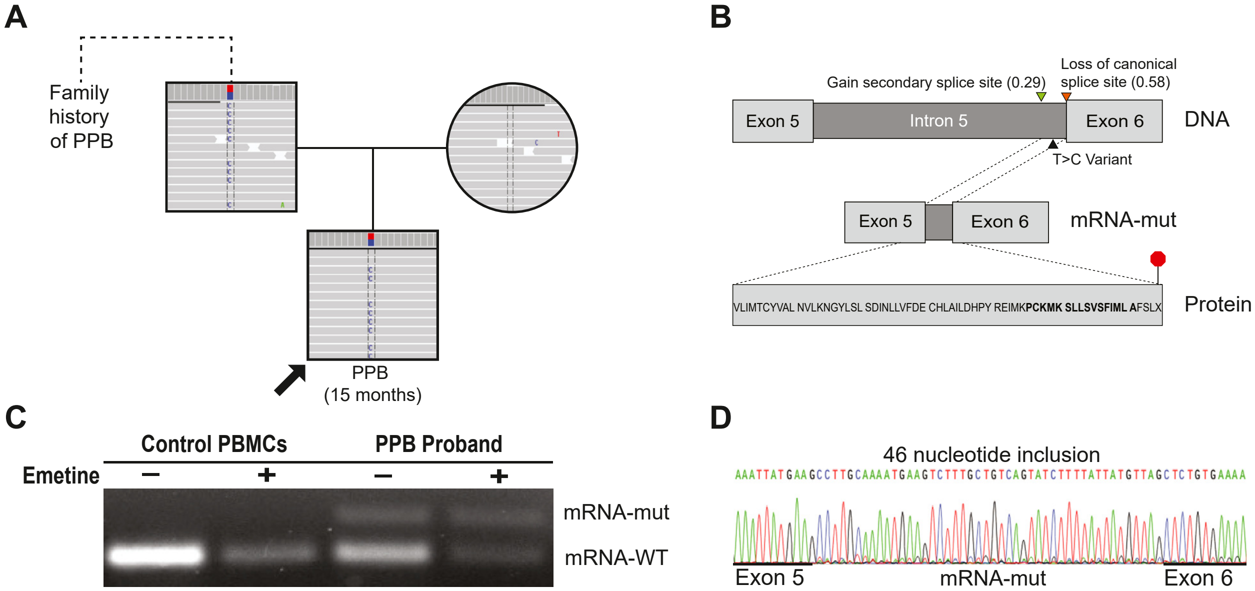
A
Family history of PPB
PPB (15 months)
B
Gain secondary splice site (0.29)
Loss of canonical splice site (0.58)
Exon 5
Intron 5
Exon 6
DNA
T>C Variant
Exon 5
Exon 6
mRNA-mut
VLIMTCYVAL NVLKNGYLSL SDINLLVFDE CHLAILDHPY REIMK**PCKMK SLLSVSFIML A**FSLX
Protein
C
Control PBMCs
PPB Proband
Emetine
–
+
–
+
mRNA-mut
mRNA-WT
D
46 nucleotide inclusion
AAATTATGAAGCCTTGCAAAATGAAGTCTTTGCTGTCAGTATCTTTTATTATGTTAGCTCTGTGAAAA
Exon 5
mRNA-mut
Exon 6

Figure 1 **Intronic *DICER1* variant identified in KidsCanSeq Cohort.** A. Pedigree showing the inheritance of said variant in the affected proband and how it fits with the family history of pleuropulmonary blastoma. B. Diagram outlining the predicted effect of the variant on splicing in the messenger RNA and how this change would affect translation, with the SpliceAI scores in parentheses and the amino acids in bold coming from the inserted nucleotides. C. Agarose gel image of the reverse transcriptase polymerase chain reaction between exon 5 and 6 of *DICER1* with and without emetine of the pleuropulmonary blastoma proband with the *DICER1* variant and a control patient without this variant. D. Sanger sequencing showing the exon 5-6 junction of the messenger RNA-mut confirming the 46-nt intron inclusion.

acceptor site 20 bp upstream of the intronic variant, resulting in a 46-bp insertion. Upon translation, this partial intronic retention would result in a PTC expected to trigger NMD of the *DICER1* mRNA (Figure 1B). RNA from viably frozen PBMCs from the PPB proband and an unrelated control proband without PPB were analyzed by RT-PCR for products spanning exons 5 to 6 (Figure 1C). Sanger analysis confirmed the addition of a 46 bp insertion (Figure 1D) in 36% of the RNA from the PPB proband, with 64% reference sequence, and none from the control proband. The intronic variant may result in abnormal splicing because of disruption of the normal splicing branch point. Because the proband's PBMCs are heterozygous for the mutant allele, these results suggest that the majority of transcripts produced from the mutant haplotype result in the frameshift event.

ClinVar variant analysis—*DICER1*

Given the finding of this potentially clinically relevant *DICER1* variant in the KCS cohort, we searched for other variants designated as B and/or LB when submitted to ClinVar with similar predicted splicing abnormalities. To this end, all 4227 *DICER1* variants in ClinVar were filtered using the methodology described in Materials and Methods—Analysis of Variants Submitted to the ClinVar Database (Supplemental Figure 1A). This resulted in detection of a single synonymous variant, *DICER1* NM_177438.3:c.5499G>A p.(?) (variant B) (Table 1, Figure 2A) that had been submitted to ClinVar once as LB. Inquiry of the submitting laboratory revealed that they had seen that variant in 2 unrelated probands: (1) an approximately 40 year old female with a history of recurrent thyroid cancer, parotid tail pleomorphic adenoma, and a parent with thyroid and colon cancer and (2) a female child with a history of multinodular goiter, macrocephaly, and learning disability and a family history of thyroid nodules, goiters, and thyroid cancer. Thus, both probands have phenotypes consistent with *DICER1*-related tumor predisposition syndrome, which encompasses PPB, multinodular goiter, thyroid tumors, and neurodevelopmental disorders.[16] Neither of these probands were found at that time to have any other P, LP, or VUS results in *DICER1* or other genes tested.

In vitro splice reporter assay

Biological samples were not available from either patient with the *DICER1* ClinVar variant (variant B). We thus decided to evaluate both variant A and B with a well-established in vitro splice reporter assay. The reporter vector contains integration of the *DICER1* exon-intron-exon sequence between 2 rat insulin exons that are driven by the Rous sarcoma virus long terminal repeat promoter. We designed fragments encompassing the reference allele for the relevant section of *DICER1*, the mutant alleles for both variant A and B and a common nearby variant identified in gnomAD as a control for each construct (Figure 3A, Supplemental Table 2). After transfection into HEK-293T cells, RNA products are quantified and sequenced to assess the effects of variation on splicing efficiency and accuracy. As shown in Figure 3B, the RNA produced for the reference vector and gnomAD-common variant

Table 1 Overview of 9 variants assessed with combined annotation-dependent depletion score, maximum SpliceAI delta score, and the relative position of predicted splice site loss or gain

Variant ID	Human Genome Variation Society (HGVS) Transcript	HGVS GRCh38 Coordinate	Source	Gene Symbol	Phenotype	CADD Score	Maximum SpliceAI Score	SpliceAI Base Change	Notes
Variant A	NM_177438.3: c.574-26A>G	NC_000014.9: g.95129658T>C	KidsCanSeq Cohort	*DICER1*	Pleuropulmonary Blastoma	20.8	0.58	20	Inherited from an unaffected father; affected paternal cousin
Variant B	NM_177438.3: c.5499G>A	NC_000014.9: g.95091231C>T	ClinVar: Invitae	*DICER1*	(1) Thyroid cancer and parotid tail pleomorphic adenoma; family history of colon and thyroid cancer (2) Multinodular goiter, macrocephaly, and learning disability; family history of thyroid nodules, goiters, thyroid cancer	18.62	0.69	31	Submitted to ClinVar twice for unrelated probands
Variant C	NM_177438.3: c.2026C>A	NC_000014.9: g.95113106G>T	ClinVar: Ambry	*DICER1*	Not provided	15.75	0.26	59	Part of a validation cohort
Variant D	NM_177438.3: c.2466T>A	NC_000014.9: g.95108064A>T	ClinVar: Ambry	*DICER1*	Not provided	14.61	0.51	31	Part of a validation cohort
Variant E	NM_177438.3: c.4077T>A	NC_000014.9: g.95099909A>T	ClinVar: Ambry	*DICER1*	Not provided	15.19	0.87	28	Part of a validation cohort
Variant F	NM_004360.5: c.2178G>T	NC_000016.10: g.68828187G>T	ClinVar: Ambry	*CDH1*	Not provided	13.39	0.40	34	Part of a validation cohort
Variant G	NM_004360.5: c.2187G>T	NC_000016.10: g.68828196G>T	ClinVar: Ambry	*CDH1*	Not provided	10.80	0.44	34	Part of a validation cohort
Variant H	NM_024675.4: c.2595A>T	NC_000016.10: g.23626389T>A	ClinVar: Invitae	*PALB2*	Female diagnosed with breast cancer with family history of breast cancer	13.44	0.74	25	Submitted to ClinVar once for 1 proband
Variant I	NM_024675.4: c.2601C>T	NC_000016.10: g.23626383G>A	ClinVar: Ambry	*PALB2*	(1) Mother and daughter pair, both diagnosed with early-onset breast cancer with family history significant for pancreatic and breast cancer (2) Female diagnosed with late onset breast cancer with family history significant for breast cancer	13.37	0.27	25	Submitted to ClinVar once for 2 unrelated families

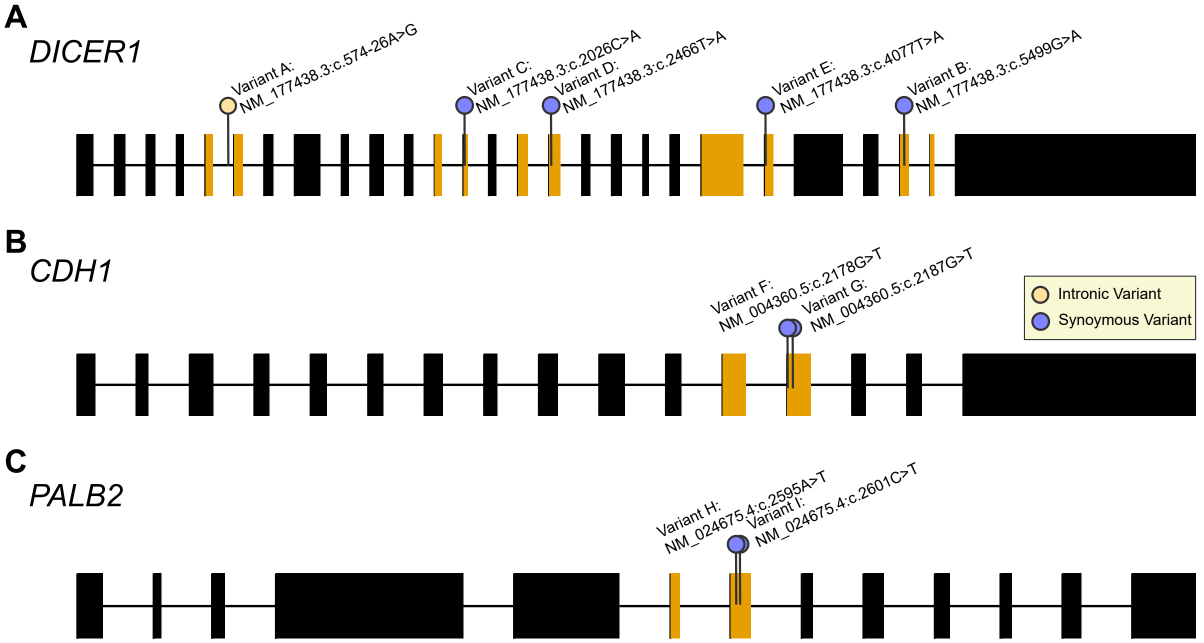
A
DICER1
Variant A: NM_177438.3:c.574-26A>G
Variant C: NM_177438.3:c.2026C>A
Variant D: NM_177438.3:c.2466T>A
Variant E: NM_177438.3:c.4077T>A
Variant B: NM_177438.3:c.5499G>A
B
CDH1
Variant F: NM_004360.5:c.2178G>T
Variant G: NM_004360.5:c.2187G>T
○ Intronic Variant
● Synoymous Variant
C
PALB2
Variant H: NM_024675.4:c.2595A>T
Variant I: NM_024675.4:c.2601C>T

Figure 2 **Location of assessed variants in cancer susceptibility genes.** (A) *DICER1*, (B) *CDH1*, and (C) *PALB2*. Colored exon pairs indicates the exon-exon junctions used in each in vitro splicing assay.

vector were consistent with the Matched Annotation from NCBI and EBI Select transcript. The variant A splicing assay recapitulated the abnormal splice products seen in the monocyte-derived RNA of the KCS cohort proband (Figure 1D) at 52% of splicing products. In addition, 2 secondary mRNAs were detected, at 30% and 18% abundance, which were not observed in the proband analysis. Both of these products also result in a frameshift (Figure 3C). Results from variant B analysis identified that 100% of the products harbored the SpliceAI-predicted 31-nucleotide exclusion of the 3′-end of exon 25 compared with normal products for the reference and gnomAD variant vector (Table 1, Figure 3A). The resulting 31-nucleotide frameshift mRNA would be expected to undergo NMD; even if truncated protein were produced, the critical *DICER1* double-stranded RNA binding domain would be disrupted. Of note, this variant has been subsequently upgraded to VUS by the submitting lab given the SpliceAI score.

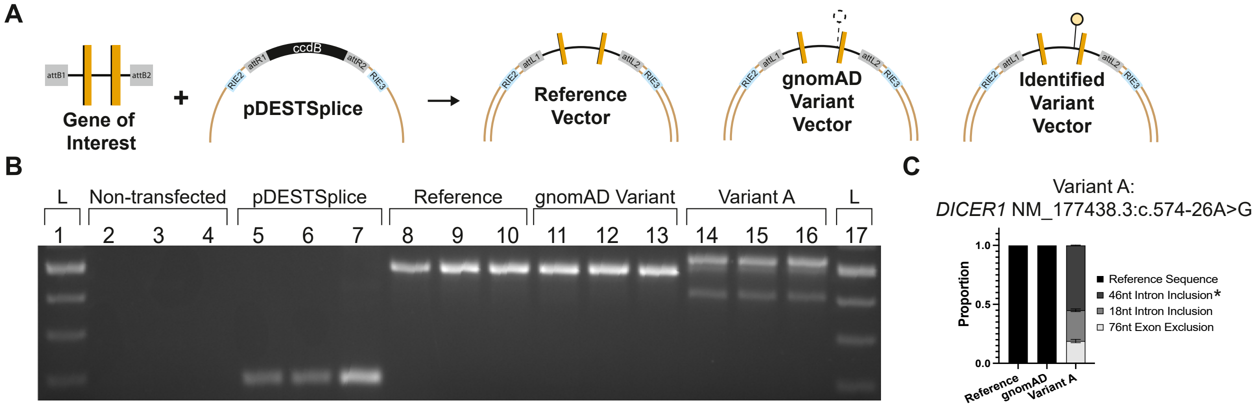

Figure 3 **Diagram and result of variant identified in KidsCanSeq Cohort using in vitro splicing assay.** A. Diagram of the vectors generated from the gene of interest using pDESTSplice, with each vector containing either the reference, a common Genome Aggregation Database variant, or the identified variant of interest. B. Results of the reverse transcriptase polymerase chain reaction of variant A, in which lanes 1 and 17 are DNA ladders (L), 2 to 4 are nontransfected controls, 5 to 7 are empty pDESTSplice, 8 to 10, is the reference vector, 11 to 13 is the vector containing the common Genome Aggregation Database variant, and 14 to 16 are the vector containing the identified variant of interest (variant A). C. Graph showing the proportion of products produced by reverse transcriptase polymerase chain reaction in the reference and variant A vector, identifying the proportion of products produced through quantification of luminescence, with the * denoting the messenger RNA-mut identified in the patient's RNA, as shown in Figure 1C.

ClinVar variant analysis—61 ClinGen expert panel genes

Given the results of the in vitro splice reporter assay, we expanded the ClinVar analysis to all 61 genes at that time with ClinGen VCEP specifications, which includes *DICER1*. We selected ClinGen VCEP genes, as ClinGen has recently published recommendations for incorporation of SpliceAI into variant classification specifications.[17] We extracted 34,188 B/LB variants across the 61 genes that were subsequently filtered using the methodology described in Materials and Methods—Analysis of Variants Submitted to the ClinVar Database (Supplemental Figure 1B). Among those, 23 variants had both SpliceAI gains and losses with scores over 0.2, CADD scores > 10, and were absent from gnomAD. Eighteen of these variants, affecting 12 genes, are predicted to lead to a frameshift when abnormal splicing occurs and should be considered for additional pathogenicity analysis (Supplemental Table 4). Four of these variants occurred in *DICER1*, including variant B and 3 submitted after our initial analysis (Table 1, Figure 2A).

To further assess evidence for missplicing, we performed in vitro splicing assays on the new *DICER1* variants, 2 variants in *CDH1* (Table 1, Figure 2B) and 2 in *PALB2* (Table 1, Figure 2C). These variants were selected because of their potential clinical importance, and for each of these variants, the full-length exon and intron sequences flanking the variant could be incorporated into the minigene splicing vector. As previously, a nearby common gnomAD variant was selected as a control (Supplemental Table 3). As shown in Figure 4B-H, the in vitro splice products for both the reference and gnomAD-common alleles were similar,

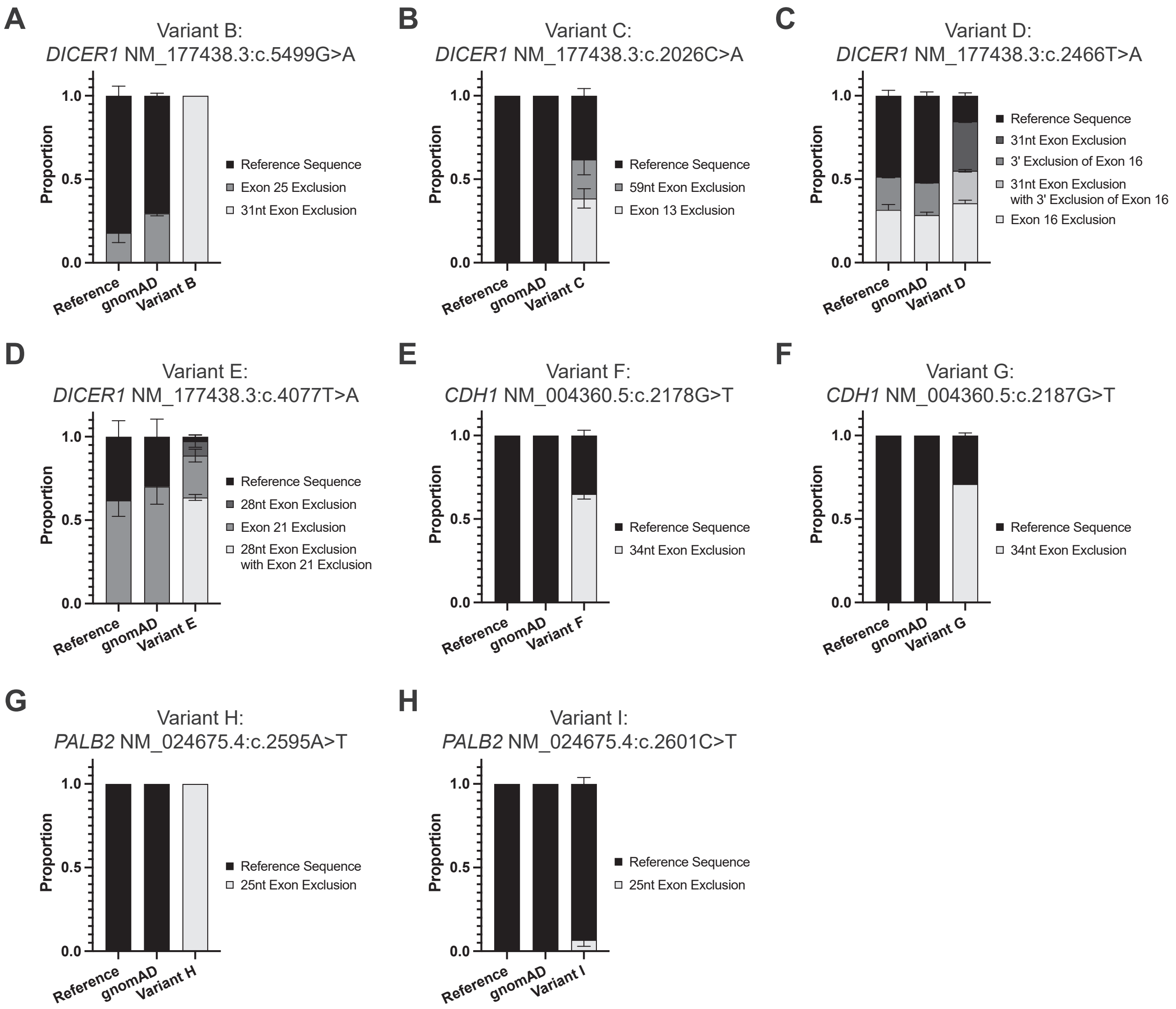

Figure 4 **Result of in vitro splicing assay on benign/likely benign variants extracted from ClinVar.** (A-H) shows results of reverse transcriptase polymerase chain reaction for variant B through variant I, displaying the proportion of products produced by reverse transcriptase polymerase chain reaction in the reference, nearby common variant, and variant vector. Photographs of the original gel electrophoresis are available in Supplemental Figure 4.

whereas the prioritized mutant alleles all revealed abnormal splice products not present in the controls to varying extent ranging from 6.8% (variant I) to 100% (variant H) (Supplemental Figure 2A). Of the abnormal products, the product predicted by SpliceAI was the largest component in all but 2 variants (variant D at 49.2% and variant I at 6.8%); however, the other abnormal products would also lead to frameshifts and subsequent PTCs (Figure 4B-H, Supplemental Figure 3).

We analyzed the SpliceAI scores for predicted loss and gain versus the proportion of mutant products and see some correlation consistent with the recent ClinGen guidance to use 0.2 as a minimum SpliceAI cutoff, which would need validation in a larger study. Comparing mean percentage of abnormal product produced with SpliceAI scores, the strongest correlation appears to be the SpliceAI donor or acceptor loss score, which could be studied further in a larger series (Supplemental Figure 2).

After completing these assays, we reached out to the submitting laboratories for any clinical information. Five of these variants, in *DICER1* and *CDH1* (variants C-G Table 1) were part of a validation cohort without any clinical information available from the submitting laboratory. Variant H in *PALB2* was found in a female proband with breast cancer diagnosed in their 50s with a family history of breast cancer from the maternal lineage. Variant I, also in *PALB2*, was found to occur in 3 patients from 2 families. In the first family, a mother and daughter pair carried the variant and were diagnosed with breast cancer in their 30s (ductal carcinoma in situ in the daughter) with family history significant for pancreatic cancer from the maternal lineage and breast cancer (including a male relative) from the paternal lineage. The third unrelated patient was a female diagnosed with late onset breast cancer in her 60s with family history significant for breast cancer from the paternal (also in a male relative) and maternal lineage. There were no other reportable variants identified for any of the 3 kindreds. The cancer diagnoses in these 3 kindreds are consistent with the phenotype of individuals with monoallelic *PALB2* loss-of-function variants[18,19] and monoallelic *BRCA1* or *BRCA2* loss-of-function variants.

Evaluation of *DICER1* VUS ClinVar submissions

In addition to the analysis of B/LB variants, we examined *DICER1* VUS that met our filtering criteria. There were 5364 VUS in *DICER1* extracted April 2023, which were subsequently filtered using the methodology described in Materials and Methods—Analysis of Variants Submitted to the ClinVar Database (Supplemental Figure 1C). Among those, 9 synonymous or intronic variants had both SpliceAI gains and losses with scores over 0.2, CADD scores > 10, and were absent from gnomAD. Since April 2023, 2 of these variants are now noted to be conflicting classifications of pathogenicity due to the submission of LP classification based on the implementation of RNA and phenotype analysis by the submitting laboratory (Supplemental Table 5). The other 7 variants are all predicted to affect splicing as noted in the ClinVar Submission Comments, but because of the lack of supporting RNA evidence, they were not able to be classified further (Supplemental Table 5).

Application of American College of Medical Genetics and Genomics classification guidelines

We applied the *DICER1* VCEP specifications (clinicalgenome.org, accessed 24 January 2024) for the 2 *DICER1* variants (A and B) in which clinical phenotype information was available. The following evidence codes were applied: 1. PS4_Supporting: the variant is absent from controls, and the gene is strongly associated with the proband disease, PPB; 2. PM2_Supporting: the variant is absent from the gnomAD v 3.1.1 database, and PVS1: Null variant in a gene in which loss of function is a known mechanism of disease. Incorporation of the in vitro splice reporter assays show that both variants result in a complete experimentally confirmed out-of-frame impact on splicing. These data applied to the PVS1 decision tree produced by Walker et al[17] and the *DICER1* VCEP result in full strength use of the code. In combination, these codes result in the designation of LP to both variants.

Discussion

We have identified potentially clinically significant variants originally classified as LB or in intronic regions not usually considered within reportable range in patients with phenotypes representative of *DICER1*-related tumor predisposition syndrome. The approach described here, which incorporates allelic frequency, SpliceAI prediction, and CADD score, identified these variants in our KCS exome data and the ClinVar database across many genes with minimal analytical labor and high specificity. Notably, we found the CADD score criteria to be useful in filtering the number of variants after SpliceAI and gnomAD were applied in the analysis of exome data from the KCS cohort; however, there was not a reduction in variant count when the CADD criteria was applied to those variants obtained from ClinVar. One explanation for this lack of contribution in filtering ClinVar variants is that variants submitted have undergone substantial filtering by the clinical labs and may already be enriched for variants over the CADD > 10 filtering threshold. In the analysis of KCS exome data, the CADD threshold excluded 12 low-scoring variants; thus, the criteria may be useful to other laboratories when analyzing other exome or unfiltered variant data sets. Incorporation of this information can result in clinically meaningful changes in variant classification. For example, variant A became a clear candidate for biological validation given that (1) there is a strong connection between the *DICER1* variant and the proband's and his relatives phenotype, (2) SpliceAI predicted both an acceptor loss and gain above 0.20, and (3) no other P variant in *DICER1* had been reported. RNA analysis and functional splicing assessment results in the designation of variant A as LP.

We focused on intronic and synonymous B and LB variants because they are not typically reported to the requesting physician and thus may not be further analyzed for pathogenicity. Interestingly, no B variants met the filtering criteria in either the exome VCF or ClinVar variant analysis. Additionally, analysis of ClinVar for non-synonymous B/LB variants returned no variants that met gnomAD allele frequency, CADD, and SpliceAI filtering criteria. We additionally recognize that by prioritizing a subset of variants in ClinVar primarily from hereditary cancer panel and KCS exome data, the scope of our investigation has been limited. These types of data sets exclude deep-intronic variants that result in aberrant splicing. For example, Fraire et al[20] used a custom capture panel including *DICER1* intronic sequences to identify 2 LP intronic variants (NM_177438.3:c.1509+16 p.(?) and c.1752+213 p.(?)) in *DICER1* potentially casual for *DICER1*-associated tumors. Additionally, it is probable that there are variants that result in in-frame alternative splicing events that affect the expression or functionality of the genes investigated in this study. Although these variants have the potential to be equally deleterious as the variants resulting in frameshift products investigated in the study, we prioritized the latter because they can be identified as deleterious without functional protein studies. This is a shortcoming of the in vitro splicing assay, in that functional effects are inferred from RNA products of a portion of the gene rather than the evaluation of the variant in the entire cDNA or demonstrated through RNA sequencing and the evaluation of protein translation.

Additionally, by prioritizing the B/LB variants, we exclude potentially putative VUS. This class of variants was selected because of the lack of reporting of B/LB variants (thus clinicians are not aware of the result) and deprioritization for RNA analysis. Of note, although only analyzed in *DICER1*, there are many other variants currently classified as VUS in ClinVar that meet these criteria and have the potential to be P or LP pending further functional assessment. All 9 *DICER1* VUS that met our criteria were submitted to ClinVar with specific language denoting their likely impact on pre-mRNA splicing and on average have higher SpliceAI scores in comparison with those 23 variants classified as B/LB (Supplemental Table 5). The B/LB variants identified in our analysis, including variant A, were on average 17 bases away from the exon boundary (Figure 5A), whereas those VUS in *DICER1* were on average 5 bases away the exon boundary and when annotated for molecular

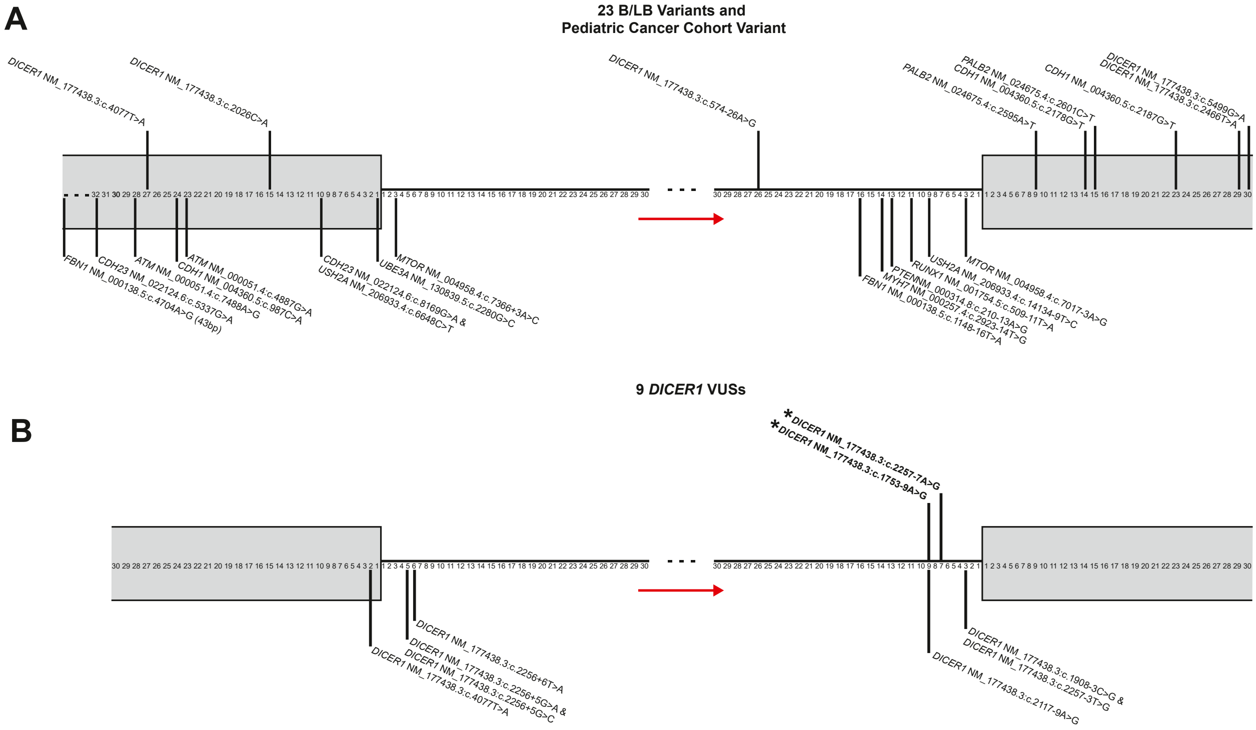
A
23 B/LB Variants and
Pediatric Cancer Cohort Variant
DICER1 NM_177438.3:c.4077T>A
DICER1 NM_177438.3:c.2026C>A
DICER1 NM_177438.3:c.574-26A>G
PALB2 NM_024675.4:c.2595A>T
PALB2 NM_024675.4:c.2601C>T
CDH1 NM_004360.5:c.2178G>T
CDH1 NM_004360.5:c.2187G>T
DICER1 NM_177438.3:c.5499G>A
DICER1 NM_177438.3:c.2466T>A
32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
FBN1 NM_000138.5:c.4704A>G (43bp)
CDH23 NM_022124.6:c.5337G>A
ATM NM_000051.4:c.7488A>G
CDH1 NM_004360.5:c.987C>A
ATM NM_000051.4:c.4887G>A
CDH23 NM_022124.6:c.8169G>A &
USH2A NM_206933.4:c.6648C>T
UBE3A NM_130839.5:c.2280G>C
MTOR NM_004958.4:c.7366+3A>C
MTOR NM_004958.4:c.7017-3A>G
USH2A NM_206933.4:c.14134-9T>C
RUNX1 NM_001754.5:c.509-11T>A
PTENNM_000314.8:c.210-13A>G
MYH7 NM_000257.4:c.2923-14T>G
FBN1 NM_000138.5:c.1148-16T>A
B
9 DICER1 VUSs
*DICER1 NM_177438.3:c.2257-7A>G
*DICER1 NM_177438.3:c.1753-9A>G
DICER1 NM_177438.3:c.2256+6T>A
DICER1 NM_177438.3:c.2256+5G>A &
DICER1 NM_177438.3:c.2256+5G>C
DICER1 NM_177438.3:c.4077T>A
DICER1 NM_177438.3:c.1908-3C>G &
DICER1 NM_177438.3:c.2257-3T>G
DICER1 NM_177438.3:c.2117-9A>G

Figure 5 Diagram of variants identified via analysis pipeline on conceptual exon-intron-exon map. A. Benign/likely benign variants and variant identified in KCS cohort mapped using base pairs (bp) from variant's respective exon boundary. Exons are denoted in gray boxes, whereas introns are horizontal black lines. Variants on top are those tested via in vitro splice assay and those on the bottom were not. The variant outside of 32 bp, have distance marked in parentheses next to Human Genome Variation Society. B. Variants of uncertain significance identified via analysis pipeline mapped using bps from respective exon boundary. Exons are denoted in gray boxes, whereas introns are horizontal black lines. Variants on top, in bold, and marked with an * and have been upgraded to conflicting classifications of pathogenicity with the addition of likely pathogenic classifications.

consequence were identified as splice/splice-region variants (Figure 5B, Supplemental Table 5). The subsequent LP submissions for 2 of these variants due to the inclusion of RNA and phenotype data highlight the prospect that other VUS identified via this pipeline are also potentially P when subsequent RNA or additional phenotype data become available. Potential misclassification of disease-relevant variants as B/LB may represent a systematic problem with the sole use of older algorithms designed to detect potential splicing defects. Variants that are farther from the splice site, either intronic or synonymous, are typically not subject to functional evaluation by clinical pipelines. As seen in variant B (2 families), variant H (1 family), and variant I (2 families), patients with no other reportable variants displayed phenotypes consistent with monoallelic loss-of-function variants in *DICER1* and *PALB2*, respectively. It is important to note that all 9 of the variants predicted by SpliceAI to cause missplicing result in the use of aberrant splice sites (at distances up to dozens of nucleotides away) that are not found in any known transcript and are outside the range of other splice prediction algorithms.[4,21] Biological validation through analysis of patient RNA and in vitro splicing assays confirm that (1) the limitations of only using a scoring algorithm that does not characterize all variants, (2) the scoring guidelines for SpliceAI from the original paper (“confidently predicted cryptic splice variants [score ≥ 0.5]”) are too conservative, and 3) variants scoring above 0.20 via SpliceAI should be considered potential mis-splicing variants consistent with the recently published ClinGen guidance.[3,17] Across the 9 variants assessed, we also identified a correlation between increased SpliceAI scores and abnormal products produced through the in vitro splicing assay, with the strongest correlation being seen with the increased SpliceAI donor or acceptor loss score (Supplemental Figure 2C). Although this is a limited analysis, splice variants do not always affect splicing in a binary manner, and it is possible that increasingly large SpliceAI scores may act as further evidence for classifying a variant’s effect on splicing (ie, although 0.2 is a useful threshold, larger values may provide even more evidence) as seen for missense predictors.[22] We have additionally demonstrated that as SpliceAI donor or acceptor loss scores decrease toward the 0.2 cutoff, the proportion of abnormal products detected by our splice reporter assays are increasingly variable. This phenomenon may not reflect the true biological consequences of these variants in patient transcriptomes.

The implementation of newer tools such as SpliceAI, demonstrates the ongoing improvement of algorithms and meta-predictors in describing potential effects of variants to shorten diagnostic odysseys for patients. Although large-scale, massively parallel splicing assays are beginning to be available to support these predictions,[23,24] the combined use of clinical RNA and DNA analysis can also reveal abnormal splicing products and aid in the rapid identification of clinically relevant variants.[25,26] Additionally, when genomic results are negative in patients with significant and specific phenotypic overlap with mendelian conditions, RNA sequencing may be an appropriate course of action to determine if aberrant RNA products are causative of the patient’s disease. In the case of the KCS proband, both panel and exome sequencing clinical pipelines did not return the intronic variant A. Our combined DNA and RNA analysis revealed the aberrant splicing product and was crucial to resolving the diagnostic odyssey of this proband. Laboratories should also consider re-evaluation of previously classified variants, including those classified as LB, with these newer algorithms to consider variant reclassification given the importance of diagnoses on proband outcomes and at-risk relatives obtaining prevention measures.

Data Availability

Genome sequencing and phenotype data for study participants that opted-in to data sharing are available for authorized access and hosted via dbGaP for the Texas KidsCanSeq study (phs002378.v1.p1). We have submitted the individual variants found to be of potential clinical relevance to ClinVar; accession IDs for each variant described can be found in Supplemental Table 4.

Acknowledgments

The authors would like to acknowledge Ambry Genetics and Invitae for providing the phenotype information included in the manuscript. The authors also acknowledge the families that agreed to participate in the KidsCanSeq study.

Funding

This study was supported by funding from the National Institutes of Health (NIH) U24 award grants U24 HG009649 to S.E.P. and U01 HG006485 to S.E.P. and D.W.P. and U01 HG007301 to G.M.C. O.R.H. was supported by the NIH F31 award CA265163. S.A.F. was supported by the NIH F31 award NIH F31 MH126628 followed by the NIH T32 award HG008961.

Author Information

Conceptualization: S.E.P., D.W.P., G.M.C.; Data Curation: S.P.R., K.L.V., S.A.F., O.R.H.; Formal Analysis: S.A.F., O.R.H., D.W.P., A.R.; Investigation: S.A.F., O.R.H.; Methodology: S.A.F., O.R.H.; Resources: S.E.P., D.W.P.; Visualization: O.R.H., S.A.F.; Writing-original draft: S.A.F., O.R.H.; Writing-review and editing: S.A.F., O.R.H., D.W.P., A.R., S.E.P., G.M.C.

Ethics Declaration

The KCS study was approved by the institutional review board of Baylor College of Medicine, which served as the central institutional review board for all participating sites (protocol number H-42376). All participants provided written informed consent.

Conflict of Interest

Dr Plon is a member of the scientific advisory panel of Baylor Genetics. All other authors declare no conflicts of interest.

Additional Information

The online version of this article (https://doi.org/10.1016/j.gimo.2024.101850) contains supplemental material, which is available to authorized users.

References

1. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424. http://doi.org/10.1038/gim.2015.30
2. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315. http://doi.org/10.1038/ng.2892
3. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting splicing from primary sequence with deep learning. *Cell*. 2019;176(3):535-548.e24. http://doi.org/10.1016/j.cell.2018.12.015
4. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*. 2004;11(2-3):377-394. http://doi.org/10.1089/1066527041410418
5. Amendola LM, Robinson JO, Hart R, et al. Why patients decline genomic sequencing studies: experiences from the CSER consortium. *J Genet Couns*. 2018;27(5):1220-1227. http://doi.org/10.1007/s10897-018-0243-7
6. Scollon S, Eldomery MK, Reuther J, et al. Clinical and molecular features of pediatric cancer patients with Lynch yndrome. *Pediatr Blood Cancer*. 2022;69(11):e29859. http://doi.org/10.1002/pbc.29859
7. Green RC, Goddard KAB, Jarvik GP, et al. Clinical sequencing exploratory research consortium: accelerating evidence-based practice of genomic medicine. *Am J Hum Genet*. 2016;98(6):1051-1066. http://doi.org/10.1016/j.ajhg.2016.04.011
8. Farek J, Hughes D, Salerno W, et al. xAtlas: scalable small variant calling across heterogeneous next-generation sequencing experiments. *Gigascience*. 2022;12:giac125. http://doi.org/10.1093/gigascience/giac125
9. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2):giab008. http://doi.org/10.1093/gigascience/giab008
10. McLaren W, Gil L, Hunt SE, et al. The Ensembl variant effect predictor. *Genome Biol*. 2016;17(1):122. http://doi.org/10.1186/s13059-016-0974-4
11. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. http://doi.org/10.1038/s41586-020-2308-7
12. 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. http://doi.org/10.1186/s13073-021-00835-9
13. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46(D1):D1062-D1067. http://doi.org/10.1093/nar/gkx1153
14. Kishore S, Khanna A, Stamm S. Rapid generation of splicing reporters with pSpliceExpress. *Gene*. 2008;427(1-2):104-110. http://doi.org/10.1016/j.gene.2008.09.021
15. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods*. 2012;9(7):671-675. http://doi.org/10.1038/nmeth.2089
16. Khan NE, Bauer AJ, Schultz KAP, et al. Quantification of thyroid cancer and multinodular goiter risk in the DICER1 syndrome: a family-based cohort study. *J Clin Endocrinol Metab*. 2017;102(5):1614-1622. http://doi.org/10.1210/jc.2016-2954
17. Walker LC, Hoya M, Wiggins GAR, et al. Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: recommendations from the ClinGen SVI Splicing Subgroup. *Am J Hum Genet*. 2023;110(7):1046-1067. http://doi.org/10.1016/j.ajhg.2023.06.002
18. Rahman N, Seal S, Thompson D, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet*. 2007;39(2):165-167. http://doi.org/10.1038/ng1959
19. Antoniou AC, Casadei S, Heikkinen T, et al. Breast-cancer risk in families with mutations in PALB2. *N Engl J Med*. 2014;371(6):497-506. http://doi.org/10.1056/NEJMoa1400382
20. Fraire CR, Mallinger PR, Hatton JN, et al. Intronic germline DICER1 variants in patients with Sertoli-Leydig cell tumor. *JCO Precis Oncol*. 2023;7:e2300189. http://doi.org/10.1200/PO.23.00189
21. Tang R, Prosser DO, Love DR. Evaluation of bioinformatic programmes for the analysis of variants within splice site consensus regions. *Adv Bioinformatics*. 2016;2016:5614058. http://doi.org/10.1155/2016/5614058
22. Pejaver V, Byrne AB, Feng BJ, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet*. 2022;109(12):2163-2177. http://doi.org/10.1016/j.ajhg.2022.10.013
23. Rhine CL, Neil C, Wang J, et al. Massively parallel reporter assays discover de novo exonic splicing mutants in paralogs of autism genes. *PLoS Genet*. 2022;18(1):e1009884. http://doi.org/10.1371/journal.pgen.1009884
24. Rong S, Neil CR, Welch A, et al. Large-scale functional screen identifies genetic variants with splicing effects in modern and archaic humans. *Proc Natl Acad Sci U S A*. 2023;120(21):e2218308120. http://doi.org/10.1073/pnas.2218308120
25. Murdock DR, Dai H, Burrage LC, et al. Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. *J Clin Invest*. 2021;131(1):e141500. http://doi.org/10.1172/JCI141500
26. Horton C, Cass A, Conner BR, et al. Mutational and splicing landscape in a cohort of 43,000 patients tested for hereditary cancer. *NPJ Genom Med*. 2022;7(1):49. http://doi.org/10.1038/s41525-022-00323-y