RESEARCH

Open Access

Benchmarking splice variant prediction algorithms using massively parallel splicing assays

Cathy Smith[1,2] and Jacob O. Kitzman[1,2]*

*Correspondence:
kitzmanj@umich.edu

[1] Department of Human
Genetics, University of Michigan
Medical School, Ann Arbor, MI
48109, USA
[2] Department of Computational
Medicine and Bioinformatics,
University of Michigan Medical
School, Ann Arbor, MI 48109,
USA

Abstract

Background: Variants that disrupt mRNA splicing account for a sizable fraction of the pathogenic burden in many genetic disorders, but identifying splice-disruptive variants (SDVs) beyond the essential splice site dinucleotides remains difficult. Computational predictors are often discordant, compounding the challenge of variant interpretation. Because they are primarily validated using clinical variant sets heavily biased to known canonical splice site mutations, it remains unclear how well their performance generalizes.

Results: We benchmark eight widely used splicing effect prediction algorithms, leveraging massively parallel splicing assays (MPSAs) as a source of experimentally determined ground-truth. MPSAs simultaneously assay many variants to nominate candidate SDVs. We compare experimentally measured splicing outcomes with bioinformatic predictions for 3,616 variants in five genes. Algorithms' concordance with MPSA measurements, and with each other, is lower for exonic than intronic variants, underscoring the difficulty of identifying missense or synonymous SDVs. Deep learning-based predictors trained on gene model annotations achieve the best overall performance at distinguishing disruptive and neutral variants, and controlling for overall call rate genome-wide, SpliceAI and Pangolin have superior sensitivity. Finally, our results highlight two practical considerations when scoring variants genome-wide: finding an optimal score cutoff, and the substantial variability introduced by differences in gene model annotation, and we suggest strategies for optimal splice effect prediction in the face of these issues.

Conclusion: SpliceAI and Pangolin show the best overall performance among predictors tested, however, improvements in splice effect prediction are still needed especially within exons.

Background

Splicing is the process by which introns are removed during mRNA maturation using sequence information encoded in the primary transcript. Sequence variants which disrupt splicing contribute to the allelic spectrum of many human genetic disorders, and it is estimated that overall, as many as 1 in 3 disease-associated single-nucleotide variants are splice-disruptive [1–7]. Splice-disruptive variants (SDVs) are most readily recognized at the essential splice site dinucleotides (GU/AG for U2-type introns) with many examples across Mendelian disorders [8–12]. SDVs can also occur at several so-called flanking noncanonical positions [13], and by some estimates, these outnumber essential splice mutations by several-fold [5, 14].

Variants beyond the splice-site motifs may be similarly disruptive but are more challenging to recognize [15]. Some of these disrupt splicing enhancers or silencers which are short motifs bound by splicing factors to stimulate or suppress nearby splice sites, to confer additional specificity, and to provide for regulated alternative splicing [16, 17]. These splicing regulatory motifs are widespread [18] and maintained by purifying selection [19], though they often feature partial redundancy and can tolerate some mutations. Some of the best-characterized cases come from genetic disorders, of which an archetypal example is spinal muscular atrophy caused by the loss of *SMN1*. Its nearly identical paralog *SMN2* cannot functionally complement its loss, due to a fixed difference that eliminates an exonic splice enhancer (ESE). The resulting exon 7 skipping [20, 21] can be targeted by antisense oligonucleotides [22] to boost SMN2 protein expression sufficiently to provide therapeutic benefit. Other cases include synonymous variants, which as a class may be overlooked, and can disrupt existing splice regulatory elements or introduce new ones, as in the case of *ATP6PA2*-associated X-linked parkinsonism [23]. Systematically delineating splice-regulatory elements and their cognate factors, and defining the grammar or "splicing code" through which they act to shape splicing patterns has been a long-standing challenge for molecular and computational biology [24].

RNA analysis from patient specimens can provide strong evidence for splice-disruptive variants, and its inclusion in clinical genetic testing can improve diagnostic yield [5, 25–28]. However, advance knowledge of the affected gene is necessary for targeted RT-PCR analysis, while RNA-seq-based tests are not yet widespread [29, 30], and both rely upon sufficient expression in the blood or other clinically accessible tissues for detection. Therefore, a need remains for reliable *in silico* prediction of SDVs during genetic testing, and a diverse array of algorithms have been developed to this end. For instance, S-Cap [31] and SQUIRLS [32] implement classifiers that use features, such as motif models of splice sites, kmer scores for splice regulatory elements, and evolutionary sequence conservation, and are trained on sets of benign and pathogenic clinical variants. Numerous recent algorithms use deep learning to predict splice site likelihoods directly from the primary sequence; SDVs can then be detected by comparing predictions for wild-type and mutant sequence. Rather than training with clinical variant sets, SpliceAI [33] and Pangolin [34] use gene model annotations to label each genomic position as true/false based on whether it appears as an acceptor or donor in a known transcript. SPANR [35] uses the primary sequence to predict percent spliced-in (PSI) measurements with training data provided by RNA-seq measurements. HAL [36] takes a distinct approach by training on a library of randomized sequences and their experimentally observed

splicing patterns, while MMSplice [37] combines the training data from HAL with features derived from primary sequence and additional modules trained on annotated splice sites and clinical variants. Finally, ConSpliceML [38] is a metaclassifier that combines SQUIRLS and SpliceAI scores with a population-based constraint metric measuring the regional depletion of predicted splice-disruptive variants among apparently healthy adults in population databases. While these are, for the most part, general-purpose short-variant predictors, other tools have been purpose-built for more specialized contexts, e.g., synonymous variants and deep intronic variants [39–41].

Given the proliferation of splicing predictors and their utility in variant interpretation, it is important to understand their performance characteristics. Previous comparisons have suggested that overall, SpliceAI represents the state-of-the art, with several other algorithms including MMSplice, SQUIRLS, and ConSpliceML showing competitive or in some cases better performance [32, 34, 38, 42–47]. However, benchmarking efforts to date primarily relied upon curated sets of clinical variants [32, 38, 42, 44–47], which are strongly enriched for canonical splice site mutations [35, 42, 47–50] likely due to the relative ease of their classification. This leaves open the question of how well these tools' performance may generalize, and whether certain tools may excel in particular contexts (e.g., for exonic cryptic splice activating mutations). A further challenge is that some of these tools' training data may partially overlap with benchmarking validation sets which risks circularity if overlapping variants are not carefully identified and removed.

Massively parallel splicing assays (MPSAs) provide an opportunity to benchmark splicing effect predictors entirely orthogonally to clinical and population variant sets. MPSAs measure thousands of variants' splicing effects in a pooled fashion: cells are transfected with a library of variants cloned into a minigene construct with deep RNA sequencing as a quantitative readout of variants' splicing outcomes. MPSAs come in several different flavors: broad MPSA screens assess many exons and measure one or a few variants' effects at each [6, 14, 51, 52], while saturation screens focus on individual exons [53–58] or motifs [36, 59] and measure the effects of every possible point variant within each target. Two broad MPSA datasets, Vex-seq [51] and MaPSy [6], were recently used to benchmark splicing effect predictors as part of the Critical Assessment of Genome Interpretation (CAGI) competition [60], and another, MFASS [14], has been used to validate a recent meta-predictor [43]. However, a limitation of benchmarking with broad MPSAs is that they may reflect an exon's overall properties while lacking the finer resolution to assess different variants within it. For instance, an algorithm could perform well by predicting SDVs within exons with weak splice sites, or with evolutionarily conserved sequence, while failing to distinguish between truly disruptive and neutral variants within each.

Here we leverage saturation MPSAs as a complementary, high-resolution source of benchmarking data to evaluate eight recent and widely used splice predictors. Algorithms using deep learning to model splicing impacts using extensive flanking sequence contexts, SpliceAI and Pangolin, consistently showed the highest agreement with measured splicing effects, while other tools performed well on specific exons or variant types. Even for the best performing tools, predictions were less concordant with measured effects for exonic variants versus intronic ones, indicating a key area of improvement for future algorithms.

Results

A validation set of variants and splice effects

We aggregated splicing effect measurements for 2230 variants from four massively parallel splice assay (MPSA) studies, focusing on saturation screens targeting all single nucleotide variants (SNVs) in and around selected exons [53, 54, 57, 58] (Fig. 1A). We also included 1386 variants in *BRCA1* from a recent saturation genome editing (SGE) study, in which mutations were introduced to the endogenous locus by CRISPR/Cas9-mediated genome editing, with splicing outcomes similarly measured by RNA sequencing [61]. Splice-disruptive variants (SDVs) and, conversely, neutral variants, were defined as specified by the respective studies, to account for their differences in gene target and methodology. For contrast with these saturation-scale datasets, we also prepared a more conventional, gene-focused benchmarking dataset by manually curating a set of 296 variants in the tumor suppressor gene *MLH1* from clinical variant databases and literature reports. In sum, this benchmarking dataset contained 3912 SNVs across 33 exons spanning six genes (Additional file 1: Fig. S1–S5; Additional file 2: Table S1).

As expected, MPSAs measured most of the possible single-nucleotide variants at each target (93.3% of SNVs) with relatively uniform coverage by exon/intron region (Fig. 1B). From the *BRCA1* SGE study, we retained only intronic or synonymous variants because missense variants' effects could be mediated via protein alteration, splicing impacts, or both. Targeted exons varied in their robustness to splicing disruption, from *POU1F1* exon 2 (10.2% SDV), to *MST1R* (also known as *RON*) exon 11 (68.4% SDV; Additional file 1: Fig. S6), reflecting both intrinsic differences between exons as well as different procedures for calling SDVs across MPSA studies. In contrast to the high coverage of the mutational space from MPSA and SGE datasets, reported clinical variants only sparsely covered the mutational space (1.6% of the possible SNVs in *MLH1* exons +/− 100 bp)
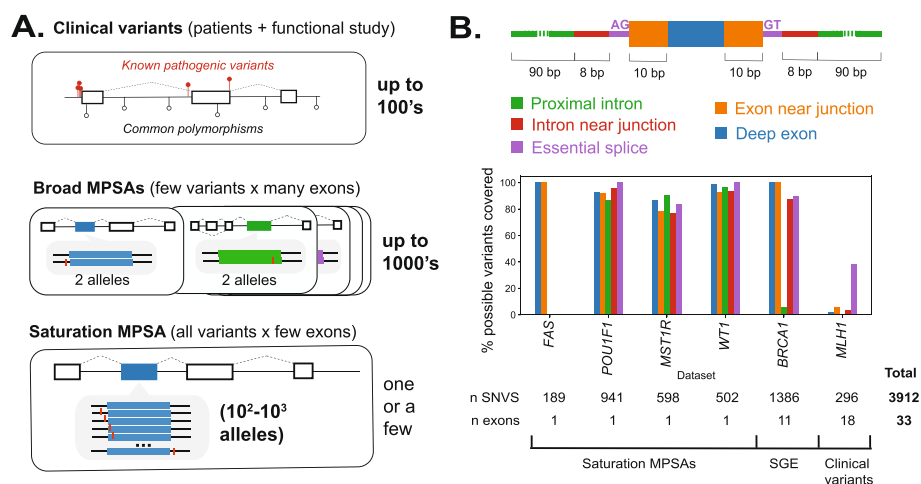

Fig. 1 Variants used for splice effect predictor benchmarking. **A** Validation sets can be drawn from pathogenic clinical variants and, conversely, common polymorphisms in frequently screened disease genes (top panel), broadly targeted massively parallel splice assays (MPSAs) interrogating a few variants across many exons (middle panel), and saturation MPSAs in which all possible variants are created for a few target exons (bottom panel). **B** Variant classes defined by exon/intron region and proximity to splice sites (upper), with the percent coverage of the possible SNVs within each variant class (denoted by color) for each dataset in the benchmark set (for *BRCA1*, missense, and stop-gain variants were excluded and not counted in the denominator)

and were heavily biased towards splice sites (59.5% of reported variants within $+/-$ 10 bp of a splice site; Additional file 1: Fig. S7). Larger clinical variant sets used to train classifiers showed a similar skew: 94.6% of the SQUIRLS training variants [32] and 88.9% of the pathogenic S-Cap training set [31] were within $+/-$ 10 bp of splice sites. Thus, MPSAs offer high coverage without the variant class biases present among clinical variant sets.

Comparing bioinformatic predictions with MPSA measured effects

We selected eight recent and widely used predictors to evaluate: HAL [36], S-Cap [31], MMSplice [37], SQUIRLS [32], SPANR [35], SpliceAI [33], Pangolin [34], and ConSpliceML [38]. Most variants (93.1%) were scored by all tools except HAL and S-Cap (which focus on only exonic variants and synonymous/proximal intronic variants, respectively). Algorithms' predictions were only modestly correlated with each other (median pairwise Pearson r between absolute values of tools' scores $= 0.58$, range: 0.04 to 0.97; Additional file 1: Fig. S8). One exception was Pangolin and SpliceAI which share similar model architectures and training sets and were almost perfectly correlated with each other ($r = 0.97$). These two were also strongly correlated with MMSplice ($r = 0.81$ and 0.80 respectively). The pattern of modest agreement across tools was not specific to exons and variants tested by MPSAs: we observed a similar degree of correlation between tools' scores across a background set of randomly sampled genomic SNVs (median pairwise $r = 0.60$; range: 0.07 to 0.93; Additional file 1: Fig. S8 and Methods). Concordance between algorithms was notably lower within exons, both in the MPSA benchmarking set variants (median pairwise $r = 0.43$) and random background set variants (median $r = 0.39$).

Agreement between predictors' scores and experimentally measured effects also varied widely by algorithm and MPSA dataset, but were similarly modest, with a median Pearson's r of 0.53 (range of -0.06 to 0.85; Additional file 1: Fig. S9). Even for a single algorithm and exon, substantial regional variability in concordance was evident (Fig. 2). For instance, at *POU1F1* exon 2, every tool other than S-Cap recapitulated the strong constraint observed by MPSA at the donor region. By contrast, at a putative exonic splicing silencer (ESS) near the alternative beta acceptor, algorithmic and measured effects were much less concordant, reflecting the difficulty of modeling variant effects outside canonical splicing motifs.

To systematically benchmark each predictor, we treated the splicing status from the experimental assays and curated clinical variant set as ground truth. We quantified the ability of each predictor to distinguish between the splice disruptive (n=1,060) and neutral (n=2852) variants in the benchmark set by taking the area under the precision-recall curve (prAUC) per classifier/gene (Fig. 3A). We next asked if classifiers' performance differed by variant type and location. Algorithms consistently performed better for intronic than for exonic variants (median prAUC for introns: 0.773; for exons: 0.419; Fig. 3B), despite a similar proportion of SDVs in exons and introns (28.4% and 25.9% SDV, respectively). This difference persisted even when removing canonical splice dinucleotide variants (Additional file 1: Fig. S10). More finely subdividing the benchmark variant set by regions (defined as in Fig. 1B) demonstrated that performance suffers farther from splice sites where the overall load of SDVs is lower (Additional file 1: Fig. S11). To summarize
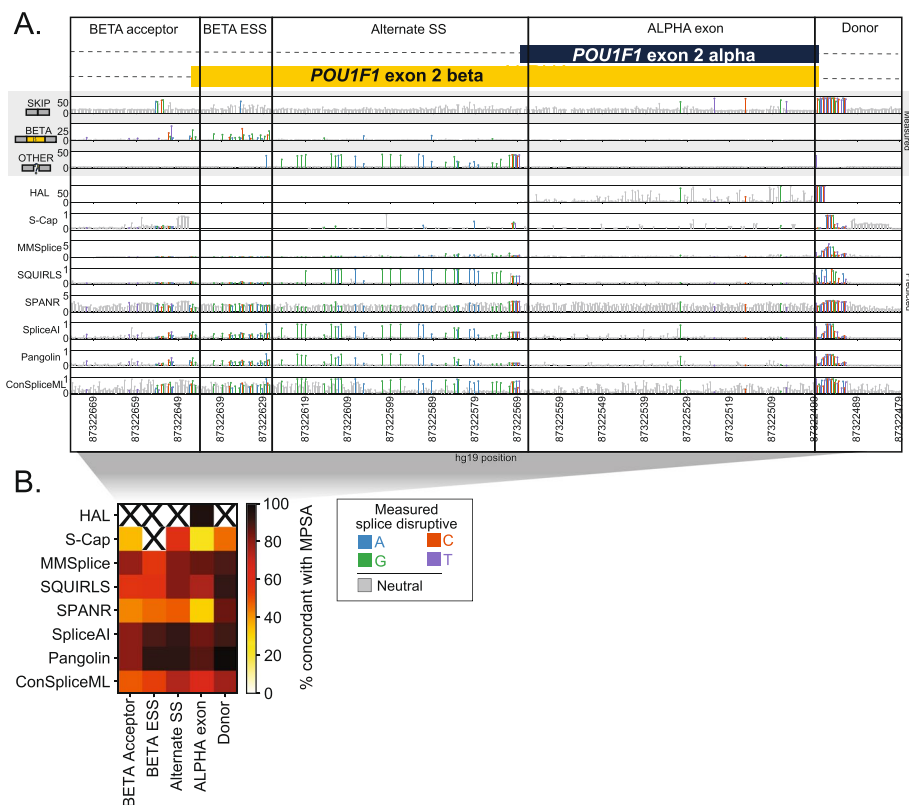
Fig. 2 Agreement between predictors and experiments varies by gene region. **A** Splicing effect tracks at alternate isoforms beta and alpha from a published MPSA [57] of *POU1F1* exon 2. Upper panel tracks (gray background) show MPSA-measured percent of exon skipping, beta exon inclusion, and other (non-alpha/beta/skip) isoform usage; lower tracks show scores from bioinformatic predictors by position. Each lollipop denotes one variant, shaded by effect in MPSA (gray: neutral, colors: SDVs, shaded by mutant base). The exon and flanking introns are split by region. **B** Heatmap showing concordance between each algorithm's binary classification of variants as SDV/neutral versus those of the MPSAs. For each algorithm, the score threshold that maximizes Youden's *J* across the full *POU1F1* dataset is used. Concordance is shown per algorithm (row) vs each region (column) at that score threshold. Regions with <10 scored variants are omitted ("X" symbols)

overall performance, we counted the number of instances in which each predictor either had the highest prAUC or was within the 95% confidence interval of the winning tool's prAUC (Fig. 3C). Every tool scored well for at least one dataset or variant class, but Pangolin and SpliceAI had the best performance most frequently (7 and 3 datasets/variant classes, respectively).

Benchmarking in the context of genome-wide prediction

In practice, a splicing effect predictor must sensitively identify SDVs while maintaining a low false positive rate across many thousands of variants identified in an individual genome. We therefore evaluated each tool's sensitivity for SDVs within our benchmark set, as a function of its genome-wide SDV call rate. We used a background set of 500,000 simulated SNVs drawn at random from in or near (+/− 100 bp) internal protein-coding exons (Additional file 1: Fig. S12; Additional file 2: Table S1). We scored these background SNVs with each tool and computed the fraction of the background set called as SDV as a function of the tool-specific score threshold. Although the true
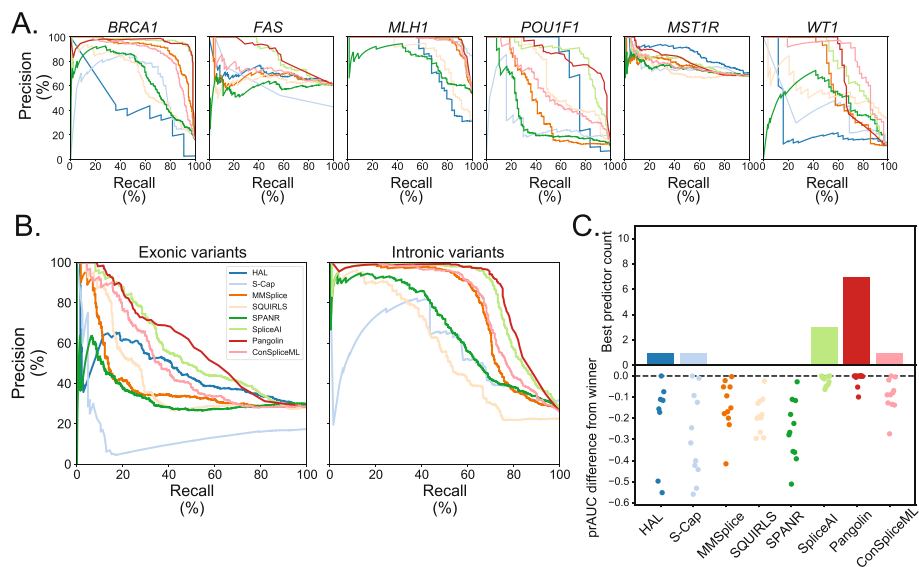
Fig. 3 Splice effect predictors' classification performance on benchmark variants. **A** Precision-recall curves showing algorithms' performance distinguishing SDVs and splicing-neutral variants in each dataset. **B** Precision-recall curves of tools' performance differentiating SDVs and splice neutral variants in exons (left) and introns (right). **C** Top panel: tally, for each algorithm, of the number of individual datasets and variant classes (defined as in Fig. 1B) for which that algorithm had the highest prAUC or was within the 95% confidence interval of the best performing tool. Bottom panel: signed difference between the best performing tool's prAUC and a given tool's prAUC; each dot corresponds to an individual dataset or variant class

splice-disruptive fraction of these background variants is unknown, we normalized algorithms to each other by taking, for each algorithm, the score threshold at which it called an equal fraction (e.g., 10%) of the genomic background set as SDV. We then computed the sensitivity across the benchmark-set SDVs using this score threshold and termed this the 'transcriptome-normalized sensitivity'. Taking SpliceAI as an example, at a threshold of deltaMax $\geq$0.06, 10% of the background set is called as SDV. Applying the same threshold (deltaMax$\geq$0.06) to *BRCA1* SGE variants in the benchmark set, SpliceAI reaches 98.2% sensitivity and 80.7% specificity (Fig. 4A).

We repeated this process, using for each algorithm the score threshold at which 10% of the background set was called as SDV, and applying this threshold to the benchmark set (Additional file 3: Table S2). Transcriptome-normalized sensitivity varied widely between algorithms, but SpliceAI, ConSpliceML, and Pangolin emerged as consistent leaders (median across datasets of 87.3%, 85.8%, and 79.9%, respectively). Mirroring the results seen entirely within the benchmarking variant set (Fig. 3), median transcriptome-normalized sensitivity was lower for exonic vs intronic variants for all tools examined by an average of 36.9%, and the same pattern remained after removing intronic variants at essential splice sites. These results were not specific to the transcriptome-wide threshold of 10%: the same three algorithms scored highly for thresholds at which 5% or 20% of the background set scored as SDV. Performance also varied by exon target (Fig. 4B); for example, many of the SDVs in *FAS* exon 6 and *MST1R* exon 11 were not detected by any algorithm at a threshold which would classify 10% of the background set as SDV. The effects measured by MPSAs in these specific exons may be particularly subtle, posing difficult targets for prediction, and suggesting that existing tools may need scoring
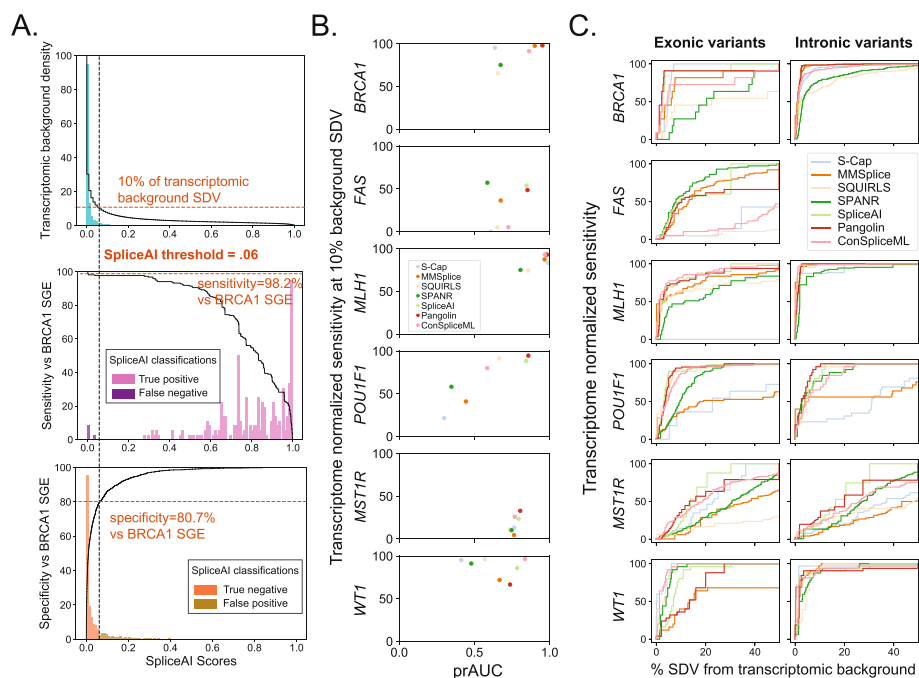
Fig. 4 Transcriptome-normalized sensitivity. **A** Example shown for SpliceAI. Upper panel shows SpliceAI scores for the 500,000 background set variants (teal histogram) and the cumulative fraction (black line) of variants above a given score threshold (10% of background set variants with deltaMax ≥0.06). Below, histograms of SpliceAI scores for *BRCA1* SGE benchmark variants, either SDVs (middle) or splicing-neutral variants (bottom), and the resulting transcriptome-normalized sensitivity and specificity at a deltaMax cutoff of 0.06. **B** Transcriptome-normalized sensitivity (at 10% background set SDV) versus within-benchmark variant set prAUC by benchmarked dataset. **C** Transcriptome-normalized sensitivity on benchmark exonic and intronic variants plotted as a function of the percent of the background variant set called SDV

thresholds tuned to specific exons or variant regions. Finally, to explore the tradeoff between SDV recall and overall call rate, we quantified the transcriptome-normalized sensitivity for SDVs in the benchmark set, as a function of percent of the background set called SDV and took the area under the resulting curve, analogous to the prAUC statistic. Again, performance was consistently lower within exons than introns, across algorithms and datasets (Fig. 4C).

Determining optimal score cutoffs

Integrating splice effect predictors into variant interpretation pipelines requires a pre-determined score threshold beyond which variants are deemed disruptive. We explored whether our benchmarking efforts could inform this by identifying the score threshold that maximized Youden's J statistic (J=sensitivity+specificity-1; Additional file 4: Table S3). For each algorithm, we first identified optimal score thresholds on each dataset individually to explore differences across genes and exons. For most tools we evaluated, ideal thresholds varied considerably across exons, regions, and variant classes, such that a threshold derived from one was suboptimal for others (Fig. 5). For some tools, including HAL and ConSpliceML, thresholds optimized on individual datasets spanned nearly the tools' entire range of scores, while for others such as SQUIRLS, SpliceAI, and Pangolin, the optimal thresholds were less variable. For the tools with consistently high
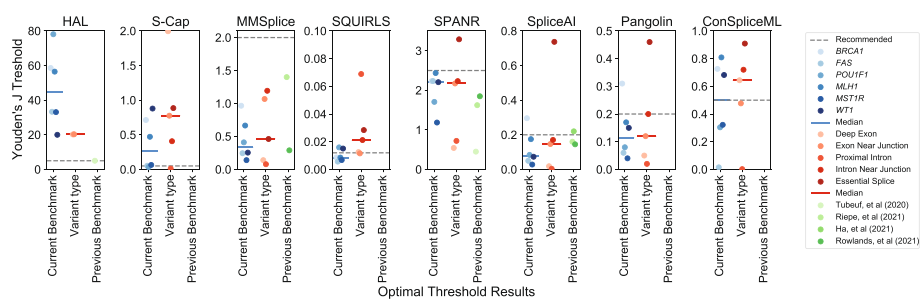
Fig. 5 Substantial variability of score thresholds by dataset and variant type. For each algorithm (panel), the score thresholds (*y*-axis) that maximized Youden's *J* is shown, across each benchmark variant dataset (blue points), by variant type (red points), and compared to previous reports (green points). Dashed gray line shows tool developers' recommended thresholds. Solid lines indicate medians

classification performance and transcriptome-normalized sensitivity — SpliceAI, Pangolin, and ConSpliceML (Figs. 3 and 4) — we found that the optimal thresholds were usually lower than the threshold recommended by the tools' authors, largely consistent with conclusions of other previous benchmarking efforts [44, 45, 47, 48]. Optimal thresholds also differed by variant class, suggesting that tuning cutoffs by variants' annotated effects, like those implemented in S-Cap, may offer some improvement for classification accuracy on variants genome-wide.

Variant effects at alternative splice sites

Alternative splicing can present challenges for variant effect prediction. Several of the tools tested here require gene model annotation, and their scores may be influenced by the inclusion or exclusion of nearby alternative isoforms in these annotations. In particular, SpliceAI and Pangolin use these annotations by default to apply a "mask" which suppresses scores from variants that either strengthen known splice sites or weaken unannotated splice sites, under the assumption that neither would be deleterious. Masking reduces the number of high-scoring variants genome-wide: among the background set, nearly one in every four splice-disruptive variants (deltaMax ≥0.2) identified without masking were suppressed (called neutral) by enabling it (Fig. 6A). Even without masking, annotation differences can introduce more subtle changes: among background set variants called by Pangolin as splice disruptive (absolute score ≥ 0.1), ~0.7% of variants called SDV with one annotation set (GENCODE) were called neutral by another (MANE Select), and vice versa (Fig. 6B). Therefore, while masking may be a necessary filter to reduce the number of variants for follow-up, it requires the provided annotation to be complete and further assumes there is no functional sensitivity to the relative balance among alternative splice forms.

We examined the effects of annotation choices and masking options at the two alternatively spliced exons in our benchmark variant set. In the first, *POU1F1*, two functionally distinct isoforms (alpha and beta) result from a pair of competing acceptors at exon 2. Alpha encodes a robust transactivator and normally accounts for ≥97% of *POU1F* expression in the human pituitary [62–65]. Beta exhibits dominant negative activity, and SDVs that increase its expression cause combined pituitary hormone deficiency [57, 66]. We focused on SpliceAI, in which the default annotation file includes only the alpha transcript. Predictions were broadly similar after updating annotations to include
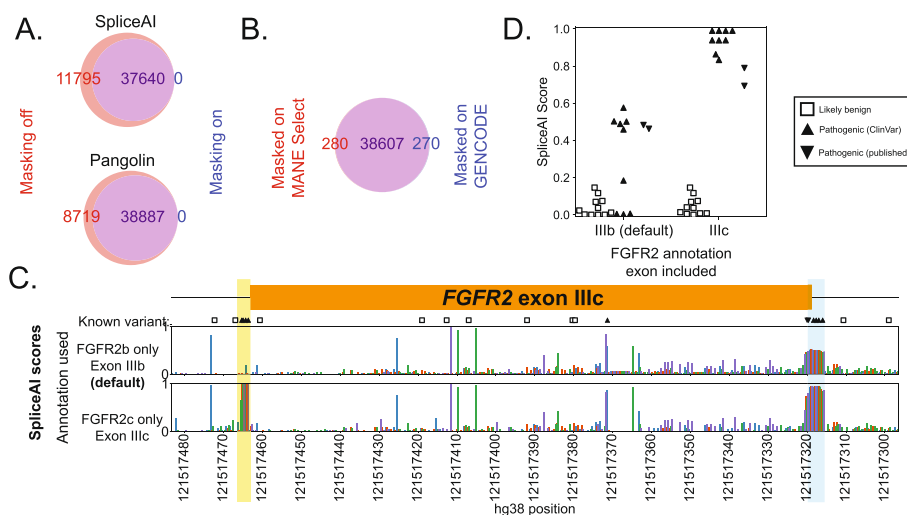
Fig. 6 Influence of masking and annotation choice. **(A)** Venn charts showing counts of background set variants (n=500,000) called as SDV only when masking is disabled (red), enabled (blue), or in both cases (overlap, purple) for SpliceAI and Pangolin. **(B)** Same for Pangolin, run with masking using two different annotation sets. **C** Tracks of SpliceAI scores (*y*-axis) for all SNVs in *FGFR2* exon IIIc showing either score using masking and default annotation (upper panel) or scores with masking and only the FGFR2c isoform (lower panel) vs hg38 position (*x*-axis), formatted as in Fig. 2. Symbols denote known benign or pathogenic variants from ClinVar or published reports. A cluster of pathogenic exon IIIc acceptor disrupting variants is missed when annotation does not include exon IIIc (yellow region), and exon IIIc donor disrupting variants have intermediate scores (blue region). **D** SpliceAI masked scores perfectly separate known pathogenic and benign variants when exon IIIc is included but not using default annotations

only the beta isoform or to include both: 13.8% (*n*=130/941) and 10.5% (*n*=99/941) of the variants, respectively, changed classifications compared to SpliceAI run with default annotations (each at an SDV cutoff of deltaMax≥0.08 which was optimal across that dataset; Additional file 1: Fig. S13). Among these were several pathogenic SDVs including c.143-5A>G which is associated with combined pituitary hormone deficiency (CPHD) [67], scored as highly disruptive by MPSA [57], and was validated in vivo by a mouse model [68]. With the default annotations (alpha isoform only) and when including both isoforms, SpliceAI scores c.143-5A>G as disruptive (deltaMax =0.21 and 0.16, respectively). However, when only the beta isoform is included, this variant is predicted neutral (deltaMax <0.001). A similar pattern emerged at a cluster of six pathogenic SDVs which disrupt a putative exonic splicing silencer which normally suppresses beta isoform expression [57]. Therefore, counterintuitively, pathogenic SDVs which act by increasing beta isoform usage go undetected when using annotation specific to that isoform.

The choice of canonical transcript may be less clear when alternative isoforms' expression is more evenly balanced, as in the case of *WT1,* a key kidney and urogenital transcription factor gene [69] covered by our benchmarking set. Exon 9 of *WT1* has two isoforms, KTS+ and KTS−, named for the additional three amino acids included when the downstream donor is used [70, 71]. In the healthy kidney, KTS+ and KTS− are expressed at a 2:1 ratio [72, 73]. Decreases in this ratio cause the rare glomerulopathy Frasier's syndrome [72–74], while increases are associated with differences in sexual development (DSD) [75]. We ran SpliceAI using annotations including KTS+ alone (its default), KTS− alone, and with both isoforms (Additional file 1: Fig. S14). A cluster of

variants, including one associated with DSD near the unannotated KTS− donor [75] (c.1437A>G), appear to weaken that donor but are masked because the KTS− donor is absent from the default annotations. Conversely, another variant (c.1447+3G>A) associated with DSD appears to increase the KTS+/KTS− ratio but is also masked because it strengthens the annotated KTS+ donor (deltaMax=0 with default annotation), and similarly scores as neutral when the annotation is updated to include both isoforms (deltaMax=0.02). That variant scores somewhat more highly (deltaMax=0.12) when only the KTS− annotation is used, but that in turn results in failure to capture several known Frasier's Syndrome pathogenic variants near the KTS+ donor [58, 72, 73, 76–79]. This case illustrates that predictors can fail even when all functionally relevant isoforms are included, because masking may suppress SDVs for which the pathogenic effects result from strengthening of annotated splice sites and disrupting the balance between alternative isoforms. This challenge was not specific to SpliceAI; for instance, Pangolin also showed poor recovery of KTS− SDVs (only 25% correctly predicted) due to a similar masking operation.

POU1F1 and *WT1* do not represent exceptional cases. Among RNA-seq junction usage data from the GTEx Consortium [65], we estimate 18.0% of all protein-coding genes (*n*=3571/19,817 genes) have at least one alternate splice site that is expressed and at least modestly used (≥20% PSI) in at least one tissue, yet is absent from SpliceAI default annotations (Additional file 1: Fig. S15). One of these is *FGFR2,* a tyrosine kinase gene with key roles in craniofacial development [80–82]. Mutually exclusive inclusion of its exons IIIb and IIIc results in two isoforms with different ligand specificities [80, 81, 83], and disruption of exon IIIc splicing causes Crouzon, Apert, and Pfeiffer Syndromes, which share overlapping features including craniosynostosis (premature cranial suture fusion) [84–87]. Pathogenic variants cluster near exon IIIc splice sites and at a synonymous site that activates cryptic donor usage within the exon [84, 86–97] (Fig. 6C). The default annotation excludes exon IIIc, causing all four pathogenic variants at its acceptor to be scored splice neutral, but when IIIc is included in the annotation, all four are predicted with high confidence (all ≥0.99; Fig. 6D, Additional file 2: Table S1). Disabling masking could capture cases such as this, but may not be a viable option in practice as reduces overall performance, and increases the number of high-scoring variants which must be reviewed [43].

Discussion

We evaluated the performance of eight splice effect predictors using a benchmark set of variants from saturation-level massively parallel splicing assays (MPSAs) across fifteen exons. By holding the sequence context constant for hundreds of variants per exon, these MPSAs afforded an opportunity to systematically evaluate how well each tool could distinguish individual variants' effects without confounding effects of differences in exons' overall characteristics. Compared to traditional validation sources such as clinical variant databases which are enriched for essential splice site mutations, these MPSA datasets had more uniform representation of variant types including those for which classification is currently challenging.

Across most exons tested, the deep learning-based tools Pangolin and SpliceAI had the best overall performance. These two were not uniformly superior, however, and other

tools excelled on certain datasets. ConSpliceML was comparably sensitive at identifying SDVs within the benchmarking set, while normalizing for genome-wide call rate, and MMSplice performed well for intronic SDVs. Even for the best performing tools, SDVs were more difficult to identify within exons compared to introns, highlighting an area for future improvement. These results are consistent with other recent splice predictor benchmarks using broad MPSAs and clinical variants, which also noted low concordance among tools [43, 45], particularly for exonic variants [47], and poorer classification performance in exons and with greater distance from splice sites [14, 32, 33, 48, 98]. As we found, SpliceAI was often but not always the top performer in these past comparisons [42–47, 98]. Together, our results suggest opportunities for metaclassifiers to better calibrate existing predictors and to leverage each within its strongest domain [38, 43].

A key issue this benchmarking study highlights is the challenge of selecting a scoring threshold for splicing predictors. This may reflect differences in exons' and genes' intrinsic vulnerability to SDVs, as a function of factors such as splice site strength [99] and wildtype exons' baseline inclusion rates [100]. For instance, most predictors fared poorly on *FAS* exon 6 and *MST1R* exon 11, both of which are intermediately included at baseline, and so may be more sensitive to splice disruption [100]. For moderately included exons such as these, more lenient thresholds may be required.

Tools' differing levels of concordance with MPSAs may reflect the fact that each was designed for a distinct prediction task. Consequently, their scores are not necessarily interchangeable. For instance, SpliceAI and Pangolin estimate the change in probability of a base being a splice acceptor/donor due to a nearby variant, while MMSplice, HAL, and SPANR each set out to predict changes in the inclusion of a given exon defined by a donor/acceptor pair. MPSAs identify the isoforms that are used and measure their steady-state abundances. These measures, though related, each have different meanings.

After identifying a splicing defect either bioinformatically or experimentally, a key downstream task — one not directly addressed by most of the tools here nor by the MPSAs – is to determine its pathogenicity. One reason is that the same change in exon inclusion may have drastically different effects from gene to gene: in some, an SDV that reduces properly spliced mRNA abundance by ~50% may be tolerated, whereas in more highly dosage-sensitive genes, an equivalently splice-disruptive variant would be highly deleterious. Supervised classifiers such as SPiCE provide one means to define thresholds at which changes in splice site strength may become pathogenic, at least for specific genes included in its training set [49]. Qualitative changes in splicing are also important to consider; SDVs that increase expression of an isoform with dominant negative effects may be deleterious even at a low level of expression as in the case of *POU1F1* exon 2 beta-promoting SDVs. As another example, while *DNM1* loss of function can result in developmental and epileptic encephalopathies, specific SDVs yield in-frame insertions which act in a dominant negative fashion and cause a particularly severe presentation [101]. Interpreting the results of bioinformatic splice effect predictions may therefore depend upon expert knowledge of the individual genes' dosage sensitivity and the functional properties of different isoforms, potentially limiting the utility of readily computed genome-wide scores. Methods such as ConSpliceML offer a means of inferring such thresholds by modeling the constraint against SDVs among healthy individuals on a regional (e.g., per exon) basis

[38] depending on individual genes' dosage sensitivity. Yet other tools take in account protein sequence and domain features to model such changes' impacts [41].

Our results also highlight the major influence of gene model annotation, a required input for many splice effect predictors. For two of the MPSA-tested exons in our benchmarking set (*POU1F1* and *WT1*), the inclusion of alternate splice forms in the annotation input altered SpliceAI's predictions across >10% of variants. Using RNA-seq data from GTEx, we conservatively project that this challenge may impact nearly one in every five genes in the human genome. Such annotation changes are inconvenient for end users and are not readily accommodated by some tools. Moreover, they may not be possible when the functionally relevant isoforms are not known in advance, though one recent tool, TRIFID, may help to define these on the basis of their appearance in proteomic data, evolutionary conservation, and other features [102]. Using the most comprehensive annotation set is not a universal fix, as illustrated by *POU1F1*, where it resulted in poorer concordance with MPSA measurements and lower specificity in recovering pathogenic variants. Some tools, including MMSplice and SQUIRLS, provide splicing effect predictions specific to all overlapping transcripts and could permit the investigation of isoform-specific effects at the cost of reviewing many additional variant scores. Recently, a new tool called SpliceMap provided a promising strategy for addressing the challenge of selecting relevant gene models [103]. SpliceMap leverages SpliceAI and MMSplice scores combined with a masking strategy that uses tissue-specific splice site usage annotation informed by GTEx. At equal recall to either of the predictors on which it relied for features (SpliceAI and MMSplice), SpliceMap substantially improved the precision for identifying rare variants in genes showing outlier patterns of splicing in RNA-seq.

Our conclusions are limited by the modest number of exons which have undergone saturation screens to date. Future efforts will benefit from additional splicing screens, as well as variant panels from large-scale clinical RNA-seq [25–27]. Another limitation is that the splicing assays we employed made certain tradeoffs in exchange for scale. Minigenes are often limited to short exons and partial flanking introns, though modestly longer sequence contexts can be accommodated [104]. We did not include deep intronic variants, which are an important source of pathogenic SDVs, and which can be interrogated by minigenes [105], but may be difficult to systematically screen given their length. Moreover, plasmid-based MPSAs cannot capture effects from transcription elongation rate or nucleosome positioning, each of which can influence splicing [106]. Some of these limitations can be addressed with emerging approaches for in situ genome engineering [61, 107]. However, as with MPSAs, those typically rely upon immortalized cancer cell lines, in which the splicing factor milieu may differ from that of the relevant tissues in vivo, and which do not necessarily match the cell types from which algorithms' training sets were derived. In the future, using more physiologically relevant cell lines [108] for these assays may allow for training and evaluation of tissue-specific splice effect predictors [34, 103, 109]. Despite these potential issues, minigene assays are often well correlated across cell lines [2, 6, 14, 57, 58] and have a sufficient track record of concordance with blood RNA analysis that they are often deemed acceptable as functional evidence during clinical variant interpretation [48, 105, 110]. Moreover, even when minigene assays misidentify a

variant's aberrant splicing outcome(s), they may still correctly flag the variant itself as splice disruptive [52, 111].

Conclusions

Here we have shown that saturation MPSAs provide an opportunity to critically evaluate the performance of computational splice effect predictors. Our results complement past benchmarking efforts using clinical variants and more broadly targeted MPSAs by testing algorithms' ability to distinguish many variants' effects within the context of several exons. This classification task resembles that faced by clinicians during variant interpretation, as there are many rare variants which do not impact splicing even in disease gene exons prone to splice disruption. We identified SpliceAI and Pangolin as the top-performing tools, but noted shortcomings including exonic variant performance, as well as practical challenges that end-users may encounter including selection of thresholds and the need for careful attention to gene model annotations. The continued growth of MPSA screens will present an opportunity to further improve splice effect predictors to aid interpretation of variants' splicing impacts.

Methods

Saturation mutagenesis datasets

Splice effect measurements were obtained for a total of 3616 variants in *POU1F1* (exon 2), *MST1R* (exon 11; also known as *RON*), *FAS* (exon 6), *WT1* (exon 9), and *BRCA1* (11 exons) from the respective studies' supplementary materials [53, 54, 57, 58, 61]. Variants were labeled as splice disruptive (SDV), intermediate, or neutral, according to the classification made by each study; intermediate effect variants (n=121) were removed. For *FAS*, SDVs were taken as those with a value of "S" (skipping) or "I" (inclusion) under the "Category" column of Supplementary Table 1; all others were taken as neutral. For *POU1F1* and *WT1*, the "splice disruptive" and 'intermediate' columns of each respective Supplementary Table 1 were taken as indicators, with all other variants considered neutral. *BRCA1* SGE measurements reflect both protein loss of function and mRNA effects, and we retained only synonymous and intronic variants so as to discard variants with effects that may be independent of splicing, and further restricted to internal coding exons. Among those, variants from Supplementary Table 1 with "LOF" in the "func. class" column and a 'mean.rna.score' value ≤ -2 were deemed SDV and the rest neutral. For *MST1R*, the minigene library spanned exons 10, 11, and 12, but as that assay did not measure the skipping of exons 10 or 12, we only included variants most likely to influence exon 11 inclusion (i.e., within exon 11 and proximal halves of its flanking introns). Intronic variants of more than 100 bp from either end of the selected exon were also discarded (n=93). We tested for significance as described by Braun et al. (heading "Significant mutation effects and synergistic interactions") using per-replicate variant scores from Supplementary Data 3 (sheets "HEK293T rep1," "HEK293T rep2," "HEK293T rep3"). Variants with $\geq$5% change in isoform frequency from WT minigenes at an FDR-corrected p<.05 were deemed as splice disruptive. The count of significant SDVs matched that listed in the original study (n=778 in total before filtering to exon 11). For MPSAs in *POU1F1*, *MST1R*, and *WT1* that reported effects upon usage of multiple isoforms, we used for each variant the isoform score that was most highly significantly

different than baseline (that is, maximum absolute *z*-score across isoforms per variant). For consistency in the direction of effect (higher measured scores denoting greater splice disruption), *BRCA1* RNA and function scores' signs were reversed. *FAS* enrichment scores were used without modification.

Manual curation of clinical *MLH1* variants

A literature search was conducted for variants assayed for splicing effects in the tumor suppressor gene *MLH1*, yielding 77 publications (publication years 1995–2021). We included only single-base substitutions and required each variant's splicing effects be supported either by RT-PCR and sequencing from patient blood-derived RNA or by mini-gene analysis. One exception is that essential splice site dinucleotide variants from Lynch syndrome patients were included without molecular evidence, as loss of the native site would be considered strong evidence of pathogenicity by ACMG guidelines [28]. Any splicing outcome other than full exon inclusion was considered pathogenic [112]. Eight variants had conflicting reports (i.e., both pathogenic and benign) and were resolved with a majority vote among the reporting publications with ties being considered pathogenic. The final dataset included 296 variants (mean: 1.8 references per variant), of which 160 were splice disruptive.

Random background variant set

We randomly drew 500,000 SNVs from within and near protein-coding genes to serve as a background set of exonic and proximal intronic variants with the potential to affect splicing. We used MANE Select canonical gene model annotations (version 1.0) [113], restricting to protein-coding transcripts with at least three coding exons. We discarded transcripts that had exons overlapping or within 100 bp of exon(s) of another transcript (on either strand), so that the variants' classification (intronic vs exonic, proximity to splice site) would not depend upon the choice of gene model; this left 79.6% of all MANE Select transcripts (n=14,618/17,631). SNVs were selected at random from internal coding exons (padded by $+/-$ 100 bp), and then these background SNVs were scored by splice effect predictors.

Scoring with eight splice effect predictors

Pangolin version 1.0.2 was run with the distance parameter set equal to the length of the scored exon (for *MLH1* and *BRCA1*: the longest exon for each gene; for background set SNVs: 300 bp) and both masking settings (off/on); reported Pangolin_max scores were used. SpliceAI version 1.3.1 was run via the python interface using a custom wrapper with the distance setting following the same process as for Pangolin to compute both masked and unmasked scores. For the transcriptomic background set, due to the high computational time to run SpliceAI, we downloaded version 1.3 precomputed scores from Illumina BaseSpace. SQUIRLS version 1.0.0 and MMSplice version 2.2.0 were both run on the command line with default settings to compute SQUIRLS score and delta logit PSI values respectively. HAL was run via the web interface (http://splicing.cs.washington.edu/SE) to predict exon-skipping effects. HAL requires a baseline percent spliced-in (PSI) value for the wildtype sequence (a parameter which has some predictive value on its own [43, 100]). For this parameter,

we used the following values: 90% for *MLH1*, 90% for *POU1F1*, 50% for *FAS*, 60% for *MST1R*, 80% for *BRCA1*, and 60% for *WT1,* based upon WT PSI values from the single exon MSPA original publications (rounded to the nearest 10%) and/or those exons' known splicing patterns [112]. For SPANR, S-Cap, and ConSpliceML, we obtained precomputed scores (SPIDEX zdelta PSI scores for SPANR; sens scores for S-Cap; ConSpliceML scores for ConSpliceML) from publicly accessible databases provided by the tools' authors. For essential splice site dinucleotide scores, S-Cap provides two models (dominant and recessive), and we selected the lowest score (most severe predicted impact). We then transformed the S-Cap scores (taking 1-x, for input scores x in [0,1]) to match the direction of effect for other tools with higher values indicating greater likelihood of splice disruption.

To ensure that differences between tools' predictions did not simply reflect their use of different gene model annotations, we selected the single MANE Select transcript model for each gene tested in the benchmarking set: ENST00000350375.7 (*POU1F1;* corresponding to the predominant isoform alpha), ENST00000452863.10 (*WT1* KTS+ isoform), ENST00000296474.8 (*MST1R*), ENST00000231790.8 (*MLH1*), ENST00000652046.1 (*FAS*), and ENST00000357654.9 (*BRCA1*). MMSplice, SQUIRLS, SpliceAI, and Pangolin all require an accompanying annotation file, and we provided each tool an identical annotation in which only the canonical transcript within the region of interest was included. Pre-computed ConSpliceML scores were selected by matching to the genomic position and relevant gene name. SQUIRLS' annotation file was not readily customizable, so we used the default hg19 ENSEMBL annotation files that it supplies. We verified that at or near the tested exons, there were no differences between the selected gene models provided to other tools and the gene models within SQUIRLS' annotations (ENST00000350375.2 for *POU1F1* alpha, ENST00000452863.3 for *WT1* KTS+, ENST00000296474.3 for *MST1R*, ENST00000231790.2 for *MLH1*, ENST00000355740.2 for *FAS*, ENST00000357654.3 for *BRCA1*). Substantive results did not change for the annotation dependent tools (MMSplice, Pangolin, SQUIRLS, and SpliceAI) when they were scored against annotation files including both alternative isoforms within *POU1F1* (beta isoform: ENST00000344265.8 for MMSplice, SpliceAI, and Pangolin and ENST00000344265.3 for SQUIRLS) and *WT1* (KTS− isoform: ENST00000332351.9 for MMSplice, SpliceAI, and Pangolin and ENST00000332351.3 for SQUIRLS). Within the testing using both alternative isoforms, for MMSplice and SQUIRLS, the most severe predicted impact from both isoforms was examined, and for Pangolin and SpliceAI, masking was based on both transcripts. For the transcriptomic background set, some variants either did not have a precomputed score for some tools, or the precomputed score entry mismatched the gene name or accession; and these variants were omitted (set to missing/NA) in the respective tools (SPANR: 1.6% of background variants excluded, SQUIRLS: 9.4%, SpliceAI: 4.1%). Pangolin and MMSplice each scored every background SDV. HAL only scores exonic variants, so all intronic variants were missing (56.5% of the background set), and S-Cap scores only some synonymous variants and variants within 50 bp of the splice sites, so 61.0% of background variants were missing.

Variant classes

To examine performance within different gene regions, we categorized variants as those in (i) essential splice site dinucleotides, (ii) intron near junction (3–10 bp from nearest exonic base), (iii) proximal intron (11–100 bp from nearest exonic base), (iv) exon near junction (<10 bp from nearest intronic base), and (v) deep exon ($\geq$ 10 bp from nearest intronic base). For variants in multiple transcripts, the category with the most severe consequence was chosen (order: essential splice > exon near junction > intron near junction > deep exon > proximal intron). We assessed the abundance of each variant class within previously curated clinical variant sets. For the S-Cap training set we combined the proportions of 5' core, 5' core extended, and 3' core variants listed from their clinically derived pathogenic set in Fig. 1C [31], and for SQUIRLS we tallied variant classes across training data without alterations [32].

Nominating annotation-sensitive alternatively spliced genes

To identify genes with alternative splice forms for which choice of annotation could influence splicing effect predictions, we obtained exon-exon junction read counts from GTEx portal (version 8). We restricted to protein-coding genes (n=19,817) and computed, for each of 54 tissues, the median junction read counts per million junction reads (junction CPM) across samples of that tissue for junctions that fell within coding portions of their respective genes. Junctions with a junction CPM $\geq$ 0.1 were considered expressed (n=14,831 genes had at least one expressed junction in at least one tissue). Next, we identified 12,124 genes where at least one expressed splice site was alternatively used in multiple junctions. Within each group of alternative junctions at a given splice site (e.g., two junctions corresponding to one donor paired with either of two different acceptors corresponding to skipping or inclusion of a cassette exon), we computed the fractional proportions of each junction's use and determined which alternate junctions were included in SpliceAI's default annotations. Fractional proportions were computed separately for each tissue. We deemed "moderately used unannotated" splice sites as any group of alternative junctions with at least one unannotated expressed junction which had $\geq$20% fractional usage in a given tissue. As a specific example of an exon sensitive to annotation selections, we scored variants in *FGFR2* against two alternative transcripts (ENST00000358487.10 for FGFR2c and ENST00000457416.2 for FGFR2b (default)) using our SpliceAI wrapper scripts [114] with masking turned on and a scoring distance equal to the length of the exon.

Statistical methods

Area under the curve metrics were calculated using the python package scikit-learn [115]. For Fig. 2, a single cutoff was selected for each tool that maximized Youden's J for identifying SDVs across the full *POU1F1* and *WT1* MPSA variant sets, respectively. Tools with fewer than ten scored variants in each pre-defined region were excluded. To compute transcriptome-normalized sensitivity for each algorithm, we first computed the score threshold $t(x)$ at which that algorithm called a given fraction x of the transcriptomic background set as disruptive (for all values of x in [0,1]). Transcriptome-normalized sensitivity was then the sensitivity for benchmark SDV

detection at this threshold: (# benchmark SDVs with score$\geq$t(x))/(# benchmark SDVs). Area under the curve was then taken for transcriptome-normalized sensitivity as a function of the background set fraction x and was computed using the scikit-learn auc function. HAL was excluded from transcriptome-normalized sensitivity analysis because its scores take on different ranges depending upon the level of exon inclusion in each wild-type exon (a required parameter for HAL).

To analyze the correlation between algorithms and MPSA measurements, the absolute value of each algorithms' scores, if signed, was taken. *FAS* was one exception to the rule: since *FAS* enrichment scores directly measured exon skipping (negative values) and exon inclusion (positive values), for signed scoring tools (HAL, MMSplice, SPANR, and Pangolin) we compared signed *FAS* scores with tools' signed values, and for the rest (SpliceAI, SQUIRLS, ConSpliceML, S-CAP), compared absolute values of the measured scores with the tools' scores. For classification performance analyses (prAUC, transcriptome-adjusted sensitivity), absolute values of tools' predicted scores were used.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-03144-z.

Additional file 1: Supplementary information. This additional file contains the supplementary figures and descriptions accompanying this manuscript.

Additional file 2: Table S1. Measured splicing scores and splicing outcomes as well as predicted scores for each of the eight algorithms evaluated for the six benchmarked datasets. Predicted scores from the eight bioinformatic tools for each of the 500,000 randomly selected exonic and near-exonic background variants. SpliceAI predictions against the FGFR2b and FGFR2c annotations for *FGFR2* exon IIIc.

Additional file 3: Table S2. Tool specific thresholds at which 5%, 10%, and 20% of the 500,000 background variants are predicted to be splice disruptive. Transcriptomic normalized sensitivity values for each algorithm and benchmarked dataset overall, within exonic variants, within intronic variants, and within intronic variants after removing variants at essential splice sites. Tables of sensitivity values are provided at transcriptome normalized cut offs of 5%, 10%, and 20% as well as the area under the curve (AUC) values.

Additional file 4: Table S3. Optimal score thresholds for each tool within the benchmarked datasets and across variant classes as defined in Figure 1B.

Additional file 5. Review history.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
J.O.K. serves as a scientific advisor to MyOme, Inc. The authors declare that there are no further competing interests.

References

1. Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc Natl Acad Sci U S A. 2011;108:11093–8.
2. Rhine CL, Cygan KJ, Soemedi R, Maguire S, Murray MF, Monaghan SF, et al. Hereditary cancer genes are highly susceptible to splicing mutations. PLoS Genet. 2018;14:e1007231.
3. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017;136:665–77.
4. Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. Loss of exon identity is a common mechanism of human inherited disease. Genome Res. 2011;21:1563–71.
5. Truty R, Ouyang K, Rojahn S, Garcia S, Colavin A, Hamlington B, et al. Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation. Am J Hum Genet. 2021;108:696–708.
6. Soemedi R, Cygan KJ, Rhine CL, Wang J, Bulacan C, Yang J, et al. Pathogenic variants that alter protein code often disrupt splicing. Nat Genet. 2017;49:848–55.
7. Scotti MM, Swanson MS. RNA mis-splicing in disease. Nat Rev Genet. 2016;17:19–32.
8. Takeshima Y, Yagi M, Okizuka Y, Awano H, Zhang Z, Yamauchi Y, et al. Mutation spectrum of the dystrophin gene in 442 Duchenne/Becker muscular dystrophy cases from one Japanese referral center. J Hum Genet. 2010;55:379–88.
9. Habara Y, Takeshima Y, Awano H, Okizuka Y, Zhang Z, Saiki K, et al. In vitro splicing analysis showed that availability of a cryptic splice site is not a determinant for alternative splicing patterns caused by +1G-->A mutations in introns of the dystrophin gene. J Med Genet. 2009;46:542–7.
10. Ramalho AS, Beck S, Penque D, Gonska T, Seydewitz HH, Mall M, et al. Transcript analysis of the cystic fibrosis splicing mutation 1525-1G>A shows use of multiple alternative splicing sites and suggests a putative role of exonic splicing enhancers. J Med Genet. 2003;40:e88.
11. Takahara K, Schwarze U, Imamura Y, Hoffman GG, Toriello H, Smith LT, et al. Order of intron removal influences multiple splice outcomes, including a two-exon skip, in a COL5A1 acceptor-site mutation that results in abnormal pro-alpha1(V) N-propeptides and Ehlers-Danlos syndrome type I. Am J Hum Genet. 2002;71:451–65.
12. Fang LJ, Simard MJ, Vidaud D, Assouline B, Lemieux B, Vidaud M, et al. A novel mutation in the neurofibromatosis type 1 (NF1) gene promotes skipping of two exons by preventing exon definition. J Mol Biol. 2001;307:1261–70.
13. Lord J, Gallone G, Short PJ, McRae JF, Ironfield H, Wynn EH, et al. Pathogenicity and selective constraint on variation near splice sites. Genome Res. 2019;29:159–70.
14. Cheung R, Insigne KD, Yao D, Burghard CP, Wang J, Hsiao YE, et al. A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. Mol Cell. 2019;73:183–194.e188.
15. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42:13534–44.
16. Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA. 2008;14:802–13.
17. Zhang XH, Chasin LA. Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev. 2004;18:1241–50.
18. Savisaar R, Hurst LD. Exonic splice regulation imposes strong selection at synonymous sites. Genome Res. 2018;28:1442–54.
19. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science. 2002;297:1007–13.
20. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet. 2002;3:285–98.
21. Kashima T, Manley JL. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. Nat Genet. 2003;34:460–3.
22. Hua Y, Vickers TA, Okunola HL, Bennett CF, Krainer AR. Antisense masking of an hnRNP A1/A2 intronic splicing silencer corrects SMN2 splicing in transgenic mice. Am J Hum Genet. 2008;82:834–48.
23. Korvatska O, Strand NS, Berndt JD, Strovas T, Chen DH, Leverenz JB, et al. Altered splicing of ATP6AP2 causes X-linked parkinsonism with spasticity (XPDS). Hum Mol Genet. 2013;22:3259–68.
24. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. Nature. 2010;465:53–9.
25. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9.
26. Landrith T, Li B, Cass AA, Conner BR, LaDuca H, McKenna DB, et al. Splicing profile by capture RNA-seq identifies pathogenic germline variants in tumor suppressor genes. NPJ Precis Oncol. 2020;4:4.
27. Horton C, Cass A, Conner BR, Hoang L, Zimmermann H, Abualkheir N, et al. Mutational and splicing landscape in a cohort of 43,000 patients tested for hereditary cancer. NPJ Genom Med. 2022;7:49.

28. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, 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;17:405–24.

29. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet. 2016;17:257–71.

30. Sheinson DM, Wong WB, Flores C, Ogale S, Gross CP. Association Between Medicare's National Coverage Determination and Utilization of Next-Generation Sequencing. JCO Oncol Pract. 2021;17:e1774–84.

31. Jagadeesh KA, Paggi JM, Ye JS, Stenson PD, Cooper DN, Bernstein JA, et al. S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. Nat Genet. 2019;51:755–63.

32. Danis D, Jacobsen JOB, Carmody LC, Gargano MA, McMurry JA, Hegde A, et al. Interpretable prioritization of splice variants in diagnostic next-generation sequencing. Am J Hum Genet. 2021;108:2205.

33. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176:535–548.e524.

34. Zeng T, Li YI. Predicting RNA splicing from DNA sequence using Pangolin. Genome Biol. 2022;23:103.

35. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347:1254806.

36. Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. Learning the sequence determinants of alternative splicing from millions of random sequences. Cell. 2015;163:698–711.

37. Cheng J, Nguyen TYD, Cygan KJ, Celik MH, Fairbrother WG, Avsec Z, et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. Genome Biol. 2019;20:48.

38. 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;23:482.

39. Gelfman S, Wang Q, McSweeney KM, Ren Z, La Carpia F, Halvorsen M, et al. Annotating pathogenic non-coding variants in genic regions. Nat Commun. 2017;8:236.

40. Shibata A, Okuno T, Rahman MA, Azuma Y, Takeda J, Masuda A, et al. IntSplice: prediction of the splicing consequences of intronic single-nucleotide variations in the human genome. J Hum Genet. 2016;61:633–40.

41. Zhang X, Li M, Lin H, Rao X, Feng W, Yang Y, et al. regSNPs-splicing: a tool for prioritizing synonymous single-nucleotide substitution. Hum Genet. 2017;136:1279–89.

42. Wai HA, Lord J, Lyon M, Gunning A, Kelly H, Cibin P, et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. Genet Med. 2020;

43. 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;13:31.

44. Riepe TV, Khan M, Roosing S, Cremers FPM, 't Hoen PA. Benchmarking deep learning splice prediction tools using functional splice assays. Hum Mutat. 2021;42:799–810.

45. Rowlands C, Thomas HB, Lord J, Wai HA, Arno G, Beaman G, et al. Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. Sci Rep. 2021;11:20607.

46. Strauch Y, Lord J, Niranjan M, Baralle D. CI-SpliceAI-Improving machine learning predictions of disease causing splicing variants using curated alternative splice sites. PLoS One. 2022;17:e0269159.

47. Ha C, Kim JW, Jang JH. Performance Evaluation of SpliceAI for the Prediction of Splicing of NF1 variants. Genes (Basel). 2021;12.

48. Tubeuf H, Charbonnier C, Soukarieh O, Blavier A, Lefebvre A, Dauchel H, et al. Large-scale comparative evaluation of user-friendly tools for predicting variant-induced alterations of splicing regulatory elements. Hum Mutat. 2020;41:1811–29.

49. Leman R, Gaildrat P, Le Gac G, Ka C, Fichou Y, Audrezet MP, et al. Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort. Nucleic Acids Res. 2018;46:7913–23.

50. Moles-Fernández A, Duran-Lozano L, Montalban G, Bonache S, López-Perolio I, Menéndez M, et al. Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations? Front Genet. 2018;9:366.

51. Adamson SI, Zhan L, Graveley BR. Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. Genome Biol. 2018;19:71.

52. Rhine CL, Neil C, Wang J, Maguire S, Buerer L, Salomon M, et al. Massively parallel reporter assays discover de novo exonic splicing mutants in paralogs of Autism genes. PLoS Genet. 2022;18:e1009884.

53. Braun S, Enculescu M, Setty ST, Cortés-López M, de Almeida BP, Sutandy FXR, et al. Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. Nat Commun. 2018;9:3315.

54. Julien P, Miñana B, Baeza-Centurion P, Valcárcel J, Lehner B. The complete local genotype-phenotype landscape for the alternative splicing of a human exon. Nat Commun. 2016;7:11558.

55. Baeza-Centurion P, Miñana B, Schmiedel JM, Valcárcel J, Lehner B. Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. Cell. 2019;176:549–563.e523.

56. Ke S, Anquetil V, Zamalloa JR, Maity A, Yang A, Arias MA, et al. Saturation mutagenesis reveals manifold determinants of exon definition. Genome Res. 2018;28:11–24.

57. Gergics P, Smith C, Bando H, Jorge AAL, Rockstroh-Lippold D, Vishnopolska SA, et al. High-throughput splicing assays identify missense and silent splice-disruptive POU1F1 variants underlying pituitary hormone deficiency. Am J Hum Genet. 2021;108:1526–39.

58. Smith CS, Burugula BB, Dunn I, Aradhya S, Kitzman JO, Yee JL. High-Throughput Splicing Assays Identify Known and Novel WT1 Exon 9 Variants in Nephrotic Syndrome. Kidney International Reports; 2023.

59. Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, et al. Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. 2011;21:1360–74.

60. Mount SM, Avsec Ž, Carmel L, Casadio R, Çelik MH, Chen K, et al. Assessing predictions of the impact of variants on splicing in CAGI5. Hum Mutat. 2019;40:1215–24.

61. Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature. 2018;562:217–22.
62. Wallis M. Evolution of the POU1F1 transcription factor in mammals: Rapid change of the alternatively-spliced beta-domain. Gen Comp Endocrinol. 2018;260:100–6.
63. Konzak KE, Moore DD. Functional isoforms of Pit-1 generated by alternative messenger RNA splicing. Mol Endocrinol. 1992;6:241–7.
64. Gordon DF, Haugen BR, Sarapura VD, Nelson AR, Wood WM, Ridgway EC. Analysis of Pit-1 in regulating mouse TSH beta promoter activity in thyrotropes. Mol Cell Endocrinol. 1993;96:75–84.
65. Consortium GT. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
66. Jonsen MD, Duval DL, Gutierrez-Hartmann A. The 26-amino acid beta-motif of the Pit-1beta transcription factor is a dominant and independent repressor domain. Mol Endocrinol. 2009;23:1371–84.
67. Takagi M, Kamasaki H, Yagi H, Fukuzawa R, Narumi S, Hasegawa T. A novel heterozygous intronic mutation in POU1F1 is associated with combined pituitary hormone deficiency. Endocr J. 2017;64:229–34.
68. Akiba K, Hasegawa Y, Katoh-Fukui Y, Terao M, Takada S, Hasegawa T, et al. POU1F1/Pou1f1 c.143-83A > G Variant Disrupts the Branch Site in Pre-mRNA and Leads to Dwarfism. Endocrinology. 2022;164
69. Hastie ND. Wilms' tumour 1 (WT1) in development, homeostasis and disease. Development. 2017;144:2862–72.
70. Mrowka C, Schedl A. Wilms' tumor suppressor gene WT1: from structure to renal pathophysiologic features. J Am Soc Nephrol. 2000;11(Suppl 16):S106–15.
71. Larsson SH, Charlieu JP, Miyagawa K, Engelkamp D, Rassoulzadegan M, Ross A, et al. Subnuclear localization of WT1 in splicing or transcription factor domains is regulated by alternative splicing. Cell. 1995;81:391–401.
72. Klamt B, Koziell A, Poulat F, Wieacker P, Scambler P, Berta P, et al. Frasier syndrome is caused by defective alternative splicing of WT1 leading to an altered ratio of WT1 +/-KTS splice isoforms. Hum Mol Genet. 1998;7:709–14.
73. Barbaux S, Niaudet P, Gubler MC, Grünfeld JP, Jaubert F, Kuttenn F, et al. Donor splice-site mutations in WT1 are responsible for Frasier syndrome. Nat Genet. 1997;17:467–70.
74. Kikuchi H, Takata A, Akasaka Y, Fukuzawa R, Yoneyama H, Kurosawa Y, et al. Do intronic mutations affecting splicing of WT1 exon 9 cause Frasier syndrome? J Med Genet. 1998;35:45–8.
75. Sirokha D, Gorodna O, Vitrenko Y, Zelinska N, Ploski R, Nef S, et al. A Novel WT1 Mutation Identified in a 46,XX Testicular/Ovotesticular DSD Patient Results in the Retention of Intron 9. Biology (Basel). 2021;10
76. Tsuji Y, Yamamura T, Nagano C, Horinouchi T, Sakakibara N, Ishiko S, et al. Systematic Review of Genotype-Phenotype Correlations in Frasier Syndrome. Kidney Int Rep. 2021;6:2585–93.
77. Miyoshi Y, Santo Y, Tachikawa K, Namba N, Hirai H, Mushiake S, et al. Lack of puberty despite elevated estradiol in a 46,XY phenotypic female with Frasier syndrome. Endocr J. 2006;53:371–6.
78. Gast C, Pengelly RJ, Lyon M, Bunyan DJ, Seaby EG, Graham N, et al. Collagen (COL4A) mutations are the most frequent mutations underlying adult focal segmental glomerulosclerosis. Nephrol Dial Transplant. 2016;31:961–70.
79. Bruening W, Bardeesy N, Silverman BL, Cohn RA, Machin GA, Aronson AJ, et al. Germline intronic and exonic mutations in the Wilms' tumour gene (WT1) affecting urogenital development. Nat Genet. 1992;1:144–8.
80. Eswarakumar VP, Horowitz MC, Locklin R, Morriss-Kay GM, Lonai P. A gain-of-function mutation of Fgfr2c demonstrates the roles of this receptor variant in osteogenesis. Proc Natl Acad Sci U S A. 2004;101:12555–60.
81. Pfaff MJ, Xue K, Li L, Horowitz MC, Steinbacher DM, Eswarakumar JVP. FGFR2c-mediated ERK-MAPK activity regulates coronal suture development. Dev Biol. 2016;415:242–50.
82. Gong SG. Isoforms of receptors of fibroblast growth factors. J Cell Physiol. 2014;229:1887–95.
83. Mohammadi M, Olsen SK, Ibrahimi OA. Structural basis for fibroblast growth factor receptor activation. Cytokine Growth Factor Rev. 2005;16:107–37.
84. Roscioli T, Elakis G, Cox TC, Moon DJ, Venselaar H, Turner AM, et al. Genotype and clinical care correlations in craniosynostosis: findings from a cohort of 630 Australian and New Zealand patients. Am J Med Genet C Semin Med Genet. 2013;163C:259–70.
85. Wilkie AO. Craniosynostosis: genes and mechanisms. Hum Mol Genet. 1997;6:1647–56.
86. Teebi AS, Kennedy S, Chun K, Ray PN. Severe and mild phenotypes in Pfeiffer syndrome with splice acceptor mutations in exon IIIc of FGFR2. Am J Med Genet. 2002;107:43–7.
87. Reardon W, Winter RM, Rutland P, Pulleyn LJ, Jones BM, Malcolm S. Mutations in the fibroblast growth factor receptor 2 gene cause Crouzon syndrome. Nat Genet. 1994;8:98–103.
88. Kan R, Twigg SR, Berg J, Wang L, Jin F, Wilkie AO. Expression analysis of an FGFR2 IIIc 5' splice site mutation (1084+3A->G). J Med Genet. 2004;41:e108.
89. Cornejo-Roldan LR, Roessler E, Muenke M. Analysis of the mutational spectrum of the FGFR2 gene in Pfeiffer syndrome. Hum Genet. 1999;104:425–31.
90. Schell U, Hehr A, Feldman GJ, Robin NH, Zackai EH, de Die-Smulders C, et al. Mutations in FGFR1 and FGFR2 cause familial and sporadic Pfeiffer syndrome. Hum Mol Genet. 1995;4:323–8.
91. Paumard-Hernández B, Berges-Soria J, Barroso E, Rivera-Pedroza CI, Pérez-Carrizosa V, Benito-Sanz S, et al. Expanding the mutation spectrum in 182 Spanish probands with craniosynostosis: identification and characterization of novel TCF12 variants. Eur J Hum Genet. 2015;23:907–14.
92. Oldridge M, Zackai EH, McDonald-McGinn DM, Iseki S, Morriss-Kay GM, Twigg SR, et al. De novo alu-element insertions in FGFR2 identify a distinct pathological basis for Apert syndrome. Am J Hum Genet. 1999;64:446–61.
93. Hollway GE, Suthers GK, Haan EA, Thompson E, David DJ, Gecz J, et al. Mutation detection in FGFR2 craniosynostosis syndromes. Hum Genet. 1997;99:251–5.
94. Del Gatto F, Breathnach R. A Crouzon syndrome synonymous mutation activates a 5' splice site within the IIIc exon of the FGFR2 gene. Genomics. 1995;27:558–9.
95. Li X, Park WJ, Pyeritz RE, Jabs EW. Effect on splicing of a silent FGFR2 mutation in Crouzon syndrome. Nat Genet. 1995;9:232–3.
96. Traynis I, Bernstein JA, Gardner P, Schrijver I. Analysis of the alternative splicing of an FGFR2 transcript due to a novel 5' splice site mutation (1084+1G>A): case report. Cleft Palate Craniofac J. 2012;49:104–8.

97. Fenwick AL, Goos JA, Rankin J, Lord H, Lester T, Hoogeboom AJ, et al. Apparently synonymous substitutions in FGFR2 affect splicing and result in mild Crouzon syndrome. BMC Med Genet. 2014;15:95.
98. Leman R, Parfait B, Vidaud D, Girodon E, Pacot L, Le Gac G, 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;43:2308–23.
99. Glidden DT, Buerer JL, Saueressig CF, Fairbrother WG. Hotspot exons are common targets of splicing perturbations. Nat Commun. 2021;12:2756.
100. Baeza-Centurion P, Miñana B, Valcárcel J, Lehner B: Mutations primarily alter the inclusion of alternatively spliced exons. Elife. 2020, 9.
101. Parthasarathy S, Ruggiero SM, Gelot A, Soardi FC, Ribeiro BFR, Pires DEV, et al. A recurrent de novo splice site variant involving DNM1 exon 10a causes developmental and epileptic encephalopathy through a dominant-negative mechanism. Am J Hum Genet. 2022;109:2253–69.
102. Pozo F, Martinez-Gomez L, Walsh TA, Rodriguez JM, Di Domenico T, Abascal F, et al. Assessing the functional relevance of splice isoforms. NAR Genom Bioinform. 2021;3:lqab044.
103. Wagner N, Çelik MH, Hölzlwimmer FR, Mertes C, Prokisch H, Yépez VA, et al. Aberrant splicing prediction across human tissues. Nat Genet. 2023;55:861–70.
104. Khan M, Cornelis SS, Pozo-Valero MD, Whelan L, Runhart EH, Mishra K, et al. Resolving the dark matter of ABCA4 for 1054 Stargardt disease probands through integrated genomics and transcriptomics. Genet Med. 2020;22:1235–46.
105. van der Klift HM, Jansen AM, van der Steenstraten N, Bik EC, Tops CM, Devilee P, et al. Splicing analysis for exonic and intronic mismatch repair gene variants associated with Lynch syndrome confirms high concordance between minigene assays and patient RNA analyses. Mol Genet Genomic Med. 2015;3:327–45.
106. Kornblihtt AR. Promoter usage and alternative splicing. Curr Opin Cell Biol. 2005;17:262–8.
107. Erwood S, Bily TMI, Lequyer J, Yan J, Gulati N, Brewer RA, et al. Saturation variant interpretation using CRISPR prime editing. Nat Biotechnol. 2022;40:885–95.
108. Becirovic E, Böhm S, Nguyen ON, Riedmayr LM, Koch MA, Schulze E, et al. In Vivo Analysis of Disease-Associated Point Mutations Unveils Profound Differences in mRNA Splicing of Peripherin-2 in Rod and Cone Photoreceptors. PLoS Genet. 2016;12:e1005811.
109. Cheng J, Çelik MH, Kundaje A, Gagneur J. MTSplice predicts effects of genetic variants on tissue-specific splicing. Genome Biol. 2021;22:94.
110. Morak M, Pineda M, Martins A, Gaildrat P, Tubeuf H, Drouet A, et al. Splicing analyses for variants in MMR genes: best practice recommendations from the European Mismatch Repair Working Group. Eur J Hum Genet. 2022;30:1051–9.
111. Dawes R, Bournazos AM, Bryen SJ, Bommireddipalli S, Marchant RG, Joshi H, et al. SpliceVault predicts the precise nature of variant-associated mis-splicing. Nat Genet. 2023;55:324–32.
112. Thompson BA, Martins A, Spurdle AB. A review of mismatch repair gene transcripts: issues for interpretation of mRNA splicing assays. Clin Genet. 2015;87:100–8.
113. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022;604:310–5.
114. Smith C; Kitzman JO. GitHub. 2023a. [https://github.com/kitzmanlab/splfxseq]
115. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Machine Learn Res. 2011;12:2825–30.
116. Smith C; Kitzman JO. GitHub. 2023b. [https://github.com/kitzmanlab/splicebench2023]
117. Smith C; Kitzman JO. Zenodo. 2023. [https://zenodo.org/records/8351879]

Publisher's Note