Variant effect on splicing regulatory elements, branchpoint usage, and pseudoexonization: strategies to enhance bioinformatic prediction using hereditary cancer genes as exemplars

Daffodil Canson[1], Dylan Glubb[1], and Amanda Spurdle[1]

[1]QIMR Berghofer Medical Research Institute

May 5, 2020

Abstract

It is possible to estimate the prior probability of pathogenicity for germline disease gene variants based on bioinformatic prediction of variant effect/s. However, routinely used approaches have likely led to the underestimation and underreporting of variants located outside donor and acceptor splice site motifs that affect mRNA processing. This review presents information about hereditary cancer gene germline variants, outside native splice sites, with experimentally validated splicing effects. We list 81 exonic variants that impact splicing regulatory elements in *BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH6* and *PMS2*. We utilized a pre-existing large-scale BRCA1 functional dataset to map functional splicing regulatory elements, assess the relative performance of different tools to predict effects of 283 variants on such elements, and develop a generic workflow to prioritize variants that may impact splicing regulatory elements. We also describe rare examples of intronic variants that impact branchpoint sites and create pseudoexons. We discuss the challenges in predicting variant effect on branchpoint site usage and pseudoexonization, and suggest strategies to improve the bioinformatic prioritization of such variants for experimental validation. Importantly, our review highlights the importance of considering impact of variants outside donor and acceptor motifs on mRNA splicing and disease causation.

Keywords

Splicing regulatory elements, ESE, ESS, branchpoint, pseudoexon, hereditary cancer genes

Introduction

Evaluating the potential functional impact of variants in Mendelian disease genes is a key component in the interpretation of their clinical significance. Disease gene databases chiefly contain nonsense, frameshift indels, and missense variants, in addition to variants that impact donor and acceptor splice site motifs. In particular, synonymous variants are often dismissed from variant curation and test reporting under the assumption that they are “silent” variants. However, these variants can still impact transcription, mRNA processing and translation (Sauna & Kimchi-Sarfaty, 2011). Further, intronic variants outside of the donor and acceptor splice site motifs are mostly disregarded in clinical testing and/or reporting due to the low sensitivity and specificity of currently available methods to predict their impact on mRNA splicing. This negative bias in recording of synonymous and intronic variants has implications for their inclusion in data analyses and functional studies in research settings.

Current variant interpretation approaches also generally ignore the fact that *all* types of exonic and intronic variants can potentially affect mRNA splicing (we will term these types of variants as being “spliceogenic”). Exonic variants initially annotated as synonymous, missense, nonsense or frameshift based on predicted codon usage can destroy, enhance or create motifs recognized by the mRNA splicing machinery (see below). Intronic variants outside the native splice sites can destroy branchpoint (BP) motifs, or create or enhance the use of cryptic sites. To improve assessment of variant pathogenicity and clinical decision-making, it is important to expand variant curation and reporting to include reliable bioinformatic prediction of spliceogenicity for variants located outside the donor and acceptor splice site motifs.

Variants outside donor and acceptor splice site motifs and impact on mRNA splicing

Precursor mRNAs are transcription products of human genes composed of exons interspersed with introns. Exon-intron boundaries are defined by multiple sequence motifs (Figure 1). Among these are the donor (5') and acceptor (3') splice site motifs based on the definitions by Burge et al (Burge, Tuschi, & Sharp, 1999): 11 bases for the donor splice site motif (from the 3 last exonic to the 8 first intronic bases); and 14 bases for the acceptor splice site motif (from the 12 last intronic to the first 2 exonic bases). The other significant motifs are BP sites and the polypyrimidine tract upstream of the 3' splice site (Z. Wang & Burge, 2008). In mature mRNAs, introns are spliced out and exons are ligated by a complex cellular machinery called the spliceosome, containing small nuclear RNAs and proteins (Z. Wang & Burge, 2008). However, precursor mRNAs can undergo alternative splicing, leading to different mature mRNAs. This process is regulated by *cis*regulatory elements including splicing enhancers and silencers that recruit various RNA-binding proteins (Z. Wang & Burge, 2008). Exonic splicing regulatory elements (SREs) include exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs). Sequence variants that alter the composition, affinity, and function of spliceosomes can lead to the improper identification of exon-intron boundaries, thereby generating mRNAs that encode a premature termination codon, or otherwise encode a dysfunctional protein (G.-S. Wang & Cooper, 2007).

Variants within the first two bases or last three bases of the exon can alter the native splice sites and inactivate them. Exonic variants that introduce sequences that are identical or closely similar to donor or acceptor splice site sequences can be spliceogenic if the *de novo*splice site motif has sufficient activity to outcompete the native splice site. Exonic variants can also induce exon skipping through ESE loss and/or ESS gain (Cartegni, Hastings, Calarco, de Stanchina, & Krainer, 2006).

Intronic variants that abrogate BP sites, commonly located within the -18 to -44 nucleotide window (Signal, Gloss, Dinger, & Mercer, 2018), can lead to exon skipping (Khan et al., 2004; Wappenschmidt et al., 2012; K. Zhang, Nowak, Rushlow, Gallie, & Lohmann, 2008), intron retention (M. Li & Pritchard, 2000), or usage of new distant 3' splice sites (Crotti et al., 2009). Deeper intronic variants, typically more than 100 nucleotides from exon-intron junctions, can lead to insertion of cryptic exons into the mature mRNA transcript by creating a new (or enhancing use of a cryptic) donor or acceptor motif, or by interfering with SREs [reviewed in (Vaz-Drago, Custódio, & Carmo-Fonseca, 2017)].

Utility of splicing prediction tools in variant interpretation

Multiple *in silico* tools have been developed to predict the impact of spliceogenic variants (Table 1), and such prediction is an important component of variant curation and interpretation processes for Mendelian disease genes. Several studies have assessed the utility of prediction tools in interpretation of variants in hereditary breast and ovarian cancer genes. These include: clinical calibration of the MaxEntScan (MES) tool to estimate the prior probability of pathogenicity of genetic variation in *BRCA1* and *BRCA2* due to impact on native donor and acceptor motifs, or the creation of exonic *de novo* donor sites (Vallée et al., 2016); assessment of the sensitivity and specificity of different MES thresholds to predict aberrant splicing using experimentally validated spliceogenic variants in *BRCA1* ,*BRCA2* , *MLH1, MSH2, MSH6* and *PMS2* (Shamsani et al., 2018); and the combined use of MES and Splice Site Finder-like, trained and validated using *in vitro* mRNA data to improve *in silico* prediction of spliceogenic variants in donor and acceptor splice site motifs (Leman et al., 2018). The combined MES and Splice Site Finder-like analysis pipeline has been previously proposed as a prioritization method for splicing analysis of *BRCA1* and*BRCA2* variants of uncertain significance (VUSs) (Houdayer et al., 2012). These studies have shown the reliability of bioinformatic tools in predicting spliceogenic variants in the donor and acceptor splice site motifs, especially those that disrupt the highly conserved dinucleotides at the 3' (AG) and 5' splice sites (GT).

In contrast, predictors of variant effects on exonic SREs or BP sites currently perform poorly (see below), which limits their utility to inform variant classification in routine diagnostics. There is currently no prediction tool specifically designed for pseudoexon-activating variants. In the following sections, we discuss: i) the spliceogenic variants outside of the donor and acceptor splice site motifs; ii) the current tools used to predict their effects on splicing and their predictive performance; and iii) combined strategies using functional studies and *in silico* tools to prioritize variants for confirmatory splicing assays.

Exonic variants can lead to loss/gain of SREs

Variants annotated as synonymous, missense, nonsense, or frameshift variants that disrupt exonic SREs have been identified in hereditary cancer genes. Most of the published splicing assays on exonic SREs have focused on variants in *BRCA1, BRCA2* and the mismatch repair genes (*MLH1* , *MSH2* , *MSH6* and *PMS2*). We have generated a comprehensive list of 81 variants in exonic SREs in these genes that resulted in exon skipping in Supplementary Table 1 (*BRCA1* , *BRCA2*) and Supplementary Table 2 (*MLH1* ,*MSH2* , *MSH6* , *PMS2*). Exonic splicing variants in other hereditary cancer genes include: a synonymous variant in the*APC* gene, NM_000038.6:c.1869G>T [p.(Arg623=)], detected in a familial adenomatous polyposis family, that leads to exon 14 skipping (Montera et al., 2001); and two nonsense *NF1* variants identified in Neurofibromatosis type 1 patients, NM_000267.3:c.6792C>A and NM_000267.3:c.6792C>G [both initially annotated as p.(Tyr2264*)] that induce skipping of exon 37 and exons 36-37 (Baralle et al., 2006; Messiaen, Callens, De Paepe, Craen, & Mortier, 1997). In addition to single nucleotide substitutions, other types of exonic variants can also disrupt SREs, such as small deletions, e.g. NM_000059.3(*BRCA2*):c.470_474del (Sanz et al., 2010), and duplications, e.g. NM_000535.6(*PMS2*):c.325dup (van der Klift et al., 2015). Some variants can act through a combination of mechanisms. For example, NM_000249.3(*MLH1*):c.840T>A [p.(Tyr280*)] and NM_000249.3(*MLH1*):c.842C>T [p.(Ala281Val)] have each been shown to disrupt an SRE and at the same time create a new donor site leading to exon skipping and partial exon deletion (Soukarieh et al., 2016).

Experimental assays can identify active exonic SREs:*BRCA1/2* as exemplars

It has been recommended that a precise and detailed map of active SREs be established for each gene of interest in order for SRE prediction to be useful in clinical diagnostics (Houdayer et al., 2012). A large proportion of missense and synonymous *BRCA1* and *BRCA2*variants are currently catalogued in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) as (likely) benign or VUS. We describe below published findings and our own bioinformatic analysis of existing experimental data, which reveal that some of these variants are located in putative exonic SREs. Variant nomenclature is based on NM_007294.3 for *BRCA1* and NM_000059.3 for *BRCA2* .

Minigene-based microdeletion assays have been used to map ESEs in BRCA2

Minigene constructs, containing a genomic segment from the gene of interest that includes the alternatively spliced exon(s) and flanking intronic regions, express pre-mRNAs. These constructs provide a rapid assay for SRE function, and effect of trans-acting factors on splicing regulation (Cooper, 2005).

Acedo et al. (2015) functionally mapped the ESE-rich regions in*BRCA2* exons 19-27 using minigene splicing assays to improve ESE predictions and facilitate identification of ESE-disrupting spliceogenic variants. Since the density of active ESEs is highest near splice sites (50 nt at both exon ends) (Fairbrother, Holste, Burge, & Sharp, 2004), they mapped functional ESEs by introducing 34 30-nt microdeletions at the ends of each exon (Acedo et al., 2015). They found a microdeletion in exon 19 and another in exon 20 that clearly affected the splicing process, and six other microdeletions in exons 19, 20, 21 and 23 that had weak effects. Three previously characterized ESE variants, c.8378G>A (exon 19), c.8969G>A (exon 23), and c.9006A>T (exon 23) (Acedo et al., 2012), lay within microdeletions shown to impact mRNA splicing, so demonstrating the utility of this strategy to locate putative ESE variants (Acedo et al., 2015). Fraile-Bethencourt et al. adapted this systematic minigene assay approach to map active ESEs in *BRCA2* exons 2-9 and 14-18 (Fraile-Bethencourt et al., 2017; Fraile-Bethencourt, Valenzuela-Palomo, Diez-Gomez, Acedo, & Velasco, 2018; Fraile-Bethencourt, Valenzuela-Palomo, Diez-Gomez, Caloca, et al., 2019; Fraile-Bethencourt, Valenzuela-Palomo, Diez-Gomez, Goina, et al., 2019). Selection of variants within the microdeletion-mapped ESEs improved the specificity of bioinformatic predictions (Fraile-Bethencourt, Valenzuela-Palomo, Diez-Gomez, Goina, et al., 2019). Results from these assays have also been useful in re-classifying variants. For example, two variants in ClinVar, c.441A>G [p.(Gln147=), likely benign] and c.451G>A [p.(Val151Ile), VUS], were designated as spliceogenic variants and re-interpreted as VUS and likely pathogenic, respectively (Fraile-Bethencourt, Valenzuela-Palomo, Diez-Gomez, Goina, et al., 2019).

Saturation genome editing experimental data are used here to map putative SREs in BRCA1

Currently, there are no studies that systematically map the active SREs in *BRCA1* exons, covering their entire lengths and all possible nucleotide substitutions. However, we took advantage of available mRNA expression data from a recently published large-scale functional analysis of *BRCA1* (Findlay et al., 2018) to identify putative SREs across multiple exons of this gene. The study of Findlay et al. applied saturation genome editing to measure the cell survival consequences of all possible single nucleotide variants in the 13 exons that encode the BRCA1 RING and BRCT protein domains, critical for its role as a tumor suppressor (Findlay et al., 2018). Specifically, near-haploid HAP1 cells were genomically edited using CRISPR-Cas9 to introduce *BRCA1* single nucleotide variants and variant abundances were quantified by targeted DNA sequencing as readout for a cell survival assay; this information was used to assign a "function score." Variants that did not affect DNA abundance were classified as "functional"; otherwise, variants were classified as "non-functional" or "intermediate" depending on the extent of DNA depletion. In total, function scores were calculated for 3,893*BRCA1* variants, and these scores were observed to accurately predict variant pathogenicity as reported to the ClinVar database. mRNA expression scores were also determined for 96% of the functionally characterized variants, and variants that were depleted in mRNA relative to DNA were interpreted to affect mRNA expression and/or processing.

From this dataset, we selected 33 *BRCA1* synonymous or missense variants in putative SREs (Table 2) based on the following criteria (Figure 2): (a) depleted in mRNA (Findlay mean RNA score < -2); (b) non-functional or intermediate function based on DNA depletion; (c) outside of the donor and acceptor splice site motifs; (d) not predicted to create *de novo* donor or acceptor sites by the MES-based Variant Effect Predictor plugin using the thresholds and decision flowchart described in Shamsani et al. (2018); and (e) predicted to alter or create SREs by at least one SRE algorithm in HSF. These bioinformatic tools were chosen because they are freely available and easy to use. The MES-based Variant Effect Predictor plugin also allows high-throughput submission, and HSF accepts multiple variant queries for analysis using 14 different SRE algorithms (Table 1) in a single platform. Although nonsense variants can also alter SREs to lead to exon skipping (Supplementary Tables 2 and 3), these were excluded because they are expected to deplete mRNA via nonsense-mediated decay.

Exons 2, 3, and 19-22 did not harbor variants that passed the above criteria. The 33 variants prioritized as likely to impact SREs, are shown in Table 2. We then mapped the location of putative SREs in exons 5, 6, 16-18, 23 and 24 of*BRCA1* by identifying SRE sequences that overlap with these 33 variants (Figure 3, Supplementary Figure 1). Notably, the putative SREs mapped to exons with at least one weak splice site (MES score < 6.2), or with moderate strength for both splice sites (MES score between 6.2 and 8.5) (Table 3). With a single exception, in exon 23, putative SREs did *not* map to exons with strong splice donors (MES score [?] 8.5) (Table 3). Since variants demonstrating minor mRNA depletion (Findlay mean RNA score between -0.5 and -2) were excluded to limit false predictions, this map is expected to capture only putative SREs with strong activity.

The validity of the mapping approach is supported by published mRNA splicing assay results that relate to variants prioritized as SRE-disrupting (Table 2). Variants c.5080G>A, c.5434C>G and c.5453A>G have been proven to lead to exon skipping in previous studies. Variants c.5080G>A and c.5123C>G are located at the same nucleotide position as three other variants for which exon skipping has been previously reported. Of these latter three variants, one was excluded from our mapping analysis since it encodes a premature termination codon, one had a minor mRNA depletion score of -0.55 above the filter of < -2, and the last was a 2 bp deletion and thus not assayed by Findlay et al. (2018). Eleven of the 33 putative SRE-disrupting variants are reported in ClinVar, where nine are catalogued based on predicted codon usage, and not annotated as spliceogenic variants; the (likely) pathogenic classification of c.5434C>G and c.5453A>G considered published splicing assay results as evidence (Table 2). Eight of these variants are currently not interpreted as (likely) pathogenic: c.5007C>T [p.(Ala1669=), likely benign], c.5044G>A [p.(Glu1682Lys), benign], c.5044G>C [p.(Glu1682Gln), VUS], c.5045A>T [p.(Glu1682Val), VUS], c.5078C>T [p.(Ala1693Val), VUS], c.5080G>A [p.(Glu1694Lys), VUS], c.5444G>C [p.(Trp1815Ser), VUS], and c.5528C>A [p.(Ala1843Glu),VUS].

The assay of Findlay et al. (2018) shows the effect of variants on mRNA levels and does not directly inform variant effect on mRNA splicing. Follow-up splicing assays would still be needed to confirm SRE-related mRNA aberrations for those variants in Table 2 without previously reported splicing assay results. The confirmatory splicing assays would provide further evidence to establish the *BRCA1* SRE map for the exons examined, as well as potentially aiding the re-interpretation of variant pathogenicity.

Bioinformatic analysis of exonic SREs

SRE predictions have poor specificity. There are several factors that contribute to the complexity of SRE prediction, including the diverse range of splicing regulatory motifs (Ke et al., 2011; X. H.-F. Zhang & Chasin, 2004) and the context-dependence of their activity (Fu & Ares Jr, 2014; Z. Wang & Burge, 2008). The surrounding sequences and their location in the gene relative to the consensus splice sites significantly impact their activity and usage. For instance, some ESS motifs, including G runs, can *promote* splicing when located in an intron (Z. Wang & Burge, 2008). Moreover, RNA secondary structure and chromatin state may also influence SRE accessibility affecting its usage (reviewed by (Fu & Ares Jr, 2014; Hnilicova & Staněk, 2011)).

There are already several datasets and prediction algorithms (Table 1) that have been used to identify SREs or test if a variant can potentially create or abolish SREs (reviewed by Grodecká, Buratti, and Freiberger (2017)). However, experimental studies have shown that these bioinformatic prediction tools have high false positive rates. For example, one of the largest studies to date (Houdayer et al., 2012) reported that predictions were confirmed for only 14% (15/108) of*BRCA1* and *BRCA2* variants predicted to alter ESEs using a combination of ESEfinder, RESCUE-ESE, PESE octamer, and HSF algorithms.

More recently, two studies have assessed both positive *and*negative predictive values of selected bioinformatic tools to determine variant effects on SREs. ΔtESRseq (using hexamer scores from Ke et al. (2011)) and $\Delta\text{HZ}_{\text{EI}}$ were reported to perform better than ΔΨ and EX-SKIP in analysis of 154 variants (including 50 spliceogenic) from select exons from five genes (Soukarieh et al., 2016). The data from this study led the authors to postulate that the predictive performance of SRE-dedicated tools varies for different genes and exons (Soukarieh et al., 2016). For example, sensitivity of ΔtESRseq ranged from 67-100% and specificity from 66-97% depending on the gene and exon (Soukarieh et al., 2016). In another evaluation of ΔtESRseq, $\Delta\text{HZ}_{\text{EI}}$, and EX-SKIP (Grodecká et al., 2017), analysis of only 20 variants (10 spliceogenic) from four genes found that ΔtESRseq had higher sensitivity (80%) but lower specificity (60%) compared to $\Delta\text{HZ}_{\text{EI}}$ and EX-SKIP (both 70% sensitivity, 70% specificity). However, given the sample sizes for these two studies (Grodecká et al., 2017; Soukarieh et al., 2016), it is difficult to have confidence in their assessment of comparative performance of bioinformatic tools.

Map of *BRCA1* putative SREs is used here to assess bioinformatic predictor performance

To extend the comparisons described above, we have evaluated the performance of ΔtESRseq, $\Delta\text{HZ}_{\text{EI}}$, and HOT-SKIP (same approach as EX-SKIP, examines all possible exonic substitutions simultaneously) (Table 4, Supplementary Table 3). We used the 33*BRCA1* variants identified as located in putative SREs (Table 2) as positive controls, and 250 non-spliceogenic variants as negative controls, as selected from all exons included in the assays of Findlay et al. (2018) (Figure 2). HSF was used in the selection of positive control variants as it incorporates several different algorithms thus capturing a more comprehensive set of SRE sequences; by design HSF could thus not be used to assess sensitivity in a comparative analysis but it tested a large proportion of negative control variants as false positives (27% specificity). Previous studies used $\Delta\text{HZ}_{\text{EI}}$ arbitrary thresholds of -20 (Soukarieh et al., 2016) and -0.5 (Grodecká et al., 2017) and ΔtESRseq cut-off of

-0.5 (Grodecká et al., 2017; Soukarieh et al., 2016). For our analysis, ΔtESRseq and ΔHZ_{EI} cut-off scores were adjusted based on serial Matthews Correlation Coefficient calculations to obtain optimal predictive values: we set -0.75 for ΔtESRseq and -5 for ΔHZ_{EI} as the cut-off scores. We set the HOT-SKIP threshold (alt/wt > 1) based on the EX-SKIP cut-off score used by Grodecká et al. (2017). Results comparing tool performance are shown in Table 4. ΔHZ_{EI} had the best performance with 76% sensitivity and 82% specificity, followed by ΔtESRseq with 73% sensitivity and 80% specificity. HOT-SKIP had the lowest sensitivity (45%) and specificity (78%). Further, as a secondary analysis, false positive variants located in exons with no mapped SRE (see Table 3) were designated as true negatives (i.e. they were not predicted to impact an SRE). This markedly improved the specificity of all three tools (Table 4).

Future use of mapped SREs in *BRCA1* to improve SRE prediction

Solving the problem of over-prediction is an important step towards the utility of SRE-dedicated bioinformatic tools in variant interpretation and clinical diagnostics. As shown in our detailed *BRCA1* SRE map (Supplementary Figure 1), there are negative control variants within the mapped SREs that are predicted by HSF to alter these motifs. Some are even located at the same nucleotide position as positive control variants. For example, c.5007C>T, categorized as non-functional and with effects on mRNA depletion (as per Figure 2), is designated a true positive since it is also predicted to create an ESS by HSF (Supplementary Figure 1); whereas, c.5007C>A and c.5007C>G have no functional impact, and are designated false positives since they were predicted to create an ESS and break an ESE, respectively (Supplementary Figure 1). ΔtESRseq, ΔHZ_{EI}, and HOT-SKIP – which combine the scores of ESEs and ESSs disrupted or created by a variant – correctly predicted c.5007C>A and c.5007C>G to have no impact on an SRE. Similar results are observed for other co-located HSF-predicted false positive variants at c.5127, c.5130, c.5430, and c.5472 (Supplementary Figure 1), where at least two of the three tools (ΔtESRseq, ΔHZ_{EI}, and HOT-SKIP) had negative calls in agreement with mRNA depletion score results. While the quantitative combined ESS-ESE scoring approach of ΔtESRseq, ΔHZ_{EI}, and HOT-SKIP appears to significantly lower the number of HSF-predicted false positives, there are still negative control variants within the mapped SREs that are predicted as impacting SREs by these three tools. Clearly, there are other factors that need to be considered to improve prediction of variant effect with mapped SREs.

The false positive variants can be studied further to gain more understanding of the structural features that prevent the usage of SREs. For false positive variants outside of the mapped SREs, the location of predicted SREs with respect to local mRNA secondary structure could also play a role e.g. inclusion of SRE in the stem of a stem-loop structure may possibly lessen the access of a corresponding RNA-binding protein (Buratti et al., 2004). In the same way, the positive control dataset of 33 variants could be assessed for structural features that enable these variants to alter mRNA expression. More information on structural patterns that influence exonic SRE activity, which can be obtained from bioinformatic analysis, may be useful in improving SRE prediction not only in *BRCA1* but also in other genes.

Prioritization model to identify SRE-disrupting variants for splicing analysis

While the current SRE-dedicated bioinformatic tools have significant limitations in terms of specificity, our results indicate that using a combination of tools can improve prediction and increase confidence in the selection of variants for confirmatory splicing assays. Drawing from the variant selection process (Figure 2), results relating to donor and acceptor splice site strength (Table 3), and the evaluation of SRE prediction tools (Table 4), we developed a generic workflow. This prioritization model uses MES for native and *de novo* splice site analysis, and HSF and ΔHZ_{EI} in series for SRE prediction, as summarized in Figure 4.

Intronic variants can abrogate a BP site or activate a pseudoexon

Only a limited number of variants in hereditary cancer genes have been reported to cause aberrant splicing through the alteration of BP sites (Table 5). Of these, experimental validation of BP abrogation has been conducted for only two *XPC* variants detected in patients with xeroderma pigmentosum, a condition that increases the risk of skin cancers: LRG_472t1:c.413-24A>G resulted in partial skipping of exon 4 skipping in patient-derived mRNA (Khan et al., 2004); and LRG_472t1:c.413-9T>A, located within the acceptor motif but *also* annotated as a BP site variant, was found to lead to complete exon 4 skipping (Khan et al., 2004). The -9 and -24 nucleotides in intron 3 of *XPC* were subsequently shown to be functional BP sites necessary for the efficient and accurate splicing of*XPC* pre-mRNA using U2 small nuclear ribonucleoprotein-BP interaction assays (Khan et al., 2010). Two substitutions predicted to alter a BP motif within the *RB1* gene, identified in patients with retinoblastoma, result in the skipping of a downstream exon (Houdayer et al., 2008; K. Zhang et al., 2008). Similarly, a substitution at the -18 nucleotide upstream of exon 5 in *BRCA1*(LRG_292t1:c.135-18T>G) resulted in a three-fold increase of the Δ5 transcript isoform in an analysis of mRNA from a hereditary breast and ovarian cancer patient (Wappenschmidt et al., 2012). This variant was not captured by *in silico* prediction methods but it is within the -18 to -44 nucleotide window of high-confidence annotated BPs (Signal, Gloss, Dinger, & Mercer, 2018). Nine other variants within the BP window in *BRCA1* , *MLH1* and *RAD51C* have been reported to lead to exon skipping (Leman et al., 2020).

Pseudoexon-activating variants have been documented mostly in rare monogenic disorders, but several examples have also been reported in genes causing hereditary cancer syndromes (Vaz-Drago et al., 2017). To date, there are at least 13 documented pseudoexon-activating variants in hereditary cancer genes (Table 5): 12 are single nucleotide variants that create a new 5' splice site or strengthen an existing cryptic 5' splice site, and the other is a 4-bp deletion that introduces an intron-splicing processing element.

Bioinformatic prediction of BP site abrogation

BP prediction tools (Table 1) have demonstrated poor specificity due to BP motif degeneracy combined with a lack of experimental data to train algorithms (Corvelo, Hallegger, Smith, & Eyras, 2010). BP characterization has lagged far behind that of 5' and 3' splice sites because of experimental difficulties in detecting BPs (Paggi & Bejerano, 2018). A large genome-wide dataset of experimentally confirmed BPs (Mercer et al., 2015) has been used to develop the BP prediction tools Branchpointer and LaBranchoR. Based on the Mercer dataset (Mercer et al., 2015), only ˜18% of human 3' splice sites have high confidence experimental BP annotations (Mercer et al., 2015; Paggi & Bejerano, 2018).

The Branchpointer BP annotations were used to attribute hundreds of clinically associated variants with changes in BP architecture, but the impact of these variants on splicing was largely uncharacterized (Signal et al., 2018). Other tools (SVM-BPfinder, BPP, RNABPS) are also available but, similar to Branchpointer and LaBranchoR, these are mainly for predicting the presence of a BP site. Namely, these tools were not designed to automatically identify spliceogenic variants, and require separate input of wild-type and variant intronic sequences for non-automated comparison of scores. Branchpointer also allows input of single nucleotide variants using rsIDs to evaluate separately the effect of reference and alternative variants on BPs (Signal et al., 2018). The use of R by Branchpointer, and python scripts by LaBranchoR and BPP, have also rendered these tools less accessible to non-bioinformatician users (Leman et al., 2020). HSF, an older and easy-to-use online splicing tool, can directly analyze an intronic variant to predict BP site abrogation; however, recent evaluations have revealed its poor performance in detecting experimentally verified BPs (Leman et al., 2020; Signal et al., 2018; Q. Zhang et al., 2017).

It is important to note that variants predicted to disrupt a BP do not necessarily induce aberrant splicing, as introns can have multiple functional BPs (Mercer et al., 2015), which adds to the complexity of predicting the spliceogenicity of a single variant in the BP window. Moreover, in the analysis of Leman et al. (2020), the use of score change to predict BP disruption by a variant was found to *not* be the best strategy to predict spliceogenic variants. According to Leman et al. (2020), the best approach would be to consider a variant as potentially spliceogenic if it is located in the BP motif regardless of score change. Performance of BPP, Branchpointer, HSF, LaBranchoR, RNABPS, and SVM-BPfinder was evaluated by checking the co-location of confirmed spliceogenic variants within predicted BP motifs, and revealed BPP as having the highest accuracy of 89.17% (Leman et al., 2020). In their positive control set of 38 spliceogenic variants, 32 variants were within BP motifs predicted by BPP, which predicted a total of 39 BP motifs (Leman et al., 2020).

Generally, the current BP prediction tools are useful in prioritizing candidate spliceogenic variants for downstream analysis through predicting their location in putative BP sites. Further, while variants reported to alter a BP site sequence generally lead to exon skipping, other types of splicing aberrations have been observed (Crotti et al., 2009; M. Li & Pritchard, 2000). Hence, the current BP prediction tools are not suitable for predicting a specific splicing effect.

Bioinformatic prediction of pseudoexon activation

Currently, there is no bioinformatic tool dedicated to prediction of pseudoexon-activating variants together with the corresponding size and/or sequence of the inserted cryptic exon. The current prediction strategy is to determine whether a deep intronic variant leads to a*de novo* splice site gain, and then separately check for a nearby pre-existing cryptic splice site of opposite polarity that could define the boundary of the new exon (Caminsky et al., 2016; Lee et al., 2017).

In the variant prioritization method of Caminsky et al. (2016), an Information Theory model was used to measure changes in splicing-relevant protein binding sites and predict whether a variant would lead to a gain or loss of a splicing motif. A total of 623 variants in hereditary breast and ovarian cancer genes were predicted to create or strengthen an intronic cryptic splice site. However, only 17 variants were prioritized as likely to create a pseudoexon due to their location within 250 nucleotides of another existing intronic site of opposite polarity and the existence of an hnRNPA1 site within five nucleotides of the acceptor of the predicted pseudoexon (Caminsky et al., 2016). However, these prioritized variants have yet to undergo splicing analysis, and so it is not possible to assess the performance of the Information Theory model.

Another workflow incorporates use of CryptSplice, a tool which extends the splice site definition of Burge et al. (1999) to capture more sequence component information (Lee et al., 2017). The donor sequences extend from seven nucleotides upstream of GT (-7) to six nucleotides downstream of GT (+6), and acceptor sequences extend from 68 nucleotides upstream of AG (-68) to 20 nucleotides downstream of AG (+20). This extended definition was previously reported to improve splice site prediction by combining the feature information of splicing signals and SREs around splice sites (J. L. Li, Wang, Wang, Bai, & Yuan, 2012). In an analysis of *CFTR* variants in cystic fibrosis patients with partly explained genetic cause for their recessively inherited disease, intronic variants underwent prioritization to detect variants that may lead to pseudoexon activation (Lee et al., 2017). Of 41 candidate intronic variants predicted to create either donor or acceptor sequences using CryptSplice, only three donor sequences were additionally predicted to activate pseudoexons by manual evaluation of the surrounding sequence for a splice site of opposite polarity (Lee et al., 2017). Two variants were shown to lead to pseudoexon insertion resulting in transcript loss due to nonsense-mediated decay; and the other, with a weakly predicted upstream acceptor, did not lead to aberrant splicing. In the same study, CryptSplice analysis of 4,685 *DKC1* unique variants present in six individuals identified five candidate donor sequences and 12 candidate acceptor sequences (Lee et al., 2017). Only one of the five candidate donors was predicted to activate a pseudoexon; while mRNA analysis provided evidence for pseudoexonization, the donor activated by this *DKC1* variant did not pair with the CryptSplice predicted acceptor, but rather with another acceptor 14 nucleotides upstream (Lee et al., 2017).

The Information Theory and CryptSplice prioritization methods for pseudoexon-activating variants did not comprehensively take into account the role of SREs, which can influence the expression of pseudoexons. To illustrate, the Information Theory model predicted that *MLH1*LRG_216t1:c.1559-1732A>T creates a new acceptor and activates a 239-bp pseudoexon due to the presence of a downstream pre-existing cryptic donor (Caminsky et al., 2016). However, our analysis of the pseudoexon sequence using HSF revealed a cluster of putative ESS octamers ((X. H.-F. Zhang & Chasin, 2004), with high relative activity and located within 30 nucleotides upstream of the cryptic donor that potentially inactivates this cryptic donor (Supplementary Figure 2). Therefore, a prediction model that incorporates both splice site motifs and the distribution of SREs within candidate pseudoexons and their flanking regions is likely to improve the accuracy of pseudoexon activation predictions.

Conclusions

In this paper, we have presented functional evidence for spliceogenic variants that are generally overlooked in clinical genetic testing and/or reporting, including: variants that affect SREs, abrogate BP sites, or activate pseudoexons. Bioinformatic analysis considering variant effects at the mRNA level may help prioritize likely functional variants currently annotated as (likely) benign or VUS for additional functional and clinical analyses. Further, clinical diagnostic laboratories may need to consider expanding their sequencing coverage and/or variant annotation to include BP window and deep intronic regions to detect additional pathogenic intronic variants, particularly when strongly indicated by patient presentation. However, improving the low performance of current predictors is a challenge due to the limited size of experimentally validated training data. Clearly, experimental studies that assess variants outside of the donor and acceptor splice site motifs for splicing mechanisms are needed to further calibrate algorithms, and to improve prediction of variant effect. We have shown that results from a published large-scale saturation genome editing experiment can be used to map SREs, to assess the performance of bioinformatic predictors, and inform development of a prioritization workflow to detect variants that impact SREs. As more such data become available, we anticipate that the expansion of training datasets will lead to improvements in approaches to predict variant effect/s. Such advances will be critical to improve the sensitivity and specificity of bioinformatic prediction of variant effect/s on SREs, BP sites, and pseudoexon usage, and thereby improve assessment of variant pathogenicity.

Author Contributions

All authors conceived the review. D.C. performed the literature review and conducted the analysis. All authors contributed to manuscript writing and completion of the final version.

Acknowledgement

D. C. is supported by a QIMR Berghofer Ailsa Zinns PhD Scholarship and UQ Research Training Tuition Fee Offset. A.B.S has been supported by NHMRC Fellowship funding (ID1061779, ID1177524). The authors thank Michael Parsons from QIMR Berghofer Medical Research Institute for his helpful advice in the use of relevant bioinformatic tools.

Conflict of Interest Statement

The authors declare no conflict of interest.

Availability of Data

The data that supports the findings of this review and analysis are available in the supplementary material of this article.

REFERENCES

Acedo, A., Hernández-Moro, C., Curiel-García, Á., Díez-Gómez, B., & Velasco, E. A. (2015). Functional Classification of BRCA2 DNA Variants by Splicing Assays in a Large Minigene with 9 Exons. *Human Mutation, 36* (2), 210-221. doi:10.1002/humu.22725

Acedo, A., Sanz, D. J., Durán, M., Infante, M., Pérez-Cabornero, L., Miner, C., & Velasco, E. A. (2012). Comprehensive splicing functional analysis of DNA variants of the BRCA2 gene by hybrid minigenes.*Breast Cancer Research, 14* (3), R87. doi:10.1186/bcr3202

Anczukow, O., Buisson, M., Leone, M., Coutanson, C., Lasset, C., Calender, A., . . . Mazoyer, S. (2012). BRCA2 deep intronic mutation causing activation of a cryptic exon: opening toward a new preventive therapeutic strategy. *Clin Cancer Res, 18* (18), 4903-4909. doi:10.1158/1078-0432.ccr-12-1100

Baralle, M., Skoko, N., Knezevich, A., De Conti, L., Motti, D., Bhuvanagiri, M., . . . Baralle, F. E. (2006). NF1 mRNA biogenesis: Effect of the genomic milieu in splicing regulation of the NF1 exon 37 region. *FEBS Letters, 580* (18), 4449-4456. doi:10.1016/j.febslet.2006.07.018

Buratti, E., Muro, A. F., Giombi, M., Gherbassi, D., Iaconcig, A., & Baralle, F. E. (2004). RNA Folding Affects the Recruitment of SR Proteins by Mouse and Human Polypurinic Enhancer Elements in the Fibronectin EDA Exon. *Molecular and Cellular Biology, 24* (3), 1387. doi:10.1128/MCB.24.3.1387-1400.2004

Burge, C. B., Tuschi, T., & Sharp, P. A. (1999). Splicing of precursors to mRNAs by the spliceosomes. In C. S. H. L. Press (Ed.), *The RNA World II* (pp. pp. 525–560). NY: Oxford University Press.

Caminsky, N. G., Mucaki, E. J., Perri, A. M., Lu, R., Knoll, J. H. M., & Rogan, P. K. (2016). Prioritizing Variants in Complete Hereditary Breast and Ovarian Cancer Genes in Patients Lacking Known BRCA Mutations. *Human Mutation, 37* (7), 640-652. doi:10.1002/humu.22972

Campos, B., Díez, O., Domènech, M., Baena, M., Balmaña, J., Sanz, J., . . . Baiget, M. (2003). RNA analysis of eight BRCA1 and BRCA2 unclassified variants identified in breast/ovarian cancer families from Spain. *Human Mutation, 22* (4), 337-337. doi:10.1002/humu.9176

Cartegni, L., Chew, S. L., & Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews Genetics, 3* , 285. doi:10.1038/nrg775

Cartegni, L., Hastings, M. L., Calarco, J. A., de Stanchina, E., & Krainer, A. R. (2006). Determinants of Exon 7 Splicing in the Spinal Muscular Atrophy Genes, SMN1 and SMN2. *The American Journal of Human Genetics, 78* (1), 63-77. doi:10.1086/498853

Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q., & Krainer, A. R. (2003). ESEfinder: a web resource to identify exonic splicing enhancers.*Nucleic Acids Research, 31* (13), 3568-3571.

Castellanos, E., Rosas, I., Solanes, A., Bielsa, I., Lázaro, C., Carrato, C., . . . on behalf of the, N. F. M. C. H.-I. C. O. I. (2013). In vitro antisense therapeutics for a deep intronic mutation causing Neurofibromatosis type 2. *European Journal Of Human Genetics, 21* (7), 769-773. doi:10.1038/ejhg.2012.261

Cavalieri, S., Pozzi, E., Gatti, R. A., & Brusco, A. (2013). Deep-intronic ATM mutation detected by genomic resequencing and corrected in vitro by antisense morpholino oligonucleotide (AMO).*European Journal Of Human Genetics, 21* (7), 774-778. doi:10.1038/ejhg.2012.266

Clendenning, M., Buchanan, D. D., Walsh, M. D., Nagler, B., Rosty, C., Thompson, B., . . . Young, J. P. (2011). Mutation deep within an intron of MSH2 causes Lynch syndrome. *Familial Cancer, 10* (2), 297-301. doi:10.1007/s10689-011-9427-0

Cooper, T. A. (2005). Use of minigene systems to dissect alternative splicing elements. *Methods, 37* (4), 331-340. doi:10.1016/j.ymeth.2005.07.015

Corvelo, A., Hallegger, M., Smith, C. W. J., & Eyras, E. (2010). Genome-Wide Association between Branch Point Properties and Alternative Splicing. *PLOS Computational Biology, 6* (11), e1001016. doi:10.1371/journal.pcbi.1001016

Coutinho, G., Xie, J., Du, L., Brusco, A., Krainer, A. R., & Gatti, R. A. (2005). Functional significance of a deep intronic mutation in the ATM gene and evidence for an alternative exon 28a. *Human Mutation, 25* (2), 118-124. doi:10.1002/humu.20170

Crotti, L., Lewandowska, M. A., Schwartz, P. J., Insolia, R., Pedrazzini, M., Bussani, E., . . . Pagani, F. (2009). A KCNH2 branch point mutation causing aberrant splicing contributes to an explanation of genotype-negative long QT syndrome. *Heart Rhythm, 6* (2), 212-218. doi:10.1016/j.hrthm.2008.10.044

Dehainault, C., Michaux, D., Pagès-Berhouet, S., Caux-Moncoutier, V., Doz, F., Desjardins, L., . . . Houdayer, C. (2007). A deep intronic mutation in the RB1 gene leads to intronic sequence exonisation.*European Journal Of Human Genetics, 15* (4), 473-477. doi:10.1038/sj.ejhg.5201787

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., & Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Research, 37* (9), e67-e67. doi:10.1093/nar/gkp215

Di Giacomo, D., Gaildrat, P., Abuli, A., Abdat, J., Frébourg, T., Tosi, M., & Martins, A. (2013). Functional Analysis of a Large set of BRCA2 exon 7 Variants Highlights the Predictive Value of Hexamer Scores in Detecting Alterations of Exonic Splicing Regulatory Elements.*Human Mutation, 34* (11), 1547-1557. doi:10.1002/humu.22428

Erkelenz, S., Theiss, S., Otte, M., Widera, M., Peter, J. O., & Schaal, H. (2014). Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. *Nucleic Acids Research, 42* (16), 10681-10697. doi:10.1093/nar/gku736

Fairbrother, W. G., Holste, D., Burge, C. B., & Sharp, P. A. (2004). Single Nucleotide Polymorphism–Based Validation of Exonic Splicing Enhancers. *PLOS Biology, 2* (9), e268. doi:10.1371/journal.pbio.0020268

Fairbrother, W. G., Yeh, R.-F., Sharp, P. A., & Burge, C. B. (2002). Predictive Identification of Exonic Splicing Enhancers in Human Genes.*Science, 297* (5583), 1007-1013.

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., . . . Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature, 562* (7726), 217-222. doi:10.1038/s41586-018-0461-z

Fraile-Bethencourt, E., Díez-Gómez, B., Velásquez-Zapata, V., Acedo, A., Sanz, D. J., & Velasco, E. A. (2017). Functional classification of DNA variants by hybrid minigenes: Identification of 30 spliceogenic variants of BRCA2 exons 17 and 18. *PLOS Genetics, 13* (3), e1006691. doi:10.1371/journal.pgen.1006691

Fraile-Bethencourt, E., Valenzuela-Palomo, A., Díez-Gómez, B., Acedo, A., & Velasco, E. A. (2018). Identification of Eight Spliceogenic Variants in BRCA2 Exon 16 by Minigene Assays. *Frontiers in Genetics, 9* (188). doi:10.3389/fgene.2018.00188

Fraile-Bethencourt, E., Valenzuela-Palomo, A., Díez-Gómez, B., Caloca, M. J., Gómez-Barrero, S., & Velasco, E. A. (2019). Minigene Splicing Assays Identify 12 Spliceogenic Variants of BRCA2 Exons 14 and 15.*Frontiers in Genetics, 10* (503). doi:10.3389/fgene.2019.00503

Fraile-Bethencourt, E., Valenzuela-Palomo, A., Díez-Gómez, B., Goina, E., Acedo, A., Buratti, E., & Velasco, E. A. (2019). Mis-splicing in breast cancer: identification of pathogenic BRCA2 variants by systematic minigene assays. *The Journal of Pathology, 248* (4), 409-420. doi:10.1002/path.5268

Fu, X.-D., & Ares Jr, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics, 15* , 689. doi:10.1038/nrg3778

Gaildrat, P., Krieger, S., Théry, J.-C., Killian, A., Rousselin, A., Berthet, P., . . . Tosi, M. (2010). The *BRCA1* c.5434C-G (p.Pro1812Ala) variant induces a deleterious exon 23 skipping by affecting exonic splicing regulatory elements. *Journal of Medical Genetics, 47* (6), 398. doi:10.1136/jmg.2009.074047

Goina, E., Skoko, N., & Pagani, F. (2008). Binding of DAZAP1 and hnRNPA1/A2 to an Exonic Splicing Silencer in a Natural BRCA1 Exon 18 Mutant. *Molecular and Cellular Biology, 28* (11), 3850-3860. doi:10.1128/mcb.02253-07

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., . . . Ast, G. (2006). Comparative Analysis Identifies Exonic Splicing Regulatory Sequences—The Complex Definition of Enhancers and Silencers. *Molecular Cell, 22* (6), 769-781. doi:10.1016/j.molcel.2006.05.008

Grodecka, L., Buratti, E., & Freiberger, T. (2017). Mutations of Pre-mRNA Splicing Regulatory Elements: Are Predictions Moving Forward to Clinical Diagnostics? *International Journal of Molecular Sciences, 18* (8). doi:10.3390/ijms18081668

Hnilicova, J., & Staněk, D. (2011). Where splicing joins chromatin.*Nucleus, 2* (3), 182-188. doi:10.4161/nucl.2.3.15876

Houdayer, C., Caux-Moncoutier, V., Krieger, S., Barrois, M., Bonnet, F., Bourdon, V., . . . Stoppa-Lyonnet, D. (2012). Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. *Human Mutation, 33* (8), 1228-1238. doi:10.1002/humu.22101

Houdayer, C., Dehainault, C., Mattler, C., Michaux, D., Caux-Moncoutier, V., Pagès-Berhouet, S., . . . Stoppa-Lyonnet, D. (2008). Evaluation of in silico splice tools for decision-making in molecular diagnosis.*Human Mutation, 29* (7), 975-982. doi:10.1002/humu.20765

Ke, S., Shang, S., Kalachikov, S. M., Morozova, I., Yu, L., Russo, J. J., . . . Chasin, L. A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Research, 21* (8), 1360-1374.

Khan, S. G., Metin, A., Gozukara, E., Inui, H., Shahlavi, T., Muniz-Medina, V., . . . Kraemer, K. H. (2004). Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced XPC mRNA levels that correlate with cancer risk. *Human Molecular Genetics, 13* (3), 343-352. doi:10.1093/hmg/ddh026

Khan, S. G., Yamanegi, K., Zheng, Z.-M., Boyle, J., Imoto, K., Oh, K.-S., . . . Kraemer, K. H. (2010). XPC branch-point sequence mutations disrupt U2 snRNP binding, resulting in abnormal pre-mRNA splicing in xeroderma pigmentosum patients. *Human Mutation, 31* (2), 167-175. doi:10.1002/humu.21166

Lee, M., Roos, P., Sharma, N., Atalar, M., Evans, T. A., Pellicore, M. J., . . . Cutting, G. R. (2017). Systematic Computational Identification of Variants That Activate Exonic and Intronic Cryptic Splice Sites.*The American Journal of Human Genetics, 100* (5), 751-765. doi:10.1016/j.ajhg.2017.04.001

Leman, R., Gaildrat, P., Gac, G. L., Ka, C., Fichou, Y., Audrezet, M.-P., . . . Houdayer, C. (2018). Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort.*Nucleic Acids Research, 46* (15), 7913-7923. doi:10.1093/nar/gky372

Leman, R., Tubeuf, H., Raad, S., Tournier, I., Derambure, C., Lanos, R., . . . Krieger, S. (2020). Assessment of branch point prediction tools to predict physiological branch points and their alteration by variants.*BMC Genomics, 21* (1), 86. doi:10.1186/s12864-020-6484-5

Li, J. L., Wang, L. F., Wang, H. Y., Bai, L. Y., & Yuan, Z. M. (2012). High-accuracy splice site prediction based on sequence component and position features. *Genet Mol Res, 11* (3), 3432-3451. doi:10.4238/2012.September.25.12

Li, M., & Pritchard, P. H. (2000). Characterization of the Effects of Mutations in the Putative Branchpoint Sequence of Intron 4 on the Splicing within the Human Lecithin:cholesterol Acyltransferase Gene.*Journal of Biological Chemistry, 275* (24), 18079-18084.

Mazoyer, S., Puget, N., Perrin-Vidoz, L., Lynch, H. T., Serova-Sinilnikova, O. M., & Lenoir, G. M. (1998). A BRCA1 Nonsense Mutation Causes Exon Skipping. *The American Journal of Human Genetics, 62* (3), 713-715. doi:10.1086/301768

McConville, C. M., Stankovic, T., Byrd, P. J., McGuire, G. M., Yao, Q. Y., Lennox, G. G., & Taylor, M. R. (1996). Mutations associated with variant phenotypes in ataxia-telangiectasia. *American journal of human genetics, 59* (2), 320-330.

Mercer, T. R., Clark, M. B., Andersen, S. B., Brunck, M. E., Haerty, W., Crawford, J., . . . Mattick, J. S. (2015). Genome-wide discovery of human splicing branchpoints. *Genome Research, 25* (2), 290-303. doi:10.1101/gr.182899.114

Messiaen, L., Callens, T., De Paepe, A., Craen, M., & Mortier, G. (1997). Characterisation of two different nonsense mutations, C6792A and C6792G, causing skipping of exon 37 in the NF1 gene. *Human Genetics, 101* (1), 75-80. doi:10.1007/s004390050590

Millevoi, S., Bernat, S., Telly, D., Fouque, F., Gladieff, L., Favre, G., . . . Toulas, C. (2010). The c.5242C>A

BRCA1 missense variant induces exon skipping by increasing splicing repressors binding.*Breast Cancer Research and Treatment, 120* (2), 391-399. doi:10.1007/s10549-009-0392-3

Montalban, G., Bonache, S., Moles-Fernández, A., Gisbert-Beamud, A., Tenés, A., Bach, V., . . . Gutiérrez-Enríquez, S. (2019). Screening of &lt;em&gt;BRCA1/2&lt;/em&gt; deep intronic regions by targeted gene sequencing identifies the first germline &lt;em&gt;BRCA1&lt;/em&gt; variant causing pseudoexon activation in a patient with breast/ovarian cancer. *Journal of Medical Genetics, 56* (2), 63. doi:10.1136/jmedgenet-2018-105606

Montera, M., Piaggio, F., Marchese, C., Gismondi, V., Stella, A., Resta, N., . . . Mareni, C. (2001). A silent mutation in exon 14 of the APC gene is associated with exon skipping in a FAP family. *Journal of Medical Genetics, 38* (12), 863. doi:10.1136/jmg.38.12.863

Nazari, I., Tayara, H., & Chong, K. T. (2019). Branch Point Selection in RNA Splicing Using Deep Learning. *IEEE Access, 7* , 1800-1807. doi:10.1109/ACCESS.2018.2886569

Pagani, F., Buratti, E., Stuani, C., Bendix, R., Dörk, T., & Baralle, F. E. (2002). A new type of mutation causes a splicing defect in ATM.*Nature Genetics, 30* (4), 426-429. doi:10.1038/ng858

Paggi, J. M., & Bejerano, G. (2018). A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *RNA, 24* (12), 1647-1658. doi:10.1261/rna.066290.118

Raponi, M., Kralovicova, J., Copson, E., Divina, P., Eccles, D., Johnson, P., . . . Vorechovsky, I. (2011). Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6. *Human Mutation, 32* (4), 436-444. doi:10.1002/humu.21458

Rouleau, E., Lefol, C., Moncoutier, V., Castera, L., Houdayer, C., Caputo, S., . . . Lidereau, R. (2010). A missense variant within BRCA1 exon 23 causing exon skipping. *Cancer Genetics and Cytogenetics, 202* (2), 144-146. doi:10.1016/j.cancergencyto.2010.07.122

Sanz, D. J., Acedo, A., Infante, M., Durán, M., Pérez-Cabornero, L., Esteban-Cardeñosa, E., . . . Velasco, E. A. (2010). A High Proportion of DNA Variants of BRCA1 and BRCA2 Is Associated with Aberrant Splicing in Breast/Ovarian Cancer Patients. *Clinical Cancer Research, 16* (6), 1957-1967. doi:10.1158/1078-0432.ccr-09-2564

Sauna, Z. E., & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics, 12* , 683. doi:10.1038/nrg3051

Shamsani, J., Kazakoff, S. H., Armean, I. M., McLaren, W., Parsons, M. T., Thompson, B. A., . . . Spurdle, A. B. (2018). A plugin for the Ensembl Variant Effect Predictor that uses MaxEntScan to predict variant spliceogenicity. *Bioinformatics, 35* (13), 2315-2317. doi:10.1093/bioinformatics/bty960

Shapiro, M. B., & Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Research, 15* (17), 7155-7174.

Signal, B., Gloss, B. S., Dinger, M. E., & Mercer, T. R. (2018). Machine learning annotation of human branchpoints. *Bioinformatics, 34* (6), 920-927. doi:10.1093/bioinformatics/btx688

Sironi, M., Menozzi, G., Riva, L., Cagliani, R., Comi, G. P., Bresolin, N., . . . Pozzoli, U. (2004). Silencer elements as possible inhibitors of pseudoexon splicing. *Nucleic Acids Research, 32* (5), 1783-1791. doi:10.1093/nar/gkh341

Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frébourg, T., . . . Martins, A. (2016). Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools. *PLOS Genetics, 12* (1), e1005756. doi:10.1371/journal.pgen.1005756

Spier, I., Horpaopan, S., Vogt, S., Uhlhaas, S., Morak, M., Stienen, D., . . . Aretz, S. (2012). Deep intronic APC mutations explain a substantial proportion of patients with familial or early-onset adenomatous polyposis. *Human Mutation, 33* (7), 1045-1050. doi:10.1002/humu.22082

Sutton, I. J., Last, J. I. K., Ritchie, S. J., Harrington, H. J., Byrd, P. J., & Taylor, A. M. R. (2004). Adult-onset ataxia telangiectasia due to ATM 5762ins137 mutation homozygosity. *Annals of Neurology, 55* (6), 891-895. doi:10.1002/ana.20139

Svaasand, E., Engebretsen, L., Ludvigsen, T., Brechan, W., & Sjursen, W. (2015). A Novel Deep Intronic Mutation Introducing a Cryptic Exon Causing Neurofibromatosis Type 1 in a Family with Highly Variable Phenotypes: A Case Study. *Hereditary Genet, 4* (3). doi:10.4172/2161-1041.1000152

Théry, J. C., Krieger, S., Gaildrat, P., Révillion, F., Buisine, M.-P., Killian, A., . . . Tosi, M. (2011). Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. *European Journal Of Human Genetics, 19* (10), 1052-1058. doi:10.1038/ejhg.2011.100

Vallée, M. P., Di Sera, T. L., Nix, D. A., Paquette, A. M., Parsons, M. T., Bell, R., . . . Tavtigian, S. V. (2016). Adding In Silico Assessment of Potential Splice Aberration to the Integrated Evaluation of BRCA Gene Unclassified Variants. *Human Mutation, 37* (7), 627-639. doi:10.1002/humu.22973

van der Klift, H. M., Jansen, A. M. L., Steenstraten, N., Bik, E. C., Tops, C. M. J., Devilee, P., & Wijnen, J. T. (2015). Splicing analysis for exonic and intronic mismatch repair gene variants associated with Lynch syndrome confirms high concordance between minigene assays and patient RNA analyses. *Molecular Genetics & Genomic Medicine, 3* (4), 327-345. doi:10.1002/mgg3.145

Vaz-Drago, R., Custódio, N., & Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. *Human Genetics, 136* (9), 1093-1111. doi:10.1007/s00439-017-1809-4

Wang, G.-S., & Cooper, T. A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics, 8* , 749. doi:10.1038/nrg2164

Wang, Z., & Burge, C. B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA, 14* (5), 802-813. doi:10.1261/rna.876308

Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M., & Burge, C. B. (2004). Systematic Identification and Analysis of Exonic Splicing Silencers. *Cell, 119* (6), 831-845. doi:10.1016/j.cell.2004.11.010

Wappenschmidt, B., Becker, A. A., Hauke, J., Weber, U., Engert, S., Köhler, J., . . . Schmutzler, R. K. (2012). Analysis of 30 Putative BRCA1 Splicing Mutations in Hereditary Breast and Ovarian Cancer Families Identifies Exonic Splice Site Mutations That Escape In Silico Prediction. *PLoS ONE, 7* (12), e50800. doi:10.1371/journal.pone.0050800

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., . . . Frey, B. J. (2015). The human splicing code reveals new insights into the genetic determinants of disease.*Science, 347* (6218), 1254806. doi:10.1126/science.1254806

Yeo, G., & Burge, C. B. (2004). Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *Journal of Computational Biology, 11* (2-3), 377-394. doi:10.1089/1066527041410418

Zhang, C., Li, W.-H., Krainer, A. R., & Zhang, M. Q. (2008). RNA landscape of evolution for optimal exon and intron discrimination.*Proceedings of the National Academy of Sciences, 105* (15), 5797. doi:10.1073/pnas.0801692105

Zhang, K., Nowak, I., Rushlow, D., Gallie, B. L., & Lohmann, D. R. (2008). Patterns of missplicing caused by RB1 gene mutations in patients with retinoblastoma and association with phenotypic expression.*Human Mutation, 29* (4), 475-484. doi:10.1002/humu.20664

Zhang, Q., Fan, X., Wang, Y., Sun, M.-a., Shao, J., & Guo, D. (2017). BPP: a sequence-based algorithm for branch point prediction.*Bioinformatics, 33* (20), 3166-3172. doi:10.1093/bioinformatics/btx401

Zhang, X. H.-F., & Chasin, L. A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. *Genes & Development, 18* (11), 1241-1250. doi:10.1101/gad.1195304

TABLES

Table 1. *In silico* predictors of variant spliceogenicity

Examples of splice site, SRE, and BP prediction tools	**Examples of splice site, SRE, and BP prediction to**
Tool	**Motifs**
Human Splicing Finder (HSF)	5' and 3' splice site motifs, SRE, BP
MaxEntScan (MES)	5' and 3' splice site motifs
Splice Site Finder	5' and 3' splice site motifs
SPiCE	5' and 3' splice site motifs
MES-based VEP plugin	5' and 3' splice site motifs
ΔtESRseq	SRE
ΔHZ_{EI}	SRE
ΔΨ	SRE
EX-SKIP	SRE
HOT-SKIP	SRE
Branchpointer	BP
LaBranchoR	BP
SVM-BPfinder	BP
Branch Point Prediction (BPP)	BP
RNA Branch Point Selection (RNABPS)	BP
SRE algorithms in Human Splicing Finder v3.1	**SRE algorithms in Human Splicing Finder v3.1**
Type of Signal	**Prediction Algorithm**
ESE	HSF: 9G8, Tra2-β
	ESEfinder: SF2/ASF, SF2/ASF(IgM), SC35, SRp40, SRp5
	RESCUE ESE hexamers
ESS	HSF hnRNP-A1
	Sironi motifs
	ESS decamers
ESE and ESS	PESX Octamers
	ESR Sequences
	EIEs & IIEs Hexamers

Table 2. Variants prioritized as likely to disrupt exonic SREs in *BRCA1* $^{+}$

Variant	**Predicted amino acid change**	**Splicing assay in literature**	**Mean RNA score (Findlay et al., 2018)**	**ClinVar Classification (28-Feb-2020)**
c.196A>T	p.(Asn66Tyr)	No record	-3.836201203	N/A
c.253G>A	p.(Glu85Lys)	No record	-2.588684303	N/A
c.257T>G	p.(Leu86Arg)	No record	-2.109971236	N/A
c.261G>T	p.(Leu87Phe)	No record	-2.338647117	N/A
c.4938C>A	p.(Val1646=)	No record	-2.200191231	N/A
c.5007C>T	p.(Ala1669=)	No record	-2.172296931	Likely benign
c.5044G>A	p.(Glu1682Lys)	No record	-3.922769452	Benign
c.5044G>C	p.(Glu1682Gln)	No record	-2.077097565	Uncertain significance
c.5045A>T	p.(Glu1682Val)	No record	-3.19611471	Uncertain significance
c.5046A>T	p.(Glu1682Asp)	No record	-2.279964014	N/A

Variant	**Predicted amino acid change**	**Splicing assay in literature**	**Mean RNA score (Findlay et al., 2018)**	**ClinVar Classification (28-Feb-2020)**
c.5047G>A	p.(Glu1683Lys)	No record	-2.845610337	N/A
c.5047G>C	p.(Glu1683Gln)	No record	-2.665637918	N/A
c.5048A>G	p.(Glu1683Gly)	No record	-2.717632186	N/A
c.5051C>T	p.(Thr1684Ile)	No record	-2.256440077	N/A
c.5054C>A	p.(Thr1685Asn)	No record	-2.926076348	N/A
c.5054C>G	p.(Thr1685Ser)	No record	-3.51332298	N/A
c.5066T>G	p.(Met1689Arg)	No record	-2.582200979	Pathogenic/Likely pathogenic
c.5078C>T	p.(Ala1693Val)	No record	-2.19424566	Uncertain significance
c.5080G>A	p.(Glu1694Lys)	YES (Houdayer et al., 2012); co-located variants c.5080G>T^{++} (Goina, Skoko, & Pagani, 2008; Mazoyer et al., 1998) and c.5078_5080del (Campos et al., 2003) lead to exon skipping	-2.93457224	Uncertain significance
c.5123C>G	p.(Ala1708Gly)	No record; co-located variant c.5123C>A§ leads to minor exon skipping (Millevoi et al., 2010; Sanz et al., 2010)	-3.080698027	N/A
c.5127A>G	p.(Gly1709=)	No record	-2.143205298	N/A
c.5130A>G	p.(Gly1710=)	No record	-2.826582097	N/A
c.5137G>T	p.(Val1713Leu)	No record	-4.026139955	N/A
c.5430G>C	p.(Val1810=)	No record	-4.027269569	N/A
c.5434C>G	p.(Pro1812Ala)	YES (Gaildrat et al., 2010; Théry et al., 2011)	-4.854622646	Pathogenic/Likely pathogenic
c.5441C>G	p.(Ala1814Gly)	No record	-5.847427008	N/A
c.5444G>C	p.(Trp1815Ser)	No record	-2.530289371	Uncertain significance
c.5445G>C	p.(Trp1815Cys)	No record	-5.45381894	N/A
c.5453A>G	p.(Asp1818Gly)	YES (Rouleau et al., 2010)	-4.429910618	Conflicting: Likely pathogenic(2); Pathogenic(3); Uncertain significance(2)
c.5472T>G	p.(Ile1824Met)	No record	-2.325563519	N/A

Variant	Predicted amino acid change	Splicing assay in literature	Mean RNA score (Findlay et al., 2018)	ClinVar Classification (28-Feb-2020)
c.5528C>A	p.(Ala1843Glu)	No record	-2.158302832	Uncertain significance
c.5546A>C	p.(Glu1849Ala)	No record	-2.541769038	N/A
c.5546A>T	p.(Glu1849Val)	No record	-2.880877731	N/A

[+] See Figure 2 for overview of prioritization process. cDNA numbering is based on NM_007294.3 transcript.

[++] c.5080G>T was assayed in Findlay et al. (2018): non-functional, depleted in mRNA (RNA score = -2.5); predicted nonsense thus excluded from list.

[§] c.5123C>A was assayed in Findlay et al. (2018): non-functional, not depleted in mRNA (RNA score = -0.55); aberrant transcript was observed as faint band in splicing assay of Sanz et. al (2010).

Table 3. MaxEntScan scores of native splice sites of*BRCA1* exons assessed by functional assays of Findlay et al. (2018)

Exon	Acceptor	Donor	Splice site strength (acceptor-donor)[+]	Mapped exonic SRE
2	4.9	10.65	LOW-HIGH	No
3	7.05	10.08	MOD-HIGH	No
5	8.19	7.84	MOD-MOD	Yes
6	4.84	8.46	LOW-MOD	Yes
16	10.02	5.91	HIGH-LOW	Yes
17	6.69	7.48	MOD-MOD	Yes
18	8.96	7.96	HIGH-MOD	Yes
19	8.78	11.08	HIGH-HIGH	No
20	9.36	9.06	HIGH-HIGH	No
21	13.07	10.77	HIGH-HIGH	No
22	8.67	9.49	HIGH-HIGH	No
23	4.86	9.33	LOW-HIGH	Yes
24	9.53	-	HIGH-	Yes

[+] LOW: score < 6.2, MODERATE: 6.2 [?] score < 8.5, HIGH: score [?] 8.5

Table 4. Evaluation of ΔtESRseq, ΔHZEI, and HOT-SKIP performance in predicting putative SRE-disrupting spliceogenic variants[+]

Bioinformatic tools (cut-off score)	Positive controls	Positive controls	Negative controls	Negative controls	Negative controls (secondary analysis) ++	Negative controls (secondary analysis) ++	Sensitivity	Specificity	Sp (se on an
	True Positive	False Negative	False Positive	True Negative	False Positive	True Negative			
ΔτΕΣΡσεχ (< -0.75)	24	9	51	199	30	220	73%	80%	88

Bioinformatic tools (cut-off score)	Positive controls	Positive controls	Negative controls	Negative controls	Negative controls (secondary analysis) ++	Negative controls (secondary analysis) ++	Sensitivity	Specificity	Sp (se on an
ΔHZ$_{\mathbf{EI}}$ ([;] -5)	25	8	45	205	23	227	76%	82%	91
HOT-SKIP (alt/wt > 1)	15	18	56	194	37	213	45%	78%	85

[+] See Figure 2 for overview of prioritization process to detect 33 putative SRE-disrupting variants, and 250 non-spliceogenic negative control variants; and Supplementary Table 3 for detailed scores and cut-offs.

[++] For secondary analysis, false positive variants located in exons with no mapped SREs were designated as true negatives.

Table 5. Spliceogenic variants in hereditary cancer genes reported to abrogate a BP site or activate a pseudoexon

Gene	**HGVS**
Spliceogenic variants experimentally validated or inferred to abrogate a BP site[+]	**Spliceogenic variants ex**
BRCA1	LRG_292t1:c.135-18T>G
BRCA1	LRG_292t1:c.135-27T>A
BRCA1	LRG_292t1:c.4358-31_4358-
BRCA1	LRG_292t1:c.4358-33T>G
BRCA1	LRG_292t1:c.5153-26A>G
BRCA1	LRG_292t1:c.5153-26A>T
BRCA1	LRG_292t1:c.5153-27_5153-
BRCA1	LRG_292t1:c.5407-25T>A
MLH1	LRG_216t1:c.1732-19T>A
RAD51C	LRG_314t1:c.572-23_572-20
RB1	LRG_517t1:c.2326-26A>C
RB1	LRG_517t1:c.2326-26A>G
XPC	LRG_472t1:c.413-9T>A
XPC	LRG_472t1:c.413-24A>G
Deep intronic variants that activate a pseudoexon[++]	**Deep intronic variants t**
APC	LRG_130t1:c.532-941G>A
APC	LRG_130t1:c.1408+731C>T
APC	LRG_130t1:c.1408+735A>T
ATM	U82828.1(ATM_v001):c.123
ATM	LRG_135t1:c.2839-581_2839
ATM	U82828.1(ATM_v001):c.399
ATM	LRG_135t1:c.5763-1050A>C
BRCA1	LRG_292t1:c.4185+4105C>
BRCA2	LRG_293t1:c.6937+594T>C
MSH2	LRG_218t1:c.212-478T>G
NF1	LRG_214t1:c.288+1137C>T

Gene	HGVS
NF2	LRG_511t1:c.1447-233T>A[§]
RB1	LRG_517t1:c.2490-1398A>G

[+] Only two *XPC* variants were experimentally validated to affect actual BP sites. The remaining spliceogenic variants were inferred to abrogate a BP site due to nucleotide position, i.e. located within the -18 to -44 nucleotide BP window, and/or positive BP site prediction.

[++] All deep intronic variants listed here create a new 5' splice site or activate a cryptic 5' splice site except LRG_135t1:c.2839-581_2839-578del, which introduces an intron-splicing processing element.

[§] Reported as c.1236-405C>T, in intron 11, with NM_000051.3 as reference transcript. However, the first base of exon 12 in the NM_000051.3 transcript is c.1803. Mutalyzer (https://mutalyzer.nl/) maps c.1236-405C>T in intron 11 of U82828.1 reference sequence.

[¶] Authors did not indicate the reference sequence and HGVS nomenclature. The published pseudoexon sequence maps to intron 18 of LRG_135. This variant was described as *ATM*IVS20-579_IVS20-576delGTAA by Coutinho et al. (2005).

[++] Reported as U82828.1:g.75117A>G, IVS28-159A>G

[++++] Reported as 5762ins137 caused by A>G variant. Published pseudoexon sequence maps to intron 38 of LRG_135. This variant was described as *ATM* IVS40-1126G>A by Coutinho et al. (2005).

[SSSS] Reported as NG_009057.1:g.74409T>A, NM_000268.3:c.1447-240T>A. However, inspection of the NG_009057.1 sequence (16-NOV-2019 version) revealed two sequence errors in the published illustration of the pseudoexon with flanking intronic regions: G deletion (+1 position) and A insertion (+4 position). Based on the publication's illustration of variant location and reported sequence of pseudoexon boundaries, the variant is located at g.74408 in NG_009057.1, 233 bp upstream of exon 14.

[PP] Reported as IVS23-1398A>G

FIGURE LEGENDS

Figure 1. Splicing motifs for exon recognition by the spliceosome. The acceptor (3') and donor (5') splice site motifs are shown. The U1 and U2 small nuclear ribonucleoprotein complexes recognize the donor splice site and the branchpoint (BP) site, respectively, while the U2AF proteins recognize the acceptor splice site and polypyrimidine tract located between the BP and 3' splice site. Diverse sets of splicing regulatory elements refine exon recognition and regulate alternative splicing. Some exonic splicing enhancers (ESEs) bind Serine/Arginine-rich (SR) proteins and stabilize the binding of U2AF to promote splice site usage. Exonic splicing silencers (ESSs) bind heterogeneous nuclear ribonucleoproteins (hnRNPs) to inhibit splice site usage. Intronic splicing regulatory elements are not shown. Donor and acceptor splice site motif figures adapted from Cartegni, Chew, and Krainer (2002).

Figure 2. Flowchart of *BRCA1* variant selection for exonic SRE mapping and evaluation of SRE predictor performance.

[+] *BRCA1* variants were from saturation genome editing experiments (SGE) of Findlay et al. (2018) covering exons 2, 3, 5, 6 and 16-24. Cut-off scores selected for determining function class and mRNA depletion are drawn from their Supplementary Table 1. For our selection of negative controls, we set a conservative RNA score cut-off [?] 0 for variants not depleted in mRNA, as some variants experimentally confirmed to have weak to moderate splicing effects (from literature and unpublished splicing assay data) were observed to have RNA scores between -0.5 and -2 (data not shown).

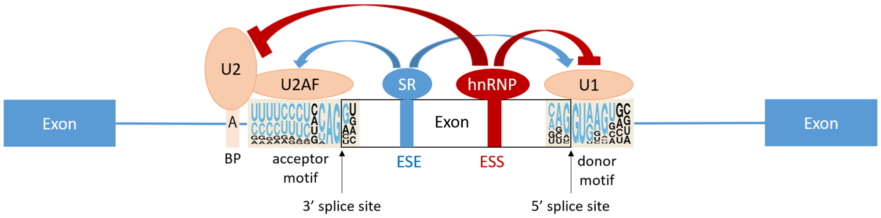
Selection of variants outside donor and acceptor splice site motifs and prediction of *de novo* splice site gain were done through MES-based Variant Effect Predictor plugin using the thresholds and decision flowchart described in Shamsani et al. (2018).

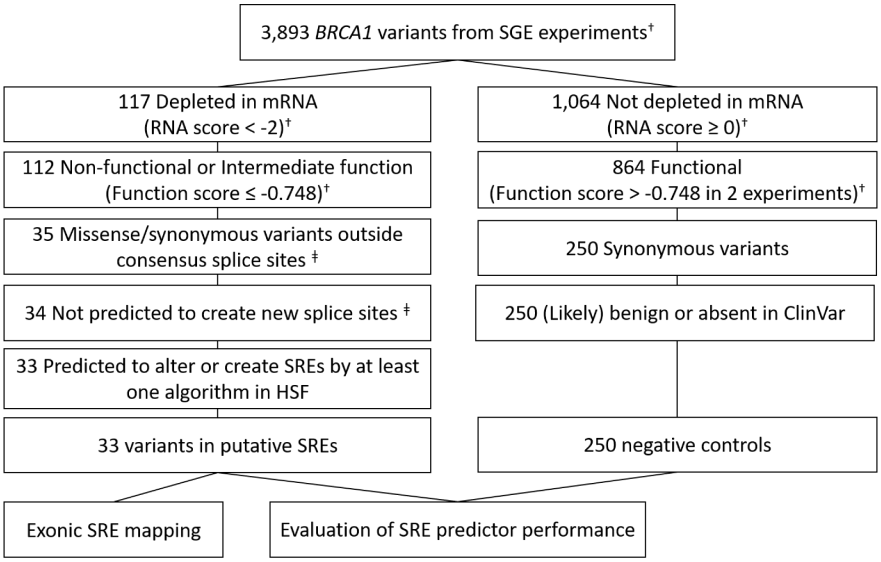
Figure 3. Map of putative exonic SREs (blue) in *BRCA1* as inferred from analysis of published functional data of Findlay et al. (2018) and Human Splicing Finder SRE predictions. Findlay et al. (2018) assessed exons 2, 3, 5, 6 and 16-24. Exons 2, 3, and 19-22 did not harbor variants that passed the selection criteria described in Figure 2. The cDNA positions are indicated in the map, and the locations of 33 putative SRE-disrupting variants are bolded and underlined. Longer blue regions reflect several overlapping SRE sequences that extend across the exon. The grey area at the 5' end of exon 16 represents a region without functional data from Findlay et al. (2018).

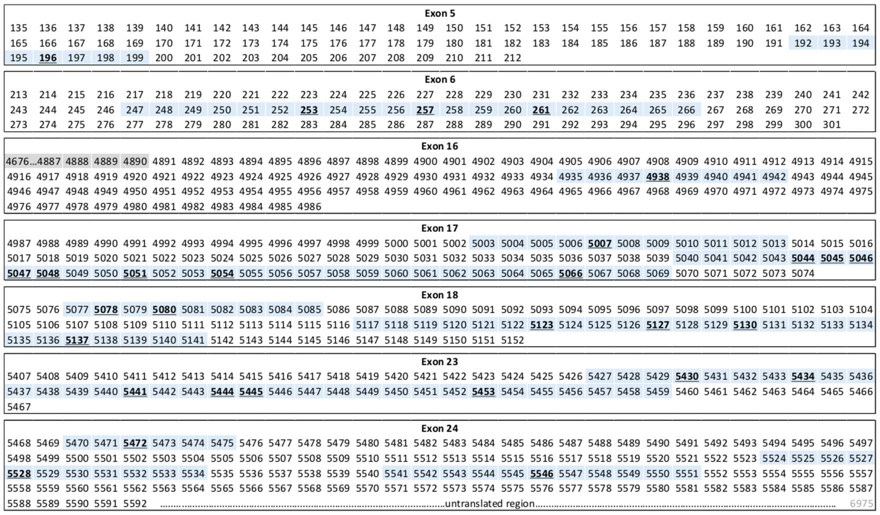
Figure 4. Prioritization model to identify SRE-disrupting variants for splicing assay. For exonic variants outside of the donor and acceptor splice site motifs, the first step is to filter out variants that are predicted to lead to *de novo* splice site gain using the MES algorithm. Then MES is used to determine the native splice site scores, to identify exons most likely to harbor ESEs. Variants in exons with low to moderate donor score (< 8.5), or with high donor MES score ([?] 8.5) *and* low acceptor score (< 6.2), are then selected as eligible for two-step SRE analysis. Positive calls in HSF are further analyzed using ΔHZ_{EI} to minimize the number of false positives.

Supplementary Figure 1. Map of putative exonic SREs in*BRCA1* exons 5, 6, 16-18, 23 and 24. The map is based on the combined analysis of results from the functional assay of Findlay et al. (2018), and SRE predictions from the algorithms in Human Splicing Finder v3.1. Variants that passed the selection criteria (see main text and Figure 2) are shown. Putative SRE locations are highlighted in blue. cDNA positions in green font indicate locations of negative control variants outside of putative SREs with false positive HSF SRE predictions. cDNA positions in red font indicate locations of negative control variants within putative SREs but with false positive HSF SRE predictions.

Supplementary Figure 2. Snapshot of Human Splicing Finder results webpage for *MLH1* LRG_216t1:c.1559-1732A>T.The Information Theory model of Caminsky et al. (2016) predicted that LRG_216t1:c.1559-1732A>T creates a new acceptor and activates a pseudoexon due to the presence of a downstream pre-existing cryptic donor, but our HSF analysis revealed a s trong putative exonic splicing silencer (PESS) cluster within 30 nucleotides upstream of the donor site of the pseudoexon (encircled in red), which potentially inactivates this cryptic donor.




3,893 *BRCA1* variants from SGE experiments†
117 Depleted in mRNA (RNA score < -2)†
1,064 Not depleted in mRNA (RNA score ≥ 0)†
112 Non-functional or Intermediate function (Function score ≤ -0.748)†
864 Functional (Function score > -0.748 in 2 experiments)†
35 Missense/synonymous variants outside consensus splice sites ‡
250 Synonymous variants
34 Not predicted to create new splice sites ‡
250 (Likely) benign or absent in ClinVar
33 Predicted to alter or create SREs by at least one algorithm in HSF
33 variants in putative SREs
250 negative controls
Exonic SRE mapping
Evaluation of SRE predictor performance



Exon 5

135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164
165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194
195	**196**	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212												

Exon 6

213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242
243	244	245	246	247	248	249	250	251	252	**253**	254	255	256	**257**	258	259	260	**261**	262	263	264	265	266	267	268	269	270	271	272
273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	

Exon 16

4676…4887		4888	4889	4890	4891	4892	4893	4894	4895	4896	4897	4898	4899	4900	4901	4902	4903	4904	4905	4906	4907	4908	4909	4910	4911	4912	4913	4914	4915
4916	4917	4918	4919	4920	4921	4922	4923	4924	4925	4926	4927	4928	4929	4930	4931	4932	4933	4934	4935	4936	4937	**4938**	4939	4940	4941	4942	4943	4944	4945
4946	4947	4948	4949	4950	4951	4952	4953	4954	4955	4956	4957	4958	4959	4960	4961	4962	4963	4964	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974	4975
4976	4977	4978	4979	4980	4981	4982	4983	4984	4985	4986																			

Exon 17

4987	4988	4989	4990	4991	4992	4993	4994	4995	4996	4997	4998	4999	5000	5001	5002	5003	5004	5005	5006	**5007**	5008	5009	5010	5011	5012	5013	5014	5015	5016
5017	5018	5019	5020	5021	5022	5023	5024	5025	5026	5027	5028	5029	5030	5031	5032	5033	5034	5035	5036	5037	5038	5039	5040	5041	5042	5043	**5044**	**5045**	**5046**
5047	**5048**	5049	5050	**5051**	5052	5053	**5054**	5055	5056	5057	5058	5059	5060	5061	5062	5063	5064	5065	**5066**	5067	5068	5069	5070	5071	5072	5073	5074		

Exon 18

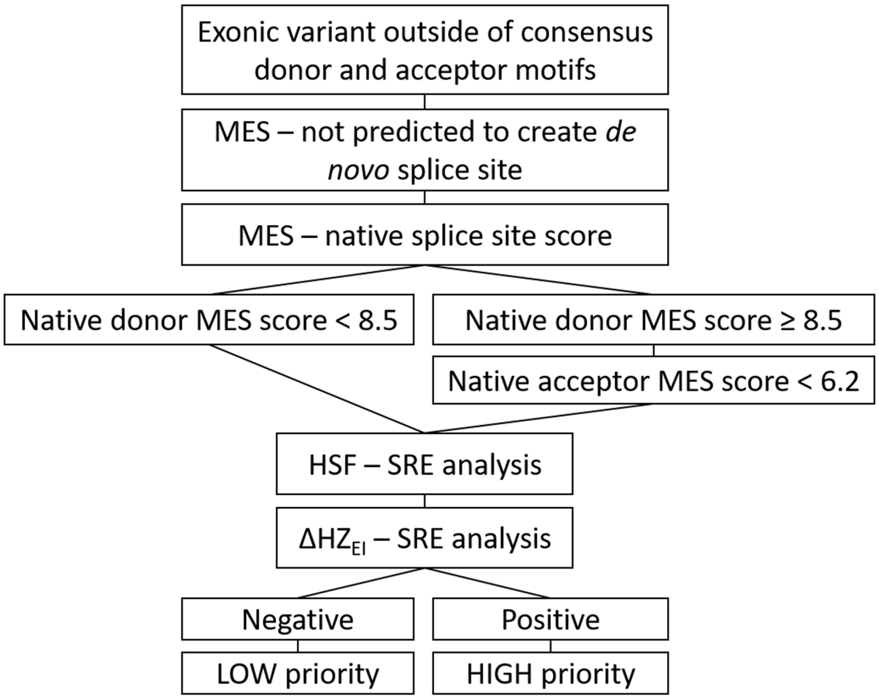
5075	5076	5077	**5078**	5079	**5080**	5081	5082	5083	5084	5085	5086	5087	5088	5089	5090	5091	5092	5093	5094	5095	5096	5097	5098	5099	5100	5101	5102	5103	5104
5105	5106	5107	5108	5109	5110	5111	5112	5113	5114	5115	5116	5117	5118	5119	5120	5121	5122	**5123**	5124	5125	5126	**5127**	5128	5129	**5130**	5131	5132	5133	5134
5135	5136	**5137**	5138	5139	5140	5141	5142	5143	5144	5145	5146	5147	5148	5149	5150	5151	5152												

Exon 23

5407	5408	5409	5410	5411	5412	5413	5414	5415	5416	5417	5418	5419	5420	5421	5422	5423	5424	5425	5426	5427	5428	5429	**5430**	5431	5432	5433	**5434**	5435	5436
5437	5438	5439	5440	**5441**	5442	5443	**5444**	**5445**	5446	5447	5448	5449	5450	5451	5452	**5453**	5454	5455	5456	5457	5458	5459	5460	5461	5462	5463	5464	5465	5466
5467																													

Exon 24

5468	5469	5470	5471	**5472**	5473	5474	5475	5476	5477	5478	5479	5480	5481	5482	5483	5484	5485	5486	5487	5488	5489	5490	5491	5492	5493	5494	5495	5496	5497
5498	5499	5500	5501	5502	5503	5504	5505	5506	5507	5508	5509	5510	5511	5512	5513	5514	5515	5516	5517	5518	5519	5520	5521	5522	5523	5524	5525	5526	5527
5528	5529	5530	5531	5532	5533	5534	5535	5536	5537	5538	5539	5540	5541	5542	5543	5544	5545	**5546**	5547	5548	5549	5550	5551	5552	5553	5554	5555	5556	5557
5558	5559	5560	5561	5562	5563	5564	5565	5566	5567	5568	5569	5570	5571	5572	5573	5574	5575	5576	5577	5578	5579	5580	5581	5582	5583	5584	5585	5586	5587
5588	5589	5590	5591	5592	……………………untranslated region……………………																								6975


Exonic variant outside of consensus donor and acceptor motifs
MES – not predicted to create *de novo* splice site
MES – native splice site score
Native donor MES score < 8.5
Native donor MES score ≥ 8.5
Native acceptor MES score < 6.2
HSF – SRE analysis
ΔHZ_{EI} – SRE analysis
Negative
Positive
LOW priority
HIGH priority