Background splicing and genetic disease

Diana Alexieva
Imperial College London
Yi Long
Imperial College London
Rupa Sarkar
Imperial College London
Hansraj Dhayan
Imperial College London
Emmanuel Bruet
Imperial College London
Robert Winston
Imperial College London
Igor Vorechovsky
University of Southampton
Leandro Castellano
University of Sussex
Nick Dibb (n.dibb@imperial.ac.uk)
Imperial College London

Research Article

Keywords: background splicing, splice site mutations, cryptic splice sites, exon skipping, pseudoexons, recursive splicing, spliceosomal mutations, splicing therapy, BRCA1, BRCA1, DMD

Posted Date: October 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-92665/v1

Version of Record: A version of this preprint was published at RNA Biology on October 15th, 2020. See the published version at https://doi.org/10.1080/15476286.2021.2024031.

Abstract

We report that low level background splicing by normal genes can be used to predict the likely effect of splicing mutations upon cryptic splice site activation and exon skipping, with emphasis on the DBASS databases, BRCA1, BRCA2 and DMD. In addition we show that background RNA splice sites are also involved in pseudoexon formation, recursive splicing and aberrant splicing in cancer. We discuss how background splicing information might inform splicing therapy.

Introduction

We previously established that cryptic splices sites (css) are already active, albeit at very low levels, in normal genes. We did this by using EST data to identify rare splice sites and then compared their positions to known css that are activated in human disease (1). However, this approach was limited to a minority of genes for which there was sufficient EST sequence data. Since that time a large amount of RNA-sequencing data has been deposited, which we reasoned would strongly increase the power of css prediction. In support of this, RNA sequencing studies have shown that normal splicing is accompanied by a background of low level or noisy splicing between a large number of hidden splice sites within introns and exons (2).

The snaptron database (http://snaptron.cs.jhu.edu/) lists all of the RNA-seq reads from over 70,000 human samples that were most probably generated by splicing (3). Here we compare the snaptron database to cryptic splice site and exon skipping databases (4,5) and conclude that background splicing reads give a good indication of the likely effect of splicing mutations upon exon skipping as well as css activation. Further comparisons show that background splice sites in normal human genes are also informative about pseudoexon formation, recursive splicing, aberrant splicing in cancer and splicing therapy.

Materials And Methods

References to experimental reports of splice site mutations that cause aberrant splicing of BRCA1, BRCA2 and DMD were obtained from the database of aberrant splice sites (DBASS) the human genome mutation database (HGMD), the Leiden Open Variation Database online (LOVD) and by searching Pubmed (4,6,7).

We also analysed splicing mutations that cause a wide range of medical syndromes by using DBASS and exon skipping databases (4,5). Databases of aberrant splicing in cancer and of recursive splicing were also analysed and are described in the text. We used the HGMD http://www.hgmd.cf.ac.uk/ac/index.php and LOVD websites https://databases.lovd.nl/shared/genes (6,7) to clarify the exon nomenclature used in original reports and the BLAT tool (8) from the UCSC website http://genome.ucsc.edu/ (9) to obtain genome reference numbers for relevant splice sites.

We then compared the above experimental databases of aberrant splicing to the Snaptron database (3), which lists all RNA sequences from over 70 000 human samples that were most probably generated by splicing. Snaptron lists major splicing events between intronic ss and alternative splice sites (ass) but there are many more examples of splicing events with much lower read numbers that are referred to here as background splicing (bss) events. Background splicing may occur between 5' or 3' intron ss with bss that are not the normal intron partner (Fig 1B,C) but also occurs within introns and exons or across these boundaries. The splice sites listed in snaptron are all canonical, GT or GC for 5'ss or AG for 3'ss. Possible non-canonical ss were filtered from snaptron in order to avoid including mRNA deletions that were not generated by splicing (3).

The experimental datasets were compared to the snaptron database by a method that is explained in Figure 1 and Table 1 for BRCA1 but also applies to all other analyses. Fig 1D illustrates that mutations of intron splice sites typically activate cryptic splice sites (css) or cause exon skipping or both (4,5). Cryptic splice sites are dormant splice sites that are activated to very high levels following the mutation of intronic splice sites, and usually occur within 1000bp of the mutated ss (4). DBASS also lists mutations that create de novo splice sites (also known as de novo css) or pseudoexons. De novo ss are new ss that are directly generated by mutations and then compete for splicing with an intron ss. De novo mutations can exist on their own or may also activate pseudoexons (Figure 2A,B).

Snaptron has four different RNA sequencing databases that can be analysed. SRAv1 (hg19) and SRAv2 (hg38) are from the sequencing read archive at NCBI and contain 41 and 83M splice junctions identified by sequencing, respectively There are also two smaller databases TCGA (hg38) and GTEx (hg38) with 37 and 29M junctions (3).

We largely analysed SRAv1 because this was the first available database and it also helped with comparisons to other hg19 databases. SRAv1 is still available but has now been superseded by SRAv2. We downloaded snaptron data for individual genes and arranged it on worksheets in a manner that allowed us to identify background splicing events that might be activated by splicing mutations, as illustrated in Fig 1D. A protocol for this is given below for BRCA1 but is applicable to all genes.

BRCA1 splicing data was downloaded from snaptron by using the link http://snaptron.cs.jhu.edu/srav1/snaptron?regions=brca1. RNA splicing data for any other gene can be obtained by changing brca1 to the required gene name ie http://snaptron.cs.jhu.edu/srav1/snaptron?regions=dmd. To access the other spliced RNA databases of snaptron srav1 can be changed to srav2, gtex or tcga. Downloaded snaptron data was then selected, copied and pasted into the spreadsheet LibreOffice Calc. We chose paste special, unformatted text, then UT-16 and tab options. Sheet 1 in the spreadsheet was copied to two further worksheets. For sheet three we re-ordered the data by selecting the highest sequencing reads in column O (with the extended data option). We then chose and copied the top rows with the highest sequencing reads into sheet 4 (the more rows that are chosen from sheet 3 the greater the number of minor alternative splicing events that can be seen). In sheet 4 it is important to take note of the gene direction along the chromosome indicated by – or + in column G (strand). If + this

means that the 5'ss are listed in column D (and 3'ss in column E) however, if column G is - then the 5'ss are listed in column E and the 3'ss in column D. The splice sites were then ordered in a 5' to 3' direction in sheet 4 either by ordering from low to high for column D (for those genes on the + strand) or from high to low for column E (for those genes on the – strand). Worksheets 1 and 2 were used separately to order the large numbers of 5'ss or 3'ss from snaptron in a 5' to 3' direction. By this means 5' or 3' intronic ss and major alternative ss can be identified from sheet 4 (see Fig 1A) and all of their partner background ss can be identified from either sheet 1 or from sheet 2 (see Figure 1B,C). Occasionally, column G of worksheet 4 contained rows with both – and + values, due to overlapping transcripts from both strands. The required transcript usually has the greatest number of reads and can also be identified from the UCSC genome browser.

It is also useful to identify all of the partners of the mutated splice site at the start of the analysis in case this site is normally involved in alternative splicing. Major alternative splicing can be seen in sheet 4 of the spreadsheet and minor alternative splicing can identified by looking at the read numbers for the partner sites of the mutated splice site (in sheet 1 or 2). If there is no or little alternative splicing the analysis proceeds as outlined in Figure 1D. If there are also alternative splice sites then these should be analysed alongside the main partner ss, as outlined in figure 1D, because these alternative splice sites may also participate significantly in aberrant splicing events. A splicing mutation of LAMP2 is a good example of this (Appendix 1).

By unrestricted search (see text) we mean searching snaptron for reads for the genome reference number of a putative splice site irrespective of its splice partners.

Results

BRCA1

We initially analysed BRCA1 as proof-of-principle because its mutational landscape in cancer is well described and includes splicing mutations that have been repeatedly analysed (10,11). We first downloaded the RNA-seq data for BRCA1 from snaptron into a spreadsheet (see Materials and Methods). This spreadsheet lists over 6000 differently spliced transcripts of BRCA1, although the large majority of these are background splicing events that are only supported by very low reads. Fig 1A lists the splicing events with the highest reads, these include intron removal and major alternative splicing events. At least 8 isoforms of BRCA1 have been identified (12) and the major alternative splice sites (but not the isoforms) can easily be identified in Figure 1A (shaded).

Figures 1B, C and D illustrate how background RNA sequencing data can be used to predict css or exon skipping events that are likely to result from splicing mutations of BRCA1 or any gene. Fig 1B examines the theoretical effect of mutation of the BRCA1 intron 5'ss 41222944 (shown in red). This might be expected to enhance the use of alternative 5'ss partners for the non-mutated 3'ss 41219713, as illustrated in figure 1D. Figure 1B lists the 5'ss partners for 3'ss 41219713 that have been identified in snaptron. As expected there are a large number of reads (148299) for splicing between 3'ss 41219713

and its normal 5'ss partner 41222944 of BRCA1 (blue shading). Other 5'ss partners of the 3'ss 41219713 are also used but at much lower background levels in wild type BRCA1 transcripts. These include single and multiple exon skipping events (yellow shading) between the 3'ss 41219713 and the 5'ss of other upstream introns (compare Figure 1A and B). In addition there are 2 reads for a rare splicing event between 3'ss 41219713 and an exonic 5'ss that is located -93 bases upstream of the normal 5'ss 41222944 and further low level reads for seven background 5'ss that are located downstream within the intron.

Mutation of BRCA1 5'ss 41222944 is known to activate a css at +69 (13-15) or at +65 (16) These two css exactly match the bss with the most supporting reads (Figure 1B, red shading). The background splicing information is therefore a good match to the slightly different experimental results of both groups.

Similarly, Figure 1C examines the possible effect of mutation of the 3'ss 41203135 (red shading) by showing all of the splicing events involving its normal partner 5'ss 41209068, as illustrated in Figure 1D. Mutation of 3'ss 41203135 is known to activate exon skipping between the normal partner 5'ss 41209068 and the downstream intronic 3'ss 41201212 plus weaker activation of the 3'css 41203127 (13,14). Figure 1C shows that these two splicing events have the most reads of the background splicing events involving the 5'ss 41209068 of the wild type BRCA1 gene.

The data for Figures 1B and C is summarised in Table 1 (rows 13 and 32), which includes all mutations of the splice sites of BRCA1. Fig S1 shows this data in full in the same format as Fig 1B, C. From the literature we identified seventeen different css that are activated by mutations of the indicated BRCA1 splice sites and Table 1 shows that 15 of these css exactly match bss of wild type BRCA1, the two exceptions are shaded in column 3 and discussed in Table S1, which also provides references. Twelve of the15 bss that match css have the highest reads of all candidate bss, as listed under column 4 and as illustrated in Figs 1B, C. Sites that are candidates for css activation are defined here as bss within 1000 bases of the intronic ss that is mutated (see Discussion).

Many of the splice site mutations of BRCA1 in Table 1 activate exon skipping rather than css and eight of the splice site mutations do both (Table 1, column 2). The ratio of css reads to exon skip reads from the background RNA sequencing data (Table 1, columns 5,6) appears to correlate with the experimental finding of whether splice site mutations activate css or exon skipping. There are six exceptions to this that are shaded as pairs in columns 5 & 6 and are discussed (Table S1). Also shaded are some possible false positive bss reads for both css activation (column 5 rows 5, 24, 31 and 35) and for a double exon skip (column 7 row 16), see Table S1 and Discussion. This data (Table 1) suggests that the effect of splice site mutations upon css activation and even exon skipping can be inferred from background splicing data. In order to test this hypothesis we undertook analyses of further experimental databases that include over 300 medical syndromes caused by splice site mutations.

DBASS, BRCA2 and DMD

We next compared the snaptron database with the database of aberrant splice sites (DBASS). DBASS lists the experimental results for splicing mutations that cause a wide range of human genetic diseases (4). Table 2 is a summary of Table S2, which is an index all of the splicing mutations in DBASS, and shows that the DBASS mutations are subdivided into those that activate aberrant 5' or 3' splice sites (DBASS5 and DBASS3) and that the most common mutations activate css but can also generate de novo css or pseudoexons.

We first compared the DBASS5 experimental results for 5' css activation with the snaptron RNA splicing data. Table S2 shows how 199 of the 459 mutations in DBASS5 that activate css were systematically chosen to cover every listed medical syndrome. We generated similar tables of background splicing to those illustrated in Fig 1A,B,C for each of the 199 mutations and compared these with the experimental results. Each analysis is summarised in single rows in Table S3 sheet 1. The background splicing tables (see Fig 1B or C) are not shown but the key results are recorded in Table S3 and the raw data can easily be generated as described in Materials and methods. Table 3 row DBASS5 summarises Table S3 sheet 1 and shows that 201 out of 237 (85%) of the 5'css identified by experiment (some mutations activate more than one css) exactly match bss in snaptron and are therefore already in use at low levels by normal genes. 150 out of 201 (75%) of the bss that match the position of css have the greatest number of supporting reads compared to other bss (Table 3, S3). Similar results were found for the analysis of the 3'css listed in DBASS3 where 97 out of 110 (87%) 3'css matched bss in snaptron (Tables 3, S2, S3).

The reason why 15% or so of the experimentally identified 5' css or 3'css did not match a background ss was usually because there were no background ss reads for comparison (Table S3). Where background ss data was available, we found that background ss did not match the experimentally reported 5' or 3' css in only 2 to 3% of cases, listed as poor matches in Tables 3 and S3. Table 3 also includes summaries for similar analyses of BRCA1 (Table 1), BRCA2 and DMD (Tables S4, S5). DBASS5* and DBASS3* of Table 3 summarise an analysis of a subcategory of css that are activated by mutations that occur outside the highly conserved regions of the normal 5' or 3'ss (Tables 2, S2, S6). The activated css of DBASS5* and DBASS3* tend to match bss with particularly high reads (Table S6, see Discussion). Overall the very large majority of css originate from bss (see Discussion) and usually the bss that is activated is the one with the most reads relative to other bss (Table 3).

Exon skipping

We next asked whether background splicing data can indicate whether splice site mutations might cause exon skipping rather than css activation. Some of the papers referenced in DBASS report whether or not exon skipping accompanied css activation (Table S3, column N). Table 4 column 1 summarises that there are 39 reports of both exon skipping and css activation and 71 reports of css activation only for the 5'ss mutations analysed in Table S3. For the reports of css activation only, the total number of background single exon skip reads from the 71 examples is 6621, which is much smaller than the total

background skip reads (251128) from the 29 reports of both css and skip activation, so confirming the correlation seen for Table 1. Similar results were found for DBASS3 (Table 4).

Table 4 also summarises an analysis of a second database of splicing mutations (Tables S7, S8) that generally cause exon skipping rather than css activation (5). Table 4 shows that we analysed 79 experimental reports of 5'ss mutations that cause exon skipping only. Of these, 71 examples have higher background splicing reads for exon skipping than reads for potential css (background ss within 1000 bases of the intronic ss). Conversely, the 71 experimental reports in DBASS5 of 5'ss mutations that only caused css activation (column 1, line 5) had higher reads for the css than for background exon skipping in 60 out of 71 examples. Similar results are found by comparing the 64 examples of 3'ss mutations that cause exon skipping only with the 18 examples of 3'ss mutations in DBASS3 that cause css activation only (Table 4). Overall these results confirm that the likely effect of splicing mutations upon css activation or exon skipping can in general be inferred from their background splicing ratios. The exceptions to this general finding are shaded in Table 4 and discussed in more detail in Tables S3 and S7. This analysis shows that when the background reads for single exon skipping are greater that the background reads for any candidate css then exon skipping preferentially occurs in response to a splice site mutation (Fig 1D).

Multiple exon skipping

Table 5 lists all experimental reports of multiple exon skipping events that we found and compares these to the background splicing reads from snaptron. We also included experiments that did not detect the multiple skipping events indicated by snaptron but used RT-PCR primers that were capable of doing so (rows 33 to 42). We did not include predictions of multiple exon skipping from snaptron where experiments were restricted to single skip analyses. The first three examples are taken from a report about the LAMP2A, B and C variants which are generated by alternative splicing from a common 5'ss and three alternative 3'ss (17). The authors report that the same mutation of the common 5'ss has different effects upon single or double exon skipping by each 3' alternative ss. It can be seen that these differences in skipping correlate well with the relevant background splicing reads (Table 5, Appendix 1). Other notable features of Table 5 include reports of double exon skips only (rows 19 and 26) or mainly double exon skipping (rows 3, 7, 9, 22 and 24) and how this correlates with the higher background reads for double skips than single exon skips in snaptron. Similarly the reports of css and triple exon skipping (row 18) and single and quadruple exon skipping (row 23) are a good match to the background splicing reads.

There are ten examples (rows 33 to 42) where the experimental results do not match the multiple exon skip predictions from snaptron and six examples (8, 12, 13, 15, 18 and 30) where there is some but not exact agreement between the experimental results and background splicing reads. There are also six css listed that did not match snaptron background ss reads. For the css of row 5, snaptron has no splicing variants with which to compare and for row 2 the css has a non-consensus sequence, which is filtered from snaptron (3). The other four non-matching css are discussed at the bottom of the source tables.

This analysis shows that high background reads for multiple exon skips is a good indication that these events will occur in response to splice site mutations.

De novo ss and pseudoexons

Table 6 summarises our comparison of the snaptron database with mutations in DBASS that generate de novo splice sites (also known as de novo css) or pseudoexons (Table 2). Here we have divided the de novo mutations into two types, created or enhanced. Created refers to a mutation that creates the GT or GC dinucleotides of a 5' de novo ss or that creates the AG dinucleotide of a 3' de novo ss. Enhanced refers to mutations that enhance already existing GT, GC or AG dinucleotides. As expected none of the 34 and 123 created de novo ss of DBASS5 or DBASS3 match bss in snaptron (Table 6, row1). Even if there were reads for the original dinucleotide these would have been filtered from this database (3). There are 95 reports of mutations that enhance de novo ss in DBASS5 (Table 2) and we analysed the first 40 medical syndromes caused by this mutation type and report that 29 of these de novo css positions exactly match bss from snaptron (Table 6, row 1). Similar results were found for mutations that generated 3' de novo ss (row 2), although a far bigger proportion of the mutations created an AG dinucleotide splice site rather than enhanced existing AG sites.

Pseudoexons are most commonly generated when a mutation that creates a 5' or 3' de novo ss also causes the activation of a partner pseudoexon ss (Fig 2A,B). The 5' and 3' de novo ss that initiate pseudoexon formation matched background ss at a similar level to the de novo mutations only (Table 6). For the 3'pss that partner the 5' de novo mutations, there is a match of 59 out of seventy one 3'pss with background ss (Table 6). Of the twelve 3'pss that did not have a match in snaptron, ten were partnered to 5' de novo sites that were created from non-GT or non-GC dinucleotides (Table S9). Table S9 also describes that 54 out of the fifty nine 3'pss that matched background ss were the nearest upstream background 3'ss to the downstream mutation that created the de novo 5' pss. Four of the five 3'pss were only marginally more distant from an inner background 3'ss and all had far more reads than the inner 3'bss (Table S9).

Table 6 row 4 shows that a smaller proportion of 5' pss matched background ss (10/22). In all cases the matching bss are the nearest of all bss to the upstream 3' de novo ss mutation (Table S9).

Pseudoexons that were created by means other than de novo css mutations (Fig 2C) had the best match to bss (Table 6 row 5, Fig 2C). These were mainly mutations within the pseudoexon, some of which are known to create splicing enhancers, but also included five mutations outside the pseudoexon that enhance the polypyrimidine tract or the branch point recognition sites for the 3'pss. In addition, some of the pseudoexons were activated by mutations of flanking 5' or 3' ss (Table S9). 25 out of 26 pairs of these pseudo splice sites matched background ss in snaptron and 48 of these 50 pss matched bss with the highest reads of all background ss within the intron in which the pseudoexon was formed (Table S9).

Spliceosome mutations and cancer

Mutations of the spliceosome, in particular of SF3B1, have been reported to activate novel aberrant splicing events in leukaemia and other cancers (18-20). We report that all tested novel cancer ss caused by SF3B1 mutations matched background ss in snaptron with relatively high read numbers (Table 7, Table S10). Nevertheless, the background read numbers for the aberrant 3' or 5' css that are activated by spliceosomal mutations are still in the order of 1000 fold less than the reads for normal intron removal, as indicated in the last column of Table 7 and see Table S10. By contrast, the rarer exon skipping or exon inclusion events that are enhanced by SF3B1 mutations have background splicing reads only 20 to 40 fold lower on average than normal intronic splicing (Table 7).

Mutations of the splicing components U2AF and SRSF2 are reported to cause quantitative rather than qualitative changes in splicing (21,22), whereas mutations of the small non-coding RNA U1 are reported to activate novel splicing events in SHH medulloblastomas (23). However, we found that 23 out of 24 of the most novel aberrant splice sites caused by U1 mutations matched background splice sites, including matches to aberrant splice sites for PTCH1, GLI2, CCND2 and PAX5, which are implicated in this cancer (Tables 7, S10). Sixteen out of the 23 css caused by U1 mutations matched background ss within the top three background reads (Tables 7, S10)

Recursive splicing

Large introns are removed in sections by a process called recursive splicing that uses internal splice sites within introns (24-27). We analysed the first 20 of over 2000 recursive splice sites discovered by a screen of the human genome (25) and Tables 8 and S11 show that all of these sites matched background ss, as would be expected (26) and that in 12/20 cases the matching background ss had the highest reads of all bss with an individual intron. Similarly, Tables 8 and S11 show that 82 and 86% of 5' and 3' recursive splices identified in human DMD introns (27) matched background ss and that the background ss with the most reads matched 3' and 5'RS from DMD introns on 23/34 and 26/36 occasions.

Discussion

We report that the large majority of css that are activated in genetic disease are already in use at low levels by normal genes and are therefore a component of background or noisy splicing (1,2). Our results also indicate that the likely effect of splice site mutations upon css activation or exon skipping (Figure 1D) can often be discerned from the pattern of background splicing by normal genes. For example the css that are experimentally reported often correspond to background ss within 1000 bases of the mutated intronic splice site that have the most reads (Table 3, 3S, Fig 1D); when exon skipping is caused by a splice site mutation this usually correlates with higher background reads for skipping compared to candidate css reads (Table 4). Table 5 shows that the experimental reports of multiple exon skipping caused by splicing mutations also correlate reasonably well with background splicing reads. Consequently an initial consideration of background splicing is likely to give a useful indication of the primer design required to investigate the full effect of a potential splice site mutation.

It should be noted that this paper does not contribute to the large body of work designed to assess whether mutations at splice sites or outside these regions are likely to impair splicing (28), nor is it informative about intron retention, which is a common aberrant effect caused by splice site mutations.

We generally restricted our css candidates to background ss within 1000 bases of the mutated canonical ss (Figure 1D), because this is observed experimentally (4) (Table S2). However, many background ss are greater than 1000 bases from a canonical ss and in about 10% of introns, these sites have the highest number of reads (Table S7). We show here that some of these background ss have facilitated pseudoexon formation (Table 6) and that some are recursive splice sites (Table 8). Sibley et al (2015) previously established that recursive splice sites and recursive exons can be identified from RNA seq data (26).

About fifteen percent of css from the cryptic splice site database DBASS did not match background ss in the snaptron database SRAv1 (see below), usually because in these cases snaptron has no or relatively few background splice site reads with which to compare. Therefore the percentage of css that do not match bss is likely to fall as RNA sequencing databases increase in size.

There might however be a higher level of false positives, ie bss within 1000 bp of a splice site mutation that are not activated as css. The numbers in brackets in column 5 of Table 1 (rows 5, 24, 31, 35) show the reads of top bss that were not activated as css, despite having higher read numbers than those bss that matched the css. Of course these non-matching bss might be identified as css in subsequent experiments, Figure 1B provides an example of this. However, for BRCA2, two out of six top bss that were not activated as css following the mutation of 6 different splice sites (out of a total of 40) have been repeatedly analysed (Table S4).

The upper limit of top bss reads that might be false positives can be estimated from Table 3 as the proportion of css that matched bss that did not have the highest reads. For DBASS5 this is 51/201 (25%) and for DBASS3 35/97 (36%). Table S3 column T gives all of the details and also indicates those non-matching bss with markedly higher reads than the bss that match css. This gives a false positive estimate of 22 out of 201 (11%) for DBASS5 and 19 out of 97 (20%) for DBASS3. An estimate of the level of possible false positives can also be made from Table 4 where 8 out of 79 reports (10%) of 5' exon skipping only, nevertheless have higher background reads for candidate 5' css than for the exon skips (for example Table 1 row 9). Similarly there are 10 out of 64 reports (16%) of likely false positive 3'css candidates (Table S7 sheet 2).

For multiple exon skipping, we suggest that the level of false positives indicated by Table 5, ten out of 42, is an upper limit. We included these ten examples because the RT-PCR primers that were used were capable of but did not apparently detect the multiple exon skips indicated by snaptron (Table 5). However, there may be other reasons why some of these skipping events, if they occurred, might not have been reported.

Six mutations listed in DBASS5 and two mutations from DBASS3 (Table S2) generate more complicated patterns of aberrant splicing than those illustrated in Figure 1D. Because most of these examples did not readily fit the format of Table S3 they are separately analysed and discussed in Appendix 1.

We largely analysed background splice sites that splice to intron 5' or 3' ss (Figure 1D). Consequently background splicing events within introns and across intronic ss were not usually considered, in order to remove less relevant background splicing. We rescreened the 15% of experimentally identified css from DBASS for which we found no match to a background ss (Table 3) without this restriction and only found one clear example that we had missed by our approach (Table S3 5'css, PKP1).

Unrestricted screening increased the bss match to enhanced 5' de novo css from 29/40 to 37/40 and from 8/10 to 10/10 for enhanced 3' de novo css (Tables 6 rows 1 and 2, Table S9, sheets 1 and 2, column O). So an unrestricted screen for background splice site matches is perhaps best for assessing whether a mutation might generate a de novo css. In addition, there are many excellent in silico programmes that can already do this effectively (29,30).

There was a better match between background ss and css than between background ss and de novo css (Tables 3 and 6). In large part this is because possible reads for de novo css that originate from non-canonical splice sites are filtered from snaptron. But also the read numbers of the background ss that matched css were relatively higher than background ss that matched de novo sites (Table S3, S9).

De novo 3' or 5' css mutations that are more distant from an intron ss often activate partner pseudo splice sites (4). For such cases we found that there was a 59/71 match between 3'pss and background ss and a lower match (10/22) between 5'pss and background ss (Table 6). The 3' and 5'pss usually matched background ss nearest to the de novo mutation (Table S9).

Pseudoexons that were not initiated by de novo ss mutations (Fig 2C) usually matched background splice sites in the host intron with the highest reads (Table 6, Table S9). Similarly, css that are activated by mutations outside the core sequence of a canonical splice site often matched background ss with relatively high reads (Table 3). We suggest that for both of these different types of aberrant splicing events, the causative mutations are relatively weak and consequently may only have a phenotypic effect through enhancement of relatively active bss.

Snaptron has four different RNA sequencing databases that can be analysed (3) and see Materials and methods. We largely analysed the first database SRAv1 but as a control we also analysed BRCA1 and BRCA2 splicing mutations using the smaller GTEx and larger SRAv2 databases (Table S12). We found 35 experimentally reported css from both BRAC1 and BRCA2 of which 29 match bss listed in SRAv1 (Table S12). Use of the larger SRAv2 database increased the number of matches to 32/35, whereas the smaller Gtex database had only 18/35 matches (Table S12). Gtex RNA is made from normal tissue samples and Table S12 shows that the ratio of intron to css reads for each of the css of BRCA1 and BRCA2 have similar values when calculated from Gtex or from the SRA databases, so demonstrating that

css usage occurs at similar frequencies in the three databases. In further support of this conclusion, the css examples that matched bss from SRA databases but not Gtex usually had particularly large intron to css read ratios (Table S12), indicating that non-matching in Gtex is due to its smaller number of sequencing reads. Therefore background splicing is a property of normal genes and not just of genes from diseased tissue.

For DBASS, the 201/237 match between 5'css and bss increases to 219/237 (92%) with SRAv2 and from 97/110 to 101/110 (92%) for 3'css (Tables 3, S3). Similarly, the 59/71 match between 3'pss and bss increases to 63/71 with SRAv2 and from 10/22 to 14/22 for 5'pss (Tables 6, S9). Seven out of eight of the new pss matches in SRAv2 are with bss that are closest to the activating de novo mutation (Table S9).

The match between bss and the experimental reports of css activation in DMD is less than average (Table 3). DMD has relatively low expression levels and the correlation between css activation and bss reads increases slightly with the use of the larger SRAv2 database from 11/22 to 14/22 exact matches, this is again below average most probably because there are still relatively few sequencing reads and therefore variants for DMD even in SRAv2 (Table S5).

Our analysis indicates that the large majority of aberrant splicing events that have been detected in cancer samples are present in the snaptron database SRAv1 (Tables 7, S10a) and the Gtex database made from normal tissue (Table S10b). Furthermore, the background ss that matched the cancer ss had relatively high reads compared to other background ss, which is perhaps reflective of the relatively low sequencing coverage of the cancer samples. Our finding that oncogenic mutations of the spliceosome enhance strong background splice sites, rather than activate entirely novel ss, is consistent with the subtle effects of the spliceosome mutations upon splice site recognition (23,31).

There is strong evidence that two genes EZH2 and BRD9 have a causal role in cancer as the result of mutations of splicing components SRSF2 and SF3B1 respectively (32,33). EZH2 and BRD9 are both inactivated by aberrant splicing events that cause the inclusion of pseudoexons with stop codons (32,33). Snaptron shows that the 3' and 5' pss of the pseudoexons of both genes match the highest or second highest background reads within their host introns at approximately 5% of normal intron splicing reads for BRD9 and 25% for EZH2 (Table S10a, sheet 9). This indicates one way that spliceosome mutations, which are likely to cause only mild changes to splicing (34), might achieve a phenotypic effect, namely by altering background or alternative splicing events that are already established at relatively high levels.

Figure 3 outlines how background splicing information might contribute to the development of splicing therapies. Fig 3A illustrates the recent use of an antisense oligonucleotide (ASO) for the possible treatment of limb-girdle muscular dystrophy type 2A (35). The ASO blocks a css at -398 that is activated by deletion of a single T base 29 bases downstream from the acceptor cleavage site of intron 6 of CAPN3. ASO treatment results in the restoration of normal intron removal plus a mild side effect of weak activation of a css at -51 (Fig 3A). The snaptron database (Fig 3B) contributes the information that the

css at -389 has by far the most reads of any background ss within this intron at 1700 reads and that the css at -51 has 11 reads (Fig 3B).

We suggest that the CAPN3 mutation is a typical example of a relatively weak splicing mutation that only has a phenotypic effect because it is able to enhance an already strong background ss. Table S6 lists other examples of inactivating 5' and 3'ss mutations that lie outside the core ss region and that activate css with relatively high background reads, these may also be amenable to the approach used by Hu et al (35). By contrast the use of oligonucleotides to block css that are activated by mutations of the highly conserved region of an intron ss might be less likely to restore intron removal and more likely to enhance a different css (36) or to induce exon skipping (37).

Antisense oligonucleotides have been developed that can induce exon skipping for the treatment of Duchenne muscular dystrophy (38-40) and Wilton et al (38) have categorised individual DMD exons according to the ease with which they can be skipped, with category 4 exons as the most difficult. We wondered whether single exon skipping is favoured when background reads for this event outnumber reads for alternative background splicing events involving the same exon (Fig 4A). Fig 4B shows this type of background splicing data for all category 4 DMD exons plus the exons that we would predict to be the most difficult to skip as indicated by their background splicing reads. As can be seen there is not a strong overlap between the experimental results and our predictions, although there are points of agreement. Targeting of exon 8 or 54 with ASOs is found experimentally to cause the skipping of both exon 8 and 9 or of both exon 54 and 55 (38). This agrees with the background splicing data which has higher reads for the downstream double exon skip than for single skips of exons 8 and 54. Targeting of exon 10 caused multiple but variable downstream exon skipping (38), which is consistent with the dominant background reads for multiple downstream skipping of this exon (Fig 4B). However, our analysis also indicates that targeting exon 43 might induce double exon skipping (Fig 4B), which isn't the case (Table S13).

Interestingly, Fig 4B also illustrates that some DMD intron splice sites have more background reads for skips of 6, 7, 9, 11 or 12 exons than for single exon skips or for potential css. There are also examples of highest background splicing reads for 3, 4 or 5 exon skips for some of the genes listed in Table S3. Multiple exon skipping is relatively untested and it would seem important to discover if skipping on this scale does occur in response to splice site mutations or to antisense oligonucleotides.

Declarations

Funding

This work was supported by the BBSRC and the Genesis Research Trust.

Competing interests

None of the authors have any competing interests

Acknowledgments

We are most grateful to Chris Wilks and Ben Langmead for helping us to analyse their snaptron database and we also acknowledge the generous help of Annemieke Aartsma-Rus, Isabella Gazzoli and Yuri Kapustin.

References

1. Kapustin, Y., Chan, E., Sarkar, R., Wong, F., Vorechovsky, I., Winston, R.M., Tatusova, T. and Dibb, N.J. (2011) Cryptic splice sites and split genes. *Nucleic Acids Res*, **39**, 5837-5844.
2. Pickrell, J.K., Pai, A.A., Gilad, Y. and Pritchard, J.K. (2010) Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet*, **6**, e1001236.
3. Wilks, C., Gaddipati, P., Nellore, A. and Langmead, B. (2018) Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples. *Bioinformatics*, **34**, 114-116.
4. Buratti, E., Chivers, M., Hwang, G. and Vorechovsky, I. (2011) DBASS3 and DBASS5: databases of aberrant 3'- and 5'-splice sites. *Nucleic Acids Res*, **39**, D86-91.
5. Divina, P., Kvitkovicova, A., Buratti, E. and Vorechovsky, I. (2009) Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping. *Eur J Hum Genet*, **17**, 759-765.
6. Stenson, P.D., Mort, M., Ball, E.V., Chapman, M., Evans, K., Azevedo, L., Hayden, M., Heywood, S., Millar, D.S., Phillips, A.D. *et al.* (2020) The Human Gene Mutation Database (HGMD((R))): optimizing its use in a clinical diagnostic or research setting. *Hum Genet*.
7. Fokkema, I.F., Taschner, P.E., Schaafsma, G.C., Celli, J., Laros, J.F. and den Dunnen, J.T. (2011) LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat*, **32**, 557-563.
8. Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res*, **12**, 656-664.
9. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996-1006.
10. Thomassen, M., Blanco, A., Montagna, M., Hansen, T.V., Pedersen, I.S., Gutierrez-Enriquez, S., Menendez, M., Fachal, L., Santamarina, M., Steffensen, A.Y. *et al.* (2012) Characterization of BRCA1 and BRCA2 splicing variants: a collaborative report by ENIGMA consortium members. *Breast Cancer Res Treat*, **132**, 1009-1023.
11. Whiley, P.J., de la Hoya, M., Thomassen, M., Becker, A., Brandao, R., Pedersen, I.S., Montagna, M., Menendez, M., Quiles, F., Gutierrez-Enriquez, S. *et al.* (2014) Comparison of mRNA splicing assay protocols across multiple laboratories: recommendations for best practice in standardized clinical testing. *Clin Chem*, **60**, 341-352.
12. Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazor, Y. *et al.* (2016) The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics*, **54**, 1.30.31-31.30.33.

13. Wappenschmidt, B., Becker, A.A., Hauke, J., Weber, U., Engert, S., Kohler, J., Kast, K., Arnold, N., Rhiem, K., Hahnen, E. *et al.* (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**, e50800.
14. Colombo, M., De Vecchi, G., Caleca, L., Foglia, C., Ripamonti, C.B., Ficarazzi, F., Barile, M., Varesco, L., Peissel, B., Manoukian, S. *et al.* (2013) Comparative in vitro and in silico analyses of variants in splicing regions of BRCA1 and BRCA2 genes and characterization of novel pathogenic mutations. *PLoS One*, **8**, e57173.
15. Baert, A., Depuydt, J., Van Maerken, T., Poppe, B., Malfait, F., Van Damme, T., De Nobele, S., Perletti, G., De Leeneer, K., Claes, K.B. *et al.* (2017) Analysis of chromosomal radiosensitivity of healthy BRCA2 mutation carriers and non-carriers in BRCA families with the G2 micronucleus assay. *Oncol Rep*, **37**, 1379-1386.
16. Scholl, T., Pyne, M.T., Russo, D. and Ward, B.E. (1999) BRCA1 IVS16+6T–>C is a deleterious mutation that creates an aberrant transcript by activating a cryptic splice donor site. *Am J Med Genet*, **85**, 113-116.
17. Di Blasi, C., Jarre, L., Blasevich, F., Dassi, P. and Mora, M. (2008) Danon disease: a novel LAMP2 mutation affecting the pre-mRNA splicing and causing aberrant transcripts and partial protein expression. *Neuromuscul Disord*, **18**, 962-966.
18. Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M. *et al.* (2011) Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*, **478**, 64-69.
19. Darman, R.B., Seiler, M., Agrawal, A.A., Lim, K.H., Peng, S., Aird, D., Bailey, S.L., Bhavsar, E.B., Chan, B., Colla, S. *et al.* (2015) Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. *Cell Rep*, **13**, 1033-1045.
20. DeBoever, C., Ghia, E.M., Shepard, P.J., Rassenti, L., Barrett, C.L., Jepsen, K., Jamieson, C.H., Carson, D., Kipps, T.J. and Frazer, K.A. (2015) Transcriptome sequencing reveals potential mechanism of cryptic 3' splice site selection in SF3B1-mutated cancers. *PLoS Comput Biol*, **11**, e1004105.
21. Ilagan, J.O., Ramakrishnan, A., Hayes, B., Murphy, M.E., Zebari, A.S., Bradley, P. and Bradley, R.K. (2015) U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res*, **25**, 14-26.
22. Zhang, J., Lieu, Y.K., Ali, A.M., Penson, A., Reggio, K.S., Rabadan, R., Raza, A., Mukherjee, S. and Manley, J.L. (2015) Disease-associated mutation in SRSF2 misregulates splicing by altering RNA-binding affinities. *Proc Natl Acad Sci U S A*, **112**, E4726-4734.
23. Suzuki, H., Kumar, S.A., Shuai, S., Diaz-Navarro, A., Gutierrez-Fernandez, A., De Antonellis, P., Cavalli, F.M.G., Juraschka, K., Farooq, H., Shibahara, I. *et al.* (2019) Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma. *Nature*, **574**, 707-711.
24. Burnette, J.M., Miyamoto-Sato, E., Schaub, M.A., Conklin, J. and Lopez, A.J. (2005) Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. *Genetics*, **170**, 661-674.

25. Kelly, S., Georgomanolis, T., Zirkel, A., Diermeier, S., O'Reilly, D., Murphy, S., Langst, G., Cook, P.R. and Papantonis, A. (2015) Splicing of many human genes involves sites embedded within introns. *Nucleic Acids Res*, **43**, 4721-4732.
26. Sibley, C.R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M.E., Hardy, J. *et al.* (2015) Recursive splicing in long vertebrate genes. *Nature*, **521**, 371-375.
27. Gazzoli, I., Pulyakhina, I., Verwey, N.E., Ariyurek, Y., Laros, J.F., t Hoen, P.A. and Aartsma-Rus, A. (2016) Non-sequential and multi-step splicing of the dystrophin transcript. *RNA Biol*, **13**, 290-305.
28. Andreoletti, G., Pal, L.R., Moult, J. and Brenner, S.E. (2019) Reports from the fifth edition of CAGI: The Critical Assessment of Genome Interpretation. *Hum Mutat*, **40**, 1197-1201.
29. Spurdle, A.B., Couch, F.J., Hogervorst, F.B., Radice, P., Sinilnikova, O.M. and Group, I.U.G.V.W. (2008) Prediction and assessment of splicing alterations: implications for clinical testing. *Hum Mutat*, **29**, 1304-1313.
30. Ohno, K., Takeda, J.I. and Masuda, A. (2018) Rules and tools to predict the splicing effects of exonic and intronic mutations. *Wiley Interdiscip Rev RNA*, **9**.
31. Escobar-Hoyos, L., Knorr, K. and Abdel-Wahab, O. (2019) Aberrant RNA Splicing in Cancer. *Annu Rev Canc Biol*, **3**, 167-185.
32. Kim, E., Ilagan, J.O., Liang, Y., Daubner, G.M., Lee, S.C., Ramakrishnan, A., Li, Y., Chung, Y.R., Micol, J.B., Murphy, M.E. *et al.* (2015) SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. *Cancer Cell*, **27**, 617-630.
33. Inoue, D., Chew, G.L., Liu, B., Michel, B.C., Pangallo, J., D'Avino, A.R., Hitchman, T., North, K., Lee, S.C., Bitner, L. *et al.* (2019) Spliceosomal disruption of the non-canonical BAF complex in cancer. *Nature*, **574**, 432-436.
34. Inoue, D., Bradley, R.K. and Abdel-Wahab, O. (2016) Spliceosomal gene mutations in myelodysplasia: molecular links to clonal abnormalities of hematopoiesis. *Gene Dev*, **30**, 989-1001.
35. Hu, Y., Mohassel, P., Donkervoort, S., Yun, P., Bolduc, V., Ezzo, D., Dastgir, J., Marshall, J.L., Lek, M., MacArthur, D.G. *et al.* (2019) Identification of a Novel Deep Intronic Mutation in CAPN3 Presenting a Promising Target for Therapeutic Splice Modulation. *J Neuromuscul Dis*.
36. Sadusky, T., Newman, A.J. and Dibb, N.J. (2004) Exon junction sequences as cryptic splice sites: implications for intron origin. *Curr Biol*, **14**, 505-509.
37. Balestra, D., Barbon, E., Scalet, D., Cavallari, N., Perrone, D., Zanibellato, S., Bernardi, F. and Pinotti, M. (2015) Regulation of a strong F9 cryptic 5'ss by intrinsic elements and by combination of tailored U1snRNAs with antisense oligonucleotides. *Hum Mol Genet*, **24**, 4809-4816.
38. Wilton, S.D., Fall, A.M., Harding, P.L., McClorey, G., Coleman, C. and Fletcher, S. (2007) Antisense oligonucleotide-induced exon skipping across the human dystrophin gene transcript. *Mol Ther*, **15**, 1288-1296.
39. Matsuo, M. (1996) Duchenne/Becker muscular dystrophy: from molecular diagnosis to gene therapy. *Brain Dev*, **18**, 167-172.

40. Aartsma-Rus, A., De Winter, C.L., Janson, A.A., Kaman, W.E., Van Ommen, G.J., Den Dunnen, J.T. and Van Deutekom, J.C. (2005) Functional analysis of 114 exon-internal AONs for targeted DMD exon skipping: indication for steric hindrance of SR protein binding sites. *Oligonucleotides*, **15**, 284-297.
41. Dhir, A. and Buratti, E. (2010) Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. *FEBS J*, **277**, 841-855.
42. Vaz-Drago, R., Custodio, N. and Carmo-Fonseca, M. (2017) Deep intronic mutations and human disease. *Hum Genet*, **136**, 1093-1111.

Tables

Tables can be found in the supplementary section.

Table 1. Comparison of the experimental effect of splice site mutations of BRCA1 with snaptron splicing data. The first two columns report the experimental results. Column one lists the mutated canonical 5' or 3' splice sites (chr17hg19). The second column indicates whether the mutation caused css activation, exon skipping or both, the position of the css relative to the canonical ss is also indicated. Columns 3-7 compare RNA sequencing data from snaptron. Column 3 indicates whether the experimentally identified css (from column 2) exactly matches a background splice site in snaptron. Column 4 shows the css rank, for example 1(4) for row 3 of this column means that snaptron identified four background splice sites within 1000 bp (upstream or downstream) of the 5'ss 41258472 and that the site that matched the experimentally identified css had the most reads. Column 5 lists the highest scoring background splice site within 1000 bases of the mutated splice site For rows 5, 24, 31 and 35 the top bss read is given in brackets next to the reads for the bss that matches the css. Columns 6 and 7 list background reads for single and double exon skipping. The shaded boxes indicate the RNA sequencing results that do not match the experimental data, these are discussed in Table S1, which also lists references.

Table 2. Summary of the different types of splicing mutations in DBASS. DBASS lists mutations from a wide variety of human genes that cause disease by disrupting splicing (4). In DBASS5 there are 459 examples of inactivating mutations of 5'ss that activate 5'css, mainly within 1000 bases of the mutation. There are also 15 examples of css that are activated by less common mutations (described here as unusual) that lie outside the nine bases of the 5'ss consensus sequence CAG/guragu. Similarly, for DBASS3 there are 182 examples of inactivating mutations of the 3'ss that activate 3'css and 41 examples of 3' css that are activated by mutations outside the 3'ss consensus sequence yyycag/G. De novo ss are the second largest category of splicing defects, these are generated directly by the mutation. Here, we have divided the de novo mutations into two types: created refers to mutations that create the GT or GC dinucleotide of the 5'ss or create the AG of the 3'ss; enhanced refers to nearby mutations that enhance already existing GT, GC or AG dinucleotides. Pseudoexon activation is caused by mutations that

generate a de novo 5'ss (for DBASS5) or de novo 3'ss (for DBASS3) and a pseudoexon is produced when this is also accompanied by the activation of a partner splice site. Unusual pseudoexon activation refers to the creation of a pseudoexon by a mutation that does not create a de novo ss and usually lies within the pseudoexon (see Figure 2).

Table 3. Summary of matches between 5' and 3' css from DBASS, BRCA1, BRCA2 and DMD (columns 1,2) with background splice sites from snaptron (columns 3 to 5). DBASS5* and DBASS3* consider the subsets of css that are activated by mutations outside the core consensus intron splice sites. Column 2 shows the total number of css analysed for each gene or database and column 3 shows the number that matched a background ss. Column 4 shows how many of the css matched a background ss with the highest reads of all bss within 1000 bases of the mutation and column 5 shows the number of css that did not match a background ss for reasons other than paucity of sequencing data (see text and supplementary tables). Data summarised from the Tables listed in column 6.

Table 4. Css activation versus exon skipping. The experimental results listed in column 1 are from Table S3 (DBASS) and an exon skip database (Table S7) and they show the numbers of reports of css activation only, exon skipping only or both in response to 5' or 3'ss mutations. Columns 2 and 3 are from snaptron and show whether or not the reads for single exon skipping exceed the reads for the background ss that matches the css (DBASS) or for the background ss (<1000b from the canonical ss) with the most reads, when no css is reported or there isn't a match. Columns 4 and 5 show the total css and single exon skip reads (Tables S3, S7). Shaded examples are discussed (Tables S3, S7, see text).

Table 5. Experimental reports of mutations that cause multi-exon skipping compared to background splicing predictions. The experimental results are listed in columns 1 to 3 and snaptron data is compared in columns 4 to 9. For shading see text.

Table 6. De novo splice sites and pseudoexons . Row 1 summarises the information in Table S9 sheet1. It shows that none of 34 de novo 5'ss GT sites created from the mutation of non-GT or GC dinucleotides match bss listed in snaptron (possible ss that have non GT or GC dinucleotides are filtered from snaptron). By contrast 29 out of 40 de novo 5'ss match snaptron ss if they were generated by mutations just outside existing GT or GC dinucleotides (Table S9 sheet 1 for details). Similarly row 2 is a summary of Table S9 sheet 2 and lists that 123 mutations that created a 3' AG css did not match snaptron whereas 8 out of 10 mutations that enhanced existing AG dinucleotides matched snaptron (row 2). Row 3 summarises Table S9 sheet 3 and reports a total of 71 mutations that generate 5' dn css and also active a partner 3' pseudo splice site. 35 of the de novo ss were created and 36 were enhanced, only the enhanced de novo sites match snaptron background ss in 31/36 cases. 59 out of the 71 partner 3' pss match snaptron bss, of these 25 are partners with created de novo 5' ss and 34 are partners with enhanced de novo splice sites. Row 4 is a summary of Table S9 sheet 4 and reports that just one of the mutations that generate 3' de novo mutations match snaptron bss, but ten of the 22 partner 5'pss do match. Row 5 is a summary of Table S9 sheet 5 and reports that 25 out of 26 examples of pseudoexon ss that are created by mutations that do not directly create the pseudo ss match snaptron bss. Rows 3, 4

and 5 summarise an analysis of all pseudoexon reports in DBASS and also from Dhir & Buratti 2010 (41) and Vaz-Drago et al 2017 **(42)** (Table S9 sheets 3,4,5).

Table 7. Aberrant splice sites in cancer match background ss. Column 1 lists two of the spliceosomal components that are mutated in cancer and column 2 describes the aberrant splicing types caused by these mutations. Column 3 shows the number of aberrant ss identified in the cancer samples that match background ss from snaptron. Column 4 shows how many times the cancer ss match background ss within the top 3 most reads. Column 5 shows how often the 5' and 3'css match background ss that are nearest to a 5' or 3 intron ss. Column 6 shows the ratio of reads for normal intron splicing divided by the reads for the aberrant splicing event, from snaptron.

Table 8. Recursive spice sites. Table 8 summarises a comparison of recursive ss identified by Kelly et al 2015 (21) and Gazzoli et al 2016 (23) with background ss listed in snaptron. Column 2 shows that the first 20 recursive splice sites listed by Kelly et al 2015 (21) matched background ss and also shows how many of the RS reported by Gazzoli et al 2016 (23) matched background ss. Column 3 shows how many times recursive ss matched the background ss within an intron with the most reads. Further details (Table S11).

Figures

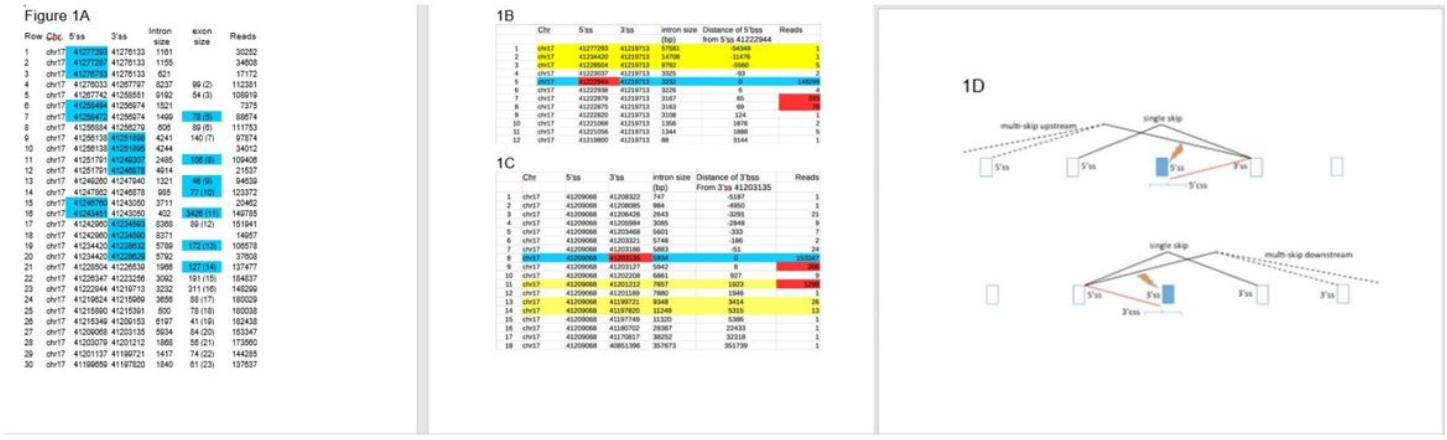


Figure 1

Figure 1A. Top BRCA1 spliced sequencing reads from snaptron identify intron ss and major ass. The 5'ss number is the first base of the intron and the 3'ss number is the last (hg19). Shading (columns 3 and 4)

shows alternative splice sites. Intron sizes are listed in column 5 and the sequencing reads for each of the splicing events are shown in column 7. The exon sizes generated from the constitutive splicing events and the most common alternative splicing events are shown in column 6, alternate exons are shaded, exon numbers are in brackets. The start codon of BRCA1 is located within the 99 bp exon at position 31276103 and the stop codon is at position 41197697 within exon 24. BRCA 1 does not have an exon 4 for historical reasons. B,C. Aberrant splicing activates background ss, particularly those with the most reads. B. The snaptron data is arranged to show all splicing events involving the 3'ss 41219713 of wild type BRCA1 (see text). Blue shading shows normal splicing, yellow shading shows background exon skipping, red shading indicates the effect of mutation of the 5'ss partner 42111944. C. All splicing events involving the 5'ss 41209068, blue and yellow shading as above, red shading indicates the effect of mutation of the normal partner 3'ss 41203135. D. These figures illustrate aberrant splicing events that are commonly enhanced by mutations of the 5' or 3'ss. The brackets reflect that most css lie within 1000 bases of the ss mutation.

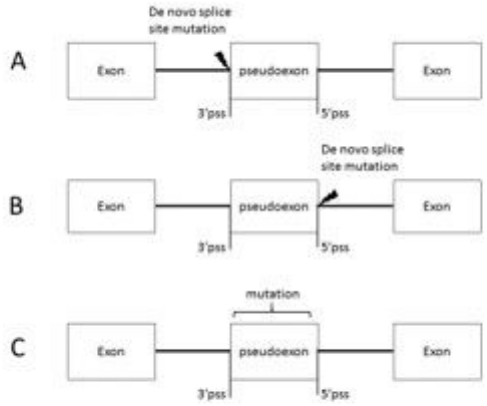
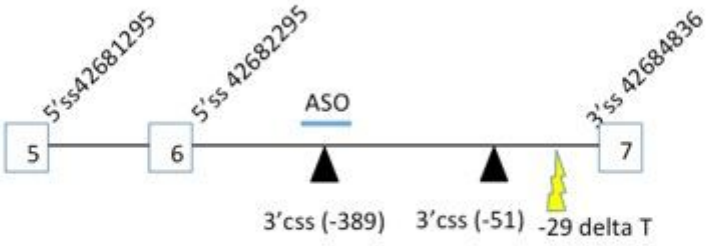


Figure 2

The three most common ways of generating a pseudoexon. A) A de novo mutation creates or enhances a 3'pseudo splice site which leads to the activation of a downstream 5' pseudo splice site. B) A de novo mutation creates or enhances a 5' pseudo splice which leads to the activation of an upstream 3' pseudo splice site. C) Mutations other than de novo splice site mutations can enhance pseudoexon usage, of these the most common mutations occur within the pseudoexon (4).

3A



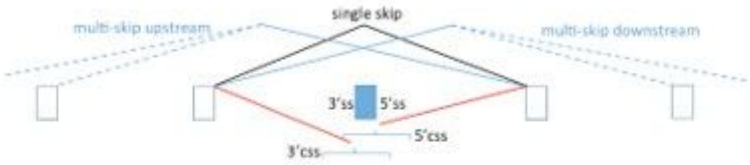
3B

Rows	5'ss (exon 6)	3'ss	Intron size bp	3'css positions	Reads
1	42682295	42683641	1347	-1194	1
2	42682295	42684036	1742	-799	5
3	42682295	42684446	2152	-389	1701
4	42682295	42684452	2158	-383	3
5	42682295	42684531	2237	-304	41
6	42682295	42684784	2490	-51	11
7	42682295	42684836	2542	0	94115
8	42682295	42684854	2560	19	1
9	42682295	42686453	4159	1618	115
10	42682295	42688997	6703	4162	10
11	42682295	42700408	18114	15573	1
12	5'ss (exon 5)				
13	42681295	42684446	3152	-389	168
14	42681295	42684531	3237	-304	13
15	42681295	42684836	3542	0	12827

Figure 3

Diagram of exons 5 to 7 of the CAPN3 gene. Hu et al 2019 (35) report that deletion of a thymidine 29 bases upstream from the 3'ss at the start of exon 7 (42684836 hg 19) strongly activates a css at -389. Exon 7 is alternatively spliced to exons 5 and 6 and the mutation caused the insertion of 389 bases into the major exon product 5,6,7 and into the minor exon product 5,7. An antisense oligo against css -389 restored normal splicing apart from the insertion of 51 bases into a small percentage of exon product 5,6,7 but not 5,7. Fig 3B shows the background and normal splicing reads for the 5'ss of exons 5 and 6 of the wild type CAPN3 gene. The row in darkest blue shows the reads for splicing between exons 6 and 7 and the lighter blue row shows the reads for splicing between exons 5 and 7. Red rows show reads between exons 5 or 6 with the 3'css -389.

Figure 4A



4B

DMD exons	Wilton et al classification	bbs classification	Notes	5'css	multiskip upstream	Single skip	multiskip downstream	3'css	Notes
2	1	1		x	x	*106	0	x	
8	4	2	Css (+5)	13	0	0	16	13	2nd skip, Css (-17)
10	4	2	Css (+272), no skip reads	1	0	0	8,82	1	3rd and 11th skip
17	4	2	Css (+479)	11	x	8*	1	x	2nd skip
20	4	3	Css (+854), 11th skip	61	82	0*	3	113	2nd skip, Css (+2),
34	4	2	11th skip	5	x	0	1	1	Css (-7), 7th skip
54	4	2		0	0	2*	0	0	
66	4	2	Css (+897)	1	0	0*	1,8	x	3rd, 12th skip
18	1	3	Css (+4), 3rd skip	1	0	0	10	11	11th skip, Css (-947)
43	2	3	Css(+4), 2nd skip	2	8	*0*	3	1	2nd skip, Css (-7)
59	3	3	2nd skip	x	13	*0	0	[redacted]	Css (-68), (+2)
69	3	3	2nd skip	x	67	4*	11	209	Css (+6), 8th skip
70	2	3	Css (-101) and (+441)	72, 210	0	0*	1	6	2nd skip, Css (+86)
78	3	3	2nd and 8th skip	x	21,58	0	0	867	Css (+90)

Figure 4

Diagram of background splicing information of possible relevance to single exon skipping by antisense oligonucleotides. Normal intron splicing is not shown. The black lines illustrate background reads for single exon skipping, the red lines background reads for potential css (background ss within 1000 bp of an intron ss) and the blue lines represent multi-exon skipping, upstream or downstream. Fig 4B. A selection from Table S13 that analyses the exons in column 1 by comparing the background ss reads for single exon skips (column 7) to upstream and downstream background splicing events involving the same exon (columns 5,6 and 8,9). This information is used to estimate the ease with which each exon might in theory be skipped by anti-sense oligonucleotides, where 1 is easy and 3 is hard (column 3). For example exon 2 is classified as 1 (column 3) because the reads for background single exon skipping are greater than alternative background reads both upstream and downstream. Column 2 shows the experimental classification of exon skipping success by anti-sense oligonucleotides (Wilton et al 2007) where 1 is highly efficient and 4 is ineffective. See text and legend to Table S13 for further details.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigS1BRCA1.ods
- TableS1brca1exceptions.docx
- TableS2index.ods

- TableS35css3css.xlsx
- TableS4BRCA2.ods
- TableS5DMD.ods
- TableS65cssunusual3cssunusual.xlsx
- TableS7exonskipping.xlsx
- TableS8.xlsx
- TableS95dn3dn.xlsx
- TableS10acancer.xlsx
- TableS10bgtexanalysis.xlsx
- TableS11recursive.xlsx
- TableS12Gtex.xlsx
- TableS13DMD.ods
- Appendix1.docx
- Tables1through8.docx