

# **Hidden Stitches: RNA Cryptic Splicing and Its Role in Human Disease**

**Niall Patrick Keegan, BSc (Hon), MPhil**



This thesis is presented for the degree of Doctor of Philosophy at Murdoch University

Centre for Molecular Medicine and Innovative Therapeutics  
Health Futures Institute

**“Sometimes scientists change their minds. New developments cause a rethink. If this bothers you, consider how much damage is being done to the world by people for whom new developments do *not* cause a rethink.”**

Pratchett, T., Stewart, I., and Cohen, J. (2002) *The Science of the Discworld* (2<sup>nd</sup> Edition), pg. 14. London: Ebury Press.

## **Thesis Declaration**

I, Niall Patrick Keegan, declare that this thesis is my own account of my own research, except where other sources are acknowledged in the appropriate format. Where the work of others has been used, their contributions are documented with a percent allocation that has been signed by each contributor and by my Principal Supervisor, Professor Sue Fletcher.

The university-supplied anti-plagiarism software *Ouriginal* has been used to ensure this work is of the appropriate standard to send for examination.

All necessary ethics and safety approvals have been obtained, including their relevant approval or permit numbers as appropriate. Details are listed within each results chapter.

This document is a thesis by publication, and all publications have been inserted as journal-formatted PDFs, designated with a chapter number. All publications are Open Access and adhere to Creative Commons licensing.

Signed:

Date: 1<sup>st</sup> of February, 2022

## **Abstract**

A great majority of human genes contain introns: tracts of mostly non-functional sequence that intervene the functional exons. When intron-bearing genes are transcribed into RNA, the introns are removed from the transcript via splicing, a process controlled by a multimolecular assembly called the spliceosome.

Although splicing is generally well-regulated, the spliceosome sometimes splices RNA transcripts at sites other than their canonical exon boundaries. This “cryptic” splicing can be a random event, part of an unidentified regulatory process, the effect of a mutation, or the result of other perturbances to the spliceosome’s normal behaviour.

In this thesis, I present four reports on the mechanisms underlying certain forms of cryptic splicing. In the first report, an analysis of pathogenic pseudoexons in the *DMD* gene reveals that each causative mutation falls into a distinct category defined by its proximity to the pseudoexon, and that many *DMD* pseudoexon splice sites are actively spliced in non-mutant cells. The second report builds on this by constructing a catalogue of over 400 pseudoexon variants from across the human transcriptome and uses this dataset to propose new and revised pseudoexon mutation categories. Like the first report, this second report also finds substantial congruence between pseudoexons and active deep intronic splice sites – including several recursive splice sites – suggesting a causal link between these phenomena.

A third report explores how some cryptic exons may provide an explanatory mechanism to connect common genetic variants with their associated population phenotypes and outlines a simple method for discovering new examples.

The fourth and final report uses RNA secondary structure modelling to explain why some antisense oligonucleotides can induce partial exon skipping through cryptic splice-site activation.



Collectively, these reports present several novel insights into the causes of cryptic splicing and offer suggestions for how future research may build upon these insights.

## **Table of Contents**

Thesis declaration .....	i
Abstract .....	ii
Table of Contents .....	iii
Statement of Authorship and Originality .....	vii
Acknowledgements .....	ix
1. Chapter 1: Introduction .....	1
2. Chapter 2: Literature Review .....	4
2.1. What is RNA? .....	5
2.2. Types of RNA .....	5
2.3. mRNA life cycle .....	5
2.3.1. Transcription .....	5
2.3.2. Splicing .....	6
2.3.3. Termination and polyadenylation .....	7
2.3.4. Transport .....	7
2.3.5. Translation .....	8
2.3.6. Decay .....	8
2.4. The spliceosome(s) .....	9
2.4.1. U1 subunit .....	10
2.4.2. U2 subunit .....	10
2.4.3. U4, U5 and U6 subunits .....	10
2.4.4. U11 and U12 subunits .....	11
2.4.5. U4atac, U5 and U6atac .....	11
2.5. ESEs and ESSes .....	11
2.5.1. SRSFs .....	12

2.5.2.	hnRNPs .....	12
2.6.	Functional deep-intron splice motifs .....	13
2.6.1.	Recursive splicing .....	13
2.6.2.	Poison exons .....	13
2.6.3.	Decoy exons .....	14
2.7.	Splicing mutations .....	15
2.8.	Therapeutic strategies for manipulating splicing .....	16
2.9.	Summary .....	18
3.	Chapter 3: Pseudoexons of the DMD Gene .....	19
3.1.	Preamble .....	20
3.2.	Citation .....	21
3.3.	Publication .....	22
3.4.	Mentions and Awards .....	41
4.	Chapter 4: Analysis of pathogenic pseudoexons reveals novel mechanisms driving cryptic splicing .....	43
4.1.	Preamble .....	44
4.2.	Citation .....	46
4.3.	Publication .....	47
5.	Chapter 5: A spotter's guide to SNPtic exons: The common splice variants underlying some SNP-phenotype correlations .....	73
5.1.	Preamble .....	74
5.2.	Citation .....	75
5.3.	Publication .....	76
6.	Chapter 6: Induction of cryptic pre-mRNA splice switching by antisense oligonucleotides .....	92
6.1.	Preamble .....	93
6.2.	Citation .....	94
6.3.	Publication .....	96

6.4. Mentions and awards .....	109
7. Chapter 7: Conclusions and Future Directions .....	110
7.1. Revisions and additions to pseudoexon mutation categories .....	111
7.2. The interrelation of pseudoexons with recursive splice sites and other deep intronic splice motifs .....	112
7.3. The “SNPtic exon” concept and its explanatory power .....	113
7.4. The mechanisms of antisense-oligomer-induced partial exon skipping and its potential as a therapeutic strategy .....	115
7.5. Limitations and broader challenges .....	116
7.6. Concluding remarks .....	117
References .....	118

## **Statement of Authorship and Originality**

This thesis and the research presented herein is my own, except where other sources and contributors are appropriately acknowledged. The list below shows the percentage contributions of all collaborators to each results chapter. E-mail assent to these credit assignments has been obtained from each contributor and uploaded to IRMA as a supporting document for this thesis.

### **Pseudoexons of the DMD Gene**

Niall P. Keegan	100%
-----------------	------

### **Analysis of pathogenic pseudoexons reveals novel mechanisms driving cryptic splicing**

Niall P. Keegan	85%
-----------------	-----

Steve D. Wilton	5%
-----------------	----

Sue Fletcher	10%
--------------	-----

### **A spotter's guide to SNPTic exons: The common splice variants underlying some SNP-phenotype correlations**

Niall P. Keegan	90%
-----------------	-----

Sue Fletcher	10%
--------------	-----

### **Induction of cryptic pre-mRNA splice switching by antisense oligonucleotides**

Kristin A. Ham	35%
----------------	-----

Niall P. Keegan	35%
-----------------	-----

Craig S. McIntosh	5%
-------------------	----

May T. Aung-Htut	5%
------------------	----

Khine Zaw	5%
-----------	----

Kane Greer	5%
------------	----

Sue Fletcher	5%
--------------	----

Steve D. Wilton	5%
-----------------	----

Student signature:

Date: 1<sup>st</sup> of February, 2022

Principal supervisor signature:

Date: 1<sup>st</sup> of February, 2022

## **Acknowledgements**

I would like to thank and acknowledge all the people who helped me on the long and tortuous journey from enrolment to submission.

My Dad: for keeping me fed with weekly pub-dinners and limitless quantities of his home-made date clusters; for knowing when I needed help but was too proud to ask for it, especially the portable air-conditioner he got me that saved me from broiling to death in my apartment during these last two Perth summers; listening to me talk about what I've been up all those times when the most interesting thing I'd done all week was re-format a spreadsheet; for unfailingly supporting my pursuit of science as a career, tempered with the occasional (justified) concern about the state of my personal finances; and for giving me the security of knowing that I'll always have my old bedroom waiting for me if worse comes to worst. Many who enter post-graduate research aren't nearly so fortunate. Love you Dad!

My supervisor, Professor Steve D. Wilton: for being a fountain of ideas and enthusiasm about not just my projects but everything going on in the lab; for consistently emphasising the importance of honest and high-quality data, of respectful lab conduct, and of developing my skills as a science communicator. A big factor in why I wanted to come back to CMMIT for my PhD was knowing that the Centre had your steady hand on the rudder. It's a testament to the lab culture that you and Sue have built that so many of the people who were with the group when I first joined are either still here or still associated in some capacity.

My co-authors on the research paper presented in Chapter 6 - Kristin, Craig, May, Khine, and Kane: This was my first real group project since my undergrad days, and anyone who's done more than one or two group projects knows how badly they can sometimes go. Thankfully, you were all a dream to work with - especially my co-first author, Kristin. Every time during the project when I stopped and thought, "Oh no, wait, what do we do about X?" I would ask you and find out you already had it covered! Special thanks too to May and Craig, for all the essential groundwork you

both did and for helping me find my way out of a mental dead-end shortly after I joined the project.

Everyone at CMMIT, but especially the "originals" from back when I first joined the group (ANRI as it was back then), in alphabetical order: Abbie, Ianthe, Kane, Kristin, Loren, May, and Russell. You've all helped me so much over the years in more ways than I can count, not just helping me build my confidence as a scientist but including me in social outings big and small. I am lucky to have you all as co-workers and even luckier to count you all as friends.

A special tip-of-the-hat all our group's international students, past and present. I stand in awe of you. Doing a PhD at all is bloody hard work, so the intelligence, resilience, and strength-of-character it takes to do a PhD in a foreign country, in a second language, and while raising a family (as many of you were/are), is nothing short of stunning. May each of you achieve the career of your dreams - you've earned it, many times over.

And Professor Sue Fletcher, my primary supervisor. It's no exaggeration to say I owe my career to you. You took a chance on me when I cold-called you for a job in 2013 with little more to my CV than a barely-passed Honours degree and a single decent presentation on facioscapulohumeral muscular dystrophy. You encouraged me to enrol in a Masters degree and were unflinching in your support all the way through, even though I sometimes wondered why, when I was so slow with lab work and my projects were constantly changing direction. And although the Masters Thesis I eventually produced could fairly be described as "okay" at best, you didn't hesitate to take me on again for a PhD - and when I got halfway through THAT and decided to basically put aside two years' work and start over with something completely different (in my defence, largely due to the practicalities of early-pandemic snap lockdowns), you still had my back, and I again found myself wondering why. Now, as I sit here writing this final part of a thesis-by-publication with four first-author papers, I'm starting to think that maybe you were onto something with this "Dr. Keegan" business. Thank you for everything.



# **Chapter 1**

## **Introduction**

The term “RNA splicing” refers to the joining together of two polynucleotide strands of ribonucleic acid into a single continuous molecule. In eukaryotic cells, splicing is an essential part of pre-mRNA maturation, as it permits non-coding intron sequences to be excised while ligating their flanking exons together into a translatable message. This exon-to-exon splicing is referred to as “canonical” splicing. However, splicing of sites other than canonical exon boundaries is observed in transcripts of many genes, and occurs as low-frequency events of unknown function, as a regulatory mechanism, or as the pathogenic consequence of mutation. These phenomena are collectively referred to as “cryptic” splicing. It is the goal of this thesis to investigate some of the less-examined forms of cryptic splicing, in order to better understand how and why they occur, and to discuss the potential applications of these findings. The remainder of this first chapter outlines and justifies the structures of the chapters to follow, and Chapter 2 provides a literature review of mRNA splicing and other relevant topics.

Chapters 3 through 6 comprise four research papers that each examine rare forms of cryptic splicing and offer new insights into their underlying molecular mechanisms. Special attention is given to categorising the genetic variants that affect cryptic splicing and discussing how these categories, some of which are novel to these reports, could improve both genetic diagnostics and the development of new RNA-modulating therapeutics.

The first of these reports, “Pseudoexons of the *DMD* Gene,” (Chapter 3) examines reported cases of exonised deep intron tracts – pseudoexons – in the *Duchenne Muscular Dystrophy* gene and attempts to categorise their associated mutations. The report also reveals new associations between *DMD* pseudoexons and putative recursive splice sites and proposes that pseudoexons that arise from recursive splice sites do so through a distinctly different pathway than other pseudoexons.

The second report, “Analysis of pathogenic pseudoexons reveals novel mechanisms driving cryptic splicing,” (Chapter 4) extends this analysis to reported pseudoexon mutations from other genes, expanding and revising existing pseudoexon mutation

categories and identifying remaining gaps in our knowledge of why pseudoexons occur. This report also discovers that numerous pseudoexon splice sites are actively spliced in non-mutant cells, including seven experimentally verified recursive splice sites.

The third report, “A spotter’s guide to SNPtic exons: The common splice variants underlying some SNP-phenotype correlations,” (Chapter 5) describes how the splicing of some low-frequency cryptic exons can be subtly modulated by the presence of common genetic variants nearby. Because inclusion of a “SNPtic” exon in a mature transcript usually alters the reading frame, SNP-related differences in inclusion frequency may explain or predict that SNP’s correlation with certain population phenotypes. The report catalogues a few known and likely examples of SNPtic exons and describes a simple method for discovering more.

The fourth report, “Induction of cryptic pre-mRNA splice-switching by antisense oligonucleotides,” (Chapter 6) describes rare cases of antisense oligonucleotides inducing partial exon skipping through activation of cryptic splice sites within canonical exons. The report models the accessibility of each pre-mRNA exon to enhancing and silencing splice factors, and how this accessibility is influenced by antisense oligonucleotide binding. A common feature of these examples of induced partial exon skipping was a shift in the relative strength of exon definition between the retained and skipped segments.

Lastly, in Chapter 7, the findings of these four reports are summarised and their broader significance is assessed.

# **Chapter 2**

## **Literature Review**

## 2.1            What is RNA?

Ribonucleic acid (RNA) is a polymer consisting of combinations of four ribonucleosides – adenine, guanine, cytosine, and uracil – linked by a sugar-phosphate backbone. Most of the RNA in cells is synthesised by RNA polymerases, enzymes that create reverse-complementary transcripts of DNA sequences, although some enzymes such as poly-(A) polymerase can extend existing RNA molecules in a DNA-sequence-independent manner (Tian and Graber 2011). Single-stranded RNA is capable of folding into complex secondary structures that can bind and interact with other molecules, allowing it to take on a variety of roles within the cell.

## 2.2            Types of RNA

There are numerous known types of RNA, and additional new forms potentially remain to be discovered (Kapranov *et al.* 2010). Perhaps the most well-studied RNA type is messenger RNA (mRNA), which encodes the peptide sequence of proteins. Other RNAs, though not translated themselves, perform essential functions in the translation process. Ribosomal RNA (rRNA) is usually the most prevalent RNA in a cell, exceeding 94% of total RNA in some cell types (Morlan *et al.* 2012) and forms the core components of the 60S and 40S ribosomal subunits, while transfer RNAs (tRNAs) are responsible for delivering individual amino acid residues into the ribosome during translation. Other types of noncoding RNA fulfill a wide variety of functions in the cell and are reviewed in detail elsewhere (Boivin *et al.* 2019; Zhang *et al.* 2019).

## 2.3            mRNA life cycle

The life cycle of a typical mRNA molecule is composed of six stages: transcription, splicing and capping, polyadenylation, transport, translation, and decay.

### 2.3.1            Transcription

Transcription of pre-mRNA begins when a molecule of RNA polymerase II (RNAPII) binds to a promoter region at the 5' end of a gene. Promoter regions are defined by the presence of various core promoter motifs in the DNA, although the exact

combination of motifs present varies from gene to gene (Juven-Gershon *et al.* 2008) and many genes have multiple alternative promoters, a significant evolutionary strategy for increasing transcript diversity (Reyes and Huber 2018). Transcription initiation can be further modulated by distal regulatory elements such as enhancers, silencers, and insulators (Maston *et al.* 2006).

Once RNAPII has bound to the antisense strand of the DNA, the elongation phase of transcription begins. RNAPII proceeds in the 3' to 5' direction along the DNA while the reverse-complementary pre-mRNA is synthesised in the 5' to 3' direction. After the polymerase moves to each new DNA nucleotide, it guides a complementary RNA nucleotide into place and ligates it to the growing molecule before moving to the next nucleotide. The rate of elongation varies greatly both among genes and within individual genes and is paused at some points to allow certain regulatory processes to take place (Saba *et al.* 2019), such as 5'-end capping (Kiledjian 2018), but under favourable conditions elongation can reach speeds of up to five kilobases per minute (Jonkers and Lis 2015).

### 2.3.2 Splicing

A vast majority of human genes contain introns (Zou *et al.* 2011). Introns are tracts of non-coding sequence that separate exons, while exons are the segments that either encode the mRNA protein message or form part of the mRNA 5' or 3' untranslated regions in the mature transcript. The first introns are thought to have arisen concomitantly with the eukaryote domain itself, although the reason for their emergence remains an open question (Rogozin *et al.* 2012). Because introns, by definition, do not encode functional peptide sequence, they must be removed from the maturing transcript before it can be translated. This process is called splicing, and generally occurs while the mRNA is being transcribed (Pandya-Jones *et al.* 2009). The existence of exons and introns also permits alternative splicing, an additional strategy for increasing transcript diversity, whereby certain exons are included in some mature transcripts but spliced out along with their flanking introns in others (Kelemen *et al.* 2013).

Splicing is performed by the spliceosome, a multimolecular assembly that recognises exon-intron boundaries by binding to sequence motifs at the 5´ and 3´ ends of introns and is also influenced by a range of other factors, including motifs within exons and introns, RNA secondary structure and GC content.

### 2.3.3 Termination and polyadenylation

Termination and polyadenylation of pre-mRNA are interlinked processes in eukaryotes, as transcription of the polyadenylation signal at the 3´ end of a gene is what triggers cleavage of the elongating transcript (Rosonina *et al.* 2006). Cleavage typically occurs 8-40 nucleotides downstream of the transcribed polyadenylation signal (Tian *et al.* 2005), but does not immediately halt RNAPII, which continues transcribing until it is dissociated by the multiprotein cleavage and polyadenylation complex (Eaton *et al.* 2020).

Pre-mRNA cleavage is followed by polyadenylation, wherein a molecule of poly-(A) polymerase appends multiple consecutive adenine ribonucleotides to the 3´ end of the transcript (Tian and Graber 2011). The length of the poly-A tail on a freshly transcribed eukaryote mRNA is typically around 250 nucleotides, although regulatory and decay processes will inevitably alter its length over the course of the transcript's lifespan (Eckmann *et al.* 2011). Just as many genes have multiple promoters and transcription start sites, alternative polyadenylation sites are also ubiquitous throughout the human transcriptome and are a major source of transcript diversity (Tian and Graber 2012; Reyes and Huber 2018).

### 2.3.4 Transport

Because eukaryotic mRNAs are transcribed in the nucleus but translated in the cytoplasm, they must undergo an intervening step of transport out of the nuclear envelope. Transport also allows an mRNA to be translated close to the site where the encoded proteins are needed, thereby minimising the transport distance for each protein molecule (Das *et al.* 2021). Mature mRNA-protein complexes (mRNPs) move from the transcription site to the nuclear membrane via diffusion before being exported to the cytoplasm through nuclear pores (Vargas *et al.* 2005; Cole and

Scarcelli 2006). Once outside the nucleus, active transport mechanisms such as cytoskeletal motors further localise some mRNAs by moving them to destinations (Czaplinski and Singer 2006) determined by localisation elements or “zipcodes” in the mRNA, usually in the 3′ UTR (Martin and Ephrussi 2009).

#### 2.3.5 Translation

Translation initiation of mature mRNAs is predominantly cap-dependent and begins with the assembly of a pre-initiation complex at the 5′ cap of the mRNA. This complex consists of the 40S ribosomal subunit, through which the pre-mRNA is ‘threaded,’ and several other factors (Giess *et al.* 2020). The complex scans the mRNA 5′ to 3′ until it detects the translation initiation start site, typically defined by the presence of the Kozak sequence (Kozak 1987), at which point the non-40S subunits are dissociated and the 60S ribosomal subunit is recruited (Giess *et al.* 2020). Cap-independent translation, a process once thought to be restricted to viruses, is also a conserved feature of some human genes, although at present it seems to be the exception rather than the rule for protein synthesis (Merrick 2004; Shatsky *et al.* 2018).

Once initiated, the ribosome proceeds to translate the open reading frame of the mRNA according to the genetic code and ceases once it reaches a stop codon, at which point the ribosomal subunits disengage. Translation can sometimes proceed past the first in-frame stop codon but known examples of this occurring naturally are very rare (Loughran *et al.* 2014). It is common for individual mRNAs to be translated multiple times and for multiple rounds of translation to occur simultaneously on a single transcript (Afonina and Shirokov 2018), although ribosome density is generally lower on transcripts with longer open reading frames (Ciandrini *et al.* 2013).

#### 2.3.6 Decay

The controlled breakdown of mRNA is essential to cell survival, as this process prevents unneeded or damaged transcripts from accumulating and allows their component nucleotides to be recycled into new transcripts (Pérez-Ortín *et al.* 2013). There are numerous mechanisms for degrading mRNA, as reviewed in detail by



Garneau *et al.* (2007), but the most well-studied and most relevant to splicing mutations is nonsense-mediated decay (NMD).

Nonsense-mediated decay is triggered when there is a disruption to the open reading frame of a mature mRNA, as defined by the presence of a premature termination codon more than 55 nucleotides 5' of the last exon-exon junction (Zhang *et al.* 1998). It is theorised that these premature stop codons are detected during the first pass of a ribosome along a mature mRNA, which normally dislodges all the exon junction complexes deposited during splicing (Le Hir *et al.* 2001). If the ribosome reaches a stop codon more than 55 nt upstream of one or more of these complexes, NMD cofactors are recruited, and the transcript is marked for RNase degradation (Nickless *et al.* 2017).

While it is well-established that apparent premature stop codons are the root cause of NMD, the development of a complete map of factors involved and their interactions is still a work-in progress, and NMD appears to be especially important in the development of the central nervous system (Lee *et al.* 2021). Even the question of whether NMD occurs within the nucleus or exclusively in the cytoplasm is not yet settled, as there is credible evidence supporting both models (Singh *et al.* 2007; Sato *et al.* 2008).

#### 2.4 The spliceosome(s)

Although it is customary to refer to 'the spliceosome' when discussing vertebrate splicing, there are two spliceosomes employed in processing nuclear transcripts – the U2 spliceosome and the U12 spliceosome, also called the major and minor spliceosomes, respectively.

The U12 (minor) spliceosome is responsible for splicing less than 1% of all introns in the human transcriptome (Olthof *et al.* 2019), while the U2 (major) spliceosome processes the remainder. Evidence from phylogenetics suggests that the origins of the U12 spliceosome are nearly as old as the eukaryote domain itself, although many eukaryote subgroups have subsequently lost this function (Russell *et al.* 2006).

Despite being utilised much less frequently than the U2 spliceosome, the U12 spliceosome exerts higher stringency in its requirements for what can and cannot be spliced (Akinyi and Frilander 2021). Minor spliceosome donor motifs and branch point motifs are much more rigidly defined than their major spliceosome equivalents (Turunen *et al.* 2013), though the minor spliceosome is more flexible in which terminal dinucleotides it permits to splice.

The core components of both spliceosomes consist of five small nuclear ribonucleoproteins (snRNPs), each comprising a single noncoding RNA bound by multiple proteins. In the major spliceosome these snRNPs are designated U1, U2, U4, U5, and U6, while those of the minor spliceosome are U11, U12, U4atac, U5, and U6atac. Only U5 is shared between the two. The functions of the snRNPs are as follows:

#### 2.4.1 U1 subunit

Recognises and binds to the 5' donor splice motif, via reverse-complementary base-pairing with part of the U1 snRNA. This is referred to as the *splicing commitment complex*, or *complex E*.

#### 2.4.2 U2 subunit

Recognises and binds to the branch point motif. Like U1, the U2 snRNP also binds its motif through reverse-complementary base pairing, but in the U2 snRNA the branch site recognition region contains several pseudouridines added through post-transcriptional modification of the uridines encoded by the U2 gene (Dönmez *et al.* 2004). These pseudouridines are thought to stabilise the snRNA-mRNA bond through improved base-stacking and create a secondary structure more conducive to the splicing steps to follow (Adachi and Yu 2014). Once both the U1 and U2 snRNPs have bound to the RNA, this is designated as *complex A*.

#### 2.4.3 U4, U5 and U6 subunits

These three snRNPs associate with each other to form a tri-snRNP prior to binding the pre-mRNA, with the U4 and U6 components binding through complementary

base-pairing of their snRNAs while the U5 component associates with U4 indirectly via their respective protein subunits (Nguyen *et al.* 2015). Binding of the tri-snRNP to the mRNA forms *complex B* and marks completion of spliceosome assembly, allowing the splicing reaction to proceed.

#### 2.4.4 U11 and U12 subunits

The minor spliceosome U11 and U12 snRNPs serve an analogous role to the U1 and U2 snRNPs of the major spliceosome, with their snRNAs base-pairing with the donor motif and branch point motifs, respectively. However, unlike U1 and U2, the U11 and U12 snRNPs associate with each other in a di-snRNP prior to binding the pre-mRNA (Jutzi *et al.* 2018). The U12 snRNA also has substantially less pseudouridylation than the U2 – just two nucleotides compared to the U2 snRNA's thirteen (Zhao *et al.* 2018), although the reason for this difference is not yet known.

#### 2.4.5 U4atac, U5 and U6atac subunits

The tri-snRNP of U4atac, U5 and U6atac functions similarly to the major spliceosome U4-U5-U6 tri-snRNP and has similar secondary and tertiary structure, despite there being substantial differences in the primary sequences of the U4 vs. U4atac and U6 vs. U6atac snRNAs (Patel and Steitz 2003).

### 2.5 ESEs and ESSes

An additional contributing factor to the definition of exons is the highly diverse class of proteins called exon splicing enhancers (ESEs) and silencers (ESSes). These bind to certain sequence motifs in RNA and either promote or inhibit the surrounding sequence from being spliced as an exon. Unlike the core elements of the spliceosome, there does not appear to be any single enhancer that guarantees exon inclusion upon binding, nor any one silencer that guarantees exclusion. Rather, their effect is cumulative and competitive; even constitutively spliced exons will typically contain at least a few silencer motifs. There are numerous reports in the literature of mutations that pathogenically alter splicing by altering the ratio of enhancer motifs to silencer motifs within exon-like sequences (Keegan *et al.* 2022).

Although it is beyond the scope of this review to discuss the characteristics of every known splice enhancer and silencer, it is worth examining the common features of the two largest families of splice modifiers: the serine/arginine-rich splicing factors and heterogenous nuclear ribonucleoprotein particles.

### 2.5.1 SRSFs

Serine/arginine-rich splicing factors (SRSFs) are proteins that belong to the eponymous SRSF protein family, so named for their content of serine and arginine repeats. Although numerous proteins are referred to as serine/arginine rich in older literature, currently only twelve are officially classed as SRSFs (SRSF1 to SRSF12), with the numbering reflecting the order of their discovery (Manley and Krainer 2010). Examples of non-SRSF serine/arginine-rich proteins include SFSWAP, TRA2B, SREK1, SUGP2, SCAF4, and CLASRP (Stelzer *et al.* 2016).

As well as their serine/arginine repeat domains, all SRSF proteins are additionally characterised by the presence of RNA-binding motifs at their N-termini (Jeong 2017), while the C-terminal SR-domains define interactions with other proteins. Generally, this entails recruitment of spliceosome snRNPs and other components if the SRSFs bind ESEs within exons, (Zhou and Fu 2013) or inhibition of splicing if they bind ESSEs in the intron surrounding a putative exon (Wagner and Frye 2021), although there are exceptions (Zhou *et al.* 2020; Wagner and Frye 2021).

### 2.5.2 hnRNPs

Heterogenous nuclear ribonucleoprotein particles (hnRNPs) are a highly diverse protein family with numerous important functions in post-transcriptional RNA processing, and the majority of hnRNP proteins play some role in RNA splicing (Geuens *et al.* 2016).

Like SRSF proteins, hnRNP proteins are also characterised by possessing one or more RNA recognition motifs that allow them to bind to specific RNA sequence motifs. Many are also enriched in glycine and tyrosine (GY) rich motifs that are implicated in the assembly of multi-protein structures (Guerussov *et al.* 2017),

analogously to SRSF serine/arginine rich regions. But unlike SRSFs, hnRNP proteins are generally characterised as exon silencers, especially of alternatively spliced exons (Guerussov *et al.* 2017), and act antagonistically to SRSFs (Pozzoli and Sironi 2005; Busch and Hertel 2012; Rahman *et al.* 2015).

## 2.6 Functional deep-intron splice motifs

The classical model of splicing characterises splicing solely as the removal of whole introns from immature RNA, coupled with ligation of the flanking exon ends.

However, the last two decades of research have seen steady growth in the body of evidence supporting the presence of functional splice motifs deep within introns, in regions previously thought to be dispensable to correct gene expression.

### 2.6.1 Recursive splicing

Recursive splicing is the name given to a form of splicing whereby an intron is removed from a maturing RNA in segments, through multiple successive splicing reactions rather than as a single lariat. Recursive splicing was first observed in *Drosophila* RNA (Burnette *et al.* 2005) but has since been shown to occur in other eukaryotes and in many human gene transcripts (Georgomanolis *et al.* 2016; Sibley *et al.* 2015; Sibley *et al.* 2016). It has been theorised that recursive splicing evolved to more efficiently process very long introns (Zhang *et al.* 2018) and/or long transcripts, though it may have other regulatory roles that have yet to be determined.

Although varying models of recursive splicing have been proposed (Pulyakhina *et al.* 2015; Gazzoli *et al.* 2016; Blazquez *et al.* 2018), the most commonly accepted involves the use of intronic ‘ratchet points.’ These strongly resemble acceptor site motifs and are processed in much the same way as an acceptor site would be, but subsequently re-splice from the reconstituted junction, effectively forming a zero-nucleotide exon. Although this implies that the acceptor-like ratchet point might not need a nearby 3′ donor site as a canonical internal exon would, evidence indicates that recursive splice sites do require a *potentially* spliceable 3′ donor site in order to properly recruit the spliceosome (Blazquez *et al.* 2018; Maita and Nakagawa 2020),

but this donor is ordinarily repressed from directly participating in splicing (Blazquez *et al.* 2018).

### 2.6.2      Poison exons

Poison exons appear to be spliced via the same motifs and mechanisms as canonical exons but serve a distinctly different role in the cell. While canonical exons in mRNA transcripts either encode peptide sequence or form part of the 5' or 3' regulatory regions, poison exons contain one or more premature stop codons. Consequently, their inclusion in the mature transcript typically prevents it from being translated to a functional protein and triggers its degradation by nonsense-mediated decay. Poison exons therefore exist as a regulatory mechanism for controlling gene expression.

Poison exons have been discovered in multiple genes, particularly genes important in neurodevelopment (Carvill and Mefford 2020; Aziz *et al.* 2021), tumorigenesis, and splicing (Leclair *et al.* 2020). However, an ongoing challenge in identifying new poison exons is empirically distinguishing the 'true' poison exons from random mis-splicing events, since many of these mis-splicings also produce exon-like inclusions with premature stop codons. Some studies use sequence conservation to screen predictions, based on the assumption that more conserved sequence is more likely to have a beneficial function, but empirical studies that knockout or knock-down splicing of suspected poison exons in live cells (Leclair *et al.* 2020; Thomas *et al.* 2020) are essential to confirm their role.

### 2.6.3      Decoy exons

Like recursive splice sites and poison exons, decoy exons are also comprised of deep intronic exon-like sequences. However, the function of decoy exons is not to be spliced themselves but to promote retention of the surrounding intron as a regulatory mechanism. As proposed by Parra *et al.* (2018), decoy exons achieve this by binding to U2AF1 and U2AF2 spliceosome components and thereby competing with the splice signal of the flanking canonical exon splice sites (see also Howard *et al.*

2018). The model of Parra *et al.* also allows for the same sites to serve a dual role as poison exons under some splicing contexts.

Decoy exons represent an even greater challenge to experimental verification than poison exons since their function is to be *almost* spliced, but nonetheless may prove to be a worthwhile field of investigation in the search for new targets for antisense splice-modulating therapies (Parra *et al.* 2020). In particular, the TANGO (Targeted Augmentation of Nuclear Gene Output) platform aims to treat haploinsufficiency genetic diseases by therapeutically suppressing poison exon inclusion in transcripts of the functional allele (Han *et al.* 2020).

## 2.7 Splicing mutations

Much of pre-mRNA splicing is directly or indirectly dependent on motifs present in the primary sequence of the pre-mRNA molecule being spliced, and most of this sequence is directly encoded by corresponding chromosomal DNA sequence. Consequently, mutations that cause single-nucleotide changes in the DNA can have much larger effects on pre-mRNA maturation if they alter or create an important splicing motif. The complexity of splicing means that there is a broad range of splicing changes that a mutation can instigate, some of which are difficult to predict.

Wimmer *et al.* (2007) originally proposed five categories of splicing mutations:

“Classical splice-site mutations leading to exon skipping (type I); single-nucleotide changes within introns, creating *de novo* splice sites leading to inclusion of a cryptic exon (type II); single-nucleotide changes within exons, creating *de novo* splice sites whose use results in the loss of a part of the exon (type III); mutations disrupting use of the splice site and resulting in activation of cryptic exonic or intronic splice sites (type IV); and exonic sequence alterations causing exon skipping (type V).”

Other researchers have since expanded upon and refined these categories (Abramowicz and Gos 2018) or independently described similar ones (Scotti and Swanson 2016). We suggest that type II mutations (cryptic/pseudoexon creation) are

unique within this group, since all other types entail modification to the splicing of an existing, canonical exon – either skipping the exon or altering one of its splice sites – while type II mutations entail creation of a (supposedly) completely new exon in a region with no previous canonical splicing activity.

The effect of splicing mutations on the mature transcript typically equates to the addition or subtraction of sequence, a change in overall splicing speed/efficiency, or some combination of these three. Because mRNA codons are three nucleotides long, sequence changes within mRNA coding regions have a two-in-three chance of causing a shift in the open reading frame, typically leading to a premature stop codon, and sequence additions may also introduce new stop codons directly. In either case, the premature stop codons are likely to induce NMD of the transcript (Nickless *et al.* 2017) and result in a decrease in expression of the gene.

Even if the reading frame is preserved, gain or loss of multiple amino acids in the middle of a protein is generally likely to be deleterious to its function, though there are numerous reported cases of frame-preserving abridging mutations having comparatively milder effects than frameshift mutations. For example, many cases of Becker muscular dystrophy, a milder form of Duchenne muscular dystrophy, arise from frame-preserving whole exon deletions in the *DMD* gene that allow translation of an abridged dystrophin isoform of variable functionality (Nicolas *et al.* 2015); and a genomic deletion of exon 3 from the *Growth Hormone Receptor* gene (*GHR*), equating to a loss of 22 amino acids from the full-length protein, is apparently mild enough in effect that it has become a common allele carried by half of the general population (Filopanti *et al.* 2011), and has even been shown to correlate with increased height and longevity in males (Ben-Avraham *et al.* 2017).

## 2.8 Therapeutic strategies to manipulate splicing

The last decade has seen an explosion of interest in drugs that can therapeutically alter pre-mRNA splicing in patients with genetic diseases. Exon-skipping antisense oligonucleotides have been the most widely pursued and most successful strategy, with four such molecules – Exondys 51 (Syed 2016), Vyondys 53 (Heo 2020),



Viltepso (Dhillon 2020), and Amondys 45 (Shirley 2021) – achieving the status of approved drugs for the treatment of Duchenne muscular dystrophy. Another early success story for splice-modulating antisense oligomers was Spinraza, a drug that promotes inclusion of an exon in *SMN2* transcripts for the treatment of spinal muscular atrophy (Prakash 2017).

Antisense oligonucleotides (AOs) are synthetic molecules with a similar molecular structure to RNA and DNA. They are designed to bind to specific sites within their target molecules – usually pre-mRNA exon or splice-motif regions in the case of exon-skipping AOs – via reverse-complementary base pairing (Li *et al.* 2018). This binding does not directly alter the pre-mRNA sequence, but through a combination of steric blocking and/or modifying the RNA's secondary structure (Kole *et al.* 2012) it affects which splice factors and other spliceosome components can bind to the exon region, thereby decreasing its frequency of inclusion in the mature transcript.

Antisense-oligonucleotide mediated exon-skipping has repeatedly demonstrated efficacy as a therapeutic strategy, and over the last decade much has been learned about optimising AO design (Le *et al.* 2017). Nonetheless, for each new AO the path from hypothesis to approved therapy is still long and fraught, as it cannot yet be predicted with any great accuracy how easily a given exon can be skipped, or what the effects of very slight changes in AO design (e.g. altering the AO length by a single base) will be. Investigations of the aetiology behind some of the more unusual exon-skipping AO effects will do much to close these gaps in our knowledge.

In addition to AOs, other constructs and small molecule libraries are increasingly being investigated as augmenters of AO efficacy. Small molecules that improve AO trafficking into and within cells have exhibited synergy with exon-skipping AOs when used in combination, (Yang *et al.* 2015; Dang *et al.* 2021), as have modified U1snRNAs (Breuel *et al.* 2019). We speculate that small molecules that modify specific RNA secondary structures (Costales *et al.* 2020) may also improve the efficacy of some AOs by relaxing RNA self-binding in and around the AO target site.

## 2.9 Summary

Splicing of pre-mRNA is a dynamic process, and RNA represents the first stage at which the unique aspects of a cell's genotype begin to manifest as a phenotype. Understanding how pre-mRNA matures to mRNA, and how this process can be altered by both mutations and antisense oligonucleotides, is therefore a task as important to human health as it is challenging to the health science researcher.

## **Chapter 3**

# **Pseudoexons of the DMD Gene**

### 3.1 Preamble

In this report, I catalogued and examined all known (at time of writing) examples of pseudoexons (PEs) and cryptic exons within a single gene, *Duchenne Muscular Dystrophy (DMD)*. Although PEs have been reported in many other genes, *DMD* was chosen as the exemplar for this report since it appeared to have had more described PEs than any other gene. Herein, I observed that mutations that increase *DMD* PE splicing primarily do so by enhancing PE splice motifs, and that the most common type of causative mutation is single nucleotide substitution. However, there were also several cases where observed mutations were substantially distal to the PEs they enhanced, and some PEs that were spliced into a small fraction of transcripts in the absence of any causative mutations.

There is increasing evidence that the transcribed introns of many human genes undergo a process of “recursive splicing,” whereby the intron is removed from the maturing transcript in successive segments rather than as a single piece. The deep intronic boundaries of these segments are sometimes referred to as “ratchet points,” and appear to engage with the spliceosome via similar sequence motifs to those seen in and around canonical exons (Sibley *et al.* 2015; Gazzoli *et al.* 2016; Georgomanolis *et al.* 2016). Based on prior observations suggesting that some PE splice sites may also be ratchet points for recursive splicing (Georgomanolis *et al.* 2016), I cross-searched the co-ordinates of the collated *DMD* PEs against a published dataset of putative recursive splice sites in *DMD* introns (Gazzoli *et al.* 2016), discovering 20 novel matches and re-confirming six additional matches discovered by Bouge *et al.* (2017). Interestingly, the PEs that matched to these sites were predominantly those with either distal mutations or no associated mutations.

This led me to postulate that there may be *two* distinct pathways to PE pathogenesis – the commonly observed proximal mutations that directly enhance the exonic motifs of PEs, and distal mutations that disrupt regulatory elements and cause ratchet points to be mis-spliced as PEs. However, the small size of the *DMD* PE dataset offered few clues as to what these regulatory elements might be, except that they predominantly seemed to occur 3′ of the PEs. I concluded that a more

comprehensive analysis of PEs in other genes could reveal more about the mechanisms underlying PE pathogenesis.

Although the scope of this report was limited to a single gene, its findings were not only novel in themselves but also provided an essential basis for the more expansive work on PEs that would follow.

### 3.2 Citation

Keegan, N.P. (2020). Pseudoexons of the DMD Gene. *J Neuromuscul Dis* 7(2), 77-95. doi: 10.3233/JND-190431.

## Review

---

# Pseudoexons of the DMD Gene

Niall P. Keegan\*

*Centre for Molecular Medicine and Innovative Therapeutics, Murdoch University and Perron Institute, Perth, Australia*

**Abstract.** The *DMD* gene is the largest in the human genome, with a total intron content exceeding 2.2Mb. In the decades since *DMD* was discovered there have been numerous reported cases of pseudoexons (PEs) arising in the mature *DMD* transcripts of some individuals, either as the result of mutations or as low-frequency errors of the spliceosome. In this review, I collate from the literature 58 examples of *DMD* PEs and examine the diversity and commonalities of their features. In particular, I note the high frequency of PEs that arise from deep intronic SNVs and discuss a possible link between PEs induced by distal mutations and the regulation of recursive splicing.

Keywords: Cryptic Splice Sites, Muscular Dystrophy, Duchenne, DNA Mutational Analysis, RNA Splicing

## CHARACTERISTICS OF THE DMD GENE

The *DMD* gene is the largest gene in the human genome. Situated on the *p*-arm of the X chromosome, *DMD* spans over 2.22Mb, more than 99% of which is intronic sequence, with the coding sequence of its largest isoform totalling 11,058 bases across 79 exons. Eight unique alternative promoters [1], alternatively spliced exons, and an alternative polyadenylation site [2] produce at least 17 *DMD* transcript variants [3], one or more of which are expressed and translated in all types of muscle as well as various other cell types throughout the body, including myoblasts, lymphocytes and retinal cells.

The *Dp427m* transcript of *DMD* encodes the muscle isoform of dystrophin, the *DMD* protein. In XY individuals, who carry just a single *DMD* copy, mutations that fully disrupt the function of the *DMD* gene (resulting in functionless or absent dystrophin protein) give rise to Duchenne muscular dystrophy, while mutations that only partially diminish the gene's function and/or quantity of product give rise to Becker Muscular Dystrophy.

---

\*Correspondence to: Niall P. Keegan, Building 390, Discovery Way, Murdoch University, WA, Australia. Tel.: +61 8 9360 6058; E-mail: npkeegan@me.com.

## THE MAJOR SPLICEOSOME

The vast majority of RNA splicing in humans is achieved via the major spliceosome, a ribonucleoprotein complex responsible for excising introns from pre-mRNA molecules [4]. In order for the spliceosome to process a transcript correctly it first must accurately recognize the transcript's exon-intron boundaries. This recognition is achieved through a network of mechanisms, including sequence-specific interaction with conserved acceptor and donor splice site motifs in the RNA, silencer and enhancer binding motifs both proximal and distal to the splice junctions, and RNA secondary structure [5, 6]. Mutations to a gene that alter the interactions of these factors with its transcripts can lead to errors in the processing of those transcripts, such as the expansion, truncation or loss of canonical exons, or the initiation of pseudoexons (PEs) within its introns [7]. These incorrectly spliced transcripts may be degraded prior to translation or may be translated to less functional or even harmful protein isoforms, with deleterious consequences for the health of the patient.

Thirty-six of the 78 introns in *DMD* are more than ten times the human median intron length of 1334bp [4], and of these 36, three are more than 100 times the median size. A transcript of this size and complexity

presents a unique challenge to the major spliceosome, and as a result is arguably more vulnerable to splicing errors such as pseudoexons.

### PSEUDOEXONS: ONE DEFINITION OF MANY

In the literature, the terms “pseudoexon” and “cryptic exon” are often applied interchangeably and inconsistently in reference to a wide range of splicing errors. For the sake of clarity, I hereby define a pseudoexon as: *Any continuous tract of a transcribed gene that: (1) does not overlap, adjoin or duplicate any sense-strand sequence of that gene’s canonical exons; (2) bears an acceptor splice site motif at its 5’ end and a donor splice site motif at its 3’ end; and, (3) via both these motifs, is spliced into a measurable proportion of the mature transcripts of that gene in at least one proband.*

Though this definition of “pseudoexon” may not agree with every prior usage of the term, it includes the majority of prior use while excluding splicing events that are better described by other terms, such as cryptic splice sites and whole exon duplications.

While some PEs are observable as rare splicing events in normal individuals, the majority are created by mutations that give the PE site an exon-like profile, resulting in the spliceosome falsely recognizing it and splicing it into an increased proportion of transcripts. When PE-splicing levels are high compared to normal splicing, these inclusions are likely to bear negative consequences for the phenotype of the affected organism, as the majority of PEs will disrupt the transcript’s open reading frame and/or encode premature stop codons. Consequently, the resulting transcript, if it is not degraded by nonsense-mediated decay, will be translated to a non-functional or truncated protein. Even in cases where a pseudoexon preserves the reading frame and does not encode a premature stop codon, it is likely that the amino acids it encodes will disrupt the secondary structure of the protein and thereby abrogate its function.

Numerous *DMD* PEs have been reported over the last few decades, perhaps more than have been described for any other single gene. When considered as a body of research, these reports comprise a unique opportunity for generating new insights into the splicing of *DMD* and other large genes.

In this review, I catalog the characteristics of 58 reported *DMD* PEs. Where possible, I describe the

origins of these rare splicing events and draw inferences from their common features.

### PSEUDOEXONS OF THE DMD GENE

Following a thorough search of the literature, I compiled a catalog of all 58 known PEs of the *DMD* gene (Table 1). In order to consistently record highly similar PEs, I adopted the criteria of *unique local sequence*: I assigned a separate catalog entry to each PE that was unique in at least one nucleotide of its sequence or splice motifs (eg. PE09 vs. PE10, PE15 vs. PE16); and listed as single entries all PEs with locally identical sequence (eg. PEs 11 and 12).

### PSEUDOEXONS ARISING WITHOUT MUTATIONS

Six *DMD* pseudoexons have been reported as low-frequency splicing events in normal cells lacking any known mutation: PEs 04, 07, 08, 11, 21 and 44. Interestingly, four of these PEs are of lengths that do not shift the reading frame of the transcript – 162bp for PE04, 357 bp for PE11, 66bp for PE21 and 84bp for PE44 – and of these four, only PEs 04 and 11 contain stop codons.

In addition to their splicing profile in normal cells, PEs 04, 07, 08 and 11 are also spliced at much higher frequencies in the cells of some patients with other *DMD* mutations. For PEs 08 and 11, this behavior was observed for the cells of only a single patient each (see below, subsection ‘Pseudoexons arising from duplications’). However, the behavior of PEs 04 and 07 – referred to in some prior literature as exons 1a [10–14] and 2a [16, 17] respectively – is somewhat more complex. Though the inclusion of PE04 in muscle cell *DMD* transcripts is rare [14], this PE is included in approximately 50% of *DMD* transcripts in lymphocytes [11], and is included at higher frequencies in both cell types as a result of a frame-shifting single nucleotide deletion in exon 5 [12] and, in a different proband, an exon 2 tandem duplication [13]. Given the frequency of its inclusion in mature *DMD* transcripts, especially in normal lymphocytes, it may be that PE04 would be better classified as a canonical, alternatively spliced exon rather than a pseudoexon. However, at the time of writing no functional role for PE04 has been conclusively determined [10]. Similarly, PE07 has also been observed as a predominant inclusion in the muscle *DMD* transcripts of multiple patients with deletions and duplications in the 5’

Table 1  
 Details of 58 known pseudoexons (PEs) arising from mutations within the human *DMD* gene. Bold, upper-case sequence denotes the PE itself, while regular lower-case sequence indicates flanking intron. Inverted sequence is italicized. Duplications and insertions are underlined. Single base transitions are shown in the form of "(X > Y)," where X is the reference nucleotide and Y is the mutant. Shapiro-Senapathy splice scores are also shown, before and after mutation where appropriate (see supplementary materials for splice score calculator). "RNA Source" lists the cell or tissue types where the PE was detected. Genomic co-ordinates pertain to human genome assembly accession GCA\_000001405.27, chromosome X, except where otherwise stated. Transcript co-ordinates pertain to *DMD* reference sequence NG\_012232.1(DMD\_v001)

PE #	Intron Sequence	Genomic Co-ordinates (chrX)	Size (bp)	Acceptor Score	Donor Score	Mutation(s)	Orientation of Mutation to PE and Splice Sites	ORF	RNA Source	Reference
<b>01</b>	1	33174185-33174333	149	71.37 -> 87.94	72.72	c.31+36947G >A	Proximal	Frameshift	Skeletal muscle	[9]
	<i>tctctcccttggttttgc(g&gt;a)gC</i>									
	<b>TTCTCAGATTCA TAGGAGACTTTCA</b>									
	<b>GTTCAGTGACCTGGAACTCAC</b>									
	<b>ATTCCTATCACCACTCTTTACTGT</b>									
<b>02</b>	1	33088244-33088348	105	62.92	81.59	c.exon3_6del	Distal (3')	Premature stop codon	Heart, skeletal muscle	[10]
	<i>AGTAACTTCCTTTTACCTGACCACC</i>									
	<b>CTGCATAGTCACAGAAGATGCAC</b>									
	<b>CTGACAAATGATCCTCAAAACAGgt</b>									
	<i>agtaatctcttttgagaag</i>									
<b>03</b>	1	33085901-33085927	27	58.89	78.81	c.exon3_6del	Distal (3')	Premature stop codon	Heart, skeletal muscle	[10]
	<i>ctcaattaaataatatttagTCAAA</i>									
	<b>GCCAGATGCAGTGGCTCACACTGT</b>									
	<b>GATCCAGAACTTTGGAGGGGGAG</b>									
	<b>GCGGATGGATCATTTGAAGTCAGGA</b>									
<b>04</b>	1	33078186-33078347	162	96.38	88.39	Multiple mutations, also observed in normal cells (see main text)	Distal (3')	Premature stop codon	Lymphocytes	[10 - 14]
	<i>gtgagaccctgtctctact</i>									
	<b>CACTAAATTTAGGAGTGCATAgta</b>									
	<i>agtattaccatagtgta</i>									
	<b>CATGGAGAAATGGTCAFTTTACC</b>									
<b>05</b>	2	33014502-33014547	46	68.17 -> 84.74	82.07	c.93+5590T>A	Proximal	Frameshift	Lymphocytes	[15]
	<i>tctgtttctttttgttaacacagGACAC</i>									
	<b>AAGGCTTTGACTGGAAATGGCATGCT</b>									
	<b>TCCTTTAAAGAAFCAAAAGTTGACTT</b>									
	<b>ATAGAGCCATTTAAAGCCCGTTGGG</b>									
<b>GAATCGCCCTCATACCTTTGTFCCAC</b>										
<b>ACAGATCCCTGTACAAGGTTCTCTG</b>										
<b>ACCITGTGtaagtaaagaatgtcac</b>										
<i>tt</i>										
<i>tctcctgtttctctacat(t&gt;a)gg</i>										
<b>TTGAAFTCTGTTCTGCAGCAACTAG</b>										
<b>TAACCCCAACAATCTGCAC</b>										
<i>ggtgag</i>										
<i>aaattcaatctccgt</i>										

(Continued)



Table 1  
(Continued)

PE #	Intron	Sequence	Genomic Co-ordinates (chrX.)	Size (bp)	Acceptor Score	Donor Score	Mutation(s)	Orientation of Mutation to PE and Splice Sites	ORF	RNA Source	Reference
06	2	tccctgtttctctctacat (t>a)gG TTGAATCTGTTCCTGCAGCAACTAG TAACCCCAACAATCTGACTGGTGAG AAATTCAAATCTCCGTAACAAATTTGC ATTCTTCTACTTCTCCCACTTCCA TTCCAAAGGGCCAGGTAGTCCAG AAACTGtaagtgctcctggttttca g	33014416- 33014547	132	68.17-> 84.74	77.7	c.93+5590T>A	Proximal	ORF preserved	Lymphocytes, skeletal muscle	[15]
07	2	tgttagtggttttttggttcagGTCCTG TGCAGGCCAAGGTTATGGCCCTAGCT GAGAAGAGGGCTCAGAGAGCCCTAG CTAAAGTTTGGTCAAGGGGAGCGTC TTGGTGGAGCCTCACTCCTGTTCAA GTAAGAAGACACAATATTCATCCT TCTGAAACAGTgtaagtcctattgtc atcg	32960208- 32960347	140	100	77.84	Multiple deletions and duplications [17]; also observed in normal cells [16]	Distal (3')	Frameshift	Skeletal muscle	[16, 17]
08	2	agttctggctctgctcctgtagGCTAG AGTGCAGTGGTGTCTCTCAACTCA ATGCCAACCTCCGCCCTCCGGGTTC AGCAACTCCTGCTCAGCCTCCC CAGTAGCTGGGATACAGgtacctg ccaactgtgcccgg	32878883- 32878980	98	72.84	76.65	c.exon2dup; also observed in normal cells	Distal (3')	Frameshift	Multiple tissue types	[13]
09	2	tgaattggaaacttccctgtagACAGA CCCTTACAGGCATGGAAGAGAAAG AATGAAATAACCAAGGATGACTTCC CAGTAGGTGGAGGATGGGAATAAT AGGAAAGAGCTGTCTTTGTCACCT ATAATTGCCATACACAACATGCACCC TAGGGTAATTTGAGAAATTAAGAA TGgtaataaatacgtttttacat ttggaacttccctgtagacagACCCCT TACAGGCATGGAAGAAAGAAATG AATAAACCAAGGATGACTTCCCAGT AGGTGGAGGATGGGAATAATTAGGA AGAAGCTGTCTTTGTCAACCTATAT TGTCATACACAACATGCACCCGTAGG GTAATTGAGAAATTAAGAAATGgt aataaatacgtttttacat	32863759- 32863915	157	70.85	73.68	c.exon8_1 dup	Distal (3')	Frameshift	Lymphocytes	[18]
10	2		32863759- 32863911	154	79.35	73.68	c.exon8_1 dup	Distal (3')	Frameshift	Lymphocytes	[18]

11	3	aatcttactctgtgcccagGCTGG AGTCAGCGACATGATCTTTGGCTCAC TGCAACCTCTGCCCTCTGGGTTCAA GCAATTTTCCCTGCCCTCGGCCTCCCA AGTAGCTGAGACTACAGGCATGTC CATCATGCTGGCTAAGTTTGTAT TTTGTAGTAGACATGGGGTTTCACCA TGTTGGCCAGGCTGGTCTTTGAGCTC CTGATCTCAATGAGCCACCTGCCCTC AGCCTCCCAAATGCTCGCTGTTG CTTTTTTTTTTTTTTTTTTTTT AAATCTAAAGTCTTATTTTCTCCT CTTTTGGTGGAAAGTGGGAGAAAT ATCAGAAATGTAACCAACATCATTC TGACATCTTTGGAGGAAAATCTAAG AGgtaaaaatggagattctctgt ttcctctttttggtggaagtGGGA GAAATATCAGAAATGTAACCAACAT CATCTTGACATCTTTGGAGAAAAT CTAAGAGGtaaaaaatggagattctctg gt	32846622- 32846978	357	81.06	79.88	c.exon8.1 dup; also observed in normal cells	Distal (3')	Premature stop codon	Lymphocytes	[18]
12	3	ttcatatgcttctgttttttcagATGCT TGtGTTTTAACTTTACTCCACCTTAA ACAATTTGAGGAGTGTGAAGGCACAGG AGACACAGAGATTTGCCCTTGAATTAA GGCAAAATAAAACCCCTGCCAGATTTT CATTTCCAAACACAGTCTCTAGACAG AG (a>g) taagagagctggcagttt g	32846622- 32846683	62	72.62	79.88	c.280A del [12], c.exon 8.11 dup [18]	Distal (3')	Frameshift	Lymphocytes	[12, 18]
13	4	atattcttaattgaaatttcag (A>C) GTGAATTTCCACACTCTCCTTTTGA AAGATTCATTTCTATGAAATTTGGGA CAGCTTCCCTAGTATGATATCCATC T (a>g) taagtatatccatcatc ttcaatttttccattta (a>c) aga GATGATCATATTTGGGATAAAGAGCG TGTTTTGGAAATCCAAACCAACCTGGT TTTAAAGTCCCAGAACCCACAGTTAC CTTTGTGACCTTTTggaagtcatct aatTTTTct	32823851- 32823982	132	90.56	74.59 -> 91.82	c.265- 463A>G	Proximal	Premature stop codon	Skeletal muscle	[19]
14	7	atattcttaattgaaatttcag (A>C) GTGAATTTCCACACTCTCCTTTTGA AAGATTCATTTCTATGAAATTTGGGA CAGCTTCCCTAGTATGATATCCATC T (a>g) taagtatatccatcatc ttcaatttttccattta (a>c) aga GATGATCATATTTGGGATAAAGAGCG TGTTTTGGAAATCCAAACCAACCTGGT TTTAAAGTCCCAGAACCCACAGTTAC CTTTGTGACCTTTTggaagtcatct aatTTTTct	32738792- 32738868	77	79.91 -> 78.04	56.78 -> 74.00	c.650- 39575A>C; c.650- 39498A>G	Proximal	Frameshift	Skeletal muscle	[20]
15	9	atattcttaattgaaatttcag (A>C) GTGAATTTCCACACTCTCCTTTTGA AAGATTCATTTCTATGAAATTTGGGA CAGCTTCCCTAGTATGATATCCATC T (a>g) taagtatatccatcatc ttcaatttttccattta (a>c) aga GATGATCATATTTGGGATAAAGAGCG TGTTTTGGAAATCCAAACCAACCTGGT TTTAAAGTCCCAGAACCCACAGTTAC CTTTGTGACCTTTTggaagtcatct aatTTTTct	32650985- 32651074	90	83.55 -> 93.28	70.04	c.961- 5925A>C	Proximal	Premature stop codon	Skeletal muscle	[21]

(Continued)

Table 1  
(Continued)

PE #	Intron	Sequence	Genomic Co-ordinates (chrX.)	Size (bp)	Acceptor Score	Donor Score	Mutation(s)	Orientation of Mutation to PE and Splice Sites	ORF	RNA Source	Reference
16	9	ttcattttcttctacctaagAGATT GATCATATTGGGATAAAGACGTGTT TTGGAATCCAACAACCTGGTTTTA AGTCCAGAAACCACAGTTACCTTT GTGACTTTGg (c>t) aagtcactc aattttctc ctttctctctctacctaagACAGG GTTGGATAGATCCAGTCGGAAGCC ATTATTCCTGCCFAAGCC[del; 11631bp]AAACCCTTTCTACAGA AATGTAAGGGCGTTCATCT TAAAGAAAAGGATGTTAATGAGCAA TGAGTCATCATCTGAAGTACAAAA CTCACTGGGATAGTAAGgtttggt actattctggttac ctctttttttttttcccccaagTGTCCT ATTTGACTCTGGAAATAAGATGGCA TATGTGAGAGTGGATAGAGAAAAGG AGGTGCTGAACAAATGAGTGGTTA TTTTCTCCCAAGTGCAGTGAAGTTT CCGTAATGTAATATATGAATGAATA GATGAACAACATAAATAAGTTATTC AGgtaattccaaatcttggaatt agttttgttttttcccccaagGCTGG AGTGAAGTGGTGTGATCTCAGCTCA CTGCAACCTCTGTCCCCCGGGTTC AAGTGAATTCCTGCCTCAGCCTC ( c>g) Tgagtagccccggaattaaaa tttttttccccctgcttagTGGAA GAGGCTATCTTCCCTGCTGTGAGG CTCATCTCTGGGTGTGTCTCAG CACATCATCATATGTAATAAGC CACCTGGTCCATTCAGCTGTATAT CCAGATTGTCAAAAATTCATACATCCC AGgttttttattcattagctttta	32650985- 32651074	90	83.55	70.28 -> 87.27	c.961- 5831C>T	Proximal	Premature stop codon	Skeletal muscle	[9]
17	11		32641751- 32641799; 32630010- 32630119	159	73.92	79.54	c.1331+2382- 1331+14010 del	Proximal	Premature stop codon	Multiple tissue types	[22 - 25]
18	11		32638732- 32638888	157	79.76	87.83	c.1336.1337del	Distal (3')	Frameshift	Lymphocytes	[26]
19	11		32626363- 32626441	79	90.08	56.56 -> 73.78	c.1332- 11909C>G	Proximal	Frameshift	Skeletal muscle	[27]
20	18		32514091- 32514222	132	68.21	67.88	c.2622+1G>A	Distal (3')	Premature stop codon	Lymphocytes	[10]

21	21	32480880-32480945	66	80.03	78.22	None	N/A	ORF preserved	Skeletal muscle	[14]
		tttatttttaaacatctctagGCAAA GGAATGTTTGTTCAGTAGACACA TAAATCTGTTCATGCTCCTTCACT CCAGAAAACTAGgtaaaactgtttgta aatggt								
22	25	32461308-32461402	95	79.15-> 95.72	70.75	c.3432+ 2036A>G	Proximal	Frameshift	Skeletal muscle	[28]
		tatctgtgcttttcttaca(a>g)G TATCACTCTGGCCATGTTCTGACTTT TGTAGCCAAATAGTGTAGGTTGTAA AAGGAAGAACAAATGGCGCTCAAGG AGAAGAAGAACGATCGGgtaa								
23	25	32461200-32461401	202	76.71	70.14-> 87.36	c.3432+ 2240A>G	Proximal	Frameshift	Skeletal muscle	[29]
		aaacaaggaaagccata atctgtgcttttcttacaagTATCA CTCTGGCCATGTTCTGACTTTGTAG CCAAATGAGTTAGTTGTTAAAAGGA AGGAACAATGGCGCTCAAGGAGAA AAGAAGACGATCGCGTAAAACAAG GAAGCCATATGTGAATATGTTTACC AATTCAGCATCCAGAGAGAAATAT GGAAATGAAGTGTAAATCTATGCAT TACAGAAATATCTACAGACAAA (a> g)taagtggtgatacactct								
24	25	32461200-32461371	172	82.72	70.14-> 87.36	c.3432+ 2240A>G	Proximal	Frameshift	Skeletal muscle	[29]
		gccatgttctgactttgtagCCAAA TCAGTTAGGTTGTTAAAAGGAAAGAA CAATGGCCCTCAAGGAGAAAGAA GACGATCGGGTAAAACAAGGAAGC CATATGTGAATATGTTACCAATTC AGCATCCAGAGAGAAATATGGAAA TGAAGTGAATCTATGCATTTACAG AAATATCTACAGACAAA (a>g)taa								
25	26	32452550-32452629	80	86.2	77.21	c.3603+ 2053G>C	Proximal	Frameshift	Skeletal muscle	[30]
		gtgtgtgatacactct cttctctctgcttatcagAAAAC TGCAATFCCCAGATA (G>C)GTCAA ATGATTTAGCCATAGTCAAGACTT TATTTGTGGTAGAGCCACAGGATT GAAGgtattttattcttattctc								

(Continued)

Table 1  
(Continued)

PE #	Intron #	Sequence	Genomic Co-ordinates (chrX)	Size (bp)	Acceptor Score	Donor Score	Mutation(s)	Orientation of Mutation to PE and Splice Sites	ORF	RNA Source	Reference
26	27	cataatctttagtggttacagATGTT GCAGTTGTTTCTCTATTTTGGT CTGTTTCAATTACTTGGTGT CCTTTGATGAATAGAAGTCTAAT TTAAATGATGCGAATGATTTCAFC TCTTTTCTTACGGgt (c>a) agtt tttaacgagatgt	32442160- 32442278	119	87.46	80.58 -> 90.39	c.3787- 843C>A	Proximal	Frameshift	Lymphocytes or skeletal muscle (unspecified)	[31]
27	29	caatcttcttctatcaacagAGCTG AATGAGTCCAGGAAGCTGGAAAT CTGCTTACAAAAAGgtgattgtgg aagagtctag	32412019- 32412063	45	90.17	83.09	c.3613delG	Distal (5')	Premature stop codon	Lymphocytes	[10]
28	34/42	caaatgttccccgtttttatagATGTT CCCGTTTATAGATGAACAAATACA AATACAATA   del1 : 78031bp   CAG CCAGAAGATATGTTGGTGCACGFT TCTGGTACCTGACCTAATCAGGCTA GCGGATAGCCCTAAACATCCACC CAAGACTCCAGGCTGGGCCATCA GAAGATGGTAGTAAAAATTTTACCAG gtaataaaaagtaagaaat gaatttgctctccaccacagATGTG ACAGACCCAGCCAAATACAAGTTGT GACCAAGACAAGTTTGTGAGTTTC ATTTACATTTGTCATGAATTTgta agtatacttcttggaat	32372085- 32372098; 32293942- 32294069	167	83.07	88.06	c.4846- 6885_6118- 6369delins ATACAATA; c.4846- 6900_4846- 6899ins17	Proximal	Frameshift	Unspecified	[32]
29	37	gtaataaaaagtaagaaat gaatttgctctccaccacagATGTG ACAGACCCAGCCAAATACAAGTTGT GACCAAGACAAGTTTGTGAGTTTC ATTTACATTTGTCATGAATTTgta agtatacttcttggaat	32360933- 32361009	77	77.72	74.58	c.5325+1740_5325+1757del	Distal (5')	Frameshift	MyoD transformed fibroblasts	[33]
30	37	tgaattgtaactatcttgca (t>g) G TATGTTGAGCTCTGGTGTGTGAAA TGTTTCTCCTTATTTGCACTCTCAG gtacttttccagttgtattt	32348692- 32348742	51	71.85 -> 88.42	77.01	c.5326- 215T>G	Proximal	Premature stop codon	Skeletal muscle	[34]
31	43	agtatcttctccaccacagGCCTG GCTTAAAGAGCTCCTGAAGGAATC ACTAAACATGGAAGGAAAACCCGG TACCAGCCACTGAGAGAACATACC AAATTTAAAGACCATCGACCTAT GAAGAACTGCCTCAACTAACAGg (c>t) aaaaataactaaaccaacat	32256577- 32256704	128	86.94	65.72 -> 82.94	c.6290+ 30954C>T	Proximal	Frameshift	Skeletal muscle	[35]

		chr4.182041664-182041743	80	90.26	99.52	Complex insertion/deletion (see cited work)	Encompassing	Frameshift	Skeletal muscle	[36]
32	43	tttttttatttattttagattttc ctgctcttcttaagacacagtgattt agaatttctgtttcagcaagagaac taagacttctttaagagaagaag gtaaagtcaagagatttaaaa tgtctccatcacccccagatggg accatgtagctcaggaaaggaagc tcaagctccactgaattctacatt atggtgagtgtataattttttc	58	76.18	86.71	c.6291-21015-6438+98743 dupins A; c.6291-21008-6291-21007ms CTCCCCTGA ACATGG	Distal (5' and 3')	Frameshift	Unspecified	[32]
34	44	tttttttttaattgcaatcagatttcc agatggacccttaagagctcagttac ggaaagtggaaaggggaaacacacag cagcagcagcaacaacaacacac aaacaacaacacacacacacacac cgaagagaaaagagactgagctgg aaaaaagtctttgaaaattttagtccct aaaaacttataattttacatttctatgc tctcctgaaatcacagaaagaggtt cggaaacttctgcttttaagaggacat ttgaa tgcga ctggagagaaacaaag cattgaaagagagagaaatgaaactc ctcagataaaagcttataatgagaa acatagttccagagagaaatgaatg ttaacttcaatttttctgctcatttc cttcaagctcaaa gca tgaagtaagt gataa tggtaagtaaatattgtattc tc	387	75.04	90.97	Complex inversion (see cited work)	Encompassing	Premature stop codon	Skeletal muscle	[37]
35	44	ttttttttttttttttttttttttttgagaggga gpttctgctcttttggccaggctgga gtgagggtggcgaatctcagctgac tgttaacctctgccccaccagggttc GAgt [del: 121167bp] aagtatg aaaaaaaaaaaa	82	85.52	62.43 -> 78.86	c.6439-55921_6912 +26400del	Proximal	Frameshift	MyoD transformed fibroblasts	[38]

(Continued)

Table 1  
(Continued)

PE #	Intron	Sequence	Genomic Co-ordinates (chrX.)	Size (bp)	Acceptor Score	Donor Score	Mutation(s)	Orientation of Mutation to PE and Splice Sites	ORF	RNA Source	Reference
36	45	ttttcttgaggatattctagGAGAA GACATACCAGTCGAGGGTTCTGGG GAGCCAGCCCTTCAAGCAATGGATT GCTGACAACAATAATGAAGAGGATTT TACTTAGAATAAATGCAGTTGATAA AAGTTTGAATGGGACACGGAAACAA GGCAGTGG (g>t) aagtggaattcc taaatt	31875031- 31875167	137	73.52	71.16 -> 88.39	c.6614+ 3310G>T	Proximal	Frameshift	Skeletal muscle	[27]
37	47	aattatggtaatccccaca (t>g)G TCAAGGTTGGCCAGGTGCAGATA ATTGAATCTGGAGAGGATCCCCC ATACTGTTCTCCTGATAATCAGtaa gtttcaacaagatctga	31879338- 31879409	72	61.61 -> 78.18	74.53	c.6913- 4037T>G	Proximal	Premature stop codon	Skeletal muscle	[27]
38	48	aaagctgctctttccgctagGGTCC TGGGATGGAAGACACATTTAGAGCAG ATCCAATTTAGAAATGAAAAGACC TTAACTTGAATTCATAGTGTGAGTA AAGCTGCAGCTGATCTGCAGACACA	31850637- 31850744	108	81.97	70.41	Complex inversion (see cited work)	Encompassing	Premature stop codon	Cultured myogenic cells	[39]
39	48	TGAgtgagtaataaagtgttggtt gacagtggtgctctatttttagGCAAT GTGAGCACGAAGACTGCTCCTCTC TTCAGGAAATGGAAATGAAATGAAGA AATGATCATGTCAACATACTGAAG ACTATTTCCAGCCAGGGGAAAAGCA AGTTCCATTTGCCAAAAGGCGgtgag gagttttgcttggat	31845152- 31845276	125	75.76	78.66	Complex inversion (see cited work)	Encompassing	Frameshift	Cultured myogenic cells	[39]
40	49	atgatatggacctttcacagGAAGT CCTCAAAGAAAATAAAGAGTTGGT TAGTCAGAGATTTTAAAAAAGAAC CAGAAAATATACCCTCTTTGAAAATA AAA GAAATCTTAAAGAAATTTGACGA AGTGAAGGAAAAGAAAAGAGGAAA GAAACTGAGCAGGGATGAACCCAGC TTCCAATCAGAACTCAGCTTGGGAG gtgagcgtaggagctggaca	31833539- 31833718	180	93.4	87.02	Complex inversion (see cited work)	Encompassing	Premature stop codon	Cultured myogenic cells	[39]

41	49	31832500- 31832659	160	87.27	75.66	Complex inversion (see cited work)	Encompassing	Frameshift	Cultured myogenic cells	[39]			
		<i>atgatgctgtgcatttggcagtttcatt</i> <i>atctgagaaaatgatgcttggaaga</i> <i>ttatataattgcccctcttgaccattttg</i> <i>ttctacagcaatttattttgtgtttaa</i> <i>ggttgggatggaaacgctttggccaa</i> <i>ttctacagcaatttattttgtgtttaa</i> <i>atcacagaaagtctcctcacacctttt</i> <i>ttctgttaccattgaaaaaatattcaa</i> <i>gagcagttaagtaaaaaacataatactacta</i> <i>tattgtaacatttctcttagagtc</i> <i>ctctcccaactttaattctgggtttat</i> <i>tggcaatgaaaatttgctcaataaa</i> <i>tgagtcagtcaaaaatagactttgttc</i> <i>atttcaggggacaaatttttgcctgt</i> <i>tattctctatttgattttcgttgct</i> <i>acaagtgaagttaaccacaggttaaat</i> <i>gagattgtaagcaat</i>											
42	49	31831990- 31832138	149	88.15	88.06	Complex inversion (see cited work)	Encompassing	Frameshift	Cultured myogenic cells	[39]			
		<i>cttctctgaaatctctcttagaatca</i> <i>tgacgccagctgaaactctgtaagaa</i> <i>gctttttccaaatgccctcttactga</i> <i>gatgccacacctctgaaagtgtttta</i> <i>cgcctcttgctgcaataaaggttaata</i> <i>ttcttaatttttt</i>											
43	49	31831041- 31831138	98	73.62	87.58	Complex inversion (see cited work)	Encompassing	Frameshift	Cultured myogenic cells	[39]			
		<i>aaatggtcttttggttacagggta</i> <i>aagagaccacaataacacctttcc</i> <i>cacttccggaggcctttggttaaac</i> <i>catgtctgcccacaggacacagagag</i> <i>cctgtgatgactggtgttttttttgg</i> <i>tgcttttatgatttttaagaaattc</i> <i>ggaggagggccaagatggccgaat</i> <i>aggaacagctccgggtctacagctcc</i> <i>cagcgtgagcaccgacagacagcggg</i> <i>tgatttctgcaatttccattctgaggt</i> <i>accgggttcattctcaata...</i>											
44	51	31752441- 31752524	84	96.74	76.2	None	N/A	ORF preserved	Skeletal muscle	[14]			
45	51	31765009- 31765013	103	79.68	68.7	c.7542+8951- 7542+8952 ins6091	Proximal	Frameshift	Skeletal muscle	[40]			

(Continued)



Table 1  
(Continued)

PE #	Intron	Sequence	Genomic Co-ordinates (chrX)	Size (bp)	Acceptor Score	Donor Score	Mutation(s)	Orientation of Mutation to PE and Splice Sites	ORF	RNA Source	Reference
46	53	tttatacctctgagggaattatagACTAC TAAAGCAGACAGATATTTTGAGGAAAT TTCAGAGGAAAGAGAAAGAAAGAGAA GAAATGAGCTGGGCTGGAGgttaagag aatggggggagaa	31678558- 31678630	73	64.29	91.82	Complex inversion (see cited work)	Encompassing	Frameshift	Skeletal muscle	[41]
47	55	ttttacttttctatttaca (a>g) A AATGGAACACACCAGAAAAACAAG AATTTGAAAGACGAGATGAGAAAAG taagtgttaattggaaaca	31609571- 31609620	50	79.61 -> 96.18	87.36	c.8217+ 18052A>G	Proximal	Frameshift	Skeletal muscle	[33]
48	55	ttttgtttttcccttttttagAGTTTC TTGCTAATGATGGGCCCAAGTTAT ATTAAGAACCTGCAAAAGTAAATTTCA ACCAATTACTTTATTCAGg (g>t) g agtcattaaattgaggt	31595572- 31595644	73	90.26	78.52 -> 95.74	c.8217 +32103G>T	Proximal	Frameshift	Skeletal muscle	[35]
49	56	catccaattttttctaccagAAATA TTAAGAAATTTGACTACAACAGTA TGAAAAAGCAATAGATTCAGTGTG TATTCATGCCAAAAGTCTCAGCAT TCTGGATGTGGAATAAACATATGG CTAAACACTGCCTTTTCTCAAAAT GCCATCAACTATTCCTCTGTTTTGT GGCTCTCAAAAagtaagtagccagat ttttat	31497079- 31497244	166	94.32	87.36	c.8391- 1419_8391- 828del; c.8391- 102_8391- 76del	Distal (5' and 3')	Frameshift	Skeletal muscle	[37]
50	60	ttgattgattatttggcagGTTGA GTCCCTCCAAGAACAGATGCCAGGG CAGATTTATTGCAGAAATACACCTG TGAGAAATAGGGGTGGAGCAGAA TTGAACAAGg (g>t) aagccatcag atcacaaat	31364155- 31364243	89	88.65	76.64 -> 93.87	c.9085- 15519G>T	Proximal	Frameshift	Skeletal muscle	[9]
51	62	ttttgttttttttttggcagGAGCT GATAGCCAGCAACCACATTCAGA AATGGAAGACAGCTGTGAATGCTTC ATTCAGGCCCA (A>G) gtaaatata ggaagaggtgt	31261663- 31261729	67	96.61	75.91 -> 88.06	c.9225- 647A>G	Proximal	Frameshift	Skeletal muscle	[9, 42]

52	62	tgtcgggtgcctcttctgttagtGTTTC CACATTGGATGGGTGAAGAAGTCCT GATAGTCGATTAATTGATCACATAAAC AAGgtca (a>g)tttatacataactg aa	31261306- 31261363	58	81.83	77.76 -> 89.70	c.9225- 285A>G	Proximal	Frameshift	Skeletal muscle [28]; Lymphocytes or skeletal muscle (unspecified) [31]	[28, 31]
53	63	tgtttccactacatctgcagTTTACA TCCTCCACTGAAGTCTGTGAACCCCT GACAGTCATCCATGAGGGTGTGGAAT CAACGCTTCCAAACTCCTGgtaat gttgataatcttgacg	31247694- 31247768	75	76.44	74.39	Unknown	Unknown	Premature stop codon	Lymphocytes [10]	[10]
54	65	tggcaactgtttctcttcttgcagATGAC ATGTGAATGCATTTCTGAATGTATAA CTTCTCTACTGACTGAAAAGTA TTTGGGACAAFTTTAACTCCTTTGA AGACCTGAGTTGCTGTTATAAAGTGG ATTTGTAAATTTTGTACTACCTTTT TCTTAAGAGGGAGAAG (a>g) taa gaaaaatctccagtgga	31208284- 31208430	147	96.18	77.17 -> 94.40	c.9563+ 1215A>G	Proximal	Premature stop codon	Skeletal muscle	[33, 43]
55	65	atbaaattttcttcaataaagGGGCA ATCTGATGAAGATCTGAGCAFTTAA GAGGGCTGAGCAGTTAGTTGCTGgt aa (t>g)tttttggcttccat ttgctttttctctctgtcagTCAGT GCCCCAGAAACACCCCTTCTTCCC AAATGCCCTGAGCACCTCCACTGgtag acattcaatcttcttct	31207099- 31207151	53	83.88	79.28 -> 91.45	c.9564- 427T>G	Proximal	Frameshift	Skeletal muscle	[19, 45]
56	67	gtgttttggttttttttcagAAGGG GTCATACTTCGTCACCCAGGCTGGA GTGCACTGGCAGATCACAGTCAT TGCAGCCTCGACCTCTGGGCTCAAG TGAATCCTCCACCTCAGCCTCCTGA GTAGCTGGACTACAGg (c>t)atg gaccgcccagcctgg	31203344- 31203394	51	81.16	63.87	Unknown	Unknown	ORF preserved	Lymphocytes [10]	[10]
57	67	gtgttttggttttttttcagAAGGG GTCATACTTCGTCACCCAGGCTGGA GTGCACTGGCAGATCACAGTCAT TGCAGCCTCGACCTCTGGGCTCAAG TGAATCCTCCACCTCAGCCTCCTGA GTAGCTGGACTACAGg (c>t)atg gaccgcccagcctgg	31201249- 31201369	121	96.18	67.75 -> 84.98	c.9807+ 2714C>T	Proximal	Frameshift	Lymphocytes or skeletal muscle (unspecified)	[31]
58	77	gtacttatggtcaattgcagATACA GAACTCCAAAAGAAATTCAAATCACCA GTACAAATCAATTAATGATGATGAT CCTGGCCCTTGAACCTGGCAGCTTG CTTACCTGCTTGGAAAGTTGCTGG CTGCCCTGTTGGCACCCCTGGGCATTT TCTTCCACATCTAAACAAGAGgta gtagagaagaagctac	31132483- 31132633	151	76.72	87.1	Unknown	Unknown	Frameshift	Lymphocytes [10]	[10]

exons of the gene – though, as with PE04, it is yet to be determined whether these inclusions indicate a functional role for PE07 [17].

### PSEUDOEXONS OF UNKNOWN ORIGIN

Underlying genomic mutations were not identified for three of the pseudoexons catalogued (PEs 53, 56 and 58). However, as these pseudoexons were exclusively detected in the RNA of specific DMD patient cells, they are believed to be pathogenic and therefore were not classed as arising *sans* mutation.

### PSEUDOEXONS ARISING FROM DELETIONS

Of the 54 known mutation-obligate *DMD* PEs, ten arose from genomic deletions: PEs 02, 03, 12, 17, 18, 27, 28, 29, 35, and 49. For some of these cases, PE initiation can easily be explained as a direct result of the deletion event bringing into conjunction tracts of sequence that, when transcribed, present a strong exon signal to the spliceosome. Pseudoexons appearing to fit this description are PEs 17, 28 and 35, though it should be noted that PE28 also has two small insertions (17bp and 8bp) near its junction site. Two additional but less obvious examples can be seen with PEs 29 and 49 – in these cases, the sequence of the pseudoexons and their splice sites are unaltered from normal individuals, but their inclusions in mature transcripts are initiated by deletions of immediately flanking intronic regions, which presumably contain essential splicing silencers.

For the remaining five deletion-initiated PEs, the link between mutation and pseudoexon is less clear. PEs 12 (i3), 18 (i11) and 27 (i29) all arose from frame-shifting deletions of one, two and one bases in exons 5, 12 and 27 respectively, and PE02 (i1) and PE03 (i1) (both from the same patient) were purportedly initiated by a deletion of exons 3 to 6. Though a more detailed explanation of these PEs may not be possible at present, they appear to support the general theory that splicing of a given *DMD* intron is often interdependent on the correct processing of distant elements of the same transcript [46].

### PSEUDOEXONS ARISING FROM DUPLICATIONS

Five *DMD* pseudoexons arose from genomic duplications: PEs 08, 09, 10, 11, and 33. PE08 (i2), which has also been observed as a low-frequency inclusion

in normal skeletal muscle RNA, was converted to a pseudoexon by a tandem duplication of exon 2. PE09 (i2), PE10 (i2) and PE11(i3) were reported in the same proband as a result of an exon 8–11 duplication, and PE33 (i43) arose from an exon 44 duplication. These cases offer further support to the theory of correct *DMD* splicing occurring through coordination of distant elements. At this point, however, it is not clear whether these PEs are induced specifically by alterations to the canonical exon order, or whether they arise from disruptions to intronic sequences that would normally act as distal pseudoexon silencers.

### PSEUDOEXONS ARISING FROM INVERSIONS

Eight *DMD* pseudoexons arose from inversion mutations: PEs 34, 38, 39, 40, 41, 42, 43, and 46. In all these cases, each PE was completely internal to the inverted region. PE34 arose from an inversion internal to intron 44 – i.e. no canonical exons were directly affected. PEs 38 to 43 (i48 and i49) were reported from a single patient with a complex inversion of exons 49 and 50, while PE46 (i53) arose in a patient with a deletion of exons 48–52 and an inversion of exon 53. It is perhaps unsurprising that such dramatic rearrangements of large tracts of transcribed sequence would result in splicing disturbances of some kind, but these cases nevertheless serve to illustrate that, in addition to recognition of canonical exons, the silencing of pseudoexons is an equally essential component of spliceosome function, and one that is likely to be achieved through orientation-dependent sequence motifs in the intron.

### PSEUDOEXONS ARISING FROM INSERTIONS

Two *DMD* pseudoexons arose from insertion mutations, PEs 32 and 45. PE32 was created by an insertion into intron 43 of two large tracts (88.0kb and 2.6kb) of intragenic sequence from chromosome 4, the PE itself originating within the larger of these two tracts, while PE45 was created by a 6096bp LINE-1 retrotransposon with a potential donor site at its 5' end inserting immediately 3' of a latent acceptor site in *DMD* intron 51.

### PSEUDOEXONS ARISING FROM SINGLE BASE-PAIR SUBSTITUTIONS

Single base-pair substitutions were the most commonly observed cause of *DMD* PEs, accounting for

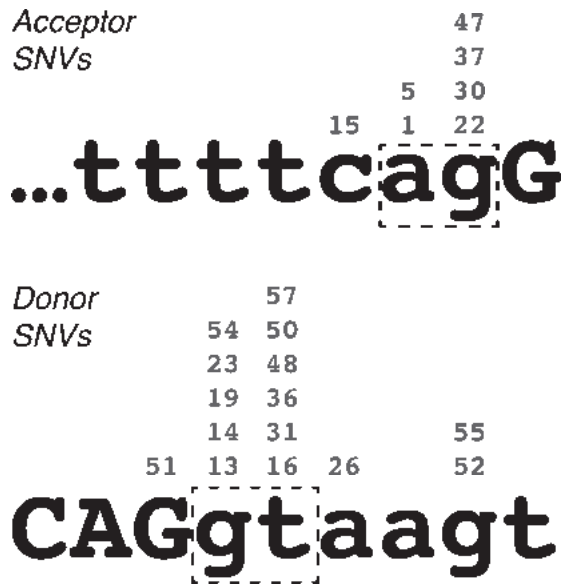


Fig. 1. Locations of pseudoexon-initiating single-nucleotide variations in the *DMD* gene, relative to acceptor and donor splice site consensus sequences. Numbers above each nucleotide indicate the exemplar pseudoexons. Lower-case letters indicate intron sequence, upper-case letters indicate exon sequence. Dash-line boxes highlight the essential “ag” and “gt” of the acceptor and donor site motifs respectively.

26 of the 58 catalogued, through 24 unique mutations. In most of these cases, the etiology of the PE appears to stem from the creation or enhancement of a latent mid-intron splice site – of the 24 unique mutations, 7 created new acceptor splice sites (PEs 01, 05/06, 15, 22, 30, 37 and 47) and 15 created new donor splice sites (PEs 13, 14, 16, 19, 23/24, 26, 31, 36, 48, 50, 51, 52, 54, 55 and 57). All 22 of these acceptor-motif and donor-motif mutations greatly enhanced the Shapiro-Senapathy splice score of the mutated site, and in every case the new nucleotide was the most common consensus base for that position in the splice site (Fig. 1). While a possible exception to this rule was noted at the acceptor site of PE14 (c.650-39575A>C), this mutation was found to be a common SNP (rs113593006, dbSNP build 151 – see ref. 44) that only marginally decreased the Shapiro-Senapathy score of the acceptor site (from 79.91 to 78.04). I therefore judged that this SNP was likely to be incidental to the pathology of this pseudoexon and did not constitute a true counterexample to the prevailing pattern of splice site enhancement.

Only two PEs arose from SNVs outside of the PE consensus splice sites, PE20 and PE25. PE20 (i18) arose from a G-to-A substitution at the first base of intron 20, suggesting that the correct splicing of these two introns may be interdependent. In this way, PE20

is qualitatively similar to PEs 12 and 18, which also arose from small mutations distal to the pseudoexon, although the mutation that initiated PE20 did not directly alter the *DMD* coding sequence. PE25 (i26) is a unique case that arose from a G-to-C substitution internal to the PE that altered the predicted binding of splicing enhancer SRp55 [30].

## PSEUDOEXONS AND RECURSIVE SPLICING

Multi-step or recursive splicing was first described in *Drosophila* in 2005 [47] and has more recently been discovered to be prevalent in the genes of the human transcriptome [48, 49], including *DMD* [50, 51]. While conventionally spliced introns are removed with a single splicing event, recursively spliced introns are excised from their maturing transcripts in two or more segments, via intronic acceptor splice sites called ‘ratchet points’. Sibley et al. [48] have also reported that recursive splicing in vertebrates is facilitated by recognition of evolutionarily conserved donor-like splice sites downstream of acceptor-like ratchet points.

Georgomanolis et al. [49] have postulated that some of the low-frequency pseudoexons observed in the transcripts of normal cells may be a natural byproduct of the spliceosome incorrectly recognizing exon-like intronic ratchet points. I suggest that this hypothesis can reasonably be extended to include mutation-induced PEs – i.e. mutations that enhance the exon-like characteristics of intronic ratchet points may thereby convert them into pathogenic PEs. Evidence supporting this hypothesis has already been described by Bouge et al. [14], who noted the alignment of six pseudoexon splice sites with six of the *DMD* intron ratchet points predicted by Gazzoli et al. [51]. Seeking to expand upon these observations, I cross-referenced the splice sites of all eligible PEs in Table 1 with all of the intronic ratchet points predicted by Gazzoli et al. This analysis excluded the splice sites of the *DMD* inversion PEs (34, 38, 39, 40, 41, 42, 43 and 46), the chromosome 4 insertion PE (32), and the *de novo* donor site for PE45, as these sites could not be sensibly compared to any part of the *DMD* reference sequence. Splice sites shared by multiple PEs (acceptor sites for PEs 5 and 6, 15 and 16, and donor sites for PEs 9 and 10, 11 and 12, and 15 and 16) were included but were counted only once each to avoid bias. Using these criteria, including the matches noted by Bouge et al. I confirmed 12 Gazzoli matches out of 47 unique acceptor sites and 14 Gazzoli matches out of 44 unique donor sites (Table 2).

Table 2  
DMD pseudoexon splice sites coinciding with recursive splicing ratchet points predicted by Gazzoli et al. (2016). Co-ordinates listed are for genomic reference sequence NC\_000023.10, as used by the cited authors. Dotted-line boxes enclose pairs of split reads that match to the same pseudoexon splice site. Asterisks (\*) indicate the six coinciding splice sites previously noted by Bouge et al. [14]

Intron	Genome positions of split read	Recursive Splicing Type	PE Match	Donor/Acceptor
1	chrX:33190805-33192302	nested	1	D
1	chrX:33096464-33229399	5'RS	4*	A
1	chrX:33038317-33096303	3'RS	4*	D
2	chrX:32978464-33038256	5'RS	7	A
2	chrX:32978004-32978325	nested	7	D
	chrX:32867937-32978325	3'RS		
2	chrX:32897097-33038256	5'RS	8	A
2	chrX:32867937-32897000	3'RS	8	D
3	chrX:32862977-32864739	3'RS	11, 12	D
18	chrX:32532339-32536125	5'RS	20	A
	chrX:32532339-32533049	nested		
18	chrX:32519959-32532208	3'RS	20	D
21	chrX:32499062-32503036	5'RS	21*	A
21	chrX:32490426-32498997	3'RS	21*	D
27	chrX:32460395-32466573	5'RS	26	A
27	chrX:32459431-32460277	3'RS	26	D
29	chrX:32430180-32456358	5'RS	27	A
	chrX:32430180-32430279	nested		
29	chrX:32430030-32430136	3'RS	27	D
43	chrX:32253321-32305646	5'RS	33	A
43	chrX:32235180-32253264	3'RS	33	D
51	chrX:31770641-31792077	5'RS	44*	A
51	chrX:31747865-31770558	3'RS	44*	D
56	chrX:31515061-31515196	3'RS	49	D
63	chrX:31265885-31279072	5'RS	53	A
63	chrX:31241238-31265811	3'RS	53	D
67	chrX:31221511-31222078	5'RS	56	A
77	chrX:31150750-31152219	5'RS	58	A
77	chrX:31144790-31150600	3'RS	58	D

Several interesting features were apparent in this set of Gazzoli-matched splice sites. Firstly, most of the matched PEs matched at both their acceptor and donor splice sites. Only PEs 1, 11/12 and 49 matched at their donor sites alone, and only PE56 matched at its acceptor site alone. Secondly, a clear bias was evident in the mutation categories of the matched PEs, as the majority of the Gazzoli-matched sites were from PEs induced either without mutations or by mutations distal to the PE and its splice motifs, PEs 1 and 26 being the only exceptions to this rule. Lastly, of the 15 PEs where inducing distal mutations were identified,

11 arose exclusively from mutations that were 3' to the PE (PEs 2, 3, 4, 7, 8, 9, 10, 11, 12, 18 and 20). Only PEs 27 and 29 were induced exclusively by 5' mutations, while PEs 33 and 49 were each induced by flanking mutations.

## PSEUDOEXONS AND RECURSIVE SPLICING REGULATION

Canonical splicing of a donor-acceptor pair is often dependent on distal regulatory elements, including

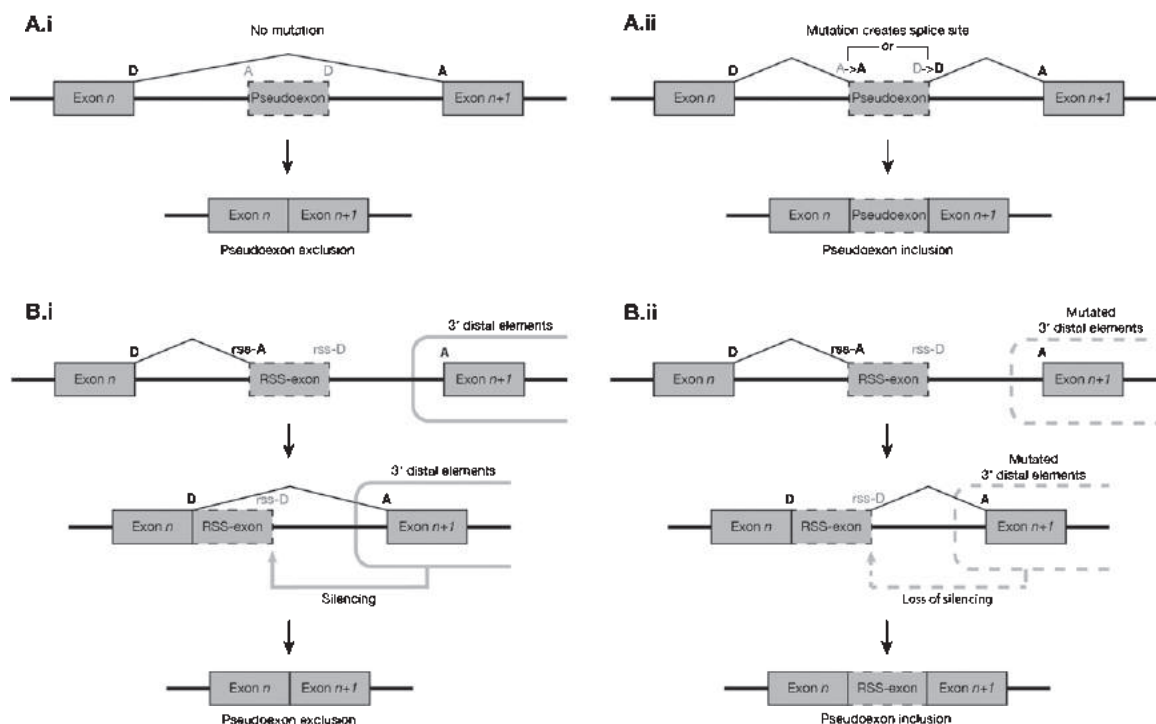


Fig. 2. Suggested model of the two most common modes of pseudoexon initiation observed in the *DMD* gene. (A) Proximal mutations at non-RS sites. (i) In the absence of mutation, a putative pseudoexon presents a weak exon-like profile to the spliceosome and is predominantly excluded from mature transcripts. (ii) The presence of a mutation, usually a splice-site-creating SNV, increases the exon-like profile of the putative pseudoexon, resulting in its inclusion in a much higher proportion of transcripts. (B) Mutations 3' distal to RS sites. (i) In the absence of mutation, the exon *n* donor site and RS-acceptor site are used to excise the 5' segment of an intron. Silencing elements distal (usually 3') to the RS-exon prevent spliceosome recognition of its donor-like motif, and the RS-exon is subsequently removed from the transcript along with the rest of the 3' intron segment. (ii) When mutations to the distal silencing elements impair their function, the intron segment 5' to the RS-exon is spliced as normal, but the RS-exon donor-like site escapes silencing and is more readily recognised by the spliceosome, leading to a much higher frequency of inclusion of the RS-exon in the mature transcript population.

but not limited to other canonical splice sites [51, 52, 53]. Mutations that alter or destroy these distal elements can impede exon definition and decrease the frequency of inclusion of the affected exons in the mature transcript [42, 43]. It may be that spliceosome recognition of recursive splice sites (which necessarily exhibit a strong exon-like profile in their local motifs) is regulated by a similar system of mostly 3' distal elements, but a system that acts to silence rather than promote the inclusion of its targets. Mutations that impaired these distal silencing elements might thereby permit an increase in the erroneous processing of recursive splice sites, converting them to PEs via a distinctly different pathway than PEs created by proximal mutation (Fig. 2). If valid, this model would explain the high coincidence of Gazzoli predicted recursive splice sites with the splice sites of PEs induced by distal mutations. However, while the mutations collated in this report may offer broad clues as to the locations of some such suppressive distal ele-

ments in the *DMD* gene, any further analysis of their common features awaits the assembly and analysis of a much larger dataset of PEs and recursive splice sites – one that will have to encompass multiple other genes besides *DMD*.

## CONCLUSIONS

The 58 *DMD* pseudoexons collated from published reports exhibit great diversity in their sizes, locations, and pathologies. Surprisingly, PEs arising either from no mutation, or from mutations distal to the pseudoexon and its splice sites, exhibited a high coincidence with predicted recursive splice sites in the *DMD* introns, suggesting that some such pseudoexons may arise from disruptions to recursive splicing regulation. This finding may represent an important new insight into the etiology of pseudoexons in *DMD* specifically and human disease genes generally.

## ACKNOWLEDGMENTS

I would like to thank the Enid and Arthur Home Scholarship and Murdoch University for the provision of funding and resources; my reviewers, for their thoroughness and insightful suggestions; and my colleagues at the Centre for Molecular Medicine and Innovative Therapeutics, especially professors Sue Fletcher and Steve Wilton, for their advice and constructive criticism.

## CONFLICT OF INTEREST

The author has no conflict of interest to declare.

## SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: <https://dx.doi.org/10.3233/JND-190431>.

## REFERENCES

- [1] Nudel U. Alternative Promoters: Duchenne Muscular Dystrophy (DMD) Gene. In eLS. (Ed.). 2005.
- [2] Feener CA, Koenig M, Kunkel LM. Alternative splicing of human dystrophin mRNA generates isoforms at the carboxy terminus. *Nature*. 1989;338(6215):509-11.
- [3] Doorenweerd N, Mahfouz A, van Putten M, Kaliyaperumal R, t'Hoen PAC, Hendriksen JGM, et al. Timing and localization of human dystrophin isoform expression provide insights into the cognitive phenotype of Duchenn muscular dystrophy. *Sci Rep*. 2017;7(12575).
- [4] Wahl MC, Will CL, Lührmann R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell*. 2009;136:701-18.
- [5] Berget SM. Exon recognition in vertebrate splicing. *J Biol Chem*. 1995;270(6):2411-4.
- [6] De Conti LM, Baralle M, Buratti E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA*. 2013;4(1):49-60.
- [7] Anna A, Gos M. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet*. 2018;59(3):253-68.
- [8] Hong X, Scofield DG, Lynch M. Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol*. 2006;23(12):2392-404.
- [9] Beroud C, Carrie A, Beldjord C, Deburgrave N, Llense S, Carelle N, et al. Dystrophinopathy caused by mid-intronic substitutions activating cryptic exons in the DMD gene. *Neuromuscular disorders: NMD*. 2004;14(1):10-8.
- [10] Zhang Z, Habara Y, Nishiyama A, Oyazato Y, Yagi M, Takeshima Y, et al. Identification of seven novel cryptic exons embedded in the dystrophin gene and characterization of 14 cryptic dystrophin exons. *Journal of human genetics*. 2007;52(7):607-17.
- [11] Roberts RG, Bentley DR, Bobrow M. Infidelity in the structure of ectopic transcripts: a novel exon in lymphocyte dystrophin transcripts. *Human mutation*. 1993;2(4):293-9.
- [12] Suminaga R, Takeshima Y, Adachi K, Yagi M, Nakamura H, Matsuo M. A novel cryptic exon in intron 3 of the dystrophin gene was incorporated into dystrophin mRNA with a single nucleotide deletion in exon 5. *Journal of human genetics*. 2002;47(4):196-201.
- [13] Tran VK, Zhang Z, Yagi M, Nishiyama A, Habara Y, Takeshima Y, et al. A novel cryptic exon identified in the 3' region of intron 2 of the human dystrophin gene. *Journal of human genetics*. 2005;50(8):425-33.
- [14] Bouge AL, Muraier E, Beyne E, Miro J, Varilh J, Taulan M, et al. Targeted RNA-Seq profiling of splicing pattern in the DMD gene: exons are mostly constitutively spliced in human skeletal muscle. *Sci Rep*. 2017;7:39094.
- [15] Yagi M, Takeshima Y, Wada H, Nakamura H, Matsuo M. Two alternative exons can result from activation of the cryptic splice acceptor site deep within intron 2 of the dystrophin gene in a patient with as yet asymptomatic dystrophinopathy. *Human genetics*. 2003;112(2):164-70.
- [16] Dwi Pramono ZA, Takeshima Y, Surono A, Ishida T, Matsuo M. A novel cryptic exon in intron 2 of the human dystrophin gene evolved from an intron by acquiring consensus sequences for splicing at different stages of anthropoid evolution. *Biochemical and biophysical research communications*. 2000;267(1):321-8.
- [17] Gualandi F, Rimessi P, TrabANELLI C, Spitali P, Neri M, Pataranello T, et al. Intronic breakpoint definition and transcription analysis in DMD/BMD patients with deletion/duplication at the 5' mutation hot spot of the DMD gene. *Gene*. 2006;370:26-33.
- [18] Ishibashi K, Takeshima Y, Yagi M, Nishiyama A, Matsuo M. Novel cryptic exons identified in introns 2 and 3 of the human dystrophin gene with duplication of exons 8-11. *The Kobe journal of medical sciences*. 2006;52(3-4):61-75.
- [19] Sedláčková J, Vondráček P, Hermanová M, Zámečník J, Hrubá Z, Haberlová J, et al. Point mutations in Czech DMD/BMD patients and their phenotypic outcome. *Neuromuscular Disorders*. 2009;19(11):749-53.
- [20] Zaum AK, Stuve B, Gehrig A, Kolbel H, Schara U, Kress W, et al. Deep intronic variants introduce DMD pseudoexon in patient with muscular dystrophy. *Neuromuscular disorders: NMD*. 2017;27(7):631-4.
- [21] Tuffery-Giraud S, Saquet C, Thorel D, Dissot A, Rivier F, Malcolm S, et al. Mutation spectrum leading to an attenuated phenotype in dystrophinopathies. *European journal of human genetics: EJHG*. 2005;13(12):1254-60.
- [22] Ferlini A, Muntoni F. The 5' region of intron 11 of the dystrophin gene contains target sequences for mobile elements and three overlapping ORFs. *Biochemical and biophysical research communications*. 1998;242(2):401-6.
- [23] Ferlini A, Galie N, Merlini L, Sewry C, Branzi A, Muntoni F. A novel Alu-like element rearranged in the dystrophin gene causes a splicing mutation in a family with X-linked dilated cardiomyopathy. *American journal of human genetics*. 1998;63(2):436-46.
- [24] Nasim MT, Chernova TK, Chowdhury HM, Yue BG, Eperon IC. HnRNP G and Tra2beta: opposite effects on splicing matched by antagonism in RNA binding. *Human molecular genetics*. 2003;12(11):1337-48.
- [25] Rimessi P, Fabris M, Bovolenta M, Bassi E, Falzarano S, Gualandi F, et al. Antisense modulation of both exonic and intronic splicing motifs induces skipping of a DMD pseudo-exon responsible for x-linked dilated cardiomyopathy. *Human gene therapy*. 2010;21(9):1137-46.
- [26] Malueka RG, Takaoka Y, Yagi M, Awano H, Lee T, Dwianingsih EK, et al. Categorization of 77 dystrophin exons into

- 5 groups by a decision tree using indexes of splicing regulatory factors as decision markers. *BMC genetics*. 2012;13:23.
- [27] Gurvich OL, Tuohy TM, Howard MT, Finkel RS, Medne L, Anderson CB, et al. DMD pseudoexon mutations: splicing efficiency, phenotype, and potential therapy. *Annals of neurology*. 2008;63(1):81-9.
- [28] Tuffery-Giraud S, Saquet C, Chambert S, Claustres M. Pseudoexon activation in the *DMD* gene as a novel mechanism for Becker muscular dystrophy. *Human mutation*. 2003;21(6):608-14.
- [29] Ikezawa M, Minami N, Takahashi M, Goto Y, Miike T, Nonaka I. Dystrophin gene analysis on 130 patients with Duchenne muscular dystrophy with a special reference to muscle mRNA analysis. *Brain & development*. 1998;20(3):165-8.
- [30] Trabelsi M, Beugnet C, Deburgrave N, Commere V, Orhant L, Leturcq F, et al. When a mid-intronic variation of DMD gene creates an ESE site. *Neuromuscular Disord*. 2014;24(12):1111-7.
- [31] Takeshima Y, Yagi M, Okizuka Y, Awano H, Zhang Z, Yamauchi Y, et al. Mutation spectrum of the dystrophin gene in 442 Duchenne/Becker muscular dystrophy cases from one Japanese referral center. *Journal of human genetics*. 2010;55(6):379-88.
- [32] Ishmukhametova A, Van Kien PK, Mechin D, Thorel D, Vincent MC, Rivier F, et al. Comprehensive oligonucleotide array-comparative genomic hybridization analysis: new insights into the molecular pathology of the DMD gene. *European Journal of Human Genetics*. 2012;20(10):1096-100.
- [33] Bovolenta M, Neri M, Fini S, Fabris M, Trabanelli C, Venturoli A, et al. A novel custom high density-comparative genomic hybridization array detects common rearrangements as well as deep intronic mutations in dystrophinopathies. *BMC genomics*. 2008;9.
- [34] Gonorazky H, Liang M, Cummings B, Lek M, Micallef J, Hawkins C, et al. RNAseq analysis for the diagnosis of muscular dystrophy. *Ann Clin Transl Neurol*. 2016;3(1):55-60.
- [35] Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med*. 2017;9(386):eaa15209.
- [36] Baskin B, Gibson WT, Ray PN. Duchenne muscular dystrophy caused by a complex rearrangement between intron 43 of the DMD gene and chromosome 4. *Neuromuscular Disorders*. 2011;21:178-82.
- [37] Khelifi MM, Ishmukhametova A, Khau Van Kien P, Thorel D, Mechin D, Perelman S, et al. Pure intronic rearrangements leading to aberrant pseudoexon inclusion in dystrophinopathy: a new class of mutations? *Human mutation*. 2011;32(4):467-75.
- [38] Greer K, Mizzi K, Rice E, Kuster L, Barrero RA, Bellgard MI, et al. Pseudoexon activation increases phenotype severity in a Becker muscular dystrophy patient. *Molecular genetics & genomic medicine*. 2015;3(4):320-6.
- [39] Madden HR, Fletcher S, Davis MR, Wilton SD. Characterization of a complex Duchenne muscular dystrophy-causing dystrophin gene inversion and restoration of the reading frame by induced exon skipping. *Human mutation*. 2009;30(1):22-8.
- [40] Goncalves A, Oliveira J, Coelho T, Taipa R, Melo-Pires M, Sousa M, et al. Exonization of an Intronic LINE-1 Element Causing Becker Muscular Dystrophy as a Novel Mutational Mechanism in Dystrophin Gene. *Genes*. 2017;8(10).
- [41] Cagliani R, Sironi M, Ciafaloni E, Bardoni A, Fortunato F, Prella A, et al. An intragenic deletion/inversion event in the DMD gene determines a novel exon creation and results in a BMD phenotype. *Human genetics*. 2004;115(1):13-8.
- [42] Juan-Mateu J, Gonzalez-Quereda L, Rodriguez MJ, Verdura E, Lazaro K, Jou C, et al. Interplay between DMD point mutations and splicing signals in Dystrophinopathy phenotypes. *PloS one*. 2013;8(3):e59916.
- [43] Deburgrave N, Daoud F, Llense S, Barbot JC, Recan D, Pécate C, et al. Protein- and mRNA-based phenotype-genotype correlations in DMD/BMD with point mutations and molecular basis for BMD with nonsense and frameshift mutations in the DMD gene. *Human mutation*. 2007;28(2):183-95.
- [44] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29:308-11.
- [45] Magri F, Del Bo R, D'Angelo MG, Govoni A, Ghezzi S, Gandossini S, et al. Clinical and molecular characterization of a cohort of patients with novel nucleotide alterations of the Dystrophin gene detected by direct sequencing. *BMC medical genetics*. 2011;12:37.
- [46] Suzuki H, Aoki Y, Kameyama T, Saito T, Masuda S, Tanihata J, et al. Endogenous Multiple Exon Skipping and Back-Splicing at the DMD Mutation Hotspot. *International Journal of Molecular Sciences*. 2016;17(10).
- [47] Burnette JM, Miyamoto-Sato E, Schaub MA, Conklin J, Lopez AJ. Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. *Genetics*. 2005;170(2):661-74.
- [48] Sibley CR, Emmett W, Blazquez L, Faro A, Haberman N, Briese M, et al. Recursive splicing in long vertebrate genes. *Nature*. 2015;521(7552):371-5.
- [49] Georgomanolis T, Sofiadis K, Papanonis A. Cutting a Long Intron Short: Recursive Splicing and Its Implications. *Front Physiol*. 2016;7:598.
- [50] Suzuki H, Kameyama T, Ohe K, Tsukahara T, Mayeda A. Nested introns in an intron: evidence of multi-step splicing in a large intron of the human dystrophin pre-mRNA. *FEBS Lett*. 2013;587(6):555-61.
- [51] Gazzoli I, Pulyakhina I, Verwey NE, Ariyurek Y, Laros JF, t Hoen PA, et al. Non-sequential and multi-step splicing of the dystrophin transcript. *RNA Biol*. 2016;13(3):290-305.
- [52] Ke S, Chasin LA. Intronic motif pairs cooperate across exons to promote pre-mRNA splicing. *Genome Biol*. 2010;11:R84.
- [53] Kim SW, Taggart AJ, Heintzelman C, Cygan KJ, Hull GH, Wang J, et al. Widespread intra-dependencies in the removal of introns from human transcripts. *Nucleic Acids Res*. 2017;45(16):9503-13.
- [54] Shapiro MB, Senapathy P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res*. 1987;15(17):7155-74.
- [55] Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res*. 2006;34(14):3955-67.



### 3.4 Mentions and awards

At time of final submission, this report has been cited in the following third-party works:

Alexieva, D., Long, Y., Sarkar, R., Dhayan, H., Bruet, E., Winston, Rm., et al. (2022) Background splicing as a predictor of aberrant splicing in genetic disease. *RNA Biol* 19(1), 256-265. doi: 10.1080/15476286.2021.2024031.

Hou, H., Wang, X., Ding, W., Xiao, C., Cai, X., Lv, W., et al. (2021). Whole-genome sequencing reveals the artificial selection and local environmental adaptability of pigeons (*Columba livia*). *Evolutionary Applications*. doi: 10.1111/eva.13284.

Liu, W., Shi, X., Li, Y., Qiao, F., and Wu, Y. (2021). The identification of a novel splicing mutation in the DMD gene of a Chinese family. *Clin Case Rep* 9(12), e05166. doi: 10.1002/ccr3.5166.

Lu, X., Han, C., Mai, J., Jiang, X., Liao, J., Hou, Y., et al. (2021). Novel Intronic Mutations Introduce Pseudoexons in DMD That Cause Muscular Dystrophy in Patients. *Front Genet* 12, 657040. doi: 10.3389/fgene.2021.657040.

Roos, D., and de Boer, M. (2021). Mutations in cis that affect mRNA synthesis, processing and translation. *Biochim Biophys Acta Mol Basis Dis* 1867(9), 166166. doi: 10.1016/j.bbadis.2021.166166.

Zhu, F., Zhang, F., Hu, L., Liu, H., and Li, Y. (2021). Integrated Genome and Transcriptome Sequencing to Solve a Neuromuscular Puzzle: Miyoshi Muscular Dystrophy and Early Onset Primary Dystonia in Siblings of the Same Family. *Front Genet* 12, 672906. doi: 10.3389/fgene.2021.672906.

This report was also awarded 3<sup>rd</sup> prize in the 2021 Perron Institute “Best Student Paper Competition” (certificate on following page).

*Best Student Paper Competition  
Third Prize*

Presented to

*Niall Keegan*

*Alan Robson*

Professor Alan Robson AO CBEWA

Chairman

17 March, 2021

*Steve Arnott*

Steve Arnott  
Chief Executive Officer

17 March, 2021

## **Chapter 4**

# **Analysis of pathogenic pseudoexons reveals novel mechanisms driving cryptic splicing**

#### 4.1 Preamble

Having completed an analysis of *DMD* pseudoexons (Chapter 3), there remained several unanswered questions about pseudoexons (PEs) in general. How many pathways exist by which mutations can induce PEs? Are the effects of these mutations understandable based on known principles of RNA splicing? And how do PEs interrelate with other forms of deep intronic splicing, such as recursive splice sites and poison exons?

A preliminary investigation revealed only a few prior attempts to analyse the common features of germline pseudoexons, and that these analyses had mostly pursued very different research questions to our own (Královičová and Vořechovský 2007; Vořechovský 2010; Dhir and Buratti 2010; Romano *et al.* 2013; Vaz-Drago *et al.* 2017). We decided that this justified assembling and curating a comprehensive pseudoexon catalogue and thus undertook a thorough search of the literature for published and empirically verified examples of PEs.

A significant time-cost to this process was the lack of precise detail offered by much of the source literature. Transcribing or deducing the sequence and mutation of a PE from a report took only minutes in most cases, but hours in others. Fortunately, however, by the time our search concluded only a handful of PEs had had to be completely excluded due to insufficient data. The second major impediment to the search process was the low-to-non-existent use of relevant keywords (“pseudoexon”, “cryptic exon” etc.) in some reports. In these cases, the problem was not the quality of data in the reports but finding the reports in the first place, as many ranked very low in search engine results despite the relevance of their content. Consequently, it is all but certain that some PEs were overlooked due to the reporting authors using terminology or keywords that completely escaped capture by the search terms we used.

Despite these obstacles, we successfully catalogued 413 examples of pseudoexons according to the criteria detailed within our report. To our knowledge, this is the largest dataset of PEs assembled to date. Our analysis of this dataset led us to

refine the parameters of some mutation categories, suggest new categories, and highlight some of the remaining gaps in our knowledge of spliceosome behaviour.

The most common mutations driving pseudoexon inclusion were those that directly created or enhanced pseudoexon splice motifs, with donor motif mutations being about four times as common as acceptor motif mutations. Collectively, the effects of the PE splice motif mutations were entirely consistent with existing models of spliceosome behaviour.

Mutations inside pseudoexons and mutations that alter pseudoexon branch-point definition have been described by others, but the parameters of these categories have proven more difficult to define than those of splice motif mutations. We found that the effects of intra-PE mutations were well-predicted by *HExoSplice*, an online mutation analysis tool built specifically for this purpose and endorsed its use in future studies (Ke *et al.* 2011; Tubeuf *et al.* 2020). Of the branch-point mutations collated in this study, most could be understood through their effects on the branch point itself, the AG-exclusion zone, or the proximity of these motifs. However, we identified a knowledge gap when it came to predicting the effects of some apparent pseudoexon branch-point mutations and suggest that these may alter binding of spliceosome element U2AF65.

In addition to proximal PE mutations – which altered the PE splice motifs, branch point motifs or internal exon definition motifs – there were also numerous examples of PE-inducing mutations that were substantially distal to the PEs. A common characteristic of most of these mutations was weakened definition in one of the canonical exons flanking the PE; or, in a few cases, weakened definition in a next-but-one neighbour exon. This category represents an expansion of one originally proposed by Dhir and Buratti (2010), and its existence indicates that definition strength of the exons flanking an intron is an important factor in suppressing the splicing of latent exon-like elements within the intron.

We also suggest several new categories for some of the remaining distal mutations, although our descriptions of these were limited by low numbers of supporting examples, while a remaining few were unique and so complex in their effects that we were unable to usefully speculate on the underlying mechanisms.

Having previously detected novel matches between pseudoexon splice sites and putative recursive splice sites in the *DMD* gene (Chapter 3), we now sought to examine whether any such matches could be found for our multi-gene PE catalogue. By cross-searching all PE coordinates against published RNAseq (Sibley *et al.* 2015) and spliced-EST datasets (Kent *et al.* 2002), we found that more than 15% of PEs underwent splicing at one or both splice sites in non-mutant cells. For most of the PEs in this 15%, it remains to be determined whether this splicing activity is functional, but in seven cases the matches were experimentally verified recursive splice sites, providing the first explicit proof of a link between PEs and recursive splicing (Blazquez *et al.* 2018; Zhang *et al.* 2018; Wan *et al.* 2021).

Our report represents substantial progress towards a comprehensive model for why and where pseudoexons arise, and how PEs interrelate with other deep intronic splicing phenomena. Although this model is far from complete, we have been able to specify certain areas where our knowledge is lacking and speculate that accelerating progress in RNA sequencing technologies will close these gaps in the very near future.

#### 4.2 Citation

Keegan, N. P., Wilton, S. D., and Fletcher, S. (2022). Analysis of Pathogenic Pseudoexons Reveals Novel Mechanisms Driving Cryptic Splicing. *Front Genet* 12, 806946. doi: 10.3389/fgene.2022.806946.



# Analysis of Pathogenic Pseudoexons Reveals Novel Mechanisms Driving Cryptic Splicing

Niall P. Keegan<sup>1,2\*</sup>, Steve D. Wilton<sup>1,2</sup> and Sue Fletcher<sup>1,2</sup>

<sup>1</sup>Centre for Molecular Medicine and Innovative Therapeutics, Health Futures Institute, Murdoch University, Perth, WA, Australia, <sup>2</sup>Centre for Neuromuscular and Neurological Disorders, Perron Institute for Neurological and Translational Science, The University of Western Australia, Perth, WA, Australia

## OPEN ACCESS

### Edited by:

Michael R. Ladomery,  
University of the West of England,  
United Kingdom

### Reviewed by:

John Conboy,  
Lawrence Berkeley National  
Laboratory, United States  
Ian Eperon,  
University of Leicester,  
United Kingdom

### \*Correspondence:

Niall P. Keegan  
N.Keegan@murdoch.edu.au

### Specialty section:

This article was submitted to  
RNA,  
a section of the journal  
Frontiers in Genetics

**Received:** 01 November 2021

**Accepted:** 09 December 2021

**Published:** 24 January 2022

### Citation:

Keegan NP, Wilton SD and Fletcher S  
(2022) Analysis of Pathogenic  
Pseudoexons Reveals Novel  
Mechanisms Driving Cryptic Splicing.  
*Front. Genet.* 12:806946.  
doi: 10.3389/fgene.2021.806946

Understanding pre-mRNA splicing is crucial to accurately diagnosing and treating genetic diseases. However, mutations that alter splicing can exert highly diverse effects. Of all the known types of splicing mutations, perhaps the rarest and most difficult to predict are those that activate pseudoexons, sometimes also called cryptic exons. Unlike other splicing mutations that either destroy or redirect existing splice events, pseudoexon mutations appear to create entirely new exons within introns. Since exon definition in vertebrates requires coordinated arrangements of numerous RNA motifs, one might expect that pseudoexons would only arise when rearrangements of intronic DNA create novel exons by chance. Surprisingly, although such mutations do occur, a far more common cause of pseudoexons is deep-intronic single nucleotide variants, raising the question of why these latent exon-like tracts near the mutation sites have not already been purged from the genome by the evolutionary advantage of more efficient splicing. Possible answers may lie in deep intronic splicing processes such as recursive splicing or poison exon splicing. Because these processes utilize intronic motifs that benignly engage with the spliceosome, the regions involved may be more susceptible to exonization than other intronic regions would be. We speculated that a comprehensive study of reported pseudoexons might detect alignments with known deep intronic splice sites and could also permit the characterisation of novel pseudoexon categories. In this report, we present and analyse a catalogue of over 400 published pseudoexon splice events. In addition to confirming prior observations of the most common pseudoexon mutation types, the size of this catalogue also enabled us to suggest new categories for some of the rarer types of pseudoexon mutation. By comparing our catalogue against published datasets of non-canonical splice events, we also found that 15.7% of pseudoexons exhibit some splicing activity at one or both of their splice sites in non-mutant cells. Importantly, this included seven examples of experimentally confirmed recursive splice sites, confirming for the first time a long-suspected link between these two splicing phenomena. These findings have the potential to improve the fidelity of genetic diagnostics and reveal new targets for splice-modulating therapies.

**Keywords:** pseudoexons, cryptic splicing, splicing mutations, recursive splicing, poison exons, genetic disease



## 1 INTRODUCTION

According to the current release of Ensembl (104.38), the 3.1 gigabases of the human genome are estimated to contain over 44,000 genes, just over 20,000 of which encode proteins (Howe et al., 2021). Based on statistics provided by Piovesan et al. (2019), protein-coding genes account for 41% of the total human genome, although exons, the transcribed segments that are retained in the mature mRNA, comprise only 4.65% of this fraction, or 1.91% of the total genome.

During and after transcription pre-mRNAs undergo splicing, a process whereby introns are excised from the pre-mRNA molecule and the ends of the flanking exons are ligated together by a multi-molecular assembly called the spliceosome. This splicing process needs to be both consistent and accurate, since an error of even a single nucleotide could render an entire transcript functionless or cause it to encode a toxic product. However, this does not mean that splicing must be perfect. Occasional splicing errors are inevitable even in healthy cells (Alexieva et al., 2021), and these aberrant transcripts are generally well-managed by error-detecting systems such as nonsense-mediated decay (NMD) (Hug et al., 2016)—but never without some energy cost to the cell.

There is, therefore, an ancient and relentless evolutionary pressure on all eukaryotes to splice pre-mRNAs as efficiently as possible. This raises the question of why evolution allows introns to persist in the first place. Much effort has been devoted to investigating this question, and there appears to be no single answer (Chorev and Carmel, 2012; Jo and Choi, 2015). Stated briefly, introns contain numerous regulatory elements that enable alternative splicing and fine control of gene expression, and the presence of these elements may have indirectly modulated the efficiency of natural selection for eukaryotic life on Earth. Recent research also reveals a regulatory role for conserved exon-like sequence elements within some introns, such as poison exons, decoy exons and recursive splice sites (Conboy, 2021). However, this does not imply that every nucleotide of every intron is of equal importance as a nucleotide of coding sequence. Pathogenic mutations are discovered within known coding regions at a rate of at least 25% (Sawyer et al., 2016), much higher than the exonic proportion of genes (4.65%), indicating that mutations in introns are generally better tolerated.

Pathogenic mutations within introns are frequently found to affect splicing. Splicing mutations account for about 9% of all identified pathogenic mutations, though this figure also includes exonic mutations with splicing impact (Stenson et al., 2017). Most pathogenic splicing mutations weaken the definition of conserved branch point, acceptor site or donor site motifs of canonical exons, resulting in skipping of whole or partial exons, or inclusion of partial or whole introns (Abramowicz and Gos 2018). In some cases, however, a mutation will cause part of an intron to be erroneously spliced into mature transcripts as if it were an exon. These inclusions, when they occur, are called pseudoexons (PEs) or cryptic exons.

Mutations that cause pathogenic PEs in germline cells could accurately be described as ‘rare but ubiquitous.’ They are rare in that they appear to account for very few unique splicing events,

yet they are also ubiquitous in the sense that they can potentially arise in any gene with at least one intron, and do not exhibit any noticeable bias towards particular genes or cell types. Consequently, any new insights into how and why pathogenic PEs arise may have implications for numerous genetic diseases.

Surprisingly, there have been very few focused studies on the general characteristics of pseudoexons. Královičová and Vořechovský (2007) examined the exonic splice enhancer and silencer (ESE and ESS) content of pathogenic PEs and found that they were intermediate between introns and canonical exons. A 2010 study by Vořechovský investigated the correlation of PEs with transposable elements and found that MIRs (Mammalian-wide interspersed repeats) and antisense *Alu* elements were statistically overrepresented in pseudoexon sequences, which they attributed to exon-like characteristics that these elements naturally possess. Dhir and Buratti (2010) directly examined the mutations that activate PEs and proposed that they could be divided into five distinct categories: splice site or branch point creation, ESE gain or ESS loss, internal deletion, intragenic inversion, and loss of a flanking canonical splice site. Subsequent observations by Romano et al. (2013) of PEs in cancer-associated genes implicitly supported Dhir and Buratti’s categories, although Romano et al., additionally noted the possibility of PEs arising through other, as-yet uncharacterised mechanisms. More recently, Vaz-Drago et al. (2017) collated additional examples of PEs fitting Dhir and Buratti’s first two categories and observed a distribution of PE sizes similar to that of internal canonical exons. Lastly, our own review of pseudoexons in the *DMD* gene (Keegan 2020) examined the coincidence of PEs with reported recursive splicing, and suggested mutation-driven exonization of recursive splice sites as an explanatory mechanism for some PEs.

Although this handful of reviews provided many useful insights into the mechanisms of PE pathogenesis, all of them (including our own) were limited by their small sample sizes. For the older reviews in particular, this impediment was largely attributable to the scarcity of pseudoexon reports available at the time they were written, although a lack of clarity in how some primary reports presented their pseudoexon data may also have contributed. In recent years, the rate of reports of new pseudoexons has continued to accelerate as new technologies make RNA analysis faster, cheaper and more accurate. Therefore, it is timely to undertake a new and comprehensive analysis of pseudoexons and their instigating mutations.

In this report, we present a catalogue of 413 germline pseudoexon variants, which were as many as we could find through a thorough search of the literature. To our knowledge, this is the first time a PE dataset of this size has been assembled. Our analysis of this data discovered that 15.7% of PEs exhibit splicing activity at one or both of their splice sites in non-mutant cells, suggesting that many reported PEs might be more accurately reclassified as mutant variants of intronic splice regulating elements within introns, such as poison exons or decoy exons. Importantly, these shared intronic splice sites include seven empirically-verified recursive splice sites, confirming for the first time a long-suspected link between these two splicing phenomena.



Additionally, while our examination of the mutations that cause PEs largely supported the observations of Dhir and Buratti (2010), the expanded sample size of our catalogue allowed us to suggest refinements and additions to their original five categories.

## 2 MATERIALS AND METHODS

Although PEs have been observed in a highly diverse range of genes and cell types, they remain a relatively rare splicing phenomenon, and it was therefore unavoidable that any analysis of their characteristics would require some degree of compromise between specificity and sample size. In this section we will outline the criteria we adopted for determining what data to include in our analyses, what data to exclude, and why.

### 2.1 Working Definition of “Pseudoexon”

Our intention for this report was not to analyse all forms of cryptic splicing, but specifically those instances where splicing of non-canonical exons in a gene increased as the result of pathogenic mutations in that gene.

In a previous report on *DMD* gene PEs (Keegan, 2020) we suggested the following definition for PEs:

*“[A pseudoexon is] any continuous tract of a transcribed gene that: 1) does not overlap, adjoin or duplicate any sense-strand sequence of that gene’s canonical exons; 2) bears an acceptor splice site motif at its 5’ end and a donor splice site motif at its 3’ end; and, 3) via both these motifs, is spliced into a measurable proportion of the mature transcripts of that gene in at least one proband.”*

We adopted a streamlined version of that definition for this report:

*A germline pseudoexon is any continuous tract of a transcribed gene that: 1) does not overlap, adjoin, or duplicate any sense-strand sequence of a canonical exon; and 2) is spliced into mature transcripts of that gene in non-cancer cells of at least one proband 3) partly or wholly due to mutation in that gene.*

This modified definition allowed for the inclusion of PEs spliced via non-canonical motifs and PEs spliced in first-exon or last-exon orientations. It also enforces the exclusion of canonical exons introduced into other genes *via* gene fusions, which could technically have been classed as PEs under the previous phrasing.

#### 2.1.1 Exclusion of Cancer Pseudoexons

While many reports have detailed the correlation of PEs and various forms of cancer, we chose to limit our analysis to only those PEs that arose from germline mutations. This was because cancer cells often exhibit idiosyncratic changes in splice factor expression, and a general relaxation of splicing stringency, in comparison to non-cancer cells (Zhang et al., 2021). As such, we judged it would not be valid to make like-for-like comparisons

between PEs in cancers and those in germline cells. However, we did include PEs arising from germline mutations in cancer-associated genes, such as *NFI* and *ATM*, as in these cases carcinogenesis appeared to be a result of PE inclusion rather than the cause of it.

#### 2.1.2 Exclusion of Pseudoexons in Non-Humans

Although PEs have been observed in other animal species (e.g., Smith et al., 2007; Gómez-Grau et al., 2017), such observations are even rarer than they are in humans, a disparity that probably stems from the lower level of interest in mutation analysis in non-human species. We determined that the inclusion of PEs from non-human species would only offer a modest increase to our study’s sample size at the cost of greatly generalising its conclusions, and therefore limited its scope to human PEs only.

#### 2.1.3 Inclusion of Pseudoexons Identified *via* Transfected Minigene/Midigene Constructs

Wherever possible, it is ideal for investigations of splicing mutations to use RNA from patient cells that natively express the gene of interest. Unfortunately, this is not a practical option for many genes. For example, the Stargardt disease gene *ABCA4* is primarily expressed in kidney and retinal cells, and it is rare that a biopsy of either of these internal tissues can be justified. Instead, many researchers utilise HEK293T cells (immortalized human embryonic kidney cells), which they transform with minigene constructs of the *ABCA4* region of interest to model the effects of the mutation. In other cases, transformed COS cells (immortalised simian kidney fibroblast-like cells) or transformed HeLa cells (immortalised cervical cancer cell line) are used for similar purposes.

We elected to include in our dataset most of the PEs identified *via* minigene constructs, provided that the minigene constructs contained, at minimum, the entire intron surrounding the PE and both the flanking canonical exons. This minimised the chance of including a PE produced by construct-specific changes to its proximal sequence elements instead of the patient’s mutation. For the same reason, we also decided that if the effects of a splicing mutation were observed in both modified and unmodified cells, only the observations from the unmodified cells would be considered.

#### 2.1.4 Inclusion of ‘Terminal-Pseudoexons’

Virtually all of the reports surveyed in this analysis described internal PEs, i.e., PEs that were spliced into a gene transcript somewhere between its canonical first and last exons. However, there were some reports that appeared to detail “terminal-pseudoexons” (tPEs) arising from mutations that caused non-canonical sequence inclusions at the 3’ ends of largely canonical transcripts. We determined that these putative tPEs would only be classed as such, and included in our catalogue, if they could meet three criteria:

- (1) They possessed a novel acceptor site.
- (2) They did not overlap or adjoin any sense-oriented canonical exon sequence.
- (3) They possessed a functional polyadenylation site.

This third criteria is perhaps the most important, as it distinguishes “true” tPEs from more common events such as incomplete splicing and/or partial intron inclusion, which typically result in rapid nonsense-mediated decay of the affected transcript. We therefore only included tPEs if the supporting RNA analysis directly confirmed a *de novo* polyadenylation site, either *via* 3' RACE or through whole transcript sequencing.

## 2.2 Quality and Method of RNA Sequencing

The primary sources collated in this report span nearly 40 years of genetics research. As such, the RNA sequencing methods used by these sources run the full gamut of technologies, from S1 nuclease mapping to Nanopore. With few exceptions, we were agnostic towards the RNA sequencing technology used, provided that the PE sequence and splice sites could be mapped to the genomic reference sequence with a high degree of certainty. The level of detail provided in most reports made this a straightforward process, especially for those that had included Sanger sequence traces of the PE splice site junctions or Varnomen-format descriptions matched to specific reference sequences. In other cases, it was necessary to deduce PE sequences from precise but indirect details, such as the stated length of the PE relative to its instigating mutation.

In some reports, the stated boundaries of one or more PEs appeared to be exceptions to the U2-type GY-AG splicing that predominates in the dataset and did not fit the established motifs of U12-type splice sites either (Turunen et al., 2013). Given the rarity of such non-canonical exon boundaries in the human transcriptome (Parada et al., 2014), we judged that it was appropriate to exercise additional scrutiny, and we only included non-canonically spliced PEs if their supporting sequence data was unequivocal.

In cases where the published detail of a PE report was insufficient to determine the PE sequence, we contacted the corresponding authors with requests for further detail and have cited those that graciously responded as “Pers Comms” where appropriate.

## 2.3 Quantity of Pseudoexon Inclusion in Mature Transcripts

Our dataset does not incorporate quantitative data for the frequency of inclusion of each PE. While it could be argued that this weakens the analyses in some respects—since PEs with 10% inclusion are treated identically to PEs with 100% inclusion—we reasoned that avoiding quantitative analysis would be the “lesser of two evils.” Collectively, this study’s source reports show enormous variation in the genes studied, the types of cells used, and the methods of RNA analysis, with many producing sequence data that was non-quantitative or at best semi-quantitative. Accurate and objective standardisation of these data would have been all but impossible and risked generating misleading conclusions.

## 2.4 Search Criteria

Literature search was performed using Google Scholar and the Murdoch University academic research portal *FindIt*. Individual searches of the following terms were conducted through both portals: “*pseudoexon*,” “*pseudo exon*,” “*cryptic exon*,” and “*deep*

*intronic*.” The first three of these terms comprise the most common descriptors for PEs, while the fourth term served as a “safety net” to return any publications of deep intronic splicing mutations that may have reported on PEs using unexpected terminology.

In addition to scrutinising each paper for useful data, we also performed searches within each paper for the key terms mentioned above and investigated any references that were cited against these mentions. This allowed us to discover additional PE reports that, for various reasons, had escaped capture by our direct searches.

Similarly, we are indebted to the authors of several previous reviews of PEs, whose works led us to additional primary sources that had eluded discovery through the above methods (Královičová and Vořechovský, 2007; Dhir and Buratti 2010; Vořechovský 2010; Romano et al., 2013; Vaz-Drago et al., 2017). We have also incorporated data for *DMD* PEs that was originally collated for a previous report (Keegan 2020).

## 2.5 Construction of Pseudoexon Catalogue

### 2.5.1 Transcribed Pseudoexon Features

The rarity of PEs, coupled with broad variability in how they were reported, made it unfeasible to automate their annotation. Therefore, all annotation of PE data was performed manually by the authors. This approach was further justified *post hoc* by discoveries of minor inconsistencies in numerous PE reports (e.g., stated splice sites or lengths that differed from those shown in the published figures), errors that would have escaped detection by any automated process and led to inclusion of inaccurate data.

For each PE, annotated data fields included the name of the affected gene using current nomenclature, as listed on the Genecards Human Gene Database (Stelzer et al., 2016); the sequence of the PE; the sequences of the flanking exons to which it was spliced; the cell or tissue type(s) in which its splicing was observed; the Varnomen cDNA code for the instigating mutation(s) if present and known (den Dunnen et al., 2016); additional notes if relevant; and citations for the primary sources of the data.

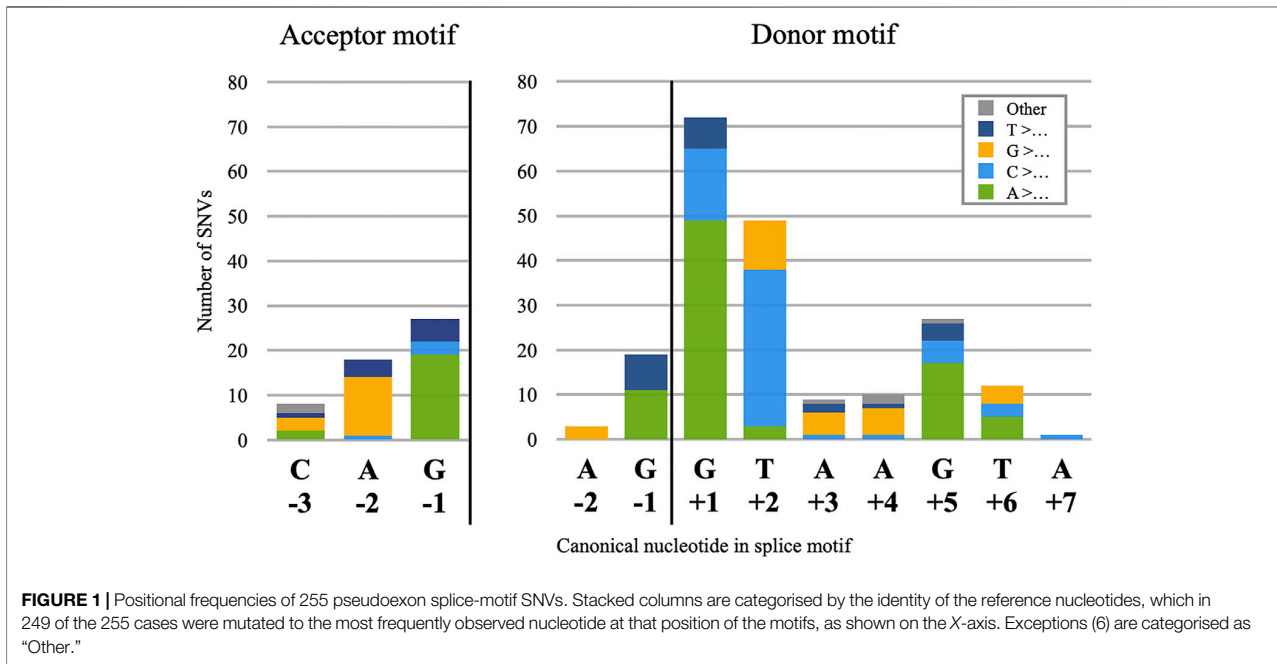
While transcript reference sequences are included for the encompassing gene of each PE, in many cases a specific refseq ID was not explicitly declared in the source. The listed reference sequences instead correspond to the lowest-numbered transcript variant (usually TV1) that matched the splicing patterns observed and have been included to allow the use of cDNA-type Varnomen mutation codes, which are more human-readable than genomic codes. Chromosomal coordinates and all other genomic features refer to the most recent human genome assembly, GRCh38.p13.

### 2.5.2 Intron Numbering

Historically, unique identifiers have been employed when referring to the introns of certain genes, such as *NFI* and *CLRN*. For consistency, we have ignored these in favour of simple 1-to-*n* numbering for all transcript variants, but caution the reader that this may create the appearance of discrepancies when referring to some cited reports.

### 2.5.3 Assignment of Unique Pseudoexon IDs

Each pseudoexon was assigned a unique ID according to the name of its gene, the number of the encompassing intron, and an



alphanumeric identifier according to its similarity to other pseudoexons in that intron. For example, two pseudoexons with completely distinct sequence were reported in the 17th intron of the *ATM* gene and were IDed as *ATM-17-1* and *ATM-17-2*, while the two PEs reported in *ATM* intron 27 were IDed as *ATM-27-1a* and *ATM-27-1b* due to sharing an acceptor site.

#### 2.5.4 Citations

While descriptions of most PEs were limited to a single report, some were reported multiple times by different research groups and to varying levels of detail. An extreme example of this is *CFTR-7-1*, a prevalent disease allele for cystic fibrosis that has been studied extensively. In the interest of clarity, we limited our citations for each PE to its earliest known report, including later reports only if they substantially added to the characterisation or were already cited for unique observations of other PEs.

#### 2.5.5 Derived Features

Maximum entropy (MaxEnt) splice site scores were calculated via the Burge Laboratory's *MaxEntScan* web-tool (Yeo and Burge, 2004). Sizes and distances were directly calculated using spreadsheet formulae.

### 3 PSEUDOEXON MUTATION CHARACTERISTICS

#### 3.1 Pseudoexon Splice Motif Mutations

##### 3.1.1 Pseudoexon Donor Motif Mutations

Donor splice site mutations are the most frequently observed cause of PE pathogenesis, comprising 210 of the 359 catalogued distinct mutations. Of these 210, 202 are single nucleotide variants (SNVs), and the frequency of mutation for each

nucleotide position relative to the donor site (**Figure 1**, right) approximates the degree of nucleotide conservation observed for that position in canonical sites (Ma et al., 2015). This may be a logical consequence of how the spliceosome binds to candidate donor sites: because the effect of nucleotide identity on spliceosome binding varies greatly across the motif, changes at the most essential positions will have the greatest effect and the best chance of "breaking through" the silencing mechanisms that would otherwise prevent detectable levels of splicing. This is corroborated by our observation that no PE in our dataset was instigated by an SNV at the donor site  $-3$  position, despite this nucleotide falling inside the donor site motif. Because the  $-3$  position is not highly conserved, any change in this nucleotide is unlikely to cause a noticeably pathogenic increase in PE inclusion.

However, we did note a single PE with a C>A SNV 7 nt 3' of the donor splice site (*MMUT-11-1b*). Although the donor site motif is traditionally considered to end at the 6th 3' nucleotide, a comprehensive analysis of human splice sites (Ma et al., 2015) reveals a substantial bias towards purines at the +7 position (59% A/G). As this *MMUT* SNV is a pyrimidine-to-purine transition, the simplest explanation of it is as a donor site mutation.

##### 3.1.2 Pseudoexon Acceptor Motif Mutations

Acceptor splice-site mutations account for 53 of the catalogued PE mutations, and all but one of those (*DMD-30-1*) were SNVs. Despite the greater size of the acceptor motif compared to the donor, pathogenic acceptor site SNVs were limited almost entirely to positions  $-2$  and  $-1$ , with only an additional eight at  $-3$  (**Figure 1**, left). As with the distribution of donor site mutations, this appears to be a result of low conservation at positions  $+1$  to  $+3$  and the low impact of individual nucleotides in the  $-20$  to  $-4$  range, with pathogenic mutations in this latter

region apparently impacting branch point definition more than they did acceptor site definition. We also noted a single case of an SNV at the PE acceptor +3 position, (*ABRAXAS1-5-1*) but chose to analyse this as an internal mutation since its effect on the acceptor splice score was negligible.

Our observations of position frequency in PE donor and acceptor splice motif SNVs generally accorded with those of Krawczak et al. (2007), who examined a much larger set of gain-of-function splice mutations; and with those of Sakaguchi and Suyama (2021), who examined a smaller dataset consisting entirely of novel PEs. We also noted that the number of transition SNVs—purine-to-purine or pyrimidine-to-pyrimidine nucleotide changes—was approximately double that of transversion mutations, at 165 and 92, respectively. This accords with prior observations of how often each mutation type is generally observed in the human genome (Jiang and Zhao 2006).

### 3.2 Pseudoexon Internal Mutations

We catalogued 37 examples of PEs caused by sequence changes between the PE splice site motifs. In five of these examples (*COL4A5-37-1*, *DMD-11-2*, *DMD-34-1*, *DMD-48-1* and *GNAS-AS1-4-1*—see **Supplementary Table S1**) the mutation was a >10 kb deletion that brought a latent acceptor-donor motif pair into conjunction. In these cases, we assumed that sheer distance between the splice sites was the chief silencing element that had been lost and did not analyse these further. Of the remaining 32 cases, there was one PE (*GLA-4-1*) caused by a 113 nt insertion, three PEs caused by 2–4 nt deletions, and 28 PEs caused by SNVs (**Supplementary Table S2**).

The positions and predicted effects of the 31 unique mutations showed much greater variability than the mutations affecting PE splice motifs. Although we could not distinguish any obvious patterns in their locations within the PEs, we noted that the primary reports consistently described these mutations as gains of ESE motifs and/or losses of ESS motifs within the PEs. Most of these assessments were made *via* an assortment of RNA motif analysis utilities, some of which are no longer available. We therefore standardised our re-analysis to a single utility, *HExoSplice*, which was designed specifically for analysing this type of mutation (Ke et al., 2011; Tubeuf et al., 2020). Because *HExoSplice* only calculates scores for SNVs, we derived scores for the deletion and insertion mutations manually by subtracting the total score for the wild-type exon from the total score of the mutant. Impressively, *HExoSplice* correctly predicted the directionality of 29 out of the 31 mutations with a net increase in the score ( $\Delta Hx$ ). There were no obvious similarities between the two SNVs with negative  $\Delta Hx$  scores (*ABCA4-30-1c* and *DMD-32-1b*). We suggest that in both these cases, the mutations may have altered binding of splicing factors specific to those genes or cells, as this kind of specific effect cannot be accurately predicted by a generalised tool such as *HExoSplice*.

### 3.3 Pseudoexon Branch Point Mutations

Just 14 of the 359 PE mutations were ultimately classified as altering pseudoexon branch point definition, and these mutations showed considerable variation in their nature and location

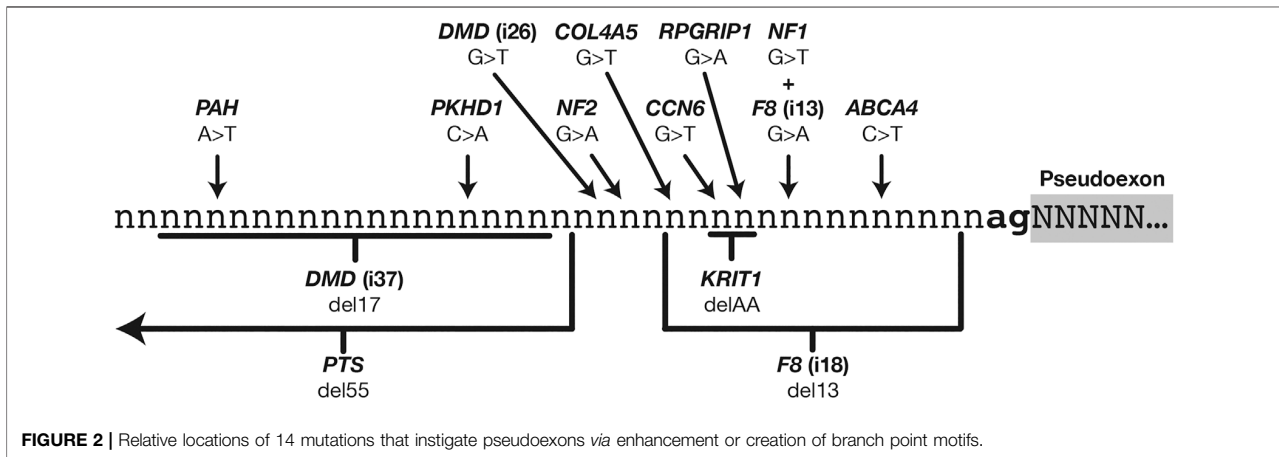
relative to the PE acceptor site (**Figure 2**). Historically, branch point mutations have been poorly understood and have proven more challenging to predict than other splicing mutations (Canson et al., 2020). Despite this variability we observed that the mutations we had catalogued were remarkably consistent in their effects on branch point characteristics (**Table 1**).

Branch point definition requires two elements: a functional branch point site (the motif for which is weakly conserved in U2 introns) and a continuous AG-exclusion zone (AGEZ) connecting the branch point to a downstream acceptor splice site. In 12 out of the 14 cases shown here, the effect of the mutation was to “close the circuit” between an acceptor site and a branch point, through some combination of improving an in-range branch site motif (as predicted by *SVM-BPfinder* with “AGEZ only” selected—see Corvelo et al., 2010), increasing the size of the AGEZ, or moving an existing branch point closer to an AGEZ (Gooding et al., 2006).

Two exceptions to this rule were *ABCA4-6-1* and *CCN6-2-1*, both instigated by SNVs that modestly increased their acceptor site scores (6.63 → 7.12 and 3.12 → 5.43, respectively) but did not affect their AGEZ size or branch point scores. Although we could defensibly have classified these as acceptor-motif mutations, we noted that every other acceptor-motif mutation fell within three nucleotides 5′ of the PE (**Figure 1**, left) and would therefore directly affect binding of the U2AF35 spliceosome component (Voith von Voithenberg et al., 2016). We reasoned that the positioning of the *ABCA4-6-1* and *CCN6-2-1* mutations within their respective polypyrimidine tracts suggested at least some interaction with other spliceosome components. A third mutation similar to these (c.639+861C>T) was reported in *GLA* (Filoni et al., 2008), although the affected exon in this case was subsequently classified as a canonical exon of transcript variant NR\_164783 (**Supplementary Table S3**). A causative mechanism was not empirically confirmed for any of these three mutations. However, a fourth mutation, similar in type but opposite in effect, was reported by van der Wal et al. (2017). This mutation (c.-32-13T > G) occurred 13 nt 5′ of exon 2 in the gene *GAA* and caused skipping of all or part of exon 2. In this case, Van der Wal et al., were able to empirically demonstrate that the splicing disruption arose from a loss of U2AF65 binding at the mutation site.

The U2AF65 protein is a U2 spliceosome component that binds pyrimidine tracts 5′ of exon acceptor sites, interacting with the U2 snRNP to facilitate branch point recognition (Valcárcel et al., 1996). It has multiple known pyrimidine-rich binding motifs (Paz et al., 2014; Drewe-Boss et al., 2018). Since the *ABCA4-6-1* and *CCN6-2-1* mutations are both SNVs that create new thymine nucleotides, we theorised that the true pathogenic effect of these mutations is creation or enhancement of a U2AF65 binding site. Unfortunately, we could not test this prediction as there is at present no *in silico* tool for predicting U2AF65 binding that incorporates all known motifs. We therefore added “new pyrimidine nucleotide” as a crude predictor of U2AF65 binding.

We also noted that these three mutations are predicted by *Splice Aid 2* (Piva et al., 2012) to create new PTBP1 (hnRNP I) binding sites. Although PTBP1 plays a complex role in splicing



**FIGURE 2** | Relative locations of 14 mutations that instigate pseudoexons via enhancement or creation of branch point motifs.

**TABLE 1** | Pseudoexon mutations that enhance pseudoexon branch point motifs.

Gene	Intron	#	Mutation	AGEZ	Max BPS	Pyr SNV?	References
ABCA4 (NM_000350.3)	6	1	c.769-784C>T	17	0.22	Yes	Sangermano et al. (2019)
CCN6 (NM_003880.4)	2	1	c.49-763G>T	24	-0.15	Yes	Garcia-Segarra et al. (2012)
COL4A5 (NM_000495.5)	29	1	c.2395+1275C>G; c.2395+1292G>T	<b>13 -&gt; 59</b>	<b>-1.28 -&gt; -0.03</b>	Yes	Wang et al. (2021)
DMD (NM_004006.3)	26	2	c.3603+820G>T	<b>16 -&gt; 42</b>	<b>0.57 -&gt; 1.00</b>	Yes	Waddell et al. (2021)
	37	2	c.5325+1740_5325+1757del	17	<b>-1.33 -&gt; -0.06 (MC)</b>	-	Bovolenta et al. (2008)
F8 (NM_000132.4)	13	2a	c.2113+461_2113+473del	74 -> 61	<b>0.83 (MC)</b>	-	Jourdy et al. (2018)
	18	2b	c.5999-798G>A	<b>8 -&gt; 65</b>	0.23	No	Pezeshkpoor et al. (2013)
KRIT1 (NM_194456.1)	6	1	c.262+132_262+133del	<b>9 -&gt; 88</b>	0.14	-	Riant et al. (2014)
NF1 (NM_001042492.3)	8	3	c.889-941G>T	<b>8 -&gt; 21</b>	-0.73	Yes	Pros et al. (2008)
NF2 (NM_000268.4)	5	1	c.516+232G>A	<b>15 -&gt; 40</b>	<b>0.06 -&gt; 3.37</b>	No	De Klein et al. (1998)
PAH (NM_000277.3)	11	1	c.1199+502A>T	34	<b>-1.64 -&gt; -0.16</b>	Yes	Jin et al. (2021)
PKHD1 (NM_138694.4)	56	1	c.8798-459C>A	46	<b>-1.42 -&gt; 1.95</b>	No	Chen et al. (2019)
PTS (NM_000317.3)	2	1b	c.163+696del55	<b>17 -&gt; 28</b>	<b>-1.83 -&gt; 1.01 (MC)</b>	-	Meili et al. (2009)
RPGRIP1 (NM_020366.4)	12	1a	c.1468-263G>C	<b>10 -&gt; 24</b>	-0.56	Yes	Jamshidi et al. (2019), Zou et al. (2021)

"AGEZ," "Max BPS," and "Pyr SNV?" changes that are predicted to enhance branch point definition are in bold text. Cases where deletion mutations moved a branch point site closer to the AGEZ/acceptor site are noted in the "Max BPS" column as "(MC)."

(Lou et al., 1999; Han et al., 2014) it is generally observed to silence nearby exons. However, in some cases PTBP1 has been shown to indirectly enhance 3' exon inclusion by antagonising splicing repressors (Paradis et al., 2007) or preventing erroneous binding of U2AF65 (Sutandy et al., 2018), so we cannot rule out a possible mechanistic role for PTBP1 in these three mutations. There is clearly a need for an accurate and comprehensive *in silico* tool that can predict the effects of all types of branch point mutations, as they remain something of a blind spot in splice mutation research.

### 3.4 Distal Mutations

All the pathogenic mutations discussed thus far have entailed some form of direct improvement to the exon-like characteristics of the PE, be it the splice sites, branch point or local enhancer/silencer balance. However, we have also catalogued 35 cases of PEs being instigated by intragenic mutations outside the span of

the PE's branch point and donor site motif, sometimes multiple canonical exons and tens of kilobases away. When viewed individually, the aetiology of these PEs may appear baffling, but examining them *en masse* reveals many important common features.

#### 3.4.1 Decreased Definition in Adjacent Canonical Exons

In their 2010 report, Dhir and Buratti proposed that loss of a canonical splice site "facing" a pseudoexon—i.e., a 5' donor site or a 3' acceptor site—may be a general mechanism of pseudoexon pathogenesis. Impressively, although their prediction was based on just five supporting examples (*BRCA2-20-1a*, *CFTR-3-1*, *IDS-3-1a*, *IDS-3-1d* and *MMUT-11-1d*) we found that it was generally supported by the characteristics of 13 of the additional mutations we collated (**Supplementary Figure S1**), albeit with refinements to the original terms of their category.



The ten examples shown in **Supplementary Figure S1A** most closely fit the terms of Dhir and Buratti's original category, as all entailed loss or weakening of a 5' donor or 3' acceptor site. This includes one mutation that appeared to weaken a 3' branch point (GAA-1-1), thereby indirectly weakening definition of the acceptor site, and one case (DGKE-5-1) where a mutation created a cryptic upstream donor site but left the original intact, indirectly weakening its definition via competitive effects. In addition, we also observed three examples of mutations that induced skipping of a whole exon (**Supplementary Figure S1B**)—including an alternative transcript variant caused by the same GAA-1-1 branch point mutation shown in **Supplementary Figure S1A**—and three examples of mutations that caused a net loss of enhancers within a flanking canonical exon (**Supplementary Figure S1C**), as indicated by their negative  $\Delta Hx$  scores. We were also interested to note that in every case where a PE was spliced to an alternative flanking splice site—whether through cryptic splice site activation or whole exon skipping - the new site always had a lower MaxEnt score than the unmutated original.

However, this pattern is further complicated by three cases of PEs instigated by mutations in “next-but-one” neighbour exons that were not directly spliced to the PE (**Supplementary Figure S1D**). These cases may be a consequence of the exons near these PEs having tightly linked splicing fates, e.g., a loss of definition in *MMUT* exon 4 exerts a similar effect as a hypothetical loss of definition in exon 5 would. An analogous example of this kind of splicing effect can be seen in cases where a mutation within a single canonical exon caused skipping of both that exon and adjacent, unmutated exons (Fisher et al., 1993; Suzuki et al., 2016). We also note that there is substantial evidence of exons in many genes being consistently spliced out of transcription order (Takahara et al., 2002; Attanasio et al., 2003; Kim et al., 2017) and speculate that this, too, may contribute to the aetiology of some highly distal pseudoexon mutations. In general, exon-like tracts in late-spliced introns could be more vulnerable to the knock-on effects of splicing changes elsewhere in the maturing transcript, while those in early-spliced introns remain insulated by sheer distance.

We also observed that one of the three mutations shown in **Supplementary Figure S1D** (*DMD-3-1a*) was an intra-exon mutation with a positive  $\Delta Hx$  score, indicating an increase in the definition strength of that exon. Although this may represent a valid counterexample to the prevailing trend of pathogenically decreased exon definition, we tentatively dismissed it as another of the inevitable minority of incorrect predictions made by the *HExoSplice* algorithm and note that one of the two other such examples we observed also occurred in the *DMD* gene (*DMD-32-1b*—see **Supplementary Table S2**).

Considered in aggregate, these observations suggest the common mechanism underlying these PEs is a comparative weakening of definition in the successfully spliced neighbouring exons; and *vice versa*, that the definition strength of exons flanking an intron can be an important mechanism for silencing latent PEs within. This hypothesis is supported by prior observations of stronger splice motifs in exons flanking large introns (Farlow et al., 2012), since larger introns could generally

be assumed to have a higher chance of containing at least one latent PE. The infrequency with which these types of PEs arise, compared to PEs generated by direct enhancement of their donor or acceptor splice sites, may be a result of general selective pressure against splice-competent elements within intron tracts. We also speculate that pseudoexons of the kind seen in **Supplementary Figure S1C** may be more common than suspected and could account for the pathogenic nature of some synonymous variants (Shi et al., 2019), perhaps having escaped detection in prior experiments due to nonsense-mediated decay (NMD) of the affected transcripts.

### 3.4.2 Novel Pseudoexon Mutation Categories

We observed 17 examples of PE mutations that bore no resemblance to Dhir and Buratti's five categories, some of which bore sufficient similarities to each other to justify new categories (**Supplementary Figure S2**).

#### *Proximity to a Directly Mutated Pseudoexon in the Same Intron*

The first of these categories comprises four cases where PEs were apparently instigated by the activation of a second PE in the same intron (**Supplementary Figure S2A**). In all four cases the mutation in the ‘primary’ PE created or enhanced a donor site, and all occurred in introns of similar sizes, ranging from 1,212 nt (*MYBPC3* intron 20) to 2,582 nt (*MYBPC3* intron 12). It is also notable that in each of these cases, the primary PE introduces a flanking splice site as strong or stronger than that of the nearby canonical exon. For example, the primary PE in *F8-12-2a* introduces a downstream acceptor site stronger (MaxEnt = 8.80) than that of exon 19 (MaxEnt = 6.91). This would appear to contradict the pattern of PEs arising from weakening of flanking exon definition. However, it may be that in these cases the intron subdivision caused by splicing of the primary PEs leads to a disruption of splicing co-ordination (Drexler et al., 2020), and a consequent loss of silencing of the secondary PE that exceeds the expected gain of silencing from the primary PEs' strong flanking splice sites. An assay of intron splicing order in cells carrying these mutations, or other similar mutations, may shed much light on the processes involved.

#### *Loss of Upstream Polyadenylation Motifs*

There was a single case of a PE arising from an intragenic region due to an upstream deletion (*EPCAM-7-1*, **Supplementary Figure S2B**). While we would otherwise hesitate to define an entire category by a single exemplar, in this case the causative mechanism is straightforward enough to justify it: Genomic deletion of the latter two exons of *EPCAM*, which necessarily entailed deletion of that gene's polyadenylation signals, permitted transcription to continue through the intergenic region and into the 3' gene *MSH2*, which shares *EPCAM*'s sense-strand orientation. By chance, this novel “intergenic intron” contained a 111 nt tract that was sufficiently exon-like to be spliced to the neighbouring canonical internal exons, namely *EPCAM* exon 7 and *MSH2* exon 2 (Ligtenberg et al., 2009).

Although this is the only example of this type of PE that we catalogued, similar PEs could occur in other cases where a

genomic polyadenylation site deletion is followed by a 3' gene with the same strand orientation and at least one intron, and where the splice sites involved are in sufficient proximity.

### Change to Proximal Intronic Splice Motifs

We catalogued six cases where PEs were instigated by mutations within the same intron but beyond the PE splice motifs and branch points (**Supplementary Figure S2C**). Three of these mutations were 3' of the PE (*FBOX38-9-1*, *MFG8-6-1*, and *NPHP3-3-1*) and three were 5' of the PE (*DMD-56-1*, *NR2E3-7-2*, and *RPGR-9-1*). Here we must clarify that although *DMD-56-1* bore deletions both 5' and 3' of the PE, the reporting authors empirically demonstrated that only the 5' deletion caused PE inclusion (Khelifi et al., 2011).

The *MFG8-6* and *NPHP3-3* mutations were both SNVs similar distances from the PE donor sites (43 and 50 nt, respectively) that created new predicted FUS (hnRNP P2) binding motifs (Piva et al., 2012). Studies in mouse cells (Ishigaki et al., 2012) have shown that FUS binding along flanking introns can regulate alternative exon splicing in neuronal cells, so it is possible that a perturbation of normal FUS binding is responsible for these PEs escaping silencing.

A similarly positioned SNV in *FBOX38-9-1* destroyed a predicted binding site for hnRNP K (Piva et al., 2012). This is consistent with a recent report demonstrating that hnRNP K depletion can lead to a widespread increase in cryptic exon inclusion, and that at least some of these cryptic exons are ordinarily silenced by hnRNP K binding within 100 nt of the 3' intron (Bampton et al., 2021).

Unfortunately, there are few such similarities to connect the three PEs with 5' distal mutations. The first, *DMD-56-1*, is caused by a 592nt deletion ending 26 nt 5' of the PE acceptor site. The authors experimentally excluded modified branch point definition as a causative factor for this PE, but despite thorough experimentation with minigene assays they could not positively identify which components of the deleted region were responsible for the PE's inclusion or what their mode of action was. The second PE with a distal 5' mutation, *NR2E3-7-2*, was instigated by an SNV 581 nt upstream that altered multiple splice factor binding sites, making it difficult to predict which, if any, are mechanistically responsible. Incidentally, this was the same mutation that created an acceptor motif AG dinucleotide in *NR2E3-7-1*, though unlike the examples discussed in the first category of *Novel Pseudoexon Mutation Categories*, these two PEs exhibit mutually exclusive splicing.

In the third PE with a distal 5' mutation, *RPGR-9-1*, a TTAAA motif is created 53 nt from the acceptor site. This motif is predicted to bind KHDRBS1 (Sam68) and/or KHDRBS3 (SLM-2), two splicing factors with high homology and similar effects on pre-mRNA splicing (Danilenko et al., 2017). In particular, KHDRBS1 has been shown to aid in the splicing of introns bearing *Alu* retrotransposon sequences (Pagliarini et al., 2020). Two such *Alu* elements occur within *RPGR* intron 9 (**Supplementary Figure S3**). Although the true pathology of this mutation is yet to be empirically determined, it may be that a disruption to KHDRBS1-mediated splicing is responsible for the *RPGR-9-1* pathogenesis.

### Unknown Mechanisms

In six cases, the connection between the identified mutation and the PE was unclear (**Supplementary Figure S2D**) and these cases bore no similarities to other catalogued examples. However, we note that *DYSF-51-1b* is an identical sequence inclusion to *DYSF-51-1a* (which arose from a PE donor site mutation), and has been observed at low levels in cells from healthy donors (Gonorazky et al., 2019); and similarly, that *DMD-3-1a* is also instigated by a 1 nt deletion in exon 5 (**Supplementary Figure S1D**).

## 3.5 Summary of Pseudoexon Mutation Analysis

Our pseudoexon catalogue, which is to date the most comprehensive ever assembled, confirms that PEs are most frequently instigated by direct mutation of their local splicing motifs; that the most frequently mutated components are the PE donor and acceptor splice motifs; and that the predominant type of instigating mutation is single nucleotide substitution. These findings support previously published observations of smaller pseudoexon datasets, which we gratefully acknowledge as secondary data sources for this catalogue. We add to this several novel classifications for rarer types of PE-instigating mutation, the most well-supported of these being mutations that weaken definition of adjacent canonical exons.

## 4 LATENT FACTORS CONTRIBUTING TO PSEUDOEXON SPLICING

Considering the complexity and stringency of vertebrate exon definition, in conjunction with the observation that single-nucleotide substitutions are the most frequent cause of PE pathogenesis, we are forced to question why these exon-like intron tracts exist in the first place. Even if the reference allele of a given PE is ultimately excluded from mature transcripts by the lack of one crucial splice motif, the presence of all the other exonic motifs might still encourage abortive "false start" activity by the spliceosome, wasting energy and unnecessarily prolonging mRNA maturation.

It may be that the latent elements of some PEs are mildly deleterious in this way but persist in the genome simply as another of evolution's myriad compromises and works-in-progress. However, we must also examine the alternative explanation that these latent elements persist due to their spliceosome interactions being benign or even beneficial, and consider the various forms these interactions may take.

### 4.1 Canonical Exon Splice Variants

The earliest reported PE to meet the criteria of this catalogue (Dobkin et al., 1983) predates the completion of the first rough draft of the human genome project by nearly 18 years (Lander et al., 2001), and the years that intervened and followed these milestones have seen numerous revisions to the official coding sequences and transcript variants of thousands of genes. An inevitable side-effect of this progress is that many splicing phenomena initially reported as PEs have subsequently been

**TABLE 2 |** Pseudoexons associated with seven confirmed intronic recursive splice sites.

Gene	Intron	#	Start	End	Size	Pseudoexon mutation(s)	ME-A	ME-D	PE RNA source	PE references	RSS RNA source	RSS references
ATM (NM_001351834.2)	27	1b	108287410	108287521	112	c.3994-159A>G (A+32)	7.71 -> 8.12	8.49	LCLs	Coutinho et al. (2005)	HBECs	Wan et al. (2021)
		1a		<b>108287438</b>	29	c.3994-193C>T (A-3)		<b>6.38</b>	LCLs; peripheral blood	Coutinho et al. (2005), Královíčová et al. (2016), Landrith et al. (2020)		
COL4A5 (NM_000495.5)	6	1	108570649	<b>108570795</b>	147	c.385-719G>A (A+46)	5.25	<b>7.51</b>	Hair bulb	King et al. (2002)	HBECs	Wan et al. (2021)
FBOX38 (NM_030793.5)	9	1	<b>148411080</b>	148411238	159	c.1093+532C>G (D+59)	<b>9.11</b>	6.57	Whole blood, lung tissue	Saferali et al. (2019)	HBECs	Wan et al. (2021)
GLA (NM_000169.3)	3	1	101401233	<b>101401347</b>	115	c.547+395G>C (D-5)	<b>5.10</b>	7.82	Whole blood	Higuchi et al. (2016)	Cerebellum, K562 cells	Blázquez et al. (2018)
MCCC2 (NM_022132.5)	10	1	<b>71636104</b>	71636167	64	c.1054G>A (e11 D-19)	<b>5.72</b>	3.24	Emetine-treated fibroblasts	Stucki et al. (2009)	HBECs	Wan et al. (2021)
NPHP3 (NM_153240.5)	3	1	<b>132717955</b>	132718117	163	c.671-996C>G (D+50)	6.50	<b>0.56</b>	Leukocytes	Larrue et al. (2020)	HBECs	Wan et al. (2021)
OCRL (NM_000276.4)	4	1	<b>129553236</b>	129553301	66	c.239-4023A>G (D+1)	<b>8.18</b>	2.68 -> 10.86	Skin fibroblasts	Rendu et al. (2017)	PA1 cells	Zhang et al. (2018)

Genomic coordinates of the recursive splice sites and their maximum entropy scores are in bold text. "ME-A" and "ME-D" refer to the Maximum Entropy scores for the acceptor and donor splice sites, respectively.

reclassified as either canonical exons or mutant splice variants thereof. In the course of assembling and curating this catalogue, we separately collated 35 such examples (**Supplementary Table S3**). These examples could not be included in any of our PE analyses since there is no meaningful distinction between them and other canonical exon splice mutations. However, they serve as a useful reminder of the difficulty in distinguishing PEs from mutant variants of as-yet-unannotated canonical exons, especially if the canonical exons are expressed at low frequencies or in unexamined cell types (Ray et al., 2020). We expect that progress in transcriptomics will eventually necessitate similar reclassification for at least some of the PEs in this catalogue.

## 4.2 Novel or Unannotated Canonical Exons

Having excluded from our catalogue those PEs that coincided with known canonical exons, we attempted to annotate additional examples of PEs that might undergo this reclassification in future. Our criteria for inclusion were 1) the PE must show evidence of splicing in normal cells for at least one of its splice sites, either in the original report, in non-cancer cell spliced expressed sequence tags (ESTs) from the UCSC Genome Browser's "Spliced ESTs" track (Kent et al., 2002) or in paired-end RNAseq data (Sibley et al., 2015); and 2) inclusion of the PE in the mature transcript must be predicted not to trigger NMD.

A total of six PEs met these criteria (**Supplementary Table S4**). In all six cases, NMD avoidance was predicted due to preservation of the open reading frame and absence of any novel stop codons. We also allowed for cases where a transcript variant with a premature stop codon may have escaped NMD due the stop codon being introduced less than 55 nt from the final splice junction of the

transcript (Zhang et al., 1998) but did not find any examples that met this criterion.

## 4.3 Poison Exons and Decoy Exons

In recent years the term "poison exons" has been steadily gaining prominence in literature related to cryptic splicing phenomena. Carvill and Mefford (2020) characterised poison exons as conserved, alternatively spliced exons containing one or more premature termination codons that are spliced into unneeded transcripts to prevent their translation and target them for nonsense-mediated decay. "Decoy" exons behave similarly but are characterised by their additional capacity to non-productively interact with adjacent canonical splice sites, thereby promoting whole intron retention (Conboy, 2021).

There is a clear overlap between the definitions of poison/decoy exons and PEs, although the phenomena are not identical. Both describe non-canonical exon inclusions that generally impair the translation of full-length, functional protein from the affected transcript; but while PEs arise due to intragenic mutations and are often deleterious to the health of the patient, poison exons are a normal component of splicing that may contribute to fine-control of gene expression and are presumably beneficial, or at the very least benign.

Given the similarities between PEs and poison exons, and the relative novelty of the latter term, the intriguing possibility emerges that some of the splicing phenomena historically reported as PEs might be better re-classified as poison exons, or splice variants thereof.

Having already determined the concurring splice site reads between our PE catalogue, and ESTs and RNAseq data (see



**Section 4.1)** we separately tabulated all those examples where evidence supported their splicing in normal cells, but which did not preserve the transcript open reading frame (**Supplementary Table S5**).

A possible reason for the high number of candidate poison exons seen in *NF1* and *DMD* is the exceptional size and high intron count of these genes. These features unavoidably entail a long transcription and maturation time, which must be reconciled with the fact that the quantity of any encoded protein that the cell needs can change dramatically in a matter of seconds. The more poison exons a transcript contains, the more possible time-points there are for interrupting the reading frame and preventing an unneeded transcript from reaching functional maturation.

Of the 413 catalogued PEs, for 65 (15.7%) we found evidence of splicing of at least one splice site in normal cells. This is a remarkably high concordance when one considers that, for the most part, splicing of putative PEs in normal cells is not something that has been systematically investigated; as such, what supporting evidence there is exists largely by chance. As RNAseq becomes more commonplace and is applied with greater sensitivity and read depth to a broader range of cell types, it may emerge that many more PEs—perhaps even a majority—originated as benign rare exons or functional exon-like intronic sites.

#### 4.4 Recursive Splice Sites

In a previous report focused on PEs in the *DMD* gene (Keegan 2020), we examined the possibility that some PEs may arise from the errant splicing of predicted recursive splice sites (RSSes). Here we sought to examine this possibility as it applies to our total set of PEs. This task was complicated by the fact that there is as yet no consensus on the precise definition of RSSes and how best to experimentally verify their presence. For example, the criteria employed by Zhang et al. (2018) required that a putative RS-exon should bear an agGT tetranucleotide at the acceptor site and that the nucleotides around the acceptor site should be highly conserved, while the approach of Wan et al. (2021) was agnostic to sequence conservation.

We searched the splice site coordinates of our PE dataset against five published datasets of recursive splice sets (Kelly et al., 2015; Sibley et al., 2015; Blázquez et al., 2018; Zhang et al., 2018; Wan et al., 2021). We did not find any matches in Kelly et al. (2015) or in the filtered results of Sibley et al. (2015), but we did discover seven matches in the filtered results of the other three reports—five in Wan et al. (2021) and one in each of Blázquez et al. (2018) and Zhang et al. (2018) (**Table 2**). To our knowledge, this is the first conclusive evidence supporting our earlier hypothesis that pathogenic PEs can arise from mutations near recursive splice sites (Keegan 2020). Additionally, we were interested to note that six of the seven recursive splice sites were also spliced as components of putative poison exons in normal cells (**Supplementary Table S5**), with *COL4A5-6-1* being the exception. This may indicate that these sites serve a dual purpose in splicing regulation, though this remains to be confirmed through functional studies.

## 5 Unique Cases and Additional Observations

### 5.1 No Known Pseudoexons are Processed by the U12 Spliceosome

The minor spliceosome, or U12 spliceosome, processes just 0.37% of all human introns (Olthof et al., 2019). Type-U12 introns can most easily be recognised by their highly conserved donor-site (UTATCCT) and branch point (CCTTUAY) motifs, and their tolerance for AT-AC terminal dinucleotides—although the latter feature is not present in all U12 introns and GT-AC, AT-AG or GT-AG terminal dinucleotide pairs are also observed (Turunen et al., 2013).

A search of the donor sites of all catalogued PEs and their 5' spliced exons discovered a single example of a UTATCCT donor site motif, at the donor site of *LHCGR-6-1b*. Although this donor site scores low as a U2 splice site (MaxEnt = 0.48), there was no type-U12 CCTTUAY branch-point motif near the acceptor site of the 3' exon 7, and the canonical 5' exon 6 did not have a type-U12 donor site. This indicates that the termini of *LHCGR* intron 6 have evolved to be removed via the predominant mode of U2 splicing. It therefore appears that this PE is spliced via the U2-spliceosome and not the U12. Therefore, we concluded that no U12-spliced PEs are reported in this dataset, although we did note that *STK11-1-1* occurs in a U2 intron that is 5' adjacent to a known U12 intron (Hastings et al., 2005).

While a type-U12 PE may yet be reported, it is unsurprising that none have been discovered thus far. The great majority of reported PEs have been observed as singletons that are directly spliced to canonical upstream and downstream exons in the mature transcript. This means that each of the two PE splicing reactions involves one neighbouring canonical splice motif that has evolved for optimal interaction with a particular spliceosome. From this we can infer that the mode of a PE's splicing will largely be determined by the splicing mode of its encompassing intron, i.e., a U2-spliced PE cannot arise within a U12 intron or vice versa. A similar hypothesis was suggested by Qu et al. (2017) in their analysis of U12 splice mutations. Because only 0.37% of human introns are type-U12 (Olthof et al., 2019) the genomic range within which a U12 PE could plausibly arise is vanishingly small. However, exceptions may occur if a mutation that prevents proper recognition of the splice motifs of a U12 intron results in cryptic U2 splicing. Madan et al. (2015) observed such cryptic U2 splicing of a U12 intron of *WDR41*, though this was the result of knockdown of the splice factor ZRSR2, rather than mutations in *WDR41* itself.

### 5.2 Pseudoexons With Non-AG Acceptor Sites Occur Rarely but Unpredictably

We catalogued four examples of PE variants with non-AG acceptor site dinucleotides. Three of these (*RBI-14-1b*, *RBI-14-1c*, and *RBI-14-1d*) arose from the 5' junction of a single LINE-1 retrotransposon insertion in *RBI* intron 14 (Rodríguez-Martín et al., 2016). These three variants share a common U2-type donor site, but each have unique non-canonical acceptor sites that were confirmed through Sanger sequencing. This LINE-

1 insertion also induced an additional PE variant with a canonical acceptor site (*RBI-14-1a*).

The fourth non-canonical PE (*NFI-39-1a*) was observed in *NFI* as the result of a donor-site-creating SNV. Like the *RBI* PEs, *NFI-39-1a* bears a canonical donor site and a non-canonical acceptor site and shares its donor site with a wholly canonical variant, *NFI-39-1b*.

A report by Parada et al. (2014), examined common features of 184 non-canonical splice sites, and the authors observed therein that the terminal dinucleotides of most non-canonical splice sites differ from the canonical AG or GY pairs by only a single nucleotide. This holds true for the *RBI* non-canonical PEs, which have CG, AT and AT respectively as their acceptor-site terminal dinucleotides, and for *NFI-39-1a*, which has a TG dinucleotide. We speculate that this one-nucleotide rule is observed because varying only a single nucleotide minimises the amount of resistance that must be overcome to “persuade” the spliceosome to cleave at a non-AG/GY dinucleotide.

Unfortunately, there are few other established hallmarks for human non-canonical exons that these PEs can be compared against. Burset et al. (2000) suggested that non-canonical splice sites may parasitically exploit the presence of nearby canonical splice motifs to recruit the spliceosome, an hypothesis supported by the alternative canonical acceptor sites observed in *RBI-14-1a* and *NFI-39-1b*. However, even if this “parasite” model accounts for spliceosome recruitment, it still begs the question of why the non-canonical splice sites are used at all when workable canonical sites are available. Similarly, although Parada et al. (2014) detected a higher density of ESEs and intronic splice enhancers around non-canonical sites, it is not valid to apply their statistical analysis to just four additional sites. Deriving a complete explanation for why these two mutations in *RBI* and *NFI* created PEs with non-canonical splice sites, when so many other similar mutations in these and other genes did not, therefore remains as a challenge for future researchers.

### 5.3 Terminal Pseudoexons are Both Rare and Difficult to Detect Without Third-Generation Sequencing Technologies

We catalogued two examples of terminal pseudoexons (tPEs), each arising from unique mutations in *ARHGEF9* (Figure 3A) and *F8* (Figure 3B). Although it is difficult to generalise from just two observations, there are obvious similarities between these cases that are worth noting. Both the *ARHGEF9* and *F8* genes are carried on the *q* arm of the X-chromosome, albeit at opposite ends (Figure 3C), and in both cases the instigating mutations entail large sequence rearrangements that moved the canonical 3' end of the gene out of splicing range of the upstream exons. In the case of *ARHGEF9-6-2*, this mutation is a balanced crossover with chromosome 18, while the *F8* gene of the second patient bears a 3.8 Mb insertion of chromosome X intergenic sequence. Notably, the region inserted into *F8* in *F8-25-1* encompasses *ARHGEF9* along with 11 other protein-coding genes (not shown). Interestingly, although the *ARHGEF9* mutation was originally described as creating two tPEs—one 5' of the breakpoint in the normal intron 6 sequence, and one in the translocated chromosome 18 sequence—we found that the first of these

terminal exons shares its polyadenylation site with *ARHGEF-IT1*, a noncoding and largely uncharacterised two-exon transcript nested within *ARHGEF* intron 6. Because it shares sequence with a canonical exon, this mutant terminal exon therefore does not meet the criteria for classification as a tPE.

The fact that tPEs are so rare in comparison to internal PEs is surprising when one considers that the defining hallmarks of a last exon are comparably well-defined to those of an internal exon, and therefore they should be expected to arise from random mutation at roughly the same frequency. The requirement for a functional acceptor site is similar in both exon types, and the requirements for polyadenylation site definition (Kaida, 2016) do not appear very much stricter than those for donor site definition. Furthermore, last exons usually contain a stop codon, a requirement that most PEs meet by default, and last exons also exhibit a much broader range of sizes than internal exons (Movassat et al., 2019).

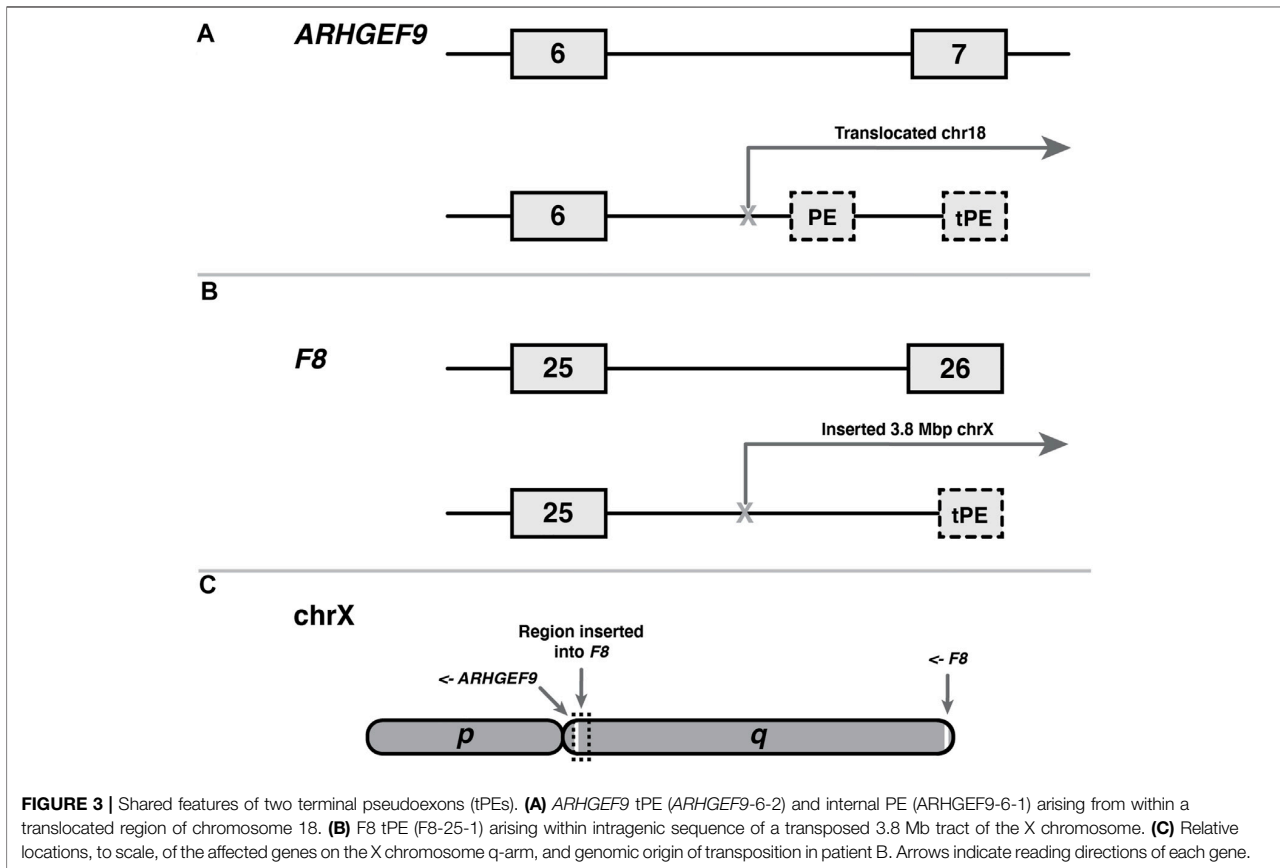
We suggest that there are at least three compounding causes for the low discovery rate of tPEs. The first is that the laboratory techniques required to confirm a tPE make them considerably more difficult to discover than internal PEs. Many internal PEs were detected serendipitously when researchers noticed unusually large products from their RT-PCRs of flanking canonical exons, but this method of discovery is only possible if the RT-PCR primer target sites are present on both sides of the mRNA insertion, and this is not the case for tPEs as they are not spliced to any 3' exon. Any RT-PCR of the canonical exons flanking a tPCR would produce either low abundance products of the expected size (if some level of normal splicing is still present) or no products at all. Even if the researcher eventually discovers the acceptor splice site of the tPE, their subsequent failure to detect an active donor site may lead them to conclude that the effect of the mutation is partial intron inclusion and arrest of splicing.

A third possible contributing factor is that terminal exon definition may be a stricter process than it appears. At the very least, the similarity of the two mutations described here suggests that the absence of competition from downstream canonical exons is a contributing factor, which is something that can only occur after large-scale sequence rearrangements such as these. Conversely, the effect of an *EPCAM* terminal exon deletion mutation (Supplementary Figure S2B) was to induce a fusion transcript with *MSH2* but no novel polyadenylation site, as in this case a latent intergenic pseudoexon combined with the chromosomal proximity of *MSH2* provided viable splicing partners.

Regardless of the true frequency of tPEs, it is worth noting that the aforementioned barriers to their detection do not apply to third generation sequencing technologies like Nanopore, which are largely agnostic in their detection of polyadenylated transcripts. As the uptake of Nanopore and other third-generation RNA sequencing technologies continues to increase, there may be a corresponding increase in the discovery rate of tPEs.

## 6 Conclusion

Pathogenic pseudoexons primarily arise from mutations that directly enhance their donor or acceptor site motifs. However, other types of instigating mutation are also observed less



frequently, but with consistent features, many of which are characterised for the first time in this report. In rare cases, the splicing pathology of a PE was highly idiosyncratic and could not be properly categorised due to a lack of similar supporting examples. These findings advance our understanding of how mutations give rise to pathogenic pseudoexons, but also highlight that our understanding is still far from complete.

We also discovered seven examples of pseudoexons that coincide with recently confirmed recursive splice sites, conclusively demonstrating that functional exon-like intron elements can be converted to pseudoexons when favourable mutations arise nearby. Although it remains to be determined how many pseudoexons arise in this way, we found that 15.7% of pseudoexons showed evidence of splicing at one or both of their splice sites in cells from healthy donors, a figure that is likely to increase further as the fidelity and quantity of RNAseq data continues to improve.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

NK conceived of the project. All authors contributed to project design and developed its rationale. NK and SF defined scope and criteria for data inclusion and exclusion and decided methods of analysis. NK performed research, catalogue curation and data analysis and wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

## FUNDING

Funding was provided by the Australian Commonwealth Government Research Training Program Scholarship.

## ACKNOWLEDGMENTS

The authors would like to thank the Australian Commonwealth Government Research Training Program and Murdoch University for the provision of the PhD stipend for NK, and funding and computing resources, and their colleagues at the Centre for Molecular Medicine and Innovative Therapeutics for advice and constructive criticism.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.806946/full#supplementary-material>

**Supplementary Figure 1** | Pseudoxons instigated by weakened definition of nearby canonical exons. Pseudoxons are shown as medium grey dash-line boxes, with flanking canonical exons in light grey and intervening introns as solid black lines. Sizes are not to scale. Dash-lines elsewhere indicate altered splicing of canonical exons. Roman numerals within pseudoxons (I–IV) indicate identical sequence inclusions (see also **Supplementary Figure S2**). Selected relevant features for each mutation are indicated where appropriate, with full details and references provided in **Supplementary Table S1**. Numbers above exon ends indicate acceptor and donor Maximum Entropy splice site scores, with “NF” indicating a site that is non-functional in the reference (NF → score) or mutant (score → NF) allele due to lacking an essential AG-GY dinucleotide. Changes to exon HExoSplice scores are shown above the affected exons as “ΔHx,” other motif changes are indicated with vertical lines and labelled as “± name-of-motif.” Vertical brackets indicate common features of enclosed splice events. Note that GAA-1-1 is shown twice due to the splicing effects of its mutation fitting both the (A,B) sub-categories.

**Supplementary Figure 2** | Pseudoxons instigated by novel or unknown mechanisms. Pseudoxons are shown as medium grey dash-line boxes, with

flanking canonical exons in light grey and intervening introns as solid black lines. Sizes are not to scale. Dash-lines elsewhere indicate altered splicing of canonical exons. Roman numeral “IV” indicates an identical sequence inclusion to one shown in **Supplementary Figure S1**. Selected relevant features for each mutation are indicated where appropriate, with full details and references provided in **Supplementary Table S1**. Numbers above exon ends indicate acceptor and donor Maximum Entropy splice site scores, with “NF” indicating a site that is non-functional in the reference (NF → score) or mutant (score → NF) allele due to lacking an essential AG-GY dinucleotide. “+D” indicates one or more gain-of-donor-motif mutations. Other motif changes are indicated with vertical lines and labelled as “± name-of-motif.” Vertical brackets indicate common features of enclosed splice events. \*Variant-induced increase in splicing of *FBXO38-9-1* was low (from 8 to 13% of *FBXO38* transcripts) but statistically significant (Saferali et al., 2019).

**Supplementary Figure 3** | A G>A substitution within *RPGR* intron 9 creates a binding motif for KHDRSB1 (Sam68), potentially altering the splicing of an adjacent pseudoxon (*RPGR-9-1*) that originates within an *AluSx* element. Image captured from *UCSC Genome Browser*, GRCh38/hg38. Location of TTAAA motif indicated with a thick vertical red line, and additional annotations are written in red text. Scale at top shows chromosome X coordinates. “GENCODE V36” track shows aligned transcripts, with gene names on the left, solid bars indicating exons and horizontal lines indicating introns. Arrows (< and >) on introns indicate transcript reading direction relative to chromosome + strand. “Repeating Elements” track shows alignments of SINE or LINE elements as solid rectangles, with darker shading indicating higher homology to the reference sequence for that element.

## REFERENCES

- Agrawal, A., Hamvas, A., Cole, F. S., Wambach, J. A., Wegner, D., Coghill, C., et al. (2012). An Intronic ABCA3 Mutation that Is Responsible for Respiratory Disease. *Pediatr. Res.* 71 (6), 633–637. doi:10.1038/pr.2012.21
- Ajeawung, N. F., Nguyen, T. T. M., Lu, L., Kucharski, T. J., Rousseau, J., Molidperee, S., et al. (2019). Mutations in ANAPC1, Encoding a Scaffold Subunit of the Anaphase-Promoting Complex, Cause Rothmund-Thomson Syndrome Type 1. *Am. J. Hum. Genet.* 105 (3), 625–630. doi:10.1016/j.ajhg.2019.06.011
- Akawi, N., Al-Gazali, L., and Ali, B. (2012). Clinical and Molecular Analysis of UAE Fibrochondrogenesis Patients Expands the Phenotype and Reveals Two COL11A1 Homozygous Null Mutations. *Clin. Genet.* 82 (2), 147–156. doi:10.1111/j.1399-0004.2011.01734.x
- Albert, S., Garanto, A., Sangermano, R., Khan, M., Bax, N. M., Hoyng, C. B., et al. (2018). Identification and Rescue of Splice Defects Caused by Two Neighboring Deep-Intronic ABCA4 Mutations Underlying Stargardt Disease. *Am. J. Hum. Genet.* 102 (4), 517–527. doi:10.1016/j.ajhg.2018.02.008
- Alexieva, D., Long, Y., Sarkar, R., Dhayan, H., Bruet, E., Winston, R., et al. (2020). *Background Splicing and Genetic Disease*. Research Square [Preprint]. doi:10.21203/rs.3.rs-92665/v1
- Alves, S., Mangas, M., Prata, M. J., Ribeiro, G., Lopes, L., Ribeiro, H., et al. (2006). Molecular Characterization of Portuguese Patients with Mucopolysaccharidosis Type II Shows Evidence that the IDS Gene Is Prone to Splicing Mutations. *J. Inher. Metab. Dis.* 29 (6), 743–754. doi:10.1007/s10545-006-0403-z
- Aminoff, M., Carter, J. E., Chadwick, R. B., Johnson, C., Gräsbeck, R., Abdelaal, M. A., et al. (1999). Mutations in CUBN, Encoding the Intrinsic Factor-Vitamin B12 Receptor, Cubilin, Cause Hereditary Megaloblastic Anaemia 1. *Nat. Genet.* 21 (3), 309–313. doi:10.1038/6831
- Anczuków, O., Buisson, M., Léoné, M., Coutanson, C., Lasset, C., Calender, A., et al. (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
- Abramowicz, A., and Gos, M. (2018). Splicing Mutations in Human Genetic Disorders: Examples, Detection, and Confirmation. *J. Appl. Genet.* 59 (3), 253–268. doi:10.1007/s13353-018-0444-7
- Ars, E., Kruyer, H., Morell, M., Pros, E., Serra, E., Ravella, A., et al. (2003). Recurrent Mutations in the NF1 Gene Are Common Among Neurofibromatosis Type 1 Patients. *J. Med. Genet.* 40 (6), e82. doi:10.1136/jmg.40.6.e82
- Ars, E., Serra, E., de la Luna, S., Estivill, X., and Lázaro, C. (2000). Cold Shock Induces the Insertion of a Cryptic Exon in the Neurofibromatosis Type 1 (NF1) mRNA. *Nucleic Acids Res.* 28 (6), 1307–1312. doi:10.1093/nar/28.6.1307
- Attanasio, C., David, A., and Neerman-Arbez, M. (2003). Outcome of Donor Splice Site Mutations Accounting for Congenital Afibrinogenemia Reflects Order of Intron Removal in the Fibrinogen Alpha Gene (FGA). *Blood* 101 (5), 1851–1856. doi:10.1182/blood-2002-03-0853
- Bach, J. E., Müller, C. R., and Rost, S. (2016). Mini-gene Assays Confirm the Splicing Effect of Deep Intronic Variants in the Factor VIII Gene. *Thromb. Haemost.* 115 (1), 222–224. doi:10.1160/TH15-05-0399
- Backers, L., Parton, B., De Bruyne, M., Tavernier, S. J., Van Den Bogaert, K., Lambrecht, B. N., et al. (2021). Missing Heritability in Bloom Syndrome: First Report of a Deep Intronic Variant Leading to Pseudo-exon Activation in the BLM Gene. *Clin. Genet.* 99 (2), 292–297. doi:10.1111/cge.13859
- Bagnall, R. D., Ingles, J., Dinger, M. E., Cowley, M. J., Ross, S. B., Minoche, A. E., et al. (2018). Whole Genome Sequencing Improves Outcomes of Genetic Testing in Patients with Hypertrophic Cardiomyopathy. *J. Am. Coll. Cardiol.* 72 (4), 419–429. doi:10.1016/j.jacc.2018.04.078
- Bagnall, R. D., Waseem, N. H., Green, P. M., Colvin, B., Lee, C., and Giannelli, F. (1999). Creation of a Novel Donor Splice Site in Intron 1 of the Factor VIII Gene Leads to Activation of a 191 Bp Cryptic Exon in Two Haemophilia A Patients. *Br. J. Haematol.* 107 (4), 766–771. doi:10.1046/j.1365-2141.1999.01767.x
- Bampton, A., Gatt, A., Humphrey, J., Cappelli, S., Bhattacharya, D., Foti, S., et al. (2021). HnRNP K Mislocalisation Is a Novel Protein Pathology of Frontotemporal Lobar Degeneration and Ageing and Leads to Cryptic Splicing. *Acta Neuropathol.* 142 (4), 609–627. doi:10.1007/s00401-021-02340-0
- Baressi, R., Di Blasi, C., Negri, T., Brugnani, R., Vitali, A., Felisari, G., et al. (2000). Disruption of Heart Sarcoglycan Complex and Severe Cardiomyopathy Caused by Beta Sarcoglycan Mutations. *J. Med. Genet.* 37 (2), 102–107. doi:10.1136/jmg.37.2.102
- Baskin, B., Gibson, W. T., and Ray, P. N. (2011). Duchenne Muscular Dystrophy Caused by a Complex Rearrangement between Intron 43 of the DMD Gene and Chromosome 4. *Neuromuscul. Disord.* 21 (3), 178–182. doi:10.1016/j.nmd.2010.11.008
- Bauwens, M., Garanto, A., Sangermano, R., Naessens, S., Weisschuh, N., De Zaeytijd, J., et al. (2019). ABCA4-associated Disease as a Model for Missing Heritability in Autosomal Recessive Disorders: Novel Noncoding Splice, Cis-Regulatory, Structural, and Recurrent Hypomorphic Variants. *Genet. Med.* 21 (8), 1761–1771. doi:10.1038/s41436-018-0420-y
- Baux, D., Vaché, C., Blanchet, C., Willems, M., Baudoin, C., Moclyn, M., et al. (2017). Combined Genetic Approaches Yield a 48% Diagnostic Rate in a Large Cohort of French Hearing-Impaired Patients. *Sci. Rep.* 7 (1), 16783. doi:10.1038/s41598-017-16846-9
- Bergougnot, A., Délétang, K., Pommier, A., Varilh, J., Houriez, F., Altieri, J. P., et al. (2019). Functional Characterization and Phenotypic Spectrum of Three



- Recurrent Disease-Causing Deep Intronic Variants of the CFTR Gene. *J. Cystic Fibrosis* 18 (4), 468–475. doi:10.1016/j.jcf.2018.10.012
- Bergsma, A. J., in 't Groen, S. L., Verheijen, F. W., van der Ploeg, A. T., and Pijnappel, W. P. (2016). In 't Groen, S.L., Verheijen, F.W., van der Ploeg, A.T., and Pijnappel, W. From Cryptic Toward Canonical Pre-mRNA Splicing in Pompe Disease: a Pipeline for the Development of Antisense Oligonucleotides. *Mol. Ther. - Nucleic Acids* 5 (9), e361. doi:10.1038/mtna.2016.75
- Bérourd, C., Carrié, A., Beldjord, C., Deburgrave, N., Llése, S., Carelle, N., et al. (2004). Dystrophinopathy Caused by Mid-intronic Substitutions Activating Cryptic Exons in the DMD Gene. *Neuromuscul. Disord.* 14 (1), 10–18. doi:10.1016/s0960-8966(03)00169-x
- Bhola, Z., Smith, M. J., Byers, H. J., Miles, E. K., Evans, D. G., and Newman, W. G. (2014). Intronic Splicing Mutations in PTCH1 Cause Gorlin Syndrome. *Fam. Cancer* 13 (3), 477–480. doi:10.1007/s10689-014-9712-9
- Blázquez, L., Azpitarte, M., Sáenz, A., Goicoechea, M., Otaegui, D., Ferrer, X., et al. (2008). Characterization of Novel CAPN3 Isoforms in white Blood Cells: an Alternative Approach for Limb-Girdle Muscular Dystrophy 2A Diagnosis. *Neurogenetics* 9 (3), 173–182. doi:10.1007/s10048-008-0129-1
- Blázquez, L., Emmett, W., Faraway, R., Pineda, J. M. B., Bajew, S., Gohr, A., et al. (2018). Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing. *Mol. Cell* 72 (3), 496–509. doi:10.1016/j.molcel.2018.09.033
- Bodle, E. E., Zhu, W., Velez-Bartolomei, F., Tesi-Rocha, A., Liu, P., and Bernstein, J. A. (2021). Combined Genome Sequencing and RNA Analysis Reveals and Characterizes a Deep Intronic Variant in IGHMBP2 in a Patient with Spinal Muscular Atrophy with Respiratory Distress Type 1. *Pediatr. Neurol.* 114, 16–20. doi:10.1016/j.pediatrneurol.2020.09.011
- Boisson, B., Honda, Y., Ajiro, M., Bustamante, J., Bendavid, M., Gennery, A. R., et al. (2019). Rescue of Recurrent Deep Intronic Mutation Underlying Cell Type-dependent Quantitative NEMO Deficiency. *J. Clin. Invest.* 129 (2), 583–597. doi:10.1172/JCI124011
- Bolduc, V., Foley, A. R., Solomon-Degefa, H., Sarathy, A., Donkervoort, S., Hu, Y., et al. (2019). A Recurrent COL6A1 Pseudoexon Insertion Causes Muscular Dystrophy and Is Effectively Targeted by Splice-Correction Therapies. *JCI Insight* 4 (6), 1. doi:10.1172/jci.insight.124403
- Bonifert, T., Karle, K. N., Tonagel, F., Batra, M., Wilhelm, C., Theurer, Y., et al. (2014). Pure and Syndromic Optic Atrophy Explained by Deep Intronic OPA1 Mutations and an Intralocus Modifier. *Brain* 137 (Pt 8), 2164–2177. doi:10.1093/brain/awu165
- Bonini, J., Varilh, J., Raynal, C., Thèze, C., Beyne, E., Audrezet, M.-P., et al. (2015). Small-scale High-Throughput Sequencing-Based Identification of New Therapeutic Tools in Cystic Fibrosis. *Genet. Med.* 17 (10), 796–806. doi:10.1038/gim.2014.194
- Bovolenta, M., Neri, M., Fini, S., Fabris, M., TrabANELLI, C., Venturoli, A., et al. (2008). A Novel Custom High Density-Comparative Genomic Hybridization Array Detects Common Rearrangements as Well as Deep Intronic Mutations in Dystrophinopathies. *BMC Genomics* 9, 572. doi:10.1186/1471-2164-9-572
- Brasil, S., Viecilli, H. M., Meili, D., Rassi, A., Desviat, L. R., Pérez, B., et al. (2011). Pseudoexon Exclusion by Antisense Therapy in 6-Pyruvoyl-Tetrahydropterin Synthase Deficiency. *Hum. Mutat.* 32 (9), 1019–1027. doi:10.1002/humu.21529
- Braun, T. A., Mullins, R. F., Wagner, A. H., Andorf, J. L., Johnston, R. M., Bakall, B. B., et al. (2013). Non-exomic and Synonymous Variants in ABCA4 Are an Important Cause of Stargardt Disease. *Hum. Mol. Genet.* 22 (25), 5136–5145. doi:10.1093/hmg/ddt367
- Brinckmann, A., Mischung, C., Bässmann, I., Kühnisch, J., Schuelke, M., Tinschert, S., et al. (2007). Detection of novelNF1 Mutations and Rapid Mutation Prescreening with Pyrosequencing. *Electrophoresis* 28 (23), 4295–4301. doi:10.1002/elps.200700118
- Broeders, M., Smits, K., Goynuk, B., Oussoren, E., van den Hout, H. J. M. P., Bergsma, A. J., et al. (2020). A Generic Assay to Detect Aberrant ARSB Splicing and mRNA Degradation for the Molecular Diagnosis of MPS VI. *Mol. Ther. - Methods Clin. Dev.* 19, 174–185. doi:10.1016/j.omtm.2020.09.004
- Bryen, S. J., Oates, E. C., Evesson, F. J., Lu, J. K., Waddell, L. B., Joshi, H., et al. (2021). Pathogenic Deep Intronic MTM1 Variant Activates a Pseudo-exon Encoding a Nonsense Codon Resulting in Severe X-Linked Myotubular Myopathy. *Eur. J. Hum. Genet.* 29 (1), 61–66. doi:10.1038/s41431-020-00715-7
- Burset, M., Seledtsov, I. A., and Solovyev, V. V. (2000). Analysis of Canonical and Non-canonical Splice Sites in Mammalian Genomes. *Nucleic Acids Res.* 28 (21), 4364–4375. doi:10.1093/nar/28.21.4364
- Bustamante, J., Aksu, G., Vogt, G., de Beaucoudrey, L., Genel, F., Chappier, A., et al. (2007). BCG-osis and Tuberculosis in a Child with Chronic Granulomatous Disease. *J. Allergy Clin. Immunol.* 120 (1), 32–38. doi:10.1016/j.jaci.2007.04.034
- Caciotti, A., Tonin, R., Mort, M., Cooper, D. N., Gasperini, S., Rigoldi, M., et al. (2018). Mis-splicing of the GALNS Gene Resulting from Deep Intronic Mutations as a Cause of Morquio a Disease. *BMC Med. Genet.* 19 (1), 183. doi:10.1186/s12881-018-0694-6
- Cagliani, R., Sironi, M., Ciafaloni, E., Bardoni, A., Fortunato, F., Prella, A., et al. (2004). An Intragenic Deletion/Inversion Event in the DMD Gene Determines a Novel Exon Creation and Results in a BMD Phenotype. *Hum. Genet.* 115 (1), 13–18. doi:10.1007/s00439-004-1118-6
- Campeau, E., Dupuis, L., Leclerc, D., and Gravel, R. A. (1999). Detection of a Normally Rare Transcript in Propionic Acidemia Patients with mRNA Destabilizing Mutations in the PCCA Gene. *Hum. Mol. Genet.* 8 (1), 107–113. doi:10.1093/hmg/8.1.107
- Canson, D., Glubb, D., and Spurdle, A. B. (2020). Variant Effect on Splicing Regulatory Elements, Branchpoint Usage, and Pseudoexonization: Strategies to Enhance Bioinformatic Prediction Using Hereditary Cancer Genes as Exemplars. *Hum. Mutat.* 41 (10), 1705–1721. doi:10.1002/humu.24074
- Caparrós-Martín, J. A., De Luca, A., Cartault, F., Aglan, M., Temtamy, S., Otaify, G. A., et al. (2015). Specific Variants in WDR35 Cause a Distinctive Form of Ellis-van Creveld Syndrome by Disrupting the Recruitment of the EvC Complex and SMO into the Cilium. *Hum. Mol. Genet.* 24 (14), 4126–4137. doi:10.1093/hmg/ddv152
- Carmody, D., Park, S.-Y., Ye, H., Perrone, M. E., Alkorta-Aranburu, G., Highland, H. M., et al. (2015). Continued Lessons from theIN5Gene: an Intronic Mutation Causing Diabetes through a Novel Mechanism. *J. Med. Genet.* 52 (9), 612–616. doi:10.1136/jmedgenet-2015-103220
- Carvill, G. L., and Mefford, H. C. (2020). Poison Exons in Neurodevelopment and Disease. *Curr Opin Genet Dev* 65, 98–102. doi:10.1016/j.gde.2020.05.030
- Cassini, T. A., Duncan, L., Rives, L. C., Newman, J. H., Phillips, J. A., Koziura, M. E., et al. (2019). Whole Genome Sequencing Reveals Novel IGHMBP2 Variant Leading to Unique Cryptic Splice-Site and Charcot-Marie-Tooth Phenotype with Early Onset Symptoms. *Mol. Genet. Genomic Med.* 7 (6), e00676. doi:10.1002/mgg3.676
- Castaman, G., Giacomelli, S. H., Mancuso, M. E., D'Andrea, G., Santacroce, R., Sanna, S., et al. (2011). Deep Intronic Variations May Cause Mild Hemophilia A. *J. Thromb. Haemost.* 9 (8), 1541–1548. doi:10.1111/j.1538-7836.2011.04408.x
- Castellanos, E., Rosas, I., Rosas, I., Solanes, A., Bielsa, I., Lázaro, C., et al. (2013). *In Vitro* antisense Therapeutics for a Deep Intronic Mutation Causing Neurofibromatosis Type 2. *Eur. J. Hum. Genet.* 21 (7), 769–773. doi:10.1038/ejhg.2012.261
- Castoldi, E., Duckers, C., Radu, C., Spiezia, L., Rossetto, V., Tagariello, G., et al. (2011). Homozygous F5 Deep-Intronic Splicing Mutation Resulting in Severe Factor V Deficiency and Undetectable Thrombin Generation in Platelet-Rich Plasma. *J. Thromb. Haemost.* 9 (5), 959–968. doi:10.1111/j.1538-7836.2011.04237.x
- Catania, A., Ardisson, A., Verrigni, D., Legati, A., Newman, J. H., Phillips, J. A., Koziura, M. E., et al. (2019). Whole Genome Sequencing Reveals Novel IGHMBP2 Variant Leading to Unique Cryptic Splice-Site and Charcot-Marie-Tooth Phenotype with Early Onset Symptoms. *Mol. Genet. Genomic Med.* 7 (6), e00676. doi:10.1002/mgg3.676
- Cavaliere, S., Pozzi, E., Gatti, R. A., and Brusco, A. (2013). Deep-intronic ATM Mutation Detected by Genomic Resequencing and Corrected *In Vitro* by Antisense Morpholino Oligonucleotide (AMO). *Eur. J. Hum. Genet.* 21 (7), 774–778. doi:10.1038/ejhg.2012.266
- Cavestro, C., Panteghini, C., Reale, C., Nasca, A., Fenu, S., Salsano, E., et al. (2021). Novel Deep Intronic Mutation in PLA2G6 Causing Early-Onset Parkinson's Disease with Brain Iron Accumulation through Pseudo-exon Activation. *Neurogenetics* 22 (4), 347–351. doi:10.1007/s10048-021-00667-0
- Chang, C. Y., Peng, C. L., Cheng, S. N., Hu, S. H., Wu, T. Y., Lin, S. Y., et al. (2019). Deep Intronic Variant c.5999-277G>A of F8 Gene May Be a Hot Spot Mutation for Mild Hemophilia A Patients without Mutation in Exonic DNA. *Eur. J. Haematol.* 103 (1), 47–55. doi:10.1111/ejh.13242

- Chatron, N., Schluth-Bolard, C., Frétygny, M., Labalme, A., Vilchez, G., Castet, S. M., et al. (2019). Severe Hemophilia A Caused by an Unbalanced Chromosomal Rearrangement Identified Using Nanopore Sequencing. *J. Thromb. Haemost.* 17 (7), 1097–1103. doi:10.1111/jth.14460
- Chen, J., Ma, N., Zhao, X., Li, W., Zhang, Q., Yuan, S., et al. (2019). A Rare Deep Intronic Mutation of PKHD1 Gene, c.8798-459 C > A, Causes Autosomal Recessive Polycystic Kidney Disease by Pseudoexon Activation. *J. Hum. Genet.* 64 (3), 207–214. doi:10.1038/s10038-018-0550-8
- Chen, X., Truong, T.-T. N., Weaver, J., Bove, B. A., Cattie, K., Armstrong, B. A., et al. (2006). Intronic Alterations inBRCA1andBRCA2: Effect on mRNA Splicing Fidelity and Expression. *Hum. Mutat.* 27 (5), 427–435. doi:10.1002/humu.20319
- Cheng, T. C., Orkin, S. H., Antonarakis, S. E., Potter, M. J., Sexton, J. P., Markham, A. F., et al. (1984). Beta-Thalassemia in Chinese: Use of *In Vivo* RNA Analysis and Oligonucleotide Hybridization in Systematic Characterization of Molecular Defects. *Proc. Natl. Acad. Sci.* 81 (9), 2821–2825. doi:10.1073/pnas.81.9.2821
- Chillón, M., Dörk, T., Casals, T., Giménez, J., Fonknechten, N., Will, K., et al. (1995). A Novel Donor Splice Site in Intron 11 of the CFTR Gene, Created by Mutation 1811+1.6kbA-->G, Produces a New Exon: High Frequency in Spanish Cystic Fibrosis Chromosomes and Association with Severe Phenotype. *Am. J. Hum. Genet.* 56 (3), 623–629.
- Chiu, Y.-H., Chang, Y.-C., Chang, Y.-H., Niu, D.-M., Yang, Y.-L., Ye, J., et al. (2012). Mutation Spectrum of and Founder Effects Affecting the PTS Gene in East Asian Populations. *J. Hum. Genet.* 57 (2), 145–152. doi:10.1038/jhg.2011.146
- Chmel, N., Danescu, S., Gruler, A., Kiritsi, D., Bruckner-Tuderman, L., Kreuter, A., et al. (2015). A Deep-Intronic FERMT1 Mutation Causes Kindler Syndrome: An Explanation for Genetically Unsolved Cases. *J. Invest. Dermatol.* 135 (11), 2876–2879. doi:10.1038/jid.2015.227
- Chorev, M., and Carmel, L. (2012). The Function of Introns. *Front. Gene* 3, 55. doi:10.3389/fgene.2012.00055
- Chorin, O., Yachevich, N., Mohamed, K., Moscatelli, L., Pappas, J., Henriksen, K., et al. (2020). Transcriptome Sequencing Identifies a Noncoding, Deep Intronic Variant in CLCN7 Causing Autosomal Recessive Osteopetrosis. *Mol. Genet. Genomic Med.* 8 (10), e1405. doi:10.1002/mgg3.1405
- Christie, P. T., Harding, B., Nesbit, M. A., Whyte, M. P., and Thakker, R. V. (2001). X-linked Hypophosphatemia Attributable to Pseudoexons of the PHEX Gene. *J. Clin. Endocrinol. Metab.* 86 (8), 3840–3844. doi:10.1210/jcem.86.8.7730
- Chung, W. Y., Cho, M. H., Gu, Y.-R., Leem, S.-H., and Cheong, H. I. (2012). Medullary Sponge Kidney Detected in the Pediatric Age. *J. Korean Soc. Pediatr. Nephrol.* 16 (2), 1. doi:10.3339/jkspn.2012.16.2.109
- Clemens, D. J., Tester, D. J., Marty, I., and Ackerman, M. J. (2020). Phenotype-guided Whole Genome Analysis in a Patient with Genetically Elusive Long QT Syndrome Yields a Novel TRDN-Encoded Triadin Pathogenetic Substrate for Triadin Knockout Syndrome and Reveals a Novel Primate-specific Cardiac TRDN Transcript. *Heart Rhythm* 17 (6), 1017–1024. doi:10.1016/j.hrthm.2020.01.012
- Clendenning, M., Buchanan, D. D., Walsh, M. D., Nagler, B., Rosty, C., Thompson, B., et al. (2011). Mutation Deep within an Intron of MSH2 Causes Lynch Syndrome. *Fam. Cancer* 10 (2), 297–301. doi:10.1007/s10689-011-9427-0
- Conboy, J. G. (2021). Unannotated Splicing Regulatory Elements in Deep Intron Space. *Wiley Interdiscip. Rev. RNA* 12 (5), e1656. doi:10.1002/wrna.1656
- Corrigan, A., Arenas, M., Escuredo, E., Fairbanks, L., and Marinaki, A. (2011). HPRT Deficiency: Identification of Twenty-Four Novel Variants Including an Unusual Deep Intronic Mutation. *Nucleosides, Nucleotides and Nucleic Acids* 30 (12), 1260–1265. doi:10.1080/15257770.2011.590172
- Corvelo, A., Hallegger, M., Smith, C. W. J., and Eyras, E. (2010). Genome-wide Association between branch point Properties and Alternative Splicing. *PLoS Comput. Biol.* 6 (11), e1001016. doi:10.1371/journal.pcbi.1001016
- Costantino, L., Claut, L., Paracchini, V., Coviello, D. A., Colombo, C., Porcaro, L., et al. (2010). A Novel Donor Splice Site Characterized by CFTR mRNA Analysis Induces a New Pseudo-exon in CF Patients. *J. Cystic Fibrosis* 9 (6), 411–418. doi:10.1016/j.jcf.2010.08.009
- Cottrell, E., Maharaj, A., Williams, J., Chatterjee, S., Cirillo, G., Miraglia Del Giudice, E., et al. (2021). Growth Hormone Receptor (GHR) 6 $\Omega$  Pseudoexon Activation: a Novel Cause of Severe Growth Hormone Insensitivity. *J. Clin. Endocrinol. Metab.* 107 (1), 401–416. doi:10.1210/clinem/dgab550
- Coutinho, G., Xie, J., Du, L., Brusco, A., Krainer, A. R., and Gatti, R. A. (2005). Functional Significance of a Deep Intronic Mutation in the ATM Gene and Evidence for an Alternative Exon 28a. *Hum. Mutat.* 25 (2), 118–124. doi:10.1002/humu.20170
- Coutinho, M. F., da Silva Santos, L., Lacerda, L., Quental, S., Wibrand, F., Lund, A. M., et al. (2012). Alu-Alu Recombination Underlying the First Large Genomic Deletion in GlcNAc-Phosphotransferase Alpha/Beta (GNPTAB) Gene in a MLII Alpha/Beta Patient. *JIMD Rep.* 4, 117–124. doi:10.1007/8904\_2011\_83
- Covello, G., Ibrahim, G. H., Bacchi, N., Casarosa, S., and Denti, M. A. (2021). *Exon Skipping via Chimeric Antisense U1 snRNAs to Correct Retinitis Pigmentosa GTPase-Regulator (RPGR) Splice Defect*. bioRxiv [preprint]. doi:10.1101/2021.06.26.449721 (Accessed September 09, 2021)
- Cummings, B. B., Marshall, J. L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A. R., et al. (2017). Improving Genetic Diagnosis in Mendelian Disease with Transcriptome Sequencing. *Sci. Transl. Med.* 9 (386), 1. doi:10.1126/scitranslmed.aal5209
- Danilenko, M., Dalglish, C., Pagliarini, V., Naro, C., Ehrmann, I., Feracci, M., et al. (2017). Binding Site Density Enables Paralog-specific Activity of SLM2 and Sam68 Proteins in Neurexin2 AS4 Splicing Control. *Nucleic Acids Res.* 45 (7), 4120–4130. doi:10.1093/nar/gkx1277
- Davis, R. L., Homer, V. M., George, P. M., and Brennan, S. O. (2009). A Deep Intronic Mutation inFGBcreates a Consensus Exonic Splicing Enhancer Motif that Results in Afibrinogenemia Caused by Aberrant mRNA Splicing, Which Can Be Corrected *In Vitro* with Antisense Oligonucleotide Treatment. *Hum. Mutat.* 30 (2), 221–227. doi:10.1002/humu.20839
- de Boer, M., van Leeuwen, K., Geissler, J., Weemaes, C. M., van den Berg, T. K., Kuijpers, T. W., et al. (2014). Primary Immunodeficiency Caused by an Exonized Retroposed Gene Copy Inserted in theCYBBGene. *Hum. Mutat.* 35 (4), 486–496. doi:10.1002/humu.22519
- De Gasperi, R., Gama Sosa, M. A., Sartorato, E. L., Battistini, S., MacFarlane, H., Gusella, J. F., et al. (1996). Molecular Heterogeneity of Late-Onset Forms of Globoid-Cell Leukodystrophy. *Am. J. Hum. Genet.* 59 (6), 1233–1242.
- De Klein, A., Riegman, P. H., Bijlsma, E. K., Heldoorn, A., Muijtjens, M., den Bakker, M. A., et al. (1998). A G-->A Transition Creates a branch point Sequence and Activation of a Cryptic Exon, Resulting in the Hereditary Disorder Neurofibromatosis 2. *Hum. Mol. Genet.* 7 (3), 393–398. doi:10.1093/hmg/7.3.393
- Dear, A., Daly, J., Brennan, S. O., Tuckfield, A., and George, P. M. (2006). An Intronic Mutation within FGB (IVS1+2076 Ag) Is Associated with Afibrinogenemia and Recurrent Transient Ischemic Attacks. *J. Thromb. Haemost.* 4 (2), 471–472. doi:10.1111/j.1538-7836.2006.01722.x
- Deburgrave, N., Daoud, F., Llense, S., Barbot, J. C., Récan, D., Peccate, C., et al. (2007). Protein- and mRNA-Based Phenotype-Genotype Correlations in DMD/BMD with point Mutations and Molecular Basis for BMD with Nonsense and Frameshift Mutations in the DMD Gene. *Hum. Mutat.* 28 (2), 183–195. doi:10.1002/humu.20422
- Dehainault, C., Michaux, D., Pagès-Berhouet, S., Caux-Moncoutier, V., Doz, F., Desjardins, L., et al. (2007). A Deep Intronic Mutation in the RB1 Gene Leads to Intronic Sequence Exonisation. *Eur. J. Hum. Genet.* 15 (4), 473–477. doi:10.1038/sj.ejhg.5201787
- den Dunnen, J. T., Dalglish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., et al. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum. Mutat.* 37 (6), 564–569. doi:10.1002/humu.22981
- den Hollander, A. I., Koenekoop, R. K., Yzer, S., Lopez, I., Arends, M. L., Voesenek, K. E. J., et al. (2006). Mutations in the CEP290 (NPHP6) Gene Are a Frequent Cause of Leber Congenital Amaurosis. *Am. J. Hum. Genet.* 79 (3), 556–561. doi:10.1086/507318
- Dericquebourg, A., Jourdy, Y., Frétygny, M., Lienhart, A., Claeysens, S., Ternisien, C., et al. (2020). Identification of New F8 Deep Intronic Variations in Patients with Haemophilia A. *Haemophilia* 26 (5), 847–854. doi:10.1111/hae.14134
- Dhir, A., and Buratti, E. (2010). Alternative Splicing: Role of Pseudoexons in Human Disease and Potential Therapeutic Strategies. *FEBS J.* 277 (4), 841–855. doi:10.1111/j.1742-4658.2009.07520.x
- Di Scipio, M., Tavares, E., Deshmukh, S., Audo, I., Green-Sanderson, K., Zubak, Y., et al. (2020). Phenotype Driven Analysis of Whole Genome Sequencing Identifies Deep Intronic Variants that Cause Retinal Dystrophies by

- Aberrant Exonization. *Invest. Ophthalmol. Vis. Sci.* 61 (10), 36. doi:10.1167/iops.61.10.36
- Dobkin, C., and Bank, A. (1985). Reversibility of IVS 2 Missplicing in a Mutant Human Beta-Globin Gene. *J. Biol. Chem.* 260 (30), 16332–16337. doi:10.1016/s0021-9258(17)36241-5
- Dobkin, C., Pergolizzi, R. G., Bahre, P., and Bank, A. (1983). Abnormal Splice in a Mutant Human Beta-Globin Gene Not at the Site of a Mutation. *Proc. Natl. Acad. Sci.* 80 (5), 1184–1188. doi:10.1073/pnas.80.5.1184
- Dominov, J. A., Uyan, Ö., McKenna-Yasek, D., Nallamilli, B. R. R., Kergourlay, V., Bartoli, M., et al. (2019). Correction of Pseudoxon Splicing Caused by a Novel Intronic Dysferlin Mutation. *Ann. Clin. Transl. Neurol.* 6 (4), 642–654. doi:10.1002/acn3.738
- Dominov, J. A., Uyan, Ö., Sapp, P. C., McKenna-Yasek, D., Nallamilli, B. R. R., Hegde, M., et al. (2014). A Novel Dysferlin Mutant Pseudoxon Bypassed with Antisense Oligonucleotides. *Ann. Clin. Transl. Neurol.* 1 (9), 703–720. doi:10.1002/acn3.96
- Drewe-Boss, P., Wessels, H.-H., and Ohler, U. (2018). omniCLIP: Probabilistic Identification of Protein-RNA Interactions from CLIP-Seq Data. *Genome Biol.* 19 (1), 183. doi:10.1186/s13059-018-1521-2
- Drexler, H. L., Choquet, K., and Churchman, L. S. (2020). Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Mol. Cell* 77 (5), 985–998. e988. doi:10.1016/j.molcel.2019.11.017
- Dwi Pramono, Z. A., Takeshima, Y., Suroono, A., Ishida, T., and Matsuo, M. (2000). A Novel Cryptic Exon in Intron 2 of the Human Dystrophin Gene Evolved from an Intron by Acquiring Consensus Sequences for Splicing at Different Stages of Anthropoid Evolution. *Biochem. Biophysical Res. Commun.* 267 (1), 321–328. doi:10.1006/bbrc.1999.1962
- Ellingford, J. M., Thomas, H. B., Rowlands, C., Arno, G., Beaman, G., Gomes-Silva, B., et al. (2019). *Functional and In-Silico Interrogation of Rare Genomic Variants Impacting RNA Splicing for the Diagnosis of Genomic Disorders*. BioRxiv. doi:10.1101/781088 (Accessed June 09, 2021)
- Engel, K., Nuoffer, J.-M., Mühlhausen, C., Klaus, V., Largiadèr, C. R., Tsiakas, K., et al. (2008). Analysis of mRNA Transcripts Improves the success Rate of Molecular Genetic Testing in OTC Deficiency. *Mol. Genet. Metab.* 94 (3), 292–297. doi:10.1016/j.ymgme.2008.03.009
- Evans, D. G., Bowers, N., Burkitt-Wright, E., Miles, E., Garg, S., Scott-Kitching, V., et al. (2016). Comprehensive RNA Analysis of the NF1 Gene in Classically Affected NF1 Affected Individuals Meeting NIH Criteria Has High Sensitivity and Mutation Negative Testing Is Reassuring in Isolated Cases with Pigmentary Features Only. *EBioMedicine* 7, 212–220. doi:10.1016/j.ebiom.2016.04.005
- Faà, V., Incani, F., Meloni, A., Corda, D., Masala, M., Baffico, A. M., et al. (2009). Characterization of a Disease-Associated Mutation Affecting a Putative Splicing Regulatory Element in Intron 6b of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Gene. *J. Biol. Chem.* 284 (44), 30024–30031. doi:10.1074/jbc.M109.032623
- Fadaie, Z., Khan, M., Del Pozo-Valero, M., Cornelis, S. S., Ayuso, C., Cremers, F. P. M., et al. (2019). Identification of Splice Defects Due to Noncanonical Splice Site or Deep-intronic Variants in ABCA4. *Hum. Mutat.* 40 (12), 2365–2376. doi:10.1002/humu.23890
- Fadaie, Z., Whelan, L., Ben-Yosef, T., Dockery, A., Corradi, Z., Gilissen, C., et al. (2021). Whole genome sequencing and in vitro splice assays reveal genetic causes for inherited retinal diseases. *NPJ Genom Med* 6 (1), 97. doi:10.1038/s41525-021-00261-1
- Falzarano, M. S., Grilli, A., Zia, S., Fang, M., Rossi, R., Gualandi, F., et al. (2022). RNA-seq in DMD Urinary Stem Cells Recognized Muscle-Related Transcription Signatures and Addressed the Identification of Atypical Mutations by Whole-Genome Sequencing. *Hum. Genet. Genomics Adv.* 3 (1), 100054. doi:10.1016/j.xhgg.2021.100054
- Fanin, M., Torella, A., Savarese, M., Nigro, V., and Angelini, C. (2015). GYG1 gene Mutations in a Family with Polyglucosan Body Myopathy. *Neurol. Genet.* 1 (3), e21. doi:10.1212/NXG.0000000000000021
- Farlow, A., Dolezal, M., Hua, L., and Schlötterer, C. (2012). The Genomic Signature of Splicing-Coupled Selection Differs between Long and Short Introns. *Mol. Biol. Evol.* 29 (1), 21–24. doi:10.1093/molbev/msr201
- Ferlini, A., Galié, N., Merlini, L., Sewry, C., Branzi, A., and Muntoni, F. (1998). A Novel Alu-like Element Rearranged in the Dystrophin Gene Causes a Splicing Mutation in a Family with X-Linked Dilated Cardiomyopathy. *Am. J. Hum. Genet.* 63 (2), 436–446. doi:10.1086/301952
- Ferlini, A., and Muntoni, F. (1998). The 5' Region of Intron 11 of the Dystrophin Gene Contains Target Sequences for Mobile Elements and Three Overlapping ORFs. *Biochem. Biophysical Res. Commun.* 242 (2), 401–406. doi:10.1006/bbrc.1997.7976
- Fernández-Rodríguez, J., Castellsagué, J., Benito, L., Benavente, Y., Capellá, G., Blanco, I., et al. (2011). A Mild Neurofibromatosis Type 1 Phenotype Produced by the Combination of the Benign Nature of a Leaky NF1-Splice Mutation and the Presence of a Complex Mosaicism. *Hum. Mutat.* 32 (7), 705–709. doi:10.1002/humu.21500
- Ferraresi, P., Balestra, D., Guittard, C., Buthiau, D., Pan-Petesh, B., Maestri, I., et al. (2020). Next-generation Sequencing and Recombinant Expression Characterized Aberrant Splicing Mechanisms and provided Correction Strategies in Factor VII Deficiency. *Haematologica* 105 (3), 829–837. doi:10.3324/haematol.2019.217539
- Filoni, C., Caciotti, A., Carraresi, L., Donati, M. A., Mignani, R., Parini, R., et al. (2008). Unbalanced GLA mRNAs Ratio Quantified by Real-Time PCR in Fabry Patients' Fibroblasts Results in Fabry Disease. *Eur. J. Hum. Genet.* 16 (11), 1311–1317. doi:10.1038/ejhg.2008.109
- Fisher, C. W., Fisher, C. R., Chuang, J. L., Lau, K. S., Chuang, D. T., and Cox, R. P. (1993). Occurrence of a 2-bp (AT) Deletion Allele and a Nonsense (G-To-T) Mutant Allele at the E2 (DBT) Locus of Six Patients with maple Syrup Urine Disease: Multiple-Exon Skipping as a Secondary Effect of the Mutations. *Am. J. Hum. Genet.* 52 (2), 414–424.
- Fitzgerald, J., Feist, C., Dietz, P., Moore, S., and Basel, D. (2020). A Deep Intronic Variant Activates a Pseudoxon in the MTM1 Gene in a Family with X-Linked Myotubular Myopathy. *Mol. Syndromol* 11 (5–6), 264–270. doi:10.1159/000510286
- Flanagan, S. E., Xie, W., Caswell, R., Damhuis, A., Vianey-Saban, C., Akcay, T., et al. (2013). Next-generation Sequencing Reveals Deep Intronic Cryptic ABCC8 and HADH Splicing Founder Mutations Causing Hyperinsulinism by Pseudoxon Activation. *Am. J. Hum. Genet.* 92 (1), 131–136. doi:10.1016/j.ajhg.2012.11.017
- Friedrich, K., Lee, L., Leistritz, D. F., Nürnberg, G., Saha, B., Hisama, F. M., et al. (2010). WRN Mutations in Werner Syndrome Patients: Genomic Rearrangements, Unusual Intronic Mutations and Ethnic-specific Alterations. *Hum. Genet.* 128 (1), 103–111. doi:10.1007/s00439-010-0832-5
- Frio, T. R., McGee, T. L., Wade, N. M., Iseli, C., Beckmann, J. S., Berson, E. L., et al. (2009). A Single-Base Substitution within an Intronic Repetitive Element Causes Dominant Retinitis Pigmentosa with Reduced Penetrance. *Hum. Mutat.* 30 (9), 1340–1347. doi:10.1002/humu.21071
- García Segarra, N., Mittaz, L., Campos-Xavier, A. B., Bartels, C. F., Tuysuz, B., Alanay, Y., et al. (2012). The Diagnostic challenge of Progressive Pseudorheumatoid Dysplasia (PPRD): a Review of Clinical Features, Radiographic Features, and WISP3 Mutations in 63 Affected Individuals. *Am. J. Med. Genet.* 160C (3), 217–229. doi:10.1002/ajmg.c.31333
- Garea, J. S. (2006). *Identification and Characterisation of CFTR Mutations at Transcript Level and Association with Disease Severity in Cystic Fibrosis* PhD Dissertation. Bern, Switzerland: Universität Bern.
- Geng, X., Liu, Y., Ren, X., Guan, Y., Wang, Y., Mao, B., et al. (2018). Novel NTRK1 Mutations in Chinese Patients with Congenital Insensitivity to Pain with Anhidrosis. *Mol. Pain* 14, 1744806918781140. doi:10.1177/1744806918781140
- Gillis, E., Kempers, M., Saleminck, S., Timmermans, J., Cheriex, E. C., Bekkers, S. C. A. M., et al. (2014). AnFBN1 Deep Intronic Mutation in a Familial Case of Marfan Syndrome: An Explanation for Genetically Unsolved Cases? *Hum. Mutat.* 35 (5), 571–574. doi:10.1002/humu.22540
- Gómez-Grau, M., Albaigés, J., Casas, J., Auladell, C., Dierssen, M., Vilageliu, L., et al. (2017). New Murine Niemann-Pick Type C Models Bearing a Pseudoxon-Generating Mutation Recapitulate the Main Neurobehavioural and Molecular Features of the Disease. *Sci. Rep.* 7, 41931. doi:10.1038/srep41931
- Gonçalves, A., Oliveira, J., Coelho, T., Taipa, R., Melo-Pires, M., Sousa, M., et al. (2017). Exonization of an Intronic LINE-1 Element Causing Becker Muscular Dystrophy as a Novel Mutational Mechanism in Dystrophin Gene. *Basel: Genes*, 8. doi:10.3390/genes8100253
- Gonorazky, H. D., Naumenko, S., Ramani, A. K., Nelakuditi, V., Mashouri, P., Wang, P., et al. (2019). Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am. J. Hum. Genet.* 104 (3), 466–483. doi:10.1016/j.ajhg.2019.01.012



- Gonorazky, H., Liang, M., Cummings, B., Lek, M., Micallef, J., Hawkins, C., et al. (2016). RNA Seq Analysis for the Diagnosis of Muscular Dystrophy. *Ann. Clin. Transl. Neurol.* 3 (1), 55–60. doi:10.1002/acn3.267
- Gooding, C., Clark, F., Wollerton, M. C., Grellscheid, S.-N., Groom, H., and Smith, C. W. (2006). A Class of Human Exons with Predicted Distant branch Points Revealed by Analysis of AG Dinucleotide Exclusion Zones. *Genome Biol.* 7 (1), R1. doi:10.1186/gb-2006-7-1-r1
- Goossens, R., van den Boogaard, M. L., Lemmers, R. J. L. F., Balog, J., van der Vliet, P. J., Willemsen, I. M., et al. (2019). Intronic SMCHD1 Variants in FSHD: Testing the Potential for CRISPR-Cas9 Genome Editing. *J. Med. Genet.* 56 (12), 828–837. doi:10.1136/jmedgenet-2019-106402
- Greer, K., Mizzi, K., Rice, E., Kuster, L., Barrero, R. A., Bellgard, M. L., et al. (2015). Pseudoexon Activation Increases Phenotype Severity in a Becker Muscular Dystrophy Patient. *Mol. Genet. Genomic Med.* 3 (4), 320–326. doi:10.1002/mgg3.144
- Grodecká, L., Kováčová, T., Kramárek, M., Seneca, S., Stouffs, K., De Laet, C., et al. (2017). Detailed Molecular Characterization of a Novel IDS Exonic Mutation Associated with Multiple Pseudoexon Activation. *J. Mol. Med.* 95 (3), 299–309. doi:10.1007/s00109-016-1484-2
- Gualandi, F., Rimessi, P., Trabanelli, C., Spitali, P., Neri, M., Patarnello, T., et al. (2006). Intronic Breakpoint Definition and Transcription Analysis in DMD/BMD Patients with Deletion/duplication at the 5' Mutation Hot Spot of the Dystrophin Gene. *Gene* 370, 26–33. doi:10.1016/j.gene.2005.11.002
- Guo, D. C., Gupta, P., Tran-Fadulu, V., Guidry, T. V., Leduc, M. S., Schaefer, F. V., et al. (2008). An FBN1 Pseudoexon Mutation in a Patient with Marfan Syndrome: Confirmation of Cryptic Mutations Leading to Disease. *J. Hum. Genet.* 53 (11–12), 1007–1011. doi:10.1007/s10038-008-0334-7
- Guo, L., Bertola, D. R., Takanohashi, A., Saito, A., Segawa, Y., Yokota, T., et al. (2019). Bi-allelic CSF1R Mutations Cause Skeletal Dysplasia of Dysosteosclerosis-Pyle Disease Spectrum and Degenerative Encephalopathy with Brain Malformation. *Am. J. Hum. Genet.* 104 (5), 925–935. doi:10.1016/j.ajhg.2019.03.004
- Gurvich, O. L., Tuohy, T. M., Howard, M. T., Finkel, R. S., Medne, L., Anderson, C. B., et al. (2008). DMDpseudoexon Mutations: Splicing Efficiency, Phenotype, and Potential Therapy. *Ann. Neurol.* 63 (1), 81–89. doi:10.1002/ana.21290
- Hamanaka, K., Miyatake, S., Koshimizu, E., Tsurusaki, Y., Mitsushashi, S., Iwama, K., et al. (2019). RNA Sequencing Solved the Most Common but Unrecognized NEB Pathogenic Variant in Japanese Nemaline Myopathy. *Genet. Med.* 21 (7), 1629–1638. doi:10.1038/s41436-018-0360-6
- Han, A., Stoilov, P., Linares, A. J., Zhou, Y., Fu, X.-D., and Black, D. L. (2014). De Novo prediction of PTBP1 Binding and Splicing Targets Reveals Unexpected Features of its RNA Recognition and Function. *Plos Comput. Biol.* 10 (1), e1003442. doi:10.1371/journal.pcbi.1003442
- Hastings, M. L., Resta, N., Traum, D., Stella, A., Guanti, G., and Krainer, A. R. (2005). An LKB1 AT-AC Intron Mutation Causes Peutz-Jeghers Syndrome via Splicing at Noncanonical Cryptic Splice Sites. *Nat. Struct. Mol. Biol.* 12 (1), 54–59. doi:10.1038/nsmb873
- Helman, G., Compton, A. G., Hock, D. H., Walkiewicz, M., Brett, G. R., Pais, L., et al. (2021). Multiomic Analysis Elucidates Complex I Deficiency Caused by a Deep Intronic Variant in NDUFB10. *Hum. Mutat.* 42 (1), 19–24. doi:10.1002/humu.24135
- Highsmith, W. E., Burch, L. H., Zhou, Z., Olsen, J. C., Boat, T. E., Spock, A., et al. (1994). A Novel Mutation in the Cystic Fibrosis Gene in Patients with Pulmonary Disease but normal Sweat Chloride Concentrations. *N. Engl. J. Med.* 331 (15), 974–980. doi:10.1056/NEJM199410133311503
- Higuchi, T., Kobayashi, M., Ogata, J., Kaneshiro, E., Shimada, Y., Kobayashi, H., et al. (2016). Identification of Cryptic Novel  $\alpha$ -Galactosidase A Gene Mutations: Abnormal mRNA Splicing and Large Deletions. *JIMD Rep.* 30, 63–72. doi:10.1007/8904\_2015\_475
- Hilgert, N., Topsakal, V., van Dinther, J., Offeciers, E., Van de Heyning, P., and Van Camp, G. (2008). A Splice-Site Mutation and Overexpression of MYO6 Cause a Similar Phenotype in Two Families with Autosomal Dominant Hearing Loss. *Eur. J. Hum. Genet.* 16 (5), 593–602. doi:10.1038/sj.ejhg.5202000
- Hiraide, T., Nakashima, M., Ikeda, T., Tanaka, D., Osaka, H., and Saito, H. (2020). Identification of a Deep Intronic POLR3A Variant Causing Inclusion of a Pseudoexon Derived from an Alu Element in Pol III-Related Leukodystrophy. *J. Hum. Genet.* 65 (10), 921–925. doi:10.1038/s10038-020-0786-y
- Holliday, M., Singer, E. S., Ross, S. B., Lim, S., Lal, S., Ingles, J., et al. (2021). Transcriptome Sequencing of Patients with Hypertrophic Cardiomyopathy Reveals Novel Splice-Altering Variants in MYBPC3. *Circ. Genom. Precis. Med.* 14 (2), e003202. doi:10.1161/CIRCGEN.120.003202
- Homolova, K., Zavadakova, P., Doktor, T. K., Schroeder, L. D., Kozich, V., and Andresen, B. S. (2010). The Deep Intronic c.903+469T>C Mutation in the MTRR Gene Creates an SF2/ASF Binding Exonic Splicing Enhancer, Which Leads to Pseudoexon Activation and Causes the cblE Type of Homocystinuria. *Hum. Mutat.* 31 (4), 437–444. doi:10.1002/humu.21206
- Horinouchi, T., Nozu, K., Yamamura, T., Minamikawa, S., Omori, T., Nakanishi, K., et al. (2018). Detection of Splicing Abnormalities and Genotype-Phenotype Correlation in X-Linked Alport Syndrome. *Jasn* 29 (8), 2244–2254. doi:10.1681/ASN.2018030228
- Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., et al. (2021). Ensembl 2021. *Nucleic Acids Res.* 49 (D1), D884–D891. doi:10.1093/nar/gkaa942
- Hug, N., Longman, D., and Cáceres, J. F. (2016). Mechanism and Regulation of the Nonsense-Mediated Decay Pathway. *Nucleic Acids Res.* 44 (4), 1483–1495. doi:10.1093/nar/gkw010
- Huizing, M., Anikster, Y., Fitzpatrick, D. L., Jeong, A. B., D'Souza, M., Rausche, M., et al. (2001). Hermansky-Pudlak Syndrome Type 3 in Ashkenazi Jews and Other Non-Puerto Rican Patients with Hypopigmentation and Platelet Storage-Pool Deficiency. *Am. J. Hum. Genet.* 69 (5), 1022–1032. doi:10.1086/324168
- Hujová, P., Souček, P., Grodecká, L., Grombřířková, H., Ravčuková, B., Kuklínek, P., et al. (2020). Deep Intronic Mutation in SERPING1 Caused Hereditary Angioedema through Pseudoexon Activation. *J. Clin. Immunol.* 40 (3), 435–446. doi:10.1007/s10875-020-00753-2
- Ikeda, H., Matsubara, Y., Mikami, H., Kure, S., Owada, M., Gough, T., et al. (1997). Molecular Analysis of Dihydropteridine Reductase Deficiency: Identification of Two Novel Mutations in Japanese Patients. *Hum. Genet.* 100 (5–6), 637–642. doi:10.1007/s004390050566
- Ikezawa, M., Minami, N., Takahashi, M., Goto, Y.-i., Miike, T., and Nonaka, I. (1998). Dystrophin Gene Analysis on 130 Patients with Duchenne Muscular Dystrophy with a Special Reference to Muscle mRNA Analysis. *Brain Dev.* 20 (3), 165–168. doi:10.1016/s0387-7604(98)00012-6
- Inaba, H., Koyama, T., Shinozawa, K., Amano, K., and Fukutake, K. (2013). Identification and Characterization of an Adenine to Guanine Transition within Intron 10 of the Factor VIII Gene as a Causative Mutation in a Patient with Mild Haemophilia A. *Haemophilia* 19 (1), 100–105. doi:10.1111/j.1365-2516.2012.02906.x
- Ing, A., Wlodaver, A., Kirschmann, D., Toledo, E., McCabe, C., Kadri, S., et al. (2021). Transcript Analysis for Variant Classification Resolution in a Child with Primary Ciliary Dyskinesia. *Cold Spring Harb. Mol. Case Stud.* 7 (1), 1. doi:10.1101/mcs.a005363
- Ishibashi, K., Takeshima, Y., Yagi, M., Nishiyama, A., and Matsuo, M. (2006). Novel Cryptic Exons Identified in Introns 2 and 3 of the Human Dystrophin Gene with Duplication of Exons 8–11. *Kobe J. Med. Sci.* 52 (3–4), 61–75.
- Ishigaki, K., Nicolle, D., Krejci, E., Leroy, J.-P., Koenig, J., Fardeau, M., et al. (2003). Two Novel Mutations in the COLQ Gene Cause Endplate Acetylcholinesterase Deficiency. *Neuromuscul. Disord.* 13 (3), 236–244. doi:10.1016/s0960-8966(02)00243-2
- Ishigaki, S., Masuda, A., Fujioka, Y., Iguchi, Y., Katsuno, M., Shibata, A., et al. (2012). Position-dependent FUS-RNA Interactions Regulate Alternative Splicing Events and Transcriptions. *Sci. Rep.* 2, 529. doi:10.1038/srep00529
- Ishii, S., Nakao, S., Minamikawa-Tachino, R., Desnick, R. J., and Fan, J.-Q. (2002). Alternative Splicing in the  $\alpha$ -Galactosidase A Gene: Increased Exon Inclusion Results in the Fabry Cardiac Phenotype. *Am. J. Hum. Genet.* 70 (4), 994–1002. doi:10.1086/339431
- Ishmukhametova, A., Van Kien, P. K., Méchin, D., Thorel, D., Vincent, M.-C., Rivier, F., et al. (2012). Comprehensive Oligonucleotide Array-Comparative Genomic Hybridization Analysis: New Insights into the Molecular Pathology of the DMD Gene. *Eur. J. Hum. Genet.* 20 (10), 1096–1100. doi:10.1038/ejhg.2012.51
- Isler, J., Rüfenacht, V., Gemperle, C., Allegri, G., and Häberle, J. (2020). Improvement of Diagnostic Yield in Carbamoylphosphate Synthetase 1 (CPS1 ) Molecular Genetic Investigation by RNA Sequencing. *JIMD Rep.* 52 (1), 28–34. doi:10.1002/jmd2.12091



- Jamshidi, F., Place, E. M., Mehrotra, S., Navarro-Gomez, D., Maher, M., Branham, K. E., et al. (2019). Contribution of Noncoding Pathogenic Variants to RPGRIP1-Mediated Inherited Retinal Degeneration. *Genet. Med.* 21 (3), 694–704. doi:10.1038/s41436-018-0104-7
- Jang, M.-A., Kim, Y.-E., Kim, S. K., Lee, M.-K., Kim, J.-W., and Ki, C.-S. (2016). Identification and Characterization of NF1 Splicing Mutations in Korean Patients with Neurofibromatosis Type 1. *J. Hum. Genet.* 61 (8), 705–709. doi:10.1038/jhg.2016.33
- Janin, A., Chanavat, V., Rollat-Farnier, P. A., Bardel, C., Nguyen, K., Chevalier, P., et al. (2020). Whole MYBPC3 NGS Sequencing as a Molecular Strategy to Improve the Efficiency of Molecular Diagnosis of Patients with Hypertrophic Cardiomyopathy. *Hum. Mutat.* 41 (2), 465–475. doi:10.1002/humu.23944
- Jiang, C., and Zhao, Z. (2006). Mutational Spectrum in the Recent Human Genome Inferred by Single Nucleotide Polymorphisms. *Genomics* 88 (5), 527–534. doi:10.1016/j.ygeno.2006.06.003
- Jin, M., Li, J.-J., Xu, G.-R., Wang, N., and Wang, Z.-Q. (2020). Cryptic Exon Activation Causes Dystrophinopathy in Two Chinese Families. *Eur. J. Hum. Genet.* 28 (7), 947–955. doi:10.1038/s41431-020-0578-z
- Jin, X., Yan, Y., Zhang, C., Tai, Y., An, L., Yu, X., et al. (2021). Identification of Novel Deep Intronic PAH Gene Variants in Patients Diagnosed with Phenylketonuria. *Hum. Mutat.* doi:10.1002/humu.24292
- Jo, B.-S., and Choi, S. S. (2015). Introns: The Functional Benefits of Introns in Genomes. *Genomics Inform.* 13 (4), 112–118. doi:10.5808/GI.2015.13.4.112
- Jourdy, Y., Janin, A., Fretigny, M., Lienhart, A., Négrier, C., Bozon, D., et al. (2018). Recurrent F8 Intronic Deletion Found in Mild Hemophilia A Causes Alu Exonization. *Am. J. Hum. Genet.* 102 (2), 199–206. doi:10.1016/j.ajhg.2017.12.010
- Juan-Mateu, J., González-Quereda, L., Rodríguez, M. J., Verdura, E., Lázaro, K., Jou, C., et al. (2013). Interplay between DMD point Mutations and Splicing Signals in Dystrophinopathy Phenotypes. *PLoS One* 8 (3), e59916. doi:10.1371/journal.pone.0059916
- Kaida, D. (2016). The Reciprocal Regulation between Splicing and 3'-end Processing. *WIREs RNA* 7 (4), 499–511. doi:10.1002/wrna.1348
- Kaimori, J.-y., Ichimaru, N., Isaka, Y., Hashimoto, F., Fu, X., Hashimura, Y., et al. (2013). Renal Transplantations from Parents to Siblings with Autosomal Recessive Alport Syndrome Caused by a Rearrangement in an Intronic Antisense Alu Element in the COL4A3 Gene Led to Different Outcomes. *CEN Case Rep.* 2 (1), 98–101. doi:10.1007/s13730-012-0049-7
- Kalscheuer, V. M., Musante, L., Fang, C., Hoffmann, K., Fuchs, C., Carta, E., et al. (2009). A Balanced Chromosomal Translocation disrupting ARHGAP9 is Associated with Epilepsy, Anxiety, Aggression, and Mental Retardation. *Hum. Mutat.* 30 (1), 61–68. doi:10.1002/humu.20814
- Kannu, P., Nour, M., Irving, M., Xie, J., Loder, D., Lai, J., et al. (2013). Paraspinal Ganglioneuroma in the Proband of a Large Family with Mild Cutaneous Manifestations of NF1, Carrying a deepNF1intronic Mutation. *Clin. Genet.* 83 (2), 191–194. doi:10.1111/j.1399-0004.2012.01882.x
- Känsäkoski, J., Jääskeläinen, J., Jääskeläinen, T., Tommiska, J., Saarinen, L., Lehtonen, R., et al. (2016). Complete Androgen Insensitivity Syndrome Caused by a Deep Intronic Pseudoexon-Activating Mutation in the Androgen Receptor Gene. *Sci. Rep.* 6, 32819. doi:10.1038/srep32819
- Kazakov, D. V., Thoma-Uszynski, S., Vanecek, T., Kacerovska, D., Grossmann, P., and Michal, M. (2009). A Case of Brooke-Spiegler Syndrome with a Novel Germline Deep Intronic Mutation in the CYLD Gene Leading to Intronic Exonization, Diverse Somatic Mutations, and Unusual Histology. *Am. J. Dermatopathol* 31 (7), 664–673. doi:10.1097/DAD.0b013e3181a05dad
- Ke, S., Shang, S., Kalachikov, S. M., Morozova, I., Yu, L., Russo, J. J., et al. (2011). Quantitative Evaluation of All Hexamers as Exonic Splicing Elements. *Genome Res.* 21 (8), 1360–1374. doi:10.1101/gr.119628.110
- Keegan, N. P. (2020). Pseudoexons of the DMD Gene. *Jnd* 7 (2), 77–95. doi:10.3233/JND-190431
- Keeratichamroen, S., Cairns, J. R., Wattanasirichaigoon, D., Wasant, P., Ngwisara, L., Suwannarat, P., et al. (2008). Molecular Analysis of the Iduronate-2-Sulfatase Gene in Thai Patients with Hunter Syndrome. *J. Inherit. Metab. Dis.* 31 (Suppl. 2), S303–S311. doi:10.1007/s10545-008-0876-z
- Kelly, S., Georgomanolis, T., Zirkel, A., Diermeier, S., O'Reilly, D., Murphy, S., et al. (2015). Splicing of many Human Genes Involves Sites Embedded within Introns. *Nucleic Acids Res.* 43 (9), 4721–4732. doi:10.1093/nar/gkv386
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The Human Genome Browser at UCSC. *Genome Res.* 12 (6), 996–1006. doi:10.1101/gr.229102
- Khan, A. O., Becirovic, E., Betz, C., Neuhaus, C., Altmüller, J., Maria Riedmayr, L., et al. (2017). A Deep Intronic CLRN1 (USH3A) Founder Mutation Generates an Aberrant Exon and Underlies Severe Usher Syndrome on the Arabian Peninsula. *Sci. Rep.* 7 (1), 1411. doi:10.1038/s41598-017-01577-8
- Khan, M., Arno, G., Fakin, A., Parfitt, D. A., Dhooge, P. P. A., Albert, S., et al. (2020a). Detailed Phenotyping and Therapeutic Strategies for Intronic ABCA4 Variants in Stargardt Disease. *Mol. Ther. - Nucleic Acids* 21, 412–427. doi:10.1016/j.omtn.2020.06.007
- Khan, M., Cornelis, S. S., De Pozo-Valero, M. I., Whelan, L., Runhart, E. H., Mishra, K., et al. (2020b). Resolving the Dark Matter of ABCA4 for 1054 Stargardt Disease Probands through Integrated Genomics and Transcriptomics. *Genet. Med.* 22 (7), 1235–1246. doi:10.1038/s41436-020-0787-4
- Khan, M., Cornelis, S. S., Khan, M. I., Elmelik, D., Manders, E., Bakker, S., et al. (2019). Cost-effective Molecular Inversion Probe-based ABCA4 Sequencing Reveals Deep-intronic Variants in Stargardt Disease. *Hum. Mutat.* 40 (10), 1749–1759. doi:10.1002/humu.23787
- Khelifi, M. M., Ishmukhametova, A., Van Kien, P. K., Thorel, D., Méchin, D., Perelman, S., et al. (2011). Pure Intronic Rearrangements Leading to Aberrant Pseudoexon Inclusion in Dystrophinopathy: a New Class of Mutations? *Hum. Mutat.* 32 (4), 467–475. doi:10.1002/humu.21471
- Khourieh, J., Rao, G., Habib, T., Avery, D. T., Lefèvre-Utile, A., Chandresris, M.-O., et al. (2019). A Deep Intronic Splice Mutation of STAT3 Underlies Hyper IgE Syndrome by Negative Dominance. *Proc. Natl. Acad. Sci. USA* 116 (33), 16463–16472. doi:10.1073/pnas.1901409116
- Kim, S. W., Taggart, A. J., Heintzelman, C., Cygan, K. J., Hull, C. G., Wang, J., et al. (2017). Widespread Intra-dependencies in the Removal of Introns from Human Transcripts. *Nucleic Acids Res.* 45 (16), 9503–9513. doi:10.1093/nar/gkx661
- King, K., Flinter, F., Nihalani, V., and Green, P. (2002). Unusual Deep Intronic Mutations in the COL4A5 Gene Cause X Linked Alport Syndrome. *Hum. Genet.* 111 (6), 548–554. doi:10.1007/s00439-002-0830-3
- Knebelmann, B., Forestier, L., Drouot, L., Quinones, S., Chuet, C., Benessy, F., et al. (1995). Splice-mediated Insertion of an Alu Sequence in the COL4A3 mRNA Causing Autosomal Recessive Alport Syndrome. *Hum. Mol. Genet.* 4 (4), 675–679. doi:10.1093/hmg/4.4.675
- Knight, S., Vulliamy, T., Morgan, B., Devriendt, K., Mason, P., and Dokal, I. (2001). Identification of Novel DKC1 Mutations in Patients with Dyskeratosis Congenita: Implications for Pathophysiology and Diagnosis. *Hum. Genet.* 108 (4), 299–303. doi:10.1007/s004390100494
- Kollberg, G., Tulinius, M., Melberg, A., Darin, N., Andersen, O., Holmgren, D., et al. (2009). Clinical Manifestation and a New ISCU Mutation in Iron-sulphur Cluster Deficiency Myopathy. *Brain* 132 (Pt 8), 2170–2179. doi:10.1093/brain/awp152
- Kossack, N., Simoni, M., Richter-Unruh, A., Themmen, A. P. N., and Gromoll, J. (2008). Mutations in a Novel, Cryptic Exon of the Luteinizing Hormone/chorionic Gonadotropin Receptor Gene Cause Male Pseudohermaphroditism. *Plos Med.* 5 (4), e88. doi:10.1371/journal.pmed.0050088
- Koster, R., Brandão, R. D., Tserpelis, D., van Roozendaal, C. E. P., van Oosterhoud, C. N., Claes, K. B. M., et al. (2021). Pathogenic Neurofibromatosis Type 1 (NF1) RNA Splicing Resolved by Targeted RNAseq. *NPJ Genom Med* 6, 95. doi:10.1038/s41525-021-00258-w
- Kráľovičová, J., Hwang, G., Asplund, A. C., Churbanov, A., Smith, C. I. E., and Vořechovský, I. (2011). Compensatory Signals Associated with the Activation of Human GC 5' Splice Sites. *Nucleic Acids Res.* 39 (16), 7077–7091. doi:10.1093/nar/gkr306
- Kráľovičová, J., Knut, M., Cross, N. C. P., and Vořechovský, I. (2016). Exon-centric Regulation of ATM Expression Is Population-dependent and Amenable to Antisense Modification by Pseudoexon Targeting. *Sci. Rep.* 6, 18741. doi:10.1038/srep18741
- Kráľovičová, J., and Vořechovský, I. (2007). Global Control of Aberrant Splice-Site Activation by Auxiliary Splicing Sequences: Evidence for a Gradient in Exon and Intron Definition. *Nucleic Acids Res.* 35 (19), 6399–6413. doi:10.1093/nar/gkm680
- Krawczak, M., Thomas, N. S. T., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., et al. (2007). Single Base-Pair Substitutions in Exon-Intron Junctions of Human

- Genes: Nature, Distribution, and Consequences for mRNA Splicing. *Hum. Mutat.* 28 (2), 150–158. doi:10.1002/humu.20400
- Kremer, L. S., Bader, D. M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., et al. (2017). Genetic Diagnosis of Mendelian Disorders via RNA Sequencing. *Nat. Commun.* 8, 15824. doi:10.1038/ncomms15824
- Kröll-Hermi, A., Ebstein, F., Stoetzel, C., Geoffroy, V., Schaefer, E., Scheidecker, S., et al. (2020). Proteasome Subunit PSMC3 Variants Cause Neurosensory Syndrome Combining Deafness and Cataract Due to Proteotoxic Stress. *EMBO Mol. Med.* 12 (7), e11861. doi:10.15252/emmm.201911861
- Kuehl, P., Zhang, J., Lin, Y., Lamba, J., Assem, M., Schuetz, J., et al. (2001). Sequence Diversity in CYP3A Promoters and Characterization of the Genetic Basis of Polymorphic CYP3A5 Expression. *Nat. Genet.* 27 (4), 383–391. doi:10.1038/86882
- Lai, C. Y., Tsai, I. J., Chiu, P. C., Ascher, D. B., Chien, Y. H., Huang, Y. H., et al. (2021). A Novel Deep Intronic Variant Strongly Associates with Alkaptonuria. *NPJ Genom. Med.* 6 (89), 89. doi:10.1038/s41525-021-00252-2
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial Sequencing and Analysis of the Human Genome. *Nature* 409 (6822), 860–921. doi:10.1038/35057062
- Landrith, T., Li, B., Cass, A. A., Conner, B. R., LaDuca, H., McKenna, D. B., et al. (2020). Splicing Profile by Capture RNA-Seq Identifies Pathogenic Germline Variants in Tumor Suppressor Genes. *Npj Precis. Onc.* 4, 4. doi:10.1038/s41698-020-0109-y
- Larrue, R., Chamley, P., Bardyn, T., Lionet, A., Gnemmi, V., Cauffiez, C., et al. (2020). Diagnostic Utility of Whole-Genome Sequencing for Nephronophthisis. *NPJ Genom. Med.* 5 (1), 38. doi:10.1038/s41525-020-00147-8
- Lassalle, F., Jourdy, Y., Jouan, L., Swystun, L., Gauthier, J., Zawadzki, C., et al. (2020). The challenge of Genetically Unresolved Haemophilia A Patients: Interest of the Combination of Whole F8 Gene Sequencing and Functional Assays. *Haemophilia* 26 (6), 1056–1063. doi:10.1111/hae.14179
- Latchman, K., Brown, J., Sineni, C. J., Ragin-Dames, L., Guo, S., Huang, J., et al. (2020). A Founder Noncoding GALT Variant Interfering with Splicing Causes Galactosemia. *Jrnl Inher. Metab. Disea* 43 (6), 1199–1204. doi:10.1002/jimd.12293
- Laugel, V., Dalloz, C., Durand, M., Sauvanaud, F., Kristensen, U., Vincent, M. C., et al. (2010). Mutation Update for the CSB/ERCC6 and CSA/ERCC8 genes Involved in Cockayne Syndrome. *Hum. Mutat.* 31 (2), 113–126. doi:10.1002/humu.21154
- Lebon, S., Minai, L., Chretien, D., Corcos, J., Serre, V., Kadhom, N., et al. (2007). A Novel Mutation of the NDUFS7 Gene Leads to Activation of a Cryptic Exon and Impaired Assembly of Mitochondrial Complex I in a Patient with Leigh Syndrome. *Mol. Genet. Metab.* 92 (1-2), 104–108. doi:10.1016/j.ymgme.2007.05.010
- Lee, H., Huang, A. Y., Wang, L.-k., Yoon, A. J., Renteria, G., Eskin, A., et al. (2020). Diagnostic Utility of Transcriptome Sequencing for Rare Mendelian Diseases. *Genet. Med.* 22 (3), 490–499. doi:10.1038/s41436-019-0672-1
- Lee, M., Roos, P., Sharma, N., Atalar, M., Evans, T. A., Pellicore, M. J., et al. (2017). Systematic Computational Identification of Variants that Activate Exonic and Intronic Cryptic Splice Sites. *Am. J. Hum. Genet.* 100 (5), 751–765. doi:10.1016/j.ajhg.2017.04.001
- Lee, W.-I., Torgerson, T. R., Schumacher, M. J., Yel, L., Zhu, Q., and Ochs, H. D. (2005). Molecular Analysis of a Large Cohort of Patients with the Hyper Immunoglobulin M (IgM) Syndrome. *Blood* 105 (5), 1881–1890. doi:10.1182/blood-2003-12-4420
- Levitus, M., Waisfisz, Q., Godthelp, B. C., Vries, Y. d., Hussain, S., Wiegant, W. W., et al. (2005). The DNA Helicase BRIP1 Is Defective in Fanconi Anemia Complementation Group J. *Nat. Genet.* 37 (9), 934–935. doi:10.1038/ng1625
- Ligtenberg, M. J. L., Kuiper, R. P., Chan, T. L., Goossens, M., Hebeda, K. M., Voorendt, M., et al. (2009). Heritable Somatic Methylation and Inactivation of MSH2 in Families with Lynch Syndrome Due to Deletion of the 3' Exons of TACSTD1. *Nat. Genet.* 41 (1), 112–117. doi:10.1038/ng.283
- Lim, B. C., Ki, C.-S., Kim, J.-W., Cho, A., Kim, M. J., Hwang, H., et al. (2010). Fukutin Mutations in Congenital Muscular Dystrophies with Defective Glycosylation of Dystroglycan in Korea. *Neuromuscul. Disord.* 20 (8), 524–530. doi:10.1016/j.nmd.2010.06.005
- Liquori, A., Vaché, C., Baux, D., Blanchet, C., Hamel, C., Malcolm, S., et al. (2016). WholeUSH2AGene Sequencing Identifies Several New Deep Intronic Mutations. *Hum. Mutat.* 37 (2), 184–193. doi:10.1002/humu.22926
- Liu, W., Han, B., Zhu, W., Cheng, T., Fan, M., Wu, J., et al. (2017). Polymorphism in the Alternative Donor Site of the Cryptic Exon of LHCGR: Functional Consequences and Associations with Testosterone Level. *Sci. Rep.* 7, 45699. doi:10.1038/srep45699
- Lo, Y.-F., Nozu, K., Iijima, K., Morishita, T., Huang, C.-C., Yang, S.-S., et al. (2011). Recurrent Deep Intronic Mutations in theSLC12A3Gene Responsible for Gitelman's Syndrome. *Cjasm* 6 (3), 630–639. doi:10.2215/CJN.06730810
- Lornage, X., Scharfner, V., Balduino, I., Biancalana, V., Willis, T., Echaniz-Laguna, A., et al. (2019). Clinical, Histological, and Genetic Characterization of PYROXD1-Related Myopathy. *Acta Neuropathol. Commun.* 7 (1), 138. doi:10.1186/s40478-019-0781-8
- Lou, H., Helfman, D. M., Gagel, R. F., and Berget, S. M. (1999). Polypyrimidine Tract-Binding Protein Positively Regulates Inclusion of an Alternative 3'-Terminal Exon. *Mol. Cell Biol* 19 (1), 78–85. doi:10.1128/MCB.19.1.78
- Lu, X., Han, C., Mai, J., Jiang, X., Liao, J., Hou, Y., et al. (2021). Novel Intronic Mutations Introduce Pseudoexons in DMD that Cause Muscular Dystrophy in Patients. *Front. Genet.* 12, 657040. doi:10.3389/fgene.2021.657040
- Lualdi, S., Pittis, M. G., Regis, S., Parini, R., Allegri, A. E., Furlan, F., et al. (2006). Multiple Cryptic Splice Sites Can Be Activated by IDS point Mutations Generating Misspliced Transcripts. *J. Mol. Med.* 84 (8), 692–700. doi:10.1007/s00109-006-0057-1
- Lucien, N., Chiaroni, J., Cartron, J.-P., and Bailly, P. (2002). Partial Deletion in the JK Locus Causing a Jknul Phenotype. *Blood* 99 (3), 1079–1081. doi:10.1182/blood.v99.3.1079
- Ma, S. L., Vega-Warner, V., Gillies, C., Sampson, M. G., Kher, V., Sethi, S. K., et al. (2015). Whole Exome Sequencing Reveals Novel PHEX Splice Site Mutations in Patients with Hypophosphatemic Rickets. *PLoS One* 10 (6), e0130729. doi:10.1371/journal.pone.0130729
- Madan, V., Kanojia, D., Li, J., Okamoto, R., Sato-Otsubo, A., Kohlmann, A., et al. (2015). Aberrant Splicing of U12-type Introns Is the Hallmark of ZRSR2 Mutant Myelodysplastic Syndrome. *Nat. Commun.* 6, 6042. doi:10.1038/ncomms7042
- Madden, H. R., Fletcher, S., Davis, M. R., and Wilton, S. D. (2009). Characterization of a Complex Duchenne Muscular Dystrophy-Causing Dystrophin Gene Inversion and Restoration of the reading Frame by Induced Exon Skipping. *Hum. Mutat.* 30 (1), 22–28. doi:10.1002/humu.20806
- Magri, F., Del Bo, R., D'Angelo, M. G., Govoni, A., Ghezzi, S., Gandossini, S., et al. (2011). Clinical and Molecular Characterization of a Cohort of Patients with Novel Nucleotide Alterations of the Dystrophin Gene Detected by Direct Sequencing. *BMC Med. Genet.* 12, 37. doi:10.1186/1471-2350-12-37
- Malueka, R. G., Takaoka, Y., Yagi, M., Awano, H., Lee, T., Dwianingsih, E. K., et al. (2012). Categorization of 77 Dystrophinexons into 5 Groups by a Decision Tree Using Indexes of Splicing Regulatory Factors as Decision Markers. *BMC Genet.* 13, 23. doi:10.1186/1471-2156-13-23
- Mameli, E., Lepori, M. B., Chiappe, F., Ranucci, G., Di Dato, F., Iorio, R., et al. (2015). Wilson's Disease Caused by Alternative Splicing and Alu Exonization Due to a Homozygous 3039-bp Deletion Spanning from Intron 1 to Exon 2 of the ATP7B Gene. *Gene* 569 (2), 276–279. doi:10.1016/j.gene.2015.05.067
- Marek-Yagel, D., Eliyahu, A., Veber, A., Shalva, N., Philosoph, A. M., Barel, O., et al. (2021). Deep Intronic Variant in the ARSB Gene as the Genetic Cause for Maroteaux-Lamy Syndrome (MPS VI). *Am. J. Med. Genet. A.* 185 (12), 3804–3809. doi:10.1002/ajmg.a.62453
- Martinez, F. J., Lee, J. H., Lee, J. E., Blanco, S., Nickerson, E., Gabriel, S., et al. (2012). Whole Exome Sequencing Identifies a Splicing Mutation in NSUN2 as a Cause of a Dubowitz-like Syndrome. *J. Med. Genet.* 49 (6), 380–385. doi:10.1136/jmedgenet-2011-100686
- Martínez, M. A., Rincón, A., Desviat, L. R., Merinero, B., Ugarte, M., and Pérez, B. (2005). Genetic Analysis of Three Genes Causing Isolated Methylmalonic Acidemia: Identification of 21 Novel Allelic Variants. *Mol. Genet. Metab.* 84 (4), 317–325. doi:10.1016/j.ymgme.2004.11.011
- Martínez-Pizarro, A., Dembic, M., Pérez, B., Andresen, B. S., and Desviat, L. R. (2018). Intronic PAH Gene Mutations Cause a Splicing Defect by a Novel Mechanism Involving U1snRNP Binding Downstream of the 5' Splice Site. *Plos Genet.* 14 (4), e1007360. doi:10.1371/journal.pgen.1007360

- Masala, M. V., Scapaticci, S., Olivieri, C., Pirodda, C., Montesu, M. A., Cuccuru, M. A., et al. (2007). Epidemiology and Clinical Aspects of Werner's Syndrome in North Sardinia: Description of a Cluster. *Eur. J. Dermatol.* 17 (3), 213–216. doi:10.1684/ejd.2007.0155
- Mayer, A. K., Rohrschneider, K., Strom, T. M., Glöckle, N., Kohl, S., Wissinger, B., et al. (2016). Homozygosity Mapping and Whole-Genome Sequencing Reveals a Deep Intronic PROM1 Mutation Causing Cone-Rod Dystrophy by Pseudoexon Activation. *Eur. J. Hum. Genet.* 24 (3), 459–462. doi:10.1038/ejhg.2015.144
- Mayer, K., Ballhausen, W., Leistner, W., and Rott, H.-D. (2000). Three Novel Types of Splicing Aberrations in the Tuberosus Sclerosis TSC2 Gene Caused by Mutations Apart from Splice Consensus Sequences. *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1502 (3), 495–507. doi:10.1016/s0925-4439(00)00072-7
- McConville, C. M., Stankovic, T., Byrd, P. J., McGuire, G. M., Yao, Q. Y., Lennox, G. G., et al. (1996). Mutations Associated with Variant Phenotypes in Ataxia-Telangiectasia. *Am. J. Hum. Genet.* 59 (2), 320–330.
- McDonnell, L. M., Mirzaa, G. M., Alcantara, D., Schwartztruber, J., Carter, M. T., Lee, L. J., et al. (2013). Mutations in STAMBP, Encoding a Deubiquitinating Enzyme, Cause Microcephaly-Capillary Malformation Syndrome. *Nat. Genet.* 45 (5), 556–562. doi:10.1038/ng.2602
- Medina, M. W., Theusch, E., Naidoo, D., Bauzon, F., Stevens, K., Mangravite, L. M., et al. (2012). RHOA Is a Modulator of the Cholesterol-Lowering Effects of Statin. *Plos Genet.* 8 (11), e1003058. doi:10.1371/journal.pgen.1003058
- Meili, D., Kráľovićová, J., Zagalak, J., Bonafé, L., Fiori, L., Blau, N., et al. (2009). Disease-causing Mutations Improving the branch Site and Polypyrimidine Tract: Pseudoexon Activation of LINE-2 and antisenseAlulacking the Poly(T)-tail. *Hum. Mutat.* 30 (5), 823–831. doi:10.1002/humu.20969
- Meischl, C., de Boer, M., Ählin, A., and Roos, D. (2000). A New Exon Created by Intronic Insertion of a Rearranged LINE-1 Element as the Cause of Chronic Granulomatous Disease. *Eur. J. Hum. Genet.* 8 (9), 697–703. doi:10.1038/sj.ejhg.5200523
- Mele, C., Lemaire, M., Iatropoulos, P., Piras, R., Bresin, E., Bettoni, S., et al. (2015). Characterization of a New DGKE Intronic Mutation in Genetically Unsolved Cases of Familial Atypical Hemolytic Uremic Syndrome. *Cjasn* 10 (6), 1011–1019. doi:10.2215/CJN.08520814
- Messiaen, L. M., and Wimmer, K. (2008). “NF1 Mutational Spectrum,” in *Monographs in Human Genetics* 16. Editor D. Kaufmann (Basel: Karger), 63–77. doi:10.1159/000126545
- Messiaen, L., and Wimmer, K. (2012). “Mutation Analysis of the NF1 Gene by cDNA-Based Sequencing of the Coding Region,” in *Advances in Neurofibromatosis Research*. Editors K. S. G. Cunha and M. Geller (Hauppauge, NY: Nova Science Publishers, Inc.), 89–101.
- Michel-Calemard, L., Dijoud, F., Till, M., Lambert, J. C., Vercherat, M., Tardy, V., et al. (2009). Pseudoexon Activation in the PKHD1 Gene: a French Founder Intronic Mutation IVS46+653A>G Causing Severe Autosomal Recessive Polycystic Kidney Disease. *Clin. Genet.* 75 (2), 203–206. doi:10.1111/j.1399-0004.2008.01106.x
- Milh, M., Pop, A., Kanhai, W., Villeneuve, N., Cano, A., Struys, E. A., et al. (2012). Atypical Pyridoxine-dependent Epilepsy Due to a Pseudoexon in ALDH7A1. *Mol. Genet. Metab.* 105 (4), 684–686. doi:10.1016/j.ymgme.2012.01.011
- Mitchell, G. A., Labuda, D., Fontaine, G., Saudubray, J. M., Bonnefont, J. P., Lyonnet, S., et al. (1991). Splice-mediated Insertion of an Alu Sequence Inactivates Ornithine delta-aminotransferase: a Role for Alu Elements in Human Mutation. *Proc. Natl. Acad. Sci.* 88 (3), 815–819. doi:10.1073/pnas.88.3.815
- Mochel, F., Knight, M. A., Tong, W.-H., Hernandez, D., Ayyad, K., Taivassalo, T., et al. (2008). Splice Mutation in the Iron-Sulfur Cluster Scaffold Protein ISCU Causes Myopathy with Exercise Intolerance. *Am. J. Hum. Genet.* 82 (3), 652–660. doi:10.1016/j.ajhg.2007.12.012
- Moles-Fernández, A., Domènech-Vivó, J., Tenés, A., Balmaña, J., Diez, O., and Gutiérrez-Enríquez, S. (2021). “Role of Splicing Regulatory Elements and,” in *Silico Tools Usage in the Identification of Deep Intronic Splicing Variants in Hereditary Breast/Ovarian Cancer Genes* (Basel: Cancers), 13. doi:10.3390/cancers13133341
- Mollin, M., Beaumel, S., Vigne, B., Brault, J., Roux-Buisson, N., Rendu, J., et al. (2021). Clinical, Functional and Genetic Characterization of 16 Patients Suffering from Chronic Granulomatous Disease Variants - Identification of 11 Novel Mutations in CYBB. *Clin. Exp. Immunol.* 203 (2), 247–266. doi:10.1111/cei.13520
- Monnier, N., Gout, J. P., Pin, I., Gauthier, G., and Lunardi, J. (2001). A Novel 3600+11.5 Kb C>G Homozygous Splicing Mutation in a Black African, Consanguineous CF Family. *J. Med. Genet.* 38 (1), E4. doi:10.1136/jmg.38.1.e4
- Monnier, N., Ferreira, A., Marty, I., Labarre-Vila, A., Mezin, P., and Lunardi, J. (2003). A Homozygous Splicing Mutation Causing a Depletion of Skeletal Muscle RYR1 Is Associated with Multi-Mincore Disease Congenital Myopathy with Ophthalmoplegia. *Hum. Mol. Genet.* 12 (10), 1171–1178. doi:10.1093/hmg/ddg121
- Montalban, G., Bonache, S., Moles-Fernández, A., Gisbert-Beamud, A., Tenés, A., Bach, V., et al. (2019). Screening of BRCA1/2 Deep Intronic Regions by Targeted Gene Sequencing Identifies the First Germline BRCA1 Variant Causing Pseudoexon Activation in a Patient with Breast/Ovarian Cancer. *J. Med. Genet.* 56 (2), 63–74. doi:10.1136/jmedgenet-2018-105606
- Morello, R., Bertin, T. K., Chen, Y., Hicks, J., Tonachini, L., Monticone, M., et al. (2006). CRTAP Is Required for Prolyl 3- Hydroxylation and Mutations Cause Recessive Osteogenesis Imperfecta. *Cell* 127 (2), 291–304. doi:10.1016/j.cell.2006.08.039
- Morrison, F. S., Locke, J. M., Wood, A. R., Tuke, M., Pasko, D., Murray, A., et al. (2013). The Splice Site Variant Rs11078928 May Be Associated with a Genotype-dependent Alteration in Expression of GSDMB Transcripts. *BMC Genomics* 14, 627. doi:10.1186/1471-2164-14-627
- Movassat, M., Forouzmand, E., Reese, F., and Hertel, K. J. (2019). Exon Size and Sequence Conservation Improves Identification of Splice-Altering Nucleotides. *RNA* 25 (12), 1793–1805. doi:10.1261/rna.070987.119
- Murdock, D. R., Dai, H., Burrage, L. C., Rosenfeld, J. A., Ketkar, S., Müller, M. F., et al. (2021). Transcriptome-directed Analysis for Mendelian Disease Diagnosis Overcomes Limitations of Conventional Genomic Testing. *J. Clin. Invest.* 131 (1), 1. doi:10.1172/JCI141500
- Nakamura, K., Du, L., Tunuguntla, R., Fike, F., Cavalieri, S., Morio, T., et al. (2012). Functional Characterization and Targeted Correction of ATM Mutations Identified in Japanese Patients with Ataxia-Telangiectasia. *Hum. Mutat.* 33 (1), 198–208. doi:10.1002/humu.21632
- Naruto, T., Okamoto, N., Masuda, K., Endo, T., Hatsukawa, Y., Kohmoto, T., et al. (2015). Deep Intronic GPR143 Mutation in a Japanese Family with Ocular Albinism. *Sci. Rep.* 5, 11334. doi:10.1038/srep11334
- Nasim, M. T., Chernova, T. K., Chowdhury, H. M., Yue, B. G., and Eperon, I. C. (2003). HnRNP G and Tra2beta: Opposite Effects on Splicing Matched by Antagonism in RNA Binding. *Hum. Mol. Genet.* 12 (11), 1337–1348. doi:10.1093/hmg/ddg136
- Navarrete, R., Leal, F., Vega, A. I., Morais-López, A., Garcia-Silva, M. T., Martín-Hernández, E., et al. (2019). Value of Genetic Analysis for Confirming Inborn Errors of Metabolism Detected through the Spanish Neonatal Screening Program. *Eur. J. Hum. Genet.* 27 (4), 556–562. doi:10.1038/s41431-018-0330-0
- Neidhardt, J., Glaus, E., Barthelmes, D., Zeitz, C., Fleischhauer, J., and Berger, W. (2007). Identification and Characterization of a Novel RPGR Isoform in Human Retina. *Hum. Mutat.* 28 (8), 797–807. doi:10.1002/humu.20521
- Nieminen, T. T., Pavicic, W., Porkka, N., Kankainen, M., Järvinen, H. J., Lepistö, A., et al. (2016). Pseudoexons Provide a Mechanism for Allele-specific Expression of APC in Familial Adenomatous Polyposis. *Oncotarget* 7 (43), 70685–70698. doi:10.18632/oncotarget.12206
- Njålsson, R., Carlsson, K., Winkler, A., Larsson, A., and Norgren, S. (2003). Diagnostics in Patients with Glutathione Synthetase Deficiency but without Mutations in the Exons of the GSS Gene. *Hum. Mutat.* 22 (6), 497. doi:10.1002/humu.9199
- Noack, D., Heyworth, P. G., Curnutte, J. T., Rae, J., and Cross, A. R. (1999). A Novel Mutation in the CYBB Gene Resulting in an Unexpected Pattern of Exon Skipping and Chronic Granulomatous Disease. *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1454 (3), 270–274. doi:10.1016/s0925-4439(99)00044-7
- Noack, D., Heyworth, P. G., Newburger, P. E., and Cross, A. R. (2001). An Unusual Intronic Mutation in the CYBB Gene Giving Rise to Chronic Granulomatous Disease. *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1537 (2), 125–131. doi:10.1016/s0925-4439(01)00065-5



- Nogueira, C., Silva, L., Marcão, A., Sousa, C., Fonseca, H., Rocha, H., et al. (2021). Role of RNA in Molecular Diagnosis of MADD Patients. *Biomedicines* 9 (5), 1. doi:10.3390/biomedicines9050507
- Novoselova, T. V., Rath, S. R., Carpenter, K., Pachter, N., Dickinson, J. E., Price, G., et al. (2015). NNTPseudoxon Activation as a Novel Mechanism for Disease in Two Siblings with Familial Glucocorticoid Deficiency. *J. Clin. Endocrinol. Metab.* 100 (2), E350–E354. doi:10.1210/jc.2014-3641
- Nozu, K., Iijima, K., Igarashi, T., Yamada, S., Kralovicova, J., Nozu, Y., et al. (2017a). A Birth of Bipartite Exon by Intragenic Deletion. *Mol. Genet. Genomic Med.* 5 (3), 287–294. doi:10.1002/mgg3.277
- Nozu, K., Iijima, K., Nozu, Y., Ikegami, E., Imai, T., Fu, X. J., et al. (2009). A Deep Intronic Mutation in the SLC12A3 Gene Leads to Gitelman Syndrome. *Pediatr. Res.* 66 (5), 590–593. doi:10.1203/PDR.0b013e3181b9b4d3
- Nozu, K., Nozu, Y., Nakanishi, K., Konomoto, T., Horinouchi, T., Shono, A., et al. (2017b). Cryptic Exon Activation in SLC12A3 Gene Leads to Gitelman Syndrome. *J. Hum. Genet.* 62 (2), 335–337. doi:10.1038/jhg.2016.129
- Nozu, K., Vorechovsky, I., Kaito, H., Fu, X. J., Nakanishi, K., Hashimura, Y., et al. (2014). X-linked Alport Syndrome Caused by Splicing Mutations in COL4A5. *Cjasn* 9 (11), 1958–1964. doi:10.2215/CJN.04140414
- Ogino, W., Takeshima, Y., Nishiyama, A., Okizuka, Y., Yagi, M., Tsuneishi, S., et al. (2007). Mutation Analysis of the Ornithine Transcarbamylase (OTC) Gene in Five Japanese OTC Deficiency Patients Revealed Two Known and Three Novel Mutations Including a Deep Intronic Mutation. *Kobe J. Med. Sci.* 53 (5), 229–240.
- Ohura, T., Narisawa, K., Tada, K., and Iinuma, K. (1999). An 84bp Insertion Found in a Propionic Acidemia Patient Is Not a Disease-Causing Mutation but a Product of Cryptic mRNA. *J. Inherit. Metab. Dis.* 22 (5), 676–677. doi:10.1023/a:1005506819699
- Olsson, A., Lind, L., Thornell, L.-E., and Holmberg, M. (2008). Myopathy with Lactic Acidosis Is Linked to Chromosome 12q23.3-24.11 and Caused by an Intron Mutation in the ISCU Gene Resulting in a Splicing Defect. *Hum. Mol. Genet.* 17 (11), 1666–1672. doi:10.1093/hmg/ddn057
- Olthof, A. M., Hyatt, K. C., and Kanadia, R. N. (2019). Minor Intron Splicing Revisited: Identification of New Minor Intron-Containing Genes and Tissue-dependent Retention and Alternative Splicing of Minor Introns. *BMC Genomics* 20 (1), 686. doi:10.1186/s12864-019-6046-x
- Osborn, M. J., and Upadhyaya, M. (1999). Evaluation of the Protein Truncation Test and Mutation Detection in the NF1 Gene: Mutational Analysis of 15 Known and 40 Unknown Mutations. *Hum. Genet.* 105 (4), 327–332. doi:10.1007/s004399900135
- Oshima, J., Yu, C. E., Piussan, C., Klein, G., Jabkowski, J., Balci, S., et al. (1996). Homozygous and Compound Heterozygous Mutations at the Werner Syndrome Locus. *Hum. Mol. Genet.* 5 (12), 1909–1913. doi:10.1093/hmg/5.12.1909
- Pagani, F., Buratti, E., Stuani, C., Bendix, R., Dörk, T., and Baralle, F. E. (2002). A New Type of Mutation Causes a Splicing Defect in ATM. *Nat. Genet.* 30 (4), 426–429. doi:10.1038/ng858
- Pagliarini, V., Jolly, A., Bielli, P., Di Rosa, V., De la Grange, P., and Sette, C. (2020). De la Grange, P., and Sette, C. Sam68 binds Alu-rich introns in SMN and promotes pre-mRNA circularization. *Nucleic Acids Res.* 48 (2), 633–645. doi:10.1093/nar/gkz1117
- Parada, G. E., Munita, R., Cerda, C. A., and Gysling, K. (2014). A Comprehensive Survey of Non-canonical Splice Sites in the Human Transcriptome. *Nucleic Acids Res.* 42 (16), 10564–10578. doi:10.1093/nar/gku744
- Paradis, C., Cloutier, P., Shkreta, L., Toutant, J., Klarskov, K., and Chabot, B. (2007). hnRNP I/PTB Can Antagonize the Splicing Repressor Activity of SRp30c. *RNA* 13 (8), 1287–1300. doi:10.1261/rna.403607
- Park, V. M., and Pivnick, E. K. (1998). Neurofibromatosis Type 1 (NF1): a Protein Truncation Assay Yielding Identification of Mutations in 73% of Patients. *J. Med. Genet.* 35 (10), 813–820. doi:10.1136/jmg.35.10.813
- Paz, I., Kostı, I., Ares, M., Jr., Cline, M., and Mandel-Gutfreund, Y. (2014). RBPmap: a Web Server for Mapping Binding Sites of RNA-Binding Proteins. *Nucleic Acids Res.* 42 (Web Server issue), W361–W367. doi:10.1093/nar/gku406
- Pérez, B., Rincón, A., Jorge-Finnigan, A., Richard, E., Merinero, B., Ugarte, M., et al. (2009). Pseudoxon Exclusion by Antisense Therapy in Methylmalonic Aciduria (MMAuria). *Hum. Mutat.* 30 (12), 1676–1682. doi:10.1002/humu.21118
- Perrault, I., Hanein, S., Gérard, X., MOUNGUENGUE, N., Bouyakoub, R., Zarhrate, M., et al. (2021). Whole Locus Sequencing Identifies a Prevalent Founder Deep Intronic RPGRIP1 Pathologic Variant in the French Leber Congenital Amaurosis Cohort. *Basel: Genes*, 12. doi:10.3390/genes12020287
- Perrin, G., Morris, M. A., Antonarakis, S. E., Boltshauser, E., and Hutter, P. (1996). Two Novel Mutations Affecting mRNA Splicing of the Neurofibromatosis Type 1 (NF1) Gene. *Hum. Mutat.* 7 (2), 172–175. doi:10.1002/(SICI)1098-1004(1996)7:2<172::AID-HUMU15>3.0.CO;2-#
- Pezeshkpoor, B., Zimmer, N., Marquardt, N., Nanda, I., Haaf, T., Budde, U., et al. (2013). Deep Intronic 'mutations' Cause Hemophilia A: Application of Next Generation Sequencing in Patients without Detectable Mutation in F8 cDNA. *J. Thromb. Haemost.* 11 (9), 1679–1687. doi:10.1111/jth.12339
- Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M. C., and Caracausi, M. (2019). Human Protein-Coding Genes and Gene Feature Statistics in 2019. *BMC Res. Notes* 12 (1), 315. doi:10.1186/s13104-019-4343-8
- Piva, F., Giulietti, M., Burini, A. B., and Principato, G. (2012). SpliceAid 2: a Database of Human Splicing Factors Expression Data and RNA Target Motifs. *Hum. Mutat.* 33 (1), 81–85. doi:10.1002/humu.21609
- Pros, E., Fernández-Rodríguez, J., Canet, B., Benito, L., Sánchez, A., Benavides, A., et al. (2009). Antisense Therapeutics for Neurofibromatosis Type 1 Caused by Deep Intronic Mutations. *Hum. Mutat.* 30 (3), 454–462. doi:10.1002/humu.20933
- Pros, E., Gómez, C., Martín, T., Fábregas, P., Serra, E., and Lázaro, C. (2008). Nature and mRNA Effect of 282 differentNF1point Mutations: Focus on Splicing Alterations. *Hum. Mutat.* 29 (9), E173–E193. doi:10.1002/humu.20826
- Puresuren, J., Fukao, T., Hasegawa, Y., Fukuda, S., Kobayashi, H., and Yamaguchi, S. (2008). Study of Deep Intronic Sequence Exonization in a Japanese Neonate with a Mitochondrial Trifunctional Protein Deficiency. *Mol. Genet. Metab.* 95 (1-2), 46–51. doi:10.1016/j.ymgme.2008.06.013
- Qu, W., Cingolani, P., Zeeberg, B. R., and Ruden, D. M. (2017). A Bioinformatics-Based Alternative mRNA Splicing Code that May Explain Some Disease Mutations Is Conserved in Animals. *Front. Genet.* 8, 38. doi:10.3389/fgene.2017.00038
- Raponi, M., Upadhyaya, M., and Baralle, D. (2006). Functional Splicing Assay Shows a Pathogenic Intronic Mutation in Neurofibromatosis Type 1 (NF1) Due to Intronic Sequence Exonization. *Hum. Mutat.* 27 (3), 294–295. doi:10.1002/humu.9412
- Rathmann, M., Bunge, S., Beck, M., Kresse, H., Tylki-Szymanska, A., and Gal, A. (1996). Mucopolysaccharidosis Type II (Hunter Syndrome): Mutation "hot Spots" in the Iduronate-2-Sulfatase Gene. *Am. J. Hum. Genet.* 59 (6), 1202–1209.
- Ray, T. A., Cochran, K., Kozlowski, C., Wang, J., Alexander, G., Cady, M. A., et al. (2020). Comprehensive Identification of mRNA Isoforms Reveals the Diversity of Neural Cell-Surface Molecules with Roles in Retinal Development and Disease. *Nat. Commun.* 11 (1), 3328. doi:10.1038/s41467-020-17009-7
- Reeskamp, L. F., Hartgers, M. L., Peter, J., Dallinga-Thie, G. M., Zuurbier, L., Defesche, J. C., et al. (2018). A Deep Intronic Variant in LDLR in Familial Hypercholesterolemia. *Circ. Genom. Precis. Med.* 11 (12), e002385. doi:10.1161/CIRCGEN.118.002385
- Reeskamp, L. F., Balvers, M., Peter, J., van de Kerkhof, L., Klaaijns, L. N., Motazacker, M. M., et al. (2021). Intronic Variant Screening with Targeted Next-Generation Sequencing Reveals First Pseudoxon in LDLR in Familial Hypercholesterolemia. *Atherosclerosis* 321, 14–20. doi:10.1016/j.atherosclerosis.2021.02.003
- Rendu, J., Montjean, R., Coutton, C., Suri, M., Chicanne, G., Petiot, A., et al. (2017). Functional Characterization and Rescue of a Deep Intronic Mutation in OCLR1 Gene Responsible for Lowe Syndrome. *Hum. Mutat.* 38 (2), 152–159. doi:10.1002/humu.23139
- Rentas, S., Rathi, K. S., Kaur, M., Raman, P., Krantz, I. D., Sarmady, M., et al. (2020). Diagnosing Cornelia de Lange syndrome and related neurodevelopmental disorders using RNA sequencing. *Genet. Med.* 22 (5), 927–936. doi:10.1038/s41436-019-0741-5
- Riant, F., Odent, S., Cecillon, M., Pasquier, L., de Baracé, C., Carney, M. P., et al. (2014). Deep Intronic KRIT1 Mutation in a Family with Clinically Silent Multiple Cerebral Cavernous Malformations. *Clin. Genet.* 86 (6), 585–588. doi:10.1111/cge.12322
- Richard, N., Abeguilé, G., Coudray, N., Mittre, H., Gruchy, N., Andrieux, J., et al. (2012). A New Deletion Ablating NESP55 Causes Loss of Maternal Imprint of

- A/BGNASand Autosomal Dominant Pseudohypoparathyroidism Type Ib. *J. Clin. Endocrinol. Metab.* 97 (5), E863–E867. doi:10.1210/jc.2011-2804
- Rimessi, P., Fabris, M., Bovolenta, M., Bassi, E., Falzarano, S., Gualandi, F., et al. (2010). Antisense Modulation of Both Exonic and Intronic Splicing Motifs Induces Skipping of a DMD Pseudo-exon Responsible for X-Linked Dilated Cardiomyopathy. *Hum. Gene Ther.* 21 (9), 1137–1146. doi:10.1089/hum.2010.010
- Rincón, A., Aguado, C., Desviat, L. R., Sánchez-Alcudia, R., Ugarte, M., and Pérez, B. (2007). Propionic and Methylmalonic Acidemia: Antisense Therapeutics for Intronic Variations Causing Aberrantly Spliced Messenger RNA. *Am. J. Hum. Genet.* 81 (6), 1262–1270. doi:10.1086/522376
- Rius, R., Riley, L. G., Guo, Y., Menezes, M., Compton, A. G., Van Bergen, N. J., et al. (2019). Cryptic Intronic NBAS Variant Reveals the Genetic Basis of Recurrent Liver Failure in a Child. *Mol. Genet. Metab.* 126 (1), 77–82. doi:10.1016/j.ymgme.2018.12.002
- Rodríguez-Martín, C., Cidre, F., Fernández-Teijeiro, A., Gómez-Mariano, G., de la Vega, L., Ramos, P., et al. (2016). Familial Retinoblastoma Due to Intronic LINE-1 Insertion Causes Aberrant and Noncanonical mRNA Splicing of the RB1 Gene. *J. Hum. Genet.* 61 (5), 463–466. doi:10.1038/jhg.2015.173
- Rodríguez-Pascual, L., Coll, M. J., Vilageliu, L., and Grinberg, D. (2009). Antisense Oligonucleotide Treatment for a Pseudoxon-Generating Mutation in the NPC1 gene Causing Niemann-Pick Type C Disease. *Hum. Mutat.* 30 (11), E993–E1001. doi:10.1002/humu.21119
- Romano, M., Buratti, E., and Baralle, D. (2013). Role of Pseudoxons and Pseudointrons in Human Cancer. *Int. J. Cell Biol* 2013, 810572. doi:10.1155/2013/810572
- Rosenberg, J. B., Newman, P. J., Mosesson, M. W., Guillin, M. C., and Amrani, D. L. (1993). Paris I Dysfibrinogenemia: a point Mutation in Intron 8 Results in Insertion of a 15 Amino Acid Sequence in the Fibrinogen Gamma-Chain. *Thromb. Haemost.* 69 (3), 217–220. doi:10.1055/s-0038-1651583
- Rump, A., Rösen-Wolff, A., Gahr, M., Seidenberg, J., Roos, C., Walter, L., et al. (2006). A Splice-Supporting Intronic Mutation in the Last Bp Position of a Cryptic Exon within Intron 6 of the CYBB Gene Induces its Incorporation into the mRNA Causing Chronic Granulomatous Disease (CGD). *Gene* 371 (2), 174–181. doi:10.1016/j.gene.2005.11.036
- Rymen, D., Lindhout, M., Spanou, M., Ashrafzadeh, F., Benkel, I., Betzler, C., et al. (2020). Expanding the Clinical and Genetic Spectrum of CAD Deficiency: an Epileptic Encephalopathy Treatable with Uridine Supplementation. *Genet. Med.* 22 (10), 1589–1597. doi:10.1038/s41436-020-0933-z
- Sabbagh, A., Pasmant, E., Imbard, A., Luscan, A., Soares, M., Blanché, H., et al. (2013). NF1 Molecular Characterization and Neurofibromatosis Type I Genotype-Phenotype Correlation: the French Experience. *Hum. Mutat.* 34 (11), 1510–1518. doi:10.1002/humu.22392
- Saferali, A., Yun, J. H., Parker, M. M., Sakornsakolpat, P., Chase, R. P., Lamb, A., et al. (2019). Analysis of Genetically Driven Alternative Splicing Identifies FBXO38 as a Novel COPD Susceptibility Gene. *Plos Genet.* 15 (7), e1008229. doi:10.1371/journal.pgen.1008229
- Sakaguchi, N., and Suyama, M. (2021). In Silico identification of Pseudo-exon Activation Events in Personal Genome and Transcriptome Data. *RNA Biol.* 18 (3), 382–390. doi:10.1080/15476286.2020.1809195
- Sangermano, R., Garanto, A., Khan, M., Runhart, E. H., Bauwens, M., Bax, N. M., et al. (2019). Deep-intronic ABCA4 Variants Explain Missing Heritability in Stargardt Disease and Allow Correction of Splice Defects by Antisense Oligonucleotides. *Genet. Med.* 21 (8), 1751–1760. doi:10.1038/s41436-018-0414-9
- Santoro, A., Cannella, S., Trizzino, A., Bruno, G., De Fusco, C., Notarangelo, L. D., et al. (2008). Mutations Affecting mRNA Splicing Are the Most Common Molecular Defect in Patients with Familial Hemophagocytic Lymphohistiocytosis Type 3. *Haematologica* 93 (7), 1086–1090. doi:10.3324/haematol.12622
- Sargent, C. A., Kidd, A., Moore, S., Dean, J., Besley, G. T., and Affara, N. A. (2000). Five Cases of Isolated Glycerol Kinase Deficiency, Including Two Families: Failure to Find Genotype-phenotype Correlation. *J. Med. Genet.* 37 (6), 434–441. doi:10.1136/jmg.37.6.434
- Sawyer, S. L., Hartley, T., Dymont, D. A., Beaulieu, C. L., Schwartztruber, J., Smith, A., et al. (2016). Utility of Whole-exome Sequencing for Those Near the End of the Diagnostic Odyssey: Time to Address Gaps in Care. *Clin. Genet.* 89 (3), 275–284. doi:10.1111/cge.12654
- Schalk, A., Greff, G., Drouot, N., Obringer, C., Dollfus, H., Laugel, V., et al. (2018). Deep Intronic Variation in Splicing Regulatory Element of the ERCC8 Gene Associated with Severe but Long-Term Survival Cockayne Syndrome. *Eur. J. Hum. Genet.* 26 (4), 527–536. doi:10.1038/s41431-017-0009-y
- Schneider, A., Maas, S. M., Hennekam, R. C. M., and Hanauer, A. (2013). Identification of the First Deep Intronic Mutation in the RPS6KA3 Gene in a Patient with a Severe Form of Coffin-Lowry Syndrome. *Eur. J. Med. Genet.* 56 (3), 150–152. doi:10.1016/j.ejmg.2012.11.007
- Schollen, E., Keldermans, L., Foulquier, F., Briones, P., Chabas, A., Sánchez-Valverde, F., et al. (2007). Characterization of Two Unusual Truncating PMM2 Mutations in Two CDG-Ia Patients. *Mol. Genet. Metab.* 90 (4), 408–413. doi:10.1016/j.ymgme.2007.01.003
- Schulz, H. L., Grassmann, F., Kellner, U., Spital, G., Rütger, K., Jäggle, H., et al. (2017). Mutation Spectrum of the ABCA4 Gene in 335 Stargardt Disease Patients from a Multicenter German Cohort-Impact of Selected Deep Intronic Variants and Common SNPs. *Invest. Ophthalmol. Vis. Sci.* 58 (1), 394–403. doi:10.1167/iovs.16-19936
- Schüssler, S., Gerhalter, T., Abicht, A., Müller-Felber, W., Nagel, A., and Trollmann, R. (2020). Rare Intronic Mutation between Exon 62 and 63 (c.9225-285A>G) of the Dystrophin Gene Associated with Atypical BMD Phenotype. *Neuromuscul. Disord.* 30 (8), 680–684. doi:10.1016/j.nmd.2020.06.003
- Schwarze, U., Hata, R.-I., McKusick, V. A., Shinkai, H., Hoyme, H. E., Pyeritz, R. E., et al. (2004). Rare Autosomal Recessive Cardiac Valvular Form of Ehlers-Danlos Syndrome Results from Mutations in the COL1A2 Gene that Activate the Nonsense-Mediated RNA Decay Pathway. *Am. J. Hum. Genet.* 74 (5), 917–930. doi:10.1086/420794
- Sedláčková, J., Vondráček, P., Hermanová, M., Zámečník, J., Hrubá, Z., Haberlová, J., et al. (2009). Point Mutations in Czech DMD/BMD Patients and Their Phenotypic Outcome. *Neuromuscul. Disord.* 19 (11), 749–753. doi:10.1016/j.nmd.2009.08.011
- Sege-Peterson, K., Chambers, J., Page, T., Jones, O. W., and Nyhan, W. L. (1992). Characterization of Mutations in Phenotypic Variants of Hypoxanthine Phosphoribosyltransferase Deficiency. *Hum. Mol. Genet.* 1 (6), 427–432. doi:10.1093/hmg/1.6.427
- Seim, I., Lubik, A. A., Lehman, M. L., Tomlinson, N., Whiteside, E. J., Herington, A. C., et al. (2013). Cloning of a Novel Insulin-Regulated Ghrelin Transcript in Prostate Cancer. *J. Mol. Endocrinol.* 50 (2), 179–191. doi:10.1530/JME-12-0150
- Sela, N., Mersch, B., Hotz-Wagenblatt, A., and Ast, G. (2010). Characteristics of Transposable Element Exonization within Human and Mouse. *PLoS One* 5 (6), e10907. doi:10.1371/journal.pone.0010907
- Serra, T., Ars, E., Ravello, A., Sánchez, A., Puig, S., Rosenbaum, T., et al. (2001). Somatic NF1 Mutational Spectrum in Benign Neurofibromas: mRNA Splice Defects Are Common Among point Mutations. *Hum. Genet.* 108 (5), 416–429. doi:10.1007/s004390100514
- Shi, F., Yao, Y., Bin, Y., Zheng, C. H., and Xia, J. (2019). Computational Identification of Deleterious Synonymous Variants in Human Genomes Using a Feature-Based Approach. *BMC Med. Genomics* 12 (Suppl. 1), 12. doi:10.1186/s12920-018-0455-6
- Sibley, C. R., Emmett, W., Blázquez, L., Faro, A., Haberman, N., Briese, M., et al. (2015). Recursive Splicing in Long Vertebrate Genes. *Nature* 521 (7552), 371–375. doi:10.1038/nature14466
- Simon, M. T., Eftekharian, S. S., Stover, A. E., Osborne, A. F., Braffman, B. H., Chang, R. C., et al. (2019). Novel Mutations in the Mitochondrial Complex I Assembly Gene NDUFAF5 Reveal Heterogeneous Phenotypes. *Mol. Genet. Metab.* 126 (1), 53–63. doi:10.1016/j.ymgme.2018.11.001
- Sjarif, D. R., Hellerud, C., Amstel, J. K. P. v., Kleijer, W. J., Sperl, W., Lacombe, D., et al. (2004). Glycerol Kinase Deficiency: Residual Activity Explained by Reduced Transcription and Enzyme Conformation. *Eur. J. Hum. Genet.* 12 (6), 424–432. doi:10.1038/sj.ejhg.5201172
- Smith, B. F., Kornegay, J. N., and Duan, D. (2007). Independent Canine Models of Duchenne Muscular Dystrophy Due to Intronic Insertions of Repetitive DNA. *Mol. Ther.* 15. doi:10.1016/s1525-0016(16)44336-4
- Smith, M. J., Bowers, N. L., Banks, C., Coates-Brown, R., Morris, K. A., Ewans, L., et al. (2020). A Deep Intronic SMARCB1 Variant Associated with Schwannomatosis. *Clin. Genet.* 97 (2), 376–377. doi:10.1111/cge.13637
- Soliman, S. E., Racher, H., Lambourne, M., Matevski, D., MacDonald, H., and Gallie, B. (2018). A Novel Deep Intronic Low Penetrance RB1 Variant in a

- Retinoblastoma Family. *Ophthalmic Genet.* 39 (2), 288–290. doi:10.1080/13816810.2017.1393828
- Spena, S., Asselta, R., Platé, M., Castaman, G., Duga, S., and Tenchini, M. L. (2007). Pseudo-exon Activation Caused by a Deep-Intronic Mutation in the Fibrinogen  $\gamma$ -chain Gene as a Novel Mechanism for Congenital Afibrinogenemia. *Br. J. Haematol.* 139 (1), 128–132. doi:10.1111/j.1365-2141.2007.06758.x
- Spier, I., Horpaopan, S., Vogt, S., Uhlhaas, S., Morak, M., Stienen, D., et al. (2012). Deep intronic APC mutations Explain a Substantial Proportion of Patients with Familial or Early-Onset Adenomatous Polyposis. *Hum. Mutat.* 33 (7), 1045–1050. doi:10.1002/humu.22082
- Spits, C., De Rycke, M., Van Ranst, N., Joris, H., Verpoest, W., Lissens, W., et al. (2005). Preimplantation Genetic Diagnosis for Neurofibromatosis Type 1. *Mol. Hum. Reprod.* 11 (5), 381–387. doi:10.1093/molehr/gah170
- Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., et al. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinformatics* 54, 1–33. doi:10.1002/cpbi.5
- Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., et al. (2017). The Human Gene Mutation Database: towards a Comprehensive Repository of Inherited Mutation Data for Medical Research, Genetic Diagnosis and Next-Generation Sequencing Studies. *Hum. Genet.* 136 (6), 665–677. doi:10.1007/s00439-017-1779-6
- Stucki, M., Suormala, T., Fowler, B., Valle, D., and Baumgartner, M. R. (2009). Cryptic Exon Activation by Disruption of Exon Splice Enhancer. *J. Biol. Chem.* 284 (42), 28953–28957. doi:10.1074/jbc.M109.050674
- Stum, M., Davoine, C.-S., Vicart, S., Guillot-Noël, L., Topaloglu, H., Carod-Artal, F. J., et al. (2006). Spectrum of HSPG2 (Perlecan) Mutations in Patients with Schwartz-Jampel Syndrome. *Hum. Mutat.* 27 (11), 1082–1091. doi:10.1002/humu.20388
- Suga, Y., Tsuda, T., Nagai, M., Sakaguchi, Y., Jitsukawa, O., Yamamoto, M., et al. (2015). Lamellar Ichthyosis with Pseudoexon Activation in the Transglutaminase 1 Gene. *J. Dermatol.* 42 (6), 642–645. doi:10.1111/1346-8138.12846
- Suminaga, R., Takeshima, Y., Adachi, K., Yagi, M., Nakamura, H., and Matsuo, M. (2002). A Novel Cryptic Exon in Intron 3 of the Dystrophin Gene Was Incorporated into Dystrophin mRNA with a Single Nucleotide Deletion in Exon 5. *J. Hum. Genet.* 47 (4), 196–201. doi:10.1007/s100380200023
- Sutandy, F. X. R., Ebersberger, S., Huang, L., Busch, A., Bach, M., Kang, H.-S., et al. (2018). *In Vitro* iCLIP-based Modeling Uncovers How the Splicing Factor U2AF2 Relies on Regulation by Cofactors. *Genome Res.* 28 (5), 699–713. doi:10.1101/gr.229757.117
- Sutton, I. J., Last, J. I. K., Ritchie, S. J., Harrington, H. J., Byrd, P. J., and Taylor, A. M. R. (2004). Adult-onset Ataxia Telangiectasia Due to ATM 5762ins137 Mutation Homozygosity. *Ann. Neurol.* 55 (6), 891–895. doi:10.1002/ana.20139
- Suzuki, H., Aoki, Y., Kameyama, T., Saito, T., Masuda, S., Tanihata, J., et al. (2016). Endogenous Multiple Exon Skipping and Back-Splicing at the DMD Mutation Hotspot. *Int. J. Mol. Sci.* 17 (10), 1. doi:10.3390/ijms17101722
- Svaasand, E. K. (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.* 04 (03), 1. doi:10.4172/2161-1041.1000152
- Takahara, K., Schwarze, U., Imamura, Y., Hoffman, G. G., Toriello, H., Smith, L. T., et al. (2002). Order of Intron Removal Influences Multiple Splice Outcomes, Including a Two-Exon Skip, in a COL5A1 Acceptor-Site Mutation that Results in Abnormal Pro- $\alpha$ 1(V) N-Propeptides and Ehlers-Danlos Syndrome Type I. *Am. J. Hum. Genet.* 71 (3), 451–465. doi:10.1086/342099
- Takeshima, Y., Yagi, M., Okizuka, Y., Awano, H., Zhang, Z., Yamauchi, Y., et al. (2010). Mutation Spectrum of the Dystrophin Gene in 442 Duchenne/Becker Muscular Dystrophy Cases from One Japanese Referral center. *J. Hum. Genet.* 55 (6), 379–388. doi:10.1038/jhg.2010.49
- Tosch, V., Vasli, N., Kretz, C., Nicot, A.-S., Gasnier, C., Dondaine, N., et al. (2010). Novel Molecular Diagnostic Approaches for X-Linked Centronuclear (Myotubular) Myopathy Reveal Intronic Mutations. *Neuromuscul. Disord.* 20 (6), 375–381. doi:10.1016/j.nmd.2010.03.015
- Tozawa, Y., Abdrabou, S. S. M. A., Nogawa-Chida, N., Nishiuchi, R., Ishida, T., Suzuki, Y., et al. (2019). A Deep Intronic Mutation of c.1166-285 T > G in SLC46A1 Is Shared by Four Unrelated Japanese Patients with Hereditary Folate Malabsorption (HFM). *Clin. Immunol.* 208, 108256. doi:10.1016/j.clim.2019.108256
- Trabelsi, M., Beugnet, C., Deburgrave, N., Commere, V., Orhant, L., Leturcq, F., et al. (2014). When a Mid-intronic Variation of DMD Gene Creates an ESE Site. *Neuromuscul. Disord.* 24 (12), 1111–1117. doi:10.1016/j.nmd.2014.07.003
- Tran, V. K., Zhang, Z., Yagi, M., Nishiyama, A., Habara, Y., Takeshima, Y., et al. (2005). A Novel Cryptic Exon Identified in the 3' Region of Intron 2 of the Human Dystrophin Gene. *J. Hum. Genet.* 50 (8), 425–433. doi:10.1007/s10038-005-0272-6
- Treisman, R., Orkin, S. H., and Maniatis, T. (1983). Specific Transcription and RNA Splicing Defects in Five Cloned  $\beta$ -thalassaemia Genes. *Nature* 302 (5909), 591–596. doi:10.1038/302591a0
- Tsuruta, M., Mitsubuchi, H., Mardy, S., Miura, Y., Hayashida, Y., Kinugasa, A., et al. (1998). Molecular Basis of Intermittent maple Syrup Urine Disease: Novel Mutations in the E2 Gene of the Branched-Chain  $\alpha$ -keto Acid Dehydrogenase Complex. *J. Hum. Genet.* 43 (2), 91–100. doi:10.1007/s100380050047
- Tubeuf, H., Charbonnier, C., Soukarieh, O., Blavier, A., Lefebvre, A., Dauchel, H., et al. (2020). Large-scale Comparative Evaluation of User-friendly Tools for Predicting Variant-induced Alterations of Splicing Regulatory Elements. *Hum. Mutat.* 41 (10), 1811–1829. doi:10.1002/humu.24091
- Tuffery-Giraud, S., Saquet, C. L., Chambert, S., and Claustres, M. (2003). Pseudoexon Activation in the DMD Gene as a Novel Mechanism for Becker Muscular Dystrophy. *Hum. Mutat.* 21 (6), 608–614. doi:10.1002/humu.10214
- Tuffery-Giraud, S., Saquet, C., Thorel, D., Disset, A., Rivier, F., Malcolm, S., et al. (2005). Mutation Spectrum Leading to an Attenuated Phenotype in Dystrophinopathies. *Eur. J. Hum. Genet.* 13 (12), 1254–1260. doi:10.1038/sj.ejhg.5201478
- Turunen, J. J., Niemelä, E. H., Verma, B., and Frilander, M. J. (2013). The Significant Other: Splicing by the Minor Spliceosome. *Wiley Interdiscip. Rev. RNA* 4 (1), 61–76. doi:10.1002/wrna.1141
- Vaché, C., Besnard, T., le Berre, P., García-García, G., Baux, D., Larrieu, L., et al. (2012). Usher Syndrome Type 2 Caused by Activation of an USH2A Pseudoexon: Implications for Diagnosis and Therapy. *Hum. Mutat.* 33 (1), 104–108. doi:10.1002/humu.21634
- Valcárcel, J., Gaur, R. K., Singh, R., and Green, M. R. (1996). Interaction of U2AF 65 RS Region with Pre-mRNA of Branch Point and Promotion Base Pairing with U2 snRNA. *Science* 273 (5282), 1706–1709. doi:10.1126/science.273.5282.1706
- Valdmanis, P. N., Belzil, V. V., Lee, J., Dion, P. A., St-Onge, J., Hince, P., et al. (2009). A Mutation that Creates a Pseudoexon in SOD1 Causes Familial ALS. *Ann. Hum. Genet.* 73 (Pt 6), 652–657. doi:10.1111/j.1469-1809.2009.00546.x
- Valero, M. C., Martín, Y., Hernández-Imaz, E., Marina Hernández, A., Meleán, G., Valero, A. M., et al. (2011). A Highly Sensitive Genetic Protocol to Detect NF1 Mutations. *J. Mol. Diagn.* 13 (2), 113–122. doi:10.1016/j.jmoldx.2010.09.002
- van den Hurk, J. A. J. M., van de Pol, D. J. R., Wissinger, B., van Driel, M. A., Hoefsloot, L. H., de Wijs, I. J., et al. (2003). Novel Types of Mutation in the Choroideremia (CHM) Gene: a Full-Length L1 Insertion and an Intronic Mutation Activating a Cryptic Exon. *Hum. Genet.* 113 (3), 268–275. doi:10.1007/s00439-003-0970-0
- van der Klift, H. M., Tops, C. M., Hes, F. J., Devilee, P., and Wijnen, J. T. (2012). Insertion of an SVA Element, a Nonautonomous Retrotransposon, in PMS2 intron 7 as a Novel Cause of Lynch Syndrome. *Hum. Mutat.* 33 (7), 1051–1055. doi:10.1002/humu.22092
- van der Wal, E., Bergsma, A. J., van Gestel, T. J. M., in 't Groen, S. L. M., Zaehres, H., Araújo-Bravo, M. J., et al. (2017). GAA Deficiency in Pompe Disease Is Alleviated by Exon Inclusion in iPSC-Derived Skeletal Muscle Cells. *Mol. Ther. - Nucleic Acids* 7, 101–115. doi:10.1016/j.omtn.2017.03.002
- van Kuilenburg, A. B. P., Meijer, J., Mul, A. N. P. M., Meinsma, R., Schmid, V., Dobritzsch, D., et al. (2010). Intragenic Deletions and a Deep Intronic Mutation Affecting Pre-mRNA Splicing in the Dihydropyrimidine Dehydrogenase Gene as Novel Mechanisms Causing 5-fluorouracil Toxicity. *Hum. Genet.* 128 (5), 529–538. doi:10.1007/s00439-010-0879-3
- Vandenbroucke, I., Callens, T., De Paepe, A., and Messiaen, L. (2002). Complex Splicing Pattern Generates Great Diversity in Human NF1 Transcripts. *BMC Genomics* 3, 13. doi:10.1186/1471-2164-3-13
- Varon, R., Gooding, R., Steglich, C., Marns, L., Tang, H., Angelicheva, D., et al. (2003). Partial Deficiency of the C-Terminal-Domain Phosphatase of RNA



- Polymerase II Is Associated with Congenital Cataracts Facial Dysmorphism Neuropathy Syndrome. *Nat. Genet.* 35 (2), 185–189. doi:10.1038/ng1243
- Vatanavicharn, N., Champattanachai, V., Liammongkolkul, S., Sawangareetrakul, P., Keeratchamroen, S., Ketudat Cairns, J. R., et al. (2012). Clinical and Molecular Findings in Thai Patients with Isolated Methylmalonic Acidemia. *Mol. Genet. Metab.* 106 (4), 424–429. doi:10.1016/j.ymgme.2012.05.012
- Vaz-Drago, R., Custódio, N., and Carmo-Fonseca, M. (2017). Deep Intronic Mutations and Human Disease. *Hum. Genet.* 136 (9), 1093–1111. doi:10.1007/s00439-017-1809-4
- Verdura, E., Schlüter, A., Fernández-Eulate, G., Ramos-Martin, R., Zulaica, M., Planas-Serra, L., et al. (2020). A Deep Intronic Splice Variant Advises Reexamination of Presumably Dominant SPG7 Cases. *Ann. Clin. Transl. Neurol.* 7 (1), 105–111. doi:10.1002/acn3.50967
- Verrier, F., Dubois d'Enghien, C., Gauthier-Villars, M., Bonadona, V., Faure-Contier, C., Dijoud, F., et al. (2018). Multiple DICER1-related Lesions Associated with a Germline Deep Intronic Mutation. *Pediatr. Blood Cancer* 65 (6), e27005. doi:10.1002/pbc.27005
- Vetrini, F., Tammaro, R., Bondanza, S., Surace, E. M., Auricchio, A., De Luca, M., et al. (2006). Aberrant Splicing in the Ocular Albinism Type 1 Gene (OAI/GPR143) Is Corrected *In Vitro* by Morpholino Antisense Oligonucleotides. *Hum. Mutat.* 27 (5), 420–426. doi:10.1002/humu.20303
- Voith von Voithenberg, L., Sánchez-Rico, C., Kang, H.-S., Madl, T., Zanier, K., Barth, A., et al. (2016). Recognition of the 3' Splice Site RNA by the U2AF Heterodimer Involves a Dynamic Population Shift. *Proc. Natl. Acad. Sci. USA* 113 (46), E7169–E7175. doi:10.1073/pnas.1605873113
- Vorechovsky, I. (2010). Transposable Elements in Disease-Associated Cryptic Exons. *Hum. Genet.* 127 (2), 135–154. doi:10.1007/s00439-009-0752-4
- Waddell, L. B., Bryen, S. J., Cummings, B. B., Bournazos, A., Evesson, F. J., Joshi, H., et al. (2021). WGS and RNA Studies Diagnose Noncoding DMD Variants in Males with High Creatine Kinase. *Neurol. Genet.* 7 (1), e554. doi:10.1212/NXG.0000000000000554
- Walker, S., Lamoureux, S., Khan, T., Joynt, A. C. M., Bradley, M., Branson, H. M., et al. (2021). Genome Sequencing for Detection of Pathogenic Deep Intronic Variation: A Clinical Case Report Illustrating Opportunities and Challenges. *Am. J. Med. Genet. A.* 185 (10), 3129–3135. doi:10.1002/ajmg.a.62389
- Walsh, T., Casadei, S., Munson, K. M., Eng, M., Mandell, J. B., Gulsuner, S., et al. (2020). CRISPR-Cas9/long-read Sequencing Approach to Identify Cryptic Mutations in BRCA1 and Other Tumour Suppressor Genes. *J. Med. Genet.* 58 (12), 850–852. doi:10.1136/jmedgenet-2020-107320
- Wan, Y., Anastasakis, D. G., Rodriguez, J., Palangat, M., Gudla, P., Zaki, G., et al. (2021). Dynamic Imaging of Nascent RNA Reveals General Principles of Transcription Dynamics and Stochastic Splice Site Selection. *Cell* 184 (11), 2878–2895. e2820. doi:10.1016/j.cell.2021.04.012
- Wang, X., Zhang, Y., Ding, J., and Wang, F. (2021). mRNA Analysis Identifies Deep Intronic Variants Causing Alport Syndrome and Overcomes the Problem of Negative Results of Exome Sequencing. *Sci. Rep.* 11 (1), 18097. doi:10.1038/s41598-021-97414-0
- Webb, T. R., Parfitt, D. A., Gardner, J. C., Martinez, A., Bevilacqua, D., Davidson, A. E., et al. (2012). Deep Intronic Mutation in OFD1, Identified by Targeted Genomic Next-Generation Sequencing, Causes a Severe Form of X-Linked Retinitis Pigmentosa (RP23). *Hum. Mol. Genet.* 21 (16), 3647–3654. doi:10.1093/hmg/dd5194
- Weisschuh, N., Mazzola, P., Bertrand, M., Haack, T. B., Wissinger, B., Kohl, S., et al. (2021). Clinical Characteristics of POC1B-Associated Retinopathy and Assignment of Pathogenicity to Novel Deep Intronic and Non-canonical Splice Site Variants. *Int. J. Mol. Sci.* 22 (10), 1. doi:10.3390/ijms22105396
- Weisschuh, N., Sturm, M., Baumann, B., Audo, I., Ayuso, C., Bocquet, B., et al. (2020). Deep-intronic Variants in CNGB3 Cause Achromatopsia by Pseudoxon Activation. *Hum. Mutat.* 41 (1), 255–264. doi:10.1002/humu.23920
- Welander, J., Larsson, C., Bäckdahl, M., Hareni, N., Sivilér, T., Brauckhoff, M., et al. (2012). Integrative Genomics Reveals Frequent Somatic NF1 Mutations in Sporadic Pheochromocytomas. *Hum. Mol. Genet.* 21 (26), 5406–5416. doi:10.1093/hmg/dd5402
- Whately, S. D., Mason, N. G., Rhodes, J. M., Stewart, M. F., Reed, P., Crowley, V., et al. (2013). Pseudoxon Activation in the HMBS Gene as a Cause of the Nonerythroid Form of Acute Intermittent Porphyria. *Clin. Chem.* 59 (7), 1123–1125. doi:10.1373/clinchem.2012.199117
- Will, K., Dörk, T., Stuhmann, M., Meitinger, T., Bertele-Harms, R., Tümmler, B., et al. (1994). A Novel Exon in the Cystic Fibrosis Transmembrane Conductance Regulator Gene Activated by the Nonsense Mutation E92X in Airway Epithelial Cells of Patients with Cystic Fibrosis. *J. Clin. Invest.* 93 (4), 1852–1859. doi:10.1172/JCI117172
- Wilson, A., Leclerc, D., Rosenblatt, D. S., and Gravel, R. A. (1999). Molecular Basis for Methionine Synthase Reductase Deficiency in Patients Belonging to the cblE Complementation Group of Disorders in Folate/cobalamin Metabolism. *Hum. Mol. Genet.* 8 (11), 2009–2016. doi:10.1093/hmg/8.11.2009
- Wilson, A., Leclerc, D., Saberi, F., Campeau, E., Hwang, H. Y., Shane, B., et al. (1998). Functionally Null Mutations in Patients with the cblG-Variant Form of Methionine Synthase Deficiency. *Am. J. Hum. Genet.* 63 (2), 409–414. doi:10.1086/301976
- Wilund, K. R., Yi, M., Campagna, F., Arca, M., Zuliani, G., Fellin, R., et al. (2002). Molecular Mechanisms of Autosomal Recessive Hypercholesterolemia. *Hum. Mol. Genet.* 11 (24), 3019–3030. doi:10.1093/hmg/11.24.3019
- Wimmer, K., Callens, T., Wernstedt, A., and Messiaen, L. (2011). The NF1 Gene Contains Hotspots for L1 Endonuclease-dependent De Novo Insertion. *Plos Genet.* 7 (11), e1002371. doi:10.1371/journal.pgen.1002371
- Xie, Z., Sun, C., Liu, Y., Yu, M., Zheng, Y., Meng, L., et al. (2020). Practical Approach to the Genetic Diagnosis of Unsolved Dystrophinopathies: a Stepwise Strategy in the Genomic Era. *J. Med. Genet.* 58 (11), 743–751. doi:10.1136/jmedgenet-2020-107113
- Xu, Y., Song, T., Li, Y., Guo, F., Jin, X., Cheng, L., et al. (2020). Identification of Two Novel Insertion Abnormal Transcripts in Two Chinese Families Affected with Dystrophinopathy. *J. Clin. Lab. Anal.* 34 (4), e23142. doi:10.1002/jcla.23142
- Yagi, M., Takeshima, Y., Wada, H., Nakamura, H., and Matsuo, M. (2003). Two Alternative Exons Can Result from Activation of the Cryptic Splice Acceptor Site Deep within Intron 2 of the Dystrophin Gene in a Patient with as yet Asymptomatic Dystrophinopathy. *Hum. Genet.* 112 (2), 164–170. doi:10.1007/s00439-002-0854-8
- Yamaguchi, H., Fujimoto, T., Nakamura, S., Ohmura, K., Mimori, T., Matsuda, F., et al. (2010). Aberrant Splicing of the Milk Fat Globule-EGF Factor 8 (MFG-E8) Gene in Human Systemic Lupus Erythematosus. *Eur. J. Immunol.* 40 (6), 1778–1785. doi:10.1002/eji.200940096
- Yamano, S., Nhamburo, P. T., Aoyama, T., Meyer, U. A., Inaba, T., Kalow, W., et al. (1989). cDNA Cloning and Sequence and cDNA-Directed Expression of Human P450 IIB1: Identification of a normal and Two Variant cDNAs Derived from the CYP2B Locus on Chromosome 19 and Differential Expression of the IIB mRNAs in Human Liver. *Biochemistry* 28 (18), 7340–7348. doi:10.1021/bi00444a029
- Yasmeen, S., Lund, K., De Paepe, A., De Bie, S., Heiberg, A., Silva, J., et al. (2014). Occipital Horn Syndrome and Classical Menkes Syndrome Caused by Deep Intronic Mutations, Leading to the Activation of ATP7A Pseudo-exon. *Eur. J. Hum. Genet.* 22 (4), 517–521. doi:10.1038/ejhg.2013.191
- Yeo, G., and Burge, C. B. (2004). Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J. Comput. Biol.* 11 (2–3), 377–394. doi:10.1089/1066527041410418
- Yuste-Checa, P., Medrano, C., Gámez, A., Desviat, L. R., Matthijs, G., Ugarte, M., et al. (2015). Antisense-mediated Therapeutic Pseudoxon Skipping in TMEM165-CDG. *Clin. Genet.* 87 (1), 42–48. doi:10.1111/cge.12402
- Zaum, A.-K., Stüve, B., Gehrig, A., Köbel, H., Schara, U., Kress, W., et al. (2017). Deep Intronic Variants Introduce DMD Pseudoxon in Patient with Muscular Dystrophy. *Neuromuscul. Disord.* 27 (7), 631–634. doi:10.1016/j.nmd.2017.04.003
- Zenteno, J. C., García-Montano, L. A., Cruz-Aguilar, M., Ronquillo, J., Rodas-Serrano, A., Aguilar-Castul, L., et al. (2020). Extensive Genic and Allelic Heterogeneity Underlying Inherited Retinal Dystrophies in Mexican Patients Molecularly Analyzed by Next-Generation Sequencing. *Mol. Genet. Genomic Med.* 8 (1), 1. doi:10.1002/mgg3.1044
- Zhang, J., Sun, X., Qian, Y., LaDuca, J. P., and Maquat, L. E. (1998). At Least One Intron Is Required for the Nonsense-Mediated Decay of Triosephosphate Isomerase mRNA: a Possible Link between Nuclear Splicing and Cytoplasmic Translation. *Mol. Cell Biol.* 18 (9), 5272–5283. doi:10.1128/MCB.18.9.5272

- Zhang, K., Nowak, I., Rushlow, D., Gallie, B. L., and Lohmann, D. R. (2008). Patterns of Missplicing Caused by RB1 Gene Mutations in Patients with Retinoblastoma and Association with Phenotypic Expression. *Hum. Mutat.* 29 (4), 475–484. doi:10.1002/humu.20664
- Zhang, X.-O., Fu, Y., Mou, H., Xue, W., and Weng, Z. (2018). The Temporal Landscape of Recursive Splicing during Pol II Transcription Elongation in Human Cells. *Plos Genet.* 14 (8), e1007579. doi:10.1371/journal.pgen.1007579
- Zhang, Y., Qian, J., Gu, C., and Yang, Y. (2021). Alternative Splicing and Cancer: a Systematic Review. *Sig Transduct Target. Ther.* 6 (1), 78. doi:10.1038/s41392-021-00486-7
- Zhang, Z., Habara, Y., Nishiyama, A., Oyazato, Y., Yagi, M., Takeshima, Y., et al. (2007). Identification of Seven Novel Cryptic Exons Embedded in the Dystrophin Gene and Characterization of 14 Cryptic Dystrophin Exons. *J. Hum. Genet.* 52 (7), 607–617. doi:10.1007/s10038-007-0163-0
- Zhu, F., Zhang, F., Hu, L., Liu, H., and Li, Y. (2021). Integrated Genome and Transcriptome Sequencing to Solve a Neuromuscular Puzzle: Miyoshi Muscular Dystrophy and Early Onset Primary Dystonia in Siblings of the Same Family. *Front. Genet.* 12, 672906. doi:10.3389/fgene.2021.672906
- Zou, G., Zhang, T., Cheng, X., Igelman, A. D., Wang, J., Qian, X., et al. (2021). Noncoding Mutation in RRGRI1 Contributes to Inherited Retinal Degenerations. *Mol. Vis.* 27, 95–106.
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.
- Copyright © 2022 Keegan, Wilton and Fletcher. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



## **Chapter 5**

**A spotter's guide to SNPTic  
exons: The common splice  
variants underlying some SNP-  
phenotype correlations**

## 5.1 Preamble

During the latter stages of data curation for the pseudoexons catalogue (Chapter 4), we re-examined the supporting evidence for a small minority of PEs with apparent non-canonical splice motifs, i.e., non-AG-GY flanking dinucleotides. Although a thorough inspection of the data revealed that some of these were indeed *bona fide* non-canonical splice events, we noted that one such PE in the *GHRL* gene (Seim *et al.* 2013) was adjacent to a common single nucleotide polymorphism (SNP) in the exact position necessary to convert its non-canonical AA acceptor dinucleotide to a canonical AG. Given the exceptional rarity of non-AG-GY splice motifs that are processed by the U2 spliceosome (Olthof *et al.* 2019 and Piovesan *et al.* 2019), it seemed that a more likely explanation for splicing of this PE was that the probands were unidentified carriers of the G-allele of the SNP.

Intriguingly, when we researched this *GHRL* SNP we found there were already published reports of its correlations with disease risk (Ando *et al.* 2006; Pabalan *et al.* 2014) but noted that these reports made no reference to differential splicing of the adjacent pseudoexon. This suggested that SNP-modulated splicing of this pseudoexon could be the missing link between the SNP and its associated phenotypes, leading us to question how many other such “SNPtic” exons might exist within our dataset.

A subsequent investigation of our dataset discovered six known SNPtic exons that were well-characterised in the original reports, five probable SNPtic exons that were not formally linked to the population phenotypes of their SNPs, and an additional five possible SNPtic exons with SNP-effects that were predicted but not yet empirically verified.

Although we had originally intended this investigation of SNPtic exons to be one component of many in our larger study of pseudoexons, it gradually became clear that this topic was unique in its implications for population genetics and required a much more thorough case-by-case discussion than did other pseudoexon phenomena. We therefore chose to present these findings separately as a methods

report, framing our set of known and putative SNPtic exons as supporting examples for a search method that could easily and fruitfully be applied by others.

We expect that SNPtic exons will soon prove to be powerful population ‘stratifiers’ in clinical trials of new genetic therapies, as such stratification can provide a mechanistic explanation for some of the variance in gene expression and therapeutic response seen in trial cohorts.

## 5.2 Citation

Keegan, N.P., and Fletcher, S. (2021). A spotter's guide to SNPtic exons: The common splice variants underlying some SNP-phenotype correlations. *Mol Genet Genomic Med*, e1840. doi: 10.1002/mgg3.1840.

## METHOD

# A spotter's guide to SNPtic exons: The common splice variants underlying some SNP–phenotype correlations

Niall Patrick Keegan<sup>1,2,3</sup>  | Sue Fletcher<sup>1,2,4</sup> <sup>1</sup>Murdoch University, Murdoch, Western Australia, Australia<sup>2</sup>Centre for Molecular Medicine and Innovative Therapeutics, Perth, Western Australia, Australia<sup>3</sup>Perron Institute, Perth, Western Australia, Australia<sup>4</sup>University of Western Australia, Perth, Western Australia, Australia**Correspondence**Niall Patrick Keegan, Murdoch University, Murdoch, WA, Australia.  
Email: n.keegan@murdoch.edu.au**Funding information**

Funding provided by the Australian Commonwealth Government Research Training Program Scholarship.

**Abstract****Background:** Cryptic exons are typically characterised as deleterious splicing aberrations caused by deep intronic mutations. However, low-level splicing of cryptic exons is sometimes observed in the absence of any pathogenic mutation. Five recent reports have described how low-level splicing of cryptic exons can be modulated by common single-nucleotide polymorphisms (SNPs), resulting in phenotypic differences amongst different genotypes.**Methods:** We sought to investigate whether additional ‘SNPtic’ exons may exist, and whether these could provide an explanatory mechanism for some of the genotype–phenotype correlations revealed by genome-wide association studies. We thoroughly searched the literature for reported cryptic exons, cross-referenced their genomic coordinates against the *dbSNP* database of common SNPs, then screened out SNPs with no reported phenotype associations.**Results:** This method discovered five probable SNPtic exons in the genes *APC*, *FGB*, *GHRL*, *MYPBC3* and *OTC*. For four of these five exons, we observed that the phenotype associated with the SNP was compatible with the predicted splicing effect of the nucleotide change, whilst the fifth (in *GHRL*) likely had a more complex splice-switching effect.**Conclusion:** Application of our search methods could augment the knowledge value of future cryptic exon reports and aid in generating better hypotheses for genome-wide association studies.**KEYWORDS**

cryptic exon, genome-wide association study, RNA splicing, single-nucleotide polymorphism

## 1 | INTRODUCTION

Since the first cryptic exon (CE), or pseudoexon, was discovered in humans in 1983 (Dobkin et al., 1983), there have been hundreds more reported examples of this

splicing phenomenon. Most CEs are detected as the result of pathogenic deep intronic mutations that directly enhance the exon-like characteristics of intron tracts not otherwise retained in mature transcripts. Because the sequences of most CEs have not evolved to preserve the

[Correction added on November 22, 2021, after first Online publication: Table 1 has been converted to Table S1.]

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2021 The Authors. *Molecular Genetics & Genomic Medicine* published by Wiley Periodicals LLC

open reading frame, CE inclusion typically introduces premature stop codons or frameshifts to the affected mRNA, resulting in non-functional transcripts and/or nonsense-mediated decay (NMD). The most common cause of CE pathogenesis is a single-nucleotide variant (SNV) in the CE or its flanking splice site motifs, usually at one of the four bases of the CE terminal dinucleotides (Romano et al., 2013; Vaz-Drago et al., 2017; Vorechovsky, 2010). Mutations that alter the binding motifs of other local splicing factors are also observed, but less frequently (Canson et al., 2020; Keegan, 2020; Tubeuf et al., 2020).

Some reports of CE pathogenesis have noted low-level CE splicing in cells that do not carry a pathogenic mutation (Braun et al., 2013; Druhan et al., 2020; Will et al., 1993). In these cases, it appears that the pathogenic mutations are not 'creating' or 'activating' a CE, but rather, dramatically enhancing the inclusion of a CE that already exists. This begs the question of why these low-frequency CEs have not been eliminated from the genome by selective pressure. Do they persist as subtle but useful regulators of gene expression, or are they merely tolerated as an unavoidable side effect of organismal complexity?

Recent research indicates that at least some low-spliced CEs are indeed functional and may be better described as 'poison exons', a spliceosomal tactic for committing unneeded transcripts to nonsense-mediated decay and thus avoiding excess translation of the encoded protein (Anko et al., 2012). However, at the time of writing, only a few poison exons have been formally characterised in a limited range of genes (Carvill & Mefford, 2020; Thomas et al., 2020).

Regardless of whether a CE serves a functional role, it can be speculated that any change in its splicing characteristics will produce a phenotypic change in corresponding directionality and severity. At one end of this spectrum are those pathogenic mutations that greatly increase CE inclusion and produce an easily observable disease phenotype; whilst at the other end are so-called 'near-neutral' variants, so slight in their effect that they would defy characterisation in a single individual. It is only when these subtle variants occur frequently in a population that statistical analysis can measure the differences amongst the carriers of each variant, and thus separate the signal from the noise (Figure 1).

Genome-wide association studies (GWASes) have used this approach to identify thousands of correlations between common genetic variants and particular phenotypes or disease risk profiles; and most germline variants examined by these studies are single-nucleotide polymorphisms (SNPs). A strict definition of the term 'SNP' refers only to germline one-nucleotide substitutions, but conventional usage of the term, which we have adopted in this report, also encompasses small deletions and insertions, and typically only refers to variants observed in at least 1% of the haploid sample population. However, despite

the great power of GWASes to discover SNP-phenotype correlations, deriving the aetiologies underlying these correlations has proved a much more challenging and laborious task (Cano-Gamez & Trynka, 2020).

Evidence indicates that the mechanism driving at least some SNP-phenotype associations is SNP-driven modulation of cryptic splicing (Stein et al., 2015). However, the effect of SNPs on the splicing of cryptic *exons* specifically is underexamined in the literature. This led us to investigate whether there may be published reports describing the components of CE-SNP pairs but not conceptually connecting them as components of a single phenomenon.

The online resource *dbSNP* (Sherry et al., 2001), accessible both directly and via the UCSC Genome Browser (Kent et al., 2002), collates the locations and frequencies of millions of SNPs across the human genome, whilst *GWAS Central* (Beck et al., 2020) serves as an international repository for GWAS data. Both are freely accessible and easily searchable. Unfortunately, however, to our knowledge an equally comprehensive database of cryptic exons does not exist. We believe that this is largely due to the sporadic nature of cryptic exon discovery over the last four decades resulting in a lack of consistency in how they are reported.

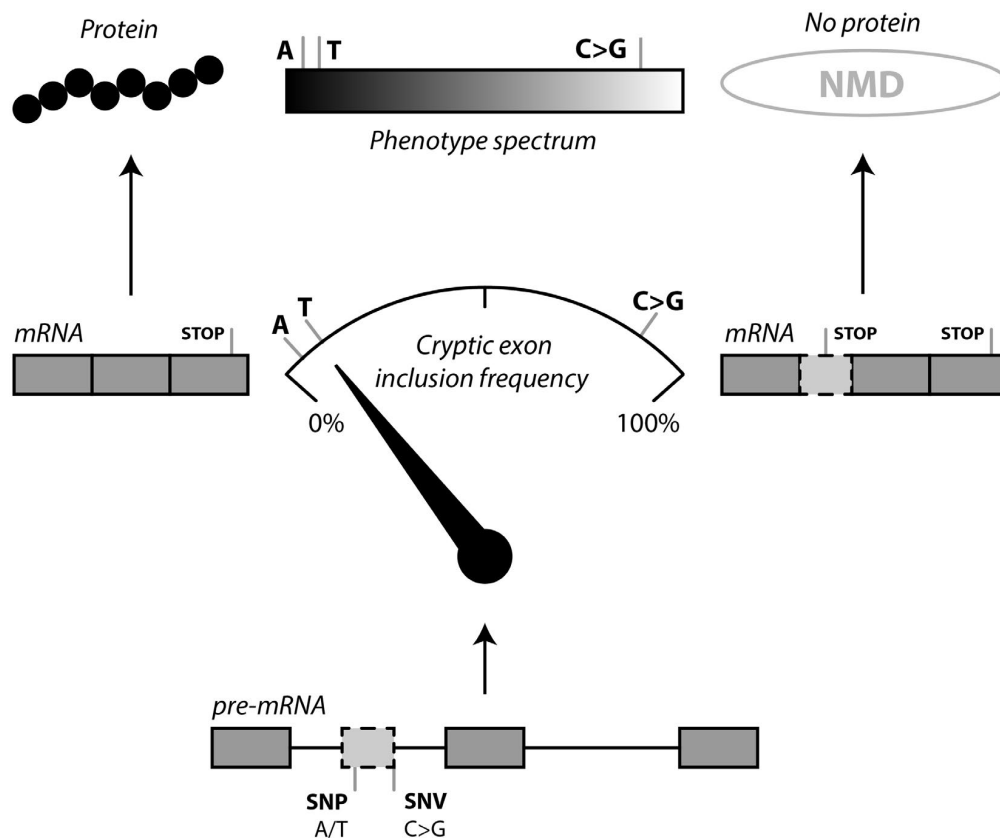
In this report, we outline our approach to discovering examples of cryptic exons likely to be subject to SNP-associated differential splicing. In the interest of clarity, we will henceforth refer to this SNP-associated differential splicing, and the cryptic exons it purportedly affects, as 'SNPtic splicing' and 'SNPtic exons', respectively; this novel term 'SNPtic' (pronounced SNIP-tick) being a portmanteau of 'SNP' and 'cryptic'.

We believe that this technique will be of great use to researchers reporting new CEs in the future, who may find it substantially adds to the information content of their publications.

## 2 | MATERIALS AND METHODS

Our strategy for search and analysis is outlined below. Henceforth, all references to the 'effect' of a SNP refer to the effect of the minor (least common) allele relative to that of the major allele.

**Cryptic exon discovery.** Using Google Scholar, we performed a thorough literature search for reported examples of cryptic exons, using the search terms 'pseudoxon', 'cryptic exon' and 'deep intronic mutation'. For each resulting report, we used the details provided therein to derive the full genomic sequence, coordinates (GRCh38/hg38) and strand identity (+ or -) of each described cryptic exon, plus 20 nucleotides of flanking sequence at each end.



**FIGURE 1** A general model of SNptic exon splicing. A cryptic exon, or CE (dashed-line box) is included in mature transcripts at frequencies that vary depending on the genotype of the carrier. Because the CE encodes a premature stop codon more than 55 nt from the final splice junction, mature transcripts that include the CE are targeted for nonsense-mediated decay (NMD, grey circle) and are not translated. If a patient carries an SNV (C>G) that greatly increases CE inclusion, NMD predominates and little protein is translated, resulting in a rare but distinct disease phenotype. Conversely, through similar mechanisms a common SNP (A>T) with a weak effect on CE splicing leads to a common but indistinct phenotype, which may only be measurable with a sufficiently powered genome-wide association study

**Cross-search for common SNPs.** The final dataset of cryptic exon coordinates was compiled into a BED file, uploaded to the UCSC Genome Browser data integrator as a custom track and cross-searched against the track 'dbSNP 153', sub-track 'Common dbSNP 153'. Results from the cross-search were exported into Microsoft Excel for further analysis.

**Alternative to steps 1. & 2.** Instead of following the methods described above, researchers investigating small numbers of cryptic exons may find it is easier to simply enable the 'Common dbSNP 153' track on the UCSC Genome Browser, and then perform serial BLAT searches of each sequence of interest whilst manually annotating the rsIDs of any coinciding SNPs.

**Filtering.** Because flanking AG-GY terminal dinucleotides appear to be almost essential for U2-type splicing,

which predominates in the human transcriptome (Parada et al., 2014), we manually excluded any cryptic exon (and associated SNPs) that did not bear these dinucleotides in at least one SNP allele.

**Search for GWAS phenotypes.** We searched the rsID of each remaining SNP both in *GWAS Central* and in Google. For each *GWAS Central* search, we considered as 'hits' only those studies that reported a  $p$ -value with baseline significance ( $p \leq .05$ ) and had a defined effect size for the searched SNP. This latter requirement was to ensure that the correct allele of the SNP was assigned to the correct phenotype. For Google results, we considered as 'hits' only those results that originated from peer-reviewed literature in which the SNP was described as being of probable significance to a particular phenotype.

**Prediction of SNP effect.** The method of analysis for each SNP depended on its position relative to the CE splice sites.

- a. **SNPs at or between positions –20 to +3 of the CE acceptor site, or at or between positions –3 to +6 of the donor site**, were analysed for their effect on the Maximum Entropy (MaxEnt) score of the corresponding motif using the *MaxEntScan* web utility (Yeo & Burge, 2004). MaxEnt was chosen based on its well-established efficacy—at the time of writing, Yeo & Burge's, 2004 report has been cited over 1600 times. However, there are numerous other splice motif scoring methods that perform comparably well (see Jian et al., 2014 for review). Most SNPs were classed as either 'More inclusion' if they increased a MaxEnt score or as 'Less inclusion' if they decreased a MaxEnt score. In cases where a SNP was predicted to alter the splicing ratio between two isoforms of a CE, it was classed as 'Splice-switching'.
- b. **For other SNPs inside the CE**, their effects were predicted using *HExoSplice* (Tubeuf et al., 2020), with a positive score indicating higher inclusion and a negative score indicating lower inclusion.
- c. **For all other SNPs outside the CE**, cryptic exon sequences corresponding to both SNP alleles were comparatively analysed via the *SpliceAid 2* web utility (Piva et al., 2012). *SpliceAid 2* automatically designates detected motifs as 'enhancers' or 'silencers' of exon inclusion, but in some cases the true effect of an RNA-bound splice factor is dependent on its orientation to the putative exon (Fu & Ares, 2014). We therefore investigated the predicted effects of any altered splice factor motifs on a case-by-case basis to determine whether they were more likely to increase or decrease inclusion.

**Categorisation.** Each resulting cryptic exon/SNP pair was categorised as:

- a. A *known* SNPTic exon, if the association between the cryptic exon and the SNP had been explicitly characterised in a prior report;
- b. A *probable* SNPTic exon, if a prior report had linked the SNP with a particular phenotype, but had not investigated differential splicing of the cryptic exon as a cause of that phenotype and
- c. A *potential* SNPTic exon, if the SNP was not significantly associated with a phenotype and had not been shown to directly affect CE splicing, but was still deemed a worthwhile candidate for further investigation due to its predicted effect on splicing of a known CE. To limit this category to the most likely examples, we included only those SNPs that altered the most highly conserved nucleotides of a cryptic exon splice motif, that is, –3 to +3 of the acceptor site or –3 to +6 of the donor site.

**Final assessment.** The expected phenotypic effects of each putative SNPTic exon were analysed and

discussed, according to both prior research on the affected gene and the fundamental principles of U2-type splicing. Additionally, the predicted changes to each gene's encoded protein sequence were calculated for each putative SNPTic exon using the ExPASy Translate Tool (Duvaud et al. 2021) and are provided as a Supplementary File in the online version of this report.

In devising this method, we were unable to account for the splicing impact of SNP-associated changes on RNA folding as, to our knowledge, there is currently no generalised method for making these types of predictions. Since it has been shown that even single nucleotide changes can affect gene expression by altering RNA secondary structures (Ritz et al., 2012; Sabarinathan et al., 2013), these types of splicing effects may well exist; although another recent report indicated that the impact of SNPs on conserved RNA structures was minimal (Kalmykova et al., 2021).

GenBank IDs of studied genes: *APC*, NG\_008481.4; *ARSB*, NG\_007089.1; *ATM*, NG\_009830.1; *CSF1R*, NG\_012303.2; *DMD*, NG\_012232.1; *F8*, NG\_011403.2; *FGB*, NG\_008833.1; *GHRL*, NG\_011560.1; *IL16*, NG\_029933.1; *LHCGR*, NG\_008193.2; *MYBPC3*, NG\_007667.1; *NF1*, NG\_009018.1; *OAS1*, NG\_011530.2; *OTC*, NG\_008471.1; *POC1B*, NG\_041783.1; *TSMF*, NG\_016971.1.

### 3 | RESULTS AND DISCUSSION

In addition to six *known* SNPTic exons, our analysis also discovered five *probable* SNPTic exons and five *potential* SNPTic exons (Tables S1 and 1), each arising within a different gene. With one exception (*OAS1-2a*, described below), the predicted reading frame effect of each CE inclusion was to introduce at least one premature stop codon more than 55 nt upstream of the transcript's final exon junction, either within the putative SNPTic exon itself or within its flanking 3' canonical exon, and would therefore be expected to induce NMD of the mature transcript (Zhang, Sun, et al., 1998; Zhang, Center, et al., 1998). There were no examples of a SNP of interest adding or altering a start or stop codon within a CE.

Therefore, except where otherwise stated, we have assumed the following general precepts: (a) Splicing of a SNPTic exon into a transcript prevents translation of the transcript and triggers its decay via NMD, (b) leading to chronically lower levels of the full-length mature transcript, (c) leading to chronically lower levels of the full-length protein and (d) leading to the observed phenotypic differences amongst different genotypes of the relevant SNP. We have applied these assumptions accordingly in discussing each putative SNPTic exon in the sections that follow.

TABLE 1 Sixteen putative SNPtic exons and their associated phenotypes

SNPtic exon	SNPs	Expected effect	SNP phenotype	Exon high-inclusion phenotype
<i>ATM</i> -27a (Known)	<b>rs609261</b> (NC_000011.10: g.108287407T>C)	Less inclusion	Lower cancer risk	Ataxia telangiectasia (poor coordination, prominent eye blood vessels and high cancer risk)
<i>F8</i> -13a (Known)	<b>rs781928603</b> (NC_000023.11: g.154947237_154947249del)	More inclusion	Mild haemophilia type A	Mild haemophilia type A
<i>IL16</i> -6a (Known)	<b>rs4778639</b> (NC_000015.10: g.81308110T>C)	More inclusion (no NMD)	Higher interleukin-16 levels in blood	Unknown
<i>LHCGR</i> -6a (Known)	<b>rs68073206</b> (NC_000002.12: g.48721568A>C)	Splice-switch (S > L)	Higher testosterone levels and higher androgen sensitivity index	Male pseudohermaphroditism
<i>OAS1</i> -2a (Known) (rs116086311)	<b>rs116086311</b> (NC_000012.12: g.112910849C>T), <b>rs34137742</b> (NC_000012.12: g.112910856C>T)	More inclusion	Higher risk of encephalitis and paralysis if infected with West Nile virus (rs34137742)	Unknown. Other <i>OAS1</i> mutations associated with higher risk of West Nile virus infection
<i>TSMF</i> -2a (Known)	<b>rs2014886</b> (NC_000012.12: g.57783654C>G)	More inclusion	PREDICTED: Higher risk of multiple sclerosis	Unknown. Other <i>TSMF</i> mutations associated with cardiomyopathy, encephalomyopathy and ataxia
<i>APC</i> -11a (Probable)	<b>rs2545162</b> (NC_000005.10: g.112822734G>A)	More inclusion	Higher colorectal cancer risk	Adenomatous polyposis (colon cancer)
<i>FGB</i> -1a (Probable)	<b>rs2227401</b> (NC_000004.12: g.15456529C>T)	Less inclusion	Higher blood fibrinogen levels	Afibrinogenemia (Persistent cerebral transient ischemic attacks, blood clots and 1/50th normal fibrinogen levels)
<i>GHRL</i> -4a (Probable)	<b>rs2075356</b> (NC_000003.12: g.10287125T>C)	Splice-switch (L > S)	Decreases cancer risk and increases bulimia risk	Unknown; other <i>GHRL</i> mutations associated with metabolic dysregulation
<i>MYBPC3</i> -12a (Probable)	<b>rs10769255</b> (NC_000011.10: g.47345820C>A)	Less inclusion	Slightly higher cognitive performance	Hypertrophic cardiomyopathy
<i>OTC</i> -9a (Probable)	<b>rs5963419</b> (NC_000023.11: g.38412940T>A)	Less inclusion	Increased risk of bipolar disorder	Hyperammonemia leading to brain damage and death

(Continues)



TABLE 1 (Continued)

SNPtic exon	SNPs	Expected effect	SNP phenotype	Exon high-inclusion phenotype
ARSB-6a (Possible)	<b>rs337836</b> (NC_000005.10: g:78884913T>C)	More inclusion	PREDICTED: Shorter stature and higher risk profile for other symptoms.	Mucopolysaccharidosis Type VI (Skeletal abnormalities, hearing and vision loss and heart disease)
CSF1R-15a (Possible)	<b>rs11952821</b> (NC_000005.10: g:150060771G>A)	More inclusion	PREDICTED: Shorter stature and increased susceptibility to cognitive decline.	Early onset HDLS, skeletal dysplasia (dwarfism) and brain malformation
DMD-2a (Possible)	<b>rs145743673</b> (NC_000023.11: g:32863915T>C)	Splice-switch (S > L)	PREDICTED: Asymptomatically lower dystrophin levels. May compound an existing BMD phenotype.	Duchenne muscular dystrophy, primarily due to DMD e8-11 duplication
NF1-36a (Possible)	<b>rs35888506</b> (NC_000017.11: g:31324211C>T)	More inclusion	PREDICTED: Higher cancer risk	Unknown; other NF1 mutations cause neurofibromatosis type 1
POC1B-9a (Possible)	<b>rs11323565</b> (NC_000012.12: g:89461145del)	More inclusion	PREDICTED: Lower visual acuity	Reduced visual acuity and contrast, photophobia

Note: Citations are shown in main text. GenBank IDs of studied genes: APC, NG\_008481.4; ARSB, NG\_007089.1; ATM, NG\_009830.1; CSF1R, NG\_012303.2; DMD, NG\_012232.1; F8, NG\_011403.2; FGB, NG\_008833.1; GHRL, NG\_011560.1; IL16, NG\_029933.1; LHCGR, NG\_008193.2; MYBPC3, NG\_007667.1; NF1, NG\_009018.1; OASI, NG\_011530.2; OTC, NG\_008471.1; POC1B, NG\_041783.1; TSEF, NG\_016971.1.

Below we have identified each SNPTic exon according to the name of the gene and the intron in which it occurs, followed by the letter 'a' to distinguish it from the preceding canonical exon. Where two splice variants exist for a single cryptic exon, we have identified each variant as 'S' or 'L' depending on whether it is the shorter or longer variant, respectively.

### 3.1 | Known SNPTic exons

#### 3.1.1 | ATM-27a

This CE in *Ataxia-Telangiectasia Mutated (ATM-OMIM #607585)* was first discovered by Coutinho et al. (2005), who also described a longer variant that shared the same acceptor site. The short variant was subsequently characterised as a SNPTic exon (*sans* use of this term) by Kralovicova et al. (2016). Remarkably, even though this SNP only slightly weakened the CE's acceptor site, Kralovicova and colleagues demonstrated that this was sufficient to cause a measurable decrease in the rate of its inclusion. This, in turn, led to the corresponding increase in translation of ATM protein; and since *ATM* is a tumour-suppressor gene (Choi et al., 2016), it is likely that this elevated ATM level explains the lower cancer risk seen in carriers of the SNP.

#### 3.1.2 | F8-13a

Unlike the other SNPs discussed in this report, which are germline substitutions of single nucleotides, the SNP in this case (*rs781928603*) is a variably sized poly-T deletion with multiple reported alternative alleles. Although the summed frequencies of these alternative alleles exceed 1%, Jourdy et al. (2018) report only on the phenotype of the *del13T* variant, the global frequency of which is not precisely defined but estimated at well below 1%. This *del13T* allele is associated with a mild haemophilia type A phenotype in males, as it induces inclusion of a CE in transcripts of *Coagulation Factor VIII (F8-OMIM #300841)*, an important blood clotting protein. However, despite being associated with increased *F8-13a* inclusion, the *del13T* allele slightly *decreases* the MaxEnt score of the CE acceptor site. Jourdy and colleagues showed that the likely reason for the splicing enhancement is a decrease in 5' silencer binding, although we suggest that shortening of the branch point AG-exclusion zone may also be a contributing factor (Wimmer et al., 2020). Interestingly, inclusion of identical CE sequence, and a mild haemophilia type A phenotype, has also been reported to result from an enhancing mutation in the CE donor site (Dericquebourg

et al., 2020), demonstrating that the major allele isoform of the *F8-13a* acceptor site is functional.

### 3.1.3 | *IL16-6a*

This CE in *Interleukin 16 (IL16-OMIM #603035)* is unique amongst the putative SNPTic exons discussed in this report, as it is the only one not to introduce a premature stop codon into the mature transcript, and therefore is not expected to promote transcript degradation via NMD. The CE was discovered in the peripheral blood RNA of 23 individuals by Sakaguchi and Suyama (2021), via bioinformatic analysis of RNA-Seq and whole-genome sequence data, and was not linked with a disease phenotype.

The SNP *rs4778639* converts the *IL16-6a* acceptor site dinucleotide from an AT to an AG and is therefore likely to be essential for splicing of the CE. This SNP was found by Sun et al. (2018) to significantly correlate with increased IL16 protein levels in blood. The CE arises in the terminal intron of *IL16* and is predicted to introduce nine additional amino acids to the IL16 peptide (see Appendix S1). This insertion interrupts the PDZ3 domain of the precursor protein (Sakaguchi & Suyama, 2021) and constitutes a substantial increase in the size of the mature protein, which is typically only 121 peptides long after caspase-3 catalysis (Zhang, Sun, et al., 1998; Zhang, Center, et al., 1998). This would presumably have a marked effect on the 3D structure, export, multimeric assembly and CD4+ recruitment activity of mature IL16 (Richmond et al., 2014), yet the haploid frequency of the causative SNP (8.37%) indicates that it is not significantly deleterious, at least for heterozygous carriers. We would welcome any future research that elucidates the true in vivo behaviour of this novel potential protein isoform.

### 3.1.4 | *LHCGR-6a(S/L)*

Like *ATM-27a*, the SNPTic exon in *Luteinising Hormone/Choriogonadotropin Receptor (LHCGR-OMIM #152790)* was discovered (Kossack et al., 2008) several years before the effect of its SNP was directly characterised (Liu et al., 2017). This cryptic exon bears two variants that have distinct donor sites but share an acceptor site. In their 2008 report, Kossack and colleagues detailed an SNV in *LHCGR-6a* that significantly increased its frequency of inclusion, resulting in a male-pseudohermaphroditism phenotype in the affected patients. The authors also showed significant inclusion of *LHCGR-6a* from the reference allele and claimed that this demonstrated its status as a bona fide exon, a claim that appears to be supported

by the high degree of conservation of *LHCGR-6a* and its flanking regions (Figure 2b). However, at the time of writing, *LHCGR-6a* has not yet been listed as a canonical exon of any official transcript variants on NCBI, and we have therefore continued to refer to it as a cryptic exon here.

Liu et al. (2017) investigated the effects of the SNP *rs68073206*, located in the donor site of *LHCGR-6aL*. Because this SNP substantially enhances this donor site, it might be expected that this would increase the NMD of inclusive transcripts and therefore be associated with a phenotype of lower male sexual development. Surprisingly, the authors discovered just the opposite—SNP carrier status was associated with higher levels of testosterone and higher androgen sensitivity, and inter-genotype differences in transcript frequencies did not follow a simple ‘zero sum’ model. Part of the reason for these counterintuitive effects may be competition between the donor sites of the long and the short isoforms, as it is unclear how much of the SNP-driven increase in *LHCGR-6aL* splicing comes at the expense of *LHCGR-6aS* splicing and how much at the expense of normal *LHCGR* splicing. The likely status of *LHCGR-6a* as a highly conserved bona fide exon suggests that its splicing may play a more complex role in *LHCGR* autoregulation.

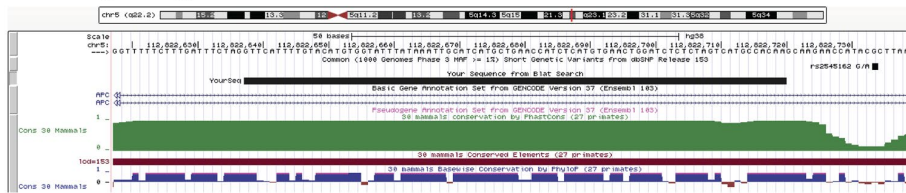
### 3.1.5 | *OAS1-2a*

This CE in *2'-5'-Oligoadenylate Synthetase 1 (OAS1-OMIM #164350)* was identified in whole blood RNA sequence from eight healthy donors by Sakaguchi and Suyama (2021). The *OAS1* gene plays an important role in the innate immune response to viruses, and a canonical splice site polymorphism near *OAS1* exon 6 has been shown to increase the risk of West Nile virus infection (Lim et al., 2009).

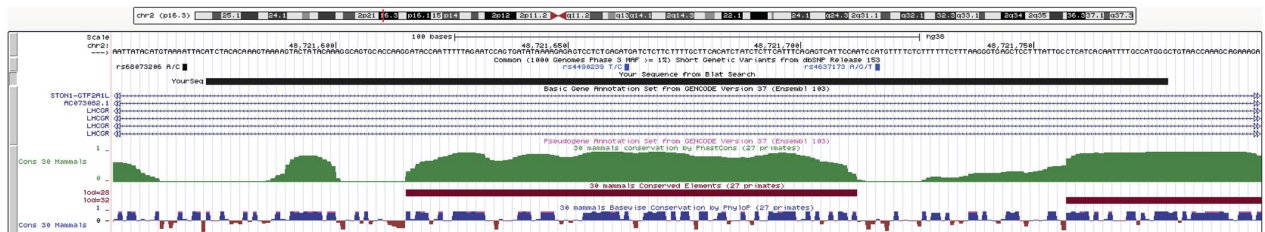
Sakaguchi and Suyama identified the *rs116086311* SNP as causative of *OAS1-2a* inclusion. The aetiology of this SNP is obvious, as it converts *OAS1-2a*'s GC donor site dinucleotide to a much stronger GT. But whilst no phenotype associations have been discovered for *rs116086311*, a second SNP 3' of the donor site, *rs34137742*, was found to be associated with a higher risk of encephalitis and paralysis following West Nile virus infection (Bigham et al., 2011). At first glance this seems counterintuitive: since the most powerful single-nucleotide splice mutations tend to be those that alter an intron terminal dinucleotide, one might expect that the strongest association would be detected for *rs116086311*, with *rs34137742* perhaps being identified as a weaker contributing factor.

However, this phenotype association can be interpreted consistently with the general model of SNPTic exon

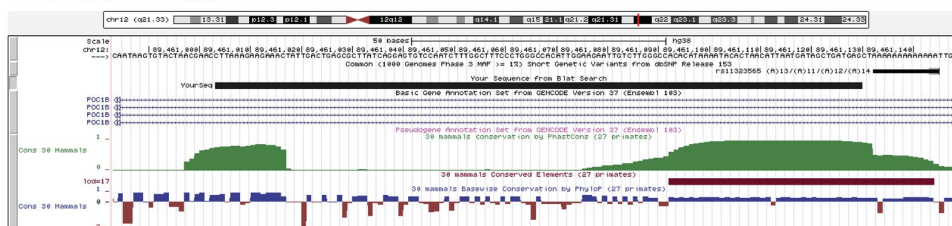
## (a) APC-11a



## (b) LHCGR-6aL



## (c) POC1B-9a



**FIGURE 2** Cryptic exons *APC-11a*, *LHCGR-6a* and *POC1B-9a* exhibit high sequence conservation. Images were captured as screenshots from the UCSC Genome Browser (Kent et al., 2002). In descending order, displayed tracks are: Base position, dbSNP 153, input sequence, ‘GENCODE V37’ (aligned transcript variants) and ‘Cons 30 Primates’. ‘The Cons 30 Primates’ track, which is erroneously labelled as ‘Cons 30 Mammals’ in the browser, displays sequence conservation data from 30 non-human primate species

splicing (Figure 1) once population genetics are considered. Firstly, the direct effect of *rs34137742* is to remove a binding motif for SRSF9, a ubiquitously expressed serine-rich splicing factor that silences upstream donor sites and enhances downstream donor sites (Cloutier et al., 2008). Loss of this motif would therefore be more permissive of *OAS1-2a* splicing. Secondly, since the *OAS1-2a* donor site dinucleotide is splice-competent in both *rs116086311* alleles (i.e. GC or GT), it is theoretically possible to observe a quantitative effect from *rs34137742* in a population independently of their *rs116086311* genotypes. Lastly, *rs34137742* has a haploid frequency of over 10.3%, compared to less than 3.5% for *rs116086311*. This means that *rs34137742* is likely to be much better represented in the sample group of any GWAS, making its phenotypic effects more easily discoverable at the population level even if they are milder than those of *rs34137742* at an individual level.

Although a disease risk phenotype has been established only for *rs34137742*, and an *OAS1-2a* splicing effect only for *rs116086311*, we suggest that the reverse may also

be true, and that these effects are a logical consequence of CE-induced NMD of *OAS1* transcripts.

3.1.6 | *TSMF-2a*

Unlike the other three ‘known’ SNPTic exons, the SNPTic exon in *Ts Translation Elongation Factor, Mitochondrial* (*TSMF-OMIM* #604723) does not have an associated phenotype and was discovered in the blood RNA of healthy individuals (Morrison et al., 2013). Morrison and colleagues suggest that this SNP may be a risk factor for multiple sclerosis (MS); but although both prior and subsequent research has supported a link between MS and other *TSMF* variants (Handel et al., 2010; Mo et al., 2019), at the time of writing, no such association has been demonstrated for this SNP. However, the authors did demonstrate that this SNP was almost entirely responsible for splicing of the SNPTic exon through conversion of the GC-donor motif to a GT-donor motif, though they

also detected low levels of splicing even in C-allele homozygotes, which fits with prior observations of U2-spliced GC-donor sites being functional but less efficient (Thanaraj & Clark, 2001). We also noted that this SNPTic exon was an exact match for 1 of the 10 CEs previously predicted by Sela et al. (2010). Other mutations in *TSMF* have been associated with cardiomyopathy, encephalomyopathy and ataxia (Smeitink et al., 2006; Emperador et al., 2016).

### 3.1.7 | Other SNPTic exons in Sakaguchi and Suyama 2021

Sakaguchi and Suyama (2021) reported 116 new CEs discovered in publicly available RNAseq data. For two of these CEs, we found evidence in the literature supporting a SNP-associated phenotype, and we have discussed these above as *OAS1-2a* and *IL16-6a*. We also noted an additional 17 CEs in the authors' report where the causative variants corresponded to common SNPs, though we were not able to find any published phenotype associations for these SNPs, nor for any other SNPs within  $\pm 20$  nt of their associated SNPTic exons. These examples are listed in Table 2, but as we have little to add to the original authors' analysis of these 17 CEs, we instead refer interested readers to investigate their report.

## 3.2 | Probable SNPTic exons

### 3.2.1 | *APC-11a*

This CE in *Adenomatous Polyposis Coli* (*APC-OMIM* #611731) was first reported as a pathogenic inclusion by Spier et al. (2012). Remarkably, three unique donor site SNVs have been reported as being causative of pathogenic *APC-11a* splicing (Nieminen et al., 2016; Spier et al., 2012). All three mutations caused a phenotype of familial adenomatous polyposis (FAP), a disease characterised by colon polyps and an elevated risk of colon cancer. Like *LHCGR-6a*, the sequence in and surrounding *APC-11a* is highly conserved (Figure 2a), supporting the case for this being an as yet unrecognised bona fide exon.

The SNP *rs2545162* is predicted to create a 3' binding motif for MBNL1, an alternative splicing regulator that has been shown to consistently enhance the splicing of exons when it binds within  $\sim 200$  nt 3' of their donor sites (Konieczny et al., 2014; Wang et al., 2012). We would therefore expect the minor allele of this SNP to increase *APC-11a* inclusion and be associated with a higher risk of FAP-like symptoms. This prediction agrees with the findings of Hildebrandt et al. (2016), who found that

*rs2545162* was significantly associated with a higher risk of colorectal cancer.

### 3.2.2 | *FGB-1a*

This pathogenic CE in *Fibrinogen Beta* (*FGB-OMIM* #134830) was first predicted by Dear et al. (2006), who identified the causative mutation in a consanguineous family, and was later confirmed and further characterised by Davis et al. (2009). The authors determined that an SNV within the CE converted a silencer motif to an enhancer, thereby substantially increasing *FGB-1a* inclusion. Consequently, the homozygous proband exhibited a phenotype of afibrinogenemia with recurrent transient ischemic attacks, whilst his two heterozygous children bore a milder phenotype of hypofibrinogenemia.

The SNP *rs2227401* is situated inside the CE and is predicted to silence its inclusion, so we would expect an associated phenotype opposite to afibrinogenemia. This is supported by two GWASes (de Vries et al., 2017; Kolz et al., 2009) that independently discovered an association between *rs2227401* and higher levels of blood fibrinogen.

### 3.2.3 | *GHRL-4a(S/L)*

Like the *LHCGR-6a* SNPTic exon, this CE in *Ghrelin* (*GHRL-OMIM* #605353) also consists of a short and a long variants, though in this case it is the donor site that is shared with two unique acceptor sites (Seim et al., 2013). Seim and colleagues observed *GHRL-4a* inclusion in multiple healthy cell types and elevated inclusion in prostate cancer cell lines. They also noted that the acceptor site of *GHRL-4aS* appeared to be non-canonical, with an AA terminal dinucleotide. However, the SNP *rs2075356* converts this AA to a canonical AG. Given the haploid frequency of this SNP (11%) compared to the frequency of bona fide non-AG acceptor sites ( $< 0.1\%$  as per Olthof et al., 2019 and Piovesan et al., 2019), we suggest that carriage of this SNP may be the more likely explanation for *GHRL-4aS* splicing.

The *rs2075356* SNP has separately been linked with a decreased risk of certain forms of cancers (Pabalan et al., 2014) and elevated risk of purging-type bulimia nervosa (Ando et al., 2006). However, whilst the *rs2075356* minor allele is likely to be essential for *GHRL-4aS* splicing, the confounding effect of competition between the *GHRL-4aS* and *GHRL-4aL* acceptor sites makes it difficult to predict how it would change the total amount of *GHRL-4a* splicing. This difficulty is compounded by the complex post-translational processing of preproghrelin peptides and the



TABLE 2 SNPtic exons caused by common SNPs ( $\geq 1\%$  haploid frequency) as reported by Sakaguchi and Suyama (2021)

Chr.	Gene	Start	End	SNP position	rsID	Varnomen
chr1-	<i>NOC2L</i>	882,137	882,244	882,250	rs111463901	NC_000001.11:g.946870C>A
chr1+	<i>RWDD3</i>	95,702,899	95,703,016	95,702,898	rs80241359	NC_000001.11:g.95237342A>G
chr5-	<i>TBCA</i>	77,026,223	77,026,280	77,026,221	rs75503375	NC_000005.10:g.77730396C>A
chr5-	<i>SRA1</i>	139,932,741	139,932,889	139,932,740	rs112703681	NC_000005.10:g.140553155T>C
chr6+	<i>ABRACL</i>	139,354,886	139,354,992	139,354,992	rs62441851	NC_000006.12:g.139033855A>G
chr7-	<i>COA1</i>	43,695,632	43,695,752	43,695,628	rs1859877	NC_000007.14:g.43656029C>T
chr10+	<i>HSD17B7P2</i>	38,654,838	38,654,939	38,654,940	rs2804645	NC_000010.11:g.38366012T>A
chr11-	<i>DHCR7</i>	71,157,568	71,157,656	71,157,567	rs75686975	NC_000011.10:g.71446521G>A
chr12+	<i>MGST1</i>	16,503,692	16,503,788	16,503,789	rs9332891	NC_000012.12:g.16350855T>G
<b>chr12+</b>	<b><i>OAS1</i></b>	<b>113,348,549</b>	<b>113,348,652</b>	<b>113,348,654</b>	<b>rs116086311</b>	<b>NC_000012.12:g.112910849C&gt;T</b>
chr14+	<i>CRIP1</i>	105,954,227	105,954,364	105,954,368	rs112661676	NC_000014.9:g.105488031G>A
<b>chr15+</b>	<b><i>IL16</i></b>	<b>81,600,452</b>	<b>81,600,478</b>	<b>81,600,451</b>	<b>rs4778639</b>	<b>NC_000015.10:g.81308110T&gt;C</b>
chr16-	<i>CNOT1</i>	58,662,843	58,663,002	58,662,841	rs28644182	NC_000016.10:g.58628937G>A
chr16-	<i>FANCA</i>	89,829,046	89,829,201	89,829,201	rs9806894	NC_000016.10:g.89762793G>A
chr17+	<i>STAT5A</i>	40,440,948	40,441,015	40,441,014	rs74875201	NC_000017.11:g.42288996G>A
chr19+	<i>CERS4</i>	8,312,329	8,312,446	8,312,447	rs12977774	NC_000019.10:g.8247563A>G
chr21-	<i>LINC00158</i>	26,758,995	26,759,072	26,758,994	rs13049048	NC_000021.9:g.25386681T>A
chr21-	<i>C21orf59</i>	33,980,707	33,980,799	33,980,705	rs111323620	NC_000021.9:g.32608395G>A
chr21+	<i>NDUFV3</i>	44,326,950	44,327,012	44,327,013	rs73905782	NC_000021.9:g.42906903A>G
chr22+	<i>APOBEC3D</i>	39,419,690	39,419,852	39,419,853	rs6001388	NC_000022.11:g.39023848T>G

Note: 'Start' and 'End' coordinates refer to human genome assembly hg19, as per cited work. SNPtic exons *OAS1*-2a and *IL16*-6a are indicated with bold text. The *APOBEC3D* SNP is not shown in the cited work but is required for splicing in addition to the published variant (Narumi Sakaguchi 2021, Pers. Comm).

varied roles they play in metabolic regulation. We therefore limit ourselves to suggesting that a focused investigation of the effects of *rs2075356* may prove to be a fruitful line of research.

### 3.2.4 | *MYBPC3*-12a

This CE in *Myosin-Binding Protein C3* (*MYBPC3*-OMIM #600958) was discovered by Bagnall et al. (2018). In this case, the patient's SNV converted the GC of *MYBPC3*-12a donor site to a stronger GT. The proband was one of a cohort of patients with hypertrophic cardiomyopathy, a disease characterised by overdevelopment of the muscle in the left ventricle of the heart, leading to a greatly elevated risk of arrhythmia and heart failure. Cardiac hypertrophy in general has also been associated with a higher risk of cognitive dysfunction in later life (Hayakawa et al., 2012).

The SNP *rs10769255* occurs inside *MYBPC3*-12a and is predicted to silence its inclusion and thereby permit increased translation of full-length *MYBPC3*. Surprisingly, in a subsequent GWAS *rs10769255* was found to correlate with higher performance in certain tests of cognitive ability (Lee et al., 2018). Although the difference in scores

attributed to the SNP was quite small, it was nonetheless determined to be highly significant due to the study's large sample size. This phenotype could be explained as a mild inverse of the elevated cognitive decline risk typically associated with hypertrophic cardiomyopathy.

### 3.2.5 | *OTC*-9a

This CE in *Ornithine Transcarbamylase* (*OTC*-OMIM #300461) was first observed as a pathogenic inclusion by Engel et al. (2008), caused by a donor site SNV. Because *OTC* is a key component in the metabolic conversion of ammonia to urea, the *OTC* deficiency caused by pathogenic inclusion of *OTC*-9a resulted in hyperammonemia, and was ultimately fatal to the affected patient, who died at a very young age due to severe cerebral oedema. Mutations with less severe effects on the quantity and function of *OTC* protein have been known to cause late-onset *OTC* deficiency, which can manifest in previously asymptomatic patients as erratic behaviour, lethargy and hyperammonemia (Hidaka et al., 2020; Rush et al., 2014).

The SNP *rs5963419* is situated within this CE and is predicted to silence its inclusion. We might therefore expect this SNP to be associated with higher *OTC* protein

levels and a benign 'hypoammonemic' phenotype, opposite to the severe hyperammonemia observed for pathogenic inclusion of *OTC-9a*. However, to date the only positive GWAS correlation for *rs5963419* is deleterious: its minor allele was found to be overrepresented in populations with bipolar disorder (Sklar et al., 2008).

A possible explanation for this is that a higher level of neuronal OTC (Bernstein et al., 2017) in carriers of this SNP may elevate the conversion of ammonia to urea in some neurons, and therefore leave less ammonia available for the conversion of glutamate into glutamine by glutamine synthetase. This could in turn result in chronically higher neuronal glutamate levels, which have been associated with bipolar disorder (Gigante et al., 2012). If this SNP had an opposite mechanism of action—that is, it increased risk of bipolar disorder by *reducing* OTC levels—then there should also be a strong and obvious correlation between bipolar disorder and late-onset hyperammonemia generally; yet we could find no reports of any such association in the literature.

### 3.3 | Potential SNPTic exons

#### 3.3.1 | *ARSB-6a*

This CE in *Arylsulfatase B* (*ARSB*—OMIM #300461) was discovered by Broeders et al. (2020) as a sporadic inclusion in both patient and healthy control RNAs from primary human fibroblasts treated with cycloheximide, an NMD inhibitor. Broeders and colleagues noted that the donor site of this CE, which bears a non-canonical AT flanking dinucleotide in the reference sequence, was not predicted by any of the algorithms they tested. However, we observed that if the SNP *rs337836* was present then this donor site dinucleotide would be converted to a canonical GT. Given that this SNP has a haploid frequency of 33%, we suggest that its presence or absence is the most likely explanation for differential *ARSB-6a* splicing between individuals.

Loss-of-function mutations in *ARSB* are typically causative of mucopolysaccharidosis type six (MPS VI), a recessive inherited disorder with a spectrum of severity and a broad range of symptoms, including skeletal abnormalities, hearing loss, vision loss and heart disease. Broeders and colleagues showed compelling evidence that the immediate effect of *ARSB-6a* inclusion is to induce NMD, as *ARSB-6a*-inclusive transcripts were almost undetectable in the RNA of cells not treated with cycloheximide. Therefore, the expected phenotype associations for this SNP would be analogous to sub-clinical MPS VI. We speculate that these might include shorter stature and an elevated risk of sleep apnoea and heart disease.

We also noted that this CE falls within the 3' UTR of *ARSB* transcript variant ENST00000565165.2 (GENCODE), although its sequence does not show significant conservation.

#### 3.3.2 | *CSF1R-15a*

This CE in *Colony-Stimulating Factor 1 Receptor* (*CSF1R*—OMIM #164770) was discovered by Guo et al. (2019), who observed it as a pathogenic splicing variant induced by an internal two-nucleotide deletion. The consanguineous proband had a severe phenotype due to being homozygous for this allele, and their symptoms included hypotonicity, focal seizures, brain malformation and mild skeletal abnormalities. In cases of other monoallelic *CSF1R* loss-of-function mutations, a phenotype of 'hereditary diffuse leukoencephalopathy with spheroids' (HDLS) is often observed, a neurodegenerative disorder with adult onset and variable presentation.

Although the SNP *rs11952821* only slightly enhances the *CSF1R-15a* acceptor site, it is comparable to the improvement induced by a SNP at the same position in *ATM-27a*, which was demonstrated to have a significant splicing effect. We would therefore expect *rs11952821* carriers to have elevated *CSF1R-15a* inclusion leading to NMD and lower full-length *CSF1R* translation, and an associated phenotype equivalent to very mild HDLS. Due to the variable presentation of classical HDLS, this phenotype could manifest as an increased general risk of neurodegenerative disease and/or a more severe prognosis when neurodegenerative symptoms are already present for other reasons.

#### 3.3.3 | *DMD-2a(S/L)*

This CE in *Duchenne Muscular Dystrophy* (*DMD*—OMIM #300377) was detected in a patient diagnosed with Duchenne muscular dystrophy (Ishibashi et al., 2006). The CE bears a short (S) and a long (L) isoforms, with a shared donor site, and two acceptor sites four nucleotides aside. Unusually, the causative mutation in this case was significantly distal on the same allele—a tandem duplication of *DMD* exons 8–11. The affected 3-year-old male (XY) patient had a characteristic Duchenne muscular dystrophy phenotype for his age, with extremely high serum creatine kinase and early signs of muscle weakness. However, because the exons 8–11 duplication already induces a reading frame shift in the *DMD* transcript, it is not possible to assign aspects of this patient's phenotype to *DMD-2a* splicing alone.

The SNP *rs145743673*, respectively, weakens and strengthens the acceptor sites of the CE short and long

isoforms, and would therefore be expected to induce splice-switching from the short to the long isoform. As with *LHCGR-6a* and *GHRL-4a*, we have refrained from predicting the effect of CE splice-switching on total transcript and protein levels. It is possible that a GWAS could detect a correlation between *rs145743673* and levels of dystrophin in normal individuals, though the rarity of the SNP (1.1%) would make this challenging, and any differences detected may be largely asymptomatic if the high variability of 'normal' dystrophin expression is any indication (Beekman et al., 2018).

### 3.3.4 | *NF1-36a*

This CE in *Neurofibromin 1 (NF1-OMIM #613113)* was first detected in the peripheral blood RNA of at least 17 healthy control individuals (Landrith et al., 2020). Although this splice variant is not yet associated with a phenotype, loss-of-function mutations in *NF1* are typically causative of type 1 neurofibromatosis (NF1), which is characterised by ubiquitous benign nerve tumours, café-au-lait skin pigmentation, neurocognitive impairment and a greatly elevated risk of cancer.

The SNP *rs35888506* converts the *NF1-36a* donor site dinucleotide from a GC to a stronger GT. It would therefore be expected to cause substantially higher inclusion of this CE, although low-level splicing of the GC allele might also be observed. We predict an associated phenotype equivalent to very mild NF1, which may be detected as elevated cancer risk and elevated risk of neurocognitive impairment.

### 3.3.5 | *POC1B-9a*

This CE in *Proteome Of Centriole Protein 1B (POC1B-OMIM #614784)* was detected in blood RNA from a compound heterozygous patient with adult-onset symptoms of reduced visual acuity, reduced visual contrast and photophobia (Weisschuh et al., 2021). Pathogenic mutations to *POC1B* generally cause some form of retinopathy, although symptoms and age of onset are highly variable. In this case, the patient's mutation destroyed the *POC1B* exon 7 donor site, resulting in variable skipping of exons 6 and 7 in addition to *POC1B-9a* inclusion. Consequently, *POC1B-9a* inclusion by itself cannot be definitively implicated in the proband's symptoms. However, like *LHCGR-6a* and *APC-11a*, *POC1B-9a* also exhibits high sequence conservation (Figure 2c), indicating that it may be a bona fide poison exon.

Similar to the *F8-13a* SNP, *rs11323565* causes a length variation in the *POC1B-9a* acceptor site poly-T tract, extending it from 12T to 13T. But unlike *F8-13a*, in this case

an expansion of the poly-T tract appears more likely to increase inclusion of the CE, as the change in AGEZ length is minimal. We would therefore predict that this SNP may be associated with diminished visual acuity in the elderly.

Our comparison of *POC1B-9a* with *F8-13a* led us to note that length variations in acceptor site poly-T tracts appear to have competing and contradictory effects on exon recognition, as such variants can simultaneously strengthen an acceptor splice motif whilst weakening branch point definition. We would welcome any further research towards reliably predicting the effects of these variants.

## 3.4 | Conclusions and recommendations

Although we discovered only five new probable SNPTic exons, we were encouraged to observe that in four of these cases, the predicted splicing effect was generally consistent with the correlated phenotype, whilst the fifth (*GHRL-4a*) was expected to cause complex splice-switching and thus neither supported nor contradicted our model. We also highlighted an additional four possible SNPTic exons; their associated SNPs may prove worthwhile targets of future GWASes.

A reviewer of this report observed that several of the SNPs in the 'Probable' and 'Possible' SNPTic exon categories fell outside of the highly conserved splice motif regions (as defined in step 5a of our search method), whilst this was true for only one (*F8-13a*) in the 'Known' category. This discrepancy may be a consequence of the fact that, prior to this report, there had not been any general attempts to match SNPTic exons with population phenotypes. Consequently, only those SNPs with the most noticeable splicing effects have been characterised, and these primarily occur in the most highly conserved splice motif nucleotides.

Whilst we hope these findings will be of interest, our primary goal in reporting them is to demonstrate proof of concept for the utility of our discovery method. In future, researchers reporting on new cryptic exons may apply this method for no cost greater than a few minutes expended on online database queries, and in doing so may discover better explanations for published results, or fruitful new lines of inquiry for their research. Antisense oligonucleotide-based skipping of NMD-inducing poison exons is already showing great promise for the treatment of heritable encephalopathies (Aziz et al., 2021), and it is possible that further discoveries of SNPTic exons will reveal additional novel antisense targets.

As innovations in RNA sequencing technology continue to accelerate the discovery of new cryptic exons and pseudoexons, so will grow the potential for making exciting new connections between this relatively small body of data and the vast number of SNP-phenotype associations already discovered by GWASes.

Addendum: Close to time of publication we identified what appear to be two additional examples of known SNPTic exons, one in the gene *Ras Homolog Family Member A (RHOA-OMIM #165390)* (Medina et al., 2012) and one in the gene *F-Box Protein 38 (FBXO38-OMIM #608533)* (Saferali et al. 2019). Although we could not include these examples in our analysis without further peer review, we wish to acknowledge the original reports as literature of interest.

## ACKNOWLEDGEMENTS

The authors thank Professor Joerg Gromoll, Professor Mikita Suyama and Narumi Sakaguchi for their helpful correspondence.

## CONFLICT OF INTEREST

The authors have declared no conflicts of interest.

## ETHICAL COMPLIANCE

As our study exclusively used published and publicly available data, it did not require approval by an ethics committee.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study.

## ORCID

Niall Patrick Keegan  <https://orcid.org/0000-0001-9475-103X>

<https://orcid.org/0000-0001-9475-103X>

Sue Fletcher  <https://orcid.org/0000-0002-8632-641X>

## REFERENCES

- Ando, T., Komaki, G., Naruo, T., Okabe, K., Takii, M., Kawai, K., Konjiki, F., Takei, M., Oka, T., Takeuchi, K., Masuda, A., Ozaki, N., Suematsu, H., Denda, K., Kurokawa, N., Itakura, K., Yamaguchi, C., Kono, M., Suzuki, T., ... Ichimaru, Y. (2006). Possible role of preproghrelin gene polymorphisms in susceptibility to bulimia nervosa. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics*, *141B*(8), 929–934. <https://doi.org/10.1002/ajmg.b.30387>
- Änkö, M.-L., Müller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J., & Neugebauer, K. M. (2012). The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biology*, *13*(3), R17. <https://doi.org/10.1186/gb-2012-13-3-r17>
- Aziz, M. C., Schneider, P. N., & Carvill, G. L. (2021). Targeting poison exons to treat developmental and epileptic encephalopathy. *Developmental Neuroscience*, *43*(3-4), 241–246. <https://doi.org/10.1159/000516143>
- Bagnall, R. D., Ingles, J., Dinger, M. E., Cowley, M. J., Ross, S. B., Minoche, A. E., Lal, S., Turner, C., Colley, A., Rajagopalan, S., Berman, Y., Ronan, A., Fatkin, D., & Semsarian, C. (2018). Whole genome sequencing improves outcomes of genetic testing in patients with hypertrophic cardiomyopathy. *Journal of the American College of Cardiology*, *72*(4), 419–429. <https://doi.org/10.1016/j.jacc.2018.04.078>
- Beck T., Shorter T., & Brookes A. J. (2020). GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. *Nucleic Acids Research*, *48*(D1), D933–D940. <https://www.doi.org/10.1093/nar/gkz895>
- Beekman, C., Janson, A. A., Baghat, A., van Deutekom, J. C., & Datson, N. A. (2018). Use of capillary Western immunoassay (Wes) for quantification of dystrophin levels in skeletal muscle of healthy controls and individuals with Becker and Duchenne muscular dystrophy. *PLoS One*, *13*(4), e0195850. <https://doi.org/10.1371/journal.pone.0195850>
- Bernstein, H. G., Dobrowolny, H., Keilhoff, G., & Steiner, J. (2017). In human brain ornithine transcarbamylase (OTC) immunoreactivity is strongly expressed in a small number of nitrergic neurons. *Metabolic Brain Disease*, *32*(6), 2143–2147. <https://doi.org/10.1007/s11011-017-0105-2>
- Bigham, A. W., Buckingham, K. J., Husain, S., Emond, M. J., Bofferding, K. M., Gildersleeve, H., Rutherford, A., Astakhova, N. M., Perelygin, A. A., Busch, M. P., Murray, K. O., Sejvar, J. J., Green, S., Kriesel, J., Brinton, M. A., & Bamshad, M. (2011). Host genetic risk factors for West Nile virus infection and disease progression. *PLoS One*, *6*(9), e24745. <https://doi.org/10.1371/journal.pone.0024745>
- Braun, T. A., Mullins, R. F., Wagner, A. H., Andorf, J. L., Johnston, R. M., Bakall, B. B., Deluca, A. P., Fishman, G. A., Lam, B. L., Weleber, R. G., Cideciyan, A. V., Jacobson, S. G., Sheffield, V. C., Tucker, B. A., & Stone, E. M. (2013). Non-exomic and synonymous variants in ABCA4 are an important cause of Stargardt disease. *Human Molecular Genetics*, *22*(25), 5136–5145. <https://doi.org/10.1093/hmg/ddt367>
- Broeders, M., Smits, K., Goynuk, B., Oussoren, E., van den Hout, H. J. M. P., Bergsma, A. J., van der Ploeg, A. T., & Pijnappel, W. W. M. P. (2020). A generic assay to detect aberrant ARSB splicing and mRNA degradation for the molecular diagnosis of MPS VI. *Molecular Therapy Methods & Clinical Development*, *19*, 174–185. <https://doi.org/10.1016/j.omtm.2020.09.004>
- Cano-Gamez, E., & Trynka, G. (2020). From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, *11*, 424. <https://doi.org/10.3389/fgene.2020.00424>
- Canson, D., Glubb, D., & Spurdle, A. B. (2020). Variant effect on splicing regulatory elements, branchpoint usage, and pseudoexonization: Strategies to enhance bioinformatic prediction using hereditary cancer genes as exemplars. *Human Mutation*, *41*(10), 1705–1721. <https://doi.org/10.1002/humu.24074>
- Carvill, G. L., & Mefford, H. C. (2020). Poison exons in neurodevelopment and disease. *Current Opinion in Genetics & Development*, *65*, 98–102. <https://doi.org/10.1016/j.gde.2020.05.030>
- Choi, M., Kipps, T., & Kurzrock, R. (2016). ATM mutations in cancer: Therapeutic implications. *Molecular Cancer Therapeutics*, *15*(8), 1781–1791. <https://doi.org/10.1158/1535-7163.MCT-15-0945>
- Cloutier, P., Toutant, J., Shkreta, L., Goekjian, S., Revil, T., & Chabot, B. (2008). Antagonistic effects of the SRp30c protein and cryptic 5' splice sites on the alternative splicing of the apoptotic regulator Bcl-x. *Journal of Biological Chemistry*, *283*(31), 21315–21324. <https://doi.org/10.1074/jbc.M800353200>



- 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. <https://doi.org/10.1002/humu.20170>
- Davis, R. L., Homer, V. M., George, P. M., & Brennan, S. O. (2009). A deep intronic mutation in FGB creates a consensus exonic splicing enhancer motif that results in afibrinogenemia caused by aberrant mRNA splicing, which can be corrected in vitro with antisense oligonucleotide treatment. *Human Mutation*, 30(2), 221–227. <https://doi.org/10.1002/humu.20839>
- de Vries, P. S., Sabater-Lleal, M., Chasman, D. I., Trompet, S., Ahluwalia, T. S., Teumer, A., Kleber, M. E., Chen, M.-H., Wang, J. J., Attia, J. R., Marioni, R. E., Steri, M., Weng, L.-C., Pool, R., Grossmann, V., Brody, J. A., Venturini, C., Tanaka, T., Rose, L. M., ... Dehghan, A. (2017). Comparison of HapMap and 1000 genomes reference panels in a large-scale genome-wide association study. *PLoS One*, 12(1), e0167742. <https://doi.org/10.1371/journal.pone.0167742>
- Dear, A., Daly, J., Brennan, S. O., Tuckfield, A., & George, P. M. (2006). An intronic mutation within FGB (IVS1+2076 a->g) is associated with afibrinogenemia and recurrent transient ischemic attacks. *Journal of Thrombosis and Haemostasis*, 4(2), 471–472. <https://doi.org/10.1111/j.1538-7836.2006.01722.x>
- Dericquebourg, A., Jourdy, Y., Fretigny, M., Lienhart, A., Claeysens, S., Ternisien, C., & Vinciguerra, C. (2020). Identification of new F8 deep intronic variations in patients with haemophilia A. *Haemophilia*, 26(5), 847–854. <https://doi.org/10.1111/hae.14134>
- Dobkin, C., Pergolizzi, R. G., Bahre, P., & Bank, A. (1983). Abnormal splice in a mutant human beta-globin gene not at the site of a mutation. *Proceedings of the National Academy of Sciences of the United States of America*, 80(5), 1184–1188. <https://doi.org/10.1073/pnas.80.5.1184>
- Druhan, L. J., Lance, A., Hamilton, A., Steuerwald, N. M., Tjaden, E., & Avalos, B. R. (2020). Altered splicing and intronic polyadenylation of CSF3R via a cryptic exon in acute myeloid leukemia. *Leukemia Research*, 92, 106349. <https://doi.org/10.1016/j.leukres.2020.106349>
- Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V., & Durinx, C. (2021). Expsy, the Swiss bioinformatics resource portal, as designed by its users. *Nucleic Acids Research*, 49(W1), W216–W227. <https://doi.org/10.1093/nar/gkab225>
- Emperador, S., Bayona-Bafaluy, M. P., Fernández-Marmiesse, A., Pineda, M., Felgueroso, B., López-Gallardo, E., Artuch, R., Roca, I., Ruiz-Pesini, E., Couce, M. L., & Montoya, J. (2016). Molecular-genetic characterization and rescue of a TSFM mutation causing childhood-onset ataxia and nonobstructive cardiomyopathy. *European Journal of Human Genetics*, 25(1), 153–156. <https://doi.org/10.1038/ejhg.2016.124>
- Engel, K., Nuoffer, J.-M., Mühlhausen, C., Klaus, V., Largiadèr, C. R., Tsiakas, K., Santer, R., Wermuth, B., & Häberle, J. (2008). Analysis of mRNA transcripts improves the success rate of molecular genetic testing in OTC deficiency. *Molecular Genetics and Metabolism*, 94(3), 292–297. <https://doi.org/10.1016/j.ymgme.2008.03.009>
- Fu, X. D., & Ares, M. Jr (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, 15(10), 689–701. <https://doi.org/10.1038/nrg3778>
- Genomes Project, & Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicke, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., ... Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. <https://doi.org/10.1038/nature15393>
- Gigante, A. D., Bond, D. J., Lafer, B., Lam, R. W., Young, L. T., & Yatham, L. N. (2012). Brain glutamate levels measured by magnetic resonance spectroscopy in patients with bipolar disorder: A meta-analysis. *Bipolar Disorders*, 14(5), 478–487. <https://doi.org/10.1111/j.1399-5618.2012.01033.x>
- Guo, L., Bertola, D. R., Takanohashi, A., Saito, A., Segawa, Y., Yokota, T., Ishibashi, S., Nishida, Y., Yamamoto, G. L., Franco, J. F. D. S., Honjo, R. S., Kim, C. A., Musso, C. M., Timmons, M., Pizzino, A., Taft, R. J., Lajoie, B., Knight, M. A., Fischbeck, K. H., ... Ikegawa, S. (2019). Bi-allelic CSF1R mutations cause skeletal dysplasia of dysosteosclerosis-pyle disease spectrum and degenerative encephalopathy with brain malformation. *American Journal of Human Genetics*, 104(5), 925–935. <https://doi.org/10.1016/j.ajhg.2019.03.004>
- Handel, A. E., Handunnetthi, L., Berlanga, A. J., Watson, C. T., Morahan, J. M., & Ramagopalan, S. V. (2010). The effect of single nucleotide polymorphisms from genome wide association studies in multiple sclerosis on gene expression. *PLoS One*, 5(4), e10142. <https://doi.org/10.1371/journal.pone.0010142>
- Hayakawa, M., Yano, Y., Kuroki, K., Inoue, R., Nakanishi, C., Sagara, S., Koga, M., Kubo, H., Imakiire, S., Aoyagi, Z., Kitani, M., Kanemaru, K., Hidehito, S., Shimada, K., & Kario, K. (2012). Independent association of cognitive dysfunction with cardiac hypertrophy irrespective of 24-h or sleep blood pressure in older hypertensives. *American Journal of Hypertension*, 25(6), 657–663. <https://doi.org/10.1038/ajh.2012.27>
- Hidaka, M., Higashi, E., Uwatoko, T., Uwatoko, K., Urashima, M., Takashima, H., Watanabe, Y., Kitazono, T., & Sugimori, H. (2020). Late-onset ornithine transcarbamylase deficiency: A rare cause of recurrent abnormal behavior in adults. *Acute Medicine & Surgery*, 7(1), e565. <https://doi.org/10.1002/ams2.565>
- Hildebrandt, M. A., Reyes, M. E., Lin, M., He, Y., Nguyen, S. V., Hawk, E. T., & Wu, X. (2016). Germline genetic variants in the Wnt/beta-catenin pathway as predictors of colorectal cancer risk. *Cancer Epidemiology, Biomarkers & Prevention*, 25(3), 540–546. <https://doi.org/10.1158/1055-9965.EPI-15-0834>
- Ishibashi, K., Takeshima, Y., Yagi, M., Nishiyama, A., & Matsuo, M. (2006). Novel cryptic exons identified in introns 2 and 3 of the human dystrophin gene with duplication of exons 8–11. *Kobe Journal of Medical Sciences*, 52(3–4), 61–75.
- Jian, X., Boerwinkle, E., & Liu, X. (2014). In silico tools for splicing defect prediction: A survey from the viewpoint of end users. *Genetics in Medicine*, 16(7), 497–503. <https://doi.org/10.1038/gim.2013.176>
- Jourdy, Y., Janin, A., Fretigny, M., Lienhart, A., Negrier, C., Bozon, D., & Vinciguerra, C. (2018). Recurrent F8 intronic deletion found in mild hemophilia A causes alu exonization. *American Journal of Human Genetics*, 102(2), 199–206. <https://doi.org/10.1016/j.ajhg.2017.12.010>
- Kalmykova, S., Kalinina, M., Denisov, S., Mironov, A., Skvortsov, D., Guigo, R., & Pervouchine, D. (2021). Conserved long-range base pairings are associated with pre-mRNA processing of human genes. *Nature Communications*, 12(1), 2300. <https://doi.org/10.1038/s41467-021-22549-7>
- Keegan, N. P. (2020). Pseudoexons of the DMD gene. *Journal of Neuromuscular Diseases*, 7(2), 77–95. <https://doi.org/10.3233/JND-190431>

- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006. <https://doi.org/10.1101/gr.229102>
- Kolz, M., Baumert, J., Gohlke, H., Grallert, H., Döring, A., Peters, A., Wichmann, E., Koenig, W., & Illig, T. (2009). Association study between variants in the fibrinogen gene cluster, fibrinogen levels and hypertension: Results from the MONICA/KORA study. *Thrombosis and Haemostasis*, *101*(2), 317–324. <https://doi.org/10.1160/Th08-06-0411>
- Konicieczny, P., Stepniak-Konicieczna, E., & Sobczak, K. (2014). MBNL proteins and their target RNAs, interaction and splicing regulation. *Nucleic Acids Research*, *42*(17), 10873–10887. <https://doi.org/10.1093/nar/gku767>
- Kossack, N., Simoni, M., Richter-Unruh, A., Themmen, A. P., & Gromoll, J. (2008). Mutations in a novel, cryptic exon of the luteinizing hormone/chorionic gonadotropin receptor gene cause male pseudohermaphroditism. *PLoS Med*, *5*(4), e88. <https://doi.org/10.1371/journal.pmed.0050088>
- Kralovicova, J., Knut, M., Cross, N. C., & Vorechovsky, I. (2016). Exon-centric regulation of ATM expression is population-dependent and amenable to antisense modification by pseudoexon targeting. *Scientific Reports*, *6*, 18741. <https://doi.org/10.1038/srep18741>
- Landrith, T., Li, B., Cass, A. A., Conner, B. R., LaDuca, H., McKenna, D. B., Maxwell, K. N., Domchek, S., Morman, N. A., Heinlen, C., Wham, D., Koptiuch, C., Vagher, J., Rivera, R., Bunnell, A., Patel, G., Geurts, J. L., Depas, M. M., Gaonkar, S., ... Karam, R. (2020). Splicing profile by capture RNA-seq identifies pathogenic germline variants in tumor suppressor genes. *NPJ Precision Oncology*, *4*, 4. <https://doi.org/10.1038/s41698-020-0109-y>
- Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghziyan, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., Fontana, M. A., Kundu, T., Lee, C., Li, H., Li, R., Royer, R., Timshel, P. N., Walters, R. K., Willoughby, E. A., ... Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, *50*(8), 1112–1121. <https://doi.org/10.1038/s41588-018-0147-3>
- Lim, J. K., Lisco, A., McDermott, D. H., Huynh, L., Ward, J. M., Johnson, B., Johnson, H., Pape, J., Foster, G. A., Krysztof, D., Follmann, D., Stramer, S. L., Margolis, L. B., & Murphy, P. M. (2009). Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man. *PLoS Path*, *5*(2), e1000321. <https://doi.org/10.1371/journal.ppat.1000321>
- Liu, W., Han, B., Zhu, W., Cheng, T., Fan, M., Wu, J., Yang, Y., Zhu, H., Si, J., Lyu, Q., Chai, W., Zhao, S., Song, H., Kuang, Y., & Qiao, J. (2017). Polymorphism in the alternative donor site of the cryptic exon of LHCGR: Functional consequences and associations with testosterone level. *Scientific Reports*, *7*, 45699. <https://doi.org/10.1038/srep45699>
- Medina M. W., Theusch E., Naidoo D., Bauzon F., Stevens K., Mangravite L. M., Kuang Y. L., & Krauss R. M. (2012). RHOA Is a Modulator of the Cholesterol-Lowering Effects of Statin. *PLoS Genetics*, *8*, (11), e1003058. <https://www.doi.org/10.1371/journal.pgen.1003058>
- Mo, X. B., Lei, S. F., Qian, Q. Y., Guo, Y. F., Zhang, Y. H., & Zhang, H. (2019). Integrative analysis revealed potential causal genetic and epigenetic factors for multiple sclerosis. *Journal of Neurology*, *266*(11), 2699–2709. <https://doi.org/10.1007/s00415-019-09476-w>
- Morrison, F. S., Locke, J. M., Wood, A. R., Tuke, M., Pasko, D., Murray, A., Frayling, T., & Harries, L. W. (2013). The splice site variant rs11078928 may be associated with a genotype-dependent alteration in expression of GSDMB transcripts. *BMC Genomics*, *14*, 627. <https://doi.org/10.1186/1471-2164-14-627>
- Nieminen, T. T., Pavicic, W., Porkka, N., Kankainen, M., Jarvinen, H. J., Lepisto, A., & Peltomaki, P. (2016). Pseudoexons provide a mechanism for allele-specific expression of APC in familial adenomatous polyposis. *Oncotarget*, *7*(43), 70685–70698. <https://doi.org/10.18632/oncotarget.12206>
- Olthof, A. M., Hyatt, K. C., & Kanadia, R. N. (2019). Minor intron splicing revisited: Identification of new minor intron-containing genes and tissue-dependent retention and alternative splicing of minor introns. *BMC Genomics*, *20*(1), 686. <https://doi.org/10.1186/s12864-019-6046-x>
- Pabalan, N. A., Seim, I., Jarjanazi, H., & Chopin, L. K. (2014). Associations between ghrelin and ghrelin receptor polymorphisms and cancer in Caucasian populations: A meta-analysis. *BMC Genetics*, *15*, 118.
- Parada, G. E., Munita, R., Cerda, C. A., & Gysling, K. (2014). A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Research*, *42*(16), 10564–10578. <https://doi.org/10.1093/nar/gku744>
- Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M. C., & Caracausi, M. (2019). Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*, *12*(1), 315. <https://doi.org/10.1186/s13104-019-4343-8>
- Piva, F., Giulietti, M., Burini, A. B., & Principato, G. (2012). SpliceAid 2: A database of human splicing factors expression data and RNA target motifs. *Human Mutation*, *33*(1), 81–85. <https://doi.org/10.1002/humu.21609>
- Richmond, J., Tuzova, M., Cruikshank, W., & Center, D. (2014). Regulation of cellular processes by interleukin-16 in homeostasis and cancer. *Journal of Cellular Physiology*, *229*(2), 139–147. <https://doi.org/10.1002/jcp.24441>
- Ritz, J., Martin, J. S., & Laederach, A. (2012). Evaluating our ability to predict the structural disruption of RNA by SNPs. *BMC Genomics*, *13*(Suppl 4), S6. <https://doi.org/10.1186/1471-2164-13-S4-S6>
- Romano, M., Buratti, E., & Baralle, D. (2013). Role of pseudoexons and pseudointrons in human cancer. *International Journal of Cell Biology*, *2013*, 1–16. <https://doi.org/10.1155/2013/810572>
- Rush, E. T., Hartmann, J. E., Skrabal, J. C., & Rizzo, W. B. (2014). Late-onset ornithine transcarbamylase deficiency: An under recognized cause of metabolic encephalopathy. *SAGE Open Med Case Rep*, *2*, 2050313X14546348. <https://doi.org/10.1177/2050313X14546348>
- Sabarinathan, R., Tafer, H., Seemann, S. E., Hofacker, I. L., Stadler, P. F., & Gorodkin, J. (2013). RNAsnp: Efficient detection of local RNA secondary structure changes induced by SNPs. *Human Mutation*, *34*(4), 546–556. <https://doi.org/10.1002/humu.22273>
- Saferali A., Yun J. H., Parker M. M., Sakornsakolpat P., Chase R. P., Lamb A., Hobbs B. D., Boezen M. H., Dai X., de Jong K., Beaty T. H., Wei W., Zhou X., Silverman E. K., Cho M. H., & Hersh C. P. (2019). Analysis of genetically driven alternative splicing identifies FBXO38 as a novel COPD susceptibility gene. *PLOS Genetics*, *15*, (7), e1008229. <https://www.doi.org/10.1371/journal.pgen.1008229>

- Sakaguchi, N., & Suyama, M. (2021). In silico identification of pseudo-exon activation events in personal genome and transcriptome data. *RNA Biology*, *18*(3), 382–390. <https://doi.org/10.1080/15476286.2020.1809195>
- Seim, I., Lubik, A. A., Lehman, M. L., Tomlinson, N., Whiteside, E. J., Herington, A. C., Nelson, C. C., & Chopin, L. K. (2013). Cloning of a novel insulin-regulated ghrelin transcript in prostate cancer. *Journal of Molecular Endocrinology*, *50*(2), 179–191. <https://doi.org/10.1530/JME-12-0150>
- Sela, N., Mersch, B., Hotz-Wagenblatt, A., & Ast, G. (2010). Characteristics of transposable element exonization within human and mouse. *PLoS One*, *5*(6), e10907. <https://doi.org/10.1371/journal.pone.0010907>
- Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. <https://doi.org/10.1093/nar/29.1.308>
- Sklar, P., Smoller, J. W., Fan, J., Ferreira, M. A. R., Perlis, R. H., Chambert, K., Nimgaonkar, V. L., McQueen, M. B., Faraone, S. V., Kirby, A., de Bakker, P. I. W., Ogdie, M. N., Thase, M. E., Sachs, G. S., Todd-Brown, K., Gabriel, S. B., Sougnez, C., Gates, C., Blumenstiel, B., ... Purcell, S. M. (2008). Whole-genome association study of bipolar disorder. *Molecular Psychiatry*, *13*(6), 558–569. <https://doi.org/10.1038/sj.mp.4002151>
- Smeitink, J. A. M., Elpeleg, O., Antonicka, H., Diepstra, H., Saada, A., Smits, P., Sasarman, F., Vriend, G., Jacob-Hirsch, J., Shaag, A., Rechavi, G., Welling, B., Horst, J., Rodenburg, R. J., van den Heuvel, B., & Shoubbridge, E. A. (2006). Distinct clinical phenotypes associated with a mutation in the mitochondrial translation elongation factor EFTs. *American Journal of Human Genetics*, *79*(5), 869–877. <https://doi.org/10.1086/508434>
- 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. <https://doi.org/10.1002/humu.22082>
- Stein, S., Lu, Z. X., Bahrami-Samani, E., Park, J. W., & Xing, Y. (2015). Discover hidden splicing variations by mapping personal transcriptomes to personal genomes. *Nucleic Acids Research*, *43*(22), 10612–10622. <https://doi.org/10.1093/nar/gkv1099>
- Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., Oliver-Williams, C., Kamat, M. A., Prins, B. P., Wilcox, S. K., Zimmerman, E. S., Chi, A. N., Bansal, N., Spain, S. L., Wood, A. M., ... Butterworth, A. S. (2018). Genomic atlas of the human plasma proteome. *Nature*, *558*(7708), 73–79. <https://doi.org/10.1038/s41586-018-0175-2>
- Thanaraj, T. A., & Clark, F. (2001). Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Research*, *29*(12), 2581–2593. <https://doi.org/10.1093/nar/29.12.2581>
- Thomas, J. D., Polaski, J. T., Feng, Q., De Neef, E. J., Hoppe, E. R., McSharry, M. V., Pangallo, J., Gabel, A. M., Belleville, A. E., Watson, J., Nkisi, N. T., Berger, A. H., & Bradley, R. K. (2020). RNA isoform screens uncover the essentiality and tumor-suppressor activity of ultraconserved poison exons. *Nature Genetics*, *52*(1), 84–94. <https://doi.org/10.1038/s41588-019-0555-z>
- Tubeuf, H., Charbonnier, C., Soukariéh, O., Blavier, A., Lefebvre, A., Dauchel, H., Frebourg, T., Gaildrat, P., & Martins, A. (2020). Large-scale comparative evaluation of user-friendly tools for predicting variant-induced alterations of splicing regulatory elements. *Human Mutation*, *41*(10), 1811–1829. <https://doi.org/10.1002/humu.24091>
- Vaz-Drago, R., Custodio, N., & Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. *Human Genetics*, *136*(9), 1093–1111. <https://doi.org/10.1007/s00439-017-1809-4>
- Vorechovsky, I. (2010). Transposable elements in disease-associated cryptic exons. *Human Genetics*, *127*(2), 135–154. <https://doi.org/10.1007/s00439-009-0752-4>
- Wang, E. T., Cody, N. A. L., Jog, S., Biancolella, M., Wang, T. T., Treacy, D. J., Luo, S., Schroth, G. P., Housman, D. E., Reddy, S., Lécuyer, E., & Burge, C. B. (2012). Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell*, *150*(4), 710–724. <https://doi.org/10.1016/j.cell.2012.06.041>
- Weisschuh, N., Mazzola, P., Bertrand, M., Haack, T. B., Wissinger, B., Kohl, S., & Stingl, K. (2021). Clinical characteristics of POC1B-associated retinopathy and assignment of pathogenicity to novel deep intronic and non-canonical splice site variants. *International Journal of Molecular Sciences*, *22*(10), 5396. <https://doi.org/10.3390/ijms22105396>
- Will, K., Stuhmann, M., Dean, M., & Schmidtke, J. (1993). Alternative splicing in the first nucleotide binding fold of CFTR. *Human Molecular Genetics*, *2*(3), 231–235. <https://doi.org/10.1093/hmg/2.3.231>
- Wimmer, K., Schamschula, E., Wernstedt, A., Traunfellner, P., Amberger, A., Zschocke, J., Kroisel, P., Chen, Y., Callens, T., & Messiaen, L. (2020). AG-exclusion zone revisited: Lessons to learn from 91 intronic NF1 3' splice site mutations outside the canonical AG-dinucleotides. *Human Mutation*, *41*(6), 1145–1156. <https://doi.org/10.1002/humu.24005>
- 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. <https://doi.org/10.1089/1066527041410418>
- Zhang, J., Sun, X. L., Qian, Y. M., LaDuca, J. P., & Maquat, L. E. (1998). At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: A possible link between nuclear splicing and cytoplasmic translation. *Molecular and Cellular Biology*, *18*(9), 5272–5283. <https://doi.org/10.1128/Mcb.18.9.5272>
- Zhang, Y., Center, D. M., Wu, D. M., Cruikshank, W. W., Yuan, J., Andrews, D. W., & Kornfeld, H. (1998). Processing and activation of pro-interleukin-16 by caspase-3. *Journal of Biological Chemistry*, *273*(2), 1144–1149. <https://doi.org/10.1074/jbc.273.2.1144>

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Keegan, N. P., & Fletcher, S. (2021). A spotter's guide to SNPtic exons: The common splice variants underlying some SNP–phenotype correlations. *Molecular Genetics & Genomic Medicine*, *00*, e1840. <https://doi.org/10.1002/mgg3.1840>

## **Chapter 6**

# **Induction of cryptic pre-mRNA splice-switching by antisense oligonucleotides**

## 6.1 Preamble

Although the previous chapters in this thesis have focused on pseudoexons, a more frequently observed form of cryptic splicing entails use of alternative donor or acceptor sites in or near canonical exons. The effect of this on the mature transcript is either the inclusion of an exon-adjacent segment of intron, or the skipping of part of the exon, depending on whether the cryptic site lies outside or within the canonical boundaries of the exon (Nelson and Green 1990; Haj Khelil *et al.* 2008).

This form of cryptic splicing typically occurs as the result of mutations in or around canonical exons (Buratti *et al.* 2011), although changes in splice factor abundance can also cause global shifts in splice site selection, as is seen in many cancers (Zhang *et al.* 2021). However, over the course of many exon-skipping antisense oligonucleotide (AO) experiments by our group, we noted that a handful of AOs caused cryptic splicing when transfected into cells expressing the target genes. This AO-induced cryptic splicing appeared to exclusively affect the targeted exons, indicating that it was not a general effect of the AO. Furthermore, the cryptic splice sites induced were all inside the affected exons and, with a single exception, were all donor sites.

While strategies for optimising AO design are progressively becoming more sophisticated (Aartsma-Rus *et al.* 2009; Mitrpant *et al.* 2009; Echigoya *et al.* 2015), they still entail an element of trial-and-error, and it is not always obvious why one AO is effective while another AO with an overlapping target site is not. We therefore sought to collectively examine whether the AOs that induced cryptic splicing had any common characteristics that might indicate the mechanisms responsible for why and where these cryptic splice sites were activated, in an attempt to generate useful new insights into effective AO design.

This was a highly collaborative project, as indicated by the number of co-authors on the report, encompassing numerous AOs, cell types and transfection methods performed by multiple researchers. My own role in the project was to conduct the bioinformatic data analysis, generate figures from this analysis, and to share

responsibility for writing and editing the report with my colleague and co-first author, Kristin Ham.

A common rule-of-thumb in exon-skipping AO design is to target regions with high densities of exon splice enhancers, in order to maximise silencing of the exon definition signal. We reasoned that, just as AOs and splice factors compete for RNA binding, so too would the local secondary structure of each RNA molecule. We therefore devised a “hybrid plot” method of visualising individual pre-mRNA exons that combined the predicted secondary structure of each exon with the predicted splice factor motifs. Then, by performing an additional round of modelling that blocked the nucleotides within the AO target sites, we could approximate the effects of AO binding on pre-mRNA local secondary structures.

In general, this analysis found that cryptic splice-site activating AOs tend to selectively shift the accessibility of exon splice enhancers and silencers in and around the pre-mRNA exon, in a way that strengthens definition of the retained segments and/or weakens definition of the skipped segments. They do this both directly, through steric blocking, and indirectly through changes in the openness of surrounding RNA secondary structure.

Our findings highlight the importance of incorporating RNA secondary structure predictions into AO design and present a novel and intuitive approach for doing so. We also suggest that partial exon skipping need not remain a curious side-effect of some AOs, but could be exploited as a valuable therapeutic approach in select cases where skipping of whole exons is not a viable strategy.

## 6.2 Citation

Ham, K.A., Keegan, N.P., McIntosh, C.S., Aung-Htut, M.T., Zaw, K., Greer, K., et al. (2021). Induction of cryptic pre-mRNA splice-switching by antisense oligonucleotides. *Sci Rep* 11(1), 15137. doi: 10.1038/s41598-021-94639-x.

An earlier version of this report (not included in this thesis) was published as the following pre-print:

Ham, K.A., Keegan, N.P., McIntosh, C.S., Aung-Htut, M.T., Zaw, K., Greer, K., et al. (2020). Cryptic U2-dependent pre-mRNA splice site usage induced by splice switching antisense oligonucleotides [PREPRINT]. *Research Square*. doi: 10.21203/rs.3.rs-144809/v1



# OPEN Induction of cryptic pre-mRNA splice-switching by antisense oligonucleotides

Kristin A. Ham<sup>1,2,4</sup>, Niall P. Keegan<sup>1,2,4</sup>, Craig S. McIntosh<sup>1,2</sup>, May T. Aung-Htut<sup>1,2</sup>, Khine Zaw<sup>1,2,3</sup>, Kane Greer<sup>1,2</sup>, Sue Fletcher<sup>1,2</sup> & Steve D. Wilton<sup>1,2</sup>✉

Antisense oligomers (AOs) are increasingly being used to modulate RNA splicing in live cells, both for research and for the development of therapeutics. While the most common intended effect of these AOs is to induce skipping of whole exons, rare examples are emerging of AOs that induce skipping of only part of an exon, through activation of an internal cryptic splice site. In this report, we examined seven AO-induced cryptic splice sites in six genes. Five of these cryptic splice sites were discovered through our own experiments, and two originated from other published reports. We modelled the predicted effects of AO binding on the secondary structure of each of the RNA targets, and how these alterations would in turn affect the accessibility of the RNA to splice factors. We observed that a common predicted effect of AO binding was disruption of the exon definition signal within the exon's excluded segment.

## Abbreviations

AO	Antisense oligonucleotide
SnRNP	Small nuclear ribonucleoproteins
5'ss	5' Splice site
3'ss	3' Splice site
Nt	Nucleotide
NMD	Nonsense mediated decay
NC	Negative control
SRSF	Serine/arginine-rich splicing factor
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer
lncRNA	Long non-coding RNA
2'-OMe PS	2'-O-Methyl modified bases on a phosphorothioate backbone
FBS	Fetal bovine serum
BLAST	Basic local alignment search tool

The process of pre-mRNA splicing is a fundamental aspect of gene regulation and function in higher eukaryotes. Pre-mRNA consists of retained regions, termed exons, that are interspersed with regions destined for excision, termed introns<sup>1</sup>. During maturation into mRNA, the introns are removed and the exons are ligated together to form a continuous message, ready to be translated into a protein, or in some cases to serve other functions as a non-coding RNA. Pre-mRNA splicing involves a multitude of splicing factors that interact with numerous splicing motifs on the transcript<sup>2</sup>. A large multi-protein complex called the spliceosome is responsible for the coordination of this complex set of transesterification reactions<sup>3</sup>.

The major form of the spliceosome is composed of five small nuclear ribonucleoproteins (snRNPs; U1, U2, U5 and U4/U6), as well as numerous non-snRNP proteins<sup>4,5</sup>. The canonical 5' splice site (5'ss) is defined by an AG|GURAGU sequence, while the 3' splice site (3'ss) is denoted by a (Yn)-YAG| sequence (where; |= exon boundary; underlined sequence identifies invariant nucleotides; R = purine; Y = pyrimidine)<sup>6</sup>. The branchpoint sequence, typically located approximately 15 to 50 nucleotides (nt) upstream from the 3'ss, is required for U2

<sup>1</sup>Centre for Molecular Medicine and Innovative Therapeutics, Health Futures Institute, Murdoch University, Perth, WA 6150, Australia. <sup>2</sup>Perron Institute for Neurological and Translational Science, Centre for Neuromuscular and Neurological Disorders, The University of Western Australia, Perth, WA 6009, Australia. <sup>3</sup>Department of Biochemistry, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand. <sup>4</sup>These authors contributed equally: Kristin A. Ham and Niall P. Keegan. ✉email: s.wilton@murdoch.edu.au



**Figure 1.** Activation of cryptic splice sites by AO-mediated splice switching in four different gene transcript targets. (a) *COL7A1* exon 15. (b) *SRSF2* exon 2. (c) *ATXN3* exon 9. (d) *USH2A* exon 13. Reverse transcription-PCR analysis after transfection with antisense oligonucleotides (AOs), at various nM concentrations indicated above the gel image. Exon splice enhancer (ESE) motifs, predicted by ESEfinder 3.0<sup>34</sup>. The color code indicates the putative binding sites for serine/arginine rich splicing factors (SRSFs). Blue boxes represent exons, lines between the boxes represent introns, dashed lines above and below represent various splicing events, pink boxes represent the portion of exon removed after the activation of a new cryptic splice site, black arrows indicate primer location and direction, coloured lines indicate AO target binding site, red polygons represent termination codons removed after cryptic splice site activation. Alternative transcript exon composition before and after *SRSF2* AO treatment. Multiple transcript isoforms noted as T### according to Ensembl. Grey arrow indicates an amplicon that could not be successfully isolated and sequenced. NC, negative control sequence synthesized as 2'-OMe PS; UT, untreated; 100 bp, 100 base pair DNA ladder; nM, nanomolar. The gel images were cropped for presentation. Full-length gel images are presented in Supplementary Fig. S3.

snRNA binding during spliceosome formation. This sequence is defined as YNCURAY (underlined sequence denotes branch formation region; bold nucleotides are highly conserved; N = any nucleotide)<sup>6</sup>. The major spliceosome (called spliceosome hereon), along with hundreds of associated splicing factors are responsible for over 95% of all splicing reactions, including the phenomenon known as alternative splicing<sup>7–10</sup>.

Alternative splicing is a process whereby multiple different transcripts and protein isoforms can arise from a single protein-coding gene and is an essential element in spatial and temporal regulation of gene expression in higher eukaryotes<sup>7</sup>. In order to achieve alternative splicing, the spliceosome must recognize and select a splice site amid a variety of alternative splice sites and branchpoints within the transcript. Typically, these splice sites are well defined and have evolutionarily conserved functions. However, on occasion, sequences usually ignored by the spliceosome can become activated as splice junctions. These are known as cryptic splice sites<sup>11</sup> and are most often activated by mutations or errors during transcription<sup>12</sup>. According to DBASS, the mutations most commonly causative of cryptic splice site activation are those that weaken canonical exon splice sites, thus redirecting the spliceosome to utilize a viable cryptic site nearby<sup>13</sup>. However, this is a relatively rare outcome of such mutations, which are generally far more likely to induce whole exon skipping<sup>14</sup>. Cryptic splice sites may be found within both exonic and intronic regions and typically include or exclude a proportion of the exon or intron<sup>12</sup>. Interestingly, recent data has shown that cryptic splice sites can also be activated by synthetic molecules such as antisense oligonucleotides.

Antisense oligonucleotides (AOs) are small, single-stranded RNA or DNA-like synthetic molecules used to modify gene expression. These AOs can be used to downregulate gene expression through RNA silencing, redirection of pre-mRNA splicing patterns, intron retention, inhibiting translation, or RNase H-induced degradation of the target gene transcript<sup>15</sup>. The sequence of maturing gene transcripts can also be altered by using AOs to induce removal or inclusion of an exon, as demonstrated by current therapeutic strategies approved for the treatment of Duchenne muscular dystrophy and spinal muscular atrophy, respectively.

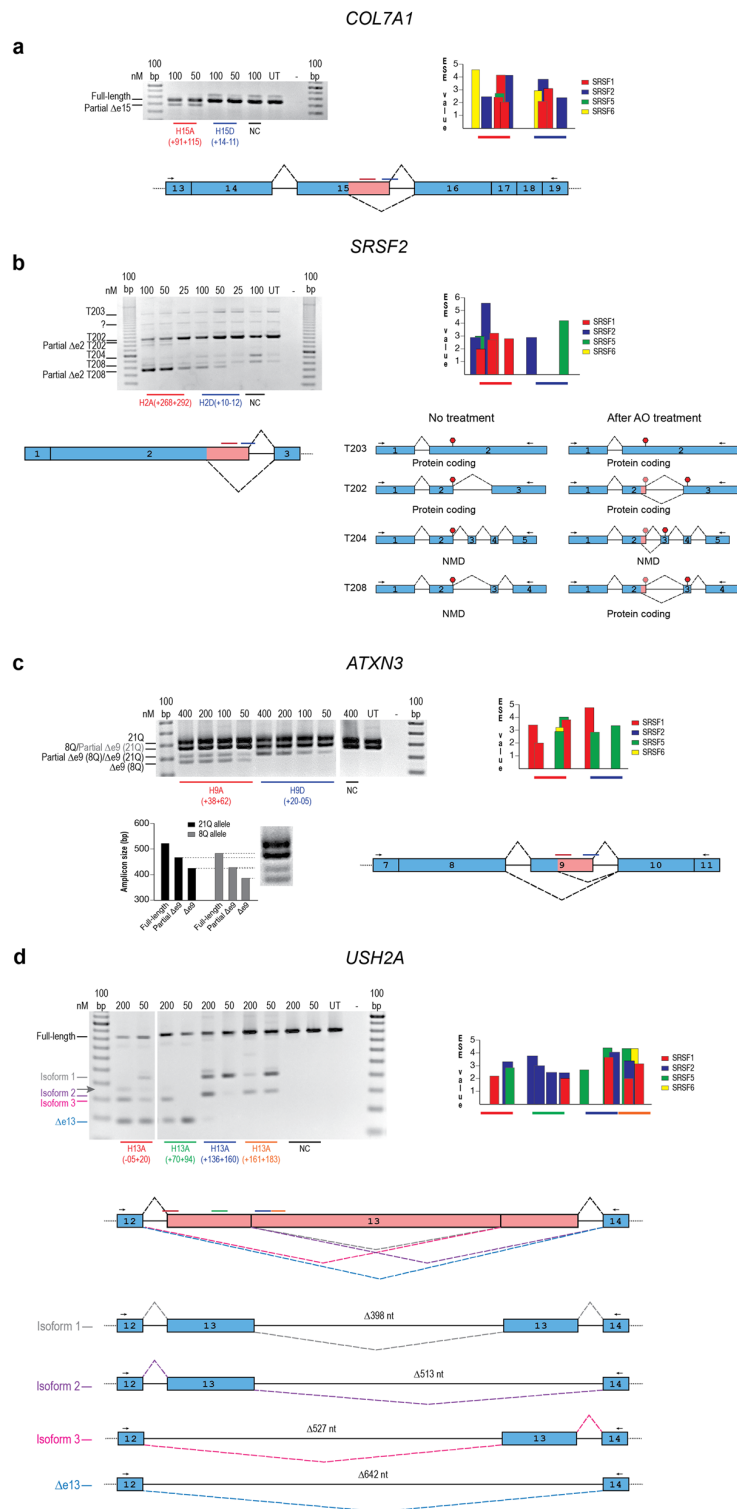
While most splice modulating AOs are designed with the intention to enhance exon selection or induce skipping of whole exons, the occasional activation of cryptic splice sites after in vitro AO treatment has also been observed. We have reported the activation of a cryptic donor splice site after treatment with an AO targeting *LMNA* pre-mRNA, promoting removal of 150 nt from the end of exon 11<sup>16</sup>. This precisely replicates the alternative *LMNA* transcript isoform that was reported to arise from recurrent pathogenic mutations within the cryptic splice motif<sup>17</sup>. Evers et al.<sup>18</sup> observed that an AO targeting exon 9 in *ATXN3* promoted a partial exon 9 skip, activating an alternative 5'ss. A partial exon 12 skip in the *HTT* transcript was also detected after treatment with an AO (World Patent WO2015053624A2); once again activating a cryptic donor splice site<sup>19,20</sup>, this time one that was previously observed to be used at low levels (3.2% of full-length) in normal human embryonic stem cells<sup>21</sup>. Lastly, we recently reported activation of two cryptic donor splice sites by AOs containing several locked nucleic acid residues, designed to enhance efficiency of exon skipping from the dystrophin transcript<sup>22</sup>.

In addition to the established roles that splice site motifs and exon enhancer and silencer motifs play in directing RNA splicing, there is increasing evidence of a similar role for RNA secondary structure<sup>23–26</sup>, and of its effect on splice factor binding<sup>27,28</sup>. While modelling the interactions of these phenomena presents a highly complex challenge, a reasonable starting point may be to assume that RNA secondary structure is generally antagonistic to splice factor binding within closed regions.

In our laboratory's quest to develop new therapeutics for debilitating genetic diseases, we have tested thousands of AOs targeted to numerous gene transcripts in a variety of cell types. We have confirmed AO-induced cryptic splicing events in the target transcripts in less than 0.2% of human cells, and only a single example in mouse cells<sup>29</sup>. In this study, we investigated the possible mechanisms by which AOs may induce cryptic splicing. We analyzed 12 AOs targeting six different human gene transcripts and found that changes to the accessibility of enhancer and silencer motifs within the transcript secondary structure appeared to play a role in many cases. The diverse nature of these changes indicates that there may be multiple pathways to inducing cryptic splicing, sometimes within a single exon.

## Results and discussion

To explore the possible mechanisms behind cryptic splice site activation, we analyzed AO-induced cryptic splicing events in six different human transcripts: *COL7A1*, *SRSF2*, *ATXN3*, *USH2A*, *HTT*, and *LMNA*. Data for *HTT* and *LMNA* were obtained from the literature and analyzed together with those from the remaining transcripts.



**Analysis of antisense oligonucleotide treatment.** *COL7A1 exon 15.* Antisense oligonucleotides (2'-O-methyl modified bases on a phosphorothioate backbone, (2'-OMe PS)) were transfected into healthy human fibroblasts as cationic lipoplexes at concentrations of 100 and 50 nM to induce skipping of exon 15 from the *COL7A1* pre-mRNA transcript, removing 144 nt from the full-length transcript (Fig. 1a). Subsequent RT-PCR analysis from exons 13 to 19 revealed both the full-length transcript and an unanticipated amplicon, smaller than full-length but larger than would be expected as a result of complete exon 15 removal. The unexplained amplicon was isolated and identified by Sanger sequencing to be missing the last 64 nucleotides from the 3' end of exon 15 (Supplementary Fig. S1). Removing 64 nt from the *COL7A1* transcript would render the cryptically spliced product out-of-frame, and therefore produce a premature termination codon in exon 16. This discovery highlights the importance of investigating unexpected splicing products after AO treatment. A new donor splice site was activated by treatment with an AO targeting *COL7A1* exon 15, H15A(+91+115), that resulted in cryptic splice site activation in 30% of the transcripts after transfection of the AO at both 100 nM and 50 nM. Treatment with this AO did not induce other aberrant splicing products. Transfection of cells with an AO covering the authentic donor splice site, H15D(+14-11), did not lead to cryptic donor site activation. Cryptic splice site activation was induced after the H15A(+91+115) AO was transfected into an immortalized human keratinocyte cell line (HaCaT) as cationic lipoplexes at concentrations of 400, 200, 100 and 50 nM, indicating that cryptic splice site activation after treatment with this AO is not cell-specific (Supplementary Fig. S2).

*SRSF2 exon 2.* Antisense oligonucleotides were transfected into healthy human fibroblasts as cationic lipoplexes at concentrations of 100, 50 and 25 nM to induce skipping of exon 2 from the *SRSF2* pre-mRNA transcript, removing 311 nt from the full-length transcript (Fig. 1b). Gel fractionation of the RT-PCR amplicons revealed several products confirmed by Sanger sequencing: full-length *SRSF2*-T204 (ENST00000452355.7); full-length *SRSF2*-T208 (ENST00000585202.5); and T208 missing 65 nt from the 3' end of exon 2. Multiple amplicons larger than 1000 nt were present that correspond to the amplicon sizes of the transcripts *SRSF2*-T203 (ENST00000392485.2) and *SRSF2*-T202 (ENST00000359995.10) (Fig. 1b). The splicing of T202 appears to be influenced by the AOs in the same manner (Fig. 1b). However, we were unable to isolate and identify various amplicons to confirm this. The AOs did not appear to cause exon skipping or cryptic donor site activation within the T203 transcript, most likely due to the T203 isoform containing only two exons, making both “unskippable”<sup>30</sup>. Cryptic splice site activation was induced after both H2A(+268+292) and H2D(+10-12) AOs were transfected into HaCaT cells and a human neuroblastoma cell line (SH-SY5Y) as cationic lipoplexes at concentrations of 400, 200, 100 and 50 nM, indicating that cryptic splice site activation after treatment with these AOs is not cell-specific (Supplementary Fig. S2).

Under normal conditions, *SRSF2* transcript isoforms T202 and T203 code for proteins while T208 and T204 undergo nonsense mediated decay (NMD). After AO treatment, the expression of the cryptically spliced T208 increased with a concomitant decrease in the full-length T202. The cryptic splicing of exon 2 removes the natural termination codon from T202, T204, and T208 and exposes a new in-frame termination codon in the following exon of each transcript (Fig. 1b).

Mammalian NMD generally follows the ‘50 nucleotide rule’, whereby termination codons more than 50 nt upstream of the final exon are determined premature and result in a reduction in mRNA abundance<sup>31</sup>. Cryptic splice site activation appears to stabilize T208 as a new termination codon is created within 50 nt of the penultimate 3' exon junction. Isoform T204 still appears to undergo NMD, as the new termination codon is exposed within the third exon of the five-exon isoform.

*ATXN3 exon 9.* Antisense oligonucleotides were transfected into healthy control human fibroblasts as cationic lipoplexes at concentrations of 400, 200, 100 and 50 nM to induce skipping of exon 9 from the *ATXN3* pre-mRNA, thereby removing 97 nt from the full-length transcript (Fig. 1c). Gel fractionation of the RT-PCR amplicons revealed two full-length product bands representing the two transcripts in the untreated sample: a larger product (533 nt) containing 21 CAG (21Q) repeats and a slightly smaller product (483 nt) containing eight CAG (8Q) repeats. Two additional smaller bands were observed in healthy human fibroblasts treated with H9A(+38+62) at all concentrations tested. The two bands were isolated and identified by Sanger sequencing (Supplementary Fig. S1). The smaller of the two amplicons was solely the result of complete exon 9 skipping from the 8Q transcript. The larger of the two amplicons is a similar size to complete exon 9 removal from the 21Q transcript. However, this amplicon was confirmed as resulting primarily from the activation of a cryptic donor site on position + 42 of exon 9, removing 55 nt from the 8Q transcript. Sanger sequencing revealed a minor secondary product with the removal of exon 9 entirely from the 21Q transcript. Treatment with H9D(+20-05) resulted in predominantly partial exon 9 skipping from the 8Q transcript and a low level of complete exon 9 skipping from the 21Q transcript.

Complete and partial exon 9 skipping was predominately observed in the 8Q compared with the 21Q transcript. Partial exon 9 skipping from the 8Q transcript and complete exon 9 skipping from the 21Q transcript produces products that differ by three nucleotides, and could not be distinguished on an agarose gel alone. Sanger sequencing confirmed that both transcripts were disproportionately represented, with lower levels of complete exon 9 skipping from the 21Q transcript. Partial exon 9 skipping from the 21Q transcript produces a product 16 nt smaller than the canonical 8Q transcript and could not be confirmed by the methods used in this study. Cryptic donor activation in the transcript with fewer CAG repeats dominates in some AO treatments but not others<sup>32,33</sup>. The CAG expansion occurs in the following exon 10, separated by a 10 kb intron from the AO target. Numerous studies assessing AO-mediated removal of exon 9 and/or exon 10 from the *ATXN3* transcript reported reduced exon skipping efficiencies the larger the expansion size. Although this phenomenon is directed more towards exon 10 removal, we speculate that the CAG repeat length may influence the cryptic splice site

Gene (exon)	Splice site	Canonical splice site sequence	Cryptic splice site sequence	HSF canonical splice site score	HSF cryptic splice site score	MaxEnt canonical splice site score	MaxEnt cryptic splice site score	Position relative to beginning of exon
<i>USH2A</i> (13)	Acceptor	ttttatcttttagGG	caacactgccagAT	88.04	80.44	8.95	-1.01	+527
	Donor	CAGgtaaga	AGTgtgagt	97.66	82.16	10.77	4.88	+129
<i>COL7A1</i> (15)	Donor	CGGgtcagg	CAGgtggct	88.19	78.49	4.01	2.97	+80
<i>ATXN3</i> (9)	Donor	AAAgtaaag	CAGgtacaa	74.37	<b>76.7</b>	1.6	<b>7.09</b>	+42
<i>SRSF2</i> (2)	Donor	TAAGtaatg	CAGgtcgcg	73.19	72.69	-0.64	<b>5.46</b>	+246
<i>HTT</i> * (12)	Donor	ATTgtaagt	CAGgtcagc	83.43	<b>92.8</b>	7.16	<b>8.54</b>	+206
<i>LMNA</i> * (11)	Donor	CAGgtgagt	CAGgtgggc	98.84	88.33	8.07	2.93	+120

**Table 1.** Comparing canonical and cryptic splice site scores using two different modeling approaches. Cryptic splice site scores that are higher than the canonical splice site scores are highlighted in bold.

usage frequency. The nature of the CAG repeat allows for numerous consecutive potential serine/arginine-rich splicing factor (SRSF) 2 (AGCAG) and SRSF5 (ACAGC) splice motifs. The fact that these positive exon selection sites are heavily repeated may influence exon 10 and potentially exon 9 selection and, therefore, susceptibility to AO-mediated exon skipping.

As *ATXN3* is ubiquitously expressed, AO-mediated cryptic splice site activation was tested in both HaCaT and SH-SY5Y cells. The number of repeats for each cell line was determined via Sanger sequencing: heterozygous for 19Q and 18Q transcripts in the HaCaT cells and homozygous for 10Q transcript in the SH-SY5Y cells. Antisense oligonucleotides were transfected as cationic lipoplexes at concentrations of 400, 200, 100 and 50 nM (Supplementary Fig. S2). The H9D(+20-05) AO targeting the donor site activated the cryptic 3'ss in both cell lines, but cryptic splice site activation was not apparent after treatment with the H9A(+38+62) AO. Although, without testing both AOs in multiple cell types from the same healthy control donor, it cannot be determined if the discrepancy in cryptic splice site activation is due to the cell type or some other factors.

*USH2A* exon 13. Antisense oligonucleotides were transfected into a Huh7 cell line as cationic lipoplexes at concentrations of 200 and 50 nM to induce skipping of exon 13 from the *USH2A* pre-mRNA transcript (Fig. 1d). Subsequent RT-PCR analysis revealed multiple unanticipated amplicons larger than expected from the removal of exon 13 in its entirety. It was confirmed by Sanger sequencing that multiple splicing events occurred: removal of the complete exon 13 (Fig. 1d  $\Delta$ e13); activation of a cryptic donor (Fig. 1d isoform 2); activation of a cryptic acceptor (isoform 3); or activation of both cryptic donor and acceptor sites within exon 13 (Fig. 1d isoform 1), after treatment with different AOs (Supplementary Fig. S1). Treatment with H13A(-05+20) and H13A(+70+94) resulted mainly in complete exon 13 exclusion, removing 642 nt from the full-length transcript (Fig. 1d  $\Delta$ e13), and the activation of a cryptic acceptor site, removing 527 nt from the full-length transcript (Fig. 1d isoform 3). Treatment with H13A(+136+160) and H13A(+161+183) resulted in the activation of a cryptic donor site, both on its own (missing 513 nt from the 3' end of exon 13; Fig. 1d isoform 2) and in conjunction with the cryptic acceptor site (missing 398 nt from the middle of exon 13; Fig. 1d isoform 1) but did not remove the entire exon 13. We were unable to isolate and identify one of the amplicons by Sanger sequencing (labelled with a grey arrow in Fig. 1d). We speculate that this amplicon is a heteroduplex, which would explain why it could not be isolated.

The *USH2A* expression profile is limited to a small subset of tissue types (eye, heart muscle, liver, and testis) that were not available for use at the time of this study. Thus AO-induced cryptic splicing was not investigated in additional cell types.

*LMNA* exon 11. Lou et al.<sup>16</sup> sought to induce cryptic splicing through AO-mediated splice-switching by designing a panel of AOs to anneal across exon 11 of the *LMNA* gene transcript in human myogenic cells. Initially, 2'-OMe PS AOs were tested at concentrations of 400, 200 and 100 nM as cationic lipoplexes. The transfection of several different AOs resulted in the cryptically spliced  $\Delta$ 150 transcript and whole exon 11 removal. Transfection of the H11A(+221+245) AO sequence resulted predominantly in  $\Delta$ 150 transcript expression and was thus synthesized as a phosphorodiamidate morpholino oligomer, producing even more specific and potent cryptic splicing activation. This finding, along with the ability of AOs containing several locked nucleic acids to activate cryptic donor splice sites from the dystrophin transcript<sup>22</sup>, highlights that cryptic splicing can be activated by AOs comprised of various backbone chemistries and sugar modifications.

*HTT* exon 12. As a potential treatment for Huntington's disease, a 2'-OMe PS AO was developed to reduce the levels of toxic huntingtin protein by activating a cryptic donor splice site, removing 135 nt from the 3' end of exon 12<sup>19,20</sup>. Antisense oligonucleotides were transfected into Huntington's disease patient-derived fibroblasts as cationic lipoplexes at various concentrations and resulted in a dose-dependent partial exon 12 skipping (150 nM 95% skipping; 25 nM 92% skipping, and 1 nM 16% skipping) except at the highest concentration where no exon skipping was evident (1000 nM 0% skipping)<sup>20</sup>.

**Analysis of splice site scores and exonic splicing enhancer motifs masked by the examined antisense oligonucleotides.** Two models were employed to calculate the scores of both the canonical and cryptic splice sites activated after AO treatment: a weight matrix model, Human Splice Finder 3.1<sup>34</sup>, and a maxi-

Gene	AO nomenclature	SRSF1 (SF2)	SRSF2 (SC35)	SRSF5 (SRp40)	SRSF6 (SRp55)
USH2A	H13A(-05+20)	1	1	1	0
	H13A(+70+94)	1	4	0	0
	H13A(+136+160)	1	1	2	0
	H13A(+161+183)	2	2	1	1
COL7A1	H15A(+91+115)	3	2	1	1
	H15D(+14-11)	2	2	0	1
ATXN3	H9A(+38+62)	3	0	2	1
	H9D(+20-05)	1	0	2	0
SRSF2	H2A(+268+292)	4	2	1	0
	H2D(+10-12)	0	1	1	0
HTT	H12A(+269+297) <sup>a</sup>	3	4	3	0
LMNA	H11A(+221+245) <sup>a</sup>	3	3	1	1

**Table 2.** Exonic splicing enhancer motifs masked by the antisense oligonucleotides examined in this study. <sup>a</sup>Not tested in this study; published results.

num entropy model, MaxEntScan<sup>35</sup>. No discernable pattern became evident using either model (Table 1), indicating splice site scores are not the only factor influencing splice site usage. Various cryptic splice site scores were higher when compared to canonical splice site scores, but again, with the small number of examples covered in this study, no pattern could be deduced. Included in Table 1 are the canonical and cryptic splice site sequences recognized by the spliceosome in the examples reported here. The CAG cryptic 3'ss is activated in the *USH2A* transcript after AO treatment. During U2-type canonical splicing of human transcripts, CAG 3'ss are more frequently used by the spliceosome than TAG 3'ss (64.55% versus 29.01%)<sup>36</sup>. Except for the *USH2A* transcript, all the studied activated cryptic 5'ss comprise the CAGgt sequence. Additionally, the canonical and cryptic 5'ss are strikingly similar in the *LMNA* example.

Exonic splicing enhancer (ESE) motifs masked by AO binding sites were tallied using ESEFinder 3.0<sup>37</sup>; (Fig. 1; Table 2). Motifs were considered when one or more motif nucleotides were masked by the targeting AO, as even partially covering a motif by two nucleotides influences splice outcome<sup>38</sup>. The examined AOs were found to consistently mask SRSF1 motifs, with exception of the AO H2D(+10-12) targeting the *SRSF2* exon 2 donor site.

The splicing factor SRSF1 is necessary for several splicing processes, including lariat formation and 5'ss cleavage<sup>39</sup>. In addition, SRSF1 assists in modulating 5'ss selection<sup>39</sup>. The addition of purified SRSF1 to cultured cells favored 5'ss located more proximally to the 3'ss while lower levels of SRSF1 favored 5'ss located distal to the 3'ss<sup>40</sup>. In our study, AOs can mask the availability of ESE motif binding sites, therefore reducing the amount of SRSF1 that can bind to the pre-mRNA. Fewer SRSF1 binding sites may drive the 5'ss preference away from the canonical splice site towards a more distal cryptic splice site.

### Analysis of AO-induced changes to exonic splicing enhancer/silencer access within cryptically spliced exons.

It is notable that all seven of the observed cryptic splice sites fell within the affected exons, between the canonical splice sites, rather than in the downstream or upstream introns. We suggest that this is a logical consequence of the 'exon definition' paradigm under which the human spliceosome is thought to operate, whereby transcript sequence between the first and last exons is processed as intron unless specifically defined as being part of an internal exon<sup>41</sup>. Because 'intron' is the default sequence identity under this paradigm, AO binding is therefore much more likely to diminish an existing exon signal than it is to spontaneously extend it.

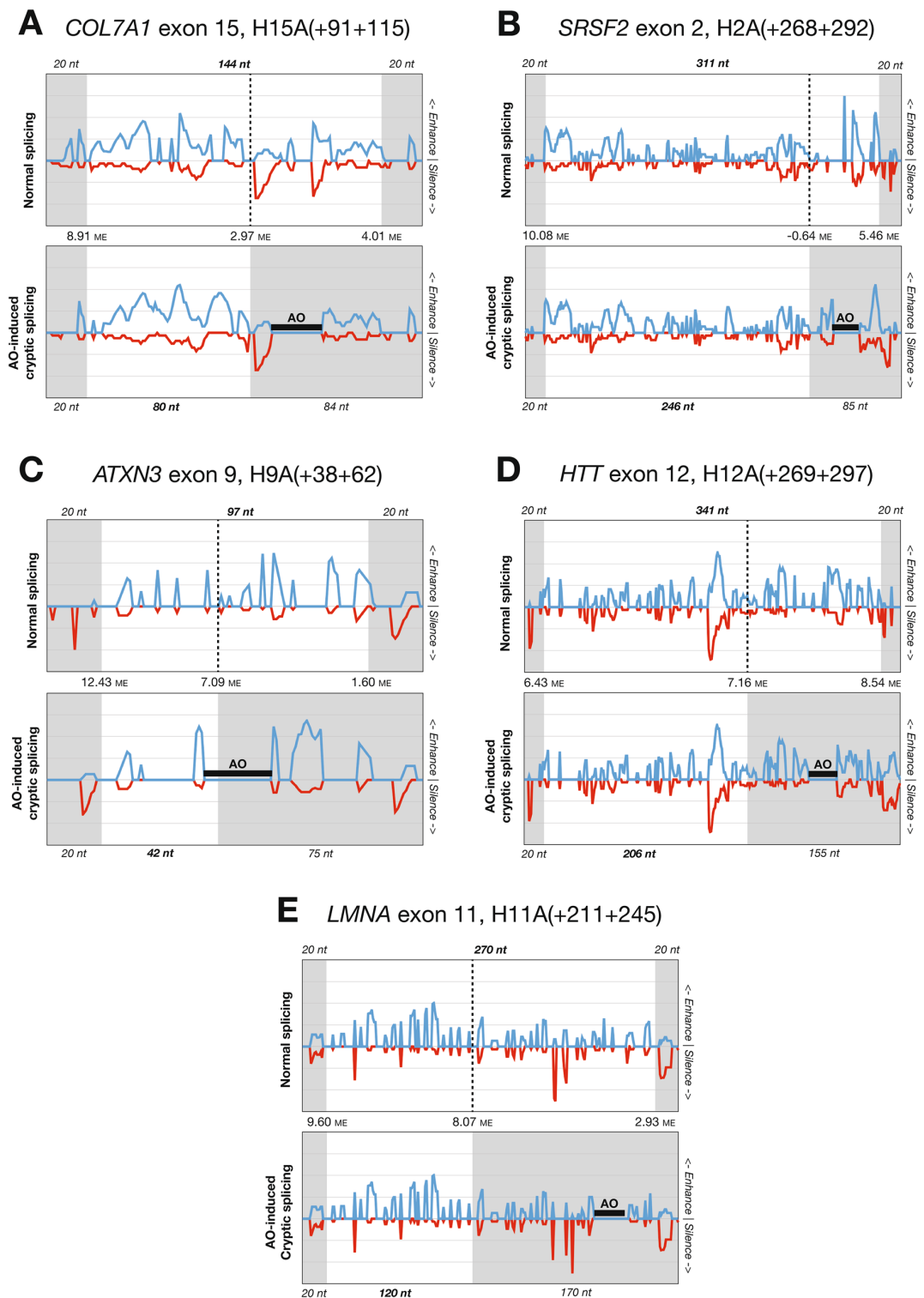
Because four of the seven cryptic splice sites had MaxEnt scores lower than their canonical counterparts, it was clear that our analysis would need to encompass other variables in order to explain the activation of these sites—specifically, those variables that could plausibly be altered by AO binding. We therefore attempted to model the effect that AO binding would have on both the local secondary structure of the transcript, and the subsequent change in accessibility to ESE and exon splicing silencer (ESS) motifs.

Two of the eleven cryptic-splice inducing AOs, SRSF2 H2D(+10-12) and ATXN3 H9D(+20-05), were excluded from this modelling, as we reasoned that simple steric blocking of the target exon donor sites was the most likely explanatory mechanism in those cases.

The ESE and ESS motifs for each cryptically spliced exon were overlaid to generate enhancer and silencer scores at each nucleotide position. These values were then "masked" by the predicted secondary structure for the exons, effectively resetting the ESE and ESS scores to zero for all nucleotides predicted to bind other nucleotides. This masking was repeated with the altered structures predicted for on-target AO binding, and the two plots were vertically aligned to allow comparison between them (Fig. 2A-E). Because the size of *USH2A* exon 13 (642 nt) made it impractical to visually compare changes in its ESE and ESS access in the same manner as for the other exons, we elected to present only the net changes in ESE and ESS access as a result of AO binding (Fig. 2F-G).

We acknowledge that there are impediments to the accuracy of this approach. Individually, HSF 3.1 and RNAfold are imperfect predictors that encompass only a fraction of the RNA interactions occurring within living cells, and neither account for more complex factors, such as RNA tertiary structure or local ribonucleo-protein context. However, despite their limitations, these two utilities have proven instrumental for numerous scientific reports over the past decade and have amassed a combined total of over 4000 citations. We therefore





**Figure 2.** Changes to predicted exon splicing enhancer/silencer (ESE/ESS) access in five examples of antisense oligonucleotide (AO)-induced cryptic splicing of canonical exons. Blue lines indicate ESE access and red lines indicate ESS access. Grey shading indicates pre-mRNA sequence excluded from the mature transcript. Region sizes and Maximum Entropy scores for cryptic and canonical splice sites are also shown.

reasoned that integrating the predictions of these two well-tested programs might prove more informative than their individual outputs.

In *COL7A1* exon 15 (Fig. 2A), AO binding was predicted to increase ESE access in the retained 5' segment, as well as directly competing with ESEs in the excised 3' segment. The net effect was a much stronger exon signal from the 5' segment that improved the profile of the cryptic donor site. This example demonstrates that blocking an authentic donor site does not automatically activate a cryptic donor site; additional elements, including secondary structure and exon and intron definition motifs, are necessary to define the exon boundary.

For *SRSF2* exon 2 (Fig. 2B), the AO directly obscured the strongest enhancer peak in the excised 3' segment and induced a moderate increase in ESE access within the retained 5' segment. We also observed that, in the absence of AO binding, the enhancer signal in the excised 3' segment of the exon was substantially stronger than in the rest of the exon. This may be a positively selected feature to ensure inclusion of this segment and avoidance of the cryptic splice site, though it is not clear why the very poor MaxEnt score of the cryptic donor is not a sufficient deterrent alone.

In *ATXN3* exon 9 (Fig. 2C), the AO binding site overlapped the cryptic donor site and caused loss of ESE access 3' of the cryptic donor and a slight increase of ESE access immediately 5' of the cryptic donor. This, combined with the much stronger MaxEnt score of the cryptic site, may have been enough to shift exon definition to the 5' region of the exon. Partial occlusion of the cryptic donor site by this AO may also explain why it induces whole exon skipping in some fibroblast transcripts (Fig. 1c), as this would sterically block spliceosome binding.

In *HTT* exon 12 (Fig. 2D), the changes in secondary structure did not clearly favor either enhancement or silencing of the excised segment. However, ESS access was increased both 5' and 3' of the canonical donor site, and this appears to have been sufficient to tip the balance towards the comparably strong cryptic donor splice site.

A similar change to *HTT* exon 12 appears to have occurred in *LMNA* exon 11 (Fig. 2E), with the exception that the cryptic donor site in this exon was much stronger than its canonical neighbor.

For *USH2A* exon 13 (Fig. 3), all four AOs induced use of varying combinations of the two canonical splice sites, an internal cryptic donor site, and an internal cryptic acceptor site. In examining the effects of the four AOs, we noted that they appeared to group together as two pairs. The first two AO sequences, H13A(-05+20) and H13A(+70+94), were targeted 5' of the cryptic donor site and predominantly induced splice-switching from the canonical to the cryptic acceptor site. Conversely, the second two AOs, H13A(+136+160) and H13A(+161+183), were both targeted 3' to the cryptic donor site and induced its activation, splice-switching away from the canonical donor site. This is consistent with our earlier observation that the orientation of the AO target site relative to the cryptic donor site appears to be a key determinant of the AO's effect. The second pair of AOs primarily act to enhance the 5' cryptic donor site, in much the same way as the examples shown in Fig. 2, while the first pair of AOs act to silence the canonical acceptor site. Both these splicing effects are further complicated by the presence of the internal cryptic acceptor site that provides an alternative partner for the canonical donor site, and by the distance between the two cryptic sites (398 nt), which allows sufficient separation for both to be activated within the same transcript (see also Fig. 1d, isoform 1).

It appears that some facets of exon definition are unique to large internal exons and that these can only be properly understood by studying splicing in similarly sized exons from other genes. Exons longer than 500 nt, such as *USH2A* exon 13, typically rely on intron definition rather than exon definition in order to achieve correct splicing, but this intron-defined splicing can become inefficient when the intron size exceeds 500 nt<sup>42,43</sup>. It is possible that sporadic splice site activation in this larger exon is partly due to the inability of the spliceosome to utilize intron definition, and thus inefficiently creates exon isoforms of less than 500 nt by activating various internal splice sites, regardless of their strength.

There is accumulating evidence that long non-coding RNA (lncRNA) plays a role in post-transcriptional modification, including splicing<sup>44</sup>. In most cases, lncRNA contains sequence motifs or scaffolds that can recruit splicing factors to promote or restrict splicing<sup>44</sup>. We cannot rule out that the introduction of AOs to the cells may have caused a disturbance to the lncRNA and led to the observed cryptic splicing. It is also possible that the AOs have become part of the splicing complex as non-coding RNA and shifted the whole paradigm. At this stage, the results are inconclusive as only the AOs targeting *COL7A1*, *SRSF2* and *HTT* showed some similarity towards lncRNAs with no mention of splicing involvement.

## Conclusions

Despite the small number of examples of AO-induced cryptic splicing, we observed considerable diversity in the etiology of this phenomenon. However, a common feature appears to be disruption of the exon definition signal.

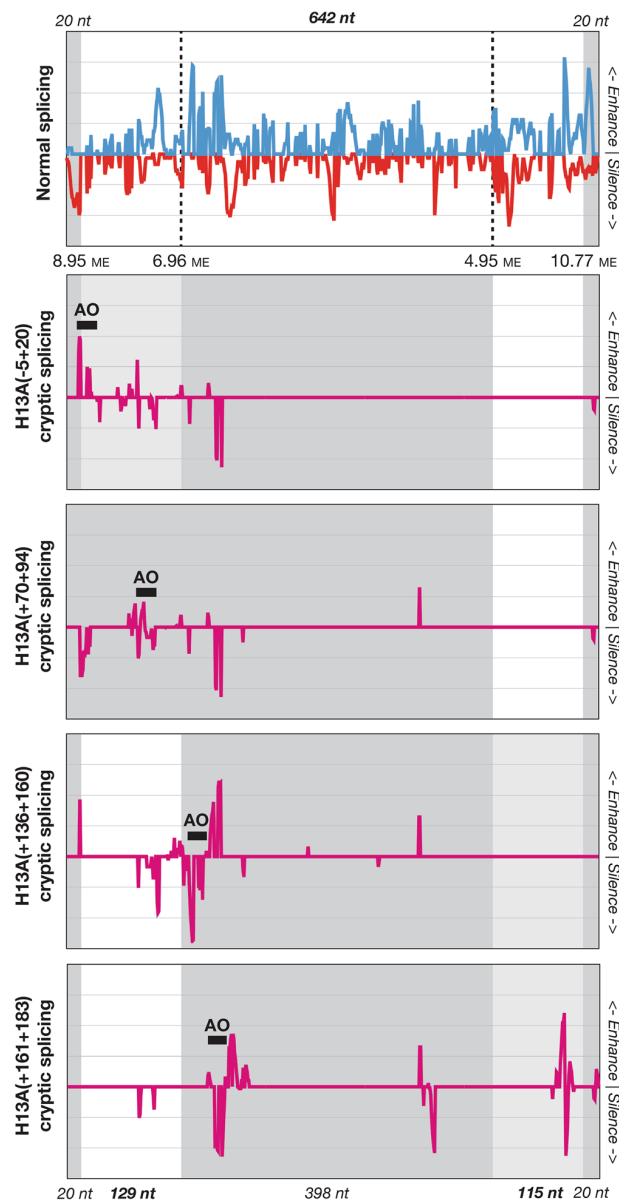
It is clear that canonical exon definition is achieved not by any single motif, but by the cumulative signal of multiple enhancers binding with regularity and consistency along the entire exon span. Furthermore, continuity of this enhancing signal appears to be just as important, if not more important, than its overall strength. This continuity is especially crucial when the exon contains a cryptic splice site, as this is often the only metric by which the spliceosome can distinguish the cryptic site from its canonical neighbor.

## Methods

All methods were carried out in accordance with relevant guidelines and regulations.

**Antisense oligonucleotides (AOs).** Antisense oligonucleotides (AOs) comprising of 2'-O-methyl modified bases on a phosphorothioate backbone (2'-OMe PS) were synthesized by TriLink BioTechnologies (San Diego, CA) or synthesized in-house on an Expedite 8909 Nucleic Acid synthesizer (Applied Biosystems, Melbourne, Australia) using the 1 μmol thioate synthesis protocol, as described previously<sup>45</sup>. After synthesis, the oligonucleotides were cleaved from the support following incubation in ammonium hydroxide for a minimum

### USH2A exon 13



**Figure 3.** Net changes to predicted exon splicing enhancer/silencer (ESE/ESS) access in four examples of antisense oligonucleotide (AO)-induced cryptic splicing of *USH2A* exon 13. Blue lines indicate ESE access and red lines indicate ESS access, purple indicates the net change in ESE and ESS access as a result of AO binding. Grey shading indicates pre-mRNA sequence excluded from the mature transcript, and pale grey indicates regions with intermediate inclusion. Region sizes and Maximum Entropy scores for cryptic and canonical splice sites are also shown.

of 24 h at room temperature. The 2'-OMe PS AOs were subsequently desalted under sterile conditions on NAP-10 columns (GE Healthcare, Sydney, Australia) according to manufacturer's instructions. The 2'-OMe PS AOs used in this study are listed in Table 3. Oligonucleotide nomenclature is based on that described by Aung-Htut et al.<sup>46</sup> and Mann et al.<sup>47</sup>, indicating the intron:exon, exon or exon:intron annealing coordinates in the target gene pre-mRNA.

**Cell culture and transfection.** All cell culture reagents were purchased from Gibco, (Thermo Fisher Scientific, Scoresby, Australia), unless otherwise stated. Primary dermal fibroblasts were derived from a healthy



Gene	AO nomenclature	Sequence (5' to 3')
USH2A	H13A(-05+20)	GCAAUGAUCACACCUAAGCCCUAAA
	H13A(+70+94)	GAGCCAUGGAGGUUACACUGGCAGG
	H13A(+136+160)	UGAAGUCCUUUGGUUCUUUUUUGC
	H13A(+161+183)	AGUUUUCUCUGCAGGUGUCACAC
COL7A1	H15A(+91+115)	CCCUCUCUCUGCCUCGAGUACCG
	H15D(+14-11)	CAGGGCCUGACCCGUUCGAGCCACG
ATXN3	H9A(+38+62)	UUCUGAAGUAAGAUUUGUACCUGAU
	H9D(+20-05)	UUUACUUUCAAAGUAGGCUUCUCG
SRSF2	H2A(+268+292)	CGCUCCUUCUCUUCAGGAGACUUG
	H2D(+10-12)	CCCAGACAUUACCAUUUCUUA
HTT	H12A(+269+297) <sup>a20</sup>	CGGUGGUGGUCUGGGAGCUGCGUGAUG
LMNA	H11A(+221+245) <sup>a16</sup>	AGGAGGUAGGAGCGGGUGACCAGAU
	Negative control	CCUCUUACCUCAGUACA AUUUAUA

**Table 3.** Information for AOs. <sup>a</sup>Not tested in this study; published results.

Cell strain	Propagation/seeding media	Transfection seeding density
Primary dermal fibroblasts	Dulbecco's modified Eagle's medium supplemented with 1% GlutaMax™-I and 10% FBS	1.8 × 10 <sup>4</sup> cells/well
Huh7	Dulbecco's modified Eagle's medium supplemented with 10% FBS	5 × 10 <sup>4</sup> cells/well
SH-SY5Y	1:1 mixture of Eagle's Minimum Essential Medium and Ham's F-12 medium supplemented with 10% FBS	7 × 10 <sup>4</sup> cells/well
HaCaT	Dulbecco's modified Eagle's medium supplemented with 10% FBS	3 × 10 <sup>4</sup> cells/well

**Table 4.** Culture conditions for cell strains used in the study. All cells were maintained at 37 °C in a 5% CO<sub>2</sub> atmosphere. Cells were seeded 24 h before transfection in a 24-well plate. Fetal bovine serum (FBS) (Scientific, Cheltenham, Australia).

volunteer after informed consent (The University of Western Australia Human Research Ethics Committee approval RA/4/1/2295; Murdoch University Human Research Ethics Committee approval 2013/156). The human hepatocarcinoma cell line, Huh7, was supplied by the Japanese Collection of Research Bioresources Cell Bank (Osaka, Japan) and purchased from CellBank Australia (Westmead, Australia). The human neuroblastoma cell line, SH-SY5Y, was supplied by ATCC (Gaithersburg, MD) and purchased from In Vitro Technologies (Canning Vale, Australia). HaCaT cells were purchased from AddexBio (San Diego, CA). Culture conditions and transfection seeding density are described in Table 4.

All cell strains were transfected with 2'-OMe PS AO-Lipofectamine 3000 (Thermo Fisher Scientific) lipoplexes in Opti-MEM (Gibco) according to the manufacturer's instructions, at various concentrations in duplicate wells, and the cells were then incubated at 37 °C in a 5% CO<sub>2</sub> atmosphere for 24 h before RNA extraction. The negative control oligomer (sequence from Gene Tools, LLC synthesized as a 2'-OMe PS AO) that targets a human beta-globin intron mutation was used as a negative transfection control.

**Molecular analysis.** After harvesting the cells, total RNA was extracted using MagMax nucleic acid isolation kit (AM1830; Thermo Fisher Scientific) according to manufacturer's instructions and included the DNase treatment step. Molecular analyses were accomplished using three different systems optimized for different gene targets. SuperScript III One-Step RT-PCR System with Platinum Taq DNA Polymerase (Thermo Fisher Scientific) was used to synthesize and amplify cDNA from 50 ng of total RNA in a single step. Nested PCR was necessary to amplify the *USH2A* transcripts. Briefly, after 20 cycles of amplification, 1 µl aliquot was removed and subjected to nested PCR for 25 cycles using AmpliTaq Gold (Thermo Fisher Scientific) and an inner primer set. For regions with a high GC-content that are more difficult to amplify, SuperScript IV First-Strand Synthesis System and random hexamers (Thermo Fisher Scientific) were used to synthesize cDNA from harvested total RNA, and approximately 50 ng of cDNA was used as a template for PCR amplification using the TaKaRa LA Taq DNA Polymerase with GC Buffer II system (Takara Bio USA, Inc., Clayton, Australia). PCR systems, conditions and primers used to assess splice modulation across the different gene transcripts are summarized in Table 5.

Amplified RT-PCR products were resolved on 2% agarose gels by electrophoresis in Tris-acetate ethylenediaminetetraacetic acid buffer, compared to a 100 bp DNA size standard (Geneworks, Adelaide, Australia). Relative transcript abundance was estimated by densitometry on images captured by the Fusion FX system (Vilber Lourmat, Marne-la-Vallée, France) using Fusion-Capt software and ImageJ (version 1.8.0\_112) software for densitometry analysis. To identify RT-PCR products, the amplicons were first isolated by bandstab<sup>48</sup>, followed by template preparation using Diffinity RapidTip for PCR Purification (Diffinity Genomics, Inc., West Henrietta, NY) and DNA sequencing, performed by the Australian Genome Research Facility Ltd. (Nedlands, Australia).

Gene target (accession numbers)	Primer orientation	Sequence (5'-3')	Length (nt)	PCR system	Cycling conditions
ATXN3 (NM_004993.6)	Exon 7F	GTCCAACAGATGCATCGACCAA	522 (21Q) 516 (19Q)	SSIII One-Step	55 °C (30 min) and 94 °C (2 min); 28 cycles of 94 °C (30 s), 55 °C (30 s) and 68 °C (1.5 min)
	Exon 11R	AGCTGCCTGAAGCATGTCTTCTT	513 (18Q) 489 (10Q) 483 (8Q)		
COL7A1 (NM_000094.4)	Exon 13F	CTTAGCTACACTGTGCGGGT	765	SSIII One-Step	55 °C (30 min) and 94 °C (2 min); 30 cycles of 94 °C (30 s), 60 °C (30 s) and 68 °C (1.5 min)
	Exon 19R	TGGGAGTATCTGGTGCTCA			
SRSF2 (XR_429913.4)	Exon 1F	CCCAGAGCTGAGGAAGCC	850	SSIV TaKaRa GC I	94 °C (1 min); 32 cycles of 94 °C (30 s), 62 °C (30 s) and 72 °C (4 min)
	Exon 4R	CTCAACTGCTACACAACCTGC			
USH2A (NM_206933.4)	Exon 12F	AAGAGTTGGATCCTGATGGCTGC	993	SSIII One-Step	55 °C (30 min) and 94 °C (2 min); 20 cycles of 94 °C (15 s), 60 °C (30 s) and 68 °C (1 min)
	Exon 15R	GACAGGTTTCATCAAGGCTCC			
	Exon 12F	CTGTAACCTGCAATACCTCTGG	837	AmpliTaq Gold	94 °C (5 min); 25 cycles of 94 °C (30 s), 60 °C (30 s) and 72 °C (1 min); 72 °C (5 min)
	Exon 14R	CAAACACACTGACCAGTCAGG			

**Table 5.** List of primers, PCR system and conditions used in this study.



**Figure 4.** Example of exonic splicing enhancer (ESE) and exonic splicing silencer (ESS) score calculations for an RNA nucleotide. An RNA nucleotide, N, indicated with a rectangular box, is assigned an ESE score as the sum of its contributions to any overlapping enhancer motifs, indicated with ‘e’ characters and blue text, and an ESS score as the sum of its contributions to any overlapping silencer motifs, indicated with ‘s’ characters and red text. The ‘Net score,’ shown in purple text, is determined as the sum of the ESE and ESS scores.

**In silico analysis.** Basic Local Alignment Search Tool (BLAST)<sup>49</sup> was used to compare amplicon sequences to the reference mRNA sequences (accession numbers: Table 5). ESEFinder 3.0<sup>34</sup> was used to evaluate ESE motifs masked by AO binding sites. Motifs were considered when one or more motif nucleotides were masked by the targeting AO. Human Splice Finder 3.1<sup>34</sup> and MaxEntScan<sup>35</sup> were employed to calculate the scores of both the canonical and cryptic splice sites activated after treatment with each AO. Sequences for each cryptically spliced exon and  $\pm 20$  nt of flanking intron were input to Human Splice Finder 3.1<sup>34</sup> which generated a JSON file with the locations of every detected ESE and ESS motif, as well as predicted acceptor and donor splice sites. Raw text from this JSON file was then imported into a custom-made spreadsheet (see Supplementary Material) that used this data to assign an ESE and an ESS score to each nucleotide of the sequence, under the following rationale:

ESE score:  $+ 1/n$  for each overlapping ESE motif, where  $n$  = ESE motif length.  
 ESS score:  $- 1/n$  for each overlapping ESS motif, where  $n$  = ESS motif length.

For example, a nucleotide that fell within two six nt ESE motifs and one eight nt ESS motif would be assigned an ESE score of 0.333 ( $2 \times 1/6$ ) and an ESS score of  $- 0.125$  ( $1 \times - 1/8$ ). An example diagram of this calculation is provided in Fig. 4.

Predicted centroid normal RNA folding was calculated for the sequence of each cryptically spliced exon with  $\pm 70$  nt flanking intron, using RNAfold<sup>50</sup> with the “avoid isolated base pairs” option. Predicted centroid

AO-induced folding was calculated for each exon using the same sequence and settings as for normal folding, but with an additional constraint mask that prohibited binding within the AO target sites.

### Data availability

All data generated or analyzed during this study are included in this published article (and its Supplementary Information file).

Received: 11 January 2021; Accepted: 14 July 2021

Published online: 23 July 2021

### References

- Ward, A. J. & Cooper, T. A. The pathobiology of splicing. *J. Pathol.* **220**, 152–163. <https://doi.org/10.1002/path.2649> (2010).
- Hang, J., Wan, R., Yan, C. & Shi, Y. Structural basis of pre-mRNA splicing. *Science* **349**, 1191–1198. <https://doi.org/10.1126/science.aac8159> (2015).
- Sperling, R. The nuts and bolts of the endogenous spliceosome. *WIREs RNA* **8**, e1377. <https://doi.org/10.1002/wrna.1377> (2017).
- Papasaïkas, P. & Valcarcel, J. The spliceosome: The ultimate RNA chaperone and sculptor. *Trends Biochem. Sci.* **41**, 33–45. <https://doi.org/10.1016/j.tibs.2015.11.003> (2016).
- Turunen, J. J., Niemela, E. H., Verma, B. & Frilander, M. J. The significant other: Splicing by the minor spliceosome. *WIREs RNA* **4**, 61–76. <https://doi.org/10.1002/wrna.1141> (2013).
- Matera, A. G. & Wang, Z. A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* **15**, 108–121. <https://doi.org/10.1038/nrm3742> (2014).
- Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18**, 437–451. <https://doi.org/10.1038/nrm.2017.27> (2017).
- Kelemen, O. *et al.* Function of alternative splicing. *Gene* **514**, 1–30. <https://doi.org/10.1016/j.gene.2012.07.083> (2013).
- Lee, Y. & Rio, D. C. Mechanisms and regulation of alternative pre-mRNA splicing. *Annu. Rev. Biochem.* **84**, 291–323. <https://doi.org/10.1146/annurev-biochem-060614-034316> (2015).
- Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The expanding landscape of alternative splicing variation in human populations. *Am. J. Hum. Genet.* **102**, 11–26. <https://doi.org/10.1016/j.ajhg.2017.11.002> (2018).
- Nelson, K. K. & Green, M. R. Mechanism for cryptic splice site activation during pre-mRNA splicing. *Proc. Natl. Acad. Sci. USA* **87**, 6253–6257. <https://doi.org/10.1073/pnas.87.16.6253> (1990).
- Haj Khelil, A., Deguillien, M., Moriniere, M., Ben Chibani, J. & Baklouti, F. Cryptic splicing sites are differentially utilized in vivo. *FEBS J.* **275**, 1150–1162. <https://doi.org/10.1111/j.1742-4658.2008.06276.x> (2008).
- Buratti, E., Chivers, M., Hwang, G. & Vorechovsky, I. DBASS3 and DBASS5: Databases of aberrant 3'- and 5'-splice sites. *Nucleic Acids Res.* **39**, D86–91. <https://doi.org/10.1093/nar/gkq887> (2011).
- Krawczak, M. *et al.* Single base-pair substitutions in exon-intron junctions of human genes: Nature, distribution, and consequences for mRNA splicing. *Hum. Mutat.* **28**, 150–158. <https://doi.org/10.1002/humu.20400> (2007).
- Aartsma-Rus, A. *et al.* Guidelines for antisense oligonucleotide design and insight into splice-modulating mechanisms. *Mol. Ther.* **17**, 548–553. <https://doi.org/10.1038/mt.2008.205> (2009).
- Luo, Y.-B. *et al.* Antisense oligonucleotide induction of progerin in human myogenic cells. *PLoS ONE* **9**, e98306. <https://doi.org/10.1371/journal.pone.0098306> (2014).
- Eriksson, M. *et al.* Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature* **423**, 293–298. <https://doi.org/10.1038/nature01629> (2003).
- Evers, M. M. *et al.* Ataxin-3 protein modification as a treatment strategy for spinocerebellar ataxia type 3: Removal of the CAG containing exon. *Neurobiol. Dis.* **58**, 49–56. <https://doi.org/10.1016/j.nbd.2013.04.019> (2013).
- Evers, M. M. *et al.* Preventing formation of toxic N-terminal huntingtin fragments through antisense oligonucleotide-mediated protein modification. *Nucleic Acid Ther.* **24**, 4–12. <https://doi.org/10.1089/nat.2013.0452> (2014).
- van Roon-Mom, W. M., Evers, M. M., Peppers, B. A., Aartsma-Rus, A. & Van Ommen, G. J. Antisense oligonucleotide directed removal of proteolytic cleavage sites, the hchwa-d mutation, and trinucleotide repeat expansions. WO201505362A2. WIPO (2014).
- Ruzo, A. *et al.* Discovery of novel isoforms of huntingtin reveals a new hominid-specific exon. *PLoS ONE* **10**, e0127687. <https://doi.org/10.1371/journal.pone.0127687> (2015).
- Zaw, K. *et al.* Consequences of making the inactive active through changes in antisense oligonucleotide chemistries. *Front. Genet.* **10**, 1249. <https://doi.org/10.3389/fgene.2019.01249> (2019).
- Jin, Y., Yang, Y. & Zhang, P. New insights into RNA secondary structure in the alternative splicing of pre-mRNAs. *RNA Biol.* **8**, 450–457. <https://doi.org/10.4161/rna.8.3.15388> (2011).
- Shilo, A., Tosto, F. A., Rausch, J. W., Le Grice, S. F. J. & Misteli, T. Interplay of primary sequence, position and secondary RNA structure determines alternative splicing of LMNA in a pre-mature aging syndrome. *Nucleic Acids Res.* **47**, 5922–5935. <https://doi.org/10.1093/nar/gkz259> (2019).
- Soemedi, R. *et al.* The effects of structure on pre-mRNA processing and stability. *Methods* **125**, 36–44. <https://doi.org/10.1016/j.ymeth.2017.06.001> (2017).
- Zhang, J., Kuo, C. C. & Chen, L. GC content around splice sites affects splicing through pre-mRNA secondary structures. *BMC Genomics* **12**, 90. <https://doi.org/10.1186/1471-2164-12-90> (2011).
- Hiller, M., Zhang, Z., Backofen, R. & Stamm, S. Pre-mRNA secondary structures influence exon recognition. *PLoS Genet.* **3**, e204. <https://doi.org/10.1371/journal.pgen.0030204> (2007).
- Saha, K. *et al.* Structural disruption of exonic stem-loops immediately upstream of the intron regulates mammalian splicing. *Nucleic Acids Res.* **48**, 6294–6309. <https://doi.org/10.1093/nar/gkaa358> (2020).
- Mitrapant, C. *et al.* Rational design of antisense oligomers to induce dystrophin exon skipping. *Mol. Ther.* **17**, 1418–1426. <https://doi.org/10.1038/mt.2009.49> (2009).
- Lee, Y. *et al.* Variants affecting exon skipping contribute to complex traits. *PLoS Genet.* **8**, e1002998. <https://doi.org/10.1371/journal.pgen.1002998> (2012).
- Hillman, R. T., Green, R. E. & Brenner, S. E. An unappreciated role for RNA surveillance. *Genome Biol.* **5**, R8. <https://doi.org/10.1186/gb-2004-5-2-r8> (2004).
- McIntosh, C. S., Aung-Htut, M. T., Fletcher, S. & Wilton, S. D. Removal of the polyglutamine repeat of ataxin-3 by redirecting pre-mRNA processing. *Int. J. Mol. Sci.* <https://doi.org/10.3390/ijms20215434> (2019).
- Toonen, L. J. A., Schmidt, I., Luijsterburg, M. S., Van Attikum, H. & Van Roon-Mom, W. M. C. Antisense oligonucleotide-mediated exon skipping as a strategy to reduce proteolytic cleavage of ataxin-3. *Sci. Rep.* **6**, 35200. <https://doi.org/10.1038/srep35200> (2016).
- Desmet, F. O. *et al.* Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37**, e67. <https://doi.org/10.1093/nar/gkp215> (2009).

35. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394. <https://doi.org/10.1089/1066527041410418> (2004).
36. Sibley, C. R., Blazquez, L. & Ule, J. Lessons from non-canonical splicing. *Nat. Rev. Genet.* **17**, 407–421. <https://doi.org/10.1038/nrg.2016.46> (2016).
37. Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q. & Krainer, A. R. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* **31**, 3568–3571 (2003).
38. Ham, K. A., Aung-Htut, M. T., Fletcher, S. & Wilton, S. D. Nonsequential splicing events alter antisense-mediated exon skipping outcome in COL7A1. *Int. J. Mol. Sci.* <https://doi.org/10.3390/ijms21207705> (2020).
39. Zuo, P. & Manley, J. L. Functional domains of the human splicing factor ASF/SF2. *EMBO J.* **12**, 4727–4737 (1993).
40. Krainer, A. R., Conway, G. C. & Kozak, D. The essential pre-mRNA splicing factor SF2 influences 5' splice site selection by activating proximal sites. *Cell* **62**, 35–42. [https://doi.org/10.1016/0092-8674\(90\)90237-9](https://doi.org/10.1016/0092-8674(90)90237-9) (1990).
41. De Conti, L., Baralle, M. & Buratti, E. Exon and intron definition in pre-mRNA splicing. *WIREs RNA* **4**, 49–60. <https://doi.org/10.1002/wrna.1140> (2013).
42. Fox-Walsh, K. L. *et al.* The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc. Natl. Acad. Sci. USA.* **102**, 16176–16181. <https://doi.org/10.1073/pnas.0508489102> (2005).
43. Sterner, D. A., Carlo, T. & Berget, S. M. Architectural limits on split genes. *Proc. Natl. Acad. Sci. USA.* **93**, 15081–15085. <https://doi.org/10.1073/pnas.93.26.15081> (1996).
44. He, R. Z., Luo, D. X. & Mo, Y. Y. Emerging roles of lncRNAs in the post-transcriptional regulation in cancer. *Genes Dis.* **6**, 6–15. <https://doi.org/10.1016/j.gendis.2019.01.003> (2019).
45. Adams, A. M. *et al.* Antisense oligonucleotide induced exon skipping and the dystrophin gene transcript: Cocktails and chemistries. *BMC Mol. Biol.* **8**, 57. <https://doi.org/10.1186/1471-2199-8-57> (2007).
46. Aung-Htut, M. *et al.* Systematic approach to developing splice modulating antisense oligonucleotides. *Int. J. Mol. Sci.* **20**, 5030. <https://doi.org/10.3390/ijms20205030> (2019).
47. Mann, C. J., Honeyman, K., McClorey, G., Fletcher, S. & Wilton, S. D. Improved antisense oligonucleotide induced exon skipping in the mdx mouse model of muscular dystrophy. *J. Gene Med.* **4**, 644–654. <https://doi.org/10.1002/jgm.295> (2002).
48. Wilton, S. D., Lim, L., Dye, D. & Laing, N. Bandstab: A PCR-based alternative to cloning PCR products. *Biotechniques* **22**, 642–645. <https://doi.org/10.2144/97224bm14> (1997).
49. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biochem.* **215**, 403–410. [https://doi.org/10.1016/s0022-2836\(05\)80360-2](https://doi.org/10.1016/s0022-2836(05)80360-2) (1990).
50. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26. <https://doi.org/10.1186/1748-7188-6-26> (2011).

### Author contributions

Conceptualization, K.A.H., N.P.K., S.D.W.; methodology, K.A.H., N.P.K., C.S.M., K.Z., K.G., S.D.W.; formal analysis, K.A.H., N.P.K., S.D.W.; investigation, K.A.H., N.P.K., C.S.M., K.Z., K.G.; writing—original draft preparation, K.A.H., N.P.K., C.S.M.; writing—review and editing, K.A.H., N.P.K., C.S.M., M.T.A-H., K.Z., K.G., S.F., S.D.W.; supervision, S.D.W., S.F., M.T.A-H.; resources, S.D.W., S.F.; funding acquisition, S.D.W., S.F. All authors have read and agreed to the published version of the manuscript.

### Funding

This work was supported by the National Health and Medical Research Council [grant number 1144791].

### Competing interests

S.D.W. is a consultant to Sarepta Therapeutics; S.D.W. and S.F. are named inventors on patents licensed through the University of Western Australia to Sarepta Therapeutics and as such are entitled to milestone and royalty payments; K.A.H., C.S.M., M.T.A-H., K.G. receive salary support from Sarepta Therapeutics. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results. N.P.K and K.Z declare no competing interests.


### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-94639-x>.

**Correspondence** and requests for materials should be addressed to S.D.W.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

#### 6.4 Mentions and awards

At time of submission, this report has been cited in the following third-party work:

Arthur, G.K., and Cruse, G. (2022). Regulation of Trafficking and Signaling of the High Affinity IgE Receptor by FcεRIβ and the Potential Impact of FcεRIβ Splicing in Allergic Inflammation. *International Journal of Molecular Sciences* 23(2). doi: 10.3390/ijms23020788.

This report was also awarded the Australian Gene and Cell Therapy Society's 2021 (July-December) "Best Paper Prize." A certificate for this award was not available at time of thesis submission, but a list of present and past winners is available on the AGCTS website at <https://agcts.org.au/best-paper-prize/>.

# **Chapter 7**

## **Conclusions and Future Directions**

The research presented in this thesis provides several important advances in our collective understanding of cryptic splicing and RNA biology in general. In this chapter, I explain the immediate impact of these advances and speculate on how future research might build upon them to generate further new insights.

### 7.1 Revisions and additions to pseudoexon mutation categories

Chapter 3 described the various mutation types reported to cause pseudoexons in the *DMD* gene, noting that mutations tend to either directly alter the PE sequence or splice motifs (proximal mutations) or occur great distances away from the PE (distal mutations). There was no gradual tapering-off of reported mutations as their distance from the PE increased – either the mutations were very close to the PEs they caused, or their distance barely appeared to matter at all. Furthermore, when the coincidence of PEs with reported recursive splice sites (Gazzoli *et al.* 2016) was examined, it became apparent that there was a strong overlap between PEs with distal mutations and PEs that matched to reported recursive splice sites\*. Our hypothesis was that a mutation of unidentified distal silencing elements could exonize recursive splice sites, and that this might represent an entirely separate pathway to PE activation.

I later discovered that this PE mutation category approximated one already proposed by Dhir and Buratti in their 2010 report, which they described as “Loss of upstream 5′ss or downstream 3′ss.” In Chapter 4, analysis of a larger dataset of PE distal mutations led us to suggest *weakened neighbour-exon definition* as a revised version of these previous categories.

We hypothesised that when mutations weaken definition of a canonical exon, this reduces the contrast between that exon and its flanking introns and causes exon-like tracts within those introns to appear relatively *more* exon-like by comparison, thereby increasing their chance of being spliced into the mature transcript. We find indirect supporting evidence for this hypothesis in the positive correlation between intron size

---

\* I am indebted to one of the report’s anonymous peer reviewers, who first drew my attention to this pattern.

and flanking splice site strength (Farlow *et al.* 2012), as larger introns generally allow more opportunities for errant splicing.

Although “weakened neighbour-exon definition” was the most well-supported novel category of PE mutation, we also outlined the defining features of three other PE mutation types. Only one of these – “*Loss of an upstream polyadenylation site*” – could be well-explained by established spliceosome mechanics and is described in detail within the report (Chapter 4). The second category, “*Change to local intronic silencers or enhancers*,” is broadly consistent with the splice-modulating effects of numerous RNA-binding proteins, but currently lacks the specificity needed to predict when these effects will produce a PE. The third category, “*Close adjacent pseudoexons*,” may be a product of disrupted intron splicing order but will likely require substantial further research before it can be properly explained.

These advances in categorizing pseudoexon mutations will be of great interest to clinical geneticists, who may find them useful in explaining the underlying pathologies of patients with novel splicing mutations, and to other RNA biologists who are developing predictive models of spliceosome behaviour.

## 7.2 The interrelation of pseudoexons with recursive splice sites and other deep intronic splice motifs

In Chapter 3, I showed how multiple *DMD* pseudoexon splice sites matched to putative recursive splice sites, suggesting a link between these phenomena. In Chapter 4, we confirmed that this pattern extended to PEs in other genes, with the caveat that only seven of the matches were confirmed recursive splice sites. This is an important finding but begs the question of the nature of the other active deep intronic splice motifs. We inferred that many of these were components of poison exons or decoy exons, but could not define which ones were which, as we currently lack a rigorous method for distinguishing these phenomena from each other.

Although we found that 15.6% of PEs bore splice sites that were active in non-mutant cells, the EST and RNAseq (Sibley *et al.* 2015) data we used for comparison



are far from a complete set of all the active splice sites in human cells. It is likely that expanding the dataset used for comparison with the PEs in this catalogue would substantially increase the number of matches, and 15.6% should therefore be viewed as a conservative estimate.

However, even this conservative estimate goes a long way towards addressing the underlying paradox of pseudoexons: *If exon definition is complex and well-regulated, why are so many pseudoexons created by mutations of only a single nucleotide?* For if many pseudoexons are ultimately found to be splice-mutant versions of recursive splice sites, poison exons, decoy exons, or other functional deep-intronic splice motifs, then this paradox ceases to be a paradox and could more usefully be restated as: *Because exon definition is complex and imperfectly regulated, many pseudoexons are created by mutations of only a single nucleotide.*

### 7.3 The “SNPtic exon” concept and its explanatory power

In Chapter 5, we outline the concept of “SNPtic” exons, collate several supporting cases from the literature, and describe a straightforward method that others can use to discover new examples. Although we anticipate that our analysis of individual SNPtic exons will be of interest to researchers studying the genes in which they arise, we believe that the SNPtic exon concept itself is the most useful outcome of that report. In this section, we discuss why this is so.

The concept of a SNPtic exon is a logical extension of the concept of a pseudoexon. Most pseudoexons arise from rare pathogenic mutations that cause a drastic increase in the splicing of an intronic region that would otherwise be spliced at low levels or not at all. SNPtic exons are mechanically similar, but their splicing is modulated by common variants (SNPs) that cause only a minor change in the degree of inclusion. Due to the subtlety of this effect on expression of the gene, carriage of the relevant SNP often does not manifest as a disease phenotype in an individual but may be statistically detectable within populations via genome-wide association studies.

Encouragingly, recent research by Ma *et al.* (2022) and Brown *et al.* (2022) has identified a clinically relevant splicing variant very similar to a SNPtic exon. In these reports, published simultaneously in *Nature*, the two groups independently analysed the splicing of a cryptic exon in the autosomal gene *UNC13A*. This cryptic exon is observed almost exclusively in cells that are functionally deficient in splicing factor TDP43, a characteristic feature of amyotrophic lateral sclerosis (ALS). Both teams reported that inclusion frequency of the *UNC13A* cryptic exon was positively associated with the minor allele of the SNP *rs12973192*, an ALS risk variant that resides within the cryptic exon and appears to increase its inclusion by altering a splice factor binding motif. The two reports disagreed on whether the splice factor affected was TDP43, with Brown *et al.* supporting this conclusion while Ma *et al.* were tentatively negative. However, both made a strong case for *rs12973192*-modulated splicing being the root cause of this SNP's associated phenotype of higher disease severity, thus marking an important new milestone in our understanding of ALS pathology.

The concept of examining RNA splicing to explain population-level SNP-phenotype correlations is not a new one – indeed, this is the underlying principle of splicing quantitative trait loci, or sQTLs (Pickrell *et al.* 2010). However, both sQTL assays and traditional GWAS typically take an agnostic approach to detecting SNP-phenotype correlations, making few *a priori* assumptions about which SNPs will be examined or which allele of a given SNP will correlate with which phenotype. This approach maximises the breadth of potential discoveries at the cost of detection sensitivity, since statistical rigor demands higher significance thresholds for higher numbers of tests.

We suggest that an approach that limits its focus to suspected SNPtic exons might achieve a more optimal balance of breadth and sensitivity. Whole-transcriptome sequencing datasets have already revealed numerous low-frequency cryptic exons in healthy cells (Sibley *et al.* 2015; Sakaguchi and Suyama 2021); and these can be cross-referenced with the known locations of common SNPs within these cryptic exons or their splice motifs. The 'polarity' of each SNP, i.e., which of its alleles is

expected to cause higher inclusion of the putative SNPtic exon, can be predicted by *MaxEnt* (Yeo and Burge 2004) or *HExoSplice* scoring (Ke *et al.* 2011; Tubeuf *et al.* 2020), while the expected consequence of inclusion of the exon is modelled by its impact on the mature transcript's reading frame.

Incorporating common-sense *a priori* assumptions about splicing into future GWAS could greatly improve their capacity to detect phenotypically important SNPs, whilst also providing a built-in mechanistic explanation for why these correlations exist. Some of the SNPtic exons thus discovered might prove to be useful new targets for exon-skipping antisense oligonucleotide therapies. Furthermore, a retrospective analysis of any SNPs that exert unexpected splicing effects may yield valuable new insights into the behaviour of the spliceosome.

#### 7.4 The mechanisms of antisense-oligomer-induced partial exon skipping and its potential as a therapeutic strategy

Chapter 6 reported on antisense-oligomer-induced partial exon skipping involving six exons of six genes. Partial exon skipping – a form of cryptic splicing – had been observed as an unintended side effect of some of the many hundreds of AOs our group has designed and tested in human cell lines. We investigated the common features of these six examples and attempted to explain their causative factors.

Although the small sample number prohibited any statistical analysis, we were able to devise a hybrid predictive model that combined RNA secondary structure with splice factor motifs. This modelling revealed that all the AOs appear to increase the difference in exon definition strength between the retained and skipped segments, through a combination of direct steric blocking of some splice factors binding sites and altering the openness of RNA structure for others. This modelling may be of use to future research into partial exon skipping.

If partial exon-skipping can be better understood and even deliberately induced, it has the potential to expand skipping therapies into genes and gene regions that had previously seemed unsuitable. In the past, potential target exons for therapeutic

skipping have been limited to pseudo/cryptic exons and canonical exons that preserve the open reading frame and that encode non-essential regions of the translated protein. This has led to many mutations, and the patients that carry them, being excluded as candidates for exon skipping therapies. Partial exon-skipping may offer hope for some of these patients – for example, skipping only part of a particular exon might preserve the reading frame of a transcript in cases where skipping of the entire exon would not. Partial exon-skipping also has the advantage of hewing much closer to the evolved splicing pattern of the gene, as it still permits recognition of an exon at the same location in the transcript, and with one of the two normal splice sites. The main impediment to partial exon-skipping is that many exons will simply not have an amenable combination of splice motifs, secondary structure, and a latent internal cryptic splice site, though *in silico* techniques such as those outlined in our report will enable better predictions about where to start looking. As such, while we are optimistic about the future of partial exon skipping, whole exon-skipping is all but certain to remain the predominant strategy for splice-modulating therapies.

### 7.5 Limitations and broader challenges

As this thesis has focused exclusively on cryptic splicing in humans, the question remains as to what extent the patterns observed here can serve as a model for cryptic splicing in other species. As noted in chapter 4, there have been reports of similar pathogenic cryptic splicing in non-human species, such as dogs (Smith *et al.* 2007), mice (Gómez-Grau *et al.* 2017), and even wheat (Howitt *et al.* 2009), and the major spliceosome itself is highly conserved within the eukaryote lineage (Rogozin *et al.* 2012). On the other hand, certain features of the human genome that are unique to our recent lineage, such as *Alu* elements (Arcot *et al.* 1995), have been shown to be a significant source of cryptic splicing events (Vořechovský 2010) and thus care must be taken to ensure their effect on human RNA processing does not bias our expectations of spliceosome behaviour for non-primates. Conversely, transposable elements that are unique to certain non-human lineages, such as the P-elements observed in *Drosophila* (Kofler *et al.* 2015), may yet prove to be a source of novel cryptic splicing events wholly distinct from those seen in humans.

Even within *Homo sapiens*, the work in this thesis has been unavoidably limited by a narrowness of focus. Most of the transcript data we analysed was originally generated through Sanger sequencing of RT-PCR products; products that were, themselves, amplified from limited regions of mature transcripts extracted from one or very few cell types. While it is generally safe to assume that the immediate vicinity of a splicing mutation or AO target site is the most likely region to be processed differently, a growing body of evidence shows that changes can often be observed elsewhere in the transcript (Keegan *et al.* 2021) and throughout the transcriptome (Flynn *et al.* 2022). The effects of splicing mutations can also differ from one cell/tissue type to another (Grandchamp *et al.* 1989; Will *et al.* 1994; Ishigaki *et al.* 2003), especially when multiple canonical transcript isoforms are involved, as seen in some pathogenic *DMD* mutations (Moizard *et al.* 2000; Neri *et al.* 2012). It is anticipated that the ongoing uptake of whole transcriptome sequencing, coupled with long-read technologies that can sequence entire transcripts (Drexler *et al.* 2020), will do much to broaden our view of how and where these enormously diverse splicing changes manifest in the cells of humans and other eukaryotes.

## 7.6 Concluding remarks

The work collated in this thesis has generated several new insights into some infrequently examined forms of cryptic splicing – primarily pseudoexons, but also the induced cryptic splicing of canonical exons. These insights were achieved through a consistent focus on categorisation. Existing categories of pseudoexon mutations were supported in some cases and modified in others, and new categories were nominated where appropriate. Novel categories of cryptic splicing, such as SNPtic exons and AO-induced partial exon-skipping, were also defined and examined for the first time as general phenomena, despite supporting examples of both being extant in published literature for many years previously.

Ironically, this focus on categorisation has only emphasised how few type-boundaries there truly are when it comes to RNA splicing. Pseudoexons, poison exons, recursive splice sites, decoy exons – all have their own distinct definitions and origins, and yet all appear to overlap in their locations and regulatory

mechanisms. Even application of the term “canonical exon” is openly subject to revision, as ongoing research confirms new transcript variants and discredits old ones.

Yet none of this implies that biological categories are futile, only that they should not be allowed to stagnate for the sake of convenience. An effective language of biology requires constant revision and should be as dynamic and responsive as the living world it describes. It is my hope that the work presented here has contributed to this effort and that it will inform and eventually be superseded by the work of researchers to follow.

## **References**

Aartsma-Rus, A., van Vliet, L., Hirschi, M., Janson, A.A., Heemskerk, H., de Winter, C.L., et al. (2009). Guidelines for antisense oligonucleotide design and insight into splice-modulating mechanisms. *Mol Ther* 17(3), 548-553. doi: 10.1038/mt.2008.205.

Abramowicz, A., and Gos, M. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet* 59(3), 253-268. doi: 10.1007/s13353-018-0444-7.

Adachi, H., and Yu, Y.T. (2014). Insight into the mechanisms and functions of spliceosomal snRNA pseudouridylation. *World J Biol Chem* 5(4), 398-408. doi: 10.4331/wjbc.v5.i4.398.

Afonina, Z.A., and Shirokov, V.A. (2018). Three-Dimensional Organization of Polyribosomes - A Modern Approach. *Biochemistry (Mosc)* 83(Suppl 1), S48-S55. doi: 10.1134/S0006297918140055.

Akinyi, M.V., and Frilander, M.J. (2021). At the Intersection of Major and Minor Spliceosomes: Crosstalk Mechanisms and Their Impact on Gene Expression. *Front Genet* 12, 700744. doi: 10.3389/fgene.2021.700744.

Ando, T., Komaki, G., Naruo, T., Okabe, K., Takii, M., Kawai, K., et al. (2006). Possible role of preproghrelin gene polymorphisms in susceptibility to bulimia nervosa. *Am J Med Genet B Neuropsychiatr Genet* 141B(8), 929-934. doi: 10.1002/ajmg.b.30387.

Arcot, S.S., Wang, Z., Weber, J.L., Deininger, P.L. and Batzer, M.A. (1995) Alu repeats: a source for the genesis of primate microsatellites. *Genomics* 29(1), 136-44. doi: 10.1006/geno.1995.1224.

Aziz, M.C., Schneider, P.N., and Carvill, G.L. (2021). Targeting Poison Exons to Treat Developmental and Epileptic Encephalopathy. *Dev Neurosci*, 1-6. doi: 10.1159/000516143.

Ben-Avraham, D., Govindaraju, D.R., Budagov, T., Fradin, D., Durda, P., Liu, B., et al. (2017). The GH receptor exon 3 deletion is a marker of male-specific exceptional longevity associated with increased GH sensitivity and taller stature. *Sci Adv* 3(6), e1602025. doi: 10.1126/sciadv.1602025.

Blazquez, L., Emmett, W., Faraway, R., Pineda, J.M.B., Bajew, S., Gohr, A., et al. (2018). Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing. *Mol Cell* 72(3), 496-509 e499. doi: 10.1016/j.molcel.2018.09.033.

Boivin, V., Faucher-Giguere, L., Scott, M., and Abou-Elela, S. (2019). The cellular landscape of mid-size noncoding RNA. *Wiley Interdiscip Rev RNA* 10(4), e1530. doi: 10.1002/wrna.1530.

Bouge, A.L., Murauer, E., Beyne, E., Miro, J., Varilh, J., Taulan, M., et al. (2017). Targeted RNA-Seq profiling of splicing pattern in the DMD gene: exons are mostly constitutively spliced in human skeletal muscle. *Sci Rep* 7, 39094. doi: 10.1038/srep39094.

Breuel, S., Vorm, M., Bräuer, A.U., Owczarek-Lipska, M. and Neidhardt, J. (2019) Combining Engineered U1 snRNA and Antisense Oligonucleotides to Improve the Treatment of a BBS1 Splice Site Mutation. *Mol Ther Nucleic Acids* 18, 123-130. doi: 10.1016/j.omtn.2019.08.014

Brown, A.L., Wilkins, O.G., Keuss, M. J., Hill, S. E., Zanovello, M., Lee, W. C. et al. (2022). TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature* 603(7899), 131-137. doi: 10.1038/s41586-022-04436-3.



Buratti, E., Chivers, M., Hwang, G., and Vořechovský, I. (2011). DBASS3 and DBASS5: databases of aberrant 3'- and 5'-splice sites. *Nucleic Acids Res* 39(Database issue), D86-91. doi: 10.1093/nar/gkq887.

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(2), 661-674. doi: 10.1534/genetics.104.039701.

Busch, A., and Hertel, K.J. (2012). Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip Rev RNA* 3(1), 1-12. doi: 10.1002/wrna.100.

Carvill, G.L., and Mefford, H.C. (2020). Poison exons in neurodevelopment and disease. *Curr Opin Genet Dev* 65, 98-102. doi: 10.1016/j.gde.2020.05.030.

Ciandrini, L., Stansfield, I., and Romano, M.C. (2013). Ribosome traffic on mRNAs maps to gene ontology: genome-wide quantification of translation initiation rates and polysome size regulation. *PLoS Comput Biol* 9(1), e1002866. doi: 10.1371/journal.pcbi.1002866.

Cole, C.N., and Scarcelli, J.J. (2006). Transport of messenger RNA from the nucleus to the cytoplasm. *Curr Opin Cell Biol* 18(3), 299-306. doi: 10.1016/j.ceb.2006.04.006.

Czaplinski, K., and Singer, R.H. (2006). Pathways for mRNA localization in the cytoplasm. *Trends Biochem Sci* 31(12), 687-693. doi: 10.1016/j.tibs.2006.10.007.

Costales, M.G., Childs-Disney, J.L., Haniff, H.S., and Disney, M.D. (2020) How We Think about Targeting RNA with Small Molecules. *J Med Chem* 63(17), 8880-8900. doi: 10.1021/acs.jmedchem.9b01927

Dang, Y., van Heusden, C., Nickerson, V., Chung, F., Wang, Y., Quinney, N. L., et al. (2021) Enhanced delivery of peptide-morpholino oligonucleotides with a small molecule to correct splicing defects in the lung. *Nucleic Acids Res* 49(11), 6100-6113. doi: 10.1093/nar/gkab488.

Das, S., Vera, M., Gandin, V., Singer, R.H., and Tutucci, E. (2021). Intracellular mRNA transport and localized translation. *Nat Rev Mol Cell Biol* 22(7), 483-504. doi: 10.1038/s41580-021-00356-8.

Dhillon, S. (2020). Viltolarsen: First Approval. *Drugs* 80(10), 1027-1031. doi: 10.1007/s40265-020-01339-3.

Dhir, A., and Buratti, E. (2010). Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. *FEBS J* 277(4), 841-855. doi: 10.1111/j.1742-4658.2009.07520.x.

Donmez, G., Hartmuth, K., and Luhrmann, R. (2004). Modified nucleotides at the 5' end of human U2 snRNA are required for spliceosomal E-complex formation. *RNA* 10(12), 1925-1933. doi: 10.1261/rna.7186504.

Drexler, H. L., Choquet, K. and Churchman, L. S. (2020). Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Mol Cell* 77(5), 985-998 e8. doi: 10.1016/j.molcel.2019.11.017.

Eaton, J.D., Francis, L., Davidson, L., and West, S. (2020). A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes. *Genes Dev* 34(1-2), 132-145. doi: 10.1101/gad.332833.119.

Echigoya, Y., Mouly, V., Garcia, L., Yokota, T., and Duddy, W. (2015). In silico screening based on predictive algorithms as a design tool for exon skipping oligonucleotides in Duchenne muscular dystrophy. *PLoS One* 10(3), e0120058. doi: 10.1371/journal.pone.0120058.

Eckmann, C.R., Rammelt, C., and Wahle, E. (2011). Control of poly(A) tail length. *Wiley Interdiscip Rev RNA* 2(3), 348-361. doi: 10.1002/wrna.56.

Farlow, A., Dolezal, M., Hua, L., and Schlotterer, C. (2012). The genomic signature of splicing-coupled selection differs between long and short introns. *Mol Biol Evol* 29(1), 21-24. doi: 10.1093/molbev/msr201.

Filopanti, M., Giavoli, C., Grottoli, S., Bianchi, A., De Marinis, L., Ghigo, E., et al. (2011). The exon 3-deleted growth hormone receptor: molecular and functional characterization and impact on GH/IGF-I axis in physiological and pathological conditions. *J Endocrinol Invest* 34(11), 861-868. doi: 10.1007/BF03346731.

Flynn, L.L., Li, R., Pitout, I. L., Aung-Htut, M. T., Larcher, L. M., Cooper, J. A. L., et al. (2022). Single Stranded Fully Modified-Phosphorothioate Oligonucleotides can Induce Structured Nuclear Inclusions, Alter Nuclear Protein Localization and Disturb the Transcriptome In Vitro. *Front Genet* 13, 791416. doi: 10.3389/fgene.2022.791416.

Garneau, N.L., Wilusz, J., and Wilusz, C.J. (2007). The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol* 8(2), 113-126. doi: 10.1038/nrm2104.

Gazzoli, I., Pulyakhina, I., Verwey, N.E., Ariyurek, Y., Laros, J.F., t Hoen, P.A., et al. (2016). Non-sequential and multi-step splicing of the dystrophin transcript. *RNA Biol* 13(3), 290-305. doi: 10.1080/15476286.2015.1125074.

Georgomanolis, T., Sofiadis, K., and Papantonis, A. (2016). Cutting a Long Intron Short: Recursive Splicing and Its Implications. *Front Physiol* 7, 598. doi: 10.3389/fphys.2016.00598.

Geuens, T., Bouhy, D., and Timmerman, V. (2016). The hnRNP family: insights into their role in health and disease. *Hum Genet* 135(8), 851-867. doi: 10.1007/s00439-016-1683-5.

Giess, A., Torres Cleuren, Y.N., Tjeldnes, H., Krause, M., Bizuayehu, T.T., Hiensch, S., et al. (2020). Profiling of Small Ribosomal Subunits Reveals Modes and Regulation of Translation Initiation. *Cell Rep* 31(3), 107534. doi: 10.1016/j.celrep.2020.107534.

Gómez-Grau, M., Albaigès, J., Casas, J., Auladell, C., Dierssen, M., Vilageliu, L., et al. (2017). New Murine Niemann-Pick Type C Models Bearing a Pseudoexon-Generating Mutation Recapitulate the Main Neurobehavioural and Molecular Features of the Disease. *Sci. Rep.* 7, 41931. doi:10.1038/srep41931.

Grandchamp, B., Picat, C., Mignotte, V. Wilson, J. H., Te Velde, K., Sandkuyi, L. et al. (1989). Tissue-specific splicing mutation in acute intermittent porphyria. *Proc Natl Acad Sci USA* 86(2), 661-4. doi: 10.1073/pnas.86.2.661.

Gueroussov, S., Weatheritt, R.J., O'Hanlon, D., Lin, Z.Y., Narula, A., Gingras, A.C., et al. (2017). Regulatory Expansion in Mammals of Multivalent hnRNP Assemblies that Globally Control Alternative Splicing. *Cell* 170(2), 324-339 e323. doi: 10.1016/j.cell.2017.06.037.

Haj Khelil, A., Deguillien, M., Moriniere, M., Ben Chibani, J., and Baklouti, F. (2008). Cryptic splicing sites are differentially utilized in vivo. *FEBS J* 275(6), 1150-1162. doi: 10.1111/j.1742-4658.2008.06276.x.

Han, Z., Chen, C., Christiansen, A., Ji, S., Lin, Q., Anumonwo, C., et al. (2020). Antisense oligonucleotides increase Scn1a expression and reduce seizures and SUDEP incidence in a mouse model of Dravet syndrome. *Sci Transl Med* 12(558). doi: 10.1126/scitranslmed.aaz6100.

Heo, Y.A. (2020). Golodirsen: First Approval. *Drugs* 80(3), 329-333. doi: 10.1007/s40265-020-01267-2.

Howard, J.M., Lin, H., Wallace, A.J., Kim, G., Draper, J.M., Haeussler, M., et al. (2018). HNRNPA1 promotes recognition of splice site decoys by U2AF2 in vivo. *Genome Res* 28(5), 689-698. doi: 10.1101/gr.229062.117.

Howitt, C.A., Cavanagh, C. R., Bowerman, A. F., Cazzonelli, C., Rampling, L., Mimica, J. L., et al. (2009) Alternative splicing, activation of cryptic exons and amino acid substitutions in carotenoid biosynthetic genes are associated with lutein accumulation in wheat endosperm. *Funct Integr Genomics* 9(3), 363-76. doi: 10.1007/s10142-009-0121-3

Ishigaki, K., Nicolle, D. Krejci, E., Leroy, J. P., Koenig, J., Fardeau, M. et al. (2003). Two novel mutations in the COLQ gene cause endplate acetylcholinesterase deficiency. *Neuromuscular Disorders* 13(3), 236-244. doi: 10.1016/s0960-8966(02)00243-2.

Jeong, S. (2017). SR Proteins: Binders, Regulators, and Connectors of RNA. *Mol Cells* 40(1), 1-9. doi: 10.14348/molcells.2017.2319.

Jonkers, I., and Lis, J.T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* 16(3), 167-177. doi: 10.1038/nrm3953.

Jutzi, D., Akinyi, M.V., Mechttersheimer, J., Frilander, M.J., and Ruepp, M.D. (2018). The emerging role of minor intron splicing in neurological disorders. *Cell Stress* 2(3), 40-54. doi: 10.15698/cst2018.03.126.

Juven-Gershon, T., Hsu, J.Y., Theisen, J.W., and Kadonaga, J.T. (2008). The RNA polymerase II core promoter - the gateway to transcription. *Curr Opin Cell Biol* 20(3), 253-259. doi: 10.1016/j.ceb.2008.03.003.

Kapranov, P., St Laurent, G., Raz, T., Ozsolak, F., Reynolds, C.P., Sorensen, P.H., et al. (2010). The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol* 8, 149. doi: 10.1186/1741-7007-8-149.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., et al. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res* 21(8), 1360-1374. doi: 10.1101/gr.119628.110.

Keegan, N.P., Wilton, S.D. and Fletcher, S. (2021) Analysis of Pathogenic Pseudoexons Reveals Novel Mechanisms Driving Cryptic Splicing. *Front Genet* 12, 806946. doi: 10.3389/fgene.2021.806946.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., et al. (2013). Function of alternative splicing. *Gene* 514(1), 1-30. doi: 10.1016/j.gene.2012.07.083.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., et al. (2002). The human genome browser at UCSC. *Genome Res* 12(6), 996-1006. doi: 10.1101/gr.229102.

Kiledjian, M. (2018). Eukaryotic RNA 5'-End NAD(+) Capping and DeNADding. *Trends Cell Biol* 28(6), 454-464. doi: 10.1016/j.tcb.2018.02.005.

Kofler, R., Hill, T., Nolte, V., Betancourt, A.J. and Schlötterer, C. (2015) The recent invasion of natural *Drosophila simulans* populations by the P-element. *Proc Natl Acad Sci USA* 112(21), 6659-63. doi: 10.1073/pnas.1500758112.

Kole, R., Krainer, A.R., and Altman, S. (2012). RNA therapeutics: beyond RNA interference and antisense oligonucleotides. *Nat Rev Drug Discov* 11(2), 125-140. doi: 10.1038/nrd3625.

Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* 15(20), 8125-8148. doi: 10.1093/nar/15.20.8125.

Královičová, J., and Vořechovský, I. (2007). Global control of aberrant splice-site activation by auxiliary splicing sequences: evidence for a gradient in exon and intron definition. *Nucleic Acids Res* 35(19), 6399-6413. doi: 10.1093/nar/gkm680.

Le, B.T., Adams, A.M., Fletcher, S., Wilton, S.D., and Veedu, R.N. (2017). Rational Design of Short Locked Nucleic Acid-Modified 2'-O-Methyl Antisense Oligonucleotides for Efficient Exon-Skipping In Vitro. *Mol Ther Nucleic Acids* 9, 155-161. doi: 10.1016/j.omtn.2017.09.002.

Le Hir, H., Gatfield, D., Izaurralde, E., and Moore, M.J. (2001). The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J* 20(17), 4987-4997. doi: 10.1093/emboj/20.17.4987.

Leclair, N.K., Brugiolo, M., Urbanski, L., Lawson, S.C., Thakar, K., Yurieva, M., et al. (2020). Poison Exon Splicing Regulates a Coordinated Network of SR Protein Expression during Differentiation and Tumorigenesis. *Mol Cell* 80(4), 648-665 e649. doi: 10.1016/j.molcel.2020.10.019.

Lee, P.J., Yang, S., Sun, Y., and Guo, J.U. (2021). Regulation of nonsense-mediated mRNA decay in neural development and disease. *J Mol Cell Biol* 13(4), 269-281. doi: 10.1093/jmcb/mjab022.

Li, D., Mastaglia, F.L., Fletcher, S., and Wilton, S.D. (2018). Precision Medicine through Antisense Oligonucleotide-Mediated Exon Skipping. *Trends Pharmacol Sci* 39(11), 982-994. doi: 10.1016/j.tips.2018.09.001.

Loughran, G., Chou, M.Y., Ivanov, I.P., Jungreis, I., Kellis, M., Kiran, A.M., et al. (2014). Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res* 42(14), 8928-8938. doi: 10.1093/nar/gku608.

Ma, X.R., Prudencio, M., Kolke, Y., Vatsavayai, S. C., Kim, G., Harbinski, F. et al. (2022). TDP-43 represses cryptic exon inclusion in the FTD-ALS gene UNC13A. *Nature* 603(7899), 124-130 (2022). doi: 10.1038/s41586-022-04424-7.

Maita, H., and Nakagawa, S. (2020). What is the switch for coupling transcription and splicing? RNA Polymerase II C-terminal domain phosphorylation, phase separation and beyond. *Wiley Interdiscip Rev RNA* 11(1), e1574. doi: 10.1002/wrna.1574.

Manley, J.L., and Krainer, A.R. (2010). A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). *Genes Dev* 24(11), 1073-1074. doi: 10.1101/gad.1934910.

Martin, K.C., and Ephrussi, A. (2009). mRNA localization: gene expression in the spatial dimension. *Cell* 136(4), 719-730. doi: 10.1016/j.cell.2009.01.044.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 7, 29-59. doi: 10.1146/annurev.genom.7.080505.115623.

Merrick, W.C. (2004). Cap-dependent and cap-independent translation in eukaryotic systems. *Gene* 332, 1-11. doi: 10.1016/j.gene.2004.02.051.

Mitropant, C., Adams, A.M., Meloni, P.L., Muntoni, F., Fletcher, S., and Wilton, S.D. (2009). Rational design of antisense oligomers to induce dystrophin exon skipping. *Mol Ther* 17(8), 1418-1426. doi: 10.1038/mt.2009.49.



Moizard, M.P., Toutain, A., Fournier, D., Berret, F., Raynaud, M., Billard, C., et al. (2000). Severe cognitive impairment in DMD: obvious clinical indication for Dp71 isoform point mutation screening. *Eur J Hum Genet* 8(7), 552-6. doi: 10.1038/sj.ejhg.5200488.

Morlan, J.D., Qu, K., and Sinicropi, D.V. (2012). Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLoS One* 7(8), e42882. doi: 10.1371/journal.pone.0042882.

Nelson, K.K., and Green, M.R. (1990). Mechanism for cryptic splice site activation during pre-mRNA splicing. *Proc Natl Acad Sci U S A* 87(16), 6253-6257. doi: 10.1073/pnas.87.16.6253.

Neri, M., Valli, E., Alfano, G., Bovolenta, M., Spitali, P., Rapezzi, C., et al. (2012). The absence of dystrophin brain isoform expression in healthy human heart ventricles explains the pathogenesis of 5' X-linked dilated cardiomyopathy. *BMC Med Genet* 13, 20. doi: 10.1186/1471-2350-13-20.

Nguyen, T.H., Galej, W.P., Bai, X.C., Savva, C.G., Newman, A.J., Scheres, S.H., et al. (2015). The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature* 523(7558), 47-52. doi: 10.1038/nature14548.

Nickless, A., Bailis, J.M., and You, Z. (2017). Control of gene expression through the nonsense-mediated RNA decay pathway. *Cell Biosci* 7, 26. doi: 10.1186/s13578-017-0153-7.

Nicolas, A., Raguenes-Nicol, C., Ben Yaou, R., Ameziane-Le Hir, S., Cheron, A., Vie, V., et al. (2015). Becker muscular dystrophy severity is linked to the structure of dystrophin. *Hum Mol Genet* 24(5), 1267-1279. doi: 10.1093/hmg/ddu537.

Olthof, A.M., Hyatt, K.C., and Kanadia, R.N. (2019). Minor intron splicing revisited: identification of new minor intron-containing genes and tissue-dependent retention and alternative splicing of minor introns. *BMC Genomics* 20(1), 686. doi: 10.1186/s12864-019-6046-x.

Pabalan, N.A., Seim, I., Jarjanazi, H., and Chopin, L.K. (2014). Associations between ghrelin and ghrelin receptor polymorphisms and cancer in Caucasian populations: a meta-analysis. *BMC Genet* 15, 118. doi: 10.1186/s12863-014-0118-3.

Pandya-Jones, A., and Black, D.L. (2009). Co-transcriptional splicing of constitutive and alternative exons. *RNA* 15(10), 1896-1908. doi: 10.1261/rna.1714509.

Parra, M., Booth, B.W., Weiszmann, R., Yee, B., Yeo, G.W., Brown, J.B., et al. (2018). An important class of intron retention events in human erythroblasts is regulated by cryptic exons proposed to function as splicing decoys. *RNA* 24(9), 1255-1265. doi: 10.1261/rna.066951.118.

Parra, M., Zhang, W., Vu, J., DeWitt, M., and Conboy, J.G. (2020). Antisense targeting of decoy exons can reduce intron retention and increase protein expression in human erythroblasts. *RNA* 26(8), 996-1005. doi: 10.1261/rna.075028.120.

Patel, A.A., and Steitz, J.A. (2003). Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol* 4(12), 960-970. doi: 10.1038/nrm1259.

Pérez-Ortín, J.E., Alepuz, P., Chavez, S., and Choder, M. (2013). Eukaryotic mRNA decay: methodologies, pathways, and links to other stages of gene expression. *J Mol Biol* 425(20), 3750-3775. doi: 10.1016/j.jmb.2013.02.029.

Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., et al. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464(7289), 768-772. doi: 10.1038/nature08872.

Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M.C., and Caracausi, M. (2019). Human protein-coding genes and gene feature statistics in 2019. *BMC Res Notes* 12(1), 315. doi: 10.1186/s13104-019-4343-8.

Pozzoli, U., and Sironi, M. (2005). Silencers regulate both constitutive and alternative splicing events in mammals. *Cell Mol Life Sci* 62(14), 1579-1604. doi: 10.1007/s00018-005-5030-6.

Prakash, V. (2017). Spinraza-a rare disease success story. *Gene Ther* 24(9), 497. doi: 10.1038/gt.2017.59.

Pulyakhina, I., Gazzoli, I., t Hoen, P.A., Verwey, N., den Dunnen, J.T., Aartsma-Rus, A., et al. (2015). SplicePie: a novel analytical approach for the detection of alternative, non-sequential and recursive splicing. *Nucleic Acids Res* 43(12), e80. doi: 10.1093/nar/gkv242.

Rahman, M.A., Azuma, Y., Nasrin, F., Takeda, J., Nazim, M., Bin Ahsan, K., et al. (2015). SRSF1 and hnRNP H antagonistically regulate splicing of COLQ exon 16 in a congenital myasthenic syndrome. *Sci Rep* 5, 13208. doi: 10.1038/srep13208.

Reyes, A., and Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res* 46(2), 582-592. doi: 10.1093/nar/gkx1165.

Rogozin, I.B., Carmel, L., Csuros, M., and Koonin, E.V. (2012). Origin and evolution of spliceosomal introns. *Biol Direct* 7, 11. doi: 10.1186/1745-6150-7-11.

Romano, M., Buratti, E., and Baralle, D. (2013). Role of pseudoexons and pseudointrons in human cancer. *Int J Cell Biol* 2013, 810572. doi: 10.1155/2013/810572.

Rosonina, E., Kaneko, S., and Manley, J.L. (2006). Terminating the transcript: breaking up is hard to do. *Genes Dev* 20(9), 1050-1056. doi: 10.1101/gad.1431606.

Russell, A.G., Charette, J.M., Spencer, D.F., and Gray, M.W. (2006). An early evolutionary origin for the minor spliceosome. *Nature* 443(7113), 863-866. doi: 10.1038/nature05228.

Saba, J., Chua, X.Y., Mishanina, T.V., Nayak, D., Windgassen, T.A., Mooney, R.A., et al. (2019). The elemental mechanism of transcriptional pausing. *Elife* 8. doi: 10.7554/eLife.40981.

Sakaguchi, N., and Suyama, M. (2021). In silico identification of pseudo-exon activation events in personal genome and transcriptome data. *RNA Biol* 18(3), 382-390. doi: 10.1080/15476286.2020.1809195.

Sato, H., Hosoda, N., and Maquat, L.E. (2008). Efficiency of the pioneer round of translation affects the cellular site of nonsense-mediated mRNA decay. *Mol Cell* 29(2), 255-262. doi: 10.1016/j.molcel.2007.12.009.

Scotti, M.M., and Swanson, M.S. (2016). RNA mis-splicing in disease. *Nat Rev Genet* 17(1), 19-32. doi: 10.1038/nrg.2015.3.

Seim, I., Lubik, A.A., Lehman, M.L., Tomlinson, N., Whiteside, E.J., Herington, A.C., et al. (2013). Cloning of a novel insulin-regulated ghrelin transcript in prostate cancer. *J Mol Endocrinol* 50(2), 179-191. doi: 10.1530/JME-12-0150.

Shatsky, I.N., Terenin, I.M., Smirnova, V.V., and Andreev, D.E. (2018). Cap-Independent Translation: What's in a Name? *Trends Biochem Sci* 43(11), 882-895. doi: 10.1016/j.tibs.2018.04.011.

Shirley, M. (2021). Casimersen: First Approval. *Drugs* 81(7), 875-879. doi: 10.1007/s40265-021-01512-2.

Sibley, C.R., Blazquez, L., and Ule, J. (2016). Lessons from non-canonical splicing. *Nat Rev Genet* 17(7), 407-421. doi: 10.1038/nrg.2016.46.

Sibley, C.R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., et al. (2015). Recursive splicing in long vertebrate genes. *Nature* 521(7552), 371-375. doi: 10.1038/nature14466.

Singh, G., Jakob, S., Kleedehn, M.G., and Lykke-Andersen, J. (2007). Communication with the exon-junction complex and activation of nonsense-mediated decay by human Upf proteins occur in the cytoplasm. *Mol Cell* 27(5), 780-792. doi: 10.1016/j.molcel.2007.06.030.

Smith, B. F., Kornegay, J. N., and Duan, D. (2007). Independent Canine Models of Duchenne Muscular Dystrophy Due to Intronic Insertions of Repetitive DNA. *Mol Ther* 15. doi:10.1016/s1525-0016(16)44336-4

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., 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. doi: 10.1002/cpbi.5.

Syed, Y.Y. (2016). Eteplirsen: First Global Approval. *Drugs* 76(17), 1699-1704. doi: 10.1007/s40265-016-0657-1.

Thomas, J.D., Polaski, J.T., Feng, Q., De Neef, E.J., Hoppe, E.R., McSharry, M.V., et al. (2020). RNA isoform screens uncover the essentiality and tumor-suppressor activity of ultraconserved poison exons. *Nat Genet* 52(1), 84-94. doi: 10.1038/s41588-019-0555-z.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* 33(1), 201-212. doi: 10.1093/nar/gki158.

- Tian, B., and Graber, J.H. (2012). Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA* 3(3), 385-396. doi: 10.1002/wrna.1116.
- Tubeuf, H., Charbonnier, C., Soukarieh, O., Blavier, A., Lefebvre, A., Dauchel, H., et al. (2020). Large-scale comparative evaluation of user-friendly tools for predicting variant-induced alterations of splicing regulatory elements. *Hum Mutat* 41(10), 1811-1829. doi: 10.1002/humu.24091.
- Turunen, J.J., Niemela, E.H., Verma, B., and Frilander, M.J. (2013). The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA* 4(1), 61-76. doi: 10.1002/wrna.1141.
- Vargas, D.Y., Raj, A., Marras, S.A., Kramer, F.R., and Tyagi, S. (2005). Mechanism of mRNA transport in the nucleus. *Proc Natl Acad Sci U S A* 102(47), 17008-17013. doi: 10.1073/pnas.0505580102.
- Vaz-Drago, R., Custódio, N., and Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. *Hum Genet* 136(9), 1093-1111. doi: 10.1007/s00439-017-1809-4.
- Vořechovský, I. (2010). Transposable elements in disease-associated cryptic exons. *Hum Genet* 127(2), 135-154. doi: 10.1007/s00439-009-0752-4.
- Wagner, R.E., and Frye, M. (2021). Noncanonical functions of the serine-arginine-rich splicing factor (SR) family of proteins in development and disease. *Bioessays* 43(4), e2000242. doi: 10.1002/bies.202000242.
- Wan, Y., Anastasakis, D.G., Rodriguez, J., Palangat, M., Gudla, P., Zaki, G., et al. (2021). Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection. *Cell* 184(11), 2878-2895 e2820. doi: 10.1016/j.cell.2021.04.012.

Will, K., Dörk, T., Stuhmann, M., Meitinger, T., Bertele-Harms, R., Tümmler, B. et al. (1994). A novel exon in the cystic fibrosis transmembrane conductance regulator gene activated by the nonsense mutation E92X in airway epithelial cells of patients with cystic fibrosis. *J Clin Invest* 93(4), 1852-9. doi: 10.1172/JCI117172.

Wimmer, K., Roca, X., Beiglbock, H., Callens, T., Etzler, J., Rao, A.R., et al. (2007). Extensive in silico analysis of NF1 splicing defects uncovers determinants for splicing outcome upon 5' splice-site disruption. *Hum Mutat* 28(6), 599-612. doi: 10.1002/humu.20493.

Yang, B., Ming, X., Cao, C., Laing, B., Yuan, A., Porter, M. A., et al. (2015). High-throughput screening identifies small molecules that enhance the pharmacological effects of oligonucleotides. *Nucleic Acids Res* 43(4), 1987-96. doi: 10.1093/nar/gkv060.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 11(2-3), 377-394. doi: 10.1089/1066527041410418.

Zhang, X., Hong, R., Chen, W., Xu, M., and Wang, L. (2019). The role of long noncoding RNA in major human disease. *Bioorg Chem* 92, 103214. doi: 10.1016/j.bioorg.2019.103214.

Zhang, X.O., Fu, Y., Mou, H., Xue, W., and Weng, Z. (2018). The temporal landscape of recursive splicing during Pol II transcription elongation in human cells. *PLoS Genet* 14(8), e1007579. doi: 10.1371/journal.pgen.1007579.

Zhang, Y., Center, D.M., Wu, D.M., Cruikshank, W.W., Yuan, J., Andrews, D.W., et al. (1998). Processing and activation of pro-interleukin-16 by caspase-3. *J Biol Chem* 273(2), 1144-1149. doi: 10.1074/jbc.273.2.1144.

Zhang, Y., Qian, J., Gu, C., and Yang, Y. (2021). Alternative splicing and cancer: a systematic review. *Signal Transduct Target Ther* 6(1), 78. doi: 10.1038/s41392-021-00486-7.

Zhao, Y., Dunker, W., Yu, Y.T., and Karijolich, J. (2018). The Role of Noncoding RNA Pseudouridylation in Nuclear Gene Expression Events. *Front Bioeng Biotechnol* 6, 8. doi: 10.3389/fbioe.2018.00008.

Zhou, Z., and Fu, X.D. (2013). Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma* 122(3), 191-207. doi: 10.1007/s00412-013-0407-z.

Zhou, Z., Gong, Q., Lin, Z., Wang, Y., Li, M., Wang, L., et al. (2020). Emerging Roles of SRSF3 as a Therapeutic Target for Cancer. *Front Oncol* 10, 577636. doi: 10.3389/fonc.2020.577636.

Zou, M., Guo, B., and He, S. (2011). The roles and evolutionary patterns of intronless genes in deuterostomes. *Comp Funct Genomics* 2011, 680673. doi: 10.1155/2011/680673.