Received: 12 July 2021 | Accepted: 12 October 2021

DOI: 10.1002/mgg3.1840

Molecular Genetics & Genomic Medicine Open Access WILEY

METHOD

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

Niall Patrick Keegan[1,2,3] | **Sue Fletcher**[1,2,4]

[1]Murdoch University, Murdoch, Western Australia, Australia

[2]Centre for Molecular Medicine and Innovative Therapeutics, Perth, Western Australia, Australia

[3]Perron Institute, Perth, Western Australia, Australia

[4]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 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 'pseudoexon', '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.

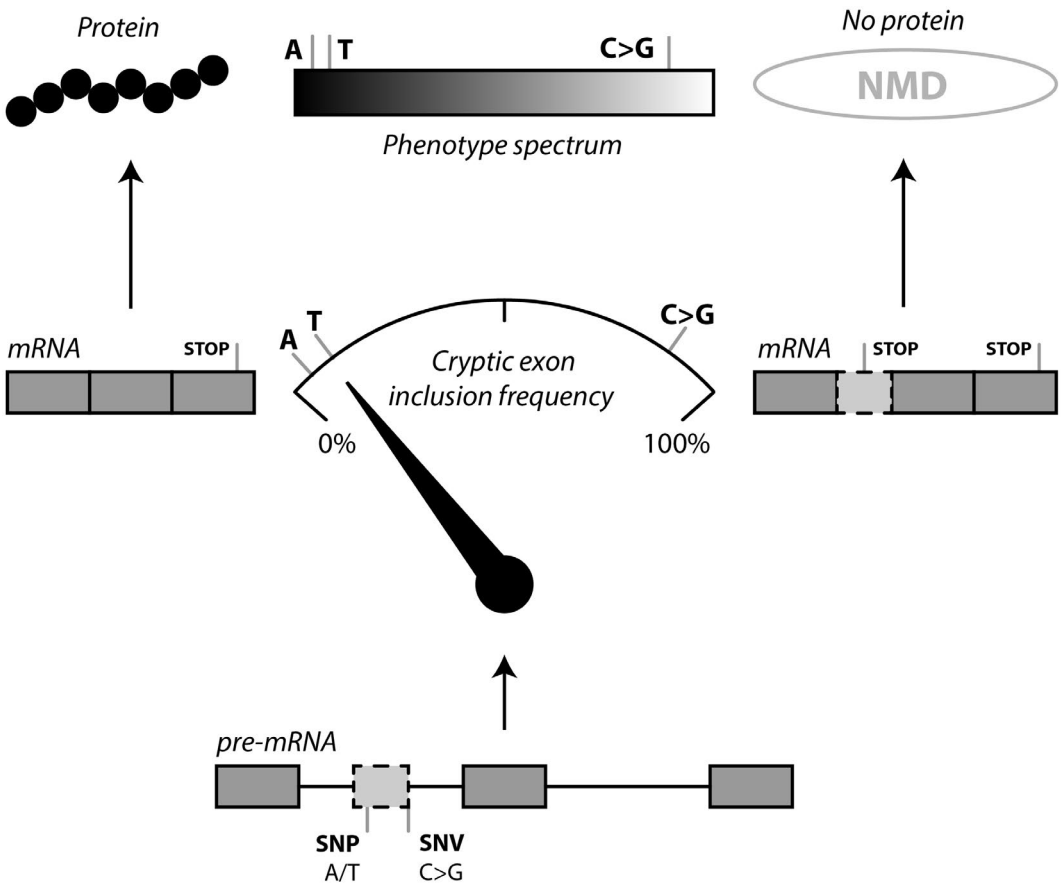

FIGURE 1 A general model of SNPtic exon splicing. A cryptic exon, or CE (dashed-line box) is included in mature transcripts at frequencies that vary depending on the genotype of the carrier. Because the CE encodes a premature stop codon more than 55 nt from the final splice junction, mature transcripts 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; *TSFM*, 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 1 and 2), 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 The splicing effects of common SNPs on 16 putative SNP-modulated cryptic ('SNPtic') exons

SNPtic exon	Transcript variant	GRCh38 coordinates	Acceptor MaxEnt	Donor MaxEnt	SNPs	MAF (%)	Δ Splice factor motifs
ATM-27a (Known)	NM_001351834.2	chr11+:108287410-108287438	8.12 -> 7.71	6.38	rs609261	48.8319	None
ttcatacttttcctct[c>t]agTCTACAGGTTGGCTGCATAGAAGAAAAAGgtagagttatttataatctt							
F8-13a (Known)	NM_000132.4	chrX-:154947100-154947221	11.28 -> 11.09	5.77	rs781928603	Rare	Loss of hnRNP C motifs, contraction of AGEZ
tcttctttttttttttt[del13T]cagACGGAGTCTCGCTCTGTAGCCCAGGCTGGAGTGCGGTGGCACGATCTCGGCTCACTGCAAGCTCCGCCACCCGGGTTCACGCCATTCTCCTGCCTCAGCCTCC CGAGTAGCTGGGACTACAGgtgcccgccaccacacccag							
IL16-6a (Known)	NM_004513.6	chr15+:81308111-81308137	3.42 -> 12.01	2.31	rs4778639	8.3666	None
atggttgtttccttttttta[t>g]GTGGACACAGAGGACTTCGTGCCAGAGgcaagatccctgtaaatatt							
LHCGR-6a (Known)	NM_000233.4	chr2-:48721573-48721779	8.60	0.48 (S), 1.96 -> 9.11 (L)	rs68073206	26.5775	None
tctttctgctttggttacagCCCATGGCAAAATTGTGATGAGGCAATAAAGGAGCTCACCCTTAAAGAAAAAAGAGAAAACATGGATTGGAATGACTCTGAAATGAAGAGATAGATGTGAAGCAA AAGAAGAGATCATCTCAGAGGACTCTCTTTTATATCACTGGATTCTAAAAATTGGTATCCTTGGTGCACTGCCTTTGTATAGTACTTTTACTTTGTGTAGATgtaa[t>g]tttacatgtataatt							
OAS1-2a (Known)	NM_016816.4	chr12+:112910744-112910847	6.72	3.14 -> 10.90 (rs116086311)	rs116086311, rs34137742	3.4944, 10.3035	Loss of SRSF9 motif (rs34137742)
tttgggttttaaaattccagATGTTATGGGATGCAGGAAGCAGCATGATCAGCAGCATTCTCTAGGTGCCAGGTTGAGAACAGGCTGTGGGGGAACCTGTAAAGAGGTTGCTGCCATAGTTCCGg [c>t]gagtga[c>t]ggtggtggctt							
TSFM-2a (Known)	NM_001172696.2	chr12+:57783615-57783652	10.93	0.07 -> 7.82	rs2014886	33.9457	None
actttttttttttttttcagACGGAGTCTTGCTCTGTGGCCCAGGCCGGCGTGCAATGg[c>t]acgatgttggctcactgc							
APC-11a (Probable)	NM_000038.6	chr5+:112822638-112822720	9.09	−0.15	rs2545162	32.0887	Gain MBNL1 motif
ggtttttctttgatttctagGTTCATTTTGTACATGTGGTATTTATAAATTGCATCATGCTGAACCATCTCATGTGAACTGGATCTCTCTAGTCATGCCACAAgcaagaaccatac[a>g]cttaat							
FGB-1a (Probable)	NM_005141.5	chr4+:154565186-154565235	9.10	8.56	rs2227401	16.1342	−1.7211 *(HExoSplice)*
tcattacacttattttacagATGAGAAAACTGGGGCACAGATAAAGCAACTTGCCCAAGGTCT[C>T]ATAGCTgtaagtcaaccctacggtca							
GHRL-4a (Probable)	NM_001302821.2	chr3-:10287028-10287326	10.21 (L), −0.85 -> 7.90 (S)	11.00	rs2075356	10.9625	None
acttccttttgattttccagGCATCAAAGAAAGGTCTATGCATATGAGAGAAAAAATACCTAGGGAAGGAGAAAAAGGAGAAGAAACAGACAAGAGAGAAAGGAGAAAGAGCTGGGAGCAGGG AAGGAATAGAAGGCCAGCAGGCCAGTGCTGAGAATGCTGGGCAGACCCCCTGCCAGGGCCTGAAATGCTGCCATTGGTGACTCAACCTTTTGTCTTCA[A>G]AACTGCAAGGGAG AAAAAAAATCTCACACTAGGGGGCACCACAGCTTTTATCAGCAATCCCATTAGAGGCTAAGGCTGAGAGGCATGAGAGTGCAAGgtaagtgtctgaaagatggt							
MYBPC3-12a (Probable)	NM_001042492.3	chr11-:47345756-47345832	2.68	1.90	rs10769255	22.2843	−0.5727 *(HExoSplice)*
ctctaccctcttctgaaaagAAATGAAGCCCG[G>A]TTAACCCTCTCCACACCCAAAAGAAAAGGGAAGAGGCCCGCTAAGCCTGGAGAGCCACACACAGgcaaagaaaagagcctggct							
OTC-9a (Probable)	NM_000531.6	chrX+:38412905-38413085	7.81	−7.41	rs5963419	36.3709	−0.7772 *(HExoSplice)*
acactgtggttcttttgtagCCAGAACACCACTGACACAGCCTGGTACCTGAGGC[C>T] CTTGAAACCAGATGTCAGATACTCCATTATAGCCACCACAGCTGAAATAGATTTCTCTACCTGCATATGTAGCAGAAGATCCAATCTCTCCGAAGACGGCTTCCACTGCATGTTG TTTGCTTCTGCACCCAAGTCTGGAGTAGATgcagctgagtgggatgcctg							

(Continues)

TABLE 1 (Continued)

SNPtic exon	Transcript variant	GRCh38 coordinates	Acceptor MaxEnt	Donor MaxEnt	SNPs	MAF (%)	Δ Splice factor motifs
ARSB-6a (Possible)	NM_000046.5	chr5-:78884914-78885128	8.48	−2.55 -> 5.63	rs337836	33.0072	None
tgtgccataccttcttctagATGCTGAGAAATTAGGAATGAACAAGTCAGTCAAGATCCTGCCTCTCAGGAAGCTGTATTCTAGTTGGGGGAGAAAGATGTTGGACAAATGAACACACAGATGA GCAAGATGACTGCCAGTTGTGATAAGTGCCAGGAAGGCAAAAAAGTATGTTGTGATAGATAGAGCTGGCTGATGTGAAGAAGATACATCAGGGAACAGGACCCAGACTCAG[a>g] tttggaaatcgtgatggtc							
CSF1R-15a (Possible)	NM_001349736.2	chr5-:150060668-150060768	10.92 -> 11.31	5.96	rs11952821	3.3946	None
ttccgtcttctccttct[c>t]agGATCCTACTGCTCCAAGTGTCAGGGGGATCCCGGTCACAGCATCCCTTAAACTCCTGGGCCCATCTCCTGGAATAGTCAGGAGCTGCACGGGCAGCTTGA Ggtataaagagagactgatag							
DMD-2a (Possible)	NM_004006.3	chrX-:32863759-32863915	−1.24 -> −1.30 (S), 5.84 -> 6.69 (L)	5.05	rs145743673	1.0861	None
tgatttggaacttcctgtag[A>G]CAGACCCTTACAGGCATGGAAGAAGAAAGAATGAATAAACCAAGGATGACTTCCCAGTAGGTGGAGGATGGGAATAATTAGGAAGAAGCTGTGTCTTGT CACCTATATTGTCCATACACAACTGCACCGTAGGGTAATTGAGAAATTATAAGAATGgtaataaatacgttttacat							
NF1-36a (Possible)	NM_001042492.3	chr17+:31324113-31324209	4.68	2.90 -> 10.65	rs35888506	37.8994	None
aatctcgctctatcgcccagGCTGGAGTGCAGTGGCACAATCTCAGCTCACTGCAAACTCCCCTCCCAGGTTCAAGCAATTCTCCTGCCTCAGCCTCCCAAGTAGCTGGGGCTACAGg[c>t] acgtgccaccacgcccag							
POC1B-9a (Possible)	NM_172240.3	chr12-:89461130-89461185	8.51 -> 9.13	8.44	rs11323565	4.6326	Longer AGEZ
tgcaattttttttttttt[insT]agCTCATCAGCTATCATTAATGTTAGTGTATTTTATGTGTGGCCCAAGACAATTCTTCCAATGTGGCCCAGGGAAAGCCAAAGATTGGACACTCCTGATAAG CGCTCAGTCAATAGTTTCTTCTTTAAGgttcgttagtacacttattg							

Note: The 'MAF (%) column' lists the haploid frequency of each minor allele according to data from the 1000 Genomes Project (Genomes Project et al., 2015), except for *rs781928603* ('Rare'), as precise data are not available for the T12 allele of this SNP. Sequence of each putative SNPtic exon sequence is shown in upper case, with flanking sequence in lower case. For SNPtic exons with short and long variants, the sequence and coordinates of the long variant are shown, and the sequence of the short variant is underlined. 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; *TSFM*, NG_016971.1.

TABLE 2 Sixteen putative SNPtic exons and their associated phenotypes

SNPtic exon	SNPs	Expected effect	SNP phenotype	Exon high-inclusion phenotype
ATM-27a (Known)	**rs609261** (NC_000011.10: g.108287407T>C)	Less inclusion	Lower cancer risk	Ataxia telangiectasia (poor coordination, prominent eye blood vessels and high cancer risk)
F8-13a (Known)	**rs781928603** (NC_000023.11: g.154947237_154947249del)	More inclusion	Mild haemophilia type A	Mild haemophilia type A
IL16-6a (Known)	**rs4778639** (NC_000015.10: g.81308110T>C)	More inclusion (no NMD)	Higher interleukin-16 levels in blood	Unknown
LHCGR-6a (Known)	**rs68073206** (NC_000002.12: g.48721568A>C)	Splice-switch (S > L)	Higher testosterone levels and higher androgen sensitivity index	Male pseudohermaphroditism
OAS1-2a (Known) (rs116086311)	**rs116086311** (NC_000012.12: g.112910849C>T), **rs34137742** (NC_000012.12: g.112910856C>T)	More inclusion	Higher risk of encephalitis and paralysis if infected with West Nile virus (rs34137742)	Unknown. Other *OAS1* mutations associated with higher risk of West Nile virus infection
TSFM-2a (Known)	**rs2014886** (NC_000012.12: g.57783654C>G)	More inclusion	PREDICTED: Higher risk of multiple sclerosis	Unknown. Other *TSFM* mutations associated with cardiomyopathy, encephalomyopathy and ataxia
APC-11a (Probable)	**rs2545162** (NC_000005.10: g.112822734G>A)	More inclusion	Higher colorectal cancer risk	Adenomatous polyposis (colon cancer)
FGB-1a (Probable)	**rs2227401** (NC_000004.12: g.154565229C>T)	Less inclusion	Higher blood fibrinogen levels	Afibrinogenemia (Persistent cerebral transient ischemic attacks, blood clots and 1/50th normal fibrinogen levels)
GHRL-4a (Probable)	**rs2075356** (NC_000003.12: g.10287125T>C)	Splice-switch (L > S)	Decreases cancer risk and increases bulimia risk	Unknown; other *GHRL* mutations associated with metabolic dysregulation
MYBPC3-12a (Probable)	**rs10769255** (NC_000011.10: g.47345820C>A)	Less inclusion	Slightly higher cognitive performance	Hypertrophic cardiomyopathy
OTC-9a (Probable)	**rs5963419** (NC_000023.11: g.38412940T>A)	Less inclusion	Increased risk of bipolar disorder	Hyperammonemia leading to brain damage and death

(Continues)

TABLE 2 (Continued)

SNPtic exon	SNPs	Expected effect	SNP phenotype	Exon high-inclusion phenotype
ARSB-6a (Possible)	**rs337836** (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)	**rs11952821** (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)	**rs145743673** (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)	**rs35888506** (NC_000017.11: g.31324211C>T)	More inclusion	PREDICTED: Higher cancer risk	Unknown; other *NF1* mutations cause neurofibromatosis type 1
POC1B-9a (Possible)	**rs11323565** (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; *OAS1*, NG_011530.2; *OTC*, NG_008471.1; *POC1B*, NG_041783.1; *TSFM*, 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 splicing (Figure 1) once population genetics are considered. Firstly, the direct effect of *rs34137742* is to remove a

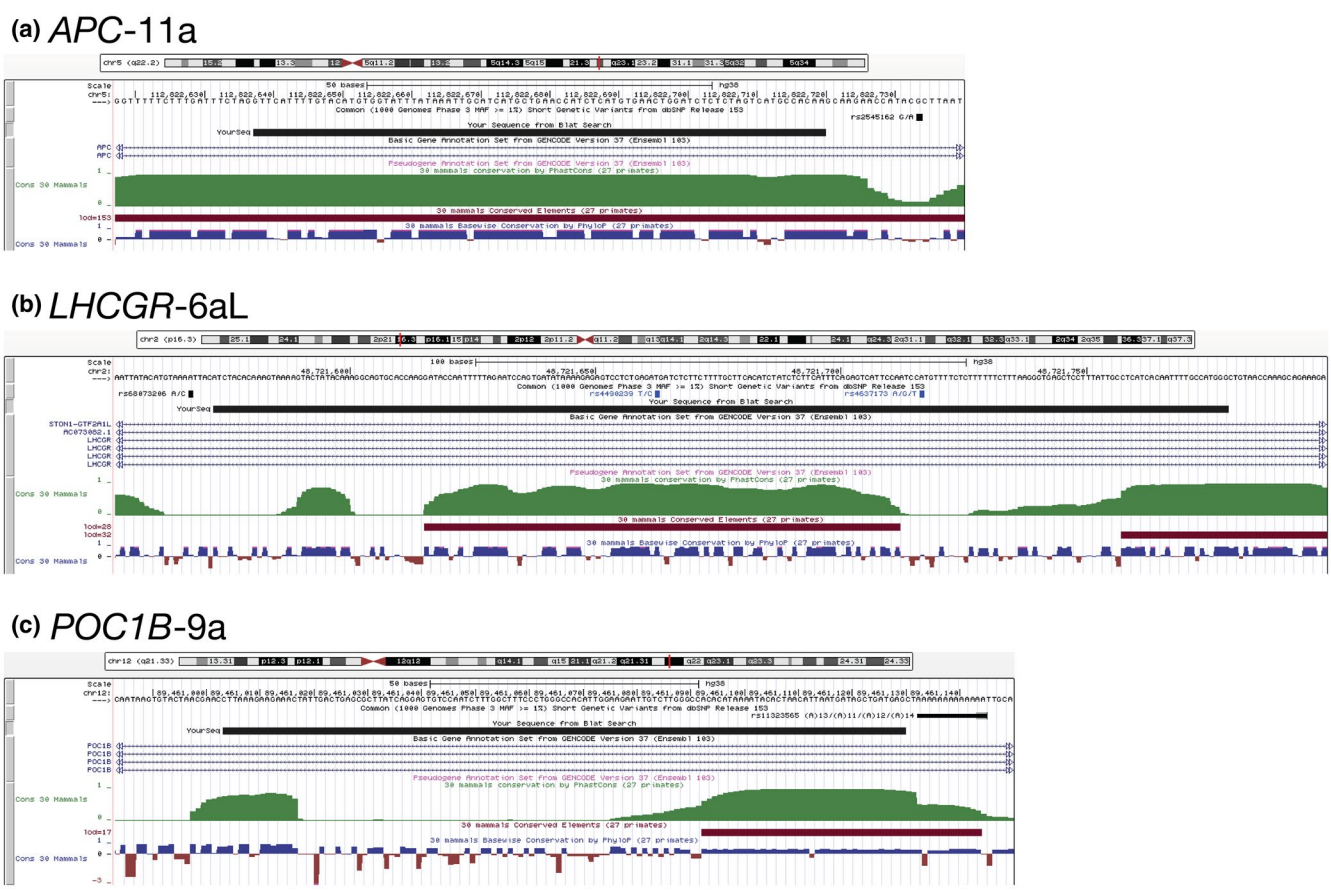

FIGURE 2 Cryptic exons *APC*-11a, *LHCGR*-6a and *POC1B*-9a exhibit high sequence conservation. Images were captured as screenshots from the UCSC Genome Browser (Kent et al., 2002). In descending order, displayed tracks are: Base position, dbSNP 153, input sequence, 'GENCODE V37' (aligned transcript 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

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 | *TSFM*-2a

Unlike the other three 'known' SNPtic exons, the SNPtic exon in *Ts Translation Elongation Factor*, *Mitochondrial* (*TSFM*–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 *TSFM* 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

TSFM 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 ±20 nt of their associated SNPtic exons. These examples are listed in Table 3, 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 ~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.

TABLE 3 SNPtic exons caused by common SNPs (≥1% haploid frequency) as reported by Sakaguchi and Suyama (2021)

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

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 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
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., Claeyssens, 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). Expasy, 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., Flicek, 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). Reccurrent 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

Konieczny, P., Stepniak-Konieczna, 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., Maghzian, 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., & Shoubridge, 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., Nkinsi, 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., Soukarieh, 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., Stuhrmann, 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