5′ splice site GC>GT and GT>GC variants differ markedly in terms of their functionality and pathogenicity

Jin-Huan Lin1,2,3 | Emmanuelle Masson1,4 | Arnaud Boulling1 | Matthew Hayden5 | David N. Cooper5 | Claude Férec1,4 | Zhuan Liao2,3 | Jian-Min Chen1

1EFS, Univ Brest, INSERM, UMR 1078, GGB, Brest, France
2Department of Gastroenterology, Shanghai Hospital, Second Military Medical University, Shanghai, China
3Shanghai Institute of Pancreatic Diseases, Shanghai, China
4CHRU Brest, Service de Génétique Médicale et de Biologie de la Reproduction, Brest, France
5Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, UK

Correspondence
Jian-Min Chen, INSERM U1078, Faculté de Médecine, Bâtiment E – 2ème étage - Bureau E201b, 22 avenue Camille Desmoulins, F-29238 Brest Cedex 3, France.
Email: jian-min.chen@univ-brest.fr

Present address
Arnaud Boulling, Université Paris-Saclay, INRAE, AgroParisTech, GABI 78350, Jouy-en-Josas, France.

Funding information
China Scholarship Council, Grant/Award Number: 201706580018; Institut National de la Santé et de la Recherche Médicale (INSERM), France

Abstract
In the human genome, most 5′ splice sites (~99%) employ the canonical GT dinucleotide whereas a small minority (~1%) use the noncanonical GC dinucleotide. The functionality and pathogenicity of 5′ splice site GT>GC (+2T>C) variants have been extensively studied but we know very little about 5′ splice site GC>GT (+2C>T) variants. Herein, we have addressed this deficiency by performing a meta-analysis of reported +2C>T “pathogenic” variants together with a functional analysis of engineered +2C>T substitutions using a cell culture-based full-length gene splicing assay. Our results establish proof of concept that +2C>T variants are qualitatively different from +2T>C variants in terms of their functionality and suggest that, in sharp contrast to +2T>C variants, most if not all +2C>T variants have no pathological relevance. Our findings have important implications for interpreting the clinical relevance of +2C>T variants and understanding the evolutionary switching between GT and GC 5′ splice sites in mammalian genomes.

KEYWORDS
+2C>T variant, +2T>C variant, 5′ splice site, full-length gene splicing assay, Human Gene Mutation Database, noncanonical GC dinucleotide

In the human genome, the vast majority (>99%) of introns are of the U2 type. Most of these (~99%) employ the canonical 5′ splice site GT dinucleotide whereas a minority (~1%) use the noncanonical 5′ splice site GC dinucleotide (Abril, Castelo, & Guigo, 2005; Burset, Seledtsov, & Solovyev, 2000, 2001; Parada, Munita, Cerda, & Gysling, 2014; Sheth et al., 2006). 5′ splice site GT>GC (or +2T>C) variants have been frequently described as causing human genetic disease (Stenson et al., 2017) and are routinely scored as splicing mutations (Mount et al., 2019). However, we have recently provided evidence that such variants in human disease genes may not invariably be pathogenic. Specifically, combining data derived from a meta-analysis of human disease-causing +2T>C variants and a cell culture-based Full-Length Gene Splicing Assay (FLGSA) of engineered +2T>C substitutions, we estimated that ~15–18% of +2T>C variants generate up to 84% wild-type transcripts (Lin et al., 2019). In another recent study, the functional effects of over 4,000 BRCA1 variants were analyzed by means of saturation genome editing (Findlay et al., 2018). Of these variants, 12 were noted to be of the +2T>C type (Chen et al., 2020); 25% (n = 3) of these +2T>C variants generated wild-type transcripts, a proportion not inconsistent with our estimated 15–18% rate. By contrast, we know very little about 5′ splice site GC>GT (or +2C>T) variants. Herein, we aimed to bridge this gap by performing a meta-analysis of “pathogenic” +2C>T variants reported to date and a functional analysis of artificially engineered +2C>T substitutions.
### Table 1: Description of the clinically identified germline 5′ splice site GC>GT (+2C>T) variants

| Gene symbol | Chr. | hg38 position | Reference allele | Variant allele | HGVS nomenclature | Original report or ClinVar accession number | Variant description and interpretation in original report or ClinVar rs number | MAF in non-Finnish Europeans | MAF across all gnomAD populations |
|-------------|------|---------------|-----------------|--------------|------------------|--------------------------------------------|---------------------------------|-----------------------------|----------------------------------|
| C1orf127    | 1    | 10949622      | G               | A            | NM_001170754.1:c.1290+2C>T | Lim et al. (2013)                             | Homozygote in a subject with autism, resulting from exome sequencing of 933 cases (MAF, 0.0011) and 869 controls; considered to be pathogenic | rs1281013                    | 0.03452                         | 0.04412                         |
| DMD         | X    | 32411750      | G               | A            | ENST000003570338:c.4233+2C>T | Griswold et al. (2015)                          | Identified in three of 2071 (MAF, 0.00072) autism spectrum disorder subjects; considered to be pathogenic | rs147474070                  | 0.0007263                         | 0.006742                         |
| THSD7B      | 2    | 137272664     | C               | T            | NM_001316349.2:c.2396+2C>T | Lim et al. (2013)                              | Homozygote in a subject with autism, resulting from exome sequencing of 933 cases (MAF, 0.0011) and 869 controls. Considered to be pathogenic | rs12622896                  | 0.009264                         | 0.02316                         |

**ClinVar-derived variants**

| Gene symbol | Chr. | hg38 position | Reference allele | Variant allele | HGVS nomenclature | Original report or ClinVar accession number | Variant description and interpretation in original report or ClinVar rs number | MAF in non-Finnish Europeans | MAF across all gnomAD populations |
|-------------|------|---------------|-----------------|--------------|------------------|--------------------------------------------|---------------------------------|-----------------------------|----------------------------------|
| COL4A4      | 2    | 227089866     | G               | A            | NM_000092.5:c.1459+2C>T | RCV000607055.1                               | One submitter; uncertain significance | rs932962404                | 0.000008835                      | 0.000004010                     |
| EYA1        | 8    | 71216690      | G               | A            | NM_001370335.1:c.1360+2C>T | VCV000163430.1                               | One submitter; uncertain significance | rs727503045                | 0.00004376                          |                                  |
| MUTYH       | 1    | 45331180      | G               | A            | NM_012222.2:c.1476+2C>T | VCV000187040.4                               | Seven submitters; six with an interpretation of “uncertain significance” and one of “likely benign” | rs140288388                | 0.00000388                        | 0.0001308                         |

**Abbreviations:** HGMD, Human Gene Mutation Database; HGVS, Human Genome Variation Society; gnomAD, the Genome Aggregation Database; MAF, minor allele frequency.

*a*In accordance with gnomAD’s exome plus genome data.
We first examined three +2C>T variants. C1orf127 c.1290+2C>T, DMD c.4233+2C>T, and THSD7B c.2396+2C>T (Table 1), logged in the Professional version of the Human Gene Mutation Database (http://www.hgmd.org; as of August 2019; Stenson et al., 2017). Both C1orf127 c.1290+2C>T and THSD7B c.2396+2C>T were identified by exome sequencing 933 subjects with autism spectrum disorders (ASD) and 869 controls; each variant was only found in a single case and in the homozygous state, corresponding to a minor allele frequency (MAF) of 0.0011 in the context of cases (Lin et al., 2013). These and other “rare complete gene knockouts” were considered to be important inherited risk factors for ASD by the original authors. However, as shown in Table 1, both variants occur at polymorphic frequencies in the non-Finnish European population and across all gnomAD populations (https://gnomad.broadinstitute.org/; Lek et al., 2016). DMD c.4233+2C>T was found in three of 2,071 ASD cases (MAF of 0.00072) but not in 904 controls of European White ancestry by means of targeted massively parallel sequencing of ASD-associated genes; it was considered to be a rare loss-of-function risk variant for ASD (Griswold et al., 2015). However, this variant has a MAF of 0.00073 in non-Finnish Europeans and a MAF of 0.00067 across all gnomAD populations (Table 1). Moreover, the DMD c.4233+2C>T variant is registered in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), wherein it is interpreted as being “Benign/Likely benign.” We then searched for +2C>T plus “mutation” or “variant” via Google (as of October 18, 2019) and found an additional three +2C>T variants, COL4A4 c.1459+2C>T, EYA1 c.1360+2C>T, and MUTYH c.1476+2C>T. They were all registered in ClinVar and almost invariably interpreted as being of “uncertain significance” (Table 1). Although the number (n = 6) of +2C>T variants identified to date is very limited, they serve to illuminate the problems of interpretation inherent to the +2C>T variants as a whole. This is due to the paucity of our knowledge about the functional effects of +2C>T variants on splicing.

Employing cell culture-based FLGSA, we have previously analyzed the functional impact of 103 engineered +2C>T variants. About 82% of these variants resulted in the generation of only aberrantly spliced transcripts (Lin et al., 2019). The other 18% of variants retained some ability to generate wild-type transcripts but none were capable of generating a wild-type transcript level equal to or higher than its wild-type counterpart (Lin et al., 2019). The accuracy and reliability of the FLGSA-derived data were supported by multiple lines of evidence. First, the findings have a strong biological basis; any +2T>C substitutions in the canonical 5′ splice site GT dinucleotides will decrease the complementarity between the 9-bp consensus sequence for the U2-type 5′ splice site (defined as sequence spanning nucleotide positions −3/+6) and the 3′-GUCCAUUCA-5′ sequence at the 5′ end of U1 small nuclear RNA (snRNA; see Lin et al., 2019 and references therein). Second, and consistent with this biological basis, the canonical 5′ splice sites in which substitution of GT by GC generated wild-type transcripts exhibit a stronger complementarity to the 5′ end of U1 snRNA than those sites whose substitutions of GT by GC did not lead to the generation of wild-type transcripts (Lin et al., 2019). Third, of the 103 engineered +2T>C substitutions, two had corresponding patient-derived in vivo splicing data reported. In both cases (i.e., HESX1 c.357+2T>C and SPINK1 c.194+2T>C), the FLGSA-derived data were in perfect agreement with the in vivo data. Fourth, in terms of the rate of GT>GC variants generating wild-type transcripts, the FLGSA-derived data agreed well with data obtained not only from disease-causing variants but also from BRCA1 variants analyzed in their natural genomic sequence contexts (Chen et al., 2020; Findlay et al., 2018; Lin et al., 2019). Fifth, the FLGSA-derived data correlated well with predictions made by SpliceAI, a recently developed artificial intelligence-based splicing prediction tool (Chen et al., 2020; Jaganathan et al., 2019). Finally, the FLGSA assay preserves better the natural genomic sequence context of the studied variant as compared to the commonly used minigene assay (Wu et al., 2017; Zou et al., 2016).

Herein, we employed the same experimental model system, FLGSA in transfected HEK293T cells, to analyze the functional impact of engineered +2C>T substitutions on splicing (Figure 1a; Table S1). All experiments were performed as previously described (Lin et al., 2019). Primer sequences pertaining to full-length gene amplification, mutagenesis and quantitative reverse-transcription polymerase chain reaction (RT-PCR) analysis are provided in Tables S2–S4, respectively. For the construction of the full-length gene expression cassettes, the pcDNA3.1/V5-His-TOPO vector was used in early experiments whereas the pcDNA3.1 vectors were used in later experiments (Table S2). Conventional qualitative RT-PCR analyses were invariably performed using primers located within the vector sequence (i.e., 5′-GGAGACCACAAGCTGGCTAGT-3′ [forward] and 5′-AGACCGAGGAGGGTTAGG-3′ [reverse] for genes cloned into the pcDNA3.1/V5-His-TOPO vector; or 5′-TATTACGACTACTATAGG-3′ [forward] and 5′-TAGAAGGCAACAGTGAGG-3′ [reverse] for genes cloned into the pcDNA3.1 vector). As for the quantitative RT-PCR analysis, a minigene was used as an internal control for quantifying the expression level of wild-type transcripts generated from either the wild-type or variant allele, and a primer pair comprising a gene-specific primer and a vector-specific primer (Table S4) were used to amplify the wild-type transcripts (refer to fig. 3 in Lin et al., 2019 for illustration).

Starting from 42 genes, we succeeded in analyzing 15 GC>GT variants from 15 different genes (Figure 1a,b; Figure S1). Before going into the results in detail, we would like to make several points. First, most of the 15 genes are characterized by alternative transcripts. Under our experimental model system, we were only able to analyze these genes in the context of the reference NM_accessions as specified in Figure 1c, whose authenticity was invariably confirmed by Sanger sequencing of the corresponding RT-PCR products from cells transfected with the respective wild-type full-length gene expression vectors. Second, as implied by the NM_accession, all these transcripts were protein coding. For each transcript, the number of exons, exonic location of the translation initiation codon, exonic location of the translation termination codon, and location of the +2C site are summarized in Table S5. In the case of four genes (i.e., APEX1, CTSL, DDT13, and EIF1AD), the +2C sites are located upstream of the translation initiation codon; in one gene (i.e., DUSP28), the +2C site is located downstream of the translation
FIGURE 1  Functional analysis of 5′ splice site GC>GT (+2C>T) substitutions by means of cell culture-based FLGSA. (a) Illustration of the experimental procedures and outcomes. (b) Conventional qualitative RT-PCR analyses of HEK293T cells transfected with full-length APEX1 and PRSS3 gene expression constructs carrying respectively the wild-type and +2C>T substitutions as examples. Wild-type transcripts (invariably confirmed by Sanger sequencing) emanating from both the wild-type and variant expression vectors are indicated by horizontal arrows. IVS, InterVening Sequence (i.e., an intron). The definition of wild-type transcripts was in accordance with NM_accessions in subpanel c. See Figure S1 for gel photographs of all 15 functionally analyzed +2C>T substitutions. (c) Details of the 15+2C>T substitutions and their effects on splicing in terms of the expression level of variant allele-derived normal transcripts relative to that of wild-type allele-derived transcripts. mRNA expression was determined using quantitative RT-PCR, and results were expressed as means ± SD from three independent transfection experiments. mRNA, messenger RNA; NED, not experimentally determined; RT-PCR, reverse-transcription polymerase chain reaction; SD, standard deviation.
termination codon-harboring exon; as for the remaining 10 genes, the +2C sites are located downstream of the translation initiation codon-harboring exons and upstream of the translation termination codon-harboring exons. In the five former cases, the potential effects of the +2C>T variants on splicing may not result in protein-altering transcripts. However, this is unlikely to affect the main conclusions of this study given that (a) we aimed to establish rules about the functional impact of +2C>T variants on splicing in an experimental model system and (b) altered functionality is a prerequisite for pathogenicity. In addition, all sense primers pertaining to full-length gene amplifications were located upstream of the respective translation initiation codons except in the case of DUSP28. In this latter case, we had to use a sense primer located downstream of the translation initiation codon (both the translation initiation codon and sense primer are located within exon 1) because of a primer design problem related to sequence context (Figure S2). This treatment was found not to affect normal splicing of the three-exon gene by means of conventional qualitative RT-PCR analysis (Figure S1). Finally, whereas the Human Genome Variation Society nomenclature (den Dunnen et al., 2016) was used to describe clinically identified variants, the traditional InterVening Sequence (IVS; i.e., an intron) nomenclature was used for the engineered substitutions in accordance with our previous publication (Lin et al., 2019). hg38 coordinates for these engineered substitutions are provided in Figure 1c.

The functional impact of the 15+2C>T substitutions on splicing was first evaluated by conventional qualitative RT-PCR analyses. It contrasted with that of the +2T>C substitutions in two respects. First, almost all the 15+2C>T substitutions generated the same transcript band(s) as their wild-type counterparts (Figure 1b and Figure S1). The only exception was TMED4, wherein the wild-type allele generated an additional shorter and barely visible RT-PCR band as compared with the substitution allele (Figure S1). We failed to establish the nature of this faint band by sequencing the gel-purified band. Second, only or almost only (i.e., cases wherein the additional bands were barely visible [e.g., CCL19]) wild-type transcripts were observed for 12 of the 15+2C>T substitutions (and their wild-type counterparts). In the remaining three cases, CDRT15, EIF1AD, and PRSS3, additional bands were clearly visible (Figure 1b; Figure S1). We attempted to sequence the corresponding gel-purified bands but did not obtain readable sequences owing to band faintness and/or nonspecific PCR amplification. Finally, we would like to emphasize that wild-type transcripts derived from the substitution alleles were also invariably confirmed by Sanger sequencing. This served to exclude the possibility that a variant may have led to the production of a transcript displaying only subtle differences as compared with the wild-type transcript. Take two of our previously FLGSA-analyzed variants for example. First, FABP7 IVS1+2T>C activated a cryptic 5’ splice site GT located 2 bp downstream of the normal one, resulting in the retention of the first 4 bp of the Intron 1 sequence (Lin et al., 2019). Second, SPINK1 c.88-1G>A resulted in the skipping of the first nucleotide of Exon 3 (Tang et al., 2019).

We then quantified the relative levels of the wild-type transcripts for the 12 +2C>T substitutions that generated only (or almost only) wild-type transcripts by means of quantitative RT-PCR analysis, with the results being summarized in Figure 1c. For the remaining three +2C>T substitutions (i.e., CDRT15 IVS2+2C>T, EIF1AD IVS1+2C>T, and PRSS3 IVS1+2C>T), the relative level of the wild-type transcripts was estimated to be approximately the same as those generated by their corresponding wild-type alleles based upon visual inspection of the relative intensities of the wild-type and additional transcripts bands (Figure S1). Thus, 87% (n = 13) of the 15+2C>T substitutions generated an equal or even higher level of normal transcripts as compared with that of their corresponding wild-type alleles (Figure 1c). As in the case of the +2T>C variants, the findings on the +2C>T variants also have a strong biological basis; any +2C>T substitutions in the noncanonical GC 5’ splice sites will increase the complementarity between the 9-bp consensus sequence for the U2-type 5’ splice site and the 3’-GUCAUUCU-5’ sequence at the 5’ end of U1 snRNA (Lin et al., 2019). Nonetheless, two +2C>T substitutions, TMED4 IVS4+2C>T and TUBB IVS4+2C>T, generated a much lower level of wild-type transcripts as compared with their wild-type alleles (Figure 1c), suggesting the possible involvement of sequence determinants for splicing beyond the short 9-bp consensus sequence motif.

The 15 genes used for the FLGSA assay were selected without consideration at the time of whether or not the corresponding +2C>T substitutions existed as polymorphisms in the general population. A retrospective search via VarSome (https://varsome.com/; Kopanos et al., 2019) and the UCSC Genome Browser (https://genome.ucsc.edu/; Kent et al., 2002) revealed that five of the 15 substitutions existed as natural variants. Four of these five variants were registered in gnomAD (https://gnomad.broadinstitute.org/; Lek et al., 2016); their allele frequencies are 0.00001596 (CDRT15 IVS2+2C>T; NM_001007530.2:c.411+2C>T), 0.00001591 (CTSL IVS2+2C>T; NM_001257973.2:c.-18+2C>T), 0.00003185 (EIF1AD IVS1+2C>T; NM_001242484.2:c.-117+2C>T), and 0.03481 (PRSS3 IVS1+2C>T; NM_002771.3:c.40+2C>T), respectively (Table S6). The first three rare variants were all annotated as “uncertain significance” whereas the last common variant was annotated as “likely benign” in accordance with the verdict of VarSome. This once again reflects the problems of interpretation associated with +2C>T variants.

Our study has a number of limitations. In common with our previous work (Lin et al., 2019), the genes studied here were confined to those with a genomic size of <8 kb for both practical and technical reasons. A significant proportion of the genes with a genomic size of <8 kb were not expressed in the transfected HEK293T cells due to cell-specific effects or for other reasons. Nonetheless, as we recently noted, the bottom line of the FLGSA assay is that (a) the genes were selected using a procedure that did not take into consideration the gene’s physiological function and expression location; (b) the genes used for the final analysis were those that underwent normal splicing in the context of their specified reference messenger RNA sequences; and (c) the generation (or not) of the wild-type transcripts from the variant alleles was observed under the same experimental conditions as for the wild-type allele (Chen et al., 2020). Consequently, any differences in terms of functional impact on splicing between wild-type and variant alleles could be directly attributable.
to the variant under study. This notwithstanding, one may argue whether findings from these comparatively small genes could be extrapolated to larger genes. While a definitive answer would require the availability of more population/patient data as well as functional data from well-validated assays, we would nevertheless favor an affirmative answer given the strong biological basis underlying our FLGSA-derived data.

As mentioned earlier, we analyzed the genes in the context of the reference NM_accessions as specified in Figure 1c. Consequently, the functional effect and pathogenic relevance of the studied +2C>T variants should only be considered within the given sequence contexts. This is an important and complex issue because (a) most of the studied genes have multiple alternative transcripts and (b) alternative transcripts may differ in terms of their temporal and spatial expression and hence may perform different functions. Herein, we use three genes for the purpose of illustration. First, APEX1 has four alternative transcripts, all comprising five exons. The corresponding genomic sequences have identical start and end nucleotide positions and identical exons 2–4. The only differences lie in the end positions of Exon 1 (Figure S3). The transcript under study, NM_080648.3, has the shortest Exon 1 sequence and the studied +2C>T variant is located within Intron 1. Thus, our constructed full-length APEX1 expression vector was potentially suitable for the analysis of all four alternative transcripts. However, only the NM_080648.3 transcript was expressed in the HEK293T cells transfected with both the wild-type and +2C>T variant expression vectors, a finding which served to illustrate the cell-specific effects on transcription. Second, PRSS3 has five alternative transcripts; the genomic sequence corresponding to the transcript under study, NM_002771.3, is the shortest one (Figure S4). The full-length PRSS3 expression vector was designed specifically for the analysis of the NM_002771.3 transcript. Consequently, we could not determine whether the +2C>T variant under study could modulate the splicing of a temporally and spatially co-expressed alternative transcript. Third, CTSL has eight alternative transcripts (Figure S5). The +2C>T variant studied here in the context of NM_001257973.2 would however result in a missense variant, p.Ala34Val (c.101C>T), in the context of all seven other alternative transcripts. Taken together, it is apparent that we should always bear in mind the issue of sequence context for the +2C>T variants.

In summary, we establish proof of concept that +2C>T variants behave quite differently from +2T>C variants in relation to the resulting splicing phenotype. Our findings suggest that, in sharp contrast with +2T>C variants, most if not all +2C>T variants have no pathological relevance. This conclusion has immediate implications for interpreting the clinical relevance of 5′ splice site GC>GT variants, most (if not all) of which would fall into the category of "benign variants" (Abou Tayoun et al., 2018; Richards et al., 2015). It should however be noted that any such conclusion should be arrived at through careful consideration of the variant’s appropriate reference sequence context. Our current study, together with our previous one (Lin et al., 2019), should also help us to understand in more detail the process of evolutionary switching between GT and GC 5′ splice sites in mammalian genomes (Abril et al., 2005).

ACKNOWLEDGMENTS

This study was presented as an oral communication at the 10th French Human and Medical Genetics Conference, Tours, January 21–24, 2020. This study was supported by the Institut National de la Santé et de la Recherche Médicale (INSERM), France; and the China Scholarship Council (No. 20170658018; J. H. L.). D. N. C. and M. H. acknowledge the financial support of Qiagen plc through a License Agreement with Cardiff University.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

J. H. L designed and performed the experiments. E. M. and A. B. assisted in performing the experiments. M. H. provided a list of GC-containing genes. D. N. C. critically revised the manuscript. C. F. and Z. L. provided funding and supervised the study. J. M. C. conceived the study, performed meta-analysis, and wrote the manuscript. All authors contributed to revision of the manuscript and approved the final manuscript.

DATA AVAILABILITY STATEMENT

All relevant data are available in this article and its the supplementary material.

ORCID

Lin-Huan Lin http://orcid.org/0000-0002-2346-5658
Emmanuelle Masson http://orcid.org/0000-0001-9777-523X
Matthew Hayden http://orcid.org/0000-0002-6272-3840
David N. Cooper http://orcid.org/0000-0002-8943-8484
Claude Férec http://orcid.org/0000-0002-2325-0710
Zhuo Liu http://orcid.org/0000-0001-8506-8159
Jian-Min Chen http://orcid.org/0000-0002-2424-3969

REFERENCES

Abou Tayoun, A. N., Pesaran, T., Di Stefano, M. T., Oza, A., Rehm, H. L., ... Biesecker, L. G. ClinGen Sequence Variant Interpretation Working Group (2018). Recommendations for interpreting the loss of function PV51 ACMG/AMP variant criterion. Human Mutation, 39(11), 1517–1524. https://doi.org/10.1002/humu.23626
Abril, J. F., Castelo, R., & Guigo, R. (2005). Comparison of splice sites in mammals and chicken. Genome Research, 15(1), 111–119. https://doi.org/10.1101/gr.3108805
Burset, M., Seledtsov, I. A., & Solovyev, V. V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Research, 28(21), 4364–4375. https://doi.org/10.1093/nar/28.21.4364
Burset, M., Seledtsov, I. A., & Solovyev, V. V. (2001). SpliceDB: Database of canonical and non-canonical mammalian splice sites. Nucleic Acids Research, 29(1), 255–259.
Chen, J. M., Lin, J. H., Masson, E., Liao, Z., Férec, C., Cooper, D. N., & Hayden, M. (2020). The experimentally obtained functional impact assessments of 5′ splice site GT>GC variants differ markedly from those predicted. Current Genomics, 21(1), 56–66. https://doi.org/10.2174/1389202921666202010141701
den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., ... Taschner, P. E. (2016). HGVS recommendations for the description of sequence variants: 2016 update. Human Mutation, 37(6), 564–569. https://doi.org/10.1002/humu.22981
mount, S. M., Avsec, Z., Carmel, L., Casadio, R., Celik, M. H., Chen, K., Lin, J. H., Tang, X. Y., Boulling, A., Zou, W. B., Masson, E., Fichou, Y., Lim, E. T., Raychaudhuri, S., Sanders, S. J., Sabo, A., MacArthur, D. G., … Daly, M. J. (2013). Rare complete knockouts in humans: Population distribution and significant role in autism spectrum disorders. Neuron, 77(2), 235–242. https://doi.org/10.1016/j.neuron.2012.12.029

Lin, J. H., Tang, X. Y., Boulling, A., Zou, W. B., Masson, E., Fichou, Y., … Chen, J. M. (2019). First estimate of the scale of canonical 5’ splice site GT>GC variants capable of generating wild-type transcripts. Human Mutation, 40(10), 1856–1873. https://doi.org/10.1002/humu.23821

Mount, S. M., Avsec, Z., Carmel, L., Casadio, R., Celik, M. H., Chen, K., … Zhao, H. (2019). Assessing predictions of the impact of variants on splicing in CAGIS. Human Mutation, 40(9), 1215–1224. https://doi.org/10.1002/humu.23869

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

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … Rehm, H. LACMG Laboratory Quality Assurance Committee (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine, 17(5), 405–424. https://doi.org/10.1038/gim.2015.30

Sheth, N., Roca, X., Hastings, M. L., Roeder, T., Krainer, A. R., & Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Research, 34(14), 3955–3967. https://doi.org/10.1093/nar/gkl556

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., … Cooper, D. N. (2017). The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Human Genetics, 136(6), 665–677. https://doi.org/10.1007/s00439-017-1779-6

Tang, X. Y., Lin, J. H., Zou, W. B., Masson, E., Boulling, A., Deng, S. J., … Chen, J. M. (2019). Toward a clinical diagnostic pipeline for SPINK1 intronic variants. Human Genomics, 13(1), 8. https://doi.org/10.1186/s13229-019-0193-7

Wu, H., Boulling, A., Cooper, D. N., Li, Z. S., Liao, Z., Chen, J. M., & Férec, C. (2017). In vitro and in silico evidence against a significant effect of the SPINK1 c.194G>A variant on pre-mRNA splicing. Gut, 66(12), 2195–2196. https://doi.org/10.1136/gutjnl-2017-313948

Zou, W. B., Boulling, A., Masson, E., Cooper, D. N., Liao, Z., Li, Z. S., … Chen, J. M. (2016). Clarifying the clinical relevance of SPINK1 intronic variants in chronic pancreatitis. Gut, 65(5), 884–886. https://doi.org/10.1136/gutjnl-2015-311168

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.