Open access • Posted Content • DOI:10.1101/829010

5’ splice site GC>GT variants differ from GT>GC variants in terms of their functionality and pathogenicity — Source link

Jin-Huan Lin, Emmanuelle Masson, Arnaud Boulling, Matthew J. Hayden ...+4 more authors

Institutions: French Institute of Health and Medical Research, Cardiff University, Second Military Medical University

Published on: 02 Nov 2019 - bioRxiv (Cold Spring Harbor Laboratory)

5' splice site GC>GT variants differ from GT>GC variants in terms of their functionality and pathogenicity

Jin-Huan Lin[1,2,3], Emmanuelle Masson[1,4], Arnaud Boulling[1], Matthew Hayden[5], David N. Cooper[5], Claude Férec[1,4], Zhuan Liao[2,3], Jian-Min Chen[1]

[1]EFS, Univ Brest, Inserm, UMR 1078, GGB, F-29200 Brest, France

[2]Department of Gastroenterology, Changhai Hospital, Second Military Medical University, Shanghai, China

[3]Shanghai Institute of Pancreatic Diseases, Shanghai, China

[4]CHRU Brest, Service de Génétique, Brest, France

[5]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom

*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

ABSTRACT

In the human genome, most 5' splice sites (~99%) employ the canonical GT dinucleotide whereas a small minority (~1%) use the non-canonical GC dinucleotide. The functionality and pathogenicity of 5' splice site GT>GC (i.e., +2T>C) variants have been extensively studied but we still know very little about 5' splice site GC>GT (+2C>T) variants. Herein, we sought to address this deficiency by performing a meta-analysis of identified +2C>T pathogenic variants together with a functional analysis of +2C>T substitutions using a cell culture-based full-length gene splicing assay. Our results establish a proof of concept that +2C>T variants are qualitatively different from +2T>C variants in terms of their functionality and pathogenicity and suggest that, in sharp contrast with +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 but might also improve our understanding of the evolutionary basis of switching between GT and GC 5' splice sites in mammalian genomes.

KEYWORDS

+2C>T variant, full-length gene splicing assay, Human Gene Mutation Database, human inherited disease, 5' splice site, +2T>C variant

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 non-canonical 5' splice site GC dinucleotide (Abril et al., 2005; Burset et al., 2000, 2001; Parada et al., 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 +2T>C substitutions, we estimated that ~15-18% of +2T>C variants generate up to 84% normal transcripts (Lin et al., 2019). In another recent study, the functional effects of over 4000 *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; 25% (n = 3) of these +2T>C variants generated wild-type transcripts, a proportion which is 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 identified "pathogenic" +2C>T variants and a functional analysis of artificially engineered +2C>T substitutions.

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 (HGMD; 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 (Lim 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 2071 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 variant is registered in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), wherein it is interpreted as being "Benign/Likely benign". A search for "+2C>T" plus "mutation" or "variant" via Google (as of October 18, 2019) yielded 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 +2T>C variants. 82% of these variants resulted in the generation of only aberrantly spliced transcripts. The other 18% of variants retained some ability to generate wild-type transcripts but none was capable of generating a transcript level equal to or higher than its wild-type counterpart (Lin et al., 2019). Herein, we employed the same experimental model system to analyze engineered +2C>T substitutions (**Fig. 1a; Supplementary Table S1**). All experiments were performed as previously described (Lin et al., 2019). Primer sequences used to amplify the full-length genes and perform mutagenesis and sequencing are available upon

request. Whereas the Human Genome Variation Society (HGVS) nomenclature (den Dunnen et al., 2016) was used for clinically identified variants, the traditional IVS (InterVening Sequence; i.e., an intron) nomenclature was used for the engineered substitutions in accordance with our previous publication (Lin et al., 2019).

We succeeded in analyzing 15 +2C>T substitutions from 15 different genes by *in vitro* expression analysis. Their functional impact contrasted with that of the +2T>C substitutions in two respects. First, all 15 +2C>T substitutions generated the same transcript band(s) as their wild-type counterparts (**Fig. 1b** and **Supplementary Fig. S1**). Second, only wild-type transcripts were observed for 12 of the 15 +2C>T substitutions (and their wild-type counterparts). We quantified the relative levels of the normally spliced transcripts for these 12 +2C>T substitutions by means of quantitative RT-PCR analysis as previously described (Lin et al., 2019), with the results being summarized in **Fig. 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 normally spliced transcripts was estimated to be approximately the same as the wild-type based upon visual inspection of the intensities of the wild-type and additional transcripts bands (**Supplementary Fig. S1**). Thus, 87% (n = 13) of the 15 +2C>T substitutions generated an equal or even higher level of normal transcripts as compared to that of their corresponding wild-type alleles (**Fig. 1c**). These findings are consistent with the expectation that any +2C>T substitutions in GC 5' splice sites will increase the complementarity between the 9-bp consensus sequence for the U2-type 5' splice site and the 3'-GUCCAUUCA-5' sequence at the 5' end of U1 snRNA (see (Lin et al., 2019) and references therein). Nonetheless, two +2C>T substitutions, *TMED4* IVS4+2C>T and *TUBB* IVS4+2C>T, generated a much lower level of normal transcripts as compared to their wild-type alleles (Fig. 1c), suggesting the possible involvement of sequence determinants for splicing beyond the short 9-bp consensus sequence motif.

In summary, we establish a proof of concept that +2C>T variants behave 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 but should also help us to understand the process of evolutionary switching between GT and GC 5' splice sites in mammalian genomes (Abril et al., 2005).

DISCLOSURE STATEMENT

The authors are unaware of any conflict of interest.

ACKNOWLEDGMENTS

J.H.L. was in receipt of a 20-month scholarship from the China Scholarship Council (No. 201706580018). This study was supported by the Institut National de la Santé et de la Recherche Médicale (INSERM), France. D.N.C. and M.H. acknowledge the financial support of Qiagen plc through a License Agreement with Cardiff University.

REFERENCES

Abril JF, Castelo R, Guigo R. Comparison of splice sites in mammals and chicken. *Genome Res* 2005; **15**: 111-119.

Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* 2000; **28**: 4364-4375.

Burset M, Seledtsov IA, Solovyev VV. SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res* 2001; **29**: 255-259.

den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat* 2016; **37**: 564-569.

Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, Janizek JD, Huang X, Starita LM, Shendure J. Accurate classification of *BRCA1* variants with saturation genome editing. *Nature* 2018; **562**: 217-222.

Griswold AJ, Dueker ND, Van Booven D, Rantus JA, Jaworski JM, Slifer SH, Schmidt MA, Hulme W, Konidari I, Whitehead PL, Cuccaro ML, Martin ER, Haines JL, Gilbert JR, Hussman JP, Pericak-Vance MA. Targeted massively parallel sequencing of autism spectrum disorder-associated

genes in a case control cohort reveals rare loss-of-function risk variants. *Mol Autism* 2015; **6**: 43.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation C. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016; **536**: 285-291.

Lim ET, Raychaudhuri S, Sanders SJ, Stevens C, Sabo A, MacArthur DG, Neale BM, Kirby A, Ruderfer DM, Fromer M, Lek M, Liu L, Flannick J, Ripke S, Nagaswamy U, Muzny D, Reid JG, Hawes A, Newsham I, Wu Y, Lewis L, Dinh H, Gross S, Wang LS, Lin CF, Valladares O, Gabriel SB, dePristo M, Altshuler DM, Purcell SM, Project NES, State MW, Boerwinkle E, Buxbaum JD, Cook EH, Gibbs RA, Schellenberg GD, Sutcliffe JS, Devlin B, Roeder K, Daly MJ. Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* 2013; **77**: 235-242.

Lin JH, Tang XY, Boulling A, Zou WB, Masson E, Fichou Y, Raud L, Le Tertre M, Deng SJ, Berlivet I, Ka C, Mort M, Hayden M, Leman R, Houdayer C, Le Gac G, Cooper DN, Li ZS, Férec C, Liao Z, Chen JM. First estimate of the scale of canonical 5' splice site GT>GC variants capable of generating wild-type transcripts. *Hum Mutat* 2019; **40**: 1856-1873.

Mount SM, Avsec Z, Carmel L, Casadio R, Celik MH, Chen K, Cheng J, Cohen NE, Fairbrother WG, Fenesh T, Gagneur J, Gotea V, Holzer T, Lin CF, Martelli PL, Naito T, Nguyen TYD, Savojardo C, Unger R, Wang R, Yang Y, Zhao H. Assessing predictions of the impact of variants on splicing in CAGI5. *Hum Mutat* 2019; **40**: 1215-1224.

Parada GE, Munita R, Cerda CA, Gysling K. A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Res* 2014; **42**: 10564-10578.

Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res* 2006; **34**: 3955-3967.

Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* 2017; **136**: 665-677.

(a)

Genes with a genomic size of <8 kb and contains at least one 5' splice site GC dinucleotide (**n = 42**)

↓ PCR amplification

Success (**n = 38**)

↓ Cloning into the expression vector

Success (**n = 32**)

↓ Transfection of each wild-type expression vector into the HEK293T cells, RT-PCR

Screening of genes for which RT-PCR analysis of transfected cells generated the band of expected size (**n = 15**)

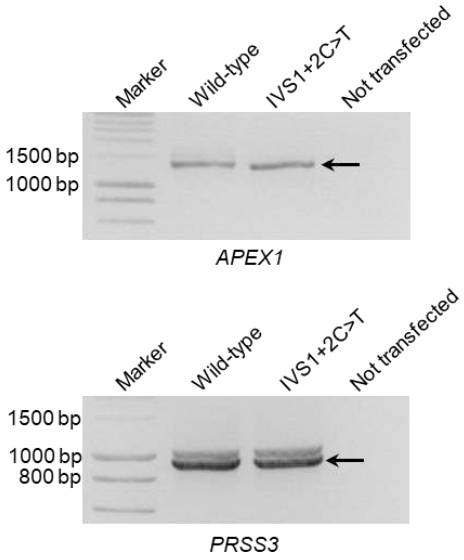
↓ Mutagenesis of each of the 15 available GC sites to GT

Success (**n = 15**)

↓ Transfection of each mutant expression vector, RT-PCR and sequencing, quantitative RT-PCR

Final dataset (15 GC>GT substitutions from 15 genes)

(b)



(c)

Gene	mRNA reference	Variant	hg38 position	% of normal expression level
APEX1	NM_080648.3	IVS1+2C>T	chr14: 20455330	103±26
CCL19	NM_006274.3	IVS3+2C>T	chr9: 34689928	85±16
CDRT15	NM_001007530.2	IVS2+2C>T	chr17: 14236280	NED
CTSL	NM_001257973.2	IVS2+2C>T	chr9: 87727704	123±25
DAZAP2	NM_001136266.2	IVS1+2C>T	chr12: 51238922	270±29
DDIT3	NM_004083.5	IVS1+2C>T	chr12: 57520416	154±16
DUSP28	NM_001033575.1	IVS2+2C>T	chr2: 240561521	238±23
EIF1AD	NM_001242484.2	IVS1+2C>T	chr11: 66002038	NED
MRPL24	NM_024540.4	IVS4+2C>T	chr1: 156738029	106±7
PIMREG	NM_001195228.2	IVS3+2C>T	chr17: 6447760	138±29
PRSS3	NM_002771.3	IVS1+2C>T	chr9: 33795615	NED
SLC39A7	NM_001288777.2	IVS2+2C>T	chr6: 33201390	126±20
TMED4	NM_182547.4	IVS4+2C>T	chr7: 44581091	58±9
TMEM31	NM_182541.2	IVS2+2C>T	chrX: 103712362	153±6
TUBB	NM_001293213.2	IVS4+2C>T	chr6: 30723433	20±3

Figure 1. Functional analysis of 5' splice site GT>GC (+2C>T) substitutions by means of cell culture-based FLGSA. (**a**) Illustration of the experimental procedures and outcomes. (**b**) 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. Normal transcripts (confirmed by sequencing) emanating from the expression vectors are indicated by arrows. IVS, InterVening Sequence (i.e., an intron). See **Supplementary Fig. 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. NED, not experimentally determined.

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	Variant description and interpretation in original report or ClinVar	rs number	MAF in non-Finnish Europeans*	MAF across all gnomAD populations*
HGMD-derived variants									
C1orf127	1	10949622	G	A	c.1290+2C>T	Homozygote in a subject with autism, resulting from exome sequencing of 933 cases (MAF, 0.0011) and 869 controls; considered to be pathogenic (Lim et al. 2013)	rs1281013	0.03452	0.04412
DMD	X	32411750	G	A	c.4233+2C>T	Identified in 3 of 2071 (MAF, 0.00072) autism spectrum disorder subjects; considered to be of pathogenic (Griswold et al. 2015)	rs147474070	0.0007263	0.0006742
THSD7B	2	137272664	C	T	c.2396+2C>T	Homozygote in a subject with autism, resulting from exome sequencing of 933 cases (MAF, 0.0011) and 869 controls. Considered to be pathogenic (Lim et al. 2013)	rs12622896	0.009264	0.02316
ClinVar-derived variants									
COL4A4	2	227089866	G	A	c.1459+2C>T	One submitter; uncertain significance (RCV000607055.1)	rs932962404	0.000008835	0.000004010
EYA1	8	71216690	G	A	c.1360+2C>T	One submitter; uncertain significance (VCV000163430.1)	rs727503045	0.000	0.00004376
MUTYH	1	45331180	G	A	c.1476+2C>T	Seven submitters; six with an interpretation of "uncertain significance" and one of "likely benign" (VCV000187040.4)	rs140288388	0.000	0.0001308

*In accordance with gnomAD's exome plus genome data.

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