Zhuan Liao ORCID iD: 0000-0001-8506-8159

Jian-Min Chen ORCID iD: 0000-0002-2424-3969

First estimate of the scale of canonical 5' splice site GT>GC variants capable of generating wild-type transcripts

Jin-Huan Lin[1,2,3†], Xin-Ying Tang[2,3†], Arnaud Boulling[1], Wen-Bin Zou[2,3], Emmanuelle Masson[1,4], Yann Fichou[1,5], Loann Raud[1], Marlène Le Tertre[1], Shun-Jiang Deng[2,3], Isabelle Berlivet[1], Chandran Ka[1,4,5], Matthew Mort[6], Matthew Hayden[6], Raphaël Leman[7], Claude Houdayer[8], Gerald Le Gac[1,4,5], David N. Cooper[6], Zhao-Shen Li[2,3], Claude Férec[1], Zhuan Liao[2,3*] and 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] CHU Brest, Service de Génétique, Brest, France

[5] Laboratory of Excellence GR-Ex, Paris, France

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

[7] Laboratoire de Biologie Clinique et Oncologique, Centre François Baclesse, Caen, France

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/humu.23821.

[8] Department of Genetics, F76000 and Normandy University, UNIROUEN, Inserm U1245, Normandy Centre for Genomic and Personalized Medicine, Rouen University Hospital, Rouen, France

An early version of this manuscript was posted in bioRxiv at https://www.biorxiv.org/content/10.1101/479493v1 (https://doi.org/10.1101/479493).

***Correspondence**

Jian-Min Chen, EFS, Univ Brest, Inserm, UMR 1078, GGB, F-29200 Brest, France.

Email: jian-min.chen@univ-brest.fr

Zhuan Liao, Department of Gastroenterology, Changhai Hospital, Second Military Medical University, Shanghai, China.

Email: liaozhuan@smmu.edu.cn.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

***Disclosure statement*:** The authors declare no conflict of interest.

Funding information

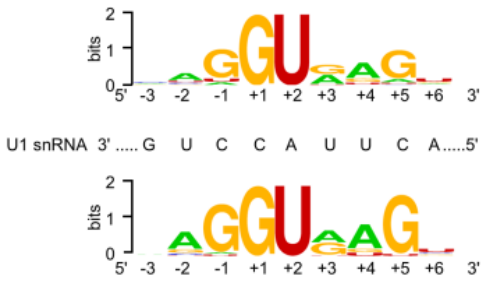
J.H.L., a joint PhD student between the Changhai Hospital and INSERM U1078, was in receipt of a 20-month scholarship from the China Scholarship Council (No. 201706580018). Support for this study came from the Institut National de la Santé et de la Recherche Médicale (INSERM) and the Etablissement Français du Sang (EFS), France; the National Natural Science Foundation of China (81470884 (to Z.L.)),

81770636 (to Z.L.) and 81700565 (to W.B.Z.)), the Shuguang Program of Shanghai (15SG33 (to Z.L.)), the Chang Jiang Scholars Program of Ministry of Education (Q2015190 (to Z.L.)), and the Scientific Innovation Program of Shanghai Municipal Education Committee (to Z.L.), China. M.M., M.H. and D.N.C. acknowledge financial support from Qiagen Inc. through a License Agreement with Cardiff University.

Abstract

It has long been known that canonical 5' splice site (5'SS) GT>GC variants may be compatible with normal splicing. However, to date, the actual scale of canonical 5'SSs capable of generating wild-type transcripts in the case of GT>GC substitutions remains unknown. Herein, combining data derived from a meta-analysis of 45 human disease-causing 5'SS GT>GC variants and a cell culture-based full-length gene splicing assay of 103 5'SS GT>GC substitutions, we estimate that ~15-18% of canonical GT 5'SSs retain their capacity to generate between 1 and 84% normal transcripts when GT is substituted by GC. We further demonstrate that the canonical 5'SSs in which substitution of GT by GC generated normal transcripts exhibit 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 normal transcripts. We also observed a correlation between the generation of wild-type transcripts and a milder than expected clinical phenotype but found that none of the available splicing prediction tools were capable of reliably distinguishing 5'SS GT>GC variants that generated wild-type transcripts from those that did not. Our findings imply that 5'SS GT>GC variants in human disease genes may not invariably be pathogenic.

Graphical Abstract



Based upon complementary data from the meta-analysis of 45 disease-causing 5'SS GT>GC variants and the cell culture-based full-length gene splicing analysis of 103 5'SS GT>GC substitutions, we have provided a first estimate of ~15-18% for the proportion of canonical GT 5'SSs that are capable of generating between 1 and 84% normal transcripts in case of the substitution of GT by GC. Given that even the retention of 5% normal transcripts can significantly ameliorate a patient's clinical phenotype, our findings imply the potential existence of hundreds or even thousands of disease-causing 5'SS GT>GC variants that may underlie relatively mild clinical phenotypes. Because 5'SS GT>GC variants can also give rise to relatively high levels of wild-type transcripts, our findings imply that 5'SS GT>GC variants may not invariably be pathogenic in disease-causative or disease-associated genes.

KEYWORDS

Canonical 5' splice site, full-length gene splicing assay, genotype and phenotype relationship, human gene mutation database, human inherited disease, non-canonical splice donor site

1. INTRODUCTION

The vast majority of eukaryotic introns are spliced by the U2 spliceosome (the only alternative U12 spliceosome is responsible for <0.5% of all introns (Parada, Munita, Cerda, & Gysling, 2014; Turunen, Niemela, Verma, & Frilander, 2013; Verma, Akinyi, Norppa, & Frilander, 2018)), which interacts with RNA sequences specifying the 5' and 3' splice sites (Papasaikas & Valcarcel, 2016; Sharp & Burge, 1997). In vertebrates, the 9-bp consensus sequence for the U2-type 5' splice site (5'SS) has traditionally been described as 5'-MAG/GURAGU-3' (where M denotes C or A, R denotes A or G and / denotes the exon-intron boundary; the corresponding nucleotide positions are denoted -3_-1/+1_+6) although in reality this consensus sequence does not reflect the true extent of sequence variability (Abril, Castelo, & Guigo, 2005; Burset, Seledtsov, & Solovyev, 2000; Mount, 1982; Roca et al., 2012; Roca, Krainer, & Eperon, 2013; Wong, Kinney, & Krainer, 2018). Base-pairing of this 9-bp sequence with 3'-GUCCAUUCA-5' at the 5' end of U1 snRNA (Figure 1A) is critical for splicing to occur (Kondo, Oubridge, van Roon, & Nagai, 2015; Kramer, Keller, Appel, & Luhrmann, 1984; Mount, Pettersson, Hinterberger, Karmas, & Steitz, 1983; Roca et al., 2013; Zhuang & Weiner, 1986). Although the GT dinucleotide in the first two intronic positions (in the context of DNA sequence) constitutes the most highly conserved portion of the U2-type 5'SS, it was reported, as early as 1983, that GC occasionally occurs in place of GT (Dodgson & Engel, 1983; Erbil & Niessing, 1983; King & Piatigorsky, 1983). Subsequent genome-wide analyses have established that this non-canonical 5'SS GC is present as wild-type in ~1% of human U2-type introns (Abril et al., 2005; Burset et al., 2000; Burset, Seledtsov, & Solovyev, 2001; Parada et al., 2014; Sheth et al., 2006). Importantly, the remaining nucleotides in these

evolutionarily fixed non-canonical GC 5'SSs exhibit a stronger complementarity to the 3'-GUCCAUUCA-5' sequence at the 5' end of U1 snRNA than those in the canonical GT 5'SSs (Figure 1A), thereby in all likelihood compensating for the decreased complementarity between the 5'SS and the 5' end of U1 snRNA due to the U to C substitution (Abril et al., 2005; Burset et al., 2000). Comparative genome analyses have also revealed frequent switching of U2-type introns from the canonical 5'SS GT subtype to the non-canonical 5'SS GC subtype during mammalian evolution (Abril et al., 2005; Churbanov, Winters-Hilt, Koonin, & Rogozin, 2008). Finally, GC has recently been ranked first among the six non-canonical 5′SSs identified by genome-wide RNA-seq analysis and splicing reporter assays (Erkelenz et al., 2018).

The finding that GC occasionally occurs instead of GT within the canonical 5'SS in some vertebrate genes implies that substitution of the canonical 5'SS GT by GC (termed a 5'SS GT>GC variant) may still allow normal splicing to occur. The first direct experimental evidence supporting such a postulate came in the late 1980s; analyses of both the splicing products of *in vitro* transcribed rabbit beta globin (*Hbb*) RNA in a HeLa cell nuclear extract and the splicing products of the *Hbb* gene transiently expressed in HeLa cells demonstrated that, of all the possible single nucleotide substitutions of the canonical 5'SS GT of the second and last intron of *Hbb*, only the substitution of T by C was compatible with normal splicing, albeit at a much reduced rate (approximately 10% of normal; see also Figure 1B) (Aebi, Hornig, Padgett, Reiser, & Weissmann, 1986; Aebi, Hornig, & Weissmann, 1987). Further supporting evidence came from the study of disease-causing 5'SS GT>GC variants, some of which were reported to generate wild-type transcripts (see below). Additionally, the activation of cryptic non-canonical 5'SS GC has also been reported

as a consequence of some disease-causing variants (Kralovicova et al., 2011; Pagani et al., 2002).

The above notwithstanding, to date, the actual scale of canonical 5'SSs capable of generating wild-type transcripts in the case of GT>GC substitutions, both in the context of the frequency of such substitutions and the level of wild-type transcripts generated by such substitutions, remains unknown owing to the intrinsic complexity of splicing (Boehm et al., 2018; De Conti, Baralle, & Buratti, 2013; Krainer, 2015; Wong et al., 2018; Zhang, Arias, Ke, & Chasin, 2009) and the lack of suitable model systems for study. This issue has important implications for medical genetics since mutant genotypes retaining even a small fraction of their normal function may differ significantly from null genotypes in terms of their associated clinical phenotypes (e.g., 5% normal *CFTR* gene expression is sufficient to prevent the lung manifestations of cystic fibrosis (Ramalho et al., 2002; Raraigh et al., 2018); for hemophilia B and other coagulation factor deficiencies, raising plasma levels above 5% normal often results in milder bleeding phenotypes (Den Uijl et al., 2011; Scalet et al., 2019)). Herein, we attempted to address this issue by employing two distinct but complementary approaches in concert.

2. MATERIALS AND METHODS

2.1. Meta-analysis of disease-causing 5'SS GT>GC variants

Human disease-causing 5'SS GT>GC variants logged in the Professional version of the Human Gene Mutation Database (HGMD; http://www.hgmd.cf.ac.uk/ac/index.php; as of June 2017) (Stenson et al., 2017) were

used as starting material. The procedure of the meta-analysis is described in Figure 1C.

2.2. Cell culture-based full-length gene splicing assay

The outline of the cell culture-based full-length gene splicing assay is illustrated in Figure 1C.

2.2.1. Amplification of full-length gene sequences

For this experiment, we focused on genes whose genomic sizes were <8 kb (from the translation initiation codon to the translation termination codon) and whose exons numbered ≥3. Long-range PCR was performed in a 25 μL reaction mixture containing 0.5 U KAPA HiFi HotStart DNA Polymerase (Kapa Biosystems), 0.75 μL KAPA dNTP Mix (300 μM final), 5 μL 5 × KAPA HiFi Buffer, 50 ng DNA, and 0.3 μM forward and reverse primers (primer sequences available upon request). The PCR program comprised an initial denaturation at 95°C for 5 min, followed by 30 cycles of denaturation at 98°C for 20 s, annealing at 66°C for 15 s, extension at 72°C for 1 min/kb, and a final extension at 72°C for 5 min. In some of the cases where the desired fragments could not be obtained, a second amplification was attempted: PCR was performed using 50 ng DNA in a 50 μL reaction mixture with 2.5 U TaKaRa LA Taq DNA polymerase (TaKaRa), 8 μL dNTP Mixture (400 μM final), 5 μL 10 × LA PCR Buffer, and 1 μM forward and reverse primers; thermal cycling conditions were initial denaturation at 94°C for 1 min, 30 cycles of denaturation at 98°C for 10 s, annealing and extension at 68°C for 1 min/kb, and a final extension at 72°C for 10 min.

2.2.2. Cloning of the amplified full-length wild-type gene sequences into the expression vector

Early experiments were performed by means of TA cloning. In those cases in which the PCR products contained multiple bands, the band of the expected size was gel purified using the QIAquick Gel Extraction Kit (Qiagen) and 3'-A overhangs added; in cases where a single and expected band was obtained, 3'-A overhangs were directly added to the PCR products amplified from the KAPA HiFi HotStart DNA Polymerase (this step was omitted for those amplified using the TaKaRa LA Taq DNA polymerase). The resulting products were cloned into the pcDNA3.1/V5-His-TOPO vector (Invitrogen) in accordance with the manufacturer's instructions. Transformation was performed using Stellar Competent Cells (TaKaRa) or XL10-Gold Ultracompetent Cells (Agilent Technologies). Transformed cells were spread onto LB agar plates with 50 µg/mL ampicillin and incubated at 37°C overnight. Plasmid constructs containing inserts in the correct orientation were selected by PCR screening using the HotStarTaq Master Mix Kit (Qiagen).

Later experiments were performed by means of in-fusion cloning. PCR products of the expected size were purified using the QIAquick Gel Extraction Kit (Qiagen) after gel electrophoresis. The purified products were cloned into *EcoR*I restriction site of the linearized pcDNA3.1(+) vector with the In-Fusion HD Cloning kit (TaKaRa) according to the manufacturer's instructions. Transformation was performed using Stellar Competent Cells (TaKaRa) or XL10-Gold Ultracompetent Cells (Agilent Technologies). Transformed cells were spread onto LB agar plates with 50 µg/mL ampicillin and incubated at 37°C overnight. Plasmid constructs containing inserts were confirmed by PCR using the HotStarTaq Master Mix Kit (Qiagen).

2.2.3. Mutagenesis

Variants were introduced into the wild-type full-length gene expression constructs by means of the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies). Mutagenesis was performed in a 25.5 µL mixture containing 1.25 U PfuUltra HF DNA polymerase, 0.5 µL dNTP mix, 2.5 µL 10× reaction buffer, 1.5 µL QuikSolution, 100 ng wild-type plasmid, and 62.5 ng each mutagenesis primer (primer sequences available upon request). The PCR program had an initial denaturation at 95°C for 2 min, followed by 18 cycles of denaturation at 95°C for 1 min, annealing at 60°C for 50 s, and extension at 68°C for 1 min/kb, and a final extension at 68°C for 7 min. The PCR products were transformed into XL10-Gold Ultracompetent cells (Agilent Technologies) after treated with *Dpn*I at 37°C for 1 h. Transformed cells were spread onto LB agar plates with 50 µg/mL ampicillin and incubated at 37°C overnight. Selected colonies were cultured overnight. Plasmids were isolated using the QIAprep Spin Miniprep Kit (Qiagen) and the successful introduction of the desired substitutions was validated by DNA sequencing with the BigDye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems).

2.2.4. Cell culture, transfection, RNA extraction, and reverse transcription

Human embryonic kidney 293T (HEK293T) and HeLa cells were cultured in the Dulbecco's modified Eagle's medium (BioWhittaker) with 10% fetal calf serum (Eurobio). 3.5×10^5 cells were seeded per well in 6-well plates 24 h before transfection. For conventional RT-PCR analyses, 1 µg wild-type or variant plasmid, mixed with 2 µL jetPEI DNA transfection reagent (Polyplus-transfection), was used for transfection per well. For real-time quantitative RT-PCR analyses, 500 ng wild-

type or variant plasmid was mixed with 500 ng pGL3-GP2 minigene for transfection (Boulling, Chen, Callebaut, & Férec, 2012; Zou, Boulling, Masamune, et al., 2016; Zou et al., 2017). Forty-eight hours after transfection, total RNA was extracted using the RNeasy Mini Kit (Qiagen). RT was performed with 200 U SuperScript III Reverse Transcriptase (Invitrogen), 500 µM dNTPs, 4 µL 5 × First-Strand Buffer, 5 mM dithiothreitol, 2.5 µM 20mer-oligo (dT), and 1 µg total RNA. The resulting complementary DNA (cDNA) were treated with 2U RNaseH (Invitrogen) to degrade the remaining RNA.

2.2.5. Conventional RT-PCR analyses and sequencing of the resulting products

Conventional RT-PCR was performed in a 25-µL reaction mixture containing 12.5 µL HotStarTaq Master Mix (Qiagen), 1 µL cDNA, and 0.4 µM each primer (5'-GGAGACCCAAGCTGGCTAGT-3' (forward) and 5'-AGACCGAGGAGAGGGTTAGG-3' (reverse) for TA cloning-obtained plasmids (both primers are located within the pcDNA3.1/V5-His-TOPO vector sequence); 5'-TAATACGACTCACTATAGGG-3' (forward) and 5'-TAGAAGGCACAGTCGAGG-3' (reverse) for in-fusion cloning-obtained plasmids (both primers are located within the pcDNA3.1(+) vector sequence)). The PCR program had an initial denaturation step at 95°C for 15 min, followed by 30 cycles of denaturation at 94°C for 45 s, annealing at 58°C for 45 s, and extension at 72°C for 1 min/kb (in the step to screen wild-type genes for which RT-PCR analysis of transfected cells generated a single or quasi-single band of expected size) or for 2 min (in the step to analyze the splicing outcomes of 5'SS GT>GC substitutions), and a final extension step at 72°C for 10 min. RT-PCR products of a single band were cleaned by ExoSAP-IT (Affymetrix). In the case of multiple bands, the band

corresponding to the normal-sized product was excised from the agarose gel and then purified by QIAquick Gel Extraction Kit (Qiagen). Sequencing primers were those used for the RT-PCR analyses. Sequencing was performed by means of the BigDye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems).

2.2.6. Quantitation of the relative level of correctly spliced transcripts from constructs with artificially introduced GT>GC substitutions

The relative level of correctly spliced transcripts in association with GT>GC substitutions that generated only wild-type transcripts (confirmed by Sanger sequencing) was determined by real-time quantitative RT-PCR analyses in accordance with Pfaffl's mathematical model (Pfaffl, 2001), essentially as described elsewhere (Boulling et al., 2012; Zou, Boulling, Masamune, et al., 2016; Zou et al., 2017). In brief, the previously constructed *GP2* minigene (Boulling et al., 2012) was employed as an internal control for this analysis. 500 ng wild-type or variant plasmid was mixed with an equal amount of minigene for co-transfection in HEK293T cells. Real-time RT-PCR analysis was performed 48 hours after transfection. After RNA extraction and reverse transcription, seven serial dilutions of the resulting cDNA (1:10, 1:20, 1:40, 1:80, 1:160, 1:320 and 1:640) were used to determine the real-time RT-PCR efficiency for each primer set. Finally, a 1:80 dilution of cDNA was used to quantify the relative expression ratio of the variant construct versus the wild-type construct, which was normalized against the co-transfected *GP2* minigene. Results were obtained from three independent transfection experiments, with each experiment being performed in three replicates.

2.3. Pictogram analysis of the 9-bp 5'SS signal sequences associated with 5'SS GT>GC variants

The 9-bp canonical 5'SS signal sequences of the currently studied disease-associated and artificially introduced GT>GC variants were extracted from the UCSC Genome Browser (https://genome.ucsc.edu/). The respective pictograms were constructed using WebLogo (http://weblogo.berkeley.edu/).

2.4. *In silico* splicing prediction

In silico splicing prediction was performed by means of Alamut® Visual v.2.11 rev. 0 (https://www.interactive-biosoftware.com/; Interactive Biosoftware, Rouen, France) and a recently reported prediction protocol, Splicing Prediction in Consensus Elements (SPiCE; https://sourceforge.net/projects/spicev2-1/) (Leman et al., 2018), under default conditions.

2.5. Relative mRNA expression levels of genes of interest in HEK293 and HeLa cells

Relative mRNA expression levels of genes of interest, represented as transcripts per million (TPM), in HEK293 and HeLa cells were obtained from the Human Protein Atlas (https://www.proteinatlas.org/) (Uhlen et al., 2015).

2.6. Variant nomenclature

Nomenclature with respect to disease-causing variants followed Human Genome Variation Society (HGVS) recommendations (den Dunnen et al., 2016). For ease of description, artificially introduced 5'SS GT>GC substitutions were named in

accordance with the traditional IVS (InterVening Sequence; i.e., an intron) nomenclature.

3. RESULTS AND DISCUSSION

3.1. Estimation by meta-analysis of disease-causing 5'SS GT>GC variants

First, we performed a meta-analysis of disease-causing 5'SS GT>GC variants logged in the Professional version of Human Gene Mutation Database (HGMD; as of June 2017) (Stenson et al., 2017), with a view to generating an "*in vivo*" dataset to estimate the scale of those 5'SS GT>GC variants capable of generating wild-type transcripts. Employing a stringent approach (Figure 1C), we identified 45 disease-causing 5'SS GT>GC variants (from 42 genes) that were informative with respect to the presence or absence of wild-type transcripts derived from the variant allele (Table 1; see Supp. Table S1 for more information including affected intron number, reference mRNA accession number, chromosomal location, hg38 coordinates, and patient-derived tissue or cells used for RT-PCR analysis, etc.). It should be noted that the assignments of "presence" or "absence" of mutant allele-derived wild-type transcripts depended upon the agarose gel evaluation of RT-PCR products as described in the corresponding original publications. Thus, we conservatively annotated an isolated case (i.e., the *PCCB* c.183+2T>C variant) which was not found to generate wild-type transcripts upon agarose gel evaluation of RT-PCR products but which was nevertheless found to generate <0.1% normal wild-type transcripts by means of quantitative RT-PCR (Desviat et al., 2006), as generating no wild-type transcripts.

The 45 informative 5'SS GT>GC variants comprised 30 homozygotes, 13 hemizygotes and 2 compound heterozygotes (Table 1). Whilst the presence or absence of wild-type transcripts derived from the variant allele was straightforward for all

homozygous or hemizygous variants studied here, the two compound heterozygotes required special treatment. In the case of the *CD3E* c.520+2T>C variant, the pathogenic *CD3E* variant in *trans* was a nonsense variant in exon 6. Sequencing of the patient-derived, normal-sized RT-PCR products failed to demonstrate the presence of the exon 6 variant, suggesting that the wild-type transcripts were derived from the c.520+2T>C allele (Soudais, de Villartay, Le Deist, Fischer, & Lisowska-Grospierre, 1993). In the case of the *PNPLA2* c.757+2T>C variant, the second *PNPLA2* variant in *trans* was a missense variant, c.749A>C (p.Gln250Pro). RT-PCR analysis detected only the c.749A>C variant mRNA in skeletal muscle from the patient, indicating the absence of detectable wild-type transcript emanating from the c.757+2T>C allele (Lin et al., 2012).

15.6% (n=7) of the 45 informative variants were found to have been capable of generating some correctly spliced transcripts (Table 1). Information on the expression level of the variant allele-derived wild-type transcripts relative to that of the wild-type transcripts from a normal control (by definition, 100%) was directly available from four of the seven original publications (i.e., *CD3E* c.520+2T>C (Soudais et al., 1993), *CD40LG* c.346+2T>C (Seyama et al., 1998), *DMD* c.8027+2T>C (Bartolo et al., 1996) and *SLC26A2* c.-26+2T>C (Hastbacka et al., 1999)), and was reported to range from 1-15% of normal in individual cases (Table 1). All three of the remaining variants generated both wild-type and aberrant transcripts (i.e., *CAV3* c.114+2T>C (Muller et al., 2006), *PLP1* c.696+2T>C (Aoyagi et al., 1999) and *SPINK1* c.194+2T>C (Kume, Masamune, Kikuta, & Shimosegawa, 2006)); based upon visual inspection of the original gel photographs, we estimate that the relative expression

level of the mutant allele-derived wild-type transcripts in these three cases would fall within the 1-15% range.

Taken together, the meta-analysis of disease-causing variants suggests that 15.6% of 5'SS GT>GC variants retained the ability to generate between 1 and 15% correctly spliced transcripts relative to their wild-type counterparts.

3.2. Estimation from the cell culture-based full-length gene splicing assay of 5'SS GT>GC substitutions

To corroborate the findings derived from the above "*in vivo*" dataset, we sought to generate an "*in vitro*" dataset of 5'SS GT>GC substitutions. In this regard, we have previously used a cell culture-based full-length gene splicing assay to analyze a series of *SPINK1* intronic variants; and this full-length gene expression system has proved itself in practice by accurately representing the in vivo situation in the context of the observed splicing pattern of a disease-causing 5'SS GT>GC variant, *SPINK1* c.194+2T>C (Zou, Boulling, Masson, et al., 2016; Zou, Masson, et al., 2016). Specifically, the full-length 7-kb *SPINK1* genomic sequence (including all four exons plus all three introns of the gene) was cloned into the pcDNA3.1/V5-His-TOPO vector (Boulling et al., 2012). The full-length gene splicing assay preserves better the natural genomic context of the studied variants as compared to the commonly used minigene splicing assay, a point of importance given the highly context-dependent and combinatorial nature of alternative splicing regulation (Fu & Ares, 2014). Moreover, the full-length gene splicing assay can be readily used to evaluate all intronic variants including those located near the first or last exons of the gene (Tang et al., 2019). Despite these advantages, the full-length gene assay cannot easily be

applied to large-sized genes owing to the technical difficulties inherent in amplifying and cloning long DNA fragments into the expression vector. Finally, it is pertinent to point out that, to functionally evaluate the impact on splicing of any given gene variant in a transient expression system, it is highly desirable to use of cell types of pathophysiological relevance owing to the tissue specificity of the splicing process in some instances (Boehm et al., 2018; De Conti et al., 2013; Wong et al., 2018; Zhang et al., 2009). However, this may not always be possible in practice, particularly if variants in multiple genes are to be analyzed in large-scale studies. For example, one recent study that measured 5'SS activity in the context of three minigenes was performed in transfected HeLa cells (Wong et al., 2018) whereas another study, that analyzed the splicing of thousands of minigene molecules, was performed in transfected HEK293 cells (Ke et al., 2018). In the present study, we used HEK293T cells for transfection as previously described (Wu, Boulling, Cooper, Li, Liao, Férec, et al., 2017; Zou, Boulling, Masamune, et al., 2016; Zou, Boulling, Masson, et al., 2016).

Bearing in mind the aforementioned advantages and disadvantages, we co-opted a cell culture-based full-length gene splicing assay (Figure 1C; Figure 2A). In brief, for various technical reasons/practical considerations, we firstly selected genes whose genomic sizes did not exceed 8 kb (from the translation initiation codon to the translation termination codon) and whose exons numbered ≥3, in order to construct full-length gene expression vectors; we then screened those genes, which had yielded a single or quasi-single band of the expected size by means of the RT-PCR analysis of transfected cells, for subsequent mutagenesis of all available 5'SS GT dinucleotides in the construct (for details of the genes selected and screened, see Supp. Table S2). In

the end, we succeeded in functionally analyzing 103 GT>GC substitutions from 30 different genes (Supp. Table S3). 18.4% (n=19) of these artificially introduced 5'SS GT>GC substitutions generated wild-type transcripts (all confirmed by Sanger sequencing; Figure 2B and Supp. Figure S1), a finding that concurs with the 15.6% value obtained from the meta-analysis of disease-causing 5'SS GT>GC variants.

Only wild-type transcripts were observed for 10 of the aforementioned 19 5'SS GT>GC substitutions (e.g., *FATE1* IVS1+2T>C in Figure 2B). In other words, no aberrantly spliced transcripts were observed in these 10 cases. It is possible that aberrantly spliced transcripts may be rendered invisible by RNA degradation mechanisms such as nonsense-mediated mRNA decay (NMD) (Lykke-Andersen & Jensen, 2015; Popp & Maquat, 2016). One way to test such a possibility is to add an NMD inhibitor such as cycloheximide (Pereverzev et al., 2015) to the cell culture medium, but this was considered to be beyond the scope of the present study. We quantified the relative level of correctly spliced transcripts for these 10 5'SS GT>GC substitutions using our previously described quantitative RT-PCR method (Boulling et al., 2012; Wu, Boulling, Cooper, Li, Liao, Chen, et al., 2017; Zou, Boulling, Masamune, et al., 2016). Here we would like to reiterate that a co-transfected minigene construct was used as an internal control in this analysis (Figure 3A), a prerequisite for obtaining accurate results. As shown in Figure 3B, the relative level of correctly spliced transcripts emanating from these 10 substitutions is remarkably similar to that observed for the disease-causing 5'SS GT>GC variants in terms of the lower bound (2-5% vs. 1-5%); however, the upper bound for the level of correctly spliced transcripts (84%) is much higher than the corresponding 15% value observed for the disease-causing 5'SS GT>GC variants (Table 1). We were initially puzzled by

this disparity, but it could be accounted for in two different ways. On the one hand, the currently analyzed disease-causing variants were likely to be biased toward those that generated either no wild-type transcripts or only a low level. On the other hand, given (i) that 5'SS GC may occur as wild-type in the human genome, (ii) the highly degenerate nature of the 5'SS splice signal sequences and (iii) the complex regulation of the splicing process *in vivo*, it is entirely possible that a 5'SS GT>GC variant may behave similarly to its original wild-type counterpart. It should however be noted that no single GC variant was found to have an identical or higher than normal splicing activity than its 5'SS GT counterpart (Figure 3B), an observation consistent with the inherently weaker binding of any 5'SS GC sites, as compared to their corresponding wild-type GT sites, to U1 snRNA.

Additionally, the single RT-PCR band of wild-type transcript size from either the wild-type *CCDC103* gene or the *CCDC103* IVS1+2T>C variant (refer to Supp. Figure S1) was revealed by Sanger sequencing to comprise the correctly spliced transcript and an alternatively spliced transcript; the level of the correctly spliced transcripts generated from the variant allele was estimated to be ~18% of that generated from the wild-type allele based upon evaluation of the corresponding sequence peak heights (Supp. Figure S2). By contrast, we did not attempt to quantify the relative expression level of correctly spliced transcripts for the remaining 8 GT>GC substitutions due to the co-presence of aberrantly spliced transcripts (e.g., *DBI* IVS2+2T>C in Figure 2B). Nonetheless, based upon the relative intensities of the wild-type and aberrant transcript bands (Figure 2B; Supp. Figure S1), we consider it unlikely that the relative expression level of correctly spliced transcripts in these cases will have fallen outside of the abovementioned, experimentally obtained, 2-84% range.

Finally, we sequenced some aberrantly spliced transcripts (n=12), which resulted from exon skipping, retention of intronic sequence or deletion of partial exonic sequences (Table 2). Most notably, the *PRSS2* IVS4+2T>C substitution activated a cryptic 5'SS GC that is located 15 bp upstream of the normal one, resulting in the deletion of the last 17 bp of exon 4 (i.e., the most abundant band generated by *PRSS2* IVS4+2T>C; Figure 2B).

3.3. Synthesis of estimates from two distinct but complementary datasets

We obtained remarkably similar findings in terms of both the frequency of 5'SS GT>GC substitutions generating wild-type transcripts (15.6% vs. 18.4%) and the relative level of mutant allele-derived wild-type transcripts at the lower bound (2-5% vs. 1-5%) from two quite distinct yet complementary datasets (i.e., the 45 human disease-causing variants vs. the 103 artificially introduced substitutions). The consistent relative level of mutant allele-derived wild-type transcripts at the lower bound across the two datasets suggested that the gel-based analytical method is sensitive enough to detect as little as ~1% of normally spliced transcripts. The apparent disparity in terms of the relative level of mutant-derived wild-type transcripts at the upper bound between the two datasets (15% vs. 84%) can however be accounted for largely by the selection bias inherent to disease-causing variants. Therefore, we estimate that some 15-18% of 5'SS GT>GC substitutions generate between 1 and 84% of wild-type transcripts.

3.4. Exploration of the mechanisms underlying the generation (or not) of wild-type transcripts by 5'SS GT>GC variants

As mentioned above, canonical GT and non-canonical GC 5'SSs in the human genome exhibit different patterns of sequence conservation, the latter showing stronger complementarity to the 3'-GUCCAUUCA-5' sequence at the 5' end of U1 snRNA (Figure 1A). We postulated that the canonical 5'SSs whose substitutions of GT by GC generated normal transcripts (termed group 1) might also exhibit stronger complementarity to the aforementioned 9-bp sequence than those sites whose substitutions of GT by GC did not lead to the generation of normal transcripts (termed group 2). We therefore extracted the 9-bp sequence tracts surrounding the corresponding groups of the 45 disease-causing 5'SS GT>GC variants (Supp. Tables S1) and those of the 103 functionally analyzed 5'SS GT>GC substitutions (Supp. Table S3). Comparison of the resulting pictograms confirmed our postulate in both contexts, the respective pictograms for the combined group 1 variants (n=26) and combined group 2 variants (n=122) being provided in Figure 4. It should be emphasized that the surrounding 9-bp sequence tract is an important (but certainly not the only) factor in determining whether or not a given 5'SS GT>GC variant will generate some wild-type transcripts. A simple example may be used to illustrate this point: the *DMD* c.8027+2T>C variant (which generates 10% of wild-type transcripts) contrasts with the *NCAPD2* c.4120+2T>C variant (which generates no wild-type transcripts) despite occurring in an identical 9-bp sequence tract, AAGGTATGA (see Supp. Table S1).

We also explored whether the creation or disruption of splice enhancer/silencer motifs by the 5'SS GT>GC variants could be associated with the generation or not of

some wild-type transcripts. To this end, we employed ESEfinder and RESUE-ESE provided by the Alamut software suite under default conditions. We were unable to draw any meaningful conclusions, primarily due to the short and degenerate nature of the splicing enhancer/silencer binding motifs.

3.5. Correlation between the retention of wild-type transcripts and a milder than expected clinical phenotype

Given that even the retention of a small fraction of normal gene function may significantly impact the clinical phenotype (Den Uijl et al., 2011; Ramalho et al., 2002; Raraigh et al., 2018; Scalet et al., 2019), we reviewed the original publications describing the seven disease-causing 5'SS GT>GC variants that generated at least some wild-type transcript (Table 1) with respect to the accompanying genotypic and phenotypic descriptions. In six cases, the variants were specifically described as being associated with mild clinical phenotypes as compared to their classical disease counterparts (see Supp. Table S1). In the remaining case (*SPINK1* c.194+2T>C), the original publication (Kume et al., 2006) was not informative in this regard; however, it is known that homozygosity for this variant causes chronic pancreatitis with variable expressivity (Ota et al., 2010) whereas null *SPINK1* genotypes cause severe infantile isolated exocrine pancreatic insufficiency (Venet et al., 2017).

The above noted apparent correlation between the retention of some wild-type transcripts and a milder than expected phenotype prompted us to postulate that 5'SS GT>GC variants previously reported to confer a milder than expected phenotype but having no supportive patient-derived transcript expression data, may have a tendency to be associated with a non-canonical 5'SS GC signal. We therefore collated a total of

six such variants (i.e., *CYB5R3* c.463+2T>C (Yilmaz, Cogulu, Ozkinay, Kavakli, & Roos, 2005), *HBB* c.315+2T>C (Frischknecht et al., 2009), *HPRT* c.485+2T>C (Hladnik, Nyhan, & Bertelli, 2008), *LAMB2* c.3327+2T>C (Wuhl et al., 2007), *LMNA* c.1968+2T>C (Bar et al., 2017) and *MTTP* c.61+2T>C (Al-Mahdili, Hooper, Sullivan, Stewart, & Burnett, 2006); Supp. Table S4). In this regard, two points require clarification. First, in two cases, patient-derived transcript expression data were available (Hladnik et al., 2008; Wuhl et al., 2007); these cases were however further explored here because the corresponding expression data were insufficiently informative for them to be listed in Supp. Table S1 (for explanations, see Sup. Table S4). Second, five of these six variants (all germline) were derived from the HGMD dataset whereas the remaining one (*LMNA* c.1968+2T>C) (Bar et al., 2017), a somatic variant, was obtained from a literature search; this somatic variant was included owing to its clear phenotypic impact. Pictogram analysis of the six corresponding 9-bp canonical 5'SSs revealed a non-canonical 5'SS GC signal (Supp. Figure S3). Notably, one of the variants affected the splice donor splice site of *HBB* intron 2 (i.e., *HBB* c.315+2T>C) (Frischknecht et al., 2009), site of the previously analyzed orthologous variant in the rabbit *Hbb* gene (Aebi et al., 1986; Aebi et al., 1987). We were able to study the effect of the *HBB* intron 2 GT>GC variant on splicing by means of the full-length gene assay and found that it had indeed retained the ability to generate normal *HBB* transcripts (Figure 5).

3.6. Prediction of the functional effect of 5'SS GT>GC variants

In a previous study, we observed that the functional effect of the *SPINK1* c.194+2T>C variant could not be accurately predicted by the widely used Alamut® software suite under default conditions (Zou et al., 2017). Herein, we extended this analysis to the 45

disease-causing 5'SS GT>GC variants as well as the 19 functionally analyzed 5'SS GT>GC substitutions that generated some wild-type transcripts. Whereas SpliceSiteFinder-like tended to predict a slightly reduced score, MaxEntScan, NNSPLICE and GeneSplicer invariably yielded no scores, for all variants tested (Table 1; Supp. Table S3). We also analyzed these variants by means of the recently developed SPiCE tool (Leman et al., 2018); all variants were invariably predicted to alter splicing.

Taken together, we conclude that none of the available splicing prediction tools were able to distinguish 5'SS GT>GC variants generating wild-type transcripts from those that did not generate wild-type transcripts, a reflection perhaps of our rather poor understanding of the rules governing the use of GC as a viable 5'SS site. This highlights the importance of functional analysis for accurate interpretation of this particular type of variation, although caution should always be exercised when extrapolating from functional analytical data to prediction of clinical phenotype.

3.7. Further experiments using the cell culture-based full-length gene splicing assay

The above notwithstanding, it is nevertheless appropriate to have reservations regarding the accuracy and reliability of data obtained from the experimental model system we adopted. To validate these data, we therefore performed additional experiments using expression plasmids available in the Brest laboratory. First, we sought to test whether 5'SS GT>GA or 5'SS GT>GG substitutions may also generate wild-type transcripts. To this end, we firstly mutated a set of 5'SS GT sites to GA and GG; and then analyzed the resulting substitutions under same experimental conditions

as per the analysis of 5'SS GT>GC substitutions. None of the 15 5'SS GT>GA substitutions or 18 5'SS GT>GG substitutions analyzed were found to generate wild-type transcripts (Table 3; Supp. Figure S4). This served to exclude the (albeit rather remote) possibility that the generation of wild-type transcripts from 5'SS GT>GC substitutions was simply spurious.

As mentioned earlier, the impact of genetic variants on splicing may be tissue- or cell-specific in some instances. To test this possibility, we analyzed 10 5'SS GT>GC substitutions that generated wild-type transcripts (Left panel, Figure 6A) and 10 5'SS GT>GC substitutions that did not generate wild-type transcripts (Right panel, Figure 6A) in HEK293T cells for full-length splicing assay in HeLa cells. We observed entirely consistent findings in the two cell lines in terms of the generation of wild-type transcripts or not (Figure 6A; see also Supp. Figures S5 and S6).

Finally, we evaluated whether the generation of wild-type transcripts by 5'SS GT>GC substitutions could be in some way related to the natural expression status of their corresponding genes in HEK293T or HeLa cells. We thus obtained the relative expression levels of the 10 genes shown in Figure 6A in the two cell lines via the Human Protein Atlas website (https://www.proteinatlas.org/). It should be noted that no data were available for HEK293T cells; we therefore used data for HEK293 cells instead (Figure 6B). The relative mRNA expression levels (represented as transcripts per million (TPM)) of the genes harboring wild-type transcript-generating 5'SS GT>GC substitutions in the two cells ranged from none to high (defined as >1000 TPM by Human Protein Atlas). This suggests that the generation of wild-type transcripts by 5'SS GT>GC substitutions is not related to the natural expression status of their corresponding genes in either HEK293T or HeLa cells.

Taken together, these three lines of evidence supported the reliability and accuracy of the data obtained from our experimental model system. It may therefore be concluded that whether a given 5'SS GT>GC variant will generate some wild-type transcripts or not is primarily dependent on the corresponding gene's sequence context, although tissue- or cell-specific splicing effects may exist in some instances.

4. CONCLUSIONS

Based upon complementary data from the meta-analysis of 45 disease-causing 5'SS GT>GC variants and the cell culture-based full-length gene splicing analysis of 103 5'SS GT>GC substitutions, we have provided a first estimate of ~15-18% for the proportion of canonical GT 5'SSs that are capable of generating between 1 and 84% normal transcripts in case of the substitution of GT by GC. Extrapolation of the 15-18% value to the entire human genome implies that in at least 30,000 U2-type introns, the substitution of 5'SS GT by GC would result in the retention of partial ability to generate wild-type transcripts. Given that even the retention of 5% normal transcripts can significantly ameliorate a patient's clinical phenotype, our findings imply the potential existence of hundreds or even thousands of disease-causing 5'SS GT>GC variants that may underlie relatively mild clinical phenotypes. Because 5'SS GT>GC variants can also give rise to relatively high levels of wild-type transcripts, our findings imply that 5'SS GT>GC variants may not invariably be pathogenic in disease-causative or disease-associated genes. We believe that our study will not only raise fresh awareness of the 5'SS GT versus 5'SS GC issue in health and disease but also stimulate new studies that aim to better predict the functional effects of 5'SS GT>GC variants detected in a clinical context.

Acknowledgements

We are grateful to the original authors who reported the disease-causing 5'SS GT>GC variants studied here. We thank Nicolas Tomat and Léhna Bouchama (Brest, France) for technical assistance.

REFERENCES

Abril, J. F., Castelo, R., & Guigo, R. (2005). Comparison of splice sites in mammals and chicken. *Genome Res, 15*(1), 111-119. doi:10.1101/gr.3108805

Adly, N., Alhashem, A., Ammari, A., & Alkuraya, F. S. (2014). Ciliary genes *TBC1D32/C6orf170* and *SCLT1* are mutated in patients with OFD type IX. *Hum Mutat, 35*(1), 36-40. doi:10.1002/humu.22477

Aebi, M., Hornig, H., Padgett, R. A., Reiser, J., & Weissmann, C. (1986). Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. *Cell, 47*(4), 555-565.

Aebi, M., Hornig, H., & Weissmann, C. (1987). 5' cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5' splice region, not by the conserved 5' GU. *Cell, 50*(2), 237-246.

Ahmed, I., Mittal, K., Sheikh, T. I., Vasli, N., Rafiq, M. A., Mikhailov, A., . . . Vincent, J. B. (2014). Identification of a homozygous splice site mutation in the dynein axonemal light chain 4 gene on 22q13.1 in a large consanguineous family from Pakistan with congenital mirror movement disorder. *Hum Genet, 133*(11), 1419-1429. doi:10.1007/s00439-014-1475-8

Al-Mahdili, H. A., Hooper, A. J., Sullivan, D. R., Stewart, P. M., & Burnett, J. R. (2006). A mild case of abetalipoproteinaemia in association with subclinical hypothyroidism. *Ann Clin Biochem, 43*(Pt 6), 516-519. doi:10.1258/000456306778904650

Aldahmesh, M. A., Mohamed, J. Y., & Alkuraya, F. S. (2012). A novel mutation in *PRDM5* in brittle cornea syndrome. *Clin Genet, 81*(2), 198-199. doi:10.1111/j.1399-0004.2011.01808.x

Allamand, V., Sunada, Y., Salih, M. A., Straub, V., Ozo, C. O., Al-Turaiki, M. H., . . . Campbell, K. P. (1997). Mild congenital muscular dystrophy in two patients

with an internally deleted laminin alpha2-chain. *Hum Mol Genet, 6*(5), 747-752.

Aoyagi, Y., Kobayashi, H., Tanaka, K., Ozawa, T., Nitta, H., & Tsuji, S. (1999). A de novo splice donor site mutation causes in-frame deletion of 14 amino acids in the proteolipid protein in Pelizaeus-Merzbacher disease. *Ann Neurol, 46*(1), 112-115.

Bar, D. Z., Arlt, M. F., Brazier, J. F., Norris, W. E., Campbell, S. E., Chines, P., . . . Gordon, L. B. (2017). A novel somatic mutation achieves partial rescue in a child with Hutchinson-Gilford progeria syndrome. *J Med Genet, 54*(3), 212-216. doi:10.1136/jmedgenet-2016-104295

Bartolo, C., Papp, A. C., Snyder, P. J., Sedra, M. S., Burghes, A. H., Hall, C. D., . . . Prior, T. W. (1996). A novel splice site mutation in a Becker muscular dystrophy patient. *J Med Genet, 33*(4), 324-327.

Biancheri, R., Grossi, S., Regis, S., Rossi, A., Corsolini, F., Rossi, D. P., . . . Filocamo, M. (2014). Further genotype-phenotype correlation emerging from two families with *PLP1* exon 4 skipping. *Clin Genet, 85*(3), 267-272. doi:10.1111/cge.12154

Boehm, V., Britto-Borges, T., Steckelberg, A. L., Singh, K. K., Gerbracht, J. V., Gueney, E., . . . Gehring, N. H. (2018). Exon junction complexes suppress spurious splice sites to safeguard rranscriptome integrity. *Mol Cell, 72*(3), 482-495 e487. doi:10.1016/j.molcel.2018.08.030

Boulling, A., Chen, J. M., Callebaut, I., & Férec, C. (2012). Is the *SPINK1* p.Asn34Ser missense mutation *per se* the true culprit within its associated haplotype? *WebmedCentral GENETICS, 3*, WMC003084 (Available at: https://www.webmedcentral.com/article_view/003084). Accessed 003014 November 002018.

Burset, M., Seledtsov, I. A., & Solovyev, V. V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res, 28*(21), 4364-4375.

Burset, M., Seledtsov, I. A., & Solovyev, V. V. (2001). SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res, 29*(1), 255-259.

Churbanov, A., Winters-Hilt, S., Koonin, E. V., & Rogozin, I. B. (2008). Accumulation of GC donor splice signals in mammals. *Biol Direct, 3*, 30. doi:10.1186/1745-6150-3-30

Das, S., Levinson, B., Whitney, S., Vulpe, C., Packman, S., & Gitschier, J. (1994). Diverse mutations in patients with Menkes disease often lead to exon skipping. *Am J Hum Genet, 55*(5), 883-889.

De Conti, L., Baralle, M., & Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA, 4*(1), 49-60. doi:10.1002/wrna.1140

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. *Hum Mutat, 37*(6), 564-569. doi:10.1002/humu.22981

Den Uijl, I. E., Mauser Bunschoten, E. P., Roosendaal, G., Schutgens, R. E., Biesma, D. H., Grobbee, D. E., & Fischer, K. (2011). Clinical severity of haemophilia A: does the classification of the 1950s still stand? *Haemophilia, 17*(6), 849-853. doi:10.1111/j.1365-2516.2011.02539.x

Desviat, L. R., Clavero, S., Perez-Cerda, C., Navarrete, R., Ugarte, M., & Perez, B. (2006). New splicing mutations in propionic acidemia. *J Hum Genet, 51*(11), 992-997. doi:10.1007/s10038-006-0068-3

Dodgson, J. B., & Engel, J. D. (1983). The nucleotide sequence of the adult chicken alpha-globin genes. *J Biol Chem, 258*(7), 4623-4629.

Dolcini, L., Caridi, G., Dagnino, M., Sala, A., Gokce, S., Sokucu, S., . . . Minchiotti, L. (2007). Analbuminemia produced by a novel splicing mutation. *Clin Chem, 53*(8), 1549-1552. doi:10.1373/clinchem.2007.089748

Erbil, C., & Niessing, J. (1983). The primary structure of the duck alpha D-globin gene: an unusual 5' splice junction sequence. *EMBO J, 2*(8), 1339-1343.

Erkelenz, S., Theiss, S., Kaisers, W., Ptok, J., Walotka, L., Muller, L., . . . Schaal, H. (2018). Ranking noncanonical 5' splice site usage by genome-wide RNA-seq analysis and splicing reporter assays. *Genome Res, 28*(12), 1826-1840. doi:10.1101/gr.235861.118

Frischknecht, H., Dutly, F., Walker, L., Nakamura-Garrett, L. M., Eng, B., & Waye, J. S. (2009). Three new beta-thalassemia mutations with varying degrees of severity. *Hemoglobin, 33*(3), 220-225. doi:10.1080/03630260903089060

Fu, X. D., & Ares, M., Jr. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet, 15*(10), 689-701. doi:10.1038/nrg3778

Fukao, T., Yamaguchi, S., Scriver, C. R., Dunbar, G., Wakazono, A., Kano, M., . . . Hashimoto, T. (1993). Molecular studies of mitochondrial acetoacetyl-coenzyme A thiolase deficiency in the two original families. *Hum Mutat, 2*(3), 214-220. doi:10.1002/humu.1380020310

Gok, F., Crettol, L. M., Alanay, Y., Hacihamdioglu, B., Kocaoglu, M., Bonafe, L., & Ozen, S. (2010). Clinical and radiographic findings in two brothers affected with a novel mutation in matrix metalloproteinase 2 gene. *Eur J Pediatr, 169*(3), 363-367. doi:10.1007/s00431-009-1028-7

Haas, J. T., Winter, H. S., Lim, E., Kirby, A., Blumenstiel, B., DeFelice, M., . . . Farese, R. V., Jr. (2012). *DGAT1* mutation is linked to a congenital diarrheal disorder. *J Clin Invest, 122*(12), 4680-4684. doi:10.1172/JCI64873

Haire, R. N., Ohta, Y., Strong, S. J., Litman, R. T., Liu, Y., Prchal, J. T., . . . Litman, G. W. (1997). Unusual patterns of exon skipping in Bruton tyrosine kinase are associated with mutations involving the intron 17 3' splice site. *Am J Hum Genet, 60*(4), 798-807.

Hastbacka, J., Kerrebrock, A., Mokkala, K., Clines, G., Lovett, M., Kaitila, I., . . . Lander, E. S. (1999). Identification of the Finnish founder mutation for diastrophic dysplasia (DTD). *Eur J Hum Genet, 7*(6), 664-670. doi:10.1038/sj.ejhg.5200361

Hermans, M. M., van Leenen, D., Kroos, M. A., & Reuser, A. J. (1997). Mutation detection in glycogen storage-disease type II by RT-PCR and automated sequencing. *Biochem Biophys Res Commun, 241*(2), 414-418. doi:10.1006/bbrc.1997.7811

Hladnik, U., Nyhan, W. L., & Bertelli, M. (2008). Variable expression of HPRT deficiency in 5 members of a family with the same mutation. *Arch Neurol, 65*(9), 1240-1243. doi:10.1001/archneur.65.9.1240

Hopp, K., Heyer, C. M., Hommerding, C. J., Henke, S. A., Sundsbak, J. L., Patel, S., . . . Harris, P. C. (2011). *B9D1* is revealed as a novel Meckel syndrome (MKS) gene by targeted exon-enriched next-generation sequencing and deletion analysis. *Hum Mol Genet, 20*(13), 2524-2534. doi:10.1093/hmg/ddr151

Humbert, C., Silbermann, F., Morar, B., Parisot, M., Zarhrate, M., Masson, C., . . . Jeanpierre, C. (2014). Integrin alpha 8 recessive mutations are responsible for bilateral renal agenesis in humans. *Am J Hum Genet, 94*(2), 288-294. doi:10.1016/j.ajhg.2013.12.017

Infante, J. B., Alvelos, M. I., Bastos, M., Carrilho, F., & Lemos, M. C. (2016). Complete androgen insensitivity syndrome caused by a novel splice donor site mutation and activation of a cryptic splice donor site in the androgen receptor gene. *J Steroid Biochem Mol Biol, 155*(Pt A), 63-66. doi:10.1016/j.jsbmb.2015.09.042

Kajihara, S., Hisatomi, A., Mizuta, T., Hara, T., Ozaki, I., Wada, I., & Yamamoto, K. (1998). A splice mutation in the human canalicular multispecific organic anion transporter gene causes Dubin-Johnson syndrome. *Biochem Biophys Res Commun, 253*(2), 454-457. doi:10.1006/bbrc.1998.9780

Ke, S., Anquetil, V., Zamalloa, J. R., Maity, A., Yang, A., Arias, M. A., . . . Chasin, L. A. (2018). Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Res, 28*(1), 11-24. doi:10.1101/gr.219683.116

King, C. R., & Piatigorsky, J. (1983). Alternative RNA splicing of the murine alpha A-crystallin gene: protein-coding information within an intron. *Cell, 32*(3), 707-712.

Kondo, Y., Oubridge, C., van Roon, A. M., & Nagai, K. (2015). Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *Elife, 4*. doi:10.7554/eLife.04986

Krainer, A. R. (2015). Splicing: still so much to learn. *RNA, 21*(4), 500-501. doi:10.1261/rna.050641.115

Kralovicova, J., Hwang, G., Asplund, A. C., Churbanov, A., Smith, C. I., & Vorechovsky, I. (2011). Compensatory signals associated with the activation of human GC 5' splice sites. *Nucleic Acids Res, 39*(16), 7077-7091. doi:10.1093/nar/gkr306

Kramer, A., Keller, W., Appel, B., & Luhrmann, R. (1984). The 5' terminus of the RNA moiety of U1 small nuclear ribonucleoprotein particles is required for the splicing of messenger RNA precursors. *Cell, 38*(1), 299-307.

Kume, K., Masamune, A., Kikuta, K., & Shimosegawa, T. (2006). [-215G>A; IVS3+2T>C] mutation in the *SPINK1* gene causes exon 3 skipping and loss of the trypsin binding site. *Gut, 55*(8), 1214. doi:10.1136/gut.2006.095752

Lagier-Tourenne, C., Tazir, M., Lopez, L. C., Quinzii, C. M., Assoum, M., Drouot, N., . . . Koenig, M. (2008). ADCK3, an ancestral kinase, is mutated in a form of recessive ataxia associated with coenzyme Q10 deficiency. *Am J Hum Genet, 82*(3), 661-672. doi:10.1016/j.ajhg.2007.12.024

Leman, R., Gaildrat, P., Gac, G. L., Ka, C., Fichou, Y., Audrezet, M. P., . . . Houdayer, C. (2018). Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort. *Nucleic Acids Res, 46*(15), 7913-7923. doi:10.1093/nar/gky372

Lin, P., Li, W., Wen, B., Zhao, Y., Fenster, D. S., Wang, Y., . . . Yan, C. (2012). Novel *PNPLA2* gene mutations in Chinese Han patients causing neutral lipid storage disease with myopathy. *J Hum Genet, 57*(10), 679-681. doi:10.1038/jhg.2012.84

Lykke-Andersen, S., & Jensen, T. H. (2015). Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol, 16*(11), 665-677. doi:10.1038/nrm4063

Martin, C. A., Murray, J. E., Carroll, P., Leitch, A., Mackenzie, K. J., Halachev, M., . . . Jackson, A. P. (2016). Mutations in genes encoding condensin complex proteins cause microcephaly through decatenation failure at mitosis. *Genes Dev, 30*(19), 2158-2172. doi:10.1101/gad.286351.116

Matsuura, T., Hoshide, R., Komaki, S., Kiwaki, K., Endo, F., Nakamura, S., . . . Matsuda, I. (1995). Identification of two new aberrant splicings in the ornithine carbamoyltransferase (*OCT*) gene in two patients with early and late onset OCT deficiency. *J Inherit Metab Dis, 18*(3), 273-282.

Moran, C. J., Walters, T. D., Guo, C. H., Kugathasan, S., Klein, C., Turner, D., . . . Muise, A. M. (2013). IL-10R polymorphisms are associated with very-early-onset ulcerative colitis. *Inflamm Bowel Dis, 19*(1), 115-123. doi:10.1002/ibd.22974

Mount, S. M. (1982). A catalogue of splice junction sequences. *Nucleic Acids Res, 10*(2), 459-472.

Mount, S. M., Pettersson, I., Hinterberger, M., Karmas, A., & Steitz, J. A. (1983). The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro. *Cell, 33*(2), 509-518.

Muller, J. S., Piko, H., Schoser, B. G., Schlotter-Weigel, B., Reilich, P., Gurster, S., . . . Walter, M. C. (2006). Novel splice site mutation in the caveolin-3 gene leading to autosomal recessive limb girdle muscular dystrophy. *Neuromuscul Disord, 16*(7), 432-436. doi:10.1016/j.nmd.2006.04.006

Nicholls, A. C., Valler, D., Wallis, S., & Pope, F. M. (2001). Homozygosity for a splice site mutation of the *COL1A2* gene yields a non-functional pro(alpha)2(I) chain and an EDS/OI clinical phenotype. *J Med Genet, 38*(2), 132-136.

Nichols, W. C., Seligsohn, U., Zivelin, A., Terry, V. H., Hertel, C. E., Wheatley, M. A., . . . Ginsburg, D. (1998). Mutations in the ER-Golgi intermediate compartment protein ERGIC-53 cause combined deficiency of coagulation factors V and VIII. *Cell, 93*(1), 61-70.

Ota, Y., Masamune, A., Inui, K., Kume, K., Shimosegawa, T., & Kikuyama, M. (2010). Phenotypic variability of the homozygous IVS3+2T>C mutation in the serine protease inhibitor Kazal type 1 (*SPINK1*) gene in patients with chronic pancreatitis. *Tohoku J Exp Med, 221*(3), 197-201.

Pagani, F., Buratti, E., Stuani, C., Bendix, R., Dork, T., & Baralle, F. E. (2002). A new type of mutation causes a splicing defect in ATM. *Nat Genet, 30*(4), 426-429. doi:10.1038/ng858

Papasaikas, P., & Valcarcel, J. (2016). The spliceosome: the ultimate RNA chaperone and sculptor. *Trends Biochem Sci, 41*(1), 33-45. doi:10.1016/j.tibs.2015.11.003

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 Res, 42*(16), 10564-10578. doi:10.1093/nar/gku744

Pereverzev, A. P., Gurskaya, N. G., Ermakova, G. V., Kudryavtseva, E. I., Markina, N. M., Kotlobay, A. A., . . . Lukyanov, K. A. (2015). Method for quantitative analysis of nonsense-mediated mRNA decay at the single cell level. *Sci Rep, 5*, 7729. doi:10.1038/srep07729

Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res, 29*(9), e45.

Popp, M. W., & Maquat, L. E. (2016). Leveraging rules of nonsense-mediated mRNA decay for genome engineering and personalized medicine. *Cell, 165*(6), 1319-1322. doi:10.1016/j.cell.2016.05.053

Rahner, N., Nuernberg, G., Finis, D., Nuernberg, P., & Royer-Pokora, B. (2016). A novel *C8orf37* splice mutation and genotype-phenotype correlation for cone-rod dystrophy. *Ophthalmic Genet, 37*(3), 294-300. doi:10.3109/13816810.2015.1071408

Ramalho, A. S., Beck, S., Meyer, M., Penque, D., Cutting, G. R., & Amaral, M. D. (2002). Five percent of normal cystic fibrosis transmembrane conductance regulator mRNA ameliorates the severity of pulmonary disease in cystic fibrosis. *Am J Respir Cell Mol Biol, 27*(5), 619-627. doi:10.1165/rcmb.2001-0004OC

Raraigh, K. S., Han, S. T., Davis, E., Evans, T. A., Pellicore, M. J., McCague, A. F., . . . Cutting, G. R. (2018). Functional assays are essential for interpretation of missense variants associated with variable expressivity. *Am J Hum Genet, 102*(6), 1062-1077. doi:10.1016/j.ajhg.2018.04.003

Rios, M., Storry, J. R., Hue-Roye, K., Chung, A., & Reid, M. E. (2002). Two new molecular bases for the Dombrock null phenotype. *Br J Haematol, 117*(3), 765-767.

Roca, X., Akerman, M., Gaus, H., Berdeja, A., Bennett, C. F., & Krainer, A. R. (2012). Widespread recognition of 5' splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides. *Genes Dev, 26*(10), 1098-1109. doi:10.1101/gad.190173.112

Roca, X., Krainer, A. R., & Eperon, I. C. (2013). Pick one, but be quick: 5' splice sites and the problems of too many choices. *Genes Dev, 27*(2), 129-144. doi:10.1101/gad.209759.112

Scalet, D., Maestri, I., Branchini, A., Bernardi, F., Pinotti, M., & Balestra, D. (2019). Disease-causing variants of the conserved +2T of 5' splice sites can be rescued by engineered U1snRNAs. *Hum Mutat, 40*(1), 48-52. doi:10.1002/humu.23680

Seyama, K., Nonoyama, S., Gangsaas, I., Hollenbaugh, D., Pabst, H. F., Aruffo, A., & Ochs, H. D. (1998). Mutations of the CD40 ligand gene and its effect on CD40

ligand expression in patients with X-linked hyper IgM syndrome. *Blood, 92*(7), 2421-2434.

Sharp, P. A., & Burge, C. B. (1997). Classification of introns: U2-type or U12-type. *Cell, 91*(7), 875-879.

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

Shimozawa, N., Nagase, T., Takemoto, Y., Suzuki, Y., Fujiki, Y., Wanders, R. J., & Kondo, N. (2002). A novel aberrant splicing mutation of the *PEX16* gene in two patients with Zellweger syndrome. *Biochem Biophys Res Commun, 292*(1), 109-112.

Smith, S. B., Qu, H. Q., Taleb, N., Kishimoto, N. Y., Scheel, D. W., Lu, Y., . . . German, M. S. (2010). Rfx6 directs islet formation and insulin production in mice and humans. *Nature, 463*(7282), 775-780. doi:10.1038/nature08748

Sobrier, M. L., Maghnie, M., Vie-Luton, M. P., Secco, A., di Iorgi, N., Lorini, R., & Amselem, S. (2006). Novel *HESX1* mutations associated with a life-threatening neonatal phenotype, pituitary aplasia, but normally located posterior pituitary and no optic nerve abnormalities. *J Clin Endocrinol Metab, 91*(11), 4528-4536. doi:10.1210/jc.2006-0426

Soudais, C., de Villartay, J. P., Le Deist, F., Fischer, A., & Lisowska-Grospierre, B. (1993). Independent mutations of the human CD3-epsilon gene resulting in a T cell receptor/CD3 complex immunodeficiency. *Nat Genet, 3*(1), 77-81. doi:10.1038/ng0193-77

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. *Hum Genet, 136*(6), 665-677. doi:10.1007/s00439-017-1779-6

Sumegi, J., Huang, D., Lanyi, A., Davis, J. D., Seemayer, T. A., Maeda, A., . . . Gross, T. G. (2000). Correlation of mutations of the *SH2D1A* gene and epstein-barr virus infection with clinical phenotype and outcome in X-linked lymphoproliferative disease. *Blood, 96*(9), 3118-3125.

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. *Hum Genomics, 13*(1), 8. doi:10.1186/s40246-019-0193-7

Tanugi-Cholley, L. C., Issartel, J. P., Lunardi, J., Freycon, F., Morel, F., & Vignais, P. V. (1995). A mutation located at the 5' splice junction sequence of intron 3 in

the p67phox gene causes the lack of p67phox mRNA in a patient with chronic granulomatous disease. *Blood, 85*(1), 242-249.

Tosetto, E., Ghiggeri, G. M., Emma, F., Barbano, G., Carrea, A., Vezzoli, G., . . . Anglani, F. (2006). Phenotypic and genetic heterogeneity in Dent's disease--the results of an Italian collaborative study. *Nephrol Dial Transplant, 21*(9), 2452-2463. doi:10.1093/ndt/gfl274

Turunen, J. J., Niemela, E. H., Verma, B., & Frilander, M. J. (2013). The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA, 4*(1), 61-76. doi:10.1002/wrna.1141

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., . . . Ponten, F. (2015). Proteomics. Tissue-based map of the human proteome. *Science, 347*(6220), 1260419. doi:10.1126/science.1260419

Venet, T., Masson, E., Talbotec, C., Billiemaz, K., Touraine, R., Gay, C., . . . Ferec, C. (2017). Severe infantile isolated exocrine pancreatic insufficiency caused by the complete functional loss of the *SPINK1* gene. *Hum Mutat, 38*(12), 1660-1665. doi:10.1002/humu.23343

Verma, B., Akinyi, M. V., Norppa, A. J., & Frilander, M. J. (2018). Minor spliceosome and disease. *Semin Cell Dev Biol, 79*, 103-112. doi:10.1016/j.semcdb.2017.09.036

Villa, A., Sironi, M., Macchi, P., Matteucci, C., Notarangelo, L. D., Vezzoni, P., & Mantovani, A. (1996). Monocyte function in a severe combined immunodeficient patient with a donor splice site mutation in the *Jak3* gene. *Blood, 88*(3), 817-823.

Vockley, J., Rogan, P. K., Anderson, B. D., Willard, J., Seelan, R. S., Smith, D. I., & Liu, W. (2000). Exon skipping in *IVD* RNA processing in isovaleric acidemia caused by point mutations in the coding region of the *IVD* gene. *Am J Hum Genet, 66*(2), 356-367. doi:10.1086/302751

Wibawa, T., Takeshima, Y., Mitsuyoshi, I., Wada, H., Surono, A., Nakamura, H., & Matsuo, M. (2000). Complete skipping of exon 66 due to novel mutations of the dystrophin gene was identified in two Japanese families of Duchenne muscular dystrophy with severe mental retardation. *Brain Dev, 22*(2), 107-112.

Wong, M. S., Kinney, J. B., & Krainer, A. R. (2018). Quantitative activity profile and context dependence of all human 5' splice sites. *Mol Cell, 71*(6), 1012-1026 e1013. doi:10.1016/j.molcel.2018.07.033

Wood-Trageser, M. A., Gurbuz, F., Yatsenko, S. A., Jeffries, E. P., Kotan, L. D., Surti, U., . . . Rajkovic, A. (2014). *MCM9* mutations are associated with ovarian failure,

short stature, and chromosomal instability. *Am J Hum Genet, 95*(6), 754-762. doi:10.1016/j.ajhg.2014.11.002

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. doi:10.1136/gutjnl-2017-313948

Wu, H., Boulling, A., Cooper, D. N., Li, Z. S., Liao, Z., Férec, C., & Chen, J. M. (2017). Analysis of the impact of known *SPINK1* missense variants on pre-mRNA splicing and/or mRNA stability in a full-length gene assay. *Genes (Basel), 8*(10). doi:10.3390/genes8100263

Wuhl, E., Kogan, J., Zurowska, A., Matejas, V., Vandevoorde, R. G., Aigner, T., . . . Zenker, M. (2007). Neurodevelopmental deficits in Pierson (microcoria-congenital nephrosis) syndrome. *Am J Med Genet A, 143*(4), 311-319. doi:10.1002/ajmg.a.31564

Yilmaz, D., Cogulu, O., Ozkinay, F., Kavakli, K., & Roos, D. (2005). A novel mutation in the *DIA1* gene in a patient with methemoglobinemia type II. *Am J Med Genet A, 133A*(1), 101-102. doi:10.1002/ajmg.a.30467

Zanni, G., Saillour, Y., Nagara, M., Billuart, P., Castelnau, L., Moraine, C., . . . Chelly, J. (2005). Oligophrenin 1 mutations frequently cause X-linked mental retardation with cerebellar hypoplasia. *Neurology, 65*(9), 1364-1369. doi:10.1212/01.wnl.0000182813.94713.ee

Zhang, X. H., Arias, M. A., Ke, S., & Chasin, L. A. (2009). Splicing of designer exons reveals unexpected complexity in pre-mRNA splicing. *RNA, 15*(3), 367-376. doi:10.1261/rna.1498509

Zhuang, Y., & Weiner, A. M. (1986). A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell, 46*(6), 827-835.

Zou, W. B., Boulling, A., Masamune, A., Issarapu, P., Masson, E., Wu, H., . . . Liao, Z. (2016). No association between *CEL-HYB* hybrid allele and chronic pancreatitis in Asian populations. *Gastroenterology, 150*(7), 1558-1560.e5. doi:10.1053/j.gastro.2016.02.071

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. doi:10.1136/gutjnl-2015-311168

Zou, W. B., Masson, E., Boulling, A., Cooper, D. N., Li, Z. S., Liao, Z., . . . Chen, J. M. (2016). Digging deeper into the intronic sequences of the *SPINK1* gene. *Gut, 65*(6), 1055-1056. doi:10.1136/gutjnl-2016-311428

Zou, W. B., Wu, H., Boulling, A., Cooper, D. N., Li, Z. S., Liao, Z., . . . Férec, C. (2017). *In silico* prioritization and further functional characterization of *SPINK1* intronic variants. *Hum Genomics, 11*(1), 7. doi:10.1186/s40246-017-0103-9

FIGURES

FIGURE 1. Background information, aims and analytical strategy of the study. (**a**) Current knowledge of the canonical 5' splice sites (5'SS) GT and non-canonical 5'SS GC in the human genome in terms of their relative abundance of U2-type introns, their corresponding 9-bp 5'SS signal sequence position weight matrices (PWM) and their associated splicing outcomes. The two PWM illustrative figures were taken from (Leman et al., 2018); an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License. (**b**) Illustration of the first experimental evidence showing that a 5'SS GT>GC substitution may retain the ability to generate wild-type transcripts, albeit at a much reduced level (~10% of normal in (Aebi et al., 1986; Aebi et al., 1987)). (**c**) Aim and analytical strategy of the study.

FIGURE 2. Qualitative analysis of 5'SS GT>GC substitutions. (**a**) Illustration of the cell culture-based full-length gene splicing assay in the context of a 5'SS GT>GC substitution generating some wild-type transcripts. The two horizontal arrows indicate the primers (both located within the vector sequence) used to amplify normally spliced transcripts (and also aberrantly spliced transcripts). F.L., full-length. (**b**) RT-PCR analyses of HEK293T cells transfected with full-length *DBI*, *FATE1* and *PRSS2* gene expression constructs carrying respectively the wild-type and 5'SS GT>GC substitutions as examples. Normal transcripts (confirmed by sequencing) resulting from two of the substitutions are indicated by arrows. IVS, InterVening Sequence (i.e., an intron). See Supp. Figure S1 for all 103 functionally analyzed 5'SS GT>GC substitutions.

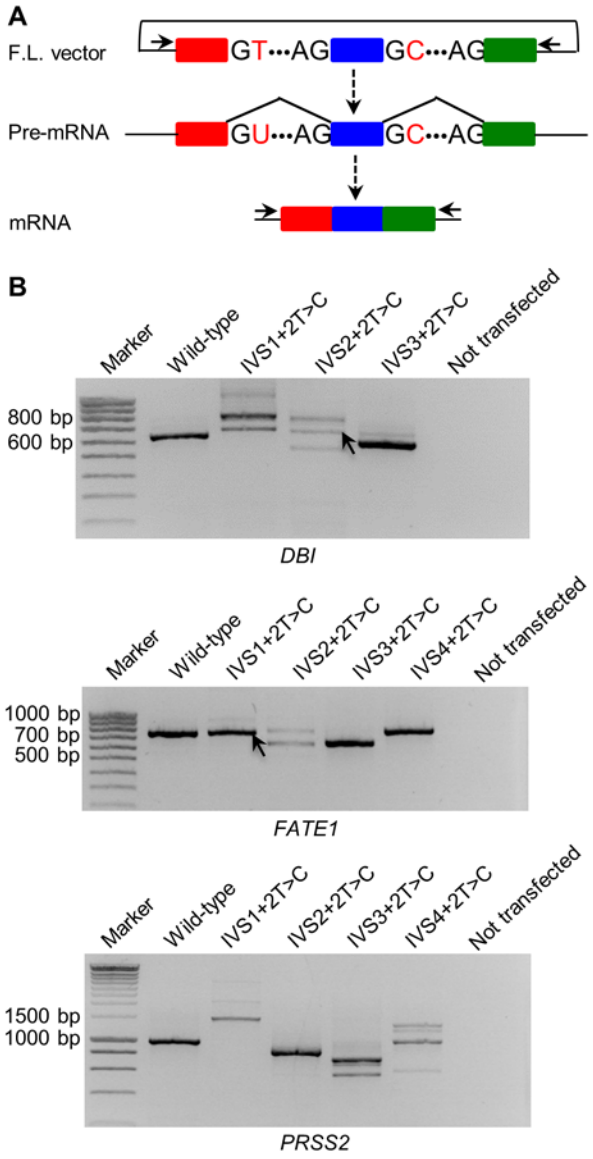


FIGURE 3. Quantitative analysis pertaining to the relative level of 5'SS GT>GC substitution-derived wild-type transcripts. (**a**) Illustration of one key feature of the quantitative RT-PCR analysis: co-transfection of a minigene expression vector with respectively the full-length wild-type target gene expression vector and the full-length variant target gene expression vector. The minigene was constructed in pGL3 (Boulling et al., 2012) whereas the target gene was constructed in either pcDNA3.1/V5-His-TOPO vector or pcDNA3.1(+). The minigene was used as an internal control for quantifying the expression level of wild-type transcripts generated from either the wild-type or variant target full-length gene. The horizontal arrows indicate the relative positions of the primers used for this purpose. Note that for amplifying the target gene sequence, either a primer pair comprising a forward vector-specific primer and a reverse gene-specific primer (as illustrated) or alternatively a primer pair comprising a forward gene-specific primer and a reverse vector-specific primer was used. This assay was performed exclusively for the 10 5'SS GT>GC substitutions that generated only wild-type transcripts. F.L., full-length. (**b**) Quantitative RT-PCR-determined expression level of the mutant allele-derived correctly spliced transcripts relative to that derived from the corresponding wild-type allele (defined as 100%) in the 10 5'SS GT>GC substitutions that generated only wild-type transcripts. Results were expressed as means ± SD from three independent transfection experiments.

FIGURE 4. Pictogram analysis of the 5'SSs under study. Comparison of the pictogram of the 122 5'SSs whose substitutions of GT by GC did not lead to the generation of normal transcripts (upper panel) and that of the 26 5'SSs whose substitutions of GT by GC generated normal transcripts (lower panel). Middle panel shows the 5' end sequence of U1 snRNA that is complementary to the 9-bp U2-type 5'SS signal sequence. 5'SS signal sequences are shown as RNA sequence.

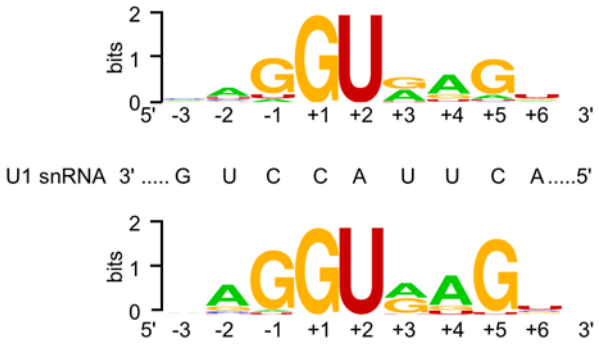


FIGURE 5. Functional characterization of the *HBB* c.315+2T>C variant. RT-PCR analyses of HEK293T cells transfected with full-length *HBB* gene expression constructs carrying respectively the wild-type and two 5'SS GT>GC substitutions. Wild-type transcripts (confirmed by sequencing) resulting from the wild-type and the IVS2+2T>C (i.e., c.315+2T>C) variant are indicated by arrows. The *HBB* c.315+2T>C variant was previously reported to be associated with a mild phenotype (Frischknecht et al., 2009). IVS, InterVening Sequence (i.e., an intron).

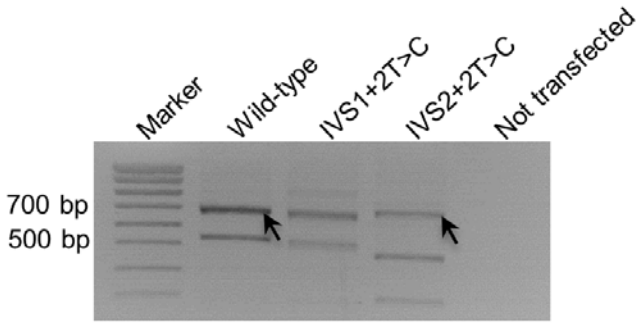


FIGURE 6. Further experiments and analysis validating the experimental model system used in the study. (**a**) Comparison of the splicing outcomes of 20 5'SS GT>GC substitutions in HEK293T and HeLa cells. For each variant, only the informative part of the gel is shown (all data in HEK293T cells were taken from Supp. Figure S1; refer to Supp. Figures S5 and S6 for data in HeLa cells). Normal transcripts (all confirmed by Sanger sequencing) are indicated by arrows. IVS, InterVening Sequence (i.e., an intron). (**b**) Relative mRNA expression levels of the 10 genes in HEK293 and HeLa cells. Data were taken from https://www.proteinatlas.org/. TPM, transcripts per million.

TABLE 1. The 45 informative disease-causing 5'SS GT>GC variants and their predicted splicing effects*

Disease	Gene	Variant	Reference	Zygosity	Generation of Wild-Type Transcripts[a]	SpliceSiteFinder-like (0-100)[b]	MaxEntScan (0-12)[b]	NNSPLICE (0-1)[b]	GeneSplicer (0-15)[b]
Dubin-Johnson syndrome	*ABCC2*	c.1967+2T>C	(Kajihara et al., 1998)	Homozygote	No	74.2→NS	7.4→NS	0.6→NS	1.8→NS
Acetoacetyl-CoA-thiolase deficiency	*ACAT1*	c.1163+2T>C	(Fukao et al., 1993)	Homozygote	No	76.8→74.1	6.9→NS	1.0→NS	NS
Ubiquinone deficiency with cerebellar ataxia	*COQ8A*	c.1398+2T>C	(Lagier-Tourenne et al., 2008)	Homozygote	No	76.9→75.2	6.4→NS	0.8→NS	4.6→NS
Analbuminaemia	*ALB*	c.1428+2T>C	(Dolcini et al., 2007)	Homozygote	No	78.9→70.8	5.6→NS	1.0→NS	NS
Androgen insensitivity syndrome	*AR*	c.2173+2T>C	(Infante, Alvelos, Bastos, Carrilho, & Lemos, 2016)	Hemizygote	No	86.5→81.1	9.7→NS	1.0→NS	4.4→NS
Dombrock null allele	*ART4*	c.144+2T>C	(Rios, Storry, Hue-Roye, Chung, & Reid, 2002)	Homozygote	No	81.8→79.9	7.8→NS	0.9→NS	4.1→NS
Menkes syndrome	*ATP7A*	c.1946+2T>C	(Das et al., 1994)	Hemizygote	No	78.6→75.5	9.5→NS	1.0→NS	3.8→NS
Meckel syndrome	*B9D1*	c.341+2T>C	(Hopp et al., 2011)	Hemizygote	No	81.0→78.7	9.4→NS	1.0→NS	8.2→NS
Agammaglobulinaemia	*BTK*	c.588+2T>C	(Haire et al., 1997)	Hemizygote	No	71.9→NS	7.5→NS	0.8→NS	4.0→NS
Cone-rod dystrophy	*C8orf37*	c.155+2T>C	(Rahner, Nuernberg, Finis, Nuernberg, &	Homozygote	No	72.2→NS	1.6→NS	0.6→NS	2.1→NS

			Royer-Pokora, 2016)						
Autosomal recessive limb girdle muscular dystrophy	*CAV3*	c.114+2T>C	(Muller et al., 2006)	Homozygote	**Yes**	83.8→81.8	10.1→NS	1.0→NS	10.1→NS
Immunodeficiency	*CD3E*	c.520+2T>C	(Soudais et al., 1993)	Compound heterozygote	**Yes** (1-5%)	83.0→78.1	8.1→NS	1.0→NS	2.9→NS
Hyper-IgM syndrome	*CD40LG*	c.346+2T>C	(Seyama et al., 1998)	Hemizygote	**Yes** (15%)	89.6→90.0	10.3→NS	1.0→NS	1.6→NS
Dent disease	*CLCN5*	c.205+2T>C	(Tosetto et al., 2006)	Hemizygote	No	84.8→82.1	10.0→NS	1.0→NS	NS
Ehlers-Danlos syndrome/Osteogenesis imperfecta	*COL1A2*	c.3105+2T>C	(Nicholls, Valler, Wallis, & Pope, 2001)	Homozygote	No	75.4→72.8	8.6→NS	0.9→NS	1.1→NS
Congenital diarrhoeal disorder	*DGAT1*	c.751+2T>C	(Haas et al., 2012)	Homozygote	No	78.6→NS	7.9→NS	1.0→NS	11.9→NS
Becker muscular dystrophy	*DMD*	c.8027+2T>C	(Bartolo et al., 1996)	Hemizygote	**Yes** (10%)	84.2→81.5	9.1→NS	1.0→NS	1.7→NS
Duchenne muscular dystrophy	*DMD*	c.9649+2T>C	(Wibawa et al., 2000)	Hemizygote	No	84.3→84.4	9.1→NS	1.0→NS	NS
Mirror movements (congenital)	*DNAL4*	c.153+2T>C	(Ahmed et al., 2014)	Homozygote	No	NS	7.4→NS	0.8→NS	9.7→NS
Glycogen storage disease 2	*GAA*	c.2331+2T>C	(Hermans, van Leenen, Kroos, & Reuser, 1997)	Homozygote	No	86.4→76.7	11.5→NS	1.0→NS	13.6→NS
Pituitary aplasia	*HESX1*	c.357+2T>C	(Sobrier et al., 2006)	Homozygote	No	80.2→70.6	6.7→NS	0.8→NS	NS

Ulcerative colitis	*IL10RA*	c.688+2T>C	(Moran et al., 2013)	Homozygote	No	73.8→NS	7.0→NS	0.8→NS	2.7→NS
Renal hypodysplasia	*ITGA8*	c.2982+2T>C	(Humbert et al., 2014)	Homozygote	No	71.9→NS	5.8→NS	0.9→NS	NS
Isovaleric acidaemia	*IVD*	c.465+2T>C	(Vockley et al., 2000)	Homozygote	No	90.3→80.5	9.2→NS	1.0→NS	4.5→NS
Immunodeficiency (severe combined)	*JAK3*	c.2350+2T>C	(Villa et al., 1996)	Homozygote	No	NS	5.8→NS	NS	6.8→NS
Muscular dystrophy (merosin deficient)	*LAMA2*	c.3924+2T>C	(Allamand et al., 1997)	Homozygote	No	79.8→77.0	8.3→NS	0.8→NS	3.4→NS
Factor V and factor VIII deficiency (combined)	*LMAN1*	c.1149+2T>C	(Nichols et al., 1998)	Homozygote	No	79.8→70.6	8.1→NS	NS	NS
Primary amenorrhea & short stature	*MCM9*	c.1732+2T>C	(Wood-Trageser et al., 2014)	Homozygote	No	NS	1.7→NS	NS	NS
Torg-Winchester syndrome	*MMP2*	c.658+2T>C	(Gok et al., 2010)	Homozygote	No	90.0→80.1	8.7→NS	1.0→NS	10.9→NS
Microcephaly	*NCAPD2*	c.4120+2T>C	(Martin et al., 2016)	Homozygote	No	84.2→81.5	9.1→NS	0.9→NS	7.1→NS
Chronic granulomatous disease	*NCF2*	c.257+2T>C	(Tanugi-Cholley et al., 1995)	Homozygote	No	84.8→84.7	9.8→NS	1.0→NS	5.3→NS
Mental retardation syndrome (X-linked)	*OPHN1*	c.154+2T>C	(Zanni et al., 2005)	Hemizygote	No	84.8→84.7	9.8→NS	1.0→NS	7.9→NS
Ornithine carbamoyltransferase	*OTC*	c.540+2T>C	(Matsuura et al., 1995)	Hemizygote	No	80.0→78.2	8.1→NS	0.6→NS	NS

deficiency									
Propionic acidaemia	*PCCB*	c.183+2T>C	(Desviat et al., 2006)	Homozygote	No	74.5→NS	8.5→NS	0.9→NS	9.7→NS
Propionic acidaemia	*PCCB*	c.1498+2T>C	(Desviat et al., 2006)	Homozygote	No	81.8→79.9	7.8→NS	0.7→NS	5.0→NS
Zellweger syndrome	*PEX16*	c.952+2T>C	(Shimozawa et al., 2002)	Homozygote	No	82.1→79.0	7.5→NS	1.0→NS	5.2→NS
Spastic tetraparesis/paraparesis	*PLP1*	c.622+2T>C	(Biancheri et al., 2014)	Hemizygote	No	86.8→77.2	10.1→NS	1.0→NS	6.2→NS
Pelizaeus-Merzbacher disease	*PLP1*	c.696+2T>C	(Aoyagi et al., 1999)	Hemizygote	**Yes**	92.6→85.9	10.0→NS	1.0→NS	6.5→NS
Neutral lipid storage disease with myopathy	*PNPLA2*	c.757+2T>C	(Lin et al., 2012)	Compound heterozygote	No	NS	8.7→NS	NS	8.3→NS
Brittle cornea syndrome	*PRDM5*	c.93+2T>C	(Aldahmesh, Mohamed, & Alkuraya, 2012)	Homozygote	No	85.3→78.5	8.2→NS	0.9→NS	10.6→NS
Diabetes (neonatal, with intestinal atresia)	*RFX6*	c.380+2T>C	(Smith et al., 2010)	Homozygote	No	78.7→NS	5.5→NS	0.6→NS	2.9→NS
Oro-facio-digital syndrome type IX	*SCLT1*	c.290+2T>C	(Adly, Alhashem, Ammari, & Alkuraya, 2014)	Homozygote	No	87.5→87.1	8.9→NS	1.0→NS	NS
Lymphoproliferative syndrome (X-linked)	*SH2D1A*	c.137+2T>C	(Sumegi et al., 2000)	Hemizygote	No	71.1→NS	7.4→NS	0.4→NS	5.9→NS
Diastrophic dysplasia	*SLC26A2*	c.-26+2T>C	(Hastbacka et al., 1999)	Homozygote	**Yes** (5%)	87.3→77.7	7.7→NS	1.0→NS	11.5→NS

Chronic pancreatitis	*SPINK1*	c.194+2T>C	(Kume et al., 2006)	Homozygote	**Yes**	82.6→72.3	11.1→NS	1.0→NS	4.0→NS

*See Supp. Table S1 for more information.

[a]Relative expression level is indicated in parentheses wherever applicable.

[b]Prediction was performed under default conditions. NS, no score.

TABLE 2. Nature of the sequenced 12 aberrantly spliced transcripts*

Gene	Reference mRNA accession number	Variant[a]	Aberrant Transcripts
DBI	NM_001079862.2	IVS1+2T>C	1. Activation of a cryptic 5'SS GT located 152 bp downstream of the normal one, resulting in the retention of the first 154 bp of the intron 1 sequence. 2. Activation of a cryptic 5'SS GT located 28 bp downstream of the normal one, resulting in the retention of the first 30 bp of the intron 1 sequence.
		IVS3+2T>C	Exon 3 skipping
FABP7	NM_001446.4	IVS1+2T>C	Activation of a cryptic 5'SS GT located 2 bp downstream of the normal one, resulting in the retention of the first 4 bp of the intron 1 sequence.
		IVS2+2T>C	Activation of a cryptic 5'SS GT located 3 bp upstream of the normal one, resulting in the deletion of the last 5 bp of exon 2.
HESX1	NM_003865.2	IVS2+2T>C	Exon 2 skipping
		IVS3+2T>	Exon 3 skipping

		C	
IL10	NM_000572.3	IVS1+2T>C	Activation of a cryptic 5'SS GT located 2 bp downstream of the normal one, resulting in the retention of the first 4 bp of the intron 1 sequence.
		IVS4+2T>C	Activation of a cryptic 5'SS GT located 19 bp upstream of the normal one, resulting in the deletion of the last 21 bp of exon 4.
PRSS2	NM_002770.3	IVS4+2T>C	Activation of a cryptic 5'SS GC located 15 bp upstream of the normal one, resulting in the deletion of the last 17 bp of exon 4.
SPINK1	NM_003122.3	IVS1+2T>C	Activation of a cryptic 5'SS GT located 138 bp downstream of the normal one, resulting in the retention of the first 140 bp of the intron 1 sequence.
UQCRB	NM_006294.4	IVS1+2T>C	Activation of a cryptic 5'SS GT located 10 bp upstream of the normal one, resulting in the deletion of the last 12 bp of exon 1.

*See Supplementary Figure S1 for the corresponding RT-PCR products.

[a]In accordance with the traditional IVS (InterVening Sequence; i.e., an intron) nomenclature.

TABLE 3. Comparison of 5'SS GT>GC, >GA and >GG substitutions with respect to the generation (or not) of wild-type transcripts

Gene	Reference mRNA accession number	Site[a]	GT>GC[b]	GT>GA[c]	GT>GG[c]
DBI	NM_001079862.2	IVS1+2T	No	No	No
		IVS2+2T	Yes	No	No
		IVS3+2T	No	No	No
FATE1	NM_033085.2	IVS2+2T	No	–[d]	No
FOLR3	NM_000804.3	IVS1+2T	–	No	No
		IVS2+2T	No	No	No
		IVS3+2T	No	No	–
		IVS4+2T	Yes	No	No
HESX1	NM_003865.2	IVS1+2T	Yes	No	No
IL10	NM_000572.3	IVS1+2T	No	–	No
		IVS2+2T	No	–	No
		IVS3+2T	Yes	–	No
		IVS4+2T	No	No	No
RPL11	NM_000975.5	IVS1+2T	No	No	–
		IVS2+2T	Yes	–	No

		IVS4+2T	No	No	No
		IVS5+2T	No	No	No
SPINK1	NM_003122.3	IVS1+2T	No	No	No
		IVS2+2T	No	No	No
		IVS3+2T	Yes	No	No

[a]In accordance with the traditional IVS (InterVening Sequence; i.e., an intron) nomenclature.

[b]See Supp. Figure S1 for original data.

[c]See Supp. Figure S4 for original data.

[d]Failure in mutagenesis.