

1 **First estimation of the scale of canonical 5' splice site GT>GC**
2 **mutations generating wild-type transcripts and their medical genetic**
3 **implications**

4
5 Jin-Huan Lin^{1,2,3†}, Xin-Ying Tang^{2,3†}, Arnaud Boulling¹, Wen-Bin Zou^{2,3}, Emmanuelle Masson¹, Yann
6 Fichou^{1,4}, Loann Raud¹, Marlène Le Tertre¹, Shun-Jiang Deng^{2,3}, Isabelle Berlivet¹, Chandran Ka^{1,4},
7 Matthew Mort⁵, Matthew Hayden⁵, Gerald Le Gac^{1,4}, David N. Cooper⁵, Zhao-Shen Li^{2,3}, Claude Férec¹,
8 Zhuan Liao^{2,3*} and Jian-Min Chen^{1*}

9
10 ¹ EFS, Univ Brest, Inserm, UMR 1078, GGB, F-29200 Brest, France
11 ² Department of Gastroenterology, Changhai Hospital, the Second Military Medical University,
12 Shanghai, China
13 ³ Shanghai Institute of Pancreatic Diseases, Shanghai, China
14 ⁴ CHU Brest, Service de Génétique, Brest, France
15 ⁵ Laboratory of Excellence GR-Ex, Paris, France
16 ⁶ Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom

17
18 * To whom correspondence should be addressed. Tel: 33(2) 98 01 81 74; Fax: 33(2) 98 01 64 74;
19 Email: jian-min.chen@univ-brest.fr.
20 Correspondence may also be addressed to Zhuan Liao. Email: liao zhuan@smmu.edu.cn.

21
22 †The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint
23 First Authors.

24 **ABSTRACT**

25 It has long been known that canonical 5' splice site (5'SS) GT>GC mutations may be compatible with
26 normal splicing. However, to date, the true scale of canonical 5'SS GT>GC mutations generating wild-
27 type transcripts, both in the context of the frequency of such mutations and the level of wild-type
28 transcripts generated from the mutation alleles, remain unknown. Herein, combining data derived from a
29 meta-analysis of 45 informative disease-causing 5'SS GT>GC mutations (from 42 genes) and a cell
30 culture-based full-length gene splicing assay of 103 5'SS GT>GC mutations (from 30 genes), we
31 estimate that ~15-18% of the canonical GT 5'SSs are capable of generating between 1 and 84% normal
32 transcripts as a consequence of the substitution of GT by GC. We further demonstrate that the
33 canonical 5'SSs whose substitutions of GT by GC generated normal transcripts show stronger
34 complementarity to the 5' end of U1 snRNA than those sites whose substitutions of GT by GC did not
35 lead to the generation of normal transcripts. We also observed a correlation between the generation of
36 wild-type transcripts and a milder than expected clinical phenotype but found that none of the available
37 splicing prediction tools were able to accurately predict the functional impact of 5'SS GT>GC mutations.
38 Our findings imply that 5'SS GT>GC mutations may not invariably cause human disease but should also
39 help to improve our understanding of the evolutionary processes that accompanied GT>GC subtype
40 switching of U2-type introns in mammals.

41
42 **Keywords:** Canonical 5' splice site, Full-length gene splicing assay, Genotype and phenotype
43 relationship, Human Gene Mutation Database, Human inherited disease, Meta-analysis, Non-canonical
44 splice donor site, U2-type intron

46 **INTRODUCTION**

47 The vast majority of eukaryotic introns are spliced by the U2 spliceosome (the only alternative U12
48 spliceosome is responsible for <0.5% of all introns [1-3]), which interacts with RNA sequences
49 specifying the 5' and 3' splice sites [4, 5]. In vertebrates, the 9-bp consensus sequence for the U2-type
50 5' splice site (5'SS) has traditionally been described as 5'-MAG/GURAGU-3' (where M denotes C or A,
51 R denotes A or G and / denotes the exon-intron boundary; the corresponding nucleotide positions are
52 denoted -3_-1/+1_+6) although in reality this consensus sequence does not reflect the true extent of
53 sequence variability [6-11]. Base-pairing of this 9-bp sequence with 3'-GUCCAUUCA-5' at the 5' end of
54 U1 snRNA (Figure 1A) is critical for splicing to occur [10, 12-15]. Although the GT dinucleotide in the
55 first two intronic positions (in the context of DNA sequence) is the most highly conserved portion of the
56 U2-type 5'SS, it was reported, as early as 1983, that GC occasionally occurs in place of GT [16-18].
57 Subsequent genome-wide analyses have established that this non-canonical 5'SS GC is present as
58 wild-type in ~1% of human U2-type introns [2, 7, 8, 19, 20]. Importantly, the remaining nucleotides in
59 these evolutionarily fixed non-canonical GC 5'SSs exhibit a stronger complementarity to the 3'-
60 GUCCAUUCA-5' sequence at the 5' end of U1 snRNA than those in the canonical GT 5'SSs (Figure
61 1A), thereby in all likelihood compensating for the decreased complementarity between the 5'SS and
62 the 5' end of U1 snRNA due to the U to C substitution [7, 8]. Comparative genome analyses have also
63 revealed frequent switching of U2-type introns from the canonical 5'SS GT subtype to the non-canonical

64 5'SS GC subtype during mammalian evolution [8, 21]. Finally, GC has recently been ranked first among
65 the six non-canonical 5'SSs identified by genome-wide RNA-seq analysis and splicing reporter assays
66 [22].

67 The finding that GC occasionally occurs instead of GT within the canonical 5'SS in some vertebrate
68 genes implies that substitution of the canonical 5'SS GT by GC (termed a 5'SS GT>GC mutation) may
69 allow normal splicing to occur. The first direct experimental evidence supporting such a postulate came
70 in the late 1980s; analyses of both the splicing products of *in vitro* transcribed rabbit beta globin (*Hbb*)
71 RNA in a HeLa cell nuclear extract and the splicing products of the *Hbb* gene transiently expressed in
72 HeLa cells demonstrated that, of all the possible single nucleotide substitutions of the canonical 5'SS
73 GT of the second and last intron of *Hbb*, only the substitution of T by C was compatible with normal
74 splicing, albeit at a much reduced rate (approximately 10% of normal; see also [Figure 1B](#)) [23, 24].
75 Further supporting evidence came from the study of disease-causing 5'SS GT>GC mutations, some of
76 which were reported to generate wild-type transcripts (see below). Additionally, the activation of cryptic
77 non-canonical 5'SS GC has also been reported as a consequence of some disease-causing mutations
78 [25, 26].

79 The above notwithstanding, to date, the scale of canonical 5'SS GT>GC mutations generating wild-
80 type transcripts, both in the context of the frequency of such mutations and the level of wild-type
81 transcripts generated by such mutations, remain unknown owing to the intrinsic complexity of splicing
82 [11, 27-29] and the lack of suitable model systems for study. This issue has important implications for
83 medical genetics since mutant genotypes retaining even a small fraction of their normal function may
84 differ significantly from null genotypes in terms of their associated clinical phenotypes (e.g., only 5% of
85 normal *CFTR* gene expression is enough to prevent the lung manifestations of cystic fibrosis [30, 31]).
86 Herein, we attempted to address this issue by employing two distinct but complementary approaches in
87 concert.

88

89 **RESULTS AND DISCUSSION**

90 **Estimation by Meta-Analysis of Disease-Causing 5'SS GT>GC Mutations**

91 First, we performed a meta-analysis of disease-causing 5'SS GT>GC mutations logged in the
92 Professional version of Human Gene Mutation Database (HGMD; as of June 2017) [32], with a view to
93 generating an "*in vivo*" dataset to estimate the scale of 5'SS GT>GC mutations generating wild-type
94 transcripts. Employing a stringent approach ([Figure 1C](#)), we identified 45 disease-causing 5'SS GT>GC
95 mutations (from 42 genes) that were informative with respect to the presence or absence of wild-type
96 transcripts derived from the mutant allele ([Table 1](#); see [Supplementary Table S1](#) for more information
97 including affected intron, reference mRNA accession number, chromosomal location, hg38 coordinate,
98 and patient-derived tissue or cells used for RT-PCR analysis, etc.). It should be noted that the
99 assignments of "presence" or "absence" of mutant allele-derived wild-type transcripts depended upon
100 the agarose gel evaluation of RT-PCR products as described in the corresponding original publications.
101 Thus, we conservatively annotated an isolated case (i.e., the *PCCB* c.183+2T>C mutation), which was
102 not found to generate wild-type transcripts on agarose gel evaluation of RT-PCR products but was

103 found to generate <0.1% normal wild-type transcripts by means of quantitative RT-PCR [33], as
104 generating no wild-type transcripts.

105 The 45 informative 5'SS GT>GC mutations comprised 30 homozygotes, 13 hemizygotes and 2
106 compound heterozygotes (Table 1). Whilst the presence or absence of wild-type transcripts derived
107 from the mutant allele was straightforward for all homozygous or hemizygous mutations included, the
108 two compound heterozygotes required special treatment. In the case of the *CD3E* c.520+2T>C
109 mutation, the pathogenic *CD3E* mutation in *trans* was a nonsense mutation in exon 6. Sequencing of
110 the patient-derived, normal-sized RT-PCR products failed to demonstrate the exon 6 mutation,
111 suggesting that the wild-type transcripts were derived from the c.520+2T>C allele [34]. In the case of
112 the *PNPLA2* c.757+2T>C mutation, the second *PNPLA2* mutation in *trans* was a missense mutation,
113 c.749A>C (p.Gln250Pro). RT-PCR analysis detected only the c.749A>C mutant mRNA in skeletal
114 muscle of the patient, indicating the absence of detectable wild-type transcript emanating from the
115 c.757+2T>C allele [35].

116 15.6% (n=7) of the 45 informative mutations were found to have been capable of generating some
117 correctly spliced transcripts (Table 1). Information on the expression level of the mutant allele-derived
118 wild-type transcripts relative to that of the wild-type transcripts from a normal control (by definition,
119 100%) was available from four of the seven original publications (i.e., *CD3E* c.520+2T>C [34], *CD40LG*
120 c.346+2T>C [36], *DMD* c.8027+2T>C [37] and *SLC26A2* c.-26+2T>C [38]), which ranged from 1-15% of
121 normal in individual cases (Table 1). All three of the remaining mutations generated both wild-type and
122 aberrant transcripts (i.e., *CAV3* c.114+2T>C [39], *PLP1* c.696+2T>C [40] and *SPINK1* c.194+2T>C
123 [41]); based upon visual inspection of the original gel photographs, the relative expression level of the
124 mutant allele-derived wild-type transcripts in these three cases could also be estimated to fall within the
125 1-15% range.

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

129

130 **Estimation from the Cell Culture-Based Full-Length Gene Splicing Assay of 5'SS GT>GC**

131 **Mutations**

132 To corroborate the findings derived from the above “*in vivo*” dataset, we sought to generate an “*in vitro*”
133 dataset of 5'SS GT>GC mutations. In this regard, we have previously used a cell culture-based full-
134 length gene splicing assay to analyze a series of *SPINK1* intronic variants including a 5'SS GT>GC
135 mutation, c.194+2T>C [42, 43]. Specifically, the full-length 7-kb *SPINK1* genomic sequence (including
136 all four exons plus all three introns of the gene) was cloned into the pcDNA3.1/V5-His-TOPO vector
137 [44]. The full-length gene splicing assay preserves better the natural genomic context of the studied
138 mutations as compared to the commonly used minigene splicing assay, a point of importance given the
139 highly context-dependent and combinatorial nature of alternative splicing regulation [45]. Moreover, the
140 full-length gene splicing assay can be readily used to evaluate all intronic variants including those
141 located near the first or last exons of the gene. Despite these advantages, the full-length gene assay
142 cannot easily be applied to large-size genes owing to the technical difficulties inherent in amplifying and

143 cloning long DNA fragments into the expression vector. Finally, it is pertinent to point out that, to
144 functionally evaluate the impact on splicing of any given gene mutation in a transient expression
145 system, it is highly desirable to use of cells of pathophysiological relevance owing to the tissue
146 specificity of the splicing process in some instances [11, 27-29]. However, this may not always be
147 possible in practice, particularly if variants in multiple genes are to be analyzed in large-scale studies.
148 For example, a recent study that measured 5'SS activity in the context of three minigenes was
149 performed in transfected HeLa cells [11]. In the present study, we used HEK293T cells for transfection
150 as previously described [42, 46].

151 Bearing in mind the aforementioned advantages and disadvantages, we employed a cell culture-
152 based full-length gene splicing assay (Figure 1C). In brief, for various technical and practical reasons,
153 we firstly selected genes whose genomic sizes did not exceed 8 kb (from the translation initiation codon
154 to the translation termination codon) and whose exons numbered ≥ 3 in order to construct full-length
155 gene expression vectors; we then screened genes, which had yielded a single or quasi-single band of
156 expected size by means of RT-PCR analysis of transfected cells, for subsequent mutagenesis of all
157 available 5'SS GT dinucleotides in the construct (for details on the selected and screened genes, see
158 Supplementary Table S2). In the end, we succeeded in functionally analyzing 103 GT>GC mutations
159 from 30 different genes (Supplementary Table S3). 18.4% (n=19) of these artificially introduced 5'SS
160 GT>GC mutations generated wild-type transcripts (all confirmed by Sanger sequencing; Figure 2 and
161 Supplementary Figure S1), a finding that concurs with the 15.6% value obtained from the meta-analysis
162 of disease-causing 5'SS GT>GC mutations.

163 Only wild-type transcripts were observed for 10 of the aforementioned 19 5'SS GT>GC mutations
164 (e.g., *FATE1* IVS1+2T>C in Figure 2B). In other words, no aberrantly spliced transcripts were observed
165 in these 10 cases. It is possible that aberrantly spliced transcripts may be rendered invisible by RNA
166 degradation mechanisms such as nonsense-mediated mRNA decay (NMD) [47, 48]. One way to test
167 such a possibility is to add an NMD inhibitor such as cycloheximide [49] to the cell culture medium,
168 although this is beyond the scope of the present study. We quantified the relative level of correctly
169 spliced transcripts for these 10 5'SS GT>GC mutations by means of our previously described
170 quantitative RT-PCR method [46, 50, 51]. Here it is pertinent to mention that a co-transfected minigene
171 construct was used as an internal control in this analysis (Figure 3A), a prerequisite to obtain accurate
172 results. As shown in Figure 3B, the relative level of correctly spliced transcripts emanating from these
173 10 mutations is remarkably similar to that observed for the disease-causing 5'SS GT>GC mutations in
174 terms of the lowest extreme (2-5% vs. 1-5%); however, the functionally obtained highest level of
175 correctly spliced transcripts (84%) is much higher than the corresponding 15% value observed for the
176 disease-causing 5'SS GT>GC mutations (Table 1). We were initially puzzled by this disparity, but this
177 could be accounted for by two considerations. On the one hand, the currently analyzed disease-causing
178 mutations were likely to be biased toward those that generated either no wild-type transcripts or only a
179 low level. On the other hand, given (i) that 5'SS GC may occur as wild-type in the human genome, (ii)
180 the highly degenerate nature of the 5'SS splice signal sequences and (iii) the complex regulation of the
181 splicing process *in vivo*, it is entirely possible that a 5'SS GT>GC mutation may behave similarly to its

182 original wild-type sequence. This notwithstanding, no single GC mutation was noted to have an identical
183 or higher normal splicing activity than its 5'SS GT counterpart (Figure 3B).

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

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

200

201 **Integrated Estimation from the Two Distinct but Complementary Datasets**

202 We obtained remarkably similar findings in terms of both the frequency of 5'SS GT>GC mutations
203 generating wild-type transcripts and the lowest relative level of mutant allele-derived wild-type
204 transcripts from two quite distinct yet complementary datasets. The consistently lowest relative level of
205 mutant allele-derived wild-type transcripts across the two datasets suggested that the gel-based
206 analytical method is sensitive enough to detect as little as ~1% of normally spliced transcripts. The
207 apparent disparity in terms of the highest relative level of mutant-derived wild-type transcripts between
208 the two datasets can however be accounted for largely by the selection bias inherent to disease-causing
209 mutations. Therefore, we estimate that some 15-18% of 5'SS GT>GC mutations generate between 1
210 and 84% of wild-type transcripts.

211

212 **Exploration of the Mechanisms Underlying the Generation or Not of Wild-Type Transcripts by** 213 **5'SS GT>GC Mutations**

214 As mentioned above, canonical GT and non-canonical GC 5'SSs in the human genome exhibit different
215 patterns of sequence conservation, the latter showing stronger complementarity to the 3'-
216 GUCCAUUCA-5' sequence at the 5' end of U1 snRNA (Figure 1A). We surmised that the canonical
217 5'SSs whose substitutions of GT by GC generated normal transcripts (termed group 1) should also
218 exhibit stronger complementarity to the aforementioned 9-bp sequence than those sites whose
219 substitutions of GT by GC did not lead to the generation of normal transcripts (termed group 2). We
220 therefore extracted the 9-bp sequence tracts surrounding the corresponding groups of the 45 disease-
221 causing 5'SS GT>GC mutations (Supplementary Tables S1) and those of the 103 functionally analyzed

222 5'SS GT>GC mutations ([Supplementary S3](#)). Comparison of the resulting pictograms confirmed our
223 postulate in both contexts, the respective pictograms for the combined group 1 mutations (n=26) and
224 combined group 2 mutations (n=122) being provided in [Figure 4](#). It should be emphasized that the
225 surrounding 9-bp sequence tract is an important (but certainly not the only) factor in determining
226 whether or not a given 5'SS GT>GC mutation will generate some wild-type transcripts. A simple
227 example may be used to illustrate this point: the *DMD* c.8027+2T>C mutation (which generates 10% of
228 wild-type transcripts) contrasts with the *NCAPD2* c.4120+2T>C mutation (which generates no wild-type
229 transcripts) despite occurring in an identical 9-bp sequence tract, AAGGTATGA (see [Supplementary](#)
230 [Table S1](#)).

231 We also explored whether the creation or disruption of splice enhancer/silencer motifs by the 5'SS
232 GT>GC mutations could be associated with the generation or not of some wild-type transcripts. To this
233 end, we employed ESEfinder and RESUE-ESE provided by the Alamut suite under default conditions.
234 We were unable to draw any meaningful conclusions, primarily due to the short and degenerate nature
235 of the splicing enhancer/silencer binding motifs.

236 237 **Correlation Between the Retention of Wild-Type Transcripts and a Milder Than Expected** 238 **Phenotype**

239 Given that even the retention of a small fraction of normal gene function may significantly impact the
240 clinical phenotype, we reviewed the original publications describing the seven disease-causing 5'SS
241 GT>GC mutations that generated at least some wild-type transcript ([Table 1](#)) with respect to the
242 accompanying genotypic and phenotypic descriptions. In six cases, the mutations were specifically
243 described as being associated with mild clinical phenotypes as compared to their classical disease
244 counterparts (see [Supplementary Table S1](#)). In the remaining case (*SPINK1* c.194+2T>C), the original
245 publication [41] was not informative in this regard; however, it is known that homozygosity for this
246 mutation causes chronic pancreatitis with variable expressivity [52] whereas null *SPINK1* genotypes
247 cause severe infantile isolated exocrine pancreatic insufficiency [53].

248 The above correlation between the retention of some wild-type transcripts and a milder than
249 expected phenotype prompted us to postulate that 5'SS GT>GC mutations previously reported to confer
250 a milder than expected phenotype but having no supportive patient-derived transcript expression data,
251 may be collectively associated with a non-canonical 5'SS GC signal. We collated a total of six such
252 mutations (i.e., *CYB5R3* c.463+2T>C [54], *HBB* c.315+2T>C [55], *HPRT* c.485+2T>C [56], *LAMB2*
253 c.3327+2T>C [57], *LMNA* c.1968+2T>C [58] and *MTTP* c.61+2T>C [59]; [Supplementary Table S4](#)). In
254 this regard, two points require clarification. First, in two cases, patient-derived transcript expression data
255 were available [56, 57]; these cases were however addressed here because the corresponding
256 expression data were insufficiently informative for them to be listed in [Supplementary Table S1](#) (for
257 explanations, see [Supplementary Table S4](#)). Second, five of these six mutations (all germline) were
258 derived from the HGMD dataset whereas the remaining one (*LMNA* c.1968+2T>C) [58], a somatic
259 mutation, was obtained from a literature search; this somatic mutation was included owing to its clear
260 phenotypic impact. Pictogram analysis of the six corresponding 9-bp canonical 5'SSs did reveal a non-
261 canonical 5'SS GC signal ([Supplementary Figure S3](#)). Notably, one of the mutations affected the splice

262 donor splice site of *HBB* intron 2 (i.e., *HBB* c.315+2T>C) [55], site of the previously analyzed
263 orthologous mutation in the rabbit *Hbb* gene [23, 24]. We were able to study the effect of the *HBB* intron
264 2 GT>GC mutation on splicing by means of the full-length gene assay and found that it had indeed
265 retained the ability to generate normal *HBB* transcripts (Figure 5).

266

267 **Prediction of the Functional Effect of 5'SS GT>GC Mutations**

268 Finally, it is important to point out that none of the splicing prediction tools were able to accurately
269 predict the functional effect of 5'SS GT>GC mutations. For example, we analyzed the 45 disease-
270 causing 5'SS GT>GC mutations as well as the 19 functionally analyzed 5'SS GT>GC mutations that
271 generated some wild-type transcripts by means of the widely used Alamut® software suite under default
272 conditions. Whereas SpliceSiteFinder-like tended to predict a slightly reduced score, MaxEntScan,
273 NNSPLICE and GeneSplicer invariably gave no scores, for all mutations tested (Table 1;
274 [Supplementary Table S3](#)).

275

276 **Conclusions**

277 Based upon complementary data from the meta-analysis of 45 disease-causing 5'SS GT>GC mutations
278 and the cell culture-based full-length gene splicing analysis of 103 5'SS GT>GC mutations, we have
279 provided a first estimate of ~15-18% for the proportion of canonical GT 5'SSs that are capable of
280 generating between 1 and 84% normal transcripts in case of the substitution of GT by GC. Extrapolation
281 of the 15-18% value to the entire human genome implies that in at least 30,000 U2-type introns, the
282 substitution of 5'SS GT by GC would result in the retention of partial ability to generate wild-type
283 transcripts. Given that even the retention of 5% normal transcripts can significantly ameliorate a
284 patient's clinical phenotype, our findings imply the potential existence of hundreds or even thousands of
285 disease-causing 5'SS GT>GC mutations that may underlie relatively mild clinical phenotypes. Given
286 that 5'SS GT>GC mutations can also give rise to relatively high levels of wild-type transcripts, our
287 findings imply that 5'SS GT>GC mutations may not invariably cause human disease. Apart from their
288 direct implications for medical genetics, our findings may also help to improve our understanding of the
289 evolutionary processes that accompanied the GT>GC subtype switching of U2-type introns in mammals
290 [8, 21].

291

292 **MATERIALS AND METHODS**

293 **Meta-Analysis of Disease-Causing 5'SS GT>GC Mutations**

294 Human disease-causing 5'SS GT>GC mutations logged in the Professional version of the Human Gene
295 Mutation Database (HGMD; <http://www.hgmd.cf.ac.uk/ac/index.php>; as of June 2017) [32] were used as
296 starting material. The procedure of the meta-analysis is described in [Figure 1C](#).

297

298 **Cell Culture-Based Full-Length Gene Splicing Assay**

299 Outline of the cell culture-based full-length gene splicing assay is illustrated in [Figure 1C](#).

300

301 *Amplification of full-length gene sequences*

302 For this experiment, we focused on genes whose genomic sizes were <8 kb (from the translation
303 initiation codon to the translation termination codon) and whose exons numbered ≥ 3 . Long-range PCR
304 was performed in a 25 μL reaction mixture containing 0.5 U KAPA HiFi HotStart DNA Polymerase (Kapa
305 Biosystems), 0.75 μL KAPA dNTP Mix (300 μM final), 5 μL 5 \times KAPA HiFi Buffer, 50 ng DNA, and 0.3
306 μM forward and reverse primers (primer sequences available upon request). The PCR program
307 comprised an initial denaturation at 95°C for 5 min, followed by 30 cycles of denaturation at 98°C for 20
308 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
309 some of the cases where the desired fragments could not be obtained, a second amplification was
310 attempted: PCR was performed using 50 ng DNA in a 50 μL reaction mixture with 2.5 U TaKaRa LA
311 Taq DNA polymerase (TaKaRa), 8 μL dNTP Mixture (400 μM final), 5 μL 10 \times LA PCR Buffer, and 1 μM
312 forward and reverse primers; thermal cycling conditions were initial denaturation at 94°C for 1 min, 30
313 cycles of denaturation at 98°C for 10 s, annealing and extension at 68°C for 1 min/kb, and a final
314 extension at 72°C for 10 min.

315

316 *Cloning of the amplified full-length wild-type gene sequences into the expression vector*

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

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

336

337 *Mutagenesis*

338 Variants were introduced into the wild-type full-length gene expression constructs by means of the
339 QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies). Mutagenesis was performed in
340 a 25.5 μL mixture containing 1.25 U PfuUltra HF DNA polymerase, 0.5 μL dNTP mix, 2.5 μL 10 \times
341 reaction buffer, 1.5 μL QuikSolution, 100 ng wild-type plasmid, and 62.5 ng each mutagenesis primer

342 (primer sequences available upon request). The PCR program had an initial denaturation at 95°C for 2
343 min, followed by 18 cycles of denaturation at 95°C for 1 min, annealing at 60°C for 50 s, and extension
344 at 68°C for 1 min/kb, and a final extension at 68°C for 7 min. The PCR products were transformed into
345 XL10-Gold Ultracompetent cells (Agilent Technologies) after treated with *DpnI* at 37°C for 1 h.
346 Transformed cells were spread onto LB agar plates with 50 µg/mL ampicillin and incubated at 37°C
347 overnight. Selected colonies were cultured overnight. Plasmids were isolated using the QIAprep Spin
348 Miniprep Kit (Qiagen) and the successful introduction of the desired mutations was validated by DNA
349 sequencing with the BigDye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems).

350

351 *Cell culture, transfection, RNA extraction, and reverse transcription*

352 Human embryonic kidney 293T (HEK293T) cells were cultured in the Dulbecco's modified Eagle's
353 medium (BioWhittaker) with 10% fetal calf serum (Eurobio). 3.5×10^5 cells were seeded per well in 6-
354 well plates 24 h before transfection. For conventional RT-PCR analyses, 1 µg wild-type or variant
355 plasmid, mixed with 2 µL jetPEI DNA transfection reagent (Polyplus-transfection), was used for
356 transfection per well. For real-time quantitative RT-PCR analyses, 500 ng wild-type or variant plasmid
357 was mixed with 500 ng pGL3-GP2 minigene for transfection [44, 46, 50]. Forty-eight hours after
358 transfection, total RNA was extracted using the RNeasy Mini Kit (Qiagen). RT was performed with 200
359 U SuperScript III Reverse Transcriptase (Invitrogen), 500 µM dNTPs, 4 µL 5 × First-Strand Buffer, 5 mM
360 dithiothreitol, 2.5 µM 20mer-oligo (dT), and 1 µg total RNA. The resulting complementary DNA (cDNA)
361 were treated with 2U RNaseH (Invitrogen) to degrade the remaining RNA.

362

363 *Conventional RT-PCR analyses and sequencing of the resulting products*

364 Conventional RT-PCR was performed in a 25-µL reaction mixture containing 12.5 µL HotStarTaq
365 Master Mix (Qiagen), 1 µL cDNA, and 0.4 µM each primer (5'-GGAGACCCAAGCTGGCTAGT-3'
366 (forward) and 5'-AGACCGAGGAGAGGGTTAGG-3' (reverse) for TA cloning-obtained plasmids (both
367 primers are located within the pcDNA3.1/V5-His-TOPO vector sequence); 5'-
368 TAATACGACTCACTATAGGG-3' (forward) and 5'-TAGAAGGCACAGTCGAGG-3' (reverse) for in-
369 fusion cloning-obtained plasmids (both primers are located within the pcDNA3.1(+) vector sequence)).
370 The PCR program had an initial denaturation step at 95°C for 15 min, followed by 30 cycles of
371 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
372 to screen wild-type genes for which RT-PCR analysis of transfected cells generated a single or quasi-
373 single band of expected size) or for 2 min (in the step to analyze the splicing outcomes of 5'SS GT>GC
374 mutations), and a final extension step at 72°C for 10 min. RT-PCR products of a single band were
375 cleaned by ExoSAP-IT (Affymetrix). In the case of multiple bands, the band corresponding to the
376 normal-sized product was excised from the agarose gel and then purified by QIAquick Gel Extraction Kit
377 (Qiagen). Sequencing primers were those used for the RT-PCR analyses. Sequencing reaction was
378 performed by means of the BigDye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems).

379

380 *Quantitation of the relative level of correctly spliced transcripts in artificially introduced GT>GC* 381 *mutations*

382 The relative level of correctly spliced transcripts in association with GT>GC mutations that generated
383 only wild-type transcripts (confirmed by Sanger sequencing) was determined by real-time quantitative
384 RT-PCR analyses, essentially as described elsewhere [44, 46, 50]. Results were from three
385 independent transfection experiments, with each experiment being performed in three replicates.

386

387 **Pictogram Analysis of the 9-bp 5'SS Signal Sequences Associated with 5'SS GT>GC mutations**

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

392

393 ***In Silico* Splicing Prediction**

394 *In silico* splicing prediction was performed by means of Alamut® Visual v.2.11 rev. 0
395 (<https://www.interactive-biosoftware.com/>; Interactive Biosoftware, Rouen, France) under default
396 conditions.

397

398 **Acknowledgements**

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

401

402 **Funding**

403 J.H.L., a joint PhD student between the Changhai Hospital and INSERM U1078, was in receipt of a 20-
404 month scholarship from the China Scholarship Council (No. 201706580018). Support for this study
405 came from the Institut National de la Santé et de la Recherche Médicale (INSERM) and the
406 Etablissement Français du Sang (EFS), France; the National Natural Science Foundation of China
407 (81470884 (to Z.L.)), 81770636 (to Z.L.) and 81700565 (to W.B.Z.)), the Shuguang Program of
408 Shanghai (15SG33 (to Z.L.)), the Chang Jiang Scholars Program of Ministry of Education (Q2015190
409 (to Z.L.)), and the Scientific Innovation Program of Shanghai Municipal Education Committee (to Z.L.),
410 China. M.M., M.H. and D.N.C. acknowledge financial support from Qiagen Inc. through a License
411 Agreement with Cardiff University.

412

413 **REFERENCES**

- 414 1. Turunen JJ, Niemela EH, Verma B, Frilander MJ. The significant other: splicing by the minor
415 spliceosome. *Wiley Interdiscip Rev RNA*. 2013;4:61-76.
- 416 2. Parada GE, Munita R, Cerda CA, Gysling K. A comprehensive survey of non-canonical splice
417 sites in the human transcriptome. *Nucleic Acids Res*. 2014;42:10564-78.
- 418 3. Verma B, Akinyi MV, Norppa AJ, Frilander MJ. Minor spliceosome and disease. *Semin Cell Dev*
419 *Biol*. 2018;79:103-12.
- 420 4. Sharp PA, Burge CB. Classification of introns: U2-type or U12-type. *Cell*. 1997;91:875-9.
- 421 5. Papasaikas P, Valcarcel J. The spliceosome: the ultimate RNA chaperone and sculptor. *Trends*
422 *Biochem Sci*. 2016;41:33-45.
- 423 6. Mount SM. A catalogue of splice junction sequences. *Nucleic Acids Res*. 1982;10:459-72.

- 424 7. Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in
425 mammalian genomes. *Nucleic Acids Res.* 2000;28:4364-75.
- 426 8. Abril JF, Castelo R, Guigo R. Comparison of splice sites in mammals and chicken. *Genome Res.*
427 2005;15:111-9.
- 428 9. Roca X, Akerman M, Gaus H, Berdeja A, Bennett CF, Krainer AR. Widespread recognition of 5'
429 splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides. *Genes*
430 *Dev.* 2012;26:1098-109.
- 431 10. Roca X, Krainer AR, Eperon IC. Pick one, but be quick: 5' splice sites and the problems of too
432 many choices. *Genes Dev.* 2013;27:129-44.
- 433 11. Wong MS, Kinney JB, Krainer AR. Quantitative activity profile and context dependence of all
434 human 5' splice sites. *Mol Cell.* 2018;71:1012-26 e3.
- 435 12. Mount SM, Pettersson I, Hinterberger M, Karmas A, Steitz JA. The U1 small nuclear RNA-
436 protein complex selectively binds a 5' splice site in vitro. *Cell.* 1983;33:509-18.
- 437 13. Kramer A, Keller W, Appel B, Luhrmann R. The 5' terminus of the RNA moiety of U1 small
438 nuclear ribonucleoprotein particles is required for the splicing of messenger RNA precursors.
439 *Cell.* 1984;38:299-307.
- 440 14. Zhuang Y, Weiner AM. A compensatory base change in U1 snRNA suppresses a 5' splice site
441 mutation. *Cell.* 1986;46:827-35.
- 442 15. Kondo Y, Oubridge C, van Roon AM, Nagai K. Crystal structure of human U1 snRNP, a small
443 nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *Elife.*
444 2015;4.
- 445 16. Dodgson JB, Engel JD. The nucleotide sequence of the adult chicken alpha-globin genes. *J Biol*
446 *Chem.* 1983;258:4623-9.
- 447 17. Erbil C, Niessing J. The primary structure of the duck alpha D-globin gene: an unusual 5' splice
448 junction sequence. *EMBO J.* 1983;2:1339-43.
- 449 18. King CR, Piatigorsky J. Alternative RNA splicing of the murine alpha A-crystallin gene: protein-
450 coding information within an intron. *Cell.* 1983;32:707-12.
- 451 19. Burset M, Seledtsov IA, Solovyev VV. SpliceDB: database of canonical and non-canonical
452 mammalian splice sites. *Nucleic Acids Res.* 2001;29:255-9.
- 453 20. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. Comprehensive splice-
454 site analysis using comparative genomics. *Nucleic Acids Res.* 2006;34:3955-67.
- 455 21. Churbanov A, Winters-Hilt S, Koonin EV, Rogozin IB. Accumulation of GC donor splice signals in
456 mammals. *Biol Direct.* 2008;3:30.
- 457 22. Erkelenz S, Theiss S, Kaisers W, Ptok J, Walotka L, Muller L, et al. Ranking noncanonical 5' splice
458 site usage by genome-wide RNA-seq analysis and splicing reporter assays. *Genome Res.* 2018
459 Oct 24. doi: 10.1101/gr.235861.118. [Epub ahead of print].
- 460 23. Aebi M, Hornig H, Padgett RA, Reiser J, Weissmann C. Sequence requirements for splicing of
461 higher eukaryotic nuclear pre-mRNA. *Cell.* 1986;47:555-65.
- 462 24. Aebi M, Hornig H, Weissmann C. 5' cleavage site in eukaryotic pre-mRNA splicing is
463 determined by the overall 5' splice region, not by the conserved 5' GU. *Cell.* 1987;50:237-46.
- 464 25. Pagani F, Buratti E, Stuni C, Bendix R, Dork T, Baralle FE. A new type of mutation causes a
465 splicing defect in ATM. *Nat Genet.* 2002;30:426-9.
- 466 26. Kralovicova J, Hwang G, Asplund AC, Churbanov A, Smith CI, Vorechovsky I. Compensatory
467 signals associated with the activation of human GC 5' splice sites. *Nucleic Acids Res.*
468 2011;39:7077-91.
- 469 27. Zhang XH, Arias MA, Ke S, Chasin LA. Splicing of designer exons reveals unexpected complexity
470 in pre-mRNA splicing. *RNA.* 2009;15:367-76.
- 471 28. De Conti L, Baralle M, Buratti E. Exon and intron definition in pre-mRNA splicing. *Wiley*
472 *Interdiscip Rev RNA.* 2013;4:49-60.
- 473 29. Boehm V, Britto-Borges T, Steckelberg AL, Singh KK, Gerbracht JV, Gueney E, et al. Exon
474 junction complexes suppress spurious splice sites to safeguard transcriptome integrity. *Mol*
475 *Cell.* 2018;72:482-95 e7.

- 476 30. Ramalho AS, Beck S, Meyer M, Penque D, Cutting GR, Amaral MD. Five percent of normal
477 cystic fibrosis transmembrane conductance regulator mRNA ameliorates the severity of
478 pulmonary disease in cystic fibrosis. *Am J Respir Cell Mol Biol.* 2002;27:619-27.
- 479 31. Raraigh KS, Han ST, Davis E, Evans TA, Pellicore MJ, McCague AF, et al. Functional assays are
480 essential for interpretation of missense variants associated with variable expressivity. *Am J*
481 *Hum Genet.* 2018;102:1062-77.
- 482 32. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation
483 Database: towards a comprehensive repository of inherited mutation data for medical
484 research, genetic diagnosis and next-generation sequencing studies. *Hum Genet.*
485 2017;136:665-77.
- 486 33. Desviat LR, Clavero S, Perez-Cerda C, Navarrete R, Ugarte M, Perez B. New splicing mutations
487 in propionic acidemia. *J Hum Genet.* 2006;51:992-7.
- 488 34. Soudais C, de Villartay JP, Le Deist F, Fischer A, Lisowska-Groszpiere B. Independent mutations
489 of the human CD3-epsilon gene resulting in a T cell receptor/CD3 complex immunodeficiency.
490 *Nat Genet.* 1993;3:77-81.
- 491 35. Lin P, Li W, Wen B, Zhao Y, Fenster DS, Wang Y, et al. Novel *PNPLA2* gene mutations in Chinese
492 Han patients causing neutral lipid storage disease with myopathy. *J Hum Genet.* 2012;57:679-
493 81.
- 494 36. Seyama K, Nonoyama S, Gangsaas I, Hollenbaugh D, Pabst HF, Aruffo A, et al. Mutations of the
495 CD40 ligand gene and its effect on CD40 ligand expression in patients with X-linked hyper IgM
496 syndrome. *Blood.* 1998;92:2421-34.
- 497 37. Bartolo C, Papp AC, Snyder PJ, Sedra MS, Burghes AH, Hall CD, et al. A novel splice site
498 mutation in a Becker muscular dystrophy patient. *J Med Genet.* 1996;33:324-7.
- 499 38. Hastbacka J, Kerrebrock A, Mokkalá K, Clines G, Lovett M, Kaitila I, et al. Identification of the
500 Finnish founder mutation for diastrophic dysplasia (DTD). *Eur J Hum Genet.* 1999;7:664-70.
- 501 39. Muller JS, Piko H, Schoser BG, Schlotter-Weigel B, Reilich P, Gurster S, et al. Novel splice site
502 mutation in the caveolin-3 gene leading to autosomal recessive limb girdle muscular
503 dystrophy. *Neuromuscul Disord.* 2006;16:432-6.
- 504 40. Aoyagi Y, Kobayashi H, Tanaka K, Ozawa T, Nitta H, Tsuji S. A de novo splice donor site
505 mutation causes in-frame deletion of 14 amino acids in the proteolipid protein in Pelizaeus-
506 Merzbacher disease. *Ann Neurol.* 1999;46:112-5.
- 507 41. Kume K, Masamune A, Kikuta K, Shimosegawa T. [-215G>A; IVS3+2T>C] mutation in the *SPINK1*
508 gene causes exon 3 skipping and loss of the trypsin binding site. *Gut.* 2006;55:1214.
- 509 42. Zou WB, Boulling A, Masson E, Cooper DN, Liao Z, Li ZS, et al. Clarifying the clinical relevance of
510 *SPINK1* intronic variants in chronic pancreatitis. *Gut.* 2016;65:884-6.
- 511 43. Zou WB, Masson E, Boulling A, Cooper DN, Li ZS, Liao Z, et al. Digging deeper into the intronic
512 sequences of the *SPINK1* gene. *Gut.* 2016;65:1055-6.
- 513 44. Boulling A, Chen JMC, I., Férec C. Is the *SPINK1* p.Asn34Ser missense mutation *per se* the true
514 culprit within its associated haplotype? *WebmedCentral GENETICS.* 2012;3:WMC003084
515 (Available at: https://www.webmedcentral.com/article_view/3084). Accessed 14 November
516 2018.
- 517 45. Fu XD, Ares M, Jr. Context-dependent control of alternative splicing by RNA-binding proteins.
518 *Nat Rev Genet.* 2014;15:689-701.
- 519 46. Zou WB, Boulling A, Masamune A, Issarapu P, Masson E, Wu H, et al. No association between
520 *CEL-HYB* hybrid allele and chronic pancreatitis in Asian populations. *Gastroenterology.*
521 2016;150:1558-60 e5.
- 522 47. Lykke-Andersen S, Jensen TH. Nonsense-mediated mRNA decay: an intricate machinery that
523 shapes transcriptomes. *Nat Rev Mol Cell Biol.* 2015;16:665-77.
- 524 48. Popp MW, Maquat LE. Leveraging rules of nonsense-mediated mRNA decay for genome
525 engineering and personalized medicine. *Cell.* 2016;165:1319-22.

- 526 49. Pereverzev AP, Gurskaya NG, Ermakova GV, Kudryavtseva EI, Markina NM, Kotlobay AA, et al.
527 Method for quantitative analysis of nonsense-mediated mRNA decay at the single cell level. *Sci*
528 *Rep.* 2015;5:7729.
- 529 50. Zou WB, Wu H, Boulling A, Cooper DN, Li ZS, Liao Z, et al. *In silico* prioritization and further
530 functional characterization of *SPINK1* intronic variants. *Hum Genomics.* 2017;11:7.
- 531 51. Wu H, Boulling A, Cooper DN, Li ZS, Liao Z, Chen JM, et al. *In vitro* and *in silico* evidence against
532 a significant effect of the *SPINK1* c.194G>A variant on pre-mRNA splicing. *Gut.* 2017;66:2195-6.
- 533 52. Ota Y, Masamune A, Inui K, Kume K, Shimosegawa T, Kikuyama M. Phenotypic variability of the
534 homozygous IVS3+2T>C mutation in the serine protease inhibitor Kazal type 1 (*SPINK1*) gene in
535 patients with chronic pancreatitis. *Tohoku J Exp Med.* 2010;221:197-201.
- 536 53. Venet T, Masson E, Talbotec C, Billiemaz K, Touraine R, Gay C, et al. Severe infantile isolated
537 exocrine pancreatic insufficiency caused by the complete functional loss of the *SPINK1* gene.
538 *Hum Mutat.* 2017;38:1660-5.
- 539 54. Yilmaz D, Cogulu O, Ozkinay F, Kavakli K, Roos D. A novel mutation in the *DIA1* gene in a
540 patient with methemoglobinemia type II. *Am J Med Genet A.* 2005;133A:101-2.
- 541 55. Frischknecht H, Dutly F, Walker L, Nakamura-Garrett LM, Eng B, Wayne JS. Three new beta-
542 thalassemia mutations with varying degrees of severity. *Hemoglobin.* 2009;33:220-5.
- 543 56. Hladnik U, Nyhan WL, Bertelli M. Variable expression of HPRT deficiency in 5 members of a
544 family with the same mutation. *Arch Neurol.* 2008;65:1240-3.
- 545 57. Wuhl E, Kogan J, Zurowska A, Matejas V, Vandevoorde RG, Aigner T, et al.
546 Neurodevelopmental deficits in Pierson (microcoria-congenital nephrosis) syndrome. *Am J*
547 *Med Genet A.* 2007;143:311-9.
- 548 58. Bar DZ, Arlt MF, Brazier JF, Norris WE, Campbell SE, Chines P, et al. A novel somatic mutation
549 achieves partial rescue in a child with Hutchinson-Gilford progeria syndrome. *J Med Genet.*
550 2017;54:212-6.
- 551 59. Al-Mahdili HA, Hooper AJ, Sullivan DR, Stewart PM, Burnett JR. A mild case of
552 abetalipoproteinaemia in association with subclinical hypothyroidism. *Ann Clin Biochem.*
553 2006;43:516-9.
- 554 60. Kajihara S, Hisatomi A, Mizuta T, Hara T, Ozaki I, Wada I, et al. A splice mutation in the human
555 canalicular multispecific organic anion transporter gene causes Dubin-Johnson syndrome.
556 *Biochem Biophys Res Commun.* 1998;253:454-7.
- 557 61. Fukao T, Yamaguchi S, Sriver CR, Dunbar G, Wakazono A, Kano M, et al. Molecular studies of
558 mitochondrial acetoacetyl-coenzyme A thiolase deficiency in the two original families. *Hum*
559 *Mutat.* 1993;2:214-20.
- 560 62. Lagier-Tourenne C, Tazir M, Lopez LC, Quinzii CM, Assoum M, Drouot N, et al. ADCK3, an
561 ancestral kinase, is mutated in a form of recessive ataxia associated with coenzyme Q10
562 deficiency. *Am J Hum Genet.* 2008;82:661-72.
- 563 63. Dolcini L, Caridi G, Dagnino M, Sala A, Gokce S, Sokucu S, et al. Analbuminemia produced by a
564 novel splicing mutation. *Clin Chem.* 2007;53:1549-52.
- 565 64. Infante JB, Alvelos MI, Bastos M, Carrilho F, Lemos MC. Complete androgen insensitivity
566 syndrome caused by a novel splice donor site mutation and activation of a cryptic splice donor
567 site in the androgen receptor gene. *J Steroid Biochem Mol Biol.* 2016;155:63-6.
- 568 65. Rios M, Storry JR, Hue-Roye K, Chung A, Reid ME. Two new molecular bases for the Dombrock
569 null phenotype. *Br J Haematol.* 2002;117:765-7.
- 570 66. Das S, Levinson B, Whitney S, Vulpe C, Packman S, Gitschier J. Diverse mutations in patients
571 with Menkes disease often lead to exon skipping. *Am J Hum Genet.* 1994;55:883-9.
- 572 67. Hopp K, Heyer CM, Hommerding CJ, Henke SA, Sundsbak JL, Patel S, et al. *B9D1* is revealed as a
573 novel Meckel syndrome (MKS) gene by targeted exon-enriched next-generation sequencing
574 and deletion analysis. *Hum Mol Genet.* 2011;20:2524-34.
- 575 68. Haire RN, Ohta Y, Strong SJ, Litman RT, Liu Y, Prchal JT, et al. Unusual patterns of exon skipping
576 in Bruton tyrosine kinase are associated with mutations involving the intron 17 3' splice site.
577 *Am J Hum Genet.* 1997;60:798-807.

- 578 69. Rahner N, Nuernberg G, Finis D, Nuernberg P, Royer-Pokora B. A novel *C8orf37* splice mutation
579 and genotype-phenotype correlation for cone-rod dystrophy. *Ophthalmic Genet.* 2016;37:294-
580 300.
- 581 70. Tosetto E, Ghiggeri GM, Emma F, Barbano G, Carrea A, Vezzoli G, et al. Phenotypic and genetic
582 heterogeneity in Dent's disease--the results of an Italian collaborative study. *Nephrol Dial*
583 *Transplant.* 2006;21:2452-63.
- 584 71. Nicholls AC, Valler D, Wallis S, Pope FM. Homozygosity for a splice site mutation of the *COL1A2*
585 gene yields a non-functional pro(α)2(I) chain and an EDS/OI clinical phenotype. *J Med*
586 *Genet.* 2001;38:132-6.
- 587 72. Haas JT, Winter HS, Lim E, Kirby A, Blumenstiel B, DeFelice M, et al. *DGAT1* mutation is linked
588 to a congenital diarrheal disorder. *J Clin Invest.* 2012;122:4680-4.
- 589 73. Wibawa T, Takeshima Y, Mitsuyoshi I, Wada H, Surono A, Nakamura H, et al. Complete
590 skipping of exon 66 due to novel mutations of the dystrophin gene was identified in two
591 Japanese families of Duchenne muscular dystrophy with severe mental retardation. *Brain Dev.*
592 2000;22:107-12.
- 593 74. Ahmed I, Mittal K, Sheikh TI, Vasli N, Rafiq MA, Mikhailov A, et al. Identification of a
594 homozygous splice site mutation in the dynein axonemal light chain 4 gene on 22q13.1 in a
595 large consanguineous family from Pakistan with congenital mirror movement disorder. *Hum*
596 *Genet.* 2014;133:1419-29.
- 597 75. Hermans MM, van Leenen D, Kroos MA, Reuser AJ. Mutation detection in glycogen storage-
598 disease type II by RT-PCR and automated sequencing. *Biochem Biophys Res Commun.*
599 1997;241:414-8.
- 600 76. Sobrier ML, Maghnie M, Vie-Luton MP, Secco A, di Iorgi N, Lorini R, et al. Novel *HESX1*
601 mutations associated with a life-threatening neonatal phenotype, pituitary aplasia, but
602 normally located posterior pituitary and no optic nerve abnormalities. *J Clin Endocrinol Metab.*
603 2006;91:4528-36.
- 604 77. Moran CJ, Walters TD, Guo CH, Kugathasan S, Klein C, Turner D, et al. IL-10R polymorphisms
605 are associated with very-early-onset ulcerative colitis. *Inflamm Bowel Dis.* 2013;19:115-23.
- 606 78. Humbert C, Silbermann F, Morar B, Parisot M, Zarhrate M, Masson C, et al. Integrin alpha 8
607 recessive mutations are responsible for bilateral renal agenesis in humans. *Am J Hum Genet.*
608 2014;94:288-94.
- 609 79. Vockley J, Rogan PK, Anderson BD, Willard J, Seelan RS, Smith DI, et al. Exon skipping in *IVD*
610 RNA processing in isovaleric acidemia caused by point mutations in the coding region of the
611 *IVD* gene. *Am J Hum Genet.* 2000;66:356-67.
- 612 80. Villa A, Sironi M, Macchi P, Matteucci C, Notarangelo LD, Vezzoni P, et al. Monocyte function in
613 a severe combined immunodeficient patient with a donor splice site mutation in the *Jak3*
614 gene. *Blood.* 1996;88:817-23.
- 615 81. Allamand V, Sunada Y, Salih MA, Straub V, Ozo CO, Al-Turaiki MH, et al. Mild congenital
616 muscular dystrophy in two patients with an internally deleted laminin alpha2-chain. *Hum Mol*
617 *Genet.* 1997;6:747-52.
- 618 82. Nichols WC, Seligsohn U, Zivelin A, Terry VH, Hertel CE, Wheatley MA, et al. Mutations in the
619 ER-Golgi intermediate compartment protein ERGIC-53 cause combined deficiency of
620 coagulation factors V and VIII. *Cell.* 1998;93:61-70.
- 621 83. Wood-Trageser MA, Gurbuz F, Yatsenko SA, Jeffries EP, Kotan LD, Surti U, et al. *MCM9*
622 mutations are associated with ovarian failure, short stature, and chromosomal instability. *Am J*
623 *Hum Genet.* 2014;95:754-62.
- 624 84. Gok F, Crettol LM, Alanay Y, Hacıhamdioglu B, Kocaoglu M, Bonafe L, et al. Clinical and
625 radiographic findings in two brothers affected with a novel mutation in matrix
626 metalloproteinase 2 gene. *Eur J Pediatr.* 2010;169:363-7.
- 627 85. Martin CA, Murray JE, Carroll P, Leitch A, Mackenzie KJ, Halachev M, et al. Mutations in genes
628 encoding condensin complex proteins cause microcephaly through decatenation failure at
629 mitosis. *Genes Dev.* 2016;30:2158-72.

- 630 86. Tanugi-Cholley LC, Issartel JP, Lunardi J, Freycon F, Morel F, Vignais PV. A mutation located at
631 the 5' splice junction sequence of intron 3 in the p67phox gene causes the lack of p67phox
632 mRNA in a patient with chronic granulomatous disease. *Blood*. 1995;85:242-9.
- 633 87. Zanni G, Saillour Y, Nagara M, Billuart P, Castelnau L, Moraine C, et al. Oligophrenin 1
634 mutations frequently cause X-linked mental retardation with cerebellar hypoplasia. *Neurology*.
635 2005;65:1364-9.
- 636 88. Matsuura T, Hoshide R, Komaki S, Kiwaki K, Endo F, Nakamura S, et al. Identification of two
637 new aberrant splicings in the ornithine carbamoyltransferase (*OCT*) gene in two patients with
638 early and late onset OCT deficiency. *J Inher Metab Dis*. 1995;18:273-82.
- 639 89. Shimozawa N, Nagase T, Takemoto Y, Suzuki Y, Fujiki Y, Wanders RJ, et al. A novel aberrant
640 splicing mutation of the *PEX16* gene in two patients with Zellweger syndrome. *Biochem*
641 *Biophys Res Commun*. 2002;292:109-12.
- 642 90. Biancheri R, Grossi S, Regis S, Rossi A, Corsolini F, Rossi DP, et al. Further genotype-phenotype
643 correlation emerging from two families with *PLP1* exon 4 skipping. *Clin Genet*. 2014;85:267-72.
- 644 91. Aldahmesh MA, Mohamed JY, Alkuraya FS. A novel mutation in *PRDM5* in brittle cornea
645 syndrome. *Clin Genet*. 2012;81:198-9.
- 646 92. Smith SB, Qu HQ, Taleb N, Kishimoto NY, Scheel DW, Lu Y, et al. Rfx6 directs islet formation
647 and insulin production in mice and humans. *Nature*. 2010;463:775-80.
- 648 93. Adly N, Alhashem A, Ammari A, Alkuraya FS. Ciliary genes *TBC1D32/C6orf170* and *SCLT1* are
649 mutated in patients with OFD type IX. *Hum Mutat*. 2014;35:36-40.
- 650 94. Sumegi J, Huang D, Lanyi A, Davis JD, Seemayer TA, Maeda A, et al. Correlation of mutations of
651 the *SH2D1A* gene and epstein-barr virus infection with clinical phenotype and outcome in X-
652 linked lymphoproliferative disease. *Blood*. 2000;96:3118-25.
- 653 95. Leman R, Gaildrat P, Gac GL, Ka C, Fichou Y, Audrezet MP, et al. Novel diagnostic tool for
654 prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro
655 studies: an international collaborative effort. *Nucleic Acids Res*. 2018;46:7913-23.
- 656

657 **Table 1. The 45 Informative Disease-Causing 5'SS GT>GC Mutations and their Predicted Splicing Effects***

Disease	Gene	Mutation	Reference	Zygoty	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	<i>ABCC2</i>	c.1967+2T>C	Kajihara et al. [60]	Homozygote	No	74.2→NS	7.4→NS	0.6→NS	1.8→NS
Acetoacetyl-CoA-thiolase deficiency	<i>ACAT1</i>	c.1163+2T>C	Fukao et al. [61]	Homozygote	No	76.8→74.1	6.9→NS	1.0→NS	NS
Ubiquinone deficiency with cerebellar ataxia	<i>COQ8A</i>	c.1398+2T>C	Lagier-Tourenne et al. [62]	Homozygote	No	76.9→75.2	6.4→NS	0.8→NS	4.6→NS
Analbuminaemia	<i>ALB</i>	c.1428+2T>C	Dolcini et al. [63]	Homozygote	No	78.9→70.8	5.6→NS	1.0→NS	NS
Androgen insensitivity syndrome	<i>AR</i>	c.2173+2T>C	Infante et al. [64]	Hemizygote	No	86.5→81.1	9.7→NS	1.0→NS	4.4→NS
Dombrock null allele	<i>ART4</i>	c.144+2T>C	Rios et al. [65]	Homozygote	No	81.8→79.9	7.8→NS	0.9→NS	4.1→NS
Menkes syndrome	<i>ATP7A</i>	c.1946+2T>C	Das et al. [66]	Hemizygote	No	78.6→75.5	9.5→NS	1.0→NS	3.8→NS
Meckel syndrome	<i>B9D1</i>	c.341+2T>C	Hopp et al. [67]	Hemizygote	No	81.0→78.7	9.4→NS	1.0→NS	8.2→NS
Agammaglobulinaemia	<i>BTK</i>	c.588+2T>C	Haire et al. [68]	Hemizygote	No	71.9→NS	7.5→NS	0.8→NS	4.0→NS
Cone-rod dystrophy	<i>C8orf37</i>	c.155+2T>C	Rahner et al. [69]	Homozygote	No	72.2→NS	1.6→NS	0.6→NS	2.1→NS
Autosomal recessive limb girdle muscular dystrophy	<i>CAV3</i>	c.114+2T>C	Muller et al. [39]	Homozygote	Yes	83.8→81.8	10.1→NS	1.0→NS	10.1→NS
Immunodeficiency	<i>CD3E</i>	c.520+2T>C	Soudais et al. [34]	Compound heterozygote	Yes (1-5%)	83.0→78.1	8.1→NS	1.0→NS	2.9→NS
Hyper-IgM syndrome	<i>CD40LG</i>	c.346+2T>C	Seyama et al [36]	Hemizygote	Yes (15%)	89.6→90.0	10.3→NS	1.0→NS	1.6→NS
Dent disease	<i>CLCN5</i>	c.205+2T>C	Tosetto et al. [70]	Hemizygote	No	84.8→82.1	10.0→NS	1.0→NS	NS
Ehlers-Danlos syndrome/Osteogenesis imperfecta	<i>COL1A2</i>	c.3105+2T>C	Nicholls et al. [71]	Homozygote	No	75.4→72.8	8.6→NS	0.9→NS	1.1→NS
Congenital diarrhoeal disorder	<i>DGAT1</i>	c.751+2T>C	Haas et al. [72]	Homozygote	No	78.6→NS	7.9→NS	1.0→NS	11.9→NS
Becker muscular dystrophy	<i>DMD</i>	c.8027+2T>C	Bartolo et al. [37]	Hemizygote	Yes (10%)	84.2→81.5	9.1→NS	1.0→NS	1.7→NS
Duchenne muscular dystrophy	<i>DMD</i>	c.9649+2T>C	Wibawa et al. [73]	Hemizygote	No	84.3→84.4	9.1→NS	1.0→NS	NS
Mirror movements (congenital)	<i>DNAL4</i>	c.153+2T>C	Ahmed et al. [74]	Homozygote	No	NS	7.4→NS	0.8→NS	9.7→NS
Glycogen storage disease 2	<i>GAA</i>	c.2331+2T>C	Hermans et al. [75]	Homozygote	No	86.4→76.7	11.5→NS	1.0→NS	13.6→NS
Pituitary aplasia	<i>HESX1</i>	c.357+2T>C	Sobrier et al. [76]	Homozygote	No	80.2→70.6	6.7→NS	0.8→NS	NS
Ulcerative colitis	<i>IL10RA</i>	c.688+2T>C	Moran et al. [77]	Homozygote	No	73.8→NS	7.0→NS	0.8→NS	2.7→NS
Renal hypodysplasia	<i>ITGA8</i>	c.2982+2T>C	Humbert et al. [78]	Homozygote	No	71.9→NS	5.8→NS	0.9→NS	NS
Isovaleric acidaemia	<i>IVD</i>	c.465+2T>C	Vockley et al. [79]	Homozygote	No	90.3→80.5	9.2→NS	1.0→NS	4.5→NS
Immunodeficiency (severe combined)	<i>JAK3</i>	c.2350+2T>C	Villa et al. [80]	Homozygote	No	NS	5.8→NS	NS	6.8→NS

Muscular dystrophy (merosin deficient)	<i>LAMA2</i>	c.3924+2T>C	Allamand et al. [81]	Homozygote	No	79.8→77.0	8.3→NS	0.8→NS	3.4→NS
Factor V and factor VIII deficiency (combined)	<i>LMAN1</i>	c.1149+2T>C	Nichols et al. [82]	Homozygote	No	79.8→70.6	8.1→NS	NS	NS
Primary amenorrhea & short stature	<i>MCM9</i>	c.1732+2T>C	Wood-Trageser et al. [83]	Homozygote	No	NS	1.7→NS	NS	NS
Torg-Winchester syndrome	<i>MMP2</i>	c.658+2T>C	Gok et al. [84]	Homozygote	No	90.0→80.1	8.7→NS	1.0→NS	10.9→NS
Microcephaly	<i>NCAPD2</i>	c.4120+2T>C	Martin et al. [85]	Homozygote	No	84.2→81.5	9.1→NS	0.9→NS	7.1→NS
Chronic granulomatous disease	<i>NCF2</i>	c.257+2T>C	Tanugi-Cholley et al. [86]	Homozygote	No	84.8→84.7	9.8→NS	1.0→NS	5.3→NS
Mental retardation syndrome (X-linked)	<i>OPHN1</i>	c.154+2T>C	Zanni et al. [87]	Hemizygote	No	84.8→84.7	9.8→NS	1.0→NS	7.9→NS
Ornithine carbamoyltransferase deficiency	<i>OTC</i>	c.540+2T>C	Matsuura et al. [88]	Hemizygote	No	80.0→78.2	8.1→NS	0.6→NS	NS
Propionic acidaemia	<i>PCCB</i>	c.183+2T>C	Desviat et al. [33]	Homozygote	No	74.5→NS	8.5→NS	0.9→NS	9.7→NS
Propionic acidaemia	<i>PCCB</i>	c.1498+2T>C	Desviat et al. [33]	Homozygote	No	81.8→79.9	7.8→NS	0.7→NS	5.0→NS
Zellweger syndrome	<i>PEX16</i>	c.952+2T>C	Shimozawa et al. [89]	Homozygote	No	82.1→79.0	7.5→NS	1.0→NS	5.2→NS
Spastic tetraparesis/paraparesis	<i>PLP1</i>	c.622+2T>C	Biancheri et al. [90]	Hemizygote	No	86.8→77.2	10.1→NS	1.0→NS	6.2→NS
Pelizaeus-Merzbacher disease	<i>PLP1</i>	c.696+2T>C	Aoyagi et al. [40]	Hemizygote	Yes	92.6→85.9	10.0→NS	1.0→NS	6.5→NS
Neutral lipid storage disease with myopathy	<i>PNPLA2</i>	c.757+2T>C	Lin et al. [35]	Compound heterozygote	No	NS	8.7→NS	NS	8.3→NS
Brittle cornea syndrome	<i>PRDM5</i>	c.93+2T>C	Aldahmesh et al. [91]	Homozygote	No	85.3→78.5	8.2→NS	0.9→NS	10.6→NS
Diabetes (neonatal, with intestinal atresia)	<i>RFX6</i>	c.380+2T>C	Smith et al. [92]	Homozygote	No	78.7→NS	5.5→NS	0.6→NS	2.9→NS
Oro-facio-digital syndrome type IX	<i>SCLT1</i>	c.290+2T>C	Adly et al. [93]	Homozygote	No	87.5→87.1	8.9→NS	1.0→NS	NS
Lymphoproliferative syndrome (X-linked)	<i>SH2D1A</i>	c.137+2T>C	Sumegi et al. [94]	Hemizygote	No	71.1→NS	7.4→NS	0.4→NS	5.9→NS
Diastrophic dysplasia	<i>SLC26A2</i>	c.-26+2T>C	Hastbacks et al. [38]	Homozygote	Yes (5%)	87.3→77.7	7.7→NS	1.0→NS	11.5→NS
Chronic pancreatitis	<i>SPINK1</i>	c.194+2T>C	Kume et al. [41]	Homozygote	Yes	82.6→72.3	11.1→NS	1.0→NS	4.0→NS

658 *See [Supplementary Table S1](#) for more information.

659 ^aRelative expression level is indicated in parentheses wherever applicable.

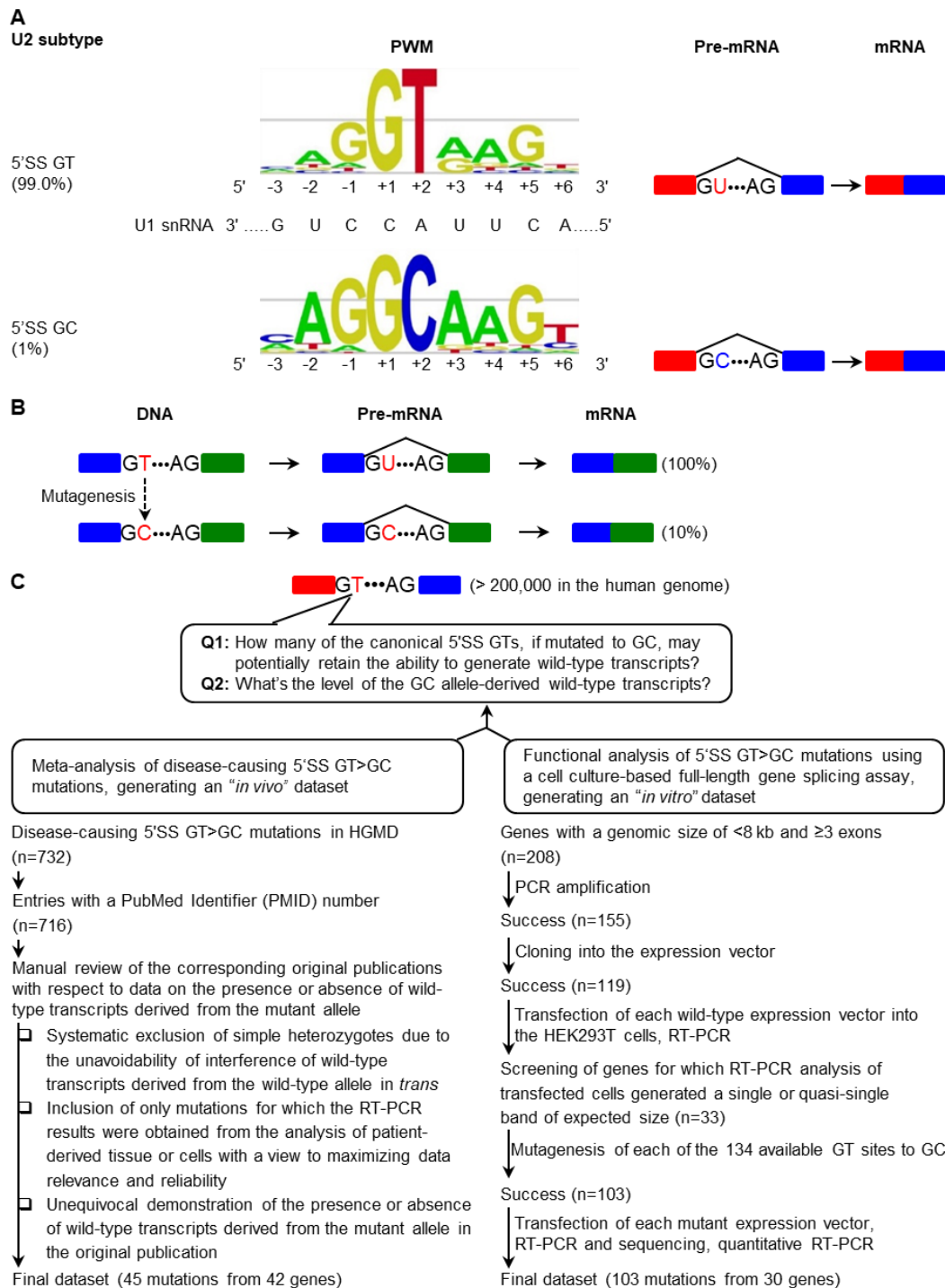
660 ^bPrediction was done under default conditions. NS, no score.

661

662 **Table 2. Nature of the Sequenced 12 Aberrantly Spliced Transcripts***

Gene	Reference mRNA accession number	Mutation	Aberrant Transcripts
<i>DBI</i>	NM_001079862.2	IVS1+2T>C	1. Activation of a cryptic 5'SS GT which is 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 which is 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
<i>FABP7</i>	NM_001446.4	IVS1+2T>C	Activation of a cryptic 5'SS GT which is 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 which is located 3 bp upstream of the normal one, resulting in the deletion of the last 5 bp of exon 1.
<i>HESX1</i>	NM_003865.2	IVS2+2T>C	Exon 2 skipping
		IVS3+2T>C	Exon 3 skipping
<i>IL10</i>	NM_000572.3	IVS1+2T>C	Activation of a cryptic 5'SS GT which is 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 which is located 19 bp upstream of the normal one, resulting in the deletion of the last 21 bp of exon 4.
<i>PRSS2</i>	NM_002770.3	IVS4+2T>C	Activation of a cryptic 5'SS GC which is located 15 bp upstream of the normal one, resulting in the deletion of the last 17 bp of exon 4.
<i>SPINK1</i>	NM_003122.3	IVS1+2T>C	Activation of a cryptic 5'SS GT which is located 138 bp downstream of the normal one, resulting in the retention of the first 140 bp of the intron 1 sequence.
<i>UQCRB</i>	NM_006294.4	IVS1+2T>C	Activation of a cryptic 5'SS GT which is located 10 bp upstream of the normal one, resulting in the deletion of the last 12 bp of exon 1.

663 *See [Supplementary Figure S1](#) for the corresponding RT-PCR products



664

665

666

667

668

669

670

671

672

673

674

675

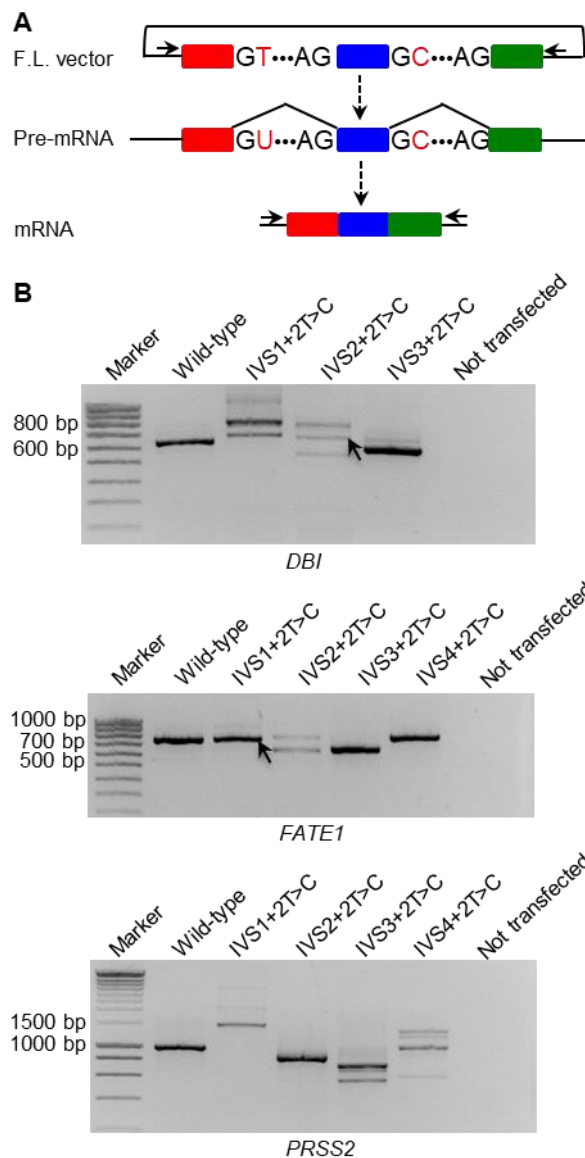
Figure 1. Background Information, Aims and Analytical Strategy of the Study

(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. 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.* 2018;46(15):7913-7923 [95] (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 mutation may retain the ability to generate wild-type transcripts, albeit at a much reduced level (~10% of normal in [23, 24]).

(C) Aim and analytical strategy of the study.

676



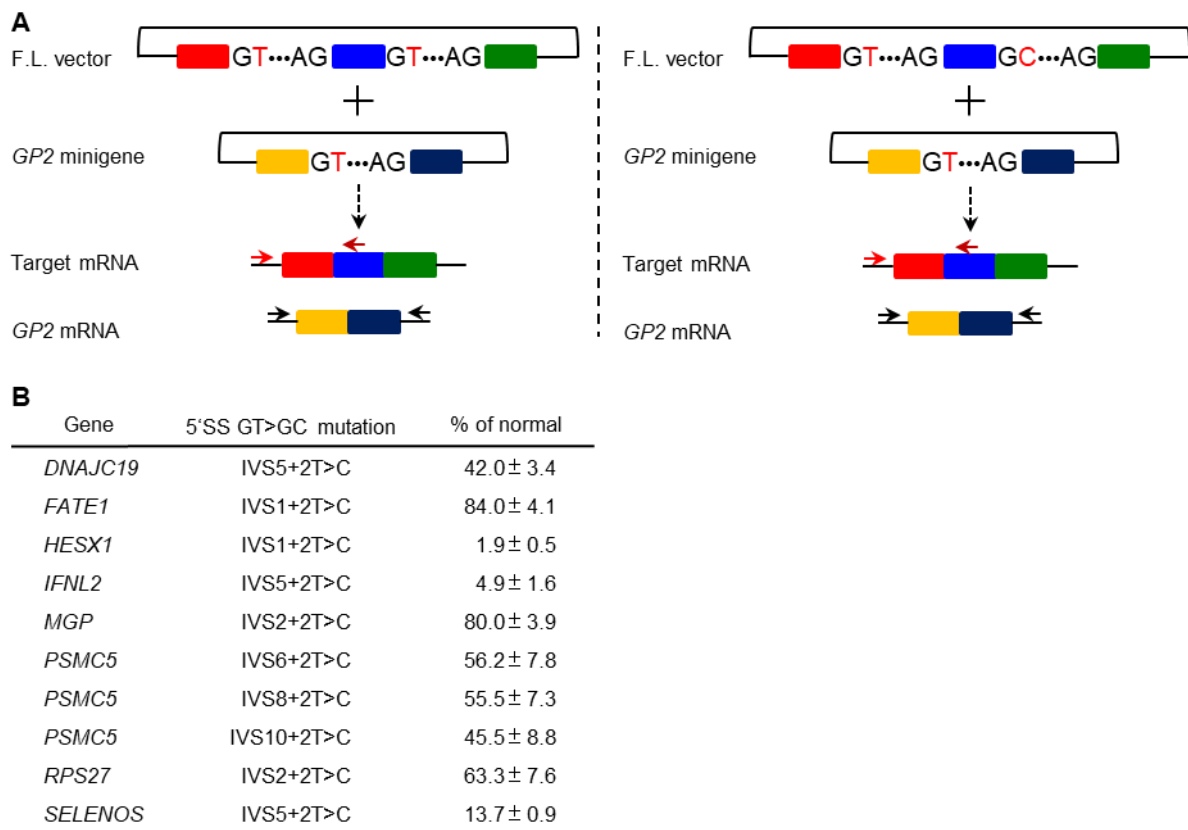
677

678 **Figure 2. Qualitative Analysis of 5'SS GT>GC Mutations**

679 (A) Illustration of the cell culture-based full-length gene splicing assay in the context of a 5'SS GT>GC
680 mutation generating some wild-type transcripts. The two horizontal arrows indicate the primers (both
681 located within the vector sequence) used to amplify normally spliced transcripts (and also aberrantly
682 spliced transcripts). F.L., full-length.

683 (B) RT-PCR analyses of HEK293T cells transfected with full-length *DBI*, *FATE1* and *PRSS2* gene
684 expression constructs carrying respectively the wild-type and 5'SS GT>GC mutations as examples.
685 Normal transcripts (confirmed by sequencing) resulting from two of the mutations are indicated by
686 arrows. IVS, InterVening Sequence (i.e., an intron). See [Supplementary Figure S1](#) for all 103
687 functionally analyzed 5'SS GT>GC mutations.
688

689



690

691

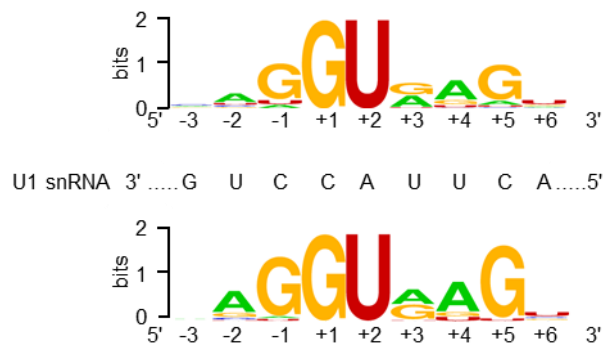
692 **Figure 3. Quantitative Analysis Pertaining to the Relative Level of 5'SS GT>GC Mutation-**
 693 **Derived Wild-Type Transcripts**

694 (A) Illustration of one key feature of the quantitative RT-PCR analysis: co-transfection of a minigene
 695 expression vector with respectively the full-length wild-type target gene expression vector and the full-
 696 length variant target gene expression vector. The minigene was constructed in pGL3 [44] whereas the
 697 target gene was constructed in either pcDNA3.1/V5-His-TOPO vector or pcDNA3.1(+). The minigene
 698 was used as an internal control for quantifying the expression level of wild-type transcripts generated
 699 from either the wild-type or variant target full-length gene. The horizontal arrows indicate the relative
 700 positions of the primers used for this purpose. Note that for amplifying the target gene sequence,
 701 either a primer pair comprising a forward vector-specific primer and a reverse gene-specific primer (as
 702 illustrated) or alternatively a primer pair comprising a forward gene-specific primer and a reverse
 703 vector-specific primer was used. This assay was performed exclusively for the 10 5'SS GT>GC
 704 mutations that generated only wild-type transcripts. F.L., full-length.

705 (B) Quantitative RT-PCR-determined expression level of the mutant allele-derived correctly spliced
 706 transcripts relative to that derived from the corresponding wild-type allele (defined as 100%) in the 10
 707 5'SS GT>GC mutations that generated only wild-type transcripts. Results were expressed as means ±
 708 SD from three independent transfection experiments.

709

710



711

712 **Figure 4. Pictogram Analysis of the 5'SSs Under Study**

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

718

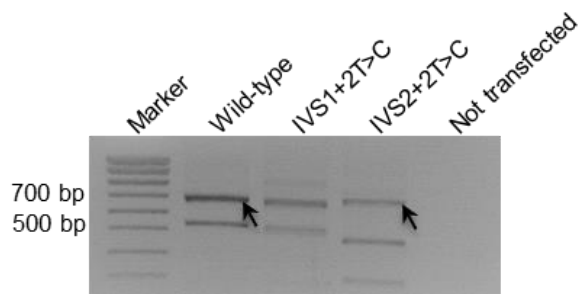
719

720

721

722

723



724

725

726 **Figure 5. Functional Characterization of the *HBB* c.315+2T>C mutation**

727 RT-PCR analyses of HEK293T cells transfected with full-length *HBB* gene expression constructs
728 carrying respectively the wild-type and two 5'SS GT>GC mutations. Wild-type transcripts (confirmed
729 by sequencing) resulting from the wild-type and the IVS2+2T>C (i.e., c.315+2T>C) mutation are
730 indicated by arrows. The *HBB* c.315+2T>C mutation was previously reported to be associated with a
731 mild phenotype [55]. IVS, InterVening Sequence (i.e., an intron).

732