

1 The experimentally obtained functional impact assessments of GT>GC 5' splice 2 site variants differ markedly from those predicted

3
4 Jian-Min Chen¹, Jin-Huan Lin^{1,2,3}, Emmanuelle Masson^{1,4}, Zhuan Liao^{2,3}, Claude Férec^{1,4},
5 David N. Cooper⁵, Matthew Hayden⁵

6
7 ¹EFS, Univ Brest, Inserm, UMR 1078, GGB, F-29200 Brest, France

8 ²Department of Gastroenterology, Changhai Hospital, Second Military Medical University,
9 Shanghai, China

10 ³Shanghai Institute of Pancreatic Diseases, Shanghai, China

11 ⁴CHRU Brest, Service de Génétique, Brest, France

12 ⁵Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, United
13 Kingdom

14
15 *Correspondence:

16 Jian-Min Chen, INSERM U1078, Faculté de médecine, Bâtiment E – 2ème étage - Bureau
17 E201b, 22 avenue Camille Desmoulins, F-29238 BREST Cedex 3, France.

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

19 20 ABSTRACT

21 GT>GC 5' splice site (or +2T>C) variants have been frequently reported to cause human
22 genetic disease. However, although we have demonstrated that GT>GC variants in human
23 disease genes may not invariably be pathogenic, none of the currently available splicing
24 prediction tools appear to be capable of reliably distinguishing those GT>GC variants that
25 generate wild-type transcripts from those that do not. Recently, SpliceAI, a novel deep
26 residual neural network tool, has been developed for splicing prediction. Methodologically
27 distinct from previous approaches that either rely on human-engineered features and/or
28 which focus on short nucleotide windows adjoining exon-intron boundaries, SpliceAI
29 assesses splicing determinants by evaluating 10,000 nucleotides of flanking contextual
30 sequence to predict the functional role in splicing of each position in the pre-mRNA
31 transcript. Herein, we evaluated the performance of SpliceAI in the context of three datasets
32 of GT>GC variants, all of which had been characterized functionally in terms of their impact
33 on mRNA splicing. The first two datasets refer to our recently described “*in vivo*” dataset of
34 45 disease-causing GT>GC variants and the “*in vitro*” dataset of 103 GT>GC substitutions.
35 The third dataset comprised 12 *BRCA1* GT>GC variants that were recently analyzed by
36 saturation genome editing. We processed all GT>GC variants using the default settings of
37 SpliceAI. Comparison of the SpliceAI-predicted and experimentally obtained functional
38 impact assessments of the analyzed GT>GC variants revealed that although SpliceAI
39 performed rather better than other prediction tools, it was still far from perfect. A key issue is
40 that the impact of GT>GC (as well as GT>GA or +2T>A) variants that generated wild-type
41 transcripts represents a quantitative change that can vary from barely detectable to almost
42 full expression of wild-type transcripts, with wild-type transcripts often co-existing with
43 aberrantly spliced transcripts. Our findings highlight the challenges that we still face in
44 attempting to accurately identify splice-altering variants.

45 46 KEYWORDS

47 full-length gene splicing assay, GT>GC variant, human genetic disease, *in silico* splicing
48 prediction, *in vitro* functional analysis, mRNA splicing, SpliceAI, 5' splice site, +2T>C variant

49 50 1. INTRODUCTION

51 Technological advances in DNA sequencing have made whole exome sequencing and even
52 whole genome sequencing increasingly practicable, especially in the clinical setting.
53 However, our ability to accurately interpret the clinical relevance of genetic variants,
54 particularly those that are rare or even private, has so far been quite limited; this represents a

55 rate-limiting step in realizing the full potential of precision medicine (Lappalainen et al., 2019;
56 Shendure et al., 2019). Functional analysis performed in a well-validated assay should
57 provide the strongest possible basis for variant classification (Richards et al., 2015; Starita et
58 al., 2017) but this is often not feasible in practice for certain types of variant. Many
59 computational algorithms have been developed with the aim of predicting the functional
60 effects of different types of genetic variant but none of them meets the exacting standards
61 required in the clinic. This is particularly true for splice-altering variants outside the obligate
62 GT and AG splice-site dinucleotides because (i) splice-altering variants can occur virtually
63 anywhere within a gene's coding or intronic sequences (Anna and Monika, 2018; Cooper et
64 al., 2009; Scotti and Swanson, 2016; Vaz-Drago et al., 2017) and (ii) splicing is a highly
65 regulated process, involving a complex interaction between *cis*-elements and *trans*-acting
66 factors (Baeza-Centurion et al., 2019; Fu and Ares, 2014; Scotti and Swanson, 2016; Shi,
67 2017; Wang and Burge, 2008).

68 Even for variants that occur within the supposedly obligate splice-site dinucleotides, we
69 may still encounter problems of interpretation. For example, variants affecting the 5' splice
70 site GT dinucleotide, which have been frequently reported to cause human genetic disease
71 (Stenson et al., 2017), are routinely scored as pathogenic splicing mutations and are usually
72 considered to be fully penetrant (Jaganathan et al., 2019; Mount et al., 2019). However, we
73 have recently provided evidence to suggest that 5' splice site GT>GC variants (henceforth
74 simply termed GT>GC variants or alternatively +2T>C variants) in human disease genes
75 may not invariably be pathogenic (Lin et al., 2019b). Specifically, combining data derived
76 from a meta-analysis of 45 human disease-causing GT>GC variants and a cell culture-based
77 Full-Length Gene Splicing Assay (FLGSA) of 103 GT>GC substitutions, we estimated that
78 ~15-18% of GT>GC variants generate between 1 and 84% wild-type transcripts (Lin et al.,
79 2019b). During this analysis, we found that none of the four most popular splicing prediction
80 tools, namely SpliceSiteFinder-like, MaxEntScan, NNSPLICE and GeneSplicer (all included
81 within Alamut® Visual; <https://www.interactive-biosoftware.com/>), were capable of reliably
82 distinguishing those GT>GC variants that generated wild-type transcripts from those that did
83 not (Lin et al., 2019b); for all variants tested, SpliceSiteFinder-like tended to predict a slightly
84 reduced score whilst the other three invariably failed to yield any score. The root of this
85 problem is twofold: Firstly, these splicing prediction tools (in common with many others)
86 focus exclusively on short local DNA sequence motifs and secondly, GC is used instead of
87 GT as the wild-type 5' splice site dinucleotide in ~1% of U2 type introns in the human
88 genome (Bursat et al., 2000; Parada et al., 2014). It follows that both GT>GC variants that
89 generate wild-type transcripts and those that do not, could in principle occur within identical
90 local sequence tracts as far as the conventional 9-bp 5' splice site consensus sequence,
91 comprising the last three bases of the preceding exon and the first six bases of the affected
92 intron (the corresponding nucleotide positions are denoted -3_-1/+1_+6), is concerned (Lin et
93 al., 2019b).

94 Recently, SpliceAI, a novel deep residual neural network tool, has been developed for
95 splicing prediction (Jaganathan et al., 2019). Methodologically distinct from previous
96 approaches that have either relied on human-engineered features and/or focused on short
97 nucleotide windows adjoining exon-intron boundaries, SpliceAI learns splicing determinants
98 directly from the primary sequence by evaluating 10,000 nucleotides of the flanking
99 sequence context to predict the role in splicing of each position in the pre-mRNA transcript.
100 Jaganathan et al. (2019) showed that SpliceAI achieved a top-*k* accuracy of 95% for pre-
101 mRNA transcripts of protein-coding genes and 84% for long intergenic noncoding RNAs
102 (lincRNAs) in the test dataset. [Top-*k* accuracy is defined as the fraction of correctly
103 predicted splice sites at the threshold where the number of predicted sites is equal to the
104 actual number of splice sites present in the test dataset] The accuracy and reliability of
105 SpliceAI was evidenced by (i) the observation that synonymous and intronic variants with
106 predicted splice-altering impact were found to be depleted in the human population, (ii) the
107 fact that 75% of these synonymous and intronic mutations were validated by RNA-seq and
108 (iii) the finding that *de novo* cryptic splice variants were enriched in patients with
109 neurodevelopmental disorders (Jaganathan et al., 2019). Herein, we sought to ascertain

110 whether SpliceAI is capable of accurately distinguishing GT>GC variants that generate wild-
111 type transcripts from those that do not.

112

113 **2. MATERIALS AND METHODS**

114 **2.1. Source of GT>GC variants**

115 Three datasets of GT>GC variants, all of which have been characterized functionally in terms
116 of their impact on splicing, were employed in this study. The first two datasets correspond to
117 our previously described “*in vivo*” dataset of 45 disease-causing GT>GC variants and the “*in*
118 *vitro*” dataset of 103 GT>GC substitutions (Lin et al., 2019b). The third dataset comprised 12
119 GT>GC variants from the *BRCA1* gene, which were extracted from a recent study that
120 prospectively analyzed the functional impact of over 4000 *BRCA1* variants by means of
121 saturation genome editing (Findlay et al., 2018).

122 In the context of the first dataset ([Supplementary Table S1](#)), the precise level of the
123 variant allele-derived wild-type transcripts was available for four of the seven disease-
124 causing GT>GC variants that generated wild-type transcripts in the corresponding original
125 publications ([Table 1](#)). For the three remaining variants (i.e., *CAV3* c.114+2T>C in (Muller et
126 al., 2006); *PLP1* c.696+2T>C in (Aoyagi et al., 1999) and *SPINK1* c.194+2T>C in (Kume et
127 al., 2006)), it is apparent from RT-PCR gel photographs in the original publications that all
128 three were associated with the generation of both wild-type and aberrantly spliced
129 transcripts. We employed ImageJ (<https://imagej.net>) to provide approximate estimates of the
130 levels of the variant allele-derived wild-transcripts for each of the three variants ([Table 1](#)).

131

132 **2.2. Variant description and nomenclature**

133 Variant description and nomenclature were in line with our previous publication (Lin et al.,
134 2019b). First, we used the term ‘variants’ to describe naturally occurring disease-causing
135 events and ‘substitutions’ to denote artificially engineered events. Second, 5’ splice site
136 GT>GC, GT>GA and GT>GG variants or substitutions were used synonymously with +2T>C,
137 +2T>A and +2T>G variants or substitutions, respectively. Third, disease-causing variants
138 were named in accordance with Human Genome Variation Society (HGVS)
139 recommendations (den Dunnen et al., 2016) whilst the traditional IVS (InterVening
140 Sequence; i.e., an intron) nomenclature was used for the engineered substitutions. Finally,
141 hg38 positions (<https://genome.ucsc.edu/>) for all variants or substitutions under study are
142 systematically provided in the various tables.

143

144 **2.3. SpliceAI prediction**

145 GT>GC variants or substitutions as well as their corresponding GT>GA and GT>GG
146 counterparts were processed (during October 2019) using the default settings of SpliceAI
147 version 1.2.1, with a custom gene annotation file containing NCBI reference sequence
148 transcript start and end coordinates. Default settings, and instruction for use of custom
149 annotation files, were taken from <https://pypi.org/project/spliceai/>.

150

151 **2.4. Performance testing**

152 Two statistical tests, a Matthews correlation coefficient (MCC) and a Receiver operating
153 characteristic (ROC) curve, were carried out on the dataset 2 substitutions assessed by
154 SpliceAI. MCC test is a correlation coefficient between the observed and predicted binary
155 classifications. For a perfect prediction, the coefficient is +1; a coefficient of 0 is no better
156 than random, and no correlation between observed and predicted yields -1 (Matthews,
157 1975). A ROC curve illustrates the diagnostic specificity and sensitivity of a binary classifier
158 system as its discrimination threshold is varied; this enables the selection of an optimum
159 threshold value. To assess the difference between the diagonal and the ROC curve
160 obtained, the area under the ROC curve is measured (AUC). An AUC of 0.5 would be a
161 random prediction whilst a perfect predictor would score 1. ROC analysis was carried out
162 using the R-based web tool easyROC (Goksuluk et al., 2016).

163 For the MCC test, a contingency table was derived from dataset 2 ([Supplementary Table](#)
164 [S2](#)) where a true positive is defined as a predicted splice altering substitution for which

165 FLGSA produced no transcript and a true negative is a substitution not predicted to alter
166 splicing and for which FLGSA produces transcript.

167

168 **2.5. Functional analysis of two GT-affecting variants**

169 The functional impact of two newly engineered GT-affecting variants in the *HESX1* gene
170 were analyzed by means of the cell culture-based FLGSA method as previously described
171 (Lin et al., 2019b).

172

173 **3. RESULTS AND DISCUSSION**

174 **3.1. Accuracy and reliability of the experimentally obtained functional assessment of 175 the GC>GT variants analyzed**

176 We employed SpliceAI to make predictions as to the splicing consequences of GT>GC
177 variants from three distinct datasets. Since the experimentally ascertained functional impact
178 of the GT>GC variants analyzed was used as the starting point for our analysis, their
179 accuracy and reliability were of critical importance. Regarding the first dataset of known
180 pathogenic variants ([Supplementary Table S1](#)), several points are worth highlighting. First, all
181 45 disease-causing variants were either homozygotes, hemizygotes or compound
182 heterozygotes, a prerequisite for determining the presence or absence of the variant allele-
183 derived normal transcripts. Second, for each variant, patient-derived tissue or cells
184 (pathologically relevant in about half of the cases) had been used to perform the RT-PCR
185 analysis that had unequivocally demonstrated the presence or absence of variant allele-
186 derived wild-type transcripts in the corresponding original publication. Third, the levels of the
187 variant allele-derived wild-type transcripts in the seven disease-causing GT>GC variants that
188 generated wild-type transcripts were very low ($\leq 15\%$ of normal; [Table 1](#)), potentially
189 explicable by the ascertainment bias inherent to all disease-causing variants. Nonetheless,
190 all seven of these variants were noted to be associated with a milder clinical phenotype than
191 would have been expected from a functionally null variant (Lin et al., 2019b), consistent with
192 other findings that even the retention of a small fraction of normal gene function can
193 significantly impact the clinical phenotype (Den Uijl et al., 2011; Ramalho et al., 2002;
194 Raraigh et al., 2018; Scalet et al., 2019).

195 In the case of the second dataset ([Supplementary Table S2](#)), the functional effects of all
196 103 engineered GT>GC substitutions (from 30 different genes) were analyzed by Full-Length
197 Gene Splicing Assay (FLGSA) in transfected HEK293T cells (Lin et al., 2019b), with all 19
198 substitutions that generated some wild-type transcripts being listed in [Table 2](#). By
199 comparison to the commonly used minigene splicing assay, FLGSA preserves better the
200 natural genomic sequence context of the studied variants (Wu et al., 2017; Zou et al., 2016).
201 The accuracy and reliability of the FLGSA-derived data can be inferred from the following
202 three lines of evidence. First, 10 GT>GC substitutions that generated wild-type transcripts
203 and 10 GT>GC substitutions that did not generate wild-type transcripts in transfected
204 HEK293T cells were further analyzed in transfected HeLa cells using FLGSA, yielding
205 entirely consistent findings in terms of whether or not wild-type transcripts were generated
206 (Lin et al., 2019b). Second, *HESX1* c.357+2T>C and *SPINK1* c.194+2T>C were the only
207 variants common to both the first and second datasets. The functional effects of these two
208 variants *in vivo* — *HESX1* c.357+2T>C generated no wild-type transcripts whereas *SPINK1*
209 c.194+2T>C generated some wild-type transcripts ([Supplementary Table S1](#)) — were
210 faithfully replicated in FLGSA ([Supplementary Table S2](#)). Third, a GT>GC variant that was
211 not present in either dataset, *HBB* c.315+2T>C, had been reported to be associated with a
212 milder hematological phenotype and it was suggested that it might have a limited impact on
213 splicing (Frischknecht et al., 2009). Using FLGSA performed in HEK293T cells, we found that
214 it generated a low level of wild-type transcripts (Lin et al., 2019b). Importantly, the
215 orthologous variant of *HBB* c.315+2T>C in the rabbit *Hbb* gene has also been found to be
216 capable of generating wild-type transcripts in two experimental model systems, namely *in*
217 *vitro* transcription of *Hbb* RNA in a HeLa cell nuclear extract and transient expression of the
218 full-length *Hbb* gene in HeLa cells (Aebi et al., 1986; Aebi et al., 1987). These
219 notwithstanding, tissue- or cell-specific factors have on some occasions impacted splicing

220 (Jaganathan et al., 2019; Pineda and Bradley, 2018), an issue that was not extensively
221 addressed in our previous study (Lin et al., 2019b). The bottom line here is that (i) the 30
222 genes used for FLGSA analysis were selected using a procedure that did not take into
223 consideration the gene's function or expression, (ii) all 30 genes underwent normal splicing in
224 the context of their reference mRNA sequences as specified in [Supplementary Table S2](#) and
225 (iii) the generation (or not) of wild-type transcripts from the engineered GC allele was
226 observed under the same experimental conditions as for the wild-type GT allele (Lin et al.,
227 2019b).

228 The third dataset was obtained courtesy of a perusal of the literature ([Table 3](#)). Recently,
229 the functional impact of all possible single nucleotide substitutions within 13 exons and
230 adjacent intronic sequences of the 23-exon *BRCA1* gene (NM_007294.3) have been
231 prospectively analyzed by means of saturation genome editing (Findlay et al., 2018). Taking
232 advantage of the essentiality of *BRCA1* in the human near-haploid cell line HAP1 (Blomen et
233 al., 2015), Findlay and colleagues used cell viability as a proxy indicator for the functional
234 consequences of the analyzed substitutions. It should be noted that the functional
235 consequences of all tested substitutions were actually evaluated in their natural genomic
236 sequence contexts. Of the ~4000 *BRCA1* single nucleotide substitutions analyzed, 12 were
237 GT>GC substitutions. Of these 12 GT>GC substitutions, one was classified as "functional",
238 two were classified as "intermediate" and the remaining nine were classified as "non-
239 functional" ([Table 3](#)). Whereas "functional" and "intermediate" were interpreted as having
240 generated wild-type transcripts, "non-functional" was interpreted as having not generated any
241 wild-type transcripts (Lin et al., 2019a). As such, 25% (n = 3) of these 12 *BRCA1* GT>GC
242 substitutions generated wild-type transcripts, a proportion largely consistent with our
243 estimated 15-18% rate. Moreover, the *BRCA1* GT>GC variant in intron 18 was shown to be
244 "functional", providing further support for our contention that GT>GC variants in human
245 disease genes may not invariably be pathogenic (Lin et al., 2019b).

246 Taken together, the experimentally obtained functional assessments of the included
247 GC>GT variants or substitutions were considered to be of high quality and appropriate for the
248 intended study.

249

250 **3.2. Selection and interpretation of SpliceAI Delta scores for analysis**

251 We processed GT>GC variants using the default settings of SpliceAI as detailed in
252 <https://pypi.org/project/spliceai/>. SpliceAI provides Delta scores (ranging from 0 to 1) for each
253 variant, thereby providing a measure of their probability of altering splicing in terms of either
254 splice donor gain, splice donor loss, splice acceptor gain, and splice acceptor loss. SpliceAI
255 also provides Delta position that conveys information specifying the location where splicing
256 differs from normal relative to the position of the associated variant. Since the GT>GC
257 variants or substitutions under study invariably affected the +2 position of the canonical 5'
258 splice site GT dinucleotides (in the context of the specified mRNA reference sequence), we
259 focused our analysis exclusively on the Delta scores of donor loss although other scores may
260 provide clues as to the nature of the resulting aberrantly spliced transcripts of splice-altering
261 variants. Thus, only the SpliceAI-predicted Delta scores of donor loss for the studied GT>GC
262 variants or substitutions are provided in [Supplementary Tables S1 and S2](#) as well as in
263 [Tables 1-3](#). Here, it is important to note two points. First, the previously studied GT>GC
264 events generated maximally 84% wild-type transcripts as compared to their wild-type GT
265 allele counterparts (Lin et al., 2019b). In other words, all these variants were associated
266 minimally with a 16% functional loss. Therefore, strictly speaking, all these previously studied
267 GT>GC events can be defined as splice-altering. Second, in those cases of GT>GC events
268 that generated wild-type transcripts, the level of wild-type transcript varied from 1-84% (Lin et
269 al., 2019b). Intuitively, whether or not a GT>GC variant capable of generating wild-type
270 transcripts is pathogenic is likely to depend at least in part upon the level of the generated
271 wild-type transcripts. Taking these points into consideration, we shall use the SpliceAI Delta
272 score of donor loss as a proxy indicator of the probability of a given GT>GC variant being
273 able to generate wild-type transcripts; variants with a Delta score above a certain cutoff value
274 will be considered not to be capable of generating wild-type transcripts whereas variants with

275 a Delta score below the cutoff value will be considered as being capable of generating wild-
276 type transcripts.

277

278 **3.3. Encouraging findings from a quick survey of the three datasets of GT>GC variants**

279 As mentioned in the Introduction, none of the four most popular splicing prediction tools,
280 SpliceSiteFinder-like, MaxEntScan, NNSPLICE and GeneSplicer, were found to be able to
281 distinguish those GT>GC variants that generated wild-type transcripts from those that did not
282 (Lin et al., 2019b). As described below, a quick survey of SpliceAI-predicted scores yielded
283 encouraging results across all three datasets of GT>GC variants.

284 First, in the context of dataset 1, the level of variant allele-derived wild-type transcripts
285 associated with the seven disease-causing GT>GC variants was at most 15% of normal
286 (Table 1). Although this low level increase in the generation of wild-type transcripts may
287 make prediction a daunting task, it is interesting to see that the two lowest Delta scores of
288 donor loss, 0.35 and 0.63, were observed in association with the two variants that generated
289 ~10% wild-type transcripts (Supplementary Table S1; Table 1). The score of 0.35 was
290 observed for the *SPINK1* c.194+2T>C variant, for which the RT-PCR analysis was performed
291 using gastric tissue from a homozygous patient with chronic pancreatitis (Kume et al., 2006).
292 Although stomach is not known to be clinically affected in chronic pancreatitis, the expression
293 data were considered to be highly reliable for two reasons. Firstly, the *in vivo* expression data
294 was confirmed by FLGSA performed in both HEK293T and HeLa cells (Lin et al., 2019b).
295 Secondly, had the *SPINK1* c.194+2T>C variant in question caused a complete functional
296 loss of the affected allele, the homozygotes should have developed severe infantile isolated
297 exocrine pancreatic insufficiency instead of chronic pancreatitis (Venet et al., 2017). The
298 score of 0.63 was observed for the *DMD* c.8027+2T>C variant, for which the detection of
299 wild-type transcripts was based upon RT-PCR analysis of disease-affected muscle tissue
300 from a hemizygous carrier with Becker muscular dystrophy (Bartolo et al., 1996).

301 As for the second dataset (Supplementary Table S2), the four lowest Delta scores of
302 donor loss (i.e., 0, 0.03, 0.05 and 0.08) were all found in substitutions that generated wild-
303 type transcripts; and 63% (n = 12) of the 19 substitutions that generated some wild-type
304 transcripts had a Delta score of <0.80 (Table 2). As for the third dataset, the lowest Delta
305 score, 0.53, was observed in association with the only “functional” *BRCA1* IVS18+2T>C
306 variant (Table 3).

307

308 **3.4. Statistical comparison of experimentally obtained functional data with SpliceAI** 309 **predictions for the 103 engineered GT>GC splice variants (dataset 2)**

310 Dataset 2 comprised 19 substitutions that generated wild-type transcripts and 84
311 substitutions that generated no wild-type transcripts. We thus performed two statistical tests,
312 a Receiver operating characteristic (ROC) curve and a Matthews correlation coefficient
313 (MCC), on the 103 substitutions assessed by SpliceAI (Supplementary Table S2) with a view
314 both to identifying an optimum threshold value and to assessing the correlation between the
315 FLGSA assay results and the SpliceAI predictions.

316 Based on an ROC analysis of 103 variants from dataset 2 (Supplementary Table S2), an
317 optimum threshold point of 0.85 was provided - similar to the threshold of 0.80,
318 recommended by SpliceAI for high precision results. A contingency table was constructed
319 (Supplementary Table S3) to calculate values for the false positive rate, specificity,
320 sensitivity, accuracy and the Matthews correlation coefficient. These are summarized in
321 Table 4, along with the AUC result obtained from the ROC analysis, the curve from which is
322 shown in Fig. 1.

323 As can be seen from Table 4, the AUC of 0.79 and the MCC score of 0.41 are indicative
324 of a good correlation between predicted and actual results. There is also a low false positive
325 rate whilst still maintaining a high accuracy and sensitivity. These results show that for
326 dataset 2, at a threshold of 0.85, SpliceAI can accurately discriminate between those GT>GC
327 substitutions which disrupt splicing and transcript production and those which do not disrupt
328 splicing and produce transcript.

329

330 **3.5. Considerable discrepancy between the predicted and experimentally obtained** 331 **functional impact assessments of GT>GC 5' splice site variants**

332 Employing 0.85 as the threshold Delta score (donor loss) to define the generation of wild-
333 type transcripts, rather variable performance between SpliceAI-predicted and experimentally
334 demonstrated functional effects of GT>GC variants were observed across the three datasets:
335 33-84% of the variants that generated wild-type transcripts and 67-89% of the variants that
336 generated no wild-type transcripts were correctly predicted by SpliceAI (Table 5).

337 The poorest performance (43% (3/7) and 33% (1/3)) was observed with datasets 1 and 3
338 variants that generated wild-type transcripts (Table 5). In the context of the seven dataset 1
339 variants that generated wild-type transcripts (a qualitative property), the relatively poor
340 performance of 43% might be related to the fact that the functional impact of these GT>GC
341 variants actually manifested as rather small quantitative changes, generating between 1-15%
342 normal transcripts (Table 1). This notwithstanding, it should be pointed out that the two
343 disease-causing variants that generated 10-15% wild-type transcripts, *CAV3* c.114+2T>C
344 (Muller et al., 2006) and *CD40LG* c.346+2T>C (Seyama et al., 1998), had Delta scores of
345 ≥ 0.9 (Table 1); and in each of these two cases, RT-PCR analysis was performed using
346 patient-derived and pathologically relevant tissue or cells. In short, it remains unclear why
347 some of the disease-causing variants that generated comparable levels of wild-type
348 transcripts were predicted to have low Delta scores (i.e., *DMD* c.8027+2T>C and *SPINK1*
349 c.194+2T>C) whereas others were predicted to have high Delta scores (i.e., *CAV3*
350 c.114+2T>C and *CD40LG* c.346+2T>C). In the context of dataset 3 substitutions that
351 generated wild-type transcripts, the precise levels of wild-type transcripts generated by the
352 two "intermediate" *BRCA1* +2T>C substitutions (both had a Delta score of ≥ 0.93 ; Table 3)
353 were unknown.

354 As for variants that did not generate wild-type transcripts, an excellent correlation rate,
355 89%, was observed with the 38 such disease-causing variants. By contrast, the performance
356 in datasets 2 and 3 variants was much lower and almost identical (68% and 67%,
357 respectively; Table 5). A fundamental difference between dataset 1 variants and the latter
358 two dataset substitutions is that all of the former were previously published whilst almost all
359 of the latter were prospectively generated. Thus, it is tempting to speculate that for most of
360 the 38 disease-causing variants that did not generate wild-type transcripts, their functional
361 effects might have been 'seen' by SpliceAI during training, thereby leading to a higher
362 correlation rate.

363 In an attempt to further understand the poor performance of dataset 2 and 3 substitutions
364 that did not generate wild-type transcripts, we opted to use the corresponding +2T>A and
365 +2T>G substitutions as controls. The underlying premise was that, based upon current
366 knowledge, +2T>A and +2T>G variants should completely disrupt normal splicing and
367 consequently have high Delta scores in virtually all cases (see also section 3.6). Here it is
368 worth mentioning that we previously employed FLGSA to analyze the functional impact of 15
369 +2T>A substitutions and 18 +2T>G substitutions, none of which generated any wild-type
370 transcripts (Lin et al., 2019b). We processed all corresponding +2T>A and +2T>G variants
371 by means of SpliceAI in the same way as for the +2T>C variants (during October 2019), the
372 resulting Delta scores for donor loss being provided in Tables 1-3 and Supplementary Tables
373 S1 and S2.

374 As shown in Supplementary Table S1, all +2T>A and +2T>G variants corresponding to
375 the 45 disease-causing +2T>C variants had very high Delta scores, ranging from 0.92 to 1.
376 By contrast, 91% (n = 94) of the +2T>A and +2T>G variants corresponding to the 103
377 dataset 2 +2T>C substitutions had a Delta score of ≥ 0.85 (Supplementary Table S2). In other
378 words, nine of the 103 +2T sites were predicted to have a Delta score of < 0.85 when
379 substituted by either A or G; and in these sites, the Delta scores are often identical for all
380 three possible substitutions (Table 6). One possible reason for lower than expected Delta
381 scores is provided in (Jaganathan et al., 2019); exons which undergo a substantial degree of
382 alternative splicing, defined as being between 10% and 90% exon inclusion averaged across
383 samples, tend towards intermediate scores (stated as between 0.35 and 0.8). We therefore
384 explored this possibility using the two sites for which all possible substitutions had the lowest

385 Delta scores (i.e., 0.59 and 0.3; [Table 6](#)) as examples. To this end, alternative transcripts of
386 the genes of interest were surveyed via <https://www.ncbi.nlm.nih.gov/gene/>.

387 All three possible single nucleotide substitutions in the *RPL11* g.23695910T (IVS5+2T in
388 accordance with NM_000975.5) site had an identical Delta score of 0.59. *RPL11* has two
389 transcripts, the other being NM_001199802.1. Nonetheless, the two transcripts have
390 common coding sequences from exons 3-6. Moreover, all three possible single nucleotide
391 substitutions in the NM_000975.5-defined *RPL11* IVS5+2T site have been previously
392 subjected to FLGSA, invariably generating no wild-type transcripts (Lin et al., 2019b). Taken
393 together, in this particular case, the lower than expected Delta scores cannot be adequately
394 explained by alternative splicing.

395 All three possible single nucleotide substitutions in the *LY6G6F* g.31708136T (IVS5+2T in
396 accordance with NM_001003693.1) site had a score of 0.3 ([Table 6](#)). NM_001003693.1-
397 defined *LY6G6F* has sequence from exons 1 to 4 in common with NM_001353334.2-defined
398 *LY6G6F-LY6G6D*, which represents naturally occurring readthrough transcription between
399 the neighboring *LY6G6F* and *LY6G6D* genes on chromosome 6 ([Supplementary Fig. S1](#)). By
400 contrast, NM_001003693.1-defined exons 5 and 6 are spliced out in NM_001353334.2-
401 defined *LY6G6F-LY6G6D*. It is likely that the use of the “*LY6G6F* IVS5+2T site” as a splice
402 site in one transcript isoform but not in the other underlies the similarly low Delta scores for
403 the three above mentioned possible single nucleotide substitutions. However, two points
404 should be emphasized here. Firstly, none of the three possible single nucleotide substitutions
405 in the context of the NM_001003693.1-defined *LY6G6F* IVS5+2T site led to the generation of
406 wild-type transcripts as evidenced by FLGSA. Whether these substitutions would lead to the
407 increased use of NM_001353334.2-defined *LY6G6F-LY6G6D* remains unclear. Secondly, all
408 three possible single nucleotide substitutions, if considered only in the context of
409 NM_001353334.2-defined *LY6G6F-LY6G6D* ([Supplementary Fig. S1](#)), may not affect gene
410 splicing at all.

411 Finally, let us turn our attention to the *BRCA1* findings in relation to NM_007294.3 ([Table](#)
412 [3](#)). The lowest Delta score of donor loss in the context of +2T>A and +2T>G variants, 0.65,
413 was observed for all three possible SNVs in the *BRCA1* IVS4+2T site. The next lowest score,
414 0.67, was observed for all three possible SNVs in the *BRCA1* IVS5+2T site ([Table 3](#)). All six
415 of these variants have been analyzed using saturation genome editing and were invariably
416 classified as “non-functional” (Jaganathan et al., 2019). Moreover, although *BRCA1* has
417 multiple transcripts, these two introns are used by all transcripts ([Supplementary Fig. S2](#)).
418 Therefore, as in the abovementioned *RPL11* case, these lower than expected Delta scores
419 cannot be adequately explained by alternative splicing.

420

421 **3.6. Additional findings**

422 We succeeded in analyzing two additional engineered GT-impacting substitutions in the
423 *HESX1* gene, IVS2+2T>A (hg38# chr3:57198751A>T) and IVS3+2T>G (hg38#
424 chr3:57198389A>C), using the cell culture-based FLGSA method. Interestingly, the
425 IVS2+2T>A substitution generated both wild-type and aberrant transcripts whereas
426 IVS3+2T>G generated only aberrant transcripts ([Fig. 2](#)). Moreover, two of the 12 *BRCA1*
427 +2T>A substitutions, IVS16+2T>A and IVS18+2T>A, were described as being “intermediate”
428 ([Table 3](#)). Although no disease-causing +2T>A variants have been found to generate wild-
429 type transcripts, GA has recently been ranked fourth in terms of its relative frequency among
430 the six noncanonical 5' splice sites identified by genome-wide RNA-seq analysis and splicing
431 reporter assays (Erkelenz et al., 2018). However, of the three +2T>A substitutions that were
432 experimentally shown to generate some wild-type transcripts, two were predicted to have a
433 Delta score of >0.85, namely 0.93 for *HESX1* IVS2+2T>A ([Supplementary Table S2](#)) and
434 0.98 for *BRCA1* IVS18+2T>A ([Table 3](#)). The other one, *BRCA1* IVS16+2T>A, was predicted
435 to have a Delta score of 0.74; but an identical score was also predicted for *BRCA1*
436 IVS16+2T>C and IVS16+2T>G, both of which were classified as “non-functional” ([Table 3](#)).
437 In short, SpliceAI appeared not to work as well for the +2T>A variants that generated wild-
438 type transcripts as for the +2T>C variants that generated wild-type transcripts.

439

440 **4. CONCLUSIONS AND PERSPECTIVES**

441 In the present study, we attempted to correlate SpliceAI-predicted and experimentally
442 obtained functional effects of GT>GC variants in the context of three independent and
443 complementary datasets. Employing data from dataset 2 substitutions, we were able to
444 propose a Delta score of donor loss, 0.85, as defining the threshold of whether or not wild-
445 type transcripts would be generated by GT>GC variants; whereas a score of ≥ 0.85 defines
446 the complete absence of wild-type transcripts, a score of < 0.85 defines the generation of at
447 least some wild-type transcripts. Subsequent use of this threshold score to correlate
448 SpliceAI-predicted and experimentally obtained functional effects of the GT>GC variants
449 revealed that SpliceAI performed better than the popular prediction tools such as
450 SpliceSiteFinder-like, MaxEntScan, NNSPLICE and GeneSplicer. However, a considerable
451 discrepancy still existed between SpliceAI-predicted and experimentally obtained functional
452 assessments in relation to GT>GC (as well as GT>GA) variants. Indeed, this discrepancy
453 serves to illuminate the challenges ahead in accurately identifying all splice-altering variants.
454 A key issue here is that the impact of GT>GC (as well as GT>GA) variants that generated
455 wild-type transcripts represents a quantitative change that can vary from barely detectable to
456 almost full expression of wild-type transcripts, with wild-type transcripts often co-existing with
457 aberrantly spliced transcripts. This is also the case for most of the splice-altering variants
458 occurring outside the essential splice site dinucleotides, whose effects “are not fully
459 penetrant and a mixture of both normal and aberrant splice isoforms are produced”
460 (Jaganathan et al., 2019). Moreover, there is also the issue of alternative splicing related to
461 tissue- or cell-specific factors. While it is clear that we are still very far acquiring a full
462 understanding of the ‘splicing code’ (Bao et al., 2019), we are of the opinion that any
463 improvement in the prioritization of splicing variants will necessitate the refinement of *in silico*
464 prediction tools by reference to *in vitro* functional assessment courtesy of the results
465 obtained from well-validated assays such as FLGSA.

467 **ACKNOWLEDGMENTS**

468 We are grateful to the original authors who reported the disease-causing 5' splice site
469 GT>GC variants studied here. J.H.L. was in receipt of a 20-month scholarship from the China
470 Scholarship Council (No. 201706580018). This study was supported by the Institut National
471 de la Santé et de la Recherche Médicale (INSERM), France. D.N.C. and M.H. acknowledge
472 the financial support of Qiagen plc through a License Agreement with Cardiff University.

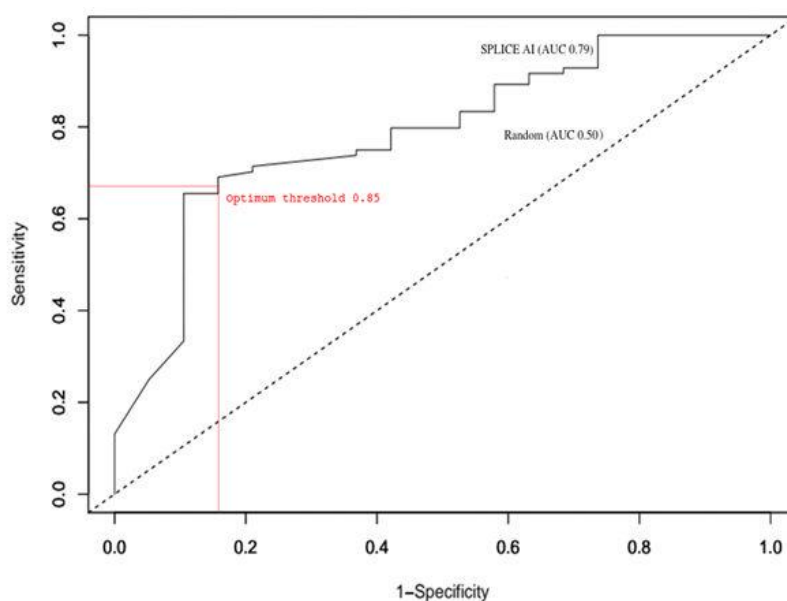
474 **REFERENCES**

- 475 Aebi M, Hornig H, Padgett RA, Reiser J, Weissmann C. Sequence requirements for splicing
476 of higher eukaryotic nuclear pre-mRNA. *Cell* 1986; **47**: 555-565.
- 477 Aebi M, Hornig H, Weissmann C. 5' cleavage site in eukaryotic pre-mRNA splicing is
478 determined by the overall 5' splice region, not by the conserved 5' GU. *Cell* 1987; **50**:
479 237-246.
- 480 Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and
481 confirmation. *J Appl Genet* 2018; **59**: 253-268.
- 482 Aoyagi Y, Kobayashi H, Tanaka K, Ozawa T, Nitta H, Tsuji S. A de novo splice donor site
483 mutation causes in-frame deletion of 14 amino acids in the proteolipid protein in
484 Pelizaeus-Merzbacher disease. *Ann Neurol* 1999; **46**: 112-115.
- 485 Baeza-Centurion P, Minana B, Schmiedel JM, Valcarcel J, Lehner B. Combinatorial genetics
486 reveals a scaling law for the effects of mutations on splicing. *Cell* 2019; **176**: 549-563
487 e523.
- 488 Bao S, Moakley DF, Zhang C. The splicing code goes deep. *Cell* 2019; **176**: 414-416.
- 489 Bartolo C, Papp AC, Snyder PJ, Sedra MS, Burghes AH, Hall CD, Mendell JR, Prior TW. A
490 novel splice site mutation in a Becker muscular dystrophy patient. *J Med Genet* 1996;
491 **33**: 324-327.
- 492 Blomen VA, Majek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, Sacco R, van
493 Diemen FR, Olk N, Stukalov A, Marceau C, Janssen H, Carette JE, Bennett KL, Colinge

- 494 J, Superti-Furga G, Brummelkamp TR. Gene essentiality and synthetic lethality in
495 haploid human cells. *Science* 2015; **350**: 1092-1096.
- 496 Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in
497 mammalian genomes. *Nucleic Acids Res* 2000; **28**: 4364-4375.
- 498 Cooper TA, Wan L, Dreyfuss G. RNA and disease. *Cell* 2009; **136**: 777-793.
- 499 den Dunnen JT, Dagleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J,
500 Roux AF, Smith T, Antonarakis SE, Taschner PE. HGVS Recommendations for the
501 Description of Sequence Variants: 2016 Update. *Hum Mutat* 2016; **37**: 564-569.
- 502 Den Uijl IE, Mauser Bunschoten EP, Roosendaal G, Schutgens RE, Biesma DH, Grobbee
503 DE, Fischer K. Clinical severity of haemophilia A: does the classification of the 1950s still
504 stand? *Haemophilia* 2011; **17**: 849-853.
- 505 Erkelenz S, Theiss S, Kaisers W, Ptok J, Walotka L, Muller L, Hillebrand F, Brillen AL,
506 Sladek M, Schaal H. Ranking noncanonical 5' splice site usage by genome-wide RNA-
507 seq analysis and splicing reporter assays. *Genome Res* 2018; **28**: 1826-1840.
- 508 Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, Janizek JD, Huang X,
509 Starita LM, Shendure J. Accurate classification of *BRCA1* variants with saturation
510 genome editing. *Nature* 2018; **562**: 217-222.
- 511 Frischknecht H, Dutly F, Walker L, Nakamura-Garrett LM, Eng B, Waye JS. Three new beta-
512 thalassemia mutations with varying degrees of severity. *Hemoglobin* 2009; **33**: 220-225.
- 513 Fu XD, Ares M, Jr. Context-dependent control of alternative splicing by RNA-binding
514 proteins. *Nat Rev Genet* 2014; **15**: 689-701.
- 515 Goksuluk D, Korkmaz S, Zararsiz G, Karaagaoglu AE. easyROC: an interactive web-tool for
516 ROC curve analysis using R language environment. *The R Journal* 2016; **8**: 213-230.
- 517 Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li
518 YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, Chow ED, Kanterakis E, Gao H, Kia
519 A, Batzoglou S, Sanders SJ, Farh KK. Predicting splicing from primary sequence with
520 deep learning. *Cell* 2019; **176**: 535-548 e524.
- 521 Kume K, Masamune A, Kikuta K, Shimosegawa T. [-215G>A; IVS3+2T>C] mutation in the
522 *SPINK1* gene causes exon 3 skipping and loss of the trypsin binding site. *Gut* 2006; **55**:
523 1214.
- 524 Lappalainen T, Scott AJ, Brandt M, Hall IM. Genomic analysis in the age of human genome
525 sequencing. *Cell* 2019; **177**: 70-84.
- 526 Lin JH, Masson E, Boulling A, Hayden M, Cooper DN, Férec C, Liao Z, Chen JM. 5' splice
527 site GC>GT variants differ from GT>GC variants in terms of their functionality and
528 pathogenicity. bioRxiv 829010; doi: <https://doi.org/10.1101/829010>. 2019a.
- 529 Lin JH, Tang XY, Boulling A, Zou WB, Masson E, Fichou Y, Raud L, Le Tertre M, Deng SJ,
530 Berlivet I, Ka C, Mort M, Hayden M, Lemar R, Houdayer C, Le Gac G, Cooper DN, Li
531 ZS, Férec C, Liao Z, Chen JM. First estimate of the scale of canonical 5' splice site
532 GT>GC variants capable of generating wild-type transcripts. *Hum Mutat* 2019b; **40**:
533 1856-1873.
- 534 Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage
535 lysozyme. *Biochim Biophys Acta* 1975; **405**: 442-451.
- 536 Mount SM, Avsec Z, Carmel L, Casadio R, Celik MH, Chen K, Cheng J, Cohen NE,
537 Fairbrother WG, Fenesh T, Gagneur J, Gotea V, Holzer T, Lin CF, Martelli PL, Naito T,
538 Nguyen TYD, Savojardo C, Unger R, Wang R, Yang Y, Zhao H. Assessing predictions of
539 the impact of variants on splicing in CAGI5. *Hum Mutat* 2019; **40**: 1215-1224.
- 540 Muller JS, Piko H, Schoser BG, Schlotter-Weigel B, Reilich P, Gurster S, Born C, Karcagi V,
541 Pongratz D, Lochmuller H, Walter MC. Novel splice site mutation in the caveolin-3 gene
542 leading to autosomal recessive limb girdle muscular dystrophy. *Neuromuscul Disord*
543 2006; **16**: 432-436.
- 544 Parada GE, Munita R, Cerda CA, Gysling K. A comprehensive survey of non-canonical
545 splice sites in the human transcriptome. *Nucleic Acids Res* 2014; **42**: 10564-10578.
- 546 Pineda JMB, Bradley RK. Most human introns are recognized via multiple and tissue-specific
547 branchpoints. *Genes Dev* 2018; **32**: 577-591.

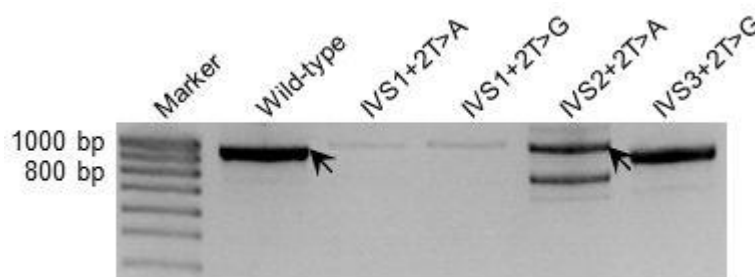
- 548 Ramalho AS, Beck S, Meyer M, Penque D, Cutting GR, Amaral MD. Five percent of normal
549 cystic fibrosis transmembrane conductance regulator mRNA ameliorates the severity of
550 pulmonary disease in cystic fibrosis. *Am J Respir Cell Mol Biol* 2002; **27**: 619-627.
- 551 Raraigh KS, Han ST, Davis E, Evans TA, Pellicore MJ, McCague AF, Joynt AT, Lu Z, Atalar
552 M, Sharma N, Sheridan MB, Sosnay PR, Cutting GR. Functional assays are essential for
553 interpretation of missense variants associated with variable expressivity. *Am J Hum
554 Genet* 2018; **102**: 1062-1077.
- 555 Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E,
556 Spector E, Voelkerding K, Rehm HL, Committee ALQA. Standards and guidelines for the
557 interpretation of sequence variants: a joint consensus recommendation of the American
558 College of Medical Genetics and Genomics and the Association for Molecular Pathology.
559 *Genet Med* 2015; **17**: 405-424.
- 560 Scalet D, Maestri I, Branchini A, Bernardi F, Pinotti M, Balestra D. Disease-causing variants
561 of the conserved +2T of 5' splice sites can be rescued by engineered U1snRNAs. *Hum
562 Mutat* 2019; **40**: 48-52.
- 563 Scotti MM, Swanson MS. RNA mis-splicing in disease. *Nat Rev Genet* 2016; **17**: 19-32.
- 564 Seyama K, Nonoyama S, Gangsaas I, Hollenbaugh D, Pabst HF, Aruffo A, Ochs HD.
565 Mutations of the CD40 ligand gene and its effect on CD40 ligand expression in patients
566 with X-linked hyper IgM syndrome. *Blood* 1998; **92**: 2421-2434.
- 567 Shendure J, Findlay GM, Snyder MW. Genomic medicine-progress, pitfalls, and promise.
568 *Cell* 2019; **177**: 45-57.
- 569 Shi Y. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nat
570 Rev Mol Cell Biol* 2017; **18**: 655-670.
- 571 Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, Shendure J, Fowler DM.
572 Variant interpretation: functional assays to the rescue. *Am J Hum Genet* 2017; **101**: 315-
573 325.
- 574 Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD,
575 Cooper DN. The Human Gene Mutation Database: towards a comprehensive repository
576 of inherited mutation data for medical research, genetic diagnosis and next-generation
577 sequencing studies. *Hum Genet* 2017; **136**: 665-677.
- 578 Vaz-Drago R, Custodio N, Carmo-Fonseca M. Deep intronic mutations and human disease.
579 *Hum Genet* 2017; **136**: 1093-1111.
- 580 Venet T, Masson E, Talbotec C, Billiemaz K, Touraine R, Gay C, Destombe S, Cooper DN,
581 Patural H, Chen JM, Férec C. Severe infantile isolated exocrine pancreatic insufficiency
582 caused by the complete functional loss of the *SPINK1* gene. *Hum Mutat* 2017; **38**: 1660-
583 1665.
- 584 Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an
585 integrated splicing code. *RNA* 2008; **14**: 802-813.
- 586 Wu H, Boulling A, Cooper DN, Li ZS, Liao Z, Chen JM, Férec C. *In vitro* and *in silico*
587 evidence against a significant effect of the *SPINK1* c.194G>A variant on pre-mRNA
588 splicing. *Gut* 2017; **66**: 2195-2196.
- 589 Zou WB, Boulling A, Masson E, Cooper DN, Liao Z, Li ZS, Férec C, Chen JM. Clarifying the
590 clinical relevance of *SPINK1* intronic variants in chronic pancreatitis. *Gut* 2016; **65**: 884-
591 886.

592



593
594 **Figure 1.** A receiver operating characteristic (ROC) curve for the SpliceAI predictions
595 generated from dataset 2 (Supplementary Table S2), with dotted diagonal line indicating a
596 random prediction (0.5 AUC) and the solid line showing SpliceAI prediction performance (0.79
597 AUC). The intersection between the two represents the optimum threshold.

598
599
600
601
602
603
604
605
606



607
608 **Figure 2.** RT-PCR analyses of HEK293T cells transfected with full-length *HESX1* gene
609 expression constructs carrying respectively the wild-type and four different nucleotide
610 substitutions. Wild-type transcripts emanating from the wild-type construct and the construct
611 containing the IVS2+2T>A substitution (confirmed by DNA sequencing) are indicated by
612 arrows. IVS2+2T>A (hg38# chr3:57198751A>T) and IVS3+2T>G (hg38# chr3:57198389A>C)
613 substitutions were newly analyzed here. IVS1+2T>A and IVS1+2T>G, which had been
614 previously analyzed (Lin et al., 2019b), are included for the sake of comparison.

615 **Table 1.** Comparison of SpliceAI-predicted and experimentally demonstrated functional effects of the seven disease-causing GT>GC (+2T>C)
 616 variants that generated wild-type transcripts

Gene symbol	mRNA reference	Chromosome	HG38 coordinate	Reference sequence	Variant ^a	% normal expression level ^b	SpliceAI Delta score of donor loss		
							+2T>C	+2T>A	+2T>G
<i>CAV3</i>	NM_001234.4	3	8733992	T	c.114+2T>C	11 ^c	0.9	1	1
<i>CD3E</i>	NM_000733.3	11	118313876	T	c.520+2T>C	1-5 ^d	0.99	0.99	0.99
<i>CD40LG</i>	NM_000074.2	X	136654432	T	c.346+2T>C	15 ^d	0.95	0.97	0.97
<i>DMD</i>	NM_004006.2	X	31657988	A	c.8027+2T>C	10 ^d	0.63	0.99	0.99
<i>PLP1</i>	NM_000533.4	X	103788512	T	c.696+2T>C	8 ^c	0.74	1	1
<i>SLC26A2</i>	NM_000112.3	5	149960981	T	c.-26+2T>C	5 ^d	0.9	0.99	0.99
<i>SPINK1</i>	NM_003122.3	5	147828020	A	c.194+2T>C ^e	10 ^c	0.35	0.99	1

617 ^aNomenclature in accordance with Human Genome Variation Society (HGVS) recommendations (den Dunnen et al., 2016).

618 ^bExpressed as the level of the variant allele-derived wild-type transcripts relative to that of the wild-type allele-derived wild-type transcripts.

619 ^cExpression level determined here by ImageJ using gel photos from the original publications.

620 ^dExpression level as described in the original publications.

621 ^eIdentical to the *SPINK1* IVS3+2T>C substitution in Table 2.

622

623 **Table 2.** Comparison of SpliceAI-predicted and experimentally demonstrated functional effects of the 19 engineered GT>GC (+2T>C)
 624 substitutions that generated wild-type transcripts

Gene symbol	mRNA reference	Chromosome	hg38 coordinate	Reference sequence	Substitution ^a	Generation of wild-type transcripts ^b	SpliceAI Delta score of donor loss		
							+2T>C	+2T>A	+2T>G
<i>CCDC103</i>	NM_213607.2	17	44899861	T	IVS1+2T>C	Yes	0.82	0.82	0.82
<i>DBI</i>	NM_001079862.2	2	119368307	T	IVS2+2T>C	Yes	0.86	1	1
<i>DNAJC19</i>	NM_145261.3	3	180985924	A	IVS5+2T>C	Yes (42%)	0.03	0.99	0.95
<i>FATE1</i>	NM_033085.2	X	151716227	T	IVS1+2T>C	Yes (84%)	0.08	0.96	1
<i>FOLR3</i>	NM_000804.3	11	72139484	T	IVS4+2T>C	Yes	0.45	1	1
<i>HESX1</i>	NM_003865.2	3	57199760	A	IVS1+2T>C	Yes (2%)	0.81	0.98	0.98
<i>IFNL2</i>	NM_172138.1	19	39269823	T	IVS5+2T>C	Yes (5%)	0.05	0.84	0.73
<i>IL10</i>	NM_000572.3	1	206770905	A	IVS3+2T>C	Yes	0.61	1	1
<i>MGP</i>	NM_000900.4	12	14884211	A	IVS2+2T>C	Yes (80%)	0.97	0.99	0.99
<i>PSMC5</i>	NM_001199163.1	17	63830503	T	IVS6+2T>C	Yes (56%)	0.31	0.98	1
			63831228	T	IVS8+2T>C	Yes (56%)	0.21	1	1
			63831618	T	IVS10+2T>C	Yes (46%)	0.83	1	1
<i>RPL11</i>	NM_000975.5	1	23692761	T	IVS2+2T>C	Yes	0	0.87	0.86
			23693915	T	IVS3+2T>C	Yes	0.74	1	1
<i>RPS27</i>	NM_001030.4	1	153991225	T	IVS2+2T>C	Yes (63%)	0.67	1	1
			153991678	T	IVS3+2T>C	Yes	0.98	1	1
<i>SELENOS</i>	NM_203472.2	15	101277340	A	IVS1+2T>C	Yes	0.81	1	1
			101274418	A	IVS5+2T>C	Yes (14%)	0.79	1	1
<i>SPINK1</i>	NM_003122.3	5	147828020	A	IVS3+2T>C ^c	Yes	0.35	0.99	1

625 ^aIn accordance with the traditional IVS (InterVening Sequence; i.e., an intron) nomenclature as previously described (Lin et al. 2019b).

626 ^bExpression level (in parentheses), determined by quantitative RT-PCR analysis, was available for all +2T>C substitutions that generated only
 627 wild-type transcripts under the experimental conditions described in (Lin et al. 2019b).

628 ^cIdentical to the *SPINK1* c.194+2T>C variant in Supplementary Table S1 and Table 1.

629

630

631

632 **Table 3.** Comparison of SpliceAI-predicted and experimentally demonstrated functional effects of all possible single nucleotide substitutions in
 633 the +2 positions of 12 *BRCA1* introns*

Intron ^a	HG38 coordinate	Reference sequence	+2T>C		+2T>A		+2T>G	
			Functional classification ^b	Delta score (donor loss)	Functional classification	Delta score (donor loss)	Functional classification	Delta score (donor loss)
2	43124015	A	Non-functional	0.9	Non-functional	0.9	Non-functional	0.9
3	43115724	A	Non-functional	0.97	Non-functional	0.98	Non-functional	0.98
4	43106454	A	Non-functional	0.65	Non-functional	0.65	Non-functional	0.65
5	43104866	A	Non-functional	0.67	Non-functional	0.67	Non-functional	0.67
15	43070926	A	Non-functional	0.99	Non-functional	0.99	Non-functional	0.99
16	43067606	A	Non-functional	0.74	Intermediate	0.74	Non-functional	0.74
17	43063872	A	Non-functional	0.9	Non-functional	0.9	Non-functional	0.9
18	43063331	A	Functional	0.53	Intermediate	0.98	Non-functional	0.98
19	43057050	A	Non-functional	0.82	Non-functional	1	Non-functional	1
20	43051061	A	Non-functional	0.9	Non-functional	0.99	Non-functional	0.99
21	43049119	A	Intermediate	0.96	Non-functional	0.99	Non-functional	0.99
22	43047641	A	Intermediate	0.93	Non-functional	0.93	Missing data	0.93

634 *Experimental data were extracted from Findlay et al. (2018).

635 ^aIn accordance with NM_007294.3.

636 ^b“Non-functional” was interpreted as meaning that no wild-type transcripts were generated whereas “functional” and “intermediate” were held to
 637 imply the generation of wild-type transcripts.

638 **Table 4.** Performance metrics of SpliceAI as a predictor for splice site disruption on 103
639 variants from dataset 2

False positive rate	True positive rate (sensitivity)	True negative rate (specificity)	Accuracy	AUC	MCC
16%	67%	84%	70%	0.79	0.41

640

641

642

643

644

645

646

647

648

649

650 **Table 5.** Overall correlation rates between SpliceAI-predicted and experimentally
651 demonstrated functional effects of the GT>GC variants in the context of three datasets*

Variants generating wild-type transcripts	
Dataset 1 (45 disease-causing variants)	43% (3/7)
Dataset 2 (103 variants analyzed by FLGSA)	84% (16/19)
Dataset 3 (12 <i>BRCA1</i> variants analyzed by saturation genome editing)	33% (1/3)
Variants generating no wild-type transcripts	
Dataset 1 (45 disease-causing variants)	89% (34/38)
Dataset 2 (103 variants analyzed by FLGSA)	68% (57/84)
Dataset 3 (12 <i>BRCA1</i> variants analyzed by saturation genome editing)	67% (6/9)

652 *Splice AI Delta score (donor loss) of 0.85 was used as the threshold value for defining the
653 generation of wild-type transcripts or not.

Table 6. Nine +2T positions for which all three possible nucleotide substitutions had a consistent SpliceAI Delta score of <0.85

Gene symbol	mRNA reference	Chromosome	hg38 coordinate	Reference sequence	SpliceAI Delta score of donor loss		
					+2T>C	+2T>A	+2T>G
<i>AURKC</i>	NM_001015878.1	19	57235060	T	0.8	0.8	0.8
<i>CCDC103</i>	NM_213607.2	17	44899861	T	0.82	0.82	0.82
<i>FABP7</i>	NM_001446.4	6	122779869	T	0.83	0.84	0.84
<i>IFNL2</i>	NM_172138.1	19	39269823	T	0.05	0.84	0.73
<i>LY6G6F</i>	NM_001003693.1	6	31708136	T	0.81	0.81	0.81
			31710420	T	0.3	0.3	0.3
<i>PSMC5</i>	NM_001199163.1	17	63830191	T	0.76	0.77	0.77
<i>RPL11</i>	NM_000975.5	1	23695910	T	0.59	0.59	0.59
<i>SELENOS</i>	NM_203472.2	15	101272762	A	0.64	0.64	0.64