Supplementary Material

The Experimentally Obtained Functional Impact Assessments of 5' Splice Site GT>GC Variants Differ Markedly from Those Predicted

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

[1]EFS, Univ Brest, Inserm, UMR 1078, GGB, F-29200 Brest, France; [2]Department of Gastroenterology, Changhai Hospital, Second Military Medical University, Shanghai, China; [3]Shanghai Institute of Pancreatic Diseases, Shanghai, China; [4]CHRU Brest, Service de Génétique Médicale et de Biologie de la Reproduction, Brest, France; [5]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, UK

Supplementary Table S1. Comparison of SpliceAI-predicted and experimentally demonstrated functional effects of the 45 disease-causing GT>GC (+2T>C) variants*

Gene symbol	mRNA reference	Chr.	HG38 coordinate	Reference sequence	Variant	Generation of wild-type transcripts[a]	SpliceAI Delta score of donor loss		
							+2T>C	+2T>A	+2T>G
ABCC2	NM_000392.5	chr10	99811604	T	c.1967+2T>C	No	0.97	0.98	0.98
ACAT1	NM_000019.3	chr11	108146361	T	c.1163+2T>C	No	0.99	1	1
COQ8A	NM_020247.4	chr1	226984237	T	c.1398+2T>C	No	0.99	1	1
ALB	NM_000477.6	chr4	73417671	T	c.1428+2T>C	No	0.95	1	1
AR	NM_000044.4	chrX	67711691	T	c.2173+2T>C	No	0.99	0.99	0.99
ART4	NM_021071.3	chr12	14842968	A	c.144+2T>C	No	0.97	0.99	0.99
ATP7A	NM_000052.6	chrX	78011254	T	c.1946+2T>C	No	0.96	0.96	0.96
B9D1	NM_015681.4	chr17	19347782	A	c.341+2T>C	No	0.99	0.99	0.99
BTK	NM_000061.2	chrX	101362171	A	c.588+2T>C	No	0.99	0.99	0.99
C8orf37	NM_177965.3	chr8	95269033	A	c.155+2T>C	No	0.92	0.92	0.92
CAV3	NM_001234.4	chr3	8733992	T	c.114+2T>C	Yes (11%)[b]	0.9	1	1
CD3E	NM_000733.3	chr11	118313876	T	c.520+2T>C	Yes (1-5%)[c]	0.99	0.99	0.99
CD40LG	NM_000074.2	chrX	136654432	T	c.346+2T>C	Yes (15%)[c]	0.95	0.97	0.97
CLCN5	NM_000084.4	chrX	50072590	T	c.205+2T>C	No	0.92	0.92	0.92
COL1A2	NM_000089.3	chr7	94426532	T	c.3105+2T>C	No	0.99	1	1
DGAT1	NM_012079.5	chr8	144318093	A	c.751+2T>C	No	0.99	0.99	0.99
DMD	NM_004006.2	chrX	31657988	A	c.8027+2T>C	Yes (10%)[c]	0.63	0.99	0.99
DMD	NM_004006.2	chrX	31206580	A	c.9649+2T>C	No	0.99	0.99	0.99
DNAL4	NM_005740.2	chr22	38780924	A	c.153+2T>C	No	0.98	0.99	0.99
GAA	NM_000152.4	chr17	80117111	T	c.2331+2T>C	No	0.94	0.97	0.97
HESX1	NM_003865.2	chr3	57198751	A	c.357+2T>C[d]	No	0.92	0.93	0.93
IL10RA	NM_001558.3	chr11	117994151	T	c.688+2T>C	No	0.96	0.97	0.97
ITGA8	NM_003638.2	chr10	15531048	A	c.2982+2T>C	No	0.95	0.99	0.99
IVD	NM_002225.3	chr15	40410799	T	c.456+2T>C	No	0.71	1	1
JAK3	NM_000215.3	chr19	17834569	A	c.2350+2T>C	No	0.71	0.97	0.99
LAMA2	NM_000426.3	chr6	129315952	T	c.3924+2T>C	No	0.97	0.97	0.97

LMAN1	NM_005570.3	chr18	59338758	A	c.1149+2T>C	No	0.97	0.97	0.97
MCM9	NM_017696.2	chr6	118827925	A	c.1732+2T>C	No	0.95	0.95	0.95
MMP2	NM_004530.5	chr16	55485429	T	c.658+2T>C	No	0.9	1	1
NCAPD2	NM_014865.3	chr12	6531078	T	c.4120+2T>C	No	0.88	1	1
NCF2	NM_001127651.2	chr1	183586893	A	c.257+2T>C	No	0.99	0.99	0.99
OPHN1	NM_002547.2	chrX	68432865	A	c.154+2T>C	No	0.81	1	1
OTC	NM_000531.5	chrX	38401430	T	c.540+2T>C	No	0.98	1	1
PCCB	NM_000532.4	chr3	136250560	T	c.183+2T>C	No	0.99	0.99	0.99
PCCB	NM_000532.4	chr3	136328859	T	c.1498+2T>C	No	1	1	1
PEX16	NM_004813.2	chr11	45910896	A	c.952+2T>C	No	0.97	0.99	0.99
PLP1	NM_000533.4	chrX	103787968	T	c.622+2T>C	No	0.95	0.96	0.96
PLP1	NM_000533.4	chrX	103788512	T	c.696+2T>C	Yes (8%)[b]	0.74	1	1
PNPLA2	NM_020376.3	chr11	823589	T	c.757+2T>C	No	0.95	0.99	0.99
PRDM5	NM_018699.3	chr4	120922514	A	c.93+2T>C	No	0.73	1	0.99
RFX6	NM_173560.3	chr6	116877954	T	c.380+2T>C	No	0.9	0.9	0.9
SCLT1	NM_144643.3	chr4	129039039	A	c.290+2T>C	No	0.95	0.95	0.95
SH2D1A	NM_002351.4	chrX	124346781	T	c.137+2T>C	No	0.96	0.98	0.98
SLC26A2	NM_000112.3	chr5	149960981	T	c.-26+2T>C	Yes (5%)[c]	0.9	0.99	0.99
SPINK1	NM_003122.3	chr5	147828020	A	c.194+2T>C[e]	Yes (10%)[b]	0.35	0.99	1

*Refer to Lin et al. [16] for the original publications that reported the 45 variants.

[a] Relative expression level is indicated in parentheses.

[b] Expression level determined here by ImageJ using gel photos from the original publications.

[c] Expression level as described in the original publications.

[d] Identical to the *HESX1* IVS2+2T>C substitution in Supplementary Table S2.

[e] Identical to the *SPINK1* IVS3+2T>C substitution in Supplementary Table S2 and Table 2.

Supplementary Table S2. Comparison of SpliceAI-predicted and experimentally demonstrated functional effects of the 103 engineered GT>GC (+2T>C) substitutions

Gene symbol	mRNA reference	Chr.	hg38 coordinate	Reference sequence	Variant[a]	Generation of wild-type transcripts[b]	SpliceAI Delta score of donor loss		
							+2T>C	+2T>A	+2T>G
ACTA1	NM_001100.3	chr1	229434002	A	IVS1+2T>C	No	0.84	0.99	0.99
			229432985	A	IVS2+2T>C	No	0.87	0.87	0.87
			229432554	A	IVS3+2T>C	No	0.26	1	0.96
			229432268	A	IVS4+2T>C	No	0.34	0.98	0.98
			229431719	A	IVS6+2T>C	No	0.28	1	1
AURKC	NM_001015878.1	chr19	57231789	T	IVS2+2T>C	No	0.92	0.96	0.96
			57232226	T	IVS3+2T>C	No	0.99	1	1
			57232682	T	IVS4+2T>C	No	0.99	0.99	0.99
			57233610	T	IVS5+2T>C	No	0.95	0.95	0.95
			57235060	T	IVS6+2T>C	No	0.8	0.8	0.8
CCDC103	NM_213607.2	chr17	44899861	T	IVS1+2T>C	Yes	0.82	0.82	0.82
			44901654	T	IVS3+2T>C	No	0.99	0.99	0.99
COX4I1	NM_001861.4	chr16	85799754	T	IVS1+2T>C	No	0.95	0.95	0.95
			85801280	T	IVS2+2T>C	No	0.92	0.93	0.93
			85805106	T	IVS3+2T>C	No	0.96	0.99	0.99
			85805866	T	IVS4+2T>C	No	0.82	1	1
COX7A1	NM_001864.3	chr19	36152391	A	IVS1+2T>C	No	0.95	0.96	0.96
			36151667	A	IVS2+2T>C	No	0.87	0.87	0.87
DBI	NM_001079862.2	chr2	119367062	T	IVS1+2T>C	No	0.99	0.99	0.99
			119368307	T	IVS2+2T>C	Yes	0.86	1	1
			119370804	T	IVS3+2T>C	No	0.87	1	1
DDIT3	NM_004083.5	chr12	57517704	A	IVS2+2T>C	No	0.98	1	1
			57517267	A	IVS3+2T>C	No	0.46	1	1
DNAJC19	NM_145261.3	chr3	180988176	A	IVS2+2T>C	No	0.94	0.95	0.95

			180985924	A	IVS5+2T>C	Yes (42%)	0.03	0.99	0.95
FABP1	NM_001443.2	chr2	88127949	A	IVS1+2T>C	No	0.96	0.99	0.99
			88126174	A	IVS2+2T>C	No	0.97	0.98	0.98
			88124492	A	IVS3+2T>C	No	0.98	0.98	0.98
FABP7	NM_001446.4	chr6	122779869	T	IVS1+2T>C	No	0.83	0.84	0.84
			122780465	T	IVS2+2T>C	No	0.64	1	1
			122781196	T	IVS3+2T>C	No	0.98	0.99	0.99
FATE1	NM_033085.2	chrX	151716227	T	IVS1+2T>C	Yes (84%)	0.08	0.96	1
			151717401	T	IVS2+2T>C	No	0.98	0.99	0.99
			151721503	T	IVS3+2T>C	No	0.99	0.99	0.99
			151721983	T	IVS4+2T>C	No	0.92	1	1
FOLR3	NM_000804.3	chr11	72136122	T	IVS2+2T>C	No	0.96	1	1
			72139151	T	IVS3+2T>C	No	0.97	0.98	0.98
			72139484	T	IVS4+2T>C	Yes	0.45	1	1
HESX1	NM_003865.2	chr3	57199760	A	IVS1+2T>C	Yes (2%)	0.81	0.98	0.98
			57198751	A	IVS2+2T>C[c]	No	0.92	0.93	0.93
			57198389	A	IVS3+2T>C	No	0.96	1	1
HLA-DRA	NM_019111.4	chr6	32442695	T	IVS2+2T>C	No	0.99	0.99	0.99
			32443468	T	IVS3+2T>C	No	0.98	0.99	0.99
			32443923	T	IVS4+2T>C	No	0.93	0.99	0.99
IFI30	NM_006332.4	chr19	18175224	T	IVS2+2T>C	No	0.25	1	0.99
			18177294	T	IVS5+2T>C	No	0.98	1	1
			18177766	T	IVS6+2T>C	No	0.62	1	1
IFNL2	NM_172138.1	chr19	39268580	T	IVS1+2T>C	No	0.78	0.98	0.98
			39268860	T	IVS2+2T>C	No	0.88	0.97	0.96
			39269231	T	IVS3+2T>C	No	0.58	1	1
			39269639	T	IVS4+2T>C	No	0.99	0.99	0.99
			39269823	T	IVS5+2T>C	Yes (5%)	0.05	0.84	0.73
IL10	NM_000572.3	chr1	206772269	A	IVS1+2T>C	No	0.56	0.99	0.96
			206771354	A	IVS2+2T>C	No	0.23	0.99	1
			206770905	A	IVS3+2T>C	Yes	0.61	1	1
			206769827	A	IVS4+2T>C	No	0.44	1	1
LENG1	NM_024316.2	chr19	54159562	A	IVS1+2T>C	No	0.22	0.99	1
			54158280	A	IVS2+2T>C	No	0.89	1	1
			54156761	A	IVS3+2T>C	No	0.87	0.87	0.87
LY6G6F	NM_001003693.1	chr6	31707789	T	IVS2+2T>C	No	0.60	0.98	0.98
			31708136	T	IVS3+2T>C	No	0.81	0.81	0.81
			31710183	T	IVS4+2T>C	No	0.76	0.9	0.9
			31710420	T	IVS5+2T>C	No	0.3	0.3	0.3
MGP	NM_000900.4	chr12	14885729	A	IVS1+2T>C	No	0.98	0.99	0.99
			14884211	A	IVS2+2T>C	Yes (80%)	0.97	0.99	0.99
PIMREG	NM_001195228.1	chr17	6444490	T	IVS1+2T>C	No	0.97	0.99	0.99
			6445406	T	IVS2+2T>C	No	0.96	0.96	0.96
			6449409	T	IVS4+2T>C	No	0.99	1	1
PRSS2	NM_002770.3	chr7	142771024	T	IVS1+2T>C	No	0.81	0.99	0.98
			142772210	T	IVS2+2T>C	No	0.98	0.98	0.98
			142773521	T	IVS3+2T>C	No	0.75	1	0.99
			142774057	T	IVS4+2T>C	No	0.97	0.97	0.97
PSMC5	NM_001199163.1	chr17	63827925	T	IVS1+2T>C	No	0.93	0.98	0.98
			63830191	T	IVS5+2T>C	No	0.76	0.77	0.77
			63830503	T	IVS6+2T>C	Yes (56%)	0.31	0.98	1
			63830937	T	IVS7+2T>C	No	0.96	0.96	0.96
			63831228	T	IVS8+2T>C	Yes (56%)	0.21	1	1
			63831427	T	IVS9+2T>C	No	0.95	0.99	0.99
			63831618	T	IVS10+2T>C	Yes (46%)	0.83	1	1
RPL11	NM_000975.5	chr1	23691831	T	IVS1+2T>C	No	0.85	0.99	0.99
			23692761	T	IVS2+2T>C	Yes	0	0.87	0.86
			23693915	T	IVS3+2T>C	Yes	0.74	1	1

			23694793	T	IVS4+2T>C	No	0.94	0.94	0.94
			23695910	T	IVS5+2T>C	No	0.59	0.59	0.59
RPS17	NM_001021.4	chr15	82540424	A	IVS1+2T>C	No	0.95	0.98	0.98
			82539979	A	IVS2+2T>C	No	0.51	0.99	0.99
			82538878	A	IVS3+2T>C	No	0.99	1	1
RPS20	NM_001023.3	chr8	56074379	A	IVS1+2T>C	No	0.98	0.99	0.99
			56074058	A	IVS2+2T>C	No	0.95	0.99	0.99
			56073693	A	IVS3+2T>C	No	0.99	1	1
RPS27	NM_001030.4	chr1	153990804	T	IVS1+2T>C	No	0.93	0.94	0.94
			153991225	T	IVS2+2T>C	Yes (63%)	0.67	1	1
			153991678	T	IVS3+2T>C	Yes	0.98	1	1
SELENOS	NM_203472.2	chr15	101277340	A	IVS1+2T>C	Yes	0.81	1	1
			101276539	A	IVS2+2T>C	No	0.96	1	1
			101274590	A	IVS4+2T>C	No	0.97	0.98	0.98
			101274418	A	IVS5+2T>C	Yes (14%)	0.79	1	1
			101272762	A	IVS6+2T>C	No	0.64	0.64	0.64
SPINK1	NM_003122.3	chr5	147831521	A	IVS1+2T>C	No	0.85	0.86	0.86
			147829597	A	IVS2+2T>C	No	0.97	0.97	0.97
			147828020	A	IVS3+2T>C[d]	Yes	0.35	0.99	1
UQCRB	NM_006294.4	chr8	96235510	A	IVS1+2T>C	No	0.97	0.99	0.99
			96231772	A	IVS3+2T>C	No	1	1	1

[a] In accordance with the traditional IVS (InterVening Sequence; i.e., an intron) nomenclature as previously described [16].

[b] Expression level (in parentheses), determined by quantitative RT-PCR analysis, was available for all +2T>C substitutions that generated only wild-type transcripts under the experimental conditions described in Lin et al. [16].

[c] Identical to the *HESX1* c.357+2T>C variant in Supplementary Table S1.

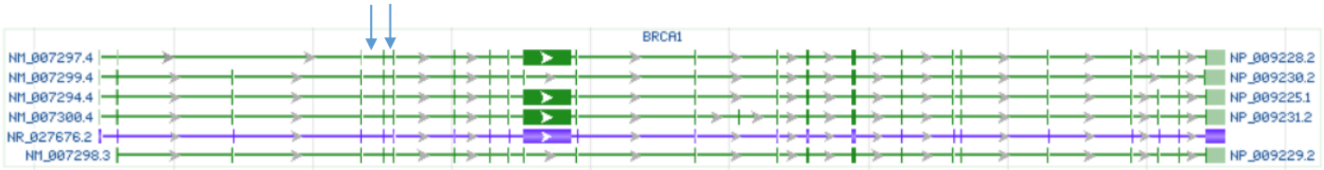
[d] Identical to the *SPINK1* c.194+2T>C variant in Supplementary Table S1 and Table 1.

Supplementary Table S3. Confusion matrix using a threshold value of 0.85 derived from Supplementary Table S2

True positive (TP): SpliceAI identifies a splicing event (>0.85) which produces no transcript	**False positive (FP):** SpliceAI identifies a splicing event (>0.85) which does produce transcript
False negative (FN): SpliceAI does not identify a splicing event (<0.85) which does not produce transcript	**True negative (TN):** Splice AI does not identify a splicing event (<0.85) which does produce transcript

TP:57	FP:3
FN:27	TN:16

Supplementary Figure S1. Transcript isoforms of the *LY6G6F* gene. The approximate position of the discussed +2T site is indicated by arrow.



Supplementary Figure S2. Transcript isoforms of the *BRCA1* gene. The two discussed introns are indicated by arrows.