Compensatory signals associated with the activation of human GC 5′ splice sites

Jana Kralovicova[1], Gyulin Hwang[1], A. Charlotta Asplund[2], Alexander Churbanov[3], C. I. Edvard Smith[2] and Igor Vorechovsky[1,*]

[1]University of Southampton School of Medicine, Division of Human Genetics, Southampton SO16 6YD, UK, [2]Department of Laboratory Medicine, Karolinska Institutet, Clinical Research Center, Karolinska University Hospital Huddinge, SE-141 86 Huddinge, Sweden and [3]Biology Department, New Mexico State University, Las Cruces, NM 88003, USA

Received February 3, 2011; Revised April 3, 2011; Accepted April 17, 2011

ABSTRACT

GC 5′ splice sites (5′ss) are present in ~1% of human introns, but factors promoting their efficient selection are poorly understood. Here, we describe a case of X-linked agammaglobulinemia resulting from a GC 5′ss activated by a mutation in *BTK* intron 3. This GC 5′ss was intrinsically weak, yet it was selected in >90% primary transcripts in the presence of a strong and intact natural GT counterpart. We show that efficient selection of this GC 5′ss required a high density of GAA/CAA-containing splicing enhancers in the exonized segment and was promoted by SR proteins 9G8, Tra2β and SC35. The GC 5′ss was efficiently inhibited by splice-switching oligonucleotides targeting either the GC 5′ss itself or the enhancer. Comprehensive analysis of natural GC-AG introns and previously reported pathogenic GC 5′ss showed that their efficient activation was facilitated by higher densities of splicing enhancers and lower densities of silencers than their GT 5′ss equivalents. Removal of the GC-AG introns was promoted to a minor extent by the splice-site strength of adjacent exons and inhibited by flanking *Alu* repeats, with the first downstream *Alu*s located on average at a longer distance from the GC 5′ss than other transposable elements. These results provide new insights into the splicing code that governs selection of noncanonical splice sites.

INTRODUCTION

Eukaryotic genes contain introns that are removed by splicing of precursor messenger RNAs (pre-mRNAs) to ensure accurate and efficient gene expression. This process is accomplished by the spliceosome, a large and highly dynamic RNA–protein complex that is assembled in a step-wise manner on each intron (1). Intron recognition is mediated by splicing signals on the pre-mRNA that include 5′ and 3′ splices sites (ss), branch point sequence (BPS), polypyrimidine tract and auxiliary sequences that promote (enhancers) or inhibit (silencers) splicing and participate in ligand binding and/or structural interactions (2–4). Unlike in lower eukaryotes, splicing recognition sequences in vertebrates are highly degenerate, particularly in mammals and especially in primates (2,5,6). This highlights the importance of the auxiliary signals for accurate intron removal and exon ligation.

The human 5′ss consensus sequence (MAG/GURAGU, M is A or C; R is G or A) spans from position −3 to position +6 relative to the exon–intron junction (/) and base-pairs with the U1 small nuclear RNA (snRNA) (7). This interaction is important for the earliest step of the spliceosome assembly, but may not be necessary in some cases (8). For example, U1 snRNA-depleted *in vitro* splicing reactions could be reconstituted with other factors, such as serine/arginine-rich (SR) proteins (9), which facilitate exon inclusion by participating in cross-exon interactions and play a more general role in coupled gene expression pathways, including transcription and translation (10,11).

Human GC 5′ss are present in ~1% of authentic introns (12–14). They accumulated extensively during mammalian evolution (15) and are particularly frequent in alternatively spliced introns (~1 in 20 human introns) as opposed to those spliced constitutionally (~1 in 200) (12). GC 5′ ss are also enriched among aberrant 5′ ss activated by pathogenic mutations relative to their GT counterparts (16). Alternative 5′ GCs are most commonly found 4 nt downstream or upstream from predominant 5′ss (17). This has been attributed to the U1 snRNP binding at the 5′ss, which frequently contains a GU dinucleotide 4 nt downstream from the dominant 5′ss (17). The GC 5′ss are

[*]To whom correspondence should be addressed. Tel: +44 2380 796425; Fax: +44 2380 794264; Email: igvo@soton.ac.uk

intrinsically weaker than their GT counterparts because the T>C substitution at intron position +2 introduces a mismatch in the U1 snRNA:5′ss pre-mRNA helix, but the rest of the 5′ss consensus is on average stronger, both in humans and in lower organisms, apparently compensating for the +2 T>C transition (13,14,18–22). Canonical GT dinucleotides at the exon–intron junction can be replaced by GC without loss of cleavage accuracy, but splicing of 5′GC mutants is considerably slower (23). Although both GC-AG and GT-AG introns are recognized by the major U2 spliceosome (23), auxiliary factors that facilitate accurate selection of weak GC 5′ss as opposed to strong GT 5′ss are poorly understood.

In this work, we report an efficient activation of weak GC 5′ss that was induced by a point mutation causing X-linked agammaglobulinemia (XLA), the first described immunodeficiency (24). We identify both *cis*- and *trans*-acting factors promoting or repressing the use of this aberrant splice site. Compilation of GC-AG introns and pathogenic 5′GC splice sites allowed us to systematically characterize auxiliary splicing motifs and elements that contribute to accurate selection of such weak pre-mRNA signals.

MATERIAL AND METHODS

BTK mutation screening

Whole blood samples of clinically diagnosed XLA patients were collected in Paxgene™ Blood RNA tubes (PreAnalytiX GmbH), which permits their storage at room temperature for up to 72 h without any degradation of RNA. RNA was extracted by using Paxgene™ Blood RNA Kit (PreAnalytiX GmbH) according to the manufacturer's protocol. Synthesis of complementary DNA (cDNA) was performed using 500 ng of total RNA, random hexamers and the First-strand cDNA synthesis kit (Roche Applied Science) according to the manufacturer's recommendations. Aberrant transcripts were visualized using reverse transcription (RT)-polymerase chain reaction (PCR) with primers C (5′-ccg gat cca tgg ccg cag tga ttc tgg a) and D (5′-gat act gcc cat cga tcc ag). Direct sequencing of aberrant transcripts and DNA was carried out with the Big Dye Terminator Cycle Sequencing kit (Applied Biosystems).

Splicing reporter constructs

The wild-type (WT) *BTK* reporter construct was cloned in NheI/EcoRI sites of pCR3.1 (Invitrogen) using primers A (*BTK*-*Nhe*I; 5′-gat c*gc tag* cac aca ggt gaa ctc cag a) and B (*BTK*-*Eco*RI; 5′-gat c*ga att* cct gga agg gat aag gga ac). The plasmids were propagated in the *Escherichia coli* strain DH5α (Invitrogen). Plasmid DNA samples were extracted as described (25). Splicing reporters with the disease-causing T>G mutation was prepared using overlap-extension PCR with primer M (5′-tgg aac acg ggc aag ttt cct t). Deletion mutants were prepared using primers *BTK*del-F (5′-caa ttt cag tag cat agc tac cta act cct) and *BTK*del-R (5′-ggt agc tat gct act gaa att gat ata tat). Mutations introduced in putative enhancers within the newly exonized segment are shown in Figure 1. All constructs were fully sequenced using vector primers PL3 and PL4 (26) and an internal primer S (5′-aat gtc tga gat ggg gaa c) to confirm the intended mutations and exclude undesired changes.

For preparation of constructs capable of splicing in nuclear extracts, we amplified a 0.3-kb segment containing *BTK* exon 3 and adjacent exonized intronic sequences with primers *BTK*-PY7-F-*Nhe*I (5′-ata *ggc tag c*ag aga aga ggc agt aag aag) and *BTK*-PY7-R-XhoI (5′-ata g*ct cga g*at atg aag gaa aga agc taa) and cloned into NheI/XhoI sites of pCR3.1 (Invitrogen). The XhoI/XbaI fragment of an *in vitro* splicing-proficient construct PY7 (27) was cloned downstream of *BTK* exon 3. Before crosslinking experiments, the hybrid *BTK*-PY7 construct was transfected into human embryonal kidney 293 T and HeLa cells to confirm the GC 5′ss activation.

Cell cultures and transfections

A pre-B cell line Nalm-6, 293 T and HeLa cells were cultured as described previously (25,28). Transient transfections were carried out using Fugene HD (Roche) according to manufacturer's recommendations. Cells were harvested 48 h post-transfection.

For RNA interference (RNAi)-mediated depletion, 293 T cells were subjected to a two-hit transfection with custom-made small interfering RNA (siRNA) duplexes (MWG Biotech) on Days 2 and 3, reporters were added on Day 4 and cells were harvested 24 h later, essentially as described (6). Briefly, mammalian cells were plated in 6- or 12-well plates to achieve ~40% confluence. The next day, HiPerFect (Qiagen) was combined with Opti-MEM medium (Invitrogen) and siRNAs (MWG Biotech). Before adding to cells, the mixture was left at room temperature for 20 min. Sequences of siRNAs targeting SR proteins, hnRNP I (PTB/nPTB) and scrambled controls were as we described (6). Sequences of hnRNP A1/A2 siRNAs were also published previously (29).

Visualization of RNA products

Total RNA was extracted with TRI Reagent (Ambion) and reverse transcribed using oligo-d(T)$_{15}$ primer, as previously described (25). Vector-specific primers PL3 and PL4 (26) were used for PCR. PCR products were separated on agarose and polyacrylamide gels and their signal intensity was measured as described (30).

Antisense oligonucleotides

Splice-switching oligonucleotides (MWG Biotech) were co-transfected into 293 T cells together with the WT and mutated *BTK* reporter constructs. Cells were harvested 48 h post-transfection. Antisense oligonucleotides (ASOs) targeted the GC 5′ss (*BTK*-5′GC, 5′-mG*mG*mA*mA*mA*mC*mU*mU*mG*mC*mC*mC*mG*mU*mG*mU*mU*mC*mC*mA) or splicing enhancers between the authentic and cryptic 5′ss (*BTK*-ESE-M2M3, 5′-mG*mU*mU*mG*mC*mA*mA*mA*mU*mU*mU*mC*mU*mU*mC *mA*mA*mA*mU*mC). Negative controls included scrambled oligonucleotides (*BTK*-5′GC-sc, 5′-mU*mC*mG*mA*mU*mG*mC*mA*mU*

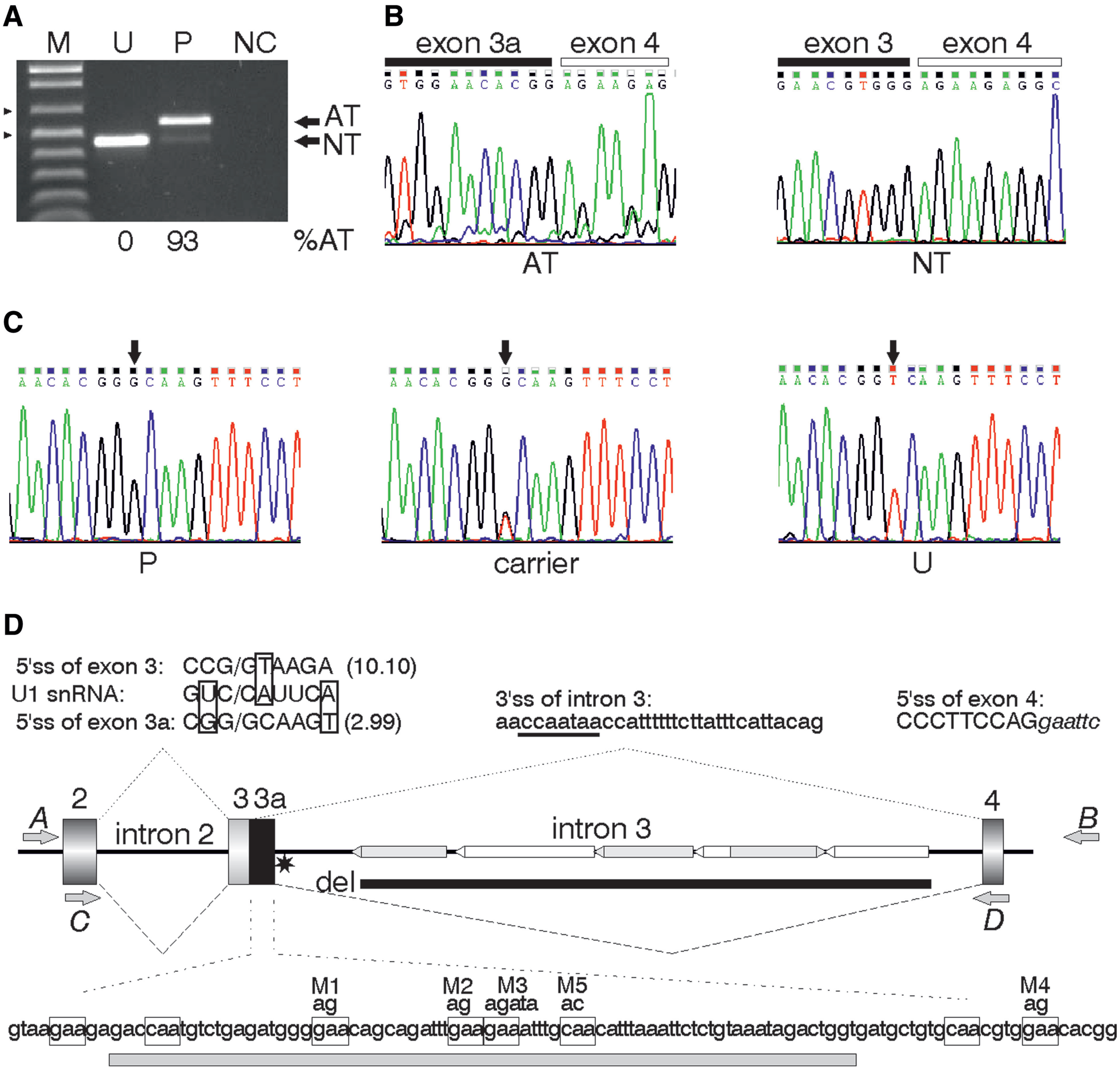
A
M
U
P
NC
AT
NT
0
93
%AT
B
exon 3a
exon 4
G T G G A A C A C G G A G A A G A G
AT
exon 3
exon 4
G A A C G T G G G A G A A G A G G C
NT
C
A A C A C G G G C A A G T T T C C T
P
A A C A C G G G C A A G T T T C C T
carrier
A A C A C G G T C A A G T T T C C T
U
D
5'ss of exon 3: CCG/GTAAGA (10.10)
U1 snRNA: GUC/CAUUCA
5'ss of exon 3a: CGG/GCAAGT (2.99)
3'ss of intron 3:
aaccaataaccatttttcttatttcattacag
5'ss of exon 4:
CCCTTCCAGgaattc
A
2
intron 2
3 3a
intron 3
4
B
C
del
D
M1
ag
M2
ag
M3
agata
M5
ac
M4
ag
gtaagaagagaccaatgtctgagatggggaacagcagatttgaagaaatttgcaacatttaaattctctgtaaatagactggtgatgctgtgcaacgtggaacacgg

Figure 1. Identification of aberrant GC 5′ss in the *BTK* gene. (**A**) Reverse transcription–PCR agarose gel with transcripts from the XLA patient 210/DACC (P) and an unaffected control (U). M, size marker (500- and 650-nt fragments are denoted by arrowheads); NC, negative PCR control. Amplification of normal (NT) and aberrant (AT) transcripts was carried out with primers C and D in exons 2 and 4. The sizes of NT and AT fragments are 457 and 564 nt, respectively. (**B**) Sequence chromatograms of AT/NT transcripts. (**C**) DNA sequencing with primers E (5′-ctg gtt gct taa tcc ctc tt) and F (5′-gag atg ttc tga ata tga agg) identified a single-nucleotide substitution in intron 3 in the XLA patient 210/DACC (P) and his heterozygous carrier mother, but not in unaffected controls. (**D**) The *BTK* minigene construct and schematics of RNA products of the wild-type and mutated alleles. Exons are shown as boxes, introns as lines. Authentic and aberrant transcripts are shown as dotted and dashed lines, respectively. The disease-causing T > G mutation is denoted by a star. The alignment of canonical and GC 5′ss with U1 snRNA is shown at the top; the improved base-pairing with U1 snRNA is boxed; the maximum entropy (ME) scores are in parentheses. Predicted BPS of exon 4 is underlined. Tranposable elements in intron 3 are denoted by horizontal bars, with *Alu*s in gray and LINEs in white; their orientation is indicated by arrowheads. Intron 3 deletion made in the minigene constructs is denoted by a thick line. GAA/CAA trinucleotides in the sequence between aberrant and authentic 5′ss are boxed at the bottom; mutations M1-M5 are shown above; a 76-nt riboprobe for UV crosslinking is shown below as a gray bar. Cloning (A, B) and amplification (C, D) primers are denoted by gray arrows.

mU*mG*mA*mG*mU*mG*mC*mC*mA*mC*mC and *BTK*-ESE-M2M3-sc, 5′-mG*mU*mA*mC*mA*mU* mU*mA*mA*mC*mU*mU*mU*mA*mC*mG*mC*mA* mU*mU), respectively. In these sequences, the letter m represents an *O*-methyl modification at the second position of a sugar residue and asterisks denote phosphorothioates of the RNA backbone. As additional controls, we employed oligonucleotides with identical modifications that target splicing regulatory motifs in *INS* intron 1 and their scrambled versions, as described (ref. 6 and Kralovicova1,J., *et al.*, manuscript in preparation). Transfections were carried out using jetPRIME (Polyplus Transfection, Inc.) according to manufacturer's recommendations.

Ultraviolet-light RNA crosslinking

We amplified a 76-nt DNA fragment (Figure 1D) with primers T7-*BTK*-F (5′-*taa tac gac tca cta tag gga* cca atg tct gag atg ggg a; T7 promoter is italicised) and *BTK*-R (5′- acc agt cta ttt aca gag aa) using M/M2 plasmid DNAs as templates. PCR products were gel purified using the MinElute kit (Qiagen). Riboprobes were transcribed using the MAXIScript kit (Ambion) in the presence of [α-^{32}P] UTP (800 Ci/mmol; PerkinElmer). RNAs were incubated in 10-μl reaction mixtures containing cytoplasmic S100 or HeLa nuclear extracts (4 C Biotech), 2 mM $MgCl_2$, 0.5 mM ATP, 20 mM creatine phosphate and 16 U of RNasin (Promega) at 30°C for 20 min. ASOs were added to the indicated final concentration. The samples were irradiated on ice with a 254-nm ultraviolet (UV) cross-linker (Ultralum) for 10 min and digested with RNase T1 (0.8 U/μl) and RNase A (0.4 U/μl) (Ambion) at 37°C for 20 min. Cross-linked proteins were resolved by 10% sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS–PAGE). Gels were dried, exposed to phosphor screens and signal intensity was evaluated using ImageQuant TL software (Amersham).

RNA secondary structure predictions

We used the RNA folding program RNAstructure (v. 5.2), which accommodates a dynamic programming algorithm on the principle of minimizing free energy and has an accuracy of ~73% for sequences <700 nt (31).

Identification of mutation-induced GC 5′ss resulting in genetic disease

Published reports of cryptic and *de novo* 5′ss were identified essentially as described (32). We searched PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi), locus specific mutation databases (http://www.hgmd.cf.ac.uk/docs/oth_mut.html), the database of aberrant 5′ss (http://www.dbass.org.uk) (16,33) and home pages of genetics journals for peer-reviewed reports of sequence-verified aberrant transcripts that resulted from natural disease-causing variants or mutations activating human GC 5′ss. All GC 5′ss were manually validated by mapping the information in the literature to sequences in the Human Genome project databases.

Validation of the GC 5′ss data set

Conserved GC introns were identified by comparing human and mouse reference mRNAs and genomic contigs (http://www.ncbi.nlm.nih.gov/RefSeq/) (Release 40) using GIGOgene engine (http://biigserver.ist.unomaha.edu/~achurban/GIGOgene/). The non-canonical splice sites were checked against the human ASTD database (http://www.ebi.ac.uk/astd/main.html; release 1.1) and against the Ensembl collection (http://www.ensembl.org; release 58). Only validated GC 5′ss were analyzed further.

RESULTS

Activation of *de novo* GC 5′ss in the *BTK* gene

Mutation screening of the gene encoding *B*ruton's *t*yrosine *k*inase (*BTK*) in male patients with XLA (Bruton disease, MIM 300300) using an RNA-based strategy revealed an aberrant product (Figure 1A). Sequencing of this cDNA fragment showed that, as compared to unaffected controls, the first 107 nt of intron 3 was inserted between exons 3 and 4 of the *BTK* mRNA (Figure 1B). A search for the underlying mutation in DNA identified a T>G transversion at the first position of the shortened intron, creating a noncanonical 5′ss (Figure 1C). The new GC-AG intron was of the U2-type as it lacked typical U12-type motifs, including BPS. Although the remaining positions of the *de novo* GC 5′ss showed a better match to U1 snRNA, the estimated strength of this 5′ss was considerably weaker than its authentic counterpart (Figure 1D, top left), yet the latter splice site was efficiently repressed *in vivo*.

Identification of disease-causing GC 5′ss and characterization of their intrinsic strength

To begin addressing the question of why a weak *de novo* 5′ss in *BTK* was used instead of the intact and strong authentic site nearby, we first looked for previously published disease-causing GC 5′ss. We identified additional 11 cases (Table 1), with only two in introns (refs. 34, 35 and this study). Apart from the GC 5′ss that employed authentic 3′ss, GC 5′ss were reported for three disease-causing cryptic exons that were activated in introns of the *ATM* (36), *NF1* (37) and *LHCGR* (38) genes.

Examination of the intrinsic strength of the GC 5′ss showed that their average maximum entropy (ME) score (39) was much lower than for their authentic GT equivalents (medians 2.33 versus 8.24; $P < 0.001$, Mann–Whitney rank sum test). This score, which best discriminated authentic and aberrant 5′ss (16), was, with one exception, always lower for GC 5′ss than for their authentic equivalent (Table 1). Interestingly, both the calculated free energy and the number of hydrogen bonds with U1 snRNA were higher in 4 of 12 GC 5′ss than in their GT counterparts, with free energy showing a significant difference between the two groups of splice sites ($P = 0.035$). This supports the compensatory role of base-pairing interactions with U1 snRNA in efficient GC 5′ss activation.

Intrinsically strong adjacent splice sites facilitate activation of weak GC 5′ss

Inspection of splice sites flanking the aberrant GC 5′ss in *BTK* intron 3 showed that the 3′ss of this intron was very strong (ME score 9.0), with a long polypyrimidine tract and a well-defined BPS (Figure 1D, top). Computation of the intrinsic strength of all disease-causing aberrant 5′ss and their authentic neighbors showed that splice sites adjacent to aberrant GC 5′ss tended to be stronger than the average (Figure 2A), with a significant difference in the means for the 5′ss of downstream but not upstream introns. This suggests that efficient inclusion of these

Table 1. Summary of aberrant GC 5′ss activated by pathogenic mutations

Gene	Phenotype	Mutation[a]	Location of aberrant 5′ss	Type of aberrant 5′ss[b]	Authentic 5′ss (maximum entropy score; base-pairs; ΔG)[c]	Aberrant 5′ss (maximum entropy score; base-pairs; ΔG)[c]	GAA/CAA density (%) in exonized/ intronized segments	Distance from the nearest TE (nt), TE family	Reference
PKP1	Ectodermal dysplasia-skin fragility (McGrath) syndrome	IVS9+1 G>A	Exon	Cryptic	CTG/GTGAGT (10.1;9;−7.2)	GAG/GCCTGT (−2.4;6;−1.4)	4.44	534, L3b	(94)
FGB	Congenital afibrinogenemia	IVS7+1 G>T	Exon	Cryptic	CTG/GTATGT (8.2;9;−3.9)	ACG/GCATGT (2.2;6;−0.1)	2.56	118, MIRb	(95)
CDKN2A	Melanoma	E1+353G>C, IVS1+1G>A	Exon	Cryptic	CAG/GTAGGA (9.8;8;−8.0)	GAG/GCGGCG (−7.4;3;−1.0)	0.98	2900, MER5A	(96)
XPA	Xeroderma pigmentosum group A	IVS1+2T>G	Exon	Cryptic	GAG/GTTTGG (3.5;7;−3.3)	CGG/GCGAGT (2.1;8;−5.6)	1.03	438, L3	(97)
ELN	Supravalvular aortic stenosis	E26+121G>A	Exon	Cryptic	CAG/GTGCAG (5.0;6;−4.5)	CAG/GCAGGT (2.5;8;−5.4)	0	–	(98)
MTHFR	Methylenetetrahydrofolate reductase deficiency	IVS5+1G>A	Exon	Cryptic	CAG/GTGAGG (10.1;8;−9.3)	AAG/GCATGC (1.8;6;−0.9)	0	343, MER5b	(99)
NF1	Neurofibromatosis, type I	E31+194G>A	Exon	Cryptic	CAG/GTATTG (8.4;7;−4.9)	TGG/GCAAGT (2.5;7;−5.4)	25.00	46, L2a_3end	(100)
SCN5A	Brugada syndrome	IVS27+3_6dup(TGGG)	Exon	*De novo*	TGG/GTGGGT (3.8;8;−8.8)	CAG/GCGAGT (2.9;8;−5.8)	3.13	381, L2a_3end	(101)
NF1	Neurofibromatosis type I	E25+17del11	Exon	*De novo*	TAA/GTAAAT (1.5;7;−2.3)	CAG/GCAAGT (3.1;8;−5.6)	6.45	277, MER2	(102)
CFTR	Cystic fibrosis	E23+156G>C	Intron	Cryptic	CAG/GTGAGC (9.6;8;−8.1)	AAG/GCAACT (1.3;6;−1.4)	6.90	1164, Charlie10	(34)
TAZ	Barth syndrome	IVS3+110G>A	Intron	*De novo*	TTG/GTGAGG (6.9;7;−7.0)	CAG/GCAAGG (3.3;7;−5.4)	1.89	275, L2	(35)
BTK	X-linked agammaglobulinemia	IVS3+108T>G	Intron	*De novo*	CCG/GTAAGA (10.1;7;−6.2)	CGG/GCAAGT (3.0;8;−5.4)	7.48	354, *Alu*S	This study

[a] Mutations (>, del, dup) are designated according to the traditional nomenclature (IVS, intervening sequence or intron; E, exon; exon or intron number is followed by the distance from 5′ss in nucleotides; del, deletion; dup, duplication).

[b] Types of aberrant splice sites are as described (16,103).

[c] Sequences flanking each disease-associated GC 5′ss are available from the Database of Aberrant 5′ Splice Sites (DBASS5) at http://www.dbass.org.uk (16,33).

aberrant exons in mRNAs may require help from flanking splice sites and that the 5′ss immediately downstream could be of a particular importance.

Comparisons of authentic GC and GT 5′ss

To see if similar differences can be found for natural human GC and GT introns, we compiled sequences of 500 natural GC 5′ss that were annotated in Ensembl/ASTD (Supplementary Table S1). The number of unique 9-nt GC 5′ ss sequences was 101, with 28 exonic and 40 intronic (excluding the first two intron positions) combinations.

As expected, the mean ME score of the validated GC 5′ss data set was significantly lower than a similar number (n = 457) of authentic GT counterparts of pathogenic GT 5′ss (2.36 versus 7.60; $P < 0.0001$, t-test). The mean number of hydrogen bonds was also lower for existing GC sites than for GT counterparts (means 6.9 versus 7.1, $P < 0.002$, t-test, Supplementary Table S1). Even more significant P values were observed for the Shapiro and Senapathy scores (75.4 versus 80.1, $P < 0.0001$) and free energy (ΔG) values (−4.0 versus −5.7, $P < 0.0001$). Although the tendency for higher scores of flanking splice-sites was similar to disease-associated GC splice sites (compare Figure 2A and B), the observed differences failed to reach strict statistical significance. This comparison suggests that, in addition to stronger affinity to U1 snRNA, efficient splicing of existing GC introns requires other, more critical signals.

Role of auxiliary splicing sequences in activation of GC 5′ss

The GC 5′ss in *BTK* intron 3 is intrinsically much weaker than its authentic equivalent, yet it was used by the majority of transcripts (Figure 1A), suggesting that it is promoted by splicing enhancers and/or the authentic neighbor is repressed by silencers. Interestingly, inspection of the newly exonized segment revealed an unusually high number of GAA and CAA trinucleotides (Figure 1D). The former elements are potent exonic enhancers while the latter motifs activate exon inclusion to a lesser extent (40–42). The GAA/CAA-containing enhancers show a bias toward single-stranded RNA segments both in natural exons (43) and in pseudoexons activated by pathogenic intronic mutations (42). The frequency of GAA elements in the newly exonized *BTK* segment was ~3-times higher than in average introns (4.67% versus 1.76%) and ~2-times higher than in existing human exons (versus 2.36%) (42). The density of CAA trinucleotides was also higher (2.80% versus 1.55%/1.94%, respectively).

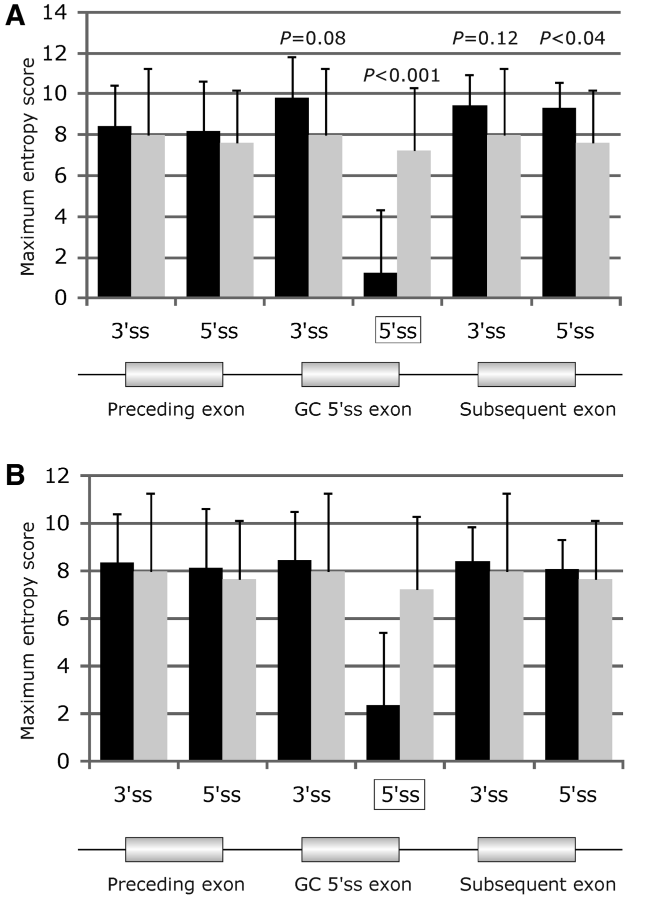

Figure 2. Comparison of the predicted strength of disease-causing GC 5′ss and their authentic counterparts. GC and GT 5′ss are denoted by black and gray columns, respectively. (**A**) Disease-causing aberrant GC 5′ss. (**B**) Authentic human GC 5′ss. The predicted strength was computed as the maximum entropy (ME) score (16,39,86). Error bars represent standard deviations (SD). Corresponding splice sites are schematically shown in the lower panels where the GC 5′ss is boxed.

Table 2 shows densities of predicted enhancers and silencers in newly exonized and intronized segments for disease-causing aberrant GC 5′ss and their authentic counterparts. The density of enhancers in the exonized sequences was higher than in existing introns and, in the

Table 2. Average densities of predicted auxiliary splicing motifs in exonized and intronized segments upon activation of GC 5′ss by pathogenic mutations

	RESCUE-ESE density	PESE density	EIE score density	SF2/ASF ESE score density	PESS density	FAS-ESS density
Exonized intronic segments	8.33	8.48	504.73	9.69	0	2.07
Intronized exonic segments	4.56	6.45	298.40	15.44	1.58	2.35
Average human introns	7.47	3.41	340.03	9.16	4.66	5.76
Average human exons	10.91	7.00	470.97	14.15	1.34	2.86

Densities of each predicted auxiliary splicing motif in newly intronized and exonized segments were obtained as described previously (25,104). Densities in existing human exons and introns were determined previously (25) and are shown here for comparison.

case of putative exonic splicing enhancers (PESEs) (44) and exon identity elements (EIEs) (45), even higher than in average exons. Conversely, the enhancer densities in intronized segments were lower than in average exons. By contrast, the density of silencers in exonized segments was lower than that in existing introns or even average human exons while intronized segments had a higher density of putative exonic splicing silencers (PESSs) (44), but not silencers determined by a fluorescence-activated screen (FAS-ESSs) (46).

At the level of individual nucleobases, intronized sequences of GC exons had a lower frequency of adenines (Figure 3A–C), which may reflect a critical importance of (di)adenosines for tertiary interactions in large RNAs and their role in exon recognition. In addition, these sequences were also cytosine- and guanine-rich, with the GC-content higher in natural GC exons as compared to GT exons (Figure 3C). We also observed a loss of uracil and gain of guanine in exonized intronic sequences (Figure 3). The frequency of non-overlapping G triplets in intronized segments was 1.41% as compared to 1.12% in exons and 1.24% in introns, but was somewhat higher for exonized segments (1.65%). Finally, Figure 4 shows the density of auxiliary splicing elements in conserved GC and GT exons. Mean densities of PESEs and putative SF2/ASF ESEs were particularly enriched in GC exons as compared to conserved GT exons, speculatively reflecting a possible preference of this SR protein for 5′ss (47,48).

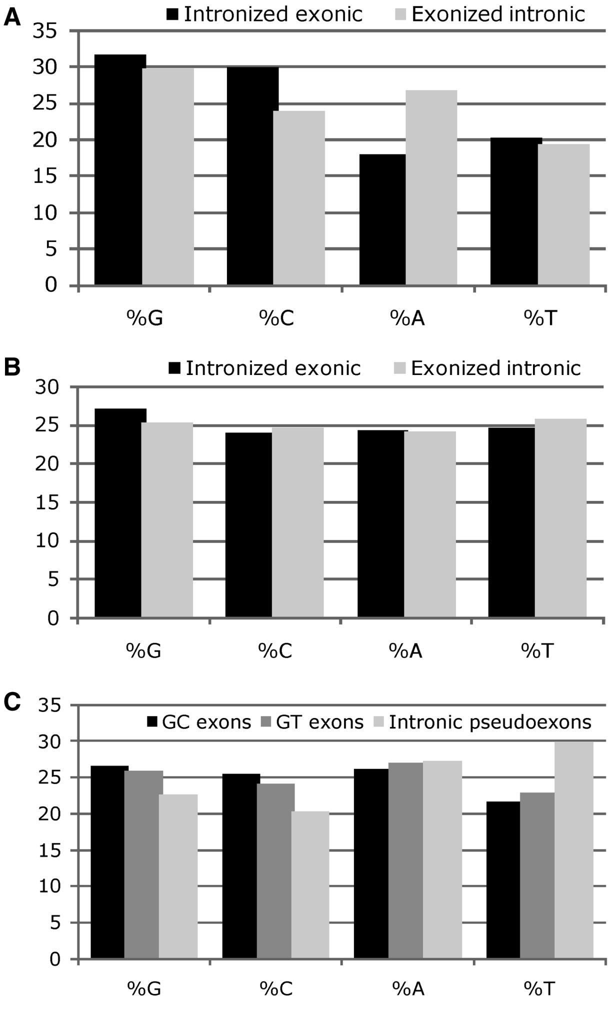

Figure 3. Nucleotide composition of intronized exonic and exonized intronic segments. (**A**) Disease-associated GC 5′ss (a total of 881 nt). (**B**) Disease-associated GT 5′ss (36 378 nt). (**C**) Authentic GC (n = 500) and GT (n = 44 232) human exons are shown together with a sample of 2309 intronic pseudoexons in the 5′ untranslated region (44) for comparison.

Identification of a splicing enhancer of the *BTK* GC 5′ss

To test if activation of the weak GC 5′ss is indeed induced by high-density GAA/CAA-type enhancers in the newly exonized segment we constructed a three-exon minigene reporter with and without the disease-causing substitution and designed five mutations (M1–M5) that reduce the number of GAAs (M1–M4) or CAAs (M5) to the level typical of average introns. We swapped the first and second positions of GA or CA dinucleotides in these elements to remove diadenosines, which are frequently involved in intra- and intermolecular interactions of large RNAs (refs. 42, 49 and 50 and references therein), while keeping the overall nucleotide composition unchanged. Splicing of the WT minigenes produced canonical mRNAs, with minor species spliced to a dual splice site (3′ss at the 5′ss of the last exon) activated by the adjacent EcoRI cloning site, both in the 293 T cells and a B-cell line Nalm-6 (Figure 5A and B, lane 1). The reporter construct with the T>G transversion (M) exhibited splicing to the GC 5′ss in ~90% primary transcripts (Figure 5A and B, lane 2), recapitulating the *in vivo* splicing pattern. A major reduction of aberrant splicing was observed for M2 and M3 mutants (Figure 5C and D), each removing one of the tandem GAA elements located in the middle of the exonized segment. This $(GAA)_2$ element was conserved in Great Apes and Old World Monkeys while the 9-nt signal of the GC 5′ss was more relaxed in *Homininae* and more

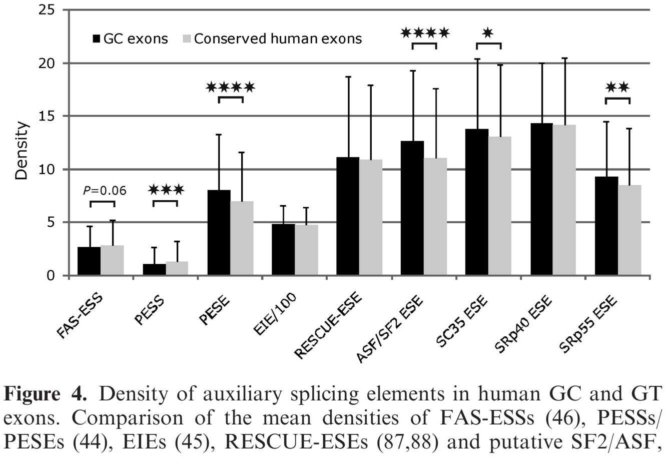

Figure 4. Density of auxiliary splicing elements in human GC and GT exons. Comparison of the mean densities of FAS-ESSs (46), PESSs/PESEs (44), EIEs (45), RESCUE-ESEs (87,88) and putative SF2/ASF, SC35, SRp40 and SRp55 ESEs (89,90) in 500 GC and 44 232 GT exons. ****$P < 0.0001$; ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$ (unpaired t-tests). Element densities were calculated for each input sequence as described previously (25); EIE densities are shown as 1/100 of their actual values. Error bars indicate SD.

stringent in lower primates due to an A>G transition (Supplementary Figure S1). Mutation M4, which was close to the GC 5′ss, had only a minor effect. Mutation M5, which removed an apparently weaker CAA element, decreased aberrant splicing by a third. By contrast, mutation M1, which is adjacent to a G triplet, increased splicing to cryptic 5′ss. Thus, the strongest support of GC 5′ss splicing was provided by the conserved $(GAA)_2$ element.

Efficient inhibition of aberrant GC 5′ss by splice-switching oligonucleotides

To reduce cryptic splicing and provide further evidence for the importance of the $(GAA)_2$-containing intronic enhancer, we have employed ASOs designed to block access of the spliceosomal machinery to the GC 5′ss and M2/M3. Uniform phosphorothioate oligoribonucleotides that had 2′-*O*-methyl modification at each base efficiently repressed the GC 5′ss in a dose-dependent manner (Figure 5E), with oligos targeting the GC 5′ss exhibiting significant repression at a final concentration of only 1 nM (Supplementary Figure S2). A major reduction of cryptic splicing was observed even in the absence of transfection reagent (lane T-, Figure 5E). By contrast, scrambled oligos and their mixture failed to inhibit the inclusion of the intron 3 segment in mRNA (Figure 5E and data not shown). At higher concentrations, each oligo as well as their scrambled counterparts induced exon 3 skipping while repressing splicing to the dual splice site that separated the last minigene exon from the cloning EcoRI site. Thus, the aberrant splicing phenotype in this XLA case was effectively repaired by chemically modified oligonucleotides targeting either the GC 5′ss itself or the enhancer, providing the first example of a successful correction of the pathogenic GC 5′ss activation.

Trans-acting factors associated with the GC 5′ss selection in *BTK* intron 3

To identify proteins that bind the critical enhancer motif, we first examined splicing of M, M2 and M3 constructs in cells depleted of/overexpressing a series of cellular factors known to influence RNA processing, as we described earlier (6). We found that depletion of SR proteins 9G8, SC35 and Tra2α/β decreased utilization of cryptic GC 5′ss on each reporter pre-mRNA (Figure 6A and B). By contrast, overexpression of these factors was associated with almost exclusive use of this splice site (Figure 6C). Overexpression of SF2/ASF promoted exon 3 skipping (Figure 6C and D). Exon 3 skipping was also observed, albeit to a lesser extent in cells overexpressing SRp20 and SC35 but not Tra2β (Figure 6C), despite efficient downregulation and overexpression of this factor in transfected cells (Figure 6E).

To identify proteins that bind the $(GAA)_2$ enhancer, we carried out UV-RNA crosslinking with the *BTK*-PY7 versions of M, M2 and M3 constructs (Figure 7A). Splicing of each construct in 293 T and HeLa cells replicated the M>M3>M2 hierarchy observed for authentic clones (compare Figures 7B and 5 C), indicating that the tandem GAA element plays a critical role also in splicing of hybrid reporters. The transcripts from each clone could be spliced in nuclear extracts (Figure 7C); however, the pattern of proteins crosslinked to uniformly labeled RNAs was similar (Figure 7D). The UV crosslinking of short, internally labelled RNA probes carrying M and M2 and spanning 71% of the exonized segment (Figure 1D, bottom) also revealed a comparable set of proteins, with the exception of a weaker 55-kDa fragment which was more pronounced for the M construct (Figure 8A). Secondary structure predictions suggested that both M and M2 maintain a relatively stable hairpin with a terminal purine tetraloop and an A/U repeat as loop-closing base pairs (Figure 8B). This motif may

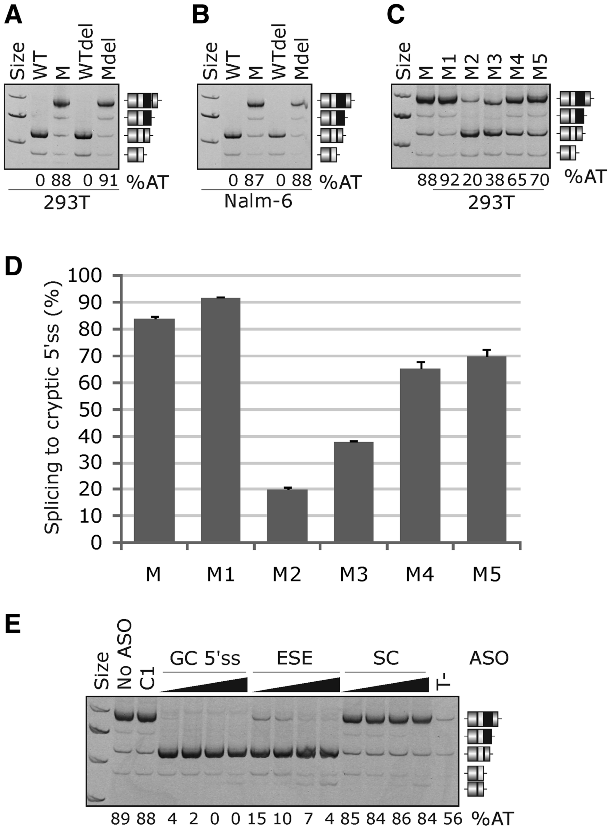

Figure 5. Efficient activation of the GC 5′ss in *BTK* intron 3 requires a GAA intronic enhancer. The splicing pattern of *BTK* minigene reporters in 293 T (**A, C**) and Nalm-6 (**B**) cell lines. Designation of the reporter constructs is at the top, RNA products are to the right and percentage of splicing to the GC 5′ss (%AT) is at the bottom. WT, wild type reporter construct; M, construct with the T > G transversion; M1–M5 mutations and deletion of interspersed repeats (Del) are shown in Figure 1D. (**D**) Summary of the triplicated transfection experiment. Error bars represent SD. (**E**) Effective repression of aberrant splicing by antisense oligonucleotides (ASOs) targeting the GC 5′ss itself or the upstream M2/M3 enhancer (ESE). The final concentration of each ASO was 20, 50, 100 and 200 nM. SC, an equimolar mix of scrambled controls of the two ASOs; C1, a generic negative control at 200 nM, T-, equimolar amounts of GC 5′ss and ESE ASOs (at 100 nM each) added to cells without any transfection reagent. No-transfection, no-template and no-RT controls are not shown. RNA products are to the right and the percentage of splicing to the GC 5′ss (%AT) is at the bottom. The lowest effective ASO concentrations were determined in a separate experiment shown in Supplementary Figure S1.

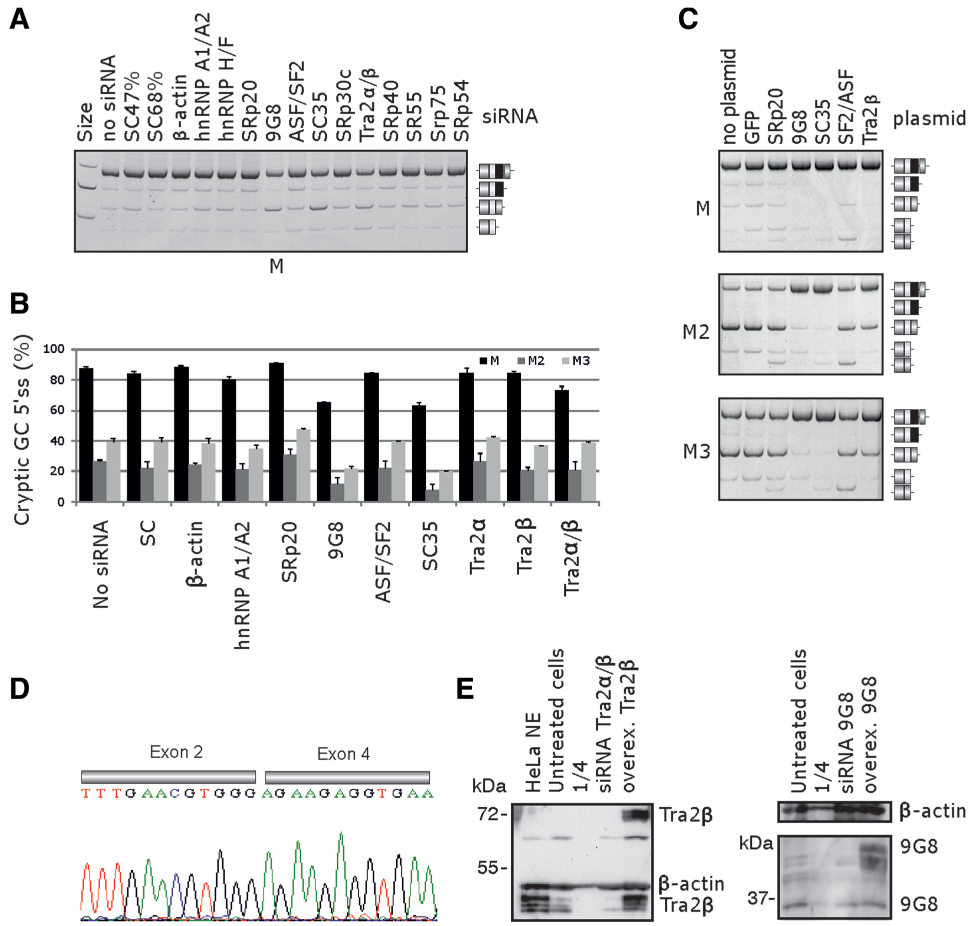

Figure 6. Splicing factors associated with promotion or repression of the GC 5′ss in *BTK* intron 3. (**A**) RNA interference-mediated depletion of proteins shown at the top in 293 T cells. SC47%, SC68%, scrambled control siRNAs with the indicated GC contents. RNA products are to the right, the reporter is at the bottom. Individual depletion of Tra2α and Tra2β proteins (the official gene symbol *TRAB*) is shown in panel B. (**B**) Summary of the triplicated transfection experiment with M, M2 and M3 constructs. Error bars represent SD. (**C**) Overexpression of SF2/ASF, SRp20, 9G8, SC35 and Tra2β in 293 T cells. *BTK* reporter constructs are to the left, RNA products to the right. GFP, plasmid expressing green fluorescent protein. Tra2β here refers to the longest isoform designated Tra2β1 (82). (**D**) Sequence chromatogram showing *BTK* exon 3 skipping in 293 T cells overexpressing SF2/ASF and SRp20. (**E**) Western blot analysis of Tra2β and 9G8 proteins. Antibodies against Tra2β and β-actin were purchased from Abcam (ab31353 and ab37063, respectively), antibodies against 9G8 (*SFRS7*) were obtained from Sigma (SAB1101226). NE, nuclear extracts; 1/4, loading of the lysate was reduced by 75% as compared to untreated 293 T cells. Overexpressed Tra2β and 9G8 were fused with GFP and T7 tags, respectively. Size markers are shown to the left. Antibodies against SC35 (ab28428) failed to yield specific bands, but the increased and decreased expression was observed using quantitative RT–PCR with RNA samples from transfected cells (data not shown).

serve as a potential binding site for poly(Y)-proteins, such as ∼57-kDa hnRNP I (PTB). However, downregulation of PTB and nPTB proteins in 293 T cells failed to alter the relative expression of RNA products (Figure 8C). Taken together, SR proteins 9G8, SC35 and Tra2α/β were associated with pathogenic activation of the GC 5′ss in *BTK* intron 3.

Transposable elements and GC 5′ss activation

BTK intron 3 is ∼2.8 kb long, but >2.0 kb of its mid-portion is formed by a continuous tract of transposable elements (TEs), consisting of several *Alu* insertions in a single long interspersed nuclear element (Figure 1D). *Alu*s in the opposite orientation would be predicted to create base-pairing interactions that may influence splicing efficiency of adjacent segments (51,52) and may promote or repress the aberant GC 5′ss. In particular, short interspersed nuclear elements (SINEs), including *Alu*s and mammalian interspersed repeats (MIRs), may provide favorable substrates for activation of splice sites and cryptic exons (42,52–54). To test if the TE stretch can influence the GC 5′ss splicing, we deleted this sequence from intron 3 and examined exogenous transcripts following transient transfections into several cell lines. However, splicing was not altered by this deletion (Figure 5A and B, lanes 3 and 4) and repeated attempts to delete one or more *Alu*Sx in the opposite orientation, predicted to form a double-stranded structure of extreme stability (−487.3 kcal/mol), were not successful.

We also examined repetitive sequences flanking each disease-associated GC 5′ss, but we found no aberrant GC 5′ss that would be activated in recognizable TE fragments. The average distance from the nearest TE (L2/L3 fragments in five cases, DNA transposons in four cases and SINEs in two cases) was ∼620 nt. Extension of this analysis to authentic GC exons and intronic sequences

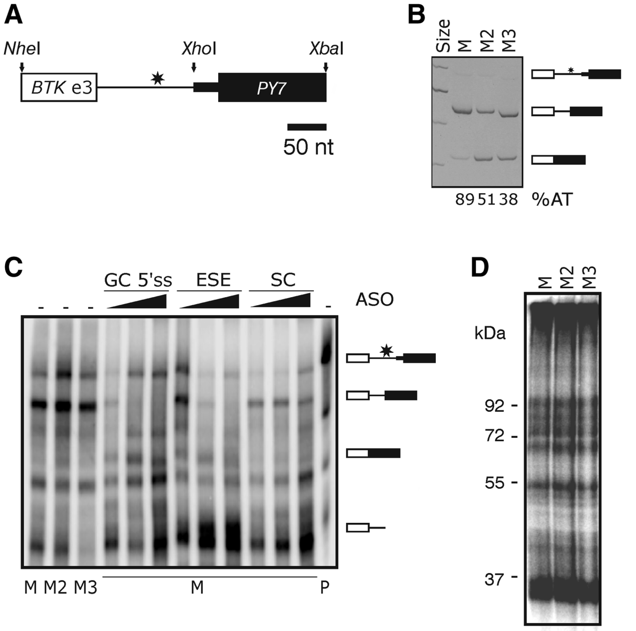

Figure 7. UV-RNA crosslinking of hybrid *BTK*-PY7 constructs. (**A**) Schematic representation of the hybrid *BTK*-PY7 reporter. The length of introns (lines) and exons (boxes) is to a scale shown at the bottom. Restriction enzymes are shown at the top. The position of GC 5′ss-activating mutation is denoted by a star. (**B**) Splicing pattern of the hybrid *BTK*-PY7 reporters in HeLa cells. (**C**) Splicing of *BTK*-PY7 reporters in HeLa nuclear extracts. Mutations are shown at the bottom; P, mock treated riboprobe; splicing products are to the right. Final concentration of ASOs was 2, 20 and 200 nM. (**D**) UV-RNA crosslinking of *BTK*-PY7 reporter constructs (at the top). The size of crosslinked proteins is to the left.

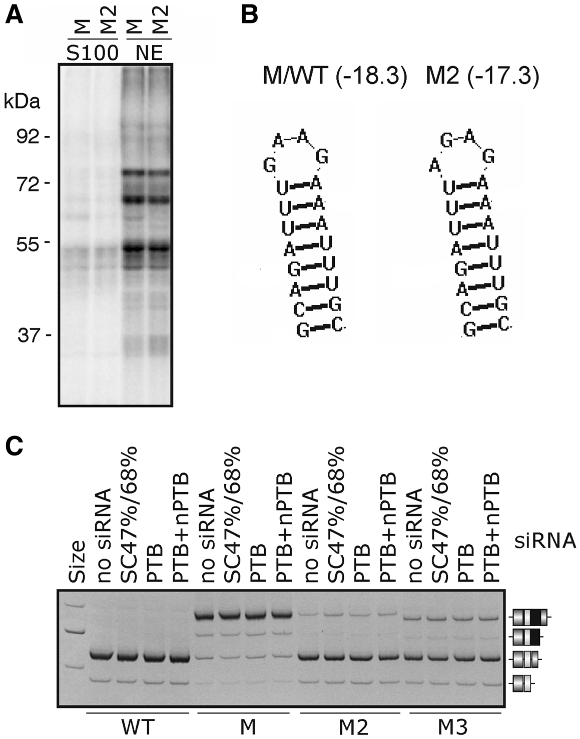

Figure 8. UV-RNA crosslinking of exonized pre-mRNA of *BTK* intron 3. (**A**) UV crosslinking of M and M2 RNAs in cytoplasmic S100 and nuclear extracts. (**B**) Predicted hairpins of M and M2 riboprobes. (**C**) Splicing of *BTK* constructs in 293 T cells depleted of PTB/nPTB. RNA products are shown schematically to the right, reporter constructs at the bottom and siRNAs at the top. SC47%/SC68%, control siRNAs with the indicated GC content. Final concentration of siRNAs targeting PTB (91) was 70 nM. Final concentration of each siRNA targeting nPTB (92) was 15 nM.

adjacent to GC 5′ss revealed several examples of TE-derived GC 5′ss (Supplementary Figure S3). Specifically, *PSD4* exon 3 and *FBXO18* exon 2 originated from sense MIRc, GC 5′ss in *PTR*, *DDX52*, *AGBL2* and *AP3M1* exons from *Alu*s, and *NAT1, GLYATL1* and *BAZ1A* exons from LTR retrotransposons. Both GC 5′ss in *PSD4* and *FBXO18* mapped to identical positions in the sense MIR consensus and colocalized with at least three GT 5′ss (42).

Analysis of the first TEs downstream of the GC 5′ss revealed that SINEs were clearly overrepresented (>60%, Figure 9A). The distance between GC 5′ss and these TEs was the largest for *Alu*s (mean 608 nt) and smallest for LTRs (206 nt) (Figure 9B). This difference was greater than would be expected by chance ($P = 0.03$, Kruskal–Wallis one-way analysis of variance on ranks), suggesting that the *Alu* proximity to splice sites may interfere with selection of weaker 5′ss, most likely through *Alu*-formed pre-mRNA structures and/or their binding properties.

DISCUSSION

Choice of authentic splice sites among a large excess of similar consensus signals in primary transcripts is governed by multiple decisions of the spliceosomal machinery that are at present poorly understood. Focusing on a group of noncanonical 5′ss that provide natural models of weak splice-site activation *in vivo*, our study has characterized compensatory signals associated with selection of pathogenic GC 5′ss. Although these 5′ss appear to signify a relaxation of splice site motifs in higher eukaryotes, which enhances regulatory potential of gene expression pathways (55), the fraction of natural GC-AG introns is much higher in some lower organisms, particularly those with high genome GC content, approaching 40% in a pelagophyte *A. anophagefferens* (56). It is not clear why this frequency is as high and if there is any special function of this peculiar choice of the 5′ intron end as compared to other organisms. Mammalian GC 5′ss can play a role in coupled splicing and translation control as well as in subcellular localization. For example, in the mouse Gli1 oncogene, a terminal transcriptional effector of the *hedgehog* signaling pathway, the GC 5′ss promoted exon 1B skipping and translation efficiency (57). Subcellular localization and degradation of *ING4*, a cell growth repressor, was controlled by alternative splicing of the GC(N)$_7$GT exon (58).

An additional level of GC 5′ss regulation can be provided by RNA editing, which has been shown to modulate choice of GT 5′ss (59) and 3′ss (60,61).

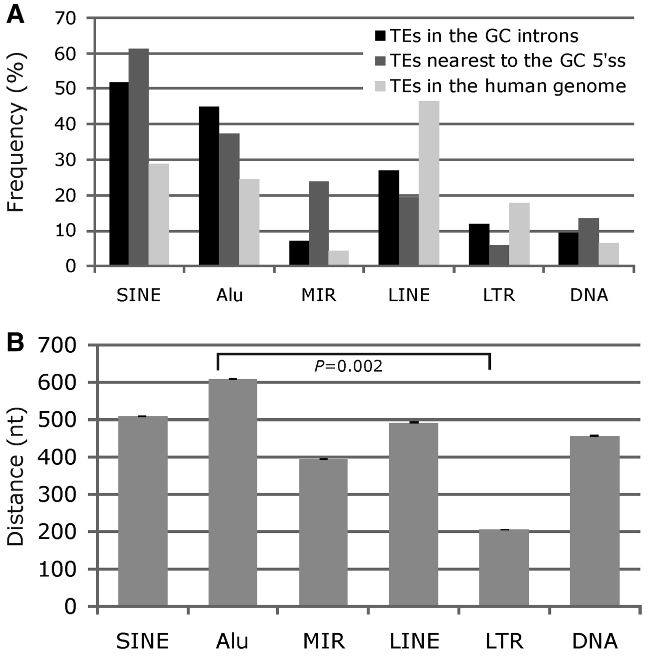

Figure 9. Transposable elements and selection of GC 5′ss. (**A**) Frequencies of TE families in 500 GC introns, among the first TEs downstream of the GC 5′ss and the overall frequency in the human genome (93) for comparison. (**B**) The average distance between the GC 5′ss and the 5′-end of the first downstream TE. TE families are shown at the bottom.

Although this has not been formally shown for GC 5′ss to our knowledge, editing of adenosine to inosine, which is recognized as guanosine by the spliceosome, might create the 5′ss GC (AC > IC) and either improve or impair the 5′ss consensus. For example, a CAG/GCAAAU > CAG/GCAAIU editing change would increase the ME score from 1.0 to 3.1, while a CAG/GCGAGU > CAG/GCGIGU change would reduce it from 2.9 to 0.8. Both adenosine-to-inosine and cytosine-to-uridine editing events may also alter auxiliary splicing sequences required for exon recognition, as was shown for an *Alu*-derived splicing enhancer in the gene coding for prelamin A (61). Because GC 5′ss are on average weaker than GT 5′ss (Figure 2) and rely more on splicing silencers and enhancers (Figure 4), editing events within these motifs would be expected to have a stronger influence on recognition of GC 5′ss than GT 5′ss. Our search of the database of human editing sites (62) has identified several examples of adenosine-to-inosine editing in the proximity of exons with GC 5′ss (Supplementary Table S2). The importance of these editing sites in splicing can be tested can be tested experimentally in future studies. Although creation of GC 5′ss by RNA editing could be excluded in our *BTK* case (Figure 1), coupled editing and alternative splicing of GC 5′ss would represent a particularly fine instrument for spatiotemporal regulation of gene expression, which warrants a more systematic investigation in the future.

The association of GC 5′ss with stronger adjacent splice sites (Figure 2) is in agreement with previous observations of splicing interdependence of neighboring introns. Exon-skipping splice-site muations often lead to deletion of more than one exon in both directions, and the same applies to intron retention events (63,64). Selection of adjacent 3′ss (30) and 5′ss (65) can also be influenced by unproductive splicing, presumably by recognizing decoy 3′ss and 5′ss to which spliceosomal components are attracted, which then facilitates recruitment of additional factors. Kinetic analyses demonstrated that an upstream competing 5′ss enhances the rate of intron-proximal splicing (65). Alternatively, decoy 5′ss may resemble auxiliary splicing elements that are abundant in exons and introns and may shape intramolecular interactions and/or binding gamut (30,65,66).

The U1 binding potential for GC 5′ss seems to be an important predictor of their activation (Supplementary Table S1), consistent with previous correlation of the observed rates of 5′ss splicing with their U1 binding potential (65). A mutant U1 activated a noncanonical cryptic 5′ss in *Schizosaccharomyces pombe* even in the presence of a wild-type sequence at the natural 5′ junction, suggesting that snRNAs redirected splicing via base-pairing (67).

Weak or low-complementarity splice sites are expected to be less efficiently included in mature transcripts. The intrinsic strength of *BTK* splice sites and auxiliary elements (Supplementary Table S3) suggested that the 3′ss of exon 14 and the 5′ss of exon 15 are exceptionally weak. Similarly, exons 16–18 lack enhancers or had an excess of silencers, which makes them more susceptible to skipping than the remaining exons. In fact, characterization of aberrant transcripts in purified pre-B leukemia cells clearly showed that these exons were preferentially skipped (68), which probably resulted from subphysiological conditions during cell purification, rather than from *BTK*-mediated processes contributing to malignant transformation. Thus, exons that are more likely to exhibit skipping in stressed cells are predictable *ab initio*, which should be taken into account in experimental studies. Recently developed online tools predicting exonic positions/substitutions that result in skipping of the exon provide merely a starting point toward more definitive and practical algorithms (69).

It was surprising to see the excess of predicted enhancers in exonized segments and their lack in intronized segments (and vice versa) with such a small sample size of disease-associated events (Table 2). For example, the EIE score density was 1.96 × higher in exonized intronic segments than for those activated as a result of aberrant GT 5′ss. Collectively, our data suggest that recognition of GC 5′ss requires a stronger set of auxiliary elements that compensate weak splice sites and promote high-level usage required for manifestation of a recognizable phenotype. These results should ultimately improve computational prediction of GC 5′ss. For example, a recently updated SpliceScan II, which outperformed other GC 5′ss sensors based on weight matrices (70), did not identify the *BTK* GC 5′ss on the mutated allele (Supplementary Figure S4). Giving more weight to auxiliary elements should improve sensitivity of prediction algorithms that attempt to identify pathogenic GC 5′ss.

In our experimental model of GC 5′ss activation, we identified 9G8, Tra2α/β, SF2/ASF and SC35 as proteins important for selection of this splice site (Figure 6A and data not shown). Tra2β has been previously shown to bind purine-rich elements, including GAA repeats (71–75) and also remains a prime candidate for binding to the predicted terminal loop (Figure 8B), probably in a complex with other SR proteins. The AGAA motif appears to be an optimal binding site of Tra2β that lacks the arginine/serine-rich domain, and was superior to NGAA or UCAA (76,77). The weaker affinity for CAA trinucleotides is likely to reflect cytidine deficiency to fix the first adenosine on the β sheet surface of Tra2β (76). This interaction may account for least partially account for a lower activity of C/A-rich splicing enhancers than their G/A-containing equivalents (41,42), although the involvement of other proteins cannot be excluded. *Drosophila Transformer 2* protein partners include *B52*/SRp55 and *Rbp1*/SRp20 (78); in our experiments SRp20 depletion promoted the GC 5′ss use in each reporter pre-mRNA (Figure 6A and B). Although site-specific labeling was not carried out, we observed no M- or M2/M3-specific proteins under splicing conditions on crosslinking gels (Figures 7D and 8A), keeping open the possibility that the loop is primarily involved in tertiary interactions to an unknown RNA receptor. In large RNAs, such contacts are often mediated by diadenosines, with the A-minor motifs docking into the minor groove of a receptor helix as probably the most frequent tertiary contacts (50,79). Secondary structure predictions of 76-nt riboprobes showed that the terminal purine tetraloop in WT/M was also maintained for M2 (Figure 8B), but not M3. This may explain a smaller size of M3 transcripts on native polyacrylamide gels (Figures 5C and 7B), as their sequence analysis excluded cryptic splicing. By contrast, although the GAAG tetraloop was predicted for the M version of *BTK*-PY7 pre-mRNA, the M2 and M3 motifs in the hybrid clone would rather reside in internal bulges/loops or double-stranded conformation (data not shown). Unlike G triplets that appear to act additively and are less context-sensitive (6,80), individual GAA/CAA elements located in the exonized segment of *BTK* had more variable effects on splice site selection (Figure 5D). This is likely to reflect structural constraints of their respective ligands or their high dependency on adjacent sequences. Finally, *BTK* exon 3 provides yet another example of an exon negatively regulated by SF2/ASF (Figure 6C and D), as reported earlier for splicing of the adenovirus IIIa pre-mRNA (81) and human substrates, including *TRA2B* exon 2 (82) and *CFTR* exon 9 (83).

Our analysis of TEs around GC 5′ss further supported a special role of *Alu*s in exonization of intronic sequences. Apart from supplying strong polypyrimidine tracts and both 3′ss and 5′ss (53,54,84), *Alu*s contain splicing enhancers (61,85) and silencers (30). The latter may be directly linked to the observed lack of these elements in the vicinity of GC 5′ss and this repression of splice site selection may be further modified by RNA editing. Although we could not test this in our *BTK* model in which *Alu*s inserted in a more ancient repeat hampered DNA manipulation (Figure 1D), future studies should address this hypothesis.

In conclusion, we have developed a new experimental model for activation of a pathogenic GC 5′ss (Figures 1 and 5), identified *cis*-elements and SR proteins positively and negatively associated with this event (Figures 5 and 6), and showed that the aberrant splicing can be corrected *in vitro* and *ex vivo* (Figure 5 and Supplementary Figure S2). Systematic characterization of both disease-causing and authentic GC 5′ss revealed a series of characteristics associated with their efficient activation, explaining why such intrinsically weak sites are used in the vicinity of strong authentic equivalents and facilitating their better prediction in human genetic disorders. Finally, the ever-increasing collection of disease-linked aberrant splice sites (33) provides a simple yet powerful means of studying intricate rules that govern human splice-site selection *in vivo*.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank D. Black, UCLA, G. Dreyfuss, HHMI, C. Smith, University of Cambridge, A. Krainer, Cold Spring Harbor Lab, M. Hastings, University of Chicago, G. Screaton, Imperial College, J. Cáceres, University of Edinburgh, and I. Eperon, University of Leicester, for providing antibodies/plasmids or useful discussions. We also thank L. Tricas, Hospital Universitario Central de Asturias, for providing the patient's blood sample, J. Rogers for critically reading the manuscript and anonymous reviewers for helpful comments.

FUNDING

The Juvenile Diabetes Research Foundation International, Diabetes UK (to I.V.); the Swedish Research Council, the European Union FP7-HEALTH-2008-223293 grant "European genetic disease diagnostics"; and the Stockholm County Council (to E.S.). Funding for open access charge: the Swedish Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Wahl,M.C., Will,C.L. and Lührmann,R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**, 701–718.
2. Burge,C.B., Tuschl,T. and Sharp,P.A. (1999) In Gesteland,R.F., Cech,T.R. and Atkins,J.F. (eds), *The RNA World*. Cold Spring Harbor Laboratory Press, New York, pp. 525–560.
3. Buratti,E., Muro,A.F., Giombi,M., Gherbassi,D., Iaconcig,A. and Baralle,F.E. (2004) RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon. *Mol. Cell. Biol.*, **24**, 1387–1400.
4. Warf,M.B. and Berglund,J.A. (2010) Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem. Sci.*, **35**, 169–178.
5. Zhang,X.H. and Chasin,L.A. (2006) Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc. Natl Acad. Sci. USA*, **103**, 13427–13432.

6. Kralovicova,J. and Vorechovsky,I. (2010) Allele-dependent recognition of the 3′ splice site of *INS* intron 1. *Hum. Genet.*, **128**, 383–400.
7. Zhuang,Y. and Weiner,A.M. (1986) A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. *Cell*, **46**, 827–835.
8. Bruzik,J.P. and Steitz,J.A. (1990) Spliced leader RNA sequences can substitute for the essential 5′ end of U1 RNA during splicing in a mammalian in vitro system. *Cell*, **62**, 889–899.
9. Crispino,J.D., Blencowe,B.J. and Sharp,P.A. (1994) Complementation by SR proteins of pre-mRNA splicing reactions depleted of U1 snRNP. *Science*, **265**, 1866–1869.
10. Zhong,X.-Y., Wang,P., Han,J., Rosenfeld,M.G. and Fu,X.-D. (2009) SR proteins in vertical integration of gene expresion from transcription to RNA processing to translation. *Mol. Cell*, **35**, 1–10.
11. Moore,M.J. and Proudfoot,N.J. (2009) Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell*, **136**, 688–700.
12. Thanaraj,T.A. and Clark,F. (2001) Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res.*, **29**, 2581–2593.
13. Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.
14. Sheth,N., Roca,X., Hastings,M.L., Roeder,T., Krainer,A.R. and Sachidanandam,R. (2006) Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.*, **34**, 3955–3967.
15. Churbanov,A., Winters-Hilt,S., Koonin,E.V. and Rogozin,I.B. (2008) Accumulation of GC donor splice signals in mammals. *Biol. Direct*, **3**, 30.
16. Buratti,E., Chivers,M.C., Kralovicova,J., Romano,M., Baralle,M., Krainer,A.R. and Vorechovsky,I. (2007) Aberrant 5′ splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res.*, **35**, 4250–4263.
17. Dou,Y., Fox-Walsh,K.L., Baldi,P.F. and Hertel,K.J. (2006) Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *RNA*, **12**, 2047–2056.
18. Jackson,I.J. (1991) A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res.*, **19**, 3795–3798.
19. Mount,S.M. (2000) Genomic sequence, splicing, and gene annotation. *Am. J. Hum. Genet.*, **67**, 788–792.
20. Kitamura-Abe,S., Itoh,H., Washio,T., Tsutsumi,A. and Tomita,M. (2004) Characterization of the splice sites in GT-AG and GC-AG introns in higher eukaryotes using full-length cDNAs. *J. Bioinform. Comput. Biol.*, **2**, 309–331.
21. Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2001) SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res.*, **29**, 255–259.
22. Abril,J.F., Castelo,R. and Guigo,R. (2005) Comparison of splice sites in mammals and chicken. *Genome Res.*, **15**, 111–119.
23. Aebi,M., Hornig,H. and Weissmann,C. (1987) 5′ cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5′ splice region, not by the conserved 5′ GU. *Cell*, **50**, 237–246.
24. Vetrie,D., Vorechovsky,I., Sideras,P., Holland,J., Davies,A., Flinter,F., Hammarström,L., Kinnon,C., Levinsky,R., Bobrow,M. *et al.* (1993) The gene involved in X-linked agammaglobulinaemia is a member of the src family of protein-tyrosine kinases. *Nature*, **361**, 226–233.
25. Kralovicova,J. and Vorechovsky,I. (2007) Global control of aberrant splice site activation by auxiliary splicing sequences: evidence for a gradient in exon and intron definition. *Nucleic Acids Res.*, **35**, 6399–6413.
26. Kralovicova,J., Houngninou-Molango,S., Krämer,A. and Vorechovsky,I. (2004) Branch sites haplotypes that control alternative splicing. *Hum. Mol. Genet.*, **13**, 3189–3202.
27. Deirdre,A., Scadden,J. and Smith,C.W. (1995) Interactions between the terminal bases of mammalian introns are retained in inosine-containing pre-mRNAs. *EMBO J.*, **14**, 3236–3246.
28. Heinonen,J.E., Smith,C.I. and Nore,B.F. (2002) Silencing of Bruton′s tyrosine kinase (Btk) using short interfering RNA duplexes (siRNA). *FEBS Lett.*, **527**, 274–278.
29. Kashima,T. and Manley,J.L. (2003) A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nat. Genet.*, **34**, 460–463.
30. Lei,H. and Vorechovsky,I. (2005) Identification of splicing silencers and enhancers in sense *Alu*s: a role for pseudo-acceptors in splice site repression. *Mol. Cell. Biol.*, **25**, 6912–6920.
31. Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
32. Kralovicova,J., Christensen,M.B. and Vorechovsky,I. (2005) Biased exon/intron distribution of cryptic and *de novo* 3′ splice sites. *Nucleic Acids Res.*, **33**, 4882–4898.
33. Buratti,E., Chivers,M.C., Hwang,G. and Vorechovsky,I. (2011) DBASS3 and DBASS5: databases of aberrant 3′ and 5′ splice sites in human disease genes. *Nucleic Acids Res.*, **39**, D86–D91.
34. Jones,C.T., McIntosh,I., Keston,M., Ferguson,A. and Brock,D.J. (1992) Three novel mutations in the cystic fibrosis gene detected by chemical cleavage: analysis of variant splicing and a nonsense mutation. *Hum. Mol. Genet.*, **1**, 11–17.
35. Sakamoto,O., Ohura,T., Katsushima,Y., Fujiwara,I., Ogawa,E., Miyabayashi,S. and Iinuma,K. (2001) A novel intronic mutation of the TAZ (G4.5) gene in a patient with Barth syndrome: creation of a 5′ splice donor site with variant GC consensus and elongation of the upstream exon. *Hum. Genet.*, **109**, 559–563.
36. Pagani,F., Buratti,E., Stuani,C., Bendix,R., Dork,T. and Baralle,F.E. (2002) A new type of mutation causes a splicing defect in ATM. *Nat. Genet.*, **30**, 426–429.
37. Pros,E., Gomez,C., Martin,T., Fabregas,P., Serra,E. and Lazaro,C. (2008) Nature and mRNA effect of 282 different NF1 point mutations: focus on splicing alterations. *Hum. Mutat.*, **29**, E173–E193.
38. Kossack,N., Simoni,M., Richter-Unruh,A., Themmen,A.P. and Gromoll,J. (2008) Mutations in a novel, cryptic exon of the luteinizing hormone/chorionic gonadotropin receptor gene cause male pseudohermaphroditism. *PLoS Med.*, **5**, e88.
39. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.
40. Yeakley,J.M., Morfin,J.P., Rosenfeld,M.G. and Fu,X.D. (1996) A complex of nuclear proteins mediates SR protein binding to a purine-rich splicing enhancer. *Proc. Natl Acad. Sci. USA*, **93**, 7582–7587.
41. Coulter,L.R., Landree,M.A. and Cooper,T.A. (1997) Identification of a new class of exonic splicing enhancers by in vivo selection. *Mol. Cell. Biol.*, **17**, 2143–2150.
42. Vorechovsky,I. (2010) Transposable elements in disease-associated cryptic exons. *Hum. Genet.*, **127**, 135–154.
43. Hiller,M., Zhang,Z., Backofen,R. and Stamm,S. (2007) Pre-mRNA secondary structures influence exon recognition. *PLoS Genet.*, **3**, e204.
44. Zhang,X.H. and Chasin,L.A. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.*, **18**, 1241–1250.
45. Zhang,C., Li,W.H., Krainer,A.R. and Zhang,M.Q. (2008) RNA landscape of evolution for optimal exon and intron discrimination. *Proc. Natl Acad. Sci. USA*, **105**, 5797–5802.
46. Wang,Z., Rolish,M.E., Yeo,G., Tung,V., Mawson,M. and Burge,C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831–845.
47. Krainer,A.R., Conway,G.C. and Kozak,D. (1990) The essential pre-mRNA splicing factor SF2 influences 5′ splice site selection by activating proximal sites. *Cell*, **62**, 35–42.
48. Eperon,I.C., Ireland,D.C., Smith,R.A., Mayeda,A. and Krainer,A.R. (1993) Pathways for selection of 5′ splice sites by U1 snRNPs and SF2/ASF. *EMBO J.*, **12**, 3607–3617.
49. Gutell,R.R., Cannone,J.J., Shang,Z., Du,Y. and Serra,M.J. (2000) A story: unpaired adenosine bases in ribosomal RNAs. *J. Mol. Biol.*, **304**, 335–354.
50. Nissen,P., Ippolito,J.A., Ban,N., Moore,P.B. and Steitz,T.A. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl Acad. Sci. USA*, **98**, 4899–4903.

51. Schwartz,S., Gal-Mark,N., Kfir,N., Oren,R., Kim,E. and Ast,G. (2009) Alu exonization events reveal features required for precise recognition of exons by the splicing machinery. *PLoS Comput. Biol.*, **5**, e1000300.
52. Lev-Maor,G., Ram,O., Kim,E., Sela,N., Goren,A., Levanon,E.Y. and Ast,G. (2008) Intronic Alus influence alternative splicing. *PLoS Genet.*, **4**, e1000204.
53. Sorek,R., Lev-Maor,G., Reznik,M., Dagan,T., Belinky,F., Graur,D. and Ast,G. (2004) Minimal conditions for exonization of intronic sequences: 5′ splice site formation in Alu exons. *Mol. Cell*, **14**, 221–231.
54. Lev-Maor,G., Sorek,R., Shomron,N. and Ast,G. (2003) The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science*, **300**, 1288–1291.
55. Ast,G. (2004) How did alternative splicing evolve? *Nat Rev. Genet.*, **5**, 773–782.
56. Iwata,H. and Gotoh,O. (2011) Comparative analysis of information contents relevant to recognition of introns in many species. *BMC Genomics*, **12**, 45.
57. Palaniswamy,R., Teglund,S., Lauth,M., Zaphiropoulos,P.G. and Shimokawa,T. (2010) Genetic variations regulate alternative splicing in the 5′ untranslated regions of the mouse glioma-associated oncogene 1, Gli1. *BMC Mol Biol.*, **11**, 32.
58. Tsai,K.W., Tseng,H.C. and Lin,W.C. (2008) Two wobble-splicing events affect ING4 protein subnuclear localization and degradation. *Exp. Cell Res.*, **314**, 3130–3141.
59. Flomen,R., Knight,J., Sham,P., Kerwin,R. and Makoff,A. (2004) Evidence that RNA editing modulates splice site selection in the 5-HT2C receptor gene. *Nucleic Acids Res.*, **32**, 2113–2122.
60. Rueter,S.M., Dawson,T.R. and Emeson,R.B. (1999) Regulation of alternative splicing by RNA editing. *Nature*, **399**, 75–80.
61. Lev-Maor,G., Sorek,R., Levanon,E.Y., Paz,N., Eisenberg,E. and Ast,G. (2007) RNA-editing-mediated exon evolution. *Genome Biol.*, **8**, R29.
62. Kiran,A. and Baranov,P.V. (2010) DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics*, **26**, 1772–1776.
63. Fisher,C.W., Fisher,C.R., Chuang,J.L., Lau,K.S., Chuang,D.T. and Cox,R.P. (1993) Occurrence of a 2-bp (AT) deletion allele and a nonsense (G-to-T) mutant allele at the E2 (DBT) locus of six patients with maple syrup urine disease: multiple-exon skipping as a secondary effect of the mutations. *Am. J. Hum. Genet.*, **52**, 414–424.
64. Schwarze,U., Starman,B.J. and Byers,P.H. (1999) Redefinition of exon 7 in the *COL1A1* gene of type I collagen by an intron 8 splice-donor-site mutation in a form of osteogenesis imperfecta: influence of intron splice order on outcome of splice-site mutation. *Am. J. Hum. Genet.*, **65**, 336–344.
65. Hicks,M.J., Mueller,W.F., Shepard,P.J. and Hertel,K.J. (2010) Competing upstream 5′ splice sites enhance the rate of proximal splicing. *Mol. Cell. Biol.*, **30**, 1878–1886.
66. Nemeroff,M.E., Utans,U., Kramer,A. and Krug,R.M. (1992) Identification of cis-acting intron and exon regions in influenza virus NS1 mRNA that inhibit splicing and cause the formation of aberrantly sedimenting presplicing complexes. *Mol. Cell. Biol.*, **12**, 962–970.
67. Alvarez,C.J. and Wise,J.A. (2001) Activation of a cryptic 5′ splice site by U1 snRNA. *RNA*, **7**, 342–350.
68. Feldhahn,N., Rio,P., Soh,B.N., Liedtke,S., Sprangers,M., Klein,F., Wernet,P., Jumaa,H., Hofmann,W.K., Hanenberg,H. *et al.* (2005) Deficiency of Bruton′s tyrosine kinase in B cell precursor leukemia cells. *Proc. Natl Acad. Sci. USA*, **102**, 13266–13271.
69. Raponi,M., Kralovicova,J., Copson,E., Divina,P., Eccles,D., Johnson,P.M., Baralle,D. and Vorechovsky,I. (2011) Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in *BRCA1* exon 5. *Hum. Mutat.*, **32**, 436–444.
70. Churbanov,A., Vorechovsky,I. and Hicks,C. (2010) A method of predicting changes in human gene splicing induced by genetic variants in context of *cis*-acting elements. *BMC Bioinformatics*, **11**, 22.
71. Ramchatesingh,J., Zahler,A.M., Neugebauer,K.M., Roth,M.B. and Cooper,T.A. (1995) A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer. *Mol. Cell. Biol.*, **15**, 4898–4907.
72. Lynch,K.W. and Maniatis,T. (1996) Assembly of specific SR protein complexes on distinct regulatory elements of the *Drosophila doublesex* splicing enhancer. *Genes Dev.*, **10**, 2089–2101.
73. Tacke,R., Tohyama,M., Ogawa,S. and Manley,J.L. (1998) Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing. *Cell*, **93**, 139–148.
74. Nasim,M.T., Chernova,T.K., Chowdhury,H.M., Yue,B.G. and Eperon,I.C. (2003) HnRNP G and Tra2beta: opposite effects on splicing matched by antagonism in RNA binding. *Hum. Mol. Genet.*, **12**, 1337–1348.
75. Hofmann,Y., Lorson,C.L., Stamm,S., Androphy,E.J. and Wirth,B. (2000) Htra2-beta 1 stimulates an exonic splicing enhancer and can restore full-length SMN expression to survival motor neuron 2 (SMN2). *Proc. Natl Acad. Sci. USA*, **97**, 9618–9623.
76. Tsuda,K., Someya,T., Kuwasako,K., Takahashi,M., He,F., Unzai,S., Inoue,M., Harada,T., Watanabe,S., Terada,T. *et al.* (2010) Structural basis for the dual RNA-recognition modes of human Tra2-{beta} RRM. *Nucleic Acids Res.*, **39**, 1538–1553.
77. Clery,A., Jayne,S., Benderska,N., Dominguez,C., Stamm,S. and Allain,F.H. (2011) Molecular basis of purine-rich RNA recognition by the human SR-like protein Tra2-beta1. *Nat. Struct. Mol. Biol.*, **18**, 443–450.
78. Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.
79. Doherty,E.A., Batey,R.T., Masquida,B. and Doudna,J.A. (2001) A universal mode of helix packing in RNA. *Nat. Struct. Biol.*, **8**, 339–343.
80. McCullough,A.J. and Berget,S.M. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol. Cell. Biol.*, **17**, 4562–4571.
81. Kanopka,A., Muhlemann,O. and Akusjarvi,G. (1996) Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature*, **381**, 535–538.
82. Stoilov,P., Daoud,R., Nayler,O. and Stamm,S. (2004) Human tra2-beta1 autoregulates its protein concentration by influencing alternative splicing of its pre-mRNA. *Hum. Mol. Genet.*, **13**, 509–524.
83. Buratti,E., Stuani,C., De Prato,G. and Baralle,F.E. (2007) SR protein-mediated inhibition of CFTR exon 9 inclusion: molecular characterization of the intronic splicing silencer. *Nucleic Acids Res.*, **35**, 4359–4368.
84. Makalowski,W., Mitchell,G.A. and Labuda,D. (1994) Alu sequences in the coding regions of mRNA: a source of protein variability. *Trends Genet.*, **10**, 188–193.
85. Lei,H., Day,I.N.M. and Vorechovsky,I. (2005) Exonization of AluYa5 in the human ACE gene requires mutations in both 3′ and 5′ splice sites and is facilitated by a conserved splicing enhancer. *Nucleic Acids Res.*, **33**, 3897–3906.
86. Vorechovsky,I. (2006) Aberrant 3′ splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res.*, **34**, 4630–4641.
87. Fairbrother,W.G., Yeh,R.F., Sharp,P.A. and Burge,C.B. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.
88. Fairbrother,W.G., Yeo,G.W., Yeh,R., Goldstein,P., Mawson,M., Sharp,P.A. and Burge,C.B. (2004) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.*, **32**, W187–W190.
89. Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q. and Krainer,A.R. (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.
90. Smith,P.J., Zhang,C., Wang,J., Chew,S.L., Zhang,M.Q. and Krainer,A.R. (2006) An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum. Mol. Genet.*, **15**, 2490–2508.
91. Wollerton,M.C., Gooding,C., Wagner,E.J., Garcia-Blanco,M.A. and Smith,C.W. (2004) Autoregulation of polypyrimidine tract

binding protein by alternative splicing leading to nonsense-mediated decay. *Mol. Cell*, **13**, 91–100.

92. Spellman,R., Llorian,M. and Smith,C.W. (2007) Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. *Mol. Cell*, **27**, 420–434.
93. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
94. Steijlen,P.M., van Steensel,M.A., Jansen,B.J., Blokx,W., van de Kerkhof,P.C., Happle,R. and van Geel,M. (2004) Cryptic splicing at a non-consensus splice-donor in a patient with a novel mutation in the plakophilin-1 gene. *J. Invest. Dermatol.*, **122**, 1321–1324.
95. Spena,S., Duga,S., Asselta,R., Malcovati,M., Peyvandi,F. and Tenchini,M.L. (2002) Congenital afibrinogenemia: first identification of splicing mutations in the fibrinogen Bbeta-chain gene causing activation of cryptic splice sites. *Blood*, **100**, 4478–4484.
96. Harland,M., Taylor,C.F., Chambers,P.A., Kukalizch,K., Randerson-Moor,J.A., Gruis,N.A., de Snoo,F.A., ter Huurne,J.A., Goldstein,A.M., Tucker,M.A. *et al.* (2005) A mutation hotspot at the p14ARF splice site. *Oncogene*, **24**, 4604–4608.
97. Tanioka,M., Budiyant,A., Ueda,T., Nagano,T., Ichihashi,M., Miyachi,Y. and Nishigori,C. (2005) A novel XPA gene mutation and its functional analysis in a Japanese patient with xeroderma pigmentosum group A. *J. Invest. Dermatol.*, **125**, 244–246.
98. Urban,Z., Michels,V.V., Thibodeau,S.N., Donis-Keller,H., Csiszar,K. and Boyd,C.D. (1999) Supravalvular aortic stenosis: a splice site mutation within the elastin gene results in reduced expression of two aberrantly spliced transcripts. *Hum. Genet.*, **104**, 135–142.
99. Goyette,P., Frosst,P., Rosenblatt,D.S. and Rozen,R. (1995) Seven novel mutations in the methylenetetrahydrofolate reductase gene and genotype/phenotype correlations in severe methylenetetrahydrofolate reductase deficiency. *Am. J. Hum. Genet.*, **56**, 1052–1059.
100. Wimmer,K., Roca,X., Beiglbock,H., Callens,T., Etzler,J., Rao,A.R., Krainer,A.R., Fonatsch,C. and Messiaen,L. (2007) Extensive *in silico* analysis of NF1 splicing defects uncovers determinants for splicing outcome upon 5′ splice-site disruption. *Hum. Mutat.*, **28**, 599–612.
101. Hong,K., Guerchicoff,A., Pollevick,G.D., Oliva,A., Dumaine,R., de Zutter,M., Burashnikov,E., Wu,Y.S., Brugada,J., Brugada,P. *et al.* (2005) Cryptic 5′ splice site activation in SCN5A associated with Brugada syndrome. *J. Mol. Cell. Cardiol.*, **38**, 555–560.
102. Ars,E., Serra,E., Garcia,J., Kruyer,H., Gaona,A., Lazaro,C. and Estivill,X. (2000) Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. *Hum. Mol. Genet.*, **9**, 237–247.
103. Roca,X., Sachidanandam,R. and Krainer,A.R. (2005) Determinants of the inherent strength of human 5′ splice sites. *RNA*, **11**, 683–698.
104. Divina,P., Kvitkovicova,A. and Vorechovsky,I. (2009) *Ab initio* prediction of cryptic splice-site activation and exon skipping. *Eur. J. Hum. Genet.*, **17**, 759–765.