Human Splicing Finder: an online bioinformatics tool to predict splicing signals.

François-Olivier Desmet, Dalil Hamroun, Marine Lalande, Gwenaëlle Collod-Béroud, Mireille Claustres, Christophe Béroud

Human Splicing Finder: an online bioinformatics tool to predict splicing signals

François-Olivier Desmet[1], Dalil Hamroun[1,2], Marine Lalande[1], Gwenaëlle Collod-Béroud[1], Mireille Claustres[1,2,3] and Christophe Béroud[1,2,3,*]

[1]INSERM, U827, [2]CHU Montpellier, Hôpital Arnaud de Villeneuve, Laboratoire de Génétique Moléculaire and [3]Université Montpellier1, UFR Médecine, Montpellier, F-34000, France

ABSTRACT

Thousands of mutations are identified yearly. Although many directly affect protein expression, an increasing proportion of mutations is now believed to influence mRNA splicing. They mostly affect existing splice sites, but synonymous, non-synonymous or nonsense mutations can also create or disrupt splice sites or auxiliary *cis*-splicing sequences. To facilitate the analysis of the different mutations, we designed Human Splicing Finder (HSF), a tool to predict the effects of mutations on splicing signals or to identify splicing motifs in any human sequence. It contains all available matrices for auxiliary sequence prediction as well as new ones for binding sites of the 9G8 and Tra2-β Serine-Arginine proteins and the hnRNP A1 ribonucleoprotein. We also developed new Position Weight Matrices to assess the strength of 5′ and 3′ splice sites and branch points. We evaluated HSF efficiency using a set of 83 intronic and 35 exonic mutations known to result in splicing defects. We showed that the mutation effect was correctly predicted in almost all cases. HSF could thus represent a valuable resource for research, diagnostic and therapeutic (e.g. therapeutic exon skipping) purposes as well as for global studies, such as the GEN2PHEN European Project or the Human Variome Project.

INTRODUCTION

Since its discovery more than three decades ago (1), mRNA splicing is the focus of many studies both in fundamental and applied research. Splicing is part of the pre-mRNA maturation process that occurs in each eukaryotic cell between mRNA transcription from DNA and its translation into protein. During this event, parts of the pre-mRNA transcripts are removed in a ribonucleoprotein complex (spliceosome) which is constituted of five essential small nuclear RNAs and more than 150 polypeptides (2,3). Depending on tissue localization and/or stage of development, pre-mRNA transcripts may be differentially spliced, allowing several transcripts to be built and thus different proteins to be synthesized from the same gene. A prime example of this phenomenon is the *Troponin T* gene for which 64 different mRNAs have been described (4). This process is called alternative splicing and it is estimated that more than 70% of human protein-coding genes are alternatively spliced (5). Understanding how splicing is regulated is thus crucial, particularly in a medical context, since genomic variations which cause aberrant splicing may represent up to 50% of all mutations that lead to gene dysfunction (6). Mutations can indeed not only alter directly the sequence that will be translated into protein, for instance, base substitutions can change a codon for an amino acid into another one or into a premature termination codon (PTC), but can also affect splicing and, as a consequence, lead to the appearance of truncated proteins or to the lack of the correct gene product.

How are exons and introns recognized during the splicing process? Exon definition (7) is the identification of splice sites located at the 5′ and 3′ ends of exon–intron–exon junctions (5′ss and 3′ss also known as donor and acceptor splice site, respectively). At 3′ end of introns, a branch point sequence and a polypyrimidine tract, which are situated upstream the 3′ss, are also used as consensus elements. These consensus sequences have probably evolved from ancestral common sequences as it has been reported for the 5′ site with the AG/guaagu prototype sequence whose eight contiguous nucleotides are complementary to nucleotides 4–11 of U1RNA (8). The divergence of splice site sequences from the prototypes has been closely associated with the creation of alternative transcripts. Moreover, in higher eukaryotes, these highly degenerated motifs can also be found in most introns, framing pseudo-exons. Pseudo-exons are intronic sequences of typical exon size that outnumber real exons and are bounded by sequences that match the 5′ and 3′

*To whom correspondence should be addressed. Tel: +33 4 67 41 53 60; Fax: +33 4 67 41 53 65; Email: christophe.beroud@inserm.fr

splicing signal requirements of an exon, but that are never considered as proper exons by the spliceosome. Furthermore, human transcripts contain many 'decoy' splice sites that are seldom used. So, while 5′ and 3′ splicing signals are mandatory for exon definition, they are not sufficient for correct splicing. In order to reliably distinguish authentic exons and splice sites from pseudo-exons and decoy splice sites, the splicing machinery must rely on auxiliary sequence features, such as intronic and exonic *cis*-elements. Among them, the Exonic Splicing Enhancers (ESEs) are the most studied. They are specific short nucleotide sequences that are targeted essentially by Serine/Argine-rich (SR) proteins which then promote exon definition (9). Conversely, the Exonic Splicing Silencers (ESSs) help the spliceosome to ignore pseudo-exons and decoy splice sites. They act as binding sites for proteins promoting exon exclusion (mainly hnRNP proteins) (10). Intronic Splicing Enhancers (ISEs) and Intronic Splicing Silencers (ISSs) are intronic *cis*-elements that play similar roles as ESEs and ESSs.

Several bioinformatics tools to study or predict splice signals have been developed and are today available online. Their approaches can vary (11) from using blastn to align a query sequence to a database of alternative splicing events and splice signals (12) to an *ab initio* prediction approach (13). Despite the quality of these tools and because of the complexity of sequence signals harbored by any mRNA sequence, new tools are needed to simultaneously identify putative donor and acceptor splice sites, branch points and *cis*-acting elements (ESE, ESS, ISE and ISS). In addition, since many human disease-causing mutations affect splicing, new bioinformatics tools should also be able to predict the consequence of mutations on splice signals. Such tool could be of great value not only for geneticists to better understand splicing events and the effect of mutations on mRNA splicing, but also for clinical researchers to design new therapeutic approaches based on splicing interference, such as the exon-skipping strategy used in Duchenne Muscular Dystrophy (DMD) (14) or gene and exon silencing through manipulation of mRNA splicing (15).

In this article, we present a new bioinformatics tool, the Human Splicing Finder (HSF) software that is freely available online (http://www.umd.be/HSF/). It includes new algorithms derived from the Universal Mutation Database (UMD) (16,17) to allow the evaluation of the strength of 5′ss, 3′ss and branch points. In addition, in order to identify *cis*-acting elements it includes already published algorithms, such as the RESCUE-ESE (18) and ESE-Finder (19) as well as new algorithms designed to use available or newly created matrices. To allow the study of virtually any human sequence, HSF includes all genes and alternative transcripts as well as intronic sequences that were extracted from the Ensembl human genome database (http://www.ensembl.org/) (20). To evaluate the predictive potential of HSF web interface (version 2.4; http://www.umd.be/HSF/), we used a set of mutations for which the effect on splicing has been experimentally demonstrated.

MATERIAL AND METHODS

Software development and database design

HSF was developed using the 4D package (4D S.A.) for data management, algorithm design and web interface. The HSF database was designed to include the introns and exons of all human genes. It was constructed from an Ensembl dataset (20) containing more than 22 000 genes and 46 000 transcripts of *Homo sapiens* (release 44, http://april2007.archive.ensembl.org) using Biomart (20). Genes were created from the crude dataset using both Ensembl transcript coordinates and sequences from the UCSC genome browser database (21). At present, HSF database only contains human genes, since matrices and tools were specifically designed for the human genome.

To study the potential effects of single nucleotide polymorphisms (SNPs) on splicing, HSF also harbors data extracted from the Ensembl Variation database (20). For this, a Perl script was developed, using Ensembl Perl API that allows HSF to directly query the Ensembl Variation database and retrieve SNPs located in human genes.

Splicing donor/acceptor sites

To predict potential 5′ss and 3′ss, we used matrices derived from Shapiro and Senapathy (22). A potential splice site is defined as an *n*-mer sequence. For each 'n' position, a weight is given to each nucleotide, based on its frequency and the relative importance of its position in the sequence motif (position weight matrices, PWM). The strength of a site is thus defined as the sum of each nucleotide's weight plus a constant (Equation 1) that is used for normalization. Only *n*-mer sequences with consensus values (CV) higher or equal to a given threshold are considered as potential 5′ or 3′ss.

Since the human 5′ consensus sequence is [C/A]AG/gt[a/g]agt, we defined the 5′ss as a 9-mer matrix. Similarly, the 3′ss was defined as a 14-mer matrix. Calculation of the strength of a potential splice site. For 5′ss $x = 9$ and for 3′ss $x = 14$.

$$Site\_strength = Base\_value + \sum_{i=1}^{x} nucleotide\_value(i)$$

HSF also includes an algorithm adapted from the MaxEnt script (23) that allows the analysis of a whole sequence. In addition, for this matrix, users can define thresholds for splice site prediction.

Branch point sequences

Since the human branch point (BP) consensus sequence is YNYCRAY (24), we defined the BP sequence as a 7×4 position weight matrix (Figure 1). The threshold for BP sequences was fixed at 67. The strength of a BP sequence was thus calculated as follows (Equation 2):

$$BPSite\_strength = Base\_value + \sum_{i=1}^{7} nucleotide\_value(i)$$

Figure 1. Branch point matrix. The size of each nucleotide is proportional to its weight in the position weight matrix. Nucleotides above the base line have positive values while nucleotides below have negative values.

Since many intronic sequences match the BP consensus sequence, we included the AG-Exclusion Zone algorithm described by Gooding *et al*. (25) to predict BP candidates. For a given intronic sequence and its intron-exon boundary, HSF searches all AG dinucleotides that are included in a 3′ss candidate sequence (threshold of 67) and therefore define the exclusion zones. As it has been shown that the BP allows the recognition of the first downstream 3′ss, HSF annotates the functional BP as the strongest candidate without a 3′-exclusion zone before the natural 3′ss.

Additionally, to take into account the steric obstruction caused by the spliceosome, we excluded BP sequences located at less than 12 nt from the exon. Finally, as most BP sequences are located between −21 and −34 nt from the exon (26), only a window of 100 bp is processed. We arbitrarily excluded the probability of having a BP motif located very far away in order to save computation time.

Matrices for splicing enhancers and silencers

To maximize the detection of auxiliary motifs, HSF integrated: (i) matrices for SR proteins (SRp40, SC35, SF2/ASF, SF2/ASF IgM/BRCA1 and SRp55) from the ESE Finder tool (19,27); (ii) sequence motifs shown to be differentially present in exons and introns, such as the RESCUE-ESE hexamers (18), the putative 8-mer ESE and ESS identified by Zhang and Chasin (28), the ESR sequences identified by Goren and co-workers (29) and the exon-identity elements (EIE) and intron-identity elements (IIE) defined by Zhang and co-workers (30). For the silencer sequences identified by Sironi and colleagues (31) and the ESS decamers (32), for which no web-based tool were available, we developed new algorithms to use the crude data.

New matrices were also created to predict hnRNP A1, Tra2-β and 9G8 protein binding motifs. These matrices were designed using published data collected from SELEX experiments and consensus sequences. Sequences were aligned with *ClustalW* (33) to generate a consensus motif. Note that these motifs were too short to be processed with *MEME* (34). The consensus sequences were then used to design PWM matrices (Figure 2).

Sequence datasets used to evaluate HSF efficiency

To evaluate the new algorithms dedicated to the prediction of 5′ss and 3′ss, we used the Ensembl database

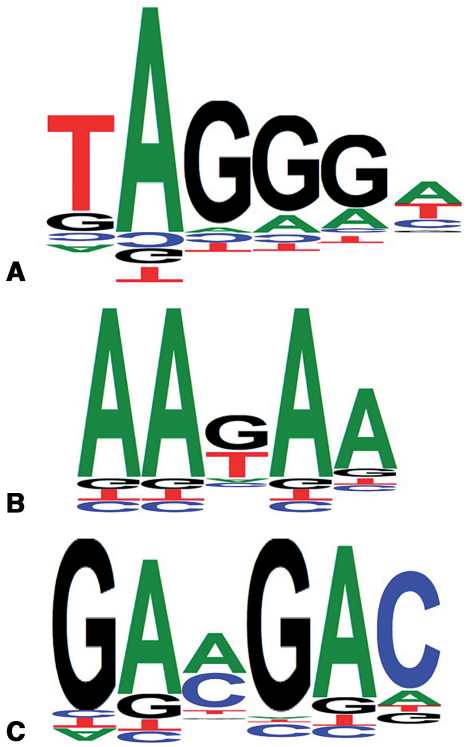

Figure 2. New position weight matrices of recognition motifs for proteins involved in splicing. (**A**) hnRNP A1; (**B**) Tra2-β and (**C**) 9G8.

(20) that contain 245.286 human exons (release 44, http://april2007.archive.ensembl.org). For BP predictions, we used a set of 14 experimentally validated BPs (Table 3). These datasets were completed by 69 intronic mutations (35–56) as well as 15 exonic mutations known to alter 5′ and 3′ss (57,58) and for whom the impact on mRNA splicing has been characterized *in vivo* or *in vitro*. To evaluate the ability to correctly predict ESE and ESS, we used a set of 20 experimentally validated mutations that affect splicing by a direct effect on ESE and/or ESS (58–66). In addition we used a set of 36 mutations previously reported to alter splicing (positive controls) and 220 SNPs (negative controls). The negative controls were extracted from the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) and corresponded to SNPs with the highest minor allele frequency and, therefore, had a minimal risk of affecting splicing. Conversely, the positive controls were chosen because experimental results showed that these mutations targeted auxiliary splicing sequence motifs. Nevertheless, in most cases the data about the exact motif and/or the protein that recognizes this motif were not available. For each mutation, we evaluated only its effects in terms of disruption of ESE or creation of ESS signals (Supplementary Table 1).

RESULTS

Web interface and database

HSF web interface was designed to maximize the perception of efficiency and easy of use by end users.

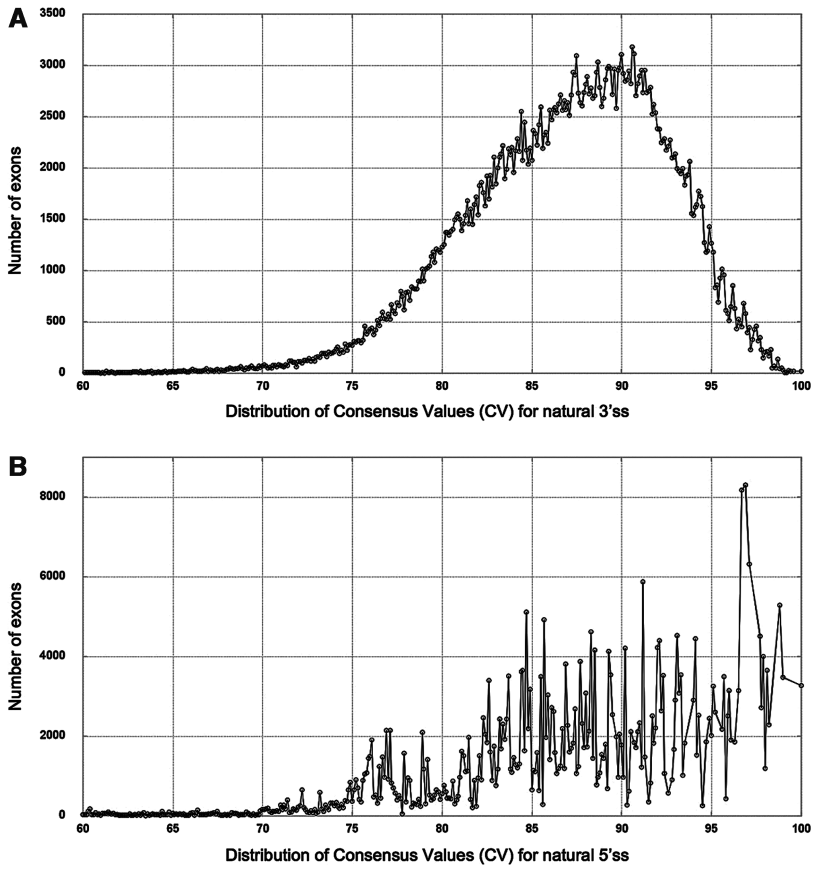

Figure 3. Distribution of CVs for (**A**) 3′ and (**B**) 5′ natural splice sites (5′ss and 3′ss). Data extracted from the Ensembl dataset (release 44, http://april2007.archive.ensembl.org) (20) using the HSF algorithm.

Only default parameters are displayed on the submission form while skilled users can easily access advanced parameters. Sequences stored in the database can be queried using either the gene symbol, the Ensembl gene ID, the Ensembl transcript ID, the RefSeq peptide ID or the consensus CDS. In addition, users can process their own sequences, either for simple sequence analysis or mutant comparison. In addition, HSF can be queried in different ways: full analysis of a sequence, comparison of a mutant and a wild-type sequence, or simultaneous analysis of several mutants related, or not, to the same transcript. In this case all mutations should be referred to sequences included in the HSF database. In order to easily study a group of mutations from different genes and transcripts, the mutation must be described by using the international nomenclature system for cDNA mutations (67) (http://www.genomic.unimelb.edu.au/mdi/mutnomen/). HSF will then check that each mutation is correctly described and automatically reconstruct the mutant allele from the wild-type sequence and the mutation name. Since only small rearrangements (i.e. substitutions, small exonic or intronic deletions and insertions, duplications and indels) provide useful information about splicing defects, large rearrangements can not be processed by HSF.

Moreover, differently from previous resources, the user can specifically analyze BP sequences or splice site motifs using HSF specific matrices and algorithms.

The main result page was divided in three areas: the reference sequence(s), various graphical displays and tables. Since mutations could have different effects related to the local context, a 'quick mutation' option allows the addition of a small rearrangement (missense, deletion, insertion, duplication, indel) to the sequence(s).

Splicing donor/acceptor sites

The new HSF algorithm to define consensus values (CV) of 5′ss or 3′ss was created to maximize the difference between wild-type (wt) active sites and mutant inactive sites. Thus, strong sites presented a CV higher than 80 and less strong sites a CV ranging between 70 and 80. Only a minor fraction of active sites showed a CV between 65 and 70 (Figure 3). The mean CV for 3′ss was 86.81 with a standard deviation of 6.33 while the mean CV for 5′ss was 87.53 with a standard deviation of 8.34. These values were calculated from more than 400 000 natural splice sites extracted from all alternative transcripts. If a mutation affects directly the CV, it is critical to consider not only the CV of the mutant splice site, but also the delta between the wt and mutant CV. To validate this algorithm, we used a set of 69 intronic mutations that affect either the canonical AG/GT splice site motifs or less conserved nucleotides (Table 1). All mutations affecting the nucleotides in canonical positions (−2, −1, +1 or +2) strongly influenced the CV value with an average

Table 1. Intronic mutations in *FBN1* (ENST00000316623), *FBN2* (ENST00000262464), *RB1* (ENST00000267163), *TGFBR2* (ENST00000295754), *MLH1* (ENST00000231790) and *MSH2* (ENST00000233146) that lead to splicing defects

Gene	Mutation	References	WT CV	Mutant CV	CV variation (%)
Mutations causing exon skipping					
FBN1	c.247 + 1G>A	(37,46,49–51,53)	82.26	55.42	−32.62[a]
FBN1	c.538 + 1G>A	(45)	83.99	57.15	−31.96[a]
FBN1	c.1468 + 5G>A	(44)	84.46	72.30	−14.40[a]
FBN1	c.3208 + 5G>T	(82)	94.98	82.66	−12.97[a]
FBN1	c.3838 + 1G>A	(52)	95.84	69.01	−28.00[a]
FBN1	c.3839 − 1G>T	(83)	87.62	58.67	−33.04[a]
FBN1	c.3964 + 1G>A	(84,85)	90.04	63.20	−29.80[a]
FBN1	c.3965 − 2A>T	(85)	89.30	60.35	−32.41[a]
FBN1	c.4459 + 1G>A	(44)	97.66	70.83	−27.47[a]
FBN1	c.4943 − 1G>C	(44)	79.77	50.82	−36.29[a]
FBN1	c.5788 + 5G>A	(35,36,38,41,43,52,54,83)	88.06	75.89	−13.82[a]
FBN1	c.6163 + 2del6	(83)	99.05	72.90	−26.40[a]
FBN1	c.6496 + 2insTG	(43)	82.21	32.05	−61.01[a]
FBN1	c.6616 + 1G>C	(86)	78.08	51.24	−34.37[a]
FBN1	c.6997 + 1G>A	(83)	92.11	65.27	−29.13[a]
FBN1	c.7205 − 2A>G	(83)	84.11	55.16	−34.42[a]
FBN1	c.7330 + 1G>A	(55)	98.02	71.18	−27.38[a]
FBN1	c.7331 − 2A>G	(40)	80.72	51.77	−35.86[a]
FBN1	c.8051 + 1G>A	(44)	92.02	65.18	−29.16[a]
FBN1	c.8051 + 5G>A	(51)	92.02	79.85	−13.22[a]
FBN1	c.8052 − 2A>G	(52)	92.86	63.92	−31.17[a]
FBN2	c.3472 + 2T>G	(48)	90.99	64.15	−28.53[a]
FBN2	c.4099 + 1G>C	(39)	91.66	64.82	−29.28[a]
FBN2	c.4222 + 5G>A	(47)	92.11	79.94	−13.21[a]
FBN2	c.4346 − 2A>T	(87)	90.91	61.96	−31.84[a]
RB1	c.264 + 4delA	(57)	91.34	84.93	−7.01[a]
RB1	c.380 + 3A>C	(57)	95.10	78.82	−17.12[a]
RB1	c.607 + 1G>T	(57)	99.05	72.21	−27.09[a]
RB1	c.939 + 4A>G	(57)	83.75	75.41	−9.96[a]
RB1	c.1049 + 2delT	(57)	76.95	57.00	−25.90[a]
RB1	c.1215 + 1G>A	(57)	85.86	59.02	−31.26[a]
RB1	c.1389 + 1G>A	(57)	82.69	55.86	−32.45[a]
RB1	c.1389 + 4A>G	(57)	82.69	74.35	−10.09[a]
RB1	c.1389 + 5G>A	(57)	82.69	70.53	−14.71[a]
RB1	c.1422 − 2A>T	(57)	86.12	57.17	−33.62[a]
RB1	c.1422 − 1G>A	(57)	86.12	57.17	−33.62[a]
RB1	c.1498 + 5G>A	(57)	82.91	70.75	−14.67[a]
RB1	c.1960 + 1G>A	(57)	94.02	67.19	−28.54[a]
RB1	c.1960 + 1delG	(57)	94.02	49.62	−47.22[a]
RB1	c.2211 + 1G>T	(57)	89.90	63.06	−29.86[a]
RB1	c.2212 − 2A>G	(57)	89.09	60.15	−32.48[a]
RB1	c.2211 + 1G>C	(57)	89.90	63.06	−29.86[a]
RB1	c.2520 + 1G>A	(57)	92.22	65.39	−29.10[a]
RB1	c.2520 + 3del4	(57)	92.22	72.26	−21.64[a]
RB1	c.2663 + 1G>A	(57)	88.37	61.54	−30.36[a]
MLH1	c.306 + 4A>G	(58)	96.07	87.73	−8.68[a]
MLH1	c.454 − 2A>G	(59)	93.59	64.64	−28.80[a]
MLH1	c.790 + 1G>A	(59)	83.28	56.45	−32.22[a]
MLH1	c.790 + 5G>T	(58)	83.28	70.97	−14.79[a]
MLH1	c.791 − 5T>G	(59)	80.80	77.17	−4.49
MLH1	c.884 + 4A>G	(58)	85.75	77.41	−9.73[a]
MSH2	c.366 + 1G>T	(59)	86.73	59.89	−30.95[a]
MSH2	c.793 − 2A>C	(59)	83.98	55.04	−34.46[a]
MSH2	c.942 + 3A>T	(59)	99.24	83.86	−15.50[a]
MSH2	c.1276 + 2T>A	(59)	84.70	57.86	−31.69[a]
MSH2	c.1386 + 1G>A	(59)	89.02	62.19	−30.13[a]
MSH2	c.2634 + 5G>T	(58)	84.41	72.09	−14.59[a]
Mutations resulting in the usage of cryptic splice sites					
FBN1	c.2293 + 2T>C	(83)	89.77	62.94	−29.89
			67.21		5′ CS (51 nt upstream)[b]
FBN1	c.3463 + 1G>A	(88)	91.34	64.50	−29.38
			88.47		5′ CS (27 nt downstream)[b]
FBN1	c.4747 + 5G>T	(42)	89.13	76.81	−13.82
			79.06		5′ CS (48 nt upstream)[b]
FBN1	c.5788 + 1G>A	(52)	88.06	61.22	−30.48
			82.64		5′ CS (33 nt downstream)[b]
RB1	c.138 − 8T>G	(57)	81.62	79.69	−2.36

(continued)

Table 1. Continued

Gene	Mutation	References	WT CV	Mutant CV	CV variation (%)
			55.35	84.29	3′ CS (7 nt upstream)[c]
RB1	c.501 − 1G>A	(57)	97.50	68.55	−29.69
			54.82	83.77	3′ CS (1 nt downstream)[c]
RB1	c.607 + 1delG	(57)	99.05	22.54	−77.24
			42.51	88.47	5′ CS (1 nt upstream)[c]
RB1	c.1815 − 2A>G	(57)	75.42	46.47	−38.39
			81.84		3′ CS (19 nt downstream)[b]
RB1	c.2107 − 2A>G	(57)	80.73	51.78	−35.86
			69.56		3′ CS (35 nt downstream)[b]
TGFBR2	c.95 − 2A>G	(56)	91.77	62.82	−31.55
			68.28		3′ CS (18 nt downstream)[b]
TGFBR2	c.1397 − 2A>G	(89)	92.32	63.38	−31.35
			84.32		3′ CS (30 nt upstream)[b]
TGFBR2	c.1397 − 1G>A	(90)	92.32	63.38	−31.35
			84.32		3′ CS (30 nt upstream)[b]

CS: cryptic site (i.e. a new splice site is created by the mutation and is used instead of the regular site). Nucleotide numbering follows the reference cDNA sequence with +1 corresponding to the A of the ATG translation initiation codon.
[a]The mutation induces exon skipping.
[b]A cryptic splice site not created by the mutation and used *in vivo* was correctly predicted by HSF.
[c]The cryptic splice site created by the mutation and used *in vivo* was correctly predicted by HSF.

reduction (ΔCV) of 31% and a standard deviation (SD) of 2.8%. Mutations affecting less conserved residues had a weaker effect with a ΔCV of 7% for the residue in position +4 and 14% for nucleotides in position +3 or +5. These results together with data from other disease-causing mutations (52,68,69) indicated that a ΔCV reduction of at least 10% for a mutation in any position or of 7% for a mutation in position +4 is likely to have a significant impact on splicing and should be further investigated.

Since a mutation can result not only in the disruption of a 5′ss or a 3′ss, but also in the creation of a new splice site, HSF evaluates the 'creation of cryptic splice sites'. As shown in Table 1 for intronic mutations, HSF correctly predicted the creation of cryptic splice sites in the *RB1* mutants c.607 + 1delG, c.138-8T>G and c.501-1G>A. Mutations in canonical sequences, such as c.95-2A>G, c.1397-2A>G and c.1397-1G>A in *TGFBR2*, c.2293 + 2T>C, c.3463 + 1G>A, c.4747 + 5G>T and c.5788 + 1G>A in *FBN1* and c.1815-2A>G, c.2107-2A>G and c.2211 + 1G>C in *RB1*, led to a more complex splicing defect in which disruption of the wt splice site was coupled to the usage of an alternative, pre-existing splice site. As mutations do not directly affect alternative splice sites, this phenomenon was not automatically investigated by HSF. Therefore, to identify the alternative splice sites, we chose in '*Select an analysis type*' the option '*Number of nucleotides surrounding the exon*' and entered the value '*100*'. In addition, we checked the advanced parameter '*Process sequence*' and selected the '*Full sequence*' option. To analyze only splice sites, we then selected in '*All or subset of matrices*' the '*Splice site matrices*' option. Using these parameters, all alternative sites were identified either as the closest and strongest alternative sites (five cases) or as the second-best sites (two cases). Overall, HSF correctly predicted the impact of mutations affecting 5′ss or 3′ss, even when complex mechanisms were involved.

In addition to splicing defects due to 5′ss or 3′ss disruption, it is well known that exonic mutations could result in the creation or activation of cryptic splice sites. As shown in Table 2, the nine mutations affecting the last base of an exon had a strong effect on the activity of the concerned 5′ss (ΔCV = 12% ± 0.7) that resulted in exon skipping or activation of a cryptic splice site. The two mutations affecting the penultimate nucleotide of an exon had a limited effect on the activity of the 5′ss (ΔCV = 5.4% ± 0.3). Indeed, these mutations were pathogenic only when a cryptic splice site was activated and therefore predictions were hazardous. Finally, exonic mutations that were distant both from the 5′ and 3′ss could activate a cryptic splice site and result in splicing defects as shown for mutations c.658C>G in *RB1*, c.1915C>T in *MSH2* and c.5985T>G in *DMD*.

Branch point sequences

We analyzed 14 BP sequences previously reported to be abolished by mutations. As shown in Table 3, 13 out of 14 BPs were correctly predicted by HSF with an average strength of 83.4 and a standard deviation of 8.6. The only discrepancy concerned the mutation localized in intron 3 of *GH1* for which the BP was predicted to be at position −26 by HSF instead of position −21. Note that in both cases, the BP was located within the c.468-37_468-16del which is responsible for the cases of autosomal dominant, isolated GH deficiency (IGHDII) in one single family and therefore additional data are needed to identify the functional BP. Among the other BP sequences, 12 were reported as targets of point mutations leading to their inactivation. In six cases, the mutation involved the critical adenosine residue, leading to a remarkable ΔBP of −29.6%. For mutations involving residues surrounding the BP, the average ΔBP was −13.9% with a SD of 3%. Taking into account the weight matrix (Figure 1) and experimental data, the threshold for BP prediction was thus set at 67.

Table 2. Exonic mutations in *DMD* (ENST00000357033), *MLH1* (ENST 00000231790), *MSH2* (ENST00000233146) and *RB1* (ENST00000267163) involved in splicing

Gene	Mutation	Position	References	WT CV	Mutant CV	CV variation (%)
DMD	c.5985T>G	Deep exonic	(91)	46.65	75.59	3′ CS (63 nt downstream)[a]
MLH1	c.677G>A	Last base	(58)	84.46	73.89	−12.52[b]
MLH1	c.882C>T	Exonic	(58)	84.46	73.89	−12.52[b]
MLH1	c.1037A>G	Penultimate base	(58)	93.04	88.19	−5.22 5′ CS (upstream[c])
MLH1	c.1038G>T	Last base	(58)	93.04	82.17	−11.68 5′ CS (upstream[c])
MLH1	c.1667G>T	Last base	(92)	85.85	74.99	−12.66 5′ CS (88 nt downstream)[a]
MLH1	c.1731G>A	Last base	(58)	93.27	82.69	−11.34
MLH1	c.1989G>T	Last base	(58)	93.22	82.35	−11.66
MSH2	c.1660A>T	Penultimate base	(58)	84.00	79.25	−5.65 5′ CS (82 nt upstream)[a]
MSH2	c.1759G>C	Last base	(58)	85.66	74.65	−12.86[b]
MSH2	c.1915C>T	Deep exonic	(59)	62.19	89.02	5′ CS (92 nt upstream)[a]
RB1	c.658C>G	Deep exonic	(57)	58.66	85.49	5′ CS (61 nt upstream)[a]
RB1	c.939G>T	Last base	(57)	83.75	72.88	−12.98[b]
RB1	c.1960G>C	Last base	(57)	94.02	83.01	−11.71[b]
RB1	c.1960G>A	Last base	(57)	94.02	83.44	−11.25[b]

CS: cryptic site (i.e. a new splice site is created by the mutation and is used instead of the regular site). Nucleotide numbering follows the reference cDNA sequence with +1 corresponding to the A of the ATG translation initiation codon.
[a]The cryptic splice site created by the mutation and used *in vivo* was correctly predicted by HSF.
[b]The mutation induces exon skipping.
[c]The cryptic splice site used *in vitro* was not clearly reported and therefore was not available for comparison.

Table 3. Branch point sequences

Gene	Intron	References	Ref BP	Ref Seq	HSF BP	HSF value
COL5A1	32	(93)	−27	ENST00000355306	−27	87.81
DYSF	31	(94)	−33	ENST00000258104	−33	93.13
FBN2	30	(95)	−24	ENST00000262464	−24	77.06
GH1	3	(96)	−21	ENST00000323322	−26	73.36
ITGB4	31	(97)	−17	ENST00000200181	−17	93.79
LCAT	4	(98)	−20	ENST00000264005	−20	95.07
LDLR	9	(99)	−25	ENST00000252444	−25	86.59
NPC1	6	(100)	−28	ENST00000269228	−28	77.41
PMM2	2	(101)	−25	ENST00000268261	−25	80.56
PMM2	7	(101)	−23	ENST00000268261	−23	72.27
RB1	23	(57)	−26	ENST00000267163	−26	75.89
TH	11	(102)	−22	ENST00000324155	−22	84.96
TSC2	38	(103)	−18	ENST00000219476	−18	67.71
XPC	3	(76)	−24	ENST00000285021	−24	82.78

For each gene the reference sequence from the Ensembl genome database (Ref Seq), the intron number (Intron) and the position of the BP identified by *in vitro* experiments (Ref BP) as well as the BP position predicted by HSF (HSF BP) and the corresponding BP value (HSF value) are shown.

Auxiliary splicing sequences: enhancers and silencers

In order to simplify the interpretation of predictions obtained with the different algorithms using weight matrices, we used a normalized range scale from 0 to 100. As a consequence, previous matrices from ESE-Finder (19,27) were modified. Nevertheless the user can define the thresholds using either the original ESE-Finder range or the new 0–100 range. In addition, when processing a single sequence and when CVs are available, HSF calculates the deviation as a percentage of the threshold. A reduced list can be obtained for each matrix by choosing the '*Only variant*' option in '*Advanced parameters*'. A color code is used for each quartile (from white to orange) to simplify the analysis. When comparing mutant sequences, HSF uses this color code to indicate the differences between the two sequences. When scalability is not possible, HSF only displays the presence of a motif.

To evaluate the sensitivity and usefulness of auxiliary splicing sequence predictions, we used a first set of genes for which 20 mutations have been reported to result in exon skipping following targeting of ESE or ESS (58–66). For each mutation, we selected the default option that allows HSF to predict modifications of ESE and/or ESS motifs using all available matrices (Table 4). For mutation c.362C>T in *ACADM* or c.4250T>A in *DMD* for which the target auxiliary sequences have been experimentally characterized (SF2/ASF and hnRNPA1 respectively), HSF correctly predicted the effect of the mutation. For other sequences, different scenarios were predicted: (i) disruption of one or more ESE without creation of an ESS, as observed for mutations c.882C>T (*MLH1*), c.362C>T (*ACADM*), c.8165C>G and

Table 4. Exonic mutations known to result in exon skipping through ESE inactivation or ESS activation

Gene	Mutation	Ref.	Motif	Ref Seq	HSF prediction
ACADM	c.362C>T	(65)	−ESE (SF2/ASF)	ENST00000370841	−9G8[i] (357_362) −SF2/ASF[e] (358_364) +EIE[h] (359_364) −SRp40[e] (359_365) − EIE[h] (360_365) +IIE[c] × 4 (359_367)
BRCA1	c.5080G>T	(64)	?	ENST00000357654	−EIE[h] (5075_5080) +SRp55[e] (5076_5081) −9G8[i] (5077_5082) −SF2/ASF[e] (5078_5085) −IIE[c] (5078_5083) +IIE[c] (5079_5084) −ESS[a](5076_5083) +hnRNPA1[d] (5080_5085)
BRCA2	c.8165C>G	(62)	−ESE	ENST00000380152	−SRp40[e] (8162_8168) −ESE[f] (8163_8168) +ESE[f] × 2 (8164_8170) −SRp55[e] (8163_8169) −SF2/ASF[e] (8165_8171) −EIE[h] × 4 (8160_8168)
BRCA2	c.5081G>T	(64)	?	ENST00000380152	+SC35[e] (5075_5082) +SRp40[e] (5080_5086) −ESE[f,h] × 2 (5080_5086) −9G8[i] (5081_5086) −ESS[a] (5078_5085)
DMD	c.4250T>A	(61)	+ESS (hnRNPA1)	ENST00000357033	+9G8[i] × 2 (4246_4251)(4248_4253) −EIE[h] (4248_4253) +ESE[f] (4250_4255) −IIE[c] × 3 (4246_4253) +hnRNPA1[d] (4249_4254)
MLH1	c.544A>G	(59)	?	ENST00000231790	+ESS[a] (537_545) 5′ss ΔCV = −6.30
MLH1	c.793C>T	(58)	?	ENST00000231790	+ESS[a] (795_802)
MLH1	c.794G>A	(58)	?	ENST00000231790	−SRp40[e] (793_799) −SC35[e] (794_801) +ESS[c] (794_799)
MLH1	c.882C>T	(58)	?	ENST00000231790	+SC35[e] (876_883) −SRp55[e] (877_882)
MLH1	c.988_990del	(58)	?	ENST00000231790	+SF2/ASF[e] (983_989) −SRp55[e] (985_990) +9G8[i] (985_990) −ESS[a] (985_992)
MSH2	c.815C>T	(58)	?	ENST00000233146	−SRp55[e] (813_818) +ESS[a] (813_820) +ESS[c] × 5 (801_819)
MSH2	c.274_276del	(58)	?	ENST00000233146	+SC35[e] (272_279) +SRp40[e] × 2 (274_285) −IIE[c] × 2 (274_280)
LAMA2	c.2230C>T	(60)	?	ENST00000354729	−SF2/ASF[e] (2226_2232) +ESS[c] (2228_2235) +IIE[c] × 2 (2229_2235) +ESS[a] (2230_2237)
NF1	c.557A>T	(66)	−ESE	ENST00000356175	−SRp55[e] (552_557) −ESE[f] (552_557) −EIE[h] × 4 (552_560) −9G8[i] (553_558) +ESS[a] × 2 (550_557) (555_562)
NF1	c.910C>T	(66)	−ESE	ENST00000356175	−9G8[i] (905_910) −EIE[h] (905_910) +ESE[f] (908_913) −ESE[f] (910_915) −ESS[a] (906_913)
NF1	c.943C>T	(66)	−ESE	ENST00000356175	−SC35[e] (941_948) −SF2/ASF[e] (943_949) −PESE[g] (942_949) −9G8[i] (938_943) +hnRNPA1[d] (943_948) +IIE[c] (942_947)

(continued)

Table 4. Continued

Gene	Mutation	Ref.	Motif	Ref Seq	HSF prediction
NF1	c.1007G>A	(66)	−ESE	ENST00000356175	+PESE[g] (1007_1014) −EIE[h] × 2 (1003_1011) +9G8[i] (1006_1011) +ESE[f] (1007_1014) −ESS[a] × 2 (1003_1011) −IIE[c] × 4 (1003_1011) +hnRNPA1[d] (1006_1011)
NF1	c.5719G>T	(66)	−ESE	ENST00000356175	−ESE[f] × 5 (5715_5724) −EIE[h] × 5 (5715_5724) −ESS[a] × 2 (5714_5725) +PESS[g] × 2 (5712_5720) +hnRNPA1[d] (5719_5724)
NF1	c.6792C>A	(66)	−ESE	ENST00000356175	+ESE[f] × 5 (6792_6797) −EIE[h] × 2 (6788_6793) (6790_6795) +Tra2−β[i] (6791_6795) −ESS[a] × 2 (6787_6794) (6792_6799)
NF1	c.6792C>G	(66)	−ESE	ENST00000356175	+ESE[f] (6792_6797) −EIE[h] × 2 (6788_6793) (6790_6795) s+Tra2−β[i] (6791_6795) −ESS[a] × 2 (6787_6794) (6792_6799) +hnRNPA1[d] (6790_6795)

+: a new site was created by the mutation; −: the motif was abolished by the mutation. Algorithms and matrices used to identify the motifs were:
[a]Silencer motifs from Sironi *et al.* (31).
[b]PESS octamers (28).
[c]IIEs (30).
[d]hnRNP motifs from HSF.
[e]ESE Finder matrices (19).
[f]RESCUE ESE hexamers (63).
[g]PESE octamers (28).
[h]EIEs (30).
[i]ESE motifs from HSF. When multiple adjacent sites were predicted, the number of sites is indicated: ×5 means that five adjacent sites were modified by the mutation. Nucleotide numbering reflects the reference cDNA sequence with +1 corresponding to the A of the ATG translation initiation codon.

c.5081G>T (*BRCA2*), c.557A>T and c.910C>T (*NF1*); (ii) creation of one or more ESS without disruption of an ESE, as shown for mutations c.544A>G and c.793C>T (*MLH1*), c.4250T>A (*DMD*) and c.6792C>G (*NF1*) and c) intermediate situation where both the disruption of one or more ESE and the creation of one or more ESS were predicted. This was observed for mutations c.5080G>T (*BRCA1*), c.794G>A and c.988_990del (*MLH1*), c.815C>T and c.274_276del (*MSH2*), c.2230C>T (*LAMA2*), c.943C>T, c.1007G>A and c.5719G>T (*NF1*). In order to evaluate the potential to differentiate 'true' ESE or ESS motifs from false positive signals, we selected a second set of 36 mutations (positive controls) and 220 SNPs (negative controls) (Supplementary Table 1). Predictions were classified in three categories: disruption of ESE motifs only (ESE), creation of ESS motifs only (ESS) or both (ESE + ESS). In addition, results were classified in two subsets: a first one (All), which included all predicted motifs, and a second one (Best), which was restricted to only one motif for each case by selecting the one recognized by the highest number of matrices.

Comparison of the three categories (ESE, ESS, and ESE + ESS) revealed a significant difference between positive and negative controls both in the 'All' ($\chi^2 = 10.05$, $P = 0.00656$) and the 'Best' subset ($\chi^2 = 11.75$, $P = 0.0028$). We then evaluated the potential of each matrix to differentiate true from false positive signals. No statistical differences were found using the Sironi, PESS, IIE, hnRNPA1 and RESCUE-ESE matrices. A statistically significant difference was found for the 'All' subset ($\chi^2 = 3.99$, $P = 0.045$), but not for the 'Best' subset ($\chi^2 = 2.47$, $P = 0.116$) with the EIE matrix. Significant results in both subsets were obtained with ESE-Finder ('All' subset: $\chi^2 = 5.17$, $P = 0.023$; 'Best' subset: $\chi^2 = 7.33$, $P = 0.0067$), the 9G8 and Tra2ß matrices from HSF ('All' subset: $\chi^2 = 9.92$, $P = 0.00164$; 'Best' subset: $\chi^2 = 9.86$, $P = 0.00169$) and PESE ('All' subset: $\chi^2 = 19.52$, $P = 9.95 \times 10^{-6}$; 'Best' subset: $\chi^2 = 13.52$, $P = 2.36 \times 10^{-4}$). The positive (PPV) and negative (NPV) predictive values as well as the sensitivity (Sv) and the specificity (Sp) of these last three matrices were then evaluated. PPV ranged from 0.22 (9G8 and Tra2ß) to 0.56 (PESE), PNV from 0.76 (PESE) to 0.95 (9G8 and Tra2ß), Sv from 0.27 (PESE) to 0.40 (9G8 and Tra2ß) and Sp from 0.88 (9G8 and Tra2ß) to 0.91 (PESE). The ESE-Finder matrix showed intermediate values in all cases.

DISCUSSION

During evolution from simple to higher eukaryotes, splicing signals evolved from well-defined motifs to degenerated sequences with the addition of new auxiliary splicing

sequences known as ESE and ESS. Although major SR proteins have been cloned and their target sites determined, much work remains to be done to understand how splice signals are recognized and splicing specificity achieved. As this complex world is progressively revealed, bioinformatics resources could play a major role in helping researchers and diagnostic laboratories to evaluate the consequence of mutations on splicing, especially because most genetic tests use DNA and not RNA samples. By giving an easy access to predictions of 5′ss, 3′ss, BP sequences as well as ESE and ESS, the HSF tool (http://www.umd.be/HSF/) fulfills this need and may assist clinicians, geneticists and researchers (70–75). By combining motifs identified with different experimental and computational approaches, it provides a common interface that can be used for sequence analysis. The inclusion of all exons and introns extracted from the Ensembl human genome database (20) allows an easy access to any sequence of human genes and thus direct comparison of virtually every mutation or SNP concerning splicing elements. Since SNPs are present at a very high frequency in the genome (1/300 bp) it could be useful to evaluate their impact in association with a mutation. We therefore included in HSF data from dbSNP using Ensembl Biomart. The user can select the 'Search for SNPs related to the analyzed sequence' option that automatically retrieves SNPs from the database. When SNPs are localized in exons, their effect on ESE and ESS motifs could help the user to better evaluate the consequence of a given mutation.

To evaluate the efficiency of the various algorithms included in HSF and its contribution to the prediction of the consequences of mutations associated with a splicing defect, we used a set of 69 intronic mutations that disrupt the 5′ss or the 3′ss and result in exon skipping and/or activation of a cryptic splice site (Table 1), and a group of 15 mutations that were previously reported to result in splicing defects by creating or activating cryptic splice sites (Table 2). HSF was able to correctly predict the disruption of the natural splice sites. Moreover, we could confirm that (i) mutations of the last nucleotide of an exon have a strong effect on the 5′ss ($\Delta CV = 12\% \pm 0.7$) resulting frequently in exon skipping or partial exonic deletion or intronic retention due to activation of a cryptic splice site; (ii) mutations of the penultimate exonic nucleotide have limited consequences on the 5′ss ($\Delta CV = 5.4\% \pm 0.3$), but they can activate a cryptic splice site, making predictions more difficult; (iii) exonic mutations distant from the 5′ and 3′ss can activate a cryptic splice site leading to partial exonic deletion. Overall these findings underline the efficiency of the HSF algorithm to predict the effect of mutations on 5′ and 3′ss. When using the HSF algorithm, the threshold for 5′ and 3′ss is 65 with a pathogenic ΔCV of -10% except for position $+4$ where it is -7%. However, in few cases when unusual splice sites are used, this algorithm could be less efficient.

BP sequences represent another essential splicing signal. When a mutation is localized in proximity of the 5′ of the 3′ss, its potential effect on a BP sequence should be examined especially when a nucleotide located at less than 85 bp from the 3′ss is targeted. In order to evaluate the HSF algorithm dedicated to the identification of BP sequences, we used 14 BP sequences inactivated by intronic mutations (Table 3). HSF correctly predicted 13 out of 14 BPs and these data allowed us to define the threshold for BP detection at 67 and the pathogenic ΔBP at -10%. Moreover, for intron 3 of *XPC*, HSF predicted a BP at position -24. However, according to Khan *et al.* (76), two BP sequences are present in this intron, one at positions -24 and another at –4. HSF could not predict the BP at position -4 simply because the HSF algorithm excludes positions -12 to -1 for BP identification because of steric obstruction caused by the spliceosome.

It has been demonstrated that two different splicing recognition mechanisms, correlated with intron length, can be used in a cell: exon definition for long and exon definition for short introns (77). Although the influence of intron length seems to be less important in humans than in other species, it should, nevertheless, be kept in mind since U12 and U2-type introns have different BP consensus sequences. In the present version of HSF (v2.4), we only focused on U2-type introns, which are by far the most abundant type in mammalian cells.

Concerning *cis*-acting elements, many works have been performed to define ESE and ESS matrices based on bioinformatics or experimental approaches (78). However, due to technical and/or conceptual bias, the various sequence sets only share partial homology. To solve this problem, HSF included all available matrices in one place. In addition, we developed new matrices to predict ESE motifs for the 9G8 and Tra2-β SR proteins and ESS motifs for the hnRNPA1 ribonucleoprotein. ESE and ESS motifs frequently overlap and therefore the identification of the specific motif/protein pair involved in a given splicing defect is difficult. This is even more complicated when considering the impact of SR and ribonucleoprotein concentration in different tissues or during development. We used a set of 20 exonic mutations known to influence splicing through ESE inactivation or ESS activation (Table 4) to evaluate the efficiency of HSF to correctly predict motifs disrupted by these mutations. We showed that when the motif/protein pairs had been previously experimentally characterized (hnRNPA1 or SF2/ASF), HSF was able to correctly predict the effects of the mutation on ESE and ESS. For most mutations, however, only the general mechanism was identified (i.e. the mutant sequence inhibits splicing in various *in vitro* reporter systems) and therefore the motif/protein couple is unknown. In these cases, HSF predicted the disruption of ESE motifs and/or the creation of ESS motifs (Table 4). In addition, to evaluate HSF efficiency to discriminate true from false positive signals, we used a second group of positive and negative controls (Supplementary Table 1). We showed that both sets could be discriminated on the basis of their overall pattern (ESE, ESS, ESE + ESS; $\chi^2 = 11.75$, $P = 0.0028$). Three matrices also gave statistically significant results: ESE-Finder ($\chi^2 = 7.33$, $P = 0.0067$), 9G8 and Tra2ß from HSF ($\chi^2 = 9.86$, $P = 0.0017$) and PESE ($\chi^2 = 13.52$, $P = 2.36 \times 10^{-4}$). Since these three matrices predict ESE motifs, these results could be associated with a bias towards the positive controls. Indeed, only few experimental validations of auxiliary sequences are

available and they are frequently initiated by predictions of ESE motifs using ESE-Finder. PESE and the 9G8/Tra2ß HSF matrices gave stronger results than ESE Finder itself and therefore can be considered efficient matrices for the identification of ESE motifs. However, predictions with other matrices, especially the hnRNPA1 matrix, should also be considered as they could provide valuable information as shown for the c.4250T>A of *DMD*. We are still in the early days of ESE and ESS motif predictions and further data are needed to select the best matrices and to define the rules for data interpretation as most mutation sets used to validate prediction tools contain mainly mutations affecting splice sites (79). Major work is also needed to ultimately address the tissue or developmental specificity.

In conclusion, the HSF tool is dedicated to the prediction of splicing signals present in any human gene using all available matrices to identify ESE and ESS and new matrices to evaluate 5′ and 3′ss and BPs. This tool is regularly updated to include new data from bioinformatics and experimental studies in order to improve predictions. Many users already have tested HSF and have stressed its value both for basic science (identification of splicing signals) and applied research or diagnostics (prediction of the pathogenic consequences of a given mutation) (70–75). In addition, new genotype-based therapies, such as the exon-skipping approach in Duchenne Muscular Dystrophy, are currently evaluated in clinical trials (international multi-center phase I/II clinical studies with PRO051 in patients with Duchenne Muscular Dystrophy – Prosensa company; http://prosensa.eu/). HSF might represent an useful tool to identify key splicing sequences in different exons (75,80) and therefore to design antisense oligonucleotides to induce exon skipping. This approach is being actively evaluated throughout the world and especially by the TREAT-NMD European network (http://www.treat-nmd.eu/home.php).

Besides these gene-specific approaches, global projects, which either aim at developing a holistic view on Genotype-To-Phenotype data (GEN2PHEN European projects; http://www.gen2phen.org/) or at improving health outcomes by facilitating the analysis of human genetic variation and its impact on human health, such as the Human Variome Project (81), might benefit from using HSF. Indeed, HSF could help to predict the theoretical impact on splicing of any sequence variation affecting a human gene.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

European Community Seventh Framework Program (FP7/2007-2013) under grant agreement number 200754—the GEN2PHEN project; The European Community Sixth Framework Program (FP6) under grant agreement number 036825; TREAT-NMD Network of Excellence. Funding for open access charge: Institut National de la Santé Et de la Recherche Médicale (INSERM).

Conflict of interest statement. None declared.

REFERENCES

1. Berget,S.M., Moore,C. and Sharp,P.A. (1977) Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Natl Acad. Sci. USA*, **74**, 3171–3175.
2. Nilsen,T.W. (2003) The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, **25**, 1147–1149.
3. Zhou,Z., Licklider,L.J., Gygi,S.P. and Reed,R. (2002) Comprehensive proteomic analysis of the human spliceosome. *Nature*, **419**, 182–185.
4. Breitbart,R.E., Nguyen,H.T., Medford,R.M., Destree,A.T., Mahdavi,V. and Nadal-Ginard,B. (1985) Intricate combinatorial patterns of exon splicing generate multiple regulated troponin T isoforms from a single gene. *Cell*, **41**, 67–82.
5. Maniatis,T. and Tasic,B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.
6. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.*, **3**, 285–298.
7. Robberson,B.L., Cote,G.J. and Berget,S.M. (1990) Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell Biol.*, **10**, 84–94.
8. Jacob,M. and Gallinaro,H. (1989) The 5′ splice site: phylogenetic evolution and variable geometry of association with U1RNA. *Nucleic Acids Res.*, **17**, 2159–2180.
9. Blencowe,B.J. (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci.*, **25**, 106–110.
10. Zhu,J., Mayeda,A. and Krainer,A.R. (2001) Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol. Cell*, **8**, 1351–1361.
11. Zhang,X.H., Leslie,C.S. and Chasin,L.A. (2005) Computational searches for splicing signals. *Methods*, **37**, 292–305.
12. Bhasi,A., Pandey,R.V., Utharasamy,S.P. and Senapathy,P. (2007) EuSplice: A unified resource for the analysis of splice signals and alternative splicing in eukaryotic genes. *Bioinformatics*, **23**, 1815–1823.
13. Churbanov,A., Rogozin,I.B., Deogun,J.S. and Ali,H. (2006) Method of predicting splice sites based on signal interactions. *Biol. Direct.*, **1**, 10.
14. Dunckley,M.G., Manoharan,M., Villiet,P., Eperon,I.C. and Dickson,G. (1998) Modification of splicing in the dystrophin gene in cultured Mdx muscle cells by antisense oligoribonucleotides. *Hum. Mol. Genet.*, **7**, 1083–1090.
15. Wilton,S.D. and Fletcher,S. (2005) RNA splicing manipulation: strategies to modify gene expression for a variety of therapeutic outcomes. *Curr. Gene. Ther.*, **5**, 467–483.
16. Beroud,C., Hamroun,D., Collod-Beroud,G., Boileau,C., Soussi,T. and Claustres,M. (2005) UMD (Universal Mutation Database): 2005 update. *Hum. Mutat.*, **26**, 184–191.
17. Beroud,C., Collod-Beroud,G., Boileau,C., Soussi,T. and Junien,C. (2000) UMD (Universal mutation database): a generic software to build and analyze locus-specific databases. *Hum. Mutat.*, **15**, 86–94.
18. 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.
19. 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.
20. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
21. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.

22. Shapiro,M.B. and Senapathy,P. (1987) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.*, **15**, 7155–7174.
23. 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.
24. Green,M.R. (1991) Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. *Annu. Rev. Cell Biol.*, **7**, 559–599.
25. Gooding,C., Clark,F., Wollerton,M.C., Grellscheid,S.N., Groom,H. and Smith,C.W. (2006) A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. *Genome Biol.*, **7**, R1.
26. Kol,G., Lev-Maor,G. and Ast,G. (2005) Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. *Hum. Mol. Genet.*, **14**, 1559–1568.
27. 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.
28. Zhang,X.H. and Chasin,L.A. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.*, **18**, 1241–1250.
29. Goren,A., Ram,O., Amit,M., Keren,H., Lev-Maor,G., Vig,I., Pupko,T. and Ast,G. (2006) Comparative analysis identifies exonic splicing regulatory sequences–The complex definition of enhancers and silencers. *Mol. Cell*, **22**, 769–781.
30. 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.
31. Sironi,M., Menozzi,G., Riva,L., Cagliani,R., Comi,G.P., Bresolin,N., Giorda,R. and Pozzoli,U. (2004) Silencer elements as possible inhibitors of pseudoexon splicing. *Nucleic Acids Res.*, **32**, 1783–1791.
32. 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.
33. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
34. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
35. Yuan,B., Thomas,J.P., von Kodolitsch,Y. and Pyeritz,R.E. (1999) Comparison of heteroduplex analysis, direct sequencing, and enzyme mismatch cleavage for detecting mutations in a large gene, FBN1. *Hum. Mutat.*, **14**, 440–446.
36. Youil,R., Toner,T.J., Bull,E., Bailey,A.L., Earl,C.D., Dietz,H.C. and Montgomery,R.A. (2000) Enzymatic mutation detection (EMD) of novel mutations (R565X and R1523X) in the FBN1 gene of patients with Marfan syndrome using T4 endonuclease VII. *Hum. Mutat.*, **16**, 92–93.
37. Schrijver,I., Liu,W., Odom,R., Brenn,T., Oefner,P., Furthmayr,H. and Francke,U. (2002) Premature termination mutations in FBN1: distinct effects on differential allelic expression and on protein and clinical phenotypes. *Am. J. Hum. Genet.*, **71**, 223–237.
38. Rommel,K., Karck,M., Haverich,A., Schmidtke,J. and Arslan-Kirchner,M. (2002) Mutation screening of the fibrillin-1 (FBN1) gene in 76 unrelated patients with Marfan syndrome or Marfanoid features leads to the identification of 11 novel and three previously reported mutations. *Hum. Mutat.*, **20**, 406–407.
39. Park,E.S., Putnam,E.A., Chitayat,D., Child,A. and Milewicz,D.M. (1998) Clustering of FBN2 mutations in patients with congenital contractural arachnodactyly indicates an important role of the domains encoded by exons 24 through 34 during human development. *Am. J. Med. Genet.*, **78**, 350–355.
40. Palz,M., Tiecke,F., Booms,P., Goldner,B., Rosenberg,T., Fuchs,J., Skovby,F., Schumacher,H., Kaufmann,U.C., von Kodolitsch,Y. *et al.* (2000) Clustering of mutations associated with mild Marfan-like phenotypes in the 3′ region of FBN1 suggests a potential genotype-phenotype correlation. *Am. J. Med. Genet.*, **91**, 212–221.
41. Nijbroek,G., Sood,S., McIntosh,I., Francomano,C.A., Bull,E., Pereira,L., Ramirez,F., Pyeritz,R.E. and Dietz,H.C. (1995) Fifteen novel FBN1 mutations causing Marfan syndrome detected by heteroduplex analysis of genomic amplicons. *Am. J. Hum. Genet.*, **57**, 8–21.
42. McGrory,J. and Cole,W.G. (1999) Alternative splicing of exon 37 of FBN1 deletes part of an 'eight-cysteine' domain resulting in the Marfan syndrome. *Clin. Genet.*, **55**, 118–121.
43. Loeys,B., Nuytinck,L., Delvaux,I., De Bie,S. and De Paepe,A. (2001) Genotype and phenotype analysis of 171 patients referred for molecular study of the fibrillin-1 gene FBN1 because of suspected Marfan syndrome. *Arch. Intern. Med.*, **161**, 2447–2454.
44. Liu,W.O., Oefner,P.J., Qian,C., Odom,R.S. and Francke,U. (1997) Denaturing HPLC-identified novel FBN1 mutations, polymorphisms, and sequence variants in Marfan syndrome and related connective tissue disorders. *Genet. Test*, **1**, 237–242.
45. Hutchinson,S., Wordsworth,B.P. and Handford,P.A. (2001) Marfan syndrome caused by a mutation in FBN1 that gives rise to cryptic splicing and a 33 nucleotide insertion in the coding sequence. *Hum. Genet.*, **109**, 416–420.
46. Halliday,D., Hutchinson,S., Kettle,S., Firth,H., Wordsworth,P. and Handford,P.A. (1999) Molecular analysis of eight mutations in FBN1. *Hum. Genet.*, **105**, 587–597.
47. Gupta,P.A., Wallis,D.D., Chin,T.O., Northrup,H., Tran-Fadulu,V.T., Towbin,J.A. and Milewicz,D.M. (2004) FBN2 mutation associated with manifestations of Marfan syndrome and congenital contractural arachnodactyly. *J. Med. Genet.*, **41**, e56.
48. Gupta,P.A., Putnam,E.A., Carmical,S.G., Kaitila,I., Steinmann,B., Child,A., Danesino,C., Metcalfe,K., Berry,S.A., Chen,E. *et al.* (2002) Ten novel FBN2 mutations in congenital contractural arachnodactyly: delineation of the molecular pathogenesis and clinical phenotype. *Hum. Mutat.*, **19**, 39–48.
49. Guo,D., Tan,F.K., Cantu,A., Plon,S.E. and Milewicz,D.M. (2001) FBN1 exon 2 splicing error in a patient with Marfan syndrome. *Am. J. Med. Genet.*, **101**, 130–134.
50. Dietz,H.C., McIntosh,I., Sakai,L.Y., Corson,G.M., Chalberg,S.C., Pyeritz,R.E. and Francomano,C.A. (1993) Four novel FBN1 mutations: significance for mutant transcript level and EGF-like domain calcium binding in the pathogenesis of Marfan syndrome. *Genomics*, **17**, 468–475.
51. Comeglio,P., Johnson,P., Arno,G., Brice,G., Evans,A., Aragon-Martin,J., da Silva,F.P., Kiotsekoglou,A. and Child,A. (2007) The importance of mutation detection in Marfan syndrome and Marfan-related disorders: report of 193 FBN1 mutations. *Hum. Mutat.*, **28**, 928.
52. Collod-Beroud,G., Le Bourdelles,S., Ades,L., Ala-Kokko,L., Booms,P., Boxer,M., Child,A., Comeglio,P., De Paepe,A., Hyland,J.C. *et al.* (2003) Update of the UMD-FBN1 mutation database and creation of an FBN1 polymorphism database. *Hum. Mutat.*, **22**, 199–208.
53. Chikumi,H., Yamamoto,T., Ohta,Y., Nanba,E., Nagata,K., Ninomiya,H., Narasaki,K., Katoh,T., Hisatome,I., Ono,K. *et al.* (2000) Fibrillin gene (FBN1) mutations in Japanese patients with Marfan syndrome. *J. Hum. Genet.*, **45**, 115–118.
54. Biggin,A., Holman,K., Brett,M., Bennetts,B. and Ades,L. (2004) Detection of thirty novel FBN1 mutations in patients with Marfan syndrome or a related fibrillinopathy. *Hum. Mutat.*, **23**, 99.
55. Attanasio,M., Lapini,I., Evangelisti,L., Lucarini,L., Giusti,B., Porciani,M., Fattori,R., Anichini,C., Abbate,R., Gensini,G. *et al.* (2008) FBN1 mutation screening of patients with Marfan syndrome and related disorders: detection of 46 novel FBN1 mutations. *Clin. Genet.*, **74**, 39–46.
56. Loeys,B.L., Chen,J., Neptune,E.R., Judge,D.P., Podowski,M., Holm,T., Meyers,J., Leitch,C.C., Katsanis,N., Sharifi,N. *et al.* (2005) A syndrome of altered cardiovascular, craniofacial, neurocognitive and skeletal development caused by mutations in TGFBR1 or TGFBR2. *Nat. Genet.*, **37**, 275–281.
57. Houdayer,C., Dehainault,C., Mattler,C., Michaux,D., Caux-Moncoutier,V., Pages-Berhouet,S., d'Enghien,C.D., Lauge,A., Castera,L., Gauthier-Villars,M. *et al.* (2008) Evaluation of in silico splice tools for decision-making in molecular diagnosis. *Hum. Mutat.*, **29**, 975–982.
58. Tournier,I., Vezain,M., Martins,A., Charbonnier,F., Baert-Desurmont,S., Olschwang,S., Wang,Q., Buisine,M.P., Soret,J., Tazi,J. *et al.* (2008) A large fraction of unclassified variants

of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. *Hum. Mutat.*, **29**, 1412–1424.
59. Auclair,J., Busine,M.P., Navarro,C., Ruano,E., Montmain,G., Desseigne,F., Saurin,J.C., Lasset,C., Bonadona,V., Giraud,S. *et al.* (2006) Systematic mRNA analysis for the effect of MLH1 and MSH2 missense and silent mutations on aberrant splicing. *Hum. Mutat.*, **27**, 145–154.
60. Di Blasi,C., He,Y., Morandi,L., Cornelio,F., Guicheney,P. and Mora,M. (2001) Mild muscular dystrophy due to a nonsense mutation in the LAMA2 gene resulting in exon skipping. *Brain*, **124**, 698–704.
61. Disset,A., Bourgeois,C.F., Benmalek,N., Claustres,M., Stevenin,J. and Tuffery-Giraud,S. (2006) An exon skipping-associated nonsense mutation in the dystrophin gene uncovers a complex interplay between multiple antagonistic splicing elements. *Hum. Mol. Genet.*, **15**, 999–1013.
62. Fackenthal,J.D., Cartegni,L., Krainer,A.R. and Olopade,O.I. (2002) BRCA2 T2722R is a deleterious allele that causes exon skipping. *Am. J. Hum. Genet.*, **71**, 625–631.
63. 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.
64. Mazoyer,S., Puget,N., Perrin-Vidoz,L., Lynch,H.T., Serova-Sinilnikova,O.M. and Lenoir,G.M. (1998) A BRCA1 nonsense mutation causes exon skipping. *Am. J. Hum. Genet.*, **62**, 713–715.
65. Nielsen,K.B., Sorensen,S., Cartegni,L., Corydon,T.J., Doktor,T.K., Schroeder,L.D., Reinert,L.S., Elpeleg,O., Krainer,A.R., Gregersen,N. *et al.* (2007) Seemingly neutral polymorphic variants may confer immunity to splicing-inactivating mutations: a synonymous SNP in exon 5 of MCAD protects from deleterious mutations in a flanking exonic splicing enhancer. *Am. J. Hum. Genet.*, **80**, 416–432.
66. Zatkova,A., Messiaen,L., Vandenbroucke,I., Wieser,R., Fonatsch,C., Krainer,A.R. and Wimmer,K. (2004) Disruption of exonic splicing enhancer elements is the principal cause of exon skipping associated with seven nonsense or missense alleles of NF1. *Hum. Mutat.*, **24**, 491–501.
67. den Dunnen,J.T. and Antonarakis,S.E. (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum. Mutat.*, **15**, 7–12.
68. Frederic,M.Y., Monino,C., Marschall,C., Hamroun,D., Faivre,L., Jondeau,G., Klein,H.G., Neumann,L., Gautier,E., Binquet,C. *et al.* (2008) The FBN2 gene: new mutations, locus-specific database (Universal Mutation Database FBN2), and genotype-phenotype correlations. *Hum. Mutat.*, **30**, 181–190.
69. Frederic,M.Y., Hamroun,D., Faivre,L., Boileau,C., Jondeau,G., Claustres,M., Beroud,C. and Collod-Beroud,G. (2008) A new locus-specific database (LSDB) for mutations in the TGFBR2 gene: UMD-TGFBR2. *Hum. Mutat.*, **29**, 33–38.
70. Frank,V., Ortiz Bruchle,N., Mager,S., Frints,S.G., Bohring,A., du Bois,G., Debatin,I., Seidel,H., Senderek,J., Besbas,N. *et al.* (2007) Aberrant splicing is a common mutational mechanism in MKS1, a key player in Meckel-Gruber syndrome. *Hum. Mutat.*, **28**, 638–639.
71. Anczukow,O., Buisson,M., Salles,M.J., Triboulet,S., Longy,M., Lidereau,R., Sinilnikova,O.M. and Mazoyer,S. (2008) Unclassified variants identified in BRCA1 exon 11: Consequences on splicing. *Genes Chromosomes Cancer*, **47**, 418–426.
72. Ng,W., Loh,A.X., Teixeira,A.S., Pereira,S.P. and Swallow,D.M. (2008) Genetic regulation of MUC1 alternative splicing in human tissues. *Br. J. Cancer*, **99**, 978–985.
73. Baala,L., Romano,S., Khaddour,R., Saunier,S., Smith,U.M., Audollent,S., Ozilou,C., Faivre,L., Laurent,N., Foliguet,B. *et al.* (2007) The Meckel-Gruber syndrome gene, MKS3, is mutated in Joubert syndrome. *Am. J. Hum. Genet.*, **80**, 186–194.
74. Habara,Y., Doshita,M., Hirozawa,S., Yokono,Y., Yagi,M., Takeshima,Y. and Matsuo,M. (2008) A strong exonic splicing enhancer in dystrophin exon 19 achieve proper splicing without an upstream polypyrimidine tract. *J. Biochem.*, **143**, 303–310.
75. Aartsma-Rus,A., van Vliet,L., Hirschi,M., Janson,A.A., Heemskerk,H., de Winter,C.L., de Kimpe,S., van Deutekom,J.C., t Hoen,P.A. and van Ommen,G.J. (2008) Guidelines for Antisense Oligonucleotide Design and Insight Into Splice-modulating Mechanisms. *Mol. Ther.*, **17**, 548–553.
76. Khan,S.G., Metin,A., Gozukara,E., Inui,H., Shahlavi,T., Muniz-Medina,V., Baker,C.C., Ueda,T., Aiken,J.R., Schneider,T.D. *et al.* (2004) Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced XPC mRNA levels that correlate with cancer risk. *Hum. Mol. Genet.*, **13**, 343–352.
77. Sharp,P.A. and Burge,C.B. (1997) Classification of introns: U2-type or U12-type. *Cell*, **91**, 875–879.
78. Chasin,L.A. (2007) Searching for splicing motifs. *Adv. Exp. Med. Biol.*, **623**, 85–106.
79. Nalla,V.K. and Rogan,P.K. (2005) Automated splicing mutation analysis by information theory. *Hum. Mutat.*, **25**, 334–342.
80. Beroud,C., Tuffery-Giraud,S., Matsuo,M., Hamroun,D., Humbertclaude,V., Monnier,N., Moizard,M.P., Voelckel,M.A., Calemard,L.M., Boisseau,P. *et al.* (2007) Multiexon skipping leading to an artificial DMD protein lacking amino acids from exons 45 through 55 could rescue up to 63% of patients with Duchenne muscular dystrophy. *Hum. Mutat.*, **28**, 196–202.
81. (2007) What is the human variome project? *Nat. Genet.*, **39**, 423.
82. Kainulainen,K., Karttunen,L., Puhakka,L., Sakai,L. and Peltonen,L. (1994) Mutations in the fibrillin gene responsible for dominant ectopia lentis and neonatal Marfan syndrome. *Nat. Genet.*, **6**, 64–69.
83. Liu,W., Qian,C., Comeau,K., Brenn,T., Furthmayr,H. and Francke,U. (1996) Mutant fibrillin-1 monomers lacking EGF-like domains disrupt microfibril assembly and cause severe marfan syndrome. *Hum. Mol. Genet.*, **5**, 1581–1587.
84. Booms,P., Cisler,J., Mathews,K.R., Godfrey,M., Tiecke,F., Kaufmann,U.C., Vetter,U., Hagemeier,C. and Robinson,P.N. (1999) Novel exon skipping mutation in the fibrillin-1 gene: two 'hot spots' for the neonatal Marfan syndrome. *Clin. Genet.*, **55**, 110–117.
85. Wang,M., Price,C., Han,J., Cisler,J., Imaizumi,K., Van Thienen,M.N., DePaepe,A. and Godfrey,M. (1995) Recurrent mis-splicing of fibrillin exon 32 in two patients with neonatal Marfan syndrome. *Hum. Mol. Genet.*, **4**, 607–613.
86. Godfrey,M., Vandemark,N., Wang,M., Velinov,M., Wargowski,D., Tsipouras,P., Han,J., Becker,J., Robertson,W., Droste,S. *et al.* (1993) Prenatal diagnosis and a donor splice site mutation in fibrillin in a family with Marfan syndrome. *Am. J. Hum. Genet.*, **53**, 472–480.
87. Wang,M., Clericuzio,C.L. and Godfrey,M. (1996) Familial occurrence of typical and severe lethal congenital contractural arachnodactyly caused by missplicing of exon 34 of fibrillin-2. *Am. J. Hum. Genet.*, **59**, 1027–1034.
88. Karttunen,L., Ukkonen,T., Kainulainen,K., Syvanen,A.C. and Peltonen,L. (1998) Two novel fibrillin-1 mutations resulting in premature termination codons but in different mutant transcript levels and clinical phenotypes. *Hum. Mutat.*, **Suppl 1**, S34–S37
89. Kosaki,K., Takahashi,D., Udaka,T., Kosaki,R., Matsumoto,M., Ibe,S., Isobe,T., Tanaka,Y. and Takahashi,T. (2006) Molecular pathology of Shprintzen-Goldberg syndrome. *Am. J. Med. Genet. A*, **140**, 104–108; author reply 109–110.
90. Loeys,B.L., Schwarze,U., Holm,T., Callewaert,B.L., Thomas,G.H., Pannu,H., De Backer,J.F., Oswald,G.L., Symoens,S., Manouvrier,S. *et al.* (2006) Aneurysm syndromes caused by mutations in the TGF-beta receptor. *N. Engl. J. Med.*, **355**, 788–798.
91. Tran,V.K., Takeshima,Y., Zhang,Z., Habara,Y., Haginoya,K., Nishiyama,A., Yagi,M. and Matsuo,M. (2007) A nonsense mutation-created intraexonic splice site is active in the lymphocytes, but not in the skeletal muscle of a DMD patient. *Hum. Genet.*, **120**, 737–742.
92. Sharp,A., Pichert,G., Lucassen,A. and Eccles,D. (2004) RNA analysis reveals splicing mutations and loss of expression defects in MLH1 and BRCA1. *Hum. Mutat.*, **24**, 272.
93. Burrows,N.P., Nicholls,A.C., Richards,A.J., Luccarini,C., Harrison,J.B., Yates,J.R. and Pope,F.M. (1998) A point mutation in an intronic branch site results in aberrant splicing of COL5A1 and in Ehlers-Danlos syndrome type II in two British families. *Am. J. Hum. Genet.*, **63**, 390–398.
94. Sinnreich,M., Therrien,C. and Karpati,G. (2006) Lariat branch point mutation in the dysferlin gene with mild limb-girdle muscular dystrophy. *Neurology*, **66**, 1114–1116.

95. Maslen,C., Babcock,D., Raghunath,M. and Steinmann,B. (1997) A rare branch-point mutation is associated with missplicing of fibrillin-2 in a large family with congenital contractural arachnodactyly. *Am. J. Hum. Genet.*, **60**, 1389–1398.
96. Vivenza,D., Guazzarotti,L., Godi,M., Frasca,D., di Natale,B., Momigliano-Richiardi,P., Bona,G. and Giordano,M. (2006) A novel deletion in the GH1 gene including the IVS3 branch site responsible for autosomal dominant isolated growth hormone deficiency. *J. Clin. Endocrinol. Metab.*, **91**, 980–986.
97. Chavanas,S., Gache,Y., Vailly,J., Kanitakis,J., Pulkkinen,L., Uitto,J., Ortonne,J. and Meneguzzi,G. (1999) Splicing modulation of integrin beta4 pre-mRNA carrying a branch point mutation underlies epidermolysis bullosa with pyloric atresia undergoing spontaneous amelioration with ageing. *Hum. Mol. Genet.*, **8**, 2097–2105.
98. Kuivenhoven,J.A., Weibusch,H., Pritchard,P.H., Funke,H., Benne,R., Assmann,G. and Kastelein,J.J. (1996) An intronic mutation in a lariat branchpoint sequence is a direct cause of an inherited human disorder (fish-eye disease). *J. Clin. Invest.*, **98**, 358–364.
99. Webb,J.C., Patel,D.D., Shoulders,C.C., Knight,B.L. and Soutar,A.K. (1996) Genetic variation at a splicing branch point in intron 9 of the low density lipoprotein (LDL)-receptor gene: a rare mutation that disrupts mRNA splicing in a patient with familial hypercholesterolaemia and a common polymorphism. *Hum. Mol. Genet.*, **5**, 1325–1331.
100. Di Leo,E., Panico,F., Tarugi,P., Battisti,C., Federico,A. and Calandra,S. (2004) A point mutation in the lariat branch point of intron 6 of NPC1 as the cause of abnormal pre-mRNA splicing in Niemann-Pick type C disease. *Hum. Mutat.*, **24**, 440.
101. Vuillaumier-Barrot,S., Le Bizec,C., De Lonlay,P., Madinier-Chappat,N., Barnier,A., Dupre,T., Durand,G. and Seta,N. (2006) PMM2 intronic branch-site mutations in CDG-Ia. *Mol. Genet. Metab.*, **87**, 337–340.
102. Janssen,R.J., Wevers,R.A., Haussler,M., Luyten,J.A., Steenbergen-Spanjers,G.C., Hoffmann,G.F., Nagatsu,T. and Van den Heuvel,L.P. (2000) A branch site mutation leading to aberrant splicing of the human tyrosine hydroxylase gene in a child with a severe extrapyramidal movement disorder. *Ann Hum. Genet.*, **64**, 375–382.
103. Mayer,K., Ballhausen,W., Leistner,W. and Rott,H. (2000) Three novel types of splicing aberrations in the tuberous sclerosis TSC2 gene caused by mutations apart from splice consensus sequences. *Biochim. Biophys. Acta*, **1502**, 495–507.