OPINION

Genomic variants in exons and introns: identifying the splicing spoilers

Franco Pagani and Francisco E. Baralle

When genome variants are identified in genomic DNA, especially during routine analysis of disease-associated genes, their functional implications might not be immediately evident. Distinguishing between a genomic variant that changes the phenotype and one that does not is a difficult task. An increasing amount of evidence indicates that genomic variants in both coding and non-coding sequences can have unexpected deleterious effects on the splicing of the gene transcript. So how can benign polymorphisms be distinguished from disease-associated splicing mutations?

The use of high-throughput screening methods has led to an exponential increase in the number of sequence variants being identified in the human genome, particularly in disease-associated genes. There is a great diversity of terms used for genome variants according to their effect or use: mutations, SNPs and SIMPLE SEQUENCE REPEATS are just a few. Here, we use the general term genomic variants (GVs) to denote single-nucleotide substitutions, or small insertions and deletions.

Researchers who attempt to track down the molecular basis of a disease almost always focus on GVs that change the sequence or expression of a candidate gene — specifically, non-synonymous substitutions in exons (that is, substitutions that change the protein sequence), substitutions in the gene promoter or substitutions at well-characterized splice sites. GVs that do not change the protein sequence (that is, silent GVs), such as synonymous GVs in exons (sGVs) and deep intronic variants, are often ignored. However, recent evidence from many laboratories indicates that silent GVs can affect splicing and as a result are often the cause of human diseases[1–14]. Splicing-affecting genomic variants (SpaGVs) can occur in exonic or intronic splicing regulatory elements that are extremely difficult to define from nucleotide sequences alone. Nonetheless, such changes can lead to catastrophic splicing abnormalities. In particular, SpaGVs can induce EXON SKIPPING, activation of CRYPTIC SPLICE SITES or can alter the fine balance of the ALTERNATIVELY SPLICED ISOFORMS that are produced and can therefore cause disease phenotypes.

We believe that many researchers still focus only on GVs that change the protein sequence and do not consider follow-up studies on SpaGVs that are shown to vary between cases and controls and that might be of equal or greater importance because of their effect on splicing. Here, we argue that all 'orphan' GVs at any candidate disease locus should be functionally assessed for their potential effect on splicing. In particular, we advocate the standard use of functional follow-up studies that are much more straightforward to conduct than is generally realized. First, we briefly summarize the pre-mRNA splicing process and how GVs can affect it, before discussing the principal categories of exonic and intronic SpaGVs. We then highlight three often-overlooked ways in which silent SpaGVs can be involved in disease phenotypes. We go on to discuss how medical geneticists can identify disease-causing SpaGVs, following with a brief conclusion in which we advocate a change in research emphasis towards a focus on elucidating the basic molecular mechanisms that are involved in splicing.

Pre-mRNA splicing, GVs and disease

To correctly identify and join together RNA sequences that code for proteins, the exons must be differentiated from the introns — that is, the large sections of non-coding RNA that separate them. There are several conserved motifs in the nucleotide sequences near the intron–exon boundaries that act as essential splicing signals. These signals — the 3′- and 5′-splice sites, a polypyrimidine tract and the branch site — are involved in the excision of introns from pre-mRNA and in the joining of the exons. In most cases, the 5′-splice site consists of a GU dinucleotide, whereas the 3′-splice site is an AG dinucleotide (FIG. 1).

Unsurprisingly, given the high accuracy and fidelity of pre-mRNA splice-site selection *in vivo*, these few essential splicing signals are not sufficient to select the correct exonic boundaries. Additional elements known as 'enhancers' and 'silencers' are needed to allow normal splicing of exonic sequences (FIG. 2). These regulatory elements can be located either in exons or introns and can be near (that is, within 50–100 bp) or far (that is, hundreds to thousands of bp) from the splice sites. These elements were originally discovered during research into alternative splicing, the process through which different splice sites are selected to generate proteins with different functions from the same gene. The need to specifically regulate alternative splicing according to cell type and developmental stage adds an additional level of complexity to the control of the splicing machinery and so made it easier to identify these extra modulatory elements. Since this original discovery, enhancers and silencers have also been identified in constitutively spliced exons, which indicates that they are a general feature of the splicing process and a target for pathogenetic mutations. (For a detailed description of the splicing process and alternative splicing, see

REF. 15.) Importantly, multiple cryptic signals are also present in the pre-mRNA that are similar to the true splicing signals. The abundance of PSEUDO SPLICE SITES and of multiple regulatory elements in thousands of bases of intronic sequences makes the splicing machinery's task of correctly identifying an exon (on average, 145 nucleotides)[16] a complex one.

GVs can cause disease by affecting both constitutive and alternative splicing. For example, a GV can cause a constitutively included exon to be skipped, leading to an aberrant mRNA and a subsequent loss of function of its translation product. The analysis of tissue samples from normal subjects frequently shows a degree of exon skipping even for constitutively spliced exons[17–20]. This basal skipping might also reflect individual variability resulting from the different genetic backgrounds that influences the concentration of splicing factors. GVs can also increase or decrease the efficiency of splicing of an alternative exon. The disruption of the tightly regulated splicing of an alternative exon in both directions will result in an imbalance in the relevant protein isoforms[21] and probably in alterations of cell biochemistry. For most genes, there is little information on patterns of alternative splicing and on the functions of the different protein isoforms that result from this process. However, such information is important for evaluating the functional significance of individual GVs that affect splicing, and it seems probable that alternatively spliced exons will be more sensitive to SpaGV-induced changes.

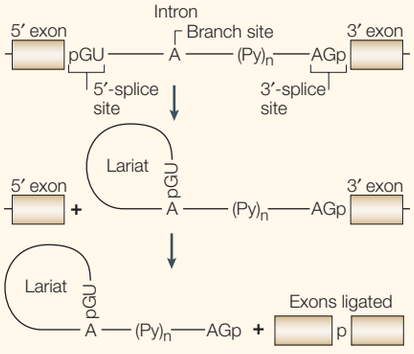


Figure 1 | **Splicing reactions and essential splicing signals.** There are several conserved motifs in the nucleotide sequences near the intron–exon boundaries that act as essential splicing signals: GU and AG dinucleotides at the exon–intron and intron–exon junctions, respectively (5′- and 3′-splice sites), a polypyrimidine tract $(Py)_n$ and an A nucleotide at the branch site. Splicing takes places in two TRANSESTERIFICATION steps. In the first step, the 2′-hydroxyl group of the A residue at the branch site attacks the phosphate at the GU 5′-splice site. This leads to cleavage of the 5′ exon from the intron and the formation of a lariat intermediate. In the following step, a second transesterification reaction, which involves the phosphate (p) at the 3′ end of the intron and the 3′-hydroxyl group of the detached exon, ligates the two exons. This reaction releases the intron, still in the form of a lariat.

In general, it is still unclear how frequently SpaGVs are involved in the onset of disease. However, studies that systematically addressed this question for two monogenic diseases showed that SpaGVs were involved in approximately 50% of ATAXIA TELANGIECTASIA and NEUROFIBROMATOSIS TYPE 1 cases[3,4], many of which would have been overlooked or not correctly classified as splicing errors because they did not affect the invariant consensus splice sites.

Exonic SpaGVs

The effects of key mutations, such as exonic insertions and deletions, nonsense substitutions and substitutions at the invariant consensus splice sites, are relatively easy to predict and will not be discussed here. Instead, we focus on substitutions located elsewhere in the exons or introns that cause aberrant splicing.

Genetic analyses in several gene systems, including ataxia telangiectasia mutated (*ATM*), neurofibromin 1 (*NF1*) and cystic fibrosis transmembrane regulator (*CFTR*) genes, have shown that many exonic genomic variants (eGVs) that are classified as neutral or missense have been subsequently found to affect the splicing process[3–5,8]. In the cases of *ATM* and *NF1* genes, 13% and 11%, respectively, of SpaGVs would have been erroneously classified as frameshift, missense or nonsense mutations if the analysis had been limited to DNA-coding sequences. Moreover, systematic SITE-DIRECTED MUTAGENESIS studies of the sequences that surround the SpaGVs have shown the extent to which the effects of GVs can be pleiotropic and unpredictable[5–7].

Exonic splicing enhancers and silencers. The widely accepted view is that if an eGV changes the splicing efficiency, it is because it affects a sequence that binds to a specific splicing factor[8,10,13]. Exon splicing enhancers (ESEs) are one class of such sequences. These enhancers are classically considered to be binding sites for serine arginine (SR) proteins, which often function to activate splicing. Several sequence motifs that act as enhancers *in vitro*, and that specific SR proteins recognize, have been used to computationally predict consensus sequences for exonic enhancers. Alignment of consensus sequences and the frequencies of the individual nucleotides at each position were used to calculate a score matrix. In some genes, changes in SR protein-binding scores have been shown to correlate with the changes in the splicing efficiency that a sGV or a missense GV induce[8,10]. Exon splicing silencers (ESSs) are less well characterized, but they seem to interact with negative regulators that often belong to the heterogenous nuclear ribonucleoprotein particles (hnRNP). So, to generate exon skipping, eGVs have been classically considered to inactivate an ESE or create an ESS (for example, see BOX 1).

The composite exonic regulatory element of splicing. This classical view of eGVs influencing exon skipping by affecting either an ESE or an ESS does not fully encompass the complexities of the process, as splicing of the *CFTR* exon 12 illustrates. In this exon, different eGVs can both increase or decrease exon skipping[5] (BOX 2). Systematic site-directed mutagenesis and *in vivo* functional splicing assay studies with hybrid minigenes identified two regulatory regions (composite exonic regulatory elements of splicing (CERES) 1 and 2) that contain overlapping enhancer and silencer functions[5]. These studies showed that different substitutions in these regions have completely different effects on splicing patterns. They also revealed that a surprising number of eGVs in these regions could cause disease through splicing aberrations: half the CERES substitutions in *CFTR* exon 12 resulted in less than 15% exon inclusion[5]. These results indicate that if we are to understand the complexities of splicing regulation, there need to be more studies of silencers and enhancers that use extended mutagenesis and *in vivo* functional assays. If such studies are performed in other genes, they could reveal that many splicing regulatory sequences that are currently viewed as pure enhancer or silencer elements are in fact composite elements like CERES. Moreover, these studies highlight the extreme sensitivity of the splicing machinery to changes induced by eGVs.

Intronic SpaGVs

Intronic SpaGVs are often located within approximately 50 bp of the splice sites but can be also found deep in an intron that is thousands of bases away from exon–intron junctions.

The flanking regions of the canonical splice sites. GVs that occur near the essential GU and AG splicing signals, 5′-splice site GVs (5′ ssGVs) and 3′-splice site GVs (3′ ssGVs) (FIG. 2b) can easily be misclassified as benign. Most sequencing of disease genes focuses on the exons, so it is not surprising that these are the most frequently identified intronic

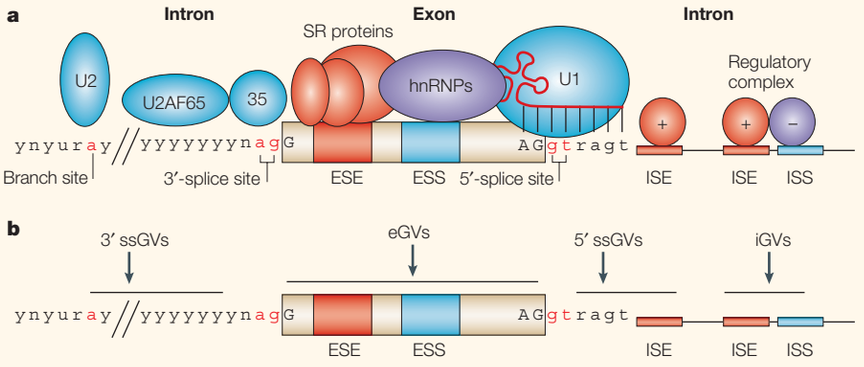


Figure 2 | **Regulatory elements in pre-mRNA splicing and GVs that can affect them. a** | The essential splicing signals that define the exon boundaries are relatively short and poorly conserved sequences. Only the GU and the AG dinucleotides that directly flank the exon (at the 3′ and 5′ ends, respectively) and the branch-point adenosine (all in red) are always conserved. In most cases, there is also a polypyrimidine tract of variable length (the consensus symbol 'y' represents a pyrimidine base — cytosine or thymine) upstream of the 3′-splice site. The branch point is typically located 18–40 nucleotides upstream from the polypyrimidine tract. Components of the basal splicing machinery bind to the consensus sequences and promote assembly of the splicing complex. This multiprotein complex, known as a spliceosome, performs the correct identification of the splicing signals and catalysis of the cut-and-paste reactions (FIG. 1). Five small nuclear ribonucleoproteins (snRNPs) and more than 100 proteins make up the spliceosome. The U1 snRNP binds to the 5′-splice site, and the U2 snRNP binds the branch site through RNA–RNA interactions. Additional enhancer and silencer elements in the exons (EXON SPLICING ENHANCER (ESE); EXON SPLICING SILENCER (ESS)) and/or introns (INTRON SPLICING ENHANCER (ISE); INTRON SPLICING SILENCER (ISS)) allow the correct splice sites to be distinguished from the many cryptic splice sites that have identical signal sequences. *Trans*-acting splicing factors can interact with enhancers and silencers and can accordingly be subdivided into two main groups: members of the serine arginine (SR) family of proteins and of the HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN PARTICLES (hnRNPs). In general, SR protein binding at ESE facilitates exon recognition whereas hnRNPs are inhibitory. Protein–protein interactions in the spliceosome that modulate the recognition of the splice sites are the probable cause of splicing inhibition or activation. **b** | Genomic variants (GVs) can affect different splicing regulatory elements, leading to aberrant splicing. Exonic GVs (eGVs) can either change the amino acid, result in synonymous GVs in exons (sGVs) or introduce a nonsense codon. Intronic GVs might be located within approximately 50 bp from the splice sites (that is, 3′-splice site GVs (ssGVs) and 5′ ssGVs) or deep in the introns (intronic GVs (iGVs)).

substitutions. However, the identity of the nucleotides that flank the 5′-GU splice signal shows some variability[22], so the effect of a particular 5′ ssGV on splicing efficiency is not always obvious. For example, whereas a G is predominantly found 5 nucleotides downstream of the 5′-splice site, approximately 25% of all normal exons have a different nucleotide at that position. So, it is not possible to predict that a specific nucleotide substitution will be deleterious if the new nucleotide is normal in other contexts[23]. On the other hand, current computer programs do not recognize many constitutively used 5′-splice sites. So, even if several 5′ ssGVs have lower scores than the wild-type counterparts, as predicted by computer programs[24], their functional effect should be directly assessed, either by studying the patient's transcript or by using a functional splicing assay[23]. The evaluation of 5′ ssGVs in flanking nucleotides further downstream of the consensus sequence is even more complex. For example, a 5′ ssGV (A to T in position +32) that is adjacent to *BRCA1* exon 22 causes skipping of this exon[25]. Human mutation databases contain several such putative splicing variants that are not characterized by transcript analysis (see, for example, the Cystic Fibrosis Mutation Database in the online links box). It is now becoming clear that additional sequences farther downstream from the 5′-splice site facilitate its recognition, such as the intronic G-triplets, which are frequently present near the 5′-splice site[26], and the polypyrimidine-rich elements that bind to the splicing factor TIA-1 (REF. 27). 5′ ssGVs might interfere with binding sites for splicing factors at these elements. The same difficulty in assessing a splicing defect might also apply for sequences that flank the 3′-splice site, in which 3′ ssGVs can occur in the highly variable polypyrimidine tract. In these cases, a direct transcript analysis and/or splicing assay should also be performed.

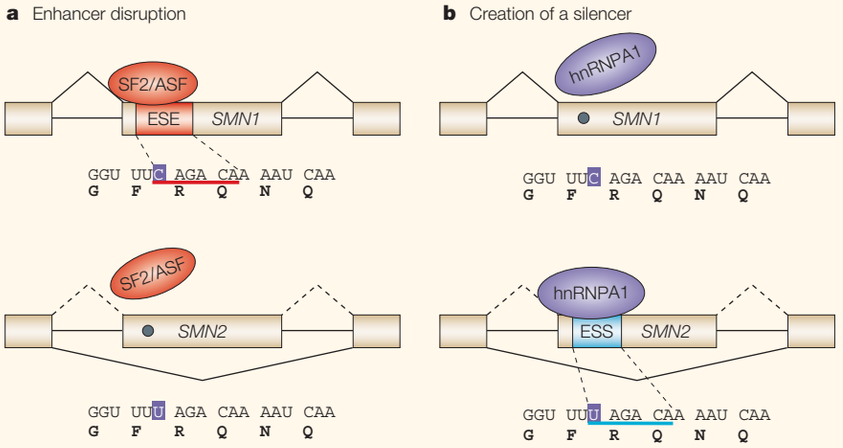
Deep intronic variants. The functional significance of GVs that are located deep in the introns (iGVs; FIG. 2b) is even more difficult to evaluate than that of GVs that are adjacent to the splice site. Despite not being close to any obvious regulatory sequences, such GVs can, for example, cause the activation of 'cryptic exons'. In these cases, most commonly the iGV creates a new splice site that defines the boundary of the cryptic exon (see, for example, REFS 28,29). Interestingly, potential 3′-splice sites that are present in more than one million repetitive Alu sequences in the human genome[16] might allow the inclusion of PSEUDO EXONS[16]. Some iGVs might lead to the partial inclusion of these repetitive Alu elements in the coding exons[30]. In addition, iGVs might affect efficient intron splicing processivity[6]. For example, a deletion of 4 bases in intron 20 of *ATM* has led to the discovery of an unknown intronic splicing element that is involved in the efficient and accurate pre-mRNA processing of long introns[6]. This element is located approximately 2 kb and 0.6 kb from the preceding and following exons, respectively. The disruption of this intronic processivity element activates a pseudo exon that contains this element. This leads to an aberrant mRNA and consequently to the disease phenotype. Site-directed mutagenesis has shown that even a single point mutation in the same region could cause this defect[6].

Hidden pathways to disease

The possibility of silent SNPs affecting splicing through the inactivation of existing *cis*-regulatory elements or the creation of new ones is at least well known, even if not generally acted on. However, the relationship between SpaGVs and the onset of disease is not necessarily as straightforward as this.

RNA secondary structure. Most RNA-binding proteins interact with ssRNA, so protein-binding target sequences are usually present in a loop in the RNA and not in a stem[31]. For this reason, GVs that modify the pre-mRNA secondary structure can affect the proper display of target RNA sequences and thereby change the splicing efficiency[32]. We know that sGVs might induce different structural folds in the pre-mRNA structure both *in vitro* and *in vivo*[33] and so might affect the recruitment of both positive and negative splicing factors[34–36]. Indeed, in the *tau* gene, 5′-ssGVs change the RNA structure, which disrupts the 5′-ss interactions and therefore causes aberrant splicing[2,37,38]. Computer-assisted methods can predict the secondary structures of naked RNAs, although the predictions must be confirmed by experimental methods[32]. However, it should be noted that it is extremely difficult with the current techniques to characterize the real RNA secondary structure in large pre-mRNA molecules that are present as ribonucleoprotein complexes *in vivo*.

Box 1 | **Does the sGV in exon 7 of *SMN2* destroy an ESE or create an ESS?**



Homozygous loss of function of the survival motor neuron 1 (*SMN1*) gene causes spinal muscular atrophy (SMA), a paediatric neurodegenerative disorder. Both normal and affected individuals have an almost identical paralogous gene, *SMN2*, that differs from *SMN1* in that it has a unique, translationally silent C→T variation in exon 7. Owing to this substitution, approximately 80% of *SMN2* mRNAs skip exon 7 and produce a truncated unstable protein. The effect of this synonymous GV (sGV) has received particular attention because reactivating the splicing of *SMN2* might represent a new strategic therapy in patients with SMA. Two models have been proposed to explain how this sGV causes exon-7 skipping. The sGV might cause the inactivation of an exon splicing enhancer (ESE) that binds the serine arginine (SR) protein SF2/ASF[61] (a) or it might create a new exon splicing silencer (ESS) that binds hnRNPA1 (REF. 9) (b). *In vitro* splicing assays and direct-binding experiments provide evidence in favour of both of these models[9,61]. For example, *in vivo* ablation of hnRNPA1 function increased the inclusion of the *SMN2* exon[9]. However, it is not clear whether the influence that hnRNPA1 has on splicing efficiency is specific to the SMN system or whether it is a general effect on alternatively spliced exons. It is also possible that both models are correct — that is, an ESE is inactivated and, simultaneously, an ESS is created. These studies emphasize the fact that the results of *in vitro* studies must be interpreted with caution. Extended mutagenesis studies *in vivo*, and perhaps studies of ablation of protein expression (such as RNA-interference experiments), should follow such *in vitro* studies to prove the real effect of an exonic genomic variant.

Modifiers of SpaGVs? Several SpaGVs do not inactivate genes completely, but give rise to a mixture of normal and aberrant transcripts[5,19,20]. It is possible that if the amount of normal transcript is below a critical threshold, possibly in specific tissues, this might cause a disease. For example, SpaGVs in *CFTR* have been associated with so-called non-classical forms of cystic fibrosis (CF). These less severe forms of the disease (see BOX 2) are restricted to a few organs and are variable in phenotype. Strikingly, both normal individuals and non-classical CF patients can have the same SpaGVs but different amounts of aberrant transcripts, indicating that a modifier gene is involved. The best-studied example is the *CFTR* UGmUn polymorphism at the 3′ end of intron 8, which modulates the amounts of aberrant exon-9 skipping[17–20]. The allele with the highest number of UG and the lowest number of U repeats (UG13U3) produces near-complete exon-9 skipping and severe CF[39]. By contrast, the alleles with lower UG and higher U repeats (UG11U7 or UG10U9) produce low amounts of exon skipping and are found in normal individuals. The intermediate alleles (such as UG11U5 or UG12U5) induce partial exon skipping and can be found in normal individuals and in patients with milder non-classical CF[17–20]. The differences in severity of the phenotype might result from different amounts of aberrant splicing that are modulated by the individual tissue concentrations of general (SR proteins and hnRNPs)[40] or specific (TDP43)[39,41] regulatory splicing factors. One interesting hypothesis to account for this pattern is that a modifier gene(s) that codes for a splicing factor(s) can change the ratio of normal to abnormal transcripts and so determine whether such SpaGVs lead to a disease.

A series of elegant studies in the mouse proved the equivalent hypothesis for an instance in which a genetic variation in a putative splicing factor influences neurological disease susceptibility[42–44]. These studies focused on a mutation that induces variable amounts of exon skipping (6–12%) in a neuronal sodium channel gene (*Scn8a*). Less severe disease is observed when approximately 12% of the transcripts are normally spliced. A mutation in a gene that encodes a homologue to the human U1C splicing factor modifies splicing *in trans* and leads to a lower amount of normal transcript (approximately 6%) and juvenile lethality.

It seems plausible that in humans too, changes in the level of expression of splicing factors and/or in their functionality might modulate the severity of diseases that are linked to SpaGVs. Certainly in humans, several splicing factors have tissue-specific expression and multiple alternatively spliced isoforms, which might lead to clinically significant differences in the amount of normal transcript that is derived from the SpaGV allele. However, in humans it will be even more difficult to identify the splicing modifiers. These striking studies in the mouse should be taken into account when assessing the effect of GVs in humans. It will be important to identify which splicing factors affect the amount of aberrant transcripts for each SpaGV. The identification of the tissue-specific expression pattern and alternatively spliced isoforms of these factors, as well as the real mechanism that underlies the splicing defect, will also be important goals.

In the future, we might be able to use microarrays not only to study tissue-specific alternative splicing patterns[45] but also to measure individual variability of splicing-factor concentrations and their pattern of alternative splicing. Such data would be a useful tool for characterizing individual tendencies for specific splicing patterns and the association of these with disease susceptibility.

The transcription–splicing connection. In human cells, transcription and splicing are coordinately regulated in the nucleus in both a temporal and a spatial fashion[46,47]. However, most studies of the effects of different GVs on splicing are done *in vitro* on preformed RNA molecules or in minigenes with heterologous promoters, so the importance of this coordinated regulation *in vivo* is often overlooked. For example, it has been shown that in an intron, the site at which an RNA polymerase II pauses during transcription can affect splice-site selection *in vivo*[48]. iGVs in these pausing sites might cause splicing alterations by changing the timing of presentation of regulatory splicing elements in the nascent transcripts[48,49]. The order of intron removal has been shown to

determine the effect of splicing mutations at a canonical splice site — that is, to determine whether an exon is skipped or an intron is retained[50,51]. SpaGVs that potentially affect the order of intron removal might eventually result in a splicing defect.

Promoter or enhancer variants should also be considered for their effect on the splicing process. Approximately one-third of natural promoter GVs have recently been found to alter gene expression to a functionally relevant extent[52]. However, promoter GVs can also modify the splicing efficiency of several alternatively spliced exons[53–57], possibly by affecting the loading of splicing factors in the polymerase complex, by modifying transcription kinetics and/or through a direct effect on polymerase II phosphorylation. If genomic variants at the promoter affect the binding of a transcription factor that modulates any of these processing events, it might change the splicing pattern of sensitive exons, thereby contributing to a disease phenotype.

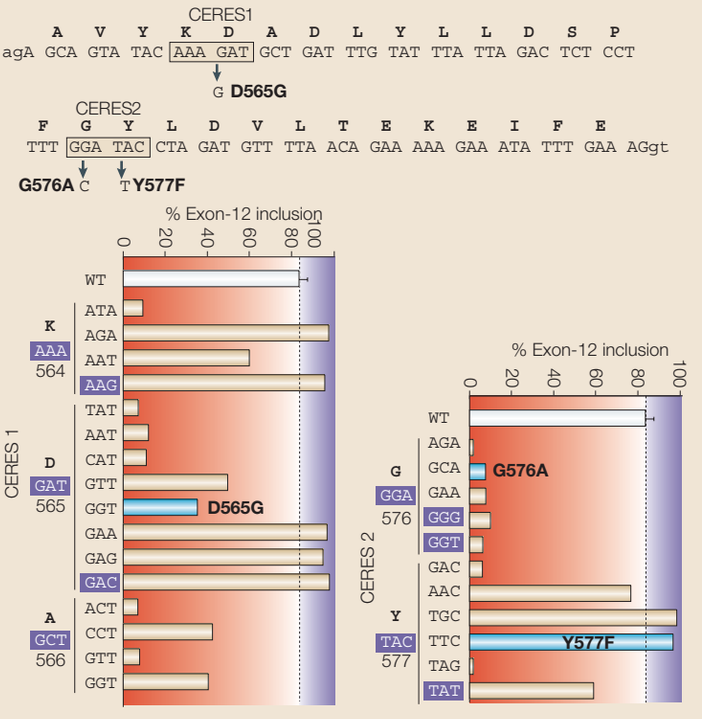
Identifying disease-causing SpaGVs

The GVs that are identified in clinical genetic screening might be simple polymorphic markers or disease-causing mutations, so a key question is how to prioritize them for further study. If we focus on their possible effect on splicing, the first difficulty of this task lies in our incomplete knowledge of the molecular mechanism involved. So, an important goal is to identify all the regulatory elements that affect splicing and the mechanisms of their action.

During genetic screening, it is a common experience to end up with a large fraction of orphan variants with an unclear pathogenetic role. These GVs — either exonic or intronic — are the best candidates for splicing evaluation. We must take into account the genetic setting in which the GVs are identified when prioritizing them. If we are looking at a clear disease-associated gene in an isolated patient, every GV should be considered as a potential SpaGV. In practical terms, once the most common mutations that are reported for that gene are excluded, every other nucleotide change that is identified should be carefully analysed for its effect on splicing. Obviously, if several affected subjects from the same family are available, the presence of a GV in their gene and not in non-affected family members increases the chance that it is a SpaGV. For complex trait diseases, prioritization of GVs for splicing evaluation will be more difficult as any one GV is only expected to have a small influence on the phenotype, and any effect on the splicing process will probably be moderate. It seems that candidate genes that are involved in complex trait diseases usually have a relatively low number of GVs (approximately 5–7 per gene considering exons and periexonic sequences[58,59]). So, if this estimation is correct, it is feasible to use functional splicing assays to systematically study the effect of common GVs on splicing.

Box 2 | **The effect of eGVs at the CERES elements in *CFTR* exon 12**

Mutations in the cystic fibrosis transmembrane regulator (*CFTR*) gene cause cystic fibrosis (CF). The disease causes pathological features of variable severity in the lungs, pancreas, sweat glands, testis, ovaries and intestine. Some patients only show evidence of the disease in a subgroup of these organs. These non-classical CF forms include late-onset pulmonary disease, male sterility owing to congenital bilateral absence of vas deferense and idiopathic pancreatitis. Among the 1,100 putative mutations reported in *CFTR* (see the Cystic Fibrosis Mutation Database in the online links box), most are missense genomic variants (GVs) or intronic genomic variants (iGVs), and for several of them, a clear disease-causative role is lacking. In *CFTR* exon 12, the GV Y577F is found in patients with classical CF. However, two other missense substitutions, D565G and G576A, are variably associated with non-classical CF. A systematic analysis of the effect on splicing of GVs in the regions in which these substitutions occur showed that both D565G and G576A tend to increase skipping of exon 12, which indicates that it could be their effect on splicing that underlies their association with non-classical CF. The upper part of the figure shows the nucleotide sequence of the *CFTR* exon 12 (in uppercase) along with the position of the two COMPOSITE EXONIC REGULATORY ELEMENTS OF SPLICING (CERES; boxed), the amino-acid sequence and the three missense substitutions. The splice sites are in lowercase. The lower histograms show the percentage of exon 12+ and exon 12– transcripts of the site-directed mutants at the two CERES that were analysed using HYBRID MINIGENES. The site-directed mutants that do not change the amino-acid code are boxed.



These data indicate that D565G and G576A lead to inefficient recognition of *CFTR* exon 12 *in vivo*, whereas Y577F does not. These missense variants are located in CERES elements. The sequence composition of the splicing regulatory elements in *CFTR* exon 12 overlaps with the codon-usage preferences and the requirements for protein function. Some of the site-directed mutants that induce a high amount of aberrant exon skipping are at the third position of the codon usage and do not change the amino-acid code. CERES-like elements have also been observed in *CFTR* exon 9 (REF. 7). It seems possible that mild non-classical CF forms are in part the result of a differential efficiency of splicing. GVs could cause such differences, along with variability in the concentrations of regulatory splicing factors among individuals and tissues[39] that lead to variable amounts of exon skipping and consequent loss of function. Adapted with permission from REF. 5 © Oxford University Press (2003).

As exonic SpaGVs overlap with protein-coding sequences, the first approach could be to take into account those exonic GVs that do not clearly show an obvious pathologic effect at the protein level, such as synonymous substitutions and conservative changes. One potentially useful approach that is already available for exonic GVs is the *in silico* prediction of regulatory elements. Prediction of ESE sequences that bind to SR proteins is useful for indicating the effect of GVs on splicing in some gene systems[10] but not in others[5,7]. In the future, the definition of consensus binding sites for other splicing factors, and the combined action of these sequences, will probably allow better computational models to be developed. In addition, statistical analysis of exon–intron and splice-site composition can be used to computationally predict which sequences have ESE activity[60]. The possible role of RNA secondary structures will also need to be considered when formulating more realistic predictive computational models. The current approach restricts gene scanning to exons and periexonic sequences: the effect of deep intronic variants can only be spotted when a sample of a patient's derived RNA is available for analysis.

At present, a splicing functional assay is the only reliable way to establish the disease-causing role of a particular SpaGV. Any genomic region of interest (that is, exons and short intronic flanking regions) that contains an orphan mutation that might cause a splicing defect can be amplified from normal and affected individuals and cloned into a minigene. The minigene plasmid is then transfected into an appropriate cell line; here, it will be transcribed by RNA polymerase II and the resulting pre-mRNA will be processed to obtain a mature mRNA. The mRNA splicing pattern is analysed mainly by reverse transcriptase (RT)-PCR with primers that are specifically designed to amplify processed transcripts derived from the minigene, to distinguish them from the cell's endogenous transcripts[6,7,9,23,50,61]. The whole process takes approximately twelve hours of a skilled technician's time and a result can be obtained in a week. This assay is extremely reproducible and suitable for routine analysis of potential SpaGVs. An alternative method that is useful to study splicing is the *in vitro* splicing assay. In this procedure, labelled pre-formed RNA molecules that are transcribed with bacterial polymerases are incubated in the presence of nuclear extracts and the resulting spliced products are resolved on polyacrylamide denaturing gel[9,10,41,48,61,62]. With this assay, the intermediates of the splicing reactions, such as the lariat, can be evaluated and more molecular mechanistic studies can be performed. However, in contrast to the hybrid minigene, the *in vitro* assay only allows relatively short sequences to be studied, which frequently contain a reduced version of the original intron. This system is not easily standardized, although there are some reliable specific heterologous assays for testing putative ESEs[62,63]. Most importantly, *in vitro* assays do not take into account the fact that transcription and splicing are intimately connected in the cell.

Glossary

ALTERNATIVELY SPLICED ISOFORMS
RNA isoforms that are generated by alternative use of splice sites, which leads to variation in which exons are included in the mRNA and subsequently translated.

ATAXIA TELANGIECTASIA
An autosomal recessive disorder that involves cerebellar degeneration, immunodeficiency, chromosomal instability, radiosensitivity and cancer predisposition.

COMPOSITE EXONIC REGULATORY ELEMENT OF SPLICING (CERES). Short exonic RNA sequences (5–12 bases) that contain overlapping enhancer and silencer sequences. The presence of such elements is indicated when scanning mutagenesis analyses reveal that different mutations at nearby positions or even at the same position have opposite effects on splicing efficiency.

CRYPTIC SPLICE SITES
Pseudo splice sites that are activated as a consequence of a mutation elsewhere in the gene.

EXON SKIPPING
Exclusion of an exon that is normally included in the mRNA.

EXON SPLICING ENHANCER AND EXON SPLICING SILENCER (ESE, ESS); INTRON SPLICING ENHANCER AND INTRON SPLICING SILENCER (ISE, ISS)
Sequences in the pre-mRNA that enhance or reduce the efficiency of splicing. In general, exonic enhancers or silencers are shorter (~6 bases) than the intronic ones, which can be hundreds of bases long.

HETEROGENOUS NUCLEAR RIBONUCLEOPROTEIN PARTICLES
(hnRNP). A class of diverse RNA-binding proteins that associate with nascent pre-mRNA.

HYBRID MINIGENE
A simplified laboratory version of a natural gene that contains one of more of the gene's exons and introns.

NEUROFIBROMATOSIS TYPE 1
An autosomal dominant disorder that is particularly characterized by cafe-au-lait spots and fibromatous tumours of the skin.

PSEUDO EXON
A pre-mRNA sequence that resembles an exon, both in its size and the presence of flanking pseudo splice sites, but that the splicing machinery does not normally recognize.

PSEUDO SPLICE SITES
Sequences that are identical to normal splice sites but that are not normally used in splicing.

SIMPLE SEQUENCE REPEAT
A sequence that consists largely of a tandem repeat of a specific *K*-mer (such as (TG)11).

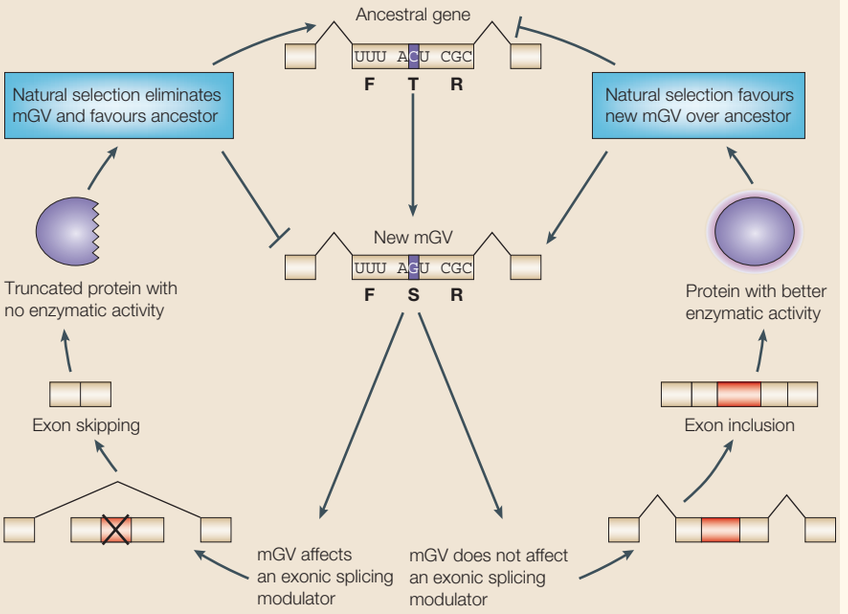
SITE-DIRECTED MUTAGENESIS
A method that is used to substitute a specific nucleotide into a DNA sequence.

TRANSESTERIFICATION
A reaction that breaks and makes chemical bonds (in this case, phosphodiester bonds) in a coordinated transfer so that energy is required.

Conclusion and perspectives

The genomic diversity and genomic pathology data that are currently available have revealed the extent of our ignorance of the basic molecular mechanisms that underlie the pre-mRNA splicing process[3,4,8]. In fact, human pathology continues to give us pointers to new and unexpected modulatory elements of splicing. However, we should not rely only on the chance identification of new splicing regulatory elements. It is worrying that many clinically relevant mutations are slipping through the net because their effect on the splicing process is not even considered. The data on the effect of silent GVs on splicing are impressive[3,5–9,11,14,38,61] and warrant a closer look at the mechanisms involved. We believe that in the rush to catalogue genomic sequences and genome sequence variants, not enough studies are being done to investigate the functional effects of the variation. Nowadays, it seems more exciting to use microarrays to look at differences between cases and controls in the transcriptome and the proteome rather than to reflect on the underlying mechanisms involved. For example, little consideration has been given to the effects of the selection of coding sequences before translation. Until now, we have been more concerned about protein-function optimization and the selective pressures that act on it, disregarding the fact that the coding sequences first have to be included in the mRNA. We hope that this article will help to make researchers who are not directly involved in the field more aware that some exon sequences have important regulatory roles before translation. We believe that the coexistence of exonic splicing regulatory elements with amino-acid coding capacity might restrict the evolutionary selection of codon variants that could improve protein function. It follows that, at least in a fraction of the exons that are present in the genome, sub-optimal protein function might be tolerated

Box 3 | **The primary selective pressure on exons is for their inclusion in mRNA**



Natural selection acts on exons in at least two ways. First, sequence variants that enhance an exon's chance of being included in the mature mRNA will be favoured over those that do not. Second, sequence variants that encode amino-acid sequences that enhance the particular function that they perform will be favoured over those that do not. The first selective pressure might be more important than the second, because exon inclusion in the final mRNA is a pre-condition for its translation. For this reason, exonic sequences cannot change freely. It follows that protein variability — the substrate of selection — is restricted to the nucleotide changes that the splicing machinery can tolerate. The model in the figure shows that the selection of a new amino acid that leads to a better enzyme can occur only if the codon substitution, caused by a missense genomic variant (mGV), does not affect an exonic regulatory element. In this model, we assume that the threonine to serine change at the catalytic site will produce a more active enzyme that has a selective advantage. However, the C to T substitution has to be compatible with the splicing machinery that identifies the exon (including exonic splicing modulators such as exon splicing enhancer (ESE), exon splicing silencer (ESS) or secondary structures). If inclusion is guaranteed, then the amino-acid change is favoured. If not, exon skipping will result in an inactive enzyme and the ancestral threonine that produces a suboptimal protein will be kept to ensure the exon-inclusion step. Alternatively, an associated variation in this exonic region can precede the codon change and create a redundant enhancer that might compensate for the loss of the original enhancer and so allow normal splicing (not shown in the model).

to allow for the persistence of sequences that are essential for exon inclusion (BOX 3).

The selective pressure on exonic splicing modulator sequences is more important than it is on the protein that the gene encodes. This seemingly obvious statement has important implications for molecular pathology studies. In particular, no GV, regardless of how innocent it looks, can be considered to be benign before its effect on splicing has been assessed.

Mutations in exonic or intronic regulatory elements that cause severe splicing defects might just be the tip of the iceberg. There might also be many GVs that cause partial splicing defects that are only pathogenic in specific tissues under the influence of a set of specific regulatory splicing factors. Similar attention should be paid to GVs that could cause defects during the initiation, elongation and termination stages of transcription. Similar to splicing, these processes are rarely considered when assessing the clinical significance of GVs. It is clear that the combined tools of genome, transcriptome and proteome analysis must be used to analyse the functional significance of GVs. The more we know about the basic function, the more probable it is that computer programmes will accurately predict *in silico* the effect of a new GV.

Franco Pagani and Francisco E. Baralle are at the International Centre for Genetic Engineering and Biotechnology, Padriciano 99, 34012 Trieste, Italy.

Correspondence to F.E.B.
e-mail: baralle@icgeb.org

doi:10.1038/nrg1327

1. Thude, H., Hundrieser, J., Wonigeit, K. & Schwinzer, R. A point mutation in the human *CD45* gene associated with defective splicing of exon A. *Eur. J. Immunol.* **25**, 2101–2106 (1995).
2. Varani, L. *et al.* Structure of *tau* exon 10 splicing regulatory element RNA and destabilization by mutations of frontotemporal dementia and parkinsonism linked to chromosome 17. *Proc. Natl Acad. Sci. USA* **96**, 8229–8234 (1999).
3. Ars, E. *et al.* Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. *Hum. Mol. Genet.* **9**, 237–247 (2000).
4. Teraoka, S. N. *et al.* Splicing defects in the ataxia-telangiectasia gene, *ATM*: underlying mutations and consequences. *Am. J. Hum. Genet.* **64**, 1617–1631 (1999).
5. Pagani, F. *et al.* New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in *CFTR* exon 12. *Hum. Mol. Genet.* **12**, 1111–1120 (2003).
6. Pagani, F. *et al.* A new type of mutation causes a splicing defect in *ATM*. *Nature Genet.* **30**, 426–429 (2002).
7. Pagani, F., Buratti, E., Stuani, C. & Baralle, F. E. Missense, nonsense and neutral mutations define juxtaposed regulatory elements of splicing in *CFTR* exon 9. *J. Biol. Chem.* **278**, 26580–26588 (2003).
8. Cartegni, L., Chew, S. L. & Krainer, A. R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet.* **3**, 285–298 (2002).
9. Kashima, T. & Manley, J. L. A negative element in *SMN2* exon 7 inhibits splicing in spinal muscular atrophy. *Nature Genet.* **34**, 460–463 (2003).
10. Liu, H. X., Cartegni, L., Zhang, M. Q. & Krainer, A. R. A mechanism for exon skipping caused by nonsense or missense mutations in *BRCA1* and other genes. *Nature Genet.* **27**, 55–58 (2001).
11. Jiang, Z. *et al.* Mutations in *tau* gene exon 10 associated with FTDP-17 alter the activity of an exonic splicing enhancer to interact with Tra2 β. *J. Biol. Chem.* **278**, 18997–19007 (2003).
12. Ishii, S., Nakao, S., Minamikawa-Tachino, R., Desnick, R. J. & Fan, J. Q. Alternative splicing in the α-galactosidase A gene: increased exon inclusion results in the Fabry cardiac phenotype. *Am. J. Hum. Genet.* **70**, 994–1002 (2002).
13. Faustino, N. A. & Cooper, T. A. Pre-mRNA splicing and human disease. *Genes Dev.* **17**, 419–437 (2003).
14. Fernandez-Cadenas, I. *et al.* Splicing mosaic of the myophosphorylase gene due to a silent mutation in McArdle disease. *Neurology* **61**, 1432–1434 (2003).
15. Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72**, 291–336 (2003).
16. Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
17. Chillon, M. *et al.* Mutations in the cystic fibrosis gene in patients with congenital absence of the vas deferens. *N. Engl. J. Med.* **332**, 1475–1480 (1995).
18. Chu, C. S., Trapnell, B. C., Curristin, S., Cutting, G. R. & Crystal, R. G. Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA. *Nature Genet.* **3**, 151–156 (1993).
19. Larriba, S. *et al.* Testicular *CFTR* splice variants in patients with congenital absence of the vas deferens. *Hum. Mol. Genet.* **7**, 1739–1743 (1998).
20. Mak, V., Jarvi, K. A., Zielenski, J., Durie, P. & Tsui, L. C. Higher proportion of intact exon 9 *CFTR* mRNA in nasal epithelium compared with vas deferens. *Hum. Mol. Genet.* **6**, 2099–2107 (1997).
21. D'Souza, I. & Schellenberg, G. D. Determinants of 4-repeat *tau* expression. Coordination between enhancing and inhibitory splicing sequences for exon 10 inclusion. *J. Biol. Chem.* **275**, 17700–17709 (2000).
22. Zhang, M. Q. Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.* **7**, 919–932 (1998).
23. Baralle, M. *et al.* Identification of a mutation that perturbs *NF1* agene splicing using genomic DNA samples and a minigene assay. *J. Med. Genet.* **40**, 220–222 (2003).
24. Roca, X., Sachidanandam, R. & Krainer, A. R. Intrinsic differences between authentic and cryptic 5′ splice sites. *Nucleic Acids Res.* **31**, 6321–6333 (2003).
25. Matsushima, M. *et al.* Mutation analysis of the *BRCA1* gene in 76 Japanese ovarian cancer patients: four germline mutations, but no evidence of somatic mutation. *Hum. Mol. Genet.* **4**, 1953–1956 (1995).

26. McCullough, A. J. & Berget, S. M. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol. Cell. Biol.* **17**, 4562–4571 (1997).
27. Forch, P. *et al.* The apoptosis-promoting factor TIA-1 is a regulator of alternative pre- mRNA splicing. *Mol. Cell* **6**, 1089–1098 (2000).
28. Metherell, L. A. *et al.* Pseudoexon activation as a novel mechanism for disease resulting in atypical growth-hormone insensitivity. *Am. J. Hum. Genet.* **69**, 641–646 (2001).
29. Christie, P. T., Harding, B., Nesbit, M. A., Whyte, M. P. & Thakker, R. V. X-linked hypophosphatemia attributable to pseudoexons of the *PHEX* gene. *J. Clin. Endocrinol. Metab.* **86**, 3840–3844 (2001).
30. Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G. The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science* **300**, 1288–1291 (2003).
31. Buratti, E. *et al.* 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 (2004).
32. Muro, A. F. *et al.* Regulation of fibronectin EDA exon alternative splicing: possible role of RNA secondary structure for enhancer display. *Mol. Cell. Biol.* **19**, 2657–2671 (1999).
33. Shen, L. X., Basilion, J. P. & Stanton, V. P. Jr. Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc. Natl Acad. Sci. USA* **96**, 7871–7876 (1999).
34. Nagel, R. J., Lancaster, A. M. & Zahler, A. M. Specific binding of an exonic splicing enhancer by the pre-mRNA splicing factor SRp55. *RNA* **4**, 11–23 (1998).
35. Shi, H., Hoffman, B. E. & Lis, J. T. A specific RNA hairpin loop structure binds the RNA recognition motifs of the *Drosophila* SR protein B52. *Mol. Cell. Biol.* **17**, 2649–2657 (1997).
36. Damgaard, C. K., Tange, T. O. & Kjems, J. hnRNP A1 controls HIV-1 mRNA splicing through cooperative binding to intron and exon splicing silencers in the context of a conserved secondary structure. *RNA* **8**, 1401–1415 (2002).
37. Grover, A. *et al.* 5′ splice site mutations in *tau* associated with the inherited dementia FTDP-17 affect a stem-loop structure that regulates alternative splicing of exon 10. *J. Biol. Chem.* **274**, 15134–15143 (1999).
38. Jiang, Z., Cote, J., Kwon, J. M., Goate, A. M. & Wu, J. Y. Aberrant splicing of *tau* pre-mRNA caused by intronic mutations associated with the inherited dementia frontotemporal dementia with parkinsonism linked to chromosome 17. *Mol. Cell. Biol.* **20**, 4036–4048 (2000).
39. Buratti, E. *et al.* Nuclear factor TDP-43 and SR proteins promote *in vitro* and *in vivo CFTR* exon 9 skipping. *EMBO J.* **20**, 1774–1784 (2001).
40. Pagani, F. *et al.* Splicing factors induce cystic fibrosis transmembrane regulator exon 9 skipping through a nonevolutionary conserved intronic element. *J. Biol. Chem.* **275**, 21041–21047 (2000).
41. Buratti, E., Brindisi, A., Pagani, F. & Baralle, F. E. Why does a variation in the number of TG repeats in *CFTR* intron 8 influence disease penetrance? *Am. J. Hum. Genet.* (in the press).
42. Buchner, D. A., Trudeau, M. & Meisler, M. H. SCNM1, a putative RNA splicing factor that modifies disease severity in mice. *Science* **301**, 967–969 (2003).
43. Nadeau, J. H. Genetics. Modifying the message. *Science* **301**, 927–928 (2003).
44. Kearney, J. A. *et al.* Molecular and pathological effects of a modifier gene on deficiency of the sodium channel *Scn8a* ($Na_v1.6$). *Hum. Mol. Genet.* **11**, 2765–2775 (2002).
45. Johnson, J. M. *et al.* Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**, 2141–2144 (2003).
46. Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416**, 499–506 (2002).
47. Proudfoot, N. J., Furger, A. & Dye, M. J. Integrating mRNA processing with transcription. *Cell* **108**, 501–512 (2002).
48. Roberts, G. C., Gooding, C., Mak, H. Y., Proudfoot, N. J. & Smith, C. W. Co-transcriptional commitment to alternative splice site selection. *Nucleic Acids Res.* **26**, 5568–5572 (1998).
49. Proudfoot, N. J. Dawdling polymerases allow introns time to splice. *Nature Struct. Biol.* **10**, 876–878 (2003).
50. Attanasio, C., David, A. & Neerman-Arbez, M. Outcome of donor splice site mutations accounting for congenital afibrinogenemia reflects order of intron removal in the fibrinogen α gene (FGA). *Blood* **101**, 1851–1856 (2003).
51. Schwarze, U., Starman, B. J. & Byers, P. H. 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 (1999).
52. Hoogendoorn, B. *et al.* Functional analysis of human promoter polymorphisms. *Hum. Mol. Genet.* **12**, 2249–2254 (2003).
53. Pagani, F., Stuani, C., Zuccato, E., Kornblihtt, A. R. & Baralle, F. E. Promoter architecture modulates *CFTR* exon 9 skipping. *J. Biol. Chem.* **278**, 1511–1517 (2003).
54. Cramer, P. *et al.* Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. *Mol. Cell* **4**, 251–258 (1999).
55. Cramer, P., Pesce, C. G., Baralle, F. E. & Kornblihtt, A. R. Functional association between promoter structure and transcript alternative splicing. *Proc. Natl Acad. Sci. USA* **94**, 11456–11460 (1997).
56. Monsalve, M. *et al.* Direct coupling of transcription and mRNA processing through the thermogenic co-activator PGC-1. *Mol. Cell* **6**, 307–316 (2000).
57. Zhang, C. *et al.* Nuclear co-activator-62 kDa/Ski-interacting protein is a nuclear matrix-associated co-activator that may couple vitamin D receptor-mediated transcription and RNA splicing. *J. Biol. Chem.* **278**, 35325–35336 (2003).
58. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
59. Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
60. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–1013 (2002).
61. Cartegni, L. & Krainer, A. R. Disruption of an SF2/ASF-dependent exonic splicing enhancer in *SMN2* causes spinal muscular atrophy in the absence of *SMN1*. *Nature Genet.* **30**, 377–384 (2002).
62. Tian, M. & Maniatis, T. A splicing enhancer complex controls alternative splicing of doublesex pre-mRNA. *Cell* **74**, 105–114 (1993).
63. Lorson, C. L. & Androphy, E. J. An exonic enhancer is required for inclusion of an essential exon in the SMA-determining gene SMN. *Hum. Mol. Genet.* **9**, 259–265 (2000).

Acknowledgements

This work was supported by the Telethon Onlus Foundation, Italy, the Fondo Investimenti per la Ricerca di Base and the Associazione Italiana per la Ricerca sul Cancro.

Competing interests statement

The authors declare that they have no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
ATM | *BRCA1* | *CFTR* | *NF1* | *Scn8a* | *SMN1* | *SMN2*
OMIM: http://www.ncbi.nlm.nih.gov/Omim
cystic fibrosis | spinal muscular atrophy
Swiss-Prot: http://ca.expasy.org/sprot
hnRNPA1

FURTHER INFORMATION

Cystics Fibrosis Mutation Database:
http://www.genet.sickkids.on.ca/*cftr*

Access to this interactive links box is free online.