Alternative pre-mRNA splicing and proteome expansion in metazoans

Tom Maniatis & Bosiljka Tasic

Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA

The protein coding sequences of most eukaryotic messenger RNA precursors (pre-mRNAs) are interrupted by non-coding sequences called introns. Pre-mRNA splicing is the process by which introns are removed and the protein coding elements assembled into mature mRNAs. Alternative pre-mRNA splicing selectively joins different protein coding elements to form mRNAs that encode proteins with distinct functions, and is therefore an important source of protein diversity. The elaboration of this mechanism may have had a significant role in the expansion of metazoan proteomes during evolution.

Metazoans display an extraordinarily broad spectrum of functional and behavioural complexity. Complex organisms probably evolved by increasing the number of components that constitute them (for example, proteins), and/or by elaborating the relationships between components (for example, regulatory networks). The size of the proteome of an organism (the complete set of proteins expressed by the genome during the life of an organism) can be expanded during evolution by increasing the number of genes, by elaborating pre-existing mechanisms that generate protein diversity, and by inventing new mechanisms. For example, the mechanism of somatic-cell DNA rearrangement, which can generate virtually unlimited diversity of antibodies and T-cell receptors, arose during vertebrate evolution.

Mechanisms that increase protein diversity in all metazoans include the use of multiple transcription start sites[1], alternative pre-mRNA splicing[2–5], polyadenylation[6], pre-mRNA editing[7], and post-translational protein modifications[8]. Among these mechanisms, alternative pre-mRNA splicing is considered to be the most important source of protein diversity in vertebrates[2,3]. Here we review the mechanisms involved in the regulation of alternative pre-mRNA splicing, and discuss the effect of this process on expansion of the proteomes of multicellular organisms.

Exon recognition in constitutively spliced pre-mRNAs

The pre-mRNA splicing reaction is carried out by spliceosomes — multicomponent ribonucleoprotein complexes containing five small nuclear RNAs (snRNAs) and a large number of associated proteins[9–11]. Spliceosomes recognize 5′ and 3′ splice sites, which are located at exon–intron boundaries (Fig. 1). The assembly of spliceosomes is a highly dynamic process, culminating in the juxtaposition of 5′ and 3′ splice sites in the catalytic core of the complex. The splicing reaction occurs via a two-step mechanism. In the first step, the 5′ end of the intron is joined to an adenine residue in the branchpoint sequence upstream from the 3′ splice site to form a branched intermediate called an intron lariat. In the second step, the exons are ligated and the intron lariat is released[12].

A fundamental problem in pre-mRNA splicing is 'exon recognition', the process by which exons are distinguished from introns, and intron–exon boundaries are accurately defined[13,14]. The average size of a human exon is 150 nucleotides, whereas introns average around 3,500 nucleotides[15], and can be as large as 500,000 nucleotides[16]. Thus, the splicing machinery must recognize small exon sequences located within vast stretches of intronic RNA. Moreover, 5′ and 3′ splice sites are poorly conserved, and introns contain large numbers of cryptic splice sites, which match the loose 5′ or 3′ splice-site consensus. Cryptic splice sites are normally avoided by the splicing machinery, but can be selected for splicing when normal splice sites are altered by mutation. Identification of the correct splice sites is achieved by virtue of their proximity to exons[13,14]. Specific sequence elements in exons known as exonic splicing enhancers (ESEs) interact with SR proteins, a family of conserved serine/arginine-rich splicing factors[14,17]. These recruit the splicing machinery to the flanking 5′ and 3′ splice sites (Fig. 1). Thus, exon sequences are under multiple evolutionary constraints, as they must be conserved not only for protein coding but also for recognition by SR proteins[18,19]. The many cryptic splice sites present in introns could also be avoided by the presence of intronic splicing silencers (ISSs) and competition with splice sites flanking the recognized exons[20].

Once exon recognition is completed, the flanking splice sites must be joined in the correct 5′→3′ order to prevent exon skipping. This is probably accomplished, at least in part, through the mechanistic coupling of transcription and splicing[21]. In this model, splicing factors, which are bound to the carboxy-terminal domain (CTD) of RNA polymerase II, interact with exons as they emerge from the exit pore of the polymerase. This interaction tethers the newly synthesized exon to the CTD until the next exon is synthesized. In large introns, many hours could pass between the synthesis of the 5′ and 3′ splice sites. Although coupling transcription to splicing can prevent exon skipping in constitutively spliced pre-mRNAs, exon skipping does occur during alternative pre-mRNA splicing. In this case, the presence or absence of regulatory proteins can determine whether an exon is recognized and subsequently included in the mature mRNA.

Mutations that interfere with proper exon recognition result in a large number of human genetic diseases[22]. Approximately 15% of the single base-pair mutations that cause human genetic diseases result in pre-mRNA splicing defects[23]. Some of these mutations interfere with the function of normal 5′ and 3′ splice sites, thereby leading to the recognition of nearby pre-existing cryptic splice sites. Others actually create new splice sites that are used instead of

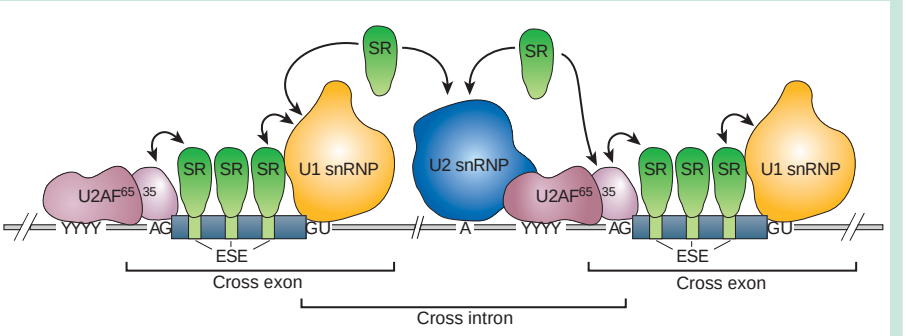

Figure 1 Exon recognition. The correct 5′ (GU) and 3′ (AG) splice sites are recognized by the splicing machinery on the basis of their proximity to exons. The exons contain exonic splicing enhancers (ESEs) that are binding sites for SR proteins. When bound to an ESE, the SR proteins recruit U1 snRNP to the downstream 5′ splice site, and the splicing factor U2AF (65 and 35 kDa subunits) to the pyrimidine tract (YYYY) and the AG dinucleotide of the upstream 3′ splice site, respectively. In turn, U2AF recruits U2 snRNP to the branchpoint sequence (A). Thus, the bound SR proteins recruit splicing factors to form a 'cross-exon' recognition complex. SR proteins also function in 'cross-intron' recognition by facilitating the interactions between U1 snRNP bound to the upstream 5′ splice site and U2 snRNP bound to the branchpoint sequence.

the normal ones. Finally, single base-pair mutations within exons can interfere with the binding of SR proteins, leading to exon exclusion from the mature mRNA[22,24,25]. For example, a translationally silent C-to-T mutation that occurs within an ESE of the human *survival of motor neuron 2* (*SMN2*) gene disrupts the binding site of the SR protein SF2/ASF and leads to exon skipping[22].

General mechanisms of alternative splicing

Alternative pre-mRNA splicing is the process by which multiple mRNAs can be generated from the same pre-mRNA by the differential joining of 5′ and 3′ splice sites. For example, exons can be extended or shortened, skipped or included, and introns can be removed or retained in the mRNA[22]. In some cases, exons are included in the mRNA in a mutually exclusive manner. Although there are very few examples in which the mechanisms of alternative splicing are known in detail, a general outline is understood. Regulatory proteins interact with specific sequences within pre-mRNAs and subsequently stimulate or repress exon recognition. These proteins bind directly to 5′ or 3′ splice sites, or to other pre-mRNA sequences called exonic or intronic splicing enhancers (ESEs or ISEs) and silencers (ESSs or ISSs). Enhancers and silencers stimulate or repress splice-site selection, respectively[9,17,22,26–28]. A common feature of proteins that regulate splicing is the presence of two functional domains, an RNA-binding domain and a protein–protein interaction domain. The best characterized RNA-binding domains are the RNA-recognition motif (RRM) and K-homology (KH) domains[29]. The three-dimensional structures of the two domains are distinct, as are the general features of the RNA-binding sites they recognize.

Most ESEs are recognized by SR proteins, which contain one or more RRM domains and an arginine/serine-rich (RS) protein–protein interaction domain[27,30]. SR proteins are essential, multifunctional splicing factors required at different steps of spliceosome assembly[17]. They are also thought to mediate cross-intron interactions between splicing factors bound to the 5′ and 3′ splice sites[17] (Fig. 1). Finally, SR proteins are required for cross-exon interactions in both constitutively and alternatively spliced pre-mRNAs[14,17,27,30].

The best characterized ESSs and ISSs are recognized by members of the heterogeneous nuclear ribonucleoprotein (hnRNP) family. hnRNP proteins are highly abundant RNA-binding proteins that lack an RS domain. Several family members contain an arginine/glycine-rich domain that may be involved in both RNA binding and interactions with other proteins. Among the best characterized members of this family are the hnRNP A1 (ref. 31) and hnRNP I (ref. 32) proteins. hnRNP proteins are diffusely distributed throughout the nucleus[33,34], unlike SR proteins, which co-localize with other splicing factors in nuclear speckles[27,30].

It is important to note that some proteins containing an RS domain function as splicing repressors, whereas certain hnRNP proteins have been shown to function as splicing activators[28]. SR proteins were originally defined by a set of common functional characteristics[17]. Subsequently, other proteins containing RS domains that do not share these characteristics were discovered, and they are referred to as SR-like proteins. For example, two recently identified SR-like proteins have been shown to function as splicing repressors, and were therefore named SR-repressor proteins[35]. We also note that there are splicing regulators that do not belong to either the SR or the hnRNP protein family. Some of these will be discussed below.

Remarkably, the role of regulatory proteins in splice-site selection can be affected by the promoter that generates the pre-mRNA[36]. Thus transcription of the same pre-mRNA from different promoters can produce distinct mRNAs. This mechanism of alternative splicing could be a consequence of the coupling between transcription and pre-mRNA splicing[21]. For example, particular SR proteins could be differentially recruited to RNA polymerase complexes assembled on different promoters, and then transferred to cognate splicing enhancers to promote the inclusion of specific exons[5].

Well characterized examples of alternative splicing

The mechanism of exon recognition in constitutively spliced pre-mRNAs provides the basis for positive and negative regulation of alternative splicing. The organization of regulatory sequences within pre-mRNAs (ESEs, ESSs, ISEs and ISSs) and the relative ratios of different regulatory proteins determine which splice sites are used in the splicing reaction. This, in turn, determines which exons are included in the mRNA[4,5,28].

The best characterized examples of regulated alternative splicing derive from studies of the *Drosophila* sex-determination pathway[37]. The key regulatory factors were first identified by genetic analyses[38,39], and then characterized in biochemical experiments. Multiple steps in this pathway are regulated by different mechanisms of alternative pre-mRNA splicing. The examples described below include both splice-site repression and activation.

The *Drosophila* female-specific protein Sex-lethal (SXL) represses male-specific 3′ splice sites in *transformer* (*tra*) and *sxl* pre-mRNAs by two distinct mechanisms. As shown in Fig. 2a, exon 2 of *tra* pre-mRNA is preceded by two alternative 3′ splice sites. The proximal and distal 3′ splice sites are used in males and females, respectively. Splicing to the male-specific 3′ splice site produces an mRNA containing a premature translational stop codon. By contrast, splicing to the female-specific 3′ splice site produces an mRNA that encodes functional TRA protein. In males, the splicing factor U2AF binds to the male-specific 3′ splice site and initiates spliceosome assembly (Fig. 2a)[37]. However in females, this splice site is bound by the female-specific splicing repressor SXL, thus blocking the binding of U2AF. Instead, U2AF binds to the female-specific 3′ splice site, and functional *tra* mRNA is produced (Fig. 2a).

SXL also regulates the alternative splicing of its own pre-mRNA, albeit by an entirely different mechanism[40]. *sxl* exon 3 is excluded

from the mRNA only in females. In males, inclusion of exon 3 introduces a premature translational stop codon. Exon 3 inclusion requires the protein SPF45, a second-step splicing factor that binds to the AG dinucleotide of the male-specific 3′ splice site (Fig. 2b). In females, SXL binds to a site adjacent to SPF45, and the two proteins interact. This interaction interferes with the activity of SPF45, and thus blocks the second step of the splicing reaction. As a consequence, exon 3 is skipped and exon 2 is spliced to exon 4, thus producing an mRNA encoding functional SXL protein. Thus, SXL blocks the first step of the splicing reaction in *tra* pre-mRNA and the second step in *sxl* pre-mRNA. *sxl* autoregulation is the only known example in which alternative pre-mRNA splicing is regulated at the second step of the splicing reaction.

While the previous examples of regulated alternative splicing involve splice-site repression, the female-specific splicing of *Drosophila doublesex* (*dsx*) pre-mRNA is the best characterized example of splice-site activation[37]. The 3′ splice site immediately upstream from exon 4 of *dsx* pre-mRNA is not recognized by the splicing machinery in males, thus leading to the exclusion of this exon (Fig. 2c). The male-specific *dsx* mRNA encodes a transcriptional repressor of female-specific genes. In females, the regulatory protein TRA promotes the cooperative binding of an SR protein, RBP1, and an SR-like protein, Transformer 2 (TRA2), to individual ESEs within exon 4 (refs 41, 42). This heterotrimeric protein complex recruits the splicing machinery to the upstream 3′ splice site, leading to the inclusion of exon 4. The female-specific *dsx* mRNA encodes a transcriptional repressor of male-specific genes.

The basic mechanisms of alternative splicing established in *Drosophila* have been shown to function in mammals. Alternative splice-site selection in mammals is also controlled by differential binding of regulatory proteins to splice sites, or to enhancer or silencer sequences within the pre-mRNA[28]. The organization of these sequences and the interplay of different regulatory proteins determine the outcome of the splicing reaction. As in *Drosophila*, regulatory proteins can exert competing influences on the splicing of a pre-mRNA.

For example, the binding of hnRNP A1 to an ESS in an exon of HIV *tat* pre-mRNA represses the inclusion of this exon in the mRNA. The binding of hnRNP A1 prevents the interaction of the SR protein SC35 with a nearby ESE[31]. Although the ESS and the ESE do not directly overlap, hnRNP A1 bound to the ESS seems to promote cooperative binding of additional hnRNP A1 proteins to adjacent exon sequences, thus spreading into the ESE. However, another SR protein, SF2/ASF, has a higher affinity for the ESE than SC35 and is therefore able to displace hnRNP A1.

One of the best characterized mammalian repressors of exon recognition is hnRNP I, also called a polypyrimidine tract-binding protein (PTB)[22,27,28,43]. This protein can repress exon inclusion by directly interfering with binding of general splicing factors to the pyrimidine tract. However, in most cases, the hnRNP I/PTB-binding sites flank the excluded exon. Two mechanisms for repression by hnRNP I/PTB have been proposed to account for this observation[32]. In the first mechanism, hnRNP I/PTB proteins bound on both sides of the exon interact with each other in such a way that the intervening exon 'loops out' and is isolated from the splicing machinery. In the second mechanism, hnRNP I/PTB cooperatively spreads across the exon, thus creating a 'zone of silencing'[22,32].

Regulation of tissue-specific alternative splicing

In addition to the ubiquitous RNA-binding proteins implicated generally in alternative splicing, a number of specific regulatory proteins have been identified, and some of them are expressed only

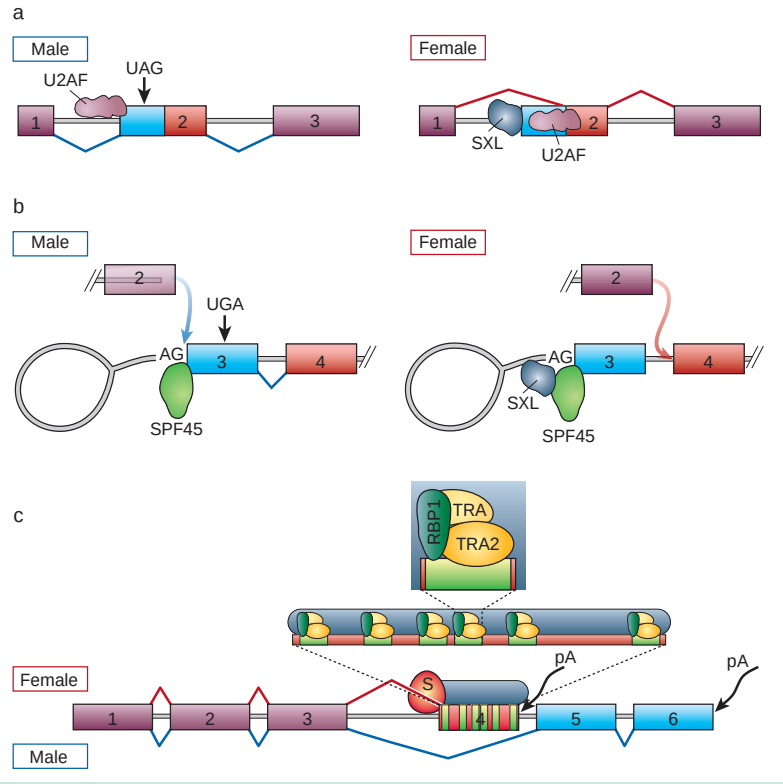

Figure 2 Regulation of alternative pre-mRNA splicing in the *Drosophila* sex-determination pathway. **a**, Alternative selection of 3′ splice sites preceding exon 2 of *tra* pre-mRNA is regulated by the SXL protein. In males, the splicing factor U2AF binds to the proximal 3′ splice site, leading to an mRNA containing a premature translational stop codon (UAG). In females, SXL binds to the proximal 3′ splice site, thus preventing the binding of U2AF. Instead, U2AF binds to the distal 3′ splice site, leading to an mRNA that encodes functional TRA protein. In all panels, the exons are indicated by coloured rectangles, while introns are shown as pale grey lines. **b**, Alternative inclusion of exon 3 of *sxl* pre-mRNA is regulated by SXL protein. In both males and females, the first step of the splicing reaction results in lariat formation at the branchpoint sequence upstream from the 3′ splice site preceding exon 3. Subsequently, the second-step splicing factor SPF45 binds to the AG dinucleotide of this splice site. In males, SPF45 promotes the second step of the splicing reaction, leading to the inclusion of exon 3. In females, SXL binds to a sequence upstream of the AG dinucleotide, interacts with SPF45 and inhibits its activity. This prevents the second step of the splicing reaction, leading to the exclusion of exon 3 and splicing of exon 2 to exon 4. Seven constitutively spliced exons are not shown. **c**, Alternative splicing of *dsx* pre-mRNA is regulated by the assembly of heterotrimeric protein complexes on female-specific ESEs. The first three exons are constitutively spliced in both sexes. In males, the 3′ splice site preceding exon 4 is not recognized by the splicing machinery, resulting in the exclusion of this exon, and splicing of exon 3 to exon 5. In females, the female-specific TRA protein promotes the binding of the SR protein RBP1, and the SR-like protein TRA2 to six copies of an ESE (indicated by green rectangles). These splicing enhancer complexes then recruit the splicing machinery to the 3′ splice site preceding exon 4, leading to its inclusion in the mRNA. In females, polyadenylation (pA) occurs downstream of exon 4, whereas in males it occurs downstream of exon 6. 'S' designates the splicing machinery.

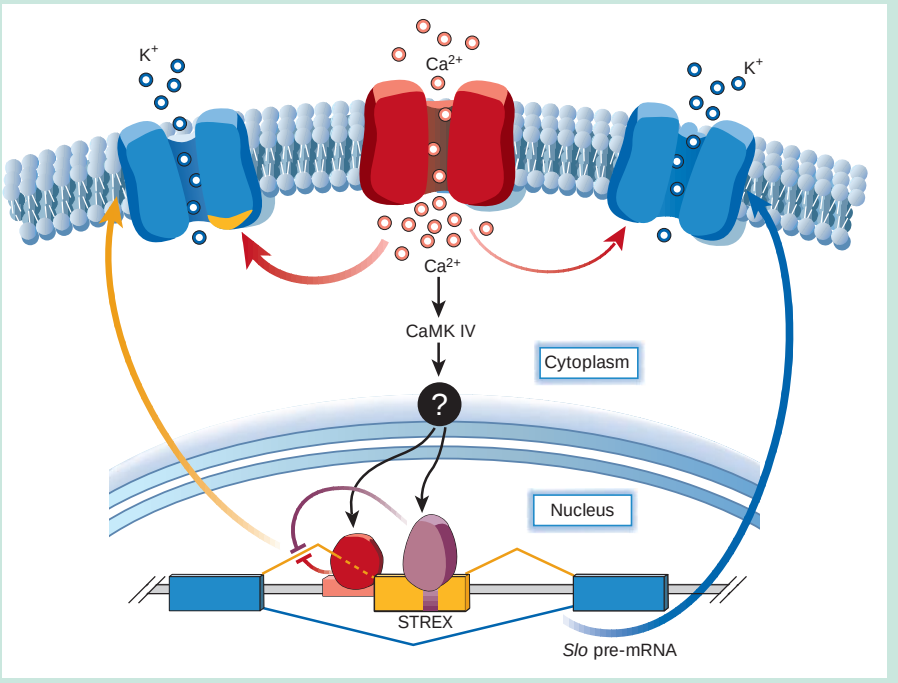

Figure 3 Inducible alternative splicing of rat *Slo* pre-mRNA. The fraction of *Slo* mRNAs containing an exon called STREX (stress axis-regulated exon, shown in yellow) is regulated by neuronal activity. Depolarization of the plasma membrane increases the intracellular Ca^{2+} concentration, leading to the activation of CaMK IV. This, in turn, is thought to result in the binding of repressor proteins to silencer sequences, one located upstream and the other within STREX. These silencers are shown as red rectangles with hypothetical repressors bound to them. In the presence of the repressors, the STREX exon is excluded from *Slo* mRNA. The alternatively spliced mRNAs encode two protein isoforms (shown in blue). The channel encoded by mRNA containing the STREX exon has an additional (yellow) domain, and is more sensitive to intracellular Ca^{2+} concentration. For clarity, only the portion of *Slo* pre-mRNA containing the STREX exon is shown.

in certain tissues. These proteins regulate alternative splicing of specific sets of pre-mRNAs.

Drosophila ELAV (for embryonic-lethal, abnormal visual system) is a pan-neuronal pre-mRNA-binding protein that regulates alternative splicing of at least three different pre-mRNAs[44]. This protein contains three RRMs, which are required for binding to uridine-rich regulatory sequences in the target pre-mRNAs. For example, ELAV binding to *neuroglian* pre-mRNA results in alternative inclusion of a neural-specific terminal exon. As yet there is no evidence that the mammalian homologues of *Drosophila* ELAV are involved in the regulation of pre-mRNA splicing[45,46].

The *Drosophila* Half pint (HFP) protein regulates alternative splicing of a subset of pre-mRNAs in the ovary[47]. One of these encodes the ovarian tumour protein, which is required for oogenesis, while another encodes a translation initiation factor. HFP is also required for constitutive splicing of the *gurken* pre-mRNA in oocytes. The mammalian orthologues of HFP are the human PUF60 protein[48] and the rat Siah-binding protein[49]. These proteins interact with the splicing factor U2AF[65], which binds to 3′ splice sites in all pre-mRNAs (Fig. 1). Although HFP targets specific pre-mRNAs, human PUF60 stimulates splicing nonspecifically in mammalian cell extracts, suggesting that it is a general splicing factor[48].

The *Drosophila* P-element transposon is active only in the germline because the splicing of its transposase pre-mRNA is repressed in somatic cells. The P-element somatic inhibitor protein (PSI) binds to the transposase pre-mRNA and inhibits removal of a specific intron in somatic cells, leading to the production of a truncated protein that represses transposition[50]. By contrast, PSI is not expressed in germ cells, and functional transposase is produced. PSI binds to 'pseudo' 5′ splice sites in transposase pre-mRNA and recruits U1 snRNP to these sites through interactions with the U1 snRNP 70K protein[50]. This leads to the formation of an abortive 5′ splice-site complex that interferes with splicing at the normal 5′ splice site. The mammalian homologue of PSI, called KSRP (for KH-type splicing regulatory protein), regulates neural-specific splicing of the human *src* pre-mRNA[51].

The mammalian splicing factor NOVA-1 (for neuro-oncological ventral antigen-1) regulates alternative pre-mRNA splicing in neurons. NOVA-1 binds to specific intronic sequences of target pre-mRNAs, and stimulates the inclusion of specific exons in the corresponding mRNAs[45]. For example, inclusion of alternatively spliced exons in the pre-mRNAs encoding subunits of the glycine and GABA (γ-aminobutyric acid) receptors requires NOVA-1.

At present, there are only a few examples of proteins that regulate tissue-specific alternative splicing, but it is likely that many more regulatory proteins, target pre-mRNAs and regulatory mechanisms will be discovered. Although many tissue-specific regulatory proteins are conserved between flies and mammals, their target pre-mRNAs can be different. This suggests a role for tissue-specific alternative splicing in the generation of species-specific traits and functions.

Inducible alternative splicing

The proteome of a cell can rapidly change in response to extracellular stimuli through complex signal-transduction pathways. Changes in protein composition can be regulated at many different levels, but have been studied primarily at the level of transcription and post-translational protein modification. Recently, several cases of inducible alternative pre-mRNA splicing have been reported.

An example from the mammalian immune system is the alternative splicing of a pre-mRNA encoding distinct isoforms of CD45, a transmembrane protein tyrosine phosphatase. *CD45* pre-mRNA is alternatively spliced in response to T-cell activation, leading to the production of proteins with distinct extracellular domains. Recent studies revealed a role for protein kinase C and Ras in the signalling pathway required for the switch in *CD45* alternative splicing[52]. Although the details of the signalling pathway are not fully understood, several lines of evidence point to SR proteins as the downstream effectors in the pathway. First, the relative levels of several SR proteins change upon T-cell activation[53]. Second, overexpression of different SR proteins in cultured cells leads to distinct patterns of alternative splicing of *CD45* pre-mRNA[53]. Finally, a T-cell-specific conditional knockout of the gene encoding the SR protein SC35 interferes with the normal pattern of *CD45* alternative splicing upon T-cell activation[54]. Based on these observations, it has been proposed that T-cell activation leads to the production of a new combination of SR proteins that bind to *CD45* pre-mRNA and change the pattern of its alternative splicing[52–54].

Similar mechanisms may control alternative splicing in the brain in response to neuronal activity[43]. For example, alternative splicing of

the rat SR-like protein Tra2-β (an orthologue of *Drosophila* TRA2) changes in response to small molecule-induced neural activity[55]. These changes in the types and levels of Tra2-β isoforms could, in turn, control the alternative splicing of other brain pre-mRNAs. A potential target of human Tra2-β is the *SMN2* pre-mRNA, because its splicing pattern can be altered by overexpression of this SR-like protein[56].

The *ania-6* pre-mRNA, which encodes a member of a new family of cyclins, provides another example of neuronal activity-dependent alternative pre-mRNA splicing. The synthesis of this protein is inducible in the adult brain of rats by cocaine and dopamine agonists, and two distinct *ania-6* mRNAs are generated by alternative splicing in response to different neurotransmitters and drugs[57]. One mRNA encodes a protein that associates with RNA polymerase II and co-localizes with splicing factors in nuclear speckles, while the other mRNA encodes a protein that localizes to the cytoplasm[57]. Thus, two functionally distinct proteins are produced from the same pre-mRNA in response to neural activity.

The only case of inducible alternative splicing in which the regulatory sequences within pre-mRNA have been identified is the neural activity-dependent alternative splicing of the rat *Slo* pre-mRNA[58]. This pre-mRNA is alternatively spliced, leading to the production of multiple protein isoforms of a calcium-dependent potassium channel[3,59]. These isoforms display distinct electrophysiological properties. A cell-culture model system was used to study the signalling pathway that induces *Slo* alternative splicing, and to identify *cis*-acting regulatory sequences required for the inclusion of a particular exon, stress axis-regulated exon (STREX), in *Slo* mRNA[58]. Depolarization of pituitary cells represses inclusion of the STREX exon in a process that requires the activity of a Ca^{2+}/calmodulin-dependent protein kinase (CaMK IV). This signalling pathway acts through an ESS in the STREX exon and the 3′ splice site upstream of STREX. Thus, it seems that depolarization of the membrane leads to the activation of CaMK IV, which in turn activates unidentified repressor proteins that bind the STREX exon and the 3′ splice site, thereby preventing STREX exon inclusion in *Slo* mRNA[58] (Fig. 3). Inclusion of the STREX exon increases the calcium sensitivity of the channel, which can modulate the electrical properties of the cell[58].

Alternative *trans*-splicing

In the examples of alternative splicing discussed above, exons located within an individual pre-mRNA are differentially joined to generate mature mRNAs (alternative *cis*-splicing). *Trans*-splicing can join exons located within separate pre-mRNAs, whereas alternative *trans*-splicing differentially uses exons within separate pre-mRNAs to produce distinct mRNAs. *Trans*-splicing was first discovered in trypanosomes and later found in organisms as complex as chordates[60]. However, this type of *trans*-splicing, called spliced leader (SL)-addition *trans*-splicing, involves specialized SL RNAs that provide 5′-terminal non-coding exons for all mRNAs[61]. SL-addition *trans*-splicing is not a source of protein diversity, as it is a constitutive process that does not lead to the production of alternatively spliced mRNAs. So far, there is no evidence for SL-addition *trans*-splicing in vertebrates.

The potential for mammalian *trans*-splicing was first demonstrated *in vitro*[14], and then shown to occur *in vivo* between separate viral pre-mRNAs, and between viral and cellular RNAs in infected cells[62,63]. Alternative *trans*-splicing has also been shown to result in exon duplications in several mammalian cellular RNAs[64–67]. However, mRNAs containing duplicated exons have not been shown to encode functional proteins, and in one case, exon duplication is not conserved in closely related organisms[68].

Although several examples of mammalian interchromosomal *trans*-splicing have been reported, none of these studies definitively proves the existence of this phenomenon[69–72]. There are also cases of intergenic *trans*-splicing between pre-mRNAs encoded by closely linked genes in mammals[73–76]. For example, several hybrid mRNAs are produced from a cluster of human cytochrome *P450 3A* genes[76] (Fig. 4a). Three of the *P450* genes (here labelled as 2, 3 and 4) are transcribed from one DNA strand, whereas one (1) is transcribed from the opposite DNA strand. Hybrid mRNAs containing exon 1 of gene 1 joined to various exons of genes 2 and 4 were detected (Fig. 4a). Because gene 1 is transcribed from one DNA strand and genes 2 and 4 from the other, the hybrid mRNAs must be generated by *trans*-splicing. However, the hybrid intergenic mRNAs are produced at levels that are orders of magnitude lower than those of intragenic mRNAs, and the existence of the endogenous proteins encoded by these hybrid mRNAs has not been demonstrated[76].

It is important to note that in every example of mammalian *trans*-splicing so far reported the pre-mRNAs also engage in *cis*-splicing. Therefore, it is possible that the observed *trans*-splicing represents 'splicing noise' resulting from low levels of splice-site pairing between separate pre-mRNAs. Thus, the functional significance of mammalian alternative *trans*-splicing remains to be established.

The only known example in which alternative *trans*-splicing is required for the production of an essential protein occurs in the *Drosophila modifier of mdg4 (mod(mdg4))* locus[77,78]. MOD(MDG4) protein isoforms, which function in the establishment or maintenance of chromatin structure, are encoded by an unusually complex genetic locus (Fig. 4b). Twenty-six alternatively spliced mRNAs encoded by this locus have been identified, and they all contain four common exons located at the 5′ end of the locus[78] (shown in red in Fig. 4b). Distinct mRNAs are generated by alternative splicing of the fourth common exon to individual downstream 'variable' exons (shown in blue or yellow in Fig. 4b). Strikingly, a number of variable exons are transcribed from the opposite DNA strand (blue exons in Fig. 4b), strongly suggesting that they are joined to the fourth common exon by *trans*-splicing.

A central question in the regulation of alternative *trans*-splicing is how splice sites located on separate pre-mRNAs are recognized by the splicing machinery and correctly joined. This is not an entirely new question, as splice sites at the ends of large introns are essentially configured *in trans*. In this case, the correct joining of splice sites is probably achieved by coupling transcription and splicing through interactions between the splicing machinery and the CTD of RNA polymerase II (ref. 21). In this mechanism, the splicing factors connect the 5′ splice site and the CTD as the long intron is being transcribed. The proximity of the eventually synthesized 3′ splice site to the CTD would ensure an interaction between splicing complexes assembled on the 5′ and 3′ splice sites.

This type of coupling cannot occur in *trans*-splicing because different RNA polymerase complexes transcribe separate *trans*-splicing precursors. Therefore, there must be another mechanism for bringing together the 5′ and 3′ splices sites. This might be achieved by the localized transcription of pre-mRNAs in the nucleus — by restricting transcription to 'gene expression factories'[21] only those pre-mRNAs transcribed in the same factory would engage in *trans*-splicing. This could explain the close linkage of *trans*-spliced exons in the *mod(mdg4)* locus, and the previously mentioned cases of intergenic *trans*-splicing. Nuclear compartmentalization could also prevent inappropriate splicing to pre-mRNAs encoded by other genes. Finally, it is possible that *trans*-splicing precursors interact through specific base pairing, or through interactions between proteins bound to each of the precursors[77,78].

Comparative genomics and alternative pre-mRNA splicing

Alternative pre-mRNA splicing is an important source of protein diversity that may have contributed to the increase in the phenotypic complexity of metazoans during evolution[2,3]. This proposal was based, in part, on the unexpectedly small difference in gene number in different organisms from yeast to humans. For example, flies have only about twice as many genes as yeast[79]. In addition, worms have more genes than flies (19,000 and 14,000, respectively), but flies are clearly more complex in their development, morphology and behaviour. Remarkably, humans have only about 35,000 genes[80,81].

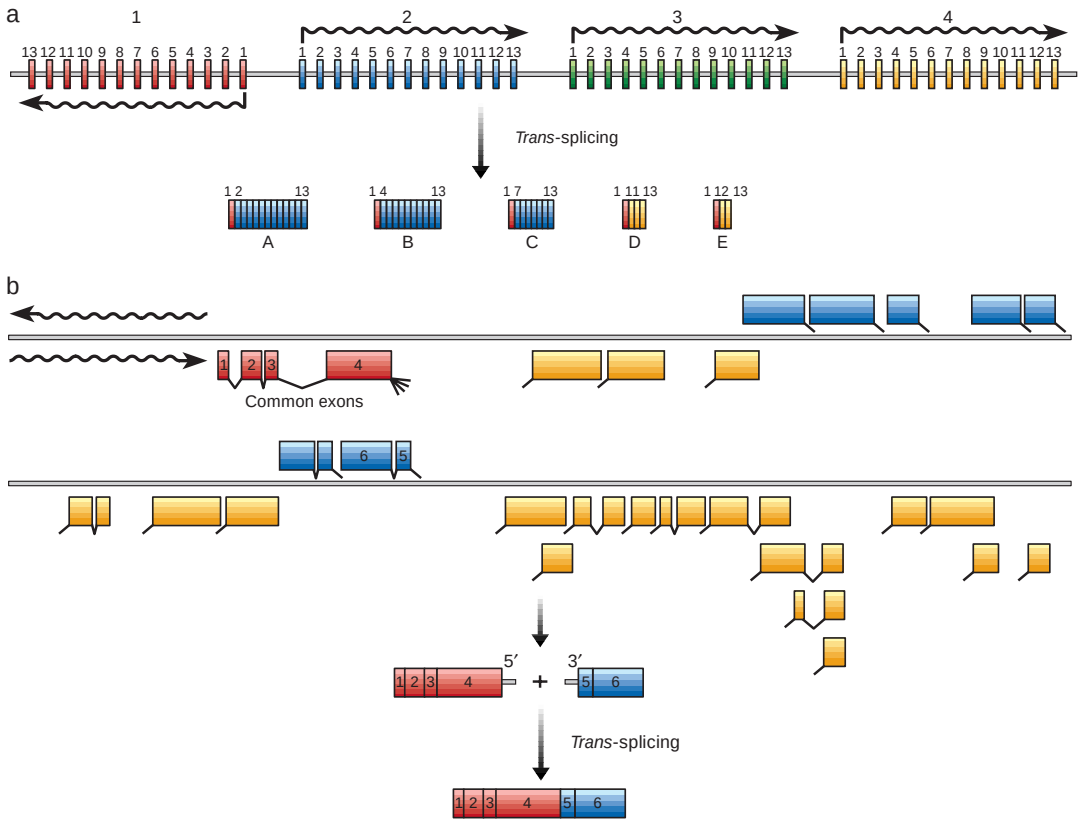

Figure 4 Alternative *trans*-splicing in mammals and flies. **a**, Four cytochrome *P450 3A* genes are arranged in a cluster that spans ~200 kilobases of genomic DNA. For simplicity, the genes have been designated 1–4 (they were designated *CYP3A43*, *CYP3A4*, *CYP3A7*, *CYP3A5*, respectively, in the original reference)[76]. Gene 1 is transcribed from one DNA strand, whereas genes 2–4 are transcribed from the opposite strand. The direction of transcription on each strand is indicated by the wavy arrows. The small coloured boxes are exons, and the pale grey lines are introns. Hybrid intergenic mRNAs are produced by *trans*-splicing between exon 1 of gene 1 and various exons in genes 2 and 4 to generate the mRNAs labelled A–E. **b**, The *Drosophila mod(mdg4)* locus spans 28 kilobases, and encodes 26 alternatively spliced mRNAs. Each mRNA is generated by splicing of four common exons (1–4, shown in red) to one of the exons located downstream (shown in yellow or blue). The yellow exons are transcribed from the same DNA strand as the common exons, while the blue exons are transcribed from the opposite DNA strand. The 5′ splice site downstream from common exon 4 is indicated by the four slanted lines. Similarly, the 3′ splice site upstream of each alternatively spliced exon is indicated by a single slanted line. An example of *trans*-splicing within this locus is shown at the bottom of the figure. Two *trans*-splicing precursors, containing exons 1–4 or 5 and 6, are transcribed from opposite DNA strands. The exons 4 and 5 are then joined by *trans*-splicing.

Although this number is still being debated[82], even the highest estimate for humans is only around threefold greater than that for worms. It is unlikely that this difference alone could explain the obvious differences in functional and behavioural complexity between invertebrates and vertebrates. The increased organismic complexity of vertebrates could be a consequence of the elaboration of mechanisms that increase proteome size and the evolution of more extensive networks of gene regulation[83].

A measure of the relative contribution of alternative splicing to proteome size could be obtained by dividing the total number of distinct full-length complementary DNAs (cDNAs) resulting from alternative splicing by the number of genes of an organism. Unfortunately, relatively few full-length cDNA sequences are available[84–86], and the annotation of several genomes, including the human genome, is still imprecise. In the absence of this information, partial cDNA sequences and expressed sequence tags (ESTs) have been used to detect alternative splicing events[87,88]. Efforts are also being made to use DNA microarrays to identify alternatively spliced mRNAs on a large scale[89–91]. These approaches have provided preliminary estimates of the percentage of genes with alternatively spliced forms ($\%G_{ASF}$) and the average number of alternatively spliced forms per gene (ASF/G). The product of these two numbers provides a quantitative estimate of the contribution of alternative splicing to proteome size.

Estimates of $\%G_{ASF}$ in humans range from 35 to 59% (refs 87, 88). However, it is generally agreed that this is an underestimate, because $\%G_{ASF}$ depends on the number of ESTs examined[92]. It is important to note that most of the detected alternative splicing events (about 80%) lead to changes in the amino acid sequence of the encoded proteins[88]. In a recent report, random sets of 650 non-redundant cDNAs were compared with 100,000 ESTs in different organisms ranging from worms and flies to humans[92]. The surprising outcome was that the estimate for $\%G_{ASF}$ is approximately the same in all organisms tested[92].

Similar values have also been obtained for ASF/G in different organisms; however, the estimates were adversely affected if the analysis included 'outlier' genes that generate extraordinarily large numbers of alternatively spliced forms (ref. 92, and J. Valcarcel and P. Bork, personal communication). For example, the *Drosophila* axon guidance receptor gene, *Dscam*, contains 95 alternatively spliced exons organized into four clusters, and it has the potential to generate over 38,000 different protein isoforms — nearly three times more proteins than the number of genes in *Drosophila*[93]. Remarkable

but less extensive examples in mammals are the previously mentioned *Slo* gene, and three *neurexin* genes, which encode proteins that may act as synaptic-cell-surface receptors or cell-adhesion molecules[94]. The *Slo* gene and the three *neurexin* genes have the potential to generate over 500 and 2,000 alternatively spliced forms, respectively[3,16].

The conclusion of this analysis is that the relative contribution of alternative splicing to the size of the proteome does not seem to be different in evolutionarily divergent metazoans if 'outlier' genes are not included in the analysis[92]. However, it is possible that the differences will be observed once the 'outlier' genes are considered. To account for their contribution to the overall estimate of ASF/G, the organization of all genes and the sequence of all full-length cDNAs are required. Thus, the relationship between alternative splicing and proteome size of different metazoans remains to be established.

Perspectives

At least two forces affect the evolution of complex organisms. On the one hand there is an extraordinary pressure to conserve the basic components of cellular machines among organisms that differ widely in complexity[80,81]. For example, the cellular machines required for different steps of gene expression, such as transcription[95], splicing[11] and mRNA export[96], are so highly conserved between yeast and humans that many components are interchangeable.

On the other hand, both the proteome size and the complexity of regulatory networks increase as more complex organisms evolve. Proteome size is determined by the number of genes in an organism, and by mechanisms that expand the coding capacity of the genome, including alternative pre-mRNA splicing. However, as discussed above, the relative contribution of alternative splicing to proteome expansion is not well understood. The organization of regulatory networks and their effect on organismic complexity are even more difficult to determine. Moreover, it is possible that a relatively small increase in the size of the proteome or a change in a key component of the regulatory networks could lead to dramatic changes in organismic complexity. If this were the case, the proteome size would not accurately correlate with organismic complexity. Rather, insights into the origins of organismic complexity would require an understanding of regulatory networks[83]. □

1. Quelle, D. E., Zindy, F., Ashmun, R. A. & Sherr, C. J. Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell* **83,** 993–1000 (1995).
2. Black, D. L. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **103,** 367–370 (2000).
3. Graveley, B. R. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17,** 100–107 (2001).
4. Goldstrohm, A. C., Greenleaf, A. L. & Garcia-Blanco, M. A. Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. *Gene* **277,** 31–47 (2001).
5. Caceres, J. F. & Kornblihtt, A. R. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.* **18,** 186–193 (2002).
6. Gautheret, D., Poirot, O., Lopez, F., Audic, S. & Claverie, J. M. Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res.* **8,** 524–530 (1998).
7. Keegan, L. P., Gallo, A. & O'Connell, M. A. The many roles of an RNA editor. *Nature Rev. Genet.* **2,** 869–878 (2001).
8. Banks, R. E. *et al.* Proteomics: new perspectives, new biomedical opportunities. *Lancet* **356,** 1749–1756 (2000).
9. Reed, R. Mechanisms of fidelity in pre-mRNA splicing. *Curr. Opin. Cell Biol.* **12,** 340–345 (2000).
10. Stevens, S. W. *et al.* Composition and functional characterization of the yeast spliceosomal penta-snRNP. *Mol. Cell* **9,** 31–44 (2002).
11. Zhou, Z., Licklider, L., Gygi, S. & Reed, R. The proteome of functional human spliceosomes. *Nature* (submitted).
12. Staley, J. P. & Guthrie, C. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92,** 315–326 (1998).
13. Berget, S. M. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270,** 2411–2414 (1995).
14. Reed, R. Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr. Opin. Genet. Dev.* **6,** 215–220 (1996).
15. Deutsch, M. & Long, M. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27,** 3219–3228 (1999).
16. Rowen, L. *et al.* Analysis of the human neurexin genes: alternative splicing and the generation of protein diversity. *Genomics* **79,** 587–597 (2002).
17. Graveley, B. R. Sorting out the complexity of SR protein functions. *RNA* **6,** 1197–1211 (2000).
18. Schaal, T. D. & Maniatis, T. Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Mol. Cell. Biol.* **19,** 261–273 (1999).
19. Mayeda, A., Screaton, G. R., Chandler, S. D., Fu, X. D. & Krainer, A. R. Substrate specificities of SR proteins in constitutive splicing are determined by their RNA recognition motifs and composite pre-mRNA exonic elements. *Mol. Cell. Biol.* **19,** 1853–1863 (1999).
20. Sun, H. & Chasin, L. A. Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.* **20,** 6414–6425 (2000).
21. Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416,** 499–506 (2002).
22. 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).
23. Krawczak, M., Reiss, J. & Cooper, D. N. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.* **90,** 41–54 (1992).
24. Valentine, C. R. The association of nonsense codons with exon skipping. *Mutat. Res.* **411,** 87–117 (1998).
25. 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).
26. Blencowe, B. J. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* **25,** 106–110 (2000).
27. Hastings, M. L. & Krainer, A. R. Pre-mRNA splicing in the new millennium. *Curr. Opin. Cell Biol.* **13,** 302–309 (2001).
28. Smith, C. W. & Valcarcel, J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.* **25,** 381–388 (2000).
29. Perez-Canadillas, J. M. & Varani, G. Recent advances in RNA-protein recognition. *Curr. Opin. Struct. Biol.* **11,** 53–58 (2001).
30. Graveley, B. R., Hertel, K. J. & Maniatis, T. SR proteins are 'locators' of the RNA splicing machinery. *Curr. Biol.* **9,** R6–R7 (1999).
31. Zhu, J., Mayeda, A. & Krainer, A. R. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol. Cell* **8,** 1351–1361 (2001).
32. Wagner, E. J. & Garcia-Blanco, M. A. Polypyrimidine tract binding protein antagonizes exon definition. *Mol. Cell. Biol.* **21,** 3281–3288 (2001).
33. Dreyfuss, G., Matunis, M. J., Pinol-Roma, S. & Burd, C. G. hnRNP proteins and the biogenesis of mRNA. *Annu. Rev. Biochem.* **62,** 289–321 (1993).
34. Krecic, A. M. & Swanson, M. S. hnRNP complexes: composition, structure, and function. *Curr. Opin. Cell Biol.* **11,** 363–371 (1999).
35. Cowper, A. E., Caceres, J. F., Mayeda, A. & Screaton, G. R. Serine-arginine (SR) protein-like factors that antagonize authentic SR proteins and regulate alternative splicing. *J. Biol. Chem.* **276,** 48908–48914 (2001).
36. 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).
37. Schutt, C. & Nothiger, R. Structure, function and evolution of sex-determining systems in Dipteran insects. *Development* **127,** 667–677 (2000).
38. Baker, B. S. Sex in flies: the splice of life. *Nature* **340,** 521–524 (1989).
39. Cline, T. W. & Meyer, B. J. Vive la différence: males vs females in flies vs worms. *Annu. Rev. Genet.* **30,** 637–702 (1996).
40. Lallena, M. J., Chalmers, K. J., Llamazares, S., Lamond, A. I. & Valcarcel, J. Splicing regulation at the second catalytic step by Sex-lethal involves 3′ splice site recognition by SPF45. *Cell* **109,** 285–296 (2002).
41. Lynch, K. W. & Maniatis, T. Assembly of specific SR protein complexes on distinct regulatory elements of the *Drosophila doublesex* splicing enhancer. *Genes Dev.* **10,** 2089–2101 (1996).
42. Hertel, K. J. & Maniatis, T. The function of multisite splicing enhancers. *Mol. Cell* **1,** 449–455 (1998).
43. Grabowski, P. J. & Black, D. L. Alternative RNA splicing in the nervous system. *Prog. Neurobiol.* **65,** 289–308 (2001).
44. Lisbin, M. J., Qiu, J. & White, K. The neuron-specific RNA-binding protein ELAV regulates neuroglian alternative splicing in neurons and binds directly to its pre-mRNA. *Genes Dev.* **15,** 2546–2561 (2001).
45. Dredge, B. K., Polydorides, A. D. & Darnell, R. B. The splice of life: alternative splicing and neurological disease. *Nature Rev. Neurosci.* **2,** 43–50 (2001).
46. Toba, G., Qui, J., Koushika, S. P. & White, K. Ectopic expression of *Drosophila* ELAV and human HuD in *Drosophila* wing disc cells reveals functional distinctions and similarities. *J. Cell Sci.* **115,** 2413–2421 (2002).
47. Van Buskirk, C. & Schupbach, T. *half pint* regulates alternative splice site selection in *Drosophila*. *Dev. Cell* **2,** 343–353 (2002).
48. Page-McCaw, P. S., Amonlirdviman, K. & Sharp, P. A. PUF60: a novel U2AF65-related splicing activity. *RNA* **5,** 1548–1560 (1999).
49. Poleev, A., Hartmann, A. & Stamm, S. A trans-acting factor, isolated by the three-hybrid system, that influences alternative splicing of the amyloid precursor protein minigene. *Eur. J. Biochem.* **267,** 4002–4010 (2000).
50. Labourier, E., Adams, M. D. & Rio, D. C. Modulation of P-element pre-mRNA splicing by a direct interaction between PSI and U1 snRNP 70K protein. *Mol. Cell* **8,** 363–373 (2001).
51. Min, H., Turck, C. W., Nikolic, J. M. & Black, D. L. A new regulatory protein, KSRP, mediates exon inclusion through an intronic splicing enhancer. *Genes Dev.* **11,** 1023–1036 (1997).
52. Lynch, K. W. & Weiss, A. A model system for activation-induced alternative splicing of CD45 pre-mRNA in T cells implicates protein kinase C and Ras. *Mol. Cell. Biol.* **20,** 70–80 (2000).
53. ten Dam, G. B. *et al.* Regulation of alternative splicing of CD45 by antagonistic effects of SR protein splicing factors. *J. Immunol.* **164,** 5287–5295 (2000).
54. Wang, H. Y., Xu, X., Ding, J. H., Bermingham, J. R. Jr & Fu, X. D. SC35 plays a role in T cell development and alternative splicing of CD45. *Mol. Cell* **7,** 331–342 (2001).
55. Daoud, R., Da Penha Berzaghi, M., Siedler, F., Hubener, M. & Stamm, S. Activity-dependent regulation of alternative splicing patterns in the rat brain. *Eur. J. Neurosci.* **11,** 788–802 (1999).
56. Hofmann, Y., Lorson, C. L., Stamm, S., Androphy, E. J. & Wirth, B. *Htra2-β1* stimulates an exonic splicing enhancer and can restore full-length *SMN* expression to *survival motor neuron 2* (*SMN2*). *Proc. Natl Acad. Sci. USA* **97,** 9618–9623 (2000).
57. Berke, J. D. *et al.* Dopamine and glutamate induce distinct striatal splice forms of Ania-6, an RNA polymerase II-associated cyclin. *Neuron* **32,** 277–287 (2001).
58. Xie, J. & Black, D. L. A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels. *Nature* **410,** 936–939 (2001).

59. Black, D. L. Splicing in the inner ear: a familiar tune, but what are the instruments? *Neuron* **20,** 165–168 (1998).
60. Nilsen, T. W. Evolutionary origin of SL-addition trans-splicing: still an enigma. *Trends Genet.* **17,** 678–680 (2001).
61. Blumenthal, T. Trans-splicing and polycistronic transcription in *Caenorhabditis elegans*. *Trends Genet.* **11,** 132–136 (1995).
62. Eul, J., Graessmann, M. & Graessmann, A. Experimental evidence for RNA trans-splicing in mammalian cells. *EMBO J.* **14,** 3226–3235 (1995).
63. Caudevilla, C. *et al.* Heterologous HIV-nef mRNA *trans*-splicing: a new principle how mammalian cells generate hybrid mRNA and protein molecules. *FEBS Lett.* **507,** 269–279 (2001).
64. Caudevilla, C. *et al.* Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. *Proc. Natl Acad. Sci. USA* **95,** 12185–12190 (1998).
65. Takahara, T., Kanazu, S. I., Yanagisawa, S. & Akanuma, H. Heterogeneous Sp1 mRNAs in human HepG2 cells include a product of homotypic *trans*-splicing. *J. Biol. Chem.* **275,** 38067–38072 (2000).
66. Frantz, S. A. *et al.* Exon repetition in mRNA. *Proc. Natl Acad. Sci. USA* **96,** 5400–5405 (1999).
67. Akopian, A. N. *et al.* Trans-splicing of a voltage-gated sodium channel is regulated by nerve growth factor. *FEBS Lett.* **445,** 177–182 (1999).
68. Caudevilla, C. *et al.* Localization of an exonic splicing enhancer responsible for mammalian natural *trans*-splicing. *Nucleic Acids Res.* **29,** 3108–3115 (2001).
69. Vellard, M. *et al.* C-myb proto-oncogene: evidence for intermolecular recombination of coding sequences. *Oncogene* **6,** 505–514 (1991).
70. Sullivan, P. M., Petrusz, P., Szpirer, C. & Joseph, D. R. Alternative processing of androgen-binding protein RNA transcripts in fetal rat liver. Identification of a transcript formed by trans splicing. *J. Biol. Chem.* **266,** 143–154 (1991).
71. Li, B. L. *et al.* Human acyl-CoA:cholesterol acyltransferase-1 (ACAT-1) gene organization and evidence that the 4.3-kilobase ACAT-1 mRNA is produced from two different chromosomes. *J. Biol. Chem.* **274,** 11060–11071 (1999).
72. Hirayama, T., Sugino, H. & Yagi, T. Somatic mutations of synaptic cadherin (CNR family) transcripts in the nervous system. *Genes Cells* **6,** 151–164 (2001).
73. Shimizu, A. & Honjo, T. Synthesis and regulation of trans-mRNA encoding the immunoglobulin epsilon heavy chain. *FASEB J.* **7,** 149–154 (1993).
74. Fujieda, S., Lin, Y. Q., Saxon, A. & Zhang, K. Multiple types of chimeric germ-line Ig heavy chain transcripts in human B cells: evidence for *trans*-splicing of human Ig RNA. *J. Immunol.* **157,** 3450–3459 (1996).
75. Chatterjee, T. K. & Fisher, R. A. Novel alternative splicing and nuclear localization of human RGS12 gene products. *J. Biol. Chem.* **275,** 29660–29671 (2000).
76. Finta, C. & Zaphiropoulos, P. G. Intergenic mRNA molecules resulting from *trans*-splicing. *J. Biol. Chem.* **277,** 5882–5890 (2002).
77. Labrador, M. *et al.* Protein encoding by both DNA strands. *Nature* **409,** 1000 (2001).
78. Dorn, R., Reuter, G. & Loewendorf, A. Transgene analysis proves mRNA trans-splicing at the complex *mod(mdg4)* locus in *Drosophila*. *Proc. Natl Acad. Sci. USA* **98,** 9724–9729 (2001).
79. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).
80. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
81. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
82. Daly, M. Estimating the human gene count. *Cell* **109,** 283–284 (2002).
83. Davidson, E. H. *Genomic Regulatory Systems* (Academic, New York, 2001).
84. Suzuki, Y., Yamashita, R., Nakai, K. & Sugano, S. DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.* **30,** 328–331 (2002).
85. Bono, H., Kasukawa, T., Furuno, M., Hayashizaki, Y. & Okazaki, Y. FANTOM DB: database of functional annotation of RIKEN mouse cDNA clones. *Nucleic Acids Res.* **30,** 116–118 (2002).
86. Kristiansen, T. Z. & Pandey, A. Resources for full-length cDNAs. *Trends Biochem. Sci.* **27,** 266–267 (2002).
87. Sorek, R. & Amitai, M. Piecing together the significance of splicing. *Nature Biotechnol.* **19,** 196 (2001).
88. Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nature Genet.* **30,** 13–19 (2002).
89. Hu, G. K. *et al.* Predicting splice variant from DNA chip expression data. *Genome Res.* **11,** 1237–1245 (2001).
90. Shoemaker, D. D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409,** 922–927 (2001).
91. Yeakly, J. M. *et al.* Profiling alternative splicing on fiber optic arrays. *Nature Biotechnol.* **20,** 1–6 (2002).
92. Brett, D., Pospisil, H., Valcarcel, J., Reich, J. & Bork, P. Alternative splicing and genome complexity. *Nature Genet.* **30,** 29–30 (2002).
93. Schmucker, D. *et al. Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101,** 671–684 (2000).
94. Missler, M. & Sudhof, T. C. Neurexins: three genes and 1001 products. *Trends Genet.* **14,** 20–26 (1998).
95. Ptashne, M. & Gann, A. *Genes & Signals* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2002).
96. Reed, R. & Hurt, E. A Conserved mRNA export machinery coupled to pre-mRNA splicing. *Cell* **108,** 523–531 (2002).

Acknowledgements

We thank P. Cramer, B. Graveley, A. Krainer, C. Nabholz and R. Reed for their comments on the manuscript, and R. Hellmiss for the illustrations.