Comparative analysis of splicing in eukaryotes

Mireya Plass Pórtulas

TESI DOCTORAL UPF / 2011

DIRECTOR DE LA TESI:

Dr. Eduardo Eyras

Departament de Ciències Experimentals i de la Salut

UNIVERSITAT POMPEU FABRA

a mi padre

Agradecimientos

A día de hoy creo que podría decir que cualquiera puede tener un doctorado. Y no es que quiera desmerecer a todos los doctores existentes. Los que quieran, de verdad, y estén dispuestos a invertir más cosas que su tiempo y esfuerzo, pueden hacerlo. O eso, o son genios, que también puede ser. Yo espero que si alguien está leyendo esto quiere decir que estoy a punto de ser doctora, pero en mi caso no soy una genio, así que si he llegado hasta aquí es gracias a toda la gente que me he encontrado en el camino y que de un modo u otro me han ayudado a conseguirlo. Estos años han sido probablemente los mejores y los peores de mi vida a la vez, así que gracias si formáis parte de los que contribuisteis a lo bueno.

Aunque suene a tópico, creo que gran parte del mérito se lo debo a mi jefe, Eduardo. Siempre es una persona a la que acudir cuando uno ya no sabe por qué camino ir (y eso que tengo una orientación terrible). Y en todos los casos, siempre sabe echar una mano, o animarte si lo necesitas. Gracias. André también ha estado siempre ahí, y con su mente crítica (especialmente conmigo), siempre ha sido (y es) un buen amigo para debatir, discutir e incluso reír (pero no hablar de fútbol). La verdad es que durante estos años me he encontrado mucha gente que ha hecho que mi doctorado, y las horas dentro y fuera del laboratorio, fueran mucho más fáciles y agradables. Con Eneritz hemos acabado compartiendo muchos ratos, y la verdad es que

siempre han estado bastante bien, así que eso cuenta. A Eneritz y Nico (y Amadís, aunque lleve menos tiempo), les agradezco los ratos compartidos y la ayuda que siempre me han dado (ya os digo yo que no han sido ni una ni dos preguntas). A Alice, por su voluntad de intentar hacerme entender que las fórmulas de los artículos significan algo (aparte de que con una fórmula parece que lo que has hecho sea muy complicado e interesante). A las chicas de mi lado, Macarena, Alice, Isabella, Núria, Inma, Eneritz, Sonja y también Steve, por los ratos en el laboratorio, hablando de cualquier cosa, y riendo un rato, que siempre va bien reírse y acordarnos que agobiarse demasiado no sirve para nada (alguien me dijo hace poco que si no conociéramos la palabra no nos agobiaríamos). Ha sido estupendo compartir con vosotras el despacho.

También tengo mucho que agradecer a mi familia: Joan, mi madre Isabel, mis hermanos Hermann (y Laura) y Guillermo, Zazil, mis tíos, mis primos... (no pongo más nombres que la lista es larga y si no tendrán que talar un árbol más); y a otros que sin ser de la familia los veo más que si lo fueran: Bea, Alice, Luca, Eric, Anna, Xavi, Blanca, Sergio, Elena, Olga. Las comidas, las cenas, las vacaciones, las ferias, los conciertos, las palomitas, la navidad, la playa, y los ratos en general han estado bien. Y siempre habéis estado por aquí (o a una llamada de distancia) cuando os he necesitado.

Y ya para acabar, sólo decir que esto del doctorado no lo empecé sola. Creo que fuimos unos cuantos los descerebrados que empezamos juntos (unos antes y unos más tarde, unos más cerca y otros más lejos), y que su compañía y amistad durante todos estos años ha sido quizás de las cosas que más recordaré. Hemos tenido comidas en la cafetería, el gallego y otros sitios. Cafés en el Laie, el bar del hospital y el parque. Cervezas en el bitácora y el Jaika y viendo el fútbol. Conquista a los pirineos, power paint y Euskadi world tour. Ya sabéis quienes sois, así que no hace falta decir más. Ahora cada cual parece que llevará un poco su camino por separado, así que espero que al menos recordéis este tiempo como lo recuerdo yo.

Abstract

Splicing is the mechanism by which introns are removed from the pre-mRNA to create a mature transcript. This process is performed by a macromolecular complex, the spliceosome, and involves the recognition of the splicing signals in the pre-mRNA. These signals are not always perfectly recognized, which allows the production of different mature transcripts from a single pre-mRNA through a process called alternative splicing. This process can be regulated by specific protein factors or by other mechanisms that affect the recognition of the splicing signals, such as the secondary structure adopted by the pre-mRNA. In this thesis we have investigated the mechanisms of splicing regulation in eukaryotes using computational approaches. Moreover, we have also studied the relationship that exists between protein factors involved in splicing regulation and splicing signals, and how they have co-evolved across species. Finally, and considering the possibilities that alternative splicing can offer from the evolutionary point of view, he have also analyzed the impact of alternative splicing in gene evolution.

Resum

L'*splicing* és el mecanisme pel qual els introns són eliminats del pre-mRNA per generar un trànscrit madur. Aquest procés és dut a terme per un complex macromolecular anomenat spliceosoma i requereix el reconeixement dels senyals d'*splicing* al pre-mRNA. Aquests senyals no són sempre identificats correctament, el que permet la producció de trànscrits diferents a partir d'un únic pre-mRNA mitjançant un procés anomenat *splicing* alternatiu. Aquest procés pot ser regulat mitjançant factors proteics específics o per altres mecanismes que alteren el reconeixement dels senyals d'*splicing* com l'estructura secundària adoptada pels pre-mRNAs. En aquesta tesi hem investigat els mecanismes de regulació de l'*splicing* en eucariotes mitjançant tècniques computacionals. També hem estudiat la relació existent entre les proteïnes que intervenen en la regulació de l'*splicing* i els senyals d'*splicing*, i com han coevolucionat en diferents espècies. Finalment, i tenint en compte les possibilitats que l'*splicing* alternatiu ofereix des del punt de vista evolutiu, també hem analitzat l'impacte de l'*splicing* alternatiu en l'evolució gènica.

Preface

When Francis Crick published the article "Split Genes and RNA splicing" in 1979 in which he summarized the revolution that splicing was producing in the molecular genetics field, probably he could not foresee the real impact that it would have in current molecular biology. Nowadays we know that splicing is responsible for changes in proteins in response to environmental conditions like heat-shock, the generation of the diversity of immunoglobulin genes and the production of soluble versions or membrane-anchored receptors. We also know that alterations in components of the splicing machinery or in the splicing signals can be responsible for genetic diseases or even cancer. Thus, splicing is somehow virtually related with all processes that exist in the cell.

Despite the amount of research performed during the last thirty years, there are still lots of aspects of the splicing mechanism to be understood. Nowadays, we may have a clear picture about the proteins that compose the spliceosome or how this complex machine roughly works. Nevertheless, we do not know the real importance of all those proteins nor of all splicing enhancers and silencers in the regulation of a particular splicing event. Yet, we aim to understand how all of them are regulated. In most of the cases, we cannot predict the splicing outcome even when we know all the players. And even if we knew, we would still need to answer lots of questions. Why all those factors are required?

Why do we have splicing? Why virtually everything depends on it? Why even though it is so important in humans, it is nearly dispensable in other species like yeast?

The advances in biology and technology have provided us with huge amount of data that can be used to try to answer these questions. Currently there are over 150 complete genomes sequenced, from bacteria to mammals. We can use the genomes of these species to learn, among other things, about splicing. For instance, we can identify splicing proteins in several species, compare them, and try to understand how these changes affect the function of these proteins. We can analyze the splicing signals of different species, and if they are different, try to understand why; or which are the implications of these differences in the regulation of this mechanism. Studying splicing with a computational approach can allow us to understand this process genome-wide, how it works, which are the players involved in the game. Furthermore, it can allow us to understand the differences across species, which can give us some insights about the origin and evolution of this mechanism and perhaps its impact on genome complexity.

Abbreviations

3'ss	3' splice site
3'UTR	3' untranslated region
5'ss	5' splice site
5'UTR	5' untranslated region
AS	alternative splicing
BS	branch site
CAGE	cap analysis of gene expression
cDNA	complementary DNA
dN / Ka	non-synonymous substitution rate
DNA	deoxyribonucleic acid
dS / Ks	synonymous substitutions rate
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
EST	expressed sequence tag
hnRNP	heterogeneous nuclear ribonucleoprotein
CLIP	Cross-linking immunoprecipitation
ISE	intronic splicing enhancer
ISS	intronic splicing silencer
mRNA	messenger RNA
ncRNA	non-coding RNA
nt	nucleotide
NMD	nonsense mediated decay
PCR	polymerase chain reaction
PESE	putative ESE
PESS	putative ESS
PPT	polypyrimidine tract
pre-mRNA	precursor messenger RNA
PTC	premature termination codon
RNA	ribonucleic acid
RNA-Seq	RNA sequencing
RRM	RNA recognition motif

RS domain	serine arginine rich domain
RT-PCR	reverse transcription PCR
SAGE	serial analysis of gene expression
SELEX	systematic evolution of ligands by exponential enrichment
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein
SR protein	serine arginine rich protein

Table of contents

I. INTRODUCTION

1. What is splicing?

Splicing is the mechanism by which introns are removed from precursor messenger RNAs (pre-mRNAs) to create mature transcripts. This mechanism was first observed in the late seventies independently by Phillip A Sharp and Richard J. Roberts when they detected, using electron microscopy, that several regions of the adenovirus 2 messenger RNA (mRNA) hybridized with DNA produced "branches" that suggested that the mRNA was not fully complementary to the DNA molecule (Berget et al., 1977; Chow et al., 1997). Since this initial observation, our knowledge on splicing has increased considerably. Splicing is no longer regarded as a mechanism solely devoted to intron removal, but rather as complex mechanism of regulation modulated by specific factors, transcription or even chromatin, that is responsible for not only the majority of protein diversity observed in higher eukaryotes but also post-transcriptional gene regulation (reviewed in Black 2003; Kornblihtt, 2007; Luco et al., 2011; Blencowe, 2006; Lareau et al., 2007a). This mechanism, in turn, has made modern biology modify its central dogma[1] (Crick, 1958; Crick, 1970)

[1] This idea was initially proposed by Francis Crick in 1956 in the letter "On protein synthesis". It was finally published in 1958 under the same name, and later revised in 1970. Epistemologically, the term dogma is used incorrectly and perhaps paradigm would be more appropriate.

where information flows from DNA to RNA and lastly to proteins to a rather intricate one (Mattick, 2003) (Figure1).

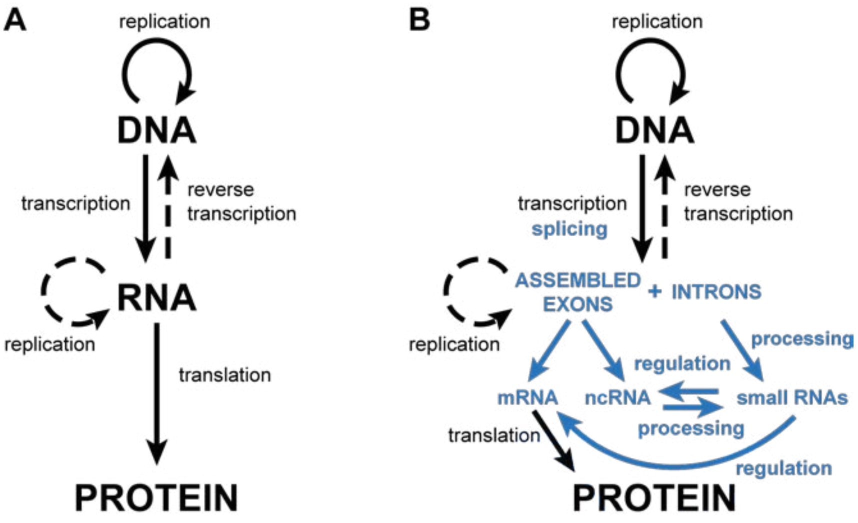


Figure 1. Schema of the Central Dogma of Biology that illustrates the flow of genetic information in eukaryotes as revised by Francis Crick in 1970 **(A)** and as we see it nowadays **(B)**. Solid arrows show general transfers; dashed arrows represent special transfers. It has to be considered that splicing can produce multiple mRNA isoforms, increasing the complexity of the schema in **B**.

In the following sections I will give a brief description on what we know about splicing and the differences in splicing across eukaryotes. Moreover, I will try to show some of the implications that splicing has from the molecular and the evolutionary point of view. Therefore, this introduction rather than being exhaustive, aims to highlight the most relevant aspects of the field from the particular point of view of the thesis that is being presented.

1.1. The splicing reaction and the spliceosome

The splicing reaction happens in the nucleus, either co-transcriptionally or post-transcriptionally (reviewed in Maniatis and Reed, 2002; Neugebauer, 2002), and requires the correct identification of exon/intron boundaries in the pre-mRNA. These boundaries are defined by the splicing signals, namely, the 5' splice site (5'ss), which identifies the beginning of the intron; the 3'ss, which the end of the intron; the branch site (BS), and the polypyrimidine tract (PPT). More details on the specific signals will be given in the next section. Biochemically, splicing consists of two transesterification reactions. In the first step, the BS attacks the 5'ss, which leads to the release of the 5' exon from the intron and the formation of a lariat intermediate. In the second step, a second transesterification reaction ligates the two exons. As a result, the exons appear together in the mature RNA and the intron is released in the form of a lariat (Padgett et al., 1984). In the cell, these signals are recognized by a macromolecular complex, the spliceosome, which catalyzes this reaction. This process requires the assembly of several spliceosome components, including the small nuclear ribonucleoprotein complexes (snRNPs) U1, U2, U3, U4/U6 and U5, which associate with the pre-mRNA in a stepwise manner. During the process, these complexes suffer several structural rearrangements in order to catalyze the splicing reaction (reviewed in Will and Lührmann, 2010) (Figure 2). There also exist another spliceosome, the minor spliceosome, which is composed of different snRNPs (U11 and U12 instead of U1 and U2, and U4atac and U6atac instead of U4 and U6) and recognizes a small subset of introns containing different splicing signals (reviewed in Tarn and Steitz, 1997). From this point onwards, I will only discuss the major spliceosome.

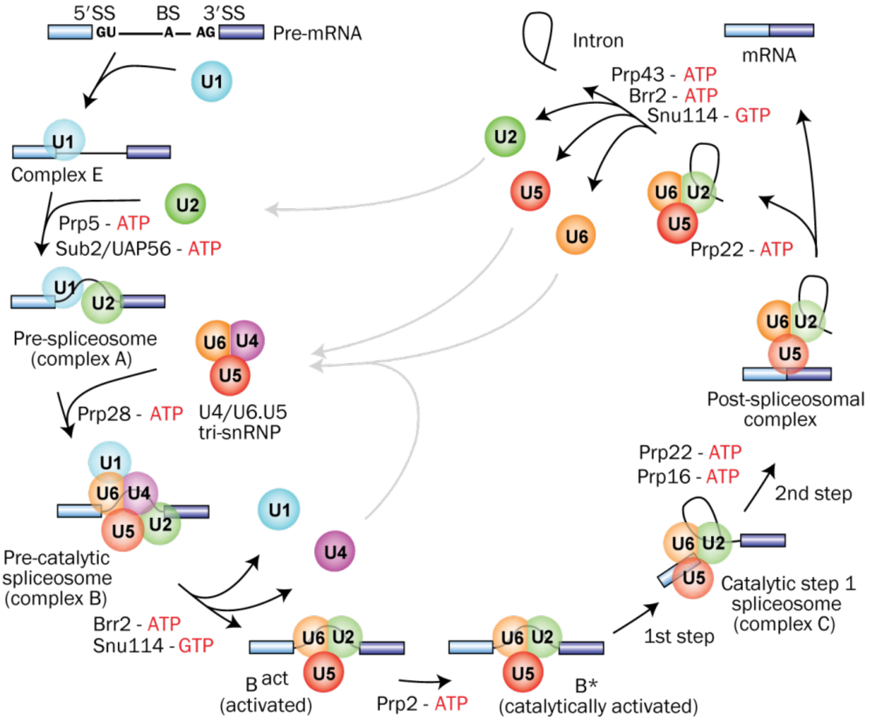


Figure 2. Spliceosome cross-intron assembly and disassembly pathway. For simplicity, only the ordered interactions of snRNPs (indicated by circles) are shown. Exon and intron sequences are represented by boxes and lines, respectively. The stages at which the evolutionarily conserved DExH/D-box RNA ATPases/helicases PRP5, SUB2/UAP56, PRP28, BRR2, PRP2, PRP16, PRP22 and PRP43, or the GTPase SNU114, act to facilitate conformational changes are indicated. Adapted from Will and Lührmann, 2010.

In addition to the spliceosome components, other proteins participate in the splicing reaction, such as splicing enhancers, which help in the recognition of the splicing signals. It is of special interest how the spliceosome performs the initial recognition of the above mentioned splicing signals, as this mechanism is one of those studied in this thesis. In higher eukaryotes, U1 snRNP recognizes the 5'ss with a direct base pairing between the U1 small nuclear RNA (snRNA) and the 5'ss (Seraphin and Rosbash, 1989). In the next step, the U2AF dimer (U2AF65 + U2AF35) performs the recognition of the PPT and the 3'ss (Zamore et al., 1992; Wu et al., 1999). This binding recruits SF1, which will contact the BS and promote the final recognition of the BS by a direct base pairing with U2 snRNA (Berglund et al., 1998) (Figure 3). Despite some differences in the components of the spliceosome (Fabrizio et al., 2009), this process is highly conserved across eukaryotes.

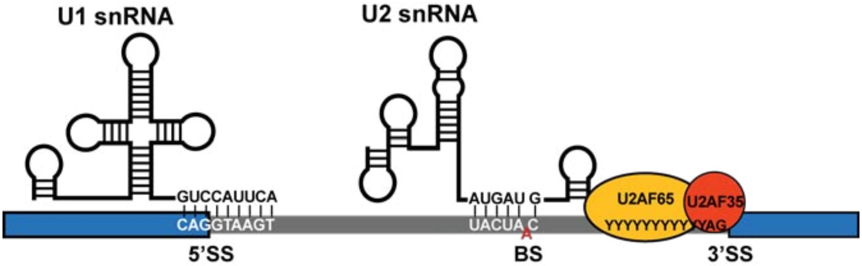


Figure 3. Initial recognition of the splicing signals by the spliceosome. Exons are represented by blue boxes. The intron is represented by a grey line. On the left, the figure shows the base pairing of the U1 snRNA with the 5'ss. On the right, U2AF65 and U2AF35 bind the PPT and the 3'ss, respectively, whereas the branch site is recognized by a direct base pairing with U2 snRNA, leaving the A unpaired.

1.2. Splicing signals

The correct identification of the splicing signals is crucial for the process of intron removal. Hence, these signals present specific sequence particularities that allow for their recognition by the spliceosome machinery. Interestingly, we can observe these dependencies across evolution, as different species have different consensus splicing signals (Schwartz et al., 2008). In the following section I explain the most important features of these signals, indicating the differences between eukaryotes.

1.2.1. The 5' splice site

The 5'ss delimits the exon/intron boundary and is characterized by the presence of a GT or a GC dinucleotide that marks the beginning of the intron. The most informative positions of the 5′ss consensus sequence are 9 nt, 3 from the end of the upstream exon, the GC or GT dinucleotide, and 4 nucleotides in the downstream intron. Regardless of variations, this sequence corresponds to the complementary sequence of U1 snRNA (Seraphin et al., 1988; Siliciano and Guthrie, 1988) (Figure 4).

1.2.2. The 3' splice site

The 3'ss delimits the intron/exon boundary at the 3′ end of the intron and consists of an AG dinucleotide, preferentially preceded by a T or a C (Mount, 1982). Despite the low information contained in the signal, only 3 nucleotides, its recognition is crucial to carry out the second step of the splicing reaction in the majority of introns (Wu et al., 1999). Interestingly, the protein responsible for the binding of the 3'ss (U2AF35) is missing in yeast, which poses an interesting question: how is the 3'ss recognized in this species?

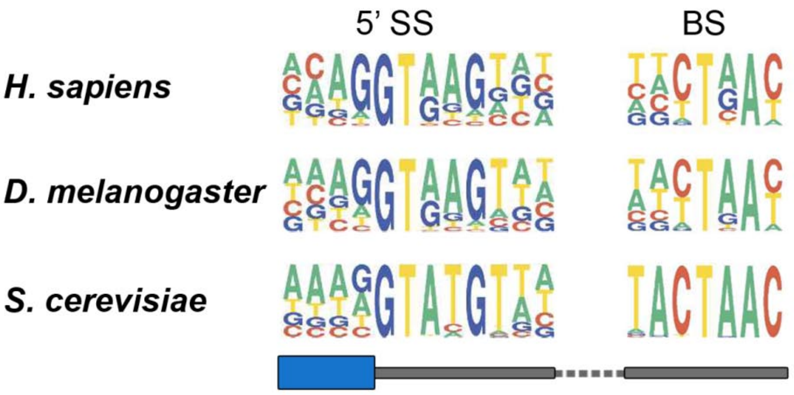


Figure 4. Comparison of 5'ss and BS sequence logos of *H. sapiens*, *D. melanogaster* and *S. cerevisiae*. The position of the upstream exon is represented by a blue box and the downstream intron by a grey line. For the 5'ss, the last four nucleotides in the exon and the first 8 nt nucleotides of the downstream intron are shown. At each position, the height of letters is proportional to the frequency of the corresponding nucleotide at the given position, and nucleotides are listed in descending order of frequency from top to bottom. Adapted from Schwartz et al., 2008.

1.2.3. The polypyrimidine tract

As its name states, this is a stretch of nucleotides enriched in pyrimidines (C and T nucleotides) present at the 3′ end of introns, upstream of the 3'ss (Mount, 1982). As mentioned in the previous section, this sequence it is bound by U2AF65 and is crucial for both, BS and 3'ss recognition (Zamore et al., 1992; Wu et al., 1999). Interestingly, the yeast U2AF65 homolog, MUD2 (Abovich et al., 1994), lacks the two RNA recognition motifs (RRMs) of U2AF65, which are responsible for the recognition of the PPT in higher eukaryotes (Zamore et al., 1992; Banerjee et al., 2004). Nevertheless, it has been shown that the PPT enhances 3'ss recognition in yeast (Patterson and Guthrie, 1991), even though its location inside introns in not always next to the 3'ss (Kupfer et al., 2004).

1.2.4. The Branch Site

In higher eukaryotes, the BS is usually located, on average, 33 nt upstream of the 3′ss (Kol et al., 2005), though there are several cases reported in which the BS is much more distant (Gooding et al., 2006). The position at which the BS is located has been shown to be important for splicing efficiency, as increasing the distance between the BS and the 3'ss reduces splicing efficiency (Cellini et al., 1986). The BS signal is characterized by the presence of an invariable A that is responsible for the nucleophilic attack on the 5'ss performed during the first step of the splicing reaction (Padgett et al., 1984). In contrast to the splice sites, the sequence of the BS is highly conserved in yeast but very variable in flies and vertebrates (Schwartz et al., 2008)

(Figure 4), though during splicing it is recognized by a direct base pairing with U2 snRNA (Black et al., 1985).

2. Alternative splicing

Splicing signals are crucial for the correct placement of the spliceosome in the pre-mRNA and the progression of the splicing reaction. Nevertheless, these signals tend to be degenerate and therefore, in some cases, the spliceosome machinery alone is not sufficient to recognize them (Schaal and Maniatis, 1999; Zhang et al., 2005). As a result, splicing is a very flexible mechanism that can be regulated, or fine tuned, in very different ways, allowing the selection of different combinations of exons from the pre-mRNA to create different mRNAs through alternative splicing (AS).

Although initially thought to be the exception rather than the rule, recent studies have estimated that up to 95% of genes are alternatively spliced in human (Pan et al., 2008). AS is a common process occurring in eukaryotes, but the frequency of this phenomenon is variable across species (Kim et al., 2007), with some species like yeast having less than 1% of transcripts alternatively spliced[2] (Yassour et al., 2009).

[2] There are several methods to estimate alternative splicing levels. More details on the techniques used can be found in Appendix A

2.1. Types of alternative splicing

According to the changes produced in the mRNA there are five main types of alternative splicing (Breitbart et al., 1987) (Figure 5):

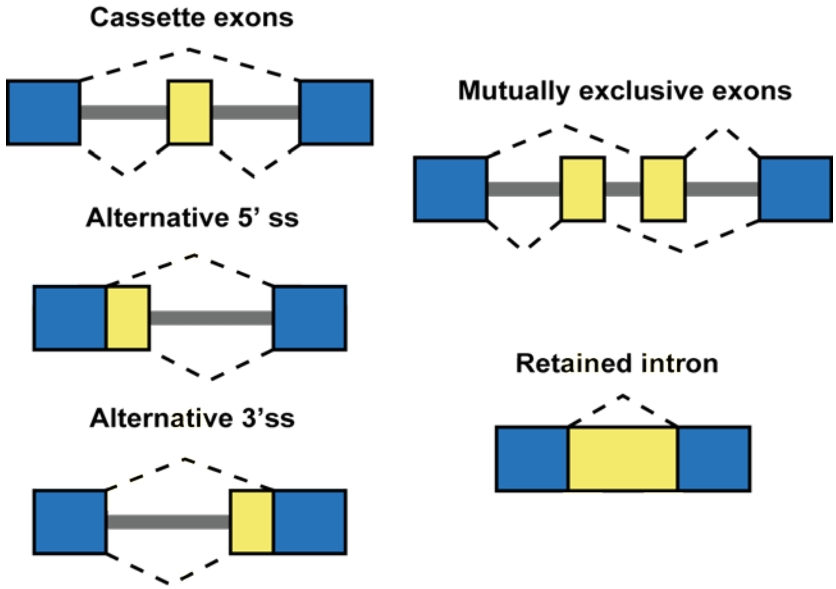


Figure 5. Common types of alternative splicing. Constitutive regions are shown in blue and alternative regions in yellow. Introns are represented with a grey line. Dashed lines indicate alternative splicing patterns.

Exon skipping: an internal exon (cassette exon) can be included or excluded from the mRNA.

Alternative 5'ss and 3'ss: an alternative 5'ss or 3'ss is used, producing longer or shorter variants of the same exon.

Intron retention: an intron that can be included or excluded in the mature transcript.

Mutually exclusive exons: two or more cassette exons that cannot appear together in the same mRNA molecule.

There are other mechanisms that can change exon composition in an mRNA without affecting splicing: alternative promoter sites, which change the first exon of an mRNA by using alternative transcription start sites (Carninci et al., 2006; Kimura et al., 2006); alternative polyadenylation sites, which change the end of the transcript (Tian et al., 2005). More complex splicing events can be obtained as a combination of the events previously described (Sammeth et al., 2008).

3. Mechanisms of splicing regulation

There are several levels at which splice site selection can be regulated. The simplest regulation can be performed directly by the spliceosome. Various studies have shown that mutation or inhibition of specific components of the spliceosome can change splicing patterns (Park et al., 2004; Pleiss et al., 2007; Saltzman et al., 2011). Additionally, the increasing evidence that splicing is coupled to transcription (reviewed in Maniatis and Reed*,* 2002; Neugebauer, 2002) has also proved that changes affecting polymerase elongation rate, such as DNA damage (Munoz et al., 2009), histone modifications (Lorincz et al., 2004), chromatin remodeling complexes (Batsche et al., 2006), or even small RNAs that alter chromatin packing (Allo et al., 2009), can influence the recognition of splice sites. More recently, it has also been demonstrated that chromatin structure is important for splice site recognition at a different level. Genome-wide analyses have shown that there is a clear correlation between nucleosome positioning and splice site location, suggesting that the positions of the nucleosomes on the pre-mRNA are important for splice site identification (Schwartz et al., 2009; Tilgner et al., 2009). Furthermore, it has also been proved that the structure adopted by the pre-mRNAs can hinder splice site recognition or enhance splicing by bringing splicing signals into close proximity (reviewed in Warf and Berglund, 2010). At a different level, splice site selection can also be triggered by specific protein factors that enhance (splicing enhancers) or decrease (splicing silencers) the recognition of the splicing signals by binding to specific sequences on the pre-mRNA (reviewed in Black, 2003).

3.1. Splicing factors

Proteins that enhance or repress the recognition of splicing signals are called splicing factors. These proteins usually bind to sequence motifs in exons or introns, which can be classified according to their function and location: intronic splicing enhancers (ISEs), intronic splicing silencers (ISSs), exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs)[3] (reviewed in Chasin, 2007). Enhancers and silencers perform

[3] More details on the techniques used to identify enhancer or silencer sequences on the pre-mRNA can be found in Appendix B

opposite functions on splice site selection. Therefore, these elements act in a combinatorial way. The balance of the competing enhancers and silencers determines the final splicing outcome (Figure 6) (reviewed in Matlin et al., 2005). Furthermore, these sequence elements are important for correctly identifying exons and distinguishing them from pseudoexons (Corvelo and Eyras, 2008), and participate not only in the regulation of alternative exons but also in recognition of constitutive exons (reviewed in Chasin, 2007).

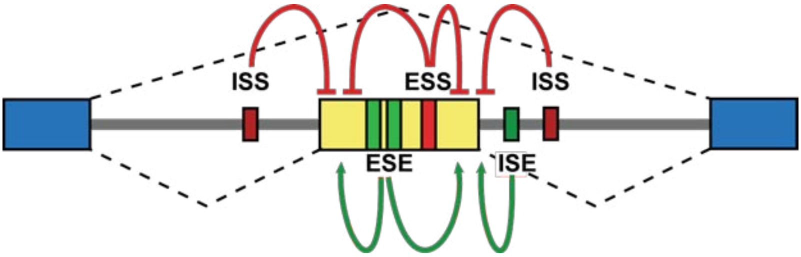


Figure 6. Combinatorial control of exon recognition. Alternative splicing patterns are indicated by dashed lines. Constitutive and alternative exons are marked by blue and yellow boxes respectively. The location of enhancers (ESE, ISE) and silencers (ESS, ISS) is indicated by green and red boxes. Enhancers can activate adjacent splice sites or antagonize silencers, whereas silencers can repress splice sites or enhancers. Exon inclusion or skipping is determined by the balance of these elements and the relative abundance of the factors that recognize them. Adapted from Matlin et al., 2005.

There are two main types of splicing factors: splicing enhancers, mainly represented by the serine arginine rich (SR) protein family, and splicing silencers, mostly represented by hnRNPs. Some of these factors are expressed in a tissue specific manner, such as Nova1 (Jensen et al., 2000), nPTB (Rahman et al., 2002), Fox-1 (Nakahata and Kawamoto, 2005), or CELF (Ladd et al., 2001), and thereby regulate splicing in specific tissues. Enhancer proteins help the recognition of splicing signals by promoting the recruitment of spliceosome components during the early stages of spliceosome assembly, mainly through protein-protein interactions (reviewed in Graveley, 2000). In contrast, silencer proteins inhibit the recognition of splicing signals by counteracting the functions of enhancer proteins or by preventing their binding to the pre-mRNA (reviewed in Martinez-Contreras et al., 2007).

3.1.1. SR proteins

SR proteins are one of the main enhancer protein families. These proteins are characterized by the presence of one or two RNA recognition motifs (RRMs) at the N-terminal end and an arginine-serine rich (RS) domain at the C-terminal end (Fu and Maniatis, 1992; Zahler et al., 1992) (Figure 7). There are at least 12 different members of this protein family in human (Manley and Krainer, 2010) (Table 1). Members of this family can be found in animals, plants and even protists, but their distribution is not even across eukaryotes, with some species having expansions of particular proteins, and others, such as yeast, lacking all of them (Barbosa-Morais et al., 2006).

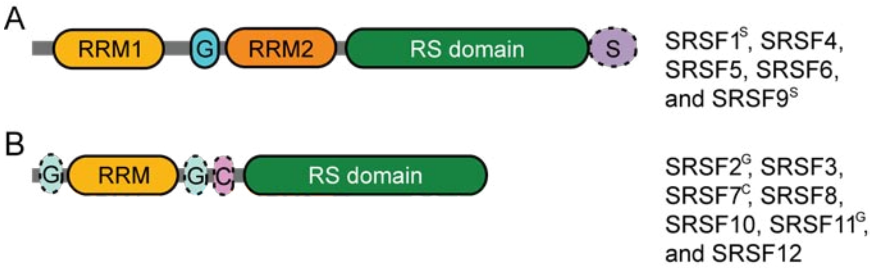

A
RRM1
G
RRM2
RS domain
S
SRSF1^{S}, SRSF4, SRSF5, SRSF6, and SRSF9^{S}
B
G
RRM
G
C
RS domain
SRSF2^{G}, SRSF3, SRSF7^{C}, SRSF8, SRSF10, SRSF11^{G}, and SRSF12

Figure 7. Structure of SR proteins. Conserved domains are marked with solid lines. Variable domains are marked with dashed lines. **(A)** SR proteins containing two RRM domains. The second RRM (RRM2) of these proteins is characterized by the presence of an invariable SWQDLKD motif. All proteins in this group also contain a glycine rich region and an RS domain. **(B)** SR proteins containing one RRM domain. All proteins in this group also contain an RS domain. The abbreviations used are: RRM, RNA Recognition Motif; RS domain, Arginine serine rich domain; G, Glycine rich region; C, CCHC-type zinc finger domain; S, region of interaction with SAFB1. Proteins containing variable groups are marked by ^{G}, ^{C} and ^{S}, respectively.

Table 1. Known SR proteins in human

Protein/gene symbol	**Aliases**
SRSF1	ASF, SF2,SRp30a
SRSF2	SC35, PR264, SRp30b
SRSF3	SRp20
SRSF4	SRp75
SRSF5	SRp40, HRS
SRSF6	SRp55, B52
SRSF7	9G8
SRSF8	SRp46
SRSF9	SRp30c
SRSF10	TARS1,SRp38, SRrp40
SRSF11	p54, SRp54
SRSF12	SRrp35

Adapted from Manley and Krainer, 2010.

A. Functions of SR proteins

SR proteins are known to participate not only in the regulation of alternative splicing but are also important for the definition of constitutive exons (reviewed in Bourgeois et al., 2004). Furthermore, they can also be involved in other cellular processes including mRNA nuclear export, mRNA stability and quality control, translation, genomic stability, and even oncongenic transformation (reviewed in Huang and Steitz, 2005; Long and Caceres, 2009; Zhong et al., 2009), highlighting their key role in post-transcriptional regulation.

As splicing enhancers, SR proteins may function in two different ways. The first possibility is that when bound to the pre-mRNA, they recruit spliceosome components through protein-protein interactions mediated by the RS domain. These interactions can be direct with spliceosomal components containing RS domains, such as U2AF65 or the U1 snRNP, or they can be mediated through other splicing factors (Wu and Maniatis, 1993). An alternative mechanism proposes that when attached to the pre-mRNA, SR proteins can stabilize the interaction between snRNAs and the splicing signals, in a mechanism that is conserved across eukaryotes (Shen et al., 2004; Shen and Green, 2004; Shen et al., 2006).

B. Regulation of SR proteins

The function of SR proteins can be regulated at different levels. On the one hand, it has been shown that the expression levels of SR proteins can be regulated by alternative splicing. Pre-mRNAs of SR proteins can produce alternative transcripts containing premature termination codons (PTCs) that trigger the degradation of the mRNAs by nonsense-mediated decay (NMD) (Ni et al., 2007; Lareau et al., 2007b). On the other hand, the function of SR proteins can be regulated by affecting their function as splicing regulators (Prasad et al., 1999) or their cellular localization (Caceres et al., 1998; Misteli et al., 1998). This can be achieved through phosphorylation or dephosphorylation of the RS domain.

3.2. RNA structures

During transcription, the pieces of RNA synthesized can fold before getting spliced. The secondary structures adopted by the pre-mRNA are very important for splicing regulation as they can promote or repress the recognition of splicing signals and the binding of splicing factors to the pre-mRNA (Figure 8). RNA structures can hinder the recognition of splicing signals by occluding them and preventing their recognition by spliceosome components. Computational genome-wide analyses have shown that conserved RNA secondary structures overlapping splice sites are related to alternative splicing (Shepard and Hertel, 2008). Besides, these structures can facilitate the recognition of splicing signals by shortening the distance between them (reviewed in Buratti and Baralle*,* 2004; Warf and Berglund, 2010). In other cases, RNA structures can regulate complex splicing patterns, as in the case of the *Dscam* gene in *Drosophila melanogaster* (Graveley et al., 2005).

All these examples may indicate that the secondary structure adopted by the pre-mRNA governs splicing. However, RNA folds co-transcriptionally (reviewed in Pan and Sosnick, 2006), and the particular structures may change when more RNA is available (reviewed in Bevilacqua and Blose, 2008; Mahen et al., 2010). Furthermore, these structures can be altered by temperature, transcription, or other factors that prevent their formation or stabilize them (reviewed in Chen, 2008; Pan and Sosnick, 2006; Warf and Berglund*,* 2010), thus providing more possibilities for splicing regulation.

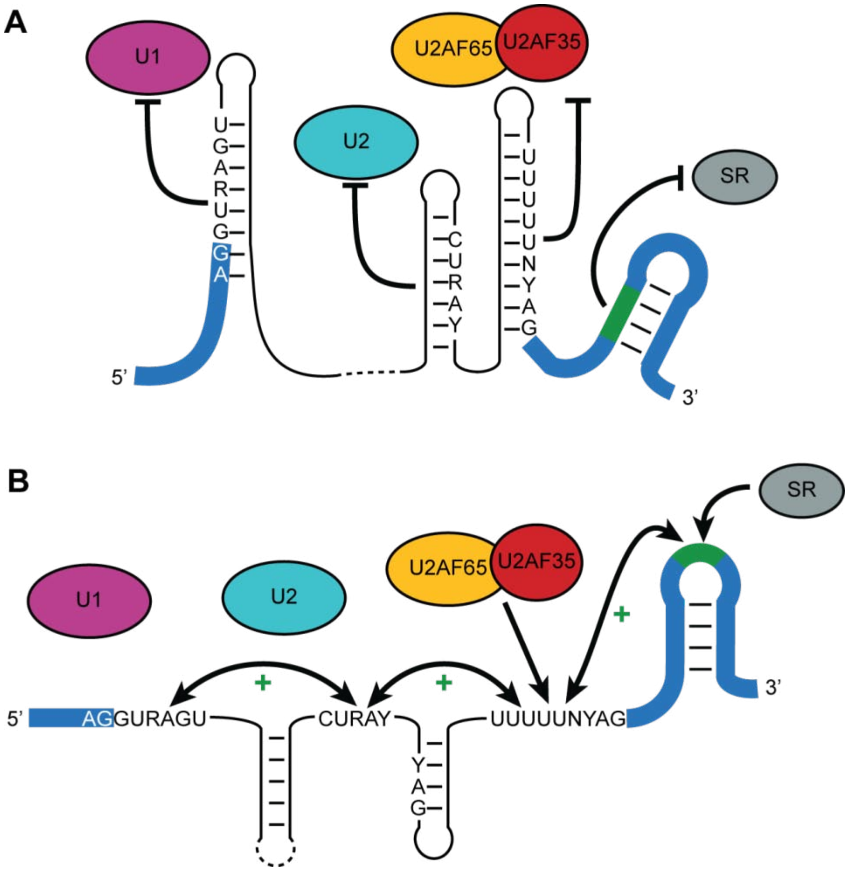


Figure 8. Role of RNA structures that directly regulate splicing. Exons and introns are represented by blue and black lines respectively. **(A)** Representative diagram of structures that inhibit splicing. Shown are stem-loops that repress binding of the U1 snRNP (pink) to the 5′ss, the U2 snRNP (light blue) to the BS, U2AF65 and U2AF35 (yellow and orange) to the 3′ss, and an SR protein (grey) to a sequence within the exon. **(B)** Representative diagram of structures that aid splicing. Depicted is a structure that brings the 5′ and 3′ splice sites into closer proximity, a stem that brings the 3′ splice site and branch-point into closer proximity, a stem that masks a cryptic 3′ splice site (denoted as YAG) and a stem that properly displays an enhancer sequence in the exon that an SR protein binds. Adapted from Warf and Berglund, 2010.

4. Functions of alternative splicing

Traditionally it was thought that the function of alternative transcripts was to increase the protein diversity of an organism (reviewed in Graveley, 2001). However, it has also been shown that, in some cases, alternative transcripts do not get translated into proteins. These transcripts have low relative abundances and usually get degraded, affecting gene regulation at post-transcriptional level (reviewed in Lareau et al., 2007a).

4.1. Impact of alternative splicing on protein evolution

Shortly after the first observation that a single pre-mRNA could produce different proteins (Berget et al., 1977; Chow et al., 1977), the idea that this process could open a range of possibilities for evolution was also proposed (Gilbert, 1978). Further research has highlighted the key role of AS in this process. Nowadays, it is known that AS is one of the major contributors to protein diversity, probably explaining the observed discrepancy between the number of genes and the complexity of an organism (Modrek and Lee, 2002; Pan et al., 2004). Several studies show that AS allows the acquisition of new protein functions by creating or adding new functional domains, or removing or disrupting existing ones (Hiller et al., 2005; Kriventseva et al., 2003; Romero et al., 2006; Xing et al., Lee, 2003). At the genomic level, this is translated into distinct types of alternative splicing events that change the exon composition of mature transcripts.

The study of how all this protein complexity can be achieved, from the molecular point of view, has interested scientists for a long time. Measures of the substitution rates at synonymous (Ks or dS) and non-synonymous (Ka or dN) sites have demonstrated that alternatively spliced regions have higher dN values, i.e. are less constrained or under relaxed purifying selection (Iida and Akashi, 2000; Xing and Lee, 2005; Chen et al., 2006; Ermakova et al., 2006), but also show higher sequence conservation, i.e. are under strong purifying selection (Philipps et al., 2004; Sorek et al., 2004). This apparent discrepancy is due to the existence of different populations of alternative exons: those exons poorly included in mRNAs have little effect on fitness and are under relaxed selection pressure, which allows them to explore new functions and therefore show high dN values. In contrast, alternative exons that are highly included in mRNAs show higher sequence conservation (Modrek and Lee, 2003).

This fact has also been related to the acquisition of regulation by these exons. Several studies have demonstrated that regulatory sequences are under selective pressure (Hurst and Pal, 2001; Orban and Olah, 2001; Carlini and Genut, 2006; Parmley et al., 2006). Furthermore, it has been suggested that poorly included alternative exons can appear by *de novo* exonization of introns or other types of non-coding sequences, such as Alu elements (Sorek et al., 2002; Modrek and Lee, 2003). At the beginning, these exons will have low amounts of regulatory sequences, will be included at low rates, and will be under relaxed selective pressure (higher dN and dS values), which will allow them to explore new functions with a minimum impact on the normal expression of the gene. However, when these alternative exons are included at high frequency in mature transcripts, or when they change from constitutive to alternative, they will be highly regulated, thus showing lower dN and dS values (Xing and Lee, 2005).

4.2. Alternative splicing as a post-transcriptional regulatory mechanism

A high proportion of alternative transcripts do not produce functional proteins. Instead, these transcripts introduce PTCs and get degraded through NMD (reviewed in Maquat, 2006). NMD is a surveillance mechanism for mRNA that avoids the production of proteins that could have a negative effect on the cell (Cali and Anderson*,* 1998; Maquat and Carmichael, 2001). Moreover, it allows modulating the final mRNA concentrations on the cell (reviewed in Lejeune and Maquat, 2005). This process is widespread across eukaryotes, although the mechanisms of action are different. In mammals, NMD is triggered in general by the presence of a STOP codon located more than 50 nt away from the last exon junction (reviewed in Maquat, 2006). In contrast, in other species such as *Arabidopsis* or yeast, NMD is triggered through a longer than expected 3' untranslated region (3'UTR) in the terminal part of genes (Conti and Izaurralde, 2005).

The relevance of AS coupled to NMD is still not clear. There are several studies that show that alternatively spliced isoforms that are targeted by NMD are conserved across species, suggesting that AS coupled to NMD is functional (Baek and Green, 2005; Pan et al., 2006). In contrast, other studies show that the frequency of the transcripts produced by AS coupled to NMD is very low, suggesting that these transcripts are the result of splicing noise. Furthermore, degradation of these transcripts may have little or no effect on the final mRNA levels (Neu-Yilik et al., 2004). Interestingly, several genes have been shown to be regulated by NMD. Among these, there are several splicing factors and RNA binding proteins that regulate their own splicing through the production of isoforms that are degraded by NMD (reviewed in Lareau et al., 2007a). Thus, even though AS coupled to NMD may not be functional in all cases, subtle changes in mRNA levels may be important in a subset of genes. Additionally, the regulation of mRNA levels by NMD could also have allowed for the appearance during evolution of more alternative spliced variants, contributing to the diversity of transcripts.

II. OBJECTIVES

There are two main objectives in this thesis. First, we want to understand how splicing has changed in eukaryotes. Second, we want to study the function of alternative splicing in eukaryotes.

The specific objectives of the thesis are:

1. Analyze the conservation of SR and SR-like proteins in eukaryotes, with a special interest in the properties of the RS domain and its implications in splicing regulation.

2. Understand the relation between SR proteins and splicing signals, specially the branch site.

3. Study the regulation of splicing in the absence of SR proteins. More specifically, the role of RNA secondary structures in 3′ss selection in yeast.

4. Understand the impact of alternative splicing as a mechanism to regulate gene expression in yeast.

5. Analyze the role of alternative splicing in sequence evolution.

6. Understand the relation between transcript structure, sequence conservation and alternative splicing.

III. RESULTS

Co-evolution of the branch site and SR proteins in eukaryotes

Mireya Plass, Eneritz Agirre, Diana Reyes, Francisco Câmara and Eduardo Eyras

Trends Genet. 2008; 24(12): 590-594.

1.2. Supplementary Material

Data sets

The genome sequences, annotations, and protein datasets were obtained for 22 different species including vertebrates (*Homo sapiens*, *Gallus gallus* and *Danio rerio*); invertebrates (*Caenorhabditis elegans* and *Drosophila melanogaster*); plants (*Arabidopsis thaliana*); protists (*Cryptosporidium parvum* and *Dictyostelium discoideum*); and 14 fungal species from 3 different groups: *Chytridiomycota* (*Batrachochytrium dendrobatidis*), *Mucoromycotina* (*Rhizopus oryzae*), and *Dykarya*. In the latter group we have the *Basidiomycota* (*Puccinia graminis*, *Cryptococcus neoformans* and *Ustilago maydis*); and the *Ascomycota* (*Schizosaccharomyces pombe, Neurospora crassa*, *Coccidiodes immitis, Aspergillus nidulans*, *Yarrowia lipolytica*, *Debaryomyces hansenii*, *Ashbya* gossypii, *Kluyveromyces lactis* and *Saccharomyces cerevisiae*) [1].

The data for *A. gossypii*, *A. nidulans*, *B. dendrobatidis*, *K. lactis*, *N. crassa, C. immitis, P. graminis*, *R. oryzae*, *S. pombe*, *C. neoformans* (JEC21), and *U. maydis* were obtained from the fungal genome resources built by J. Stajich (http://fungal.genome.duke.edu/) and from the Broad Institute Project page (http://www.broad.mit.edu/annotation/fgi/). The gene annotations and the protein prediction sets for *R. oryzae*, *P. graminis* and *B. dendrobatidis* were generated with the gene prediction program Geneid [2]. We also used the annotations for *A. thaliana* from The Arabidopsis Information Resource [3], *C. elegans* from wormbase version 150 [4], *C. parvum* from CryptoDB.org release 3.6 [5], *D. hansenii* and *Y. lypolitica* from the Genolevures project web page [6], *D. discoideum* from NCBI [7], *D. melanogaster* from flybase [8] and *H. sapiens*, *D. rerio*, *G. gallus*, and *S. cerevisiae* from Ensembl version 43, Ensembl version 48 and Ensembl version 49 respectively [9].

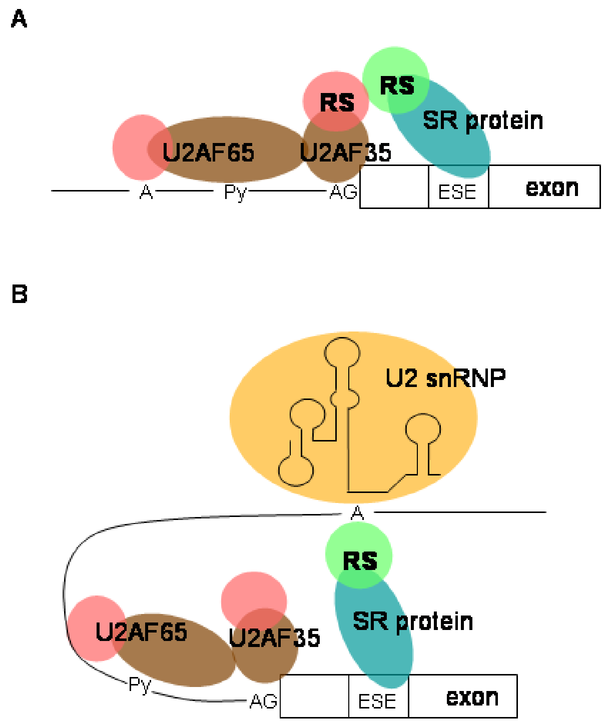


Figure S1. Modes of interaction of SR proteins through its RS domain. (A) An SR protein bound to an exon can recruit components of the spliceosome machinery and facilitate their binding to the pre-mRNA. This interaction is performed trough the interaction of the RS domains of the SR and SR-related proteins [10]. **(B)** An SR protein bound to an exon can directly contact the BS and stabilize its interaction with the U2 snRNA in a non specific manner [11-14].

Table S1. Datasets. Characteristics of the datasets used in the analyses. For those species with alternative splicing annotation, only constitutive introns were used.

Species	Gene number*	Intron number	Mean intron length	Median intron length	Mean introns per gene
D. discoideum	7687 (10522)	15614	167.96	105	2.03
C. parvum	268 (4278)	420	117.83	62	1.57
A. thaliana	20810 (27029)	110635	158.25	98	5.32
H. sapiens	18042 (23224)	152589	5023.86	1445	8.46
G. gallus	14001(16736)	118444	1717.02	749	8.46
D. rerio	16385 (21322)	113438	2104.24	808	6.92
D. melangaster	10708 (14039)	37387	632.67	69	3.49
C. elegans	19509 (20069)	99978	282.84	63	5.12
R. oryzae	10279 (12744)	34965	104.15	59	3.40
B. dendrobatidis	8283 (8956)	28921	158.77	95	3.50
P. graminis	40724 (50757)	96343	185.64	117	2.36
C. neoformans	6183 (6652)	32567	85.42	55	5.27
U. maydis	2462 (6522)	4900	126.95	95	1.99
S. pombe	206 (4970)	431	82.53	57	2.09
A. nidulans	8430 (9541)	25585	101.58	63	3.03
C. immitis	9542 (11640)	30258	209.78	102	3.17
N. crassa	9795 (12788)	20375	312.04	129	2.08
Y. lipolytica	658 (6521)	722	269.80	212	1.10
D. hansenii	326 (6896)	344	167.89	89	1.05
A. gossypii	171 (4726)	174	113.73	63	1.02
K. lactis	126 (5331)	127	345.94	273	1.01
S. cerevisiae	263 (6698)	272	235.95	143	1.03

*number of spliced genes. Between parentheses we give the total number of genes in the annotation.

Building of Hidden Markov Models (HMMs) and search for splicing factors

We compiled a set of known splicing factors including SR proteins, SR-related proteins, hnRNPs, and other proteins known to be involved in splicing (Table S2) from Swissprot database and from the literature [15]. Proteins were grouped into families and from each group we built a hidden Markov model (HMM) for each of the RNA binding domains (RRMs and KH-type) using

Hmmer (http://hmmer.janelia.org) [16]. For proteins with no RNA binding domain (e.g. SRm300) we built an HMM for the entire protein. HMMs for the non-SR proteins but with RRMs were used as negative control for the homology search. A list of all the proteins used to build the HMMs can be found in Table S2.

Table S2. List of proteins used to create the HMMs. We build independent HMMs for each of the protein domains.

Protein subset	Proteins used for building the HMMs	
SR proteins	ASF SRp30C SRp20 9G8 SRp40 SRp55 SRp75 SC35 SRp46	RY1 SRm300 Topo I-B P54 SRp86
SR-related proteins	U2AF35 U2AF65 MUD2 NPL3	TRA2 RNPS1 U1-70K
hnRNPs	Musashi hnRNP-A hnRNP-C hnRNP-D hnRNP-F-H hnRNP-E hnRNP-G	hnRNP-I hnRNP-K hnRNP-L hnRNP-M hnRNP-R hnRNP-K
Other proteins	CUG ELAV U1A U1C	FUSE TIA1 U2B'' U2A

We searched for SR proteins in all species using the program *hmmsearch* [16] and extracted a list of proteins containing candidate RNA binding domains. To define a protein homolog we took into account the score of the HMM and the domain structure conservation (Figure S2). The order and the number of domains were required to be conserved in the query protein.

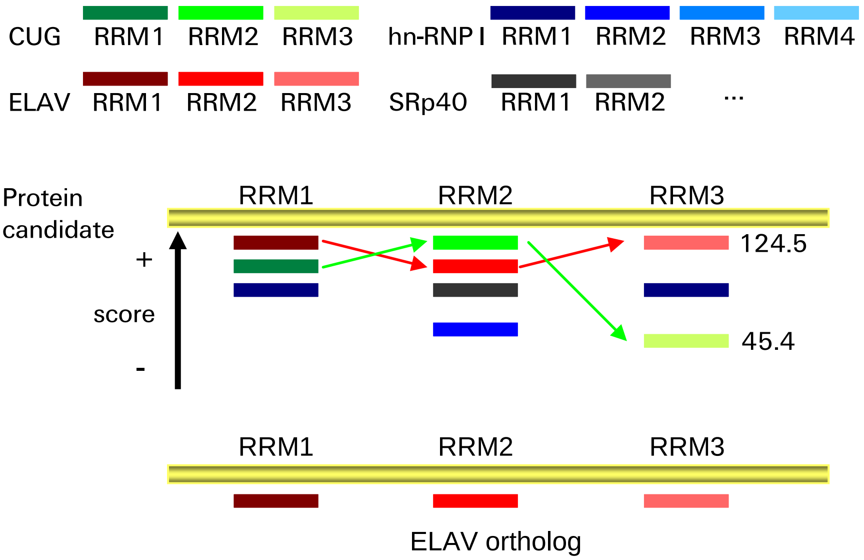


Figure S2. Homology assignation method. We consider sequence conservation and domain structure conservation. First, we score all the RRM domains of our query protein with the HMMs built for the different domains of known splicing factors. For each of the target proteins, we sum the score of collinear domain hits in our query target (e.g. score CUG RRM1 + score CUG RRM2 + score CUG RRM3) and we give a value for each of the target proteins (e.g., CUG = 45.4; ELAV =124.5). We consider the protein to be orthologous to the protein for which the sum of collinear domains is maximal. In the example, we would define our query protein as orthologous to ELAV.

Accordingly, we labeled a protein as ortholog if it had collinear hits for a multi domain protein or a single hit for a single domain protein, with a global score equal or higher than 100. When no clear ortholog could be identified, we took the best candidate and built a tree with the RNA binding domains of the possibly related proteins to ensure the correct classification (Figures S4, S5, S6 and S7). Additionally, we verified that the functional sites in the RRM domains [17,18] were conserved.

In several cases, the protein was not complete or was not found in the protein prediction set for a given species. We therefore used a combination of Exonerate [19] and GeneWise [20] to search for an SR protein candidate directly in the genome sequence. All the candidate homologs were independently verified using BLAST [21] against GenBank NR [7] to make sure they did not correspond to any other RNA-binding-domain containing protein.

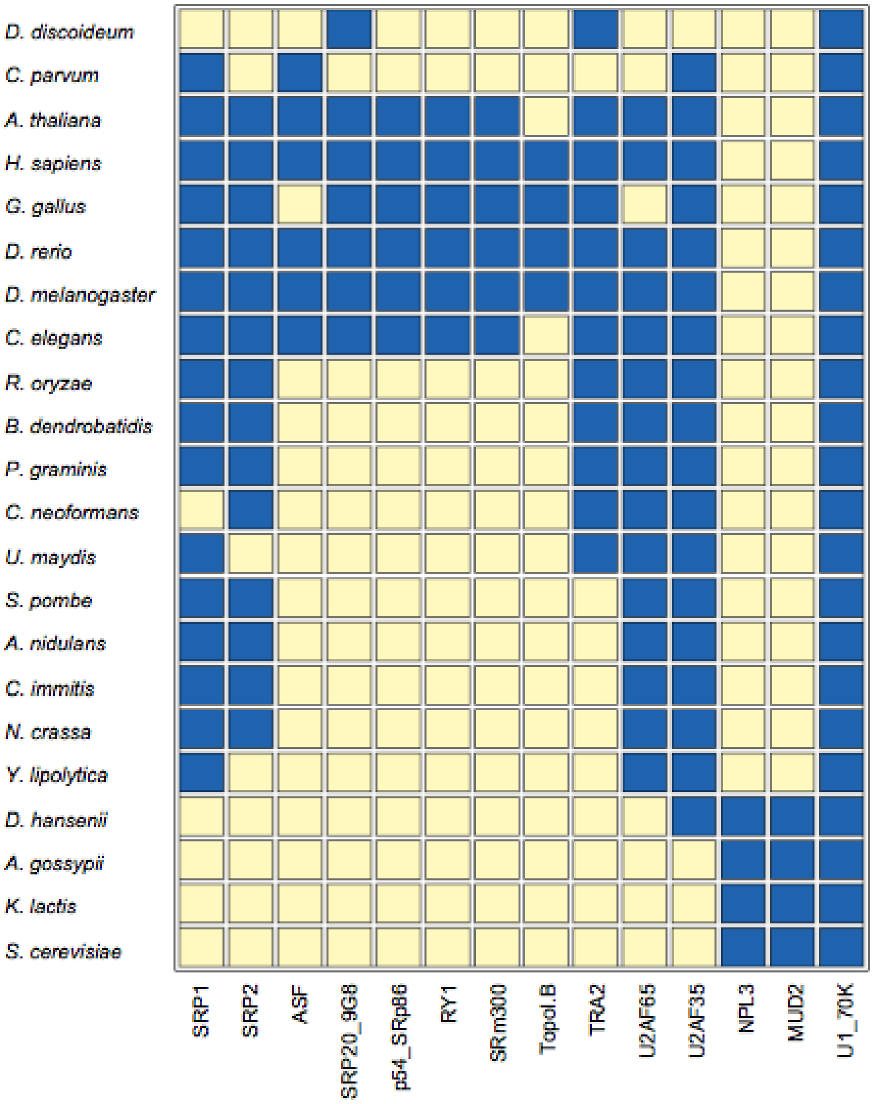
Maximum parsimony tree building

Maximum parsimony (MP) trees were built using the Close-Neighbour-Interchange algorithm with search level 3. The initial trees were obtained with random addition of sequences using 10 replicates. The trees (Figures S4, S5, S6 and S7) were drawn to scale, with branch lengths calculated using the average pathway method, and the units used were the number of changes over the whole sequence. All positions containing gaps and missing data were eliminated from the dataset. The percentage of tree-replicates, from a total of 1000 replicates, in which the associated taxa clustered together in the bootstrap test are shown next to the branches. Analyses were conducted using MEGA4 [22].

Splicing factor protein homologs

A summary of all the proteins identified can be seen in Figure S3. Most of the species tested have proteins homologous to the *S. pombe* SR proteins SRP1, which has 1 RRM domain and is homologous to SC35 and SRp46 in human; and SRP2, which encodes two RRM domains and is homologous to the proteins of the SRp40 family (i.e. SRp40, SRp55 and SRp75) (Figure S4). One of the exceptions was *C. neoformans*, for which we could not find SRP1. In fact, no SRP1 homolog was found in other closely related species: *C. neoformans* strain CH99, *Cryptococcus gattii* strain R265 and strain WM276, and *Coprinus cinereus*, which belong to the same subphylum of fungi (*Agaricomycotina*). Interestingly, SRP1 is not essential in fungi [23] and has not been implicated in splicing [24]. In contrast, SRP2 is essential [25] and has been implicated in the regulation of splicing [24]. Moreover, *C. neoformans* seems to have alternative splicing [26]. Thus, although SRP1 was probably lost in this lineage, splicing can still be regulated.

In the case of *R. oryzae* the SRP1 homolog identified has an inversion in the order of the domains, as the RS domain is in the N-terminal of the protein instead of in the C-terminal region. The protein cannot be unambiguously placed in the domain tree (Figure S5), yet we considered this as putative homolog as the closest hits were from SRP1 (SC35) homologs. For *U. maydis* we found an SRP2 homolog but the second RRM domain was right at edge of the sequence scaffold, so there is no available information on the C-terminal region. Interestingly, *R. oryzae* has two non-exact copies of SRP2 (Figure S4) and of TRA2 (Figure S5). Thus these are probably recent duplications.



	SRP1	SRP2	ASF	SRP20_9G8	p54_SRp86	RY1	SRm300	TopoI.B	TRA2	U2AF65	U2AF35	NPL3	MUD2	U1_70K
D. discoideum	□	□	□	■	□	□	□	□	■	□	□	□	□	■
C. parvum	■	□	■	□	□	□	□	□	□	□	■	□	□	■
A. thaliana	■	■	■	■	■	■	■	□	■	■	■	□	□	■
H. sapiens	■	■	■	■	■	■	■	■	■	■	■	□	□	■
G. gallus	■	■	□	■	■	■	■	■	■	□	■	□	□	■
D. rerio	■	■	■	■	■	■	■	■	■	■	■	□	□	■
D. melanogaster	■	■	■	■	■	■	■	■	■	■	■	□	□	■
C. elegans	■	■	■	■	■	■	■	□	■	■	■	□	□	■
R. oryzae	■	■	□	□	□	□	□	□	■	■	■	□	□	■
B. dendrobatidis	■	■	□	□	□	□	□	□	■	■	■	□	□	■
P. graminis	■	■	□	□	□	□	□	□	■	■	■	□	□	■
C. neoformans	□	■	□	□	□	□	□	□	■	■	■	□	□	■
U. maydis	■	□	□	□	□	□	□	□	■	■	■	□	□	■
S. pombe	■	■	□	□	□	□	□	□	□	■	■	□	□	■
A. nidulans	■	■	□	□	□	□	□	□	□	■	■	□	□	■
C. immitis	■	■	□	□	□	□	□	□	□	■	■	□	□	■
N. crassa	■	■	□	□	□	□	□	□	□	■	■	□	□	■
Y. lipolytica	■	□	□	□	□	□	□	□	□	■	■	□	□	■
D. hansenii	□	□	□	□	□	□	□	□	□	□	■	■	■	■
A. gossypii	□	□	□	□	□	□	□	□	□	□	□	■	■	■
K. lactis	□	□	□	□	□	□	□	□	□	□	□	■	■	■
S. cerevisiae	□	□	□	□	□	□	□	□	□	□	□	■	■	■

Figure S3. Splicing factor map. The two-colour heat-map indicates presence (dark blue) or absence (light yellow) of SR and SR-related proteins in each species.

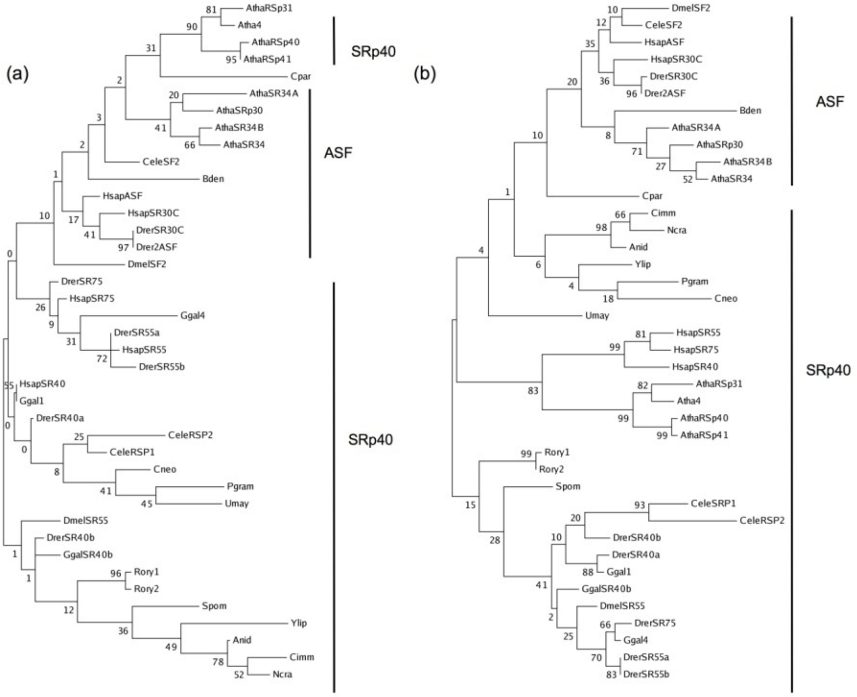


Figure S4. Maximum Parsimony Tree for SRP2 and ASF homologs. The trees are built with the RRM1 **(a)** and RRM2 **(b)** of the SRP2 and ASF known proteins and candidate homologs. We label each protein using a 4 letter short form of the species names (e.g. *H. sapiens* = Hsap). For metazoan and plants we use the protein name when available. For fungi and *C. parvum* we just use the short form of the species name. We label the groups in the tree according to their proximity to known proteins.

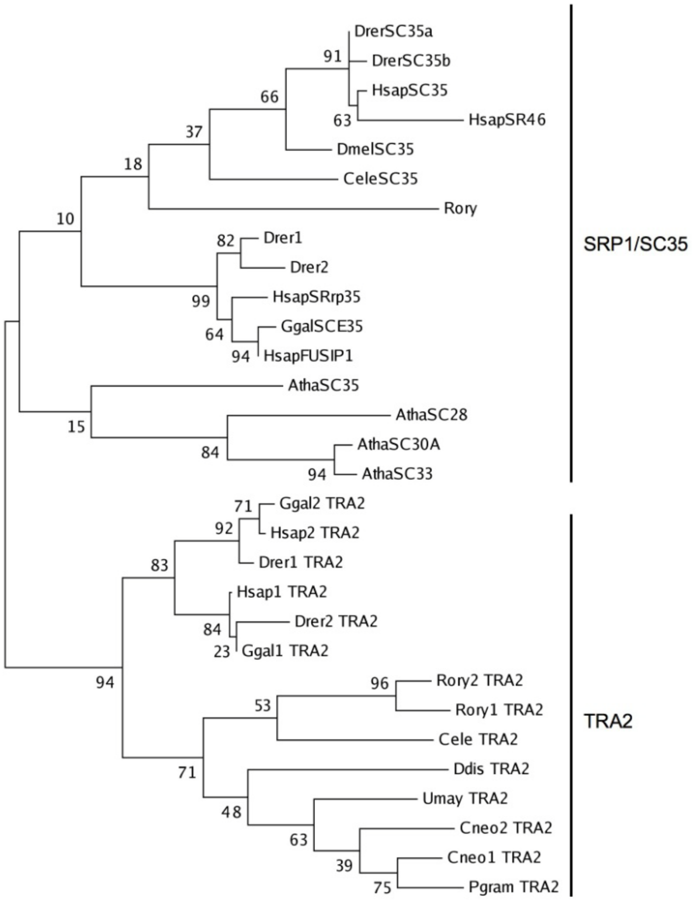


Figure S5. Maximum Parsimony Tree for TRA2 and SC35 homologs. The tree has been built with the RRMs from the SC35 and TRA2 known proteins and candidate homologs. TRA2 homologs are labelled, whereas SC35 and SRp46 homologs are indicated with the gene name if available or, like in the case of *R. oryzae*, with an identifier for the species, e.g. Rory. We label the groups in the tree according to their proximity to known proteins.

Another exception is the Saccharomycetaceae family, which has no SR proteins. Instead, they have an NPL3 homolog, whose RRMs domains are related to those from SRP2 (Figure S6). We also verified that all the Saccharomycetaceae species that have been sequenced to date have NPL3 but no SR proteins (data not shown).

All the species with SR protein homologs also have homologs to U2AF35 and U2AF65. In contrast, the Saccharomycetaceae have homologs to the *S. cerevisiae* protein MUD2, which has an analogous function to the U2AF dimer. Interestingly, *D. hansenii* has a possible homolog to U2AF35 protein, but we were unable to identify a good candidate to the U2AF65 protein. All the species analyzed have homologs of U1-70K protein. However, we could not identify any homolog of TRA2 in the ascomycetes.

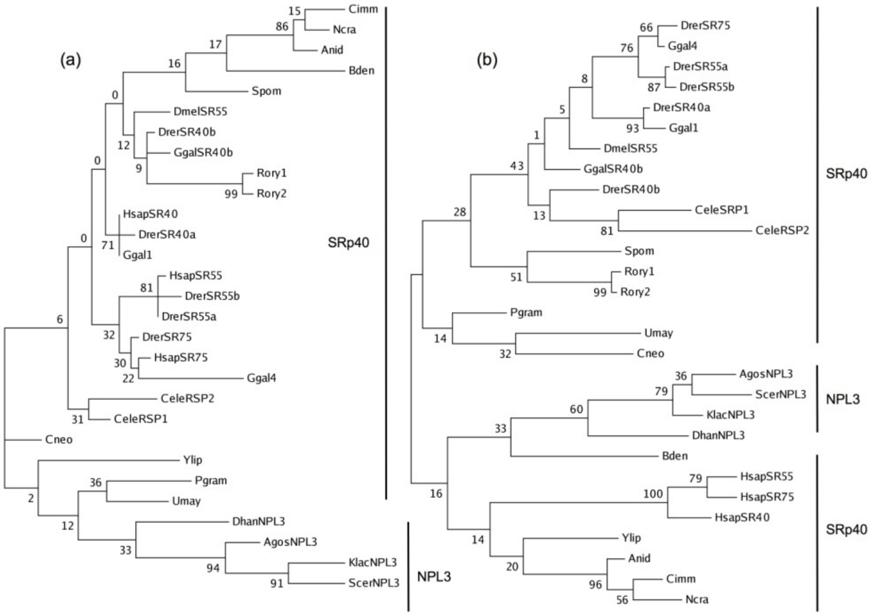


Figure S6. Maximum Parsimony Tree for SRP2 and NPL3 homologs. The trees are built with the RRM1 **(a)** and RRM2 **(b)** of the SRP2 and NPL3 homologs. Fungal homologs of SRP2 are denoted by a 4-letter form of the species name (e.g. *Y. lipolytica* = Ylip). For metazoans and plants we add to this the protein name when available. We label the groups in the tree according to their proximity to known proteins.

The analysis of protists revealed a more divergent pattern. *D. discoideum* has homologs for TRA2 (Figure S5) and SRp20

(Figure S7). In contrast, *C. parvum* has two SRP1 homologs (Figure S5) and one ASF homolog (Figure S4). These differences with animals and fungi agree with other divergent properties, like the lack of U2AF65 in both species and of U2AF35 in *D. discoideum* [27].

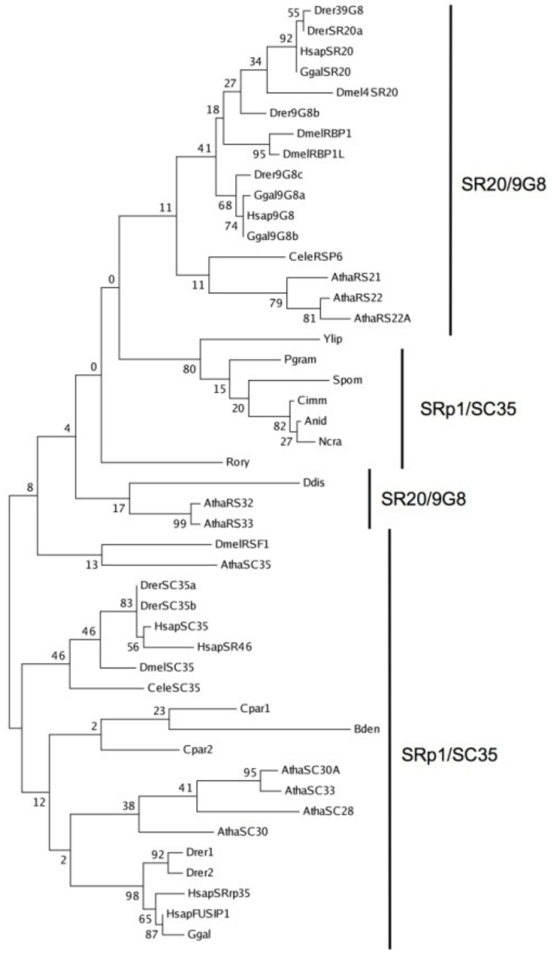


Figure S7. Maximum Parsimony Tree for SC35 and 9G8 homologs. The tree has been built with the RRM from the SC35 (and SRp46) and 9G8 (and SRp20) known proteins and candidate homologs. For metazoans and plants we use the protein name when available. For fungi and protists we just use the short form of the species name. We label the groups in the tree according to their proximity with known proteins.

Regarding the other known SR-protein families: RY1, SRm300, TopoI-B and p54/SRp86, we cannot detect any of these outside metazoans and plants. In particular, our search based on HMMs yielded homologues for all of them in vertebrates and *D. melanogaster*. On the other hand, only P54/SRp86, SRm300 and RY1 are also present in *C. elegans* and *A. thaliana*.

We could not find homologs for any of them in protists or fungi. As RY1, SRm300 and TopoI-B proteins have no RRMs, we initially built the HMM using the alignment of the complete protein. We subsequently refined the search using a shorter more conserved region of the protein to build the HMM. This second refined search did not yield any homolog either in fungi or protists. The search in metazoans resulted in additional proteins: one SRm300 homolog for *G. gallus* (ENSGALP00000002982), one RY1 homolog in *D. rerio* (ENSDARG00000035625) and one p54 homolog in *A. thaliana* (AT3G23900-TAIR-G), which were not found in previous genome-wide searches [15].

The p54/SRp86 proteins, which contain one single RRM in human, present an interesting evolutionary pattern. There are p54/SRp86 homologs in *C. elegans*, *D. melanogaster*, *D. rerio* and *A. thaliana* with two RRMs instead of one. One of the RRMs maintain a strong similarity to the RRM in either the human SRp86 (*D. rerio*) or the human p54 *(C. elegans, D. melanogaster,* *A. thaliana*). The other RRM is more divergent and has no similarity with any other known RRMs.

RS domain definition and RS domain repeat analysis

The RS domain can be easily identified in metazoans as a region of high percentage of RS repeats. We observe that the C-terminal region of SR protein homologs in fungi and protists is also arginine-rich, but they cannot be labeled as RS domains as they have very few SR/RS repeats. Accordingly, we defined the analogous RS domain (or region of interest) for each protein as the C-terminal region outside the RRM. To calculate the percentage of repeats we therefore scanned this region with a sliding window covering 30 amino acids and counted the occurrences of RX or XR, where X = S, D, E and G. For each window we calculated the density of repeats as the number of repeats found over the maximum number of possible repeats. Finally, the reported density for each protein is then the maximum obtained using the sliding window. This value reports the maximum density of a given type of repeat achieved in the C-terminal of a protein and therefore indicates the potential to have a given function associated to that type of repeat. With this calculation we also allow for the same protein to have regions of high density for different types of repeats.

We found that whereas all metazoan SRP2 homologs show a high content of RS repeats, this is much lower in fungi in general

(Figure S8A). Moreover, for SRP1 the trend is not as strong as for SRP2 (Figure S8A). For the TRA2 homologs we measured the percentage of repeats at the N- and C- terminus (Figure S8C). Remarkably, U1-70K homologs show similar densities of RS repeats in all the species, whereas U2AF35 and U2AF65 homologs in fungi tend to have lower densities of RS repeats (Figure S8D).

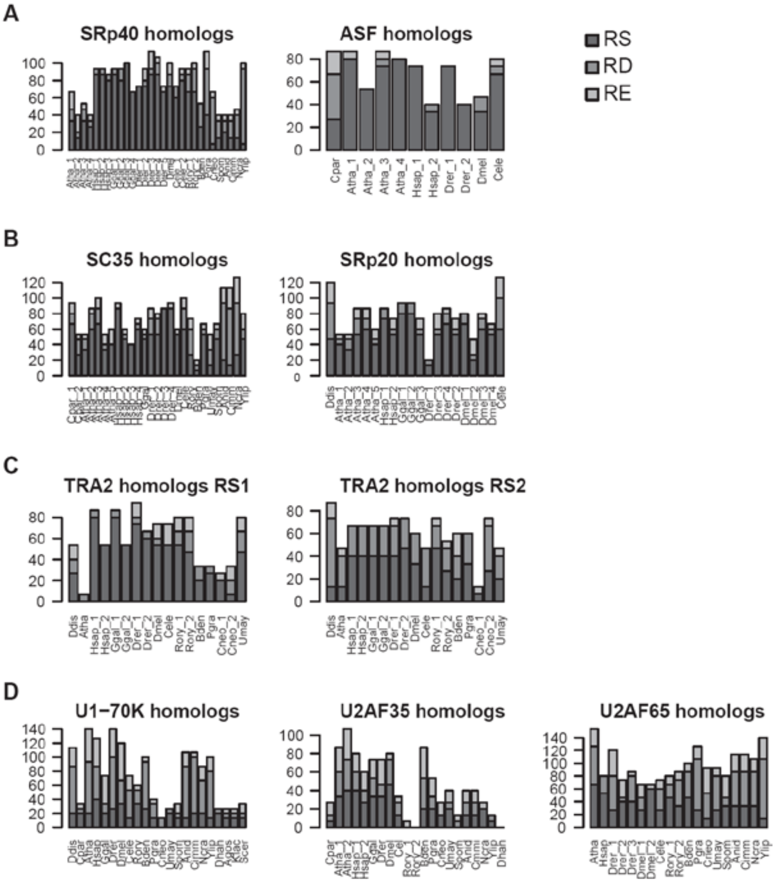


Figure S8. Repeat composition at the C-terminus of SR and SR-related proteins in different species. (A) Composition of repeats of SR proteins containing two RRM domains. **(B)** Composition of the RS domain for SR proteins containing one RRM domain. **(C)** Repeat composition of the N-terminus (RS1) and C-terminus (RS2) of the SR-related protein TRA2. **(D)** Repeat composition for U1-70K, U2AF36 and U2AF65.

U1 and U2 snRNA sequence search

We searched for the U1 snRNA and U2 snRNA sequences in all analyzed species. We compiled a set known U1 and U2 snRNAs from Rfam [28]. For the species analyzed we downloaded a representative U1 snRNA and U2 snRNA, either *seed* or *full member*, if available. In the case of the *full members*, we selected the one having the highest score. For those species without an U2 snRNA in Rfam, we looked for them in the genomic sequence using a multi-step approach: we initially performed a BLAST search [21] using a set of known U2 snRNA sequences: *H. sapiens* (M19204, X59360), *S. cerevisiae* (M14625), *S. pombe* (M23361) and *C. neoformans* (AE017345.1). For each of the target species, we kept the best 3 hits for each query and extracted the sequence from each candidate plus 200 nt on each side. These sequences were then analyzed with the *cmsearch* program from the Infernal package using the covariance model (CM) built for the U2 snRNA *seed members* in Rfam [28]. Because the CM model takes into account both sequence and secondary structure information, we manually inspected all hits to select those having both sequence and secondary structure conservation (Figure S9).

To identify U1 snRNAs, we followed the same approach as for the U2 snRNAs, but we extracted 400 on each side of the BLAST hit. We searched using the known U1 snRNAs from *H. sapiens* (V00591), *D. melanogaster* (K00787), *S. pombe*, (m29062), *A.*

thaliana (AY222070) and *S. cerevisiae* (M17205). This search only allowed us to identify the *B. dendrobatidis* U1 snRNA homolog. We therefore scanned the entire genome of the other species for the sequence that bind the 5′ss in the U1 snRNA (ACTTACC) allowing up to one mutation in the sequence. For each hit, we extracted 100 nt upstream and 800 nt downstream from the sequence and ran the *cmsearch* program. With this approach we were able to identify the U1 snRNA sequences for *A. nidulans*, *C. immitits*, *C. parvum*, *D. discoideum*, *D. hansenii*, *N. crassa*, *P. graminis*, *R. oryzae*, and *Y. lipolytica* (Figure S10 and S11). However, we could not find a candidate for *C. neoformans* and *U. maydis*. For these two species we repeated the analysis allowing up to two mutations in the 5′ss binding site, and included other three *Cryptococcus* species (*C. neoformans* strain CH99, *C. gattii* strain R265, and *C. gattii* strain WM276) and another basidiomycete, *Coprinus cinereus*. In none of these species we were able to find a good U1 snRNA candidate. Thus, these species may have an U1 snRNA with a secondary structure divergent from the known ones that cannot be found using this approach.

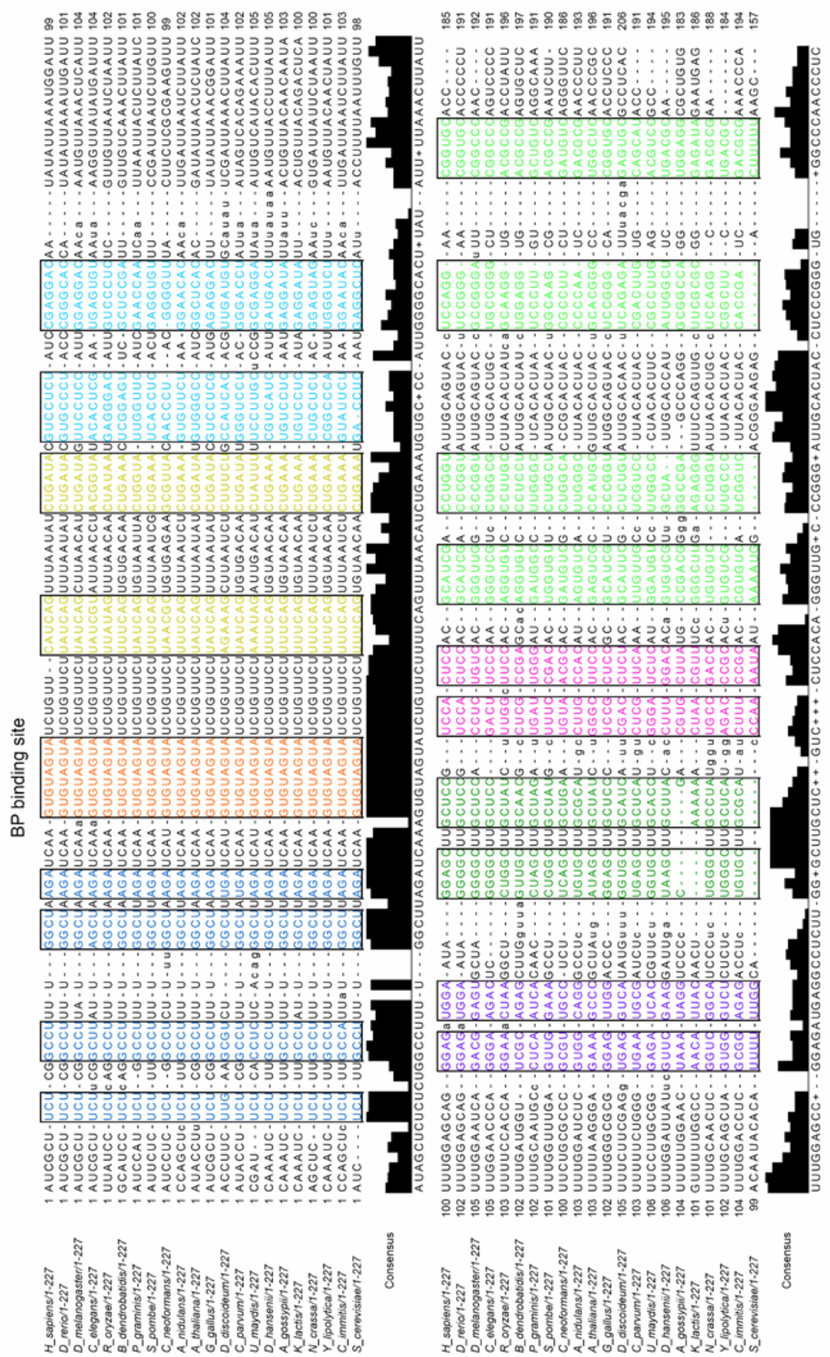


Figure S9. Multiple sequence alignment of U1 snRNA homologs. The alignment was produced with the *cmalig* program. The boxed areas of the same color show the complementary regions in the secondary structure.

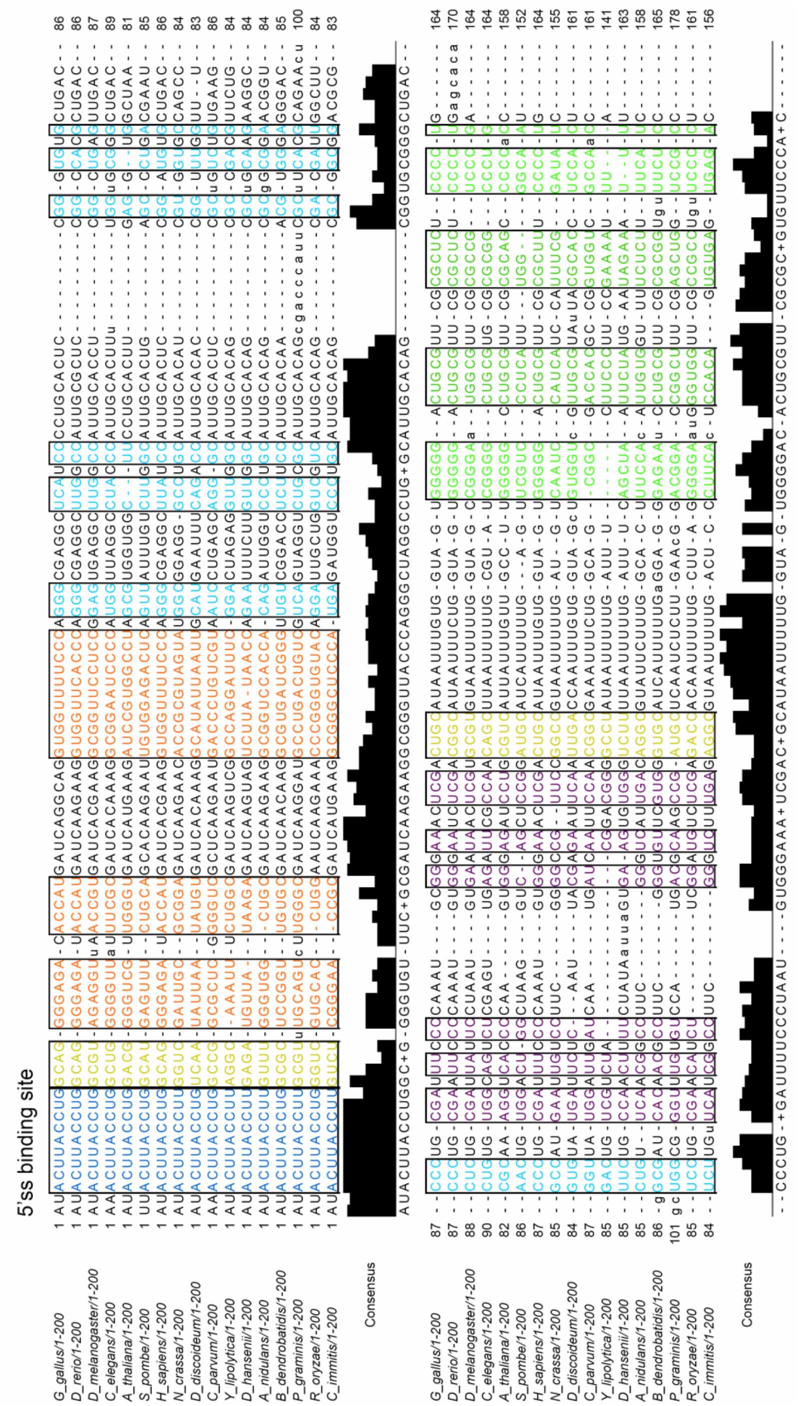


Figure S10. Multiple sequence alignment of U1 snRNA homologs. The alignment was produced with the *cmalig* program. The boxed areas of the same color show the complementary regions in the secondary structure.

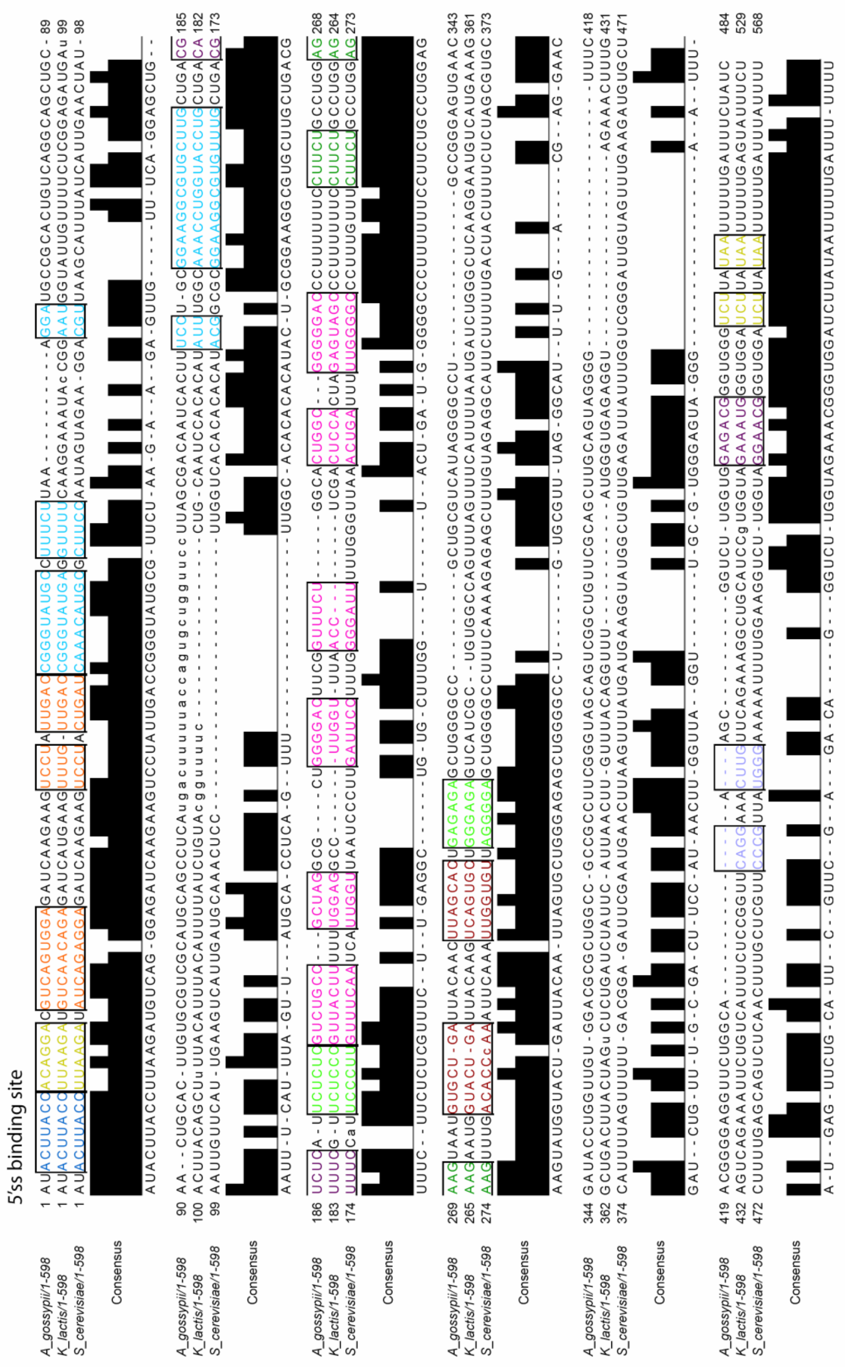


Figure S11. Multiple sequence alignment of U1 snRNA yeast like homologs. The alignment was produced with the *cmalig* program. The boxed areas of the same color show the complementary regions in the secondary structure.

Branch site signal conservation and relationship with U2 snRNA

All the found copies of U2 snRNA have an identical binding site to the BS: GUGUAGUA (Figure S9). Using this sequence, we defined a set of putative BS for each of the species. For each species we selected all introns with canonical splice sites (GT/AG) and without ambiguous nucleotides in the sequence. Furthermore, for those species for which alternative splicing information was available, we selected only constitutive introns (introns flanked by constitutive splice sites). This was done to avoid confusing the BS with other signals possibly involved in the regulation of different splicing events. In order to locate putative BSs, we looked for 9-mers in the 100 nt upstream of the 3′ splice site with the minimum number of nucleotide differences with respect to the motif having the canonical base-pairing to the U2 snRNA (TACTAACAC). If several putative signals with the minimal amount of mutations were identified, we kept the one closest to the 3′ss. If the intron was shorter than 100 nt, we scanned the entire intron. These motifs were defined with an invariable A at position 6 and up to 5 mutations relative to canonical BS motif.

The resulting BS sequences are highly degenerated in most of the species except in some of the fungi. To quantify the variability of the signals, we calculated the total relative entropy between the real signals and the motifs found in a randomized sequence set. The total relative entropy, $H(P,Q)$, is calculated from the probability distribution of nucleotides, $a = 1, 2, 3, 4$, at each position, $i = 1, \dots, N$, in the real signal, $P(X_{a,i})$, relative to the distribution in a randomized set, $Q(X_{a,i})$:

$$H(P,Q) = \sum_{i=1}^{N} \sum_{a=1}^{4} P\left(X_{a,i}\right) log \frac{P(X_{a,i})}{Q(X_{a,i})}$$

Random introns were produced in the following way: we extracted the 100 nt upstream of the acceptor site for each intron of the original dataset to define the random sequences. If the intron was shorter than 100 nt, we took all the intron sequence. Then, we shuffled the nucleotide sequence and look for the BS using the protocol described above. This procedure was repeated 10 times.

The median values of the relative entropy for most of the species are close to zero, showing that the putative BSs are nearly identical to the random ones. Only the ascomycetes show highly conserved BS.

For each predicted BS, we calculated the binding energy to the U2 snRNA. We used the program RNAcofold from the Vienna RNA package (http://www.tbi.univie.ac.at/~ivo/RNA/) [29]. This program calculates the free energy of the base pairing between two RNA sequences. To calculate the energy, we forced a complete pairing of all the nucleotides in one sequence with the corresponding nucleotide in the other sequence. In the case of the BS, we also forced the program to not pair the A from the branch site with any nucleotide from the other sequence. The energy of the base pairing depends on the complementarity between both sequences, the length of the sequence, and the sequence composition. If the energy is high (negative but close to zero), the base pairing is very unstable, because the complementarity of the sequences is poor. Conversely, if the energy is very negative, the base pairing is much more stable. In the defined set of species we see that there is a clear difference between the species from the different groups analyzed. Higher eukaryotes and protozoa show very unstable BSs (*G. gallus* median BS energy is -0.2), with energy distributions similar to the random ones. In contrast, the ascomycetes show more stable BS, specially the saccharomycetes (*K. lactis* median BS energy is -3.3). The other fungal species, *R. oryzae*, *B. dendrobatidis* and the basidiomycetes, have intermediate values between these two groups (*B. dendrobatidis* median energy = -0.5; *N. crassa* = -2.2) (Figure S12A). These results are consistent with our analysis showing that those species that have highly conserved BS also have BSs with the most stable base pairing with U2 snRNA.

5′ss signal conservation and relationship with U1 snRNA

For each of the analyzed species we extracted the 5′ss signals flanking the introns from the defined dataset and obtained the weight matrix models for them. Furthermore, we measured the relative entropy values for the real sites compared to a set of random signals. To define random 5′ ss we scanned the genomic sequences of each of the species (both strands) and kept all the sequences matching “NNNGTNNNN” where N can be A, T, G or C. From this set, we took a random sample of the same size as the set of real 5′ ss. This procedure was repeated 10 times. We observed that those species having highly conserved BS also have highly conserved 5′ss. Nevertheless, 5′ss signals are more conserved than the BS signals.

To measure the stability of the base pairing of 5′ ss signals with the U1 snRNA, we searched U1 snRNA sequence homologs in the set of species to analyze. In contrast to what happens with U2 snRNA, the sequence and the structure of the U1 snRNA is more divergent. There are two types of U1 snRNA: a short one, with an average length of 161 nt, which is the common form in metazoans and most of fungi (Figure S10), and a yeast specific U1 snRNA with an average length of 553 nt, which is the common form in the Saccharomycetaceae order (Figure S11). Our results support the observation that these two types of U1 snRNA only share some parts of the secondary structure that is common to all U1 snRNAs [28,30]. Interestingly, *Y. lipolytica* has an U1 snRNA similar to the metazoan one but not to the Saccharomycetaceae species.

We calculated the binding energies for U1 snRNA base paring with the 5′ss as described before. Contrary to what we expected, those species having more degenerate 5′ ss tend to have more stable base pairing with the U1 snRNA than those having highly conserved signals (Figure S12B). However, we see that there is not a clear relationship between these two variables. As explained before, the free energy values of the base pairing between two sequences depend directly on the sequence content. Specifically, pairing between A and T has a lower free energy than pairing between C and G. Thus, we wanted to elucidate if the observed energetic differences between different species were influenced by a sequence bias (i.e. some organisms having an AT enrichment) in either the 5'ss signal or the U1 snRNA sequence. Our results show that the observed energetic differences among different species are not due to a bias in 5'ss sequences because energies for the pairing between the U1 snRNA and the 5'ss signals from the random datasets are similar across the analyzed species (Figure S12B). Considering the existence of sequence variations in the region of the U1 snRNA that binds to the 5′ss, we standardized the observed energetic values for the pairing of the 5'ss and the U1 snRNA in the analyzed organisms. For each species, free energy values were scaled between 0 and 1, corresponding to the minimal and the theoretical maximal free energies respectively. We found no differences between the real energetic distributions and the standardized ones (data not shown). This demonstrates that the observed energetic differences could not be due to variations in the U1 snRNA sequences.

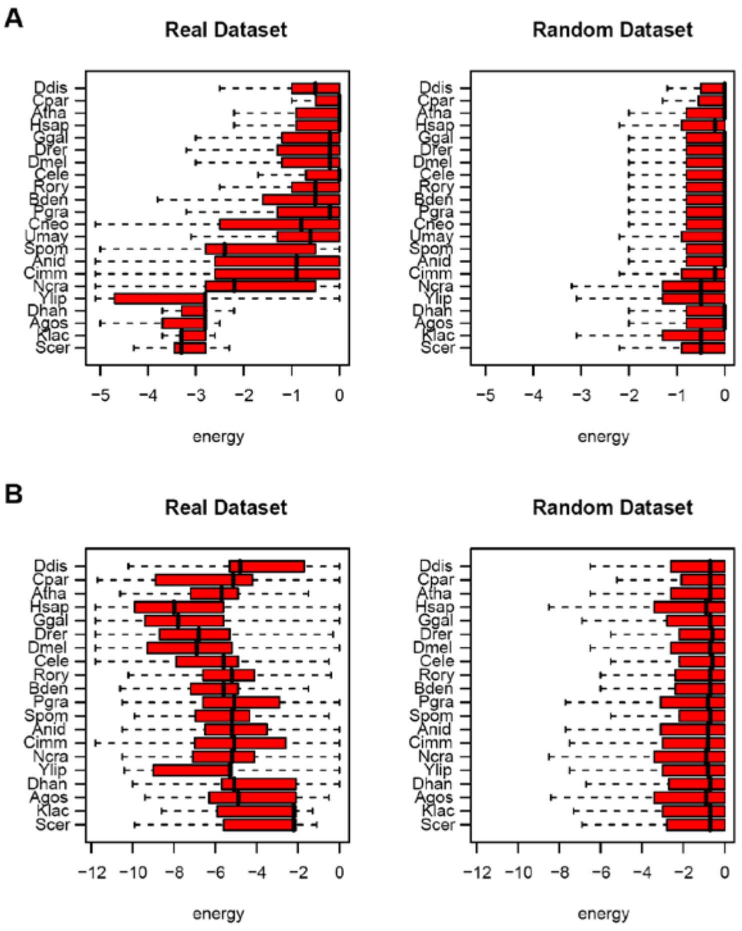


Figure S12. Distribution of Energies. Distribution of the energy of binding for branch site **(A)** and 5′ splice site **(B)** signals. The energy values on the X axis are minimum free energy values expressed in kcal/mol. The free energy values for the BS and 5′ signals (left) are compared with a random set (right).

The polypyrimidine tract (PPT)

An in depth comparative analysis of PPTs was carried out recently [27]. In this work a "PPT enrichment index" was calculated as the ratio between the strength in real introns with respect to randomized introns. All organisms showed a significant bias for having PPTs at the 3′ end of introns. The strongest bias was found for metazoans, whereas most fungi had very weak PPT bias. Two exceptions were *S. cerevisiae* and *K. lactis*, which showed stronger bias than the other fungi. Plants and Protozoa had intermediate PPTs. In the same work it was also found that the strength of the PPT correlated to changes in key residues in the splicing factor recognizing it, U2AF. Since we cover a similar range of species as in Schwartz et al. 2008 [27], we expect that a similar analysis would yield identical results.

Using an alternative approach for the PPT analysis [31-33], we searched for PPTs in the set of species analyzed. Similarly to our analysis of BSs, we also considered the prediction of PPTs in a set of randomized sequences. The method from Schwartz et al. 2008 scans only the 50nt from the 3′ end and accepts only predictions ending in the last 10nt. Thus, we can obtain a measure similar to the enrichment index from Schwartz et al. 2008 [27] by considering the ratio of the proportion of introns with PPTs ending within the last 5nt of the introns, which we call STRONG, in the real data set over the same proportion in the random dataset. Figure S13 shows these values. This result recapitulates the behavior observed in [27] for PPTs: a general enrichment (value > 1), with the highest values for metazoans and the lowest for fungi. Among fungi *K. lactis* and *S. cerevisiae* are an exception, as they have higher values than any other fungi. The case of *Y. lipolytica* is singular as it lacks PPTs between the BS and the 3′ ss [27].

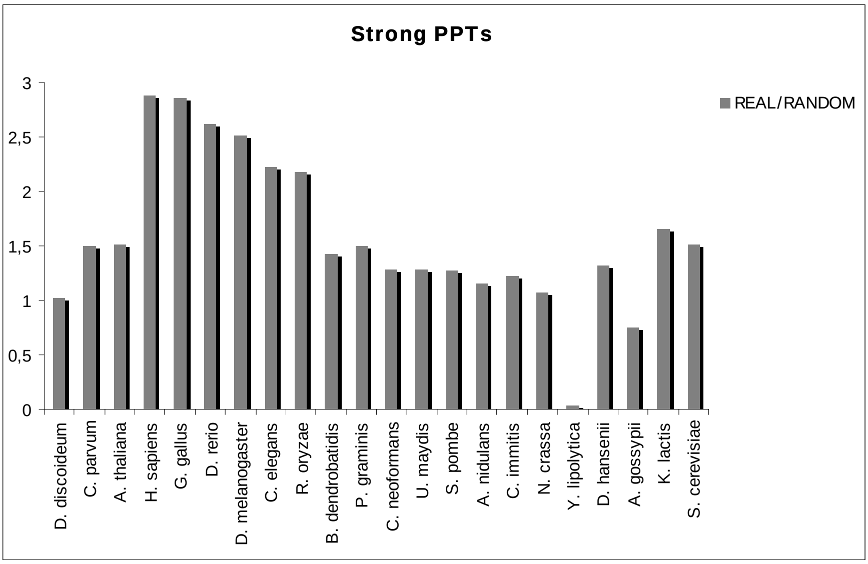


Figure S13. Ratio of the predicted PPTs ending within the 5nt of the introns (STRONG) in real versus randomized introns. For each species we show the value of dividing the fraction of real introns with strong PPTs over the fraction of randomized introns with strong PPTs.

These results and results from Schwartz et al. do not show any clear direct correlation between the SR-proteins and the properties of the PPT signal. However, for the species *R. oryzae*, *B. dendrobatidis* and *P. graminis*, which were not included in Schwartz et al. 2008, we find a bias similar to metazoans. Interestingly, these are also the fungi that have more SR protein homologues and higher content of RS repeats than any other fungi.

It is known that fungal species tend to have PPTs upstream of the BS [34]. There are singular cases like *Y. lipolytica*, which lacks any PPT downstream of the BS but has a C-rich signal upstream of the BS [27]. We have carried out analysis of the predicted PPTs according to their position relative to the predicted BS (using our method). For each set of introns (real and random) we labeled the predicted PPTs as upstream if they started before the predicted BS for that intron or downstream if they started after the BS. For each species we calculated the ratio between the number of real and random PPTs for the upstream and downstream cases (Figure S14). For the randomized introns, all species have approximately the same proportion of upstream PPTs (58,8% on average) and downstream PPTs (41.2% on average).

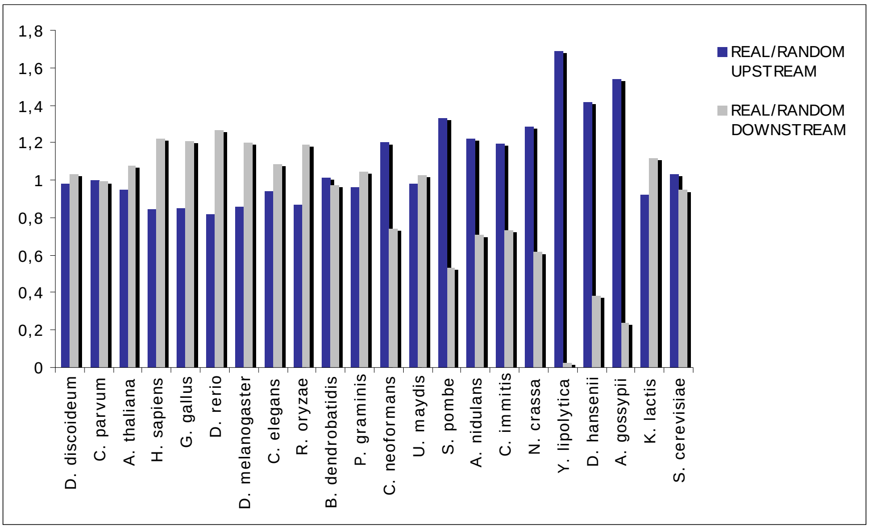


Figure S14. Ratio of the number of PPTs in real versus randomized introns, separated by the relative position with respect to the branch site (BS): upstream or downstream. A value of 1 means that randomized introns have on average the same number upstream (or downstream) introns than in the real introns. This ratio accounts for the differences in sequence background between species. A bar well above or below 1 represents a bias towards a overrepresentation or underrepresentation, respectively, of PPTs in the corresponding side of the BS.

Metazoans show a trend towards downstream PPTs and fungi have a general trend to have more upstream PPTs. Interestingly, *R. oryzae* shows a metazoan-like behaviour, and *B. dendrobatidis* and *P. graminis* show no positional bias. The rest of fungi show a strong bias towards PPTs upstream of the BS, except for *K. lactis* and *S. cerevisiae,* which seem to also have a balance between upstream and downstream no different from random. Our analysis also reproduces the singular case of *Y. lipolytica* mentioned above. Finally, protozoa and plants show no positional bias like *B. dendrobatidis* and *P. graminis*.

Density of Exonic Splicing Enhancers

Both modes of action described for SR proteins require their binding to Exonic Splicing Enhancer motifs (ESE). To measure the potential to bind SR proteins, we calculated the coverage of different sets of predicted ESEs [35-37] in the 50bp next to each splice sites. We discarded all those exons which length was smaller than 50 nt. Intriguingly, we found for all species a similar distribution of densities for the ESE set [37] (Figure S15), with the only exception of *D. discoideum*, which shows lower density of binding sites. Using the set of binding sites obtained from SELEX experiments [35,36] we found that all SR protein binding sites appear at similar levels in all species (Figure S15). Interestingly, the fungal species with no SR protein homologs have a density of binding sites similar to those with SR proteins. Moreover, the percentage of exons for which we mapped each subset of motifs is also very similar across species (Table S3). Thus, assuming that the binding specificity for SR proteins has remained similar across evolution, they would have the potential to bind to exons from all the species analyzed. Since we have found evidence that SR proteins are ancestral to eukaryotes, this analysis will show that they had the potential to bind in exons whether or not they were actually involved in splicing in the ancestral eukaryote.

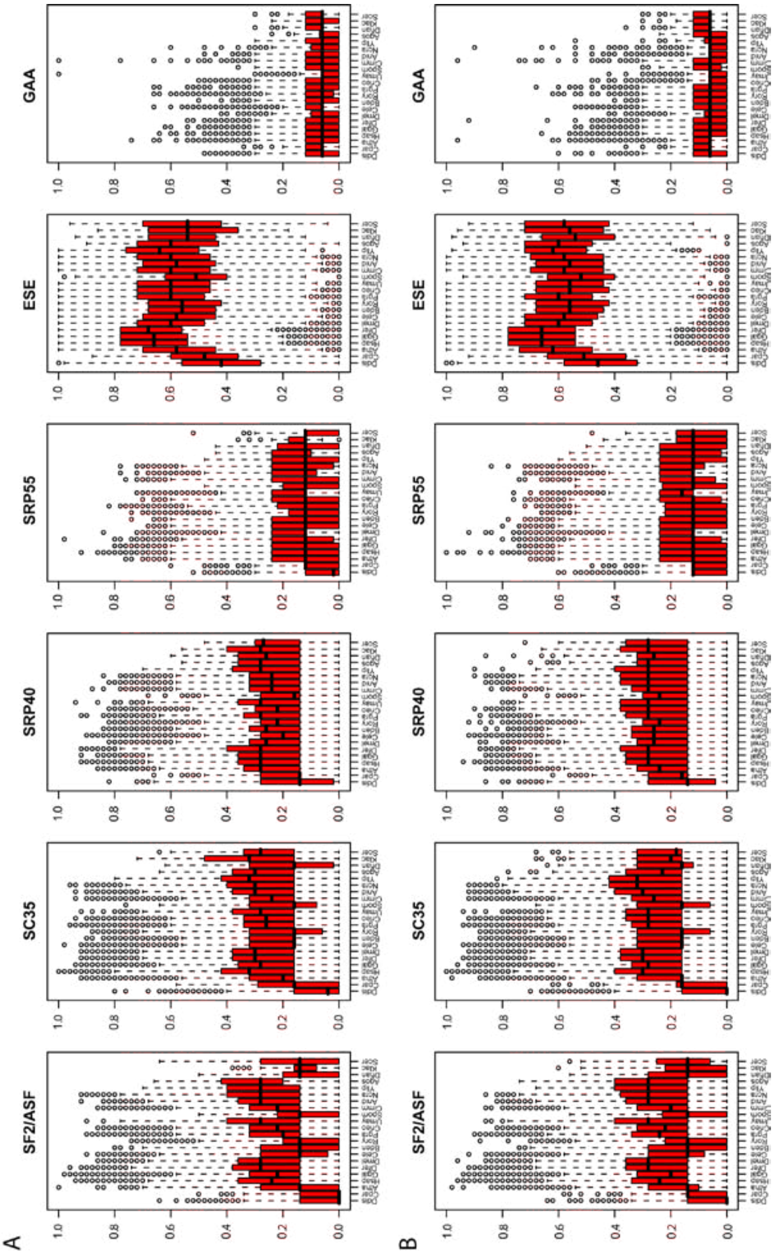


Figure S15. Density of Exonic Splicing Enhancers in different species. The plot shows the distributions per species of the density of ESEs in exons, calculated over the 50bp next to the 5′ splice site **(A)** and the 3′ splice site **(B)** discarding the 3bp next to the splice site. We show the coverage for a set of SR protein binding sites obtained from SELEX experiments (SF2/ASF, SC35, SRP40, SRP55) [36], the set of predicted ESEs from [37], and for the motif GAA, which is known to bind TRA2 and some SR proteins [35].

Table S3. Percentage of exons for which we have mapped ESEs.

	DONOR						ACCEPTOR					
Species	SF2/ASF	C35	SRP40	SRP55	ESE	GAA	SF2/ASF	SC35	SRP40	SRP55	ESE	GAA
D. discoideum	30.06	50.93	75.04	50.14	96.94	67.46	34.06	48.47	76.07	53.88	98.28	73.22
C. parvum	47.76	68.87	82.06	62.53	99.74	78.10	54.46	69.55	86.88	58.42	99.01	76.98
A. thaliana	74.94	84.49	89.47	70.23	99.76	70.48	77.99	82.57	88.02	70.64	99.85	77.67
H. sapiens	85.52	88.64	90.67	73.06	99.94	66.77	84.86	88.33	90.45	72.46	99.94	70.29
G. gallus	84.50	86.82	90.67	73.91	99.95	70.58	84.03	86.95	90.53	73.24	99.95	73.49
D. rerio	88.88	89.18	92.56	75.50	99.95	65.50	88.70	89.12	91.08	75.64	99.95	67.80
D. melangaster	88.54	88.98	88.84	79.97	99.84	61.69	88.55	89.23	88.50	80.53	99.88	61.47
C. elegans	76.00	80.29	85.94	69.45	99.60	77.26	77.14	82.13	88.68	72.46	99.66	74.02
R. oryzae	72.06	81.49	88.31	72.81	99.77	67.27	72.72	81.20	89.24	74.24	99.79	66.52
B. dendrobatidis	65.43	76.90	86.86	65.94	99.58	75.37	66.46	76.64	87.45	68.25	99.59	72.04
P. graminis	84.52	86.65	89.61	67.96	99.81	69.15	85.98	86.81	89.97	68.85	99.78	67.43
C. neoformans	85.89	87.14	86.82	70.31	99.78	69.34	85.31	87.46	91.30	75.18	99.73	58.74
U. maydis	90.24	88.10	90.24	82.24	99.71	57.81	91.14	88.49	90.70	83.13	99.67	54.56
S. pombe	61.86	78.81	85.31	68.08	98.02	74.01	70.93	77.19	82.46	68.92	99.75	75.69
A. nidulans	84.50	85.93	87.96	73.84	99.77	71.39	84.18	86.30	89.51	75.91	99.76	67.49
C. immitis	89.31	89.68	88.31	77.86	99.76	63.99	89.33	89.90	90.96	80.02	99.73	60.64
N. crassa	90.78	89.38	86.79	75.57	99.81	60.09	90.12	90.50	90.97	77.60	99.76	55.33
Y. lipolytica	91.26	90.98	91.53	74.04	100.00	66.94	90.15	91.62	91.03	74.56	100.00	63.24
D. hansenii	98.31	94.92	86.44	77.97	98.31	69.49	89.08	91.95	90.23	75.86	99.43	70.69
A. gossypii	70.68	75.19	88.72	66.92	100.00	78.20	70.91	78.79	89.39	68.48	99.70	78.79
K. lactis	75.68	94.59	97.30	78.38	100.00	72.97	74.40	84.00	94.40	68.00	100.00	76.80
S. cerevisiae	66.28	81.40	87.21	62.79	100.00	81.40	77.90	84.64	89.14	61.42	100.00	76.40

References

1. Hibbett, D. S., Binder, M., Bischoff, J. F., Blackwell, M., Cannon, P. F., Eriksson, O. E., Huhndorf, S., James, T., Kirk, P. M., Lucking, R., et al. 2007. A higher-level phylogenetic classification of the Fungi. *Mycol. Res.* **111:** 509-547.

2. Parra, G., Blanco, E., and Guigo, R. 2000. GeneID in Drosophila. *Genome Res.* **10:** 511-515.

3. Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., et al. 2008. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36:** D1009-14.

4. Rogers, A., Antoshechkin, I., Bieri, T., Blasiar, D., Bastiani, C., Canaran, P., Chan, J., Chen, W. J., Davis, P., Fernandes, J., et al. 2008. WormBase 2007. *Nucleic Acids Res.* **36:** D612-7.

5. Heiges, M., Wang, H., Robinson, E., Aurrecoechea, C., Gao, X., Kaluskar, N., Rhodes, P., Wang, S., He, C. Z., Su, Y., et al. 2006. CryptoDB: a Cryptosporidium bioinformatics resource update. *Nucleic Acids Res.* **34:** D419-22.

6. Sherman, D., Durrens, P., Iragne, F., Beyne, E., Nikolski, M., and Souciet, J. L. 2006. Genolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts. *Nucleic Acids Res.* **34:** D432-5.

7. Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., et al. 2008. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36:** D13-21.

8. Wilson, R. J., Goodman, J. L., Strelets, V. B., and FlyBase Consortium. 2008. FlyBase: integration and improvements to query tools. *Nucleic Acids Res.* **36:** D588-93.

9. Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., et al. 2008. Ensembl 2008. *Nucleic Acids Res.* **36:** D707-14.

10. Black, D. L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72:** 291-336.

11. Shen, H. and Green, M. R. 2006. RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. *Genes Dev.* **20:** 1755-1765.

12. Shen, H. and Green, M. R. 2004. A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. *Mol. Cell* **16:** 363-373.

13. Shen, H., Kan, J. L., and Green, M. R. 2004. Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Mol. Cell* **13:** 367-376.

14. Izquierdo, J. M. and Valcarcel, J. 2006. A simple principle to explain the evolution of pre-mRNA splicing. *Genes Dev.* **20:** 1679-1684.

15. Barbosa-Morais, N. L., Carmo-Fonseca, M., and Aparicio, S. 2006. Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res.* **16:** 66-77.

16. Eddy, S. R. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755-763.

17. Birney, E., Kumar, S., and Krainer, A. R. 1993. Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Res.* **21:** 5803-5816.

18. Bourgeois, C. F., Lejeune, F., and Stevenin, J. 2004. Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA. *Prog. Nucleic Acid Res. Mol. Biol.* **78:** 37-88.

19. Slater, G. S. and Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31.

20. Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and Genomewise. *Genome Res.* **14:** 988-995.

21. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403-410.

22. Tamura, K., Dudley, J., Nei, M., and Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24:** 1596-1599.

23. Gross, T., Richert, K., Mierke, C., Lutzelberger, M., and Kaufer, N. F. 1998. Identification and characterization of srp1, a gene of fission yeast encoding a RNA binding domain and a RS domain typical of SR splicing factors. *Nucleic Acids Res.* **26:** 505-511.

24. Webb, C. J., Romfo, C. M., van Heeckeren, W. J., and Wise, J. A. 2005. Exonic splicing enhancers in fission yeast: functional conservation demonstrates an early evolutionary origin. *Genes Dev.* **19:** 242-254.

25. Lutzelberger, M., Gross, T., and Kaufer, N. F. 1999. Srp2, an SR protein family member of fission yeast: in vivo characterization of its modular domains. *Nucleic Acids Res.* **27:** 2618-2626.

26. Loftus, B. J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I. J., Fraser, J. A., et al. 2005. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science* **307:** 1321-1324.

27. Schwartz, S. H., Silva, J., Burstein, D., Pupko, T., Eyras, E., and Ast, G. 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res.* **18:** 88-103.

28. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. 2005. Rfam: annotating non-

coding RNAs in complete genomes. *Nucleic Acids Res.* **33:** D121-4.

29. Hofacker, I. L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31:** 3429-3431.

30. Kretzner, L., Krol, A., and Rosbash, M. 1990. Saccharomyces cerevisiae U1 small nuclear RNA secondary structure contains both universal and yeast-specific domains. *Proc. Natl. Acad. Sci. U. S. A.* **87:** 851-855.

31. Coolidge, C. J., Seely, R. J., and Patton, J. G. 1997. Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res.* **25:** 888-896.

32. Clark, F. and Thanaraj, T. A. 2002. Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.* **11:** 451-464.

33. Gooding, C., Clark, F., Wollerton, M. C., Grellscheid, S. N., Groom, H., and Smith, C. W. 2006. A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. *Genome Biol.* **7:** R1.

34. Kupfer, D. M., Drabenstot, S. D., Buchanan, K. L., Lai, H., Zhu, H., Dyer, D. W., Roe, B. A., and Murphy, J. W. 2004. Introns and splicing elements of five diverse fungi. *Eukaryot. Cell.* **3:** 1088-1100.

35. Tacke, R., Tohyama, M., Ogawa, S., and Manley, J. L. 1998. Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing. *Cell* **93:** 139-148.

36. Liu, H. X., Zhang, M., and Krainer, A. R. 1998. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev.* **12:** 1998-2012.

37. Stadler, M. B., Shomron, N., Yeo, G. W., Schneider, A., Xiao, X., and Burge, C. B. 2006. Inference of splicing regulatory activities by sequence neighborhood analysis. *PLoS Genet.* **2:** e191.

RNA secondary structure regulates 3′ splice site selection in yeast

Mireya Plass, Pedro G. Ferreira,
and Eduardo Eyras

2.1. Manuscript

RNA secondary structure regulates 3′ splice site selection in yeast

Mireya Plass[1], Pedro G. Ferreira[2], and Eduardo Eyras[3]

[1] Computational Genomics, Universitat Pompeu Fabra, 08003, Barcelona, Spain
[3] Centre de Regulació Genòmica, Dr. Aiguader 88, 08003 Barcelona, Spain
[4] Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain (ICREA).
Corresponding author: eduardo.eyras@upf.edu

ABSTRACT

Alternative splicing is the mechanism by which different combinations of exons in the pre-mRNA give raise to several mature mRNAs. In higher eukaryotes, this process is usually regulated by protein factors, which bind to the pre-mRNA and affect the recognition of splicing signals. Yeast species lack many of the regulatory splicing factors present in metazoans, and therefore, it was considered until recently that they did not have any alternative splicing. In this work, we show that RNA secondary structure is important for 3′ splice site (3′ss) recognition in yeast through a process that is conserved across yeast species. Moreover, using the properties of pre-mRNA sequences and information about the secondary structure they can adopt, we have built a model that correctly predicts over 90% of 3′ss in yeast. This method also allows predicting

alternative 3′ss. These alternative 3′ss rather than increasing the protein repertoire in yeast, they introduce premature termination codons (PTCs), which would trigger the degradation of these mRNAs by non-sense mediated decay (NMD).

INTRODUCTION

Splicing is the mechanism by which introns are removed from the pre-mRNA to create the mature transcripts. In higher eukaryotes this process involves, apart from the core machinery of the spliceosome, many auxiliary factors, e.g. SR proteins or hnRNPs, which can enhance or block the recognition of splicing signals (reviewed in Jurica and Moore, 2003). These factors allow the modulation of the splicing reaction and thus, the existence of alternative splicing. In contrast to what happens in higher eukaryotes, yeast species do not have as many auxiliary factors (Schwartz et al., 2008; Plass et al., 2008). This reduces the number of regulatory mechanisms and makes splicing to be more dependent on *cis* acting elements. In the case of *Saccharomyces cerevisiae*, the rules for 5′ splice site (5′ss) and branch site (BS) recognition are well understood (reviewed in Madhani and Guthrie, 1994). In contrast, there is still controversy about the mechanisms involved in 3' splice site (3′ss) recognition. Yeast lacks the U2AF heterodimer, which is crucial for 3′ss recognition in higher eukaryotes (Wu et al., 1999); hence, in theory, any CAG, TAG, or AAG (HAGs) placed at a right distance from the branch site (BS) could function as a 3′ss. A scanning mechanism from the BS onwards has been proposed for 3′ss selection (Smith et al., 1993), although not always the first AG downstream of the BS is used. Additionally, several *cis* acting factors have been found to influence 3′ss selection in yeast. For instance, a U-rich tract can allow the usage of a more distant AG (Patterson and Guthrie 1991), the 5′ss sequence can guide 3′ss choice (Goguel and Rosbash, 1993), the exonic sequence just after the AG can help identify the true splice site (Crotti and Horowitz, 2009), and distance from the BS also constitutes a critical element (Cellini et al., 1986; Luukkonen and Seraphin, 1997). However, how 3′ss selection works is still not fully understood.

One of the possible elements that can allow regulation of 3′ss selection is the secondary structure adopted by the nascent pre-mRNA. Previously, it has been shown that the structure adopted by pre-mRNAs can affect splice site recognition in human (Shepard and Hertel, 2008). In yeast, RNA secondary structure has been shown to influence 5′ss recognition by shortening the 5′ss-BS distance (Rogic et al., 2008) or by sequestering splice site signals (Deshler and Rossi, 1991; Goguel et al., 1993). Interestingly, the structures adopted by the pre-mRNA can be subject to modulation, as changes in transcription rate of the RNA polymerase or temperature can affect their formation and stability (Pan and Sosnick, 2006; Mahen et al., 2010; Chen, 2008) and, consequently, could regulate 3′ss selection in yeast.

The correct identification of splice sites is one of the most challenging aspects of gene annotation. Current wealth of genomic data provides an opportunity for integration by applying statistical learning methods in order to extract patterns and thereby new biological insight and testable hypotheses. These computational methodologies, generally called Machine Learning (ML) methods (Mitchell, 1997), work by reading input data in order to estimate relationships or probabilities for a number of observables, thereby learning properties that characterize the classification values. ML methods have been employed in a variety of biological problems to build predictive models (Larranaga et al., 2006). For instance, a recent application of an ML algorithm has proved successful to predict the splicing properties of exons from a large number of sequence features (Barash et al., 2010). A class of generally used ML methods are Support Vector Machines (SVM). SVMs are supervised learning algorithms that, given a dataset represented as a set of features and a binary classification (i.e positive and negative cases), find the combination of features that provides an optimal separation between the instances of the two classes (see e.g Ben-Hur et al., 2008). SVMs are widely used in computational biology and have been shown to achieve high accuracy in a variety of problems, including the prediction of splice sites (Yamamura and Gotoh, 2003; Sun et al., 2003; Zhang et al., 2003; Sonnenburg et al., 2007) and alternative exons (Dror et al., 2005).

In this work, we investigate the role of RNA secondary structure in 3′ss selection in yeast and integrate sequence and secondary structure information of the pre-mRNA to generate a model to predict 3′ss signals. Our results show that RNA structure plays a crucial role in 3′ss selection in yeast, and that this mechanism is conserved across yeast species. Moreover, using the SVM model, we show that 3′ss selection is more dynamic than previously thought and suggest that RNA secondary structure is not only relevant for the recognition of constitutive 3′ss but also plays a role in the regulation of mRNA expression, as it allows the usage of alternative 3′ss that would trigger NMD.

MATERIALS AND METHODS

Datasets

Details on the datasets used can be found in Supplementary Material.

Branch Site Prediction

To predict branch sites, introns were scanned for NNNTRACNN motifs up to 200 nt upstream from the 3′ss, and those with the smallest Hamming distance to the TACTRACNN sequence were predicted as BS. When several motifs with identical Hamming distances were found, an additional selection based on potential base pairing to U2 snRNA was applied using RNAcofold

(Hofacker, 2009), forcing an unpaired branch site A. If several motifs had the same potential, the one closer to the 3′ss was selected.

Secondary structure prediction

For each intron, we recovered the sequence between the BS and the 3′ss, discarding both signals. From this region, we further removed the first eight nucleotides after the BS A, as previous experiments show that these nucleotides cannot be part of a secondary structure (data not shown). In the selected region, we predicted a putative secondary structure using the program RNAfold from the Vienna package (Hofacker, 2009) with default parameters.

Effective distance measure

We defined the distance between the BS and any 3′ss as the number of nucleotides between the A of the BS and the 3′ss, including only the latter. Using this definition, TACTAACACNNNN|TAG would give a distance of 10 nt. We defined the effective BS-3′ss distance as the linear distance (in nucleotides) between the BS and the 3′ss after removing the secondary structure. More specifically, we removed all the bases that were part of a structured region, and the 2 bases corresponding to the beginning and the end of the structured region were added substituting each structured region.

Accessibility measurement

Accessibility is defined as the probability of a nucleotide not being base-paired with any other nucleotide, i.e. one minus the pair probability. We calculated pair probabilities using the program RNAfold (Hofacker, 2009). The accessibility A_k of a HAG at position k is calculated as the average of the accessibilities of each of the nucleotides a in the HAG, $a(w,i)$, $j = k,\ k+1,\ k+2,$ in four different windows $w$of lengths $d,\ d+5,\ d+10,\ d+15,$ where d is the BS-HAG distance:

$$A_k = \frac{1}{4}\frac{1}{3}\sum_{w}\sum_{i=k}^{k+2} a(w,i)$$

Support Vector Machine classifier

To identify candidate alternative 3′ss we built an SVM with a linear kernel using the program Gist2.3 (Pavlidis et al., 2004) (http://svm.sdsc.edu). In this case we wanted to classify HAGs into positive (functional 3′ss) and negatives (non-functional 3′ss). We considered the positive set to be all the annotated 3′ss in our dataset (282), and the negative set, all the HAGs labeled as intronic (97) or exonic (11527) (see Supplementary Material), and whose effective distance was smaller than 52 nt, as this is the maximum BS-3′ss effective distance for a 3′ss to be recognized (Meyer et al., unpublished). We used the intronic and exonic HAGs as a negative set as we expect that the majority of them would be true negatives, as there is no evidence of their usage. However, we could expect that a small subset of them might be possible alternative 3′ss. In our approach these would be false positives. Thus we reasoned, as explained below, that using the predictive model at a very low false positive rate (FPR), would allow us to obtain the small fraction of cases that would be candidate alternative 3′ss.

In order to avoid a biased training due to the unbalanced nature of the training datasets, a total of 10000 SVM models were calculated, sampling randomly for each one 200 positive and 200 negative cases. Next, each of the SVM models was used to score all other HAGs not used for training (11506) and classifying them as functional or non-functional 3′ss according to their score, using zero as a cutoff value. Using this approach, each HAG was classified as positive or negative approximately 10000 times. Since the scores of the SVM models cannot be compared between experiments, to combine the SVM models, we defined a score (*score 1*) for each HAG to be the proportion of SVM models (out of the 10000) for which the HAG was classified as positive, using a cutoff value of zero. Additionally, for each HAG we defined a second score (*score2*), which was defined as the proportion of models in which a HAG was classified as positive, but at a fixed FPR. That is, for each of the

10000 SVMs, we produced two classifications: one with zero cutoff value, to produce the *score1*; and a second one, for which the cutoff value was set such that only 0.5% of the non-annotated HAGs were classified as functional 3′ss, i.e. only 0.5% of false positives were allowed per SVM model, to produce *score2*. This second scoring scheme for HAGs, *score2*, ensures that the classification was made at a fixed FPR of 0.5%.

The list of features selected to build the Support Vectors of each of the 3′ss analyzed are detailed in the Supplementary Material.

Analysis of *S. cerevisiae* RNA-Seq reads

We used RNA-seq data from two studies (Yassour et al., 2009, Nagalakshmi et al., 2008) (Supplementary Table 2). Reads were mapped against *S. cerevisiae* genome (SGD July 2009) (Engel et al., 2010) using GEM (http://gemlibrary.sourceforge.net), allowing 2 mismatches and with default parameters. Reads that did not map to the genome were then used to find candidate splice junctions using GEM split-mapper (http://gemlibrary.sourceforge.net). This tool tests all possible mapping combinations by splitting the reads into two parts. In this case, for reads of length 36, the split-point ranges between 10 and 27. Moreover, the consensus motifs GT-AG and GC-AG for the splice-site dinucleotides were provided to GEM split-mapper to narrow down the search space of the mapping. Additionally, the split-mapping was done using a maximum of 1 mismatch position in the read sequence. Only reads with 1 split-mapping (uniquely split-mapped) were selected. The last two columns of Supplementary Table 2 provide the number of all (unique + non-unique) and unique split-mapped reads obtained for each RNA-Seq dataset. We clustered reads according to the split-map positions (start and end of the putative intron) and calculated for each cluster the total number of reads and the number of non-redundant sequences (reads with different sequence).

Blast search of alternative splicing products

We obtained the mRNA sequences and the protein sequences, when applicable, resulting from the usage of the predicted alternative 3′ss. For all the predicted proteins (alternative 3′ss in the coding region), we looked for homologous sequences in the non-redundant protein database (nr) from RefSeq (Pruitt et al., 2007) using blastp (Altschul et al., 1997), imposing that the variable region was part of an alignment. In the cases in which we predicted alternative 3′ss in non-coding genes (snRN17A and snRN17B), we looked for homologous sequences in nr database using blastn (Altschul et al., 1997).

RESULTS

Yeast introns contain few intronic HAGs

We built a dataset of canonical introns from *S. cerevisiae* and for each of them we predicted the BS signal (see Supplementary Material). In this set, composed of 282 introns, there are 44 introns containing a total of 100 intronic HAGs that could be functional (Supplementary Material). This number is significantly lower than the 226 cases that would be expected by chance (chi-square test p-value = 1.71e-17), indicating a selective pressure against HAGs upstream of annotated 3′ss.

RNA structures shorten BS-3′ss distance

We predicted RNA secondary structures between the BS and the 3′ss for each intron in the dataset (see Materials and Methods). These predictions and further analyses described are accessible at http://regulatorygenomics.upf.edu/Software/Yeast_Introns/. We found that 113 introns from the set (40%) contain a structure in this region. Interestingly, all those introns with a BS-3′ss distance larger than 45 nt included a structure (Figure 1a). Interestingly, when computing the effective BS-3′ss distance (see Materials and Methods), the resulting length distribution is not significantly different from the one of introns without structures (Wilcoxon signed-rank test p-value = 3.4e-01; Figure 1a). As a consequence, the effective BS-3′ss distance is never larger than 45 nt. This result suggests that 45 nt may be the maximum distance for efficient splicing in *S. cerevisiae*. Similar results are observed in the other yeast species analyzed (see Supplementary Material and Supplementary Figure 1). Importantly as well, while random sequences also show this behavior, their effective distance is significantly longer (see Supplementary Material and Supplementary Figure 2).

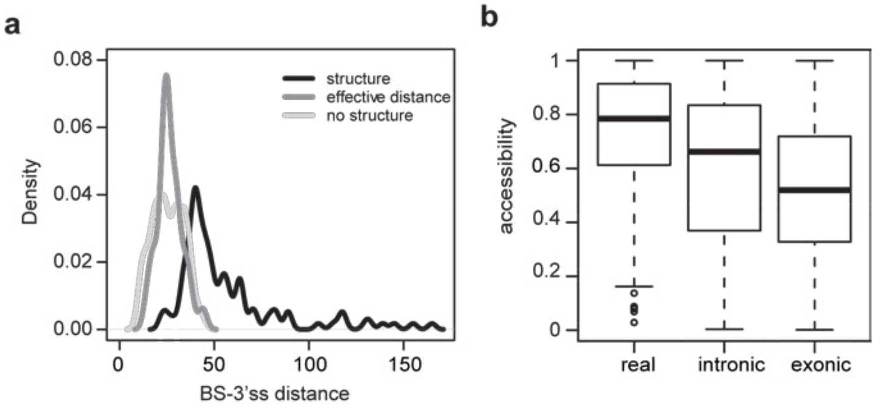


Figure 1. Position and accessibility of 3' splice sites. (a) Distribution of *S. cerevisiae* intron lengths. Introns are separated in two categories, those that contain a secondary structure in this region (black); and those that do not (light gray). The effective distance of introns containing a secondary structure is also shown (dark gray line). **(b)** Boxplot diagram showing accessibility values for annotated and cryptic (intronic and exonic) 3′ss. Dashed lines indicate the value distribution between the maximum and minimum (thin horizontal lines). Boxes include 50% of the values. Thick lines indicate median values and outliers are shown as open circles.

RNA secondary structure blocks the recognition of cryptic 3′ss

To evaluate the possibility that RNA folds occlude cryptic 3′ss from the spliceosome, we calculated the accessibility of all intronic, exonic and annotated HAGs (see Materials and Methods and Supplementary Material). Our data show that accessibility values are significantly higher for real than for cryptic 3′ss (Figure 1b). Significantly, real 3′ss are predicted to be more accessible than intronic HAGs, which is not expected in random sequences (Wilcoxon signed-rank test *p*-value real vs intronic = 5.50e-05; real *vs* exonic = 2.2e-16). These results, replicated in other yeast species analyzed (Supplementary Figure 3), are consistent with a role of RNA structures in favouring the selection of real 3′ss.

Alternative splicing prediction using a Support Vector Machine (SVM)

The previous results suggest that RNA structure may be important for understanding 3′ss selection in yeast. Therefore, we built an SVM classifier using as positive set all HAGs annotated as real 3′ss, and as negatives all cryptic 3′ss, i.e. all non-annotated intronic and exonic HAGs (see Supplementary Material). The sequence features considered for the classification were the splice site sequence, the pyrimidine content between the BS and the 3′ss, and the distance to the polypyrimidine tract (PPT). Additionally, we considered the accessibility of the candidate 3′ss, which is related to the secondary structure of the pre-mRNA. We also considered the effective distance between the BS and the HAG as a filter, as we have recently shown that there is a maximum effective distance beyond which the HAG is never used as 3′ss (Meyer et al., unpublished). Using the scoring schema *score1* (see Materials and Methods), the overall performance of the SVM classifier is very good (AUC = 0.981; Figure 2a and b).

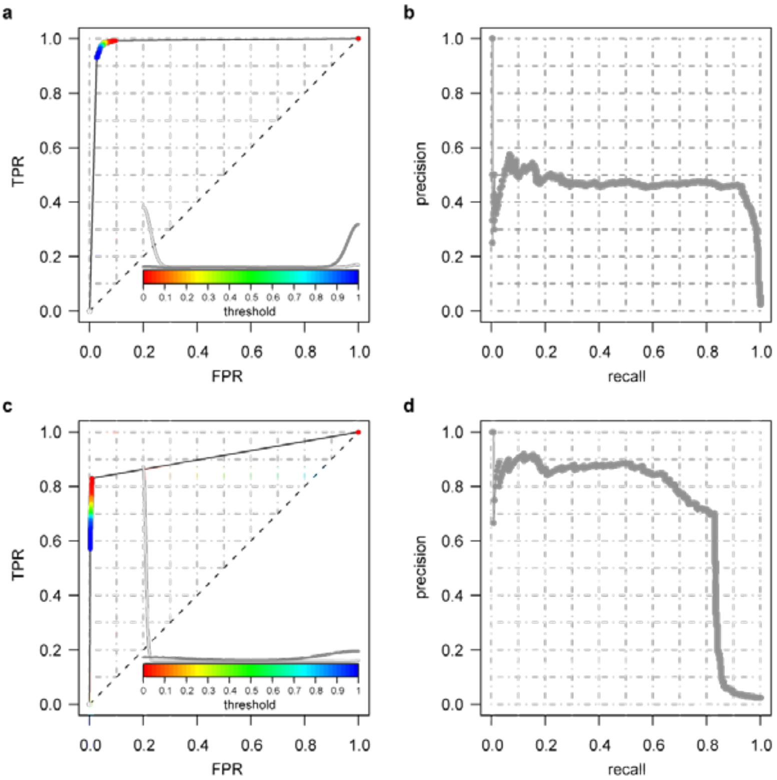


Figure 2. Evaluation of SVM classifier scoring methods. Receiving Operating Characteristic (ROC) curves of the SVM classifier using *score1* **(a)** or *score2* **(c)**. For each threshold of the score, the x-axis represents the true positive rate (TPR), i.e the proportion of positive cases that are correctly predicted; and the y-axis represents the false positive rate (FPR), i.e. the proportion of negative cases that are predicted as positive. The distribution of values for positive cases (dark grey) and negative cases (light grey) together with the color scale for the different thresholds used can be seen at the bottom of the figure. Precision-recall curves of the SVM classifier using *score1* **(b)** or *score2* **(d)**. Here the x-axis represents the recall or TPR and the y-axis represents the precision, i.e. the proportion of predicted cases that are actually positive for a given threshold.

Moreover, using a threshold of 1, i.e. selecting only HAGs that were classified as positive by all SVM models, our method is able to predict correctly 93% of real 3'ss (263/282) with only 2% of false positives (311/11624). On the other hand, the positive predictive value (PPV) of the method, i.e. the proportion of true positives from those predicted as positive, is 0.45 (Figure 2b), since we obtain more false positives than true positives.

We considered the problem of predicting alternative 3'ss as the classification as positive of cryptic sites originally labeled as negative. In this classification problem, we gave greater relevance to the false positive rate (FPR), rather than to the true positive rate (TPR); as for this problem the objective is not to recover as many annotated 3'ss as possible, but to select potential new 3'ss with high specificity. With scoring scheme *score1* we were predicting only a small percentage of negative cases as positives (2% of false positives). However, we decided to obtain a more conservative estimate and change the scoring schema of the SVM models so that we could get a set of predictions with smaller FPR and a higher PPV. With this purpose, we designed the scoring schema *score2* (see Materials and Methods). Using *score2*, the overall performance of the SVM classifier is slightly worse than using *score1* (AUC = 0.9122), but we get a better separation of positive and negative cases (Figure 1c). We selected a threshold of 0.983 for *score2*, i.e. we selected only those HAGs that were classified HAGs as positive by 98.3% of the 10000 SVM models at a FPR of 0.5%. With this threshold we obtained a high PPV (0.83), keeping a reasonable amount of true positives (TPR = 0.64) (Figure 1d) and predicting only a small fraction of cryptic HAGs (34 cases; FPR = 0.0029) as positives. These HAGs represent the subset that is most similar to the set of annotated ones and thus, we considered them as candidate alternative 3'ss (Table 1).

Validation of predicted alternative 3'ss

We used reads from two published RNA-Seq experiments (Yassour et al., 2009; Nagalakshmi et al., 2008) to validate the candidate alternative 3'ss predicted with the SVM (Table 2). We mapped the reads to the yeast genome allowing for mappings across splice-junctions using GT/AG and GC/AG as possible splice sites (see Materials and Methods). Interestingly, we found a direct relation between SVM *score2* and the proportion of cases validated by RNA-Seq reads (Figure 3a and b). In the case of non-annotated HAGs, the percentage of cases that can be validated at any SVM score cut-off is much lower than for real 3'ss (Figure 3b). Additionally, when we considered the *score2* threshold to be 0.983, i.e. we considered only those non-annotated HAGs that are candidate alternative 3'ss, 7 out of the 34 predicted cases (~20%) are validated by RNA-Seq reads (Table1). This represents more than a 50-fold enrichment over all HAGs predicted as negative (43 cases validated with RNA-Seq reads, i.e. 0.4% of all negative cases).

Table 1. Alternative 3'ss candidates

AG name	Gene Name	AG Type	SVM Score	Number of reads	Splicing evidence
chrII:366501-366582:-:YBR062C_28	-	I-1	1	1	reads
chrII:592412-592763:-:YBR181C_41	RPS26B	E1	1	0	NO
chrII:60190-60693:-:YBL087C_48	RPL23A	E1	1	0	NO
chrIII:107034-107110:+:YCL005W-A_47	VMA9	E1	1	0	NO
chrIII:107034-107110:+:YCL005W-A_54[‡]	VMA9	E2	0.984	0	NO
chrIV:1103808-1103890:+:YDR318W_31[†‡]	MCM21	E1	0.998	0	NO
chrIV:1236836-1237601:+:YDR381W_42	YRA1	E2	1	0	NO
chrIV:399360-399482:+:YDL029W_25	ARP2	E1	1	1	reads
chrIV:431385-431470:-:YDL012C_52	-	E1	1	0	NO
chrIV:629904-630171:+:YDR092W_38	UBC13	I-7	1	1	reads
chrIV:65308-65378:+:YDL219W_46	DTD1	E1	1	0	NO
chrIX:47699-47760:+:YIL156W-B_21	-	E	1	4	reads
chrV:184169-184676:-:YER014C-A_17	BUD25	I-8	1	0	NO
chrV:396807-397277:+:YER117W_56[†]	RPL23B	E1	1	0	EST
chrVI:221256-221402:-:YFR031C-A_39	RPL2A	E1	0.999	0	NO
chrVI:64599-64919:-:YFL034C-A_24	RPL22B	E1	1	0	NO*
chrVII:31427-31578:-:YGL251C_29	HFM1	I-4	1	0	NO*
chrVII:555835-556311:+:YGR034W_45	RPL26B	E1	0.999	0	NO
chrVIII:251158-251250:+:YHR076W_27	PTC7	E1	1	0	NO
chrX:580340-581044:+:YJR079W_21	-	I-1	1	0	NO
chrXI:158622-158971:+:YKL156W_40	RPS27A	E1	1	0	NO
chrXI:83004-83079:+:YKL190W_25	CNB1	E1	1	0	NO
chrXII:550461-550576:-:YLR202C_32	-	I-3	1	0	NO
chrXII:564457-564515:-:YLR211C_40[†‡]	-	E1	0.983	0	NO
chrXII:786616-786712:+:YLR329W_16	REC102	I-1	1	0	NO
chrXII:819331-819777:+:YLR344W_36	RPL26A	E1	1	0	NO
chrXIV:185493-185587:+:YNL246W_14	VPS75	I-3	1	0	NO
chrXIV:185493-185587:+:YNL246W_26	VPS75	I-1	1	2	NO
chrXIV:557612-557685:+:YNL038W_27	GPI15	E1	1	1	reads
chrXIV:622947-623288:+:YNL004W_32	HRB1	E2	1	0	NO*
chrXV:242441-242503:-:YOL047C_20	-	E1	1	0	NO
chrXV:780122-780278:+:snR17a_15[‡]	SNR17A	E1	1	0	NO
chrXVI:138725-138863:+:YPL218W_22	SAR1	E1	1	1	reads
chrXVI:281373-281502:-:snR17b_15[‡]	SNR17B	E1	1	0	NO
chrVII:1084890-1085037:+:YGR296W_42[T]	YFR1-3	E1	0.994	0	NO
chrXIV:5932-6079:-:YNL339C_42[T]	YFR1-6	E1	0.995	0	NO
chrXVI:5841-5988:-:YPL283C_42[T]	YFR1-7	E1	0.995	0	NO

[‡] cases that do not introduce PTC

[†] cases predicted as 3'ss only at 37ºC

[T] cases predicted as 3'ss only at 22ºC

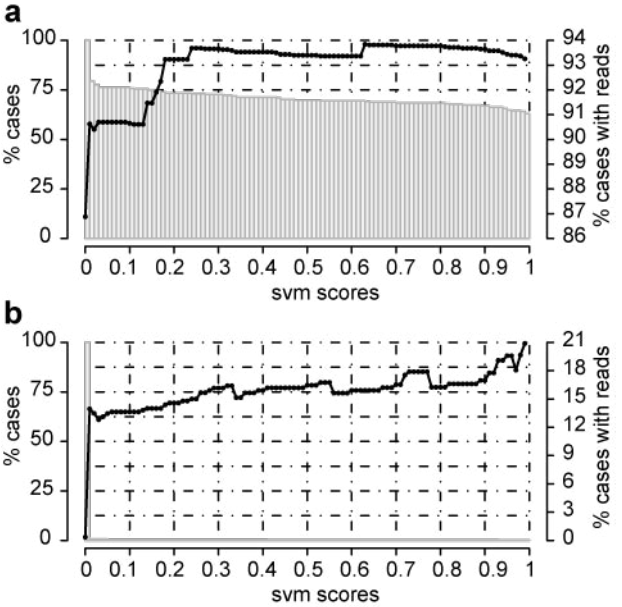


Figure 3. Cumulative distribution of HAGs that are validated by RNA-Seq reads, for annotated 3′ss **(a)** and for cryptic 3′ss **(b)**. On the x-axis the values of *score2* threshold used as cut-off are shown. The left y-axis represents the percentage of HAGs that have a *score2* higher or equal to that given on the x-axis (grey bars). The right y-axis represents the percentage of cases with a *score2* higher or equal to that given on the x-axis and that can be validated using RNA-Seq reads (black line).

Identification of constitutive and alternative 3′ss in 5′UTR regions

There are very few known 5′UTR introns in Yeast (Engel et al., 2010). SGD only contains 24 5′UTR annotations in coding genes, all of them containing 5′UTR introns. These introns, which we did not use for training, represent an independent set on which to validate the SVM model. For each of these introns we predicted the BS, located all intronic, exonic, and annotated HAGs as before, and extracted the features described previously

(Supplementary Material). Applying the SVM classifier using *score1* we can correctly predict 91% of the real 3′ss (22/24) and predict ~3% of false positives (29/1084). Using *score2* and the threshold defined above for predicting alternative 3′ss (0.983), we were able to correctly predict 17 out of 24 (71%) known 3′ss, and only 4 of the cryptic ones (0.4%) as positives. Interestingly, 3 out of these 4 possible alternative 3′ss have reads supporting them (Table 3). This represents a ~74 fold enrichment compared to all cryptic 3′ss in 5′UTR regions, as only 11 of the 1084 negative cases have RNA-Seq evidence (~1%). In one particular case, corresponding to an intronic HAG in *RPS22B* gene, we found 78 reads validating the alternative 3′ss, suggesting that some of the predicted cases can have an impact on mRNA regulation. These results confirm that our SVM model is able to distinguish real from false 3′ss and that can be used to predict new alternative 3′ss.

Table 2. Alternative 3′ss candidates from annotated 5′UTR introns

AG name	Gene Name	AG Type	SVM Score	Number of reads	Splicing evidence
chrXI:166405-166492:+:YKL150W_10	MCR1	I-1	0.9968	1	reads
chrXI:166405-166492:+:YKL150W_28 †	MCR1	E1	0.9961	0	NO
chrXI:93317-93470:-:YKL186C_13	MTR2	I-3	1	2	reads
chrXII:855877-856433:+:YLR367W_15	RPS22B	I-1	1	78	reads

† cases predicted as 3′ss only at 37ºC

Effects of temperature on 3′ss selection

We have shown that the secondary structure adopted by the pre-mRNA affects 3′ss selection and can be used to predict alternative 3′ss. However, the impact of these structures on splice site selection can change if the structures are altered. One of the elements affecting secondary structures is the temperature. Thus, as our previous predictions were made at 37ºC, we analyzed the possible impact of temperature change on 3′ss selection by checking the properties of all HAGs at 22ºC. Our results show that, in the case of annotated 3′ss, the maximum effective distance found is the same at both conditions even though the effective length distributions differ (Wilcoxon signed rank test p-value < 0.001; Figure 4a). As expected, at this temperature the accessibility of HAGs from all categories is lower when compared to the predictions at 37ºC (statistical significant differences of accessibility values at 37ºC compared to 22ºC for real 3′ss, Wilcoxon signed-rank test p-value = 4.086e-05, and for exonic HAGs, Wilcoxon signed rank test p-value < 2.2e-16. No significant differences were found for intronic HAGs) (Figure 4b). This suggests that the ability of the spliceosome to recognize 3′ss may be dependent on the temperature, i.e. temperature specific splicing events. To test this hypothesis we rebuilt the SVM classifier using the properties of HAGs at 22ºC. In this case, we selected a *score2* threshold of 0.994, such as we would obtain the same FPR as before (FPR =

0.0029). Using this threshold, 31 of the 34 alternative 3′ss predictions in coding regions (Table 1), and 3 of the 4 alternative 3′ss in 5′UTR regions (Table 2) are shared in the two conditions, showing that the effect of temperature on 3′ss selection is not very strong in the selection of alternative 3′ss.

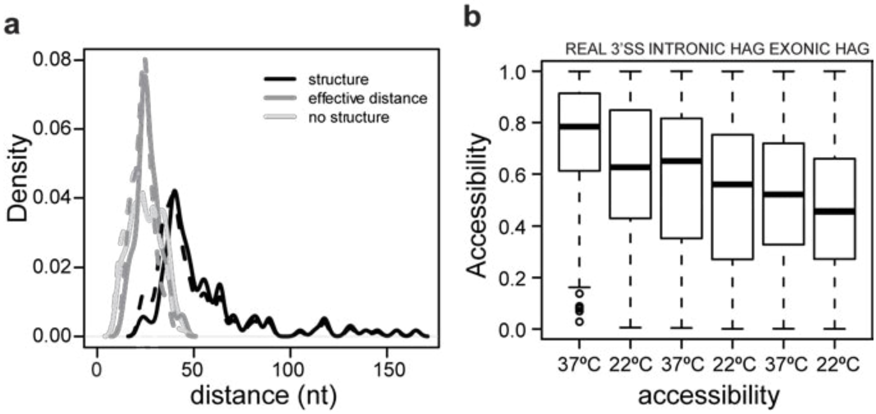


Figure 4. (a) Comparison of BS-3′ss length distribution at 37ºC (continuous lines) and 22ºC (dashed lines). The plot shows the length distribution of BS-3′ss without a secondary structure (light grey line) and for those with a predicted secondary structure (black line). In the cases in which a secondary structure was predicted, the distribution of the effective distances is also shown (grey line). **(b)** Box plots representing accessibility values distribution at 37ºC and 22ºC for real 3′ss, intronic HAGs and exonic HAGs. Accessibility values are shown on the y-axis, which vary between 0 (always covered by a secondary structure) and 1 (never covered by a secondary structure).

Function of the predicted alternative 3′ss

We checked whether the usage of the alternative 3′ss predicted would introduce PTCs that would trigger the degradation of the resulting transcripts by NMD. In fact, 32 of the 35 candidate alternative 3′ss in coding regions introduce a PTC (Table 1). In these cases, the 3′UTR is enlarged on average by 578 nt. Hence, the PTCs will possibly trigger NMD. This hypothesis is also supported by the fact that we could not find homologs of the mRNAs or translated products resulting from the usage of the candidate alternative 3′ss (see Supplementary Material). In the 3 cases in which no PTC is introduced (Table 1), the alternative 3′ss introduce a deletion in the final protein that does not appear in the homologous proteins and corresponds to a conserved region, therefore suggesting an alteration or lost of functionality of the proteins containing the deletion (see Supplementary Material and Supplementary Figures 5, 6 and 7).

Discussion and Conclusions

The analyses described allow to establish a general role of RNA secondary structure in 3′ss selection. We have shown that the RNA secondary structure adopted by nascent pre-mRNAs have a crucial role in 3'ss selection and is more widespread than previously thought. We have seen that for over a third of yeast introns the RNA secondary structure modulates splicing selection. On the one hand, RNA structures regulate the accessibility of a given 3'ss to the spliceosome (Figure 1 and 4), as we see that accessibility is higher in real 3′ss compared to cryptic 3′ss. On the other hand, RNA structures ensure a correct distance between the BS and the 3'ss (Figure 1 and 4). Distant 3'ss have always a secondary structure that brings them closer to the BS. We have predicted this to occur at distances of at least

45 nt in naturally occurring introns. This maximum effective distance is significantly lower than what has been reported in other studies, both *in vitro* in HeLa extracts (Smith et al., 1993) or *in vivo* in yeast (Cellini et al., 1986), and compared to the random distribution.

Furthermore, we have used the properties derived from the RNA structure to build a computational method, based on a support vector machine (SVM) model to predict 3′ss in yeast. Using this method we are able to correctly recover 93% of real 3′ss with only 2% of false positives. Moreover, we can apply the SVM model to an independent dataset of 5′UTR introns obtaining similar TPR and FPR (TPR = 0.91; FPR = 0.26). The data used for training the SVM model includes only information extracted from the sequence in the BS-3′ss region and the downstream exon, and the secondary structure that the sequence can adopt. Thus, our results suggest that these features may be sufficient to identify real 3′ss in yeast. Accordingly, in the majority of the cases, 3′ss selection may not require the presence of auxiliary factors, which is consistent with the fact that splicing factors that enhance splicing seem to be missing in yeast (Plass et al., 2008; Schwartz et al., 2008), and may rely solely on the sequence surrounding the 3′ss and its secondary structure. Interestingly, as the structural properties of the sequence surrounding the 3′ss can change with temperature (Chen, 2008), these changes have an impact in the selection of 3′ss. Indeed, the comparison of the properties of HAGs at 22ºC and 37ºC show that the structural properties of the sequence upstream of alternative 3′ss change with temperature (Figure 4b). These changes affect the prediction outcome of ~10% of the alternative 3′ss predicted using *score2*. Additionally, we have shown that the accessibility of 3′ss is higher at 37ºC (Figure 3b), which is in agreement with the fact that at high temperatures secondary structures present lower stabilities. Therefore, high temperatures will facilitate the usage of alternative 3′ss that may be hidden at lower temperatures, allowing the regulation of 3′ss in a temperature dependent manner (Meyer et al., unpublished).

The validation of the alternative 3′ss by RNA-Seq reads shows that the usage of the predicted splice sites is lower than that of the corresponding annotated 3′ss. This low usage of the alternative 3′ss suggests that either they are used at very low levels, e.g. only under very specific conditions, or that they introduce a PTC that would then trigger the degradation of the resulting transcripts possibly by NMD. In fact, 32 of the 35 candidate alternative 3′ss in coding regions introduce a PTC. In these cases, the 3′UTR is enlarged on average by 578nt, which is much longer than the average of 144 nt for the 3′UTR length in yeast genes (Graber et al., 1999). In yeast, NMD is triggered by PTCs that create long 3′UTRs (Amrani et al., 2004). Moreover, it′s also known that the mRNA levels of those genes that have 3′UTR longer than average are regulated by NMD (Kebaara and Atkin, 2009). Thus, our findings suggest a possible role of these alternative 3′ss in the regulation of mRNA levels by NMD. Furthermore, the hypothesis that these alternative 3′ss may produce NMD is also supported by the fact that we could not find homologs of the mRNAs or translated products resulting from the usage of the candidate alternative 3′ss.

Recent papers have shown that NMD coupled to the production of splicing variants to regulate mRNA levels is a regulatory mechanism more extended than previously thought (Baek and Green, 2005; Pan et al., 2008; Sayani et al., 2008). Moreover, it has also been shown that the fact that lots of alternative splicing events trigger NMD caused the underestimation of alternative splicing levels in several species. The results provided in this work are in agreement with these ideas, as we are able to predict previously unknown alternative splicing events that would trigger NMD. We also show that in the case of yeast, the regulation of 3′ss selection can be performed independently of external factors and that the properties of the sequence are enough to define the splicing outcome for the majority of 3′ss.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25,* 3389-3402.

Amrani, N., Ganesan, R., Kervestin, S., Mangus, D.A., Ghosh, S., and Jacobson, A. (2004). A faux 3'-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. Nature *432,* 112-118.

Baek, D., and Green, P. (2005). Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. Proc. Natl. Acad. Sci. U. S. A. *102,* 12813-12818.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465,* 53-59.

Ben-Hur, A., Ong, C.S., Sonnenburg, S., Scholkopf, B., and Ratsch, G. (2008). Support vector machines and kernels for computational biology. PLoS Comput. Biol. *4,* e1000173.

Cellini, A., Felder, E., and Rossi, J.J. (1986). Yeast pre-messenger RNA splicing efficiency depends on critical spacing requirements between the branch point and 3' splice site. EMBO J. *5,* 1023-1030.

Chen, S.J. (2008). RNA folding: conformational statistics, folding kinetics, and ion electrostatics. Annu. Rev. Biophys. *37,* 197-214.

Crotti, L.B., and Horowitz, D.S. (2009). Exon sequences at the splice junctions affect splicing fidelity and alternative splicing. Proc. Natl. Acad. Sci. U. S. A. *106,* 18954-18959.

Deshler, J.O., and Rossi, J.J. (1991). Unexpected point mutations activate cryptic 3' splice sites by perturbing a natural secondary structure within a yeast intron. Genes Dev. *5,* 1252-1263.

Dror, G., Sorek, R., and Shamir, R. (2005). Accurate identification of alternatively spliced exons using support vector machine. Bioinformatics *21,* 897-901.

Engel, S.R., Balakrishnan, R., Binkley, G., Christie, K.R., Costanzo, M.C., Dwight, S.S., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Hong, E.L.*, et al.* (2010). Saccharomyces Genome Database provides mutant phenotype data. Nucleic Acids Res. *38,* D433-6.

Goguel, V., and Rosbash, M. (1993). Splice site choice and splicing efficiency are positively influenced by pre-mRNA intramolecular base pairing in yeast. Cell *72,* 893-901.

Goguel, V., Wang, Y., and Rosbash, M. (1993). Short artificial hairpins sequester splicing signals and inhibit yeast pre-mRNA splicing. Mol. Cell. Biol. *13,* 6841-6848.

Graber, J.H., Cantor, C.R., Mohr, S.C., and Smith, T.F. (1999). Genomic detection of new yeast pre-mRNA 3'-end-processing signals. Nucleic Acids Res. *27,* 888-894.

Hofacker, I.L. (2009). RNA secondary structure analysis using the Vienna RNA package. Curr. Protoc. Bioinformatics *Chapter 12,* Unit12.2.

Jurica, M.S., and Moore, M.J. (2003). Pre-mRNA splicing: awash in a sea of proteins. Mol. Cell *12,* 5-14.

Kebaara, B.W., and Atkin, A.L. (2009). Long 3'-UTRs target wild-type mRNAs for nonsense-mediated mRNA decay in Saccharomyces cerevisiae. Nucleic Acids Res. *37,* 2771-2778.

Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armananzas, R., Santafe, G., Perez, A., and Robles, V. (2006). Machine learning in bioinformatics. Brief Bioinform *7,* 86-112.

Luukkonen, B.G., and Seraphin, B. (1997). The role of branchpoint-3' splice site spacing and interaction between intron terminal nucleotides in 3' splice site selection in Saccharomyces cerevisiae. EMBO J. *16,* 779-792.

Madhani, H.D., and Guthrie, C. (1994). Dynamic RNA-RNA interactions in the spliceosome. Annu. Rev. Genet. *28,* 1-26.

Mahen, E.M., Watson, P.Y., Cottrell, J.W., and Fedor, M.J. (2010). mRNA secondary structures fold sequentially but exchange rapidly in vivo. PLoS Biol. *8,* e1000307.

Mitchell, T.M. (1997). Machine Learning. (The Mc-Graw-Hill Companies,)

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320,* 1344-1349.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40,* 1413-1415.

Pan, T., and Sosnick, T. (2006). RNA folding during transcription. Annu. Rev. Biophys. Biomol. Struct. *35,* 161-175.

Patterson, B., and Guthrie, C. (1991). A U-rich tract enhances usage of an alternative 3' splice site in yeast. Cell *64,* 181-187.

Pavlidis, P., Wapinski, I., and Noble, W.S. (2004). Support vector machine classification on the web. Bioinformatics *20,* 586-587.

Plass, M., Agirre, E., Reyes, D., Camara, F., and Eyras, E. (2008). Co-evolution of the branch site and SR proteins in eukaryotes. Trends Genet. *24,* 590-594.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *35,* D61-5.

Rogic, S., Montpetit, B., Hoos, H.H., Mackworth, A.K., Ouellette, B.F., and Hieter, P. (2008). Correlation between the secondary structure of pre-mRNA introns and the efficiency of splicing in Saccharomyces cerevisiae. BMC Genomics *9,* 355.

Sayani, S., Janis, M., Lee, C.Y., Toesca, I., and Chanfreau, G.F. (2008). Widespread impact of nonsense-mediated mRNA decay on the yeast intronome. Mol. Cell *31,* 360-370.

Schwartz, S.H., Silva, J., Burstein, D., Pupko, T., Eyras, E., and Ast, G. (2008). Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. *18,* 88-103.

Shepard, P.J., and Hertel, K.J. (2008). Conserved RNA secondary structures promote alternative splicing. RNA *14,* 1463-1469.

Smith, C.W., Chu, T.T., and Nadal-Ginard, B. (1993). Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. Mol. Cell. Biol. *13,* 4939-4952.

Sonnenburg, S., Schweikert, G., Philips, P., Behr, J., and Ratsch, G. (2007). Accurate splice site prediction using support vector machines. BMC Bioinformatics *8 Suppl 10,* S7.

Sun, Y.F., Fan, X.D., and Li, Y.D. (2003). Identifying splicing sites in eukaryotic RNA: support vector machine approach. Comput. Biol. Med. *33,* 17-29.

Wu, S., Romfo, C.M., Nilsen, T.W., and Green, M.R. (1999). Functional recognition of the 3' splice site AG by the splicing factor U2AF35. Nature *402,* 832-835.

Yamamura, M., and Gotoh, . (2003). Detection of the splicing sites with Kernel method approaches dealing with nucleotide doublets. Genome Informatics *14,* 426.

Yassour, M., Kaplan, T., Fraser, H.B., Levin, J.Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., Gnirke, A.*, et al.* (2009). Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc. Natl. Acad. Sci. U. S. A. *106,* 3264-3269.

Zhang, X.H., Heller, K.A., Hefter, I., Leslie, C.S., and Chasin, L.A. (2003). Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. Genome Res. *13,* 2637-2650.

2.2. Supplementary Material

Yeast intron datasets

We downloaded the annotation and genomic sequence of *S. cerevisiae* from the Saccharomyces Genome Database (SGD July 2009) (Engel et al., 2010). We then extracted all introns from chromosomal genes (327) and kept only those that had length > 0 nt, canonical splice sites (GT or GC at the 5′ss and AG at the 3′ss) and did not have any ambiguous nucleotide (N) in the sequence, obtaining a final set of 282 introns.

We used Galaxy (Goecks et al., 2010) to extract the homologous regions to the *S.* cerevisae introns in 5 *Saccharomyces* species (*S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, and *S. castellii*). From these, we extracted the genomic alignments for the yeast species provided by UCSC (Fujita et al., 2011) and kept only those containing canonical splice sites and no ambiguous nucleotides in the sequence (Supplementary Table 1). For each of the homologous introns obtained, we did independent BS predictions applying the same method used for *S. cerevisiae*. Subsequently, we built pairwise alignments between each of the putative introns and the *S. cerevisiae* homologous introns using PRANK (Loytynoja et al., 2008). In the website http://regulatorygenomics.upf.edu/Software/Yeast_Introns/ we show all *S. cerevisiae* introns together with their homolog counterparts identified in the other yeast species (those that contained the BS aligned with the *S. cerevisiae* BS).

HAGs dataset

For each intron we collected the set of all HAGs (AAG, TAG and CAG) located between 10 nt downstream of the BS and the end of the downstream exon and classified them as *real* if they were annotated 3′ss, *intronic* if they were not annotated as 3′ss and were located between the BS and the annotated 3′ss; and *exonic* if they were located in the downstream exon and not annotated as 3′ss. Both exonic and intronic HAGs were considered as negatives to build the SVM and the annotated 3′ss were considered as positives.

S. cerevisiae 5′UTR intron dataset

To build the 5′UTR intron dataset we extracted a set of 24 5′ UTR exons from SGD (Engel et al., 2010). These introns were not included in the set used to build the SVM as they do not have an annotated upstream exon and the available annotation of the downstream exon is not always compatible with the intron, i.e. the annotated downstream exon starts downstream of the annotated 3′ss). In the cases in which the downstream exon was not compatible with the intron, we extended the exon length so that the start would coincide with the annotated 3′ss. All these introns contain canonical splice sites and a branch site sequence predicted as described in the Materials and Methods section. From this set we then collected all HAGs located between the BS and the end of the downstream exon resulting in 24 annotated 3′ss, 9 intronic HAGs and 1075 exonic HAGs with an effective distance <= 51 nt

Effective distance for random sets

To assess the significance of our findings regarding the effective distance we generated two sets of random sequences to compare to. First, for each of the annotated 3′ss we took the sequence between the BS and the 3′ss, discarding the first 8 nt after the Branch Site A. The first random set was composed of 1000 randomized sequences for each of the original sequences extracted, therefore maintaining sequence content and length distribution. To generate the second set, for each of the real sequences, we extracted 1000 random sequences of the same length from the genome, therefore maintaining only the same length distribution. Then, for each of these sequences we did an RNA structure prediction using the program RNAfold from the Vienna package (Hofacker, 2009) with default parameters and measured the effective distance. For each of the 2000 random datasets (1000 sets composed of 282 randomized sequences and 1000 sets composed of 282 random sequences) we measured the maximum effective length obtained in each of them. Both distributions are significantly different from the real dataset (Supplementary Figure 2): the empirical p-values for the comparisons to randomized introns and random introns are 0.002 and 0.008, respectively.

Features selected to build the SVM

The list of features selected to build the Support Vectors of each of the 3′ss analyzed are the following:

Splice site sequence: We scored each of the splice sites, AAG, CAG and UAG, using the $\log_2$-rate of the their frequency in the set of annotated 3′ss relative to their frequency in the set of unannotated HAGs.

Distance to the BS: For each HAG we measured the distance to the predicted BS. We defined the distance between the BS and any HAG as the number of nucleotides between the A of the BS and the HAG, including the last position. Using this definition, TACTAACACNNNNNTAG would represent a distance of 10 nt.

Relative accessibility: For each HAG, the relative accessibility $A_k^{(R)}$ was calculated by normalizing the square of the accessibility (A_k) to the maximum accessibility of a HAG in the intronic or exonic region around the same annotated 3′ splice site:

$$A_k^{(R)} = \frac{A_k^2}{\max_j\{A_j\}}$$

Polypyrimidine content: The polypyrimidine content was measured as the proportion of pyrimidines in the region between 7 nt downstream from the A of the BS to the nucleotide upstream from the analyzed 3′ss.

Distance to the PPT: We used the heuristic method defined in (Clark and Thanaraj, 2002) to predict polypyrimidine tracts between the BS and the HAG being evaluated. In case that more than one PPT was predicted for a given HAG, we kept the closest one. We defined the distance to the PPT as the number of nucleotides between the end of the PPT and the HAG, without including them. The score for this feature was defined as the $\log_{10}$ of this distance. When no PPT could be identified by the method, the maximum distance possible was taken, that is equal to the distance between the BS and the 3′ss minus 6 (we removed the nucleotides belonging to the BS after the A and the HAG).

Blast search of alternative splicing products

We obtained the mRNA sequences and the protein sequences, when applicable, resulting from the usage of the predicted alternative 3′ss. For all the predicted proteins (alternative 3′ss in the coding region), we looked for homologous sequences in the non-redundant protein database (nr) from RefSeq (Pruitt et al., 2007) using blastp (Altschul et al., 1997), imposing that the variable region was part of an alignment. In the cases in which we predicted alternative 3'ss in non-coding genes (snRN17A and snRN17B), we looked for homologous sequences in nr database using blastn (Altschul et al., 1997).

Search of homologous proteins

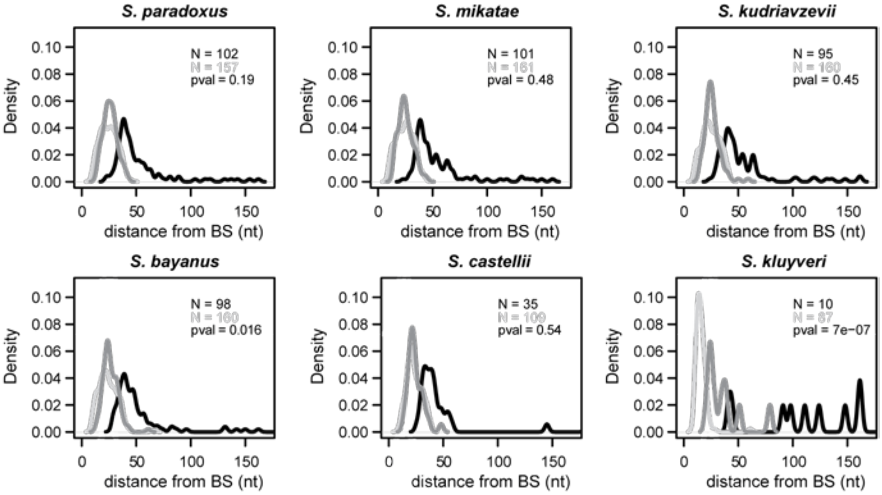
We wanted to check whether the protein products resulting from the usage of the predicted alternative 3'ss that did not introduce a PTC could be functional, i.e. chrIV:1103808-1103890:+:YDR318W_31, chrIII:107034-107110:+:YCL005W-A_54, and chrXII:564457-564515:-:YLR211C_40. In order to do so, we looked for homologous proteins of the translated results of the alternative 3'ss events that were inside coding genes and did not introduced a PTC. First, we identified putative homologous proteins using blastp against Uniprot database (UniProt Consortium, 2011), and kept all matches obtained. We then selected the possible homologous proteins based on the percentage of identity and the length of the alignment using the curve calculated for protein pairs with known structure (Rost, 1999). The threshold applied is described by the formula:

$$p(n) = n + 480 \cdot L^{-0.32 \cdot (1+e^{-L/1000})}$$

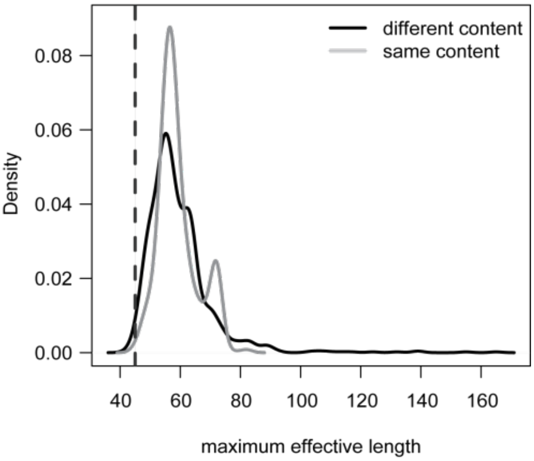
where L is the amount of nucleotides aligned between two proteins; p the cut-off percentage of identical residues over the L aligned residues; and n describes the distance in percentage points from the curve. For 99% true positives, $n = 5$ (Rost, 1999). Applying this method, we are able to retrieve 31 homologous proteins for YCL005W-A (Supplementary figure 4a), 13 for YDR318W (Supplementary Figure 4b), and 16 for YLR211C (Supplementary Figure 4c). For each of the sets of homologous proteins, we performed multiple sequence alignments using T-coffee with default parameters (Notredame et al., 2000). In all three cases we find that the deletion produced by the usage of the alternative 3′ss is not present in the majority of the other species and moreover corresponds to a conserved region, suggesting that the deleted part is important for protein function. The conservation of the predicted proteins with the closest homologs is shown in Supplementary Figures 5, 6 and 7.

Additionally, we also did secondary structure predictions with psipred (McGuffin et al., 2000) for the 3 predicted protein candidates. In all three cases the deletion falls inside a predicted alpha-helix, which suggests that the structure of the resulting proteins changes and therefore, it could affect protein function. Taking together all these pieces of evidence, we cannot claim that the protein products resulting from the usage of the candidate 3′ss predicted will be functional.

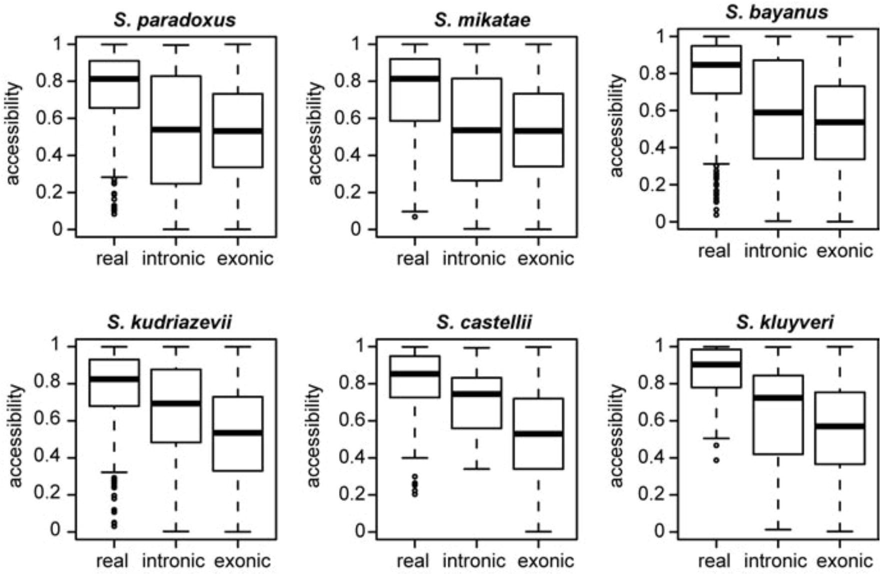
Supplementary Figures and Tables



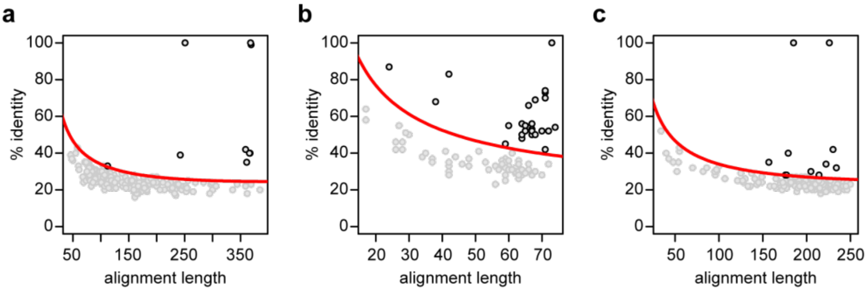
Supplementary Figure 1. Distribution of BS-3′ss distances in the homologous yeast species analyzed. Introns are separated in two categories, those that contain a secondary structure in this region (black), and those that do not (light gray). The effective distance of introns containing a secondary structure is also shown (dark gray line). For each species, the number of introns in each category is shown (black, introns with a secondary structure; light grey, introns without a secondary structure). For each species, the p-value of the comparison of length distributions of introns is also given.



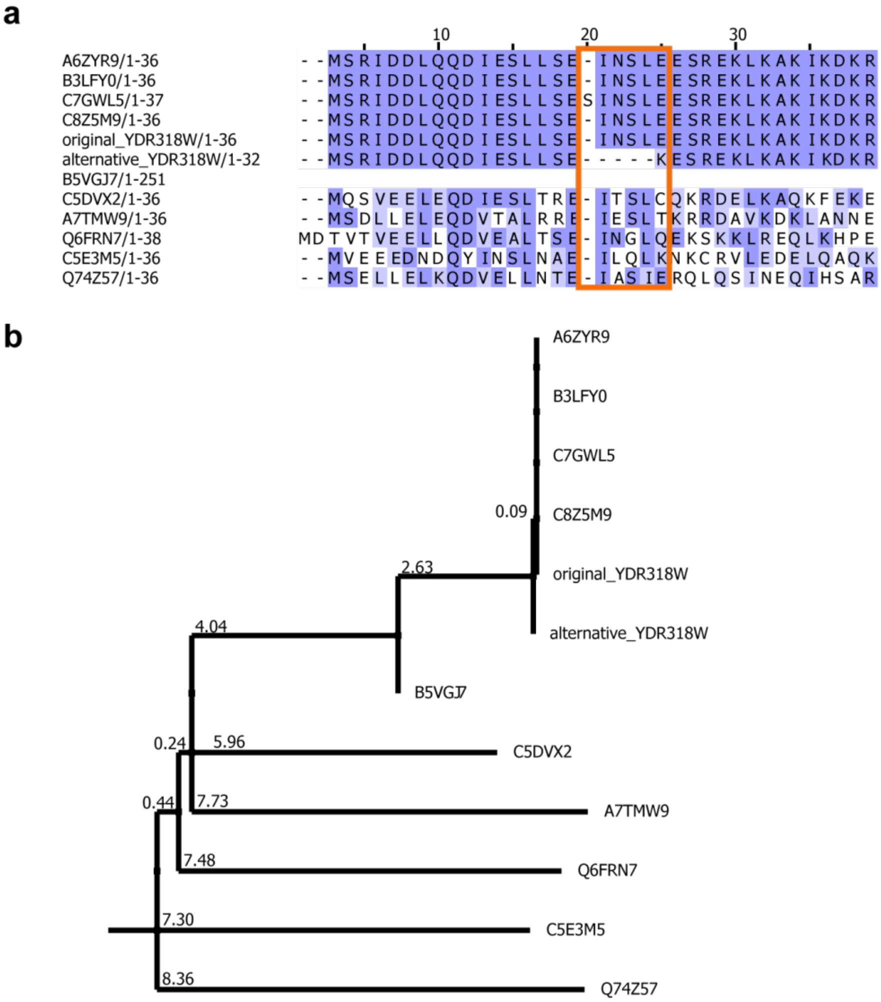
Supplementary Figure 2. Maximum length distribution of effective distances for random sets with the same (grey) and different (black) sequence content. The dashed line marks the maximum effective distance observed in real introns.



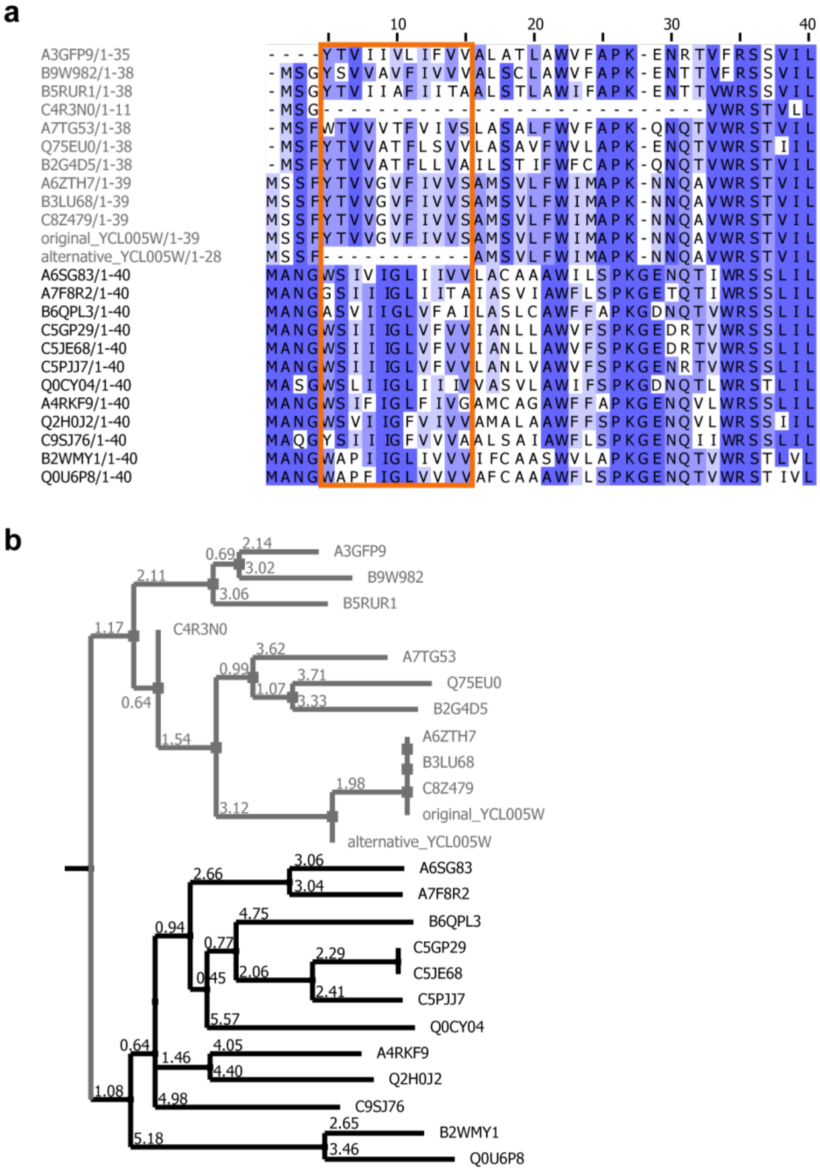
Supplementary Figure 3. Boxplot diagrams showing the accessibility values of annotated and cryptic (intronic and exonic) 3′ss for all the homologous yeast species analyzed.



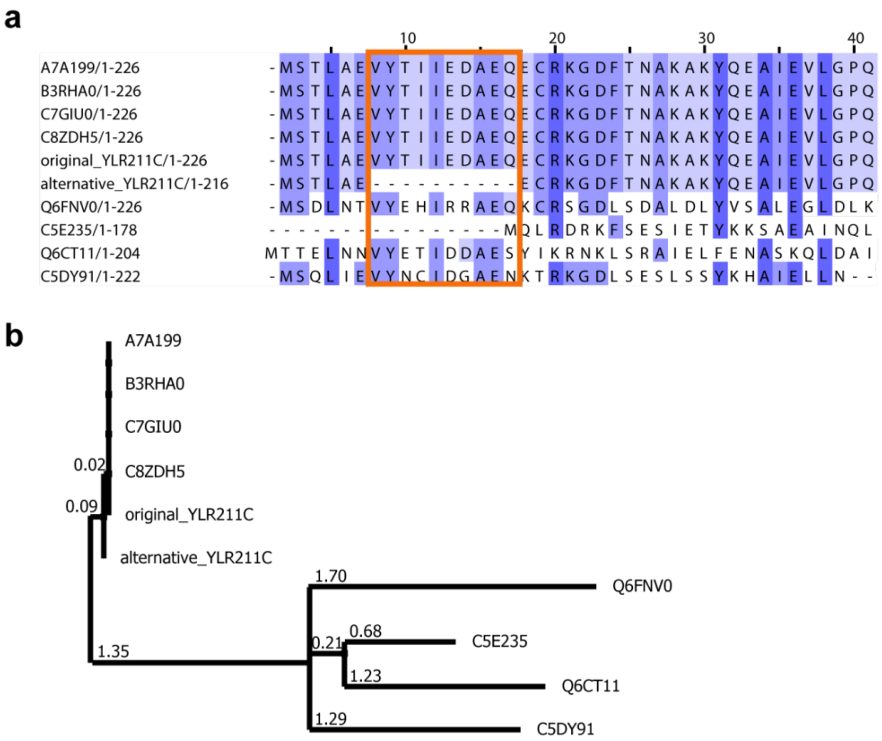
Supplementary Figure 4. Representation of percentage identity (y-axis) versus alignment length (x-axis) of the homologous proteins identified for YCL005W-A **(a)**, YDR318W **(b)**, and YLR211C **(c)**. The red line shows the curve defined by Burkhard Rost (Rost, 1999) to determine structural homolgs with a 99% of true positives. Grey dots represent discarded proteins (false structural homologs) whereas black circles represent the selected proteins (true structural homologs).


a
10
20
30
A6ZYR9/1-36 --MSRIDDLQQDIESLLSE-INSLEESREKLKAKIKDKR
B3LFY0/1-36 --MSRIDDLQQDIESLLSE-INSLEESREKLKAKIKDKR
C7GWL5/1-37 --MSRIDDLQQDIESLLSESINSLEESREKLKAKIKDKR
C8Z5M9/1-36 --MSRIDDLQQDIESLLSE-INSLEESREKLKAKIKDKR
original_YDR318W/1-36 --MSRIDDLQQDIESLLSE-INSLEESREKLKAKIKDKR
alternative_YDR318W/1-32 --MSRIDDLQQDIESLLSE----KESREKLKAKIKDKR
B5VGJ7/1-251
C5DVX2/1-36 --MQSVEELEQDIESLTRE-ITSLCQKRDELKAQKFEKE
A7TMW9/1-36 --MSDLLELEQDVTALRRE-IESLTKRRDAVKDKLANNE
Q6FRN7/1-38 MDTVTVEELLQDVEALTSE-INGLQEKSKKLREQLKHPE
C5E3M5/1-36 --MVEEEDNDQYINSLNAE-ILQLKNKCRVLEDELQAQK
Q74Z57/1-36 --MSELLELKQDVELLNTE-IASIERQLQSINEQIHSAR
b
A6ZYR9
B3LFY0
C7GWL5
0.09
C8Z5M9
2.63
original_YDR318W
4.04
alternative_YDR318W
B5VGJ7
0.24
5.96
C5DVX2
0.44
7.73
A7TMW9
7.48
Q6FRN7
7.30
C5E3M5
8.36
Q74Z57

Supplementary Figure 5. (a) Extract of the alignment of YDR318W homologous proteins identified, including the alternative protein predicted. The alignment has been edited with Jalview (Waterhouse et al., 2009). The sequences are colored according to the Neighbor-Joining (NJ) tree based on the % identity of the proteins **(b)**. The orange box shows the deletion introduce by the alternative 3′ss predicted.



Supplementary Figure 6. **(a)** Extract of the alignment of YLC005W-A homologous proteins identified, including the alternative protein predicted. The alignment has been edited with Jalview (Waterhouse et al., 2009). The sequences are colored according to the Neighbor-Joining (NJ) tree based on the % identity of the proteins **(b)**. For clarity purposes, proteins are divided into two groups according to the NJ tree (grey and black groups), and the percentage identity used for coloring them takes into account the identity within a given group. The orange box shows the deletion introduce by the alternative 3′ss predicted.


a
10
20
30
40
A7A199/1-226 -MSTLAEVYTIIEDAEQECRKGDFTNAKAKYQEAIEVLGPQ
B3RHA0/1-226 -MSTLAEVYTIIEDAEQECRKGDFTNAKAKYQEAIEVLGPQ
C7GIU0/1-226 -MSTLAEVYTIIEDAEQECRKGDFTNAKAKYQEAIEVLGPQ
C8ZDH5/1-226 -MSTLAEVYTIIEDAEQECRKGDFTNAKAKYQEAIEVLGPQ
original_YLR211C/1-226 -MSTLAEVYTIIEDAEQECRKGDFTNAKAKYQEAIEVLGPQ
alternative_YLR211C/1-216 -MSTLAE----------ECRKGDFTNAKAKYQEAIEVLGPQ
Q6FNV0/1-226 -MSDLNTVYEHIRRAEQKCRSGDLSDALDLYVSALEGLDLK
C5E235/1-178 -----------------MQLRDRKFSESIETYKKSAEAINQL
Q6CT11/1-204 MTTELNNVYETIDDAESYIKRNKLSRAIELFENASKQLDAI
C5DY91/1-222 -MSQLIEVYNCIDGAENKTRKGDLSESLSSYKHAIELLN--
b
A7A199
B3RHA0
C7GIU0
0.02
C8ZDH5
0.09
original_YLR211C
alternative_YLR211C
1.70
Q6FNV0
0.21
0.68
C5E235
1.35
1.23
Q6CT11
1.29
C5DY91

Supplementary Figure 7. (a) Extract of the alignment of YLR211C homologous proteins identified, including the alternative protein predicted. The alignment has been edited with Jalview (Waterhouse et al., 2009). The sequences are colored according to the Neighbor-Joining (NJ) tree based on the % identity of the proteins **(b)**. For clarity purposes, the homologs that were not aligned in the region of interest were deleted from the alignment and from the tree. The orange box shows the deletion introduce by the alternative 3′ss predicted.

Supplementary Table 1. Homologous introns identified in yeast species.

Species name	Nº of introns
S. cerevisiae	282
S. paradoxus	259
S. mikatae	262
S kudriavzevii	255
S. bayanus	258
S. castellii	150

Supplementary Table 2. RNA-Seq datasets.

Dataset	Study	Read Length	Reads	Reads not mapped to the genome	All split-mapped reads	Unique split-mapped reads
HS	Yassour et al. 2009	36	11776251	2662310	28085	23801
YPD-t0	Yassour et al. 2009	36	13932371	3461274	54177	45857
YPD-t15	Yassour et al. 2009	36	12118043	2833818	50907	43944
WT	Nagalakshmi et al. 2008	33	29912517	15525631	72777	60471

Dataset abbreviations: *HS* heat shock (37°C), *YPD-t0* Yeast Peptone Dextrose time 0 (22°C); YPD-t15 Yeast Peptone Dextrose time15 (22°C); WT wild type (30°C)

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25,* 3389-3402.

Clark, F., and Thanaraj, T.A. (2002). Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Hum. Mol. Genet. *11,* 451-464.

Engel, S.R., Balakrishnan, R., Binkley, G., Christie, K.R., Costanzo, M.C., Dwight, S.S., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Hong, E.L.*, et al.* (2010). Saccharomyces Genome Database provides mutant phenotype data. Nucleic Acids Res. *38,* D433-6.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A.*, et al.* (2011). The UCSC Genome Browser database: update 2011. Nucleic Acids Res. *39,* D876-82.

Goecks, J., Nekrutenko, A., Taylor, J., and Galaxy Team. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. *11,* R86.

Hofacker, I.L. (2009). RNA secondary structure analysis using the Vienna RNA package. Curr. Protoc. Bioinformatics *Chapter 12,* Unit12.2.

Loytynoja, A., and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science *320,* 1632-1635.

McGuffin, L.J., Bryson, K., and Jones, D.T. (2000). The PSIPRED protein structure prediction server. Bioinformatics *16,* 404-405.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320,* 1344-1349.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. *302,* 205-217.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *35,* D61-5.

Rost, B. (1999). Twilight zone of protein sequence alignments. Protein Eng. *12,* 85-94.

UniProt Consortium. (2011). Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. *39,* D214-9.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics *25,* 1189-1191.

Yassour, M., Kaplan, T., Fraser, H.B., Levin, J.Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., Gnirke, A.*, et al.* (2009). Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc. Natl. Acad. Sci. U. S. A. *106,* 3264-3269.

Differentiated evolutionary rates in alternative exons and the implications for splicing regulation

Mireya Plass and Eduardo Eyras

BMC Evol. Biol. 2006; 6: 50.

3.2. Supplementary Material

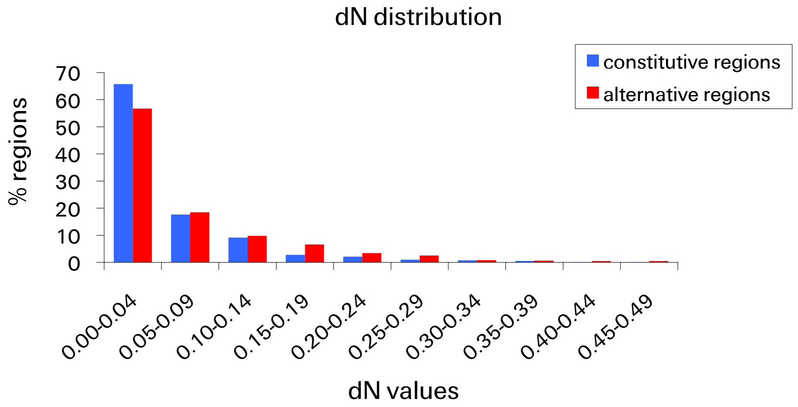


Figure S1. Distribution of the non-synonymous substitution rate (dN) for constitutive and alternative regions.

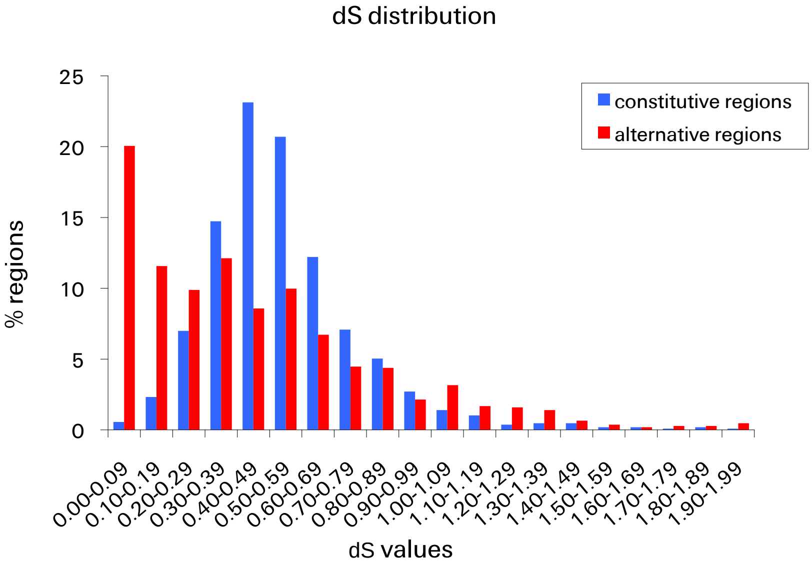


Figure S2. Distributions of the synonymous substitution rate (dS) for constitutive and alternative regions.

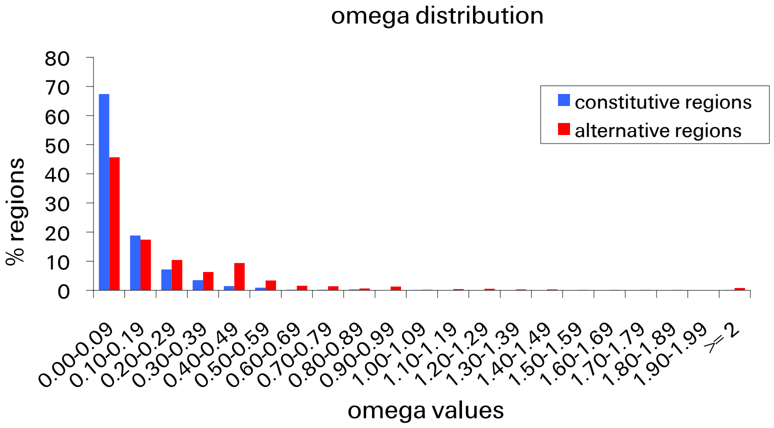


Figure S3. Distribution of the values of Omega (=dN/dS) for the alternative and constitutive regions. Alternative (red) and constitutive (blue) exons have significantly different distributions (p-value < 2.2e-16).

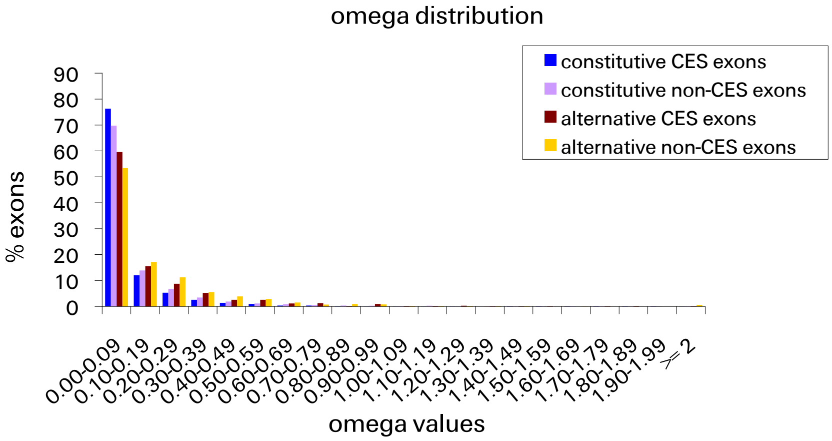


Figure S4. Distribution of omega (=dN/dS) for each of the four subsets of orthologous exons: constitutive and alternative exons with (CES) or without (non-CES) conservation of the exonic structure.

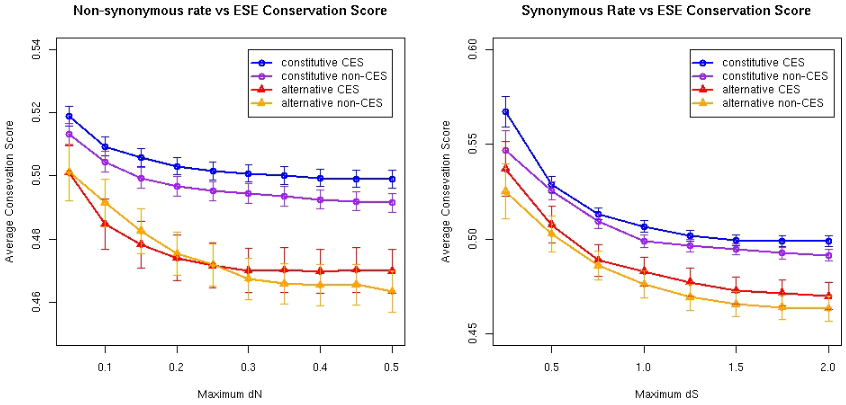


Figure S5. Correlation of the ESE conservation score with the (left) non-synonymous (dN) and (right) synonymous (dS) divergence for each of the four exon-groups.

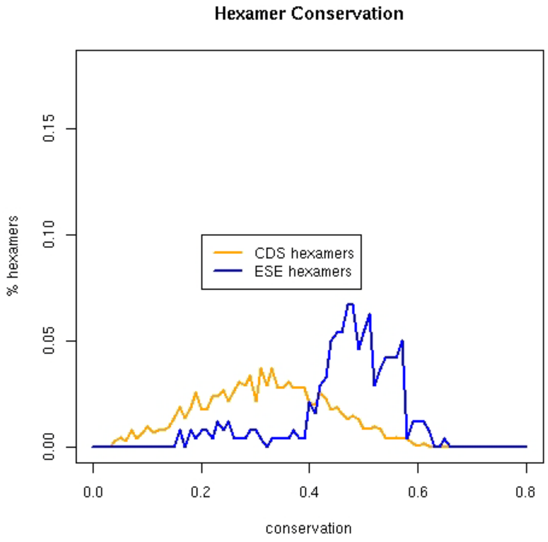


Figure S6. Distribution of the conserved hexamers for two exon data sets: hexamers in the CDS of single exon genes (orange) and ESE hexamers from our set of constitutive and alternative exons. On the x-axis, the conservation is given as the fraction of the occurrences of the hexamers in human that is exactly conserved in mouse. The y-axis represents the proportion of hexamers with a given conservation.

Density of Exonic Enhacers in Alternative and Constitutive exons

We compared the density of ESEs in constitutive and alternative exons. For each gene we calculate the difference in the proportion of bases covered by ESEs in constitutive and alternative exons:

$$\left.\frac{bp_{ESEs}}{bp_{exons}}\right|_{const} - \left.\frac{bp_{ESEs}}{bp_{exons}}\right|_{alt}$$

We found a higher density of ESEs in constitutive exons. The mean of the differences is 0.016. A check of the difference using paired t-test gives a p-value = 6.273e-05, and a 95% confidence interval [0.008496871, 0.024703948], which is not overlapping 0. From this we can conclude that constitutive exons have a slightly higher density of ESEs.

Further, we plotted the density of ESEs in alternative and constitutive exons separated by CES and non-CES exons (Figure S8). This average density was plotted for each exon subset, at different minimum percentage identity values. Slicing the data in this way, we can view the differences between the sets, and how these differences change with the conservation. We observe that constitutive exons have in general higher density of ESEs than alternative exons.

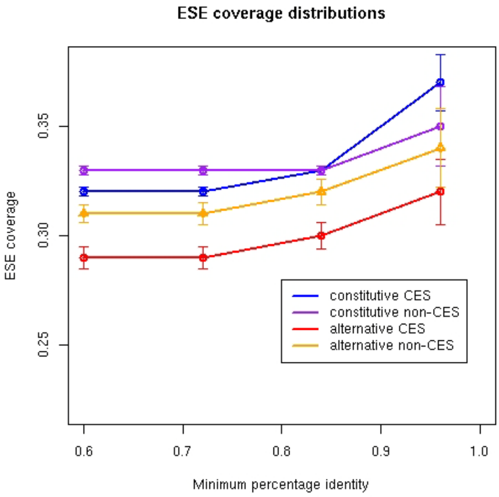


Figure S7. Correlation of the ESE density versus the minimum percentage identity conservation of exon sequences. The ESE density is measured as the fraction of the exon length in human that is covered by ESEs.

Testing the influence of biases in the results

We wanted to test whether there are biases in our dataset and whether these could influence the results that we present in our paper. For this work, we had classified our exon set according to whether they appear in a transcript with an exonic structure that is conserved (CES) or not conserved (non-CES) between human and mouse. In order to test the influence of possible biases we have considered the genes to which these exons belong, and separated them into two sets: those containing conserved exons (that we call CES-exon-containing genes) and those containing non-CES exons (that we call nonCES-exon-containing genes), and considered the distributions of the number of exons per gene, gene length, and difference in the number of transcripts between gene orthologs. We found that there are some characteristics more typical of genes containing non-CES exons, but none of these properties influence the results we present in our manuscript. A detailed explanation of this analysis is given below.

Dependencies with the number of exons per gene

We compared the distributions of the number of exons per gene for CES-exon-containing genes and nonCES-exon-containing genes. We observe that nonCES-exon-containing genes are more frequent in the range of 22 or more exons per gene, whereas CES-exon-containing genes are more frequent in the rage of less than 22 exons per gene (see Figure 1).

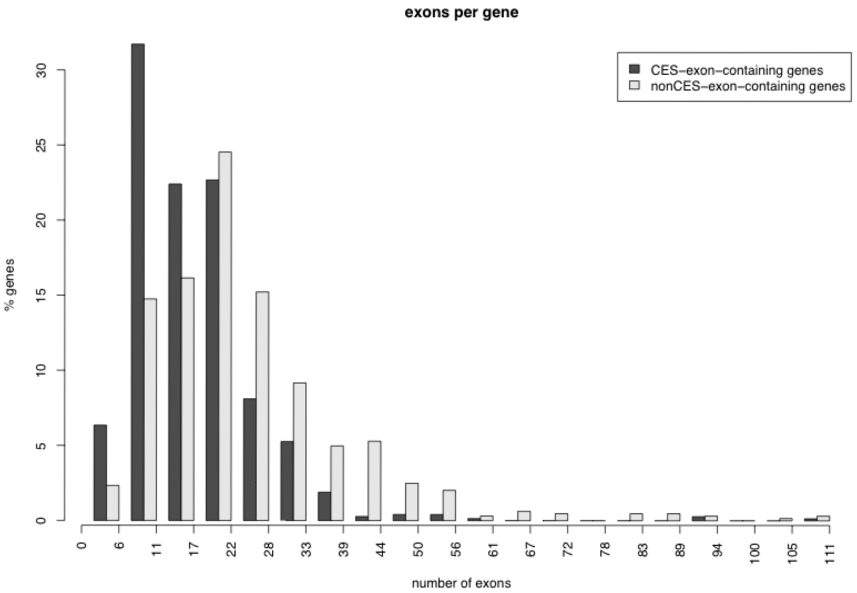


Figure 1. Distributions of the number of exons per gene for CES-exon-containing genes and nonCES-exon-containing genes.

To test whether these biases have any influence in the differences of dN and dS that we observe for the four different exon subtypes (constitutive CES, constitutive non-CES, alternative CES and alternative non-CES), we did a equal-sized random sampling of exons from this distribution. More specifically, from the distribution of the number of exons per gene we considered the following 5 bins:

number of exons-per-gene	constitutive CES exons	constitutive non-CES exons	alternative CES exons	alternative non-CES exons
3-11	942	169	291	112
12-17	915	353	215	110
18-23	1293	715	233	222
24-29	580	690	111	142
30-40	767	836	108	241

In the table we include the number of exons of each exon-subset present in each of these bins. These bins account for the 82.8% of the total number of exons considered in the paper.

From each bin, and from each exon-subtype we sampled 20 exons at random, hence 100 exons for each exon subtype, and calculated the average dN and average dS values for each subtype. This random sampling and average calculation was repeated 10000 times. Figure 2 shows the distribution of the average dN values from this 10000 samplings for each exon subtype. We observe the same behaviour reported in the manuscript: non-CES exons have higher dN than their CES counterparts. In particular, alternative non-CES exons have on average the highest dN values, whereas constitutive CES exons have on average the lowest dN values.

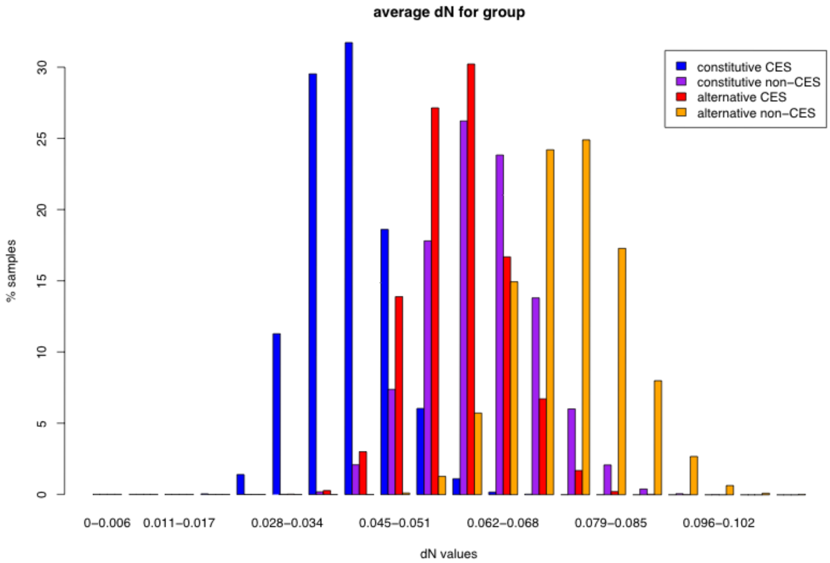


Figure 2. Distribution of the average dN values for the four exon sets (constitutive CES, const. non-CES, alternative CES and alt. non-CES), obtained from an equal-sized random sampling of equivalent bins of the exons-per-gen distribution.

Figure 3 shows the distribution of the average dS values for each exon subtype, and reflect the same pattern described in the manuscript: CES exons have lower average dS values. In particular, alternative CES exons have on average the lowest dS values, whereas constitutive non-CES exons on average the highest dS values.

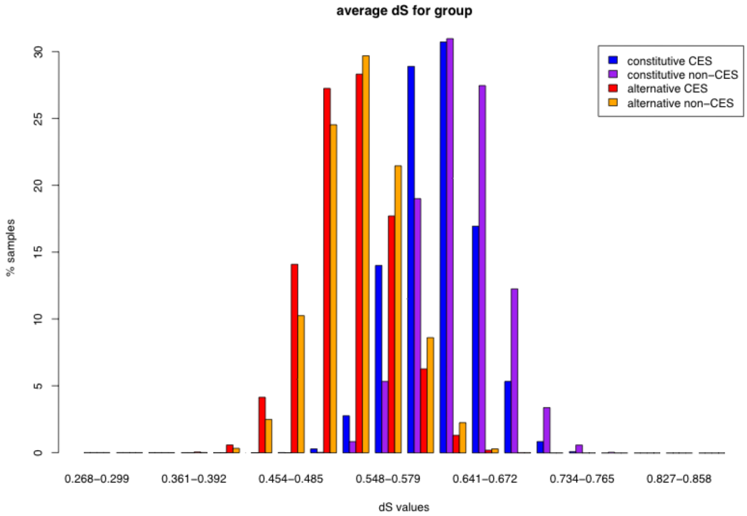


Figure 3. Distribution of the average dS values for the four exon sets (constitutive CES, const. non-CES, alternative CES and alt. non-CES), obtained from an equal-sized random sampling of equivalent bins of the exons-per-gen distribution.

We can conclude that the number of exons per gene does not affect our results.

Dependencies with the gene length

The gene-length distributions follow the same trend as for the number of exons per gene. Short genes are more frequently CES-exon-containing than nonCES-containing, and long genes are more frequently nonCES-exon-containing than CES-exon-containing ones (see Figure 4).

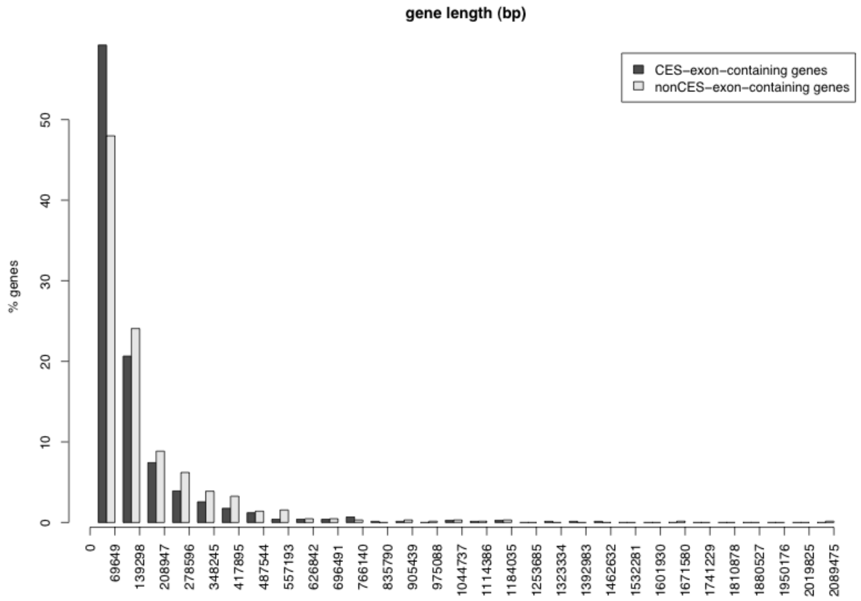


Figure 4. Distributions of the gene-lengths for CES-exon-containing genes and nonCES-exon-containing genes.

We performed the same random sampling procedure of equal-sized exon subsets as before, now using the gene-length distribution. We considered the following bins:

gene lengths (bp)	constitutive CES exons	constitutive non-CES exons	alternative CES exons	alternative non-CES exons
0-72051	2286	1391	529	508
72052-144102	1100	1023	240	283
144103-216153	449	479	69	113
216154-288203	397	573	62	142
360254-648458	279	535	49	65

In the table we also give the number of exons for each subtype. These bins account for the 96.8% of the total number of exons used in the paper.

As before, from each bin and for each exon-subtype, we sampled 20 exons at random, hence 100 exons for each subtype. This was repeated 10000 times, and each time, the average dN and dS for each subtype was calculated. We obtained the same results as before: non-CES exons have on average higher dN values (see Figure 5) and CES-exons have on average lower dS values. (see Figure 6). We therefore conclude that the gene length does not affect our results.

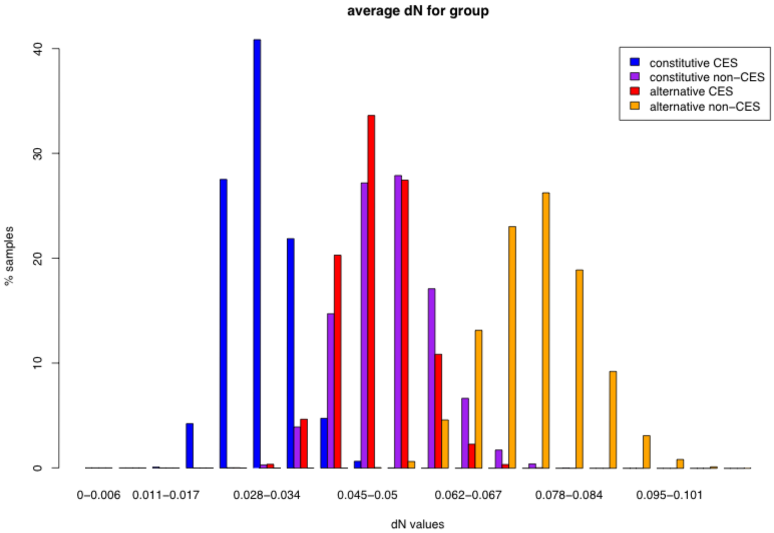


Figure 5. Distribution of the average dN values for the four exon-subtypes (constitutive CES, const. non-CES, alternative CES and alt. non-CES), obtained from an equal-sized random sampling of equivalent bins of the gene-length distribution.

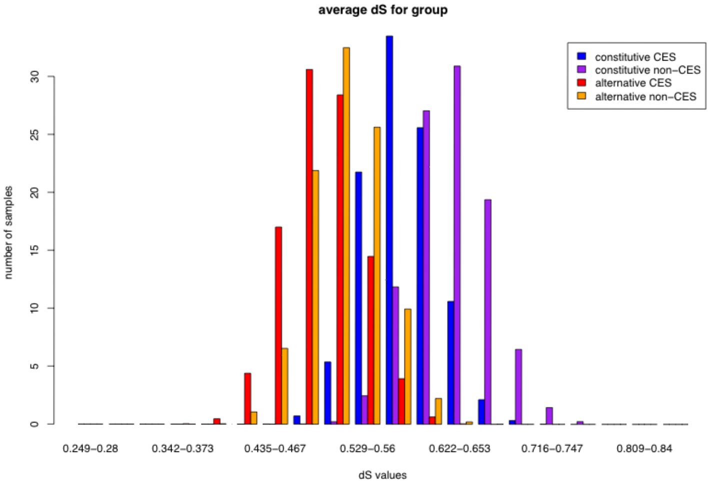


Figure 6. Distribution of the average dS values for the four exon-subtypes (constitutive CES, const. non-CES, alternative CES and alt. non-CES), obtained from an equal-sized random sampling of equivalent bins of the gene-length distribution.

Dependencies with the differences in the number of transcripts per gene between orthologous pairs

We also looked at the possible dependencies with the difference in the number of transcripts in human and mouse gene orthologs. For each pair of human-mouse gene orthologs, we calculated the distribution of the differences in the number of transcripts. The distributions for CES-exon-containing and nonCES-exon-containing genes are shown in Figure 7.

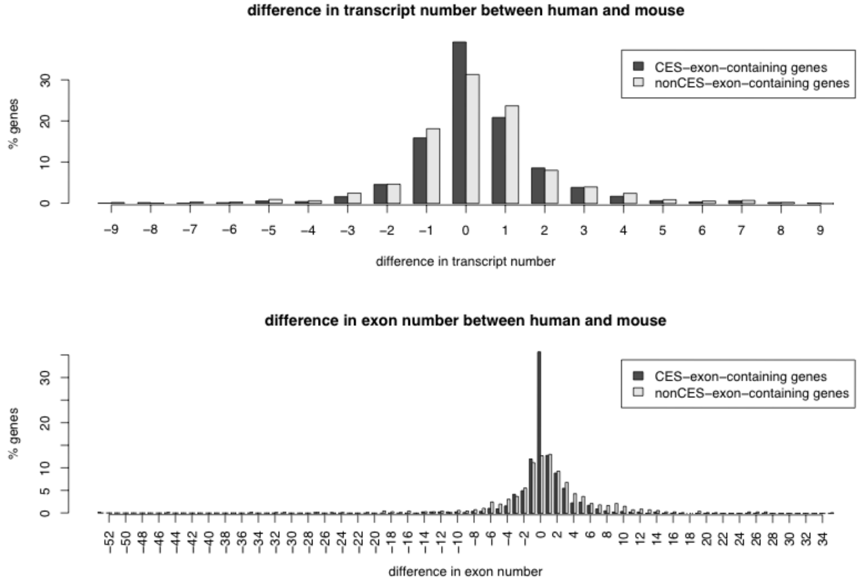


Figure 7. Distributions of the differences in the number of transcripts (above) and in the number of exons (below) between human-mouse orthologous gene-pairs. The x-axis is calculate subtracting the number in mouse to the number in human.

To test whether our results are influenced by the cases in which there is a big difference in the number of transcripts between orthologous genes, we calculated the distributions of dN and dS for orthologous pairs that have the same number of transcripts (see Figure 8). For this subset, the CES-exon-containing genes have on average 1.8 transcripts (median 2), and the nonCES-exon-containing genes have on average 2.3 transcripts (median 2).

We observe (see Figure 8) that the exons distribute with the same general trend as reported in the manuscript: constitutive CES exons have lower dN values, alternative non-CES exons have higher dN values, alternative CES exons have lower dS values and constitutive non-CES have higher dS values.

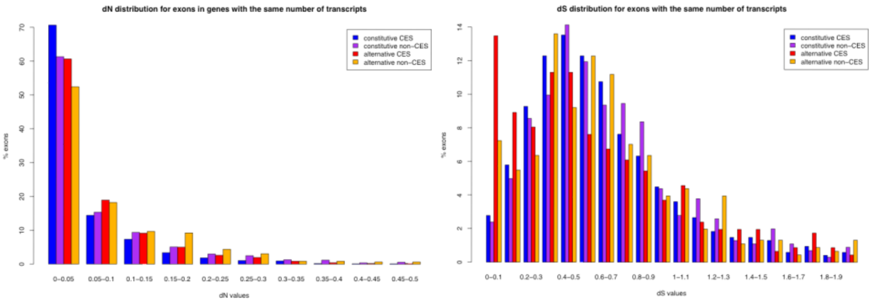


Figure 8. Distribution of the dN and dS values for exons (separated in four exon subtypes) in genes orthologs with the same number of transcripts per gene.

We looked at the distribution of the differences in the number of exons per gene between human and mouse orthologs (see Figure 7). We clearly see that orthologous genes with the same number of exons contain more frequently CES exons. This, however, is an expected feature of our classification: we want to distinguish between cases where the exonic structure varies and cases where it does not. This variation is expected to correlate with gene orthologs with exons that are species specific. Thus orthologous genes with the same number of exons are more likely to share most of the exonic structures. We also note that we are considering only coding exons in our analyses.

We calculated the dN and dS distributions for the different exon-subsets in the case where the orthologous genes have the same number of exons (see Figure 9). We observe that the exons distribute with the same general trend as reported in the manuscript: constitutive CES exons have lower dN values, alternative non-CES exons have higher dN values, alternative CES exons have lower dS values and constitutive non-CES have higher dS values.

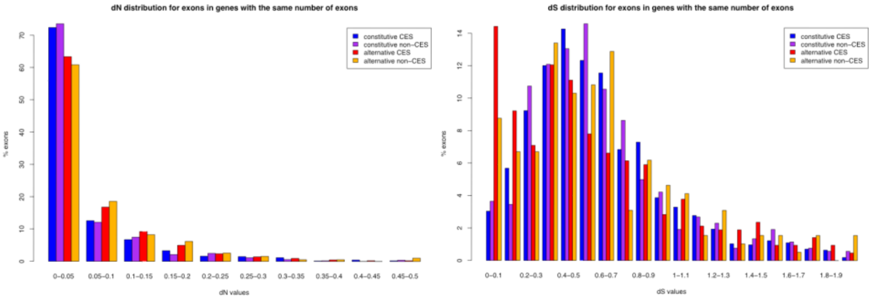


Figure 9. Distribution of the dN and dS values for exons (separated in four subtypes) in genes orthologs with the same number of exons per gene.

Conclusions

We have sliced our data taking into account that a number of genes contain one or more transcripts which exonic structure is not conserved in mouse. The present tests show that these genes are quite often long and with many exons. We also show that, however, these features do no influence the findings reported in our article. We therefore can expect that long genes with many exons are more prone to vary in exonic structure with respect to their orthologs. However, independently of the type of gene in which this variation is more frequently observed, it is the variation itself what correlates with a difference in sequence conservation. Thus we can conclude that the sequence properties of alternative exons depend on contextual factors. A subset of alternative exons has higher sequence conservation than average, and a different subset has higher dN than average, and these subsets strongly correlate with exons in conserved and non-conserved exonic structures, respectively.

IV. DISCUSSION

Each of the research articles presented in this thesis already includes a discussion section. Therefore, here I will present a global discussion of the points presented in the objectives section.

1. Conservation of SR and SR-like proteins in eukaryotes and implications in splicing regulation

In the first work presented (Plass et al., 2008), we analyzed the conservation of SR and SR-related proteins in 22 eukaryotic species including metazoans, plants, fungi and protists. This analysis extends previous works (Barbosa-Morais et al., 2006; Schwartz et al., 2008) and provides interesting insights about the evolution and expansion of SR and SR-related proteins in eukaryotes. On the one hand, we showed that SR proteins are widely spread in fungal species and only those fungi from the Saccharomycetaceae family (*K. lactis* and *S. cerevisiae*) lack SR protein homologs. On the other hand, we found that some fungal species, like *Rhizopus oryzae*, have more than a copy of some SR proteins. Moreover, in all the species analyzed we see that the lack of SR proteins is related with the presence of MUD2 and NPL3 in the same species. MUD2 is homologous to U2AF65 and is involved in the recognition of the BS (Abovich et al., 1994), although it lacks the RS domain. In metazoans, U2AF65 interacts with U2AF35, and this interaction is important for 3′ss recognition. The lack of U2AF35 in the members of the

Saccharomycetaceae family, which contain MUD2, may indicate a difference in 3′ss recognition in these species. Conversely, NPL3 is an RNA binding protein with RRMs similar to those from SRP2 (Plass et al., 2008) but without an RS domain. Interestingly, this protein has been shown to promote splicing in yeast by facilitating the co-transcriptional recruitment of U1 snRNP and U2 snRNP components in early steps of spliceosome assembly (Kress et al., 2008), similarly to what other mammalian SR proteins do (Blencowe et al., 1999; Bourgeois et al., 2004; Hertel and Graveley, 2005). However, the function of this protein differs from that of mammalian SR proteins since it is not able to enhance the recognition of suboptimal splicing signals (Kress et al., 2008). In mammals, this recognition is mediated through the RS domain (Wu and Maniatis, 1993; Shen et al., 2004; Shen and Green, 2004). Moreover, it has been shown that mammalian SR proteins inserted in *S. cerevisiae* are able to recognize suboptimal splicing signals, showing that the mechanism is conserved in yeast (Shen et al., 2006). This gives extra evidence that RS domain of NPL3 does not function as mammalian RS domains.

2. Relation between SR proteins and the BS

It has been shown that SR proteins are able to enhance the recognition of suboptimal splicing signals by stabilizing the interaction between the snRNA and the pre-mRNA through the

RS domain (Shen et al., 2004; Shen and Green, 2004), and that this mechanism is conserved across eukaryotes (Shen et al., 2006). We investigated whether there is a relation between splicing signals and the RS domain across species (Plass et al., 2008). The hypothesis was that, if the binding between the snRNAs and the signals in the pre-mRNA was strong enough, there would be no requirement for the presence of RS domains to enhance the recognition of suboptimal splicing signals. In contrast, if the signals were suboptimally recognized, we would expect more SR proteins with functional RS domains. Our results confirmed this hypothesis, as we found that those species with more conserved splicing signals are also those containing fewer or no SR proteins (Plass et al., 2008). Furthermore, in previous works it was demonstrated that the function of the RS domain depends on the content of SR repeats (Graveley et al., 1998; Philipps et al., 2003) and the phosphorylation state of the domain (Prasad et al., 1999). Our results show that the presence of RS repeats in SRP2 is inversely correlated with the energy of the binding of the U2 snRNA to the BS, suggesting the co-evolution of the signal and the amount of SR repeats in the RS domain (Plass et al., 2008). Interestingly, species with more conserved BS signal present higher RD and RE repeats. These dinucleotides mimic the function of phosphorylated SR repeats and can function as RS domains (Cartegni and Krainer, 2003; Philipps et al., 2003). However, these domains are potentially inactive, similarly to RS domains that are inactive at high phosphorylation states. This provides further evidence suggesting that these domains may not function like the RS domain of eukaryotes.

3. Role of RNA secondary structures in 3′ss selection in yeast

We have shown that RNA secondary structure is important for 3′ss selection in yeast, and that this mechanism of regulation is conserved in other yeast species. In yeast, RNA promotes 3′ss selection by maintaining the right distance between the BS and the 3′ss. Moreover, secondary structures contribute to the proper selection of real 3′ss by preventing the recognition of cryptic ones placed between the BS and the 3′ss. These functions of RNA structures had already been shown in specific cases (reviewed in Warf and Berglund, 2010), but we have shown that it is widespread in yeast. We demonstrate that RNA structures are important for 3′ss selection in the majority of yeast introns, suggesting that the role of RNA secondary structures is more important in splicing than previously thought (Meyer et al., unpublished). Furthermore, we also demonstrate that RNA structures allow regulating alternative splicing in the absence of other splicing enhancers like SR proteins.

Using this idea, we designed a computational method to model 3′ss selection in yeast taking into account the roles of RNA secondary structure in shortening BS-3′ss distances and in hiding cryptic 3′ss. Our model is able to predict correctly over 90% of annotated 3′ss. Moreover, it can also be used to predict alternative 3′ss to be tested experimentally. These results suggest that our model is able to encapsulate the information required to identify used 3′ss in yeast, and that alternative splicing is more frequent than previously thought. Additionally, by comparing the predictions obtained at two different temperatures, we also show that 3′ss selection by RNA structures can be modified by temperature. Interestingly, it has been shown that several mRNAs change their splicing pattern or the levels of mature mRNA after heat-shock (Yassour et al., 2009). Therefore, RNA structure may be a common mechanism regulating alternative splicing in those events.

4. Understanding the impact of AS as gene regulator in yeast

32 of the 37 alternative 3′ss that we have predicted using our computational method introduce a PTC. These transcripts containing long 3′UTRs will be degraded through NMD (Amrani et al., 2004) and therefore, will not contribute to expand the yeast proteome. Moreover, the usage of the alternative 3′ss predicted is very low. Previous genome-wide studies on alternative splicing have highlighted the fact that lots of alternative transcripts are expressed at very low levels (Pan et al., 2006). The majority of these transcripts, rather than creating a

truncated protein, get degraded by NMD (reviewed in Lareau et al., 2007a). However, if the concentration of these alternative transcripts is very low, the effect that the degradation would have is minimal, and probably will not affect the final mRNA levels (Neu-Yilik et al., 2004; Pan et al., 2006). Thus, it is still controversial whether such alternative splicing events may have any effect on gene regulation or, in contrast, they are just splicing noise. In some cases, it has been demonstrated that alternative splicing coupled to NMD is an important regulatory mechanism, even if it happens at low levels. SR proteins are one of the clearest examples. The mRNA levels of SR proteins get regulated through NMD (Lareau et al., 2007b). Consequently, changes in the levels of mature protein can have further effects in the splicing patterns of lots of other proteins, which are regulated by differential concentrations of splicing enhancers and silencers in the cell (reviewed in Matlin et al., 2005). Taking all this into account, without further research, we cannot know which will be the actual functional impact of the predicted alternative 3′ss.

5. Analyzing the role of AS in sequence evolution

Three of the predicted alternative 3′ss in yeast produce mRNAs that will get translated into potentially functional proteins. In all three cases, the usage of an alternative 3′ss produces a deletion in the original protein that is not present in other species and affects the secondary structure of the protein. These results suggest that the resulting proteins will not be functional or at least will have an impaired function. However, it was previously suggested that alternative splicing events producing non-functional protein isoforms, when they are not very frequent, might allow the evolution of new protein functions with a low impact in the overall function of the protein (Modrek and Lee, 2003).

To analyze further the possibility that alternative splicing could have an impact on protein evolution, we analyzed the sequence properties of human cassette exons (Plass and Eyras, 2006). Our hypothesis was that alternative exons are expected to have more non-synonymous mutations, i.e. mutations affecting the coded amino acids, than constitutive exons. Furthermore, we also expected that these non-synonymous mutations would be more frequent in exons that are poorly included in final transcripts than in highly included exons, and hence they would have a low impact on the function of the final protein. Our analyses verified these hypotheses. The amount of non-synonymous mutations in alternative exons is higher than in constitutive exons, and it decreases with the inclusion of the alternative exon measured in ESTs (Plass and Eyras, 2006). This result is consistent with previous works showing that alternative exons are important for protein evolution (Iida and Akashi, 2000; Xing and Lee, 2005; Chen et al., 2006; Ermakova et al., 2006).

6. Understanding the relation between transcript structure, sequence conservation and AS

Contrary to the result that alternative exons contribute largely to protein evolution, it had also been demonstrated that alternative exons have higher sequence conservation than constitutive exons (Modrek and Lee, 2003; Sugnet et al., 2004; Philipps et al., 2004). We hypothesized that this apparent inconsistency could be explained by the presence of different exon populations containing different amount of splicing regulatory elements that are under purifying selection. Therefore, by studying the conservation of the transcript structure, which is related with splicing, we would be able to identify these exon populations. In our analyses, we compared dS values of constitutive and cassette exons with or without conservation of transcript structure. In agreement with previous studies, we found that at any identity threshold, dS values are lower for cassette exons than for constitutive exons (Plass and Eyras, 2006). This higher conservation had been related with a higher density of regulatory motifs involved in splicing control, which are likely under purifying selection (Hurst and Pal, 2001; Orban and Olah, 2001; Carlini and Genut, 2006; Parmley et al., 2006). Moreover, it is known that splicing events can be coordinated within a transcript (Liu et al., 2001), and that not all possible exon combinations are seen in mRNAs, suggesting a relation between transcript structure and splicing regulation. To understand better the impact of splicing regulatory motifs in sequence conservation and its relation with transcript structure, we analyzed dS and dN values of alternative exons present in transcripts with conserved and non-conserved transcript structures at different minimum inclusion levels based on EST data. We found that at any inclusion level, exons without a conserved transcript structure have higher dN and dS values than their counterparts with conserved exonic structure. We also observed that dS and dN values of alternative exons approach the corresponding values of constitutive exons at high inclusion levels (Plass and Eyras, 2006). These results suggest that the differences in dS and dN between constitutive and alternative exons depend on both the conservation of the transcript structure and the inclusion level. Moreover, they also reconcile the apparent discrepancy in the properties of alternative exons reported previously in the literature: cassette exons that are highly conserved are those that are highly included in transcripts and with a conserved exonic structure, suggesting that conserved splicing regulatory elements are responsible for this high sequence conservation. In contrast, alternative exons with low sequence conservation are those that are included in fewer transcripts, particularly in those without a conserved exonic structure, and may contribute to the evolution of new protein functions. Therefore, the contradicting properties described of alternative exons are related to different alternative exon populations.

To validate the relation between sequence conservation and splicing regulation, we also analyzed the conservation of sets of regulatory elements independently predicted in human and mouse. Interestingly, ESEs are more conserved in constitutive than in alternative exons. This conservation is also higher for exons with conserved exonic structure (Plass and Eyras, 2006). This result is in agreement with previous analyses of dN and dS values. Furthermore, we found that intronic regions surrounding alternative exons are more conserved than those surrounding constitutive exons, suggesting that regulation of alternative splicing may be subject to the presence of conserved intronic elements as well (Sorek and Ast, 2003; Yeo et al., 2005).

V. CONCLUSIONS

1. The distribution of SR and SR-like proteins is not even across eukaryotes. Some species contain one or more copies of specific proteins whereas others have none.

2. The amount of splicing enhancers found in an organism is related with the conservation of the splicing signals. Moreover, species with more conserved splicing signals have fewer SR proteins or none.

3. The repeat composition of RS domains varies across species. RS domains with high density of SR repeats are observed in species with suboptimal splicing signals.

4. There is an indirect relation between the number of SR repeats in RS domains and the energy of the binding of U2 snRNA to the BS, suggesting a co-evolution of the RS domain and the BS signal.

5. The structure adopted by the pre-mRNA is important for 3′ss recognition in the majority of yeast introns as it maintains the BS and 3′ss at the right distance and hinders the recognition of cryptic 3′ss.

6. Including information from RNA secondary structures into a statistical model to predict 3′ss in yeast allows identifying new 3′ss that can be validated with RNA-Seq reads.

7. The alternative 3′ss validated in yeast are expressed at low levels and are probably degraded through NMD. Nevertheless,

the real impact in the final mRNA levels cannot be resolved without further research.

8. Splicing regulatory signals present in the pre-mRNA impose an extra layer of sequence and structure conservation at the transcript level.

9. The synonymous and non-synonymous substitution rates of an exon have a strong dependence on the inclusion level of the exon and on the conservation of the exonic structure of the transcript it belongs to.

10. The contradicting properties that have been associated to alternative exons, i.e. higher conservation and relaxed selection pressure, can be reconciled if both conservation of transcript structure and inclusion of the exon in transcripts are taken into account.

VI. APPENDICES

Appendix A. Methods to estimate alternative splicing levels

There are several methods to estimate the relative levels of transcripts produced by a single pre-mRNA. Initial genome-wide techniques included the usage of expressed sequence tags (ESTs) and microarrays. Other high-throughput methods were also developed to obtain a global census of RNA molecules, such as serial analysis of gene expression (SAGE) (Velculescu et al., 2000), cap analysis of gene expression (CAGE) (Kodzius et al., 2006), and PCR-based platforms (Brosseau et al., 2010; Hsiao et al., 2010). With the development of new sequencing technologies, RNA sequencing (RNA-seq) has now become the most powerful tool to do such measurements.

ESTs

ESTs are small pieces of DNA sequence that are generated by sequencing expressed genes. First, the RNA of interest is purified from the cell and afterwards converted into cDNA and cloned into a bacterial vector. Next, pieces from one or both ends of the cDNA clones, usually of around a few hundred nucleotides, are sequenced, producing the ESTs (reviewed in Nagaraj et al., 2007). ESTs can give an idea of the particular transcriptome in a cell type and therefore, are very valuable to investigate the relative abundance of alternative transcripts or to discover new transcripts.

ESTs can be used to detect new splicing variants by aligning them directly to the genome and for reconstructing the transcripts they come from (Eyras et al., 2004; Kim et al., 2005; Xing et al., 2006). Moreover, they can also be used to measure the relative abundance of known exons or transcripts. The basic way to use ESTs to measure alternative splicing is to calculate the inclusion level of exons. The inclusion level is defined as the fraction of ESTs validating an exon over the total amount of EST overlapping the genomic locus. Using this measure, it is easy to distinguish constitutive exons, i.e. they appear in all ESTs, from alternative exons, i.e. they only appear in a subset of the ESTs (see for instance Mironov et al., 1999; Brett et al., 2000; Modrek et al., 2001). Moreover, ESTs can also be used to compare relative inclusion of exons by, for instance, comparing the inclusion levels of an exon across tissues or cell types (see for instance Schmitt et al., 1999). For a full review of the usage of ESTs to detect alternative splicing see (Kim and Lee, 2008).

As for any other method, the usage of ESTs has several limitations due to problems arising from both the experimental procedures to obtain them and the bioinformatics analyses used for processing. The main bias in ESTs stems from the variability in the protocols used for creating EST libraries and the low coverage of ESTs in some cell lines or tissues, allowing little comparisons across libraries. Moreover, it has to be considered that PCR and sequencing steps may introduce errors in the sequence. EST databases can also be contaminated with genomic sequences, sequences coming from chimeric transcripts created from PCR artifacts or other RNA/DNA fragments, which will produce wrong predictions. Likewise, during bioinformatics processing there can be mapping problems. Paralogous genes and repetitive sequences can produce multiple alignments of a single EST. Besides, genomic variability and errors in EST sequences can result in wrong alignments (reviewed in Modrek and Lee, 2002; Nagaraj et al., 2007; Kim and Lee, 2008).

Microarrays

Microarrays have also been widely used to measure alternative splicing. In this case, the microarray is prepared with known sequences of interest, which will be hybridized with the RNA or cDNA samples of interest. Later, the array data is analyzed using different strategies, according to the platform and the type of microarray. Microarrays are very useful to measure changes in splicing patterns across tissues or developmental stages, as they are cheaper than sequencing ESTs and very sensitive (Johnson et al., 2003). The problems of microarrays may come from cross-hybridization of RNAs to probes. Additionally, quantification of relative abundance of whole transcripts is difficult. Alternative splicing events are tested independently, and if several transcripts share the same event, it is impossible to know which transcript it belongs to. Besides, microarrays require a prior design of probes. Consequently, they can only detect events related with the probes designed and cannot detect new splicing variants (reviewed in Calarco et al., 2007).

There are several types of microarrays used to measure alternative splicing, including tiling arrays, exon arrays (with single or multiple probes per exon), exon junction arrays, or arrays combining exon probes with exon junction probes (Figure 9) (reviewed in Calarco et al., 2007).

Tiling arrays

Tiling arrays are a type of microarrays in which the probes are designed to cover long contiguous regions of the genome. This is achieved by designing overlapping probes with a fix length that are separated by a fix distance (Figure 9A). The resolution of these arrays depends on the length of the probes and the spacing between them.

The aim of tilling arrays is to detect transcription genome-wide. Several studies have used these arrays to analyze changes in AS in a genome-wide manner, in specific tissues, or in particular development stages (see for instance Kampa et al., 2004; Stolc et al., 2004).

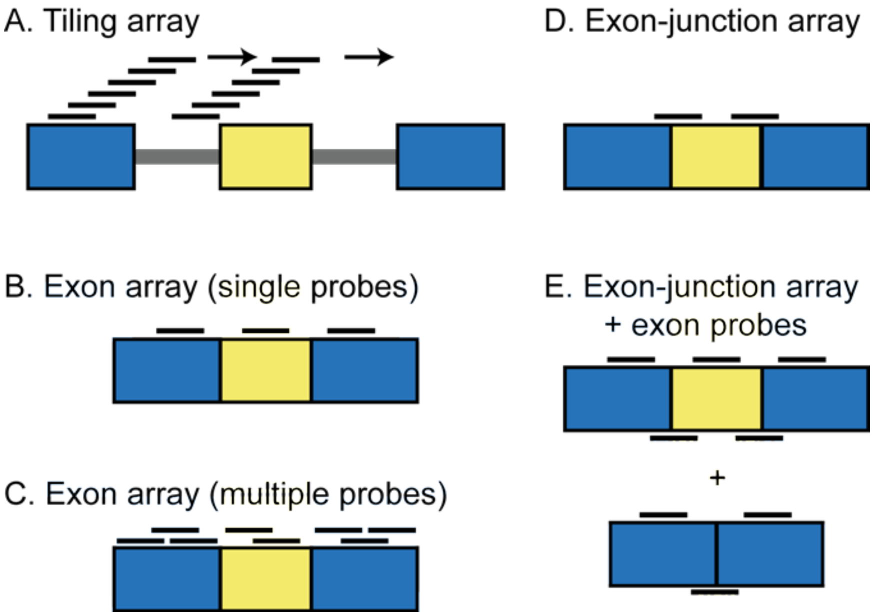


Figure 9. Types of microarrays used to monitor AS. Constitutive and alternative cassette exons are represented by blue and yellow boxes respectively. Oligonucleotide probes, which typically are anchored to glass slides, are marked with black lines. The types of microarrays illustrated are **(A)** genomic tiling arrays; **(B)** single probe exon array; **(C)** multiple probes per exon array; **(D)** junction-specific probes for included exons; **(E)** combination of exon and junction specific probes for included and skipped exons.

Exon arrays

Exon arrays contain one or more exon-probes from a gene (known or predicted). Therefore, they can be used to investigate the differences in abundance of exons in several conditions (Figure 9B and 9C) (see for instance Langer et al., 2010).

Exon junction arrays

Exon junction arrays are more interesting to study AS. In this case, the probes are designed to cover the junction between two particular exons, thus allowing the detection of particular splicing events, such as exon skipping, or alternative 5′ or 3′ss (Figure 9D). In combination with exon probes, these arrays provide better results on the relative abundance of the different isoforms, as junction probes allow distinguishing differences in transcription from differences in splicing (Figure 9E) (see for instance Clark et al., 2002; Johnson et al., 2003; Pan et al., 2004; Fehlbaum et al., 2005).

RNA-Seq

Recently developed high-throughput sequencing technologies, cheaper and faster than Sanger sequencing, have allowed better transcriptome characterization, and thus, better analysis of alternative splicing. In particular, RNA-Seq is based on sequencing short fragments of RNA. The amount of reads sequenced in each experiment is proportional to the original number of molecules in the cell, allowing direct quantification of gene expression (reviewed in Wang et al., 2009). In the simplest protocol, a population of RNAs is purified, fragmented into smaller pieces, and converted to a library of cDNA fragments, which will be later sequenced. For each sample, millions of short reads are obtained, providing enough data to analyze transcript expression more precisely than with previous methods (reviewed in Ozsolak and Milos, 2011).

Several works have used RNA-Seq to identify AS events in various species including human, yeast or even plasmodium

(Sultan et al., 2008; Filichkin et al., 2010; Sorber et al., 2011). Subsequent experiments obtained more accurate measurements using paired-end reads. These reads come from sequencing both ends of an RNA that has been purified and cloned to cDNA. Thus, it has allowed a more effective reconstruction of RNA variants and a better measurement of relative expression levels (Guttman et al., 2010; Katz et al., 2010; Trapnell et al., 2010).

RNA-Seq presents biases, some specific and some common to other high throughput sequencing technologies, like mapping problems due to the presence of polymorphisms, paralogs or repeats in the reference genome. Additionally, fragmentation methods can bias the distribution of reads along the gene and the relative abundance of reads from RNAs of different lengths in the final sample. Moreover, during the PCR step to transform RNAs into cDNAs, sequences can be amplified from already existing cDNAs. This will result in reads that are complementary to the original RNAs or in an increase of particular reads, which will not be proportional to the amount of RNAs in the original sample (reviewed in Wang et al., 2009). Nevertheless, the methods required to quantify precisely different isoforms taking into account existing biases are still an active area of research (Mortazavi et al., 2008; Hiller et al., 2009; Guttman et al., 2010; Katz et al., 2010; Richard et al., 2010; Trapnell et al., 2010; Wang et al., 2010a), suggesting that these problems will be solved in the next few years.

Appendix B. Techniques used to identify splicing regulatory motifs

To understand how splicing works it is crucial to identify all *cis* acting elements involved in its control. In this case, both computational and experimental approaches have been used successfully (reviewed in Chasin, 2007).

Computational methods

Several methods have been designed to perform statistical analysis of genomic data in order to identify motifs that enhance of repress splicing. The idea is that by comparing a set of exons with expected regulated splicing with a control set, motifs related with this regulation could be identified by relative enrichment of nucleotide words of a given length. In general, the sets of exons selected for comparison are expected to be enriched in specific splicing regulators, e.g. splicing enhancers, or be characterized by a lack of them, e.g. unspliced exons. For instance, in one of the first approaches to identify ESEs, the exon set selected consisted of constitutive exons with weak splice sites, which were expected to be enriched in enhancers to compensate the weakness of the splice sites and maintain their constitutiveness. In the same example, constitutive exons with strong splice sites and introns were selected as control set, assuming that these regions would be depleted of ESEs (Fairbrother et al., 2002). The motifs identified by comparing these two sets can be then tested individually to check their ability to enhance or repress splicing and therefore, validate the method. The most important computational methods used to identify splicing regulators are described in the table below.

Table 2. Computational methods to identify splicing regulators.

Author	Method	Motif set
Fairbrother et al., 2002	Identify motifs that are overrepresented in exons with weak splice sites *vs* exons with strong splice sites and introns.	Hexamers functioning as exonic splicing enhancers.
Zhang and Chasin, 2004	Identify motifs that are either overrepresented (PESEs) or underrepresented (PESSs) in internal non-coding exons compared to pseudoexons identified in the 5′UTR region of the same genes and 5′ UTR regions from single exon genes.	Octamers functioning as exonic splicing enhancers (PESEs) or as exonic splicing silencers (PESSs).
Sironi et al., 2004	Identify motifs overrepresented in pseudoexons compared to their flanking intronic regions and annotated exons.	Hexamers functioning as exonic splicing silencers.
Goren et al., 2006	Identify pairs of codons that appear more frequently together than expected by chance and have higher conservation scores in wobble positions.	Hexamers functioning as enhancers or silencers.

Experimental techniques

The experimental techniques have been focused towards the identification of RNA sequences that functionally regulate splicing. Initially, the use of methods involving Systematic Evolution of Ligands by Exponential Enrichment (SELEX) (Ellington and Szostak, 1990) allowed the identification of both, consensus sequences bound by splicing enhancers (Binding SELEX) and sequences that enhance splicing *in vivo* or *in vitro* (Functional SELEX). Another method, based on splicing reporter assays, has also been very useful in the identification of exonic splicing silencers. Interestingly, as it does not perform an enrichment step like SELEX, it allows identifying ESSs with variable strength (Wang et al., 2004). More recently, novel techniques based on purifying proteins bound to RNA coupled to high-throughput sequencing (CLIP-Seq), has allowed the identification of binding sites of more splicing regulators and the identification of their target genes.

Binding SELEX

In protein binding SELEX, a pool of RNA molecules containing short randomized sequences are incubated with a purified RNA binding protein or an RNA binding domain. This process is repeated several times to enrich the RNA molecules in those having high affinity for the RNA binding protein. Using this protocol, the consensus sequences for several SR proteins, hnRNPs and other splicing factors have been identified (Tacke and Manley, 1995; Abdul-Manan and Williams, 1996; Buckanovich and Darnell, 1997; Wang et al., 1997; Cavaloc et al., 1999; Amarasinghe et al., 2001; Kim et al., 2003; Hui et al., 2005).

Functional SELEX

In this technique, a pool of random short RNA sequences goes through several steps of isolation and amplification. In each step, it is tested the ability of the selected sequences to enhance splicing in splicing assays. The enhancer sequences will be selected for the next round. Using this method, the sequences that promoted splicing in response to specific splicing factors, including the SR proteins SRSF1, SRSF2, SRSF5 and SRSF6 were identified (Liu et al., 1998; Smith et al., 2006).

CLIP

The UV Cross Linking and immunoprecipitation (CLIP) method was initially developed to identify *in vivo* RNAs bound by Nova protein in the brain. After purification of the RNAs with an antibody specific for the protein, these RNAs were sequenced and the motif YCAY was identified as the consensus binding site for Nova (Ule et al., 2003). This method coupled to high-throughput sequencing (CLIP-Seq or HITS-CLIP) has allowed identifying binding sites and RNA targets of several splicing factors, including Nova (Licatalosi et al., 2008), FOX2 (Yeo et al.,

2009) and SRSF1 (Sanford et al., 2009). A novel version of this technique that allows single nucleotide resolution, iCLIP, has recently been used to identify targets of hnRNP C (Konig et al., 2010), TIA1 and TIAL1 (Wang et al., 2010b), and TDP-43 (Tollervey et al., 2011).

Appendix C. List of publications

Allo, M., Buggiano, V., Fededa, J.P., Petrillo, E., Schor, I., de la Mata, M., Agirre, E., Plass, M., Eyras, E., Elela, S.A., et al. (2009). Control of alternative splicing through siRNA-mediated transcriptional gene silencing. Nat. Struct. Mol. Biol. 16, 717-724.

Bovine Genome Sequencing and Analysis Consortium, Elsik, C.G., Tellam, R.L., Worley, K.C., Gibbs, R.A., Muzny, D.M., Weinstock, G.M., Adelson, D.L., Eichler, E.E., Elnitski, L., et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324, 522-528.

Koscielny, G., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Riethoven, J.J., Nardone, F., Stanley, E., Fallsehr, C., Hofmann, O., Kull, M., et al. (2009). ASTD: The Alternative Splicing and Transcript Diversity database. Genomics 93, 213-220.

Plass, M., Agirre, E., Reyes, D., Camara, F., and Eyras, E. (2008). Co-evolution of the branch site and SR proteins in eukaryotes. Trends Genet. 24, 590-594.

Plass, M., and Eyras, E. (2006). Differentiated evolutionary rates in alternative exons and the implications for splicing regulation. BMC Evol. Biol. 6, 50.

VII. REFERENCES

Abdul-Manan, N., and Williams, K.R. (1996). hnRNP A1 binds promiscuously to oligoribonucleotides: utilization of random and homo-oligonucleotides to discriminate sequence from base-specific binding. Nucleic Acids Res. 24, 4063-4070.

Abovich, N., Liao, X.C., and Rosbash, M. (1994). The yeast MUD2 protein: an interaction with PRP11 defines a bridge between commitment complexes and U2 snRNP addition. Genes Dev. 8, 843-854.

Allo, M., Buggiano, V., Fededa, J.P., Petrillo, E., Schor, I., de la Mata, M., Agirre, E., Plass, M., Eyras, E., Elela, S.A., et al. (2009). Control of alternative splicing through siRNA-mediated transcriptional gene silencing. Nat. Struct. Mol. Biol. 16, 717-724.

Amarasinghe, A.K., MacDiarmid, R., Adams, M.D., and Rio, D.C. (2001). An in vitro-selected RNA-binding site for the KH domain protein PSI acts as a splicing inhibitor element. RNA 7, 1239-1253.

Amrani, N., Ganesan, R., Kervestin, S., Mangus, D.A., Ghosh, S., and Jacobson, A. (2004). A faux 3'-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. Nature *432,* 112-118.

Baek, D., and Green, P. (2005). Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. Proc. Natl. Acad. Sci. U. S. A. *102,* 12813-12818.

Banerjee, H., Rahn, A., Gawande, B., Guth, S., Valcarcel, J., and Singh, R. (2004). The conserved RNA recognition motif 3 of U2 snRNA auxiliary factor (U2AF 65) is essential in vivo but dispensable for activity in vitro. RNA *10,* 240-253.

Barbosa-Morais, N.L., Carmo-Fonseca, M., and Aparicio, S. (2006). Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. Genome Res. *16,* 66-77.

Batsche, E., Yaniv, M., and Muchardt, C. (2006). The human SWI/SNF subunit Brm is a regulator of alternative splicing. Nat. Struct. Mol. Biol. *13,* 22-29.

Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc. Natl. Acad. Sci. U. S. A. *74,* 3171-3175.

Berglund, J.A., Abovich, N., and Rosbash, M. (1998). A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. Genes Dev. *12,* 858-867.

Bevilacqua, P.C., and Blose, J.M. (2008). Structures, kinetics, thermodynamics, and biological functions of RNA hairpins. Annu. Rev. Phys. Chem. *59,* 79-103.

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annu. Rev. Biochem. *72,* 291-336.

Black, D.L., Chabot, B., and Steitz, J.A. (1985). U2 as well as U1 small nuclear ribonucleoproteins are involved in premessenger RNA splicing. Cell *42,* 737-750.

Blencowe, B.J. (2006). Alternative splicing: new insights from global analyses. Cell *126,* 37-47.

Blencowe, B.J., Bowman, J.A., McCracken, S., and Rosonina, E. (1999). SR-related proteins and the processing of messenger RNA precursors. Biochem. Cell Biol. *77,* 277-291.

Bourgeois, C.F., Lejeune, F., and Stevenin, J. (2004). Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA. Prog. Nucleic Acid Res. Mol. Biol. *78,* 37-88.

Breitbart, R.E., Andreadis, A., and Nadal-Ginard, B. (1987). Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. Annu. Rev. Biochem. *56,* 467-495.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. (2000). EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett. *474,* 83-86.

Brosseau, J.P., Lucier, J.F., Lapointe, E., Durand, M., Gendron, D., Gervais-Bird, J., Tremblay, K., Perreault, J.P., and Elela, S.A. (2010). High-throughput quantification of splicing isoforms. RNA *16,* 442-449.

Buckanovich, R.J., and Darnell, R.B. (1997). The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. Mol. Cell. Biol. *17,* 3194-3201.

Buratti, E., and Baralle, F.E. (2004). Influence of RNA secondary structure on the pre-mRNA splicing process. Mol. Cell. Biol. *24,* 10505-10514.

Caceres, J.F., Screaton, G.R., and Krainer, A.R. (1998). A specific subset of SR proteins shuttles continuously between the nucleus and the cytoplasm. Genes Dev. *12,* 55-66.

Calarco, J.A., Saltzman, A.L., Ip, J.Y., and Blencowe, B.J. (2007). Technologies for the global discovery and analysis of alternative splicing. Adv. Exp. Med. Biol. *623,* 64-84.

Cali, B.M., and Anderson, P. (1998). mRNA surveillance mitigates genetic dominance in Caenorhabditis elegans. Mol. Gen. Genet. *260,* 176-184.

Carlini, D.B., and Genut, J.E. (2006). Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J. Mol. Evol. *62,* 89-98.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C*., et al.* (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. *38,* 626-635.

Cartegni, L., and Krainer, A.R. (2003). Correction of disease-associated exon skipping by synthetic exon-specific activators. Nat. Struct. Biol. *10,* 120-125.

Cavaloc, Y., Bourgeois, C.F., Kister, L., and Stevenin, J. (1999). The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. RNA *5,* 468-483.

Cellini, A., Felder, E., and Rossi, J.J. (1986). Yeast pre-messenger RNA splicing efficiency depends on critical spacing requirements between the branch point and 3' splice site. EMBO J. *5,* 1023-1030.

Chasin, L.A. (2007). Searching for splicing motifs. Adv. Exp. Med. Biol. *623,* 85-106.

Chen, F.C., Wang, S.S., Chen, C.J., Li, W.H., and Chuang, T.J. (2006). Alternatively and constitutively spliced exons are subject to different evolutionary forces. Mol. Biol. Evol. *23,* 675-682.

Chen, S.J. (2008). RNA folding: conformational statistics, folding kinetics, and ion electrostatics. Annu. Rev. Biophys. *37,* 197-214.

Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell *12,* 1-8.

Clark, T.A., Sugnet, C.W., and Ares, M.,Jr. (2002). Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science *296,* 907-910.

Conti, E., and Izaurralde, E. (2005). Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. Curr. Opin. Cell Biol. *17,* 316-325.

Corvelo, A., and Eyras, E. (2008). Exon creation and establishment in human genes. Genome Biol. *9,* R141.

Crick, F. (1970). Central dogma of molecular biology. Nature *227,* 561-563.

Crick, F. (1958). On protein synthesis. Symp. Soc. Exp. Biol. *12,* 138-163.

Ellington, A.D., and Szostak, J.W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature *346,* 818-822.

Ermakova, E.O., Nurtdinov, R.N., and Gelfand, M.S. (2006). Fast rate of evolution in alternatively spliced coding regions of mammalian genes. BMC Genomics *7,* 84.

Eyras, E., Caccamo, M., Curwen, V., and Clamp, M. (2004). ESTGenes: alternative splicing from ESTs in Ensembl. Genome Res. *14,* 976-987.

Fabrizio, P., Dannenberg, J., Dube, P., Kastner, B., Stark, H., Urlaub, H., and Luhrmann, R. (2009). The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. Mol. Cell *36,* 593-608.

Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science *297,* 1007-1013.

Fehlbaum, P., Guihal, C., Bracco, L., and Cochet, O. (2005). A microarray configuration to quantify expression levels and relative abundance of splice variants. Nucleic Acids Res. *33,* e47.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.K., and Mockler, T.C. (2010). Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res. *20,* 45-58.

Fu, X.D., and Maniatis, T. (1992). Isolation of a complementary DNA that encodes the mammalian splicing factor SC35. Science *256,* 535-538.

Gilbert, W. (1978). Why genes in pieces? Nature *271,* 501.

Gooding, C., Clark, F., Wollerton, M.C., Grellscheid, S.N., Groom, H., and Smith, C.W. (2006). A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. Genome Biol. *7,* R1.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. (2006). Comparative analysis identifies exonic splicing regulatory sequences–The complex definition of enhancers and silencers. Mol. Cell *22,* 769-781.

Graveley, B.R. (2005). Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. Cell *123,* 65-73.

Graveley, B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends Genet. *17,* 100-107.

Graveley, B.R. (2000). Sorting out the complexity of SR protein functions. RNA *6,* 1197-1211.

Graveley, B.R., Hertel, K.J., and Maniatis, T. (1998). A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. EMBO J. *17,* 6747-6756.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C.*, et al.* (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. *28,* 503-510.

Hertel, K.J., and Graveley, B.R. (2005). RS domains contact the pre-mRNA throughout spliceosome assembly. Trends Biochem. Sci. *30,* 115-118.

Hiller, D., Jiang, H., Xu, W., and Wong, W.H. (2009). Identifiability of isoform deconvolution from junction arrays and RNA-Seq. Bioinformatics *25,* 3056-3059.

Hiller, M., Huse, K., Platzer, M., and Backofen, R. (2005). Creation and disruption of protein features by alternative splicing – a novel mechanism to modulate function. Genome Biol. *6,* R58.

Hsiao, T.H., Lin, C.H., Lee, T.T., Cheng, J.Y., Wei, P.K., Chuang, E.Y., and Peck, K. (2010). Verifying expressed transcript variants by detecting and assembling stretches of consecutive exons. Nucleic Acids Res. *38,* e187.

Huang, Y., and Steitz, J.A. (2005). SRprises along a messenger's journey. Mol. Cell *17,* 613-615.

Hui, J., Hung, L.H., Heiner, M., Schreiner, S., Neumuller, N., Reither, G., Haas, S.A., and Bindereif, A. (2005). Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. EMBO J. *24,* 1988-1998.

Hurst, L.D., and Pal, C. (2001). Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet. *17,* 62-65.

Iida, K., and Akashi, H. (2000). A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene *261,* 93-105.

Jensen, K.B., Dredge, B.K., Stefani, G., Zhong, R., Buckanovich, R.J., Okano, H.J., Yang, Y.Y., and Darnell, R.B. (2000). Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. Neuron *25,* 359-371.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302,* 2141-2144.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., Tammana, H., and Gingeras, T.R. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res. *14,* 331-342.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods *7,* 1009-1015.

Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. *35,* 125-131.

Kim, N., and Lee, C. (2008). Bioinformatics detection of alternative splicing. Methods Mol. Biol. *452,* 179-197.

Kim, N., Shin, S., and Lee, S. (2005). ECgene: genome-based EST clustering and gene modeling for alternative splicing. Genome Res. *15,* 566-576.

Kim, S., Shi, H., Lee, D.K., and Lis, J.T. (2003). Specific SR protein-dependent splicing substrates identified through genomic SELEX. Nucleic Acids Res. *31,* 1955-1961.

Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H.*, et al.* (2006). Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. Genome Res. *16,* 55-65.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., and Carninci, P. (2006). CAGE: cap analysis of gene expression. Nat. Methods *3,* 211-222.

Kol, G., Lev-Maor, G., and Ast, G. (2005). Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. Hum. Mol. Genet. *14,* 1559-1568.

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. *17,* 909-915.

Kornblihtt, A.R. (2007). Coupling transcription and alternative splicing. Adv. Exp. Med. Biol. *623,* 175-189.

Kress, T.L., Krogan, N.J., and Guthrie, C. (2008). A single SR-like protein, Npl3, promotes pre-mRNA splicing in budding yeast. Mol. Cell *32,* 727-734.

Kriventseva, E.V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M.S., and Sunyaev, S. (2003). Increase of functional diversity by alternative splicing. Trends Genet. *19,* 124-128.

Kupfer, D.M., Drabenstot, S.D., Buchanan, K.L., Lai, H., Zhu, H., Dyer, D.W., Roe, B.A., and Murphy, J.W. (2004). Introns and splicing elements of five diverse fungi. Eukaryot. Cell. *3,* 1088-1100.

Ladd, A.N., Charlet, N., and Cooper, T.A. (2001). The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. Mol. Cell. Biol. *21,* 1285-1296.

Langer, W., Sohler, F., Leder, G., Beckmann, G., Seidel, H., Grone, J., Hummel, M., and Sommer, A. (2010). Exon array analysis using re-defined probe sets results in reliable identification of alternatively spliced genes in non-small cell lung cancer. BMC Genomics *11,* 676.

Lareau, L.F., Brooks, A.N., Soergel, D.A., Meng, Q., and Brenner, S.E. (2007). The coupling of alternative splicing and nonsense-mediated mRNA decay. Adv. Exp. Med. Biol. *623,* 190-211.

Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., and Brenner, S.E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature *446,* 926-929.

Lejeune, F., and Maquat, L.E. (2005). Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. Curr. Opin. Cell Biol. *17,* 309-315.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., Darnell, J.C., and Darnell, R.B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456,* 464-469.

Liu, H.X., Cartegni, L., Zhang, M.Q., and Krainer, A.R. (2001). A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. Nat. Genet. *27,* 55-58.

Liu, H.X., Zhang, M., and Krainer, A.R. (1998). Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev. *12,* 1998-2012.

Long, J.C., and Caceres, J.F. (2009). The SR protein family of splicing factors: master regulators of gene expression. Biochem. J. *417,* 15-27.

Lorincz, M.C., Dickerson, D.R., Schmitt, M., and Groudine, M. (2004). Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. Nat. Struct. Mol. Biol. *11,* 1068-1075.

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., and Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. Cell *144,* 16-26.

Mahen, E.M., Watson, P.Y., Cottrell, J.W., and Fedor, M.J. (2010). mRNA secondary structures fold sequentially but exchange rapidly in vivo. PLoS Biol. *8,* e1000307.

Maniatis, T., and Reed, R. (2002). An extensive network of coupling among gene expression machines. Nature *416,* 499-506.

Manley, J.L., and Krainer, A.R. (2010). A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). Genes Dev. *24,* 1073-1074.

Maquat, L.E. (2006). NMD in mammalian cells: A history. In Nonsense-Mediated mRNA Decay, Maquat, L. E. ed., (Georgetown: Landes Bioscience) pp. 46-58.

Maquat, L.E., and Carmichael, G.G. (2001). Quality control of mRNA function. Cell *104,* 173-176.

Martinez-Contreras, R., Cloutier, P., Shkreta, L., Fisette, J.F., Revil, T., and Chabot, B. (2007). hnRNP proteins and splicing control. Adv. Exp. Med. Biol. *623,* 123-147.

Matlin, A.J., Clark, F., and Smith, C.W. (2005). Understanding alternative splicing: towards a cellular code. Nat. Rev. Mol. Cell Biol. *6,* 386-398.

Mattick, J.S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. Bioessays *25,* 930-939.

Mironov, A.A., Fickett, J.W., and Gelfand, M.S. (1999). Frequent alternative splicing of human genes. Genome Res. *9,* 1288-1293.

Misteli, T., Caceres, J.F., Clement, J.Q., Krainer, A.R., Wilkinson, M.F., and Spector, D.L. (1998). Serine phosphorylation of SR proteins is required for their recruitment to sites of transcription in vivo. J. Cell Biol. *143,* 297-307.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nat. Genet. *30,* 13-19.

Modrek, B., and Lee, C.J. (2003). Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat. Genet. *34,* 177-180.

Modrek, B., Resch, A., Grasso, C., and Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res. *29,* 2850-2859.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5,* 621-628.

Mount, S.M. (1982). A catalogue of splice junction sequences. Nucleic Acids Res. *10,* 459-472.

Munoz, M.J., Perez Santangelo, M.S., Paronetto, M.P., de la Mata, M., Pelisch, F., Boireau, S., Glover-Cutter, K., Ben-Dov, C., Blaustein, M., Lozano, J.J.*, et al.* (2009). DNA damage regulates alternative splicing through inhibition of RNA polymerase II elongation. Cell *137,* 708-720.

Nagaraj, S.H., Gasser, R.B., and Ranganathan, S. (2007). A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform *8,* 6-21.

Nakahata, S., and Kawamoto, S. (2005). Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities. Nucleic Acids Res. *33,* 2078-2089.

Neugebauer, K.M. (2002). On the importance of being co-transcriptional. J. Cell. Sci. *115,* 3865-3871.

Neu-Yilik, G., Gehring, N.H., Hentze, M.W., and Kulozik, A.E. (2004). Nonsense-mediated mRNA decay: from vacuum cleaner to Swiss army knife. Genome Biol. *5,* 218.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., and Ares, M.,Jr. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. *21,* 708-718.

Orban, T.I., and Olah, E. (2001). Purifying selection on silent sites – a constraint from splicing regulation? Trends Genet. *17,* 252-253.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. *12,* 87-98.

Padgett, R.A., Konarska, M.M., Grabowski, P.J., Hardy, S.F., and Sharp, P.A. (1984). Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. Science *225,* 898-903.

Pan, Q., Saltzman, A.L., Kim, Y.K., Misquitta, C., Shai, O., Maquat, L.E., Frey, B.J., and Blencowe, B.J. (2006). Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev. *20,* 153-158.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40,* 1413-1415.

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T.R., Morris, Q.D., Frey, B.J., and Blencowe, B.J. (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol. Cell *16,* 929-941.

Pan, T., and Sosnick, T. (2006). RNA folding during transcription. Annu. Rev. Biophys. Biomol. Struct. *35,* 161-175.

Park, J.W., Parisky, K., Celotto, A.M., Reenan, R.A., and Graveley, B.R. (2004). Identification of alternative splicing regulators by RNA interference in Drosophila. Proc. Natl. Acad. Sci. U. S. A. *101,* 15974-15979.

Parmley, J.L., Chamary, J.V., and Hurst, L.D. (2006). Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol. Biol. Evol. *23,* 301-309.

Patterson, B., and Guthrie, C. (1991). A U-rich tract enhances usage of an alternative 3' splice site in yeast. Cell *64,* 181-187.

Philipps, D., Celotto, A.M., Wang, Q.Q., Tarng, R.S., and Graveley, B.R. (2003). Arginine/serine repeats are sufficient to constitute a splicing activation domain. Nucleic Acids Res. *31,* 6502-6508.

Philipps, D.L., Park, J.W., and Graveley, B.R. (2004). A computational and experimental approach toward a priori identification of alternatively spliced exons. RNA *10,* 1838-1844.

Plass, M., Agirre, E., Reyes, D., Camara, F., and Eyras, E. (2008). Co-evolution of the branch site and SR proteins in eukaryotes. Trends Genet. *24,* 590-594.

Plass, M., and Eyras, E. (2006). Differentiated evolutionary rates in alternative exons and the implications for splicing regulation. BMC Evol. Biol. *6,* 50.

Pleiss, J.A., Whitworth, G.B., Bergkessel, M., and Guthrie, C. (2007). Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components. PLoS Biol. *5,* e90.

Prasad, J., Colwill, K., Pawson, T., and Manley, J.L. (1999). The protein kinase Clk/Sty directly modulates SR protein activity: both hyper- and hypophosphorylation inhibit splicing. Mol. Cell. Biol. *19,* 6991-7000.

Rahman, L., Bliskovski, V., Reinhold, W., and Zajac-Kaye, M. (2002). Alternative splicing of brain-specific PTB defines a tissue-specific isoform pattern that predicts distinct functional roles. Genomics *80,* 245-249.

Richard, H., Schulz, M.H., Sultan, M., Nurnberger, A., Schrinner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., Haas, S.A., and Yaspo, M.L. (2010). Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. Nucleic Acids Res. *38,* e112.

Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z., and Dunker, A.K. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc. Natl. Acad. Sci. U. S. A. *103,* 8390-8395.

Saltzman, A.L., Pan, Q., and Blencowe, B.J. (2011). Regulation of alternative splicing by the core spliceosomal machinery. Genes Dev. *25,* 373-384.

Sammeth, M., Foissac, S., and Guigo, R. (2008). A general definition and nomenclature for alternative splicing events. PLoS Comput. Biol. *4,* e1000147.

Sanford, J.R., Wang, X., Mort, M., Vanduyn, N., Cooper, D.N., Mooney, S.D., Edenberg, H.J., and Liu, Y. (2009). Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome Res. *19,* 381-394.

Schaal, T.D., and Maniatis, T. (1999). Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. Mol. Cell. Biol. *19,* 261-273.

Schmitt, A.O., Specht, T., Beckmann, G., Dahl, E., Pilarsky, C.P., Hinzmann, B., and Rosenthal, A. (1999). Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. Nucleic Acids Res. *27,* 4251-4260.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. Nat. Struct. Mol. Biol. *16,* 990-995.

Schwartz, S.H., Silva, J., Burstein, D., Pupko, T., Eyras, E., and Ast, G. (2008). Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. *18,* 88-103.

Seraphin, B., Kretzner, L., and Rosbash, M. (1988). A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site. EMBO J. *7,* 2533-2538.

Seraphin, B., and Rosbash, M. (1989). Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. Cell *59,* 349-358.

Shen, H., and Green, M.R. (2006). RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. Genes Dev. *20,* 1755-1765.

Shen, H., and Green, M.R. (2004). A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. Mol. Cell *16,* 363-373.

Shen, H., Kan, J.L., and Green, M.R. (2004). Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. Mol. Cell *13,* 367-376.

Shepard, P.J., and Hertel, K.J. (2008). Conserved RNA secondary structures promote alternative splicing. RNA *14,* 1463-1469.

Siliciano, P.G., and Guthrie, C. (1988). 5' splice site selection in yeast: genetic alterations in base-pairing with U1 reveal additional requirements. Genes Dev. *2,* 1258-1267.

Sironi, M., Menozzi, G., Riva, L., Cagliani, R., Comi, G.P., Bresolin, N., Giorda, R., and Pozzoli, U. (2004). Silencer elements as possible inhibitors of pseudoexon splicing. Nucleic Acids Res. *32,* 1783-1791.

Smith, P.J., Zhang, C., Wang, J., Chew, S.L., Zhang, M.Q., and Krainer, A.R. (2006). An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum. Mol. Genet. *15,* 2490-2508.

Sorber, K., Dimon, M.T., and Derisi, J.L. (2011). RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. Nucleic Acids Res.

Sorek, R., and Ast, G. (2003). Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res. *13,* 1631-1637.

Sorek, R., Ast, G., and Graur, D. (2002). Alu-containing exons are alternatively spliced. Genome Res. *12,* 1060-1067.

Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G., and Shamir, R. (2004). A non-EST-based method for exon-skipping prediction. Genome Res. *14,* 1617-1623.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M.F., Rifkin, S.A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P.E., Bussemaker, H.J., and White, K.P. (2004). A gene expression map for the euchromatic genome of Drosophila melanogaster. Science *306,* 655-660.

Sugnet, C.W., Kent, W.J., Ares, M.,Jr, and Haussler, D. (2004). Transcriptome and genome conservation of alternative splicing events in humans and mice. Pac. Symp. Biocomput. 66-77.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D.*, et al.* (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science *321,* 956-960.

Tacke, R., and Manley, J.L. (1995). The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. EMBO J. *14,* 3540-3551.

Tarn, W.Y., and Steitz, J.A. (1997). Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. Trends Biochem. Sci. *22,* 132-137.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. *33,* 201-212.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcarcel, J., and Guigo, R. (2009). Nucleosome positioning as a determinant of exon recognition. Nat. Struct. Mol. Biol. *16,* 996-1001.

Tollervey, J.R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., Konig, J., Hortobagyi, T., Nishimura, A.L., Zupunski, V.*, et al.*

(2011). Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. Nat. Neurosci.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28,* 511-515.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302,* 1212-1215.

Velculescu, V.E., Vogelstein, B., and Kinzler, K.W. (2000). Analysing uncharted transcriptomes with SAGE. Trends Genet. *16,* 423-425.

Wang, J., Dong, Z., and Bell, L.R. (1997). Sex-lethal interactions with protein and RNA. Roles of glycine-rich and RNA binding domains. J. Biol. Chem. *272,* 22227-22235.

Wang, L., Xi, Y., Yu, J., Dong, L., Yen, L., and Li, W. (2010). A statistical method for the detection of alternative splicing using RNA-seq. PLoS One *5,* e8529.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10,* 57-63.

Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N.M., Rot, G., Zupan, B., Curk, T., and Ule, J. (2010). iCLIP predicts the dual splicing effects of TIA-RNA interactions. PLoS Biol. *8,* e1000530.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119,* 831-845.

Warf, M.B., and Berglund, J.A. (2010). Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem. Sci. *35,* 169-178.

Will, C.L., and Lührmann, R. Spliceosome Structure and Function. Cold Spring Harbor Perspectives in Biology

Wu, J.Y., and Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. Cell *75,* 1061-1070.

Wu, S., Romfo, C.M., Nilsen, T.W., and Green, M.R. (1999). Functional recognition of the 3' splice site AG by the splicing factor U2AF35. Nature *402,* 832-835.

Xing, Y., and Lee, C. (2005). Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Proc. Natl. Acad. Sci. U. S. A. *102,* 13526-13531.

Xing, Y., Xu, Q., and Lee, C. (2003). Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. FEBS Lett. *555,* 572-578.

Xing, Y., Yu, T., Wu, Y.N., Roy, M., Kim, J., and Lee, C. (2006). An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. Nucleic Acids Res. *34,* 3150-3160.

Yassour, M., Kaplan, T., Fraser, H.B., Levin, J.Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., Gnirke, A.*, et al.* (2009). Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc. Natl. Acad. Sci. U. S. A. *106,* 3264-3269.

Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.D., and Gage, F.H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat. Struct. Mol. Biol. *16,* 130-137.

Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T., and Burge, C.B. (2005). Identification and analysis of alternative splicing events conserved in human and mouse. Proc. Natl. Acad. Sci. U. S. A. *102,* 2850-2855.

Zahler, A.M., Lane, W.S., Stolk, J.A., and Roth, M.B. (1992). SR proteins: a conserved family of pre-mRNA splicing factors. Genes Dev. *6,* 837-847.

Zamore, P.D., Patton, J.G., and Green, M.R. (1992). Cloning and domain structure of the mammalian splicing factor U2AF. Nature *355,* 609-614.

Zhang, X.H., and Chasin, L.A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev. *18,* 1241-1250.

Zhang, X.H., Kangsamaksin, T., Chao, M.S., Banerjee, J.K., and Chasin, L.A. (2005). Exon inclusion is dependent on predictable exonic splicing enhancers. Mol. Cell. Biol. *25,* 7323-7332.

Zhong, X.Y., Wang, P., Han, J., Rosenfeld, M.G., and Fu, X.D. (2009). SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. Mol. Cell *35,* 1-10.