HOW DID ALTERNATIVE SPLICING EVOLVE?

Gil Ast

Abstract | Alternative splicing creates transcriptome diversification, possibly leading to speciation. A large fraction of the protein-coding genes of multicellular organisms are alternatively spliced, although no regulated splicing has been detected in unicellular eukaryotes such as yeasts. A comparative analysis of unicellular and multicellular eukaryotic 5′ splice sites has revealed important differences — the plasticity of the 5′ splice sites of multicellular eukaryotes means that these sites can be used in both constitutive and alternative splicing, and for the regulation of the inclusion/skipping ratio in alternative splicing. So, alternative splicing might have originated as a result of relaxation of the 5′ splice site recognition in organisms that originally could support only constitutive splicing.

An average human gene is 28,000 nucleotides long and consists of 8.8 exons of ~120 nucleotides that are separated by 7.8 introns[1]. Although the exons are relatively small and embedded within large intron sequences, the splicing machinery recognizes the exons with remarkable precision, removes the introns from the pre-mRNA molecule and ligates the exons to form a mature mRNA. The large number of exons per gene enables the splicing machinery to splice-in different sets of exons from a single pre-mRNA, generating different types of mRNA from a single gene. Bioinformatic analysis indicates that 35–65% of human genes are involved in alternative splicing[2,3], which contributes significantly to human proteome complexity and explains the numerical disparity between the low number of human protein-coding genes (~26,000) and the number of human proteins, the latter of which is estimated to be more than 90,000 (REFS 2,4).

Alternative splicing is important and widespread in some animal groups — but where does it come from? Our understanding of its origins has been limited until recently. However, since the decoding of exon–intron structure of genes in many organisms, and their mode of alternative splicing, two theories have now been proposed — one sequenced based, the other *trans*-factor based. Here, I suggest an evolutionary process for the appearance of alternative splicing, in which the ancestral 5′ splice site (5′ss) signal that only supported constitutive splicing accumulated mutations. The effect of the mutations was to sub-optimize that site, allowing it to be used in alternative splicing as well. The 5′ss that only supports constitutive splicing is found in lower eukaryotic organisms (mostly unicellular organisms such as yeast), whereas the one that supports alternative splicing is found in higher eukaryotic cells (mostly of multicellular organisms). So, might there be a link between the higher orders of complexity in higher organisms and alternative splicing?

Department of Human Genetics and Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel.
e-mail: gilast@post.tau.ac.il
doi:10.1038/nrg1451

An evolutionary overview

The vast majority of introns in eukaryotic gene families are unlikely to have been derived from the most recent common ancestral genes, but were gained subsequently, leading to the formation of multi-intron genes[5,6]. The appearance of multi-intron genes probably predated that of alternative splicing, and constitutive splicing probably predated exon skipping. So, alternative splicing might have originated from multi-intron genes with no alternative splicing, through DNA mutations and/or the evolution of splicing regulatory proteins.

Although there are introns in the genomes of most eukaryotes, alternative splicing is prevalent only in multicellular eukaryotes. The yeast *Saccharomyces cerevisiae* has introns in only ~3% of its genes (~253 introns), and

only 6 genes have 2 introns[7]. By contrast, in another yeast, *Schizosaccharomyces pombe*, 43% of the genes contain introns, with many of them containing multiple introns[8]. However, no alternative splicing has been described in this organism[7]. Unlike the introns in mammals, which are relatively long, the yeast introns are short: only 40–75- nucleotides long in *S. pombe* and 270- nucleotides long in *S. cerevisiae*[7].

The recent sequencing of the genomes of many organisms and their mRNA has facilitated a large-scale analysis of the intron–exon structure and mode of splicing for many genes in a given organism. Evolutionary conservation of a certain sequence among different organisms indicates that the conserved sequences are under PURIFYING SELECTION pressure and might have important functions. So, comparative analysis has recently provided important insights into the ways in which alternative and constitutive sites vary, giving us hints about the steps involved in the evolution of alternative splicing.

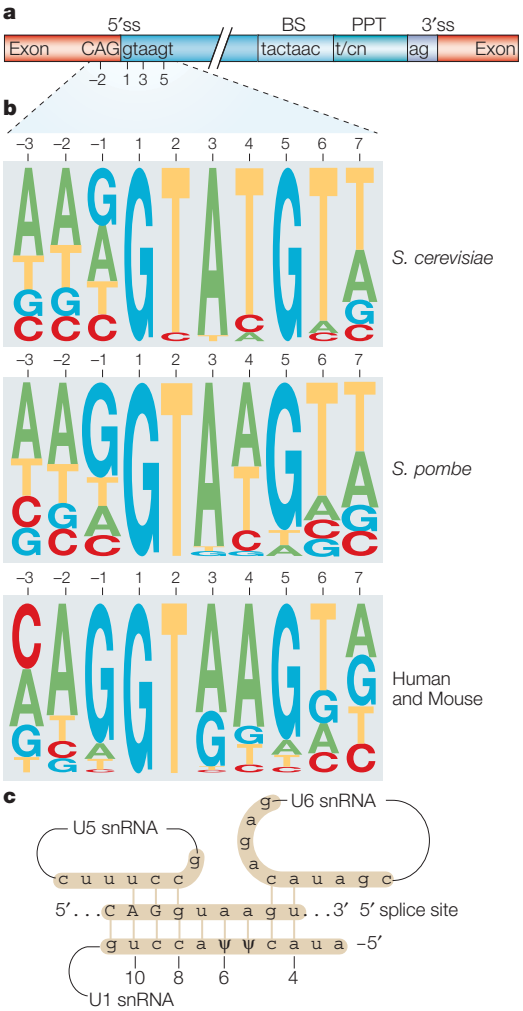

Figure 1 | **The 5′ splice site of yeasts, humans and mice.** **a** | An example of a pre-mRNA intron. The consensus sequences of the 5′ splice site (5′ss), 3′ splice site (3′ss), branch site (BS), and polypyrimidine tract (PPT) are shown. Relative positions upstream and downstream of the 5′ss are indicated underneath and exons are shown in red. **b** | Profiles of 253 *Saccharomyces cerevisiae*, 4,697 *Schizosaccharomyces pombe* and 49,778 human 5′ss are compared. There is no significant difference between the human and mouse 5′ss profile[11]. **c** | Base paring between the 5′ss and U1, U5 and U6 snRNA. Positions of U1 snRNA are shown underneath. Ψ indicates pseudo-uridine. Upper and lower case indicate exonic and intronic sequences, respectively.

PURIFYING SELECTION
Selection against deleterious alleles that arise in a population, preventing their increase in frequency and assuring their eventual disappearance from the gene pool.

SR PROTEINS
A group of highly conserved, serine- and arginine-rich splicing regulatory proteins in metazoans.

hnRNP PROTEINS
A large set of proteins that bind to pre-mRNA.

BRANCH SITE
A splicing signal located upstream of the 3′ end of the intron.

Models for the evolution of alternative splicing

There are at least two possible evolution models of alternative splicing. The first model emphasizes change as a result of mutations in DNA sequences, whereas the second emphasizes the evolution of splicing regulatory factors. The first model suggests that the production of weak splice sites would provide an opportunity for the splicing machinery to skip an internal exon during several splicing events. This gives the cell the potential to produce a new transcript with, perhaps, a new function(s), without compromising the original repertoire of transcripts produced by the gene[9]. It has been shown that alternative exons possess weaker splice sites than constitutively spliced exons[10–13], which allows for suboptimal recognition of exons by the splicing machinery and leads to alternative splicing.

The second model argues that the evolution of splicing regulatory factors, such as SR PROTEINS and hnRNP PROTEINS, applies selective pressure on constitutively spliced exons to become alternative. For example, the binding of SR proteins in proximity to a constitutively spliced exon weakens the selection of that exon, leading to alternative splicing. This releases the selective pressure from the splice sites, resulting in mutations that weaken those splice sites. So, according to this model we should not look at the linear sequence of the pre-mRNA molecule, but rather at the evolution of RNA and protein factors that are involved in the splicing-machinery regulation (this is an adaptation of the Lenny Moss model on evolution of transcription factors[14]).

It is important to realize that the two models do not necessarily contradict one another. The splicing regulatory factor model has not received much experimental attention and remains a possibility only.

The basics of intron splicing

Only four short sequences define an intron: the exon–intron junction at the 5′ and 3′ end of introns (5′ss and 3′ss); the BRANCH-SITE sequence located upstream of the 3′ss; and the polypyrimidine tract located between the 3′ss and the branch site[15] (FIG. 1a). All types of pre-mRNA splicing take place within the spliceosome — a large complex composed of five small nuclear RNA (snRNA) molecules (U1, U2, U4, U5 and U6 snRNA) and as many as 150 proteins[16–18]. Each of the five snRNAs assemble with proteins to form small nuclear ribonuclear protein complexes (snRNP). A coordinated binding of the five snRNPs with the splice signals of the pre-mRNA results in the removal of each intron and the ligation of the flanking exons[15,19,20].

A growing list of spliceosomal proteins provides the basis for positive and negative regulation of constitutive and alternative splicing, which can affect regulation of

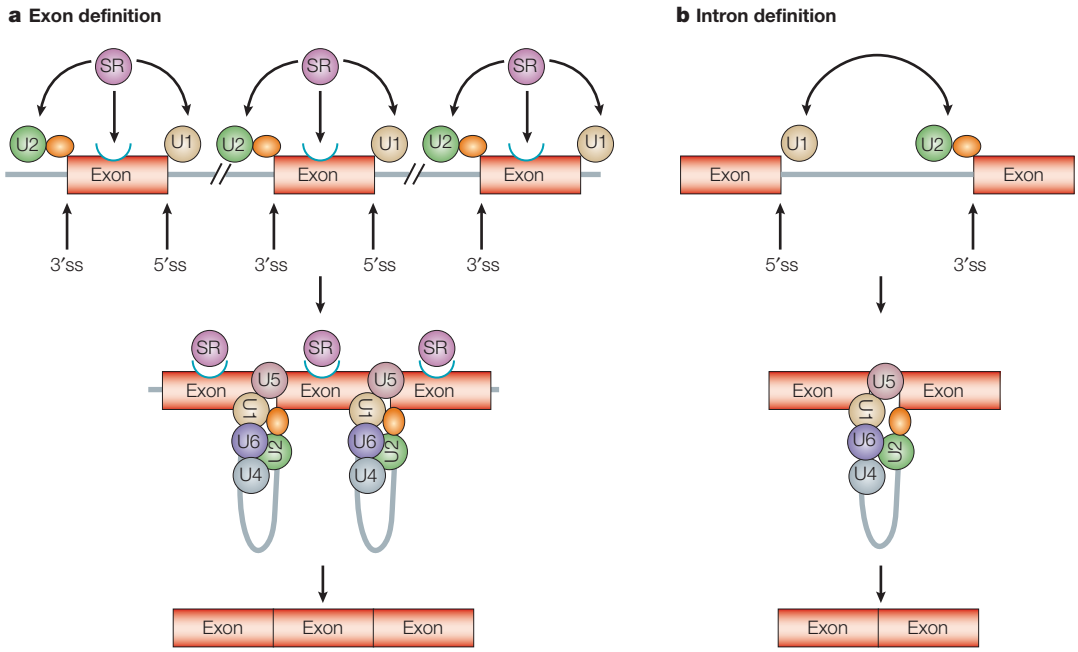

Figure 2 | **Exon and intron definition. a** | Exon definition: SR proteins (purple) bind to exonic splicing enhancers (ESE; blue), recruiting U1 to the downstream 5′ss and the splicing factor U2AF (orange) to the upstream polypyrimidine tract and the 3′ss. U2AF then recruits U2 to the branch site. Therefore, when the SR proteins bind the ESE, they promote formation of a 'cross-exon' recognition complex by placing the basal splicing machinery in the splice sites that flanked the same exon. **b** | Intron definition: the binding of U1 to the upstream 5′ splice site (ss) and U2AF and U2 to the downstream polypyrimidine tract and branch site, respectively, of the same intron. Therefore, intron definition selects pairs of splice sites located on both ends of the same intron, and SR proteins can also mediate this process[20,21,80].

cell cycle, developmental stage, sex determination or a response to an external stimulus[21–23]. In fact, aberrant regulation of alternative splicing has been implicated in an increasing number of human diseases, including cancer[2,22,24,25].

The mRNA splicing mechanism is well conserved throughout evolution and seems to originate from autocatalytic GROUP II INTRONS[26]. The five spliceosomal snRNPs and an unknown number of proteins form the backbone of this conserved mechanism — the basal machinery. Exon and intron recognition is achieved in metazoans by multiple weak degenerate signals, resulting in a network of interactions across exons and/or introns — known as exon definition (ED) and intron definition (ID), respectively[27]. Both exons and introns contain short, degenerate binding sites for splicing regulatory proteins, that is, exonic/intronic splicing enhancers/silencers (ESE, ESS, ISE and ISS). When bound to these short sequences, the SR proteins regulate the binding of the basal machinery to the corresponding splice sites and therefore, are required for both constitutive and alternative splicing[4,21,28]. For example, binding of a SR protein to an ESE can influence both the recruitment of U1 snRNP to the downstream 5′ss and a protein of the basal machinery, U2AF, to the upstream polypyrimidine tract[4,20,21,28]. So, binding of the SR protein to the ESE promotes formation of a 'cross-exon' recognition complex, termed the ED complex[27,29,30]. This complex is found only in metazoans. In unicellular eukaryotes such as yeasts, and for some introns in metazoans, there is another recognition mechanism — the ID mechanism — that defines pairs of splice sites located on both ends of the same intron[27,31] (FIG. 2).

mRNA splicing seems to be controlled at two interconnected levels: the basal and the regulatory levels. The mechanisms by which RNA polymerase II (the basal machinery) and different sets of transcription factors (the regulatory system) control the temporal and spatial activation of each gene are likely to share certain conceptual similarities. ID seems to be the ancient mechanism that allows the recognition of introns embedded in large exonic sequences, which is the case for most of the introns in lower eukaryotic cells[27,31]. The ED mechanism can identify relatively short exon sequences (~120 nucleotides) located within large intron sequences, which is the case for most of the exons in higher eukaryotic cells[1,27,31]. Indeed, mutations in splice sites that are selected via the ID system lead to activation of cryptic splice sites located upstream or downstream of the mutated site. Mutations in splice sites, which are introduced by the ED system, cause a complete cessation of splicing of the exons and lead to exon skipping, which is also the most prevalent form of alternative splicing[21,27,29,30,32].

Types of alternative splicing. There are five major forms of alternative splicing. Exon skipping, also known as cassette exon, accounts for 38% of the alternative splicing events conserved between human and mouse genomes. Alternative 3′ss and 5′ss account for

GROUP II INTRONS
Autocatalytic introns that are found in lower eukaryotic and prokaryotic organisms. These introns posses enzymatic properties that enable them to remove themselves from RNA precursor and ligate the flanking exons.

18% and 8% of the conserved events, respectively. Intron retention is responsible for less than 3% of the alternative splicing events that are conserved in human and mouse genomes. Finally, there are more complex events that account for the remaining 33% of the alternative splicing events and include mutually exclusive events, alternative transcription start sites and multiple polyadenylation sites[35] (FIG. 3).

Characteristics of an alternatively spliced exon. Alternatively spliced exons possess certain features that distinguish them from constitutively spliced ones. Conserved alternatively spliced exons are usually flanked by intronic sequences that are found in both human and mouse genomes, a feature only rarely found in constitutively spliced exons[36].

In the case of exon skipping, both intron regions that flank the exon are conserved, and for alternative 5′ and 3′ splicing events, the conservation is greater near the alternative splice site[35,36]. These conserved intronic sequences are probably involved in the regulation of alternative splicing. Alternative exons that are conserved between the human and mouse possess other characteristics. For example, they tend to be smaller and their length (in nucleotides) is divisible by 3, which distinguish them from constitutively spliced exons[37,38].

Although alternatively spliced exons possess unique features, alternatively and constitutively spliced genes have similar amino-acid usage, indicating that, overall, alternatively and consitutively spliced genes share a high degree of similarity[39]. In fact, in 66% of the alternatively spliced genes the longer form is ancestral, whereas the shorter form is associated mostly with exon skipping. *De novo* emergence of exons, rather than exon duplication, accounts for the other 34% of alternatively spliced genes[40]. This indicates that constitutively spliced exons become alternatively spliced exons (mostly by exon skipping) through evolution.

Defining an intron. The *S. pombe* introns have a degenerate branch-site consensus sequence, CURAY (where R is purine and Y is pyrimidine), similar to that found in mammals. However, in *S. cerevisiae*, the branch-site sequence is highly conserved (UACUAAC). The polypyrimidine-tract architecture is also different in the two organisms[41]. In *S. pombe*, the distance between the 3′ss and the branch-point sequence is particularly short, with an average length of 11 nucleotides. Approximately 75% of the introns in this region contain a polypyrimidine tract[34,42], similar to the polypyrimidine tract found in many vertebrate genes. Because there is no alternative splicing in either *S. pombe* or *S. cerevisiae* (at least, in a classical sense), the differences between the branch-site and polypyrimidine-tract sequences cannot, by themselves, explain how alternative splicing evolved. The degenerate branch site in *S. pombe* and metazoans presumably weakens the binding between this site and the splicing factor — the snRNA U2. This weakening is probably linked to a more predominant function of the polypyrimidine tract in 3′ss selection in metazoans[43]. Nonetheless, additional studies are needed to determine how the splicing machinery defines the 3′ end of introns as strong or weak sites.

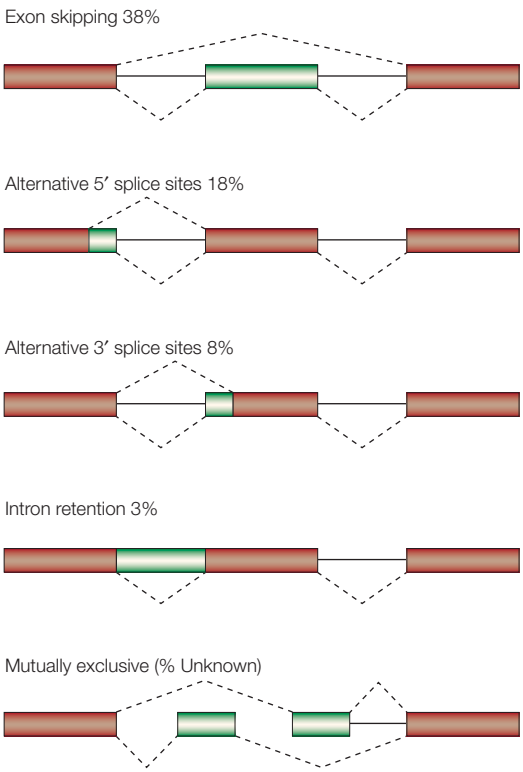

Figure 3 | **Types of alternative splicing.** In all five examples of alternative splicing, constitutive exons are shown in red and alternatively spliced regions in green, introns are represented by solid lines, and dashed lines indicate splicing activities. Relative abundance of alternative splicing events that are conserved between human and mouse transcriptomes are shown above each example (in % of total alternative splicing events[35]).

The splicing factors, the snRNAs and branch-site consensus sequence of *S. pombe* are similar to those of mammals, rather than to *S. cerevisiae*, which, together with the presence of SR-like proteins and proteins that are involved in ED in *S. pombe*, but not in *S. cerevisiae*[33,34], indicate that *S. pombe* is 'on the verge' of acquiring alternative splicing capabilities. So, why does *S. pombe* stick with default splicing instead of taking the plunge? It might be that the extremely short introns that are spliced only through ID are the missing link[31]. ED of multicellular organisms could provide an opportunity for the splicing machinery to skip an internal exon on several splicing events[27].

The age of the exon. Human–mouse comparative analysis revealed that alternative splicing is often associated with recent exon creation and/or loss[44,45]. So, alternative splicing has the potential to create species-specific alternatively spliced exons. Two processes are known to create new exons that are often alternatively spliced: exon duplication[46,47] and exonization of intronic sequences[46,48–50].

Young, alternatively spliced exons that are at an early stage in their evolution have the potential to provide the

minimal conditions required to regulate their splicing pattern. More than 5% of the alternatively spliced internal exons in the human genome are derived from *Alu* elements[50]. *Alu* elements are short primate-specific RETROTRANSPOSONS, of which humans have ~1.4 million copies, more than 500,000 of which are located in introns[12,48]. As far as we know, all alternatively spliced *Alu* exons were created exclusively through the exonization of intronic elements. An examination of minimal conditions that lead an intronic element to become an alternatively spliced exon revealed that, remarkably, the only selective pressure was creation or maintenance of weak splice sites that flank the alternatively spliced *Alu* exon[12,48,50].

Comparative approaches

Comparative analyses of splice sites between organisms that only have constitutive splicing and those that also have alternative splicing provide important clues regarding the minimal conditions required for alternative splicing.

Unicellular organisms — *S. pombe* and *S. cerevisiae* — are estimated to have diverged into two separate lineages about 370 million years ago and from metazoa more than 1,000 million years ago[51]. In the case of humans and mice, 75–130 million years has passed since divergence from the common ancestor[52,53]. Comparative analysis of *S. pombe*, *S. cerivisiae*, human and mouse splicing factors shows a higher degree of similarity between *S. pombe* and mammals than between *S. pombe* and *S. cerevisiae*[34]. There are, therefore, many significant differences among *S. pombe*, *S. cerevisiae* and mammals, with regard to both the number of introns per gene and the ability to support alternative splicing.

Comparative analysis of 5′ss. The 5′ss reveals major differences among *S. cerevisiae*, *S. pombe* and mammals. In *S. cerevisiae*, the first six intronic nucleotides are well conserved (GTATGT). There are several deviations from that sequence between *S. cerevisiae*, *S. pombe* and mammals (FIG. 1b): the conservation level of the G at position –1 increases from 37% to 55% to 80% between *S. cerevisiae*, *S. pombe* and humans, respectively (there was no significant difference between the human and mouse 5′ss profile[11]; supplementary S1 (figure and table) shows the per position conservation of each nucleotide among the 4 species). There is a gradual change at position 4, from T in *S. cerevisiae*, to an A or a T in *S. pombe* and to predominantly A in mammals. The conservation level of A at position –3 decreases from 53% in *S. cerevisiae*, to 45% in *S. pombe* and 34% in mammals. Also, only ~16% and 46% of C at position –3 and A at position –2, respectively, is conserved in *S. cerevisiae* and *S. pombe* 5′ss, whereas 36% and 64%, respectively, is conserved in mammals. Finally, the conservation level of A, G, and T at positions 3, 5 and 6, respectively, decreases from *S. cerevisiae* to *S. pombe* to mammals.

Although a comparison of additional multicellular organisms is needed to examine whether this is a *bona fide* evolutionary process, the four-species-comparison shown in FIG. 1b indicates that the level of conservation of three positions in the intronic portion of the 5′ss (positions 3, 4, and 6) decreases in the order of *S. pombe*, *S. cerivisiae*, mice and humans, whereas conservation of the last three positions of the exon (positions –1 and –3) increases in the order *S. cerevisiae*, *S. pombe*, mice and humans[11,33,54].

Conservation of intron removal. A similar molecular mechanism removes introns from pre-mRNA in all eukaryotes[15]. During mRNA splicing, three snRNAs can base pair with the 5′ss. U1 snRNA forms base pairs across the intron–exon junction (potentially base pairing at positions –3 to 6). The conservation of the exonic portion of the 5′ss in vertebrates allows U1 to base pair with that region[11,54]. However, in yeasts, this region is less conserved. Therefore, the base pairing of U1 with the exonic portion in *S. cerevisiae* — although demonstrated experimentally[55] — is probably not of principal regulatory importance in U1/5′ss binding.

Before the first catalytic step of splicing, U1 is replaced by U5 and U6; the invariant loop of U5 snRNA can potentially base pair with positions –3 to 1, and likewise, U6 snRNA can potentially base pair with positions 5 and 6 (FIG. 1C). An A or C at position –3 can base pair with U5 or U1 respectively, indicating that the decrease in the conservation level of A and appearance of a conserved C at position –3 between *S. cerevisiae*, *S. pombe* and mammals is indicative of the expansion of the U1/5′ss binding to the exonic portion of the 5′ss[11]. Furthermore, the base pairing of the invariant loop of U5 with the exonic portion of the 5′ss, which is essential for the second catalytic step of splicing in *S. cerevisiae*, is dispensable for *in vitro* splicing in human nuclear extract[56,57]. Despite the differences in the 5′ss between yeasts and mammals, there is no change in the sequence of the 5′ end of U1 snRNA gene (the binding site to the 5′ss) among these organisms[58,59].

Differences in U1/5′ss binding. The rigid 5′ss sequence in *S. cerevisiae* provides six potential sites that can base pair with U1, all located in the intronic portion of the 5′ss (a U·Ψ pairing of the U in position 4 of the 5′ss with the Ψ in position 5 of U1 snRNA is considered as one pairing; where Ψ is pseudo-uridine). In metazoans, however, the 5′ss provides nine potential positions for U1 binding, but only seven are involved in base pairing with a typical 5′ss[11]. U1/5′ss pairing in metazoans includes Watson–Crick base pairings (G·C and A·T), as well as non-Watson–Crick pairings (G·U and U·U). The seven nucleotides that are involved in base pairing with U1 in an average 5′ss are presumably a combination of nucleotides that maintains the base pairing of U1 above a certain minimal number; 5–6 nucleotides can provide the minimal 5′ss signal and also the minimal binding site for U1, but without surpassing a certain maximum that might lead to a strong U1/5′ss binding that reduces the efficiency of the splicing reaction (>8 pairings)[60].

The extension of the conserved sequence to the exonic portion of the 5′ss in metazoans is directly linked to U1 binding to both the exonic and intronic portions of the 5′ss, as shown both experimentally and by using

RETROTRANSPOSONS
A mobile genetic element; its DNA is transcribed into RNA, which is reverse-transcribed into DNA and then is inserted into a new location in the genome.

bioinformatics tools. A 'see-saw' effect can be observed, in which a higher number of base pairings of U1 snRNA with the exonic portion of the 5′ss is linked to a lower number of base pairings with the intronic portion, and *vice versa*. Also, a dependency of positions –1 and –2 on 5 indicates that the conservation of the exonic portion of the 5′ss is directly related to U1 binding[11,61].

Hard- and soft-wired organisms

Alternative splicing might be one of the ways in which organisms evolve in a more rapid and dynamic fashion; and hard- and soft-wired organisms are defined as those without and with this ability, respectively[62]. Let us examine this model with respect to the molecular differences between 5′ss of hard- and soft-wired organisms, such as yeasts and mammals.

There have been several reports of alternative splicing in unicellular organisms: for example, in *S. cerevisiae*[63,64], *S. pombe*[65], *Plasmodium falciparum* and *Dictyostelium discoideum*[66]. Most of the above cases are unspliced mRNA (that is, intron retention) — they involve shuttling of an unspliced pre-mRNA from the nucleus to the cytoplasm. Intron retention is, however, only a minor form of alternative splicing in multicellular organisms (less than 3% of the cases). Exon skipping, which is the prevalent form of alternative splicing in multicellular organisms, was not reported in unicellular organisms. The alternative splicing events listed above, therefore, might represent isolated cases and/or mis-splicing, indicating that alternative splicing is rare or non-existent in yeasts, whereas a large fraction of the protein-coding genes of multicellular organisms are alternatively spliced.

The absence (or rarity) of alternative splicing in unicellular organisms does not necessarily mean this is the primitive state; it could, for example, be a derived state; in yeast it could reflect streamlining of the genome or the lifestyle. Alternative splicing is not only an opportunity but also a risk (for example, of mis-splicing, inefficiency and extra genetic burden) and therefore, is perhaps a luxury that the fast-growing unicellular organisms cannot afford. *S. cerevisiae* might have lost a key protein that is required for dealing with multiple introns. Although *S. pombe* would still have this protein and therefore not be forced to dispense with multiple introns, both yeasts would have dispensed with the alternatively splicing machinery. It was proposed that the presence of introns in a minority of yeast genes, and always either at the 5′ or 3′ end of the genes, means that, in the past, all the genes had introns. However, the efficient recombination in yeast erased the introns by GENE CONVERSION from cDNA except where there was not enough homology to allow for recombination[67], or when the intron retained its function[68]. For this reason, I compare yeast and mammal 5′ss from the point of view of organisms without and with alternative splicing ability.

Every base pair counts

The 5′ss of yeast provides six rigid, constitutive positions of base pairing with U1 (all located in the intronic portion of the 5′ss), which presumably mark the location of the splice site at the 5′ end of a STEM STRUCTURE with U1. The average number of base pairings with U1 in mammals is, on average, one greater than that in yeast. However, recent findings indicated that there is a complex mechanism by which the sequence of the 5′ss (especially at positions 3 and 4) and its base pairing with U1 govern both alternative and constitutive splicing and the amount of skipping/inclusion at that site[12]. Positions 3 and 4 are located between two regions that form strong base pairing with U1; in almost 80% of the 5′ss analysed, positions –1 to 2 and position 5 are GGT and G, respectively. The base pairing of positions –1 to 2 with U1 probably provide the anchor for U1 binding to the 5′ss, which is then anchored again by base pairing with G in position 5. This structure 'traps' positions 3 and 4 from both sides with strong Watson–Crick pairings and allows the positions to form non-Watson–Crick pairings (such as G·U and U·U). (A non-Watson–Crick pairing can only form when it is adjacent to a Watson–Crick pairing[69].) The hierarchy of pairing, A·T > G·T > T·T ≠ C·T, regulates the level of usage of this splice site in mRNA splicing, with A·T pairs encouraging constitutive usage, G·T and T·T pairs supporting different levels of skipping/inclusion and C·T pairs leading mostly to exon skipping[12]. The plasticity of a 5′ss, therefore, lies in the type of pairing with U1 (FIG. 4). Indeed, different HEMIASCOMYCETOUS YEASTS have deviations from the canonical GTATGT 5′ss found in *S. cerevisiae*: GTAAGT, GTGAGT and GTAGGT (REF. 33).

These deviations affect only positions 3 and 4, and although one Watson–Crick pairing (A·T) is maintained, the other becomes a non-Watson–Crick pairing (G·T or T·T). One exception to this is the GTAAGT 5′ss for which both positions form a Watson–Crick pairing. This highlights an interesting point: the importance of Ψ·U pairing in mRNA splicing between position 4 of the 5′ss of *S. cerevisiae* and position 5 of U1 was recently reported[70], which indicates the importance of a non-Watson–Crick pairing in either position 3 or 4, presumably to trigger unwinding of U1/5′ss binding — an essential step in mRNA splicing.

Other positions might also contribute to the plasticity of the 5′ss. Positions 6 and –2 are located adjacent to a prominent site that forms a G·C pairing with U1, and might, therefore, form a non-Watson–Crick pairing with U1. Substituting T with C (at position 6 of the 5′ss of exon 20 of the human *IKBKAP* gene; inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase complex-associated protein) causes exon 20 to shift from constitutive to alternative splicing. This leads to familial dysautonomia (FD), an autosomal recessive congenital neuropathy[11]. The shift to alternative splicing, owing to mutation at position 6, indicates the involvement of that position in generating plasticity in the 5′ss. Although the mutation in this 5′ss is deleterious, it is probably part of the general nature of mutations to induce genomic diversity — a constant quest to find the best conditions for the organism's survival in a given environment.

Based on these findings, we can assume that some of the mutations that accumulated in constitutively spliced 5′ss lead to inactivation of that site and to exon loss —

GENE CONVERSION
A non-reciprocal recombination process that causes one sequence to be converted into the other.

STEM STRUCTURE
A region of base pairing between two stands of RNA or DNA.

HEMIASCOMYCETOUS YEASTS
A group of yeast that includes S. cerevisiae and at least 13 other yeasts species that have a small genome size and a low frequency of introns.

a process that generates species-specific transcripts[44]. Others might lead to the sub-optimal recognition of that site (a weak splice site) causing that exon to become alternatively spliced. This type of mutation could cause genetic disorders, presumably by reducing the concentration of the protein in the cell below a crucial level, as in the case of FD. The newly created alternatively spliced exon could also present an advantage to the organism. In this case, the additional transcript acquires a new function(s) or, at least, it is not deleterious. This is a way of enriching the transcriptome and enhancing the coding capacity and regulatory versatility of the genome with new isoforms, without compromising the integrity and original repertoire of the transcriptome and its resulting proteome.

Some of the weak splice sites require regulatory sequences (exonic/intronic splicing enhancer/suppressor) that are located outside the splice site, to which regulatory proteins (such as SR and hnRNP proteins) can bind and determine if that splice site is used in either a constitutive or alternative manner[19,21]. The idea that alternatively spliced exons contain splice sites that have weaker binding than constitutive sites was recently supported by experimental evidence: the free energy of U1 binding to constitutively spliced exons is –6.53 kcal mol^{-1} (REF. 11), compared to –5.2 kcal mol^{-1} in alternatively spliced exons[12], indicating that constitutively spliced mammalian 5′ss bind more tightly to U1 snRNA than alternatively spliced exons (lower free energy indicates stronger binding).

An alternatively spliced exon that inserts or removes an entire sequence from a protein, without compromising the integrity of the reading frame of the region located downstream of that exon, is less likely to be deleterious. Exons whose length is a multiple of three nucleotides are therefore candidates for alternative splicing, because skipping that exon does not change the reading frame of the downstream sequence. Furthermore, it has recently been reported that alternatively spliced exons (which are part of the reading frame) display a bias towards multiples of three nucleotides[37,38].

Extension of 5′ss conservation

Why is the exonic portion of the 5′ss well conserved in metazoans but less so in yeasts? One can argue for an evolutionary model based on conservation that has been shaped by the splicing machinery requirements.

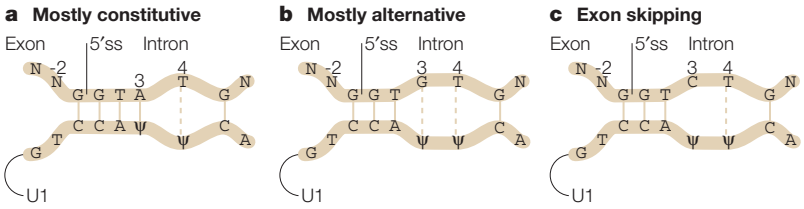

Figure 4 | **Base paring between different types of 5′ss and U1 snRNA.** Positions 3 and 4 of the 5′ss are different between the panels; stacking energy is illustrated by distances between nucleotides that base pair with U1 snRNA. Solid and dashed lines indicate Watson–Crick and non-Watson–Crick pairing, respectively. 'N' indicates unspecified nucleotides, Ψ indicates pseudo-uridine. **a** | Mostly constitutive. **b** | Mostly alternative. **c** | Exon skipping.

The 5′ss of the ancestral introns closely resembles those that we find in hard-wired unicellular organisms, such as yeasts. The first six positions of the intron are well conserved and provide a strong splice-site signal to the splicing machinery. This strong signal can support only constitutive splicing. But as mutations accumulate in that 5′ss during evolution, the pairing between certain positions and U1 change from Watson–Crick to non-Watson–Crick. For these mutations to give an advantage to the organism, they must not be deleterious, they must be located no more than one position away from a Watson–Crick base pairing with U1 and — if they weaken the strength with which the 5′ss binds U1 to below a certain minimal value — they must be compensated for by other mutations that strengthen U1 binding. The compensatory mutations could be in positions 7 and 8 or –1 to –3, which are the only two regions that have the potential to form additional base pairing with the 5′ end of U1 snRNA. In *S. cerevisiae*, positions 7 and 8 can compensate for the loss of a base pair, which is probably only the case in mammals for a minor subset of introns[11,71]. However, substitutions have mainly occurred at positions –1 to –3 — presumably to enhance marking the location of the intron–exon junction from the edge of a region that base pairs with U1 in yeasts — to between two nucleotides that base pair tightly with U1.

So, the exonic portion of the 5′ss is under two evolutionary constraints — conservation of protein coding and also of splice-site signal. Therefore, we can assume that in almost 80% of the 5′ss in mammals, selective pressure has led to a mutated G at position –1. In 64% and 35% of 5′ss an A and a C at positions –2 and –3 was substituted, respectively[11,54]. We can also assume that the mutations were not detrimental to the organism's survival.

There might be an alternative explanation for the conservation of the exonic portion of the 5′ss. According to the 'proto-splice' site model, introns can be inserted into a target sequence of (C/A)AGG, so that (C/A)AG and G become the flanking 5′ and 3′ exonic sequences[72]. It is tempting to assume that the conservation of the exonic portion of the 5′ss is related to the invasive nature of introns and not to the splicing mechanism. Presumably introns can invade multiple exonic sites, but will finally 'settle down' in sites that best support their splicing mode — those containing a (C/A)AG in the exonic portion of the 5′ss[73]. However, the most compelling evidence for the importance of the exonic portion of the 5′ss in mRNA splicing came from the exonization of *Alu* elements. In most *Alu* exons, the selected 5′ss contain CAG in the exonic portion, except in 4% when it is TAG (REF. 12). Because *Alu* exons originated from exonization of intronic sequences and are not related to the insertion of introns into a target sequence, the conservation of the exonic portion of the 5′ss of exonized *Alu* elements therefore, is directly related to the splicing machinery, probably to U1 binding.

The substitution to G at position –1 is probably the most prominent because it provides three constitutive positions that can base pair with U1 (–1 to 2). The free energy that is obtained from binding of U1 to these

three nucleotides is lower (indicating stronger binding) than the sum of each base pair alone (STACKING ENERGY). Mutations in the last nucleotide of the exon are least deleterious when the intron is located between codons (phase 0), so that position –1 is the WOBBLE SITE. This allows substitution to G without affecting the type of amino-acid synthesis from that codon in a large portion of cases. 64% of alternatively spliced exons end in phase 0 (REFS 37,74,75). Selection for A and C at positions –2 and –3 was observed only when the next nucleotide could also base pair with U1, indicating the need for a minimum of at least two adjacent positions that base pair with U1 (REF. 11). The fact that, in humans and mice, mutation at position 5 reduces the likelihood for other mutations at positions –1 and –2, and *vice versa*, indicates that the conservation reflects the binding of one splicing factor across the 5′ss motif — which can only be U1 snRNA. This is not the case in *S. cerevisiae*, in which conservation of AAG in the last three exonic nucleotides is low[11,33,54].

The invariant loop of U5 snRNA can base pair with the last nucleotides of the exonic portion of the 5′ss during mRNA splicing. In *S. cerevisiae*, the invariant loop is essential for the second step of splicing, although in humans it is dispensable for the entire reaction[57,58]. This indicates that the extension of U1 into the exonic portion of the 5′ss in humans and mice (and presumably in all metazoans) relinquishes the need for U5 snRNA base pairing with this region in 5′ss selection. We can therefore argue for gradual molecular evolutionary changes that turn an intronic 5′ss into an exonic–intronic site. Such an extension can provide both a signal that is sufficient for the recognition of that site by the splicing machinery and the plasticity that is needed for regulated splicing. Therefore, the plasticity is, in part, the sum of the binding affinity of U1 snRNA, and probably of other splicing factors to that site, which determines whether it is a strong or weak site. Furthermore, increased flexibility owing to mispairing at the base of the domain that forms RNA–RNA interactions downstream of the 3′ss (IBS3-EBS3) was recently reported as the reason for alternative 3′ss in autocatalytic group II introns[76]. This lends further support to the hypothesis that alternative splicing evolved by increasing the flexibility of RNA based interactions — such as U1/5′ss binding.

Weakening strong splice sites

The only selective pressure that was found in the exonization of *Alu* elements was in creating or maintaining weak splice sites, which supports the hypothesis that alternative splicing evolved by turning strong splice sites into weak sites.

Alternative splicing of the transposable element *Restless* is the only known case of exon skipping in fungi. The regulation of this alternative splicing is similar to that of most of the *Alu* exons, namely, through weak 5′ and 3′ splice sites that regulate the inclusion/skipping ratio[12,78]. The *Restless* cassette exon and some *Alu* exons (such as exon 8 of the *ADARB2* gene see Online links) contain similar splicing regulatory sequences — for example, a 3′ss motif of **G**AGACAG led to the selection of the distal AG (underlined). In this motif, there is a delicate interplay between the two AGs. The G at position –7 (bold) suppresses the selection of the proximal AG. However, the proximal AG is essential for the weakening of the selection of the distal AG, and so maintains alternative splicing[48].

The other splicing regulatory sequence is the 5′ss. It is a weak site because the intron begins with GC and alternative splicing is maintained owing to the unpairing of U1 with the C at position 2 of the 5′ss[12]. It is important to note that more than 98% of human introns begin with GT, whereas only ~0.7% begin with GC. The latter were shown to be frequently involved in alternative splicing and probably evolved as a result of a T to C mutation of position 2 of a canonical GT 5′ss[1,12,78,79].

The regulation of the *Restless* cassette exon seems to be the most ancient form of controlling the exon inclusion/skipping ratio in alternative splicing. Remarkably, it is almost identical to the way by which new alternatively spliced exons are regulated in the human genome — it depends almost solely on the sequence composition of the 3′ and 5′ss. This similarity further supports the hypothesis that alternative splicing might have originated by relaxation of the splice site recognition.

Conclusions

Based on what is known so far, we can predict that the appearance of multi-intron genes predated the appearance of alternative splicing. Alternative splicing probably evolved following a combination of mutations in splice sites that generated sub-optimal recognition of the sites by the basal splicing machinery (such as U1 binding to the 5′ss), evolution of protein splicing factors that can identify short exons in the multi-intron genes and the placement of the basal splicing machinery in the flanking splice sites across the same exon (ED). Sub-optimal recognition of the exon subsequently generated exon skipping, which is the prevalent form of alternative splicing. Other types of alternative splicing, such as alternative 5′ and 3′ss — although not necessarily intron retention and mutually exclusive events — are probably a specific adaptation of that mechanism. Such conclusions are based on the observations that, in exon skipping, both intron regions that flank the exon are conserved, and for alternative 5′ and 3′ splicing events, the conservation is limited to the alternative splice site[35,36]. These observations suggest that alternative 5′ and 3′ss are a subgroup of the prevalent form of alternative splicing — exon skipping.

What is required for a constitutively spliced exon to become alternatively spliced and to be conserved as an alternatively spliced exon among different organisms? There are certain features that characterize alternatively spliced conserved exons: conservation of both intron regions flanking the exon, a smaller size and divisibility by three. The conservation of the flanking intronic sequences suggests that part of the splicing regulatory sequences for that exon needs to reside outside the exon sequence. The smaller size of the exon is probably part of the sub-optimal recognition of that exon by the ED system, which presumably

STACKING ENERGY
Energy contributions from base pair stacking.

WOBBLE SITE
Pairing between the codon and anticodons of tRNA at the last codon position. Wobble enables the anticodon base to form hydrogen bonds with bases other than those in standard base pairs.

works best for internal exons of ~120 nucleotides. Finally, the divisibility by three ensures that the removal of the exon will not change the reading frame for the rest of the protein.

The evolutionary aspects of the splicing regulatory proteins, SR and ED proteins, gain little attention. SR-like proteins that were found in *S. pombe*, but not in *S. cerevisiae*[31], raise many questions. What is the original function of SR-like proteins in hard-wired organisms? When do SR proteins begin to influence splice-site selection? Do SR proteins first influence selection of a given exon, which subsequently reduces the selective pressure from the hard-wired splice sites, allowing them to mutate rapidly? Can molecular evolution of the SR protein explain, by itself, how alternative splicing evolved? All these questions remain to be addressed.

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nature Genet.* **30**, 13–19 (2002).
3. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. Frequent alternative splicing of human genes. *Genome Res.* **9**, 1288–1293 (1999).
4. Woodley, L. & Valcarcel, J. Regulation of alternative pre-mRNA splicing. *Brief. Funct. Genomic. Proteomic.* **1**, 266–277 (2002).
5. Qiu, W. G., Schisler, N. & Stoltzfus, A. The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol. Biol. Evol.* **21**, 1252–1263 (2004).
The authors suggest that introns were inserted to genes during evolution.
6. Coghlan, A. & Wolfe, K. H. Origins of recently gained introns in *Caenorhabditis*. *Proc. Natl Acad. Sci. USA* **101**, 11362–11367 (2004).
7. Barrass, J. D. & Beggs, J. D. Splicing goes global. *Trends Genet.* **19**, 295–298 (2003).
8. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880 (2002).
9. Graur, D. & Li, W.–H. *Fundamentals of Molecular Evolution* (Sinauer Associates, Sunderland, 1999).
10. Stamm, S., Zhang, M. Q., Marr, T. G. & Helfman, D. M. A sequence compilation and comparison of exons that are alternatively spliced in neurons. *Nucleic Acids Res.* **22**, 1515–1526 (1994).
11. Carmel, I., Tal, S., Vig, I. & Ast, G. Comparative analysis detects dependencies among the 5′ splice-site positions. *RNA* **10**, 828–840 (2004).
Comparative analysis of 50,000 human and mouse homologous 5′ss reveal that the 5′ss provide 9 potential positions for U1 binding, but only 7 are involved in base pairing with a typical 5′ss.
12. Sorek, R. *et al.* Minimal conditions for exonization of intronic sequences; 5′ splice site formation in *Alu* exons. *Mol. Cell* **14**, 221–231 (2004).
A demonstration that the type of base pairing between the 5′ss and U1 regulates alternative versus constitutive splicing and also controls the inclusion/skipping ratio in alternative splicing.
13. Lear, A. L., Eperon, L. P., Wheatley, I. M. & Eperon, I. C. Hierarchy for 5′ splice site preference determined *in vivo*. *J. Mol. Biol.* **211**, 103–115 (1990).
14. Moss, L. *What Genes Can't Do* (MIT Press, Cambridge, USA., 2003).
15. Brow, D. A. Allosteric cascade of spliceosome activation. *Annu. Rev. Genet.* **36**, 333–360 (2002).
16. Hartmuth, K. *et al.* Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method. *Proc. Natl Acad. Sci. USA* **99**, 16719–16724 (2002).
17. Zhou, Z., Licklider, L. J., Gygi, S. P. & Reed, R. Comprehensive proteomic analysis of the human spliceosome. *Nature* **419**, 182–185 (2002).
18. Jurica, M. S. & Moore, M. J. Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell* **12**, 5–14 (2003).
19. Graveley, B. R. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17**, 100–107 (2001).
20. Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416**, 499–506 (2002).
21. Cartegni, L., Chew, S. L. & Krainer, A. R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet.* **3**, 285–298 (2002).
22. Stoilov, P. *et al.* Defects in pre-mRNA processing as causes of and predisposition to diseases. *DNA Cell Biol.* **21**, 803–818 (2002).
23. Pick, M., Flores-Flores, C. & Soreq, H. From brain to blood: alternative splicing evidence for the cholinergic basis of mammalian stress responses. *Ann. NY Acad. Sci.* **1018**, 85–98 (2004).
24. Hastings, M. L. & Krainer, A. R. Pre-mRNA splicing in the new millennium. *Curr. Opin. Cell Biol.* **13**, 302–309 (2001).
25. Nissim-Rafinia, M. & Kerem, B. Splicing regulation as a potential genetic modifier. *Trends Genet.* **18**, 123–127 (2002).
26. Sharp, P. A. "Five easy pieces". *Science* **254**, 663 (1991).
27. Berget, S. M. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270**, 2411–2414 (1995).
28. Caceres, J. F. & Kornblihtt, A. R. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.* **18**, 186–193 (2002).
29. Wagner, E. J. & Garcia-Blanco, M. A. Polypyrimidine tract binding protein antagonizes exon definition. *Mol. Cell Biol.* **21**, 3281–3288 (2001).
30. Faustino, N. A. & Cooper, T. A. Pre-mRNA splicing and human disease. *Genes Dev.* **17**, 419–437 (2003).
31. Romfo, C. M., Alvarez, C. J., van Heeckeren, W. J., Webb, C. J. & Wise, J. A. Evidence for splice site pairing via intron definition in *Schizosaccharomyces pombe*. *Mol. Cell Biol.* **20**, 7955–7970 (2000).
The authors show that that although *S. pombe* has SR-like proteins and factors involved in exon definition, there is no alternative splicing in that organism.
32. Lareau, L. F., Green, R. E., Bhatnagar, R. S. & Brenner, S. E. The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol.* **14**, 273–282 (2004).
33. Bon, E. *et al.* Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res.* **31**, 1121–1135 (2003).
34. Kuhn, A. N. & Kaufer, N. F. Pre-mRNA splicing in *Schizosaccharomyces pombe*: regulatory role of a kinase conserved from fission yeast to mammals. *Curr. Genet.* **42**, 241–251 (2003).
35. Sugnet, C. W., Kent, W. J., Ares, M. Jr & Haussler, D. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.* 66–77 (2004).
The authors showed that alternative 5′ and 3′ splicing events contain intronic sequences that are conserved between humans and mice near the alternative splice site.
36. Sorek, R. & Ast, G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13**, 1631–1637 (2003).
This paper demonstrates that alternatively spliced exons that are in humans and mice contain conserved intron regions that flank the conserved exon.
37. Sorek, R., Shamir, R. & Ast, G. How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20**, 68–71 (2004).
Alternative exons that are conserved between humans and mice tend to be smaller and their length (in nucleotides) is divisible by 3, which distinguishes them from constitutively spliced exons.
38. Resch, A., Xing, Y., Alekseyenko, A., Modrek, B. & Lee, C. Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res.* **32**, 1261–1269 (2004).
The authors show that there is a bias of alternatively spliced exons (which are part of the reading frame) to be multiples of 3.
39. Zhuang, Y., Ma, F., Li-Ling, J., Xu, X. & Li, Y. Comparative analysis of amino acid usage and protein length distribution between alternatively and non-alternatively spliced genes across six eukaryotic genomes. *Mol. Biol. Evol.* **20**, 1978–1985 (2003).
40. Kondrashov, F. A. & Koonin, E. V. Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet.* **19**, 115–119 (2003).
The authors suggest that constitutively spliced exons become alternatively spliced exons (mostly by exon skipping) during evolution.
41. Burge, C. B., Tuschl, T. & Sharp, P. A. in *The RNA World*, 525–560 (Cold Spring Harbor Laboratory Press, New York, 1999).
42. Prabhala, G., Rosenberg, G. H. & Kaufer, N. F. Architectural features of pre-mRNA introns in the fission yeast *Schizosaccharomyces pombe*. *Yeast* **8**, 171–182 (1992).
43. Buvoli, M., Mayer, S. A. & Patton, J. G. Functional crosstalk between exon enhancers, polypyrimidine tracts and branchpoint sequences. *EMBO J.* **16**, 7174–7183 (1997).
44. Modrek, B. & Lee, C. J. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genet.* **34**, 177–180 (2003).
Human–mouse comparative analysis reveals that alternative splicing is often associated with recent exon creation and/or loss. So, alternative splicing has the potential ability to create species-specific alternatively spliced exons.
45. Nurtdinov, R. N., Artamonova, I. I., Mironov, A. A. & Gelfand, M. S. Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum. Mol. Genet.* **12**, 1313–1320 (2003).
46. Kondrashov, F. A. & Koonin, E. V. Origin of alternative splicing by tandem exon duplication. *Hum. Mol. Genet.* **10**, 2661–2669 (2001).
Exon duplication could be a major route for generating functional diversity during the evolution of multicellular eukaryotes.
47. Letunic, I., Copley, R. R. & Bork, P. Common exon duplication in animals and its role in alternative splicing. *Hum. Mol. Genet.* **11**, 1561–1567 (2002).
About 10% of all genes contain tandemly duplicated exons that might have a significant role in the rapid evolution of eukaryotic genes.
48. Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G. The birth of an alternatively spliced exon: 3′ splice-site selection in *Alu* exons. *Science* **300**, 1288–1291 (2003).
This paper reveals a mechanism that governs 3′ splice-site selection of *Alu* exons during alternative splicing and identifies mutations that activated the exonization of a silent intronic *Alu*.
49. Makalowski, W., Mitchell, G. A. & Labuda, D. *Alu* sequences in the coding regions of mRNA: a source of protein variability. *Trends Genet.* **10**, 188–193 (1994).
50. Sorek, R., Ast, G. & Graur, D. *Alu*-containing exons are alternatively spliced. *Genome Res.* **12**, 1060–1067 (2002).
This paper indicates that internal exons that contain an *Alu* sequence are predominantly, if not exclusively, alternatively spliced.
51. Sipiczki, M. Where does fission yeast sit on the tree of life? *Genome Biol.* **1**, 1011.1–1011.4 (2000).
52. Yang, Z. & Yoder, A. D. Estimation of the transition/transversion rate bias and species sampling. *J. Mol. Evol.* **48**, 274–283 (1999).
53. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
Sequencing of the mouse genome reveals that most of the genes present in humans and mice are orthologues and the majority of these genes share the same intron/exon arrangement, as well as a high degree of conservation in homologous exon sequences.
54. Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA* **98**, 11193–11198 (2001).
Splice-site analysis among different organisms.
55. Seraphin, B. & Kandels-Lewis, S. 3′ splice site recognition in *S. cerevisiae* does not require base pairing with U1 snRNA. *Cell* **73**, 803–812 (1993).
56. O'Keefe, R. T. Mutations in U5 snRNA loop 1 influence the splicing of different genes *in vivo*. *Nucleic Acids Res.* **30**, 5476–5484 (2002).
57. Segault, V. *et al.* Conserved loop I of U5 small nuclear RNA is dispensable for both catalytic steps of pre-mRNA splicing in HeLa nuclear extracts. *Mol. Cell. Biol.* **19**, 2782–2790 (1999).

58. Kretzner, L., Rymond, B. C. & Rosbash, M. *S. cerevisiae* U1 RNA is large and has limited primary sequence homology to metazoan U1 snRNA. *Cell* **50**, 593–602 (1987).
59. Zhuang, Y. & Weiner, A. M. A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. *Cell* **46**, 827–835 (1986).
60. Lund, M. & Kjems, J. Defining a 5′ splice site by functional selection in the presence and absence of U1 snRNA 5′ end. *RNA* **8**, 166–179 (2002).
This paper indicates that more than eight base pairs between U1 and the 5′ss will reduce the efficiency of the splicing reaction.
61. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
62. Herbert, A. The four Rs of RNA-directed evolution. *Nature Genet.* **36**, 19–25 (2004).
63. Davis, C. A., Grate, L., Spingola, M. & Ares, M. Jr. Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Res.* **28**, 1700–1706 (2000).
64. Vilardell, J., Chartrand, P., Singer, R. H. & Warner, J. R. The odyssey of a regulated transcript. *RNA* **6**, 1773–1780 (2000).
65. Okazaki, K. & Niwa, O. mRNAs encoding zinc finger protein isoforms are expressed by alternative splicing of an in-frame intron in fission yeast. *DNA Res.* **7**, 27–30 (2000).
66. Muhia, D. K. *et al.* Multiple splice variants encode a novel adenylyl cyclase of possible plastid origin expressed in the sexual stage of the malaria parasite *Plasmodium falciparum*. *J. Biol. Chem.* **278**, 22014–22022 (2003).
67. Fink, G. R. Pseudogenes in yeast? *Cell* **49**, 5–6 (1987).
68. Dahan, O. & Kupiec, M. The *Saccharomyces cerevisiae* gene *CDC40/PRP17* controls cell cycle progression through splicing of the *ANC1* gene. *Nucleic Acids Res.* **32**, 2529–2540 (2004).
69. Freund, M. *et al.* A novel approach to describe a U1 snRNA binding site. *Nucleic Acids Res.* **31**, 6963–6975 (2003).
An indication that a non-Watson–Crick paring can be formed only when it is adjacent to a Watson–Crick pairing.
70. Libri, D., Duconge, F., Levy, L. & Vinauger, M. A role for the Ψ-U mismatch in the recognition of the 5′ splice site of yeast introns by the U1 small nuclear ribonucleoprotein particle. *J. Biol. Chem.* **277**, 18173–18181 (2002).
71. Nandabalan, K., Price, L. & Roeder, G. S. Mutations in U1 snRNA bypass the requirement for a cell type-specific RNA splicing factor. *Cell* **73**, 407–415 (1993).
72. Dibb, N. J. & Newman, A. J. Evidence that introns arose at proto-splice sites. *EMBO J.* **8**, 2015–2021 (1989).
73. Sverdlov, A., Rogozin, B., Babenko, V. N. & Koonin V. E. Reconstruction of ancestral protosplice sites. *Curr. Biol.* **14**, 1505–1508 (2004).
74. Long, M. & Deutsch, M. Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. *Mol. Biol. Evol.* **16**, 1528–1534 (1999).
75. Sverdlov, A. V., Rogozin, I. B., Babenko, V. N. & Koonin, E. V. Evidence of splice signal migration from exon to intron during intron evolution. *Curr. Biol.* **13**, 2170–2174 (2003).
76. Robart, A. R., Montgomery, N. K., Smith, K. L. & Zimmerly, S. Principles of 3′ splice site selection and alternative splicing for an unusual group II intron from *Bacillus anthracis*. *RNA* **10**, 854–862 (2004).
77. Kempken, F. & Windhofer, F. Alternative splicing of transcripts of the transposon *Restless* is maintained in the foreign host *Neurospora crassa* and can be modified by introducing mutations at the 5′ and 3′ splice sites. *Curr. Genet.* **46**, 59–65 (2004).
78. Farrer, T., Roller, A. B., Kent, W. J. & Zahler, A. M. Analysis of the role of *Caenorhabditis elegans* GC-AG introns in regulated splicing. *Nucleic Acids Res.* **30**, 3360–3367 (2002).
79. Thanaraj, T. A. & Clark, F. Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res.* **29**, 2581–2593 (2001).
80. Reed, R. Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr. Opin. Genet. Dev.* **6**, 215–220 (1996).

Acknowledgements

I would like to thank I. Carmel, N. Sela and A. Goren for the assembly of the 5′ss datasets, and A. Weiner, M. Kupiec, E. V. Koonin and an anonymous referee for many helpful comments. G. A. is supported by grants from the Israel Academy of Science and, in part, by grants from the Israel Cancer Association, Familial Dysantonomia Hope, the MOP (research and development), India-Israel and the chief scientist of the Israel Health Ministry.

Competing interests statement

The author declares that he has no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

Entrez: http://www.ncbi.nih.gov/Entrez/
IKBKAP | *ADARB2*

FURTHER INFORMATION

Saccharomyces **Genome Database:** http://genome-www.stanford.edu/Saccharomyces/
Schizosaccharomyces pombe **Gene Database:** http://www.genedb.org/genedb/pombe/index.jsp

SUPPLEMENTARY INFORMATION

See online article: S1 (figure and table)
Access to this links box is available online.