

Chapter 2

Diversity and Evolution of Spliceosomal Systems

Scott William Roy and Manuel Irimia

Abstract

The intron–exon structures of eukaryotic nuclear genomes exhibit tremendous diversity across different species. The availability of many genomes from diverse eukaryotic species now allows for the reconstruction of the evolutionary history of this diversity. Consideration of spliceosomal systems in comparative context reveals a surprising and very complex portrait: in contrast to many expectations, gene structures in early eukaryotic ancestors were highly complex and “animal or plant-like” in many of their spliceosomal structures has occurred; pronounced simplification of gene structures, splicing signals, and spliceosomal machinery occurring independently in many lineages. In addition, next-generation sequencing of transcripts has revealed that alternative splicing is more common across eukaryotes than previously thought. However, much alternative splicing in diverse eukaryotes appears to play a regulatory role: alternative splicing fulfilling the most famous role for alternative splicing—production of multiple different proteins from a single gene—appears to be much more common in animal species than in nearly any other lineage.

Key words Spliceosomal introns, Evolution, Alternative splicing, Eukaryotes, Convergence

1 Similarities and Differences in the Spliceosomal System Across Species

Chapter 1 summarized the splicing reaction, describing a large number of the key features of the spliceosomal intron splicing machinery (the spliceosome) as well as the target of this machinery—the introns and more broadly the pre-mRNA transcripts themselves. The vast majority of our understanding of these topics comes from decades of study of a relatively small number of model species—in particular *S. cerevisiae*. More recently, genomic and transcriptomic sequencing of diverse species has allowed comparisons of these features between more eukaryotic lineages. These studies have ranged across approaches, topics, species, and conclusions, showing both differences and similarities in a wide variety of spliceosome-related phenomena. Surprisingly, given this diversity, the most important points of these studies may be largely summarized in two clear concepts: (1) *the spliceosomal system is ancestral, specific, and (nearly) universal to eukaryotes; and (2) the*

spliceosomal system shows phylogenetically complex patterns across eukaryotes, indicating recurrent transformation in diverse eukaryotes. We devote the next two sections to these two observations.

1.1 The Spliceosomal System is Ancestral, Specific, and (Nearly) Universal to Eukaryotes

Every fully sequenced nuclear genome from a eukaryotic organism contains both spliceosomal introns and recognizable spliceosomal components [1] (although *see* [2] for the one reported possible exception and [3] for the one known qualified exception). Moreover, the core features that define introns are also (nearly) completely conserved [4, 5]. The vast majority of known introns in every studied species begin with a donor site showing complete or partial complementarity to a standard U1 RNA sequence, in particular a 5' "GT" dinucleotide, and nearly all introns in all studied species end with a 3' terminal "AG" (e.g., Fig. 1). Available evidence suggests that the structure of the branchpoint sequence is also conserved across nearly all species: a region base pairing with the U2 RNA, with a "looped out" adenosine residue that performs the first nucleophilic attack. Also widespread across studied species is the polypyrimidine tract located somewhere within the 3' end of the intron, although more diversity is found for this signal [5]. These observations about different species' intronic sequences interleave with observations about the core spliceosomal RNA components: U1–U6 snRNAs have been found across a wide variety of eukaryotes [6], with generally well-conserved RNA secondary structures and strict conservation of regions involved in base pairing between different snRNAs as well as between snRNAs and corresponding regions of pre-mRNA transcripts. Thus, all available evidence points to a highly conserved core spliceosomal reaction present in a wide variety of studied eukaryotes. Since the organisms known to share these features include representatives of all major known eukaryotic groups (or kingdoms), this implies that the spliceosome and spliceosomal introns were present in the eukaryotic ancestor and that the spliceosomal system has been retained in all or nearly all species through eukaryotic evolution.

On the other hand, no sequenced prokaryotic organism contains spliceosomal introns or any recognizable component of a spliceosome, indicating that the spliceosomal system is specific to eukaryotes. Interpretation of this second finding has been more contentious. The simplest interpretation is that the spliceosomal system, including a recognizably modern core splicing machinery and intron sequence characteristics, arose in the last common ancestor of eukaryotes (the modern "Introns-Late" hypothesis [7]). This interpretation mirrors findings that many cellular structures and processes are ancestral and specific to eukaryotes, suggesting a general interpretation that the lineage leading to the last ancestor of eukaryotes experienced an unmatched degree of fundamental cell and molecular structural innovation, including the rise of the spliceosomal system. While many authors have concluded that this hypothesis is by far the more likely alternative, this

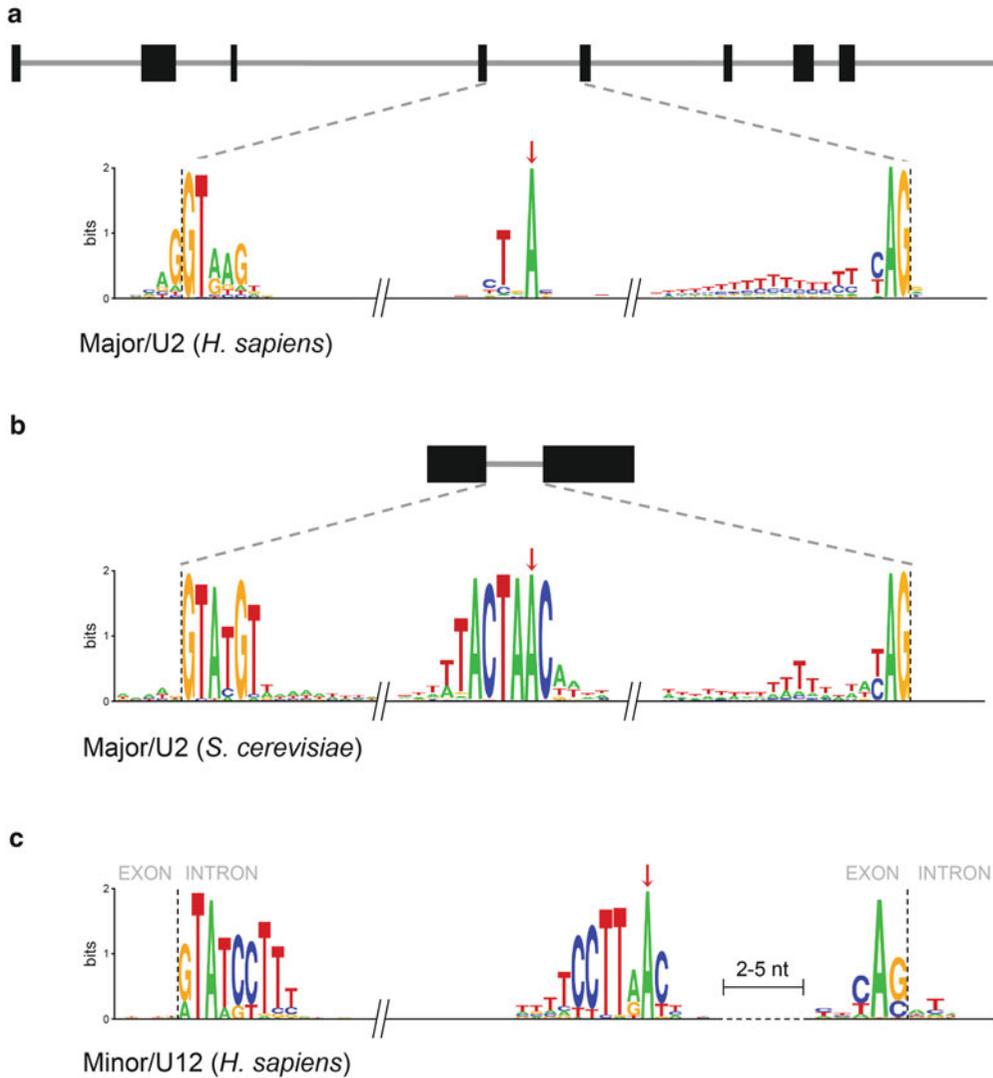


Fig. 1 Intron–exon structures and sequences of U2 (major) and U12 (minor) spliceosomal introns. **(a)** Human genes have frequent and long introns (*lines*) and correspondingly short exons (*boxes*). Human U2-type introns (accounting for >99 % of all human introns) have relatively little sequence homogeneity across intron sequences at the 5′ splice site (*left*), branchpoint (*center*), and 3′ splice site (*right*). **(b)** Introns in the model yeast *S. cerevisiae* are rarer and shorter, and exons longer, with much higher levels of homogeneity at core splice sites. **(c)** In contrast to U2 introns in most species, rare U12 introns show high levels of sequence homogeneity even in species where U2 introns show little homogeneity

perspective has failed to win over a variety of researchers who continue to favor the hypothesis that a system with at least some similarities to the modern spliceosomal system (for instance, high intron density) is even much older than early eukaryotes. Supporters of this “Introns-Early” perspective posit that introns were common in the ancestors of eukaryotes and prokaryotes and have been secondarily lost in both bacteria and archaea [8, 9].

1.2 Phylogenetically Complex Patterns of the Spliceosomal System in Eukaryotes

In stark contrast to this general conservation of the core splicing reaction and its associated machinery, early indications showed that many other aspects of the intron–exon structures of eukaryotic genomes are highly variable across species. Perhaps most striking is the difference in intron numbers. Intron number varies by many orders of magnitude per genome ([10]; Figs. 1 and 2). Whereas human genic transcripts are interrupted by an average of ~8.5 introns, *S. cerevisiae* genes contain only 0.05 introns on average, and extensive next-generation RNA sequencing of the protistan parasite *Trypanosoma brucei* has continually confirmed only two introns in this species' genome [11, 12]. The simplest explanation for these differences would be that intron number had been low in ancestral eukaryotes, with a single massive expansion leading to high intron numbers in one subset of eukaryotes (or alternatively, a single instance of massive loss from an intron-rich eukaryotic ancestor). In this case, we would expect to see high intron numbers to be characteristic of a group of related organisms: for example, in the case of massive expansion in a single event, all intron-rich species would be related. Instead, a very complex pattern is observed, with neither intron-rich nor intron-poor species forming a coherent phylogenetic group (Fig. 2). Very intron-poor organisms (say, with <0.1 introns per gene on average; blue in Fig. 2) are found in diverse eukaryotic groups whose most recent common ancestor is the last common ancestor of all eukaryotes. The same is true of intron-rich species: species with intron densities of at least a few introns per gene are found in disparate groups [4]. This pattern alone implies many different episodes of dramatic genomic change between states in which genomes are alternately nearly intronless or riddled with introns.

Intron length is also highly variable, with intron length distributions ranging widely across species. Median intron lengths range from 19 nts in the nucleomorph (a “mini” green algal nucleus) of the chlorarachniophyte protist *Bigeloviella natans* up to some 2 kb in humans (Fig. 2). Other aspects of the intron length distribution are very different across species as well—whereas the introns in the *B. natans* NM are nearly all within a few nucleotides in length (18–21 nts), human intron lengths are highly diverse, ranging from a few dozen to nearly one million nts. Moreover, intron length distributions can vary between closely related lineages. For example, the introns of tapeworms are sharply distributed around two main lengths (36 and 73 nt), whereas the related animal parasite *Schistosoma* shows only introns of 36 nt [13], implying either gain or loss or transformation of the 73 nt intron type across these species' history. Indeed, intron length distributions may differ significantly even between different classes of introns within a single genome, as recently reported for mammalian introns with different GC content [14].

Different organisms also show striking differences in their sequence characteristics. Particularly clear differences exist in the

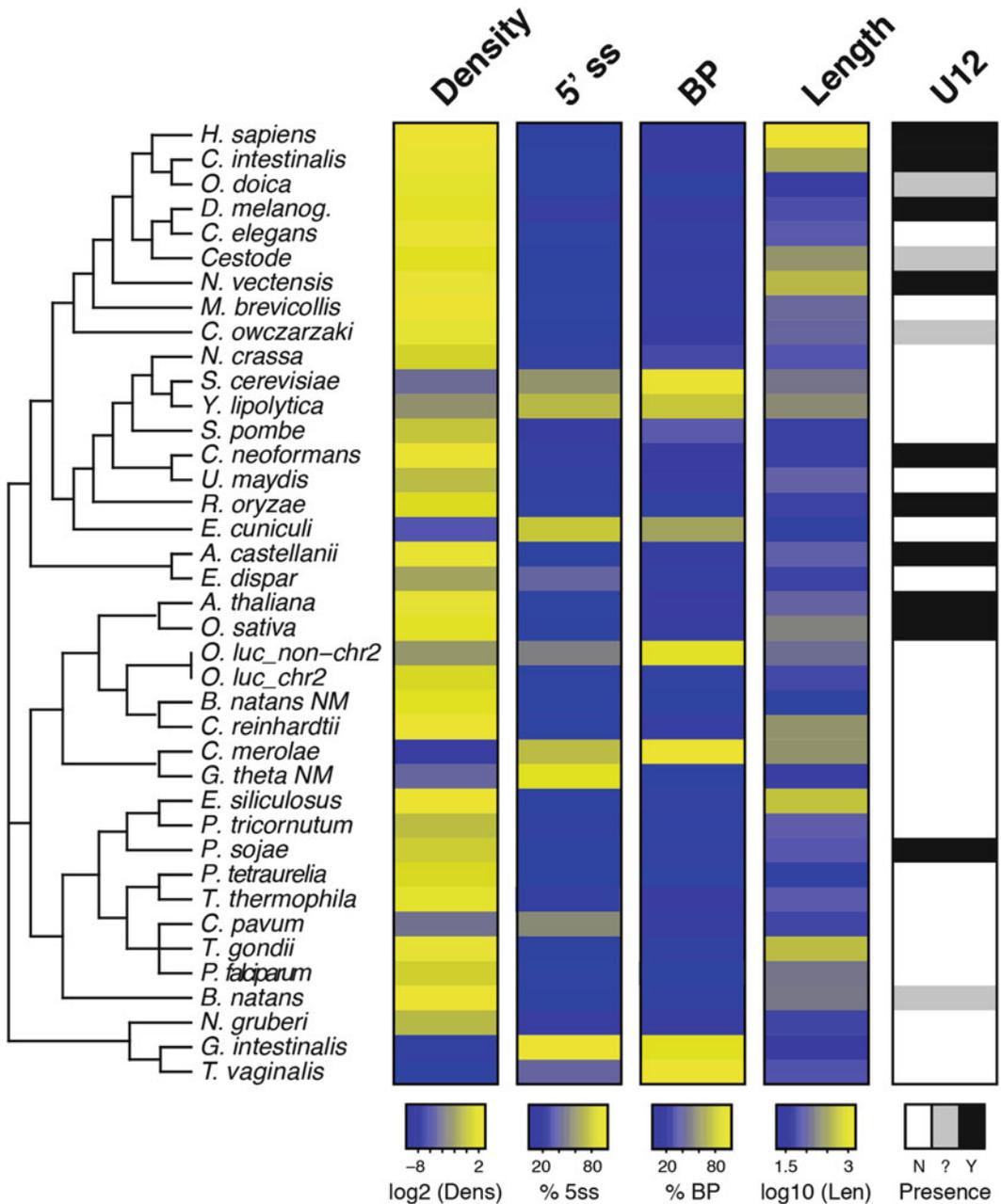


Fig. 2 Diversity of intron–exon structures across eukaryotes. Depicted are as follows: (1) intron density, in number of introns per gene; (2) the probability that two random introns have the same 5′ splice site beyond the canonical GT (in positions 3–6); (3) the fraction of introns exhibiting the exact same seven nucleotide branchpoint motif; (4) median intron length; and (5) presence/absence of minor/U12-type introns and associated splicing machinery

degree of “regularity” of core sequence motifs across introns within a species. For example, whereas nearly all introns in all species maintain significant complementarity between the 5′ splice site

sequence and the U1 snRNA, this is accomplished in very different ways. In the model baker's yeast *S. cerevisiae*, this complementarity is packed into a strongly conserved hexamer region at the very beginning of the intron: some three-quarters of *S. cerevisiae* introns share the same tetramer sequence downstream of the canonical GT (i.e., positions +3 to +6, GTATGT), and nearly all remaining introns have a motif with a single nucleotide difference from this sequence (Fig. 1a). In stark contrast, exonic regions immediately upstream of the intron sequence (e.g., -3 to -1) do not show much preferential complementarity to the U1 sequence: base pairing is largely restricted to the beginning of the intron. On the other hand, human introns' base pairing to the U1 is less concentrated in the intronic 5' splice site, with most introns having intron-U1 base pairs spread out across an extended region spanning both sides of the 5' splice site. This flexibility of base pairing is reflected in a great diversity of core 5' splice site sequences (Fig. 1b). One simple way of quantifying this diversity is to calculate the probability that two random introns from a species will have the same extended splice site sequence (positions +3 to +6). For instance, two random *S. cerevisiae* introns will have the same 5' splice site nearly 58 % of the time, compared to 5.5 % of the time for human introns (Fig. 2).

Comparative genomics reveals similarly pronounced differences for other features of the core spliceosomal sequences. Whereas *S. cerevisiae* uses a highly regular extended branchpoint sequence (ACTAAC, where A is the branchpoint A) with exact complementarity to the corresponding U2 region, human branchpoint sequences are extremely diverse, to the extent that different sites can be used as branchpoints in a single intron [15]. Among characterized branchpoint sequences, the probability that two human introns share the same branchpoint motif is <1 %, whereas for *S. cerevisiae* the probability is 94 % (Fig. 2; [16]). Regularity of the position of the branchpoint relative to the 3' end of the intron is also qualitatively different across species: the probability that two random introns have the exact same branchpoint position is <2 % in humans [16] (and is even low, 2 %, in *S. cerevisiae*), but is 67 % in the yeast species *Yarrowia lipolytica* (93 % of introns have the branchpoint A 6 nts (80 %) or 7 nts (13 %) upstream of the 3' splice site.) As with intron number, species with regular and heterogeneous splicing signals are entwined on the evolutionary tree (Fig. 2; [4, 5]).

Species also show important differences in mechanisms and patterns of splicing. For example, while some components of the spliceosome—most notably the core snRNAs—are (nearly) universally conserved across species, other splicing factors show very different patterns. For instance, a new splicing factor involved in regulating the alternative splicing (AS) of a large number of genes in *Drosophila* was shown to have arisen in *Drosophila* ancestors by duplication of an ancestral factor and functional divergence [17].

This divergence included acquisition of new RNA sequence binding preferences and new biological functions (regulation of AS of dozens of genes in the testes). In other cases, proteins that are evolutionarily old may have acquired new splicing functions (i.e., non-splicing factors have become splicing factors) in specific lineages. One potentially interesting case may involve the splicing factor Nova. Nova is an important AS factor in metazoans [18–20], but Nova plant homologs may be involved in defense mechanisms against RNA viruses [21]. However more data on Nova and other deeply splicing factors in diverse eukaryotic lineages are necessary to confidently reconstruct the evolutionary history of the functions of auxiliary splicing factors.

2 Reconstructing the Evolutionary History of Spliceosomal Systems

Understanding the origins of the diversity of spliceosomal systems not only is interesting in its own right but is an indispensable starting point in understanding the evolution of key splicing innovations in specific lineages (for instance, alternative splicing in animals, see below), since the evolutionary history constrains hypotheses about the possible sets of evolutionary steps leading to these innovations. Therefore, we turn next to results of reconstructions of the evolutionary history of spliceosomal systems.

2.1 The Evolution of the Spliceosome(s)

Crucial to understanding the evolution of spliceosomal systems is understanding the history of the components of the spliceosome. A variety of comparative studies have confirmed that the majority of central and secondary spliceosomal proteins appear to date to the last common ancestor of all eukaryotes [1], completing the portrait of ancestral eukaryotes as having contained a recognizably modern spliceosomal system with a complex spliceosome splicing a large number of introns through a recognition system likely utilizing a diversity of intronic and exonic signals [22]. However, the spliceosomal machinery also appears to have undergone various elaborations in different lineages. In particular, animals and plants appear to have experienced an increase in the number of SR proteins (a family of splicing proteins with diverse core and auxiliary roles in splicing) and other accessory proteins by processes that are likely to have involved both duplication of SR proteins and evolution of new splicing roles for ancestral non-spliceosomal proteins [23, 24]. On the other hand, other lineages have seemingly lost some of the ancestral spliceosomal components, usually in association with massive intron loss. For instance, several human spliceosomal proteins seem to have no ortholog in the *S. cerevisiae* spliceosome [25].

Another question concerns the relative prevalence of intron definition and exon definition. While ultimately detailed molecular

experiments are necessary to determine the mechanism of splicing of a given intron in a given species, the fact that the two different mechanisms tend to lead to different types of splicing variation in transcripts allows us to make educated guesses. Because in exon definition a spliceosome assembles across the length of an exon, failure of the spliceosome to assemble tends to lead to failure to “splice in” that exon, yielding exclusion of an exon in a transcript (called “exon skipping”). On the other hand, failure of a spliceosome to assemble across the length of an intron, in intron definition, tends to lead to failure to “splice out” that intron, leading to intron inclusion. These expected differences apply not only to splicing “errors” (nonfunctional splicing variants) but also to functional AS, since regulation of functional splicing generally occurs through modulation of spliceosomal assembly. Thus the relative incidence of exon skipping and intron retention in a species can yield insights into whether the species splices using exon definition, intron definition, or both mechanisms.

The largest many-species survey of splicing to date mapped available EST data from 42 species to their corresponding genomes to identify splicing variation [26]. They found that for the vast majority of species, levels of splicing variation were far lower than is found in characterized animals. They also found that the mode of splicing variation in most groups of organisms differed from that in animals: whereas animals use extensive exon skipping, nearly all nonanimal species studied had a higher incidence of intron retention. More recent studies of individual species have complicated the issue in plants, which appear to exhibit relatively frequent (and functional) exon skipping [27, 28]; however, the general pattern has held: the major mode of splicing variation in most species is intron retention. These results suggest that the vast majority of eukaryotic lineages primarily splice by intron definition and thus that intron definition is the ancestral mode of intron recognition, with exon definition arising during the evolution of animals (and perhaps, independently, in other lineages [29, 30]).

2.1.1 Notes on the U12 Spliceosomal System

Given the central focus of the book, we have focused on the “major” or “U2” spliceosome and its associated introns. U2 introns make up the vast majority of introns (typically >99 %) in all studied species. However, in some species there also exists a second separate spliceosome which is responsible for splicing of a small subset of introns. This second system (both machinery and associated introns) is referred to as the “U12” or “minor” system, after one of the four separate snRNAs that form the core of the U12 spliceosome. Termed U11, U12, U4atac, and U6atac, these components roughly correspond respectively to the U1, U2, U4, and U6 snRNAs of the major spliceosome (also called the U2 spliceosome). The U5 snRNA is involved in both spliceosomal systems. Spliceosomal proteins show a more complex pattern, with some

proteins showing specificity for either the U2 or U12 spliceosome and others being associated with both systems. Splicing signals of the U12 system broadly correspond to those in the U2 system, with important and intriguing differences. Relative to U2 introns, U12 introns show more flexibility at core splice sites (with both GT...AG and AT...AC boundaries observed) but less flexibility at extended 5' splice site and branchpoint signals (Fig. 1c; [33]). U12 branchpoints also show more conserved and more 3' proximal positions (Fig. 1c), the latter of which is likely related to the general lack of a 3' polypyrimidine tract. The evolutionary origins and functional importance of this remarkable “dual” spliceosomal system remain matters of debate.

Comparative genomics has revealed the broad contours of the evolutionary history of the U12 system. First, the U12 spliceosomal system (both U12-specific components and U12 introns) is found in a variety of very distantly related eukaryotic lineages, in a pattern that strongly suggests presence of a U12 system in the ancestor of all eukaryotes [6, 31]. Second, comparison of orthologous genes has revealed a large number of apparent cases of U12-to-U2 conversions, but few cases of U2-to-U12 conversion [32, 33]. Perhaps relatedly, whereas the U2 spliceosomal system has shown remarkable resilience across species (with no clear case of complete loss of the U2 system known), the U12 system appears to have been lost completely dozens of times independently through eukaryotic evolution, with ancestral U12 introns being either deleted from genomes or converted into U2 introns (Fig. 2) [6].

2.2 The Evolution of Spliceosomal Introns

In this section we will discuss various studies that have reconstructed the evolution of the three major intron features outlined above: intron density, intron sequence, and intron length. Before we proceed, however, it is worthwhile to clearly distinguish between two aspects of an intron: intron position and intron sequence. “Intron sequence” refers to the specific sequence of nucleotides of a specific intron (i.e., the region removed from RNA transcripts). “Intron position” is defined with reference to the final pre-mRNA transcript sequence—that is, the position of the junction between two flanking exons following intron removal (Fig. 3). In many lineages, these two traits of an intron show very different, even opposed, modes of evolution. Consistent with their removal from transcripts and subsequent degradation, most intron sequences evolve quickly, primarily by classic “micro” mutations (base pair substitutions and small indels or transposable element insertion and deletions). A change in intron position, by contrast, involves either gain or loss of an entire intron (and thus gain/loss of an intron position [34]) or intron sliding (a poorly understood and debated mutation or series of mutations leading to movement of an intron along the sequence of a gene [35, 36]). In some lineages, such intron loss and gain mutations are quite rare (see

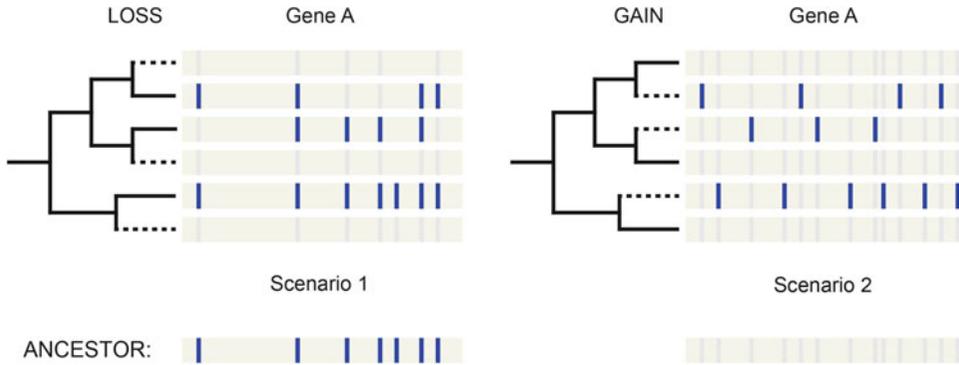


Fig. 3 Intron position comparisons reveal ancestral intron density. Illustrations are given for the cases in which (1) intron positions are shared across species, revealing the presence of introns in the ancestor (*Scenario 1*), or (2) intron positions are largely different across species, revealing that modern introns have been inserted since the common ancestor of the species (*Scenario 2*). In each case, the *gray boxes* represent aligned coding sequence (i.e., after intron removal), with the *blue vertical lines* representing intron positions (i.e., the position of the intronic sequence before removal). In the accompanying phylogenies, *dotted lines* represent lineages undergoing pronounced change, whether primarily intron loss (*on the left*) or intron gain (*on the right*)

below): in this case intron sequences generally evolve quickly, while intron positions evolve very slowly.

2.2.1 Intron Density

In the simplest case, the dramatic differences in intron–exon structures observed across all species (Fig. 2) could be explained by a single process—either intron loss (deletion) or gain (creation)—acting through eukaryotic evolution. It became clear relatively early on that the situation was not so simple. Study of two duplicated insulin genes in rat showed that one copy had lost an intron [37], while restriction of some introns in the triose-phosphate isomerase gene to one or a few related species provided strong evidence for intron gain [38]. With both processes demonstrated, debate turned to distinguishing the two processes’ relative roles and importance in evolution and to reconstruct intron density in ancestral genes.

The most common comparative approach to infer intron gain/loss and reconstruct ancestral states is relatively straightforward (Fig. 3). If an ancestor of two modern organisms had few introns, and the introns in each organism have been created since their divergence, we might expect that the intron positions in these two species—that is, the positions at which the introns interrupt the coding sequence—would have little or no correspondence above random chance (Fig. 3, right). By contrast, if the ancestor had a large number of introns, and if these introns have not been lost, we would expect to find introns in the same position—that is, they would interrupt the coding portion of genes at corresponding (homologous) positions (Fig. 3, left). Closely following on the

availability of the first full and partial genome sequences, a few studies sought to compare intron positions across species to probe intron loss and gain dynamics. By comparing intron positions in 1,560 pairs of homologous genes in humans and mouse, we found nearly complete intron correspondence (>99 % of human introns were matched by an intron at the exact same position in mouse), indicating that both intron loss and gain can be very slow in some lineages [34]. At a much deeper level, genomic sequencing of a handful of genes from jakobid protists showed that intron positions in these deeply diverged organisms showed surprising correspondence to intron positions in homologs from very distantly related eukaryotes, with half found at the exact homologous position in the gene [39]. An eight-species study also showed a high percentage of exact intron position correspondence over long evolutionary distances, with, for instance, a quarter of intron positions corresponding between humans and *Arabidopsis* [40].

While these studies would seem to indicate that many modern introns are very old, another possibility is that these coinciding intron positions in different species are just that: coincidences, with introns being inserted into identical (homologous) positions multiple times independently. However, direct tests from a set of “natural biological” experiments, in which introns are known to have been independently inserted into homologous genes in different organisms, found few correspondences [41–43]. These observations suggest that a large fraction of the observed coincident positions reflect true ancestral introns that have been retained in modern species, indicating that early eukaryotic ancestors were relatively intron rich (i.e., at the least, genes in early eukaryotic ancestors had one or a few introns per gene).

In the past few years, a series of statistical models of increasing sophistication (taking into account the possibility of convergent intron insertion and differences in rates of loss and gain across sites and across lineages), as well as ever-expanding comparative genomic databases, have been used to estimate ancestral intron densities [44–51]. Nearly all of these studies have estimated that intron densities in early eukaryotic ancestors were high by modern standards, falling within the range of modern animal species [52, 53]. Additional studies of intron loss and gain across different groups of organisms have further clarified the evolutionary history, leading to a general picture that most eukaryotic lineages experience very few intron gains (and generally more intron loss, ranging from slightly and dramatically more [54–57]). However, a growing number of exceptional lineages have been reported, in which intron gain is an active and ongoing process, potentially “replenishing” relatively intron-poor organisms with a large number of new introns [58–61].

2.2.2 Intron Structures: Splicing Sequence Motifs

As mentioned above, eukaryotic organisms differ considerably in their splicing motifs, ranging from the highly homogeneous 5' splice site and branchpoint site sequences and branchpoint positions found in the yeast *Yarrowia lipolytica* to the heterogeneous structures characterizing human intron sequences. Notably, as discussed in more detail elsewhere in this book, these differences seem to involve a greater reliance on auxiliary splicing signals (generally lying in proximal regions of introns and exons) by species with heterogeneous core splicing signals. For instance, in humans, the boundaries of exons (i.e., exonic regions near intron–exon boundaries) are enriched in certain sequence motifs, which affect splicing by serving as “exonic splicing enhancers” (ESEs) by binding spliceosomal proteins and promoting splicing at the neighboring splice site [62]. By contrast, in species such as *S. cerevisiae*, ESEs are thought to not play a major role in splicing—intron recognition signals are concentrated in the core intronic splicing motifs.

What is the history of these recognition systems and splicing motifs? Initially it was often assumed that the “simpler” system of *S. cerevisiae* was ancestral and that increased complexity of mechanism arose in animals [63]. Widespread genomic evidence allowed for the possibility to test this notion. We studied full-genome intron complements from 50 diverse eukaryotic species to reconstruct the evolution of intron sequences and recognition [4]. First, we examined 5' splice signals. We found that 5' splice sites are heterogeneous in most species and that cases such as *S. cerevisiae* represent exceptions. For nearly all species studied, the probability that two random introns use the same hexamer splice site was <5 % (Fig. 2). However, there were a few clear exceptions, with several distantly related species showing a much higher level of homogeneity. Viewed on the evolutionary tree, these exceptional lineages fall within much larger phylogenetic groups of species with more typical splice signals. This phylogenetic pattern suggests that ancestral splice site sequences were heterogeneous and that the several species or groups of species with homogeneous splice sites evolved independently.

Even more unexpectedly, scrutiny of the specific lineages that have acquired homogeneous signals revealed that they were exactly the same lineages known to have very low modern intron densities (<0.1 introns per gene, blue in Fig. 2), with no known exceptions. Together these patterns indicate that early eukaryotic ancestral genes were roughly “animal-like” in their intron–exon structures, with high intron densities and heterogeneous 5' splice sites, and that at several times through evolution, different lineages have experienced massive intron loss tightly coupled to the evolution of homogeneous 5' splice site signals.

We and others also studied 3' intron sequences [5, 64]. First, we studied branchpoint motifs. Because branchpoints in some species can be so diverse as to be difficult to identify computationally

[15, 65], we used a different metric: the fraction of introns that exhibited the same branchpoint-like sequence motif (i.e., a motif with the potential to base pair with the U2 snRNA with a protruding A nucleotide). For most organisms, we found no single dominating branchpoint motif, indicating heterogeneous branchpoint sequences (Fig. 2). However, again, a small subset of organisms including *S. cerevisiae* exhibited homogeneous branchpoints, with a majority of introns having the same clear branchpoint-like sequence [5]. This subset of organisms proved to be a subset of the studied intron-poor species. Thus low intron density appears to be closely associated with, but not sufficient for, the evolution of homogeneous branchpoint signals.

Finally, we studied the stretch of intronic nucleotides just upstream of the 3' splice site. Again, for most species we found no clear motif preference (with the exception of a weak polypyrimidine tract). However a few species showed a clear preferred extended 3' splice site, which was found to represent a branchpoint motif falling at a regular distance from the 3' terminus—that is, the branchpoint is “anchored” to the 3' end of the intron at a highly constrained distance [5]. These species proved to be a subset of species that have homogeneous branchpoint motifs. In total, then, these studies may be summarized as follows: all intron-poor lineages have homogeneous 5' splice sites, a subset of which have homogeneous branchpoints, a subset of which have homogeneous 3' splice sites owing to anchoring of the homogeneous branchpoint at a specific position a few nucleotides upstream of the 3' terminus.

This unexpectedly clear pattern is still not well understood. The most obvious hypothesis would be that these changes in the recognition signals are associated with changes in the spliceosome. This hypothesis initially defied direct testing until a natural experiment presented itself, in the form of the sequenced genomes of multiple species from an evolutionarily old group of related algae. Each species' genome showed striking differentiation in intron density across genomic regions: in contrast to genes in most of the genome, which have very few introns (~0.1 per gene), the genes on one chromosome have much higher intron densities (around two introns per gene) [66]. Scrutiny of the genome sequence revealed a single set of core spliceosomal components [5], indicating that there is no evidence that entirely separate spliceosomes are responsible for splicing in the two genomic regions: thus if changes in the spliceosome are responsible for (or closely associated with) changes in splice signals, we would expect introns in both regions of the genome to show similar levels of splice signal homogeneity. Instead, the genomic regions show clear differentiation along the exact lines expected from the across-species comparisons: introns in the intron-rich region of the genome show very heterogeneous splice signals and no recognizable branchpoints, while introns in

the intron-poor majority of the genome have homogeneous 5' splice sites and branchpoint sequences [5]. The differences in intron number and splice motif homogeneity are found across distantly related species likely spanning many millions of years of evolution; thus, this association is long-lived, not transient.

Another issue involves the evolution of ESEs, which are abundant in animal genomes but absent or nearly absent from *S. cerevisiae*. ESEs were initially recognized at the genome-wide level by identifying sequence motifs that were overrepresented in the portions of exons near intron–exon boundaries relative to more distant portions of exons, and overrepresented near intronic splice sites that were “weak” (i.e., had low predicted binding to spliceosomal uRNAs), and which were subsequently confirmed by in vitro and in vivo studies to affect splicing [67, 68]. To test whether a similar signal existed in diverse other eukaryotes, Warnecke and coauthors [67] sought motifs that were overrepresented near exon–intron boundaries relative to interior regions of exons. They found putative ESE motifs in most studied intron-rich eukaryotes, but no evidence for ESEs in studied intron-poor species. This again suggested that the animal-like state (considerable reliance on ESEs for splicing) was ancestral to eukaryotes and that the spliceosomal systems in intron-poor lineages such as *S. cerevisiae* have been altered through evolution.

In total, then, comparative studies of intronic and exonic sequences over long evolutionary distances within eukaryotes support a model in which ancestral eukaryotes had “animal-like” intron–exon structures, with frequent introns spliced by use of a combination of diffuse motifs including frequent ESEs and heterogeneous core splicing motifs. Over the course of evolution, many lineages have changed significantly, shedding the vast majority of their introns, evolving homogeneous core splicing motifs, and significantly decreasing dependence on auxiliary splicing motifs such as ESEs.

2.2.3 Intron Length

The third feature of introns that shows striking diversity is intron length. Introns show a wide variety of lengths both within and between organisms, with lengths spanning multiple orders of magnitude. Studies across many eukaryotic organisms, particularly whole genome sequencing projects, have shown that the vast majority of species have relatively short introns, often with a peak around 60 nucleotides. While it is difficult to directly reconstruct intron length over long evolutionary distances, as introns appear to readily expand and contract along with genome size [69–71], this clear preference for generally short intron length across eukaryotes suggests that it represents the ancestral condition (although it has been suggested that the most ancestral introns, presumably evolved from self-splicing group II introns, may have been much longer, perhaps around 2,000 nts [53]).

Against this backdrop of generally short introns, several lineages show very different patterns. On the one hand, many different lineages from very different groups (animals [72, 73], relatives of green algae [74], and ciliates [75]) have evolved very short introns with median lengths around 20 nts. The clearest exception at the other end of the spectrum is some animals, particularly mammals [76], in which many species have median intron lengths ranging from a couple hundred to a couple thousand nucleotides. It seems likely that there are other lineages with generally long introns yet to be discovered, particularly given that (1) the correspondence between intron and genome size suggests that organisms with long introns would tend to have large genomes; (2) genome sequencing efforts tend to be biased specifically against organisms with large genomes, because of technical difficulties of sequencing and annotation.

3 Diversity and Evolution of Alternative Splicing

Up to this point, we have focused on differences in the genomic structures and in the splicing machinery and intron recognition mechanisms. We now briefly turn to the ways that these structures are used to generate transcriptional diversity by differential splicing of transcripts of the same gene, that is, alternative splicing (AS). The types, mechanisms, and functions of AS will be discussed extensively in Chapters 4 and 5, so here we confine our discussion to AS in the broader context of intron and genome evolution.

The most well-known function of AS is to generate multiple proteins with distinct functional properties from a single gene. However, decades of research have made clear that other forms of splicing diversity in which some transcript variants do not encode proteins are very common. Many genes in animals harbor alternatively spliced “poison exons” whose inclusion in transcripts leads to disruption of the protein-coding sequence [77]. Many of these transcripts are rapidly degraded by the nonsense-mediated decay (NMD) machinery; the fates of others remain obscure, however, the lack of an extended protein-coding region suggests these transcripts are unlikely to encode proteins. Such nonprotein coding variation is usually referred to “unproductive” AS, in contrast to “productive” or multi-protein AS [78]. It is important to point out that very clear evidence exists for functional roles for many of these cases of unproductive splicing: much unproductive splicing is evolutionarily conserved and/or regulated across environmental conditions, development, life cycles, or tissue or cell types [77, 79]. However, it is also likely that nonfunctional splicing errors that lead to transcript diversity with no function also occur (even if it is the case that confidently classifying a given AS event as either nonfunctional variation or functional nonproductive AS can be

technically different). Thus in the following we distinguish between three types of AS: productive, unproductive, and nonfunctional.

AS is an extremely important and active process in animals, with the vast majority of multi-exon genes undergoing AS in diverse animal species (e.g., an estimated 95 % in humans [80, 81] and 60 % in fruit fly [82]). Animal AS uses a wide variety of mechanisms including single exon skipping, coordinated splicing of groups of exons, mutually exclusive splicing of pairs (or sets) of exons, alternative 5' and 3' splice sites, and intron retention [83]. AS is involved in a wide array of biological processes from sex determination to development to negative autoregulation and generates both productive and unproductive transcripts (*see* Chapters 4 and 5 for further examples).

Initial studies of nonanimal eukaryotes found a dearth of animal-like productive AS. In comparison to the thousands of cases of productive AS uncovered by transcriptomic studies in animals, for a long time no productive AS was known in *S. cerevisiae*, and cases in other species were only few and far between. Both reason and evidence suggest that AS would be facilitated by a variety of features of animals' intron–exon structures: (1) Large numbers of introns provide many opportunities for AS. (2) Heterogeneous intron boundaries, with associated differences in the strength of base pairing with the spliceosomal RNAs, allow for the possibility of regions for which recognition by the spliceosome might be “borderline”—leading to non-constitutive splicing of these regions. (3) Utilization of a variety of heterogeneous splicing signals—exonic and intronic splicing regulators, in addition to core splicing signals—allows for the possibility of regulating local splicing by regulation of the splicing factors that bind subsets of these signals. (4) Long introns increase opportunities for novel alternative exon creation [84–86] and are associated with AS in vertebrates [76].

The fact that these features each differ considerably between AS-rich animals and the model organism for splicing, *S. cerevisiae*, initially suggested that a wholesale remodeling of gene structures had occurred in animals roughly coincident with a rise of ubiquitous AS. However, as discussed above, genomic-era studies have shown that the story is quite different from this: many of the features associated with AS in animals—frequent introns, heterogeneous splicing boundaries, introns with lengths exceeding “minimal” intron lengths, and utilization of auxiliary splicing signals—are not specific to animals, but are in fact quite common in modern eukaryotes as well as characteristic of eukaryotic ancestors [22]. Thus, the hypothesis that widespread productive AS in animals is “due” to these features, a hypothesis still commonly invoked in passing in publications, is strongly rejected, since these features are common in organisms with little or no productive AS.

Furthermore, more recently, transcriptomic studies have opened up questions about the incidence of AS in diverse eukaryotic organisms. Initially it was thought by some authors that AS was absent or very rare in unicellular species [63]. However, genomic and transcriptomic data has greatly changed that picture. Perhaps the clearest case involves splicing of ribosomal protein-coding genes in *S. cerevisiae* [87, 88]. Introns in *S. cerevisiae* are massively overrepresented in ribosomal protein-coding genes, with half of the introns in the genome packed into only a few percent of the genes. A series of studies have shown that many ribosomal protein-coding gene (RPG) introns are regulated in response to environmental changes to produce either spliced protein-coding or unspliced sterile transcripts. This apparent regulatory role for RPG introns suggests that overrepresentation of introns in RPGs reflects selection favoring retention and/or creation of specifically these introns. This would in turn imply that at least half of introns in *S. cerevisiae* have been retained through evolution due to functional AS.

Other studies have begun to suggest that AS plays important roles in a wide variety of eukaryotes. Transcriptomic studies have found between several dozen and several hundred apparent cases of AS in the genomes of nearly all species studied to date, including diverse fungi [89–91], plants [27, 92–94], apicomplexans [95], cryptophytes [96], green algae [97], ciliates [98], and amoebozoa [99] (although studies of two other protists have drawn the opposite conclusion [100]). Nearly all of these studies have found a preponderance of intron retentions, with far smaller numbers of exon skipping events (and often intermediate numbers of alternative splice sites), even in plants [101]. These observations suggest that intron retention has predominated through eukaryotic history in diverse organisms. The one clear exception described so far is the chlorarachniophyte *Bigeloviella natans* [96], which shows striking levels of both intron retention and exon skipping, the latter only comparable to AS levels in the human cortex, which exhibits the highest levels of AS described so far [102].

In total, then, genomic and transcriptomic data have painted a very different picture of the history of AS (productive and otherwise) in animals. Features of animal intron–exon structures (long and frequent introns with diverse splicing signals) are not closely associated with animal-type AS, and AS is far from exclusive to animals, being found across phylogenetically and biologically diverse eukaryotic organisms. The one remaining feature of animal genomes that may still be rare in other organisms is exon definition. Therefore, it has been suggested that the evolution of exon definition, together with the specific expansion of SR proteins and other splicing factors, may be behind the transition from intron retention to exon skipping at the origin of animals [29].

4 Summary

A comparative perspective on spliceosomal systems of diverse eukaryotes paints a surprising portrait: ancestral eukaryotic genes were riddled with introns characterized by heterogeneous splice signals, requiring two distinct complex spliceosomes for intron removal and quite possibly involving some level of functional regulatory alternative splicing, likely dominated by intron retention. Since that time, different lineages have experienced very different evolutionary trajectories ranging from nearly complete intron loss to intron length expansion and episodic intron creation. The one feature of animal gene structures that remains as clearly exceptional is the widespread production of multiple proteins from one gene, although recent findings in *B. natans* suggest that animals may not be entirely alone in this characteristic.

References

- Collins L, Penny D (2005) Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol* 22:1053–1066
- Andersson JO, Sjögren AM, Horner DS et al (2007) A genomic survey of the fish parasite *Spironucleus salmonicida* indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. *BMC Genomics* 8:51
- Lane CE, van den Heuvel K, Kozera C et al (2007) Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. *Proc Natl Acad Sci USA* 104:19908–19913
- Irimia M, Penny D, Roy SW (2007) Co-evolution of genomic intron number and splice sites. *Trends Genet* 23:321–325
- Irimia M, Roy SW (2008) Evolutionary convergence on highly-conserved 3' intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS Genet* 4:e1000148
- Dávila LM, Rosenblad MA, Samuelsson T (2008) Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucleic Acids Res* 36:3001–3010
- Koonin EV (2006) The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct* 1:22
- Vibrantovski M, Sakabe N, Oliveira R et al (2005) Signs of ancient and modern exon-shuffling are correlated to the distribution of ancient and modern domains along proteins. *J Mol Evol* 61:341–350
- Penny D, Hoepfner MP, Poole AM et al (2009) An overview of the introns-first theory. *J Mol Evol* 69:527–540
- Logsdon J (1998) The recent origins of spliceosomal introns revisited. *Curr Opin Genet Dev* 8:637–648
- Siegel TN, Hekstra DR, Wang X et al (2010) Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. *Nucleic Acids Res* 38:4946–4957
- Kolev NG, Franklin JB, Carmi S et al (2010) The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. *PLoS Pathog* 6:e1001090
- Tsai IJ, Zarowiecki M, Holroyd N et al (2013) The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* 496(7443):57–63
- Amit M, Donyo M, Hollander D et al (2012) Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep* 1:543–556
- Kol G, Lev-Maor G, Ast G (2005) Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. *Hum Mol Genet* 14:1559–1568
- Gao K, Masuda A, Matsuura T et al (2008) Human branch point consensus sequence is yUnAy. *Nucleic Acids Res* 36:2257–2267
- Taliaferro JM, Alvarez N, Green RE et al (2011) Evolution of a tissue-specific splicing network. *Genes Dev* 25:608–620

18. Brooks AN, Yang L, Duff MO et al (2011) Conservation of an RNA regulatory map between *Drosophila* and mammals. *Genome Res* 21:193–202
19. Irimia M, Denuc A, Burguera D et al (2011) Stepwise assembly of the nova-regulated alternative splicing network in the vertebrate brain. *Proc Natl Acad Sci USA* 108:5319–5324
20. Jensen KB, Dredge BK, Stefani G et al (2000) Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron* 25:359–371
21. Fujisaki K, Ishikawa M (2008) Identification of an *Arabidopsis thaliana* protein that binds to tomato mosaic virus genomic RNA and inhibits its multiplication. *Virology* 380:402–411
22. Roy SW, Irimia M (2009) Splicing in the eukaryotic ancestor: form, function and dysfunction. *Trends Ecol Evol* 24:447–455
23. Barbosa-Morais NL, Carmo-Fonseca M, Aparicio S (2006) Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res* 16:66–77
24. Reddy AS, Shad AG (2011) Plant serine/arginine-rich proteins: roles in precursor messenger RNA splicing, plant development, and stress responses. *Wiley Interdiscip Rev RNA* 2:875–889
25. Plass M, Agirre E, Reyes D et al (2008) Co-evolution of the branch site and SR proteins in eukaryotes. *Trends Genet* 24:590–594
26. McGuire A, Pearson M, Neafsey D et al (2008) Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol* 9:R50
27. Marquez Y, Brown JW, Simpson C et al (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Res* 22:1184–1195
28. Carvalho RF, Feijão CV, Duque P (2012) On the physiological significance of alternative splicing events in higher plants. *Protoplasma* 250(3):639–650
29. Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* 11:345–355
30. Ram O, Ast G (2007) SR proteins: a foot on the exon before the transition from intron to exon definition. *Trends Genet* 23:5–7
31. Russell AG, Charette JM, Spencer DF et al (2006) An early evolutionary origin for the minor spliceosome. *Nature* 443:863–866
32. Burge CB, Padgett RA, Sharp PA (1998) Evolutionary fates and origins of U12-type introns. *Mol Cell* 2:773–785
33. Alioto TS (2007) U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res* 35:D110–D115
34. Roy SW, Fedorov A, Gilbert W (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci USA* 100:7158–7162
35. Tarrío R, Ayala FJ, Rodríguez-Trelles F (2008) Alternative splicing: a missing piece in the puzzle of intron gain. *Proc Natl Acad Sci USA* 105:7223–7228
36. Rogozin IB, Lyons-Weiler J, Koonin EV (2000) Intron sliding in conserved gene families. *Trends Genet* 16:430–432
37. Perler F, Efstratiadis A, Lomedico P et al (1980) The evolution of genes: the chicken preproinsulin gene. *Cell* 20:555–566
38. Logsdon J Jr, Tyshenko M, Dixon C et al (1995) Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc Natl Acad Sci USA* 92:8507–8511
39. Archibald J, O'Kelly C, Doolittle W (2002) The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. *Mol Biol Evol* 19:422–431
40. Rogozin I, Sverdlov A, Babenko V et al (2005) Analysis of evolution of exon–intron structure of eukaryotic genes. *Brief Bioinform* 6:118–134
41. Roy SW, Penny D (2007) A very high fraction of unique intron positions in the intron-rich diatom *Thalassiosira pseudonana* indicates widespread intron gain. *Mol Biol Evol* 24:1447–1457
42. Ahmadinejad N, Dagan T, Gruenheit N et al (2010) Evolution of spliceosomal introns following endosymbiotic gene transfer. *BMC Evol Biol* 10:57
43. Yoshihama M, Nakao A, Nguyen HD et al (2006) Analysis of ribosomal protein gene structures: implications for intron evolution. *PLoS Genet* 2:e25
44. Roy SW, Gilbert W (2005) Complex early genes. *Proc Natl Acad Sci USA* 102:1986–1991
45. Csuros M (2006) On the estimation of intron evolution. *PLoS Comput Biol* 2:e84
46. Csuros M (2008) Malin: maximum likelihood analysis of intron evolution in eukaryotes. *Bioinformatics* 24:1538–1539

47. Csurös M (2005). Likely scenarios of intron evolution. In: Third RECOMB Satellite workshop on comparative genomics. Springer LNCS 3678, p 47–60
48. Csurös M, Rogozin IB, Koonin EV (2008) Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach. *Mol Biol Evol* 25:903–911
49. Nguyen H, Yoshihama M, Kenmochi N (2005) New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Comput Biol* 1:e79
50. Carmel L, Wolf YI, Rogozin IB et al (2007) Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res* 17:1034–1044
51. Carmel L, Rogozin IB, Wolf YI et al (2009) A maximum likelihood method for reconstruction of the evolution of eukaryotic gene structure. *Methods Mol Biol* 541:357–371
52. Rogozin IB, Carmel L, Csuros M et al (2012) Origin and evolution of spliceosomal introns. *Biol Direct* 7:11
53. Koonin EV (2009) Intron-dominated genomes of early ancestors of eukaryotes. *J Hered* 100:618–623
54. Roy SW, Irimia M, Penny D (2006) Very little intron gain in *Entamoeba histolytica* genes laterally transferred from prokaryotes. *Mol Biol Evol* 23:1824–1827
55. Roy SW, Penny D (2006) Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. *Mol Biol Evol* 23:2259–2262
56. Stajich JE, Dietrich FS, Roy SW (2007) Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biol* 8:R223
57. Coulombe-Huntington J, Majewski J (2007) Intron loss and gain in *Drosophila*. *Mol Biol Evol* 24:2842–2850
58. Li W, Tucker AE, Sung W et al (2009) Extensive, recent intron gains in *Daphnia* populations. *Science* 326:1260–1262
59. Worden AZ, Lee JH, Mock T et al (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* 324:268–272
60. van der Burgt A, Severing E, de Wit PJGM et al (2012) Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. *Curr Biol* 22(13):1260–1265
61. Roy SW, Irimia M (2012) Genome evolution: where do new introns come from? *Curr Biol* 22:R529–R531
62. Lim KH, Ferraris L, Filloux ME et al (2011) Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci USA* 108:11093–11098
63. Ast G (2004) How did alternative splicing evolve? *Nat Rev Genet* 5:773–782
64. Schwartz S, Silva J, Burstein D et al (2008) Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res* 18:88–103
65. Tolstrup N, Rouze P, Brunak S (1997) A branch point consensus from *Arabidopsis* found by non-circular analysis allows for better prediction of acceptor sites. *Nucleic Acids Res* 25:3159–3163
66. Vault D, Lepère C, Toulza E et al (2012) Metagenomes of the picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS One* 7:e39648
67. Warnecke T, Parmley JL, Hurst LD (2008) Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol* 9:R29
68. Fairbrother WG, Yeh R-F, Sharp PA et al (2002) Predictive identification of exonic splicing enhancers in human genes. *Science* 297:1007–1013
69. McLysaght A, Enright AJ, Skrabanek L et al (2000) Estimation of synteny conservation and genome compaction between pufferfish (*Fugu*) and human. *Yeast* 17:22–36
70. Deutsch M, Long M (1999) Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Res* 27:3219–3228
71. Moriyama EN, Petrov DA, Hartl DL (1998) Genome size and intron size in *Drosophila*. *Mol Biol Evol* 15:770–773
72. Aruga J, Odaka YS, Kamiya A et al (2007) *Dicyema Pax6* and *Zic*: tool-kit genes in a highly simplified bilaterian. *BMC Evol Biol* 7:201
73. Ogino K, Tsuneki K, Furuya H (2010) Unique genome of dicyemid mesozoan: highly shortened spliceosomal introns in conservative exon/intron structure. *Gene* 449:70–76
74. Gilson PR, Su V, Slamovits CH et al (2006) Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc Natl Acad Sci* 103:9566–9571
75. Russell CB, Fraga D, Hinrichsen RD (1994) Extremely short 20–33 nucleotide introns are the standard length in *Paramecium* tetraurelia. *Nucleic Acids Res* 22:1221–1225

76. Gelfman S, Burstein D, Penn O et al (2012) Changes in exon–intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome Res* 22:35–50
77. Lewis BP, Green RE, Brenner SE (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci USA* 100:189–192
78. Lareau LF, Brooks AN, Soergel DAW et al (2007) The coupling of alternative splicing and nonsense mediated mRNA decay. In: Blencowe BJ, Graveley BR (eds) *Alternative splicing in the postgenomic era*. Landes Bioscience and Springer Science&Business Media, Austin, TX, pp 190–211
79. Lareau LF, Inada M, Green RE et al (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* 446:926–929
80. Wang ET, Sandberg R, Luo S et al (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470–476
81. Pan Q, Shai O, Lee LJ et al (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40:1413–1415
82. Graveley BR, Brooks AN, Carlson JW et al (2011) The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471:473–479
83. Irimia M, Blencowe BJ (2012) Alternative splicing: decoding an expansive regulatory layer. *Curr Opin Cell Biol* 24:323–332
84. Irimia M, Rukov JL, Penny D et al (2008) Widespread evolutionary conservation of alternatively spliced exons in *Caenorhabditis*. *Mol Biol Evol* 25:375–382
85. Irimia M, Rukov JL, Roy SW et al (2009) Quantitative regulation of alternative splicing in evolution and development. *Bioessays* 31:40–50
86. Roy M, Kim N, Xing Y et al (2008) The effect of intron length on exon creation ratios during the evolution of mammalian genomes. *RNA* 14:2261–2273
87. Pleiss JA, Whitworth GB, Bergkessel M et al (2007) Rapid, transcript-specific changes in splicing in response to environmental stress. *Mol Cell* 27:928–937
88. Parenteau J, Durand M, Morin G et al (2011) Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell* 147:320–331
89. Yin Y, Yu G, Chen Y et al (2012) Genome-wide transcriptome and proteome analysis on different developmental stages of *Cordyceps militaris*. *PLoS One* 7:e51853
90. Zhao C, Waalwijk C, de Wit PJ et al (2013) RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen *Fusarium graminearum*. *BMC Genomics* 14:21
91. Wang B, Guo G, Wang C et al (2010) Survey of the transcriptome of *Aspergillus oryzae* via massively parallel mRNA sequencing. *Nucleic Acids Res* 38:5075–5087
92. Campbell MA, Haas BJ, Hamilton JP et al (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics* 7:327
93. Iida K, Seki M, Sakurai T et al (2004) Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. *Nucleic Acids Res* 32:5096–5103
94. Ner-Gaon H, Halachmi R, Savaldi-Goldstein S et al (2004) Intron retention is a major phenomenon in alternative splicing in *Arabidopsis*. *Plant J* 39:877–885
95. Sorber K, Dimon MT, DeRisi JL (2011) RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic Acids Res* 39:3820–3835
96. Curtis BA, Tanifuji G, Burki F et al (2012) Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* 492:59–65
97. Labadorf A, Link A, Rogers MF et al (2010) Genome-wide analysis of alternative splicing in *Chlamydomonas reinhardtii*. *BMC Genomics* 11:114
98. Xiong J, Lu X, Zhou Z et al (2012) Transcriptome analysis of the model protozoan, *Tetrahymena thermophila*, using Deep RNA sequencing. *PLoS One* 7:e30630
99. Glöckner G, Golderer G, Werner-Felmayer G et al (2008) A first glimpse at the transcriptome of *Physarum polycephalum*. *BMC Genomics* 9:6
100. Jaillon O, Bouhouche K, Gout J-F et al (2008) Translational control of intron splicing in eukaryotes. *Nature* 451:359–362
101. Wang B-B, Brendel V (2006) Molecular characterization and phylogeny of U2AF35 homologs in plants. *Plant Physiol* 140:624–636
102. Barbosa-Morais NL, Irimia M, Pan Q et al (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338:1587–1593



<http://www.springer.com/978-1-62703-979-6>

Spliceosomal Pre-mRNA Splicing

Methods and Protocols

Hertel, K.J. (Ed.)

2014, XI, 427 p. 66 illus., 34 illus. in color., Hardcover

ISBN: 978-1-62703-979-6

A product of Humana Press