Published in final edited form as:

Trends Genet. 2008 November ; 24(11): 525–528. doi:10.1016/j.tig.2008.09.002.

Primordial spliceosomal introns were probably U2-type

Malay Kumar Basu, Igor B. Rogozin, and Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA

Abstract

The two types of eukaryotic spliceosomal introns, U2 and U12, possess different splice signals and are excised by distinct spliceosomes. The nature of the primordial introns remains uncertain. A comparison of the amino acid distributions at insertion sites of introns that retained their positions throughout eukaryotic evolution with the distributions for human and *Arabidopsis thaliana* U2 and U12 introns reveals close similarity with U2 but not U12. Thus, the primordial spliceosomal introns were, most likely, U2-type.

The two types of spliceosomal introns and the enigma of primordial introns

Introns are excised from pre-mRNAs at acceptor and donor splice sites. This process is mediated by the spliceosome, a complex assembly of small nuclear ribonucleoprotein particles (snRNPs) and heterogeneous nuclear ribonucleoprotein particles (hnRNPs) that is conserved throughout the eukaryotic world [1–3]. There are two classes of introns, U2-type and U12-type, which are excised by two distinct spliceosomes in eukaryotic nuclei [4,5]. During the splicing process, the components of a spliceosome establish specific interactions with parts of the intron and its flanking exons to ensure accurate and efficient splicing [6]. These essential interactions are supported by conserved nucleotide sequence motifs (splice signals) that flank the splice junctions from both the intronic and the exonic sides and are specific to the intron class (Box 1).

Box 1

U2 and U12 introns and the major and minor spliceosomes

The splicing process includes specific interactions between components of a spliceosome and parts of the intron and its flanking exons that ensure accurate and efficient splicing [6]. In the major spliceosome, the U1 snRNP recognizes the donor splice site and the U5 snRNP recognizes the acceptor site. The (A,C)AG|GU(A,G)AGU consensus (the first two nucleotides of an intron are underlined) is present in the donor splice sites and is partially complementary to the 5' end of U1 small nuclear RNA (snRNA); this interaction is a major requirement for splicing. The motif CAG|G (the last two nucleotides of the intron are underlined), which is preceded by a polypyrimidine tract, is typical of the acceptor splice site [6,14,15].

The minor spliceosome catalyses the removal of an atypical class of spliceosomal introns (U12-type) from eukaryotic pre-mRNAs. U12 introns have been originally recognized on the basis of their unusual terminal dinucleotides: |AT at the donor splice site and AC| at the acceptor splice site [16,17]. A closer examination of the sequences of these introns revealed several features that distinguish them from U2 introns, including conservation of unusual signals at the donor splice site (|ATATCCTT) and immediately upstream of the

acceptor splice site (TCCTTAAC 10–15 bases from the splice junction) [16,17]. Subsequently, it has been shown that some |GT-AG| introns (the consensus of U2 introns) are also excised by the U12 spliceosome. The U12 spliceosome was first identified and characterized in animals, in which it was found to contain several unique RNA constituents that share structural similarity with and seem to be functionally analogous to the snRNAs contained in the major spliceosome [18,19]. The U12 spliceosome contains several specific, low-abundance snRNPs, namely, U11, U12, U4atac and U6atac and the U5 snRNP, which is present also in the major spliceosome [20].

Major and minor spliceosomal components and both type of introns are present in animals, plants, fungi and at least several unicellular eukaryotes [21]. Given that several of their characteristic constituents are present in representative organisms from all eukaryotic supergroups, both U2 and U12 spliceosomes evolved before the radiation of the supergroups (i.e. at the earliest stages of eukaryotic evolution) [4,12,21]. This conclusion is supported by the recent demonstration that the positions of U12 introns are conserved in orthologous genes from human and *Arabidopsis* to an even greater extent than the positions of U2 introns [11].

The U12-type introns are the minor class of spliceosomal introns in eukaryotic genomes (<1%) [7]. However, the paucity of U12 introns in extant genomes does not rule out the possibility that introns of this type were substantially more abundant at the early stages of eukaryotic evolution or even that U12 introns are the ancestral form of spliceosomal introns, especially given that unidirectional conversion of U12 to U2 introns is apparent in genomic comparisons and that many lineages of eukaryotes have lost U12 introns altogether [4]. Thus, the scenario in which the ancestral introns were U12-type, but subsequent amelioration led to the current excess of U2 introns, is not unrealistic.

Inferring the nature of the primordial introns from protosplice site comparison

To gain insight into the nature of primordial introns, we analyzed putative protosplice sites of ancient introns that retained their positions throughout the course of the evolution of eukaryotes. The idea is to determine whether the primordial protosplice sites correspond to those of U12 or U2 introns. Protosplice sites [8,9] are thought to comprise specific targets for intron insertion into coding sequences of eukaryotic genes. The existence of protosplice sites is indicated by the conservation of nucleotides flanking the splice junctions (Figure 1a,b). In principle, these consensus nucleotides could be remnants of the original protosplice sites or could have evolved convergently after intron insertion. The existence of protosplice sites has been addressed directly by examining the context of introns inserted within codons encoding amino acids that are conserved in all eukaryotes and that, accordingly, are not subject to selection for splicing efficiency [10]. Evidence has been presented that introns are either predominantly inserted into specific protosplice sites, which have the consensus sequence (A/C)AG||Gt, or are inserted randomly but preferentially fixed at such sites [10]. The U12 protosplice sites are distinct from the U2 protosplice sites and have the CT||ATA consensus sequence (Figure 1c,d). This sequence is conserved in human and *Arabidopsis thaliana*, indicating that it has not changed since the divergence of plants and animals from their last common ancestor.

We analyzed the distributions of amino acids in intron-containing sites in which the amino acid is conserved in the sequences of orthologous proteins from eight eukaryotes and five prokaryotes (Supplementary Table S1 and Supplementary Materials and Methods in the

supplementary material online), that is, sites that are subject to extreme evolutionary constraints (hereafter called invariant sites). Such constraints operating at the level of amino acids imply that selection for splicing efficiency had no substantial impact on the intron insertion signal. Thus, this signal, at least to the extent that it covers conserved nucleotides within the respective codon, must have remained intact since the time of intron insertion at an early stage of eukaryotic evolution. Ancient introns, in this case, were defined as those in which positions are conserved in at least two of three major eukaryotic lineages (plants, animals plus fungi, and apicomplexa). All 197 ancient introns (53% of the intron positions that are conserved between animals and plants) found at the invariant sites were of the U2-type.

Putative protosplice sites can be inferred by analyzing amino acid frequency distributions at intron-containing sites [10]. We compared the amino acid distributions at the putative ancient protosplice sites that are derived from the invariant site analysis with the distributions at the sites containing U2 and U12 introns in human and *Arabidopsis* genes. The distributions of amino acids at intron-containing invariant sites were highly non-uniform (Figure 2). Introns occur in three phases, that is, the location of an intron can occur within or between codons; introns of phase 0, 1 and 2 are located between two codons, after the first position in a codon and after the second position, respectively. Each phase has a distinct set of over-represented conserved amino acids (Figure 2a,d and Supplementary Table S2). This effect is especially pronounced for phase 1 in which 71% of primordial introns are located within glycine codons (G|GN) (Figure 2c). This pattern is similar to that seen for U2 introns in phase 1, in which 37% of human introns and 40% of *Arabidopsis* introns are located within glycine codons (Figure 2c), in agreement with the inference that at least a substantial fraction of ancient introns was U2-type. The excess of glycine in the case of ancient introns is a straightforward consequence of the over-representation of glycine in invariant positions (Supplementary Figure S1).

Comparison of the distributions of amino acids that harbor human and *Arabidopsis* U2 and U12 introns revealed an insignificant negative correlation (Supplementary Table S3). This is not unexpected when taking into account the difference between the U2 and U12 inferred protosplice sites (Figure 1). To compare the protosplice sites of primordial introns with U2 and U12 protosplice sites, we employed multiple regression analyses using frequencies of invariant amino acids containing ancient introns as a dependent variable and frequencies of amino acids containing human and *Arabidopsis* U2 or U12 introns as independent variables (Table 1).

A strong and statistically significant positive correlation between the putative ancient protosplice sites and U2 protosplice sites from human and *Arabidopsis* was found both for the raw numbers of amino acids and for normalized values (Table 1 and Supplementary Table S4), thereby explaining a substantial part (>0.64) of the sequence variance of the ancient protosplice sites. This result indicates that most, if not all, of the analyzed primordial introns were U2-type at the time of their insertion at an early stage of eukaryotic evolution rather than being the result of U12 to U2 conversion. It should be noted that this finding in itself is not dependent on the excess of U2 introns in extant genes or even in conserved intron positions but, rather, comes from an unbiased analysis of invariant intron-containing sites.

With respect to the possibility of massive losses of U12 in early eukaryotes, we have shown in a separate recent study that positions of U12 introns are even more strongly conserved between humans and *Arabidopsis* than positions of U2 introns [11]. Therefore, it seems unlikely that all primordial U12 introns have been lost, so at least a substantial majority of the primordial introns probably were of the U2-type.

We cannot rule out the (formal) possibility that a minor fraction of U12 introns was present during the early stages of eukaryotic evolution, although there was no correlation between ancient protosplice sites and U12 protosplice sites (Table 1). We attempted to estimate the sensitivity of the multiple regression analysis using a sampling procedure. Mixtures of U2 and U12 protosplice sites with different proportions of each type (e.g. 10% U12 protosplice sites and 90% U2 protosplice sites) were generated and used as pseudo-ancestral protosplice sites. The results of this simulation show (Supplementary Figure S2) that even a 10% admixture of U12 introns yielded a correlation coefficient value that was significantly lower than the value observed with the real data (Supplementary Table S5). The number of known U12 introns is too small to enable a more precise estimate but the results strongly indicate that, if U12 introns were present among the primordial introns, their fraction was, at best, similar to that in modern genomes. The conclusions of this analysis should be interpreted with caution considering that the invariant sites that are informative for inferring the features of primordial introns comprise but a small fraction of the conserved intron positions and, also, that the statistics on the discrimination between U2 and U12 protosplice sites is weak. Nevertheless, as shown earlier, we currently have no indication of the existence of primordial U12 introns, whereas the evidence in support of primordial U2 introns is clear.

Concluding remarks

The origin of the two types of spliceosomal introns remains a matter of conjecture.

The first scenario to be proposed involved a fission–fusion model in which the two types of introns and the two distinct spliceosomes were combined in the ancestral eukaryote as a result of a fusion of two ancient lineages [4]. However, there seems to be little, if any, independent evidence in support of such a fusion. Perhaps, a more realistic hypothesis holds that U2 and U12 introns descend from two separate invasions of group II self-splicing introns (retroelements) into eukaryotic genes [12]. The present results seem to be compatible with this scenario but indicate a specific succession of the two putative waves of invasion. The U2 introns would be the first to populate the genes to substantial intron densities, followed by the later (but still antedating the radiation of eukaryotic supergroups [13]) invasion of U12 introns that was much more limited in scale owing to the paucity of niches available for insertion of new introns.

Acknowledgments

We thank Ravi Sachidanandam for providing the dump of the SpliceRack (<http://katahdin.cshl.edu:9331/SpliceRack>) database. The authors' research is supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health.

References

1. Maniatis T, Reed R. The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. *Nature*. 1987; 325:673–678. [PubMed: 2950324]
2. Padgett RA, et al. Splicing of messenger RNA precursors. *Annu Rev Biochem*. 1986; 55:1119–1150. [PubMed: 2943217]
3. Collins L, Penny D. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol*. 2005; 22:1053–1066. [PubMed: 15659557]
4. Burge CB, et al. Evolutionary fates and origins of U12-type introns. *Mol Cell*. 1998; 2:773–785. [PubMed: 9885565]
5. Will CL, Luhrmann R. Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol Chem*. 2005; 386:713–724. [PubMed: 16201866]
6. Burge, CB., et al. Splicing of precursors to mRNAs by the spliceosomes. In: Gesteland, RF., et al., editors. *The RNA World II*. 2. Cold Spring Harbor Laboratory Press; 1999. p. 525-560.

7. Sheth N, et al. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.* 2006; 34:3955–3967. [PubMed: 16914448]
8. Dibb NJ. Proto-splice site model of intron origin. *J Theor Biol.* 1991; 151:405–416. [PubMed: 1943150]
9. Dibb NJ, Newman AJ. Evidence that introns arose at proto-splice sites. *EMBO J.* 1989; 8:2015–2021. [PubMed: 2792080]
10. Sverdlov AV, et al. Reconstruction of ancestral protosplice sites. *Curr Biol.* 2004; 14:1505–1508. [PubMed: 15324669]
11. Basu MK, et al. U12 intron positions are more strongly conserved between animals and plants than U2 intron positions. *Biol Direct.* 2008; 3:19. [PubMed: 18479526]
12. Lynch M, Richardson AO. The evolution of spliceosomal introns. *Curr Opin Genet Dev.* 2002; 12:701–710. [PubMed: 12433585]
13. Keeling PJ, et al. The tree of eukaryotes. *Trends Ecol Evol.* 2005; 20:670–676. [PubMed: 16701456]
14. Mount SM. A catalogue of splice junction sequences. *Nucleic Acids Res.* 1982; 10:459–472. [PubMed: 7063411]
15. Rogozin IB, Milanesi L. Analysis of donor splice sites in different eukaryotic organisms. *J Mol Evol.* 1997; 45:50–59. [PubMed: 9211734]
16. Hall SL, Padgett RA. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J Mol Biol.* 1994; 239:357–365. [PubMed: 8201617]
17. Jackson IJ. A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res.* 1991; 19:3795–3798. [PubMed: 1713664]
18. Tarn WY, Steitz JA. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron *in vitro*. *Cell.* 1996; 84:801–811. [PubMed: 8625417]
19. Hall SL, Padgett RA. Requirement of U12 snRNA for *in vivo* splicing of a minor class of eukaryotic nuclear pre-mRNA introns. *Science.* 1996; 271:1716–1718. [PubMed: 8596930]
20. Patel AA, Steitz JA. Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol.* 2003; 4:960–970. [PubMed: 14685174]
21. Russell AG, et al. An early evolutionary origin for the minor spliceosome. *Nature.* 2006; 443:863–866. [PubMed: 17051219]

Figure 1.

Protosplice sites of U2 and U12 introns. Negative numbers indicate the nucleotide positions in the exon immediately preceding the splice junction and the positive numbers indicate the nucleotide positions in the exon immediately after the splice junction. **(a)** Position-specific information content (sequence logo) for the complete set of human U2 splices sites. **(b)** Position-specific information content (sequence logo) for the complete set of *Arabidopsis* U2 splices sites. **(c)** Position-specific information content (sequence logo) for the complete set of human U12 splices sites. **(d)** Position-specific information content (sequence logo) for the complete set of *Arabidopsis* U12 splices sites.

Figure 2. Comparison of amino acid frequencies at protosplice site of primordial, U2 and U12 introns. The distributions are shown separately for each intron phase. 'Phase aa0' and 'Phase 0aa' denote phase 0 introns located immediately downstream and immediately upstream of a conserved amino acid, respectively.

Table 1Multiple regression analyses of the protosplice sites of the primordial introns with U2 and U12 introns^a

Type of introns	Partial correlation coefficient β	P-value
Human U2	<u>0.874</u>	<10 ⁻⁶
Human U12	0.025	0.66
<i>Arabidopsis</i> U2	<u>0.827</u>	<10 ⁻⁶
<i>Arabidopsis</i> U12	-0.008	0.91
Human <i>Arabidopsis</i> U2	<u>0.870</u>	<10 ⁻⁶
Human <i>Arabidopsis</i> U12	0.019	0.74

^aFrequencies of conserved amino acids in sites containing ancient introns were used as the dependent variable and the frequencies of amino acids in sites containing U2 and U12 introns in human and *Arabidopsis* genes were used as independent variables. The partial correlation analysis was performed using the R package (<http://www.r-project.org/>). Bold underlined numbers show statistically significant correlations. For this analysis, raw numbers of amino acids were used; Human+*Arabidopsis* is the sum of the raw numbers for the two species.