Molecular Cell, Vol. 2, 773–785, December, 1998, Copyright ©1998 by Cell Press

Evolutionary Fates and Origins of U12-Type Introns

Christopher B. Burge,* Richard A. Padgett,† and Phillip A. Sharp*‡
*Center for Cancer Research and
Department of Biology
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
†Department of Molecular Biology
Lerner Research Institute
Cleveland Clinic Foundation
Cleveland, Ohio 44195

Summary

U2-type and U12-type introns are spliced by distinct spliceosomes in eukaryotic nuclei. A classification method was devised to distinguish these two types of introns based on splice site sequence properties and was used to identify 56 different genes containing U12-type introns in available genomic sequences. U12-type introns occur with consistently low frequency in diverse eukaryotic taxa but have almost certainly been lost from *C. elegans*. Comparisons with available homologous sequences demonstrate subtype switching of U12 introns between termini of AT-AC and GT-AG as well as conversion of introns from U12-type to U2-type and provide evidence for a fission/fusion model in which the two splicing systems evolved in separate lineages that were fused in a eukaryotic progenitor.

Introduction

A novel class of eukaryotic nuclear pre-mRNA introns was first recognized by Jackson (1991) and Hall and Padgett (1994) on the basis of their unusual terminal dinucleotides: members of this class had /AT at the 5′ splice site (the shill [/] indicates the splice junction) and AC/ at the 3′ splice site, breaking the rule observed in the vast majority of spliceosomal introns, which have /GT at the 5′ splice site and AG/ at the 3′ splice site. Close examination of these introns revealed several properties that distinguish them from other introns (Hall and Padgett, 1994), including conservation of unusual 8- or 9-nucleotide segments at the 5′ splice site (/ATATC CTT) and immediately upstream of the 3′ splice site (TCCTTAAC 10–15 bp from the splice junction). On the basis of sequence complementarity, it was proposed that these segments might interact with portions of two low-abundance small nuclear RNAs (snRNAs) of unknown function, U11 and U12, respectively (Hall and Padgett, 1994). Since then, evidence for these interactions has been obtained, and it has been shown conclusively that this class of introns is spliced by a novel spliceosome with several novel low-abundance small nuclear RNA components (U11, U12, U4atac, and U6atac), as well as the U5 snRNA present also in the major spliceosome (Hall and Padgett, 1996; Tarn and Steitz, 1996a, 1996b). This class of introns and the associated spliceosome are present in the nuclei of vertebrates, insects, and plants (Wu et al., 1996).

The secondary structures and interactions between the snRNAs of the U12-type spliceosome are analogous to those of the snRNAs in the U2-type spliceosome, and the two spliceosomes appear to assemble in similar cycles (Hall and Padgett, 1996; Tarn and Steitz, 1996a, 1996b; reviewed by Tarn and Steitz, 1997; Burge et al., 1998). The recognition of the branch site of U12-type introns by U12 snRNA, with the potential to form an RNA duplex with a bulged adenosine, is analogous to the previously characterized interaction of U2 snRNA with the branch site of U2-type introns (Query et al., 1994). Base pairing interactions of the 5′ splice sites of U12-type introns with U11 and U6atac snRNAs (Tarn and Steitz, 1996a, 1996b; Kolossova and Padgett, 1997; Incorvaia and Padgett, 1998) are also analogous to those described previously for the 5′ splice sites of U2-type introns with U1 and U6 snRNAs (Madhani and Guthrie, 1994). Introns spliced by the U12 spliceosome coexist in the same genes with introns spliced by the major spliceosome, and the two spliceosomes may share common protein components.

Introns with AT-AC terminal dinucleotides can be assigned to one of two groups that correspond to the two types of spliceosome: one group matches the consensus /ATATCCTTT at the 5′ splice site with variations tolerated only at positions +7, +8, and +9, while the other group matches the consensus [AC]AG/AT[AG] AGT ([AC] indicates A or C, etc.) with at most one or two variations. Several examples from the first group including human *P120* intron F (Hall and Padgett, 1996; Tarn and Steitz, 1996a, 1996b) and human sodium channel *SCN4A* intron 2 (Wu and Krainer, 1996) are spliced by the U12 spliceosome in vitro. Examples from the second group including the human sodium channel *SCN5A* intron 25 and human *SCN4A* intron 21 (Dietrich et al., 1997; Wu and Krainer, 1997) are spliced by the U2-type spliceosome in vitro. Thus, for AT-AC introns, the type of spliceosome used in intron excision can be inferred simply from inspection of the 5′ splice site sequence. Known examples of both types of AT-AC introns have been tabulated by Wu and Krainer (1997) and Dietrich et al. (1997).

More recently, it has been shown that a U12-type intron with AT-AC termini can be mutated to termini of GT-AG and still be spliced in a U12-dependent fashion (Dietrich et al., 1997). In fact, it appears that the majority of naturally occurring introns spliced by the U12-type spliceosome have GT-AG termini (Dietrich et al., 1997; Sharp and Burge, 1997). These introns, exemplified by human *ADPRT* intron 22, generally match the 5′ splice site and branch site consensus patterns characteristic of U12-type AT-AC introns with the exception of the terminal nucleotides, suggesting that similar rules are applicable for the processing of this type of intron.

We have developed an objective method based on splice site and branch site sequence properties to predict the type of spliceosome used in excision of GT-AG

[‡] To whom correspondence should be addressed (e-mail: sharppa@mit.edu).

introns. Application of this method to available annotated genomic sequences increased the total number of nonredundant U12-type introns known to about 60. Studies of these U12-type introns and homologous gene sequences characterized the frequency of U12-type introns in different taxa, identified examples of subtype switching between termini of AT-AC and GT-AG, and revealed a novel mode of intron evolution: conversion from U12-type to U2-type splicing. Evidence relating to the evolutionary history of the two spliceosomes was also obtained, leading to the proposal that the two systems evolved in separate lineages that were united by a hypothetical cell fusion or endosymbiosis event in a common ancestor of higher eukaryotes.

Results

Classification of U12-Type versus U2-Type Introns

The first objective was to develop a reliable and objective method to distinguish U12-type from U2-type introns on the basis of sequence properties. As discussed above, this task is relatively straightforward for introns with AT-AC termini. But since the 5′ splice site sequences of U2-type GT-AG introns are notoriously variable in sequence (much more so than for U2-type AT-AC introns), it was necessary to use a somewhat more complex model of the 5′ splice site sequence and to incorporate U12-type branch site sequence properties as well. Because the 5′ splice sites of all U12-type introns are recognized through base pairing with U11 and U6atac snRNAs and the branch sites are recognized by base pairing with U12 snRNA (Hall and Padgett, 1996; Tarn and Steitz, 1996a, 1996b; Dietrich et al., 1997; Kolossova and Padgett, 1997; Incorvaia and Padgett, 1998), the sequence constraints for splicing by the U12 spliceosome are likely to be very similar for introns with GT-AG and AT-AC termini. Therefore, a minimally redundant set of known U12 AT-AC introns was used to construct weight matrix models (see Experimental Procedures) of the 5′ splice site and branch site sequence properties of U12 introns (displayed in Figures 1A and 1C) and simply altered at the +1 position of the 5′ splice site to permit G nucleotides. Corresponding 5′ splice site and 3′ splice site region weight matrices derived from a nonredundant set of 1683 human U2-type GT-AG introns (Figures 1B and 1D) were then used to define normalized log–odds scores, which measure the U12 versus U2 propensity of any 5′ and 3′ splice site sequences (see Experimental Procedures). These scores were calculated for both the known U2-type and U12-type AT-AC introns and for annotated GT-AG introns from GenBank using a program called U12Scan (C. B. B. and P. A. S., unpublished data).

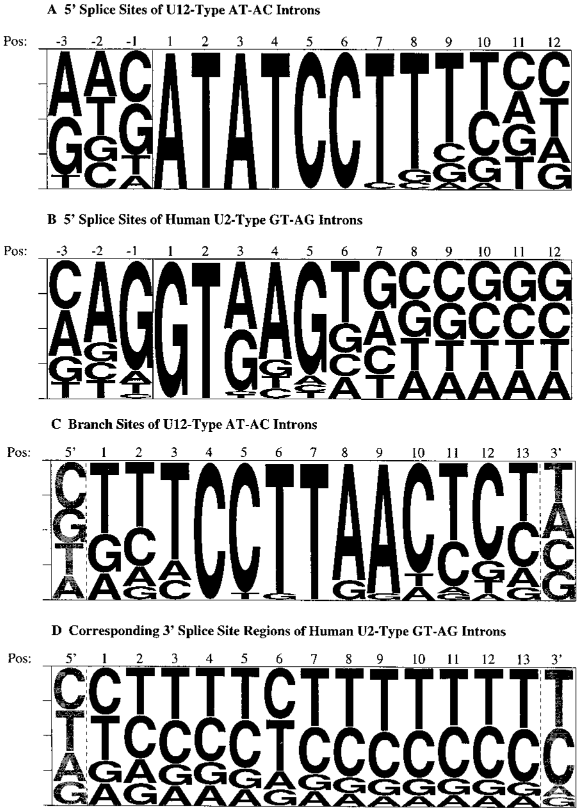

Figure 1. Composition of Splice Sites in U12-Type AT-AC and U2-Type GT-AG Introns

Splice site composition for U12-type AT-AC and U2-type GT-AG introns is displayed using the Pictogram program representation (C. B. B. and P. A. S., unpublished data), in which the frequencies of the four base types at a sequence position are represented by the relative heights of the corresponding letters, in descending order of frequency. Data for a set of 21 minimally redundant U12-type AT-AC introns from a variety of organisms (both animals and plants) are represented in (A) and (C); corresponding data for a nonredundant set of human U2-type GT-AG introns (http://www.molecule.org/cgi/content/full/2/6/773/DC1/6) are represented in (B) and (D).

(A and B) At the 5′ splice site, positions −3, −2, and −1 are the last 3 bp of the exon; positions 1–12 are the first 12 bp of the intron; the splice junction is represented by a vertical line.

(C) Putative branch sites were identified in the U12-type introns by searching the regions [−26,−8] relative to the 3′ ss for matches to the U12-type consensus TCCTTAAC (the branch adenosine is underlined). Alignment of these segments yielded the extended branch point consensus TTTCCTTAACYCY, positions of which are numbered 1–13 in the figure. The average composition of the regions upstream (5′) and downstream (3′) of this 13 bp segment in the region (−28,−5) are illustrated in gray at the left and righthand sides of the panel. The U12-type AT-AC introns used and the branch site locations are listed at http://www.molecule.org/cgi/content/full/2/6/773/DC1/7.

(D) Corresponding 13 bp segments of the U2-type introns were chosen at random from the region (−28,−5) according to the frequency with which this segment was used as the branch site in the U12-type introns.

U12Scan associates with each intron a vector (*x*,*y*) in the plane whose coordinates are the 5′ and 3′ splice site scores of the intron, respectively. The distribution of the points corresponding to all complete vertebrate GT-AG introns annotated in GenBank Release 107 is shown in Figure 2. The overall distribution of these normalized score vectors, reflecting predominantly U2-type introns, appears to follow roughly a standard (circular) bivariate normal distribution with mean vector (0,0), variance (1,1), and zero covariance (Pearson correlation = covariance = 0.03). Statistical outliers relative to this distribution may then be defined as points (*x*,*y*) with test statistic $t^2 = x^2 + y^2 > C$.We chose the criterion $t^2 > 20$ (represented by the dashed line in the figure), corresponding to a threshold for which an outlier is expected

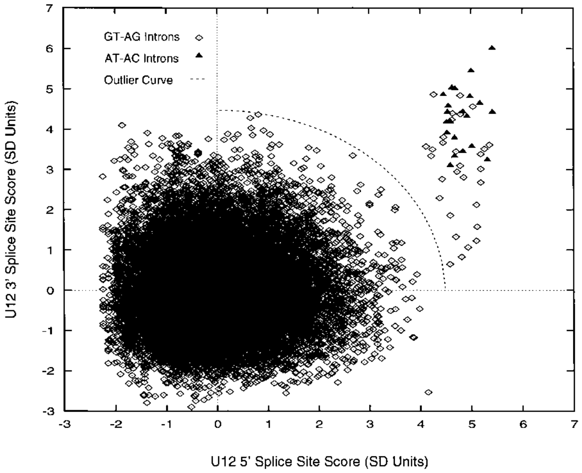

Figure 2. Discrimination of U12-Type versus U2-Type Introns

Weight matrix descriptions of the 5′ splice site (positions −3 to +9) and branch site (positions 1–13) of U12-type AT-AC and human U2-type GT-AG introns were derived from the data of Figure 1 and were used to determine normalized log–odds scores for the 5′ splice site and 3′ splice site of an intron as described in Experimental Procedures. Each vertebrate annotated complete GT-AG intron in GenBank Release 107 is represented by a diamond in the figure at (x,y) where x is the U12 5′ splice site score, and y is the U12 3′ splice site score of the intron (see Experimental Procedures). For reference, the scores of the U12-type AT-AC described in the legend to Figure 1 are also plotted (represented by closed triangles). The dashed line represents a portion of the curve $x^2 + y^2 = 20$, which is used to define outliers as described in Results. The five available vertebrate U2-type AT-AC introns (http://www.molecule.org/cgi/content/full/2/6/773/DC1/8) have scores that fall in the central cluster (not visible).

to occur less than once in 20 samples of size $N = 1000$ (Barnett and Lewis, 1994, Table XXXI). The main central cluster in the figure contains over 99.5% of the data, including many known U2-type GT-AG introns and the five known vertebrate U2-type AT-AC introns, but no known U12-type introns. All U12-type AT-AC introns (represented by closed triangles) are outliers relative to this distribution, falling in the upper right hand corner of the figure, amid a small group of GT-AG outliers with similar scores. Thus, we describe any intron that is an outlier in the first quadrant as a U12-positive intron.

Several lines of evidence suggest that all U12-positive GT-AG introns are in fact spliced by the U12 spliceosome. First, many of these introns are intermingled with the U12-type AT-AC introns in Figure 2, implying that they match the U12 consensus patterns as well as the known U12-type AT-AC introns. Second, several such introns occur in related genes at homologous positions to known U12-type AT-AC introns (Table 2). Finally, in several cases these GT-AG introns are homologous to each other and occur in species that diverged over one hundred million years ago (e.g., human versus *Xenopus laevis*, human versus *Fugu rubripes*—see Table 2), implying the conservation of the extended U12-specific 5′ splice site and branch site consensus sequences over long periods of time (see http://www.molecule.org/cgi/content/full/2/6/773/DC1/1). For these reasons, all U12-positive introns are provisionally considered U12-type introns.

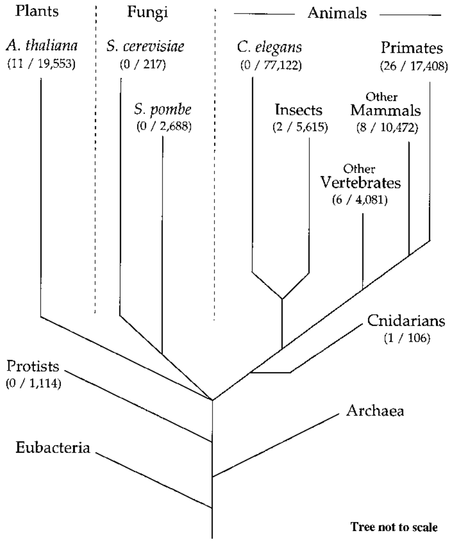

Figure 3. Phylogenetic Distribution and Frequency of U12-Type Introns

The phylogenetic relationships among selected taxonomic groups, based loosely on the 18S rRNA trees of Van de Peer and De Wachter (1997). Below each organism or taxon is listed the number of U12-type introns identified by the U12Scan program (see Experimental Procedures) followed by the total number of complete introns searched. For *S. cerevisiae*, the annotated complete genome was used (Mewes et al., 1997). In all other cases, GenBank Release 107 (June, 1998) was searched. All U12-positive introns (first quadrant outliers—see Figure 2) were considered to be U12-type. In the category of protists, all nonmetazoan invertebrates from the GenBank flatfile gbinv.seq are included. No U12-type introns were found in any fungal or nematode species. Besides *Arabidopsis*, U12-type introns have been reported in other crucifers (Wu et al., 1996). Eubacteria and archaea lack spliceosomal introns.

Phylogenetic Distribution and Frequency of U12-Type Introns

To determine the species distribution of U12-type introns and to identify additional examples of this type, we also searched invertebrate and plant sections of GenBank Release 107 (June, 1998) using similar criteria to distinguish U12-type from U2-type introns (Figure 3). Eleven distinct U12-type introns were found in genes from *Arabidopsis thaliana*, confirming previous observations of this system in higher plants (Wu et al., 1996). U12-type introns are clearly present in multiple classes of vertebrates (at least fish, amphibians, birds, and mammals), as well as multiple insect species (*Drosophila melanogaster* and the silkworm *Bombyx mori*) and in Cnidarians (jellyfish). However, U12 introns (and the associated snRNAs—see below) are absent from the yeast *Saccharomyces cerevisiae* based on the complete annotated genome (Mewes et al., 1997) and are almost certainly absent from the nematode *Caenorhabditis elegans*, based on the nearly complete genomic sequence. No U12-type introns were found in *Schizosaccharomyces pombe* sequences or in any protist, but genomic sequence data are rather incomplete for these organisms. The number of U12-positive introns identified and the total number of introns searched are shown in Figure 3. From these data, the proportion of U12-type introns appears to be consistently very low across all taxa, with the highest frequencies observed in primates (26/17,408 = 0.0015) and nonmammalian vertebrates (6/4081 = 0.0015, or about one U12-type intron for every 700 introns) among taxa with multiple U12-type introns.

A Reference Set of U12-Type Introns

Augmenting these U12-type introns with others identified using the ExonScan program (see Experimental Procedures) and from published sources resulted in the identification of a total of 60 nonredundant U12-type introns, of which a third or more are novel relative to the sets of U12-type introns described previously (Sharp and Burge, 1997; Tarn and Steitz, 1997; Wu and Krainer, 1997). All of the putative U12-type introns identified matched the minimal 5′ splice site consensus /[AG]TATCC except intron 2 of human *CACNA1F* (Table 1), which has /GCATCC-AG/ termini and is homologous to several U12-type GT-AG introns (Table 2). Interestingly, all but one of the U12-type GT-AG introns identified had a non-G nucleotide at position −1 (where G is preferred in U2-type introns—Figure 1B), suggesting that this mutation might be deleterious for U12-type splicing or might often lead to aberrant U2-type splicing; the sole exception is intron 1 of *XPG*.

In cases where clearly homologous U12 introns were found in multiple closely related genes, such as human and *Fugu* Huntington's disease gene, one representative gene (GenBank entry) was chosen as a standard based on the quality of the sequence data and/or annotation. This reference set of U12-type introns/genes is summarized in Table 1. The annotated exon/intron structures of most of these genes derives from comparison with cDNA sequences. However, in a few cases gene structure has been inferred from comparison with homologous proteins (e.g., *Arabidopsis* villin and viviparous isologs); in these cases, the degree of sequence similarity to known proteins is generally quite high. From Table 1, about three quarters (44/60) of U12-type introns have GT-AG terminal dinucleotides. An intron that falls between codons is said to have phase 0; after the first base of a codon, phase 1; and after the second base of a codon, phase 2. In a recent compilation of human introns (predominantly U2-type), Long et al. (1998) found that 44% had phase 0, 36% phase 1, and 20% phase 2. Interestingly, the phase distribution of the 60 nonredundant U12-type introns listed in Table 1 is 14 phase 0 (23%), 22 phase 1 (37%), and 24 phase 2 (40%), similar between plant and animal genes, but significantly different by a χ^2 test from the phase frequencies reported by Long et al. for U2-type introns ($P < 0.001$, 2 df).

To identify homologous genes sequenced at the genomic level, the translated amino acid sequence of each gene from Table 1 was searched against the GenBank DNA sequence database (releases 106 and 107) using TBLASTN 2.0 (Altschul et al., 1997) with default parameters. Whenever a BLAST hit was identified that was significant at the level $E < 10^{-6}$, the sequence of the coding regions flanking the known U12-type intron (the query) was aligned by hand to the homologous genomic sequence (the subject) to determine the relative locations of introns in the two sequences. Sequences that could not be reliably aligned in the region of the U12-type intron were excluded from further consideration. Three types of results were typically obtained in these comparisons: a U12-type intron of the same phase was located at the identical position in the alignment in the subject genomic sequence (32 examples); a U2-type intron of the same phase was located at the identical alignment position in the genomic sequence (12 examples); or no intron was found at nearby positions in the subject genomic sequence (approximately 100 examples in animals and plants). These three types of results are discussed in separate sections below.

Conservation of U12-Type Introns and Switching between AT-AC and GT-AG Subtypes

Examples of the first type, conserved U12-type introns, are listed in Table 2; alignments are shown at http://www.molecule.org./cgi/content/full/2/6/773/DC1/1. Many U12-type introns are conserved between mouse and human homologs, three are conserved between different classes of vertebrates (*CMP*, Huntington's disease gene, and *RPL1*), and the Prospero U12-type intron is conserved between an insect and a vertebrate. Strikingly, the second intron of the sodium channel α subunit gene is conserved between mammals and the jellyfish *Polyorchis penicillatus*, lineages that diverged at least 600–800 million years ago (Spafford et al., 1998). In addition, homologous U12-type introns are present in multiple (paralogous) sodium and calcium channel α subunit genes (Wu and Krainer, 1997), implying the conservation of these introns at least since the duplications that gave rise to these gene families (Strong et al., 1993).

Known U12-type introns can be grouped into two subtypes: those with GT-AG (or GC-AG) terminal dinucleotides and those that have AT-AC (or AT-AA or AT-AG) termini. The data in Table 2 show a few examples of pairs of homologous U12-type introns that belong to different subtypes (see also Dietrich et al., 1997), suggesting that the two subtypes may interconvert at some rate over evolutionary time scales, a process we call subtype switching. Specific examples include the second intron of sodium channel genes (AT-AC in mammals, GT-AG in jellyfish), the Prospero intron (AT-AC in *Drosophila*, GT-AG in *Fugu*), and the first/second calcium channel intron, for which AT-AG as well as AT-AC and GT-AG terminal dinucleotide pairings are observed among the different human isoforms. This last example suggests a plausible mechanism by which subtype switching could occur (AT-AC $\leftrightarrow$ AT-AG $\leftrightarrow$ GT-AG) without requiring two specific simultaneous mutations (probably extraordinarily rare). The Prospero U12-type intron is unusual in that subtype switching appears to have been accompanied by movement of the 3′ splice site by three bases (see legend to Table 2).

Through a BLASTN search of the GenBank DNA sequence database with the human U6atac snRNA sequence (Tarn and Steitz, 1996b), we identified putative U6atac snRNA genes in *Drosophila* and *Arabidopsis* genomic sequences, both significant at $E < 1.1 \times 10^{-5}$, which have conserved promoter elements and can be folded into similar secondary structures (GenBank accession numbers AC004423 and AB006702, respectively). The *Arabidopsis* gene, at least, is expressed as a small

Table 1. Reference Set of Nonredundant U12-Type Introns

No.	Gene	Organism	Intron	of	Termini	Phase	Function of Protein
1	*ActL*	Human	5	12	GT-AG	2	Actin-like
2	*ADPRT*	Human	22	22	GT-AG	2	POLY-ADP ribose polymerase
3	*AOX1*	Human	2	34+	GT-AG	1	Aldehyde oxidase
4	*ARP*	*Arabidopsis*	1	2	GT-AG	0	Acidic ribosomal prot. P1 isolog
5	*AZGP1*	Rat	1B	2+	GT-AG	2	Zn-α2-glycoprotein (MHC1)
6	*BAP*	Human	6	9	AT-AC	0	B cell receptor assoc. prot.
7	*CACNA1F*	Human	2	47	GC-AG	2	Ca^{+2} channel α subunit
8	*CACNA1F*	Human	16	47	GT-AG	1	Ca^{+2} channel α subunit
9	*CANP*	Human	5	9	GT-AG	0	Ca^{+2}-act. neutral protease
10	*CDK5*	Mouse	9	11	AT-AC	2	Cyclin-dependent kinase 5
11	*CK2*β	*Arabidopsis*	4	4	GT-AG	1	Casein kinase II β subunit
12	*CLCN6*	Human	5	22	GT-AG	1	Putative Cl$^-$ channel
13	*CMP*	Chicken	7	7	AT-AC	1	Cartilage matrix protein
14	*c-Raf-1*	Human	7	16+	GT-AG	1	MAP-KKK family prot. kinase
15	*DMC1*	*Arabidopsis*	9	14	GT-AG	2	RecA homol.; meiotic recomb.
16	*DMC1*	*Arabidopsis*	14	14	AT-AC	0	RecA homol.; meiotic recomb.
17	*DUB*	*Arabidopsis*	11	15	GT-AG	1	Deubiquitinating enzyme
18	*E2F1*	Human	4	6	AT-AC	2	Transcription factor; oncoprotein
19	*ERK2*	Mouse	2	8	GT-AG	2	MAP-kinase family prot. kinase
20	*FHIT*	Human	5	10	GT-AG	1	Fragile histidine triad
21	*G5p*	*Arabidopsis*	7	19	AT-AA	1	Similar to yeast Sac1
22	*GT334*	Human	12	22	GT-AG	2	aka TMEM1; sim. to EHOC-1
23	*GT335*	Human	6	6	AT-AC	0	aka HES1; unknown function
24	*HISS*	*Fugu*	4	12	GT-AG	0	Histidyl-tRNA synthetase
25	*HPS*	Human	16	19	AT-AC	2	Hermansky-Pudlak syndrome
26	*HD*	*Fugu*	66	66	GT-AG	2	Huntington's disease homolog
27	*HUPF1*	Human	8	22	GT-AG	1	Type 1 RNA helicase
28	*INSIG1*	Human	2	4	GT-AG	0	Insulin-induced protein 1
29	*LON*	*Arabidopsis*	11	19	GT-AG	2	Lon protease
30	Luminidep.	*Arabidopsis*	10	12	GT-AG	1	Regulates flowering time
31	*Met-8604*	Human	4	7	GT-AG	0	Unknown function
32	*Met-AP*	*Arabidopsis*	13	14	GT-AG	1	Methionine aminopeptidase
33	*MSH3*	Human	6	23	AT-AA	1	DNA repair; MutS homolog
34	*MYH9*	Human	6	39	GT-AG	1	Nonmuscle myosin type A
35	*OCRL*	Human	8	10	GT-AG	1	Lowe oculocerebrorenal synd.
36	*P120*	Human	7	14	AT-AC	1	Proliferating cell nucleolar prot.
37	*PBGD*	Human	8	13	GT-AG	0	Porphobilinogen deaminase
38	*PLC*β*3*	Human	16	30	GT-AG	2	Phospholipase C-β-3
39	*PLC*δ*1*	Hamster	3	15+	GT-AG	2	Phospholipase C-δ-1
40	*PTEN*	Human	1	8+	GT-AG	1	Dual-spec. protein phosphatase
41	Prospero	*Drosophila*	2	3	AT-AC	0	Homeodomain; neurogenesis
42	*R05D3.2* hom.	*Fugu*	12	14	GT-AG	2	Similar to *C. elegans* R05D3.2
43	*RPB5*	Human	5	6	GT-AG	2	RNA Pol II 23 kDa protein
44	*RPL1a*	*X. laevis*	3	9	GT-AG	0	Ribosomal protein L1a
45	*SAPK4*	Human	8	11	GT-AG	1	Stress-activated protein kinase 4
46	*SCN4A*	Human	2	23	AT-AC	2	Skeletal muscle Na$^+$ channel α
47	*SmE*	Human	1	4+	GT-AG	0	Core snRNP protein E
48	*SPS*	Mouse	6	10	AT-AC	0	Spermine synthase
49	*SPS*	Mouse	10	10	GT-AG	2	Spermine synthase
50	*SRPK1*	Human	6	15	GT-AG	1	SR protein kinase 1
51	*STK11*	Human	2	8	AT-AC	2	Serine/threonine kinase 11
52	*TCP1*	Mouse	5	11	GT-AG	2	T complex polypeptide 1
53	*TFIIS.oA*	*X. laevis*	6	9	AT-AC	1	Tx. elongation factor TFIIS
54	*TSPY*	Rat	1	5+	GT-AG	0	Testicular prot.; Y chromosome
55	*Tum-P35B*	Mouse	8	8	GT-AG	1	tum-Transplant. antigen P35B
56	Villin isolog	*Arabidopsis*	1	21	GT-AG	2	Similar to Villin
57	Vivip.-1 iso.	*Arabidopsis*	4	11	GT-AG	2	Similar to Viviparous-1
58	*XDH1*	*B. mori*	2	7	GT-AG	2	Xanthine dehydrogenase
59	*XPG*	Mouse	1	14	GT-AG	1	Xeroderma pigmentosum gr. G
60	*XPG*	Mouse	13	14	AT-AC	2	Xeroderma pigmentosum gr. G

The reference set of U12-type introns used in this study are numbered in alphabetical order, followed by the name of the gene and organism, the number of the U12-type intron within the gene, the total number of introns in the gene, the terminal intron dinucleotides, the phase of the intron, and a short description of the function of the encoded protein. The GenBank accession number or reference corresponding to each numbered table entry are listed at http://www.molecule.org/cgi/content/full/2/6/773/DC1/5.

RNA (G. Shukla and R. A. P., unpublished data). In contrast, BLASTN searches with human U6atac and other available vertebrate U12-specific snRNA sequences (Montzka and Steitz, 1988; Tarn et al., 1995; Tarn and Steitz, 1996b) failed to identify homologous genes in *C. elegans* despite the nearly complete genomic sequence,

Table 2. Homologous U12-Type Introns

Reference U12-Type Intron		Homologous U12-Type intron				
No.	Termini	Organism	Gene	Intron	Termini	Accession No.
6	AT-AC	Mouse	*BAP*	6	AT-AC	AC002397
7	GT-AG	*Fugu*	*CACNL*	1	GT-AG	AF026198
7	GT-AG	Human	*CACNL1A1*	2	AT-AG	Z26257
7	GT-AG	Human	*CACNL1A2*	2	GT-AG	D43747
7	GT-AG	Human	*CACNL1A3*	1	GT-AG	U50666
7	GT-AG	Human	*CACNL1A4*	1	AT-AC	Z80114
8	GT-AG	*Fugu*	*CACNL*	13	GT-AG	AF026198
8	GT-AG	Human	*CACNL1A1*	15	GT-AG	Z26257
8	GT-AG	Human	*CACNL1A2*	16	GT-AG	D43747
8	GT-AG	Human	*CACNL1A3*	13	GT-AG	U50666
8	GT-AG	Human	*CACNL1A4*	16	GT-AG	Z80114
12	GT-AG	Mouse	*CLCN6*	5	GT-AG	AF030104
13	AT-AC	Human	*CMP*	7	AT-AC	M55681
25	AT-AC	Mouse	*HPS*	15	AT-AC	U78966
26	GT-AG	Human	*HD*	66	GT-AG	L27416
33	AT-AA	Mouse	*REP3*	6	AT-AC	L10300
34	GT-AG	Human	*MyHC*	5	GT-AG	AF001548
41	AT-AC	*Fugu*	Prospero	—	GT-AG	U40760
44	GT-AG	*X. laevis*	*RPL1b*	3	GT-AG	X67691
44	GT-AG	*X. tropicalis*	*RPL1t*	3	GT-AG	X67692
44	GT-AG	Human	*RPL1h*	3	GT-AG	X72205
45	GT-AG	Human	*SAPK4h*	—	GT-AG	Z95152
46	AT-AC	Human	*SCN5A*	2	AT-AC	*
46	AT-AC	Human	*SCN8A*	2	AT-AC	*
46	AT-AC	Mouse	*SCN8A*	2	AT-AC	U59964
46	AT-AC	*P. penicill.*	*SCN1*	2	GT-AG	AF047379
48	AT-AC	Human	*SPS*	6	AT-AC	U53331
49	GT-AG	Human	*SPS*	10	GT-AG	U53331
54	GT-AG	Human	*TSPY*	1	GT-AG	M98524
54	GT-AG	Bovine	*TSPY*	1	GT-AG	X74028
58	GT-AG	*B. mori*	*XDH2*	2	GT-AG	AB005911
59	GT-AG	Human	*XPG*	1	GT-AG	X71341

U12-type introns that are homologous to the reference U12-type introns from Table 1 are listed. The first two columns list the number (from Table 1) and terminal dinucleotides of the reference U12-type intron; the remaining columns refer to the homologous U12-type intron. The last column gives the GenBank accession number for each sequence; the asterisk indicates sequences from George et al. (1995) that are not in GenBank. The gene labeled *SAPK4h* is a putative homolog of *SAPK4* identified in the same genomic contig as *SAPK4* by TBLASTN search. In all cases, the homologous U12-type intron had the same phase as the reference intron and in all but one case appeared to be located at the corresponding position relative to the coding sequence. The sole exception was the Prospero intron, for which the 3′ splice site in *Fugu* appears to be 3 base pairs 3′ of the location of the 3′ splice site in *Drosophila*.

providing further evidence that this system is absent from nematodes. The yeast *S. cerevisiae* also lacks the snRNA genes specific to the U12 spliceosome.

Conversion of U12-Type Introns to U2-Type Introns

Surprisingly, 12 examples of U2-type introns were found that occur at positions homologous in terms of codon location and phase to U12-type introns (Table 3). Representative alignments showing the homologous locations of three such intron pairs are shown in Figure 4; alignments for the other pairs are given at http://www.molecule.org./cgi/content/full/2/6/773/DC1/2. How can the occurrence of so many pairs of U2-type and U12-type introns at homologous positions be explained? One possibility is that such pairs result from instances of intron loss followed by coincidental insertion of an intron of the other type at the identical codon position. Because introns are distributed only about once per 100–200 base pairs (bp) of coding sequence in most of the organisms represented in the table, such loss and coincidental insertion are expected to occur between distantly related genes with a frequency of at most 0.5%–1%, so such an explanation is extremely unlikely to account for more than a few of the examples in Table 3.

Instead, it is likely that intron conversion has occurred, that is, mutational conversion of U12-type introns to splicing by the U2-type spliceosome. Such a process was previously conjectured by Dietrich et al. (1997), and they showed that the spliceosome used for excision of an intron in vitro could be switched from U12-type to U2-type by introduction of specific splice site mutations, albeit with a change in the location of the 3′ splice site. However, the data summarized in Table 3 provide the first convincing evidence that this process has actually occurred in evolution. These naturally occurring introns also differ from the experimentally constructed example in that the splice site locations appear to have remained unaltered by the conversion process (the sole exception is described in the legend to Table 3). Interestingly, all of the U12-type introns that have apparently converted to U2-dependent splicing are of the GT-AG subtype (Table 3). As noted previously (Dietrich et al., 1997; Sharp

Table 3. U12-Type Introns that Have Converted to U2-Type

Reference U12-Type Intron		Homologous U2-Type Intron			
No.	Termini	Organism	Gene	Intron	Accession No.
4	GT-AG	*Arabidopsis*	*ARP* homol.	1	AB010068
4	GT-AG	*S. pombe*	*rpa1*	1	AL022070
7	GT-AG	*C. elegans*	*C48A7.1*	2	U61951
8	GT-AG	*C. elegans*	*Unc-2*	11	U25119
18	GT-AG	*C. elegans*	*F43C1.2*	4	Z46937
34	GT-AG	Human	*MYH6*	7	X52889
34	GT-AG	Chicken	*MyH*	7	J02714
34	GT-AG	*C. elegans*	*Myo-3*	4	X08067
34	GT-AG	*C. elegans*	*Unc-54*	4	J01050
40	GT-AG	*C. elegans*	Tensin homol.	1	AF036706
58	GT-AG	Mouse	*XDH*	3	X75129
58	GT-AG	*C. elegans*	*F55B11.1*	2	Z83318

U2-type introns that are homologous to the reference U12-type introns from Table 1 are listed. The left half of the table gives the number (from Table 1) and terminal dinucleotides of the reference U12-type intron; the right half gives data for the homologous U2-type intron. In all cases, the homologous U2-type intron had the same phase as the reference intron and in all but one case appeared to be located at the homologous position within the coding sequence. The exception was the *C. elegans* F55B11.1 gene, for which the 5′ splice site appears to be 6 bp 5′ of the location of the 5′ splice site of the U12-type intron in *B. mori XDH1*.

and Burge, 1997), the much more highly constrained splice site sequences of U12-type introns strongly suggest that the reciprocal process (conversion from U2-type to U12-type) would be extremely improbable, that is, intron conversion is a one-way street.

Loss of U12-Type Introns

Examples of approximately 100 genes lacking introns at or near the site of a U12-type intron in homologous plant or animal genes, suggesting intron loss, are listed at http://www.molecule.org/cgi/content/full/2/6/773/DC1/3. The absence of an intron of either type at a position that is homologous to that of a U12-type intron could be explained either by intron loss in one lineage or intron insertion in the other. However, in the absence of any evidence for mobility of U12-type introns in contemporary organisms, this difference is attributed primarily to intron loss, probably by the same processes

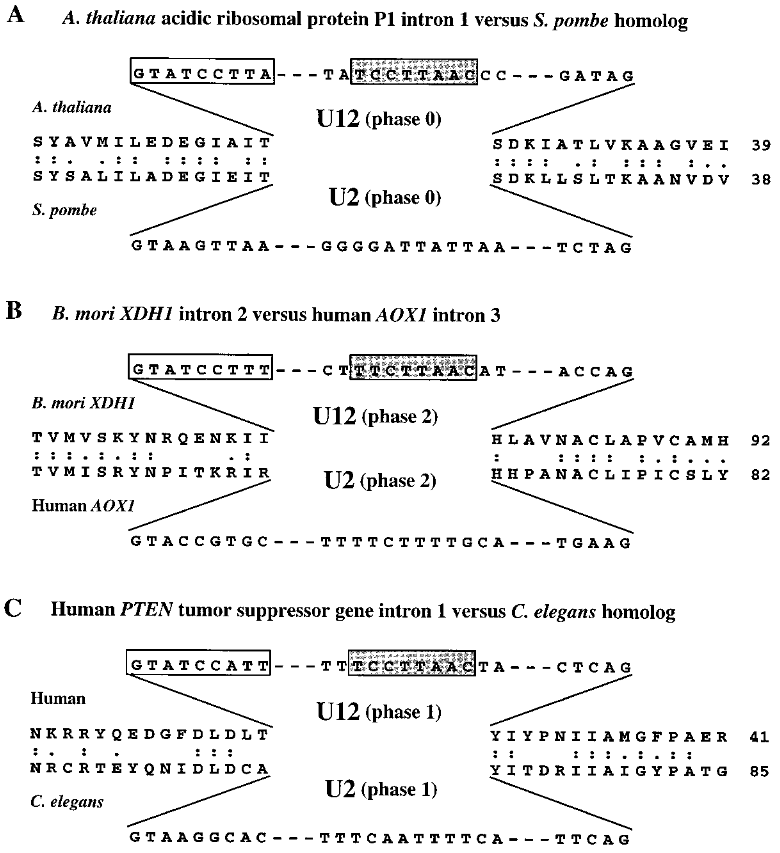
A
A. thaliana acidic ribosomal protein P1 intron 1 versus *S. pombe* homolog
GTATCCTTA---TATCCTTAACCC---GATAG
A. thaliana
U12 (phase 0)
SYAVMILEDEGIAIT
SDKIATLVKAAGVEI 39
SYSALILADEGIEIT
SDKLLSLTKAANVDV 38
U2 (phase 0)
S. pombe
GTAAGTTAA---GGGGATTATTAA---TCTAG
B
B. mori XDH1 intron 2 versus human *AOX1* intron 3
GTATCCTTT---CTTTCTTAACAT---ACCAG
B. mori XDH1
U12 (phase 2)
TVMVSKYNRQENKII
HLAVNACLAPVCAMH 92
TVMISRYNPITKRIR
HHPANACLIPICSLY 82
U2 (phase 2)
Human AOX1
GTACCGTGC---TTTTCTTTTGCA---TGAAG
C
Human *PTEN* tumor suppressor gene intron 1 versus *C. elegans* homolog
GTATCCATT---TTTCCTTAACTA---CTCAG
Human
U12 (phase 1)
NKRRYQEDGFDLDLT
YIYPNIIAMGFPAER 41
NRCRTEYQNIDLDCA
YITDRIIAIGYPATG 85
U2 (phase 1)
C. elegans
GTAAGGCAC---TTTCAATTTTCA---TTCAG

Figure 4. Homologous U12-Type and U2-Type Introns

Examples of U2-type introns that occur at positions homologous to those of U12-type introns. In each case, the locations of the U12-type and U2-type introns are shown above and below the aligned translated flanking exons, respectively, together with the phase (codon position) of the intron. Residue positions in the complete proteins are shown at right. The distinctive U12-type 5′ splice site and branch site sequences are indicated by unshaded and shaded boxes, respectively; the corresponding portions of the U2-type introns are shown below. In the alignments, identical residues are indicated by double dots; similar residues that score positively in the BLOSUM62 amino acid substitution matrix (Henikoff and Henikoff, 1993) are indicated by single dots. See Tables 1 and 3 for accession numbers. (A) The U12-type intron occurs between the indicated Thr and Ser codons of *Arabidopsis ARP*. (B) The U12-type intron occurs in the indicated Ile codon of *XDH1*. Note that human *AOX1* has a nonhomologous U12-type GT-AG intron (phase 1) at residue 35 (not shown). (C) The U12-type intron occurs in the indicated Tyr codon of human *PTEN*.

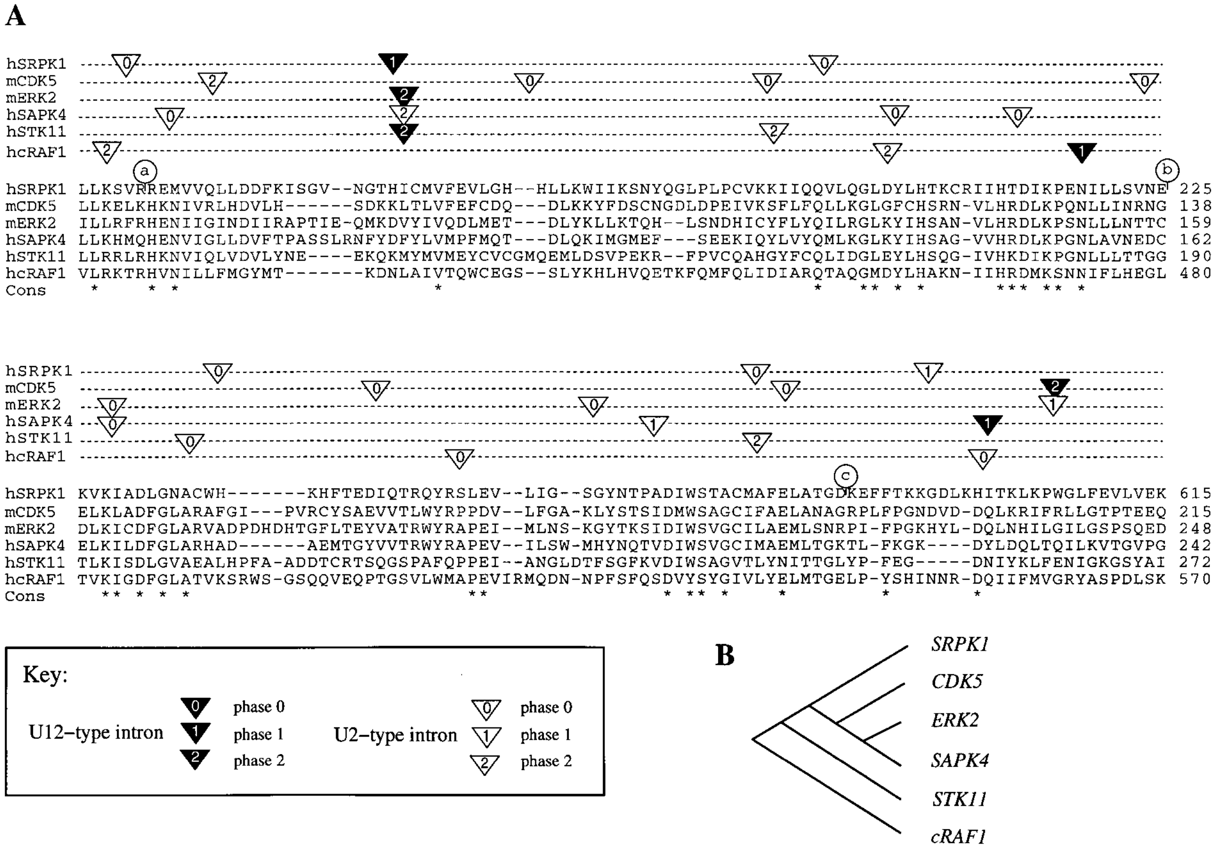
A
hSRPK1
mCDK5
mERK2
hSAPK4
hSTK11
hcRAF1
a
b
hSRPK1 LLKSVRREMVVQLLDDFKISGV--NGTHICMVFEVLGH--HLLKWIIKSNYQGLPLPCVKKIIQQVLQGLDYLHTKCRIIHTDIKPENILLSVNE 225
mCDK5 LLKELKHKNIVRLHDVLH------SDKKLTLVFEFCDQ---DLKKYFDSCNGDLDPEIVKSFLFQLLKGLGFCHSRN-VLHRDLKPQNLLINRNG 138
mERK2 ILLRFRHENIIGINDIIRAPTIE-QMKDVYIVQDLMET---DLYKLLKTQH--LSNDHICYFLYQILRGLKYIHSAN-VLHRDLKPSNLLLNTTC 159
hSAPK4 LLKHMQHENVIGLLDVFTPASSLRNFYDFYLVMPFMQT---DLQKIMGMEF---SEEKIQYLVYQMLKGLKYIHSAG-VVHRDLKPGNLAVNEDC 162
hSTK11 LLRRLRHKNVIQLVDVLYNE----EKQKMYMVMEYCVCGMQEMLDSVPEKR--FPVCQAHGYFCQLIDGLEYLHSQG-IVHKDIKPGNLLLTTGG 190
hcRAF1 VLRKTRHVNILLFMGYMT-------KDNLAIVTQWCEGS--SLYKHLHVQETKFQMFQLIDIARQTAQGMDYLHAKN-IIHRDMKSNNIFLHEGL 480
Cons
c
hSRPK1 KVKIADLGNACWH-------KHFTEDIQTRQYRSLEV--LIG--SGYNTPADIWSTACMAFELATGDKEFFTKKGDLKHITKLKPWGLFEVLVEK 615
mCDK5 ELKLADFGLARAFGI---PVRCYSAEVVTLWYRPPDV--LFGA-KLYSTSIDMWSAGCIFAELANAGRPLFPGNDVD-DQLKRIFRLLGTPTEEQ 215
mERK2 DLKICDFGLARVADPDHDHTGFLTEYVATRWYRAPEI--MLNS-KGYTKSIDIWSVGCILAEMLSNRPI-FPGKHYL-DQLNHILGILGSPSQED 248
hSAPK4 ELKILDFGLARHAD------AEMTGYVVTRWYRAPEV--ILSW-MHYNQTVDIWSVGCIMAEMLTGKTL-FKGK----DYLDQLTQILKVTGVPG 242
hSTK11 TLKISDLGVAEALHPFA-ADDTCRTSQGSPAFQPPEI--ANGLDTFSGFKVDIWSAGVTLYNITTGLYP-FEG-----DNIYKLFENIGKGSYAI 272
hcRAF1 TVKIGDFGLATVKSRWS-GSQQVEQPTGSVLWMAPEVIRMQDN-NPFSFQSDVYSYGIVLYELMTGELP-YSHINNR-DQIIFMVGRYASPDLSK 570
Cons
Key:
U12-type intron
phase 0
phase 1
phase 2
U2-type intron
phase 0
phase 1
phase 2
B
SRPK1
CDK5
ERK2
SAPK4
STK11
cRAF1

Figure 5. U12-Type and U2-Type Introns in Six Mammalian Protein Kinase Genes

(A) The positions of U2-type and U12-type introns in six mammalian protein kinase genes that contain a U12-type intron are illustrated: human *SRPK1*, mouse *CDK5*, mouse *ERK2*, human *SAPK4*, human *STK11*, and human *cRaf1*. Intron locations were derived from the GenBank sequence annotation (accession numbers given at http://www.molecule.org/cgi/content/full/2/6/773/DC1/5) and are displayed schematically at the location of the nearest encoded amino acid(s) in the flanking exons. Amino acid positions in the complete protein sequence are shown at right. The amino acid sequences were aligned as described in the text. Residues that are conserved in 5 or more of the 6 kinase proteins are indicated by asterisks. The circled letters a, b, and c indicate positions where the human *SRPK1* protein has a large insertion relative to the other five sequences.

(B) The evolutionary relationships between the six genes illustrated in (A) are shown, based on the maximum parsimony analysis of Hanks and Hunter (1995). Branch lengths are not to scale.

of reverse transcription and gene conversion that are thought to cause loss of U2-type introns (Boeke et al., 1985; Derr and Strathern, 1993).

Since the presence of the U12-type splicing system in both higher animals and Cnidarians, as well as plants (Figure 3), implies that this system was present at one time in the ancestor of nematodes, the complete absence of U12-type introns from homologous genes in *C. elegans* is particularly noteworthy. Of the 12 U12-type GT-AG introns for which a homologous gene region in *C. elegans* was found, 7 had a U2-type intron at the homologous position, consistent with intron conversion (Table 3), and 5 had no nearby intron, consistent with intron loss. For U12-type introns with AT-AC termini, three homologous gene regions in *C. elegans* were found, and all lacked introns at the corresponding position (see http://www.molecule.org/cgi/content/full/2/6/773/DC1/3). These observations suggest that U12-type introns with AT-AC termini have been eliminated from nematodes primarily through intron loss, while conversion and loss have both played significant roles in elimination of U12-type GT-AG introns.

Homologous and Nonhomologous U12-Type Introns in Distantly Related Paralogous Genes

To identify paralog relationships (common ancestry through gene duplication) between genes containing U12-type introns, an all against all comparison of the proteins encoded by genes from Table 1 was performed using BLASTP 2.0 (Altschul et al., 1997) installed locally. The results identified four pairs or groups of paralogous genes: aldehyde oxidase (*AOX1*) and xanthine dehydrogenase (*XDH1*); sodium and calcium channel α subunits; phospholipase C β-3 and phospholipase C δ-1; and the protein kinase genes *CDK5*, *c-Raf-1*, *ERK2*, *SAPK4*, *SRPK1*, and *STK11*. Alignments of these genes showed that the U12-type sodium channel intron is clearly homologous to the first U12-type calcium channel intron, as are the U12-type introns in *ERK2* and *STK11* (Figure 5). Surprisingly, however, given the extreme rarity of U12-type introns, these alignments also revealed several examples of U12-type introns at nonhomologous positions in paralogous genes. In one example, phospholipase C β-3 and phospholipase C δ-1, the U12-type introns are far apart (one near the 5′ end of the gene, the other near the middle). In the *AOX1/XDH1* comparison, the two U12-type introns are quite close to each other (109 bp apart) but have different phases and are located at clearly nonhomologous positions (see Figure 4B and its legend). In this case, the presence of a U2-type intron in human *AOX1* at a position homologous to that of the U12-type intron in *XDH1* (Figure 4B) strongly implies that the common ancestor of these paralogs had two U12-type introns.

To further investigate the history of the U12-type introns found in protein kinase genes, the six proteins were aligned to the Pfam eukaryotic protein kinase (pkinase) hidden Markov model (Sonnhammer et al., 1997) using the Pfam web server (http://pfam.wustl.edu), and the locations of all introns in the relevant regions were identified based on the GenBank annotation (Figure 5). In the figure, at least three or four distinct (nonhomologous) U12-type introns appear to be present. Particularly interesting is the region near the upper left of the figure where phase 2 U12-type introns in mouse *ERK2* and human *STK11* and a U2-type phase 2 intron in human *SAPK4* occur at identical alignment positions. The simplest explanation for this pattern is that the *SAPK4* intron resulted from conversion of an ancient U12-type intron to U2-type. The presence in *SAPK4* of a U12-type intron at a distinct position hundreds of base pairs away (Figure 5) implies that the ancestor of this group of protein kinases had at least two U12-type introns. The occurrence of a phase 1 U12-type intron in human *SRPK1* at an adjacent codon position is curious and suggests that an unusual intron sliding or phase shifting event may have taken place. A study of U2-type introns by Stoltzfus et al. (1997) found no evidence for widespread intron sliding (lateral movement to nearby codon locations) but did not rule out its occasional occurrence.

Discussion

The data presented clearly document two significant types of evolutionary phenomena: subtype switching of U12 introns between AT-AC and GT-AG terminal dinucleotides, and conversion of U12-type introns to splicing by the U2-type system. The latter finding suggests that U12-type introns may once have been far more common than they are today and raises questions about the evolutionary relationship between the two spliceosomes.

How might conversion from a U12-type to a U2-type intron take place? Since the 5′ splice site consensus (/GTATCCTTT or /ATATCCTTT) is so highly conserved in U12-type introns, even more than the branch site, it is likely that mutations of this sequence are often involved in the conversion process. For example, a C to G mutation at the universally conserved +5 position of the 5′ splice site (from /GTATCC to /GTATGC) of a U12-type GT-AG intron abolished splicing by the U12-type spliceosome and activated the site for splicing by the U2-type spliceosome in vitro and in vivo (Dietrich et al., 1997). Interestingly, all of the examples we found of U12-type introns that have converted to U2-type had GT-AG terminal dinucleotides. The apparently very low (or nil) frequency of conversion of U12-type AT-AC introns to U2-type probably reflects a requirement for multiple simultaneous mutational changes to create a functional U2-type intron. Therefore, a common pathway for the evolution of a U12-type AT-AC intron may involve subtype switching to a U12-type GT-AG intron, followed by later conversion to a U2-type GT-AG intron. As noted previously (Dietrich et al., 1997; Sharp and Burge, 1997), the much more highly conserved 5′ splice site and branch point sequences of U12-type introns compared to the variability of U2-type splice sites imply that intron conversion is likely to occur only in the U12 to U2 direction.

An issue that remains in the proposed conversion process is whether the 3′ end of a U12-type intron could be recognized by the U2-type splicing system. The trinucleotide YAG/ at the 3′ splice site (Y = C or T) is common to both splicing systems. Further, mammalian U2-type branch site sequences are highly variable (Nelson and Green, 1989), suggesting that many U12-type introns may contain a potentially functional U2-type branch site. Finally, the presence of a tract of pyrimidine nucleotides just upstream of the 3′ splice site is an important feature in the splicing of U2-type introns (Reed, 1989). In introns undergoing conversion, this feature might be provided by the U12 branch site (extended consensus: TTTCCTTAACYCY), which is quite rich in pyrimidines and is typically located at a distance of 10–16 bp upstream of the 3′ splice junction, similar to the location of the polypyrimidine tract in some U2-type introns. These three observations suggest that consensus U12-type introns may often contain a rudimentary U2-type 3′ splice site.

Strong evidence has been presented that conversion of U12-type introns to U2-type introns has occurred often in animals. Two examples suggest that this process has occurred also in plants and fungi, so intron conversion may be universal in eukaryotes. However, the rate of intron conversion has probably not been constant in different lineages, since the specificity of the U2-type splicing system differs between different groups of eukaryotes (e.g., Lopato et al., 1996). In particular, while the consensus sequences of U12-type introns appear to be the same in diverse organisms (Sharp and Burge, 1997, Table 2), the 5′ splice sites of *C. elegans* U2-type introns are somewhat different from human, with a bias toward T at positions +7, +8, and +9 (data not shown) that is not seen in human U2-type introns (Figure 1B). Since the U12-type consensus also has T at these three positions, a greater proportion of mutations in U12-type introns may have created functional U2-type introns in the nematode lineage, leading to a higher rate of intron conversion. This higher rate of intron conversion, coupled with a constant rate of intron loss, offers a plausible explanation for the apparent absence of U12-type introns and the U12-type spliceosome from nematodes.

Evolutionary History of the Two Spliceosomes

The occurrence of analogous structures or enzymatic activities in living organisms is often the result of common ancestry/homology but may sometimes also result from convergent evolution in the presence of similar functional constraints. Whether the two spliceosomes are homologous, and if so, whether they are related by speciation or by intragenomic duplication, is a question

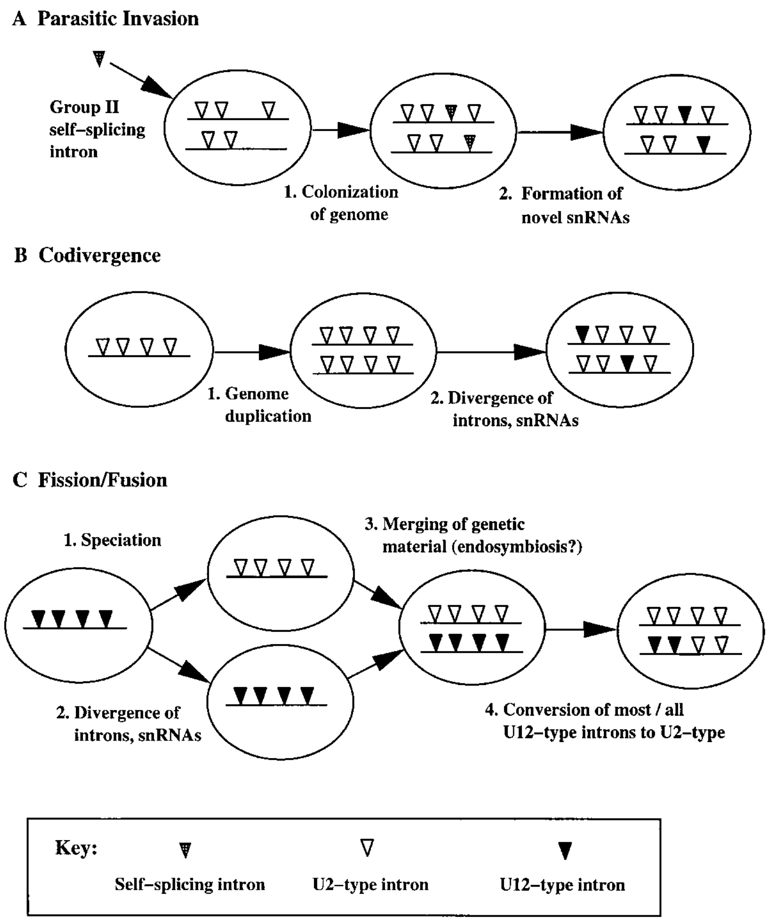
A Parasitic Invasion
Group II self–splicing intron
1. Colonization of genome
2. Formation of novel snRNAs
B Codivergence
1. Genome duplication
2. Divergence of introns, snRNAs
C Fission/Fusion
1. Speciation
2. Divergence of introns, snRNAs
3. Merging of genetic material (endosymbiosis?)
4. Conversion of most / all U12–type introns to U2–type
Key:
Self–splicing intron
U2–type intron
U12–type intron

Figure 6. Models for the Evolution of the Spliceosomes

The three models for the evolution of the U2-type and U12-type spliceosomes discussed in the text are illustrated. Organisms are represented by ovals, the genetic material is represented by horizontal lines, and introns are represented by triangles. Different types of introns (U2-type, U12-type, and self-splicing) are distinguished by shading. In a set of genes chosen on the basis of containing at least one U12-type intron, the convergent evolution and codivergence models predict that the number Z that contain two U12-type introns should have approximately a binomial distribution with parameters N = the number of genes in the set, and $p = 1 - (1 - x)^{m-1}$, where m is the mean number of introns per gene in the set, and x is the proportion of introns in the genome that are U12-type. Among the $N = 56$ genes containing at least one U12-type intron listed in Table 1, the average number of introns is $m = 15$. The value of the frequency x is not known precisely, but using the highest value observed in any taxon with multiple U12-type introns (0.0015 in primates; Figure 3) should result in a conservative statistical test. Setting $x = 0.0015$, the probability P for Z to be greater than or equal to 4 (the actual number of genes observed with two U12-type introns) is less than 0.03, allowing the convergent evolution and codivergence models to be rejected. Note that this analysis is still valid if one assumes some rate of random insertion of U2-type (and/or U12-type) introns into the genome.

of significant interest. These three possible scenarios for the evolution of the two spliceosomes are illustrated in Figure 6.

A possible nonhomologous origin of the two spliceosomes is the convergent evolution model illustrated in Figure 6A, in which a parasitic group II self-splicing intron entered the genome of a common precursor of animals and plants, which already had U2-type introns and a developed spliceosome, and was inserted into a number of genes. In this model, the self-splicing RNA structure of this intron fragmented to give rise to a new set of U12-type snRNAs that could excise these introns in *trans*, as may have occurred in the evolution of the U2-type splicing system (Sharp, 1991).

By contrast, the codivergence and fission/fusion models assume that the U12 and U2 systems are homologous (Tarn and Steitz, 1997). The codivergence model (Figure 6B) proposes that the entire set of snRNAs (with the possible exception of U5) was duplicated in a primordial organism, followed by coordinated divergence of two subsets of introns, together with the two sets of snRNA (and protein) components.

In the fission/fusion model (Figure 6C), the divergence between the two systems is envisioned to have occurred following speciation of two separate lineages, each with a single spliceosome and intron type, followed by a merging of the genetic material, perhaps through endosymbiosis, in a progenitor of higher eukaryotes. Under this model, the second (U12-type) spliceosome was retained to splice out the introns from the genes acquired through endosymbiosis, so that at one point the two types of introns may have had roughly similar frequencies in the genome, but over time the vast majority of these introns were either lost or converted to the U2-type system. Those snRNA and protein components that had not diverged extensively, for example, perhaps U5 snRNA and some protein factors common to both spliceosomes, became redundant following the union of the two lineages so that only one functional version of each was retained. An alternative explanation for the presence of U5 snRNA in both spliceosomes (Tarn and Steitz, 1997) is that this factor arose as a general splicing cofactor after the unification of the two spliceosomes.

All three of the models described above involve unusual types of evolutionary events, and yet one of these scenarios must be close to the truth. Evidence for homology of the two systems, arguing against the convergent evolution model, comes from the extensive similarities in secondary structures and interactions between the set of U11, U12, U4atac, and U6atac snRNAs of the U12 spliceosome and the set of U1, U2, U4, and U6 snRNAs of the U2 spliceosome, and there are also similarities in promoter structure between analogous snRNA genes (Tarn and Steitz, 1996b). However, U11 and U12 are not detectably similar in sequence to the analogous snRNAs U1 and U2 (Montzka and Steitz, 1988), and the degree of direct sequence similarity between U4atac and U6atac snRNAs and their counterparts U4 and U6 is only between 30% and 40% (Tarn and Steitz, 1996b), insufficient to definitively prove common ancestry. Further evidence relating to this issue will likely come from identification of the protein components of the U12 spliceosome.

Interestingly, the three models differ in their predictions about the distribution of U12-type introns in contemporary genes. The convergent evolution and codivergence models predict that U12-type introns were originally (and should still be) randomly distributed in the genome. On the other hand, the fission/fusion model predicts that there are two classes of genes, distinguished by their origins. One class of genes, derived from the U2-only organism, should not contain U12-type introns. Members of the other class of genes, derived from the U12-only organism, had only U12-type introns at one time, but with the passage of time most (or all) of these introns were either lost or converted to U2-type. The key point is that members of this latter class should presently have a distinctly higher proportion of U12-type introns than in the genome as a whole. In the available set of 56 genes that contain one or more U12 introns (Table 1), a subgroup of four genes contains two U12-type introns. Since only about 1 out of every 700 introns in modern genomes is U12-type, if these introns were distributed randomly among all genes, the mostly likely outcome would be 1 or 0 genes with multiple U12-type introns in this pool of 56. In fact, the probability that 4 or more genes out of 56 would have two U12-types is $P < 0.03$ (see legend to Figure 6), inconsistent with the convergent evolution and codivergence models. Thus, currently available gene sequence data support a fission/fusion origin of the eukaryotic nuclear genome.

Besides the fission/fusion model, another possible explanation for these data is that the U12-type introns in the four genes that contain two such introns (calcium channel α subunit, *DMC1*, spermine synthase, and *XPG*) might not represent distinct ancient introns but were instead derived by partial gene duplications early in the evolution of these genes. However, this explanation appears unlikely. First, based on sequence comparisons, only the calcium channel gene appears to be derived by duplication of subunits (Strong et al., 1993), and in this case the two U12-type introns occur at clearly distinct positions within the duplicated units. Second, in each of the four genes the two U12-type introns have different phases, again arguing against a common origin by partial gene duplication.

Given the extreme rarity of U12-type introns, it is quite surprising that multiple examples of paralogous genes (genes related by intragenomic duplication) were found that contain nonhomologous U12-type introns. The occurrence of many such paralog pairs provides further support for the fission/fusion model, since these genes would represent ancestral U12-only genes in which different sets of U12 introns happened to be lost/converted in the two lineages after gene duplication. For example, it is very likely that the common ancestor of *AOX1* and *XDH1* contained two nearby U12-type introns. This possibility is particularly strong in the case of the protein kinases, for which at least three or four distinct U12-type introns are present (Figure 5).

If one accepts that all or most of the genes containing U12-type introns may have come from a distinct ancestral organism, it is natural to ask whether they share any common features suggestive of the nature of this organism or of its contributions to modern eukaryotes. Previously, several groups have shown that most of the core eukaryotic proteins involved in DNA replication/repair, transcription, and translation are more similar to orthologous genes from archaea than from eubacteria (reviewed by Brown and Doolittle, 1997). Recently, Rivera et al. (1998) have extended this finding to include most or all genes involved in information processing, as opposed to operational genes such as those involved in energy metabolism, fatty acid and phospholipid biosynthesis, and so forth, for which the eukaryotic genes are often more similar to eubacteria. These and other studies based on protein sequence data strongly suggest a chimeric origin for the eukaryotic nuclear genome (reviewed by Gupta and Golding, 1996).

Interestingly, a large proportion of the genes that contain U12-type introns encode proteins in the category of information processing, including proteins involved in DNA replication/repair (DMC1, MSH3/REP3, and XPG), transcription (E2F1, Prospero, RPB5, and TFIIS), RNA processing (HUPF1, SmE, and SRPK1), and translation (ARP, HisS, P120, and RPL1a), with few operational proteins such as porphobilinogen synthase. In several specific cases, detailed phylogenetic reconstructions have confirmed a closer relationship of these eukaryotic proteins to the archaea: examples include DMC1 (Brendel et al., 1997), HisS (Feng et al., 1997), TFIIS, and TCP1 (Brown and Doolittle, 1997). The phylogenetic origins of eukaryotic porphobilinogen synthase are murkier, with archaeal and eubacterial sequences intermixed in phylogenetic trees (Feng et al., 1997). Thus, both protein sequence comparisons and the distribution of U12-type introns argue for a chimeric origin of the eukaryotic nucleus, and these phenomena may reflect the same evolutionary event.

Experimental Procedures

Definition of U12 Splice Site Scores

Weight matrix descriptions (Staden, 1984) of U12-type splice site sequences were derived from the data described in the legend to Figure 1 as follows. At each position *i* of the 5′ splice site, the count $N_j^{(i)}$ of each nucleotide type *j* was determined, and the probability $p_j^{(i)}$ of base *j at position i* was set equal to $p_j^{(i)} = (N_j^{(i)} + 1/4)/(N_j^{(i)} + 1)$, where $N^{(i)}$ is the number of available sequences. This amounts to adding the equivalent of one sequence of random composition as a pseudocount to compensate for the small data set size of available U12-type introns. Assuming independence between positions, the U12 5′ splice site probability, $P_{5'ss}^{U12}(X)$, for a 5′ splice site sequence *X* is then defined as

$$P_{5'ss}^{U12}(X) = \prod_{i=-3}^{9} p_{X_i}^{(i)}$$

where X_i indicates the nucleotide at position *i* relative to the splice junction. The U2 5′ splice site probability, $P_{5'ss}^{U2}$, is defined similarly, based on data from U2-type introns (see legend to Figure 1).

Using the 13 positions of the extended U12 branch site consensus (Figure 1C) to derive position-specific probabilities leads to an analogous definition for the U12 3′ splice site probability, $P_{3'ss}^{U12}(X)$, which may be calculated for any 13 bp segment *X*. The product of the probabilities from the corresponding regions of U2-type 3′ splice sites (see legend to Figure 1) defines the U2 3′ splice site probability, $P^{U2}_{3'ss}(X)$. To score a 3′ splice site, the values of $P^{U12}_{3'ss}(X)$ and $P^{U2}_{3'ss}(X)$ are calculated for each 13 bp segment *X* in the range $(-28,-5)$ relative to the 3′ splice site, corresponding to a branch site to 3′ splice site distance of 9–20 bp (Sharp and Burge, 1997), and the maximum values $P_{\max}{}^{U12}_{3'ss}$ and $P_{\max}{}^{U2}_{3'ss}$, respectively, are retained. The U12 5′ splice site and 3′ splice site log–odds ratios are then defined as $L^{U12}_{5'ss} = \log_2(P^{U12}_{5'ss}/P^{U2}_{5'ss})$ and $L^{U12}_{3'ss} = log_2(P_{\max}{}^{U12}_{3'ss}/P_{\max}{}^{U2}_{3'ss})$, respectively. Values of these log–odds ratios were calculated for each annotated complete GT-AG intron in the vertebrate partitions of GenBank (release 107), and the sample mean and variance determined. Finally, these log–odds ratios were normalized to z scores by subtracting off the appropriate sample mean and dividing by the sample standard deviation to give the U12 5′ splice site and 3′ splice site scores, $S^{U12}_{5'ss}$ and $S^{U12}_{3'ss}$. These scores, which are distributed roughly (but not precisely) as normal(0,1) random variables, provide a convenient measure of the U12 propensity of an intron and can be used as the x and y coordinates of a point in the plane corresponding to each intron (Figure 2). The use of log–odds ratios in discrimination is common and well justified statistically. Normalizing these values by dividing by the sample standard deviation tends to equalize the relative contributions of the 5′ and 3′ splice site sequences to the score vector of an intron and yields roughly a standard bivariate normal distribution for which standard tests for outliers are available (see Results).

Database Searches for U12-Type Introns

To identify U12-type introns, the flat files from GenBank releases 106 (April, 1998) and 107 (June, 1998) were retrieved by anonymous ftp from ftp.ncbi.nlm.nih.gov/pub/genbank and searched using the U12Scan and ExonScan computer programs (C. B. B. and P. A. S., unpublished data) on a Silicon Graphics O2 workstation. The U12-Scan program uses annotated CDS (coding sequence) and intron features to identify the locations of introns in a GenBank sequence and then scores the splice site sequences as described above. (Only introns completely contained in a single GenBank sequence are searched by U12Scan.) The ExonScan program is similar but uses annotated exon features to identify 5′ and 3′ splice sites and so is capable of identifying putative U12-type introns that have been submitted as separate (incomplete) GenBank entries covering the upstream and downstream exons. Both programs can also check for apparent annotation errors: a few of the U12-type introns listed were identified by allowing for up to a 6–base pair difference relative to the annotated splice site locations—these cases are described at http://www.molecule.org/cgi/content/full/2/6/773/DC1/4.

Acknowledgments

We thank P. S. McCaw, B. Panning, and T. H. Tuschl for helpful comments on the manuscript, C. C. Allende for helpful information about casein kinase II β, and M. Siafaca for her assistance. This work was supported by United States Public Health Service MERIT Award R37-GM34277 and Grant RO1-AI32486 from the National Institutes of Health (to P. A. S.). R. A. P. is supported by grant RO1-GM55105 from the National Institutes of Health. C. B. B. is supported by the National Science Foundation under a 1998 NSF/Alfred P. Sloan postdoctoral fellowship in Molecular Evolution.

Received August 7, 1998; revised September 29, 1998.

References

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

Barnett, V., and Lewis, T. (1994). Outliers in Statistical Data, Third Edition (New York: John Wiley and Sons).

Boeke, J.D., Garfinkel, D.J., Styles, C.A., and Fink, G.R. (1985). Ty elements transpose through an RNA intermediate. Cell *40*, 491–500.

Brendel, V. , Brocchieri, L., Sandler, S.J., Clark, A.J., and Karlin, S. (1997). Evolutionary comparisons of RecA-like proteins across all major kingdoms of living organisms. J. Mol. Evol. *44*, 528–541.

Brown, J.R., and Doolittle, W.F. (1997). Archaea and the prokaryote-to-eukaryote transition. Microbiol. Mol. Biol. Rev. *61*, 456–502.

Burge, C.B., Tuschl, T.H., and Sharp, P.A. (1998). Splicing of precursors to mRNAs by the spliceosomes. In The RNA World II, R.F. Gesteland, T. Cech, and J.F. Atkins, eds. (Cold Spring Harbor, NY: Cold Spring Laboratory Press), in press.

Derr, L.K., and Strathern, J.N. (1993). A role for reverse transcripts in gene conversion. Nature *361*, 170–173.

Dietrich, R.C., Incorvaia, R., and Padgett, R.A. (1997). Terminal dinucleotide sequences do not distinguish between U2- and U12-dependent introns. Mol. Cell *1*, 151–160.

Feng, D.-F., Cho, G., and Doolittle, R.F. (1997). Determining divergence times with a protein clock: update and reevaluation. Proc. Natl. Acad. Sci. USA *94*, 13028–13033.

George, A.L., Iyer, G.S., Kleinfield, R., Kallen, R.G., and Barchi, R.L. (1993). Genomic organization of the human skeletal muscle sodium channel gene. Genomics *16*, 598–606.

Gupta, R.S., and Golding, G.B. (1996). The origin of the eukaryotic cell. Trends Biochem. Sci. *21*, 166–171.

Hall, S.L., and Padgett, R.A. (1994). Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. J. Mol. Biol. *239*, 357–365.

Hall, S.L., and Padgett, R.A. (1996). Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. Science *271*, 1716–1718.

Hanks, S.K., and Hunter, T. (1995). The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. FASEB J. *9*, 576–596.

Henikoff, S., and Henikoff, J.G. (1993). Performance evaluation of amino acid substitution matrices. Proteins *17*, 49–61.

Incorvaia, R., and Padgett, R.A. (1998). Base pairing with U6atac is required for 5′ splice site activation of U12-dependent introns in vivo. RNA *4*, 709–718.

Jackson, I.J. (1991). A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Res. *19*, 3795–3798.

Kolossova, I., and Padgett, R.A. (1997). U11 snRNA interacts in vivo with the 5′ splice site of U12-dependent (AU-AC) introns. RNA *3*, 227–233.

Long, M., de Souza, S.J., Rosenberg, C., and Gilbert, W. (1998). Relationship between “proto-splice sites” and intron phases: evidence from dicodon analysis. Proc. Natl. Acad. Sci. USA *95*, 219–223.

Lopato, S., Mayeda, A., Krainer, A.R., and Barta, A. (1996). Pre-mRNA splicing in plants: characterization of Ser/Arg splicing factors. Proc. Natl. Acad. Sci. USA *93*, 3074–3079.

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

Mewes, H.W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S.G., et al. (1997). Overview of the yeast genome. Nature *387*, 7–65.

Montzka, K.A., and Steitz, J.A. (1988). Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc. Proc. Natl. Acad. Sci. USA *85*, 8885–8889.

Nelson, K.K., and Green, M.R. (1989). Mammalian U2 snRNP has a sequence-specific RNA-binding activity. Genes Dev. *3*, 1562–1571.

Query, C.C., Moore, M.J., and Sharp, P.A. (1994). Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model. Genes Dev. *8*, 587–597.

Reed, R. (1989). The organization of 3′ splice-site sequences in mammalian introns. Genes Dev. *3*, 2113–2123.

Rivera, M.C., Jain, R., Moore, J.E., and Lake, J.A. (1998). Genomic evidence for two functionally distinct gene classes. Proc. Natl. Acad. Sci. USA *95*, 6239–6244.

Sharp, P.A. (1991). “Five easy pieces.” Science *254*, 663.

Sharp, P.A., and Burge, C.B. (1997). Classification of introns: U2-type or U12-type. Cell *91*, 875–879.

Sonnhammer, E.L., Eddy, S.R., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins *28*, 405–420.

Spafford, J.D., Spencer, A.N., and Gallin, W.J. (1998). A putative voltage-gated sodium channel α subunit (PpSCN1) from the hydrozoan jellyfish, *Polyorchis penicillatus*: structural comparisons and evolutionary considerations. Biochem. Biophys. Res. Comm. *244*, 772–780.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. *12*, 505–519.

Stoltzfus, A., Logsdon, J.M., Jr., Palmer, J.D., and Doolittle, W.F. (1997). Intron "sliding" and the diversity of intron positions. Proc. Natl. Acad. Sci. USA *94*, 10739–10744.

Strong, M., Chandy, K.G., and Gutman, G.A. (1993). Molecular evolution of voltage-sensitive ion channel genes: on the origins of electrical excitability. Mol. Biol. Evol. *10*, 221–242.

Tarn, W.-Y., and Steitz, J.A. (1996a). A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. Cell *84*, 801–811.

Tarn, W.-Y., and Steitz, J.A. (1996b). Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. Science *273*, 1824–1832.

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

Tarn, W.-Y., Yario, T.A., and Steitz, J.A. (1995). U12 snRNA in vertebrates: evolutionary conservation of 5′ sequences implicated in splicing of pre-mRNAs containing a minor class of introns. RNA *1*, 644–656.

Van de Peer, Y., and De Wachter, R. (1997). Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA. J. Mol. Evol. *45*, 619–630.

Wu, Q., and Krainer, A.R. (1996). U1-mediated exon definition interactions between AT-AC and GT-AG introns. Science *274*, 1005–1008.

Wu, Q., and Krainer, A.R. (1997). Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs. RNA *3*, 586–601.

Wu, H.J., Gaubier-Comella, P., Delseny, M., Grellet, F., Van Montagu, M., and Rouze, R. (1996). Non-canonical introns are at least 10(9) years old. Nat. Genet. *14*, 383–384.