Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion

Nuno L. Barbosa-Morais,[1,2] Maria Carmo-Fonseca,[2] and Samuel Aparício[1,3,4]

[1]*University of Cambridge, Department of Oncology, Hutchison-MRC Research Centre, Cambridge CB2 2XZ, United Kingdom;* [2]*Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Av. Prof. Egas Moniz, 1649-028 Lisboa, Portugal*

Although more than 200 human spliceosomal and splicing-associated proteins are known, the evolution of the splicing machinery has not been studied extensively. The recent near-complete sequencing and annotation of distant vertebrate and chordate genomes provides the opportunity for an exhaustive comparative analysis of splicing factors across eukaryotes. We describe here our semiautomated computational pipeline to identify and annotate splicing factors in representative species of eukaryotes. We focused on protein families whose role in splicing is confirmed by experimental evidence. We visually inspected 1894 proteins and manually curated 224 of them. Our analysis shows a general conservation of the core spliceosomal proteins across the eukaryotic lineage, contrasting with selective expansions of protein families known to play a role in the regulation of splicing, most notably of SR proteins in metazoans and of heterogeneous nuclear ribonucleoproteins (hnRNP) in vertebrates. We also observed vertebrate-specific expansion of the CLK and SRPK kinases (which phosphorylate SR proteins), and the CUG-BP/CELF family of splicing regulators. Furthermore, we report several intronless genes amongst splicing proteins in mammals, suggesting that retrotransposition contributed to the complexity of the mammalian splicing apparatus.

[Supplemental material is available online at www.genome.org.]

In most eukaryotes, functional messenger RNAs (mRNAs) are produced by accurately removing noncoding sequences (introns) from precursors (pre-mRNAs) in a process termed "RNA splicing." The spliceosome, a large multicomponent ribonucleoprotein complex, carries out this intron excision (Burge et al. 1999; Jurica and Moore 2003). Extensive genetic and biochemical studies in a variety of systems have revealed that the spliceosome contains five essential small RNAs (snRNAs), each of which functions as an RNA–protein complex called a small nuclear ribonucleoprotein (snRNP). Each snRNP comprises one of these five snRNAs bound stably to two classes of proteins, i.e., Sm proteins, which are present in all snRNPs, and specific proteins that are uniquely associated with only one snRNP (Luhrmann et al. 1990). Higher eukaryotes have two distinct types of spliceosomes. The major or U2-type spliceosome, which catalyzes the removal of most introns, is composed of U1, U2, U4, U5, and U6 snRNPs. The minor or U12-type spliceosome, which recognizes <1% of all human introns, comprises U11, U12, U4atac, U5, and U6atac snRNPs (Patel and Steitz 2003). In addition to snRNPs, splicing requires many non-snRNP protein factors. Recent improved methods to purify spliceosomes coupled with advances in mass spectrometry have revealed that the spliceosome may be composed of as many as 300 distinct proteins (Jurica and Moore 2003; Nilsen 2003).

The initial events of spliceosome assembly require recognition of specific sequences located at the 5′ and 3′ splice sites, which define the intron boundaries. In metazoans, however, the splice site sequences are only weakly conserved and although introns are excised with a high degree of precision, at least 74% of human genes encode alternatively spliced mRNAs (Johnson et al. 2003). Alternative splicing is the process by which multiple mRNAs can be generated from the same pre-mRNA by the differential joining of 5′ and 3′ splice sites. Alternative splicing produces multiple mRNAs encoding distinct proteins, thus expanding the coding capacity of genes and contributing to the proteomic complexity of higher organisms (Brett et al. 2002; Maniatis and Tasic 2002; Black 2003).

In general, alternative splicing is regulated by protein factors that recognize and associate with specific RNA sequence elements either to enhance or to repress the ability of the spliceosome to recognize and select nearby splice sites (Smith and Valcarcel 2000; Maniatis and Tasic 2002). The multiplicity of protein–protein and protein–RNA interactions that modulate the association of the spliceosome with the pre-mRNA is thought to control alternative splicing (Caceres and Kornblihtt 2002; Graveley 2002; Black 2003).

The evolutionary history of the splicing machinery has not been fully elucidated, in part because appropriate near-complete genome sequences have only recently become available. The recent sequencing and annotation of the genomes of the Japanese puffer fish, *Fugu rubripes* (Aparicio et al. 2002), and the sea squirt, *Ciona intestinalis* (Dehal et al. 2002), allows us now to fill that gap with fiducial branches of distant vertebrates and chordates, respectively, providing an opportunity to exhaustively look at splicing factors in those species and extend our knowledge about their evolution. In this study, we report a semiautomated computational pipeline designed to identify and annotate splicing factors in representative species of eukaryotes.

Results

Pipeline-assisted annotation of splicing factors

Although recent reports have identified up to 300 distinct proteins associated with the spliceosome (Rappsilber et al. 2002;

[3]Present address: BC Cancer Agency, Vancouver V5Z 1L3, Canada.
[4]Corresponding author.
E-mail saparicio@bccrc.ca; fax 1 604-675-8219.
Article published online ahead of print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.3936206.

16:66–77 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

Zhou et al. 2002), many of these new proteins have not yet been shown to function in splicing and, therefore, they cannot be considered as bona fide splicing factors (Jurica and Moore 2003). In this study, we limited our analysis to proteins for which there is experimental evidence of their involvement in splicing. Our first goal was to enumerate and annotate the genes encoding spliceosomal proteins in the genomes of human, pufferfish, *Ciona*, the budding yeast *Saccharomyces cerevisiae*, the fission yeast *Schizosaccharomyces pombe*, the plant *Arabidopsis thaliana*, and several species of archaebacteria and protozoa (see Methods; Fig. 1; Supplemental Table S1).

Although the availability of "raw" and "first pass annotated" genomes (for example the ones in Ensembl [Hubbard et al. 2002]) is proving indispensable for genome-wide studies, detailed analyses are still hampered by the fact that most databases are "contaminated" with erroneous annotation. In many cases, the current algorithms used in completely automated gene-building pipelines unreliably predict features such as short exons. The algorithms are particularly ineffective with repetitive protein motifs, such as those in RS (arginine-serine-rich) domains, responsible for the protein–protein interactions of SR (Ser-Arg) proteins (important splicing factors—see below). The goal of our semiautomated pipeline was to search ab-initio the raw genomic sequence of representative eukaryotes and thus to complement pre-existing annotations, even though these acted as a seed for the pipeline. This approach demanded manual inspection and validation of the results. We therefore visually inspected a total of 1894 putative spliceosomal proteins across eukaryotic genomes, and we manually curated 224 sequences (12%). The results are listed in Supplemental Table S2. Despite the effort made to manually correct sequences, errors and uncertainties remain, especially for genes poorly supported with EST evidence, and this reduces the precision of the phylogenetic analysis (namely for Parsimony methods) and the consistency of tree topology between different methods of phylogenetic inference (all of the trees can be found in the Supplemental materials). We were unable to correct completely 388 proteins (20%) of ambiguous sequence. We identified only five putative splicing factors (all from *Fugu*) that had no previously annotated gene locus. We also report three factors (from Zebrafish) that were annotated in older versions of Ensembl, but do not appear in version 30. In the process of manual curation, we have identified 83 putative pseudo-genes that Ensembl annotates as active genes in human and mouse (Supplemental Table S3; see below), indicating that automated annotation is oversensitive.

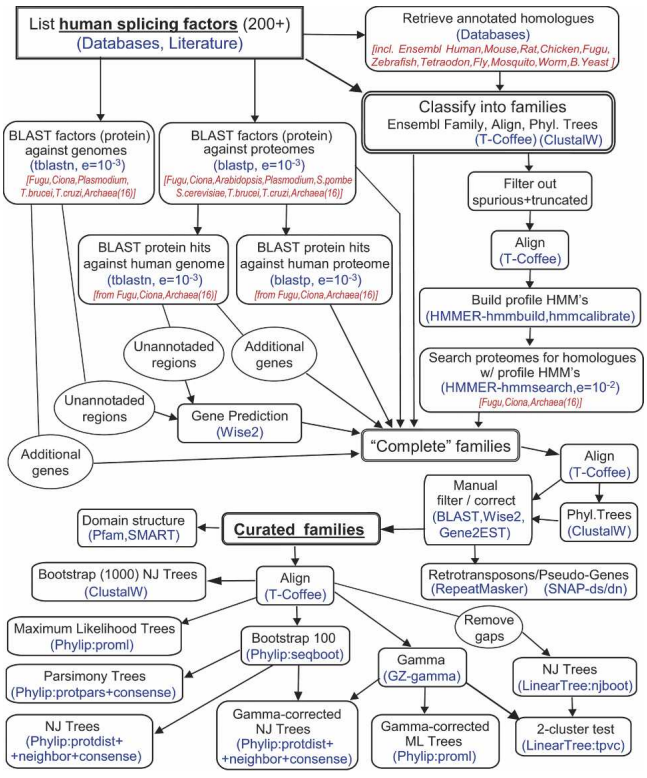

Figure 1. Schematics of the computational pipeline flow. Sources, software, and parameters are represented in blue and species in red.

Selective expansion of splicing regulatory protein families

Having enumerated all currently known splicing proteins, we asked whether major patterns of protein family expansion were evident between different animal phyla. We looked at the genes encoding the seven Sm protein families that associate with all the snRNAs, the Lsm protein families that associate with the U6 snRNA, and several snRNP-specific protein families. Most of these spliceosomal components show apparent one:one orthology mapping (or numerical concordance in the occurrence of paralogs) between vertebrates, invertebrates, and unicellular eukaryotes, consistent with previous reports (see Will and Luhrmann 2001). In contrast, we observed a different evolutionary pattern of the minor spliceosome U11/U12-snRNP proteins; they are absent from protozoa, trypanosomes, yeasts, and the nematode worm *Caenorhabditis elegans* (Table 1), but present in *Arabidopsis*, consistent with the identification of U12-dependent introns in this plant (Zhu and Brendel 2003).

In addition to snRNPs, the spliceosome comprises many non-snRNP protein factors, including DExD/H-box proteins, SR proteins, and hnRNP proteins. DExD/H-box proteins constitute a prominent family of core splicing factors. Genetic studies in *S. cerevisiae* have implicated eight DExD/H-box proteins in splicing (Staley and Guthrie 1998). Each of these conserved proteins (Prp2p, Prp16p, Prp22p, Prp43p, Brr2, Prp5p, Prp28p, Sub2p) is required for pre-mRNA splicing. Seven additional DExD/H-box proteins were recently found associated with mammalian spliceosomes (Jurica and Moore 2003). As shown in Table 1, no major expansion of the DExD/H-box gene family occurred during evolution.

The SR proteins, characterized by their typical RS domain containing repeated Arg/Ser dipeptides, are essential factors required for both constitutive and alternative splicing (Maniatis and Tasic 2002). Our results show that metazoans contain nine families of SR proteins, six of which have two or more members in mammals, whereas in unicellular eukaryotes there are only one or two SR protein genes (Table 2). Thus, the diversity of SR proteins seems to have emerged with multicellularity. Consistent with previous reports, we found no SR proteins in budding yeast, but two proteins in fission yeast (Tacke and Manley 1999; Kaufer and Potashkin 2000), and we confirmed the existence of 19 SR protein genes in *Arabidopsis* (Kalyna and Barta 2004; Reddy 2004).

The hnRNP proteins are a large group of molecules identified by their association with unspliced mRNA precursors (hnRNAs). The hnRNP proteins A, C, F, G, H, I (also termed PTB),

Table 1. Compilation of U11/U12 snRNP and DExD/H-box (DEAD) proteins identified in the analyzed genomes

Family	Human	Mouse	Rat	Chicken	*Fugu*	Zebrafish	Tetraodon	*Ciona*
U11/U12-20	UB20	UB20	UB20	UB20	UB20	UB20a UB20b	UB20	
U11/U12-25	UB25	UB25	UB25		UB25	UB25	UB25	UB25
U11/U12–31	UB31	UB31	UB31	UB31	UB31	UB31	UB31	UB31
U11/U12-35	UB35	UB35	UB35	UB35	UB35		UB35	UB35
U11/U12-48	UB48	UB48	UB48	UB48	UB48	UB48	UB48	UB48
U11/U12-65	UB65	UB65	UB65	UB65	UB65	UB65	UB65	UB65
ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS
DDX26	DDX26	DDX26	DDX26	DDX26	DDX26	DDX26 DX26b	DDX26	DDX26
	DD26B	DD26B		DD26B	DD26B	DD26B	DD26B	
DDX39	DDX39 BAT1	DDX39 BAT1	DDX39 BAT1	BAT1	DDX39	DDX39 BAT1		DDX39
DDX3XY	DDX3X DDX3Y	DDX3X DDX3Y	DDX3X	DDX3	DDX3a DDX3b	DDX3	DDX3	DDX3
DDX46	DDX46	DDX46	DDX46	DDX46	DDX46	DDX46	DDX46	DDX46
DDX48	DDX48	DDX48	DDX48	DDX48	DDX48	DDX48	DDX48	DDX48
DHX15	DHX15	DHX15	DHX15	DHX15	DHX15	DHX15	DHX15	DHX15
DHX16	DHX16	DHX16	DHX16	DHX16	DHX16	DHX16	DHX16	DHX16
DHX35	DHX35	DHX35	DHX35	DHX35	DHX35		DHX35	DHX35
DHX38	DHX38	DHX38	DHX38	DHX38	DHX38	DHX38	DHX38	DHX38
DHX8	DHX8	DHX8	DHX8	DHX8a DHX8b	DHX8a DHX8b	DHX8a DHX8b	DHX8	DHX8
DHX9	DHX9	DHX9	DHX9		DHX9	DHX9	DHX9a DHX9b	DHX9
KIAA0052	K052	K052	K052	K052	K052	K052	K052	K052
P68p72	DDX5 DDX17	DDX5 DDX17	DDX5 DDX17	DDX5 DDX17	DDX5 DD17a DD17b	DDX5 DDX17	DDX5 DDX17	p68
U5-100[a]	DDX23	DDX23			DDX23	DDX23	DDX23	DDX23
U5-200[a]	U5200	U5200	U5200	U5200	U5200	U5200	U5200	U5200

Family	Fly	Mosquito	*C. elegans*	*Arabidopsis*	*S. pombe*	*S. cerevisiae*	*Plasmodium*	*T. cruzi*
U11/U12-20	UB20							
U11/U12-25				UB25a UB25b				
U11/U12–31		UB31		UB31				
U11/U12-35		UB35		UB35				
U11/U12-48								
U11/U12-65	UB65	UB65						
ABS	ABS	ABS	ABS	ABSa ABSb			ABS	
DDX26	DDX26		DDX26					
DDX39	WM6	DDX39	DDX39	DDX39	UAP56	SUB2	DDX39	DDX39
DDX3XY	DDX3	DDX3	DDX3a DDX3b	DDX3a DDX3b DDX3c	DED1	DED1 DBP1		DDX3a DDX3b DDX3c DDX3d
DDX46	DDX46	DDX46	DD46a DD46b	DD46a DD46b	PRP11	PRP5	DDX46	DDX46
DDX48	DDX48	DDX48	DD48a DD48b	IF4Aa IF4Ab	EIF4A	FAL1	EIF	DDX48
DHX15	DHX15	DHX15	DHX15	DH15a DH15b	DHX15	PRP43	DHX15	DH15a DH15b DH15c
DHX16		DHX16	DHX16	DH16a DH16b	CDC28			
DHX35	DHX35	DHX35	DHX35	DHX35				
DHX38	DHX38	DHX38	DHX38	DHX38	PRP16	PRP16	DHX38	DHX38
DHX8	DHX8	DHX8	DHX8	DHX8	DHX8	PRP22	DHX8	DHX8
DHX9	MLE	DHX9	DHX9					
KIAA0052	K052	K052	K052	K052a K052b	K052a K052b	MTR4	K052	K052a K052b
P68p72	DDXP	DDXP	DDXP	RH20 RH30	DBP2	DBP2	DDXP	DDXPa DDXPb DDXPc
U5-100[a]	DDX23	DDX23	DDX23	DDX23	PRP28	PRP28	DDX23	
U5-200[a]	U5200	U5200	U5200 U5Hyp	U520a U520b	BRR2	BRR2	U5200	U520a U520b

Detailed identification of each gene is provided in Table S2. Small termination characters identify species/phylum specific duplications.
[a]Families annotated as snRNP specific.

Table 2. Compilation of SR proteins identified in the analyzed genomes

Family	Human	Mouse	Rat	Chicken	*Fugu*	Zebrafish	Tetraodon	*Ciona*	Fly	Mosquito	*C. elegans*	*Arabidopsis*	*S. pombe*	*Plasmodium*	*T. cruzi*
9G8-SRp20	9G8 SR20	9G8 SR20	9G8 SR20	9G8a 9G8b SR20	9G8 SR20a SR20b	9G8a 9G8b 9G8c SR20a SR20b	9G8 SR20	9G8a 9G8b SR20	9G8 RBP1 RBP1L RSF1	9G8 RBP1 RSF1	RSP6 RSPY	RS21 RS22 RS22A RS32[a] RS33[a]			
p54	p54 SR86	p54 SR86	p54 SR86	p54 SR86	p54a p54b SR86	p54 SR86	p54a p54b p54c	SR86a SR86b	p54	p54	p54				
RY1 SC35	RY1 SC35 SR46	RY1 SC35	RY1 SC35	RY1 SC35	RY1 SC35a SC35b	 SC35a SC35b	RY1	RY1 SC35	RY1 SC35	RY1 SC35	RY1 SC35	RY1 SC28[a] SC30[a] SC30A[a] SC33[a] SC35	SRP1[b]		
SRm300 SRp30c-ASF	SR300 ASF SR30C	SR300 ASF SR30C	SR300 ASF SR30C	 ASF	SR300 ASFa ASFb SR30C	SR300 ASFa ASFb SR30C	SR300 ASF	SR300 ASFa ASFb	SR300 SF2	SR300 SF2	SRRM2 SF2	SR45 RS31A[a] SR34 SR34A SR34B SRp30		SF[b]	
SRp40-55-75	SR40 SR55 SR75	SR40 SR55 SR75	SR40 SR55 SR75	SR40a SR40b SR55 SR75	SR40a SR40b SR55 SR75	SR40a SR40b SR55a SR55b SR75	SR40a SR40b	SR40a SR40b SR40c SR55	SR55	SR40	RSP1 RSP2 RSP5	RSp31[a] RSp40[a] RSP41[a]	SRP2[b]		SR1[b]
TopoI-B	T1B	T1B	T1B	T1B	T1Ba T1Bb	T1Ba T1Bb	T1Ba T1Bb		T1B	T1B					
Tra2	Tra2A Tra2B	Tra2A Tra2B	Tra2A Tra2B	Tra2A Tra2B	Tra2B	Tra2A Tra2B	Tra2B	Tra2	Tra2	Tra2a Tra2b	Tra2				

Detailed identification of each gene is provided in Table S2. Small termination characters identify species/phylum specific duplications. None of the analyzed SR protein genes was found for *Saccharomyces cerevisiae.*

[a]*Arabidopsis*-specific SR proteins, technically considered orthologs of the human proteins in the same family (reciprocal BLAST hit) but exhibiting a considerably lower degree of identity with the human factor than their *Arabidopsis* paralogs.

[b]SR proteins in unicellular eukaryotes can be considered common homologs of all the SR proteins in metazoans; here we include them in the same families of their technical human orthologs (reciprocal BLAST hit).

Table 3. Compilation of hnRNP proteins identified in the analyzed genomes

Family	Human	Mouse	Rat	Chicken	*Fugu*	Zebrafish	Tetraodon	*Ciona*	Fly	Mosquito	*C. elegans*	*Arabidopsis*	*S. pombe*	*T. cruzi*[a]
hnRNP-A	ROA0 ROA1 ROA2 ROA3	ROA0 ROA1 ROA2 ROA3	ROA1 ROA2	ROA0 ROA1 ROA3	ROA0 ROA1 ROA3	ROA0a ROA0b ROA0c ROA3	ROA0 ROA1 ROA3	ROA1a ROA1b ROA3	RO87F RO97D ROA1	RO87F	ROAa ROAb	ROAa ROAb ROAc		
hnRNP-C	RLY RLYL ROC ROCL	RLY RLYL ROC	RLY RLYL ROC	RLY RLYL	RLYLa RLYLb ROCa ROCb	RLYa RLYb RLYLa RLYLb ROCa ROCb	RLYLa RLYLb ROC	ROC						
hnRNP-D-U2	ROAB ROD0 RODL	ROAB ROD ROD0 RODL	ROAB ROD0 RODL	ROABa ROABb RODL	ROABa ROABb ROD0	ROAB ROD0 RODL	ROABa ROABb ROD0	ROAB	RO40 ROD	RO40	RODU2			RODa? RODb?
hnRNP-E	PCB1 PCB2 PCB3 PCB4	PCB1 PCB2 PCB3 PCB4	PCB1 PCB2 PCB3 PCB4	PCB3	PCB2a PCB2b PCB3a PCB3b PCB4a PCB4b	PCB2 PCB3	PCB2 PCB3a PCB3b	PCB	PCB	PCB	PCB			
hnRNP-F-H	GRSF1 ROF ROH1 ROH2 ROH3	GRSF1 ROF ROH1 ROH2 ROH3	GRSF1 ROF ROH1 ROH2 ROH3	GRSF1 ROH1 ROH3	GRSF1 ROFH ROH3	GRSF1 ROFHa ROFHb ROH3	GRSF1 ROFH ROH3	ROFHa ROFHb	ROFH	ROFH	ROFHa ROFHb	ROFHa ROFHb		ROFHa? ROFHb?
hnRNP-G	ROG ROGT	ROG ROGT	ROG ROGT	ROG	ROG	ROG	ROG							
hnRNP-I	PTB1 PTB2 ROD1	PTB1 PTB2 ROD1 smPTB	PTB1 PTB2 ROD1 smPTB	PTB1 PTB2 ROD1	PTB1a PTB1b PTB2 ROD1	PTB1a PTB1b PTB2a PTB2b ROD1	PTB1 PTB2 ROD1	PTBa PTBb	PTB	PTB	PTB	PTBa PTBb PTBc		PTBa? PTBb? PTBc?
hnRNP-K	ROK	ROK	ROK		ROKa ROKb	ROKa ROKb	ROKa ROKb	ROK	ROK	ROK	ROK			
hnRNP-L	ROL ROLH	ROL ROLH	ROL ROLH		ROL ROLHa ROLHb	ROLa ROLb	ROL ROLH	ROL	ROL	ROL	ROL			
hnRNP-M	Myel ROM	Myel ROM	Myel ROM	Myel ROM	Myel ROM	Myel ROM	Myel ROM	ROM	ROM	ROM	ROM			ROM?
hnRNP-R	ROQ ROR	ROQ ROR	ROQ ROR	ROQ ROR	ROQa ROQb ROR	ROQ ROR	ROQa ROQb ROR	ROQR	ROQR	ROQR	ROQR	ROQRa ROQRb ROQRc		
hnRNP-U	E1BA ROU ROUHY	E1BA ROU ROUHY	E1BA ROU ROUHY		E1BA ROU ROUH	E1BAa E1BAb ROU0 ROUa ROUb	E1BA ROU ROUHY	ROUa ROUb	ROU	ROU	ROU			
Musashi	MUS1 MUS2	MUS1 MUS2	MUS1 MUS2	MUS1 MUS2	MUS1	MUS1 MUS2a MUS2b	MUS1		MUSa MUSb	MUS	MUS		MUS	

Detailed identification of each gene is provided in Table S2. Small termination characters identify species/phylum specific duplications. None of the analyzed hnRNP genes was found for *Saccharomyces cerevisiae* and *Plasmodium falciparum.*

[a]Proteins signed with '?' are technically orthologs (reciprocal BLAST hit) but the large evolutionary distance (and low sequence similarity) and the absence of experimental data does not allow us to classify them as functional homologs.

Table 4. Evolution of miscellaneous splicing regulatory proteins

Family	Human	Mouse	Rat	Chicken	*Fugu*	Zebrafish	Tetraodon	*Ciona*	Fly	Mosquito	*C. elegans*	*Arabidopsis*	*S. pombe*	*S. cerevisiae*	*Plasmodium*	*T. cruzi*
CLK	CLK1	CLK1	CLK1	CLK2	CLK2a	CLK2	CLK2a	CLK	CLK	CLK	CLK	CLK	CLK	CLK	CLK	CLKa
	CLK2	CLK2	CLK2	CLK3	CLK2b	CLK4	CLK2b					AFC1				CLKb
	CLK3	CLK3	CLK3	CLK4	CLK4		CLK4					AFC2				
	CLK4	CLK4	CLK4									AFC3				
CUG	CUG1	CUG1	CUG1	CUG1	CUG1	CUG1	CUG2a	CUG1	CUGa	CUGa	CUG	RBPa				
	CUG2	CUG2	CUG2	CUG2	CUG2a	CUG2	CUG2b	ETR3	CUGb	CUGb		RBPb				
	CUG3	CUG3	CUG3	CUG3	CUG2b	CUG3	CUG3a									
	CUG4	CUG4	CUG4	CUG4	CUG3a	CUG4	CUG3b									
	CUG5	CUG5	CUG5	CUG5	CUG3b	CUG5	CUG4									
	CUG6	CUG6	CUG6	CUG6	CUG4		CUG5									
					CUG5											
ELAV	ELAV1	ELAV1	ELAV1	ELAV1	ELV1a	ELV1a	ELAV1	ELAVa	ELAVa	ELAVa	ELAV					
	ELAV2	ELAV2	ELAV2	ELAV3	ELV1b	ELV1b	ELAV2	ELAVb	ELAVb	ELAVb						
	ELAV3	ELAV3	ELAV3	ELAV4	ELAV3	ELAV2	ELAV3		ELAVc	ELAVc						
	ELAV4	ELAV4	ELAV4		ELAV4	ELAV3	ELAV4									
					ELAV	ELAV4										
						ELAV										
FUSE	FUSE1	FUSE1	FUSE2	FUSE1	FUSE1	FUSE1	FUSE1	FUSE	PSI	PSI	FUSEa	KHa				
	FUSE2	FUSE2	FUSE3	FUSE2	FUSE2	FUS2a	FUSE2				FUSEb	KHb				
	FUSE3	FUSE3		FUSE3	FUSE3	FUS2b	FUS3a				FUSEc	KHc				
						FUSE3	FUS3b				FUSEd					
											FUSEe					
SRPK	MSSK1	MSSK1	MSSK1	SRPK1	MSSK1	MSSK1	SRK1a	SRPK	SRPK	SRPK	SRPK	MSSKa	SRPK	SRPK	SRPK	SRPK
	SRPK1	SRPK1	SRPK1	SRPK2	SRPK1	SRK1a	SRK1b					MSSKb				
	SRPK2	SRPK2	SRPK2	SRPK	SRPK2	SRK1b	SRK1c					MSSKc				
					SRPKa	SRPK2	SRPK2					SRPKa				
					SRPKb	SRPKa	SRPKa					SRPKb				
						SRPKb	SRPKb									
						SRPKc	SRPKc									

Detailed identification of each gene is provided in Table S2. Small termination characters identify species/phylum specific duplications.

and M have been implicated in the regulation of splicing (Black 2003). We find that a single *S. pombe* protein shows significant sequence homology to hnRNPs, whereas 13 gene families are found in metazoans (Table 3). For each invertebrate hnRNP in *Ciona*, insects, or worms, there are, on average, three co-orthologs in the vertebrates human, mouse, and *Fugu*. *Ciona* has 16 hnRNP genes, whereas human has 37. Thus, a striking expansion of hnRNP protein gene families occurred in vertebrates.

Interestingly, gene families encoding additional splicing regulators have also expanded during the evolution of primitive metazoans into vertebrates (Table 4). These include the CLK (CDC-like) and SRPK (SR-protein-specific) kinases that phosphorylate SR proteins, modulating their function in splicing; the CUGBP (CUG-binding) and ETR-like proteins (CELF) implicated in tissue-specific and developmentally regulated alternative splicing and the alternative splicing regulators FUSE (far upstream element binding), and Elav (embryonic lethal abnormal visual) proteins (for a recent review see Black 2003).

Since genome duplication is known to have occurred at the vertebrate stem (Mazet and Shimeld 2002; McLysaght et al. 2002), we performed a phylogenetic analysis, using rate-linearized trees (see Methods) to determine whether the splicing factor family expansions are coincident with that duplication. Despite some topological inconsistencies between the different methods of phylogenetic inference, the evolutionary trees we generated are most consistent with the model that hnRNP genes underwent one or two rounds of duplication just after the divergence of vertebrates (Fig. 2A,B) and urochordates.

Furthermore, analysis of the teleost radiation, and of *Arabidopsis* revealed several localized gene duplications in *Fugu*, the zebrafish *Danio rerio*, *Tetraodon* (all teleosts), and *Arabidopsis*. These results are consistent with the currently accepted models proposing additional rounds of whole-genome duplication in ray-finned and lobe-finned fish, before teleost radiation (Amores et al. 1998; Aparicio et al. 2002; Christoffels et al. 2004), and the propensity of angiosperms to become polyploid (Simillion et al. 2002; Bowers et al. 2003). Thus, teleost fish and plants tend to have more copies of splicing genes than do mammals (Tables 1–4). However, there is no evidence for additional selective expansion of any particular family of splicing proteins in these organisms, beyond that which had occurred in the stem organism.

The domain evolution of splicing factors

Our data show conservation of the protein domain structure of splicing factors across species, and we found no evidence for domain shuffling. We observed no trend for gain or loss of domains in families of splicing factors, as has occurred in other nuclear protein families (for example, in the Polycomb and Trithorax protein families) (Ringrose and Paro 2004). We checked, for example, whether the expansion of SR protein families coincided with the appending of RS domains onto general RNA-

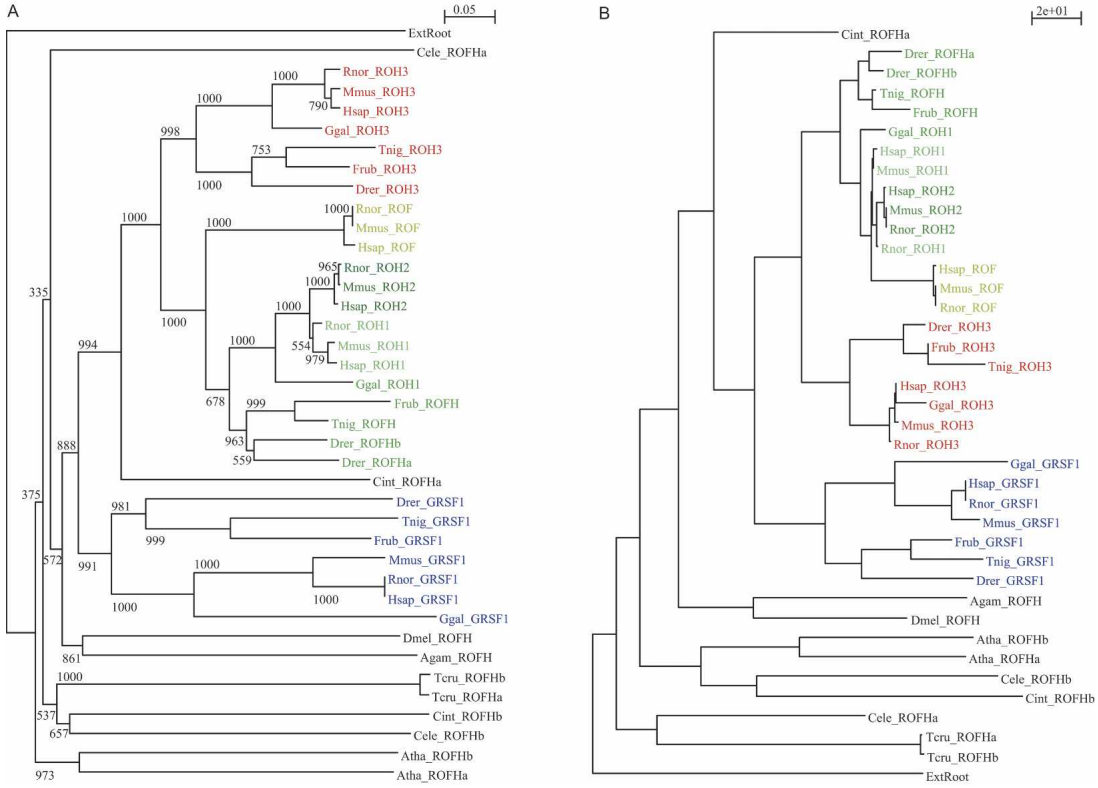

Figure 2. Evolutionary relationship among the protein members of hnRNP F/H family in several eukaryotes, i.e., human (Hsap), mouse (Mmus), rat (Rnor), chicken (Ggal), *Fugu* (Frub), zebrafish (Drer), *Tetraodon* (Tnig), *Ciona intestinalis* (Cint), fruit fly (Dmel), mosquito (Agam), *C. elegans* (Cele), *Arabidopsis* (Atha), and *Trypanosoma* (Tcru). Vertebrate factors are highlighted in blue, red, and shades of green. (*A*) Rooted Neighbor-Joining phylogenetic tree generated using ClustalW (1000 bootstraps), based on amino-acid alignment generated by T-Coffee. Bootstrap values are shown. Branch lengths are scaled in arbitrary units. (*B*) Rooted Gamma-corrected Maximum-Likelihood phylogenetic tree generated using GAMMA and the Phylip program Proml, based on amino-acid alignment generated by T-Coffee. Branch lengths are scaled in arbitrary units.

binding splicing factors. In species without SR proteins, we found no relevant homology with SR protein RNA recognition motifs (RRMs). Each factor seems to have evolved as a whole and its domains have evolved together (Fig. 3). Similarly, for the hnRNP families that are expanded in vertebrates, the motif structures are generally conserved (Fig. 4). One exception is hnRNP H3, which in mammals and chicken appears to have lost the first of the three RRM's that are common to its paralogs.

Retrotransposition and identification of putative novel splicing factors and pseudogenes in mammals

The absence of introns from mammalian genes is often indicative of retrotransposition, where a spliced mRNA is reverse transcribed into DNA and integrates back into the genome. Retrotransposition appears to have contributed as a general mechanism of gene duplication amongst mammals. We found that, with the exception of *U2AF*^{26} (a mammalian splicing factor [Shepard et al. 2002] that diverged from *U2AF*^{35} before vertebrates' radiation and is likely to have been lost by defunctionalization in teleosts) and *Sm N*, all of the mammalian specific factors *SRp46*, *U2AF1-RS1* and *hnRNPs C-like*, *E1*, *smPTB* and *G-T* are intronless, whereas their closer paralogs are multiexonic. *SRp46*, *U2AF1-RS1*, *hnRNP E1*, and *hnRNP G-T* have previously been reported to be retrotransposons (Soret et al. 1998; Makeyev et al. 1999; Elliott et al. 2000; Wang et al. 2004), which is consistent with our data. We therefore propose that retrotransposition contributed to generate the diversity of the splicing machinery observed in mammals.

We found evidence for an additional seven mouse putative intronless genes that appear to have no frame disruption in their coding sequences and for which we find evidence for transcription (Supplemental Table S4) and/or have an outstandingly high ratio of synonymous/nonsynonymous substitutions when compared with the closest active paralogue. Six of these putative intronless genes are annotated in Ensembl, but one of the genes is located in an unannotated genomic region. Two putative intronless genes exhibit transcript sequences equal to their closest paralogs. Whether these are novel functional splicing genes in mouse or very recent pseudogenes remains an open question.

In addition, we identified 107 human and 90 mouse putative pseudogenes (Supplemental Tables S3 and S5), none being found in other phyla. Of these, 30 human and 53 mouse pseudogenes are annotated as putative functional genes in Ensembl (Table S3). The majority (~80%) of all the analyzed intronless genes/pseudogenes contain evidence for surrounding LINE1 or LTR (long terminal repeat) sequences (repeats associated with transposable elements [Kazazian Jr. 2004]) and are therefore likely to be retrotransposons. Some families of Sm proteins and the hnRNP-A family contain particularly large numbers of retrotransposons (Supplemental Tables S3 and S5).

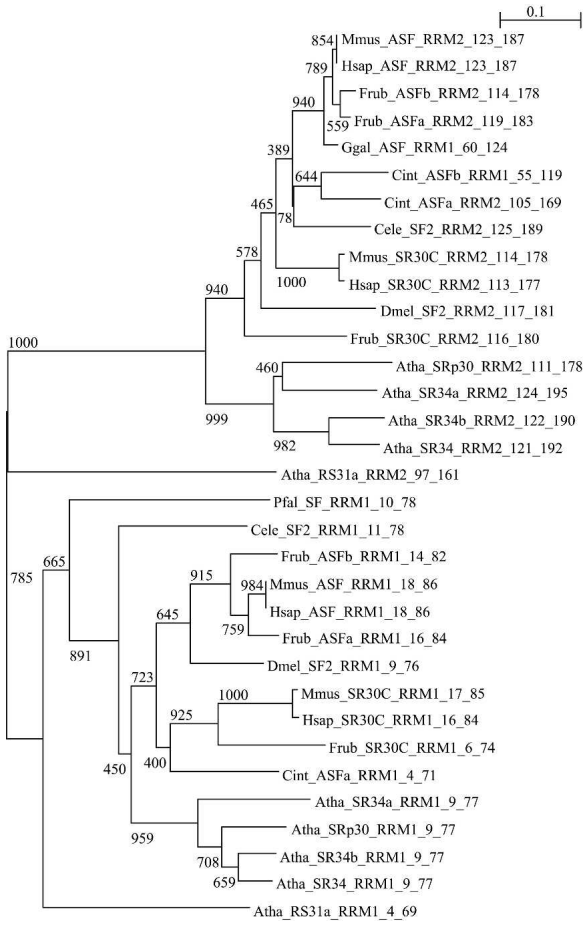

Figure 3. Evolutionary relationship among the RNA-recognition motifs (RRM) of members of the family SRp30c-ASF for several eukaryotes, i.e., human (Hsap), mouse (Mmus), chicken (Ggal), *Fugu* (Frub), *Ciona* (Cint), fruit fly (Dmel), *C. elegans* (Cele), *Arabidopsis* (Atha), and *Plasmodium* (Pfal) (for simplicity only one rodent, one teleost, and one insect are shown). Amino-acid positions of each domain within the protein are also indicated in the domain identification. The unrooted Neighbor-Joining phylogenetic tree was generated using ClustalW (1000 bootstraps) based on amino-acid alignment generated by T-Coffee. Bootstrap values are shown. Branch lengths are scaled in arbitrary units. RRM1 in Ggal_ASF and Cint_ASFb corresponds to RRM2 in the other proteins as their sequences are truncated in the N-terminal. Pfal_SF is found to have only one RRM. Atha_RS31A can be technically considered an ortholog of the Hsap_SR30C (reciprocal BLAST hit) but exhibits a considerably lower degree of identity (36%) with the human factor than its *Arabidopsis* paralogs (e.g., 53% for Atha_SRp30).

Discussion

Here we report a systematic comparison of the genes encoding the splicing machinery across diverse phyla. We designed a semi-automated computational pipeline to identify and annotate spliceosomal proteins that will also assist in the rapid reannotation of new splicing proteins as genomic sequences are updated. Our analysis shows differential gene family expansions across the eukaryotic lineage, with a disproportionate expansion of hnRNP proteins in vertebrates.

Although the origin of introns remains unknown, current data strongly indicate that introns and a spliceosome sufficient for their excision were present in the last common ancestor of eukaryotes (Johnson 2002; Collins and Penny 2005). Introns have been discovered in eukaryotes as primitive as the single-celled parasite *Giardia lamblia* (Nixon et al. 2002) and its close relative *Carpendiemona membranifera* (Simpson et al. 2002), and a core spliceosomal protein gene (*Prp8*) is remarkably conserved between metazoans and the deep-branching protist *Trichomonas vaginalis* (Fast and Doolittle 1999). Our finding that genes encoding snRNP proteins are generally conserved in animals, *Arabidopsis*, yeasts, trypanosomes, and *Plasmodium* is consistent with previous reports (for review, see Will and Luhrmann 2001). Our

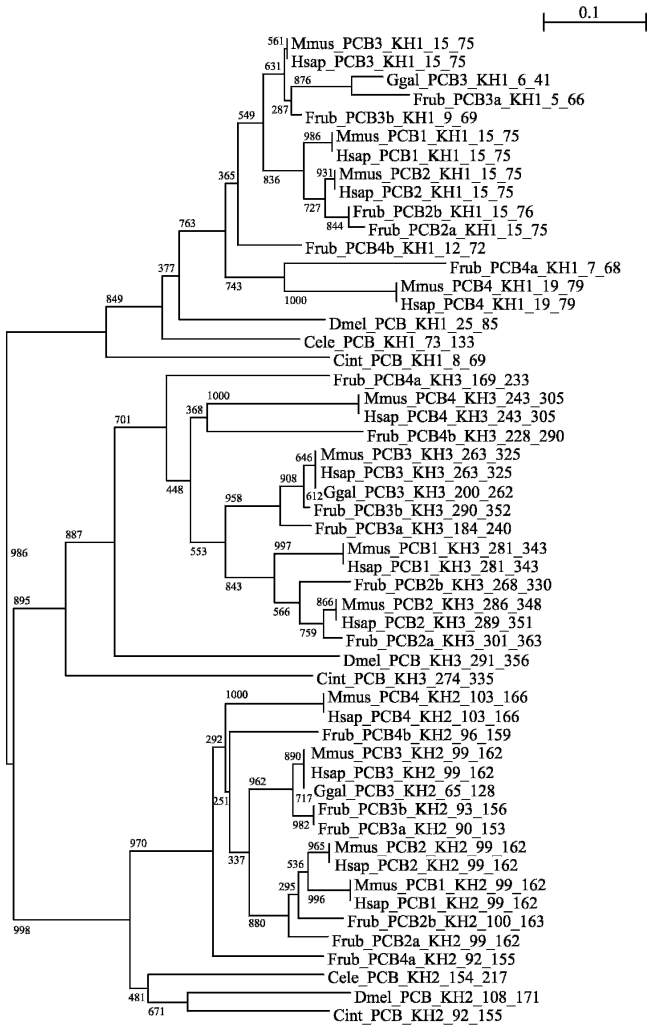

Figure 4. Evolutionary relationship among the RNA-binding K-Homology (KH) domains of members of the family hnRNP-E/PCB for several metazoans, i.e., human (Hsap), mouse (Mmus), chicken (Ggal), *Fugu* (Frub), *Ciona* (Cint), fruit fly (Dmel), and *C. elegans* (Cele) (for simplicity only one rodent, one teleost, and one insect are shown). Amino-acid positions of each domain within the protein are also indicated in the domain identification. The unrooted Neighbor-Joining phylogenetic tree was generated using ClustalW (1000 bootstraps) based on amino-acid alignment generated by T-Coffee. Bootstrap values are shown. Branch lengths are scaled in arbitrary units.

observation that *Plasmodium*, trypanosomes, yeasts, and *C. elegans* lack U11/U12 protein homologs is also in agreement with the hypothesis that the minor (U12-dependent) spliceosome was absent from the "first eukaryote" (Collins and Penny 2005).

In contrast to the conservation of snRNP protein genes, our analysis reveals that metazoans have many more genes implicated in the regulation of splicing than unicellular eukaryotes. Most probably, splicing regulatory proteins evolved as a consequence of whole-genome duplications that occurred at the vertebrate stem (Mazet and Shimeld 2002; McLysaght et al. 2002). According to the "classical" model for selective retention of gene family duplicates (Ohno 1970; Force et al. 1999; Nei and Rooney 2004), one of the duplicate genes retained the original function, while the other accumulated mutations that eventually conferred an advantageous new function (neofunctionalization).

We provide surprising evidence that retrotransposition introduced an additional level of diversity to the mammalian splicing machinery. Despite the fact that the majority of retrotransposons are nonfunctional (Goncalves et al. 2000), and that intronless genes may be transcribed less efficiently than their intron-containing homologs (Le Hir et al. 2003), we identified several retrotransposed genes, specific to mammals, encoding multifunctional RNA-binding proteins. These include *SRp46* (Soret et al. 1998), *hnRNP E1* (Leffers et al. 1995; Ostareck-Lederer et al. 1998; Krecic and Swanson 1999; Reimann et al. 2002; Bandiera et al. 2003; Persson et al. 2003; Antony et al. 2004; de Hoog et al. 2004; Morris et al. 2004), *hnRNP G-T* (Elliott et al. 2000; Nasim et al. 2003), *smPTB* (Gooding et al. 2003), and *U2AF1-RS1* (Wang et al. 2004). We also identified seven mouse putative novel active retrotransposed genes, paralogs of *NHP2-like, U1C, LSm6, LSm7, SmD2, SmG,* and *U2AF*^{35}.

Although splicing of introns from pre-mRNAs occurs in practically all eukaryotes, alternative splicing is important and widespread only in multicellular organisms. The yeast *S. cerevisiae* has introns in only ~3% of its genes and only six genes with more than one intron (Barrass and Beggs 2003). Although in the fission yeast *S. pombe*, 43% of the genes are spliced, with many of them containing multiple introns (Wood et al. 2002), no regulated alternative splicing has been detected in this organism or in any other unicellular eukaryote (Barrass and Beggs 2003; Ast 2004).

There are two current models to explain the evolution of alternative splicing, which are not mutually exclusive (Ast 2004). One is based on the accumulation of mutations that make splice sites suboptimal (or "weaker"), providing an opportunity for the splicing machinery to skip that site. In the second model, the evolution of splicing regulatory factors that either enhance or inhibit the binding of the splicing machinery to constitutive splice sites, it argues, releases the selective pressure from that sequence, resulting in mutations that weaken the splice sites. Our results clearly support this second model, which so far has not received much experimental attention. The choice of splice site is thought to be regulated by altering the binding of the spliceosome to the pre-mRNA. This is achieved by RNA-binding proteins that associate with nonsplice site sequences, located either in exons or introns. The best-characterized families of splicing regulators are SR proteins and hnRNP proteins (for review, see Black 2003). In vitro selection experiments have identified optimal binding sequences for different SR and hnRNP proteins, but the binding sites for a given family member can be fairly degenerate. Moreover, regulatory proteins can act as either splicing activators or repressors, depending on where in the pre-mRNA they bind. We propose, therefore, that the evolution of novel members of splicing regulatory protein families permitted the diversification of their canonical binding sites in pre-mRNAs, giving the cell the potential to produce new transcripts by altering splice choices. This hypothesis may be testable by correlating functional specificity of individual factors for their splice isoforms with the cognate recognition sequences in different species.

Methods

The key steps in our pipeline are illustrated in Figure 1.

All of the human splicing factors (Supplemental Table S6) and homologs annotated for other species were listed and their protein sequences were retrieved. Grouping into families was per-

formed based on full-length homology, functional domains composition, and the Ensembl Protein Family classification (Hubbard et al. 2002) (v30, http://www.ensembl.org). For each family, spurious and truncated proteins were identified and removed manually, and all of the remaining members were aligned with T-Coffee (Notredame et al. 2000) (default parameters). The alignment was used to build a profile HMM (Hidden Markov Model), using HMMER (Eddy 1998) (hmmbuild, hmmcalibrate), with which the proteomes of *Fugu*, *Ciona*, and 16 species of Archaea were searched (hmmsearch, e-value = 10^{-2}). In parallel, all of the human splicing factors were BLASTed (Altschul et al. 1990) (tblastn, BLOSUM62 matrix, SEG filter on, e-value = 10^{-3}) against the genomes of *Fugu*, *Ciona*, Archaea, *Plasmodium*, *Trypanosomas*, and proteomes of the previous species plus *A. thaliana*, *S. pombe*, and *S. cerevisiae*. Gene prediction was carried out in hit unannotated genomic regions, using Wise2 (http://www.ebi.ac.uk/Wise2). A reciprocal BLAST between the protein hits and the human genome and proteome (blastp, BLOSUM62 matrix, SEG filter on, e-value = 10^{-3}) was performed. Gene predictions were again made for hit unannotated genomic regions in Human.

The obtained members of each "complete" family were aligned and a phylogenetic tree was built with ClustalW (Thompson et al. 1994). The families of factors with relevant annotated function benefited from further curation, i.e., removal of false homologs and redundancies, correction of truncated and missannotated proteins, assessment of the likelihood of splice sites. This curation was assisted by BLAST, Wise 2, and EST searches, carried out in the Gene2EST BLAST Server (Gemund et al. 2001) (EMBL, http://woody.embl-heidelberg.de/gene2est) and the NCBI BLAST website (http://www.ncbi.nlm.nih.gov/BLAST, blastn, low complexity filter on), relying on the GenBank/dbEST database (v147.0) (Boguski et al. 1993; Benson et al. 2004).

The same approach was used to identify putative retrotransposons and discriminate pseudo-genes (based on the appearance of frame disruptions like cryptic stop codons and frameshifts introduced by missing or extra nucleotides in the conserved coding region). This procedure was complemented with the estimation of the ratio ds/dn of synonymous/nonsynonymous substitutions (using SNAP—http://www.es.embnet.org/Doc/SNAP/) and the identification of LINEs and LTR elements by searching the involving genomic sequences (1.2 kb upstream and downstream of the putative transcribed sequence) with RepeatMasker (http://www.repeatmasker.org, default parameters).

New alignments were built for the resulting curated families and, for each family, the functional domain composition of its members was compared. The domain organization of proteins relied on the Pfam database (Bateman et al. 2002) (http://www.sanger.ac.uk/Software/Pfam, version 16.0), the HMMER program hmmpfam and the SMART tool (Letunic et al. 2004) (http://smart.embl-heidelberg.de, version 4.0).

We then performed the phylogenetic analysis of all the families by generating bootstrapped Neighbor-Joining (NJ) trees with ClustalW (1000 bootstraps). Alternatively, we bootstrapped our alignments using the Phylip (Felsenstein 1989) program Seqboot (100 bootstraps). Then rooted and bootstrapped NJ and Parsimony trees were built using the Phylip program's Neighbor (preceded by Protdist) and Protpars, respectively. In both cases we generated the consensus trees with Phylip program Consense. We also created rooted Maximum-Likelihood (ML) trees using the Phylip program Proml. For details on tree rooting, see Supplemental Table S7. We did the molecular clock analysis following a procedure similar to that adopted by Christoffels et al. (2004) (Supplemental Table S8).

Supplemental Table S1 summarizes the sources for the whole genomes and predicted proteomes used in our search. The automated searches relied on BioPerl (Stajich et al. 2002) (v1.30, http://www.bioperl.org) and Ensembl Perl modules on a Linux platform. All of the phylogenetic trees and alignments can be found in the Supplemental materials.

Acknowledgments

We thank Christopher W.J. Smith and Juan Valcárcel for critical discussions, and Carol Featherston for editing of the manuscript. We are also grateful to Ângela Relógio and Christine Gemünd (EMBL) for the splicing factors' database and Claudia Ben-Dov, Carlos Caldas, Alan Christoffels, João Ferreira, Filipa Gallo, Margarida Gama-Carvalho, Inês Mollet, Teresa R. Pacheco, Søren Steffensen, Natalie Thorne, and Damian Yap for valuable contributions to this work. This work was supported by the Wellcome Trust, UK and Fundação para a Ciência e a Tecnologia, Portugal (Fellowship SFRH/BD/2914/2000).

References

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

Amores, A., Force, A., Yan, Y.L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.L., et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* **282:** 1711–1714.

Antony, A., Tang, Y.S., Khan, R.A., Biju, M.P., Xiao, X., Li, Q.J., Sun, X.L., Jayaram, H.N., and Stabler, S.P. 2004. Translational upregulation of folate receptors is mediated by homocysteine via RNA-heterogeneous nuclear ribonucleoprotein E1 interactions. *J. Clin. Invest.* **113:** 285–301.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297:** 1301–1310.

Ast, G. 2004. How did alternative splicing evolve? *Nat. Rev. Genet.* **5:** 773–782.

Bandiera, A., Tell, G., Marsich, E., Scaloni, A., Pocsfalvi, G., Akintunde Akindahunsi, A., Cesaratto, L., and Manzini, G. 2003. Cytosine-block telomeric type DNA-binding activity of hnRNP proteins from human cell lines. *Arch. Biochem. Biophys.* **409:** 305–314.

Barrass, J.D. and Beggs, J.D. 2003. Splicing goes global. *Trends Genet.* **19:** 295–298.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30:** 276–280.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2004. GenBank: Update. *Nucleic Acids Res.* **32:** D23–D26.

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

Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. 1993. dbEST—database for "expressed sequence tags". *Nat. Genet.* **4:** 332–333.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422:** 433–438.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. 2002. Alternative splicing and genome complexity. *Nat. Genet.* **30:** 29–30.

Burge, C., Tuschl, T., and Sharp, P. 1999. Splicing of precursors to mRNAs by the spliceosomes. In *The RNA world*, 2nd ed., (eds. R. Gesteland, T. Cech, and J. Atkins), pp. 525–560. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Caceres, J.F. and Kornblihtt, A.R. 2002. Alternative splicing: Multiple control mechanisms and involvement in human disease. *Trends Genet.* **18:** 186–193.

Christoffels, A., Koh, E.G., Chia, J.M., Brenner, S., Aparicio, S., and Venkatesh, B. 2004. *Fugu* genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol. Biol. Evol.* **21:** 1146–1151.

Collins, L. and Penny, D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.* **22:** 1053–1066.

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D.M., et al. 2002. The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. *Science* **298:** 2157–2167.
de Hoog, C.L., Foster, L.J., and Mann, M. 2004. RNA and RNA binding proteins participate in early stages of cell spreading through spreading initiation centers. *Cell* **117:** 649–662.
Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755–763.
Elliott, D.J., Venables, J.P., Newton, C.S., Lawson, D., Boyle, S., Eperon, I.C., and Cooke, H.J. 2000. An evolutionarily conserved germ cell-specific hnRNP is encoded by a retrotransposed gene. *Hum. Mol. Genet.* **9:** 2117–2124.
Fast, N.M. and Doolittle, W.F. 1999. *Trichomonas vaginalis* possesses a gene encoding the essential spliceosomal component, PRP8. *Mol. Biochem. Parasitol.* **99:** 275–278.
Felsenstein, J. 1989. PHYLIP—Phylogeny inference package (Version 3.2). *Cladistics* **5:** 164–166.
Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.
Gemund, C., Ramu, C., Altenberg-Greulich, B., and Gibson, T.J. 2001. Gene2EST: A BLAST2 server for searching expressed sequence tag (EST) databases with eukaryotic gene-sized queries. *Nucleic Acids Res.* **29:** 1272–1277.
Goncalves, I., Duret, L., and Mouchiroud, D. 2000. Nature and structure of human genes that generate retropseudogenes. *Genome Res.* **10:** 672–678.
Gooding, C., Kemp, P., and Smith, C.W. 2003. A novel polypyrimidine tract-binding protein paralog expressed in smooth muscle cells. *J. Biol. Chem.* **278:** 15201–15207.
Graveley, B.R. 2002. Sex, agility, and the regulation of alternative splicing. *Cell* **109:** 409–412.
Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30:** 38–41.
Johnson, P.J. 2002. Spliceosomal introns in a deep-branching eukaryote: The splice of life. *Proc. Natl. Acad. Sci.* **99:** 3359–3361.
Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302:** 2141–2144.
Jurica, M.S. and Moore, M.J. 2003. Pre-mRNA splicing: Awash in a sea of proteins. *Mol. Cell* **12:** 5–14.
Kalyna, M. and Barta, A. 2004. A plethora of plant serine/arginine-rich proteins: Redundancy or evolution of novel gene functions? *Biochem. Soc. Trans.* **32:** 561–564.
Kaufer, N.F. and Potashkin, J. 2000. Analysis of the splicing machinery in fission yeast: A comparison with budding yeast and mammals. *Nucleic Acids Res.* **28:** 3003–3010.
Kazazian Jr., H.H. 2004. Mobile elements: Drivers of genome evolution. *Science* **303:** 1626–1632.
Krecic, A.M. and Swanson, M.S. 1999. hnRNP complexes: Composition, structure, and function. *Curr. Opin. Cell. Biol.* **11:** 363–371.
Leffers, H., Dejgaard, K., and Celis, J.E. 1995. Characterisation of two major cellular poly(rC)-binding human proteins, each containing three K-homologous (KH) domains. *Eur. J. Biochem.* **230:** 447–453.
Le Hir, H., Nott, A., and Moore, M.J. 2003. How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci.* **28:** 215–220.
Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P., and Bork, P. 2004. SMART 4.0: Towards genomic data integration. *Nucleic Acids Res.* **32:** D142–D144.
Luhrmann, R., Kastner, B., and Bach, M. 1990. Structure of spliceosomal snRNPs and their role in pre-mRNA splicing. *Biochim. Biophys. Acta* **1087:** 265–292.
Makeyev, A.V., Chkheidze, A.N., and Liebhaber, S.A. 1999. A set of highly conserved RNA-binding proteins, αCP-1 and αCP-2, implicated in mRNA stabilization, are coexpressed from an intronless gene and its intron-containing paralog. *J. Biol. Chem.* **274:** 24849–24857.
Maniatis, T. and Tasic, B. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418:** 236–243.
Mazet, F. and Shimeld, S.M. 2002. Gene duplication and divergence in the early evolution of vertebrates. *Curr. Opin. Genet. Dev.* **12:** 393–396.
McLysaght, A., Hokamp, K., and Wolfe, K.H. 2002. Extensive genomic duplication during early chordate evolution. *Nat. Genet.* **31:** 200–204.
Morris, B.J., Adams, D.J., Beveridge, D.J., van der Weyden, L., Mangs, H., and Leedman, P.J. 2004. cAMP controls human renin mRNA stability via specific RNA-binding proteins. *Acta Physiol. Scand.* **181:** 369–373.
Nasim, M.T., Chernova, T.K., Chowdhury, H.M., Yue, B.G., and Eperon, I.C. 2003. HnRNP G and Tra2β: Opposite effects on splicing matched by antagonism in RNA binding. *Hum. Mol. Genet.* **12:** 1337–1348.
Nei, M. and Rooney, A.P. 2004. Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39:** 121–152.
Nilsen, T.W. 2003. The spliceosome: The most complex macromolecular machine in the cell? *Bioessays* **25:** 1147–1149.
Nixon, J.E., Wang, A., Morrison, H.G., McArthur, A.G., Sogin, M.L., Loftus, B.J., and Samuelson, J. 2002. A spliceosomal intron in *Giardia lamblia*. *Proc. Natl. Acad. Sci.* **99:** 3701–3705.
Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302:** 205–217.
Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Heidelberg, Germany.
Ostareck-Lederer, A., Ostareck, D.H., and Hentze, M.W. 1998. Cytoplasmic regulatory functions of the KH-domain proteins hnRNPs K and E1/E2. *Trends Biochem. Sci.* **23:** 409–411.
Patel, A.A. and Steitz, J.A. 2003. Splicing double: Insights from the second spliceosome. *Nat. Rev. Mol. Cell. Biol.* **4:** 960–970.
Persson, P.B., Skalweit, A., Mrowka, R., and Thiele, B.J. 2003. Control of renin synthesis. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **285:** R491–R497.
Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. 2002. Large-scale proteomic analysis of the human spliceosome. *Genome Res.* **12:** 1231–1245.
Reddy, A.S. 2004. Plant serine/arginine-rich proteins and their role in pre-mRNA splicing. *Trends Plant Sci.* **9:** 541–547.
Reimann, I., Huth, A., Thiele, H., and Thiele, B.J. 2002. Suppression of 15-lipoxygenase synthesis by hnRNP E1 is dependent on repetitive nature of LOX mRNA 3′-UTR control element DICE. *J. Mol. Biol.* **315:** 965–974.
Ringrose, L. and Paro, R. 2004. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu. Rev. Genet.* **38:** 413–443.
Shepard, J., Reick, M., Olson, S., and Graveley, B.R. 2002. Characterization of U2AF(6), a splicing factor related to U2AF(35). *Mol. Cell. Biol.* **22:** 221–230.
Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M., and Van de Peer, Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.* **99:** 13627–13632.
Simpson, A.G., MacQuarrie, E.K., and Roger, A.J. 2002. Eukaryotic evolution: Early origin of canonical introns. *Nature* **419:** 270.
Smith, C.W. and Valcarcel, J. 2000. Alternative pre-mRNA splicing: The logic of combinatorial control. *Trends Biochem. Sci.* **25:** 381–388.
Soret, J., Gattoni, R., Guyon, C., Sureau, A., Popielarz, M., Le Rouzic, E., Dumon, S., Apiou, F., Dutrillaux, B., Voss, H., et al. 1998. Characterization of SRp46, a novel human SR splicing factor encoded by a PR264/SC35 retropseudogene. *Mol. Cell. Biol.* **18:** 4924–4934.
Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12:** 1611–1618.
Staley, J.P. and Guthrie, C. 1998. Mechanical devices of the spliceosome: Motors, clocks, springs, and things. *Cell* **92:** 315–326.
Tacke, R. and Manley, J.L. 1999. Determinants of SR protein specificity. *Curr. Opin. Cell. Biol.* **11:** 358–362.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Wang, Y., Joh, K., Masuko, S., Yatsuki, H., Soejima, H., Nabetani, A., Beechey, C.V., Okinami, S., and Mukai, T. 2004. The mouse Murr1 gene is imprinted in the adult brain, presumably due to transcriptional interference by the antisense-oriented U2af1-rs1 gene. *Mol. Cell. Biol.* **24:** 270–279.
Will, C.L. and Luhrmann, R. 2001. Spliceosomal UsnRNP biogenesis, structure and function. *Curr. Opin. Cell. Biol.* **13:** 290–301.
Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., et al. 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415:** 871–880.
Zhou, Z., Licklider, L.J., Gygi, S.P., and Reed, R. 2002. Comprehensive proteomic analysis of the human spliceosome. *Nature* **419:** 182–185.
Zhu, W. and Brendel, V. 2003. Identification, characterization and molecular phylogeny of U12-dependent introns in the *Arabidopsis thaliana* genome. *Nucleic Acids Res.* **31:** 4561–4572.

Web site references

http://www.ensembl.org; Ensembl.
http://www.ebi.ac.uk/Wise2; Wise2—Intelligent algorithms for DNA searches (EBI).
http://woody.embl-heidelberg.de/gene2est; Gene2EST BLAST Server.
http://www.ncbi.nlm.nih.gov/BLAST; NCBI BLAST.
http://www.es.embnet.org/Doc/SNAP; SNAP.pl (Synonymous Nonsynonymous Analysis Program).
http://www.repeatmasker.org; RepeatMasker.
http://www.sanger.ac.uk/Software/Pfam; Pfam—Protein families database of alignments and HMMs.
http://smart.embl-heidelberg.de; SMART—Simple Modular Architecture Research Tool.
http://www.bioperl.org; BioPerl.

Received March 15, 2005; accepted in revised form August 9, 2005.