J Mol Evol (2012) 75:214–228
DOI 10.1007/s00239-012-9533-0

Complex Evolutionary Relationships Among Four Classes of Modular RNA-Binding Splicing Regulators in Eukaryotes: The hnRNP, SR, ELAV-Like and CELF Proteins

Yue Hang Tang · Siew Ping Han · Karin S. Kassahn · Adam Skarshewski · Joseph A. Rothnagel · Ross Smith

Received: 27 February 2012 / Accepted: 9 November 2012 / Published online: 24 November 2012

Abstract Alternative RNA splicing in multicellular organisms is regulated by a large group of proteins of mainly unknown origin. To predict the functions of these proteins, classification of their domains at the sequence and structural level is necessary. We have focused on four groups of splicing regulators, the heterogeneous nuclear ribonucleoprotein (hnRNP), serine–arginine (SR), embryonic lethal, abnormal vision (ELAV)-like, and CUG-BP and ETR-like factor (CELF) proteins, that show increasing diversity among metazoa. Sequence and phylogenetic analyses were used to obtain a broader understanding of their evolutionary relationships. Surprisingly, when we characterised sequence similarities across full-length sequences and conserved domains of ten metazoan species, we found some hnRNPs were more closely related to SR, ELAV-like and CELF proteins than to other hnRNPs. Phylogenetic analyses and the distribution of the RRM domains suggest that these proteins diversified before the last common ancestor of the metazoans studied here through domain acquisition and duplication to create genes of mixed evolutionary origin. We propose that these proteins were derived independently rather than through the expansion of a single protein family. Our results highlight inconsistencies in the current classification system for these regulators, which does not adequately reflect their evolutionary relationships, and suggests that a domain-based classification scheme may have more utility.

Keywords RNA-binding protein · Splicing regulators · Protein evolution · Protein classification

Yue Hang Tang and Siew Ping Han contributed equally to this study.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-012-9533-0) contains supplementary material, which is available to authorized users.

Y. H. Tang · S. P. Han · A. Skarshewski · J. A. Rothnagel · R. Smith (✉)
School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, QLD, Australia
e-mail: ross.s@uq.edu.au

Y. H. Tang
e-mail: henry.tang@uq.edu.au

S. P. Han
e-mail: uqshan1@uq.edu.au

A. Skarshewski
e-mail: a.skarshewski@gmail.com

J. A. Rothnagel
e-mail: j.rothnagel@uq.edu.au

Present Address:
S. P. Han
Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, Australia

K. S. Kassahn
ARC Centre of Excellence in Bioinformatics, Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia
e-mail: k.kassahn@uq.edu.au

Present Address:
K. S. Kassahn
Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia

Abbreviations

hnRNP	Heterogenous nuclear ribonucleoprotein
SR	Serine–arginine
ELAV	Embryonic lethal, abnormal vision
CELF	CUG-BP and ETR-like factor
RBP	RNA-binding protein
RRM	RNA recognition motif
aRRM	Atypical RRM

Introduction

Proteome expansion and diversity are the result of various genetic events including gene duplication, gene recombination and sequence and structural divergence (Chothia and Gough 2009; Vogel and Chothia 2006). The size of the protein repertoire has increased in parallel with the biological complexity of organisms, especially during metazoan evolution (Chothia and Gough 2009; Vogel and Chothia 2006; Vogel et al. 2003; Kirschner and Gerhart 1998). Protein families that show significant expansion in multicellular organisms are often involved in the regulation of gene expression and in signal transduction (Aravind and Subramanian 1999; Rubin et al. 2000), where such protein expansions correlate with an increase in organismal complexity, the expansion is termed 'progressive' (Chothia and Gough 2009; Vogel and Chothia 2006; Vogel et al. 2003; Kirschner and Gerhart 1998).

Many proteins are modular in structure, formed by a combination of discrete units or 'domains' that fold independently (Chothia and Gough 2009; Lupas et al. 2001; Yang and Bourne 2009; Ponting and Russell 2002). Domains are essential building blocks for proteins; they are defined as structural and functional units. Multi-domain proteins combine two or more different domains, normally present in multiple repeats, to provide various, sophisticated functions. Domains can also be considered as major evolutionary units, which can be grouped into domain families or superfamilies based on their homology at the sequence, structure and/or function level (Chothia and Gough 2009; Gough 2005; Todd et al. 2001; Graumann and Marahiel 1996).

The RNA-binding heterogeneous nuclear ribonucleoproteins (hnRNPs), serine–arginine (SR) proteins, embryonic lethal, abnormal vision (ELAV)-like proteins and CUG-BP and ETR-like factors (CELF) proteins are examples of modular protein families with regulatory roles that have undergone progressive expansion. These RNA-binding proteins (RBPs) play a primary role in binding to nascent transcripts and regulate post-transcriptional events including alternative splicing, thereby enhancing proteomic diversity in multicellular organisms (Vogel and Chothia 2006; Black 2003; Keren et al. 2010; Chen and Zheng 2009; Hsu et al. 2011). As important regulators of pre-mRNA splicing, these RBPs bind to a variety of specific sequences in introns and exons within the nascent transcript and regulate alternative splicing through multiple protein-RNA and protein–protein interactions (Cartegni et al. 2002; Black 2003; Matlin et al. 2005; Stamm et al. 2005). Recent studies have further demonstrated the complex control of alternative splicing through the enhancing and silencing effects by different members of these groups of RBPs. Indeed, recent genome-wide studies have revealed a position-dependent RNA splicing map for the hnRNPs, in which the control of alternative splicing is mediated by specific cooperation of the different hnRNPs to either promote or inhibit alternative splicing (Blanchette et al. 2009; Huelga et al. 2012). Protein–protein interactions between hnRNPs and other RBPs, such as U2AF65, are also important for discriminating between RNAs and play a role in U2AF-mediated recruitment of the U2 small nuclear ribonucleoprotein complex to the nascent transcript (Tavanez et al. 2012). Antagonist effects on alternative splicing have also been long-studied within hnRNPs and between hnRNPs and SR proteins. One recent example is the regulation of exon 11 of the insulin receptor gene, where hnRNP F and SRSF1 compete with hnRNP A1 for the binding site to promote or inhibit exon 11 inclusion (Talukdar et al. 2011). Alternative splicing could also act as a regulatory step for RBPs since the inclusion/exclusion of alternative exons could significantly modify their functions; e.g. inclusion of the alternative exon 1b in hnRNP A3 decreases the affinity towards the A2 regulatory-element binding sequence as well as influencing translation initiation (Han et al. 2010a).

A common feature shared by the hnRNP, SR, ELAV-like and CELF proteins that facilitates their role in alternative splicing is the presence of one or more copies of an RNA-binding domain which mediates the recognition and binding to pre-mRNAs. One of the most common and best characterized RNA-binding domains is the RNA recognition motif (RRM) which has a characteristic β_1-α_1-β_2-β_3-α_2-β_4 secondary structure and contains two conserved motifs, RNP-1 and RNP-2, that are important for RNA interaction (Maris et al. 2005). These domains, together with protein–protein interaction domains, enhance the functional specificity of these splicing proteins (Fig. 1) (Maniatis and Tasic 2002; Singh and Valcarcel 2005; Lunde et al. 2007). In addition, some of these RBPs contain quasi-RRMs or RRM homologues, referred to here as atypical RRMs (aRRMs), which are structurally divergent forms that resemble the RRM in overall topology but have highly degenerated RNP motifs (Dominguez and Allain 2006; Shepard and Hertel 2009).

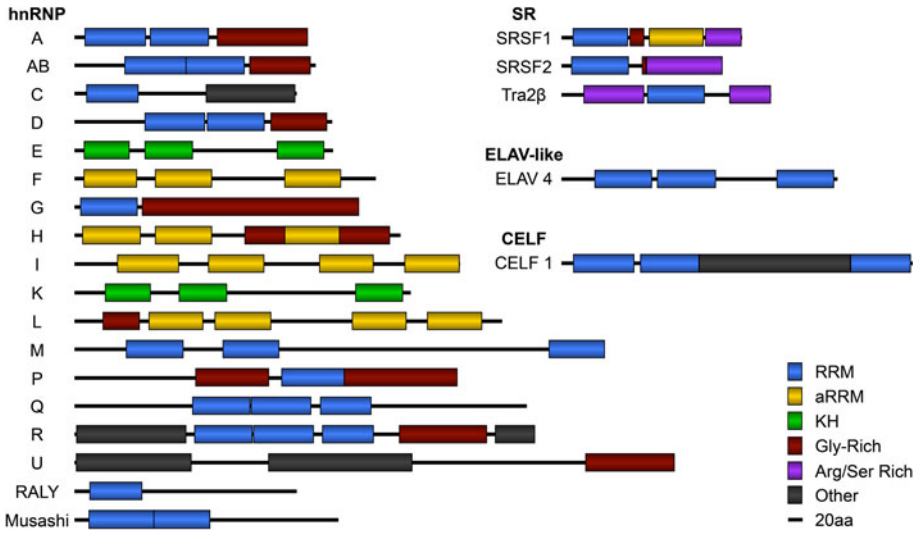


Fig. 1 RRMs and the KH (K homology) domains are two of the most abundant RNA-binding motifs that mediate protein–RNA interaction in hnRNPs and other RBPs. Other domains that lead to increasing modularity of these splicing proteins and functional diversity are also present. However, there are no domains that are common across all of the hnRNPs, whereas RRM domain is prevalent *across* all groups of splicing proteins. Domains illustrated are based on Uniprot definition as defined by PROSITE. (aRRM—atypical RRM)

The criteria that have been used to classify these RBPs into separate families are markedly different. The primary basis for the description of hnRNPs as a protein family has been their presence in a large protein complex that assembles on nascent mRNA (Dreyfuss et al. 1993) and their immunopurification by monoclonal antibodies that recognize hnRNP C, the founding member of this group of proteins (Piñol-Roma et al. 1988; Han et al. 2010b). In contrast, the SR proteins have been defined by the presence of arginine–serine–rich (RS) domains in combination with one or more RRMs (Long and Caceres 2009; Shepard and Hertel 2009) while the ELAV-like proteins, which are expressed in all metazoans, are defined as a family based on homology to the *Drosophila* ELAV protein (Good 1995; Pascale and Govoni 2012). The CELF proteins, which have been described as distantly related to ELAV-like proteins, have also been grouped based on the sequence similarity (Ladd et al. 2001; Dasgupta and Ladd 2012). Thus, the hnRNPs, SR proteins, and ELAV-like/CELF proteins have been primarily defined by their epitopes, domain composition, and gene homology, respectively.

Previous evolutionary studies have revealed a selective expansion of different groups of splicing regulatory factors, including hnRNP, SR, ELAV-like and CELF proteins, most notably during the evolution of vertebrates (Barbosa-Morais et al. 2006). For example, the number of hnRNP proteins is strikingly different between unicellular organisms e.g. *Saccharomyces pombe* (one protein) and multicellular organisms e.g. *Homo sapiens* (37 proteins), and ELAV-like proteins expanded from one protein in *Caenorhabditis. elegans* to four paralogues in vertebrates (Barbosa-Morais et al. 2006; Busch and Hertel 2012). This expansion in the different splicing regulators correlates with the increase in complexity during vertebrate evolution and may reflect the difference in alternative splicing regulation in different species (Barbosa-Morais et al. 2006; Busch and Hertel 2012). Moreover, advances in experimental methods such as mass spectrometry-based proteomic studies and bioinformatics have led to the identification of novel RBPs (Hsu et al. 2011). Hence, accurate exploitation of evolutionary relationships within protein families is essential, as comparisons based on homologies at the sequence and structural level are the key to ascribing functions to newly discovered proteins (Ponting and Russell 2002; Todd et al. 2001; Hsu et al. 2011). However, growing data indicate that the complicated sequence-structure–function relationships have led to the functional promiscuity of many protein members and have blurred the separation between different splicing protein families. In particular, structural redundancy at the domain level has conveyed similar functions between proteins with significantly divergent sequences. Therefore, it is important to understand the evolution of these proteins, especially at the domain level, in order to understand the complex connections between protein sequence, structure and function (Devos and Valencia 2000; Yang and Bourne 2009; Ponting and Russell 2002).

Thus, we have investigated phylogenetic relationships within and across groups of these proteins from a broad range of metazoans (*Trichoplax adhaerens, Nematostella vecetensis, C. elegans, Ciona intestinalis, Strongylocentrotus purpuratus, Daphnia pulex, Drosophila melanogaster*, *Xenopus tropicalis*, *Mus musculus* and *H. sapiens*). The topologies of hnRNP phylogenetic trees suggested that the diversification of alternative splicing factors may have occurred early in or before metazoan evolution and that the genes encoding these proteins were generated by domain duplication and domain acquisition events. We found a significant degree of overlap between the sequences of the different groups of splicing regulatory proteins, and on this basis we propose an alternative approach to the classification of the RBPs studied.

Methods

Sequence Data Sets

Protein sequences for *H. sapiens*, *M. musculus*, *X. tropicalis*, *D. melanogaster*, *D. pulex, S. purpuratus, C. intestinalis*, *C. elegans, N. vecetensis and T. adhaerens* were obtained from Genbank (Maglott et al. 2005), UniProtKB (Consortium 2012; Magrane and Consortium 2011) and Ensembl v51 (Hubbard et al. 2009) using keyword and sequence-based search strategies. Orthologues were identified using the Ensembl orthology predictions and UniProtKB BLAST searches. Orthologue sequences identified via UniProtKB BLAST search were further crosschecked by performing Genbank BLASTp searches, and sequences identified from both UniProtKB BLAST and Genebank BLASTp with a target ID <20 % were discarded. While more sequence data were available for metazoans such as *Branchiostoma floridae*, *Lottia gigantean* or *Capitella teleta*, it was difficult to confidently identify orthologues in these species. As most sequences from these species are either partial or hypothetical/predicted sequences, they were omitted from this study. Data sets were manually curated to remove misannotated or incomplete sequences. Domain boundaries were defined based on Uniprot domain annotations (Jain et al. 2009) unless otherwise stated. RNP motifs were defined based on (Birney et al. 1993), and identified by aligning RRMs using ClustalW2 (Larkin et al. 2007) with optimisation of gap penalties and manual adjustment of motif boundaries to improve alignments. The final dataset included 172 sequences, representing 25 hnRNP proteins, three SR proteins, one ELAV-like protein and two CELF proteins across up to ten metazoan species with representatives from basal metazoan (cnidarian and placozoan) to bilaterian invertebrates and bilaterian vertebrates. Accession numbers of sequences used in the analyses are provided in Online Resource 1.

Protein Similarity Searches

A useful approach for protein classification and characterisation is an all-versus-all BLASTp search to identify sequence similarities across the complete protein space. We thus performed BLASTp (Version 2.2.2) searches for hnRNP sequences from each species against the NCBI RefSeq databases (Pruitt et al. 2007) for that species. The following parameters were used to include lower hit sequences: 1) setting the number of targets to 1,000 and 2) the expectant statistical significance threshold at 1,000. These parameters allow identification of both closely and distantly related proteins. Similarity matrices were constructed using the negative logarithm of the resultant E-values (Enright et al. 2002).

Visualisation of Pairwise Sequence Relationships

To help visualise pairwise sequence relationships we used a modified version of uPEPperoni (www.upep.info). In brief, pairwise amino acid sequence alignments were performed using the Needleman-Wunsch algorithm, which aligns protein sequences based on their similarities as measured by the BLOSUM62 similarity matrix (Needleman and Wunsch 1970; Henikoff and Henikoff 1992). The degree of sequence identity was determined using a sliding window approach and calculating the percentage of amino acid matches (identical = +1, conservative substitution = +0.5) within a ten-amino acid window surrounding the target amino acid. The resulting values for contiguous amino acids were concatenated and converted into colour gradients, which were visualised as heatmaps. This approach allowed us to visually inspect regions of sequence conservation across a heterogeneous set of proteins.

Phylogenetic Analyses

To better understand the evolutionary ancestry of regulatory splicing proteins we performed phylogenetic analyses of all 190 RRM domains in the dataset. Phylogenetic trees were constructed from amino acid sequence alignments generated as described above using MrBayes v3.1 (Ronquist and Huelsenbeck 2003; Huelsenbeck and FR 2005). Phylogenetic analyses using sequence alignments generated with MUSCLE software (Edgar 2004b, 2004a) produced comparable results (data not shown). To determine the evolutionary model with the highest support from our data, we used model jumping between fixed-rate amino acid models. Bayesian phylogenetic analyses were performed by running 5 million generations and four chains (three heated, one cold), allowing gamma-distributed rate variation across sites and a proportion of invariable sites. Two simultaneous independent runs were started from random trees and the first 250,000 generations discarded as burn-in. Trees were sampled every hundred generations from two million generations. An example of the Markov chain simulation was illustrated in Online Resource 2. Convergence was judged using the standard deviation of split frequencies and the plot of log likelihoods. Consensus trees were visualised using TreeView (Page 1996). Collapsed trees were generated using Interactive Tree of Life (Letunic and Bork 2007, 2011). The RRMs of hnRNP I and L were excluded from the analysis as their inclusion lowered the posterior probabilities across the tree topology, and their evolutionary relationships to the other hnRNPs could not be resolved with any confidence.

Maximum Likelihood phylogenetic analyses were performed using the software PhyML (Guindon and Gascuel 2003) and the JTT substitution model with an estimated

proportion of invariable sites and gamma shape parameter estimated from the data, and four substitution rate categories. We used five random starting trees and subtree pruning and regrafting (SPR) to search tree space. Both tree topology and branch lengths were optimized to maximize the likelihood. Branch support was estimated using 100 bootstrap replicates and by calculating an approximate likelihood branch support.

RNA-Binding Sequence Motifs

In order to investigate the motif conservation involved in RNA-binding, we extracted the RNP-1 and RNP-2 region from the sequence alignments as mentioned above, and used the Weblogo programme (http://weblogo.berkeley.edu/logo.cgi) (Crooks et al. 2004; Schneider and Stephens 1990) to generate a graphical representation of consensus RNA-binding motifs. This approach allowed us to compare individual RNP sequences from different groups of splicing proteins to the consensus motif.

Results

BLASTp Protein Similarity Searches

To examine the protein similarities between the hnRNPs and other RBPs and to identify possible homologues that may not be annotated as RBPs, we blasted the known full-length hnRNPs against the online global protein database from NCBI. In comparison to SR, ELAV-like and CELF proteins, there has been a striking expansion of hnRNPs in multicellular organisms (Barbosa-Morais et al. 2006). Hence, we wished to know more about the significance of the sequence similarity that is shared between different hnRNP protein molecules. Summaries of the full-length protein similarity levels were constructed from the negative (−) log E-values obtained by BLASTp searches of each hnRNP protein against the proteome of its species of origin. Table 1 lists the −log E-values for human hnRNP complete sequences. As expected, highly significant overall sequence similarities were found among hnRNPs that are paralogues or have similar domain architectures, such as the hnRNPs A/B, AB and D (two RRMs plus GRD domains) and the hnRNPs F/H (three aRRM domains), as demonstrated by −log E-values exceeding 50. Surprisingly, the similarity levels between different hnRNP groups that share similar architecture (e.g. A/B/D and F/H) were extremely low, with −log E-values that rarely exceeded 10. Furthermore, several hnRNPs had no detectable similarities to other hnRNPs, such as hnRNPs E/K and U, which lack RRMs. Some hnRNPs such as hnRNPs I/L, which have highly atypical RRMs, only displayed limited sequence similarities by comparison. Such differences in the sequence similarity levels based on −log E-values in relation to the overall protein domain architectures were also observed in mouse, frog, fly, sea squirts and worms.

In addition, we found that non-hnRNPs, such as SR, ELAV-like and CELF proteins, often displayed greater overall sequence similarities to the hnRNPs than the

Table 1 BLASTp similarity matrix of hnRNPs from human

Target \ Query	A0	A1	A2	A3	AB	C	D	E1	E2	F	G	GT	H1	H2	H3	I	K	L	M	P	Q	R	RALY	U	Msi
A0	177.7	61.0	58.5	61.3	28.2		27.0			-2.3	7.3	6.7	-2.9			-2.9		-2.1	0.3		6.0	5.7			43.5
A1	63.2	200.0	92.0	101.7	40.4		41.7			-2.7	8.5	8.5	-1.7	-2.1	-1.5				1.0	4.0	5.5	6.3	-1.9		41.7
A2	61.3	94.4	200.0	92.0	42.0		41.3			1.1	10.4	9.4	-2.8		-2.1				1.1	2.3	4.7	6.2	-0.9		47.2
A3	64.0	104.0	90.7	200.0	38.5		40.7				9.0	8.1			-1.4				0.8	2.1	4.0	7.2	-0.9		46.5
AB	30.2	41.0	40.7	39.5	200.0	0.4	88.7				7.4	7.1	-0.5	-0.9					1.7	-0.7	6.3	6.0	0.1		51.4
C	-1.0	-2.7			0.3	200.0	0.8			-2.4	6.0	4.0				-2.1					0.3	1.1	58.2		
D	28.5	41.7	40.2	40.7	90.0	1.0	200.0			-0.8	8.3	7.5	0.0	0.0	-1.6				1.7	-2.3	8.0	7.4	1.4		47.0
E1								200.0	177.3								17.7								
E2								177.4	200.0								20.2								
F	-0.8				2.0		-0.5			200.0			200.0	200.0	103.4	-2.8			-2.5		-2.5	-2.6			
G	6.7	9.2	11.1	9.5	8.0	6.0	8.3				200.0	102.2				0.2		-0.6	3.1	5.0	5.0	2.7	5.0		10.3
GT	8.0	9.1	10.4	9.0	7.7	4.3	7.7				104.0	200.0				-0.2			3.0	6.0	4.4	5.4	3.7		10.1
H1	-2.3	-2.6	-2.8		0.1		0.3			200.0			200.0	200.0	104.2	-2.5			-2.0	-2.1	-2.2	-2.0			
H2	-2.6	-2.6	-2.2		-0.3		0.5			175.7			200.0	200.0	102.0	-2.5			-2.4	-2.2	-2.0	-1.8			
H3		-2.6			-1.3		-1.8			50.5			54.7	54.4	200.0	-2.7			-1.7		-2.0	-1.8			
I	-2.6					-1.5				-2.6	0.5	-0.3	-2.5	-2.5	-2.2	200.0		34.0	-1.0		1.5	2.7	-0.2		
K								18.3	20.7								200.0			-2.8		-2.6			
L	-1.0	-1.6		-2.6							-0.6	-2.4	-3.0			36.5		200.0			-0.4	0.0			
M	1.8	2.4	1.9	2.7	3.2		3.0			-2.1	3.7	3.5	-1.7	-2.0	-0.8	-0.8			200.0	0.4	1.4	0.8			0.5
P2	0.2	5.0	-2.6	3.7	0.0		-1.0			-1.3	5.0	5.5	-2.6	-3.0	-1.3	-2.3			-0.3	200.0	1.7	2.2	-2.3	-2.4	1.2
Q	7.7	6.0	5.7	6.7	7.4	1.1	9.1			-2.6	5.0	4.2	-2.1	-1.9	-1.8	1.6		-0.1	0.8	-2.9	200.0	200.0	2.4		
R	6.7	6.0	5.7	7.2	6.5	-0.9	8.4			-2.6	3.0	1.6	-2.1	-1.9	-2.4	-0.4		0.2	-1.3	-1.4	200.0	200.0	0.2		6.7
RALY			-2.4		-0.2	51.5	0.1				4.0	2.7				-1.0					0.6	1.0	200.0		
U		-2.8	-2.2	-2.0	-1.6		-2.9	-2.8			-2.9									-1.1		-2.4		200.0	
Msi	41.2	43.2	46.1	42.4	48.7		44.4				11.0	10.5							1.2	1.6		5.1			200.0

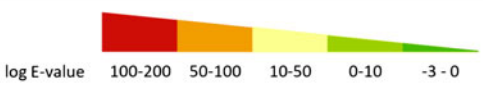
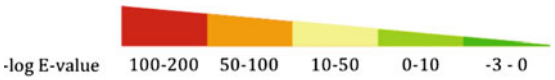


Table 2 Similarity scores between selected hnRNPs and non-hnRNPs

	-log E-value				
hnRNP A1	SRSF2	ELAV-like 1	ELAV-like 3	ELAV-like 4	CELF1
	5.5	8.5	9.2	10.2	10
hnRNP C	SRSF1	SRSF2	SRSF4	SRSF5	SRSF6
	3.1	1.1	5.7	5	5.2
hnRNP G	SRSF2	Tra2β	ELAV-like 3	ELAV-like 4	CELF1
	11.2	13.5	7.2	7.5	6.5



hnRNPs did among themselves (Table 2). There were thus significant overlaps in the sequence similarity space between these groups of proteins, and sequence similarity searches alone were unable to recover groupings consistent with current splicing regulatory protein nomenclature.

Pairwise Full-Length Sequence Comparisons

While these RBPs often share a common RNA-binding domain such as the RRM for function, the inter-domain linker region can act as a major determinant of the sequence-specific affinity towards nascent transcripts (Finger et al. 2004; Shamoo et al. 1995). Therefore, to further elucidate patterns of sequence conservation between these RBPs, full-length sequence comparisons were performed for selected hnRNPs against other hnRNP, SR, ELAV-like and CELF proteins (Fig. 2, additional heatmaps in Online Resource 3–6). For proteins that showed low levels of sequence similarity in the BLASTp searches, regions of identity were generally limited to the RRMs (e.g. A1 vs. D and A1 vs. ELAV-like 2, Fig. 2). Furthermore, proteins that showed insignificant levels of sequence similarity by BLASTp were found to possess no more than 15 % identity in any part of their sequence (e.g. hnRNP A1 vs. C, Fig. 2). On the other hand, high level of similarities in the linker region between the RNA-binding domains were only observed between paralogues, such as hnRNPs A1, A2 and A3 (Fig. 2 and Online Resource 3–6). For many proteins the RRM domains were the only conserved regions indicative of homology both within hnRNPs and between the three groups of splicing proteins, making these domains ideal candidates for further phylogenetic analyses.

Phylogenetic Analyses

Amino acid sequences of human hnRNP RRMs were aligned with ClustalW2, and phylogenetic trees were generated with the alignment using Bayesian inference. Most clades were supported with posterior probabilities usually exceeding 0.7 (Fig. 3). In many cases, we found that RRM domains from paralogues were more similar to each other than to other RRM domains within the gene of interest. For example, the hnRNPs A/B/D each contain two RRM domains. The first RRMs (RRM 1) formed a monophyletic clade, as did the second RRMs (RRM 2). The same pattern was observed for the three RRM domains of hnRNPs F/H and the two RRM domains of Q/R. While the three F/H hnRNPs (F, H1 and H2) share a similar domain architecture and all possess three RRM domains,

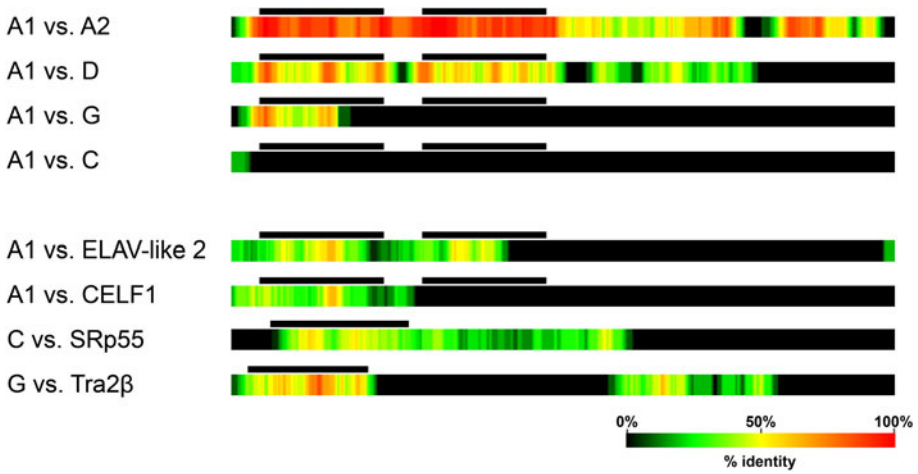


Fig. 2 Heatmaps of pairwise comparisons of hnRNP sequences with other hnRNP or non-hnRNP RBP sequences. *Black bars* over the heatmaps correspond to positions of RRM domains. The sequences identity ranged from 0 % (*green*) to 100 % (*red*). Proteins with similar overall domain architecture (e.g. A1 and A2) show high sequence similarities, especially within the RRMs region (90–100 %). By contrast, there is only 15 % identity in the N-terminus between hnRNPs A1 and C. Higher identity was also observed within the RRM regions of hnRNPs and other RBPS, with low identity outside the RRMs. More heatmaps can be viewed in Online Resources 3–6

hnRNP H3 only has two RRM domains. This may be explained by a simple loss-of-domain in the lineage leading to hnRNP H3 or a more complicated loss-and-acquisition-of-domains in the ancestral gene from which these hnRNPs derived. Nevertheless, these data indicate that the RRMs of the hnRNPs A/B/D and those of the hnRNPs F/H originated from tandem duplications of an ancestral RRM in the gene lineage from which these paralogues were derived. In contrast, the RRMs of hnRNPs Q/R were placed on different major branches. The observed splits were consistent across all other bilaterian species studied including mouse, frog, fly, sea squirts and worms (data not shown). Results from Maximum Likelihood phylogenetic analyses corroborated the results of the Bayesian analyses (Online Resource 7) with only minor differences in topology, and analyses using alignments generated from MUSCLE software yielded comparable results. Importantly, all major splits were in concordance in these analyses.

Next, we investigated the evolutionary relationships among different groups of splicing proteins from the ten species studied (Fig. 4, Online Resource 8 and 9). Expansion of the splicing regulatory factors in pluricellular organisms suggested that these protein factors may have played a defining role in metazoan evolution (Barbosa-Morais et al. 2006). Therefore, we have identified orthologues in ten metazoan species, three species from vertebrates and seven from invertebrates. No ELAV-like or CELF orthologues were found in unicellular eukaryotes such as *S. pombe* while hnRNP and SR proteins have only one or two representatives, and hence these species were omitted from this study. Again, we utilized ClustalW2 to perform multi-sequence alignment of the RRMs retrieved from all ten species for Bayesian inference analysis. In general, RRMs were found to be more conserved between orthologues than between these splicing protein groups within a species (Fig. 4a, Online Resource 8). For example, the corresponding hnRNP A1 and Musashi 1 RRMs in all ten metazoan species were grouped together in different analyses. Such close relationships between orthologue RRMs points to possible duplication and diversification of these domains before the last common ancestor of metazoans studied here.

There were several unexpected phylogenetic relationships among the different groups of splicing factors. First, while RRM 1 of SRSF1, an SR protein, formed a monophyletic clade with the sole RRM of hnRNP C and that of another SR protein (SRSF2), RRM 2 of SRSF1 formed a monophyletic clade with the RRMs of hnRNP F/H. Importantly, across different analyses performed, RRM 1 of SRSF1 and the RRMs of hnRNP C and SRSF2

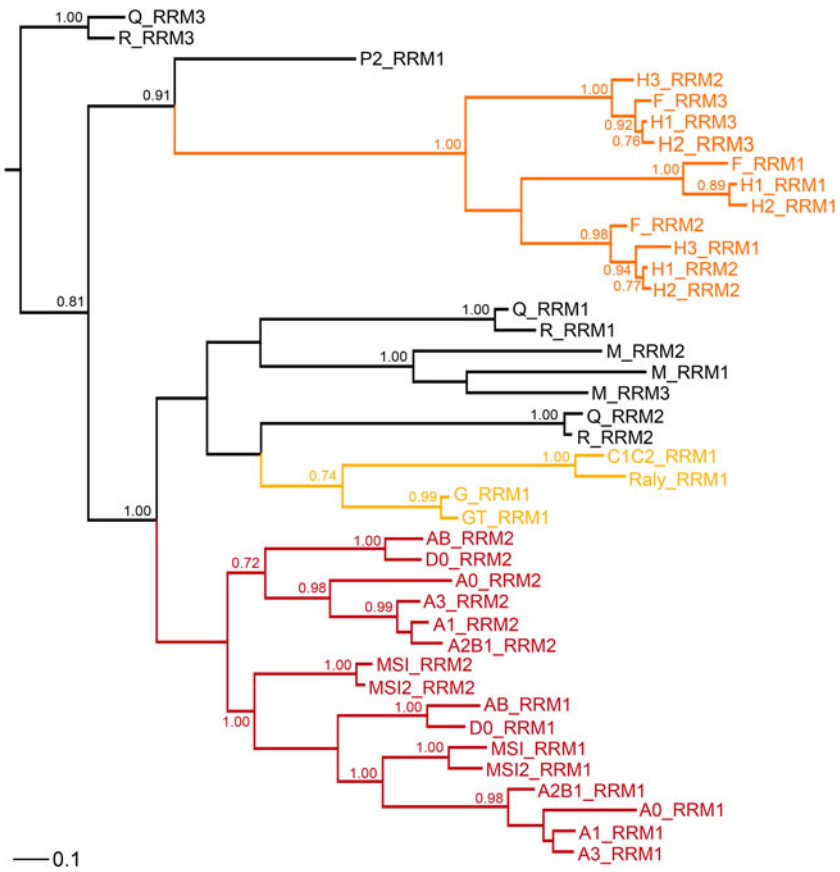


Fig. 3 Bayesian phylogenetic tree of human hnRNP RRMs. *Values above nodes* indicate *posterior* probabilities of the bipartition. The tree was arbitrarily rooted *midway* along the branch separating the RRM 3 of hnRNPs Q and R from the other RRMs

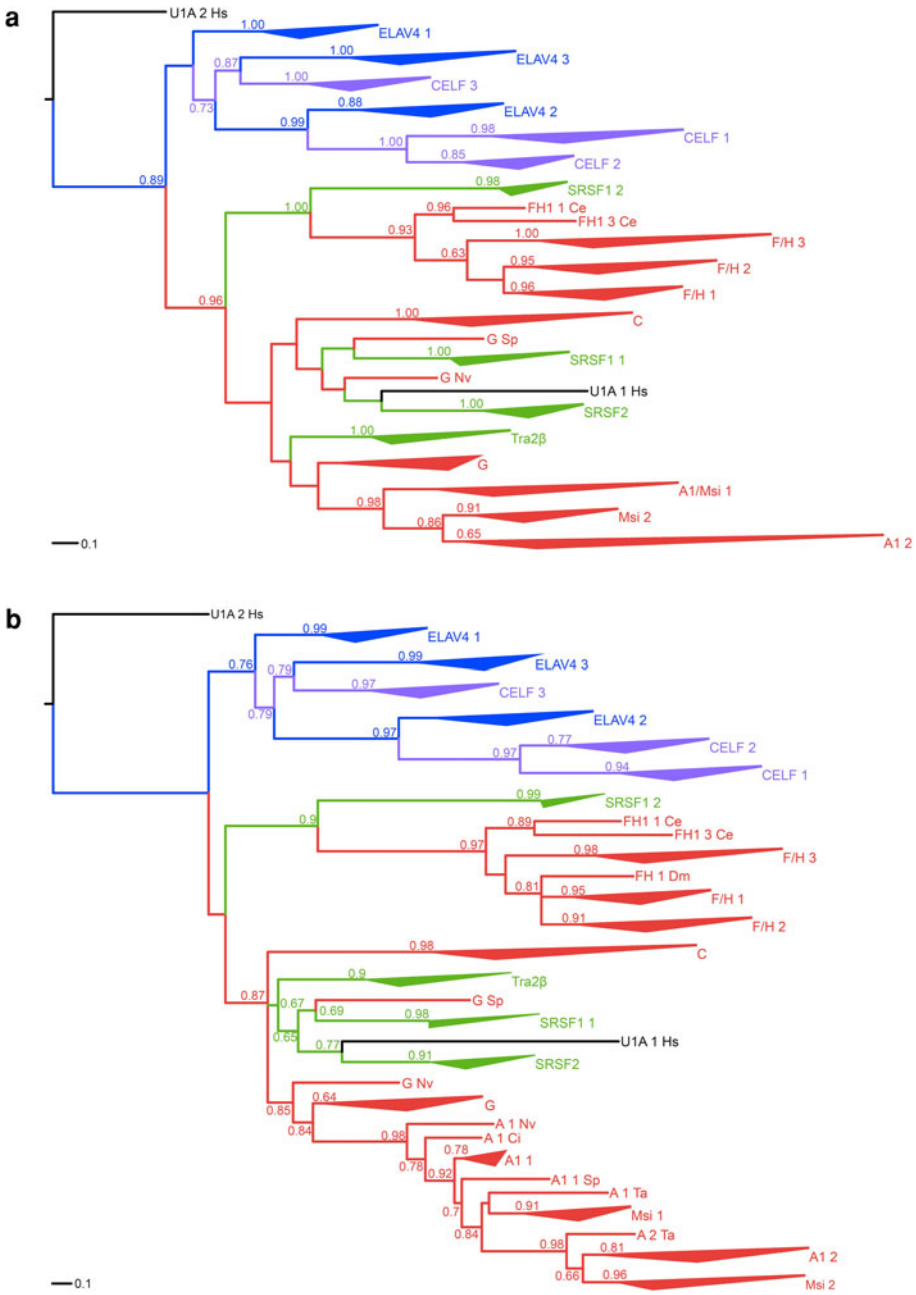


Fig. 4 Collapsed phylogenetic trees of RRMs across the hnRNP (*red*), SR (*green*), ELAV-like (*blue*) and CELF (*purple*) proteins from various species derived by **a** Bayesian inference and **b** Maximum Likelihood. The collapsed branches are shown as a triangle. The *top* side of the *triangle* shows the distance to the farthest branch, while the *bottom* side shows the distance to the *closest* branch. *Hs—H. sapiens*, *Mm—M. musculus*, *Xt—X. tropicalis*, *Dm—D. melanogaster*, *Dp—D. pulex*, *Sp—S. purpuratus*, *Ci—C. intestinalis*, *Ce—C. elegans*, *Nv—N. vecetensis and Ta—T. adhaerens*. *Values above nodes* indicate *posterior* probabilities or bootstrap value over 0.6 for the *bipartition*. The tree was arbitrarily rooted *midway* along the branch separating RRM 2 of U1A from the other RBPs. Collapsed trees were generated using the Interactive Tree of Life programme available at http://itol.embl.de/

consistently formed a monophyletic clade, while RRM 2 of SRSF1 was promiscuous, associating with the RRMs of hnRNPs F/H or ELAV-like/CELF proteins. However, the posterior probability for the split between the RRM of SRSF2 and the RRMs of hnRNP C and SRSF1 was low, and the relative position of SRSF2 could not be ascertained with confidence. Second, the RRM domains of the hnRNPs were paraphyletic in these comparisons, with the RRM of hnRNP F/H being more closely related to those of SR proteins than to those of other hnRNPs (posterior probabilities supporting this paraphyly being 0.96 and 1). Similarly, the RRM domains of SR proteins were paraphyletic, as the clade also contained the RRMs of hnRNPs A/C/F/H/G. The RRMs of ELAV-like and CELF proteins are also paraphyletic, with RRMs 1 and 2 of CELF proteins forming a monophyletic clade with the second RRM of ELAV-like proteins, while RRM 3 of CELF is more closely related to the first and third RRMs of ELAV-like proteins. The observation that the RRMs from the different protein groups are intermixed suggests that the nomenclature for these splicing proteins is discordant with their evolutionary origins, which show complex patterns of domain gain across different splicing factor groups. Again, comparable analyses using Maximum Likelihood approaches corroborated the results from Bayesian analyses with only minor differences in tree topology (Fig. 4b, Online Resource 9).

In separate analyses that used sequence alignments generated from the MUSCLE alignment software, there are only minor differences in tree topologies compared to the tree presented in Fig. 4. Thus, the RRMs of the different groups of splicing regulators have mixed evolutionary origins, and this was consistent across metazoan species.

RNP Consensus Sequence Motifs

RRM domains show a highly conserved overall domain topology consisting of β_1-α_1-β_2-β_3-α_2-β_4 secondary structure elements within which there are two conserved sequence motifs in the β_3 and β_1 strands: an octamer RNP-1 and a hexamer RNP-2 (Fig. 5) (Birney et al. 1993). As the RNP motifs are critical for the RNA-binding properties of RBPs, we examined their patterns of conservation by generating consensus motifs for RNP-1 and RNP-2 sequences from the RRMs of hnRNP, SR, ELAV-like and CELF proteins (Fig. 6). The multi-species consensus motifs were generally similar across the groups of splicing regulators. In RNP-1, the most conserved features are aromatic residues at positions 3, 5 and 8, hydrophobic residues at positions 4 and 6, glycine at positions 2 and 4, and an acidic residue at position 1. In RNP-2, the most conserved features are hydrophobic residues at positions 1, 3 and 5, and aromatic residues at position 2. This is in agreement with structural studies that have revealed that aromatic rings at positions 3 and 5 of RNP-1 and position 2 of RNP-2 contribute to base-stacking interactions with RNA (Maris et al. 2005). However, there were minor variations between the RNP consensus sequences. RNP-1 residues 1 and 3 are more poorly conserved in SR and ELAV-like proteins, respectively, and residue 3 tends to be either hydrophobic (cysteine) or aromatic (phenylalanine) in CELF proteins. RNP-2 residues 4 and 5 tended to be glycines in hnRNPs, polar residues in ELAV-like proteins, and a mixture of both in SR and CELF proteins. Similar consensus sequences were obtained in analyses of these splicing proteins from individual species, which suggests that these motifs have been conserved across metazoa.

The strong conservation of RNP consensus motifs between the different groups of splicing regulators suggests that there has been selection pressure against non-conservative substitutions and insertions that would affect RNA-binding. Nevertheless, a number of splicing proteins contain atypical RNP motifs and employ alternative modes of RNA-binding (Dominguez and Allain 2006; Tintaru et al. 2007). These atypical RNP motifs contain mutations at positions that are critical for interaction with RNA. For example, the conserved phenylalanine at position 5 in RNP-1 is replaced by a valine in SRSF1 and SRSF6 proteins (Fig. 7), and this substitution is conserved across species. As these RNPs are well-conserved across the species studied, this suggests that these aRRMs adopted alternative modes of RNA-binding early in the evolution of splicing factors. Importantly, hydrophobic residues that make up the core of the domain (Maris et al. 2005) have been conserved at positions 4 and 6 of RNP-1 and positions 1, 3 and 5 of RNP-2. This indicates that there has been particularly strong selection pressure against non-conservative mutations of residues that would affect RRM domain secondary structure, while mutations at other positions are associated with alternative modes of RNA-binding.

Discussion

In this study, we have examined the sequences and phylogenetic relationships among four diverse groups of RBPs, the hnRNP, SR, ELAV-like and CELF proteins. Our goal was to identify evolutionary, structural and functional

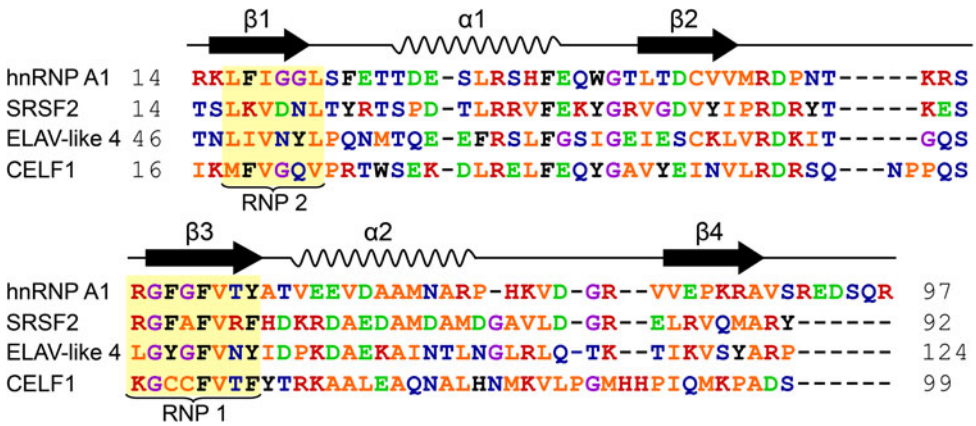


Fig. 5 Conservation of overall domain topology and RNP motifs within RRMs from representative human hnRNP, SR, ELAV-like and CELF proteins. *Colours* indicate acidic (*red*), basic (*blue*), aromatic (*black*), glycine (*purple*), other hydrophobic (*yellow*) and polar residues (*green*). Positions of secondary structures are based on (Birney et al. 1993)

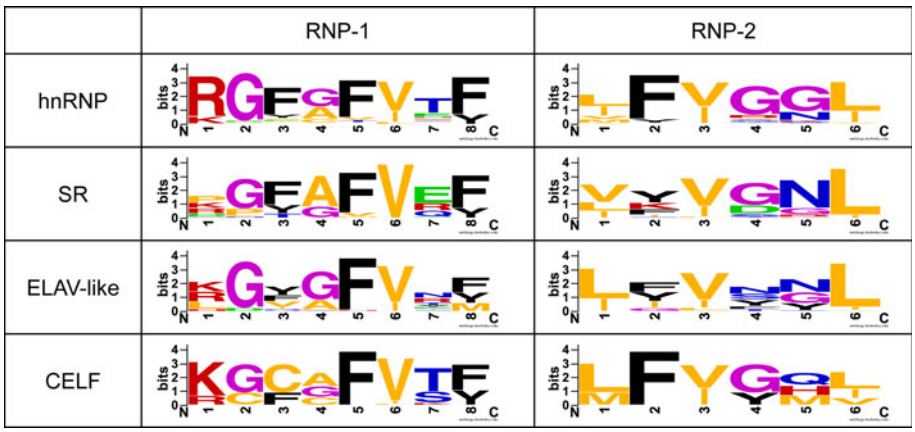


Fig. 6 RNP consensus motifs of RRMs from different groups of RBPs *across* ten metazoan species. The *colour* scheme follows that in Fig. 5. Note that atypical RRMs (hnRNPs F/H and I/L) were not included. hnRNP C was included under SR-like proteins based on the results of phylogenetic analyses presented in Fig. 4. Consensus motifs were generated using the Weblogo programme available at http://weblogo.berkeley.edu/logo.cgi

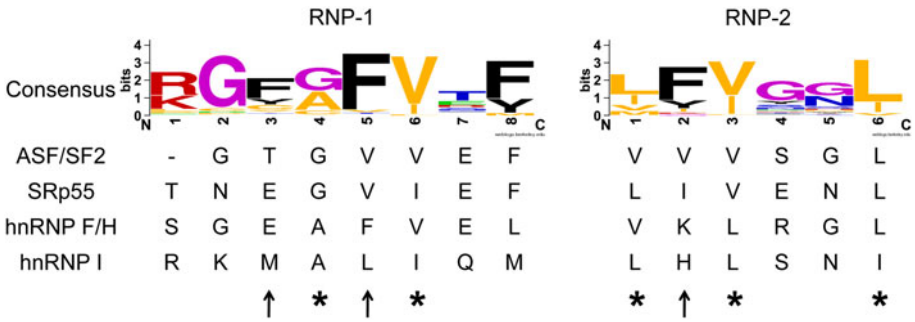


Fig. 7 Comparison of degenerate RNP motifs in aRRMs against consensus RNP motifs. The *colour* scheme follows that in Fig. 5. *Arrows* indicate positions of aromatic residues that are important for RNA-binding in standard RRMs. *Asterisks* indicate positions of conserved hydrophobic residues that are important for maintaining RRM secondary structure

relationships between these groups of proteins that are thought to be closely related. This study builds upon the study by Birney et al. (Birney et al. 1993) in which phylogenetic analyses were conducted with selected RRMs across a wide range of RBP types. Here, we have adopted a complementary approach, focusing on four groups of RBPs and including a greater number of representatives from a wide range of metazoan species to examine the intra- and inter-group relationships of RBPs in more detail. We found that some hnRNPs share greater sequence similarity with SR, ELAV-like and CELF proteins than with other hnRNPs, which are mainly manifested within the shared modular domains. We were unable to detect any regions of sequence conservation that would indicate homology between hnRNPs E/K, I/L and U and the other hnRNPs. Thus, the description of the hnRNPs as a protein family in the sense of evolutionarily related sequences is not supported by our analyses. The extremely low level of sequence similarity between some of the hnRNPs suggests that they have different evolutionary origins, or that their sequences have diverged greatly from a shared ancestral sequence. Similarly, nor can the SR proteins be separated into distinct protein families. On the other hand, given the high resemblance in the overall domain architecture and the fact that CELF proteins have been described as distantly related to the ELAV-like proteins, it is not surprising to see that these two protein groups formed a monophyletic clade. Yet, the ELAV-like and CELF proteins have strikingly different high-affinity binding sequences. Although both ELAV-like and CELF proteins are widely distributed, ELAV-like proteins are predominant in neurons, and CELF proteins are highly abundant in striated muscle tissues (Dasgupta and Ladd 2012; Pascale and Govoni 2012). Inclusion of additional species as sequences become available will help further elucidate the complex relationships among these four groups of RBPs. While it is possible that some clades currently identified in this study as monophyletic will become paraphyletic as more species are included, this will provide further evidence for the early diversification of these RBPs through several rounds of

domain loss and acquisition. Hence, our study suggests that the RBPs are an ancient group of proteins of evolutionarily heterogeneous origin that diverged rapidly early in or before metazoan evolution via domain duplication and the acquisition of domains from more-distantly related proteins. Importantly, since their early diversification, these RBPs have been subject to strong selective pressures to maintain domain architectures, protein fold and RNA-binding motif sequences. As a result, we find that overall structure, especially the RNA-binding domains within orthologues are highly conserved, while paralogues have greatly diverged in sequence to the extent that for some RBP proteins, there is no detectable homology to other RBPs.

Despite the limited overall sequence similarity among many of the RBPs studied, their RRMs are structurally and functionally similar, adopting a canonical RRM fold. For instance, the RRMs of hnRNPs A1 and C, which are highly divergent in sequence, share a common structure and have comparable RNA-binding interfaces (Wittekind et al. 1992; Ding et al. 1999). Differences in RNA-binding properties, e.g. in sequence, affinities and specificities (Singh and Valcarcel 2005; Dreyfuss et al. 1993; Martinez-Contreras et al. 2007), are the result of structural differences in the conformation of helices, loops and linker regions within the RRM fold (Maris et al. 2005). Our results suggest that early diversification of RRM domains has resulted in domain sequences and combinations that are well-conserved across a wide range of metazoan species, with orthologues showing strong conservation in the sequences of their RRM domains. This suggests that there has been a brief but intensive period of evolutionary innovation that resulted in novel RNA-binding properties while general RNA-binding characteristics were maintained. This diversification is thought to have allowed the RBPs to rapidly adopt functional niches within emerging metazoan RNA processing pathways (Anantharaman et al. 2002). As a result, RBPs play multiple, often redundant or overlapping, roles in co- and post-transcriptional processing (Singh and Valcarcel 2005; Long and Caceres 2009; Dreyfuss et al. 2002; Samson 2008). This high degree of functional overlap between RBPs in central gene expression pathways complicates their classification into functionally distinct groups.

The high level of conservation of RNP motifs across multiple species called our attention to their degenerate counterparts in aRRMs, which have alternative modes of RNA-binding. In RRM 2 of SRSF1, mutations within the degenerate RNP motifs have no effect on interactions with RNA, and key RNA-binding residues are present in the α_1 helix and β_2 strand instead (Tintaru et al. 2007). Atypical RRMs 1 and 2 of hnRNPs F/H bind RNA via an additional β hairpin that is not present in typical RRMs (Dominguez and Allain 2006). The RNA-binding region of hnRNP I, which is more commonly known as polypyrimidine tract binding protein, is significantly extended compared to that of canonical RRMs (Simpson et al. 2004). Furthermore, our phylogenetic analyses demonstrate that the RRMs of hnRNP F/H and the second RRM of SRSF1 are separated by long branches from the other RRMs. If these aRRMs arose from the same ancestor as the canonical RRMs, their sequences have diverged to the point where the mechanism for RNA-binding has been completely altered.

An alternative explanation for the origin of these aRRMs might be that they arose from evolutionarily unrelated sequences. For example, the β_1-α_1-β_2-β_3-α_2-β_4 topology is common among protein superfamilies, and can thus be obtained via stepwise additions or deletions of α-helices or β-strands from a large number of root motifs (Efimov 1997). Furthermore, $\alpha\beta$ folds exhibit significant structural overlap with over 20 % of other protein folds, which implies that they result from dominant folding pathways or represent especially stable structural configurations (Harrison et al. 2002). It is hence possible that the RRM domain has been invented multiple times in evolution. The high level of sequence divergence and evolutionary distance between the RBPs makes it difficult to reject either hypothesis. This highlights the fact that structural similarities need not necessarily imply a common evolutionary ancestry, and cautions against the assumption of functional and evolutionary relationships based on the presence of common structural domains (Skolnick et al. 2009).

In addition, there is evidence that different RBPs were created by distinct evolutionary mechanisms, resulting in mixed or reticulate evolutionary relationships among these modular proteins. For instance, in most RBPs, tandem duplication of RRM domains has occurred after gene creation; this is reflected in our observation that RRMs occupying the same position in orthologues are more closely related than RRMs occupying different positions in the same protein. By contrast, hnRNPs Q/R and SRSF1 acquired RRMs from multiple sequences with different evolutionary histories. The placement of the RRMs of hnRNPs Q/R and SRSF1 in different branches suggests that these proteins may have arisen from the fusion of genes encoding different RRMs, or by acquiring a domain from an otherwise unrelated sequence. Protein domains constitute independent evolutionary units, and domain duplication and fusion are important sources for the generation of multi-domain proteins (Ponting and Russell 2002; Vogel et al. 2004).

The complex evolutionary relationships among the hnRNP, SR, ELAV-like and CELF proteins have further complicated accurate transfer of functional annotation between protein family members. For instance, CELF1 was first identified as hNAb50, a novel member of the hnRNP

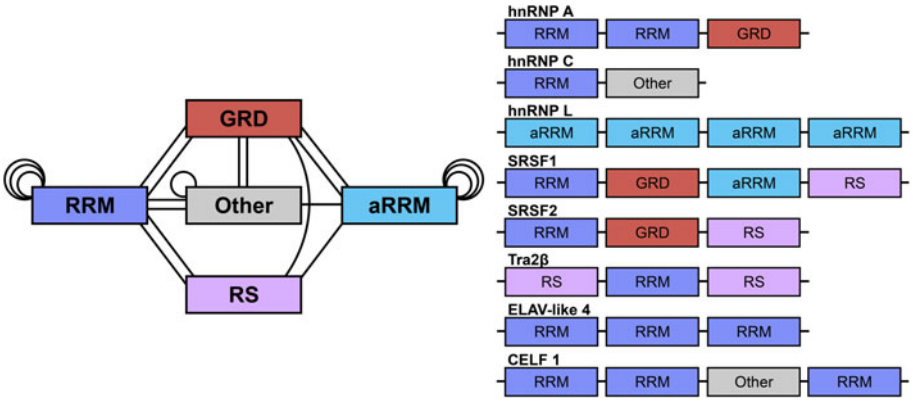


Fig. 8 Domain graph of RBPs based on phylogenetic analyses. Domains are represented as *boxes*, with connection lines signifying common memberships that link the different domains in a protein. Each *line* represents one linker region between the domains from one direction. RRMs were assigned to the classical RRM (RRM) or atypical RRM (aRRM). Domain combinations of selected RBPs are also presented (GRD—glycine-rich domain; RS—arginine/serine-rich domain; Other—other type of domain)

proteins, which binds to myotenin protein kinase (Mt-PK) transcript and was later classified as a CELF protein based on sequence homology (Ladd et al. 2001; Timchenko et al. 1996). Moreover, an increasing number of novel splicing factors are now identified using new methodologies such as mass spectrometry-based proteomic analysis (Kasyapa et al. 2005; Chen et al. 2007; Rappsilber et al. 2002; Zhou et al. 2002; Ben-Dov et al. 2008). However, since experimental identification of these novel proteins is time-consuming, various computational methods have been developed to help identify and characterize novel RBPs based on the currently available evolutionary information (Hsu et al. 2011). Our study has highlighted the limitations in functional predictions based on protein family nomenclature and calls for caution when attempting to extrapolate functional information between proteins with low sequence identity (Todd et al. 2001; Devos and Valencia 2000). Given the importance of accurate exploitation of evolutionary relationships, in recent years new approaches have been suggested for better classification of these RBPs. For example, Manley and Krainer have proposed more precise criteria for defining SR proteins based entirely on their sequence properties (Manley and Krainer 2010). Our heatmaps revealed that the RRM is the most highly conserved region among RBPs, and our phylogenetic trees were highly supported at most branches, which shows that an RRM-based approach to the classification of the RBPs may help resolve the inconsistencies in the grouping of these RBPs with mixed evolutionary histories. RBPs often contain other domains, such as glycine-rich and RS domains. However, these domains are comprised of highly repetitive sequences of low complexity, and are not amenable to phylogenetic analysis. Even so, their presence and position within proteins can provide information about the evolutionary history of these proteins. Given the modularity of RBPs, which has led many researchers to intuitively characterise them based on domain composition (Singh and Valcarcel 2005; Lunde et al. 2007; Dreyfuss et al. 2002; Biamonti and Riva 1994), a domain-centric approach may be a natural choice for describing the evolutionary relationships among these modular proteins.

Figure 8 illustrates how the protein structures in Fig. 1 and the relationships between RRMs shown in Fig. 4 could be presented in a domain graph, which captures the frequency of domains and domain combinations occurring within this limited set of RBPs. For example, the graph in Fig. 8 shows that RRMs and aRRMs have frequently undergone domain duplications, creating RBPs that contain up to four RRMs as in hnRNPs I and L. The RRMs and aRRMs are frequently associated with auxiliary domains (such as GR in hnRNPs or RS in SR proteins) and other specific types of domain, or linker region conserved within the groups of proteins (e.g. ELAV-like and CELF). Alternatively, the auxiliary domains could associate with each other. Given the challenges associated with the phylogenetic analyses of this heterogeneous group of proteins, a more practical approach to their classification may be to describe each RBP as a sum of its domains, more accurately reflecting the modular nature of these proteins. Since domains often represent functional units, this description would also encapsulate the functional properties of each RBP.

Conclusion

The examination of phylogenetic relationships between the hnRNP, SR, ELAV-like and CELF proteins has provided new insights into their evolution. Despite early and extensive diversification of their protein sequences, these RBPs have maintained a high level of structural conservation, particularly within critical RNA-binding motifs. We have highlighted issues with classification of these

proteins, which are complicated by the structural and functional overlaps between the different groups. In addition, our phylogenetic analyses show that the RRMs are ancient domains that have diverged markedly from each other. In contrast, RRMs encoded by orthologue genes have been subject to strong selection pressures. This is indicative of extensive diversification before the last common ancestor of the metazoans studied here, after which RRM sequences became highly constrained. We conclude that the current nomenclature of hnRNPs and their current classification with respect to other RBPs does not adequately reflect patterns of sequence homology and the evolutionary history of the RBPs, and propose that a domain-centric approach may be more suitable for the study of these highly modular proteins. Delineation of the evolutionary relationships among these splicing regulators is critical in advancing our understanding of how transcriptional complexity and regulation, which underpins much of the biological complexity of higher metazoa, has evolved.

Acknowledgments We thank Dr. Sam Lukowski for useful feedback and suggestions, and Prof. Mark Ragan (Institute for Molecular Biosciences at the University of Queensland) for providing the computer resources for the phylogenetic analyses. Phylogenetic analyses were performed on the NCI National Facility at the Australian National University. This study was supported by grant ID631551 from the Australian National Health and Medical Research Council and Grant ID455870 from the Cancer Council of Queensland to RS and JAR. YHT and SPH were supported by University of Queensland Research Scholarships. We acknowledge financial support from the ARC Centre of Excellence in Bioinformatics (ARC CEO348221), and an award under the Merit Allocation Scheme on the NCI National Supercomputing Facility.

References

Anantharaman V, Koonin EV, Aravind L (2002) Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res 30(7):1427–1464

Aravind L, Subramanian G (1999) Origin of multicellular eukaryotes—insights from proteome comparisons. Curr Opin Genet Dev 9(6):688–694

Barbosa-Morais NL, Carmo-Fonseca M, Aparicio S (2006) Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. Genome Res 16(1):66–77

Ben-Dov C, Hartmann B, Lundgren J, Valcarcel J (2008) Genome-wide analysis of alternative pre-mRNA splicing. J Biol Chem 283(3):1229–1233

Biamonti G, Riva S (1994) New insights into the auxiliary domains of eukaryotic RNA binding proteins. FEBS Lett 340(1–2):1–8

Birney E, Kumar S, Krainer AR (1993) Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. Nucleic Acids Res 21(25): 5803–5816

Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72:291–336

Blanchette M, Green RE, MacArthur S, Brooks AN, Brenner SE, Eisen MB, Rio DC (2009) Genome-wide analysis of alternative pre-mRNA splicing and RNA-binding specificities of the *Drosophila* hnRNP A/B family members. Mol Cell 33(4): 438–449

Busch A, Hertel KJ (2012) Evolution of SR protein and hnRNP splicing regulatory factors. Wiley Interdiscip Rev RNA 3(1):1–12

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3(4):285–298

Chen L, Zheng S (2009) Studying alternative splicing regulatory networks through partial correlation analysis. Genome Biol 10(1):R3

Chen YI, Moore RE, Ge HY, Young MK, Lee TD, Stevens SW (2007) Proteomic analysis of in vivo-assembled pre-mRNA splicing complexes expands the catalog of participating factors. Nucleic Acids Res 35(12):3928–3944

Chothia C, Gough J (2009) Genomic and structural aspects of protein evolution. Biochem J 419(1):15–28

Consortium U (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40 (Database issue):D71–75

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14(6):1188–1190

Dasgupta T, Ladd AN (2012) The importance of CELF control: molecular and biological roles of the CUG-BP, Elav-like family of RNA-binding proteins. Wiley Interdiscip Rev RNA 3(1): 104–121

Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41(1):98–107

Ding J, Hayashi MK, Zhang Y, Manche L, Krainer AR, Xu RM (1999) Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. Genes Dev 13(9):1102–1115

Dominguez C, Allain FH (2006) NMR structure of the three quasi RNA recognition motifs (qRRMs) of human hnRNP F and interaction studies with Bcl-x G-tract RNA: a novel mode of RNA recognition. Nucleic Acids Res 34(13):3634–3645

Dreyfuss G, Matunis MJ, Pinol-Roma S, Burd CG (1993) hnRNP proteins and the biogenesis of mRNA. Annu Rev Biochem 62:289–321

Dreyfuss G, Kim VN, Kataoka N (2002) Messenger-RNA-binding proteins and the messages they carry. Nat Rev Mol Cell Biol 3(3):195–205

Edgar RC (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113

Edgar RC (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5): 1792–1797

Efimov AV (1997) Structural trees for protein superfamilies. Proteins 28(2):241–260

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30(7):1575–1584

Finger LD, Johansson C, Rinaldi B, Bouvet P, Feigon J (2004) Contributions of the RNA-binding and linker domains and RNA structure to the specificity and affinity of the nucleolin RBD12/NRE interaction. Biochemistry 43(22):6937–6947

Good PJ (1995) A conserved family of elav-like genes in vertebrates. Proc Natl Acad Sci U S A 92(10):4557–4561

Gough J (2005) Convergent evolution of domain architectures (is rare). Bioinformatics 21(8):1464–1471

Graumann P, Marahiel MA (1996) A case of convergent evolution of nucleic acid binding modules. BioEssays 18(4):309–315

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52(5):696–704

Han SP, Kassahn KS, Skarshewski A, Ragan MA, Rothnagel JA, Smith R (2010a) Functional implications of the emergence of alternative splicing in hnRNP A/B transcripts. RNA 16(9):1760–1768

Han SP, Tang YH, Smith R (2010b) Functional diversity of the hnRNPs: past, present and perspectives. Biochem J 430(3):379–392

Harrison A, Pearl F, Mott R, Thornton J, Orengo C (2002) Quantifying the similarities within fold space. J Mol Biol 323(5): 909–926

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89(22): 10915–10919

Hsu JB, Bretana NA, Lee TY, Huang HD (2011) Incorporating evolutionary information and functional domains for identifying RNA splicing factors in humans. PLoS ONE 6(11):e27567

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P (2009) Ensembl 2009. Nucleic Acids Res 37 (Database issue):D690–697

Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, Yan BY, Donohue JP, Shiue L, Hoon S, Brenner S, Ares M Jr, Yeo GW (2012) Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep 1(2): 167–178

Huelsenbeck JP, Ronquist F (2005) Bayesian Analysis of Molecular Evolution using MrBayes. In: Nielsen R (ed) Statistical methods in molecular evolution. Springer, Newyork, pp 183–232

Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 10:136

Kasyapa CS, Kunapuli P, Cowell JK (2005) Mass spectroscopy identifies the splicing-associated proteins, PSF, hnRNP H3, hnRNP A2/B1, and TLS/FUS as interacting partners of the ZNF198 protein associated with rearrangement in myeloproliferative disease. Exp Cell Res 309(1):78–85

Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11(5):345–355

Kirschner M, Gerhart J (1998) Evolvability. Proc Natl Acad Sci U S A 95(15):8420–8427

Ladd AN, Charlet N, Cooper TA (2001) The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. Mol Cell Biol 21(4):1285–1296

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947–2948

Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23(1):127–128

Letunic I, Bork P (2011) Interactive tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res 39 (Web Server issue):W475–478

Long JC, Caceres JF (2009) The SR protein family of splicing factors: master regulators of gene expression. Biochem J 417(1):15–27

Lunde BM, Moore C, Varani G (2007) RNA-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol 8(6):479–490

Lupas AN, Ponting CP, Russell RB (2001) On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 134(2–3):191–203

Maglott D, Ostell J, Pruitt KD, Tatusova T (2005) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 33 (Database issue):D54–58

Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011:bar009

Maniatis T, Tasic B (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature 418(6894):236–243

Manley JL, Krainer AR (2010) A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). Genes Dev 24(11):1073–1074

Maris C, Dominguez C, Allain FH (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J 272(9):2118–2131

Martinez-Contreras R, Cloutier P, Shkreta L, Fisette JF, Revil T, Chabot B (2007) hnRNP proteins and splicing control. Adv Exp Med Biol 623:123–147

Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol 6(5):386–398

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453

Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12(4):357–358

Pascale A, Govoni S (2012) The complex world of post-transcriptional mechanisms: is their deregulation a common link for diseases? Focus on ELAV-like RNA-binding proteins. Cell Mol Life Sci 69(4):501–517

Piñol-Roma S, Choi YD, Matunis MJ, Dreyfuss G (1988) Immunopurification of heterogeneous nuclear ribonucleoprotein particles reveals an assortment of RNA-binding proteins. Genes Dev 2(2):215–227

Ponting CP, Russell RR (2002) The natural history of protein domains. Annu Rev Biophys Biomol Struct 31:45–71

Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35 (Database issue):D61–65

Rappsilber J, Ryder U, Lamond AI, Mann M (2002) Large-scale proteomic analysis of the human spliceosome. Genome Res 12(8): 1231–1245

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: bayesian phylogenetic inference under mixed models. Bioinformatics 19(12):1572–1574

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH, Lewis S (2000) Comparative genomics of the eukaryotes. Science 287(5461):2204–2215

Samson ML (2008) Rapid functional diversification in the structurally conserved ELAV family of neuronal RNA binding proteins. BMC Genomics 9:392

Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18(20):6097–6100

Shamoo Y, Abdul-Manan N, Williams KR (1995) Multiple RNA binding domains (RBDs) just don't add up. Nucleic Acids Res 23(5):725–728

Shepard PJ, Hertel KJ (2009) The SR protein family. Genome Biol 10(10):242

Simpson PJ, Monie TP, Szendroi A, Davydova N, Tyzack JK, Conte MR, Read CM, Cary PD, Svergun DI, Konarev PV, Curry S, Matthews S (2004) Structure and RNA interactions of the N-terminal RRM domains of PTB. Structure 12(9):1631–1643

Singh R, Valcarcel J (2005) Building specificity with nonspecific RNA-binding proteins. Nat Struct Mol Biol 12(8):645–653

Skolnick J, Arakaki AK, Lee SY, Brylinski M (2009) The continuity of protein structure space is an intrinsic property of proteins. Proc Natl Acad Sci U S A 106(37):15690–15695

Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H (2005) Function of alternative splicing. Gene 344:1–20

Talukdar I, Sen S, Urbano R, Thompson J, Yates JR 3rd, Webster NJ (2011) hnRNP A1 and hnRNP F modulate the alternative splicing of exon 11 of the insulin receptor gene. PLoS ONE 6(11):e27869

Tavanez JP, Madl T, Kooshapur H, Sattler M, Valcarcel J (2012) hnRNP A1 proofreads 3′ splice site recognition by U2AF. Mol Cell 45(3):314–329

Timchenko LT, Miller JW, Timchenko NA, DeVore DR, Datar KV, Lin L, Roberts R, Caskey CT, Swanson MS (1996) Identification of a (CUG)n triplet repeat RNA-binding protein and its expression in myotonic dystrophy. Nucleic Acids Res 24(22): 4407–4414

Tintaru AM, Hautbergue GM, Hounslow AM, Hung ML, Lian LY, Craven CJ, Wilson SA (2007) Structural and functional analysis of RNA and TAP binding to SF2/ASF. EMBO Rep 8(8):756–762

Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307(4):1113–1143

Vogel C, Chothia C (2006) Protein family expansions and biological complexity. PLoS Comput Biol 2(5):e48

Vogel C, Teichmann SA, Chothia C (2003) The immunoglobulin superfamily in *Drosophila* melanogaster and *Caenorhabditis elegans* and the evolution of complexity. Development 130(25): 6317–6328

Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA (2004) Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol 14(2):208–216

Wittekind M, Gorlach M, Friedrichs M, Dreyfuss G, Mueller L (1992) 1H, 13C, and 15 N NMR assignments and global folding pattern of the RNA-binding domain of the human hnRNP C proteins. Biochemistry 31(27):6254–6265

Yang S, Bourne PE (2009) The evolutionary history of protein domains viewed by species phylogeny. PLoS ONE 4(12):e8378

Zhou Z, Licklider LJ, Gygi SP, Reed R (2002) Comprehensive proteomic analysis of the human spliceosome. Nature 419(6903): 182–185