Author's personal copy

Review

RNA–protein interactions *in vivo*: global gets specific

Minna-Liisa Änkö* and Karla M. Neugebauer

Max Planck Institute of Cell Biology and Genetics, Dresden, Germany

RNA-binding proteins (RBPs) impact every process in the cell; they act as splicing and polyadenylation factors, transport and localization factors, stabilizers and destabilizers, modifiers, and chaperones. RNA-binding capacity can be attributed to numerous protein domains that bind a limited repertoire of short RNA sequences. How is specificity achieved in cells? Here we focus on recent advances in determining the RNA-binding properties of proteins *in vivo* and compare these to *in vitro* determinations, highlighting insights into how endogenous RNA molecules are recognized and regulated. We also discuss the crucial contribution of structural determinations for understanding RNA-binding specificity and mechanisms.

Introduction

Eukaryotic genomes encode hundreds of proteins with the capacity to bind RNA. RNA binding can be conferred by a variety of protein motifs, including RNA recognition motifs (RRMs), K homology (KH) domains and zinc fingers [1]. Importantly, cellular RNA is not naked but is associated with proteins in ribonucleoprotein (RNP) complexes, which comprise most functional forms of RNA. RBPs that bind mRNA can change gene expression output at different steps of RNA metabolism, including capping, pre-mRNA splicing, editing, polyadenylation, RNA export, RNA stability, and translation (Figure 1). Mutations in RBPs or in the *cis* sequences that disrupt RNA interactions with RBPs can cause disease [2]; moreover, disruption of RBP expression compromises the viability of cells and organisms [2,3]. These phenotypes underscore the central role of RNA–protein interactions in cellular processes. However, only a few RBPs have been studied extensively and many are only predicted based on sequence similarity. Thus, the functions of most RBPs remain elusive. *In vitro* studies have shown that RBPs generally bind to relatively degenerate and/or short sequence motifs. However, these motifs do not contain enough information to predict the binding sites or RNA targets of RBPs through simple sequence analysis (Figure 2a). Therefore, central questions in RNA biology are: what are the RNA targets of RBPs; and how is their *in vivo* binding specificity achieved?

Recent developments in high-throughput technologies have allowed, for the first time, the identification of RNA targets of RBPs in a genome-wide manner. These methods have provided a relatively unbiased, global approach for the derivation of *in vivo* binding sites. In this review we focus on the mechanisms of RBP binding to single-stranded RNA, because RBP binding to double-stranded RNA is not sequence specific. The field is currently in a 'cataloguing phase', generating lists of target RNAs and binding sites for RBPs. In addition, these methods have revealed novel functions for several RBPs, highlighting the importance of RNA–protein interactions in cells. Complementary to the global approach, structural studies are providing crucial information on the mechanistic details of how RBPs contact target RNAs and can resolve discrepancies in RNA-binding sites.

How are RNA–protein interactions determined?

Before the breakthrough of genome-wide methods, especially next-generation sequencing, the study of RNA–protein interactions was limited to *in vitro* assays, such as mutagenesis studies of selected RNA substrates or expression of minigenes or model transgenes in tissue culture cells. Since 1990, numerous techniques have been developed to identify the RNA sequences bound by proteins, both *in vitro* and *in vivo*. Each technique has advantages, disadvantages and caveats due to experimental bias. The key *in vitro*, *in silico* and *in vivo* approaches are summarized in Boxes 1–3. One of the first unbiased approaches to RNA–protein interactions was SELEX (systematic evolution of ligands by exponential enrichment; Box 1) [4], which allows the identification of high-affinity RNA-binding preferences of a given RBP. Because SELEX is performed with purified proteins (or parts of proteins) and synthetic short RNAs, it leaves the question about *in vivo* binding specificity unanswered. How do the identified consensus binding motifs relate to the *in vivo* binding specificity of RBPs? More recently, RNAcompete and next-generation SELEX have been used to evaluate the effect of RNA secondary structure on RBP binding (Box 1) [5,6]. However, *in vitro* methods do not take into account the effect of the complex cellular environment where RNA–protein interactions normally take place. In cells, the sequence with the highest *in vitro* binding affinity may not be the preferred binding site, and weaker interactions may be biologically relevant and even advantageous. For instance, the release of splicing factors from target sites may be just as functionally important as initial binding for the splicing reaction to occur. Furthermore, the contribution of RBP cofactors to RNA-binding affinity is often excluded *in vitro* but is likely to play an important role *in vivo*.

What is the best way to uncover RNA-binding sites and targets *in vivo*? RNA targets and the specific nucleic acids important for binding sometimes emerge through forward

Corresponding authors: Änkö, M-L. (minna-liisa.anko@monash.edu); Neugebauer, K.M. (neugebau@mpi-cbg.de)

* *Current address:* Australian Regenerative Medicine Institute, Monash University, Melbourne, Australia.

0968-0004/$ – see front matter © 2012 Elsevier Ltd. All rights reserved. doi:10.1016/j.tibs.2012.02.005

Author's personal copy

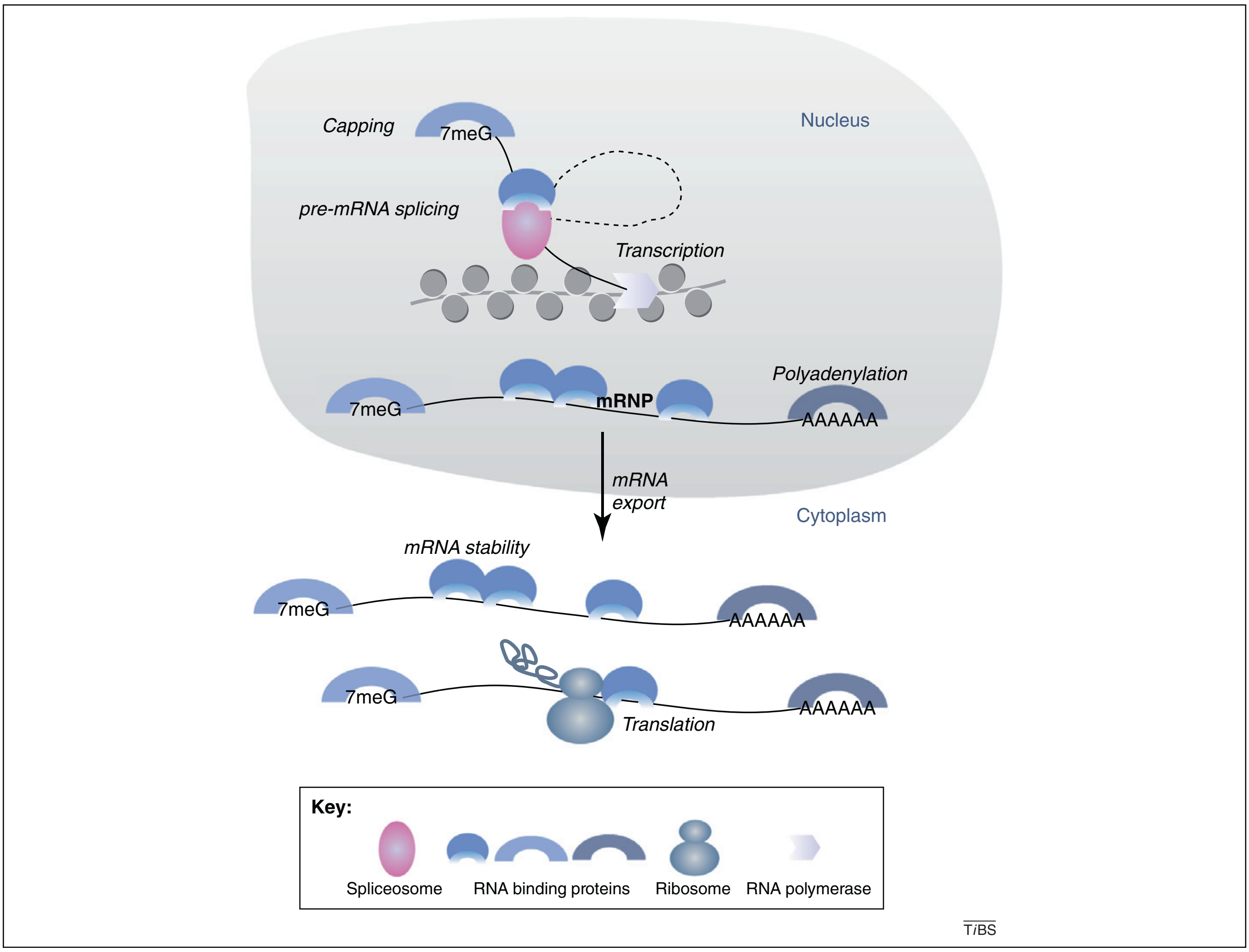

Figure 1. RNA-binding proteins (RBPs) are involved in many steps of gene expression. Cellular RNA is associated with proteins in ribonucleoprotein (RNP) complexes, which comprise most functional forms of RNA. The processes where RBPs can affect the gene expression output are presented in italics.

genetic analysis. For example, mutagenesis and phenotypic screening in *Caenorhabditis elegans* uncovered nucleotides in the *fem-3* mRNA that are required for binding of regulatory factors such as FBF [7,8]. However, this situation is relatively rare and it is a labour-intensive way to identify both the protein and the RNA-binding partners. Instead, knockdown and overexpression of RBPs followed by microarray analysis has emerged as a common strategy for obtaining a global view of endogenous RNA targets. This approach has been used extensively to investigate splicing factors, with the development of splicing-sensitive microarray platforms [9]. However, this type of analysis has the caveat that changes in gene expression levels or in exon inclusion do not necessarily require a direct interaction between the protein investigated and the RNA affected. An application of gene expression microarrays that addresses the RNA–protein interaction more directly is RNA immunoprecipitation followed by microarray analysis (RIP-chip) [10]. RIP-chip provides information about the composition of RNPs but is also not restricted to direct interactions because a protein can be an RNP component through bridging protein–protein interactions. Nevertheless, identification of RNA targets of some RBPs, using either RIP-chip or global profiling after knockdown or overexpression, has led to the derivation of binding sites through motif analysis (Box 2).

The development of *in vivo* crosslinking with UV light, followed by immunoprecipitation (CLIP) and next-generation sequencing, has permitted the identification of direct RNA-binding sites of proteins globally [11–14]. Box 3 explains the principles of CLIP and related methods. Some variants of the method even allow the determination of the binding sites at close to one nucleotide resolution. Although UV crosslinking introduces biases due to differences in crosslinking efficiencies between different nucleotides, the CLIP method has proven to be a powerful tool to determine *in vivo* binding sites and discover functional mechanisms of RBPs. CLIP and other related high-throughput technologies have been reviewed recently [9,15–17].

Structures of RBPs bound to RNA are less broadly available than genome-wide RNA-binding data; yet these studies are indispensable for understanding how proteins interact with the wide variety of different RNA molecules in cells. Structural studies of RBPs bound to RNA molecules focus on the RNA-binding pocket of the protein

Author's personal copy

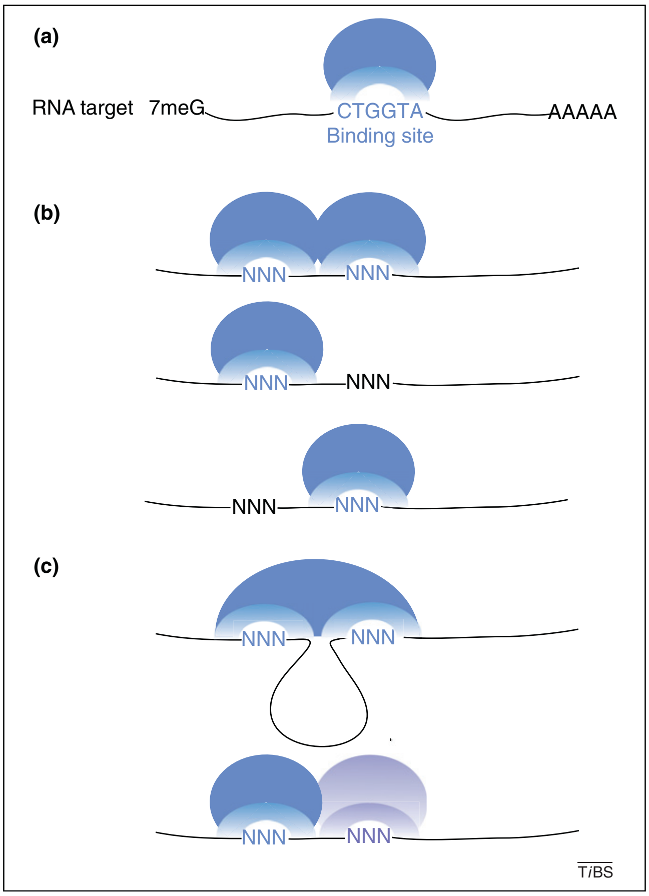

Figure 2. RNA-binding specificity: protein binding to target RNAs and binding sites. **(a)** An mRNA-binding protein (blue) is shown bound to its target mRNA (black line), which is named according to the gene ID. The specific sequence at which the RNA-binding protein (RBP) binds is called the binding site (blue nucleotide sequence). From the analysis of numerous binding sites, a consensus motif can be derived. **(b)** It is commonly seen by crosslinking and immunoprecipitation (CLIP) that the RBP binds some but not all potential binding sites (NNN) that match its consensus sequence, indicating additional determinants of RNA-binding specificity. Alternatively, the protein may be detected by CLIP at many binding sites that map to the same gene; however, it is possible that the RBP is not bound at both sites within the same RNA molecule simultaneously. **(c)** RBP with multiple RNA-binding domains can bring together RNA sequences that are far apart in the primary sequence. In an analogous way, RNA binding to nearby sites in the same transcript can induce RBP multimerization.

domain and reveal the mechanism of interaction in great detail. Such structures complement the genome-wide approaches by providing insight into how empirically determined binding sites relate to protein domain composition and structural requirements for sequence recognition.

RNA-binding domains: common modules for RNA recognition

RBPs interact with RNA through a variety of protein domains. Although single domains can be sufficient for RNA binding, many RBPs have multiple RNA-binding domains (Figure 3). One of the most frequently occurring RNA interaction domains is the RRM, which occurs 487 times in ~20 000 annotated human protein-coding genes [1]. RRMs are present in Ser/Arg-rich (SR) proteins, heterogeneous nuclear RNP (hnRNP) proteins, splicing factors such as Rbfox proteins, Elav proteins and several components of the core spliceosomal machinery, including U2 auxiliary factor 65 (U2AF65) and U1A. In addition to splicing factors, RRMs are found in many other RBPs involved in nuclear processes, such as the poly-A-binding protein (PABP) and the cap-binding protein (CBP20). RRMs can adopt very different binding specificities: CBP20 recognizes a single nucleotide (m(7)GpppG-cap) [18] and PABP binds to a polyA sequence [19]; by contrast, splicing factor consensus motifs are generally degenerate sequences of four to eight nucleotides [20]. The KH domain also binds RNA and occurs 95 times in the human genome [1]; it was first identified in the transcriptional regulator hnRNP K, which binds single-stranded DNA as well as RNA [21]. KH domains are found in many splicing regulators including branchpoint-binding protein SF1, Nova-1 and -2 (onconeural ventral antigen 1 and 2), Sam68 and FMRP, as well as in exosome subunits, and proteins involved in RNA stability and translational control [22]. Similar to RRMs, the target sequences for different KH motifs are variable: SF1 specifically recognizes the conserved branchpoint sequence but notably does not recognize all branchpoints equally [23], Nova proteins bind a four-nucleotide consensus motif, and FMRP may not have a consensus sequence [24]. Zinc fingers are relatively small protein domains that coordinate zinc or other metal ions in the binding pocket. Zinc fingers bind RNA and/or DNA [25] and are emerging as yet another versatile RNA interaction domain that occurs at least 167 times in the human genome [1]. The binding specificity of a zinc finger domain depends on the number of zinc fingers within the domain and the amino acid composition. Examples of RBPs with zinc finger domains are U2AF35, which binds together with U2AF65 at the branchpoint/3′ splice site, the SR protein SRSF7 (9G8), and muscleblind-like splicing factors

Box 1. *In vitro* determination of RNA-binding specificity

SELEX (systematic evolution of ligands by exponential enrichment) is designed to determine the sequence of high-affinity DNA or RNA ligands that bind to a protein or other molecule of interest [4,65]. Beginning with a large oligonucleotide library of random sequences and primers for amplification at the 5′ and 3′ ends, the protein of interest is allowed to bind to the oligonucleotide pool; bound sequences are selected, usually by affinity chromatography or gel shift, and reamplified and reselected through many rounds by PCR. Finally, the resulting pool of bound RNA molecules is sequenced to generate consensus information. The many rounds of selection yield high-affinity binding sites, generally in the nanomolar or subnanomolar range. The caveats to SELEX include: (1) many RBPs are expressed at micromolar concentrations in cells; nanomolar targets are not necessarily the only ones bound by the protein in cells; (2) full-length RNA molecules may have a secondary structure that promotes or excludes protein binding. Recent advances on the original SELEX protocol include the avoidance of constant primer regions that can introduce biases, such as secondary structure formation that influences binding [66]. Importantly, ‘next-generation SELEX’ (which paradoxically does not use next-generation sequencing) and RNAcompete can query both genomically derived sequences and local RNA structures by tiling gene regions or sequences on custom microarrays [5,6]. The microarray is used to generate a library of short-input RNAs, which are subjected to a selection procedure as in standard SELEX. Reprobing of the tiling array gives an apparent binding affinity for all sequences. In terms of extrapolating to cells, the caveats to these latter variations of SELEX is that they are limited to the queried sequences, the only secondary structures evaluated are based on local base-pairing interactions, and the assay still relies on binding purified proteins to the library under *in vitro* conditions.

Author's personal copy

Box 2. RIP approaches followed by *in silico* determination of RNA-binding specificity

A common way of identifying RNA targets of RBPs is immunoprecipitation of the RBP followed by microarray analysis (RIP-chip) or next-generation sequencing (RIP-seq). RIP-based approaches were first developed for polyA-binding protein (PABP) to determine which mRNAs were undergoing active translation [67]. Since then, many RBPs have been subject to RIP analysis for RNA target identification [53,54,56,58,68–70].

From a pool of unaligned sequences obtained by RIP, binding motifs can be derived using MEME (multiple expectation maximization for motif elicitation) algorithms to identify shared motifs [71]. This was particularly successful for members of the Puf family of RBPs in yeast and fruit flies, where it was possible to assemble pools of tens to hundreds of target RNA sequences for each family member; binding motifs predicted by MEME were well validated [68,72]. More recently, mRNA targets of GLD-1 were identified in *Caenorhabditis elegans*, using RIP-chip; binding motifs were similarly derived and validated [70]. These examples make it seem relatively trivial to derive relevant sequence motifs, using prediction. However, size matters: metazoan pre-mRNAs are 10 times longer than mRNAs and increase the complexity of the computational task if, for example, the RBP binds introns. Prediction from pooled sequences puts the burden of proof on validation, because the inference that a binding site occurs somewhere along the transcript is indirect. Indeed, some of the targets identified by RIP may be indirect because of protein–protein interactions within the RNP.

Box 3. *In vivo* determination of RNA-binding specificity

Methods for identifying direct interactions between RNA and protein *in vivo* rely on crosslinking of cells, tissue samples or extracts with UV light [73]. UV light penetrates cells and induces covalent bonds between RNA bases and amino acid side chains at a distance of only several ångströms; the covalent bond permits stringent washing. Unlike formaldehyde, UV light does not induce protein–protein crosslinks, making UV preferable for addressing direct binding. UV crosslinking is relatively inefficient but that can be overcome by using sufficient starting material. In crosslinking and immunoprecipitation (CLIP), fragmentation of the purified RNA permits sequencing of so-called ‘tags’. In the original protocol, fragmented RNA served as the substrate for reverse transcription and concatemerization for Sanger sequencing [58]. To increase the depth of reads, the protocol has been adapted for next-generation sequencing in high-throughput sequencing (HITS)-CLIP [11]. A further advance in cloning strategy and library preparation, called iCLIP, permits identification of the actual crosslink site with a resolution of a few nucleotides [12]. A variation of CLIP, photoactivatable ribonucleoside-enhanced (PAR)-CLIP, uses the incorporation of photoreactive ribonucleoside analogues, such as 4-thiouridine (4-SU) and 6-thioguanosine (6-SG), into RNA transcripts synthesized in living cells [14]. The advantage of PAR-CLIP is that it may increase the efficiency of crosslinking [14,74]. Potential biases in crosslinking to specific amino acids and nucleotides, such as uracil in CLIP or the photoreactive species in PAR-CLIP, may influence detectability and/or the derivation of binding motifs [92]. PAR-CLIP has been a powerful tool for the identification of RNA targets, such as miRNAs bound to their cognate mRNA targets [14]. In this case, maximization of target identification was the aim, not deriving consensus sequence motifs.

It is worth noting that CLIP results will be influenced by the complexity of the RNAs present in the sample. In other words, transcripts not expressed by the particular cells or tissues under study are not queried by CLIP. An annotated gene that is not a target may simply not be expressed or expressed at very low levels. The same transcript may indeed be a target in another cell type.

(Mbnl1-3). In addition to RRMs, KH domains and zinc fingers, RNA binding can be conferred by the SWAP domain in suppressor-of-white-apricot homolog splicing factor [26] and by the PIWI domain in the miRNA processing protein, Ago1 [27]. There are also double-stranded RNA-binding motifs, such as those in the protein Staufen, that mediate binding to double-stranded RNAs [28]. Taken together, numerous protein domains contained in hundreds of genes have the potential to bind RNA, and we may not be fully aware of all of the protein domains that can do so.

Structural insights into RBP–RNA interactions

Genome-wide *in vivo* UV crosslinking studies have identified splicing factor consensus binding sequences that are usually short and degenerate (Table 1). Interestingly, the *in vivo* consensus sequences seem to be very similar to the *in vitro* determined motifs for all factors tested so far, which validates the utility of *in vitro* approaches (Box 1). The degeneracy of consensus sequences raises the question of how the same RNA interaction domain can recognize a set of different yet related sequences. Structural studies help to interpret genome-wide data, by providing clues as to how the various RNA-binding domains physically contact RNA. Structures of RRMs bound to short RNAs have revealed that only a few nucleotides are in fact in direct contact with the protein (see below). Similarly, only four nucleotides fit into the RNA-binding pocket of a KH domain [22]. This implies that the actual consensus sequences may be even shorter than those determined by *in vitro* and genome-wide methods, where analysis parameters tend to search for longer motifs (hexa- to octamers). Furthermore, consensus sequences retrieved from genome-wide data represent the combined binding preference for a large and complex set of RNA targets. Intriguingly, some structural studies of splicing factors bound to RNA indicate that a single RRM domain can utilize a variety of mechanisms to accommodate binding of a large number of different RNA molecules that share only limited sequence similarity [29]. The SR protein SRSF3 (SRp20) has one RRM that binds the consensus sequence CNNC, based on the crystal structure; this is much shorter than the motifs retrieved from *in vitro* and *in vivo* studies (Table 1). Although these four nucleotides are contacted by the RRM, only one nucleotide is specifically recognized (5′ C) [30]. Similarly, the RRM of SRSF2 (SC35) has been shown in several studies to bind very different target sequences (Table 1). A recent report explains this, by showing that RNA can adopt either *syn* or *anti* conformation to fit into to the binding pocket [29]. This is in contrast to other RRM-containing proteins, such as sex lethal (Sxl), HuD and U1A, where the RNA-binding pocket can accommodate several different nucleotides [31–33]. Thus, either the RRM or the RNA can provide the flexibility needed to achieve specificity despite the degeneracy of the consensus sequence.

Both *in vivo* and *in vitro* studies of Nova-1 and -2 have identified repeats of YCAY tetramers as Nova consensus target sequences, which Nova recognizes through its KH domains. The structure of the KH3 domain in a complex with RNA showed that Nova binds YCAY within a loop region of a stem–loop structure [34]. However, typical of KH domain-containing proteins, Nova has three KH domains that can all interact with RNA. The clustering

Author's personal copy

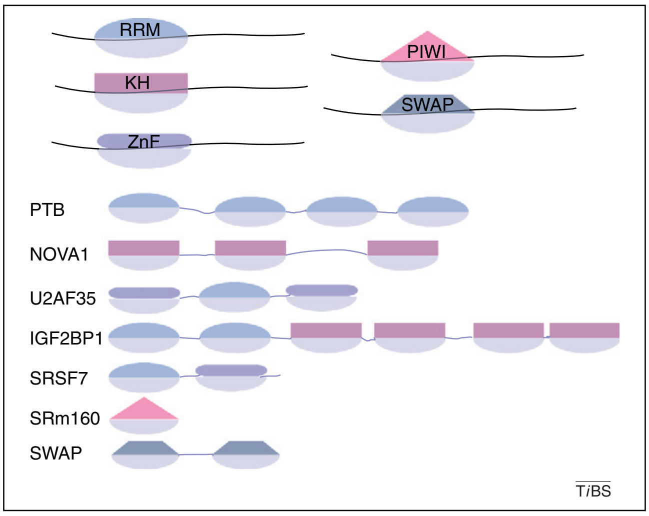

Figure 3. RNA binding can be conferred by a variety of domains. Top panel: RNA-binding proteins (RBPs) interact with RNA through a variety of distinct protein domains. Bottom panel: a single domain is sufficient for RNA binding, but many RBPs have a combination of multiple RNA-binding domains.

of KH domains seems to increase both the affinity and the binding specificity of the protein to RNA [22]. The array of KH domains might explain why the genome-wide studies have identified YCAY repeats surrounding regulated exons, because each of the Nova KH domains could bind to a one of the repeat sites. The structure of the first two KH domains of Nova in a complex with a YCAY repeat RNA was recently solved [35]. Comparison of the Nova KH3 domain and KH1/2 binding specificity within the stem–loop structure showed that, although the YCAY motif is similarly recognized by the KH3 and KH1/2 domains, the RNA hairpin structure contributes to binding affinity and specificity. This difference in the binding specificities of individual domains of Nova is similar to that observed for PTBP1 (also called PTB, hnRNP I), one of the best-studied splicing factors. PTBP1 has four RRMs that bind pyrimidine tracts but contact a slightly different set of nucleotides and have slightly different consensus sequences [36]. Thus, the consensus motifs of multidomain RBPs are averages from several binding domains contacting a variety of RNA molecules (Figure 2b,c). Accordingly, computational analyses generate short, degenerate consensus sequence motifs that occur in the genome frequently [37], but only a fraction of these sites are in fact occupied by an RBP as determined by CLIP [38]. In other words, not every possible consensus sequence motif is bound by an RBP in cells, indicating that additional features contribute to binding specificity. Therefore, computational prediction of binding sites based on sequence information alone remains an extremely difficult task.

RNA interactions in combination

How do RNA-binding motifs adapt to the longer RNA that is the actual cellular target? As mentioned above, many RBPs have clusters of RNA-binding domains, and some have combinations of different types of domains (Figure 3). For instance, IGF2-binding protein 1 (IGF2BP1) has two RRMs and four KH domains. What are the functional consequences of such an array of motifs? What is the contribution of each domain to the function of the RBP? Structural studies indicate that clustering of domains may provide increased binding specificity and/or affinity. Based on CLIP studies, splicing factor binding sites are often clustered, indicating that several relatively low-affinity binding sites could constitute a larger binding surface. However, CLIP data are derived from a population of cells, and thus binding site clusters do not necessarily mean that multiple copies of the RBP or multiple domains of the RBP are bound to a single RNA molecule at the same time (Figure 2b). Another possibility is that binding site clusters reflect flexibility of binding, that is, the RBP may slide between nearby low-affinity binding sites.

Many reports suggest that PTPB1 forms homodimers [39], although it has been shown that free PTBP1 in solution is a monomer [40,41]. Interestingly, PTBP1 CLIP detected both monomeric and dimeric PTBP1 bound to

Table 1. Comparison of RNA-binding specificities determined *in vitro* and *in vivo*

RBP	*In vivo* consensus	Species	Binding site preference	*In vitro* consensus	Reference
SRSF1	GAAGAA	Human	ORF and Intron	RGAAGAAC AGGAC[A/G][G/A]AGC	[38] [75]
SRSF3	CU-rich	Mouse	ORF and Intron	[A/U]C[A/U][A/U]C CTC[T/G]TC[C/T]	M-L Anko (unpublished) [76,77]
SRSF4	GA-rich	Mouse	ORF and Intron	GAAGGA	[78] (M-L Anko, unpublished)
PTBP1	UUCUCU	Human	Intron	UCUU	[39,42,48]
NOVA-1 and -2	YCAY	Mouse	Intron	UCAY	[11,58,79]
RBFOX1	nd	na	na	UGCAUG	[80]
RBFOX2	UGCAUG	Human	Intron	UGCAUG	[47,80]
TDP-43	UG-repeat	Human	Intron	UG-repeat	[49,81]
hnRNP C	U-repeat	Human	Intron	U-repeat	[12,82]
TIA1, TIAL1	UUUA, AUUUU	Human	Intron	(A)U-rich	[51,83]
HuR	U-repeat	Human	Intron	U-repeat	[5,84,85]
SF1	ACUNAC	Human	Intron	UACUAAC	[23,86]
PUM2	UGUANAUA	Human	3′UTR	UGUA(AUC)AUA	[14,87]
QKI	A[C/U]UAAY	Human	Intron	ACUAAY	[14,88]
IGF2BP1-3	CAUH (H=A,C,U)	Human	ORF	nd	[14]
FMRP	nd	Mouse	ORF, 3′UTR	G-quartets	[24,89]
TRA2beta	AGAAGA	Mouse	ORF and Intron	$(GAA)_n$	[90,91]

R, Purine; Y, pyrimidine; nd, not determined; na, not available.

Author's personal copy

RNA [42], indicating that the bound RNA may tether PTBP1 molecules together (Figure 2c). Binding specificity was not affected by dimerization because both monomeric and dimeric PTBP1 had the same consensus sequence motif [42]. Similarly, Nova was found to be a monomer in solution but, through complex inter- and intramolecular interactions between the KH domains, the RNA-bound form may generate higher-order complexes [35]. The zinc finger-containing Mbnl1 protein also shows flexibility in RNA binding by adopting different conformations to recognize RNAs with varying sequence configurations [43]. The cluster of zinc fingers in Mbnl1 has been proposed to cause RNA looping through intra- or intermolecular interaction of different Mbnl1 domains [44]. Recently, a similar looping mechanism was proposed for Nova [35] and PTBP1 [45], in which the tandem binding motifs provide an RNA interaction surface. The intra- and intermolecular interactions may also change the configuration of the RNA and, for instance, loop out an exon or bring two distant sites together to regulate splicing outcome. This evidence indicates that RNA harbouring a pair of sequence elements targeting an RBP may undergo looping (Figure 2c); thus, RNA looping may be a general mechanism for RBPs to bring together distant RNA regions or exclude stretches of RNA.

In addition to forming homodimeric complexes, RBPs can interact with coregulatory proteins and with each other. A classic example is Raver1, a multifunctional protein with three RMMs that can modulate PTBP1 function. According to the proposed model, Raver1 forms bridging interactions between RNA-bound PTBP1 proteins and brings the PTBP1–RNA complexes into a conformation that allows regulation of exon inclusion [46]. An Rbfox2 CLIP study also showed clusters of PTBP1 consensus motifs close to the Rbfox2 binding sites, indicating that these two proteins may either antagonize or cooperate to regulate the same splicing events [47]. Nova1 also interacts with PTBP1, as well as with Rbfox1. Therefore, it is interesting to speculate that RBPs can compete for binding sites, interact by looping and/or form bridged protein–protein interactions to determine splicing outcomes. Similarly, other coregulator proteins might function in assisting complex formation between distantly bound proteins.

The CLIP studies have shown that, in addition to the sequence information, the position of a sequence motif relative to the regulated exon plays an important role. By combining CLIP data with splicing sensitive microarray data and/or RT-PCR validation, it has been possible to construct so-called RNA maps that predict the effect of splicing factor binding on exon inclusion [12,48–51]. Depending on the position of binding relative to the regulated exon, the splicing factor can either enhance or repress inclusion. This type of positional effect has been observed for several factors, including PTBP1 [42,48], Nova-1 [11,50], TIA/TIAR [51], Rbfox2 [47], TDP-43 [49] and hnRNP C [12]. In addition to positional information, other parameters, such as the phosphorylation status, may determine whether the RBP acts as a positive or a negative regulator of exon inclusion [52]. Mechanisms such as RNA looping through intra- and intermolecular interaction may provide mechanistic explanations for the observed positional effects.

Integration of binding with function

It appears that several RBPs are required to cooperate for functional regulation of one RNA molecule. Comprehensive studies comparing multiple RBPs and their targets have provided evidence that the same RNA can be bound and regulated by multiple factors [53,54]. To date, no *in vivo* study has examined all of the proteins that bind and influence the fate of a single mRNA. An *in vitro* characterization by mass spectrometry of all proteins that copurify with a single species of spliced mRNA yielded ~45 proteins [55], showing that individual mRNPs can be expected to contain a diverse set of proteins that can impact the life of mRNA.

A currently more tractable and equally intriguing question is: how many RNAs does one RBP bind and regulate? Because of technical differences, as well differences in model systems and antibodies used, direct comparison between different factors is not yet possible. Based on the RBPs analyzed to date, it appears that most factors have some thousands of binding sites in hundreds of genes (see Table 1 for references). For example, PTBP1-binding clusters were found in almost half of annotated human genes [42]. Whether this reflects true differences in binding sites or just depth of sequencing and whether all the binding sites are functionally significant are still open questions. In the case of PTBP1, ~28% of binding events were associated with alternative splicing. However, this does not mean that the rest would not be functionally significant because PTBP1, along with other RBPs, has been associated with several different steps in gene expression [36,53]. Some of the factors analyzed are ubiquitously expressed (e.g. SR proteins, PTBP1, hnRNP C) whereas others show restricted expression in one or a few tissues (e.g. Nova1 and 2, Rbfox2, TPD-43); however, the number of binding sites does not seem to correlate generally with tissue specificity. Interestingly, RBPs often seem to regulate functionally related sets of genes or even complete pathways [38,53,54,56–58].

Genome-wide analyses have additionally revealed many novel functions for RBPs. For example, splicing factors have been shown to be multifunctional RBPs participating in many steps of gene expression, including miRNA processing and nuclear export [56,59–62]. An important question is whether the interactions of splicing factors with mature mRNA or miRNAs is mechanistically identical to the interactions with pre-mRNA. Binding interactions are likely to be the same, but the mode of recruitment and activity may be different. Again an interesting example is PTBP1: in its function in internal ribosome entry site (IRES)–mediated translation initiation, it acts as an RNA chaperone by helping IRES to achieve its correct conformation [36]. Moreover, U2AF65 undergoes conformational changes in the arrangement of its tandem RRMs in response to the strength of the polypyrimidine tract recognized; the shift from open to closed conformation reflects the nucleotide composition of the polypyrimidine tract and correlates directly with the strength of splicing regulation conferred [63]. Although RNA-binding capacity is addressed by genome-wide studies, such as CLIP, further insights into temporal relationships or the context of different biological processes await the

Author's personal copy

combination of these data with detailed biochemical and structural analyses.

Concluding remarks

In-depth analysis of RNA–protein interactions is crucial to understand how the gene expression output of cells is regulated. Yet hundreds of RBPs are encoded by genomes and only a few have been analyzed. Progress towards identifying *in vivo* RNA targets and binding sites has revealed that *in vitro* approaches to determining binding specificity is largely valid, but consensus binding motifs alone cannot reveal endogenous targets. Furthermore, structural studies have shown that, even though consensus binding sequences are degenerate, single nucleotide mutations can either abolish or create a recognition site for an RBP. Splicing factor–RNA interactions are thus far the best-studied examples of RNA–protein interactions in cells. Although structural and genome-wide studies have deepened our knowledge of how RNA binding relates to splicing regulation, we are still far from a complete understanding of these processes in cells. Efforts to construct a splicing code are underway but further genome-wide data and mechanistic insights are necessary to allow accurate predictions about splicing regulation in a given cell at a given time [64]. The overall aim is to be able to predict the functional consequences of numerous protein–RNA interactions taking place sequentially or simultaneously on a given target RNA. How these models will adapt to the variety of cell types within tissues and organs remains a challenge.

Acknowledgments

We thank Michael Hiller and members of our laboratory for helpful discussions and apologize to those authors whose work we could not cite owing to space limitations. We are grateful for support from the Max Planck Society (to K.M.N.), the Deutsche Forschungsgemeinschaft (NE909/3-1 to K.M.N.), the Helsingin Sanomain Foundation, the Sigrid Juselius Foundation and the Jane and Aatos Erkko Foundation (to M-L.Ä.).

References

1 Letunic, I. *et al.* (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.* 37, D229–D232

2 Cooper, T.A. *et al.* (2009) RNA and disease. *Cell* 136, 777–793

3 Long, J.C. and Caceres, J.F. (2009) The SR protein family of splicing factors: master regulators of gene expression. *Biochem. J.* 417, 15–27

4 Tuerk, C. *et al.* (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249, 505–510

5 Ray, D. *et al.* (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotech.* 27, 667–670

6 Reid, D.C. *et al.* (2009) Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence. *RNA* 15, 2385–2397

7 Ahringer, J. and Kimble, J. (1991) Control of the sperm-oocyte switch in *Caenorhabditis elegans* hermaphrodites by the fem-3 3′ untranslated region. *Nature* 349, 346–348

8 Zhang, B. *et al.* (1997) A conserved RNA-binding protein that regulates sexual fates in the *C. elegans* hermaphrodite germ line. *Nature* 390, 477–484

9 Blencowe, B.J. *et al.* (2009) Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes Dev.* 23, 1379–1386

10 Baroni, T.E. *et al.* (2008) Advances in RIP-chip analysis: RNA-binding protein immunoprecipitation-microarray profiling. *Methods Mol. Biol.* 419, 93–108

11 Licatalosi, D.D. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456, 464–469

12 Konig, J. *et al.* (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* 17, 909–915

13 Ule, J. *et al.* (2005) CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* 37, 376–386

14 Hafner, M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141, 129–141

15 Witten, J.T. and Ule, J. (2011) Understanding splicing regulation through RNA splicing maps. *Trends Genet.* 27, 89–97

16 Licatalosi, D.D. and Darnell, R.B. (2010) RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* 11, 75–87

17 Hafner, M. *et al.* (2010) PAR-CLIP – a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J. Vis. Exp.* 41, pii: 2034

18 Calero, G. *et al.* (2002) Structural basis of m7GpppG binding to the nuclear cap-binding protein complex. *Nat. Struct. Mol. Biol.* 9, 912–917

19 Adam, S.A. *et al.* (1986) mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. *Mol. Cell Biol.* 6, 2932–2943

20 Blencowe, B.J. (2006) Alternative splicing: new insights from global analyses. *Cell* 126, 37–47

21 Matunis, M.J. *et al.* (1992) Characterization and primary structure of the poly(C)-binding heterogeneous nuclear ribonucleoprotein complex K protein. *Mol. Cell Biol.* 12, 164–171

22 Valverde, R. *et al.* (2008) Structure and function of KH domains. *FEBS J.* 275, 2712–2726

23 Corioni, M. *et al.* (2011) Analysis of in situ pre-mRNA targets of human splicing factor SF1 reveals a function in alternative splicing. *Nucleic Acids Res.* 39, 1868–1879

24 Darnell, J.C. *et al.* (2011) FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* 146, 247–261

25 Traci, M. and Tanaka, H. (2005) Multiple modes of RNA recognition by zinc finger proteins. *Curr. Opin. Struct. Biol.* 15, 367–373

26 Denhez, F. and Lafyatis, R. (1994) Conservation of regulated alternative splicing and identification of functional domains in vertebrate homologs to the *Drosophila* splicing regulator, suppressor-of-white-apricot. *J. Biol. Chem.* 269, 16170–16179

27 Song, J-J. *et al.* (2004) Crystal structure of argonaute and its implications for RISC slicer activity. *Science* 305, 1434–1437

28 Barber, G.N. (2009) The NFARs (nuclear factors associated with dsRNA): evolutionarily conserved members of the dsRNA binding protein family. *RNA Biol.* 6, 35–39

29 Daubner, G.M. *et al.* (2011) A syn-anti conformational difference allows SRSF2 to recognize guanines and cytosines equally well. *EMBO J.* 31, 162–174

30 Hargous, Y. *et al.* (2006) Molecular basis of RNA recognition and TAP binding by the SR proteins SRp20 and 9G8. *EMBO J.* 25, 5126–5137

31 Jovine, L. *et al.* (1996) Two structurally different RNA molecules are bound by the spliceosomal protein U1A using the same recognition strategy. *Structure* 4, 621–631

32 Wang, X. and Tanaka Hall, T.M. (2001) Structural basis for recognition of AU-rich element RNA by the HuD protein. *Nat. Struct. Mol. Biol.* 8, 141–145

33 Handa, N. *et al.* (1999) Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. *Nature* 398, 579–585

34 Lewis, H.A. *et al.* (1999) Crystal structures of Nova-1 and Nova-2 K-homology RNA-binding domains. *Structure* 7, 191–203

35 Teplova, M. *et al.* (2011) Protein-RNA and protein-protein recognition by dual KH1/2 domains of the neuronal splicing factor Nova-1. *Structure* 19, 930–944

36 Auweter, S. and Allain, F. (2008) Structure-function relationships of the polypyrimidine tract binding protein. *Cell Mol. Life Sci.* 65, 516–527

37 Wang, J. *et al.* (2005) Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Res.* 33, 5053–5062

38 Sanford, J.R. *et al.* (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.* 19, 381–394

Author's personal copy

39 Perez, I. *et al.* (1997) Multiple RRMs contribute to RNA binding specificity and affinity for polypyrimidine tract binding protein. *Biochemistry* 36, 11881–11890

40 Monie, T.P. *et al.* (2005) The polypyrimidine tract binding protein is a monomer. *RNA* 11, 1803–1808

41 Petoukhov, M.V. *et al.* (2006) Conformation of polypyrimidine tract binding protein in solution. *Structure* 14, 1021–1027

42 Xue, Y. *et al.* (2009) Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol. Cell* 36, 996–1006

43 Cass, D. *et al.* (2011) The four Zn fingers of MBNL1 provide a flexible platform for recognition of its RNA binding elements. *BMC Mol. Biol.* 12, 20

44 Teplova, M. and Patel, D.J. (2008) Structural insights into RNA recognition by the alternative-splicing regulator muscleblind-like MBNL1. *Nat. Struct. Mol. Biol.* 15, 1343–1351

45 Lamichhane, R. *et al.* (2010) RNA looping by PTB: evidence using FRET and NMR spectroscopy for a role in splicing repression. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4105–4110

46 Rideau, A.P. *et al.* (2006) A peptide motif in Raver1 mediates splicing repression by interaction with the PTB RRM2 domain. *Nat. Struct. Mol. Biol.* 13, 839–848

47 Yeo, G.W. *et al.* (2009) An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat. Struct. Mol. Biol.* 16, 130–137

48 Llorian, M. *et al.* (2010) Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nat. Struct. Mol. Biol.* 17, 1114–1123

49 Tollervey, J.R. *et al.* (2011) Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci.* 14, 452–458

50 Ule, J. *et al.* (2006) An RNA map predicting Nova-dependent splicing regulation. *Nature* 444, 580–586

51 Wang, Z. *et al.* (2010) iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol.* 8, e1000530

52 Feng, Y. *et al.* (2008) Phosphorylation switches the general splicing repressor SRp38 to a sequence-specific activator. *Nat. Struct. Mol. Biol.* 15, 1040–1048

53 Hogan, D.J. *et al.* (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol.* 6, e255

54 Änkö, M-L. *et al.* (2010) Global analysis reveals SRp20- and SRp75-specific mRNPs in cycling and neural cells. *Nat. Struct. Mol. Biol.* 17, 962–970

55 Merz, C. *et al.* (2007) Protein composition of human mRNPs spliced in vitro and differential requirements for mRNP protein recruitment. *RNA* 13, 116–128

56 Gama-Carvalho, M. *et al.* (2006) Genome-wide identification of functionally distinct subsets of cellular mRNAs associated with two nucleocytoplasmic-shuttling mammalian splicing factors. *Genome Biol.* 7, R113

57 Ule, J. *et al.* (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science* 302, 1212–1215

58 Keene, J.D. and Tenenbaum, S.A. (2002) Eukaryotic mRNPs may represent posttranscriptional operons. *Mol. Cell* 9, 1161–1167

59 Zhong, X-Y. *et al.* (2009) SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. *Mol. Cell* 35, 1–10

60 Blanchette, M. *et al.* (2004) Genome-wide analysis reveals an unexpected function for the *Drosophila* splicing factor U2AF50 in the nuclear export of intronless mRNAs. *Mol. Cell* 14, 775–786

61 Michlewski, G. and Caceres, J.F. (2010) Antagonistic role of hnRNP A1 and KSRP in the regulation of let-7a biogenesis. *Nat. Struct. Mol. Biol.* 17, 1011–1018

62 Sawicka, K. *et al.* (2008) Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem. Soc. Trans.* 36, 641–647

63 Mackereth, C.D. *et al.* (2011) Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature* 475, 408–411

64 Barash, Y. *et al.* (2010) Deciphering the splicing code. *Nature* 465, 53–59

65 Ellington, A.D. and Szostak, J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature* 346, 818–822

66 Jarosch, F. *et al.* (2006) In vitro selection using a dual RNA library that allows primerless selection. *Nucleic Acids Res.* 34, e86

67 Tenenbaum, S.A. *et al.* (2000) Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc. Natl. Acad. Sci. U.S.A.* 97, 14085–14090

68 Gerber, A.P. *et al.* (2006) Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U.S.A.* 103, 4487–4492

69 Hieronymus, H. and Silver, P.A. (2003) Genome-wide analysis of RNA-protein interactions illustrates specificity of the mRNA export machinery. *Nat. Genet.* 33, 155–161

70 Wright, J.E. *et al.* (2011) A quantitative RNA code for mRNA target selection by the germline fate determinant GLD-1. *EMBO J.* 30, 533–545

71 Bailey, T.L. *et al.* (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369–W373

72 Gerber, A.P. *et al.* (2004) Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol.* 2, E79

73 König, J. *et al.* (2012) Protein-RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet.* 13, 77–83

74 Kishore, S. *et al.* (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods* 8, 559–564

75 Tacke, R. and Manley, J.L. (1995) The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *EMBO J.* 14, 3540–3551

76 Cavaloc, Y. *et al.* (1999) The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA* 5, 468–483

77 Schaal, T.D. and Maniatis, T. (1999) Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. *Mol. Cell. Biol.* 19, 1705–1719

78 Zheng, Z.M. *et al.* (1997) Structural, functional, and protein binding analyses of bovine papillomavirus type 1 exonic splicing enhancers. *J. Virol.* 71, 9096–9107

79 Jensen, K.B. *et al.* (2000) The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain. *Proc. Natl. Acad. Sci. U.S.A.* 97, 5740–5745

80 Auweter, S.D. *et al.* (2006) Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. *EMBO J.* 25, 163–173

81 Buratti, E. and Baralle, F.E. (2001) Characterization and functional implications of the RNA binding properties of nuclear factor TDP-43, a novel splicing regulator of CFTR Exon 9. *J. Biol. Chem.* 276, 36337–36343

82 Gorlach, M. *et al.* (1994) The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein C proteins. *J. Biol. Chem.* 269, 23074–23078

83 Le Guiner, C. *et al.* (2001) TIA-1 and TIAR activate splicing of alternative exons with weak 5′ splice sites followed by a U-rich stretch on their own pre-mRNAs. *J. Biol. Chem.* 276, 40638–40646

84 Lebedeva, S. *et al.* (2011) Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol. Cell* 43, 340–352

85 Mukherjee, N. *et al.* (2011) Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell* 43, 327–339

86 Berglund, J.A. *et al.* (1998) A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. *Genes Dev.* 12, 858–867

87 Galgano, A. *et al.* (2008) Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system. *PLoS ONE* 3, e3164

88 Galarneau, A. and Richard, S. (2005) Target RNA motif and target mRNAs of the Quaking STAR protein. *Nat. Struct. Mol. Biol.* 12, 691–698

89 Darnell, J.C. *et al.* (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell* 107, 489–499

90 Grellscheid, S. *et al.* (2011) Identification of evolutionarily conserved exons as regulated targets for the splicing activator Tra2β in development. *PLoS Genet.* 7, e1002390

91 Tacke, R. *et al.* (1998) Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing. *Cell* 93, 139–148

92 Kramer, K. *et al.* (2010) Mass spectrometric analysis of proteins cross-linked to 4-thio-uracil- and 5-bromo-uracil-substituted RNA. *Int. J. Mass Spectrosc.* 304, 184–194