

# A Step-by-Step Guide to Study Protein–RNA Interactions

Antoine Cléry<sup>\*ab</sup>, Laurent Gillioz<sup>a</sup>, Cristina K. X. Nguyen<sup>a</sup>, and Frédéric H.-T. Allain<sup>\*a</sup>

**Abstract:** Protein–RNA complex formation is at the center of RNA metabolism and leads to the modulation of protein and RNA functions. We propose here a step-by-step guide to investigate these interactions including the identification of the protein and RNA parts involved in complex formation, the determination of the affinity of the complex and the characterization of the protein–RNA interface at amino acid and nucleotide level. Moreover, we briefly review the methods that are the most often used to obtain this information using primarily examples from our lab and finally mention what we perceive as the next challenges in the field.

**Keywords:** Affinity · Protein–RNA complex · Protein–RNA interface · RBD · RBP



Prof. **Frédéric Allain** received his PhD in structural biology in 1996 in the lab of Prof. Gabriele Varani (MRC-LMB, Cambridge) before moving to the US for a post-doc in the lab of Prof. Juli Feigon and as a research associate in the lab of Prof. Doug Black (University of California Los Angeles). In 2001, he became an assistant professor at ETH Zurich and received tenure in 2007. Dr. **Antoine Cléry** received his PhD in molecular biology in 2006 in the lab of Dr. Christiane Branlant (Université de Lorraine, Nancy), joined the Allain's lab for a post doc in structural biology in 2007 and is now a permanent senior scientist of the lab. **Cristina Nguyen** and **Laurent Gillioz** are two PhD students of the lab. The Allain lab is specialized in the structure determination of protein–RNA complexes using Nuclear Magnetic Resonance (NMR) to understand their involvement in RNA metabolism.

## 1. Introduction

RNA binding proteins (RBPs) are at the center of the RNA world in cells. By interacting with RNA, RBPs control all RNA metabolism steps, *e.g.* mRNA localization, stability, degradation, splicing, editing, translation, long and small non-coding RNA processing.<sup>[1]</sup> It is therefore not surprising that many connections to diseases like cancer, neurological disorders or muscular atrophies have been reported with loss, mutated or overexpressed RBPs.<sup>[2]</sup> RBPs can bind ssRNA or dsRNA, specifically or non-specifically, using one or several RNA binding domains (RBD). Understanding the exact mode of interaction of these RBPs with RNA is therefore absolutely required to characterize their functions and possibly find new therapeutic strategies against RBP-causing diseases. Although recent methods emerged to study RBP binding to RNA *in vivo*,<sup>[3]</sup> RNA binding sequences identified *in vitro* with

approaches like SELEX, RBNS, RNAcompete or RIP-Chip and modes of protein–RNA interactions determined with structures are often in good agreement with the data obtained in cells. We will briefly describe these techniques herein.

## 2. What is an RNA Binding Protein?

### 2.1 Identification of an RBP

An RBP is a protein that interacts with RNA. The interaction of the protein of interest has then to be tested with RNA (a putative binding site or a degenerated RNA sequence) using one of the methods described later in section 4 (*e.g.* Nucleic Magnetic Resonance, Electrophoretic Mobility Shift Assay or Isothermal Titration Calorimetry). An *in silico* analysis of the primary sequence of the protein is often initially performed to identify putative RNA binding motifs (RBMs) or protein sub-domains that can then be expressed in isolation to test their potential interaction with RNA using the same methods. Vice versa, some RBPs are identified based on the presence of a putative RBD and later characterized experimentally for binding RNA. More recently, UV cross-linking followed by mass-spectrometry allowed the identification of many new RBPs, a great majority of them not containing known RBDs (Fig. 1A).<sup>[4]</sup> Once an interaction has been identified, the next step consists in identifying the part of the protein that is responsible for this interaction.

### 2.2 The RNA-binding Domains

Apart from protein disordered regions that can be involved in non-specific RNA interactions,<sup>[5]</sup> several RBDs with a characteristic fold have been identified so far (Fig. 1B,C). Among RBDs interacting with ssRNA molecules, the so-called RNA Recognition Motif (RRM) is the most frequent RBD found in humans. RBMs are about 90 amino acids long and are characterized by a canonical fold made of four anti-parallel  $\beta$ -strands forming a  $\beta$ -sheet surface that is packed against two  $\alpha$ -helices.<sup>[6]</sup> They usually recognize 2 to 8 nucleotides using primarily the  $\beta$ -sheet, which contains on its surface several conserved aromatic residues interacting *via* stacking with RNA bases. However, additional parts of the domain can also interact with RNA, *i.e.*  $\alpha$ -helix, loops, N- and C-termini (*i.e.* hnRNP F qRRM1 and SRSF1 pseudo-RRM in Fig. 1C).<sup>[7]</sup> Zinc-

\*Correspondence: Dr. A. Cléry, Prof. F. H.-T. Allain  
E-mail: aclery@mol.biol.ethz.ch; allain@mol.biol.ethz.ch

<sup>a</sup>Institute for Molecular Biology and Biophysics, Department of Biology, ETH Zürich, CH-8093 Zürich; <sup>b</sup>Biomolecular NMR spectroscopy platform, ETH Zurich, Zurich

fingers are the second most frequent RBDs. They adopt a variety of folds pinned by a Zn<sup>2+</sup> ion. They typically recognize 2 to 3 nucleotides *via* H-bonds involving primarily the protein backbone and the Watson-Crick edge of the nucleobases (*i.e.* the FUS zinc-finger in Fig. 1C). The hnRNP K homology (KH) domains form the third main family of RBDs interacting with ssRNA. They are approximately 70 amino acids long and can adopt two folds. The type I has a βααββα topology and is characterized by a β-sheet composed of three anti-parallel β-strands packed against three α-helices. A RNA-binding cleft is formed by α1 and α2 helices connected by a ‘GXXG loop’ containing the (I/L/V)-I-G-X-X-G-X-X-(I/L/V) conserved motif, the β2-strand and the β2-β3 loop, which is variable in length. The KH type II fold differs from the type I by a αββααβ topology and a characteristic β-sheet in which the central strand (β2) is parallel to β3 and anti-parallel to β1. KH domains can be extended allowing the binding of additional nucleotides as seen with *C.elegans* GLD-1 (Fig. 1C).

A second category of RBDs is represented by the dsRBDs, which contain approximately 70 amino acids and exhibit a conserved αβββα fold. These domains interact along one face of a regular A-form helix structure and can cover up to 16 bp spanning two consecutive minor grooves separated by a major groove. In most of the cases, dsRBDs use residues from the α1 helix and β1-β2 loop to contact the minor grooves and the N-terminus of the α2-helix with the preceding loop to bind the major groove.<sup>[8]</sup>

These domains were first described as recognizing preferentially the double-stranded RNA shape *via* interactions with the 2'-OH groups of RNA riboses. However, the structure of ADAR2 dsRBDs bound to RNA revealed that each domain binds to a single register with a few sequence-specific contacts in the minor grooves (Fig. 1C).<sup>[9]</sup>

Typically, all these individual domains bind RNA with a nano- to low micromolar affinity. However, they are often present in several copies in proteins either as multiple of the same (*i.e.* PTB with four RRM) or as a combination of several types (*i.e.* FUS with RRM, Zinc-finger and RGG motifs) (Fig. 1B), which contributes to increase the affinity and the specificity of interaction of RBPs with RNA.<sup>[6]</sup> To this list of canonical RBDs can be added the unconventional RBDs SLBP and NHL, which were also proposed to bind specifically to RNA.<sup>[10]</sup> The minimal region of the RBPs that is required for RNA binding is often identified by investigating the RNA binding ability of different truncated versions of the protein of interest (see below).

### 3. How to Identify the Sequence Bound by a RBP?

#### 3.1 *In vitro*

Once the part of the protein involved in RNA binding has been identified, it is possible to investigate whether the RBD interacts specifically with RNA and possibly identify the motif that is rec-

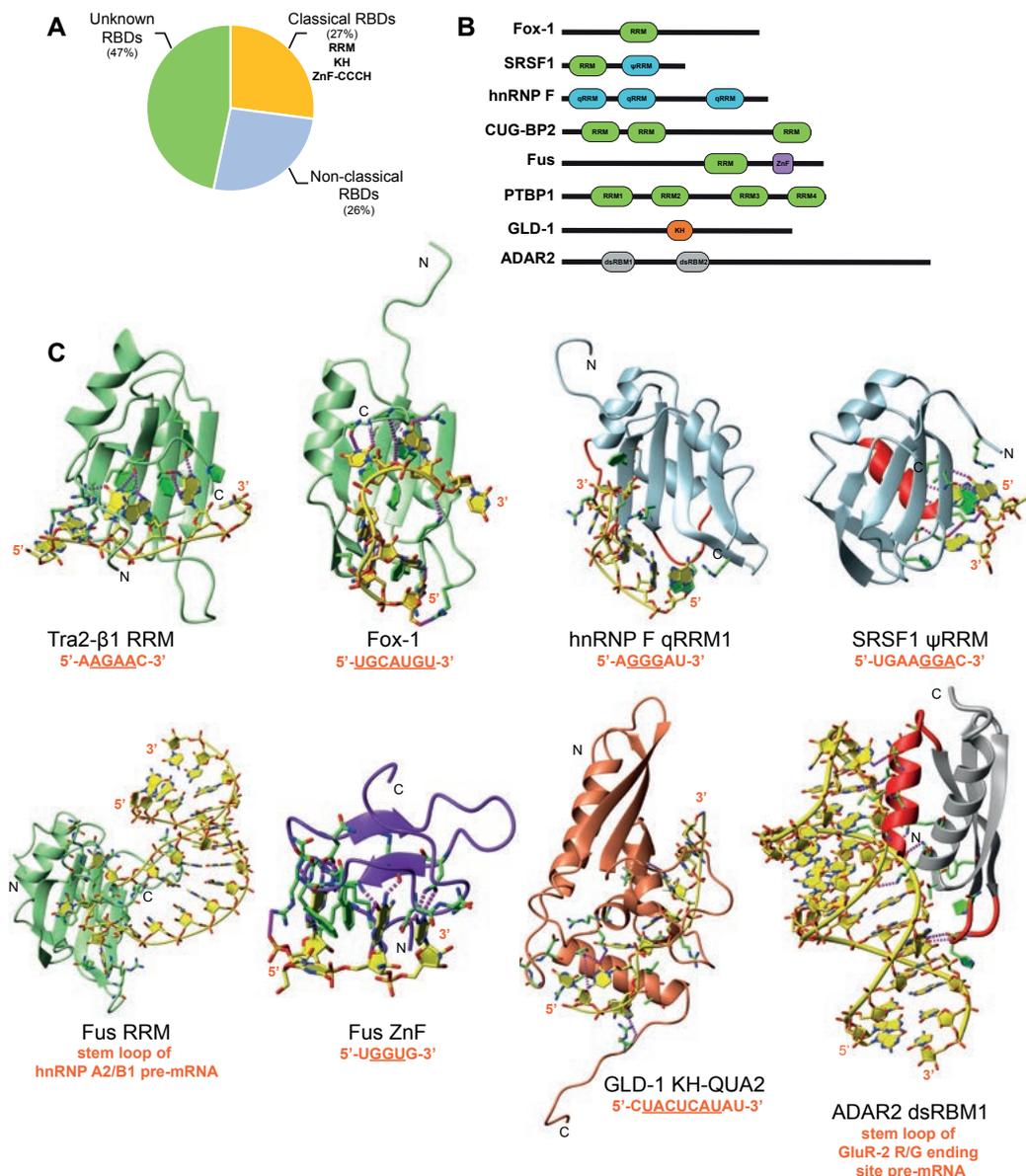


Fig. 1. Representative examples of RBPs and their mode of interaction with RNA. (A) Graph representing the percentage of Classical RBDs (yellow), Non-classical RBDs (blue) and Unknown RBDs (green) identified in the interactome capture performed in HeLa cells by Castello and collaborators.<sup>[36]</sup> (B) Schematic representation of RBPs studied in our lab harbouring one or several RBDs (RRM, ψRRM, qRRM, ZnF, KH and dsRBM). (C) Structures illustrating different modes of interaction of RBDs with RNA: the canonical mode of interaction of Tra2-β1<sup>[34a]</sup> Fox-1<sup>[37]</sup> and Fus RRM<sup>[38]</sup> with RNA, the non-conventional binding mode of quasi-RRM1 of hnRNP F (qRRM1)<sup>[39]</sup> and the pseudo-RRM of SRSF1 (ψRRM)<sup>[34b]</sup> and the complexes formed with FUS ZnF,<sup>[38]</sup> GLD-1 KH-QUA2,<sup>[40]</sup> ADAR2 dsRBM<sup>[9]</sup> and RNA. The structure of the complex is shown in ribbon (protein backbone) and stick (RNA) representation. The protein backbone is shown and heavy atoms are shown in orange (P atoms), yellow (C atoms for RNA), green (C atoms for protein), red (O atoms) and blue (N atoms). Hydrogen bonds are represented by purple dashed lines.

ognized by the domain. If no RNA target was predicted for the protein of interest, its interaction can be first tested with polyA, polyC, polyG or polyU sequences to investigate whether a clear sequence preference can be observed. If one nucleotide is preferentially recognized, it can be used as a starting point to then investigate the preferred nucleotide on each side using a modified version of the scaffold-independent analysis developed in the Ramos lab.<sup>[11]</sup> A RBD binds in general between 2 and 8 nucleotides. Therefore, a length of 8 nucleotides can typically be used. For example, if a cytosine is preferred, the binding of the protein to NNNCANN, NNNCGNN, NNNCUNN, NNNCCNN (N is for A, C, G or U) RNA molecules will be tested to identify the nucleotide that is preferentially bound at the n+1 position using NMR (described in section 4). Then the position n-1 can be investigated and one can continue the sequential selection at each position of the sequence until no apparent preference is observed anymore. To use this approach, the protein has to be <sup>15</sup>N-labeled and concentrated to at least 0.1 mM. We used such an approach to characterize the sequence-specificity of the YTH RNA binding domain.<sup>[12]</sup>

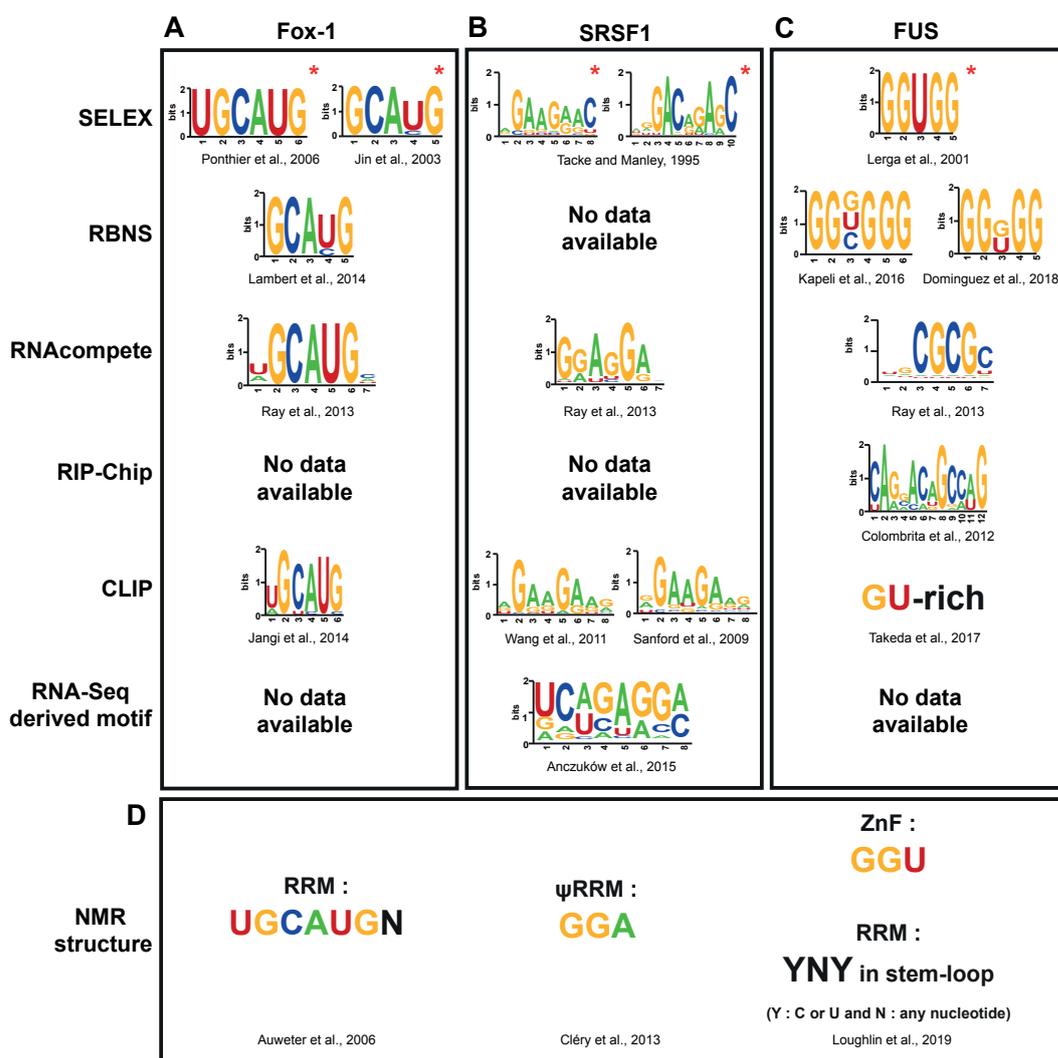
Another strategy consists in using a pool of completely randomized RNA sequences in which each sequence combination is present in multiple copies and selecting those having the highest affinity for the protein of interest. This method is called SELEX for Systematic Evolution of Ligands by EXponential enrichment.<sup>[13]</sup> The protein is first incubated with the pool of random RNA sequences and then immobilized by affinity chromatography. RNAs that are bound to the protein are extracted and converted to DNA by RT-PCR using a primer containing the T7 promoter sequence.

These DNA matrices can then be used for *in vitro* transcription using the T7 RNA polymerase to generate a new pool of RNAs having some affinity for the protein. The protein is then again incubated with these RNAs and by competition those having the highest affinity will bind preferentially to the protein. By performing successive cycles of selection, the pool of selected RNAs will progressively be enriched in RNAs having a high affinity for the protein. Typically, after 5 to 10 cycles of selection, the selected RNAs are converted to DNA and after HT-sequencing a consensus sequence is determined. SELEX consensus sequence for several RBPs like Fox-1, SRSF1 or FUS are shown in Fig. 2.<sup>[14]</sup>

Another approach called RNA bind-n-seq (RBNS)<sup>[15]</sup> consists in mixing three different concentrations of GST-tagged RBP ranging from low nM to low  $\mu$ M to a pool of random RNA sequences (40 nucleotides in length) to enable the detection of RNA motifs having strong, medium and weak but still significant affinity for the protein of interest. After pull-down of the selected RNAs on streptavidin beads and HT-sequencing, a consensus binding motif is derived from these experiments. In addition, the use of 40 nucleotide-long RNA sequences allows them to adopt secondary structures that can be predicted by thermodynamic RNA folding. It allows the user to investigate the effect of RNA secondary structure on protein binding. RBP binding consensus motifs identified with Fox-1 and FUS are shown in Fig. 2. The advantage of RNA bind-n-seq over SELEX is that a larger spectrum of sequences can be detected by this high throughput sequencing method.

Finally, a last method called RNAcompete<sup>[10,16]</sup> was also extensively tested on different systems and consists in mixing the RBP of interest with a pool of non-randomized RNA sequences (about

Fig. 2. Comparison of the consensus binding motifs identified for human Fox-1, SRSF1 and FUS using different methods. (A) A very conserved UGCAUG consensus sequence was systematically identified with FOX-1.<sup>[14a,c,15,41]</sup> (B) Slightly different GA-rich motifs were identified with SRSF1 depending on the method used.<sup>[14d,42]</sup> (C) Highly degenerated GU-rich motifs were selected with FUS with the different approaches.<sup>[14b,43]</sup> (D) NMR allows the identification of the minimal and exact motif bound by each RBD.<sup>[34b,37,38]</sup> \*motif generated from published dataset with <https://weblogo.berkeley.edu/>



40 nucleotides in length). Interacting RNAs are selected by pull-down of the tagged protein, labeled with Cy3 or Cy5 and identified by hybridization on a custom microarray. Bioinformatics is then used to derive a 7-mer RNA binding sequence consensus. This approach has similar advantages compared to the RNA bind-n-seq (RBNS) approach in terms of spectrum of sequences (see Fig. 2). Although RNAcompete is clearly more direct as sequencing is not needed, this method requires access to custom-made arrays.

All these methods have been shown to work efficiently on large-scale studies. However, the last three methods described above often result in a consensus sequence containing a repetition of the motif bound by the RBD, especially if the motif is small (2–4 nucleotides). Indeed, multiplicity of the same motif leads to an increase of the global affinity of the RBD for the RNA molecule, the so-called avidity effect. A clear advantage of the NMR-based approach is the possibility to identify the real minimal motif bound by the domain of interest. If the RBD binds multiple sites on the same RNA molecule, severe line broadening of the RNA resonances will be observed due to chemical exchange while the surface of interaction of the protein will not be enlarged compared to a smaller RNA. We experienced several examples of such effects in the past with PTB, CUG-BP2 or hnRNP C.<sup>[17]</sup>

### 3.2 *In vivo*

An alternative to these methods, especially when the protein cannot be produced *in vitro*, consists in identifying natural targets bound by the protein in cells. This does not always yield RNAs having the highest affinity for the protein and therefore leads to some difficulties in the identification of a clear consensus binding motif.

RNA immunoprecipitation (RIP)<sup>[18]</sup> was developed from a cell extract containing intact RNPs and consists in purifying the RNA bound to the protein of interest by affinity chromatography using specific antibodies or a tagged version of the protein. After several washing steps, the interacting RNAs are identified by hybridization to microarrays (RIP-chip)<sup>[19]</sup> or high-throughput sequencing (RIP-seq).<sup>[20]</sup> This method works in native conditions, however interactions between RNA and protein are often not strong enough to be conserved up to the end of this protocol. Therefore, an alternative is to use ‘crosslinking and immunoprecipitation’ (CLIP).<sup>[21]</sup> It includes a UV crosslinking step at 254 nm directly in living cells or tissues to covalently link the protein and the RNA before the lysis of the cells. Importantly, a direct contact between an amino acid and a nucleobase is required to obtain this cross-link. RNases are then added to the extract to obtain short RNA fragments and disrupt RNA-mediated protein–protein interactions. The goal of this method is to keep only the RNA parts that interact with the protein of interest. The immunoprecipitation of these protein–RNA fragments can then be performed and short RNA adaptors are ligated to the selected RNAs to allow their RT-PCR amplification. The amplified products are finally purified by electrophoresis, extracted and subjected to HT-sequencing. Although several variations of this protocol were developed, the main principle stays the same.<sup>[3]</sup> CLIP was successfully used to identify the RNA targets of many RBPs (Fig. 2). However, it sometimes leads to data of variable specificity depending on the difference in stringencies of RBP purification and on the stability of the RBP–RNA complexes.<sup>[6]</sup> In addition, this approach does not allow the identification of motifs bound by each RBD of multiple RBD-containing protein or the characterization of a network of several proteins interacting on a single RNA molecule. Therefore, the best approach still consists in combining *in vivo* and *in vitro* methods to decipher as solidly as possible the binding consensus sequence, which can be ultimately understood at the atomic level when the structure of a complex has been determined (Fig. 1C). As shown in Fig. 2, a very similar consensus binding sequence was obtained using different approaches with Fox-1, which contains a single RRM. Nevertheless, this convergence was not observed

with RBPs containing multiple RBDs like SRSF1 or FUS. One should therefore be cautious to not overinterpret RNA consensus sequences derived from these different methods.

## 4. How to Measure the Affinity of a RBP for RNA and Determine the Stoichiometry of Interaction?

### 4.1 Label-free Methods

As soon as the RBD and its associated targeted RNA motif have been identified, a measure of the affinity of this interaction becomes possible. Several methods can be used to this end. One of the most common and robust methods is Isothermal Titration Calorimetry (ITC).<sup>[22]</sup> The RNA or protein is gradually injected into the sample cell that contains the protein or RNA molecule, respectively, and the heat that is either absorbed (endothermic) or released (exothermic) upon interaction is measured by a microcalorimeter until the binding reaction reaches equilibrium. The quantity of heat that is measured is proportional to the binding energy and after data processing provides information on the dissociation constant ( $K_D$ ), stoichiometry, enthalpy and entropy of the complex formation (Fig. 3).<sup>[23]</sup>

Contrary to ITC that works with the two interacting partners in solution, surface plasmon resonance (SPR)<sup>[24]</sup> requires the immobilization of one of the two biomolecules on a gold support using either a biotinylated RNA or a tagged-protein; the other binding partner being injected in solution and passed over the surface through a series of flow cells. A polarized light is directed against the support and the angle of the reflected light is measured upon protein–RNA interactions. This angle depends on changes in association and dissociation of these biomolecules and allows the measurements of  $k_{on}$  and  $k_{off}$  kinetics values, from which one can derive the  $K_D$  of the complex (Fig. 3).

A similar approach called switchSENSE has recently been developed and was also shown to allow study of protein–RNA interactions.<sup>[25]</sup> Single-stranded DNA molecules are attached on a gold surface electrode and an alternative current will induce their oscillation. Long RNA sequences (around 60 nucleotides) containing in the 3'-extremity a sequence fully complementary to the 48 nucleotides ssDNA attached to the electrodes and in the 5'-part the protein binding motif, can be transcribed *in vitro*. These RNAs are then hybridized to the ssDNAs by base complementarity and subjected to an oscillation for which the speed will depend on the size and dynamic radius of the complex. Alternatively, proteins can be covalently attached to a ssDNA complementary to the DNA attached on the electrode and hybridized to the chip using the same strategy. As for SPR, this technology allows the determination of  $k_{on}$ ,  $k_{off}$  and  $K_D$  values. However, one of the advantages of the switchSENSE approach consists in the possibility to obtain these kinetics parameters using two independent strategies. Indeed, a fluorescent molecule is present at the 3'-extremity of the ssDNA attached to the electrode and its signal can either decrease by quenching or be enhanced upon binding of the protein to the hybridized RNA molecule. Therefore, kinetics parameters can either be determined by following the switching speed of the ssDNA–RNA duplexes or the variation in fluorescence intensity upon protein binding. This offers more flexibility on the different systems that can be used with this method (*e.g.* small molecule–protein/RNA, protein–protein, protein–RNA, small or large complexes). Moreover, contrary to SPR, switchSENSE allows the estimation of the hydrodynamic radius of biomolecules, which can be used to follow for example the multimerization of a protein on multiple binding sites.<sup>[26]</sup> Finally, conformational changes can also be observed upon binding (Fig. 3).

### 4.2 Methods Requiring Biomolecule Labeling

One of the most popular and oldest methods is the electrophoretic mobility shift assay (EMSA).<sup>[27]</sup> It consists in the vis-

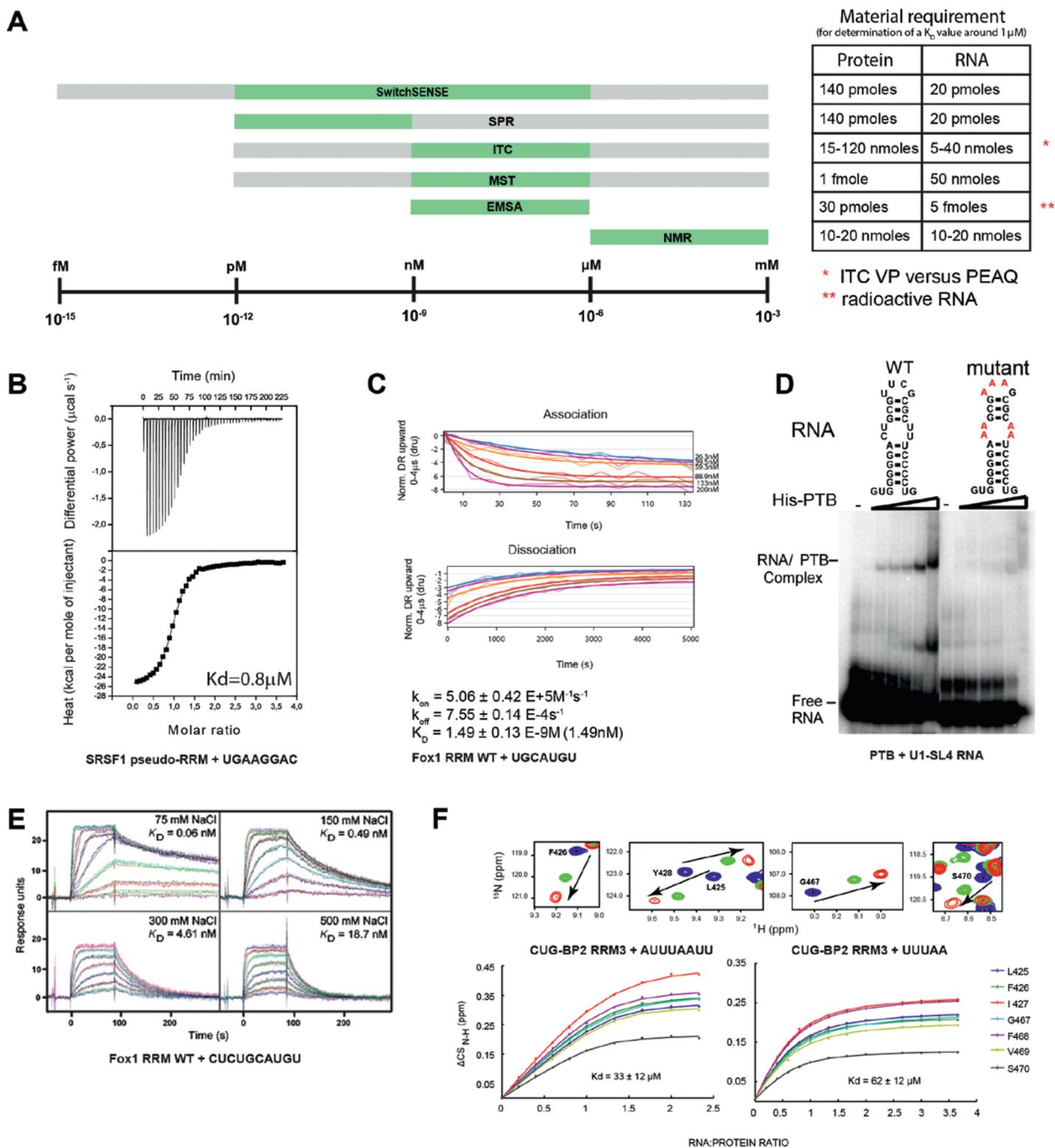


Fig. 3. Comparison of the main methods used to measure RBP affinity for RNA. (A) Overview of the range of affinities covered by the different techniques used to measure affinity of RBPs with RNA. The affinity range that was previously tested and worked in our lab is displayed in green. The table shows the amount of protein and RNA required for the determination of a  $K_D$  value around 1  $\mu\text{M}$ . (B) Determination of the  $K_D$  of SRSF1 pseudo-RRM in complex with 5'-UGAAGGAC-3' RNA using the ITC (Isothermal Titration Calorimetry) method.<sup>[34b]</sup> (C) Determination of the  $K_D$  of Fox-1 RRM in complex with 5'-UGCAUGU-3' RNA using the switchSENSE method.<sup>[25]</sup> (D) Analysis of the binding of PTB to wild-type and mutated versions of U1-SL4 RNA using EMSA.<sup>[44]</sup> (E) Determination of the  $K_D$  of Fox-1 RRM in complex with 5'-CUCUGCAUGU-3' RNA using the Surface Plasmon Resonance (SPR) method at different salt concentrations.<sup>[37]</sup> (F) Determination of the  $K_D$  of CUG-BP2 RRM3 in complex with 5'-AUUUAAU-3' and 5'-UUUAA-3' RNA using NMR.<sup>[17c]</sup>

ualization of protein–RNA complex formation through changes in migration speed during gel electrophoresis. RNA is usually radioactively labeled and pre-incubated with increasing protein concentrations to form complexes, which are then loaded on a non-denaturing acrylamide gel. The speed of migration of the pro-

tein–RNA complex is usually slower compared to the free RNA, which allows the user to distinguish them and calculate a percentage of complex formation at the different protein concentrations tested. Although  $K_D$  values can be determined, this method is not the most accurate, requires labeling of the RNA or protein and is

more often used to compare mutated versions of protein or RNA with the WT complex. Yet, taking advantage of the radioactive labeling, it is a very sensitive method requiring little material compared to ITC or NMR for example (Fig. 3).

Microscale thermophoresis (MST) is based on the observation of the directed movement of fluorescently labeled biomolecules in a temperature gradient.<sup>[28]</sup> The displacement of these molecules depends on their size, effective charge and solvation entropy. As at least one of these parameters changes upon protein–RNA interactions, it is possible to detect them and determine a corresponding  $K_D$  value. MST is performed in solution inside capillaries in close-to-native conditions. The primary advantage of this technology is the use of a very little amount of biomolecules, but it often requires fluorescent labeling of one of the component (Fig. 3).

Finally, NMR is a method that can be very useful to measure very low affinities ( $K_D > 10 \mu\text{M}$ ). In this condition, protein and RNA are typically exchanging between free and bound states in the sub-millisecond time scale and an averaged signal between the free and bound forms of the protein or the RNA are then detected by NMR. This state is described as a fast exchange regime. With  $^{15}\text{N}$ -labeled protein,  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectra are typically recorded to follow the signal of each amide resonance of the protein during an RNA titration for example. Some peaks will then gradually move from the position observed in the free state of the protein to the position observed in a fully RNA-bound state (Fig. 3). The magnitude of the chemical shift change can be plotted against the RNA concentrations and used to extract the  $K_D$ .<sup>[29]</sup> This approach requires milligram amounts of the material but allows accurate measurement of the affinity for  $K_D$  in the micro- to millimolar range. Requirements and limits of each method are described in Tables 1 and 2.

## 5. How to Identify the Interface between an RBP and RNA?

Once the affinity of the RNA binding domain and the target RNA has been measured, the next step in the investigation of this interaction is to characterize the protein–RNA interface. Of course, structure determination of the protein–RNA complex will often provide the ultimate answer regarding the molecular basis of the interaction. Yet, this process can be long and tedious and is sometimes not needed if one primarily wishes to identify the

protein–RNA interface to mutate the complex and test its functionality.

If the protein is smaller than 30 kDa and can be concentrated to 0.5 to 1 mM, NMR will be the method of choice to identify the part of the protein involved in RNA binding. Using  $^{15}\text{N}$ - $^{13}\text{C}$  labeled proteins, 3D experiments such as HNCA, CBCACONH and HNCACB can be recorded allowing the assignment of the protein backbone amides. By recording successive  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectra of the protein in the presence of increasing amount of RNA (see above), one can then identify the residues for which the environment of the backbone amide changed during the RNA titration. The largest chemical shift perturbations are usually a good evidence that these residues are involved in the complex formation either directly or indirectly (Fig. 4A). Therefore such residue-specific mapping can be used to perform subsequent mutagenesis and derive a low resolution 3D model of the complex using docking programs like HADDOCK.<sup>[30]</sup>

If the size of the protein is larger than 30 kDa or the protein not amenable for NMR investigations (*e.g.* problems of solubility at high concentration), another method developed recently in our lab in collaboration with the Aebersold lab can be used instead: CLIR-MS/MS.<sup>[31]</sup> The method allows deciphering protein–RNA interactions at single amino acid and nucleotide resolution. It requires the synthesis of  $^{13}\text{C}$ ,  $^{15}\text{N}$  segmentally labeled RNA, UV-crosslinking and LC-MS/MS. Isotopically labeled RNA segments are synthesized and ligated to flanked non-labeled RNAs (Fig. 4B). These RNAs are then mixed with an equivalent amount of fully unlabeled RNA (1:1 ratio) and this mix is used to form a complex with the proteins of interest (or possibly cell extract), which are then UV cross-linked at 254 nm. After mild protease and RNase digestions following a protocol established earlier by the Urlaub lab,<sup>[32]</sup> RNA–protein contact sites can be accurately identified by tandem MS. The photo-crosslinking site can be identified in the form of a modified peptide with an attached nucleotide or short oligonucleotide (up to 3 nucleotides). The targeted RNA sequence can be identified precisely by the fact that the cross-linked product with the 50% isotopically labeled RNA results in a doublet of approximately equal intensity separated by the mass difference between the labeled and the unlabeled nucleotide (or oligoribonucleotide). For example, for a  $^{13}\text{C}$ , $^{15}\text{N}$  labeled uracil, the mass shift is 11 Da. The mass of the nucleotide being sub-

Table 1. Comparison of methods used to identify sequence specificity of protein–RNA complexes.

	<i>in vitro</i> methods				<i>in vivo</i> methods	
	NMR	SELEX	RBNS	RNAcompete	RIP-Chip	CLIP
<b>Labeling</b>	$^{15}\text{N}$ -labeled protein < 30KDa			Cy3 or Cy5 labeling of captured RNAs		
<b>Source of the RNA</b>	Specific RNA sequence	Random RNA sequences	Biotinylated random RNA sequences (40 nucleotides)	Non-random RNA sequences (40 nucleotides)	Natural RNA targets from cell extract	Natural RNA targets in cell
<b>General features</b>	Identifies unique minimal motif preference in an iterative process ( $^1\text{H}$ $^{15}\text{N}$ HSQC)	Enrichment of high-specificity motif by affinity selection followed by RT and HT-sequencing	Enrichment of high-specificity motif by streptavidin pulldown of biotinylated RNA and HT-sequencing	Pull-down of tagged protein and hybridization on custom Agilent 244K microarray	Immunoprecipitation of specific RBP from cell extract followed by hybridization to microarray	UV-crosslinking and immunoprecipitation of specific RBP followed by RNase digestion, RT-PCR and HT-sequencing

Table 2. Comparison of methods used to determine affinities and stoichiometry of protein–RNA complexes.

	Label-free methods			Label-based methods		
	ITC	SPR	Switchsense	EMSA	MST	NMR
<b>Labeling</b>				Labeling of the RNA :  <sup>32</sup> P-radioactive / Fluorescence / biotin	One of the interaction partners is fluorescently labeled	<sup>15</sup> N-labeled protein
<b>Duration of the experiment</b>	45 min to several hours depending on K <sub>D</sub> values	30 min to several hours depending on kon and koff values	30 min to several hours depending on kon and koff values	2–3 hrs	Few minutes	2–3 hrs
<b>Affinity range</b>	pM–mM	pM–mM	fM–mM	nM–μM  Not accurate	pM–mM	μM–mM
<b>Immobilization</b>	None	Surface based immobilization via biotin	Surface based immobilization via base-pairing	In gel matrix	None	None
<b>General features</b>	Measurement of affinity and thermodynamic parameters based on enthalpy changes	Measurement of affinity and time-resolved kinetics based on refractive index	Measurement of affinity and time-resolved kinetics based on fluorescence	K <sub>D</sub> derived from ratio of bound state to free state on gel	K <sub>D</sub> derived from directed movement of fluorescently labeled molecule in temperature gradient	K <sub>D</sub> derived from chemical shift differences in the <sup>1</sup> H <sup>15</sup> N HSQC spectrum

tracted, the total mass should correspond to the mass of the cross-linked peptide that can be mapped to the exact protein site.<sup>[31]</sup> This original approach allowed the mapping of each of the four RRM of PTB on the EMCV RNA (Fig. 4C).<sup>[31]</sup> Moreover, peptide–RNA cross-links could be used as structural constraints to quickly determine atomic models of the protein–RNA interface of each RRM bound to their respective RNA binding site.<sup>[31]</sup> Compared to classical structure determinations, this CLIR-MS/MS approach is much faster and requires less material by several orders of magnitude.<sup>[31]</sup>

## 6. In vivo Validation of the Protein–RNA Interactions Characterized in vitro

Once the details about the protein–RNA interaction have been characterized *in vitro*, it is important to check their relevance *in vivo*. A good strategy consists in mutating the residues and nucleotides involved in the interaction and test their effect on a function that depends on this interaction. For example, a translation repression assay<sup>[33]</sup> or an *in vivo* splicing assay.<sup>[34]</sup> Such assays often showed clear correlation between the effect of the mutations on the protein–RNA interaction *in vitro* and the functionality of the protein *in vivo*. Alternatively, studies can also be performed in cell extracts.<sup>[35]</sup>

## 7. Conclusion and Perspectives

In this article, we show that a large panel of methods are available to study protein–RNA interactions with their *pro* and *cons*. *In vitro* and *in vivo* investigations are complementary, often support each other and are both required to decipher the mode of interaction of RBPs. While the approaches described here are robust and for the most part fairly rapid and applicable to most individual protein–RNA complex, in order to fully characterize RNA metabolism in cells, we will need soon to obtain information on networks of protein–RNA interactions involving sometimes sev-

eral dozens of biomolecules bound at once. Use of methodologies such as cryogenic electron microscopy (cryo-EM) and CLIR-MS/MS will need to become routine on these systems to investigate such questions.

## Acknowledgments

We thank the SNF-NCCR RNA & Disease for financial support.

Received: March 4, 2019

- [1] A. Re, T. Joshi, E. Kulberkyte, Q. Morris, C. T. Workman, *Methods Mol. Biol.* **2014**, 1097, 491, DOI: 10.1007/978-1-62703-709-9\_23.
- [2] a) A. E. Brinegar, T. A. Cooper, *Brain Res.* **2016**, 1647, 1, DOI: 10.1016/j.brainres.2016.02.050; b) K. E. Lukong, K. W. Chang, E. W. Khandjian, S. Richard, *Trends Genet.* **2008**, 24, 416, DOI: 10.1016/j.tig.2008.05.004.
- [3] F. C. Y. Lee, J. Ule, *Mol. Cell* **2018**, 69, 354, DOI: 10.1016/j.molcel.2018.01.005.
- [4] M. W. Hentze, A. Castello, T. Schwarzl, T. Preiss, *Nature Rev.* **2018**, 19, 327, DOI: 10.1038/nrm.2017.130.
- [5] a) S. Calabretta, S. Richard, *Trends Biochem. Sci.* **2015**, 40, 662, DOI: 10.1016/j.tics.2015.08.012; b) A. Srivastava, S. Ahmad, M. M. Gromiha, *Int. J. Mol. Sci.* **2018**, 19, DOI: 10.3390/ijms19061595.
- [6] T. Afroz, Z. Cienikova, A. Clery, F. H. Allain, *Methods Enzymol.* **2015**, 558, 235, DOI: 10.1016/bs.mie.2015.01.015.
- [7] A. Clery, M. Blatter, F. H. Allain, *Curr. Opin. Struct. Biol.* **2008**, 18, 290, DOI: 10.1016/j.sbi.2008.04.002.
- [8] G. Masliah, P. Barraud, F. H. Allain, *Cell Mol. Life Sci.* **2013**, 70, 1875, DOI: 10.1007/s00018-012-1119-x.
- [9] R. Stefl, F. C. Oberstrass, J. L. Hood, M. Jourdan, M. Zimmermann, L. Skrisovska, C. Maris, L. Peng, C. Hofr, R. B. Emeson, F. H. Allain, *Cell* **2010**, 143, 225, DOI: 10.1016/j.cell.2010.09.026.
- [10] D. Ray, K. C. H. Ha, K. Nie, H. Zheng, T. R. Hughes, Q. D. Morris, *Methods* **2017**, 118–119, 3, DOI: 10.1016/j.ymeth.2016.12.003.
- [11] B. Beuth, M. F. Garcia-Mayoral, I. A. Taylor, A. Ramos, *J. Am. Chem. Soc.* **2007**, 129, 10205, DOI: 10.1021/ja072365q.
- [12] D. Theler, C. Dominguez, M. Blatter, J. Boudet, F. H. Allain, *Nucleic Acids Res.* **2014**, 42, 13911, DOI: 10.1093/nar/gku1116.
- [13] C. Tuerk, L. Gold, *Science* **1990**, 249, 505.



- Pitsch, D. L. Black, F. H. Allain, *Science* **2005**, *309*, 2054, DOI: 10.1126/science.1114066.
- [18] L. O. Penalva, S. A. Tenenbaum, J. D. Keene, *Methods Mol. Biol.* **2004**, *257*, 125, DOI: 10.1385/1-59259-750-5:125.
- [19] J. D. Keene, J. M. Komisarow, M. B. Friedersdorf, *Nat. Protoc.* **2006**, *1*, 302, DOI: 10.1038/nprot.2006.47.
- [20] N. Cloonan, A. R. Forrest, G. Kolle, B. B. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan, S. M. Grimmond, *Nat. Methods* **2008**, *5*, 613, DOI: 10.1038/nmeth.1223.
- [21] J. Ule, K. B. Jensen, M. Ruggiu, A. Mele, A. Ule, R. B. Darnell, *Science* **2003**, *302*, 1212, DOI: 10.1126/science.1090095.
- [22] T. Wiseman, S. Williston, J. F. Brandts, L. N. Lin, *Anal. Biochem.* **1989**, *179*, 131.
- [23] A. L. Feig, *Biopolymers* **2007**, *87*, 293, DOI: 10.1002/bip.20816.
- [24] J. Hwang, S. Nishikawa, *J. Biomol. Screen.* **2006**, *11*, 599, DOI: 10.1177/1087057106288491.
- [25] A. Clery, T. J. M. Sohler, T. Welte, A. Langer, F. H. T. Allain, *Methods* **2017**, *118-119*, 137, DOI: 10.1016/j.ymeth.2017.03.004.
- [26] M. M. P. Denichenko, A. Cléry, T. Welte, J. Biran, O. Shimshon, G. D. Barnabas, M. Danan-Gotthold, S. Kumar, E. Yavin, E. Y. Levanon, F. H. Allain, T. Geiger, G. Levkowitz, R. Karni, *Nat. Commun.* **2019**, in press.
- [27] L. M. Hellman, M. G. Fried, *Nat. Protoc.* **2007**, *2*, 1849, DOI: 10.1038/nprot.2007.249.
- [28] S. Duhr, D. Braun, *Proc. Nat. Acad. Sci. USA* **2006**, *103*, 19678, DOI: 10.1073/pnas.0603873103.
- [29] L. Fielding, *Curr. Top. Med. Chem.* **2003**, *3*, 39.
- [30] C. Dominguez, R. Boelens, A. M. Bonvin, *J. Am. Chem. Soc.* **2003**, *125*, 1731, DOI: 10.1021/ja026939x.
- [31] G. Dorn, A. Leitner, J. Boudet, S. Campagne, C. von Schroetter, A. Moursy, R. Aebersold, F. H. Allain, *Nat. Methods* **2017**, *14*, 487, DOI: 10.1038/nmeth.4235.
- [32] K. Kramer, T. Sachsenberg, B. M. Beckmann, S. Qamar, K. L. Boon, M. W. Hentze, O. Kohlbacher, H. Urlaub, *Nat. Methods* **2014**, *11*, 1064, DOI: 10.1038/nmeth.3092.
- [33] J. Hennig, C. Militti, G. M. Popowicz, I. Wang, M. Sonntag, A. Geerlof, F. Gabel, F. Gebauer, M. Sattler, *Nature* **2014**, *515*, 287, DOI: 10.1038/nature13693.
- [34] a) A. Clery, S. Jayne, N. Benderska, C. Dominguez, S. Stamm, F. H. Allain, *Nat. Struct. Mol. Biol.* **2011**, *18*, 443, DOI: 10.1038/nsmb.2001; b) A. Clery, R. Sinha, O. Anczukow, A. Corrionero, A. Moursy, G. M. Daubner, J. Valcarcel, A. R. Krainer, F. H. Allain, *Proc. Nat. Acad. Sci. USA* **2013**, *110*, E2802, DOI: 10.1073/pnas.1303445110; c) A. Moursy, F. H. Allain, A. Clery, *Nucleic Acids Res.* **2014**, *42*, 6659, DOI: 10.1093/nar/gku244.
- [35] G. Nicastro, M. F. Garcia-Mayoral, D. Hollingworth, G. Kelly, S. R. Martin, P. Briata, R. Gherzi, A. Ramos, *Nat. Struct. Mol. Biol.* **2012**, *19*, 1282, DOI: 10.1038/nsmb.2427.
- [36] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B. M. Beckmann, C. Strein, N. E. Davey, D. T. Humphreys, T. Preiss, L. M. Steinmetz, J. Krijgsveld, M. W. Hentze, *Cell* **2012**, *149*, 1393, DOI: 10.1016/j.cell.2012.04.031.
- [37] S. D. Auweter, R. Fasan, L. Reymond, J. G. Underwood, D. L. Black, S. Pitsch, F. H. Allain, *EMBO J.* **2006**, *25*, 163, DOI: 10.1038/sj.emboj.7600918.
- [38] F. E. Loughlin, P. J. Lukavsky, T. Kazeeva, S. Reber, E. M. Hock, M. Colombo, C. Von Schroetter, P. Pauli, A. Clery, O. Muhlemann, M. Polymenidou, M. D. Ruepp, F. H. Allain, *Mol. Cell* **2019**, *73*, 490, DOI: 10.1016/j.molcel.2018.11.012.
- [39] C. Dominguez, J. F. Fiset, B. Chabot, F. H. Allain, *Nature Struct. Mol. Biol.* **2010**, *17*, 853, DOI: 10.1038/nsmb.1814.
- [40] G. M. Daubner, A. Brummer, C. Tocchini, S. Gerhardy, R. Ciosk, M. Zavolan, F. H. Allain, *Nucleic Acids Res.* **2014**, *42*, 8092, DOI: 10.1093/nar/gku445.
- [41] a) M. Jangi, P. L. Boutz, P. Paul, P. A. Sharp, *Genes Devel.* **2014**, *28*, 637, DOI: 10.1101/gad.235770.113; b) D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecnas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, T. R. Hughes, *Nature* **2013**, *499*, 172, DOI: 10.1038/nature12311.
- [42] a) O. Anczukow, M. Akerman, A. Clery, J. Wu, C. Shen, N. H. Shirole, A. Raimier, S. Sun, M. A. Jensen, Y. Hua, F. H. Allain, A. R. Krainer, *Mol. Cell* **2015**, *60*, 105, DOI: 10.1016/j.molcel.2015.09.005; b) J. R. Sanford, X. Wang, M. Mort, N. Vanduy, D. N. Cooper, S. D. Mooney, H. J. Edenberg, Y. Liu, *Genome Res.* **2009**, *19*, 381, DOI: 10.1101/gr.082503.108; c) X. Wang, L. Juan, J. Lv, K. Wang, J. R. Sanford, Y. Liu, *BMC Genomics* **2011**, *12 Suppl 5*, S8, DOI: 10.1186/1471-2164-12-S5-S8.
- [43] a) C. Colombrita, E. Onesto, F. Megiorni, A. Pizzuti, F. E. Baralle, E. Buratti, V. Silani, A. Ratti, *J. Biol. Chem.* **2012**, *287*, 15635, DOI: 10.1074/jbc.M111.333450; b) D. Dominguez, P. Freese, M. S. Alexis, A. Su, M. Hochman, T. Palden, C. Bazile, N. J. Lambert, E. L. Van Nostrand, G. A. Pratt, G. W. Yeo, B. R. Graveley, C. B. Burge, *Mol. Cell* **2018**, *70*, 854, DOI: 10.1016/j.molcel.2018.05.001; c) K. Kapeli, G. A. Pratt, A. Q. Vu, K. R. Hutt, F. J. Martinez, B. Sundararaman, R. Batra, P. Freese, N. J. Lambert, S. C. Huelga, S. J. Chun, T. Y. Liang, J. Chang, J. P. Donohue, L. Shiue, J. Zhang, H. Zhu, F. Cambi, E. Kasarskis, S. Hoon, M. Ares, Jr., C. B. Burge, J. Ravits, F. Rigo, G. W. Yeo, *Nat. Commun.* **2016**, *7*, 12143, DOI: 10.1038/ncomms12143; d) J. I. Takeda, A. Masuda, K. Ohno, *Gene* **2017**, *618*, 57, DOI: 10.1016/j.gene.2017.04.008.
- [44] S. Sharma, C. Maris, F. H. Allain, D. L. Black, *Mol. Cell* **2011**, *41*, 579, DOI: 10.1016/j.molcel.2011.02.012.