Advances in the characterization of RNA-binding proteins

Domenica Marchese, † Natalia Sanchez de Groot, † Nieves Lorenzo Gotor, Carmen Maria Livi, and Gian G. Tartaglia*

From transcription, to transport, storage, and translation, RNA depends on association with different RNA-binding proteins (RBPs). Methods based on next-generation sequencing and protein mass-spectrometry have started to unveil genome-wide interactions of RBPs but many aspects still remain out of sight. How many of the binding sites identified in high-throughput screenings are functional? A number of computational methods have been developed to analyze experimental data and to obtain insights into the specificity of protein–RNA interactions. How can theoretical models be exploited to identify RBPs? In addition to oligomeric complexes, protein and RNA molecules can associate into granular assemblies whose physical properties are still poorly understood. What protein features promote granule formation and what effects do these assemblies have on cell function? Here, we describe the newest in silico, in vitro, and in vivo advances in the field of protein–RNA interactions. We also present the challenges that experimental and computational approaches will have to face in future studies. © 2016 The Authors. WIREs RNA published by Wiley Periodicals, Inc.

INTRODUCTION

Since ‘the central dogma’ was formulated (the genetic information flows from DNA to protein passing through RNA), our knowledge on the transcriptome has progressed enormously.1,2 We now know that genes do not code for just a single protein but produce a number of variants: the activity of nuclear RNAs (snRNAs and snoRNAs) and several splicing factors generates several messenger RNAs (mRNAs) with different lengths and exon compositions (Figure 1).2 In addition to protein production, RNA participates in other essential processes such as mRNAs expression regulation (microRNAs, miRNA; small interfering RNAs, siRNA; long noncoding RNAs, lncRNAs)1,2 and genome protection by transposon silencing (PIWI-interacting RNA, piRNA).3 RNAs also perform a number of structural and functional tasks, as in the case of rRNAs, which constitute 60% of the ribosome,4 and tRNAs, which carry the amino acids to the ribosome during the translation process.1,2

Every aspect of RNA life, from birth (polymerases) to degradation (nucleases), involves protein binding (Figure 1). A correct interplay between RNAs and RNA-binding proteins (RBPs) is crucial for the development of cellular processes: miRNAs, for instance, require Argonaute proteins to reach target mRNAs as well as piRNAs associate with PIWI proteins to form silencing complexes that protect the germline genome from transposons.1,2

† These authors contributed equally to this article.
*Correspondence to: gian.tartaglia@crg.es
1 Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain
2 Universitat Pompeu Fabra (UPF), Barcelona, Spain
3 IFOM Foundation, FIRC Institute of Molecular Oncology Foundation, Milan, Italy
4 Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Conflict of interest: The authors have declared no conflicts of interest for this article.
Moreover, the interplay between RBPs and RNAs can promote formation of membrane-less organelles (e.g., P-bodies, stress granules), whose composition can be rapidly adapted to the cell state and the environment conditions (e.g., stress conditions). Mutations in RBPs, aberrant interactions or altered RNA processing have been related to a number of human diseases, from neurological disorders to cancer.

Protein–RNA interactions have been studied using a range of quantitative approaches such as electrophoretic mobility shift assay (EMSA), fluorescent anisotropy/polarization, Förster resonance
energy transfer (FRET),

surface plasmon resonance (SPR)

or, more recently, microscale thermophoresis (MST)

and stochastic optical reconstruction microscopy (STORM).

Although these approaches have proved to be powerful to assess the specificity and affinity of protein–RNA interactions, they only allow the study of single or few molecular interactions at a time.

The recent development of large-scale quantitative methods, exploiting next-generation sequencing and protein mass spectrometry, contributes to the genome-wide identification of RBPs, RNA targets and cofactors. Deep-sequencing approaches combined with RBPs immunoprecipitation as well as in vitro methods, such as Systematic Evolution of Ligands by Exponential Enrichment SELEX, revealed the binding ability of a number of RBPs and showed that many RBPs bind to thousands of transcripts.

In parallel to experimental advances, a number of in silico methods have been developed to predict protein–RNA interactions and to detect binding-sites. Computational tools are particularly useful to predict potential ribonucleoprotein associations and to narrow down a list of interaction partners for experimental validation. For instance, some RBPs recognize particular nucleotide sequences, whereas others bind to the backbone or double-stranded RNA. When modeling protein–RNA interactions, many factors should be considered, including secondary structures, folding and characteristics of binding interfaces. Especially physicochemical properties of amino acids, such as structural disorder and polarity, are relevant to characterize the RNA-binding ability of proteins. Indeed, recent studies reported that in addition to classical RNA-binding domains other regions found in ribosomal proteins, translation elongation factors, zinc fingers as well as structurally disordered parts participate in contacting transcripts.

Here we review the most recent experimental and computational advances for the detection of protein–RNA interactions and introduce new challenges for future developments in the field.

**EXPERIMENTAL METHODS FOR DETECTION OF PROTEIN–RNA INTERACTIONS**

Transcripts are never naked and form complexes with partner proteins in ribonucleoprotein particles from their birth to their degradation (Figure 1). The assembly of functional complexes and the delivery to final destination involves progression through a series of intermediate complexes and subcellular localizations. For instance, Cajal bodies are sites of noncoding ribonucleoprotein particles maturation, where assembly factors gather to accelerate complicated biochemical reactions. Complexes containing mRNAs often undergo remodeling as they travel from the site of transcription to the cytoplasm where they are translated.

The advent of sequencing technologies, together with the introduction of various cross-linking chemistries, has enabled the development of new high-throughput methods for the simultaneous detection of hundreds to thousands interactions in a single experiment. The methods can be classified into ‘protein-centric’ approaches, which reveal RNAs bound to a known protein, and ‘RNA-centric’ approaches, which characterize proteins interacting with an RNA of interest.

**Protein-Centric Approaches**

The predominant protein-centric methods are based on protein immunoprecipitation from cell lines or tissues, and detection of the co-purified RNAs (Table 1 and Figure 2).

RNA immunoprecipitation (RIP) implies the purification of RNA–protein interactions in native conditions by using a protein-specific antibody, and detection of interacting RNAs by either microarray (RIP-chip) or sequencing (RIP-seq). Despite the genome-wide potential, the method has practical limitations. Indeed, it is prone to detect nonspecific interactions due to the nonphysiological formation of protein–RNA complexes in solution. As a consequence, interactions identified using native purification methods often require additional validation.

RIP limitations have been overcome with the introduction of cross-linking and denaturing methods, namely cross-linking and immunoprecipitation (CLIP). CLIP combines UV cross-linking of RBPs to their cognate RNA molecules with stringent purification of protein–RNA complexes that are resolved and size-selected on an SDS-PAGE before proceeding to high-throughput sequencing of cDNA library (HITS-CLIP). Photoactivatable ribonucleoside enhanced CLIP (PAR-CLIP) and individual-nucleotide resolution CLIP (iCLIP) are approaches derived from modified cross-linking or library-preparation protocols that allow the identification of cross-linking sites at a single-nucleotide resolution.

CLIP has enabled the characterization of RNA binding profiles of several RBPs involved in neurologic disorders and cancer, thus helping in understanding their role in disease. The first CLIP experiment,
<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Challenges</th>
<th>Disease-Related? RBP/RNA</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Protein-centric</strong></td>
<td><strong>In vivo</strong></td>
<td>RIP</td>
<td>Genome-wide</td>
<td>ELAVL1 (epilepsy, cancer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Applicable to tissues</td>
<td>CELF1 (myotonic dystrophy)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>High background noise</td>
<td>IGF2BP1 (cancer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Possible artifacts</td>
<td>ADAR (Dyschromatosis symmetrica, spastic paraplegia)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low resolution</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>iCLIP</td>
<td>Genome-wide</td>
<td>CELF4 (epilepsy, hyperactivity)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eCLIP</td>
<td>High specificity</td>
<td>ELAVL1 (epilepsy, cancer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>False negative due to low cross-linking efficiency</td>
<td>FMR1 (Fragile-X mental retardation, autism spectrum disorders)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HiTS-CLIP</td>
<td>Time consuming</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PAR-CLIP</td>
<td>Challenging set-up</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>iCLIP</td>
<td>Not applicable to tissues (PAR-CLIP)</td>
<td></td>
</tr>
<tr>
<td><strong>In vitro</strong></td>
<td>RNA-compete</td>
<td>SEQRS</td>
<td>Large-scale</td>
<td>ELAVL4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RBNNS</td>
<td>Measurement of interaction affinity and specificity</td>
<td>NOVA1 (POMA) HNRNPA0 (cancer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RNA-MaP</td>
<td>Analysis of multiple proteins</td>
<td>MBNL1 (myotonic dystrophy)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HiTS-RAP</td>
<td></td>
<td>CELF1 (myotonic dystrophy)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MITOMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RNA-centric</strong></td>
<td><strong>In vitro</strong></td>
<td>TRAP/RAT</td>
<td>Relatively easy and flexible</td>
<td>TPS3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RaPID</td>
<td>Easy purification protocols</td>
<td>HRAS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RiboTrap</td>
<td></td>
<td>MYC (cancer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RNA-assisted chromatography</td>
<td></td>
<td>BCL2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Protein microarray</td>
<td>Large-scale</td>
<td>PWRN1 (Prader-Willi syndrome)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Relatively easy and fast</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>In vivo</strong></td>
<td>MS2-BioTRAP</td>
<td>Fast and easy set-up</td>
<td>Relatively easy and fast</td>
<td>TPS3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Large-scale</td>
<td>Challenges associated with cell transfection</td>
<td>HRAS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low amounts of starting material</td>
<td>Possible artifacts due to tag</td>
<td>MYC (cancer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Study of protein--RNA interactions</td>
<td>Study of protein--RNA interactions under physiological conditions</td>
<td>BCL2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ChIRP</td>
<td>Time and cost consuming</td>
<td>MALAT1 (cancer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHART</td>
<td>Large amounts of starting material required</td>
<td>NEAT1 (cancer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RAP-MS</td>
<td>Challenging set-up</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interactome capture</td>
<td>Study of protein--RNA interactions under physiological conditions</td>
<td>Possible artifacts (positive and false negative) due to cross-linking</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Identification of unknown RBPs</td>
<td>Time consuming</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Challenging set-up</td>
</tr>
</tbody>
</table>

RIP, RNA Immunoprecipitation, HiTS-CLIP, High-throughput sequencing of RNA isolated by cross-linking immunoprecipitation, PAR-CLIP, Photoactivatable ribonucleoside enhanced CLIP, iCLIP, individual-nucleotide resolution Cross-Linking and ImmunoPrecipitation, SEQRS, in vitro selection high-throughput sequencing of RNA and sequence specificity landscape, RBNNS, RNA Bond-n-Seq, RNA-MaP, RNA on a massively parallel array, HiTS-RAP, High-throughput sequencing RNA affinity profiling, RNA-MITOMI, RNA mechanically induced trapping of molecular interactions, TRAP, Tandem RNA affinity purification, RAT, RNA affinity in tandem, RaPID, RNA-binding protein purification and identification, MS2-BioTRAP, MS2 in vivo biotin tagged RNA affinity purification, ChIRP, Chromatin isolation by RNA purification, CHART, Capture hybridization analysis of RNA targets, RAP-MS, RNA antisense purification-mass spectrometry.
published by Ule et al., shed light on the role of NOVA 1 and 2 proteins in paraneoplastic opsoclonus-myoclonus-ataxia (POMA), an autoimmune neurologic disease characterized by abnormal motor inhibition.\textsuperscript{55} NOVA proteins were found to regulate alternative splicing of RNAs encoding multiple components of inhibitory synapses, such as GABA\textsubscript{2} receptor and GIRK2, which mediate slow inhibitory postsynaptic potentials, the K\textsuperscript{+} voltage-gated channel KCNQ3 or the nicotinic acetylcholine receptors \(\beta2\) and \(\alpha2\), highly represented in the GABAergic interneurons.\textsuperscript{55}

The combination of RIP-chip and CLIP approaches has allowed the identification of targets of the mammalian ELAVL family whose four members, ELAVL1,2,3 and 4, (also known as HuR, HuB, HuC, and HuD) are implicated in different cancers and neurological diseases.\textsuperscript{56,57} In particular, ELAVL1, the only member of the family expressed in both neuronal and non neuronal cells, has been found particularly abundant in breast, ovary, colon, and brain cancers\textsuperscript{58} associated with poor prognosis.\textsuperscript{59} RIP-chip and CLIP have identified a transcriptome-wide list of targets (including Cyclin D1, E1, A2 and B1, EGF, c-Myc, p27, COX-2, and BRCA1) and indicated that ELAVL1 functions as a major hub for regulating RNA metabolism in the cell at different levels, from pre-mRNA alternative splicing to mature mRNA stability and microRNA biogenesis.\textsuperscript{57,60} The other three members, ELAVL2–4 (nELAVl), are highly enriched in neurons and show a unique hierarchical expression during cortical development.\textsuperscript{61} HITS-CLIP on mice cortical tissue has revealed that nELAVl regulated transcripts are mainly involved in synaptic cytoskeleton assembly and disassembly, amino acid and sugar biosynthetic pathways. In particular, nELAVl have been found to regulate the alternative splicing of the gene coding for the glutaminase enzyme, the major responsible for the synthesis of the excitatory neurotransmitter glutamate.\textsuperscript{56}

The complex protocol and bioinformatics required for data analysis represent the main disadvantage of CLIP approaches (Figure 2). As a matter of fact, high experimental failure rates are reported. Recently, an enhanced CLIP (eCLIP) protocol has been developed.\textsuperscript{34} In addition to reducing hands-on time to as few as 4 days, eCLIP dramatically decreases required library amplification by 1000 fold and enhances the rate of success at generating libraries with high usable reads percentages across diverse RBPs, maintaining the single-nucleotide resolution of previous methods. In addition, paired-matched input controls improve the signal-to-noise ratio for the discovery of authentic binding sites. 102 eCLIP experiments for 73 diverse RBPs have been generated (available at https://www.encodeproject.org), providing an unprecedented source of data for the study of protein–RNA functional networks. Despite its successful applications, a major concern when using CLIP is that UV-induced cross-linking is still poorly understood at the biophysical level. Only a very small percentage (1–5\%) of protein–RNA complexes present in the cells can be efficiently cross-linked, and it is not clear which types of interactions might be unseen.\textsuperscript{62} For instance, several RBP families do not directly associate with nucleic acid bases but interact with other elements (i.e., the sugar phosphate backbone) showing low cross-linking efficiency. Moreover, UV only cross-links direct protein–RNA interactions but it does not capture interactions occurring with protein complexes, thus providing just part of the information.

The study of \textit{in vivo} interactions is limited to contacts formed in a certain cell type and at a specific time point. To better understand the physico-chemical properties controlling protein–RNA interactions, a number of new methods allow the \textit{in vitro} screening of interactions between proteins and libraries of randomly generated RNA sequences by combining the use of microarray and microfluidic platforms with molecule fluorescent labeling and RNA sequencing technologies. While \textit{in vitro} evolution \text{SELEX} has bias towards highest-affinity targets,\textsuperscript{56} other methods enable the characterization of lower-specificity and medium-range affinities with single proteins or multiprotein complexes.\textsuperscript{63} Incubation of RNA libraries with a protein of interest immobilized on an affinity matrix is followed by fluorescent labeling of selected RNAs and hybridization to a microarray, in the case of \textit{RNA-compete} (Figure 2), or deep sequencing in the case of \textit{in vitro} selection high-throughput sequencing of RNA and sequence specificity landscape (\text{SEQRS}) and RNA Bind-n-Seq (\text{RBNS}).\textsuperscript{25,36} These methods are often applied to identify \textit{consensus} elements of RBPs (see also \textit{Computational Methods for Detection of Protein–RNA Interactions}).

Quantitative analysis of RNA on a massively parallel array (RNA-MaP) and high-throughput sequencing RNA affinity profiling (HiTS-RAP), consist in the \textit{in situ} synthesis of RNA libraries inside an \textit{Illumina} sequencing flow cell followed by incubation with a fluoresently labeled protein and quantification of molecular interactions.\textsuperscript{37,38} These approaches, together with RBNS, provide quantitative measurements of dissociation constants (Kd) through the use of multiple protein concentrations. Moreover, methods relying on the use of fluorescent proteins, such as...
**Protein-centric methods**

(a) *In vivo* approaches include native purification protocols (RNA Immunoprecipitation, RIP) and denaturing protocols (Cross-linking and Immunoprecipitation, CLIP). In the first case, RNAs bound to a specific protein are immunoprecipitated from cell lysate in native conditions by using a protein-specific antibody and, after wash and protein removal by proteinase K treatment, RNAs are reverse transcribed and identified through RNA sequencing. In the second case, cells are UV cross-linked to freeze protein–RNA complexes. RNA is digested to obtain fragments of a defined size and the obtained complexes are immunoprecipitated and resolved on an SDS-PAGE. After isolation from membrane and proteinase K digestion, the RNA fragments are reverse transcribed and sequenced. In red and yellow are represented simulated enrichment values of a target and nontarget RNA respectively (lower part of the panel).

(b) As an example of *in vitro* approaches, the schematic workflow of RNA compete is represented. RNA libraries are generated by *in vitro* transcription. Transcripts are incubated with the protein of interest immobilized on an affinity matrix (e.g., streptavidin-biotin tag system) and the bound fragments are then fluorescently labeled and detected by hybridization on a microarray platform.

**RNA-centric methods**

(c) *In vivo* approaches for the identification of proteins bound to an RNA of interest often derive from methods used for the identification of genomic DNA loci targeted by noncoding RNAs. Cells are cross-linked and lysed. Chromatin is sheared and protein–RNA–DNA complexes are pull-down by using biotinylated oligos complementary to the sequence of the RNA. After RNA digestion, proteins can be identified by western blot or mass spectrometry analysis. (d) *In vitro* approaches commonly exploit the use of RNA tags to immobilize the RNA of interest onto an affinity matrix. Upon incubation with cell lysate, proteins bind to the immobilized RNA. After washes, the protein–RNA complexes are eluted from the matrix and proteins are characterized by western blot or mass spectrometry.

---

**FIGURE 2** Schematic representation of experimental methods for the identification of protein–RNA interactions. **Protein-centric methods.** (a) *In vivo* approaches include native purification protocols (RNA Immunoprecipitation, RIP) and denaturing protocols (Cross-linking and Immunoprecipitation, CLIP). In the first case, RNAs bound to a specific protein are immunoprecipitated from cell lysate in native conditions by using a protein-specific antibody and, after wash and protein removal by proteinase K treatment, RNAs are reverse transcribed and identified through RNA sequencing. In the second case, cells are UV cross-linked to ‘freeze’ protein–RNA complexes. RNA is digested to obtain fragments of a defined size and the obtained complexes are immunoprecipitated and resolved on an SDS-PAGE. After isolation from membrane and proteinase K digestion, the RNA fragments are reverse transcribed and sequenced. In red and yellow are represented simulated enrichment values of a target and nontarget RNA respectively (lower part of the panel). (b) As an example of *in vitro* approaches, the schematic workflow of RNA compete is represented. RNA libraries are generated by *in vitro* transcription. Transcripts are incubated with the protein of interest immobilized on an affinity matrix (e.g., streptavidin-biotin tag system) and the bound fragments are then fluorescently labeled and detected by hybridization on a microarray platform. **RNA-centric methods.** (c) *In vivo* approaches for the identification of proteins bound to an RNA of interest often derive from methods used for the identification of genomic DNA loci targeted by noncoding RNAs. Cells are cross-linked and lysed. Chromatin is sheared and protein–RNA–DNA complexes are pull-down by using biotinylated oligos complementary to the sequence of the RNA. After RNA digestion, proteins can be identified by western blot or mass spectrometry analysis. (d) *In vitro* approaches commonly exploit the use of RNA tags to immobilize the RNA of interest onto an affinity matrix. Upon incubation with cell lysate, proteins bind to the immobilized RNA. After washes, the protein–RNA complexes are eluted from the matrix and proteins are characterized by western blot or mass spectrometry.
the previously mentioned RNA-MaP and HiTS-RAP, as well as the RNA mechanically induced trapping of molecular interactions (RNA-MITOMI), provide effective visualization of multiple proteins simultaneously, thus revealing the effects of protein partners. These approaches have a great range of applicability, especially considering that RNA metabolism is regulated by multiple RBPs to form functional particles (e.g., small nuclear ribonucleoproteins, telomerases, ribosomal subunits, UTR-regulatory complexes, etc).

RNA-Centric Approaches

RNA-centric methods aim to identify RBPs targeting a single RNA of interest. The majority of the existing methods exploit tagged RNAs as a bait to capture and characterize all proteins bound to it by mass spectrometry analysis (Table 1 and Figure 2). In vitro synthesized RNA can be chemically tagged through the incorporation of modified ribonucleotides that contain biotin, fluorescent dyes, or, alternatively, natural or artificial aptamers (e.g., S1, D8, MS2 hairpin loop, etc.), can be incorporated. In addition to their in vitro application (e.g., TRAP, RAT, RapID, Ribotrapping, RNase-assisted RNA chromatography), RNA tagging can be applied in vivo, as in the example of MS2-BioTRAP method, where a MS2 hairpin loop tagged RNA and the bacteriophage MS2 coat protein fused to a protein tag (e.g., HB or streptavidin), are co-expressed to capture the RNA of interest by exploiting the high affinity interaction between the MS2 protein and the MS2 hairpin loop. All these methods are relatively flexible, and their in vivo applicability allows the study of protein–RNA interactions in physiological conditions. Nevertheless, the incorporation of a tag into the RNA bait may alter its secondary structures and possibly the formation of ribonucleoprotein complexes. In addition, MS2-BioTRAP is only applicable in easy-to-transfect cells and the overexpression of at least one of the two interacting molecules might lead to experimental artifacts.

Capture hybridization analysis of RNA targets (CHART), Chromatin isolation by RNA purification (ChiRP) and RNA antisense purification (RAP) are methods designed to identify DNA and proteins targeted by noncoding-RNAs in vivo (Figure 2). These approaches involve cell cross-linking and pull-down of the RNA of interest using short biotinylated oligodeoxynucleotides that are complementary to the endogenous RNA. After reversion of cross-linking, the RNA-associated DNA and proteins are identified by sequencing and mass spectrometry (MS) analysis, respectively. The design of antisense oligonucleotides with high affinity to accessible single-stranded regions of RNAs is often a challenging step in these approaches. Moreover, these protocols are time consuming and usually require large amounts of starting material in order to generate sufficient product for MS.

CHART was first applied for the characterization of protein interactors of the human long non-coding RNAs NEAT1 (nuclear-enriched abundant transcript 1) and MALAT1 (metastasis-associated lung adenocarcinoma transcript 1) that localize to nuclear speckles and paraspeckles, respectively. Speakles and paraspeckles are nuclear bodies in close proximity to each other, comprising many RBPs, including splicing factors. Interestingly, paraspeckles have been implicated in sequestering RNAs that respond to cellular stress. Although NEAT1 and MALAT1 have specific binders, the overlap of their protein interactome suggested potential redundancy or cooperation in regulating nuclear organization around nuclear bodies.

ChiRP and RAP-MS have been applied to the discovery of the protein interactome of Xist (X-inactive specific transcript), an lncRNA required for X chromosome inactivation (XCI) of one of the two X chromosomes in female cells, enabling dosage compensation between XX females and XY males. Through an RAP-MS approach, McHugh and colleagues recently characterized ten Xist-specific interactors in mouse embryonic stem cells (mESC). One of them, SAF-A (scaffold attachment factor-A, also known as HNRNPU) was previously shown to be required for tethering Xist to the inactive X chromosome in differentiated cells, while five of these proteins are implicated in transcriptional repression, chromatin regulation, and nuclear organization. These results partially overlap with a ChiRP-MS based study conducted on the same year by Chu and colleagues, in which 81 Xist interactors were described. These proteins are mainly involved in chromatin modification, nuclear matrix and RNA remodeling pathways. Notably, this analysis reveals two sets of proteins that interact with Xist in a developmentally regulated manner, shedding light on a potential step-wise assembly of Xist binding proteins from the pluripotent state to cell differentiation.

A number of other methods based on the in vitro or in vivo screening of protein libraries have been developed. As an example, protein microarrays have been recently used to test the binding of proteins with a specific RNA in vitro. The RNA is fluorescently labeled and hybridized on a protein chip. This method allows the screening of thousands of proteins in less than a day, and relatively small amounts of RNA are required. On the other hand,
the quality of the data is less robust because it strongly depends on the proteins conditions on the array (i.e., folding, RBDs accessibility, post-translational modifications, etc).

Finally, the interactome capture, a method developed in 2012 by Castello et al., allows the simultaneous recovery and characterization of the whole proteome associated with mRNAs in living cells. This approach, originally applied to HeLa and Hek293 cells, and lately to mouse embryonic stem cells and yeast, led to the discovery of new potential RBPs lacking the canonical RNA-binding domains but enriched in features such as protein disorder and repetitive low complexity amino acid regions. The majority of these candidate RBPs have also enzymatic activity, and, while the RNA binding capacity of some of them has been validated, further evidence should be provided for the large majority of these proteins.

Future Experimental Challenges

Despite the enormous progress done in the last few decades, it is clear that each of the experimental methods offers a partial, and in some cases potentially erroneous, view on protein–RNA interactions. In this scenario, two main questions arise: to what extent in vitro interactions are relevant in vivo? Does in vivo binding imply functionality?

As regulatory regions interact with competing RBPs and RNAs, only a small percentage of in vitro determined binding sites is actually occupied in vivo. Very often, binding regions become inaccessible due to mRNA localization into granular assemblies (see section Beyond the Protein–RNA Complex: Membrane-Less Organelles). It should be also mentioned that RNAs lacking annotated binding sites are found to interact in vivo because of indirect interactions or as a result of molecule sequestration in subcellular compartments, where high concentrations favor low affinity or unspecific interactions.

Most importantly, in vivo occupancy not always indicates functionality. As a matter of fact, to unravel the complexity of the RNA biology, it will be essential to develop new tools for the integration of interaction data with global functional assays, such as for instance Ribosome profiling.

Computational Methods for Detection of Protein–RNA Interactions

A large number of computational methods address the problem of characterizing RNA partners of specific RBPs. MEME, RBPmap, SeAMotE, and RNAcontext identify motifs enriched in RNA targets. The use of sequence motifs allows discovery of targets in RNA datasets, but ab initio predictions of RNA interactions are not possible without previous experimental knowledge on the RBPs of interest.

A different class of computational approaches aims to identify RNA-binding residues and RNA-binding regions using primary-, secondary- or tertiary-structure information. Methods such as Struct-NB, PRIP, SPOT-Seq, and OPRA predict the RNA-binding ability by identifying regions in the protein surface that accommodate nucleotide chains. As three-dimensional structures are needed to perform the calculations, these approaches are limited by the existence of available templates. Yet, using a library of 1164 nonredundant protein–RNA complexes (95% sequence identity cutoff) and the folding recognition technique SPARKS X, the SPOT-Seq-RNA approach has been used to characterize 2418 novel RBPs in the human proteome of which 291 are reported in interactome capture experiments (see RNA-centric methods in section Experimental Methods for Detection of Protein–RNA Interactions).

Algorithms relying on primary-structure features have a clear advantage over tertiary-structure methods that require three-dimensional references to compare structures and atomic details to study interactions. Methods based on primary-structure exploit evolutionary information (i.e., conservation of specific residues in sequence alignments), secondary-structure propensities and information on physicochemical properties of amino acids (e.g., hydrophobicity). The binding elements are often classified using machine-learning methods such as Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB). For further information about the algorithms, we refer the Reader to reviews describing features and techniques in detail. In the following text we will provide a short description of the most used algorithms (Table 2) and their published performances.

BindN predicts DNA- and RNA-binding residues through a SVM trained on hydrophobicity, dissociation constant (pKa), and molecular mass of amino acids. The method was validated on a dataset of 100 protein–RNA complexes characterized through X-ray crystallography and nuclear magnetic resonance (NMR), and shows a cross-validation accuracy of 0.69. Using a slightly different codification of physicochemical properties and integrating them with evolutionary information, the BindN+ algorithm reaches an accuracy to of 0.78. The evolutionary information is derived from a position-specific scoring matrix (PSSM) generated by PSI-
BLAST,99 that searches against a nonredundant database of protein sequences. With the same approach, 
Pprint100 combines evolutionary information from PSSMs with SVMs, thus predicting RNA-binding sites with an accuracy of 0.81. Similarly, RNA-ProB101 employs a smoothed PSSM encoding scheme to predict RNA-binding sites on proteins and shows an accuracy of 0.89. Slightly different types of physico-chemical properties are combined in RNABindR,27 that exploits accessible surface area, sequence-specific entropy, hydropathy, secondary structure properties, and electrostatics. In the testing phase, RNABindR identifies interface residues with an accuracy of 0.85. Adding the contribution of HomPRIP,55 a sequence homology-based method for prediction of RNA-binding sites, RNABindR+ shows an accuracy of 0.83 on a larger benchmark set of 200 proteins.80 The working principles of these algorithms are based on different assumptions: the conservation, physicochemical properties or topological properties of the binding site. Hence, it is challenging to compare the algorithms; their applicability depends on the kind of dataset that needs to be analyzed. For more details, the Reader can refer to comparative studies that describe the tools in more detail.92,102

While BindN,68 Pprint,100 RNAProB,101 and RNABindR,27 predict the RNA-binding ability of individual amino acids without considering the sequence context, HMMER82 is designed to perform homology searches in protein sequences using Hidden Markov Models and multiple sequence alignments of domain families. HMMER functionality is limited to previously annotated domain-families, as the tool does not perform de novo detection of binding-domains. A recent method, catRAPID signature,83 identifies RNA-binding regions by considering physicochemical properties that are present in known RBPs.83 catRAPID signature exploits properties such as hydrophobicity, secondary structure, disorder, and burial.83 Each feature defines a unique signature, or profile, containing position-specific information arranged in sequential order from the N- to the C-terminus.28,103 In addition to the RNA-binding score, catRAPID signature predicts regions contacting RNA. On a test set of mouse proteins harboring noncanonical RNA-binding domains,70 catRAPID signature shows an accuracy of 0.71 in predicting the RNA-binding ability. When applied to newly discovered RBPs, the algorithm discriminates RBPs from nonRBPs with an area under the receiver operating curve of 0.76.29,69 In Table 3, we show the performances of catRAPID signature,83 BindN+,78 Pprint,100 and RNABindR+80 to identify nonclassical RNA-binding domains.70

We also compared the prediction performances of BindN,95 BindN+,78 Pprint,79 RNAProB,101 RNABindR+,80 and catRAPID signature83 using structural data from 90 folds, 126 families and 100 superfamilies retrieved from the SCOP database102 (Figure 3). In all the classifications, catRAPID}

**TABLE 2 | List of Computational Methods for the Identification of Protein–RNA Interactions**

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Examples</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binding motif (RNA)</td>
<td>MEME</td>
<td>de novo binding site discovery</td>
<td>High-throughput data are required as input</td>
<td>Bailey et al.94</td>
</tr>
<tr>
<td></td>
<td>SeAMotE</td>
<td></td>
<td>Sequence complexity is a limitation</td>
<td>Agostini et al.76</td>
</tr>
<tr>
<td>Binding residue</td>
<td>Pprint</td>
<td>Evolutionary information</td>
<td>RNA-binding domains cannot be identified</td>
<td>Kumar et al.79</td>
</tr>
<tr>
<td></td>
<td>BindN+</td>
<td></td>
<td></td>
<td>Wang et al.78</td>
</tr>
<tr>
<td></td>
<td>RNABindR+</td>
<td></td>
<td></td>
<td>Wala et al.80</td>
</tr>
<tr>
<td>Domain (protein)</td>
<td>HMMER</td>
<td>Domain recognition</td>
<td>Annotation of RNA-binding domains are required</td>
<td>Finn et al.95</td>
</tr>
<tr>
<td>% signature</td>
<td>catRAPID</td>
<td>Annotation of RNA-binding domains are not required</td>
<td>Single amino acid resolution has not been implemented</td>
<td>Livi et al.83</td>
</tr>
<tr>
<td>RNA–protein interaction</td>
<td>catRAPID</td>
<td>Runs on high-throughput data</td>
<td>RNA &lt; 1200 nt</td>
<td>Bellucci et al.96</td>
</tr>
<tr>
<td></td>
<td>RPISeq</td>
<td>High sensitivity</td>
<td>Protein &lt; 750 aa</td>
<td>Agostini et al.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low specificity</td>
<td>Muppirala et al.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Max 100 sequences per run</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 3 | Performances on Detecting RNA-Binding Regions**

<table>
<thead>
<tr>
<th>Method</th>
<th>ACCa</th>
<th>sensb</th>
<th>specc</th>
<th>precd</th>
</tr>
</thead>
<tbody>
<tr>
<td>catRAPID</td>
<td>0.67</td>
<td>0.76</td>
<td>0.60</td>
<td>0.65</td>
</tr>
<tr>
<td>BindN+</td>
<td>0.38</td>
<td>0.37</td>
<td>0.39</td>
<td>0.38</td>
</tr>
<tr>
<td>Pprint</td>
<td>0.47</td>
<td>0.49</td>
<td>0.45</td>
<td>0.49</td>
</tr>
<tr>
<td>RNABindR+</td>
<td>0.48</td>
<td>0.53</td>
<td>0.42</td>
<td>0.48</td>
</tr>
</tbody>
</table>

We analyzed 102 proteins containing nonclassical RNA-binding domains and 102 without annotated RNA-binding domains with catRAPID signature and three other algorithms.27,83,100 The performances are measured using a. accuracy, b. sensitivity, c. specificity, and d. precision.
Methods to predict RNA binding sites. We calculated performance of BindN, Pprint, RNAproB, and catRAPID signature on a set of proteins whose RNA-binding sites have been validated through X-ray and NMR techniques. As in a recent work, three protein classes ‘fold,’ ‘family,’ and ‘superfamily’ were retrieved from SCOP. Performances were estimated using the formula (sensitivity + specificity)/2 on 90 folds, 126 families, and 100 superfamilies (details at http://service.tartaglialab.com/static_files/shared/documentation_signature.html).

signature and BindN+ show the best performances recognizing contacts identified at a distance cutoff of 3.5 Å. We note that while BindN, Pprint, RNAproB, and RNABindR + have been designed to predict the RNA-binding ability of individual residues in protein sequences, catRAPID signature identifies binding regions (Table 2).

Only a few methods address the question whether a protein of interest interacts with a specific RNA exploiting the properties of both interaction partners, catRAPID estimates the interaction potential through van der Waals, hydrogen bonding and secondary structure propensities of both protein and RNA sequences (total of 10 features per molecule) and allows identification of binding partners with high confidence (training accuracy = 0.89; testing accuracy = 0.78; see also Cirillo et al. Table 2). catRAPID identified the interaction of the N-terminus of Fragile Mental Retardation Protein FMRP with dendritic nontranslatable brain cytoplasmic RNA BC1 before crystallographic studies determined the presence of tandem Agenet and KH domains. Similarly, catRAPID predictions for the physical binding of heterogeneous nuclear ribonucleoprotein U SAF-A with the long noncoding Xist have been recently confirmed by ChiRP-MS experiments. In addition, catRAPID can be used to perform proteome- and transcriptome-wide calculations, which is useful to complement experimental approaches, such as interactome capture (see also Experimental Methods for Detection of Protein–RNA Interactions) and to identify target partners. Owing to the complexity of the conformational space, a fragmentation procedure, based on division of polypeptide and nucleotide sequences into overlapping regions, is used for RNAs longer than 1000 nucleotides and proteins longer than 750 amino acids.

Another method called RPISeq predicts interactions combining protein and RNA features in SVM and RF classifiers. In RPISeq, RNA sequences are encoded by frequency of nucleotide tetrads (i.e., 4-mer combinations of \([A,C,G,U]\), while protein sequences are represented using 3-mer of 7 amino acid types (\([A,G,V]\), \([I,L,F,P]\), \([Y,M,T,S]\), \([H,N,Q,W]\), \([R,K]\), \([D,E]\), and \([C]\)). More specifically, the RNA is represented by a \(4 \times 4 \times 4\) or \(256\)-dimensional vector, in which each feature represents the 4-mer normalized frequency appearing in a RNA sequence (e.g., CCAU, AUUG, and GACA). The protein instead is encoded by a \(7 \times 7 \times 7\) or \(343\)-dimensional vector, where each element of the vector corresponds to the normalized frequency of the corresponding triple of amino acids in the sequence. The protein and RNA vectors serve as input for the SVM and RF to predict whether the protein–RNA pair interacts. RPISeq shows significantly high performances in the training phase (accuracy = 0.89) and high sensitivity / low specificity when applied to other sets.

**Future Computational Challenges**

catRAPID and other computational tools predict the interaction propensity of a protein–RNA pair to interact, dismissing other proteins involved in the physical binding. As a matter of fact, RBPs often form complexes and bind together to their target RNAs (see also ‘RIP’ section in Experimental Methods for Detection of Protein–RNA Interactions). A recent computational approach developed to predict transcription factor binding sites suggests that additional information retrieved from protein–protein interactions improves the performances of current predictions. Thus, as done for DNA-binding proteins, integration of protein–protein network layers could boost considerably interaction predictions. We envisage that network-based methods will shed light on the complex life of newly discovered RBPs, their dynamics and role in human diseases.
A number of reports indicate that RBPs with many partners are enriched in structural disorder.\textsuperscript{114,115} The lack of stable tertiary structure increases the ability of RBPs to interact with multiple partners\textsuperscript{116} and promotes the formation of transient ribonucleoprotein complexes, which are discussed in section Beyond the Protein–RNA Complex: Membrane-Less Organelles. Moreover, some RBPs have been found to be enriched in glycine, arginine, lysine, and tyrosine motifs\textsuperscript{29} as well as a number of other patterns\textsuperscript{71} that could promote transient interactions. Thus, to reach a complete and biologically relevant understanding of protein–RNA interactions, complex formation should be studied as a dynamic process. Data based on techniques such as surface plasmon resonance could allow measurement of interactions in real time providing both equilibrium and kinetic information for the development of new methods.\textsuperscript{16}

**BEYOND THE PROTEIN–RNA COMPLEX: MEMBRANE-LESS ORGANELLES**

RBPs and RNAs interact through different types of contacts that facilitate not only the formation of oligomeric complexes but also other types of dynamic assemblies.\textsuperscript{31} Interestingly, a number of RBPs associate through weak contacts present in structurally disordered regions\textsuperscript{117} to trigger a process known as liquid-to-liquid phase separation.\textsuperscript{118} Through liquid-liquid demixing, ribonucleoprotein assemblies form distinct compartments in the cytoplasm and nucleoplasm\textsuperscript{119} that in vivo can act as membrane-less organelles.\textsuperscript{5–7} The liquid nature of these assemblies allows rapid diffusion favoring chemical reactions on biological timescales.

Two of the most common macromolecular assemblies are processing-bodies (P-bodies) and stress granules.\textsuperscript{7,120} These assemblies are associated with mRNA degradation and storage, respectively, and are built in response to different environmental stimuli.\textsuperscript{8} P-bodies and stress granules are conserved throughout evolution and appear in yeast, plant, nematode, fly, and mammalian cells.\textsuperscript{121}

Ribonucleoprotein assemblies have special physicochemical properties that can be investigated using computational approaches such as the multiclever-Machine.\textsuperscript{122,123} With respect to the yeast proteome (globular proteins with sequence redundancy < 40%), the multicleverMachine predicts that proteins forming P-bodies and stress granules are depleted in hydrophobicity and enriched in structural disorder and RNA-binding ability (p-values < 10\textsuperscript{−5}; Fisher’s exact test; Figure 4), which is in agreement with experimental evidence.\textsuperscript{72,124} Hydrophobicity and structural disorder highly discriminate the two sets (areas under the ROC curve AUC in the range of 0.70–0.80; calculations available at http://www.tartaglialab.com/cs_multi/confirm/1207/c51f6c3f3/).

An important property of the membrane-less organelles is the wide spectrum of structural states,\textsuperscript{126} which is promoted by the different type of interactions between protein and RNA components. The RNA content influences viscoelasticity of assemblies, as well as exchange rate with the surrounding milieu, fusion and fission kinetics.\textsuperscript{127} As a matter of fact, the protein–RNA contacts can be quite diverse: strong to stabilize RNAs from synthesis to degradation or weak to bind RNAs at a precise moment and place.\textsuperscript{128} A better understanding of the physical characteristics of ribonucleoprotein components is crucial to investigate their function and dysfunction.

**RBP Assemblies and Disease**

Current literature indicates that the structure of a ribonucleoprotein assembly depends on a delicate equilibrium between protein and RNA components that is regulated by the quality control machinery (e.g., chaperones).\textsuperscript{129} Incorrect assembly/disassembly of such macrostructures can jeopardize molecular cell homeostasis. Intriguingly, RBPs found in liquid phase-separated assemblies are enriched in structural disorder that is also associated with a high risk of misfolding and the formation of toxic protein aggregates.\textsuperscript{126,129} Neurons appear to be especially susceptible to failures in proteostasis, a major source of neurodegenerative diseases. Indeed, several dementias and motor-neuron diseases are associated with accumulation of disordered RBPs, such as TAR-DNA-binding protein 43 in Amyotrophic Lateral Sclerosis (ALS) or Ataxin 1 in Ataxia.\textsuperscript{130}

In this context, chaperones perform the important task to keep protein homeostasis and prevent disease. They are upregulated upon cellular stress (e.g., diabetes-induced [glyc]oxidation stresses) to limit the accumulation of damage\textsuperscript{131} via folding assistance, blocking aberrant interactions, disaggregating proteins, and facilitating protein degradation.\textsuperscript{132–136} We speculate that the ability to form phase-separated ribonucleoprotein assemblies requires a tightly regulated control machinery, that in stress conditions and aging can be compromised and drive to formation of toxic aggregates. Thus, chaperones can be applied to develop therapeutic interventions, from controlling the initial aggregative events in the ER, to blocking the assembly of cytotoxic aggregates.
Future Challenges in the Characterization of RBP Assemblies

A major challenge in the future will be to identify the components of each ribonucleoprotein assembly and pinpoint the defects associated with RBP dysfunction.\(^{128,137}\) Currently the characterisation of solid phase separated aggregates is possible.\(^{138}\) However, the study of the liquid phase assemblies is more complex due to their fast exchange with the cell milieu, i.e., their rapid dissolution impeding its isolation.

Solid aggregates appear in many diseases, including Huntington’s, Creutzfeldt-Jakob, and Alzheimer’s disease. This suggests a common hallmark\(^ {139}\) and the existence of common pathways for pharmacotherapeutic targeting.\(^ {137,140}\) In this context, it is crucial to understand the cellular mechanisms controlling health conditions as well as the factors triggering toxicity. Interestingly, the metastable nature of the liquid assemblies offers the possibility to use endogenous biochemical pathways to reverse pathological states.\(^ {137}\) As a matter of fact, prion protein neurodegeneration can be reduced by inhibition of RNA granules pathway and stimulation of protein synthesis.\(^ {140}\)

CONCLUSION

RBPs functions include protection, modification, and transport of transcripts to their translation or

---

**FIGURE 4** | Features of yeast proteins forming ribonucleoprotein assemblies. Using physicochemical properties, the multicleverMachine approach\(^ {122,123}\) discriminates P-bodies and stress granules from other globular proteins. The datasets employed in this analysis comprise 52 P-body and 62 stress granule proteins, previously reported in experimental works,\(^ {124}\) as well as five random sets of 100 globular proteins from Astral SCOPe 2.05 (<40% sequence identity).\(^ {125}\) Three specific properties are reported: (a) nucleic acid binding, (b) hydrophobicity, and (c) disorder propensity. For each feature, enrichment or depletion of a set is indicated with a specific color: green indicates that proteins contained in P-bodies or stress granules are enriched with respect to proteome subsamples; red means depletion with respect to the proteome subsamples; yellow indicates no significant differences between the sets. As observed in previous experimental works,\(^ {63,104}\) proteins found in P-bodies and stress granules are more structurally disordered and prone to bind nucleic acids as well as less hydrophobic (p-values <10\(^ {-5}\); Fisher’s exact test). Sets and results can be accessed at http://www.tartaglialab.com/cs_multi/confirn/1207/c51e6cf3f/ where additional statistical analyses are also reported.

<table>
<thead>
<tr>
<th>Nucleic acid binding</th>
<th>(a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glojbar set1</td>
<td>Glojbar set2</td>
</tr>
<tr>
<td>P-body</td>
<td></td>
</tr>
<tr>
<td>Stress granules</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hydrophobicity</th>
<th>(b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glojbar set1</td>
<td>Glojbar set2</td>
</tr>
<tr>
<td>P-body</td>
<td></td>
</tr>
<tr>
<td>Stress granules</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disorder propensity</th>
<th>(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glojbar set1</td>
<td>Glojbar set2</td>
</tr>
<tr>
<td>P-body</td>
<td></td>
</tr>
<tr>
<td>Stress granules</td>
<td></td>
</tr>
</tbody>
</table>
degradation sites (Figure 1). The scenario is particularly complex considering the fact that not only RBPs regulate RNAs but also RNAs regulate RBPs.\textsuperscript{108} Experimental techniques aiming to reveal RNA–protein interactions will have to uncover more details of such interactions (see ‘protein-centric’ and ‘RNA-centric’ experimental methods). The study of RBPs will benefit from new computational approaches taking into account information of protein–protein networks (see section Computational Methods for Detection of Protein–RNA Interactions).

An important aspect to consider while building computational models is that many proteins exhibit RNA-binding activity without having canonical RNA-binding domains.\textsuperscript{30,71,117} As observed for protein–protein interactions, structurally disordered regions bind with high specificity and low affinity.\textsuperscript{143} Hence, noncanonical domains, which are enriched in structural disorder, could promote low affinity interactions with RNA molecules.\textsuperscript{108} Accordingly, computational methods predict that the RNA-binding ability of structural disorder is more pronounced in noncanonical RBPs, indicating that unfolded regions promote RNA interactions.\textsuperscript{108} With respect to the RNA-binding ability of full-length proteins, the contribution of disorder is high at low-interaction propensities, which suggests that a large number of transcripts can be targeted by noncanonical domains.\textsuperscript{108} How many other RBPs lacking canonical binding domains exist in nature?\textsuperscript{69,83} The repertoire of RBPs needs to be expanded in order to link transcriptomic with metabolomic properties,\textsuperscript{144} which will involve the development of new computational strategies.\textsuperscript{80,83}

Most importantly, future computational and experimental investigations will have to focus on noncanonical protein–RNA complexes.\textsuperscript{145} Indeed, transient interactions, promoted by structurally disordered regions in RBPs, often induce phase-separation in the nucleus or cytoplasm (see section Beyond the Protein–RNA Complex: Membrane-Less Organelles).\textsuperscript{6} The functionality of phase-separated protein–RNA assemblies is still under investigation but appearance of solid protein–RNA aggregates has been linked with devastating diseases, such as for instance ALS (see section RBP Assemblies and Disease).\textsuperscript{137,140} Thus, studying protein–RNA assemblies will be essential to better understand the etiopathogenesis of specific diseases and to design new therapeutic strategies.

**ACKNOWLEDGMENTS**

The research leading to this work has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), through the European Research Council, under grant agreement RIBOMY-LOME_309545 (Gian Gaetano Tartaglia), and from the Spanish Ministry of Economy and Competitiveness (BFU2014-55054-P). We also acknowledge support from AGAUR (2014 SGR 00685) and the Spanish Ministry of Economy and Competitiveness ‘Centro de Excelencia Severo Ochoa 2013–2017’ (SEV-2012-0208).

**REFERENCES**

10. Nussbacher JK, Batra R, Lagier-Tourenne C, Yeo GW. RNA-binding proteins in neurodegeneration:


Characterization of RNA-binding proteins


88. Zhao H, Yang Y, Janga SC, Kao CC, Zhou Y. Prediction and validation of the unexplored RNA...


111. Cirillo D, Botta-Orfila T, Tartaglia GG. By the company they keep: interaction networks define the binding ability of transcription factors. *Nucl Acids Res* 2015, 43:e125.


118. Bolognesi B, Lorenzo N, Dhar R, Cirillo D, Baldrighi M, Tartaglia G, Lehner B. A concentration-dependent liquid phase separation can cause toxicity upon


