Published in final edited form as:

Curr Opin Genet Dev. 2013 February ; 23(1): 20–28. doi:10.1016/j.gde.2013.01.003.

Multi-Disciplinary Methods to Define RNA-Protein Interactions and Their Regulatory Networks

Manuel Ascano^{*}, Stefanie Gerstberger^{*}, and Thomas Tuschl[#]

Howard Hughes Medical Institute and Laboratory for RNA Molecular Biology, The Rockefeller University, 1230 York Ave, New York, NY, 10065, United States

Abstract

The advent of high-throughput technologies including deep-sequencing and protein mass spectrometry is facilitating the acquisition of large and precise data sets towards the definition of post-transcriptional regulatory networks. While early studies that investigated specific RNA-protein interactions in isolation laid the foundation for our understanding of the existence of molecular machines to assemble and process RNAs, there is a more recent appreciation of the importance of individual RNA-protein interactions that contribute to post-transcriptional gene regulation. The multitude of RNA-binding proteins (RBPs) and their many RNA targets has only been captured experimentally in recent times. In this review, we will examine current multidisciplinary approaches towards elucidating RNA-protein networks and their regulation.

Introduction

The challenges of understanding post-transcriptional gene regulation are manifold. They include the cataloging of RBPs and their respective RNA binding sites, interpretation of cell-type-specific differences in RBP and target RNA isoforms, and attributing a regulatory or biochemical function to the interactions. Since RNA is a moving target inside cells, consideration of its subcellular localization and those of its interacting proteins are relevant to their encoded function. Upon transcription of a single gene, tens to thousands of RNA copies are synthesized and undergo a series of maturation steps that occur in different locations within the cell. Therefore, RNA-based regulation is dependent on the concentration of the target RNA and its potential regulators, in contrast to transcriptional regulation, where the number of targeting sites on DNA is at a first approximation independent of the expression program of the cell (Figure 1).

Messenger RNAs (mRNAs) are subject to various maturation steps that involve protein binding and result in formation of mRNA ribonucleoprotein particles (mRNPs), whose respective compositions remain poorly defined albeit are known to control constitutive and alternative mRNA splicing, 5'-methyl-guanosine capping and 3' polyadenylation, export out of the nucleus, and ultimately loading onto ribosomes for protein translation. mRNA-binding proteins (mRBPs) were shown to have hundreds to thousands of different mRNA targets and individual mRNAs may harbor many sites occupied in a spatial and temporal

© 2013 Elsevier Ltd. All rights reserved.

[#]Corresponding author: ttuschl@rockefeller.edu.

^{*}These authors contributed equally to this work.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

order as they mature and decay inside cells. For a given mRNA-binding protein (mRBP), in order to elicit its respective function a sufficient fraction of the cellular copies of its target transcripts are required to be occupied. RNA binding proteins have been implicated as posttranscriptional modulators of gene networks in a wide range of human diseases including cancer. To understand how dysregulation of posttranscriptional networks can lead to human diseases, the breadth of RBP targets and their specific binding sites need to be determined. The capture and functional interpretation of RNA-mRBP interactions at the transcriptome-wide level can be staggering and implicitly argues for the need to combine focused biochemical and structural characterization with large-scale systems-biology approaches.

Deep-sequencing of RNA cDNA libraries from crosslinked RNPs

Covalent crosslinking of RNAs to proteins by exposure to short- or long-wavelength ultraviolet (UV) light coupled with deep-sequencing of cDNA libraries derived from the isolated crosslinked RNAs allows investigators to identify the RNA targets of RBPs with previously unprecedented scale and depth. The majority of methods for identification of binding sites of RBPs employ the use of an enrichment protocol, either by immunoprecipitation of the RBP of interest[1–3] or affinity purification of an ectopically expressed tagged form of the RBP, e.g. by a His-tag[4,5]. The methods are also distinguished by different strategies for separating signal from background RNA inevitably present in these approaches. Crosslinking and immunoprecipitation (CLIP) methods coupled with deep-sequencing were shown to be suitable for transcriptome-wide mapping of the binding sites of RBPs or microRNPs in living cells and uncovered post-transcriptional regulatory networks in cells and tissues[6]. CLIP methods yield sequencing reads corresponding to RNA segments at or near the binding sites. Binding sites typically contain a defined, often short RNA sequence and/or a structural element, known as the RNA recognition element (RRE), which specifically binds to amino acid residues located in the RNA-binding domain (RBD) of an RBP. Isolation and characterization of crosslinked mRNA fragments by sequencing represents an advantage over earlier methods such as RIP-chip[7] and RIP-seq[8] that determine enrichment of full-length target mRNAs in IP versus lysate but do not directly isolate the binding site. The absence of covalent crosslinks also requires the use of less stringent, non-denaturing RNP isolation protocols yielding more background. Nevertheless, RIP-chip or RIP-seq without crosslinking are valuable methods as they can provide a measure of the affinity of an RBP towards its mRNA targets. Combining RIP-chip with CLIP allows for a detailed analysis relating binding site information with overall mRNA-binding affinity offering a starting point for the characterization of interesting post-transcriptionally regulated mRNAs. Such a combined approach has been successfully applied to identify targets and begin to address the mechanism of target selectivity for a number of RBPs[9,10], including HuR, a protein implicated in numerous cancers. Elevated expression of HuR has been detected in various cancers with poor outcome prognosis[11]. The combination of RIP-chip and CLIP identified a transcriptome-wide list of targets and indicated that HuR functions at several levels of post-transcriptional regulation including pre-mRNA processing and mature mRNA stability. Further, an analysis of the proximity of HuR with miRNA binding sites indicated that HuR binding sites that were within 10 nt or fewer of a miRNA target region, prevented miRNA-dependent repression.

Protein mass spectrometry from crosslinked RNPs

The identification of specific RNA-interacting proteins from cell lysates is accomplished by mass spectrometry. Early approaches used conventional chromatography[12,13], or affinity purification using antibodies raised to known mRBPs to identify proteins present in

mRNPs[14]. Alternatively, RNA conjugates can be incubated with lysates for *in vitro* RNP assembly, then affinity-purified to identify interacting proteins[15]. In order to distinguish proteins that directly contacted RNA from proteins that are only present in assembled RNPs due to protein-protein interactions, RNA-protein covalent bonds are formed by UV-crosslinking to enable harsher purification steps required for removal of non-covalently bound proteins[16]. Highly purified crosslinked RNP products were also used to precisely map the amino acids directly contacting RNA[17,18]. Crosslinked RNPs are typically treated upon cell lysis with RNases to obtain smaller size (20–50 nt) RNA fragments, which are radiolabeled upon immunopurification of the specific RNPs and resolved by SDS-PAGE fractionation for its detection and isolation.

The identification of RBPs specifically binding to a given target mRNA segment represents a challenging biochemical problem, because in cell lysates, all RBPs will simultaneously compete for the target RNA in the absence of an otherwise temporally and spatially organized process. As an example, an intronic sequence RNA, which is normally retained in the nucleus may be bound by cytoplasmic RBPs in lysates, and vice versa. Furthermore, abundant RBPs with less sequence specificity or short RREs effectively compete with less abundant more specifically binding RBPs and yield high background binding. To address this background problem, SILAC mass spectrometry approaches were recently developed[19,20]. In these assays, RNA bait and reference segments of about 100–200 nt are incubated in distinct isotope-labeled cell lysates, followed by the determination of isotope-enrichment of proteins bound to the RNA bait versus the reference. These approaches allow for medium-high throughput of hundreds of RNA segments, which can be useful to scan the sequence of a selected mRNA and find both its potential interactors and regulatory regions.

In contrast to RNP assembly using cell lysates, *in vivo* assembled RNPs avoid potential biochemical artifacts. This possibility prompted the development of RNA tags that are joined to and therefore co-expressed with the RNA of interest[21–24]. The most common approach introduces one or more copies of an RNA-encoded aptamer, which is recognized by a protein (e.g. MS2 coat protein or streptavidin) or small molecule (e.g. streptomycin) that is immobilized on chromatography matrix. More recently, SILAC was coupled to RNA-based aptamer approaches as a means of identifying and quantitating affinity-purified and *in vivo* assembled RNPs[25]. Furthermore, significant advances in the visualization of mRNAs in living cells can be attributed to the use of *in vivo* RNA tags[26,27].

In an effort to catalog and discover new mRBPs, recent modern mass spectrometric approaches were applied to determine the protein composition of UV-crosslinked polyadenylated (polyA) cellular RNAs (Figure 2A). Crosslinked mRNPs were isolated under protein-denaturing conditions from cultured human cells using oligoT beads[28,29]. The Hentze group identified 860 crosslinked proteins in HeLa cells, while the Landthaler group identified 797 in HEK293 cells, 554 of which were jointly identified, i.e. corresponding to a 64% or 70% overlap respectively. A comparison to a table composed of 1,783 human genes encoding known RBPs or proteins annotated to contain RBDs, showed that the majority of the Hentze and Landthaler identified proteins, 618 (72%) and 624 (76%), respectively were previously known. Within this category, 480 proteins were jointly identified corresponding to a 78% and 77% overlap, respectively. The abundance and specificity of expression for overlapping and non-overlapping RBPs were comparable between both groups, indicating that differences in the experimental methods were responsible for the discrepancy. The overlap between proteins newly identified as candidate RNA-binding was 74 proteins, corresponding to a 31% and 43% overlap between the Hentze and Landthaler datasets, respectively. Given that abundance and specificity of expression for the novel RNA-binding protein candidates was similar and comparable to the

canonical RBPs, the approximately 2-fold reduced overlap between candidate RBPs compared to known RBPs may suggest an increased false positive rate or differences in the bioinformatic analyses when setting cut-offs for defining novel RBPs. Both laboratories performed further validation studies for a fraction (<5%) of their candidate RBPs, thereby providing additional RNA-binding evidence (Table 1). Some of these novel RBPs belong to large ubiquitously expressed protein families whose other members were not captured by their methods, which is difficult to rationalize based on the expectation that specific protein domains, presumably conserved among family members, are required for binding RNA. A domain analysis of the full set of candidate RBPs of both studies showed an accumulation of nucleotide (ATP, GTP) and nucleotide analog (e.g. NAD) domains, which may be cautiously interpreted to indicate that these domains also interact with longer RNAs, possibly at their termini, or might imply that their activity is modulated by interacting with mRNAs.

Some of the novel proteins contain repetitive, low complexity amino acid repeat motifs, such as arginine glycine (RG) rich repeats which have been previously demonstrated in RNA binding interactions[30]. It is plausible that repetitive amino acid repeat motifs could confer RNA-binding[28] although such a model does not fully address why family members of newly identified RBPs were not found. Another explanation may be that these proteins evolved further to contain unique RNA-binding surfaces within largely pre-existing domain structures, or vice versa, RNAs evolved their sequence to modulate the activity of these specific proteins. This would imply that some RREs could be as complex as RNA aptamers to specifically bind to a given protein. For such complex and unique interactions to be identified by mass spectrometry, it would imply that the target mRNA was present at very high copy numbers.

Structural elucidation of RNA-protein interactions

The molecular understanding of RNA-protein interaction is derived from the elucidation of the solution or crystal structure of RBPs with a cognate target. It is important that such studies be put into perspective with large-scale sets of data that can either agree with the solved isolated structure, or prompt the need for additional study. Structures of the PUM and NOVA protein families are known examples where the molecular contacts observed in RNA co-crystal structures correspond to the RREs deduced from large-scale natural target RNA studies. Extensive characterization of PUM2 targets by microarray[31] and CLIP[3] studies largely confirmed earlier biochemical assays[32] and structural results[33]. PUM RBPs appear unique among RBPs, as each RBD repeat can contact a single nucleotide base of an RNA molecule of the consensus sequence (UGUANAUA)[34]. PUM proteins bind their targets highly selectively and are among the most sequence-specific RBPs. The NOVA proteins are composed of three KH domains and its RREs are composed of YCAY (Y= U or C)[35]. X-ray crystallography was first performed using the third KH (KH3) domain and an RRE-comprising RNA target[36,37], but a recent crystal structure of the KH1 and 2 domains[38] showed that these other domains also recognized YCAY elements. These findings were consistent with UV254 CLIP experiments[2], and further indicated that RNA targets with repeated YCAY elements can be bound by a single RBP.

For many RBPs, structural results are yet to be fully-resolved with transcriptome-wide studies. For example, the KH1 and 2 domains of FMRP, which adapt a typical KH domain protein structure[39] and are therefore expected to bind a 4-nt RRE typical of other KH domain proteins, were initially shown to bind to a kissing complex RNA[40], that has not been demonstrated to occur within natural targets[10]. No structure has yet been solved for FMRP with a KH-specific target RNA ligand.

The RRE for members of the insulin-like growth factor 2 mRNA-binding protein (IGF2BP) family has been identified as CAUH (H= A, C, or U) by PAR-CLIP. RIP-chip experiments, although they don't have the same site-specific resolution, found multiple CA-dinucleotide-containing motifs enriched[41]. Another group obtained an RNA ligand from SELEX experiments binding to the KH3 and 4 domain of IGF2BP1, which contain MCAY (M= A or C) but also an additional G-rich element[42]. They were able to obtain NMR solution structures of two complexes, in which the order of the MCAY and G-rich elements could be interchanged[43]. In all studies, it appears that a CA dinucleotide is prominent among the targets identified for IGF2BP proteins, albeit the breadth of its functionally relevant targets remains unclear. Given that IGF2BP proteins are known to regulate gene expression of proliferation and differentiation pathways, it is not surprising that two members (IGF2BP1 and IGF2BP3) have been found highly expressed in some aggressive cancers[44]. Thus a more clear understanding of the IGF2BP mRNA targets should be helpful in defining the role of these RBPs in human disease.

The RREs of many mRBPs comprise only short sequences that can frequently occur by chance in longer RNAs, which in turn is consistent with the many target sites identified in transcriptome-wide studies. Even though these RREs are short, they are nevertheless recognized sequence-specifically, as many biochemical follow-up studies have shown. Some RBPs, however, depend on specific cellular mechanisms that direct their association to only a subset of RRE-containing cellular targets. Gemin2 is such a protein factor that directs the specific assembly of Sm protein rings on snRNAs at their RRE composed of AUUUUUA, which is also present in many other RNAs[45]. Gemin 2 binds the pentameric, pre-assembled Sm protein core on snRNAs in an snRNA-dependent cooperative manner. hnRNP and SR proteins are also known to bind short degenerate sequences[46] cooperatively, to achieve specificity.

Additional complexity is added by the many RBPs that contain combinations of RNA-binding domains in modular organization. The contribution of individual RBDs, and their relation to each other in target recognition remains unknown. Multidisciplinary approaches combining broad-scale studies and conventional biochemistry *in vivo* and *in vitro* are therefore essential for the elucidation of binding specificity of multi-RBD RBPs, and to determine whether basic rules of RNA-protein interactions and dominance of specific RBDs can be established.

The transition from discovery to functional annotation

Following the discovery of specific binding sites on mRNAs for a given RBP, functional assays are required to assess the impact of binding on posttranscriptional regulation. These assays traditionally comprise capturing changes in mRNA stability or alternative splicing. More recently, a new assay to address the role of RBPs on mRNA translation was introduced termed ribosome profiling[47]. It consists of an improvement of a ribosome footprinting approach[48] by using deep-sequencing, which recovers mRNAs from translating ribosomes for inferring rates of protein synthesis[49,50] (Figure 2B). Nuclease-digested ribosome-associated mRNAs are biochemically fractionated by density centrifugation and oligonucleotide size selection, converted to a cDNA library and deep-sequenced. Total RNA, as normalization control, is similarly processed for sequencing and the ratio between recovered ribosome protected mRNA fragments and the density of sequence reads of the same mRNAs from the control are compared. Prior to lysing cells, addition of different drugs can be used to stall ribosome translation, at initiation or elongation, and enable precise mapping of (alternative) translation initiation sites and measurement of translation kinetics.

Ribosome profiling gives a measurement of actively translating RNAs and provides a sensitive tool to assess changes to the nascent proteome within different cell types or conditions. The approach also identified a class of short highly-translated ORFs termed short polycistronic ribosome-associated RNAs (sprcRNAs), within previously annotated regions of long-intergenic non-coding RNAs (lincRNAs)[49]. Since the metric used for ribosome profiling studies normalize ribosome protected fragment sequences over RNAseq data across transcripts, it is not well understood if 5'-UTR scanning by the ribosome is discernible from true sites of active translation. Neither the functions of many lincRNAs are well characterized nor were protein products of sprcRNAs isolated.

Given that alternative 3' UTRs can dictate the repertoire of RBPs that associate and regulate mRNAs, the detailed annotation of the 3' ends of mRNAs represents an area of active research. Alternative cleavage and polyadenylation of the 3' UTRs of mRNAs is closely tied to cellular function and growth and various methods were developed to capture and sequence the 3' UTRs of mRNAs, typically beginning with an oligoT-based approach[51–54]. Shortening of 3'UTRs and usage of alternative polyadenylation sites have been implicated as mechanisms to evade posttranscriptional gene regulation in cancer. Understanding the regulation of alternative 3' UTRs of mRNAs in higher eukaryotes should provide additional and important determinants for assessing the potential regulatory impact of RBPs on its targets in human disease.

The path of gene expression, from transcription to translation, is a multi-step and ordered series of regulated events for which branched pathways are only now being uncovered. Unlike transcription, most post-transcriptional responses are rarely over 2-to-4 fold changes, at least within the resolution of current assays. Nevertheless, many genetic diseases have been identified where mutation of RBPs or related processing factors, most of which are ubiquitously expressed, still result in a large spectrum of diseases. Prominent among these are neurological, neuromuscular disorders, metabolic and mitochondrial diseases and human cancers. One of the more challenging problems with many of these diseases and the RBPs implicated, are the striking tissue-specific effects they have despite knowing that the RBPs of interest have a greater breadth of expression than the affected tissues of a given disease. RBPs, like FMR1, have pleiotropic syndromic clinical manifestations that extend beyond its well-documented importance to human cognition and memory. With respect to cancer, developmentally regulated RBPs such as LIN28 family members (LIN28A and LIN28B) or MSI1 (Musashi) are found frequently overexpressed in certain types of tumors. LIN28 expression is developmentally regulated and is normally expressed in embryonic cell types. The upregulation of LIN28A or LIN28B correlates with advanced tumors and poor prognosis[55]. MSI1 is normally expressed in multi-potent stem cells in the brain, breast, and hair follicles but found overexpressed in medulloblastomas, chronic myelogenous leukemia, and acute myeloid leukemia[56–58]. Recent work with Musashi indicates that it can regulate critical genes involved in cancer-related pathways. It is therefore plausible that many RBPs affect the same gene networks across different cellular contexts and that such gene pathways, when dysregulated, manifest as different seemingly unrelated phenotypes (or are compensated in a tissue-specific fashion). To the extent that basic modes of post-transcriptional gene regulation can be studied in any cellular system, the phenotypic impact is variable and remains subjective to transcriptomic contexts. Accordingly, such complexities of varying transcriptome and post-transcriptional regulatory networks must be investigated with an eye towards network analysis and systems-oriented multi-disciplinary strategies, in conjunction with focused validation studies.

References

1. König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol.* 2010; 17:909–915. [PubMed: 20601959]
- 2*. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature.* 2008; 456:464–469. The first published report of UV RNA-protein crosslinking methods combined with next generation RNA sequencing methods to obtain genome-wide binding sites of RNA binding proteins. [PubMed: 18978773]
- 3*. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano PhDM Jr, Jungkamp A-C, Munschauer M, et al. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell.* 2010; 141:129–141. Identification of genome-wide RNA targets by PAR-CLIP for the RNA binding proteins PUM2, QKI, IGF2BP, AGO(1–4), and TNRC6(A–C). [PubMed: 20371350]
4. Wang Z, Kayikci M, Briese M, Zarnack K, Luscombe NM, Rot G, Zupan B, Curk T, Ule J. iCLIP Predicts the Dual Splicing Effects of TIA-RNA Interactions. *PLoS Biol.* 2010; 8:e1000530. [PubMed: 21048981]
5. Granneman S, Kudla G, Petfalski E, Tollervey D. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc Nat Acad Sci.* 2009; 106:9613–9618. [PubMed: 19482942]
- 6**. Ascano, PhDM., Jr; Hafner, M.; cekan, P.; Gerstberger, S.; Tuschl, T. Identification of RNA-protein interaction networks using PAR-CLIP. *WIREs RNA.* 2012; 3:159–177. This review gives a comparative overview of CLIP-based methods for the identification of the *in vivo* target spectrum of RNA-binding proteins and their genome wide RNA-protein interacting sites. [PubMed: 22213601]
- 7**. Morris AR, Mukherjee N, Keene JD. Systematic analysis of posttranscriptional gene expression. *Wiley Interdiscip Rev Syst Biol Med.* 2010; 2:162–180. A comprehensive review about the number of studies on RNA binding proteins for which genome-wide identification of target sites of RNA interacting proteins by RNA-protein crosslinking and non-crosslinking next generation sequencing methods have been carried out. [PubMed: 20836020]
- 8*. Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song J-J, Kingston RE, Borowsky M, Lee JT. Genome-wide Identification of Polycomb-Associated RNAs by RIP-seq. *Mol Cell.* 2010; 40:939–953. First demonstration of the use of RIP-seq for identification of RNA targets. [PubMed: 21172659]
9. Mukherjee N, Corcoran DL, Nusbaum JD, Reid DW, Georgiev S, Hafner M, Ascano PhDM Jr, Tuschl T, Ohler U, Keene JD. Integrative Regulatory Mapping Indicates that the RNA-Binding Protein HuR Couples Pre-mRNA Processing and mRNA Stability. *Mol Cell.* 2011; 43:327–339. [PubMed: 21723170]
10. Ascano M, Mukherjee N, Bandaru P, Miller JB, Nusbaum JD, Corcoran DL, Langlois C, Munschauer M, Dewell S, Hafner M, et al. FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature.* 2012; 492:382–386. [PubMed: 23235829]
11. Abdelmohsen K, Gorospe M. Posttranscriptional regulation of cancer traits by HuR. *WIREs RNA.* 2010; 1:214–229. [PubMed: 21935886]
12. Garcia-Blanco MA, Jamison SF, Sharp PA. Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. *Genes Dev.* 1989; 3:1874–1886. [PubMed: 2533575]
13. Patton JG, Mayer SA, Tempst P, Nadal-Ginard B. Characterization and molecular cloning of polypyrimidine tract-binding protein: a component of a complex necessary for pre-mRNA splicing. *Genes Dev.* 1991; 5:1237–1251. [PubMed: 1906036]
14. Dreyfuss G, Choi YD, Adam SA. Characterization of heterogeneous nuclear RNA-protein complexes in vivo with monoclonal antibodies. *Mol Cell Biol.* 1984; 4:1104–1114. [PubMed: 6204191]

15. Sharma S. Isolation of a sequence-specific RNA binding protein, polypyrimidine tract binding protein, using RNA affinity chromatography. *Methods Mol Biol.* 2008; 488:1–8. [PubMed: 18982280]
16. Steen H, Jensen ON. Analysis of protein-nucleic acid interactions by photochemical cross-linking and mass spectrometry. *Mass Spectrom Rev.* 2002; 21:163–182. [PubMed: 12476441]
17. Williams KR, Konigsberg WH. Identification of amino acid residues at interface of protein-nucleic acid complexes by photochemical cross-linking. *Meth Enzymol.* 1991; 208:516–539. [PubMed: 1779846]
18. Urlaub H, Hartmuth K, Kostka S, Grelle G, Luhrmann R. A general approach for identification of RNA-protein cross-linking sites within native human spliceosomal small nuclear ribonucleoproteins (snRNPs). Analysis of RNA-protein contacts in native U1 and U4/U6.U5 snRNPs. *J Biol Chem.* 2000; 275:41458–41468. [PubMed: 11006293]
- 19*. Butter F, Scheibe M, Mörl M, Mann M. Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc Nat Acad Sci.* 2009; 106:10626–10631. This group took advantage of SILAC as a method to quantify the enrichment of specific RBPs on >100 nt sized target RNAs for the identification and quantitation of RNA-interacting proteins. [PubMed: 19541640]
20. Scheibe M, Butter F, Hafner M, Tuschl T, Mann M. Quantitative mass spectrometry and PAR-CLIP to identify RNA-protein interactions. *Nucl Acids Res.* 2012; 40:9897–9902. [PubMed: 22885304]
21. Bardwell VJ, Wickens M. Purification of RNA and RNA-protein complexes by an R17 coat protein affinity method. *Nucl Acids Res.* 1990; 18:6587–6594. [PubMed: 1701242]
22. Zhou Z, Reed R. Purification of functional RNA-protein complexes using MS2-MBP. *Curr Prot Mol Biol.* 2003; Chapter 27(Unit 27.3)
23. Windbichler N, Schroeder R. Isolation of specific RNA-binding proteins using the streptomycin-binding RNA aptamer. *Nat Protocols.* 2006; 1:637–640.
24. Walker SC, Scott FH, Srisawat C, Engelke DR. RNA affinity tags for the rapid purification and investigation of RNAs and RNA-protein complexes. *Methods Mol Biol.* 2008; 488:23–40. [PubMed: 18982282]
- 25*. Tsai BP, Wang X, Huang L, Waterman ML. Quantitative profiling of *in vivo*-assembled RNA-protein complexes using a novel integrated proteomic approach. *Mol Cell Proteomics.* 2011; 10:M110.007385. To identify *in vivo* assembled RNA-protein complexes, this group inserted MS2 recognition stem loops within an RNA target of interest and co-expressed the MS2 coat protein C-terminally tagged with an *in vivo* biotinylation site (HTBH) to allow for capture and tandem affinity purification of assembled complexes. Cells, prior to RNP isolation, were incubated in stable isotope-specific amino acid media (SILAC), allowing for more effective identification of RNA-binding proteins.
26. Tyagi S. Imaging intracellular RNA distribution and dynamics in living cells. *Nat Meth.* 2009; 6:331–338.
27. Bertrand E, Chartrand P, Schaefer M, Shenoy SM, Singer RH, Long RM. Localization of ASH1 mRNA particles in living yeast. *Mol Cell.* 1998; 2:437–445. [PubMed: 9809065]
- 28**. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, et al. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell.* 2012; 149:1393–1406. In combination with Baltz et al. this publication uses UV-crosslinking methods to isolate RNA-binding proteins at a proteome-wide level. This report discusses prion-like low-complexity regions in mRNA interacting proteins, for which they found a significant enrichment. [PubMed: 22658674]
- 29**. Baltz AG, Munschauer M, Schwanhaussner B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M, et al. The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Mol Cell.* 2012; 46:674–690. In combination with Castello et al. this report uses PAR-CLIP UV-crosslinking to broadly identify known and novel RNA-binding proteins. The authors also give a transcriptome-wide map of RNA interacting sites on mRNA transcripts and identify putative RNA *cis*-regulatory elements with disease-associated polymorphisms. [PubMed: 22681889]
30. Kiledjian M, Dreyfuss G. Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. *EMBO J.* 1992; 11:2655–2664. [PubMed: 1628625]

- 31*. Gerber AP, Luschnig S, Krasnow MA, Brown PO, Herschlag D. Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in *Drosophila melanogaster*. *Proc Nat Acad Sci*. 2006; 103:4487–4492. This crystallographic study dissects the modular recognition of the Pumilio repeats on the amino acid - RNA interaction level. [PubMed: 16537387]
32. White EK, Moore-Jarrett T, Ruley HE. PUM2, a novel murine puf protein, and its consensus RNA-binding site. *RNA*. 2001; 7:1855–1866. [PubMed: 11780640]
33. Wang X, Zamore PD, Hall TM. Crystal structure of a Pumilio homology domain. *Mol Cell*. 2001; 7:855–865. [PubMed: 11336708]
34. Wang X, McLachlan J, Zamore PD, Hall TM. Modular recognition of RNA by a human pumilio-homology domain. *Cell*. 2002; 110:501–512. [PubMed: 12202039]
35. Buckanovich RJ, Darnell RB. The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. *Mol Cell Biol*. 1997; 17:3194–3201. [PubMed: 9154818]
36. Lewis HA, Chen H, Edo C, Buckanovich RJ, Yang YY, Musunuru K, Zhong R, Darnell RB, Burley SK. Crystal structures of Nova-1 and Nova-2 K-homology RNA-binding domains. *Cell*. 1999; 7:191–203.
37. Jensen KB, Musunuru K, Lewis HA, Burley SK, Darnell RB. The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain. *Proc Nat Acad Sci*. 2000; 97:5740–5745. [PubMed: 10811881]
38. Teplova M, Malinina L, Darnell JC, Song J, Lu M, Abagyan R, Musunuru K, Teplov A, Burley SK, Darnell RB, et al. Protein-RNA and Protein-Protein Recognition by Dual KH1/2 Domains of the Neuronal Splicing Factor Nova-1. *Cell*. 2011; 19:930–944.
39. Valverde R, Pozdnyakova I, Kajander T, Venkatraman J, Regan L. Fragile X mental retardation syndrome: structure of the KH1-KH2 domains of fragile X mental retardation protein. *Structure*. 2007; 15:1090–1098. [PubMed: 17850748]
40. Darnell JC, Fraser CE, Mostovetsky O, Stefani G, Jones TA, Eddy SR, Darnell RB. Kissing complex RNAs mediate interaction between the Fragile-X mental retardation protein KH2 domain and brain polyribosomes. *Genes Dev*. 2005; 19:903–918. [PubMed: 15805463]
41. Jonson L, Vikesaa J, Krogh A, Nielsen LK, Hansen TV, Borup R, Johnsen AH, Christiansen J, Nielsen FC. Molecular composition of IMP1 ribonucleoprotein granules. *Mol Cell Proteomics*. 2007; 6:798–811. [PubMed: 17289661]
42. Chao JA, Patskovsky Y, Patel V, Levy M, Almo SC, Singer RH. ZBP1 recognition of β -actin zipcode induces RNA looping. *Genes Dev*. 2010; 24:148–158. [PubMed: 20080952]
- 43*. Patel VL, Mitra S, Harris R, Buxbaum AR, Lionnet T, Brenowitz M, Girvin M, Levy M, Almo SC, Singer RH, et al. Spatial arrangement of an RNA zipcode identifies mRNAs under post-transcriptional control. *Genes Dev*. 2012; 26:43–53. This report focuses on the NMR solution structure of a portion of the IGF2BP1 in order to determine how it contacts RNA. The work done here should be compared to transcriptome-wide studies of IGF2BP1, which also report RNA recognition elements. [PubMed: 22215810]
44. Bell JL, Wächter K, Mühleck B, Pazaitis N, Köhn M, Lederer M, Huttelmaier S. Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? *Cell Mol Life Sci*. 2012; 1007/s00018-012-1186-z
- 45*. Zhang R, So BR, Li P, Yong J, Glisovic T, Wan L, Dreyfuss G. Structure of a Key Intermediate of the SMN Complex Reveals Gemin2's Crucial Function in snRNP Assembly. *Cell*. 2011; 146:384–395. This crystallographic study gives insight into the assembly process of the Sm protein core mediated by the SMN complex. This study explains mechanistically how the mutation in SMN causing spinal muscular atrophy (SMA) disrupt snRNP biogenesis and thereby leads to SMA pathophysiology. [PubMed: 21816274]
46. Singh R, Valcárcel J. Building specificity with nonspecific RNA-binding proteins. *Nat Struct Mol Biol*. 2005; 12:645–653. [PubMed: 16077728]
- 47*. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, NY)*. 2009; 324:218–223. A milestone for the translation field: The first report of ribosome profiling yielding

genome-wide ribosome occupation sites on mRNA transcripts. The use of next generation sequencing gives a transcriptome snapshot of active translation sites.

48. Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc Nat Acad Sci*. 2003; 100:3889–3894. [PubMed: 12660367]
- 49*. Ingolia NT, Lareau LF, Weissman JS. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell*. 2011; 147:789–802. Ribosome profiling report on ribosome initiation versus stalling and translation. The authors find that mRNAs display translation at upstream alternative open reading frames. Unexpectedly, previously reported long noncoding RNAs were also found associated with polysomes, suggesting that they too are translated. [PubMed: 22056041]
50. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*. 2010; 466:835–840. [PubMed: 20703300]
51. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*. 2008; 320:1643–1647. [PubMed: 18566288]
- 52*. Mayr C, Bartel DP. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell*. 2009; 138:673–684. This study shows that cancer cell lines often express shorter 3'UTRs isoforms resulting from usage of alternative polyadenylation sites (APA). The APA events resulted in increased mRNA stability and protein amounts, thereby giving a proliferative advantage to these cancer cell lines. [PubMed: 19703394]
- 53*. Berg MG, Singh LN, Younis I, Liu Q, Pinto AM, Kaida D, Zhang Z, Cho S, Sherrill-Mix S, Wan L, et al. U1 snRNP Determines mRNA Length and Regulates Isoform Expression. *Cell*. 2012; 150:53–64. This report describes the importance of the high abundance of U1 snRNP splicing factor during acute bursts of transcriptional activation. U1 shortage leads to effects in mRNA length and isoform expression. [PubMed: 22770214]
54. Jan CH, Friedman RC, Ruby JG, Bartel DP. Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. *Nature*. 2011; 469:97–101. [PubMed: 21085120]
55. Thornton JE, Gregory RI. How does Lin28 let-7 control development and disease? *Trends Cell Biol*. 2012; 22:474–482. [PubMed: 22784697]
- 56*. Vo DT, Subramaniam D, Remke M, Burton TL, Uren PJ, Gelfond JA, de Sousa Abreu R, Burns SC, Qiao M, Suresh U, et al. The RNA-binding protein Musashi1 affects medulloblastoma growth via a network of cancer-related genes and is an indicator of poor prognosis. *Am J Pathol*. 2012; 181:1762–1772. This study uses RIP-chip to identify mRNAs targets directly bound by Musashi and analyzes expression profiles of medulloblastoma samples to identify gene networks regulated by Musashi. [PubMed: 22985791]
57. Kharas MG, Lengner CJ, Al-Shahrour F, Bullinger L, Ball B, Zaidi S, Morgan K, Tam W, Paktinat M, Okabe R, et al. Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. *Nat Med*. 2010; 16:903–908. [PubMed: 20616797]
58. Ito T, Kwon HY, Zimdahl B, Congdon KL, Blum J, Lento WE, Zhao C, Lagoo A, Gerrard G, Foroni L, et al. Regulation of myeloid leukaemia by the cell-fate determinant Musashi. *Nature*. 2010; 466:765–768. [PubMed: 20639863]

Figure 1. The stoichiometric challenge of post-transcriptional regulatory factors

Regulation of gene expression at the transcriptional level is limited to the number of encoded target sites and their accessibility in the genome. In contrast, regulation of gene expression at the posttranscriptional level is dependent on the number of possible target sites across the transcriptome and their respective stoichiometric concentrations. Hence posttranscriptional gene regulation is highly dependent on the gene expression program of particular cell types and tissues and regulatory effects variably change with concentration of target isoforms and other RBPs and RNA regulatory factors.

Figure 2. Schematic of current RNA-protein investigation methodologies

(A) and (B) Schematic protocols for the RNA crosslinking followed by mass spectrometry for the genome wide identification of RBPs by Castello et al. (A) and Baltz et al. (B). *In vivo* UV crosslinking captures proteins directly bound to RNA in their native conformation. After UV crosslinking, mRNAs are enriched by polyA purification under denaturing conditions to remove proteins not directly bound to RNA. Washes and nuclease treatment prepares the samples for mass spectrometry. Castello et al. use LC-MS/MS whereas Baltz et al. use quantitative SILAC mass spectrometry. (C) Venn diagrams analyzing the two studies for the overlap of total proteins, canonical RBPs, and novel RBPs. A selection of the novel RBPs validated by secondary biochemical assays are highlighted in the boxes, validated by either Castello et al. (purple) or Baltz et al. (gold). (D) Experimental scheme for ribosome profiling. After cell lysis and DNase digest, RNA digest yields RNA fragments protected by ribosomes. rRNAs are removed by hybridization to rRNA probes, mRNA fragments are size fractionated and prepared for deep-sequencing.

Table 1

List of experimentally validated novel RBPs by the Landthaler (129) or Hentze (128) group, for which neither protein nor domains within the proteins have ever been linked to RNA interaction. Protein family members are defined by Ensembl homology and are color-coded based on whether they were overlapping in both datasets (orange), found in either Castello et al. (purple) or Baltz et al. (gold), or not found at all (black). Parenthesis indicates closely related family members.

Gene	Dataset	Description	Domain	Protein Family	Known pathways
ALDOA	[28]	Aldolase A, fructose-bisphosphate	Glycolytic	(ALDOA, ALDOC), ALDOB	Metabolic enzyme
ALKBH5	[29]	AlkB, alkylation repair homolog 5 (E. coli)	2OG-FeII Oxygenase	ALKBH5	2OG oxygenase, homologs in DNA methylation repair
API5	[29]	Apoptosis inhibitor 5	API5	API5	Anti-apoptotic protein
BTF3	[29]	Basic transcription factor 3	NAC	(BTF3, BTF3L4)	Transcription factor
CSNK1E	[29]	Casein kinase 1, epsilon	Pkinase	(CSNK1A1, CSNK1A1L), (CSNK1G1, CSNK1G2, CSNK1G3), (CSNK1D, CSNK1E), (TTBK1, TTBK2)	Ser/Thr kinase
ECH1	[28]	Enoyl CoA hydratase 1, peroxisomal	ECH	ECH1, ECHDC1, ECHDC2, ECHDC3, ECHS1, AUH, EHHADH, HADH, HADHA	Metabolic enzyme
EDF1	[29]	Endothelial differentiation- related factor 1	MBF1, HTH_3	EDF1	Transcriptional coactivator
ENO1	[28]	Enolase 1, (alpha)	Enolase	(ENO1, ENO2, ENO3), ENO4	Metabolic enzyme
FAM32A	[28]	Family with sequence similarity 32	DUF1754	FAM32A	.
IFIT5	[29]	Interferon-induced protein with tetratricopeptide repeats 5	TPR_12, TPR_11, TPR_1	(IFIT1, IFIT1B, IFIT2, IFIT3, IFIT5)	Anti-viral response, interferon induced
MARK2	[28]	MAP/microtubule affinity- regulating kinase 2	Pkinase, KA1	(MARK1, MARK2, MARK3, MARK4), MELK, NIM1, (NUAK1, NUAK2), (PRKAA1, PRKAA2), (SIK1, SIK2, SIK3), SNRK	Ser/Thr kinase, regulator of cell polarity and stability of microtubule
MYBBP1A	[29]	MYB binding protein (P160) 1a	DNA_pol_phi	MYBBP1A	Transcription factor
PKM2	[28]	Pyruvate kinase, muscle	PK, PK_C	PKM2, PKLR	Metabolic enzyme
PPIA	[28]	Peptidyl/prolyl isomerase A (cyclophilin A)	Pro_isomerase	PPIA, PPIAL4A, PPIAL4B, PPIAL4C, PPIAL4D, PPIAL4G, PPIE, PPIF, RANBP2	Cyclophilin A, peptidyl-prolyl cis-trans isomerase (PPIase) family, protein chaperone
PP1B	[28]	Peptidyl/prolyl isomerase B (cyclophilin B)	Pro_isomerase	PP1B, PP1C, PP1H, PP1L1, PP1L2, PP1L3, PP1L4, CWC27, PPWD1	cyclophilin B, peptidyl-prolyl cis-trans isomerase (PPIase) family, protein chaperone
PRDX1	[29]	Peroxisredoxin 1	AhpC-TSA, 1-cysPrx_C	(PRDX1, PRDX2), PRDX3, PRDX4	Metabolic enzyme, antioxidant enzyme
SHMT2	[28]	Serine hydroxymethyltransferase 2 (mitochondrial)	SHMT	(SHMT1, SHMT2)	Metabolic enzyme

Gene	Dataset	Description	Domain	Protein Family	Known pathways
USP10	[29]	Ubiquitin specific peptidase 10	UCH, PAM2	USP10	De-ubiquitination protease