RNA and proteins: Mutual respect [version 1; referees: 3 approved]

Kathleen B. Hall
Washington University School of Medicine in St. Louis

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs
Please let us know how this document benefits you.

Recommended Citation
https://digitalcommons.wustl.edu/open_access_pubs/5788
RNA and Proteins: Mutual Respect [version 1; referees: 3 approved]

Kathleen B. Hall

Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St Louis, MO, 63110, USA

Abstract

Proteins and RNA are often found in ribonucleoprotein particles (RNPs), where they function in cellular processes to synthesize proteins (the ribosome), chemically modify RNAs (small nucleolar RNPs), splice pre-mRNAs (the spliceosome), and, on a larger scale, sequester RNAs, degrade them, or process them (P bodies, Cajal bodies, and nucleoli). Each RNA–protein interaction is a story in itself, as both molecules can change conformation, compete for binding sites, and regulate cellular functions. Recent studies of Xist long non-coding RNP, the U4/5/6 tri-small nuclear RNP complex, and an activated state of a spliceosome reveal new features of RNA interactions with proteins, and, although their stories are incomplete, they are already fascinating.
**Introduction**

RNA molecules in the cell are rarely naked. Rather, proteins are bound to them in some arrangement consistent with their regulation, protection from nucleases, transport, or formation of ribonucleoprotein particles (RNPs). A 2014 compendium of RNA-binding proteins in humans\(^1\) concluded that 7.5% of 20,500 known protein-coding genes are found in RNPs or bound to mRNAs, where they regulate RNA metabolism. This is likely to be an underestimate, since their structural heterogeneity makes them difficult to identify *de novo*.

The recent discovery of a plethora of non-coding RNAs\(^2\) in cells has invigorated investigation of proteins that bind to RNA. New methods of probing the proteins in a transcriptome have allowed simultaneous identification of a protein and its RNA-binding site. Typically, these are crosslinking-immunoprecipitation (CLIP) experiments\(^3\). Intact cells can be irradiated with ultraviolet (UV) light or treated with formaldehyde to crosslink proteins to RNA, then the complexes are purified from the milieu by immunoprecipitation. To identify proteins bound to mRNAs, cellular UV RNA–protein crosslinking is followed by isolation of all poly(A)-RNA\(^4\). Alternatively, proteins bound to a specific RNA could be recovered by annealing biotin-oligonucleotides complementary to the RNA and selective purification by streptavidin\(^5\). Proteins bound to RNAs could then be identified by mass spectrometry. Several groups applied this method to identify mRNA-binding proteins in human cell lines, mouse embryonic stem cells (ESCs), and *Saccharomyces cerevisiae* yeast cells (reviewed in Gerstberger et al.\(^1\)).

Assuming that there are indeed more than 1,500 RNA-binding proteins in human cells, books will be written about them and their roles in RNA biology. Here, I focus on recent advances that reveal the variety and mystery of RNPs.

**Xist, the RNA that inactivates an X chromosome**

Xist is a long non-coding RNA (lncRNA) that is responsible for transcriptional silencing of one of two X chromosomes in female cells\(^6\). There are approximately 200 Xist molecules bound to a single X chromosome, and each 18 kb of Xist is bound by proteins (Figure 1). Proteins could participate in any aspect of its biology: Xist has to associate with the X chromosome, then spread along it, and finally inhibit RNA polymerase II (Pol II) transcription. After more than twenty years of efforts to identify those proteins, the power of mass spectroscopy has been applied to proteins crosslinked *in cellulo* to Xist.

Two research groups have recently published compendia of Xist-bound proteins. Each group first crosslinked RNA to protein *in cellulo*, selected Xist through oligonucleotide-directed annealing, then used quantitative mass spectrometry to identify bound proteins. An overall comparison of their results shows great similarity but also some curious and intriguing differences. Table 1 and Table 2 list the most abundant proteins recovered from each study.

The groups of Heard and Chang\(^1\) identified 81 proteins *in toto* bound to Xist. Using formaldehyde, they crosslinked proteins to Xist in three different mouse cell types: a male ESC line containing...

---

**Figure 1. Xist wraps around nucleosomes in the X chromosome.** Approximately 200 Xist molecules bind to an X chromosome, spread along it, and inhibit RNA polymerase II from transcribing the DNA. Xist is bound by many proteins at unknown sites and with unknown stoichiometry, which subsequently interact with each other through disordered regions or structured domains. RNA is shown as a yellow/orange strand and protein linkers as blue strands. RRM, RNA recognition motif.

**Table 1. Top 15 Xist-binding proteins from Chl-MS recovery in mouse cells\(^1\).**

<table>
<thead>
<tr>
<th>Crosslinked proteins</th>
<th>In order of abundance</th>
<th>Protein structural motifs</th>
<th>Length*(number of amino acids)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hnRNP M</td>
<td>1</td>
<td>3 RRM</td>
<td>728</td>
</tr>
<tr>
<td>hnRNP U (Saf-A)</td>
<td>2</td>
<td>RGG, KH, acidic region, DNA binding</td>
<td>793</td>
</tr>
<tr>
<td>hnRNP K</td>
<td>3</td>
<td>3 KH, proline-rich</td>
<td>463</td>
</tr>
<tr>
<td>hnRNP A2/B1</td>
<td>4</td>
<td>2 RRM, RGG, glycine-rich</td>
<td>353</td>
</tr>
<tr>
<td>MYEF2</td>
<td>5</td>
<td>2 non-canonical RRM, homology to hnRNP M4</td>
<td>591</td>
</tr>
<tr>
<td>hnRNP A1</td>
<td>6</td>
<td>2 RRM, glycine-rich, RGG</td>
<td>320</td>
</tr>
<tr>
<td>DDX5</td>
<td>7</td>
<td>DEAD box protein</td>
<td></td>
</tr>
<tr>
<td>Spen (SHARP)</td>
<td>8</td>
<td>3 RRM, SPOC</td>
<td>3,640</td>
</tr>
<tr>
<td>RBM XL1</td>
<td>9</td>
<td>RRM</td>
<td></td>
</tr>
<tr>
<td>hnRNP AB</td>
<td>10</td>
<td>2 RRM</td>
<td></td>
</tr>
<tr>
<td>hnRNP D (AUF1)</td>
<td>11</td>
<td>2 RRM</td>
<td>355</td>
</tr>
<tr>
<td>hnRNP L</td>
<td>12</td>
<td>4 RRM, glycine-rich</td>
<td>589</td>
</tr>
<tr>
<td>hnRNP A3</td>
<td>13</td>
<td>2 RRM, glycine-rich</td>
<td>379</td>
</tr>
<tr>
<td>hnRNP C</td>
<td>14</td>
<td>1 RRM, acidic rich</td>
<td>293</td>
</tr>
<tr>
<td>TARDBP (TDP-43)</td>
<td>15</td>
<td>2 RRM, glycine-rich, DNA-binding protein</td>
<td>414</td>
</tr>
</tbody>
</table>

*Many proteins have isoforms with varying lengths; the longest variant in *Homo sapiens* is listed.
an inducible Xist gene, an epiblast stem cell line, and trophoblast stem cells. Each cell type represents one stage of Xist expression. Combining all datasets, three proteins were identified as being most abundant: heterogeneous nuclear RNP (hnRNP) K, hnRNP U, and hnRNP M. In addition, a detailed examination of Xist 5′-0.9 kb sequence revealed several localized proteins. In particular, SPEN (aka SHARP) was found to be necessary for transcriptional silencing.

There is a preponderance of hnRNP proteins. These heterogeneous nuclear ribonucleoproteins are abundant in metazoan cells, where they are mostly found in the nucleus. A recent review of them traced their ancestry, concluding that there are 13 families, each with isoforms or variants. For example, hnRNP A has four homologues in humans (A0, A1, A2, and A3), while hnRNP M has two (MYEF2 and hnRNP M). These proteins typically use RNA recognition motifs (RRMs) to bind RNA, while their other domains engage in protein–protein interactions. Several are involved in pre-mRNA splicing, where they repress splice site selection (hnRNP A) or regulate exon inclusion (hnRNP I). hnRNP I (aka polypyrimidine tract binding protein 1 [PTB1]) also facilitates translation from internal ribosome entry sites (IRES). In Xist, hnRNP functions in Xist are unknown, with the exception of hnRNP U (aka Saf-A), which facilitates Xist localization on chromatin.

In contrast, a group of investigators headed by Guttman took a different approach to finding Xist proteins during transcriptional silencing. After Xist induction in mouse ESCs, cells were UV-crosslinked, Xist RNP was recovered with long antisense oligonucleotides, and Xist proteins were identified by mass spectrometry. Two batches of mouse ESCs were cultured, one in 15N- and one in 14N-media to allow quantification by mass spectrometry (SILAC). Among their ten most abundant proteins, they found SHARP (SPEN) and RMD15, two proteins related in their architecture (they are SPEN family proteins). They also recovered six hnRNP proteins (Table 2). These are exciting findings. In a curious coincidence, SHARP has another life in a nuclear RNP with the steroid receptor RNA activator (SRA)4. SRA is a lncRNA that co-regulates the transcription of nuclear receptors. Bound to SRA, SHARP represses SRA transcription when it recruits histone deacetylase5. Does it carry out a similar task in Xist?

In fact, McHugh et al. found that SHARP was required for the inhibition of Pol II transcription at sites where Xist was bound. The mechanism of inhibition could lie in the recruitment of SMRT and/or HDAC3. HDAC3 is a histone deacetylase that is thought to be responsible for transcriptional repression by changing chromatin structure. Loss of SHARP, LBR, or hnRNP U in knockdown experiments was sufficient to eliminate silencing, but each protein appears to have unique contributions. The role of the other seven proteins was not tested directly, but since each binds directly to Xist, they could have functions in localization, recruitment of other enzymes, stabilization, etc. (for example, binding to Polycimb repressive complex 2 [PRC2]).

The identification of LBR bound to Xist explains localization of the Xist-X chromosome to the nuclear lamina. Transmembrane helices anchor LBR to the lamina, while its tail contacts Xist. Positioning of Xist-X on the lamina changes the structure of the DNA and facilitates protein-mediated spreading of the Xist molecules along the length of the chromatin.

Rather than discovering unknown proteins, these investigations have re-discovered known proteins. They present a new challenge: to understand why they are particularly useful in the Xist context and how their use, and corresponding abundance, is modulated according to developmental stage or cell lineage. The general challenge is not only to understand how proteins use their RNA-binding domains and intervening sequences and disordered tails to control formation of RNP but must also account for their temporal exchange.

**RNA recognition motifs**

A striking feature of proteins bound to Xist is the recurring use of tandem RRM domains. There are certainly advantages to this scheme, since affinity and specificity can be modulated by increasing the number of contacts between RNA and protein. However, neither Xist-binding sites for its associated proteins nor their binding stoichiometry are known. These biochemical characterizations are important to understand how they select their target sites on

---

**Table 2. Mouse embryonic stem cells: top 10 Xist-binding proteins from RAP-MS.**

<table>
<thead>
<tr>
<th>Crosslinked proteins</th>
<th>In order of abundance</th>
<th>Protein structural motifs</th>
<th>Length (number of amino acids)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHARP (SPEN)</td>
<td>1</td>
<td>3 RRM, SPOC</td>
<td>3,640</td>
</tr>
<tr>
<td>RBM15</td>
<td>2</td>
<td>3 RRM, SPOC</td>
<td>969</td>
</tr>
<tr>
<td>MYEF2 (hnRNP M)</td>
<td>3</td>
<td>3 RRM, homology to hnRNP M4</td>
<td>591</td>
</tr>
<tr>
<td>CELF1</td>
<td>4</td>
<td>3 RRM</td>
<td>486</td>
</tr>
<tr>
<td>hnRNP C</td>
<td>5</td>
<td>1 RRM</td>
<td>313</td>
</tr>
<tr>
<td>LBR</td>
<td>6</td>
<td>Chromatin-interaction domain, transmembrane region, lamin-interacting domain</td>
<td>626</td>
</tr>
<tr>
<td>SAF-A (hnRNP U)</td>
<td>7</td>
<td>RGG, SPRY domain, ATPase domain</td>
<td>793</td>
</tr>
<tr>
<td>RALY (hnRNP C)</td>
<td>8</td>
<td>1 RRM</td>
<td>312</td>
</tr>
<tr>
<td>hnRNP M</td>
<td>9</td>
<td>3 RRM</td>
<td>729</td>
</tr>
<tr>
<td>PTBP1 (hnRNP I)</td>
<td>10</td>
<td>4 RRM</td>
<td>555</td>
</tr>
</tbody>
</table>

hnRNP, heterogeneous nuclear ribonucleoprotein particle; RAP-MS, RNA antisense purification-mass spectrometry; RRM, RNA recognition motif; SILAC, stable isotope labeling by amino acids in culture; SPOC, Spen paralog and ortholog C-terminal domain.
the RNA, how they bind to Xist in the milieu of other RNAs in the cell, and how they hang onto the RNA while they also bind to other cellular compartments or recruit other proteins.

RRMs are the most common structural motif used in eukaryotes to bind RNA (Figure 2) and are estimated to be found in 225 human genes. When RRMs are present in multiples, deciphering the contributions of each RRM to the whole can be quite difficult. A recent biophysical study of two tandem RRMs revealed how they partition function.

U2 auxiliary factory (U2AF) is a heterodimer of U2AF65 and U2AF35, which in pre-mRNA splicing aids in the recognition of a 3′ splice site. U2AF65 has two RRMs (RRM1 and RRM2) that bind polypyrimidine tracts, but U2AF35 has a single UHM, a “U2AF homology motif”, that is structurally homologous to an RRM. RRM1 and RRM2 are tethered by a short linker (≈20 amino acids) that allows them to undergo relative motion and orientation. Since they bind to polypyrimidine tracts of variable length and sequence, they must be able to expand or contract to span the site.

The Sattler and Lamb laboratories collaborated on a comprehensive study of the spatiotemporal disposition of U2AF65 RRM1 and RRM2 and their role in RNA binding. von Voithenbergen et al. showed that RRM1 and RRM2 undergo dynamic exchange between a closed or open orientation at equilibrium (Figure 2). In the closed state, RRM1 and RRM2 do not bind RNA, but when the conformation is open, a polypyrimidine tract can bind. If binding is weak (i.e. the polypyrimidine tract is too short or contains multiple purine nucleotides), the exchange between open and closed states is relatively unperturbed. If RNA binding is tight, RRM1 and RRM2 will be trapped in an open state. Thus, the RNA shifts the equilibrium of U2AF RRM1 and RRM2 between open and closed states in an example of conformational selection.

These experiments were conducted using single pair Förster resonance energy transfer (spFRET) that observed single molecules, each containing a donor and acceptor fluorophore. One fluorophore was attached to either RRM, such that the open and closed orientations were distinguished by the FRET efficiency. Combining measurements of fluorophore lifetimes with spFRET facilitated temporal characterization of exchange between open and closed states. In experimental conditions, free RRM1 and RRM2 occupied an open state ~67% of the time. Addition of RNA trapped RRM1 and RRM2 in the open conformation 90% of the time.

U2AF65 and U2AF35 have been the subject of many biochemical and structural investigations, since they are essential proteins for pre-mRNA splicing. In particular, experimental studies of protein–protein interactions between U2AF and other proteins have identified sites where interactions occur. These latest experiments revealed a mechanism of protein–protein interaction involving the UHM of U2AF35 and U2AF65 RRM1 and RRM2. A combination of nuclear magnetic resonance (NMR) structure and dynamics experiments identified the binding site of U2AF35 UHM to be a surface of U2AF65 RRM1. Binding of the UHM to RRM1 shifts the RRM1 and RRM2 conformational equilibrium to the open state, thereby favoring RNA binding. The authors suggest that allosteric drives the RRM1 and RRM2 conformational switch. Allosteric modulation of binding is a powerful mechanism to provide discrimination and affinity, but, by its nature, it is almost impossible to anticipate and cannot be gleaned from static structures.

Figure 2. Regulation of RNA recognition motif (RRM) binding to RNA. A. An RRM has a four-stranded anti-parallel β-sheet, with two α-helices on one side. RNA often sits on the surface of the β-sheet. B. The two RRMs of U2 auxiliary factor (U2AF) exhibit closed/open transitions at equilibrium, but only in the open state can RNA bind. Binding of the U2AF homology motif (UHM) from U2AF35 to RRM1 shifts the equilibrium to favor the open state, which facilitates RNA binding. The C-terminal tail of U2AF65 contacts the U2AF35 UHM. ESE, exonic splicing enhancer.
Many RNA-binding proteins are modular, with an RNA-binding domain, intervening sequences, and disordered tails. Here, U2AF uses two proteins to regulate splicing; other examples include the Sxl-Unr heterodimer that regulates translation via interactions between Sxl RRM and a Unr cold-shock domain, while the SR protein (serine-arginine) SRSF1 is regulated by phosphorylation of its RS tail that blocks intramolecular interaction with its RRMs. Regulation by intermolecular and intramolecular interactions adds another level of complexity to RNA-binding proteins.

The spliceosome and its small nuclear ribonucleoprotein particles

It is estimated that 94% of all human genes contain introns, thereby providing protein isoform diversity. The process of removing introns and joining exons is carried out by the spliceosome, a multi-component and dynamic assembly of RNP. A great challenge in the field of pre-mRNA splicing has been to understand how the spliceosome is physically able to carry out the concerted transesterification reactions of the splicing chemistry to yield mRNAs.

The spliceosome consists of five small nuclear RNPs (snRNPs) that dynamically associate with each other and with pre-mRNA. The major snRNP uses U1, U2, U4, U5, and U6 snRNPs in the process of splicing. Each snRNP contains a single RNA (snRNA) and multiple proteins, but while U1 and U2 snRNPs are independent, U4 and U6 form a di-snRNP. The tri-snRNP is recruited to a bona-fide intron and is then remodeled, losing U4 snRNP and leaving U5 and U6 snRNPs to form the active spliceosome.

The goal of snRNP rearrangement is to allow and facilitate snRNA conformational rearrangements in the spliceosome to produce the active site for catalysis. Rearrangements of pre-mRNA and snRNAs to prepare and position them for catalysis are mainly accomplished by protein helicases. Several recent studies have delved into the details of Brr2 regulation. In a series of papers from the Wahl lab, Brr2 structure and function were addressed by crystallography and biochemistry. The goal of Brr2 in the tri-snRNP is to maintain stability. As biochemistry experiments of Brr2 show, there is a plug domain at the N-terminus of Brr2’s long N-terminal region (NTR). This plug folds back over the entrance of the helicase to block access of the U4/U6 snRNA duplex to the active site of Brr2. This is a unique intramolecular regulatory device, and more experiments are required to understand how it is directed to this position (and how it is displaced).

In the tri-snRNP, Brr2 sits on the Jab1 domain of Prp8, but its orientation and contacts change during activation of the particle. In contrast to the structure of the human tri-snRNP, in a structure of yeast tri-snRNP, a single-stranded region of U4 snRNA occupies the RNA-binding tunnel of Brr2 (PDB ID 3jcr). This state of the tri-snRNP, illustrated in Figure 3, might represent its structure as an autonomous particle before it joins the pre-spliceosome, where U4 and U6 snRNAs are still base-paired to each other. If so, then proteins and RNAs in the tri-snRNP must rearrange to present U4 and/or U6 tails to the helicase active site.

There is another competitive inhibitor of Brr2. Prp8’s Jab1 domain has a C-terminal disordered tail that sneaks into the RNA tunnel of Brr2 to compete with U4 snRNA. The intramolecular plug interaction and Prp8 Jab1 cooperate to inhibit unwinding. Removing the Jab1 tail activates Brr2 helicase activity; Brr2 without its intramolecular plug also has enhanced activity. Do both inhibitors operate in the isolated tri-snRNP?

Brr2 remains in the spliceosome after U4 snRNP has been expelled from the spliceosome. It is seen in a structure of yeast-activated spliceosome, which is defined by the loss of U1 and U4 snRNPs and rearrangements of the remaining snRNAs to interact with each other and with pre-mRNA. A cryo-EM structure of activated yeast spliceosomes (Brr2) shows Brr2 perched on Prp8’s Jab1 domain, with its helicase activity blocked by both inhibitor interactions (PDB ID 5lqw). In an illustration from this structure, U2, U5, and U6 snRNAs are remote from Brr2 (Figure 4). Although not clear from the perspective of Figure 4, Prp8 is entwined with other proteins and the snRNAs in this complex, even as it binds Brr2.
Figure 3. Two tri-small nuclear ribonucleoprotein particle (snRNP) structures trap different states of Brr2. A. Human tri-snRNP cryo-electron microscopy (cryo-EM) at 7 Å resolution shows Brr2 sitting on Prp8 (PDB ID 3jcr). A U4/U6 snRNA duplex is visible. Sm and Lsm rings are pink; other proteins are white. B. In a yeast tri-snRNP complex, (PDB ID 5GAN), U4 snRNA is threaded through Brr2 in the RNA-binding tunnel. These structures might correspond to the tri-snRNP in the nucleus (A) and the tri-snRNP poised for activation by Brr2 as it joins the pre-spliceosome (B). Visualized with visual molecular dynamics (VMD).

Figure 4. Yeast-activated (B<sup>act</sup>)(PDB ID 5LQW; cryo-electron microscopy [cryo-EM] 5.8Å). Brr2 has separated U4 and U6 small nuclear RNAs (snRNAs), and U4 small nuclear ribonucleoprotein particle (snRNP) has been expelled from the spliceosome. Brr2 is bound to the Jab1 domain of Prp8. All 27 proteins are shown in surface representation; most are colored white. Visualized with visual molecular dynamics (VMD).
As the spliceosome progresses through its cycle, there are many short RNA duplexes that need to be unwound. The other seven SF2 RNA helicases are recruited to the spliceosome when they are needed, and then they dissociate. Brm2 remains with the spliceosome until it has completed a splicing cycle, but there are no data suggesting that it is active at any time other than in the conversion from pre-spliceosome to B form. If it is not required for its helicase activity, perhaps its long NTR contributes something to splicing. Brm2 is reported to contribute to catalysis, to stabilize U5 and U6 in the spliceosome, and to assist in the final disruption of the spliceosome and release of ligated exons. If these states of the spliceosome could be trapped for structural studies, Brm2 might be captured in action.

The spliceosome is composed of hundreds of proteins, many of which simply bind RNA, but others actively remodel it. In the past year, spliceosome structures have revealed connections between RNA and proteins that explain previous observations but also raise new questions. This year, structures of the spliceosome C/C* complex show another helicase, prp16, at work on remodelling new questions. This year, spliceosome structures have revealed connections between RNA and the release of ligated exons. Slowly, this RNA enzyme is giving up its secrets.

Conclusions
There is a need to not only understand specific RNPs but also define general rules of engagement, since RNA–protein interactions dominate RNA biology. Indeed, the most mysterious are the membrane-less organelles that contain RNAs and proteins. These conglomerates of RNAs bound by RNA-binding proteins are variously thought to be centers of RNA processing, degradation, transcription, and exchange: P bodies and stress granules in the cytoplasm and nucleoli, Cajal bodies, speckles, and PML bodies in the nucleus. A current model is that disordered domains of the proteins form a fluid matrix that allows a flux of molecules through these liquid droplets. It is a sure bet that these droplets will be objects of intense scrutiny for years to come.

Competing interests
The author declares that she has no competing interests.

Grant information
The author(s) declared that no grants were involved in supporting this work.

Acknowledgements
I thank my reviewers for their thoughtful reading and perceptive comments.


Open Peer Review

Current Referee Status: ✔️ ✔️ ✔️

Editorial Note on the Review Process
F1000 Faculty Reviews are commissioned from members of the prestigious F1000 Faculty and are edited as a service to readers. In order to make these reviews as comprehensive and accessible as possible, the referees provide input before publication and only the final, revised version is published. The referees who approved the final version are listed with their names and affiliations but without their reports on earlier versions (any comments will already have been addressed in the published version).

The referees who approved this article are:

Version 1

1 Shinichi Nakagawa, RNA Biology Laboratory, Faculty of Pharmaceutical Sciences, Hokkaido University, Sapporo, Japan
   *Competing Interests*: No competing interests were disclosed.

1 Markus Wahl, Department of Biology, Chemistry, Pharmacy, Freie Universität Berlin, Berlin, Germany
   *Competing Interests*: No competing interests were disclosed.

1 Xiang-Dong Fu, Department of Cellular & Molecular Medicine, University of California San Diego, San Diego, CA, USA
   *Competing Interests*: No competing interests were disclosed.