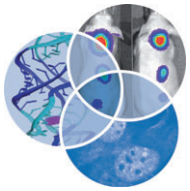
Opinion



Advances and challenges in the detection of transcriptome-wide protein–RNA interactions

Emily C. Wheeler,[1,2,3] Eric L. Van Nostrand[1,2,3] and Gene W. Yeo[1,2,3,4,5]*

RNA binding proteins (RBPs) play key roles in determining cellular behavior by manipulating the processing of target RNAs. Robust methods are required to detect the numerous binding sites of RBPs across the transcriptome. RNA-immunoprecipitation followed by sequencing (RIP-seq) and crosslinking followed by immunoprecipitation and sequencing (CLIP-seq) are state-of-the-art methods used to identify the RNA targets and specific binding sites of RBPs. Historically, CLIP methods have been confounded with challenges such as the requirement for tens of millions of cells per experiment, low RNA yields resulting in libraries that contain a high number of polymerase chain reaction duplicated reads, and technical inconveniences such as radioactive labeling of RNAs. However, recent improvements in the recovery of bound RNAs and the efficiency of converting isolated RNAs into a library for sequencing have enhanced our ability to perform the experiment at scale, from less starting material than has previously been possible, and resulting in high quality datasets for the confident identification of protein binding sites. These, along with additional improvements to protein capture, removal of nonspecific signals, and methods to isolate noncanonical RBP targets have revolutionized the study of RNA processing regulation, and reveal a promising future for mapping the human protein-RNA regulatory network.

How to cite this article:
WIREs RNA 2018, 9:e1436. doi: 10.1002/wrna.1436

INTRODUCTION

Our appreciation of the importance of RNA processing in the maintenance of cellular homeostasis has increased significantly in recent years.[1–3] RNA binding proteins (RBPs) interact with their target RNAs to affect the creation, localization, and function of each RNA molecule in the cell.[4–6] Disruption of these protein–RNA interactions by mutations in RBPs has been implicated in many diseases including neurodegeneration[7] and cancer.[8] Therefore, identifying the RNA targets of specific RBPs is important for deciphering the molecular mechanisms of RBP-mediated diseases.

Here, we discuss current technologies that identify RBP targets and present technical challenges that need to be addressed in the future. In particular, we focus on three major areas of active research in the identification of high confidence, transcriptome-wide RNA binding sites: (1) advantages and disadvantages of UV and other crosslinking methods in RBP:RNA complex capture (Figure 1), (2) recent efforts to optimize the efficiency of converting isolated RNA molecules into cDNA fragments for high-throughput

*Correspondence to: geneyeo@ucsd.edu

[1]Department of Cellular and Molecular Medicine, University of California at San Diego, La Jolla, CA, USA

[2]Stem Cell Program, University of California at San Diego, La Jolla, CA, USA

[3]Institute for Genomic Medicine, University of California at San Diego, La Jolla, CA, USA

[4]Molecular Engineering Laboratory, A*STAR, Singapore

[5]Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

E.V.N. and G.W.Y. are co-founders of Eclipse BioInnovations Inc.

sequencing (Figure 2), and (3) how incorporation of controls and normalization strategies enables proper ranking of binding sites and removal of artifacts (Figure 3). Although significant challenges still prevent the comprehensive identification and quantification of all RNA binding events in a cell, recent technical innovation has greatly enhanced our ability to identify RNA binding targets of hundreds of different RBPs from less than a million cells. With these improvements, RNA researchers are poised to dramatically expand the range of biological questions that can be addressed using transcriptome-wide profiling of protein–RNA interactions.

THE DEVELOPMENT OF TRANSCRIPTOME-WIDE DISCOVERY METHODS FOR PROTEIN–RNA INTERACTIONS

As the need to identify transcriptome-wide protein–RNA interactions continues to grow, it is important for researchers to be aware of both the benefits and shortcomings of methods designed to capture and sequence protein-bound RNAs.

RNA Immunoprecipitation

RNA immunoprecipitation (RIP) followed by quantification on a microarray (RIP-chip) or with high-throughput sequencing (RIP-seq) was the earliest method, and has been used to provide insights into the biological function of RBPs.[9,10] RIP methods involve cell/tissue lysis and immunoprecipitation of native RNA–RBP complexes with an antibody recognizing the RBP of interest. With the use of a total input RNA quantification from the lysate prior to immunoprecipitation, RIP provides a quantifiable binding score for enrichment of specific target RNAs.[9] Since RNA–protein interactions are not stabilized by a covalent crosslink, the stringency of the wash conditions needs to be carefully adjusted to ensure that direct protein–RNA interactions are maintained, while nonspecifically bound RNAs are washed away. Thus, RNAs bound to RBPs with low-affinity may not be recovered. In addition, kinetically unstably bound RBPs may dissociate from their RNA targets and re-associate with other RNAs during harsh lysis conditions such as sonication.[11,12] Milder lysis conditions appear to not lead to post-lysis interactions.[9] Ultimately the extent of post-lysis re-association likely depends on the kinetic properties of the RNA–protein interaction, the lysis and immunoprecipitation conditions,[13] and the relative abundances of the RBP and its RNA targets.[12]

Binding-site-resolution RIP

Historically, RIP experiments could only identify binding events at the whole transcript-level as the lack of an RNase digestion step prohibited the identification of specific binding sites within a transcript. Recently, a digestion-optimized RIP-seq method ('DO-RIP') was developed to investigate protein–RNA interactions with binding-site resolution.[13] Using HuR as a proof of concept, the authors identified digestion conditions that resulted in the isolation of RBP-protected RNA fragments of 20–70 nucleotides, allowing binding-site mapping at high resolution. By using control samples that include total input RNA and a negative control immunoprecipitation (IP) with nonspecific antibodies, DO-RIP quantitatively identifies transcriptome-wide protein binding sites by assigning relative enrichment scores that rank the protein occupancy of identified sites.

Crosslinking Followed by IP

Crosslinking followed by immunoprecipitation (CLIP) was developed to enable more stringent purification of protein–RNA complexes.[14] CLIP reduces the recovery of off-target RNAs indirectly associated with the RBP of interest by permitting more stringent washes of the protein–RNA complex, which disrupts protein–protein interactions. Immunoprecipitated RNAs are then treated with an optimized concentration of RNase to create short RBP-protected RNA fragments 20–70 nucleotides in length. CLIP was the first high-throughput method to identify transcriptome-wide protein binding sites on RNAs and has been used for over a decade to profile many RBPs.[14–16]

STABILIZATION OF PROTEIN–RNA INTERACTIONS BY CROSSLINKING

The implementation of UV-crosslinking was an important breakthrough enabling higher stringency IP and identification of binding sites at high resolution. UV-crosslinking at 254 nm wavelength only creates a covalent bond between an amino acid residue and the RNA base if they are in very close proximity, a constraint that is typically only met with specific, direct interactions.[17] While this selectivity is a key strength of the CLIP approach, even favorable interactions are inefficiently captured. Crosslinking yields generally range from <1% to 5% using standard low-pressure mercury lamps, which typically

deliver irradiances on the order of 1 mW/cm^2.[18] Boosting energy beyond the standard 400 mJ/cm^2 used in CLIP not only gives higher yields, but also increases inter-/intrastrand crosslinks and phosphodiester backbone breaks[19] that may interfere with library generation and the ability to map sequenced fragments to the genome. Pulsed lasers capable of emitting 266 nm radiation at >10^6 W/cm^2 on a nanosecond timescale induce crosslinking via two-photon excitation of nucleobases, a process distinct from the monophotonic mechanism of low-intensity irradiation.[17] This approach has been reported to achieve crosslinking efficiencies of >50% for purified RNA–protein complexes in solution[20] and it remains to be seen whether its RNA products are suitable for high-throughput sequencing.

Inherent Biases of 254 nm UV-crosslinking

Although commonly used due to its simple procedure and applicability to unmodified cells or tissues, standard UV-crosslinking does have known biases that could affect interpretation of downstream results. First, *in vitro* biochemical studies suggest that there are biases in crosslinking efficiency for specific nucleotides and amino acid residues: notably, pyrimidines are more photoactivatable than purines. Similarly, while all amino acids are viable substrates for crosslinking, their reactivity is highly variable (with Cys, Lys, Phe, Trp, and Tyr residues crosslinking with the highest efficiencies and His, Glu, and Asp crosslinking with moderate efficiency).[17] In addition, it is theorized that RBPs that bind double-stranded RNAs crosslink particularly poorly because the deep and narrow groove of A-form helical structures generally preclude access of amino acids to the nucleotide. Thus, depending on the nature of the RBP–RNA interaction, crosslinking will capture some more efficiently than others, while still others are missed entirely. These limitations have led to the development of alternative methods to stabilize protein–RNA interactions, increasing efficiency and widening the scope of protein–RNA interactions amenable to capture (summarized in Figure 1 and detailed below).

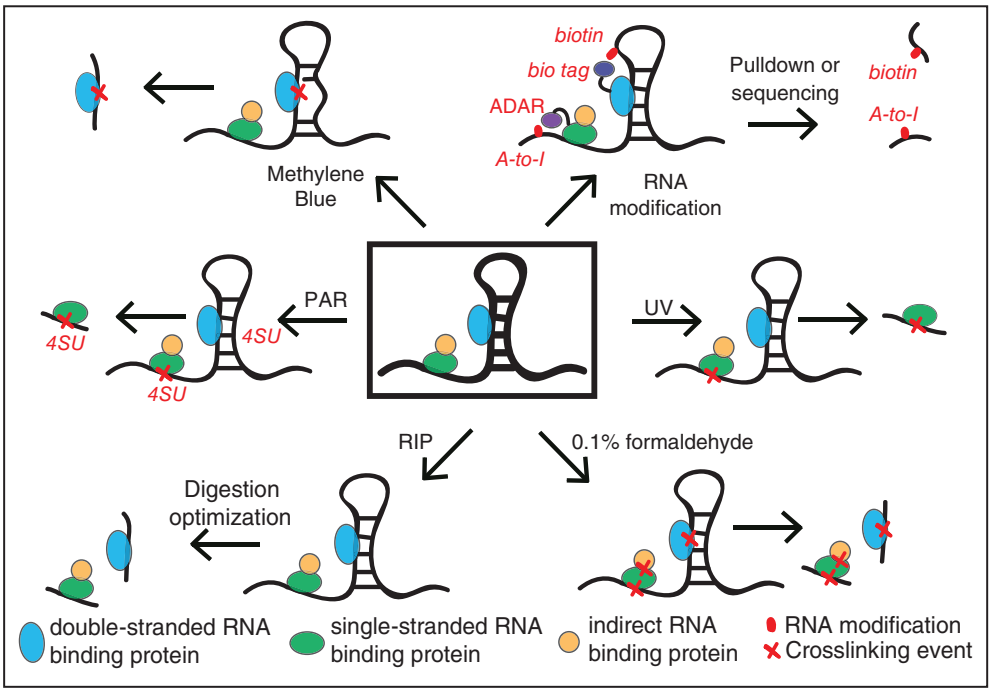


FIGURE 1 | Methods to capture protein–RNA interactions. Different techniques are required to capture single-stranded (green), double-stranded (blue), and indirect (yellow) RNA interactions. Crosses (X) in red mark RNA sites that are crosslinked to the RNA binding protein. Right: UV treatment at 254 nm preferentially captures binding in single-stranded regions. Bottom right: 0.1% formaldehyde treatment captures all protein–protein and protein–RNA interactions. Bottom left: RNA immunoprecipitation (RIP) uses a native pulldown (no crosslinking) to capture binding events with antibody selection. Optimized RNA digestion conditions can reveal specific binding sites with RIP. Left: Photoactivatable ribonucleoside (PAR) analog treatment increases UV crosslinking efficiency at 365 nm. Top left: Methylene blue intercalates between the bases of double-stranded RNA to allow crosslinking in double-stranded regions in the presence of visible light. Top right: Protein–RNA interaction sites are marked by exogenous RNA modifications. This requires creating a fusion protein to modify RNA near binding sites with biotinylation (BioTag-BirA) or A-to-I RNA editing (ADAR).

Crosslinking with Nucleoside Analogs

Photoactivatable ribonucleoside CLIP (PAR-CLIP) uses metabolic RNA labeling in cell culture to incorporate UV-reactive nucleoside analogs, such as 4-thiouridine (4sU) or 6-thioguanosine (6sG), which crosslink at a wavelength of 365 nm[21] (Figure 1). RNA yield after crosslinking is increased by the UV-reactive nucleosides compared to traditional UV-crosslinking,[21] which has driven high adoption of the method.[22–24] PAR-CLIP has the added benefit that as many as 70% of reads contain a T to C mutation after reverse transcription (RT) at the crosslinking site of 4sU,[21] which provides additional confidence in identified binding sites. 4sU does not appear to impair global RNA or protein synthesis in cells at concentrations of up to 100 μM, that is, the concentration used in PAR-CLIP[25] and has been used for decades for site-specific RNA–protein crosslinking studies and, more recently, to measure RNA synthesis and decay rates in vivo.[26] In their initial PAR-CLIP study, Hafner et al.[21] reported that mRNA abundance profiles of HEK293 cells treated with up to 1 mM 4sU or 6sG for 12 h were not grossly perturbed, but Burger et al.[27] subsequently showed in several cell lines that 50 μM 4sU for 6 h inhibits ribosome biogenesis, triggers a nucleolar stress response and causes proliferation defects. As another potential limitation of the method, only the supplied nucleotide will be available for crosslinking, so crosslinking sites are limited to sequence regions containing the supplied nucleotide. Lastly, RNAs with high turnover rates may be overrepresented due to greater incorporation of the nucleoside analog on a short time scale.

Non-UV Techniques to Crosslink Protein-bound RNA

Given that UV-crosslinking requires close association between an amino acid and a nucleotide residue, it is likely inefficient for dsRNA binders that interact with structural features of the backbone and have low accessibility of the nucleotide base.[28] hiCLIP[29] and CLASH[30] are two methods that combine UV-crosslinking with an RNA–RNA ligation step to specifically identify double stranded RNA (dsRNA) regions that interact with an RBP of interest. However, less than 2% of sequenced reads using these methods contain an RNA–RNA crosslinking event between the two interacting strands of RNA, highlighting the inefficiency of capturing dsRNA interactions with UV-crosslinking. Thus, there remains a need for the development of improved capture methods to profile RBPs that do not interact directly with a nucleobase.

Intercalating Agents

One potential approach to specifically capture dsRNA–RBP interactions is with treatment of methylene blue. Methylene blue intercalates between the bases of dsRNA to open up the RNA structure and allow crosslinking of dsRBPs in the presence of visible light[31] (Figure 1). This method is highly specific for RBP-dsRNA interactions and is estimated to have 10–15% efficiency *in vitro*.[31] Methylene blue treatment could be used in combination with 254 nm UV crosslinking *in vivo* to capture both single and double-stranded RNA interactions, or performed in independent experiments to distinguish single and dsRNA interactions of the same protein.

Chemical Crosslinking

Another alternative to UV is to perform crosslinking with low concentrations (0.1%) of formaldehyde, a reversible crosslinking agent (Figure 1).[32] Formaldehyde has long been used to map protein–DNA interactions, and was recently used to profile both RNA binding of proteins that have classically been characterized as chromatin modifiers[32] as well as direct RNA-chromatin interactions.[33] However, as formaldehyde also crosslinks protein–protein interactions, the concern of indirect crosslinking to large ribonucleoprotein complexes (e.g., ribosomes, RNA polymerase II, P-bodies, or other RNP granules) has generally limited its use for studying RBPs. Additionally, at low formaldehyde concentration (10-fold lower than field standard for ChIP), the question of overall crosslinking efficiency has yet to be addressed.

Marking Binding Sites through Direct RNA Modification

The limitations of the above crosslinking methods have led researchers to explore noncrosslinking based approaches to profile RBP–RNA interactions by direct modification of target RNAs. One such method, TRIBE, fuses the deaminase domain of ADAR to the RBP of interest to cause ectopic RNA editing near the sites of RBP binding[34] (Figure 1). These editing sites can be detected by direct sequencing of RNAs compared to an endogenous, untagged control to identify ADAR-RBP fusion-dependent RNA editing events. A major advantage of this technique is that the editing readout is sequencing based (as A-to-I changes modify the cDNA sequence), rather than IP based, eliminating the requirement of crosslinking and IP.[34] However, substrate specificity inherent to the ADAR deaminase domain could impose a bias in the selection and efficiency of

deamination at specific binding sites. Additionally, this method requires expression of the RBP-ADAR fusion protein, very deep sequencing, and a robust computational method for identifying editing sites.

Another approach utilizes *in vivo* RNA modification by fusing the RBP of interest to a 'bio' tag that recruits BirA ligase to biotinylate nearby RNAs[35] (Figure 1). Biotinylated RNAs can then be purified in stringent conditions with a streptavidin pull-down to isolate and sequence RNA sequences located near protein binding sites. Further work will be necessary to explore RBPs that would be suitable for this approach, and whether there are biases in biotinylation frequency.

METHODS TO SELECT FOR A PROTEIN OF INTEREST

Antibody-based Immunoprecipitation

The RBP of interest is enriched in a CLIP experiment by immunoprecipitation with a monoclonal or polyclonal antibody. While not all RBPs have an IP-grade antibody, many groups are working to generate databases of validated antibodies to expand the pool of RBPs that can be immunoprecipitated.[36,37] Proteins for which there is no validated IP-grade antibody requires fusion to a peptide tag for IP at either the N or C terminus. This approach has been widely used in cell culture and in model organisms by both overexpression of a transgene or modification of the endogenous genomic locus. Regardless of the antibody used, a control IP with an IgG isotype-only control should be performed in parallel to provide a specificity metric for the IP. Additionally, there is variation in the specificity of antibodies raised against different epitope tags, and therefore it is important when using tags to perform a negative control IP on a sample that does not contain the tagged protein.[38]

Increasing the Purity of Protein–RNA Complexes

CLIP protocols include a protein gel purification step in part because washing of the immunoprecipitated protein–RNA complex does not remove all indirect protein interactions; the strength of the protein-antibody interaction limits the stringency of washing. Other strategies have been developed to increase the specificity of the IP and remove the need for additional gel purification of the complexes. Singh et al. used a double IP strategy with two different antibodies in succession to increase the specificity and purity of the final sample.[28] The double antibody selection without gel purification produced a library with the same purity as a single IP followed by gel purification. This method has the added benefit that it can either be applied to a single protein with two antibodies, or a protein complex in which two different components of the complex are targeted for IP.

More recent iterations of tagging strategies are designed to increase the strength of protein capture to that of a covalent interaction, to allow for extremely harsh washing and complete removal of noncovalently linked protein and RNA species. Protein purification tags (such as the HIS, Bio/BirA,[39] or TAP-TAG)[40] can be inserted next to a protein at the genomic locus with CRISPR and allow for stringent protein purification conditions. The recently developed Halo Tag has extremely strong and highly specific interactions to the HaloLink resin for protein capture, although its large size (297 amino acids), can potentially disrupt nearby protein–protein interactions in an endogenous setting.[41] These and other tagging strategies hold promise for both increasing the signal to noise ratios in CLIP-seq data and eliminating the need for gel purification of the protein–RNA complexes.

ELIMINATING RADIOACTIVITY FROM CLIP METHODS

Replacing Radiolabeling with Fluorescence

Preparation of a CLIP library requires the optimization of RNA fragmentation after IP to ensure that the resulting RNA fragments are long enough to be uniquely mappable to the reference genome, but short enough to identify binding sites at high resolution. Historically, this has been performed by titrating RNase, resolving the RBP-RNA complexes by SDS PAGE gels, and autoradiographic visualization of radiolabeled RNA. RNase treatment results in a characteristic smear above the size of the protein of interest. As an additional control for IP specificity, at the highest RNase conditions all digested fragments would resolve to a single band at the size of the protein of interest. However, radiolabeling posed an inconvenience for widespread adoption of CLIP, and recent advances have eliminated it. For example, irCLIP uses an infrared-dye-conjugated and biotinylated RNA adapter that can be imaged with a digital fluorescence imager.[42] This adapter enables visualization of RNAs under different digestion conditions with a digital scanning readout that does not require radioactive materials.

Standardized RNA Fragmentation

RBPs bind RNAs of varying size, and different cell types contain varying levels of endogenous RNases.

As a result, RNA trimming should be optimized for each experiment. For large-scale enhanced CLIP (eCLIP) experiments performed as part of the ENCODE (https://www.encodeproject.org) efforts, we explored the requirement of customized fragmentation conditions for every CLIP experiment. In a single cell type, we tested a wide range of RNase concentrations for two RBPs representative of the extremes of RNA target lengths: RBFOX2, which binds intronic regions within pre-mRNAs that can be hundreds of kilobases in length, and SLBP, which exclusively binds the 3′ untranslated regions (UTR) of the ~150 nt intronless histone mRNAs. The total number of binding clusters and their distribution across and within genic regions (i.e., intronic regions, coding sequence, 5′ and 3′ UTRs) were surprisingly robust to the extent of RNA digestion.[43] Therefore, once endogenous levels of RNase have been accounted for in the cell/tissue type of interest, a single, optimized concentration of RNase is often appropriate to yield informative binding profiles for most RBPs.

IMPROVING THE RECOVERY RATE OF RNAs PREPARED FOR SEQUENCING

Measuring PCR Duplication

In early CLIP protocols, the low amount of RNA recovered after crosslinking and immunoprecipitation, coupled with inefficiencies in enzymatic reactions during library preparation, led to the need for many cycles of PCR amplification to generate sufficient material for sequencing. This often resulted in libraries of low sequence complexity, that is, containing a large fraction of duplicated reads. We estimated the duplication rate from publicly available CLIP datasets and found that on average, a staggering 83.8% of sequenced reads were flagged as PCR duplicates.[43] To illustrate, a standard sequencing library requires ~100 fmoles (~6×10^{10}) DNA molecules. If 25 PCR cycles are required to produce this amount, and one assumes 80% PCR efficiency, then the initial unamplified library only contained ~25,000 molecules $(6 \times 10^{10})/(1.8^{25})$. Therefore, only up to 25,000 reads in the final sequenced library will have originated from unique molecules and all additional reads are attributable to PCR duplicated sequences. To precisely quantify and remove PCR duplicated reads, adaptors containing short random sequences are ligated to the fragments during library preparation to uniquely tag each RNA fragment prior to PCR amplification. These unique molecular identifiers (UMIs) enable accurate classification of unique and duplicated reads by comparing the mapped genomic coordinates of reads that contain the same UMI.[16,44,45]

Addressing RT Termination at Crosslink Sites

Increasing the total yield of recovered cDNA has led to some of the greatest improvements in library complexity. In the first generation CLIP protocols, such as HITS-CLIP,[15] both 5′ and 3′ RNA adapters are ligated prior to RT (Figure 2). After first-strand cDNA synthesis, the sample is PCR amplified with primers complementary to a portion of the adapters ligated at the ends of the RNA. Since UV-crosslinking chemically modifies the nucleotide bridging the RBP-RNA crosslink, RT enzymes are prone to termination at the crosslink site, and as many as 80% of the resulting cDNA products lacked the 5′ adapter and therefore lacked the 5′ primer binding site.[46] These sequences fail to be PCR amplified and are thus lost from the sequencing library. Individual nucleotide resolution CLIP (iCLIP) and subsequent methods (eCLIP, irCLIP) addressed this issue by performing the second adapter ligation after RT, such that the second adapter is ligated to all cDNA fragments regardless of the RT termination site (Figure 2). Because a subset of reads terminate at the RT stops, this approach has the added advantage that the end of a portion of reads, after genome mapping, mark crosslink sites and therefore enable binding sites to be identified at single nucleotide resolution.[16]

Increasing Enzymatic Reaction Efficiency

In addition to recovering more RNAs by adding the second adapter after the RT step, improvements in enzymatic reaction efficiencies have greatly increased the yield of unique fragments prior to PCR amplification. Zarnegar et al. optimized the reaction condition for each step of library preparation by using an infrared-dye-conjugated adapter to quantify the amount of RNA before and after each reaction with a dot blot.[42] The resulting irCLIP libraries have a much higher yield of unique library fragments and therefore require a small number of PCR cycles to maintain a high complexity library. Similarly, the eCLIP method employed highly optimized reaction conditions to improve the overall yield of sequencing libraries by ~1000-fold over libraries generated with the iCLIP method.[43] These improvements have now made it possible to reliably generate informative CLIP libraries from limited amounts of starting material (less

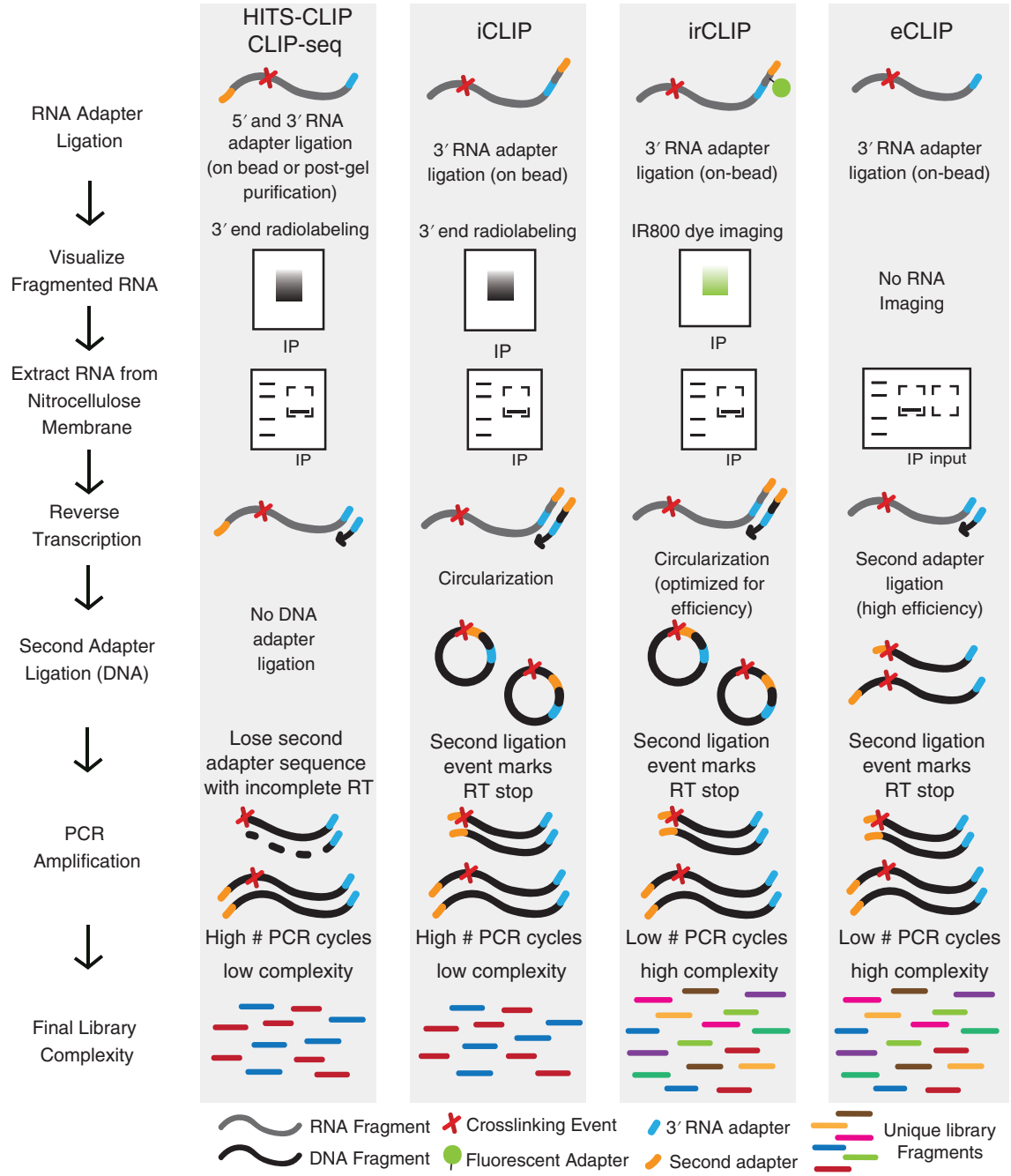


FIGURE 2 | Comparison of CLIP-seq library preparation protocols. HITS-CLIP/CLIP-seq perform adapter ligation on both ends of the RNA while other protocols only ligate an RNA adapter on the 3′ end. RNA digested fragments are visualized with radiolabeling (HITS-CLIP/CLIP-seq and iCLIP), infrared dye imaging (irCLIP), or not at all (eCLIP). RNA is then transferred and isolated from a nitrocellulose membrane. In the case of eCLIP, a size-matched input sample is excised to control for background. For all, reverse transcription (RT) generates cDNA using the 3′RNA adapter as the priming site. As the RT enzyme commonly terminates at crosslinking sites, there will be a mixture of full length and truncated fragments. iCLIP and irCLIP have a circularization step to put the second adapter on the DNA fragment. Ligation in irCLIP has been highly optimized to improve efficiency. eCLIP uses a second ligation step that has also been optimized for efficiency. HITS-CLIP/CLIP-seq fragments that have incomplete RT cannot PCR amplify due to the loss of the second adapter sequence. For all, cDNA fragments are then PCR amplified to generate enough material for sequencing. Recent methods (irCLIP, eCLIP) have routinely high complexity libraries generated from a low number of PCR cycles.

than one million cells),[42] and at scale (hundreds of RBPs).[43]

CONTROLLING FOR BACKGROUND RNA SIGNAL

As RNAs are dynamic in their expression, subcellular localization, and structural features, accurate quantification of protein bound RNAs has been quite challenging in CLIP-seq experiments. In contrast, the fixed number of copies of DNA elements in the nucleus allows for a much simpler background for robust quantifications of DNA binding from ChIP-seq experiments. Therefore, it is crucial to perform quality checks on the identified binding sites. Control experiments such as a nonspecific antibody selection, or an input sample to normalize background abundance can be used to increase the confidence of identified binding sites.

Ranking Significance of Identified Binding Sites

Binding sites are identified computationally as genomic regions that contain a pileup of reads mapping to that specific site (ie. peaks). After binding sites have been called, they can be quality checked by a variety of metrics. If a nonspecific antibody selection resulted in enough RNA to sequence, sites that were called as peaks in that sample can be flagged as background sites. Some CLIP protocols, such as CLIP-seq/HITS-CLIP, preserve crosslink-induced mutations (CIMS), which can be used to identify the specific nucleotide residue that was crosslinked.[47] Other methods, such as iCLIP, irCLIP, and eCLIP capture the crosslinking site at RT termination as marked by the end of the sequencing read. If a binding motif is known or validated for the protein of interest, peaks can be filtered to those that contain the motif.[15] And perhaps most importantly, if multiple biological replicates or antibodies targeting different epitopes for the same RBP are used, peaks that are reproducible across replicates make up the list of most confident binding sites.

Input Control for Background

Washing and gel purification of the protein–RNA complex removes the majority of proteins and RNAs bound nonspecifically. However, washing still does not result in a perfectly pure sample and it is possible that some peaks identified are due to a contaminating RBP, or sticky RNA that is contaminating the IP. To address this, the eCLIP protocol has incorporated a size-matched input to capture nonspecific, background RNAs that are sequenced in a CLIP experiment (Figure 3). The size-matched input contains 2% of the cell lysate that has been crosslinked, run on a gel, transferred to a membrane, and cut at the same size range as the IP (up to 75 kDa above the protein of interest). This input fraction has the same crosslinking, fragmentation, ligation, and amplification biases as the IP sample. Therefore it is an important control for highly abundant, sticky RNAs that are bound by many proteins and get called as binding sites in many CLIP experiments.[45] An enrichment score of reads in the IP relative to the size-matched input for a given peak location provides a metric for the specificity of a binding event relative to background (Figure 3). This metric enabled quantification of enriched binding even at abundant transcripts commonly considered as artifacts across CLIP experiments, such as MALAT1.[43] It is important to note that the enrichment score calculated here is not the same as what has been historically used for RIP experiments. The main difference is that the input in a RIP sample is total RNA, whereas the input used for eCLIP contains only the pool of crosslinked

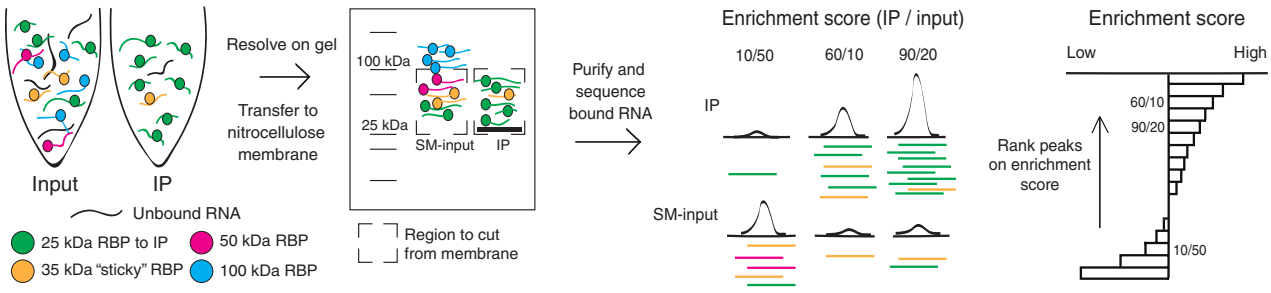


FIGURE 3 | Quantification of background signal with size-matched input (SM-Input). The 2% of lysate is taken prior to IP as the input sample. 'Sticky' RBPs (yellow) are not completely purified away and contaminate the IP sample. The input and IP are run in parallel on the protein gel and extracted from the nitrocellulose membrane at the same size range. Called peaks are then normalized by dividing the number of reads in the IP by the number of reads in the input to remove signal coming from background RNAs (yellow). The enrichment score is a rank-based metric for specificity of binding.

RNAs that are bound by proteins that run within the same size range on the gel. Therefore the eCLIP input captures the experimental background in the individual CLIP experiment, rather than providing an enrichment score relative to total RNA.

A technical limitation that remains in CLIP-seq experiments is the ability to address *in vivo* protein binding occupancy on transcripts. Quantification of binding occupancy with CLIP methods is confounded by IP efficiency, crosslinking efficiency, and the inability to simultaneously measure RBP binding and RNA abundance in a sample. Therefore, if two peaks are identified with the same enrichment score on two independent transcripts, it is not yet possible to determine which one correlates to a higher stoichiometric binding occupancy at its respective site. However, this question can be critical to addressing the functional relevance of binding sites. For example, if one transcript has 20 copies in a cell and all 20 are bound by a protein, that is likely a different mode of regulation than a binding site on a transcript that has 200 copies in a cell, 20 of which are bound by protein. It remains an open challenge to identify modifications to the CLIP procedure that would enable such quantitative estimates of binding occupancy.

THE NEXT STEP: IDENTIFICATION OF FUNCTIONAL BINDING SITES

With the plethora of binding sites identified by RIP and CLIP-related technologies, there is an urgent need to evaluate and prioritize the function of these binding sites. Key questions that need to be addressed include: (1) Are all binding sites functional? (2) What are the properties of functional binding sites? (3) How do we define, measure, or experimentally validate functional binding sites? Orthogonal assays to measure RBP function can address these questions.[48] For example, a splicing map can be generated from an analysis of alternative splicing to compare binding location relative to exons that are alternatively spliced upon knockdown of a protein of interest.[49] Emerging high-throughput screening techniques using CRISPR genome editing, tethering,[50] and other methods to assay RNA processing allows for the manipulation of specific binding sites to evaluate function. To test the function of a given binding site, antisense oligonucleotides (ASOs) can be introduced to inhibit protein binding,[51] or CRISPR genome-editing can be used to modify the binding site and test the resulting behavior of the endogenous RNA target. Tethering strategies using MS2 hairpins or RNA-guided Cas9[52] can be used to probe RBP functions, such as splicing or mRNA degradation, by tethering the protein to a previously unregulated site. Comprehensive RBP-RNA interaction maps annotated with regions where binding is associated with a validated cellular function should serve as a valuable tool for engineering synthetic RNAs to direct proper packaging, expression, and behavior of particular RNAs for therapeutic use.

CONCLUSION

The ability to identify transcriptome-wide RBP binding sites with CLIP and RIP technologies has played a critical role in our ability to understand molecular mechanism of RBP function. In the past few years, dramatic improvements in library preparation efficiencies and removal of radioactivity from CLIP have led to more widespread adoption of CLIP. The incorporation of an input sample has greatly increased the ability to distinguish true binding sites from background signal in an experiment. Future improvements are needed to: (1) increase the efficiency of protein–RNA capture with methods other than UV-crosslinking, (2) increase the strength of protein capture to allow for higher stringency washes without the need for gel purification, and (3) develop orthogonal assays to determine functional binding sites. These improvements will enhance our ability to identify and interpret functionally relevant RNA elements across the transcriptome.

ACKNOWLEDGMENTS

We would like to thank all members of the Yeo Lab for review and critique of this manuscript with particular acknowledgment to Julia Nussbacher, Stefan Aigner, Mark Perelis, and Ryan Marina. This work was supported by grants from the NIH (HG004659 and NS075449) to GWY. E.C.W. is supported by grants from the University of California, San Diego, Genetics Training Program (T23, GM008666) and the NSF Graduate Research Fellowship Program. E.L.V.N. is a Merck Fellow of the Damon Runyon Cancer Research Foundation (DRG-2172-13) and is supported by a K99 grant from NIH (HG009530).

REFERENCES

1. Tartaglia GG. The grand challenge of characterizing ribonucleoprotein networks. *Front Mol Biosci* 2016, 3:24.
2. Kudinov AE, Karanicolas J, Golemis EA, Boumber Y. Musashi RNA-binding proteins as cancer drivers and novel therapeutic targets. *Clin Cancer Res* 2017, 23:2143–2153.
3. Castello A, Fischer B, Hentze MW, Preiss T. RNA-binding proteins in Mendelian disease. *Trends Genet* 2013, 29:318–327.
4. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet* 2014, 15:829–845.
5. Muller-McNicoll M, Neugebauer KM. How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nat Rev Genet* 2013, 14:275–287.
6. Hurt JA, Robertson AD, Burge CB. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res* 2013, 23:1636–1650.
7. Nussbacher JK, Batra R, Lagier-Tourenne C, Yeo GW. RNA-binding proteins in neurodegeneration: Seq and you shall receive. *Trends Neurosci* 2015, 38:226–236.
8. Saez B, Walter MJ, Graubert TA. Splicing factor gene mutations in hematologic malignancies. *Blood* 2017, 129:1260–1269.
9. Tenenbaum SA, Carson CC, Lager PJ, Keene JD. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc Natl Acad Sci USA* 2000, 97:14085–14090.
10. Keene JD, Komisarow JM, Friedersdorf MB. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nat Protoc* 2006, 1:302–307.
11. Mili S, Steitz JA. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA* 2004, 10:1692–1694.
12. Riley KJ, Yario TA, Steitz JA. Association of argonaute proteins and microRNAs can occur after cell lysis. *RNA* 2012, 18:1581–1585.
13. Nicholson CO, Friedersdorf MB, Keene JD. Quantifying RNA binding sites transcriptome-wide using DO-RIP-seq. *RNA* 2016, 23:32–46.
14. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP identifies Nova-regulated RNA networks in the brain. *Science* 2003, 302:1212–1215.
15. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 2008, 456:464–469.
16. Konig 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.
17. Meisenheimer KM, Koch TH. Photocross-linking of nucleic acids to associated proteins. *Crit Rev Biochem Mol Biol* 1997, 32:101–140.
18. Darnell RB. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *WIREs RNA* 2010, 1:266–286.
19. Pashev IG, Dimitrov SI, Angelov D. Crosslinking proteins to nucleic acids by ultraviolet laser irradiation. *Trends Biochem Sci* 1991, 16:323–326.
20. Budowsky EI, Axentyeva MS, Abdurashidova GG, Simukova NA, Rubin LB. Induction of polynucleotide-protein cross-linkages by ultraviolet irradiation. Peculiarities of the high-intensity laser pulse irradiation. *Eur J Biochem* 1986, 159:95–101.
21. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jr., Jungkamp AC, Munschauer M, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 2010, 141:129–141.
22. Kishore S, Gruber AR, Jedlinski DJ, Syed AP, Jorjani H, Zavolan M. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome Biol* 2013, 14:R45.
23. Maatz H, Kolinski M, Hubner N, Landthaler M. Transcriptome-wide Identification of RNA-binding protein binding sites using photoactivatable-ribonucleoside-enhanced crosslinking immunoprecipitation (PAR-CLIP). *Curr Protoc Mol Biol* 2017, 118:27.6.1–27.6.19.
24. Hamilton MP, Rajapakshe KI, Bader DA, Cerne JZ, Smith EA, Coarfa C, Hartig SM, McGuire SE. The landscape of microRNA targeting in prostate cancer defined by AGO-PAR-CLIP. *Neoplasia* 2016, 18:356–370.
25. Melvin WT, Milne HB, Slater AA, Allen HJ, Keir HM. Incorporation of 6-thioguanosine and 4-thiouridine into RNA. Application to isolation of newly synthesised RNA by affinity chromatography. *Eur J Biochem* 1978, 92:373–379.
26. Miller MR, Robinson KJ, Cleary MD, Doe CQ. TU-tagging: cell type-specific RNA isolation from intact complex tissues. *Nat Methods* 2009, 6:439–441.
27. Burger K, Muhl B, Kellner M, Rohrmoser M, Gruber-Eber A, Windhager L, Friedel CC, Dolken L, Eick D. 4-Thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. *RNA Biol* 2013, 10:1623–1630.

28. Singh G, Ricci EP, Moore MJ. RIPiT-Seq: a high-throughput approach for footprinting RNA:protein complexes. *Methods* 2014, 65:320–332.
29. Sugimoto Y, Vigilante A, Darbo E, Zirra A, Militti C, D'Ambrogio A, Luscombe NM, Ule J. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. *Nature* 2015, 519:491–494.
30. Kudla G, Granneman S, Hahn D, Beggs JD, Tollervey D. Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc Natl Acad Sci U S A* 2011, 108:10010–10015.
31. Liu ZR, Wilkie AM, Clemens MJ, Smith CW. Detection of double-stranded RNA-protein interactions by methylene blue-mediated photo-crosslinking. *RNA* 1996, 2:611–621.
32. G Hendrickson D, Kelley DR, Tenen D, Bernstein B, Rinn JL. Widespread RNA binding by chromatin-associated proteins. *Genome Biol* 2016, 17:28.
33. Engreitz JM, Sirokman K, McDonel P, Shishkin AA, Surka C, Russell P, Grossman SR, Chow AY, Guttman M, Lander ES. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell* 2014, 159:188–199.
34. McMahon AC, Rahman R, Jin H, Shen JL, Fieldsend A, Luo W, Rosbash M. TRIBE: hijacking an RNA-editing enzyme to identify cell-specific targets of RNA-binding proteins. *Cell* 2016, 165:742–753.
35. He A, Pu WT. Genome-wide location analysis by pull down of in vivo biotinylated transcription factors. *Curr Protoc Mol Biol* 2010, Chapter 21:Unit 21.0.
36. Bjorling E, Uhlen M. Antibodypedia, a portal for sharing antibody and antigen validation data. *Mol Cell Proteomics* 2008, 7:2028–2037.
37. Sundararaman B, Zhan L, Blue SM, Stanton R, Elkins K, Olson S, Wei X, Van Nostrand EL, Pratt GA, Huelga SC, et al. Resources for the comprehensive discovery of functional RNA elements. *Mol Cell* 2016, 61:903–913.
38. Van Nostrand EL, Gelboin-Burkhart C, Wang R, Pratt GA, Blue SM, Yeo GW. CRISPR/Cas9-mediated integration enables TAG-eCLIP of endogenously tagged RNA binding proteins. *Methods* 2016, 118-119:50–59.
39. de Boer E, Rodriguez P, Bonte E, Krijgsveld J, Katsantoni E, Heck A, Grosveld F, Strouboulis J. Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. *Proc Natl Acad Sci USA* 2003, 100:7480–7485.
40. Gerace E, Moazed D. Affinity purification of protein complexes using TAP tags. *Methods Enzymol* 2015, 559:37–52.
41. Daniels DL, Mendez J, Mosley AL, Ramisetty SR, Murphy N, Benink H, Wood KV, Urh M, Washburn MP. Examining the complexity of human RNA polymerase complexes using HaloTag technology coupled to label free quantitative proteomics. *J Proteome Res* 2012, 11:564–575.
42. Zarnegar BJ, Flynn RA, Shen Y, Do BT, Chang HY, Khavari PA. irCLIP platform for efficient characterization of protein-RNA interactions. *Nat Methods* 2016, 13:489–492.
43. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* 2016, 13:508–514.
44. Darnell JC, Van Driesche SJ, Zhang C, Hung KY, Mele A, Fraser CE, Stone EF, Chen C, Fak JJ, Chi SW, et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* 2011, 146:247–261.
45. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res* 2017, 27:491–499.
46. Sugimoto Y, Konig J, Hussain S, Zupan B, Curk T, Frye M, Ule J. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol* 2012, 13:R67.
47. Moore MJ, Zhang C, Gantman EC, Mele A, Darnell JC, Darnell RB. Mapping argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat Protoc* 2014, 9:263–293.
48. Van Nostrand EL, Huelga SC, Yeo GW. Experimental and computational considerations in the study of RNA-binding protein-RNA interactions. *Adv Exp Med Biol* 2016, 907:1–28.
49. Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, Gaasterland T, Blencowe BJ, Darnell RB. An RNA map predicting Nova-dependent splicing regulation. *Nature* 2006, 444:580–586.
50. Bos TJ, Nussbacher JK, Aigner S, Yeo GW. Tethered function assays as tools to elucidate the molecular roles of RNA-binding proteins. *Adv Exp Med Biol* 2016, 907:61–88.
51. Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, Liang TY, Stark TJ, Gehman LT, Hoon S, et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* 2013, 20:1434–1442.
52. Nelles DA, Fang MY, O'Connell MR, Xu JL, Markmiller SJ, Doudna JA, Yeo GW. Programmable RNA tracking in live cells with CRISPR/Cas9. *Cell* 2016, 165:488–496.

