

CLIP and complementary methods

Markus Hafner¹, Maria Katsantoni², Tino Koster³, James Marks¹, Joyita Mukherjee^{4,5}, Dorothee Staiger³, Jernej Ule^{4,5,6*}, Mihaela Zavolan²

Affiliations:

¹RNA Molecular Biology Group, National Institute of Arthritis and Musculoskeletal and Skin Diseases, Bethesda, MD 20892

²Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, Basel, CH 4056

³RNA Biology and Molecular Physiology, Faculty of Biology, Bielefeld University, Bielefeld, Germany

⁴The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK

⁵Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, Queen Square, London, WC1N 3BG, UK

⁶Department of Molecular Biology and Nanobiotechnology, National Institute of Chemistry, Hajdrihova 19, SI-1001 Ljubljana, Slovenia

Author Contributions:

Introduction (D.S.,J.U.), Experimentation (M.H.,J.Ma.,J.Mu.,J.U.), Results (M.Z.,M.K.,J.U.), Applications (M.H.,J.Ma.,D.S.,T.K.,J.Mu.,J.U.), Reproducibility and data deposition (M.Z.,M.K.,J.U.), Limitations and optimizations (J.U.,J.Mu.), Outlook (J.U.,D.S.). Oversight of Primer (J.U.)

Index:

Abstract

[H1] Introduction

[H1] Experimentation

[H2] Protein-centric methods

[H3] Original CLIP, and its adaptation to high-throughput sequencing

[H3] Individual-nucleotide resolution CLIP, infrared CLIP and enhanced CLIP

[H3] Photoactivatable ribonucleoside-enhanced CLIP
[\[H3\] Mapping of RNA secondary structures interacting with RBPs using CLIP](#)

[H3] Targets of RNA-binding proteins identified by editing

[H3] Proximity-CLIP

[H2] RNA-centric methods

[H3] RNA affinity proteome capture

[H3] RNA-directed proximity-based proteome labelling

[H1] Results

[H2] CLIP analysis workflow

[H3] Extrinsic and intrinsic background in CLIP experiments

[H3] Peak identification

[H3] Characterizing RBP binding motifs

[H3] Regulatory grammar

[H2] Assessing the specificity of CLIP

[H2] Assessing the sensitivity of CLIP

[H1] Applications

[\[H2\] Cell culture models](#)

[\[H2\] Model organisms](#)

[\[H2\] Plants](#)

[H2] Development and disease

[\[H2\] Protein- and RNA-centric approaches yield complementary insights](#)

[H1] Reproducibility and data deposition

[H2] Reproducibility and comparative value of CLIP data

[H2] Data resources

[H1] Limitations and optimizations

[H2] RBP-specific challenges in CLIP data analysis

[H2] Challenges of RNA-centric methods

[H3] RNA affinity capture methods

[H3] Proximity-based methods

[H2] Challenges and opportunities in identifying the RNA binding sites

[H1] Outlook

Acknowledgements

Highlighted references:

References:

Figures and legends

Figure 1: Overview of the general CLIP workflow

Figure 2: Overview of primary CLIP variants and TRIBE

Figure 3: Overview of CLIP analysis.

[Figure 4: CLIP applications in model organisms.](#)

[Tables and boxes](#)

Table 1: Characteristics of the commonly used proximity enzymes

Table 2: Available peak detection software

Box 1: Purification of RBP-RNA complexes in CLIP

Supplementary Table: Key methods to identify protein partners of a specific RNA

[H1] Glossary

Abstract

During transcription, RNA molecules start assembling with proteins into ribonucleoprotein (RNP) complexes. The dynamic RNP assembly, largely directed by cis-acting elements on the RNA, coordinates all processes in which the RNA is involved. Here we discuss crosslinking and immunoprecipitation (CLIP) and complementary, proximity-based methods that have been developed for identifying the cis-acting sites bound by a specific RNA-binding protein on endogenous RNAs. We also discuss RNA-centric methods that identify the protein partners of a specific RNA. We review the main variants of these experimental methods and the strategies for their optimization and quality assessment. We summarize the main challenges of computational CLIP data analysis, how they handle various sources of background and how to identify functionally relevant binding regions. We outline the various applications of CLIP, and the databases of available data. We discuss the prospect of integrating the increasing amounts of data obtained by CLIP with complementary methods to gain a comprehensive view of RNP assembly and remodeling, unravel the spatial and temporal dynamics of RNPs in specific cell types and subcellular compartments, and understand how defects in RNPs that can lead to disease. Finally, we present open questions, directions for further development and applications.

[H1] Introduction

Ribonucleoprotein (RNP) complexes are key to every step in RNA processing and function. Once transcription is initiated, proteins start to interact with nascent RNAs. The protein complement decorating an RNA molecule dynamically changes in space and time, orchestrating RNA processing and function in the nucleus and cytoplasm¹. Understanding the roles these RNA-binding proteins (RBPs) play requires methods that identify the set of RNAs that they bind in cells at specific developmental, activity or disease states. This primer is focused on methods relying on crosslinking and immunoprecipitation (CLIP) that identify specific RNA sites crosslinked to RBPs by UV light². Additionally, the primer covers complementary approaches, in particular

proximity-based and RNA-centric methods. These methods offer a great opportunity to be integrated with CLIP to reveal the architecture and functions of specific RNPs.

The best understood interactions of RBPs with RNAs are those mediated via structurally defined RNA binding domains^{3,4}, but recent studies are also uncovering interactions mediated by **intrinsically disordered regions (IDRs)**¹. The most common domain is the RNA recognition motif (RRM), composed of about 80 amino acids that typically consist of four antiparallel beta-strands and two alpha-helices with side chains that stack with up to four RNA bases. The heterogeneous nuclear RNP K-homology (KH) domain is composed of about 70 amino acids that typically recognize four nucleotides in single-stranded RNA mostly through hydrophobic interactions. The double-stranded RNA binding domain (dsRBD) recognizes mainly the sugar phosphate backbone, but can achieve specificity by recognising the shape of the A-form RNA helix or forming sequence-specific contacts with the edge of RNA bases in the minor groove⁵. Whereas a single RNA binding domain displays limited sequence specificity, RBPs are often modular, comprising more than one RNA binding domain of the same type or combining multiple types. A prime example of exquisite specificity through multiple domains are the Pumilio proteins. The PUM homology domain consists of eight repeats, each of which interacts with one nucleotide in the eight nucleotide long recognition sequence. Moreover, RBPs further increase their RNA specificity by interacting with each other upon RNA binding, thus assembling into RNPs¹.

A number of methods can characterize the RNA interactions that coordinate RNP assembly. These approaches can be protein-centric, describing the whole compendium of RNA sites bound by a specific RBP in a biological sample, or RNA-centric, identifying the RNA-bound proteome. The most common protein-centric strategies are based on immunopurification of an RBP and associated RNAs, and can be broadly categorized as RNA immunoprecipitation (RIP) approaches or CLIP approaches. RIP approaches mostly purify the RNA-protein complexes under native conditions and without further stabilization, so that only low stringency can be applied during IP^{6,7}. In some studies, formaldehyde is used for crosslinking RNA and bound proteins in the cell^{8,9}. However, formaldehyde also crosslinks proteins, so its use in RIP likely also leads to isolation of transcripts bound by RBPs interacting with the IPed protein.

Currently, CLIP techniques are most widely used (Fig. 1). CLIP relies on irradiation of cells by UV light, where proteins in the immediate vicinity of the irradiated bases are irreversibly crosslinked to the RNA by a covalent bond¹⁰. Although the efficiency of UV light is much lower than the one of formaldehyde, the covalent crosslinks upon UV irradiation allows stringent purification of the RNA-protein complexes. This is followed by a series of steps to determine the direct interactions of a specific protein across the transcriptome. While RIP traditionally enriches for full length bound RNAs, CLIP uses a limited RNase treatment of crosslinked RNPs to isolate RNA fragments occupied by the RBP. Sequencing and computational analysis of these fragments then helps identify the binding sites of an RBP, which in turn reveals critically important details of RBP function, such as the location of binding sites relative to other *cis*-acting elements or RBP binding sites.

The development of high-throughput sequencing of RNA isolated by CLIP (HITS-CLIP) has enabled a transcriptome-wide view of RNA binding sites¹¹. CLIP techniques have been further

developed to identify crosslink sites with nucleotide resolution either through analysis of mutations in reads (photoactivatable ribonucleoside-enhanced CLIP, or PAR-CLIP)¹², or by capturing cDNAs that terminate at the crosslinked peptide during reverse transcription (individual-nucleotide resolution CLIP, or iCLIP)¹³. Concomitant with the refinements of the experimental protocols, the development of dedicated bioinformatics workflows has allowed the determination of binding sites and consensus motifs that elucidate the language of posttranscriptional regulation¹⁴.

Of the wide range of protein-centric methods developed, this Primer focuses on experimental and computational aspects of CLIP methods that have either been broadly adapted or have produced widely-used data. We also cover the identification of RBP binding sites by tagging RBPs with enzymes that naturally act on RNA where the resulting RNA modifications can be identified by high throughput sequencing¹⁵, as well as the use of subcellular compartment-specific proximity labelling to study localised transcriptomes¹⁶. Finally, we discuss the applications of these techniques in multiple model organisms to obtain a systems-level view of RNP assembly and dynamics, and review strategies for method optimization and quality assessment of the data. For comprehensive reviews of additional protein-centric methods, we refer the readers to a number of recent reviews^{2,17,18}.

In addition to the protein-centric methods, we also review the RNA-centric methods that identify the proteome bound to a specific RNA, with the aim to encourage integrative use of these methods to gain a complementary view of RNP assembly and remodelling. We don't cover the studies that identify the full spectrum of candidate RBPs bound to polyadenylated mRNAs, or to all types of RNAs, as these have been recently reviewed elsewhere¹. These studies have greatly expanded our knowledge of proteins capable of forming UV crosslinks with RNA, which has identified not only known RBPs, but also proteins without known RNA-binding domains or functions, suggesting that a wider range of proteins, especially those with **IDRs**, are capable of RNA binding. The RNA binding roles of this wide array of proteins can now be validated and investigated further by using the methods presented in this primer.

[H1] Experimentation

[H2] Protein-centric methods

All CLIP-based methods to determine the binding landscape of RBPs transcriptome-wide share the following core workflow. First, RNAs and interacting proteins are irreversibly crosslinked by UV light (UVC at $\lambda = 254$ nm, or UVA/B at $\lambda = 312-365$ nm for PAR-CLIP, see below) in intact cells. UV crosslinking energy and conditions need to be adapted to whether cell monolayers, a suspension of dissociated tissue¹⁹, or whole tissue, such as worms²⁰ and plants^{21,22}, are used. For tissues that can't easily be dissociated because they are too hard - which includes most adult mammalian tissues, plants, or post-mortem human tissues, frozen tissue can be ground in liquid nitrogen to a fine powder and crosslinked on dry ice^{22,23}. Next, RNAs are trimmed to short fragments by carefully optimised partial RNase digestion and the crosslinked RNP of interest is

stringently purified from the mixture using immunoprecipitation (IP) or other methods² (Box 1). The RNPs are then further fractionated by denaturing polyacrylamide electrophoresis (SDS-PAGE), and crosslinked RNA fragments released by digestion of the RBP, usually by Proteinase K. The yield of RNA fragments is typically in the low ng range and thus needs to be converted into cDNA for high-throughput sequencing using protocols optimised to work with limited amount of short RNAs. Sequenced reads are mapped to the genome and clusters of overlapping reads representing possible binding sites are computationally separated from the usually high levels of background^{12,24,25}. In order to reveal sites that are most likely functional, i.e. conferring posttranscriptional gene regulatory effects, the list of binding sites can be sorted according to various criteria such as relative RBP occupancy, which can be understood as the fraction of all instances of a binding site occupied by the RBP at the time of crosslinking²⁶.

Each variant of CLIP differentiates itself by a unique approach to one or more of the above-mentioned steps. We describe the differences among primary variants below, with further comparisons and additional variants being covered elsewhere². We do not intend to advocate one variant over another, but the provided information can help researchers to make an informed choice of their preferred CLIP variant. It is clear that RBPs differ greatly in their crosslinking efficiencies between each other, and the efficiency differs when UVC, 4SU-induced UVA/B or formaldehyde crosslinking are used^{27,28}. More studies are needed to determine how these relative efficiencies are affected by the way proteins are recruited to RNA (direct or indirect via other RBPs), by specific features of RNA motifs bound by the RBPs, and by the type of contacts between specific amino acids and RNA. Insights into these questions are bound to arise by comparative analyses of the increasing amounts of available data, especially by comparisons of data for same RBPs produced with multiple methods.

[H3] Original CLIP, and its adaptation to high-throughput sequencing

While UV crosslinking of RNA and interacting proteins, with and without RNA labeling with photoreactive nucleosides, has long been used to study protein-RNA contacts of ribosomal RNA (rRNA) or of viral RNAs in bacterial and human cells²⁹⁻³³, this approach was combined with sequencing as the starting point in the development of CLIP that crosslinked cells with UVC light (Fig. 2)¹⁰. UVC preferentially crosslinks RBPs to uridines and to a lesser degree guanosines as revealed by mass spectrometry³⁴ or mutational analysis of the sequenced cDNA³⁵⁻³⁷. Following mild RNase digestion and purification of the selected RBP, RNA fragments are ligated to a 3' adapter and radiolabelled, to visualize and aid purification of the crosslinked RNP after SDS-PAGE and membrane transfer¹⁹ (Box 1). Crosslinked RNA fragments are recovered, ligated to a 5' adapter, converted into cDNA by reverse transcription, and amplified by PCR, similar to the standard protocols developed for miRNA characterization³⁸. However, here the reverse transcriptase needs to read across a major roadblock formed by the oligopeptide attached to the crosslinked nucleotide to reach the 5' adapter. Premature termination can result in a bias towards contaminating non-crosslinked sequences in resulting cDNA libraries. Computational tools have therefore been developed to leverage a low but consistent mutation signature at such events^{24,39,40}. CLIP was adapted to high-throughput sequencing in HITS-CLIP¹¹ by adding additional sequences to the PCR primers required for Illumina sequencing¹¹. Moreover, CRAC

(crosslinking and analysis of cDNAs)³⁹ has introduced affinity-based purification under denaturing conditions as an alternative to IP, which has been particularly valuable in yeast.

[H3] Individual-nucleotide resolution CLIP, infrared CLIP and enhanced CLIP

Individual-nucleotide resolution CLIP (iCLIP)¹³, infrared CLIP (irCLIP)⁴¹, and enhanced CLIP (eCLIP)⁴² differ from the original CLIP in their purification and cDNA library preparation strategies (Fig. 2, Box 1). They take advantage of the tendency of reverse transcriptase to terminate at the crosslinked nucleotide, which yields cDNAs with a 5' end mapping to the first nucleotide downstream of the crosslinking site. This increases the sensitivity of the method by efficiently and more comprehensively amplifying cDNAs produced from crosslinked RNAs, removes the bias towards non-crosslinked reads, and allows identification of crosslink sites with nucleotide-level resolution by analysing cDNA truncations. To introduce primer binding sites for cDNA library amplification, iCLIP uses a cDNA circularisation approach, similar to the most common ribosome footprinting protocol⁴³; reverse transcription is primed with a long DNA oligonucleotide containing both PCR primer sites, and the cDNA products are circularised using thermostable RNA ligases that also act on DNA⁴⁴. At least 18 later variants of CLIP have adopted the approach to amplify truncated cDNAs²; some, such as irCLIP, use cDNA circularisation similarly to iCLIP, while others, such as eCLIP and iCLIP2⁴⁵, use a highly concentrated T4 RNA ligase 1 to ligate a DNA adapter to the 3' end of the cDNA.

[H3] Photoactivatable ribonucleoside-enhanced CLIP

PAR-CLIP¹² shares the cDNA library construction strategy with CLIP¹⁹ but differs in UV crosslinking. In a first step, cultured cells are incubated with nucleosides modified with an exocyclic thione group, specifically 4-thiouridine (4SU) or 6-thioguanosine (6SG), which are then incorporated into nascent RNAs (Fig. 2). The exocyclic thione group increases the photoreactivity of the base, allowing crosslinking with lower energy UVA/B light than other CLIP methods ($312 \leq \lambda \leq 365$ nm). When using 4SU, crosslinked amino acids are attached to position 4 of the base and change its base-pairing properties, while unmodified uridines crosslink at position 5, which leaves their **Watson-Crick face [G]** intact⁴⁶. Crosslinked 4SU preferentially pairs with guanosine during reverse transcription, resulting in a characteristic T-to-C transition in the sequenced cDNA (or a G-to-A transition when using 6SG)¹². This may simplify data analysis as enrichment of such transitions at specific genomic regions indicates *bona fide* interaction sites and helps in the determination of the precise location and strength of interaction with the RBP (see analysis section below).

[H3] Mapping of RNA secondary structures interacting with RBPs using CLIP

Some RBPs, including the intensely studied Staufe family, or the Argonaute proteins at the heart of miRNA or piRNA small RNA directed silencing pathways bind RNA at double-stranded

sequence elements. However, standard CLIP assays will only reveal one of the bound strands, thus losing information on the nature of the RNA-RNA interaction. All major CLIP variants were retooled to preserve the sequence of RNA-RNA hybrids around the interaction site of the RBP of interest to include an intermolecular ligation step to link the two RNA fragments bound to the RBP after the limited RNase digestion step using RNA ligases. Argonoute HITS-CLIP⁴⁷, CLASH⁴⁸ (crosslinking and sequencing of hybrids) and modified PAR-CLIP⁴⁹ were used to sequence miRNA:target chimeras, and hiCLIP⁵⁰ (RNA hybrid and iCLIP) revealed a prevalence of long-range intramolecular RNA duplexes bound by human STAU1 protein. These are complementary to the many additional methods that profile RNA structures on a transcriptomic scale by chemical-based approaches or by mapping RNA-RNA contacts¹⁷. CLIP has recently been integrated with one such chemical-based approach, selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) to reveal the hydrogen bonds at RNA-Protein interfaces⁵¹.

[H3] Targets of RNA-binding proteins identified by editing

Enzymatic tagging has been developed for transcriptome-wide identification of endogenous RBP interaction sites without requiring crosslinking, biochemical IP or complicated cDNA library preparation steps. An example is Targets of RNA-binding proteins identified by editing (TRIBE)¹⁵ that is conceptually related to DNA adenine methyltransferase identification (DamID), a method that identifies chromatin protein-bound regions by fusing them to the Dam methyltransferase and identifying the methylation sites⁵². TRIBE relies on transgenic expression of the RBP of interest fused to the catalytic domain of RNA-editing enzyme ADAR, which catalyses adenosine-to-inosine (A-to-I) conversions nearby the RBP interaction sites, or its hyperactive mutant (HyperTRIBE)⁵³. These sites are revealed by an excess of A-to-G mutations in libraries that are prepared as standard RNA-seq libraries (Fig. 2). Among the distinct advantages of TRIBE over CLIP approaches are its minimal number of manipulation steps, which allow for the use of small numbers of cells and the possibility to express the RBP-ADARcd fusion protein in a cell-type specific manner in model organisms to reveal the RBP interactomes in precisely-defined subpopulations of cells. Disadvantages are that very deep sequencing is necessary to capture sufficient editing signal to call interaction sites and that C-terminal or N-terminal fusions of ADARcd may compromise localisation and activity of some RBPs, their ectopic expression *in vivo* requires optimisation to ensure proper cell-type specific expression patterns and avoid excessive levels of RBP-ADARcd fusion protein levels, which can obscure target sites and lead to toxicity by hypermodification of RNA. Recently, an additional approach, termed STAMP (Surveying Targets by APOBEC Mediated Profiling), has been developed by tagging RBPs with APOBEC⁵⁴. APOBEC enzymes access cytosines in single-stranded RNA and produce clusters of edits, thus it leads to increased coverage of mutations compared to regions, which relies on ADAR-mediated editing of the relatively infrequent RNA duplexes containing a bulged mismatch¹⁵. This higher likelihood of encountering APOBEC1 cytosine substrates increases the sensitivity of STAMP and enables it to be coupled with single-cell capture.

[H3] Proximity-CLIP

CLIP experiments are typically performed from total cells under single conditions and thus yield limited insights into the spatiotemporal dynamics of RNA metabolism. One key aspect of post-transcriptional gene regulation is the controlled localization of mature transcripts and their precursors within the cell. Subcellular RNA localisation can be studied using biochemical fractionation or microscopy^{55,56}. Nevertheless, biochemical fractionation is limited to the analysis of a small number of relatively large cellular structures, and advanced RNA-FISH (fluorescent *in situ* hybridization) techniques, such as seqFISH⁵⁷ and merFISH⁵⁸, which allow localisation of multiple RNAs in a single cell, are constrained by long experiment times, complex data analysis, and probe design.

A number of recently-developed techniques overcome some of these obstacles by performing compartment-specific labelling and analysis of RNA and/or proteins. In one type of approach, genetically encoded photosensitizers localised to specific compartments mediate the oxidation of proximal guanosines by generating reactive oxygen species after irradiation with visible light⁵⁹⁻⁶¹. Photosensitised guanosines can then be coupled with reactive amino-group-containing probes to isolate and quantify localised RNA. Proximity-CLIP¹⁶ and the closely related technique APEX-seq⁶²⁻⁶⁴ allow the determination of RNA distribution to specific subcellular locations. Both techniques rely on the biotinylation of RNAs and proteins by the engineered ascorbic acid peroxidase protein APEX2⁶⁵, a tool widely used to quantify the localised proteome⁶⁶ (Table 1). To allow subcellular-compartment specific biotinylation, APEX2 is typically fused to specific localisation elements. Proximity-CLIP relies on the assumption that all cellular RNAs interact with RBPs throughout their life-cycle⁶⁷, and thus, proteins that are biotinylated in a specific compartment are isolated with streptavidin from UV-crosslinked cells, thereby enriching compartment-specific transcripts. Prior to protein biotinylation, nascent transcripts are labelled with either 4SU and 6SG, and crosslinked to interacting RBPs with 312-365 nm UV light, analogous to PAR-CLIP. The compartment-specific proteome, including crosslinked RNPs, are then isolated on streptavidin beads and, following a mild RNase digestion, crosslinked RNA fragments are isolated and sequenced. The characteristic mutations in the cDNA resulting from the use of photoreactive nucleosides reveal crosslinked sequences. A distinctive feature of Proximity-CLIP is that the sequencing of RBP protected footprints not only allows for profiling of localised RNAs, but also for the identification of protein-occupied, and thus possibly regulatory, *cis*-acting elements on RNA. In contrast to APEX-seq, this approach provides a snapshot of regulatory elements on RNA that are occupied in the examined compartments.

[H2] RNA-centric methods

Regulation of any specific RNA is coordinated by RBPs that directly bind to the RNA, as well as by additional proteins that assemble with the RNP through protein-protein interactions. To unravel the composition of full RNPs, RNA-centric methods are needed to complement the information on the direct protein-RNA contacts revealed by the protein-centric approaches. Such methods comprehensively identify proteins that assemble on a given RNA, using two broad categories: RNA affinity-capture purification, or proximity-based protein labelling.

[H3] RNA affinity proteome capture

RNA affinity proteome capture methods are mainly *in vitro* approaches based on either tagging the endogenous RNA or modifying *in vitro* transcribed or synthesized RNA at the 3' or 5' or both ends with biotin or similar small molecules⁶⁸ and immobilizing them on solid surfaces such as streptavidin beads (Supplementary Table 1). Cellular extracts are then added on the immobilized beads and proteins bound to the labelled probes are washed and eluted from the beads for proteomic analysis by boiling mostly in a 2% SDS containing elution buffer. An alternative approaches is to tag an RNA of interest with virus-derived heterogeneous RNA stem-loops or aptamers like MS2⁶⁹, PP7⁷⁰, S1⁷¹, Cys4⁷², D8⁷³, or similar heterogeneous aptamers such as those that mimic tobramycin⁷⁴ or streptomycin⁷⁵. While choosing the aptamer, one has to consider the binding affinity with the cognate ligand, keeping in mind that for highly enriched RNPs, a lower binding affinity aptamer-ligand interaction can be sufficient to pull down high enriched interactors, and will give less background with more specific elution. After lysing the cells expressing the tagged RNA of interest, the lysates are passed through beads containing the respective substrates. These are stringently washed, which can include applying a competitive binder, and the proteins are eluted for mass spectrometry analysis.

Post-lysis reorganization of RNPs⁷⁶ may result in detection of false positive association of RBPs with specific RNA baits. To avoid this concern, several approaches crosslink RNPs in cultured cells by UV with or without photoreactive nucleosides or chemically with formaldehyde prior to cell lysis (Supplementary Table 1). For example, CHART (capture hybridization of analysis of RNA targets) allowed mapping of interaction sites and proteins bound to the *Drosophila* roX2 RNA⁷⁷; RNA antisense purification (RAP) was used to identify the interactome of XIST⁷⁸ and NORAD⁷⁹; comprehensive identification of RBPs by MS⁸⁰ (ChIRP-MS) also systematically identified mouse Xist interacting proteins; *in vivo* interactions by pulldown of RNA (vIPR) studied proteins interacting with *C. elegans* gld1 RNA⁸¹. During the recent COVID-19 public health emergency caused by Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), RAP and ChIRP-MS were immediately applied to identify host and viral RBPs interacting with the SARS-CoV-2 RNA genome^{82,83}.

[H3] RNA-directed proximity-based proteome labelling

RNA-directed proximity-based methods investigate the proteome on specific RNA in its native cellular context without the need for crosslinking, which is particularly important to uncover transient interactions, and to study RNPs from poorly soluble cellular compartments like chromatin, peroxisomes or Golgi body that are prone to precipitate during affinity-capture methods. In these methods, a proximity enzyme is recruited to a specific RNA to covalently modify the proteins located in vicinity of the RNA (Supplementary Table 1). An RNA can be tagged with MS2 or similar aptamers, and various types of proximity enzymes can be tagged with the corresponding loop-binding protein that recruits them to the RNA (Table 1). In these cases, the RNA is usually expressed from a reporter plasmid together with BoxB stem loop and also expressing BASU (a mutant version of BirA*, engineered from *Bacillus subtilis*) together with BoxB stem loops binding λN peptide allowing RNA-protein interaction detection (RaPID)¹⁸ or the RNA of interest can also be tagged endogenously in the case of approaches such as RNA-BioID⁸⁴.

Alternatively, a modified CRISPR/Cas system can be used to recruit an enzyme to an endogenous RNA by tagging the enzyme with a Cas variant, and using CRISPR RNAs that are antisense to the RNA of interest⁸⁵. The excess pool of catalytically active enzymes not docked to the tagged RNA can produce unspecific signal in these methods, but this can be improved by using split proximity-based RNA assisted tools such as split APEX2, where two inactive APEX2 subunits are reconstituted to restore peroxidase activity upon physical colocalization⁸⁶.

[H1] Results

[H2] CLIP analysis workflow

Although all CLIP variants aim to capture individual binding sites of RBPs with nucleotide-level resolution, the precise steps of the complex experimental approach determine the products that are obtained and, consequently, the computational analysis that is necessary for revealing the binding sites. First, quantification of CLIP reads can be complicated due to the presence of PCR duplicates resulting from non-uniform amplification of different sequences during the PCR steps. Careful optimization of PCR cycle numbers⁸⁷ and the use of unique molecular identifiers (UMIs) in cDNAs produced by iCLIP and most current CLIP variants can mitigate introduction of these artifacts². Computational tools, such as iCount¹³, an expectation-maximization (EM)-based algorithm⁸⁸ or UMI-tools⁸⁹ take advantage of the presence of UMIs to quantify the number of unique cDNAs in the library while taking into account the abundance of each UMI and sequencing errors. Second, cDNA mutation and/or truncation patterns, caused by the reverse transcriptase reading past the crosslinked nucleotides or truncating at them, are exploited by many computational tools to achieve nucleotide level-resolution in the identification of crosslinks and corresponding binding sites. However, as both readthrough and mutation are stochastic and have relatively low rates, a sufficient number of cDNAs is needed to identify individual crosslink sites, which may be an issue for targets with low expression levels.

Beyond these initial pre-processing steps, workflows for CLIP data analysis generally cover the following main steps: peak identification from individual samples, combined analysis of replicates to identify reproducible peaks, and finally, meta-analysis of the peaks to identify binding motifs, relationships between binding sites and transcript landmarks, and to infer the functional consequences of binding (Fig 3). We provide a summary of tools for binding site identification (peak detection) that were described or updated in the past five years (Table 2). Descriptions of software for finding motifs or predicting RBP binding sites, of peak finding tools applicable only to a restricted set of targets (e.g. of miRNAs) or published over five years ago can be found in other recent reviews^{14,90}.

[H3] Extrinsic and intrinsic background in CLIP experiments

Like all high-throughput methods, CLIP reads originate from a large number of RNAs, even when the RBP is thought to have few RNA partners, as for example, the histone RNA stem loop-binding protein. Why large numbers of distinct RNAs are represented in CLIP libraries is only partially understood. For example, it has become apparent that functional outcomes require interactions

with a high total residence time of the RBP on the RNAs, on regions that vary in length, sequence composition, etc. among RBPs. Thus, binding regions that accumulate a high number of reads, either narrow 'sites' or broader RNA subsequences are likely functionally-relevant⁹¹, while the more isolated, transient interactions may represent an 'intrinsic' background in CLIP experiments. There is no absolute distinction between stable vs. transient interactions, and the functionality of these modes of interaction differs between RBPs. For example, CLIP of MEG3 in *C. elegans* showed that its function depends on non-specific interactions across the full transcripts⁹². Thus, thought needs to be given to what may constitute intrinsic background for different RBPs.

On the other hand, the antibodies that are used to immunoprecipitate RBPs can have limited selectivity, leading to some contamination of the sample with additional RBPs and their bound RNAs. Fragments of abundant RNAs may also be carried through the sample preparation. The quality control and purification of the RBP-RNA complexes of interest on the SDS-PAGE gel is important in analysing and mitigating these two sources of 'extrinsic' background, and the way this step is implemented can vary between CLIP protocols (Box 1). It is advisable that control experiments, using IgG-bound beads or antibody-bound beads with RBP-knockout material, are prepared in parallel, barcoded and pooled before sequencing, so that their similarity to parallel experiments can be used to assess data specificity.

[H3] Peak identification

Peak identification is an important step that serves to identify regions where the RBP has high occupancy, thereby representing direct and likely functionally relevant interactions, from various types of background. In ChIP-seq, IP with beads lacking antibody is used to generate a background sample for peak calling. In CLIP experiments, however, it is more challenging to generate experimental 'background' samples. For example, when performing CLIP with beads that lack the antibody, the signal on SDS-PAGE is negligible, yielding 100-fold fewer reads if sequenced, which is insufficient for background modelling¹³. Therefore, a more common approach is to computationally remove the 'intrinsic background' of very transient interactions by identifying regions with a high density of reads or crosslinking-induced features relative to other regions within the corresponding genes (same intron, coding sequence or 3'UTR) that have similar properties, such as expression level. The crosslinking-induced features could be mutations, insertions/deletions, or truncations, depending on the experiment (Fig. 3a-b). They are generally assumed to take on (zero-truncated) binomial distributions and can be used as parts of hidden Markov models that may also include additional features, such as the coverage of the sites and sequence biases due to technical issues or due to the specificity of the RBP⁹³⁻⁹⁵. For truncation-based CLIP variants, experimental approaches have also been established to estimate the frequency of cDNA truncation^{96,97}.

Peaks of diagnostic features can also result from the contamination with 'extrinsic background', as explained earlier²⁵, especially in the case of RBPs that crosslink poorly to RNA, when the signal is more easily overridden by co-precipitating RNPs. Addressing this type of 'background' is possible by producing additional experimental samples. For instance one can use the abundance of each RNA region estimated from RNA-seq to identify those regions that are in high abundance and yield a large number of CLIP reads in spite of their low occupancy by the RBP. Outliers are

identified with respect to a negative binomial distribution (or the zero-inflated negative binomial) whose parameters are determined from the background sample. These distributions capture the fact that the variance in coverage is generally larger than the mean, as would be expected from a simple sampling of the reads¹⁴. A related approach to assess the background experimentally has been taken in eCLIP, where a size-matched input (SMInput) is generated by leaving out the IP, but otherwise performing all the steps of the protocol⁴². The importance of background samples was illustrated in eCLIP by the example of the stem-loop binding protein, where only 1.2% of the peaks identified from the foreground sample were enriched over the background SMInput⁴².

While current approaches to experimentally assess the background do increase the proportion of functionally relevant binding sites among the called peaks, it remains unclear whether the ranking of these sites is quantitatively related to their functionality and whether new biases are introduced. The SMInput sample in eCLIP differs from the IP sample in unintended ways, being likely dominated by RNAs crosslinked to abundant RBPs, which may not be the same RBPs that contaminate specific CLIP data due to their interactions with the RBP-of-interest. Conversely, in some cases the SMInput may be dominated by the RBP-of-interest itself, which would result in the foreground signal becoming erroneously assigned to the background, precluding the identification of RBP's binding sites. Returning to the example of the stem-loop-binding protein 88% of the peaks enriched over background were located in the expected loci of histone mRNAs⁴², but whether the enrichment of a peak reflects the residence time of the protein on the mRNA is not known. RNA-seq may also introduce bias as the different protocols that are in use (e.g. using poly(A) selection or ribosomal RNA depletion) impact the gene/transcript expression estimates. Poly(A) selection enriches for fully-processed RNAs, thereby depleting introns. Even within a gene, the coverage of introns among nuclear RNAs varies from intron to intron, depending on the time taken for the transcription, splicing and degradation. Moreover, the delay between transcription and co-transcriptional splicing leads to increased coverage towards the 5' end of long introns⁹⁸, which is common in genes expressed in the brain⁹⁸⁻¹⁰⁰. Such issues suggest that it will be important to obtain data that can accurately estimate the local coverage of intronic regions in order to model enrichment of intronic CLIP peaks.

Finally, most RBPs are localized to specific cellular compartments, where the abundance of RNAs they encounter may be quite different from the average abundance inferred from total RNA. Thus, it would be valuable to develop models for normalization of CLIP data based on the subcellular RNA abundance that each RBP encounters, employing estimates of subcellular RNA abundance provided by RNA-seq from cellular sub-fractions, APEX-seq and/or Proximity-CLIP. Finally, it is important to be aware that a gain in specificity via increased stringency of peak calling can lead to a drop in sensitivity, as discussed later. All of these considerations make it clear that a rigorous benchmarking of methods for background modeling in CLIP experiments is necessary.

[H3] Characterizing RBP binding motifs

Once the binding peaks have been identified, the immediate question is to uncover the sequence and/or structure specificity of the protein. Traditionally, **position-specific weight matrices (PWMs)** **[G]** have been used to represent the sequence specificity of nucleic acid binding proteins, whether

transcription factors or RBPs (Fig. 3c). PWMs indicate the relative frequency with which individual nucleotides are observed among the binding sites of an RBP, which, in turn, can be related to the contribution of individual nucleotides in the binding site to the energy of interaction with the RBP, and thereby to the affinity of this interaction. PWMs can be inferred from sequences obtained in CLIP experiments with readily available computational tools^{101–103}. A key assumption of PWMs is that nucleotides in the binding site contribute independently to the energy of RBP-RNA interaction. This assumption started to be questioned as high throughput binding data (e.g. from protein microarrays) became available. It has been argued that much more parameter-rich models (derived for example through machine learning approaches) are necessary to quantitatively explain measurements of affinity of protein-nucleic acid interactions^{104–106}. However, other studies that explicitly modeled confounding experimental factors concluded that PWMs are sufficient to quantitatively explain the binding data, at least for the majority of transcription factors¹⁰⁷. In the case of RBPs, PWMs also explain relatively well both CLIP data and *in vitro* measured affinities of interaction^{108,109}. Furthermore, a detailed analysis of Gld-1 binding in *C. elegans* found that a biophysical model that included the PWM-defined specificity of the Gld-1 RBP along with the predicted structural accessibility of binding sites in RNAs was able to explain the relative enrichment of binding sites in CLIP, alleviating the need for a more parameter-rich model¹¹⁰.

As hinted above, RNA-RBP interactions are likely more complex than the interactions of transcription factors with DNA. The accessibility of binding sites, modulated through the RNA secondary structure that further depends on RNA modifications¹¹¹, plays an important role in RBP-RNA interactions. Once the secondary structure around CLIP binding sites was explicitly examined^{112,113}, it became apparent that the recognition of RBP binding motifs by RBPs may require a specific structural context, rather than a single-stranded conformation, leading to models that simultaneously infer the sequence-structure preference of RBPs^{114–116}. These allow identification of sites that were missed in CLIP experiments, due, for example, to the low expression level of the RNAs¹¹⁴. Similarly, machine learning approaches have been deployed to increase the depth of miRNA binding site identification starting from Argonaute-CLIP data¹¹⁷. Biophysical approaches to the *ab initio* prediction of molecular interactions can also pinpoint potential false negatives of CLIP experiments, as well as provide insights into the interaction propensities that ultimately determine the location of binding sites in RNAs¹¹⁸. CLIP data provides the opportunity to infer biophysical models of RNA-RNA interactions in the context of ribonucleoprotein complexes, such as the ternary miRNA-mRNA-Argonaute protein complex¹¹⁹, models that can predict *in vitro* measured affinity interactions with surprising accuracy¹²⁰.

Many tools take into account crosslinking-induced mutations to call the RBP binding sites and determine the sequence (and structure) specificity of the RBP^{35,94,95,115}. Annotation of the putative binding sites (location with respect to various landmarks such as splice sites, functional category of the gene, etc.) as well as binding data for RBPs other than the one used in the experiment can be further incorporated to improve the accuracy of binding site identification^{121,122} (and benchmarks within). A drawback is that these phenomenological models do not have a clear mechanistic interpretation. Furthermore, increasing goodness of fit by adding additional parameters is not always desirable. As new approaches are proposed, it will be important to start assessing how they compare with respect to the balance between the goodness of fit and the number of parameters.

[H3] Regulatory grammar

The final step in deciphering CLIP data is to uncover the regulatory grammar of the RBP binding sites, including the spatial relationship of RBP binding sites to important transcript categories (coding/non-coding, repeats, snoRNAs, rRNAs etc.) and landmarks (exons, introns, exon/intron boundaries, translation start/stop sites)¹²³. The binding site data are combined with data from perturbation experiments (knockdown or overexpression of the RBP of interest) to generate 'RNA maps' reflecting the functional impact of binding sites located in different transcript regions¹²⁴. RNA maps can be used to assess the quality of data sets or analysis methods, because the shape and amplitude of positional signals should have a direct correspondence to the accuracy of the method¹⁴. Computational modeling of changes in expression of transcript isoforms upon perturbation of individual RBPs provide complementary information regarding the RBP binding motifs that are involved, their location within transcripts and their functions in individual steps of RNA processing¹²⁵. As the number of RBPs studied by CLIP continues to increase, direct comparisons of the binding site profiles in the genome are starting to reveal regulatory complexes and competition between RBPs. Both of these are reflected in multiple proteins binding to closely-spaced sites in the RNA, while the data from perturbation experiments helps resolve the nature of the interactions between RBPs^{123,126,127}.

[H2] Assessing the specificity of CLIP

When compared to related methods such as RIP or ChIP-seq, CLIP has a unique in-built capacity for experimental quality control of specificity. The visualization of size-separated protein-RNA complexes and appropriate negative controls helps to estimate the likely sources of extrinsic background before proceeding to cDNA library generation. The initial CLIP publication already set high standards of specificity, as evident by the absence of extrinsic background in negative control (control serum) and the >20x enrichment of binding motifs within Nova CLIP reads compared to control¹⁰. Fusion of affinity-tags to the studied RBP further increases specificity by allowing the use of stringent, denaturing purification conditions². Nevertheless, data specificity for the IP-based variants of CLIP can vary depending on the quality of the antibody and the effort put into optimizing the conditions. When studying a new RBP by CLIP, several steps require routine optimization, including the RNase fragmentation and the IP conditions, which need to be adjusted to variations in RNase stocks, crosslinking efficiencies of RBPs, the stability of their interactions with other RBPs, and the type of cells or tissue used^{19,37}.

As these optimizations are carried out to variable extents across the many labs employing CLIP, the need for computational assessment of CLIP data specificity has increased, in order to facilitate integration of the large number of collected datasets. A simple, qualitative view is provided by the crosslink distribution across RNA types (Fig. 3C); nuclear and cytoplasmic RBPs tend to have most crosslinks in introns and exons, respectively, and, in cases where the dominant RNA binding partners are known, these are expected to rank highly in the data. Nevertheless, aside from being only qualitative, this assessment of specificity can be misleading when the studied RBP interacts and co-purifies with other RBPs that have similar localization and RNA partners, which are a likely source of extrinsic background.

The second approach is to compare the enrichment of sequence motifs in CLIP data with their affinities for the purified RBP as determined by biophysical methods. For example, systematic motif enrichment data from *in vitro* binding assays has started to become available^{112,128}. Often but not always, the *in vivo* identified binding sites resemble the highest-affinity motifs derived from *in vitro* methods such as SELEX^{129,130}, RNA Bind'n'seq¹²⁸ and RNAcompete¹¹². A challenge of this approach is that the *in vitro* assays have biases of their own, for example they often examine binding of individual domains rather than full-length protein, which furthermore lack post-translational modifications and the context of other proteins. They also tend to study binding to short RNA sequences, while *in vivo* RBPs can assemble on long RNAs with more complex secondary structures. As more CLIP data for the same RBPs becomes available, it will be informative to compare the extent of same motif enrichment across datasets to better understand the origin of differences in their specificities.

However, for many RBPs little orthogonal knowledge is available to instruct the anticipated results, and other approaches are needed to assess specificity. Binding motifs can be identified *de novo* from CLIP data and the extent of their enrichment provides some measure of data quality. For example, a comparison of publicly available data for polypyrimidine tract binding protein 1 revealed that, while all CLIP variants show enrichment of similar motifs, the extent of the enrichment varies dramatically, indicating major variations in data specificity³⁷. There are several caveats to *de novo* motif discovery from CLIP data, as factors unrelated to the studied RBP may result in enrichment of specific sequence motifs, such as the aforementioned nucleotide preferences of UV crosslinking or sequence biases of RNases and RNA ligases used to join adapters to the ends of RNA fragments^{24,36,37,87}. One way to decrease such technical biases is by producing parallel datasets for diverse RBPs from the same type of biological material, and then deriving motifs unique for each RBP after correcting for the features that are in common to different RBPs^{12,35,100,131}.

Another recently employed approach to assess the validity of *de novo* motifs is through the analysis of sites overlapping with heterozygous single-nucleotide polymorphisms (SNPs), where an imbalance of CLIP cDNAs mapping to the two alleles indicates that sequence variation affects the crosslinking efficiency³⁵. Such an allelic imbalance in a binding motif can indicate that it contributes to affinity of the studied RBP to the site, but it can also have alternative causes. First, if CLIP data contain an extrinsic background of co-IPed RBPs, allelic imbalance is equally expected at motifs bound by any of these other RBPs. Second, allelic imbalance can result from the technical biases of CLIP listed above, especially the nucleotide preferences of crosslinking.

Finally, as discussed earlier, enrichment of CLIP peaks can be assessed around regulated and unregulated RNA elements (i.e. RNA map) to inform on the 'functional specificity' of data, which can yield comparative specificity assessment for multiple CLIP data of a specific RBP¹³². As multiple CLIP datasets are becoming available for additional RBPs, their analysis with orthogonal data will be a valuable tool to gain a more comprehensive comparative estimate of functional data specificity. Ultimately, experiments to support the functionality of a binding site can be designed by perturbing the site, such as via mutations of cis-acting elements in minigene reporters or CRISPR-mediated mutations of the endogenous gene.

[H2] Assessing the sensitivity of CLIP

Sensitivity of CLIP refers to its capacity to comprehensively identify the relevant RNA sites bound by the studied RBP. Such sensitivity depends on the complexity of cDNA library, i.e., the number of unique cDNAs that are sequenced. The cDNA complexity has increased by orders of magnitude with the adaptation of high-throughput sequencing, and by increased efficiency of the cDNA library preparation steps². Yet, the capacity to prepare a library of high complexity depends on the characteristics of the RBP, especially its abundance and UV crosslinking efficiency. Moreover, sensitivity of cDNA libraries with comparable complexity can vary in dependence on specificity, because increased external background will decrease the proportion of signal for the RBP-of-interest. For example, CLIP libraries for PTBP1 that had similar cDNA complexity resulted in different numbers of identified binding peaks³⁷ and different capacity to identify binding sites around regulated exons¹³², which we refer to as ‘functional sensitivity’. Moreover, the choice of the peak calling method strongly affected the functional sensitivity of the same PTBP1 CLIP data¹⁴. This highlights the need for combined analysis of data specificity and sensitivity when making conclusions in regards to the pros and cons of the experimental variants of CLIP, and of the various computational approaches to data analysis.

[H1] Applications

[H2] Cell culture models

CLIP experiments have been carried out using various model organisms, including mammalian cell culture⁴², yeast³⁹, mice¹¹, flies¹³³, worms^{20,134} and plants^{21,22} (Fig. 4). Below, we discuss applications of CLIP techniques in selected systems with distinctive considerations, and pros and cons for the applications. Due to several practical reasons, cultured cells (transformed cell lines, primary cells, and stem cells) have been the most widely used with more than 2,500 different datasets deposited on the Gene Expression Omnibus at the time of writing. First, only ~7% of RBPs are expressed either in a tissue-specific manner or show strong tissue-specific expression bias, mainly in the germline and to a lesser extent neuronal tissues^{135,136}, while the rest tend to be expressed across most cell types¹³⁷. Therefore, cultured cells endogenously express many RBPs, which allows CLIP to be carried out with antibodies against the endogenous RBP and also to capture biologically relevant RNAs, with the caveat that some RBP targets may be absent in a culture model. Second, cultured cells are easily genetically tractable, allowing for epitope-tagging of RBPs for stringent purification, introduction of transgenically expressed cell-type-specific RBPs, and/or adding a clinically or functionally important mutation that could be lethal in an animal model. Third, cell culture allows for multiple RBPs to be studied in a comparative manner in the context of the same transcriptome. The same principles apply to single-cell organisms, such as yeast, which is genetically tractable and easy to work with, but may nevertheless be difficult for use with CLIP experiments due to its lower crosslinking efficiency³⁹. Finally, with cultured cells, material is typically not limiting. Nevertheless, although the use of cultured cells provided valuable insights into mechanisms of posttranscriptional regulation even of ectopically expressed RBPs¹³⁸, certain key bound transcripts and interacting proteins may be expressed in a cell-type specific

manner themselves. Furthermore, the binding repertoire of RBPs regulating biological processes such as developmental transitions or circadian timekeeping may also be best studied in the organismal context, as described in the next section.

[H2] Model organisms

CLIP/HITS-CLIP^{10,11}, iCLIP¹⁰⁰, PAR-CLIP^{20,139} and eCLIP¹⁴⁰ have all been successfully used with mouse, fly, and worm models. These studies provided useful insights into the roles of RBPs in various aspects of mRNA biogenesis and regulation during neuronal development and function¹³⁵, as well as specialised functions such as transposon silencing in human and mouse brain¹⁴¹, and the piRNA pathway in mouse testes or fly embryos^{142–144}. Animal models present unique challenges for the application of CLIP techniques. First, most tissues require mechanical dissociation of fresh or frozen tissue prior to UV crosslinking^{10,88}. In the case of PAR-CLIP, modified nucleotides must be delivered to the cells of interest prior to crosslinking. This can be typically accomplished by injection, but also by use of transgenic animals expressing uracil phosphoribosyltransferase in a cell-type specific manner to allow the conversion of thiouracil into thiouridine (TU-tagging)¹⁴⁵. Second, lethal mutations in RBPs can only be studied if introduced in a conditional manner. Finally, if a specific antibody for immunoprecipitation of the RBP is not available, expression of an epitope-tagged version of the RBP in a transgenic animal is required, which typically takes more effort compared to cell cultures. Nevertheless, by epitope-tagging the RBP of interest in specialized cell types¹⁴⁶ analogous to TRIBE¹⁵. This approach, employed by cTag-CLIP, revealed the interactome of Nova2, Pabpc1, or Fmrp in various cell types, including neuronal subsets of mouse brain^{147–149}.

[H2] Plants

Investigating RNP composition in higher plants is made difficult by several technical challenges. In contrast to mammalian cell cultures, plant cell cultures cannot be cultivated in monolayers and are of rather limited use for CLIP techniques. Instead, CLIP experiments have been performed in transgenic *Arabidopsis* plants expressing epitope-tagged RBPs^{21,22}. Despite the presence of UV-absorbing pigments and secondary metabolites such as chlorophyll and flavonoids, UVC-based crosslinking was successfully applied to whole plants^{21,22}. Another obstacle in plants is the rigid cell wall that requires mechanical force and harsh denaturing conditions for efficient cell lysis¹⁵⁰. Moreover, the large amounts of endogenous RNases present in the plant vacuole require the use of RNase inhibitors to prevent extensive RNA degradation during extract preparation as reported for pancreatic tissue. To ensure a controlled RNase treatment to fragment RNA, RNase treatment was performed after immunoprecipitation of the RNA-protein complexes, rather than in the lysate²².

Genome-wide binding data from HITS-CLIP have been obtained in plants for HLP1, a protein with similarity to mammalian HNRNPA/B²¹. In *hlp1* mutant plants, a shift from proximal to distal polyadenylation sites was observed for more than 2000 transcripts. As HLP1 binds to about a fifth of these aberrantly polyadenylated transcripts close to the polyadenylation site *in vivo*, HLP1 has been implicated in regulating alternative polyadenylation of these transcripts. In particular,

aberrant polyadenylation of transcripts involved in flowering time control may explain the delayed transition to flowering in the *hlp1* mutant²¹.

The first plant iCLIP study was performed for the hnRNP-like *Arabidopsis thaliana* glycine-rich RNA-binding protein 7 (*AtGRP7*)²². Among the *AtGRP7* binding partners were transcripts that are expressed specifically in inner cell layers of the leaf, demonstrating that UV light penetrates deep into the tissue. Overall, *AtGRP7* binds to U/C rich motifs mainly in the 3' untranslated regions of its targets. Cross-referencing RNA-seq data of mutants and overexpression lines revealed that *AtGRP7* predominantly down-regulates its binding partners. In particular, it dampens the peak expression of circadian clock regulated transcripts, in line with the function of *AtGRP7* as a slave oscillator transducing timing information from the circadian clock to rhythmic transcripts within the cell¹⁵¹.

Many new protein candidates to be studied by CLIP have emerged from proteomic studies that identify proteins that UV crosslink to polyadenylated RNAs in multiple *Arabidopsis* tissues. These studies have been performed in etiolated seedlings to increase the efficiency of UV crosslinking, as in higher plants chlorophyll biosynthesis is strictly dependent on light¹⁵², as well as in leaf protoplasts, cells without a cell wall¹⁵³, cell suspension cultures, and leaves of adult plants^{154,155}. These studies identified over 1100 candidate RBPs in total, but only few were found by all studies^{155,156}. This may partly be attributed to the differing developmental stages and tissues investigated and partly to the different protocols and levels of stringency used. As in non-plant species¹⁵⁷, a recurrent theme of these studies was that many proteins without known RNA-binding domains or without a link to RNA biology are identified¹⁵²⁻¹⁵⁵. Among those were photosynthesis-related proteins and plant photoreceptors that don't yet have any known role in RNA-based regulation, and therefore it is imperative to validate their RNA-binding activity by other methods such as CLIP¹⁵⁶.

[H2] Development and disease

RBPs play a myriad of important roles in development and diseases^{1,137}. CLIP has been valuable in unravelling the mechanisms behind these roles in specific biological contexts, as it can identify the endogenous protein-RNA interactions within unmodified cells and tissues. The first applications of CLIP were to unravel the roles of tissue-specific RBPs that regulate alternative splicing, such as Nova proteins in the brain. The high specificity of CLIP was essential to define the binding sites in low-abundant RNAs such as introns, which led to an unexpected finding that splicing regulators can have many thousands of high-affinity binding sites in introns^{10,11}. Sites located close to alternative exons coordinate splicing in a highly position-dependent manner that can be described by an RNA map^{11,124}. Moreover, most binding sites locate far from exons, and such sites often repress splicing of cryptic exons, such as those emerging from transposable elements^{126,141}, or recruit splicing factors to “decoy” sites that repress splicing of a nearby exon through competition with “bona fide” splice sites¹⁵⁸. CLIP can be used also to study RBPs that are parts of large RNPs, such as the core spliceosomal component PRPF8, which was used to interrogate how a phenomenon called ‘recursive splicing’ is regulated by the exon junction complex, with particular importance for appropriate splicing regulation in the brain¹⁵⁹.

CLIP has also been used to study a broad range of RBPs with roles in the regulation of RNA transport, stability and translation. For example, HITS-CLIP study of Fragile X mental retardation protein (FMRP) revealed its binding to a subset of transcripts across their entire coding length, which was suggested to result from its dual interactions with the ribosome and the mRNA that could be important for its regulation of local translation at the synapse⁸⁸. Finally, CLIP can be performed from postmortem human tissues, which can be used to interrogate pathology-related changes in protein-RNA interactions. For example, study of brain tissue from patients with pathological aggregates of TDP-43, an RBP implicated in multiple neurodegenerative diseases, demonstrated its increased binding to a non-coding RNA NEAT1¹⁶⁰. NEAT1 assembles multiple RBPs, including TDP-43, into a **biomolecular condensate** called 'paraspeckles'¹⁶¹. Interestingly, TDP-43 in turn regulates the 3' end processing of NEAT1 mRNA, which leads to cross-regulation between NEAT1 and TDP-43 that was shown to contribute to the exit from pluripotency in mouse embryonic stem cells¹⁶². Such cross-regulation between RNAs and RBPs is likely a common phenomenon, as it is becoming clear that just as RBPs regulate their RNA partners, RNAs can also act as regulators of their bound RBPs, as was shown for the case of Vault RNA-dependent regulation of proteins involved in autophagy¹⁶³.

Finally, CLIP is increasingly used in pathogen research, such as studies of RNA interaction profiles of bacterial RBPs¹⁶⁴, and studies of how viral infection remodels the RNA interactome of host and viral RBPs. For instance, study of miRNAs crosslinked to Ago indicated that miRNAs encoded by Kaposi's sarcoma-associated herpesvirus (KSHV) may function by competing with host miRNAs for Ago¹⁶⁵, and a later study using cross-linking, ligation, and sequencing of hybrids (CLASH) additionally identified over 1,400 cellular mRNAs that are targeted and might be regulated by the KSHV miRNAs¹⁶⁶. Moreover, study of HIV-1 Gag uncovered dramatic changes in its RNA-binding properties that occur during virion genesis and contribute to viral packaging¹⁶⁷, study of APOBEC3 proteins showed how their RNA binding ensures their effective encapsidation into HIV-1 virus as part of the host's defense¹⁶⁸, and study of Poly-C binding protein 2 (PCBP2) provided support for its roles in hepatitis C virus-infected cells¹⁶⁹. These studies also provided computational solutions for parallel analysis of human and user-definable nonhuman transcriptomes. Most recently, CLIP has been used to identify human RNAs that are bound by the proteins encoded by the SARS-CoV-2 genome, such as non-structural proteins (NSP)¹⁷⁰ or Nucleocapsid protein¹⁷¹, which helped to show how these RBPs alter the gene expression pathways to suppress host defenses. Conversely, CLIP of host RBPs was used to show their binding to SARS-CoV-2 RNAs, which contributes to host defense strategies⁸³. Much more work remains to be done with CLIP and complementary approaches to understand the complex cross-regulation between the RBPs and RNAs of pathogens and their hosts modulates the pathogenicity.

[H2] Protein- and RNA-centric approaches yield complementary insights

When used in combination, protein- and RNA-centric approaches can lead to particularly transformative insights into the mechanisms of RNP assembly and function. One example is the study of NORAD lncRNA, where RAP-MS was used first to identify its interaction with RBMX and several other proteins, the RNA binding sites of which were then mapped with CLIP, which showed how NORAD assembles an RNP that links proteins involved in DNA replication or

repair⁷⁹. Another example is the study of XIST lncRNA, where its bound RBPs were first identified through RNA-centric methods by several studies^{78,80}, and later studied by CLIP to show how XIST seeds a heteromeric RNP condensate that is required for heritable gene silencing¹⁷². Most recently, the host RBPs bound to SARS-CoV-2 RNAs were first identified by RAP-MS, and then studied further with CLIP to map their direct interactions with the SARS-CoV-2 RNA in infected human cells⁸³. These studies show that complementary data from these these approaches open an opportunity to build computational models that position each RBP at its bound cis-acting RNA elements along an RNA, and thus understand how protein-RNA and protein-protein interactions act combinatorially to drive the assembly and remodeling of RNPs on full RNAs.

A question that is particularly pertinent to the field of RNA localization is how RNPs form dynamic condensates, often referred to as 'RNP granules', that regulate RNA transport and local translation in response to signalling¹⁷³. Understanding RNP assembly and dynamics in these contexts is particularly challenging, as it is mediated both by direct protein-RNA interactions and protein-protein interactions, mediated both by structure domains and **IDRs**, which often coordinate condensation of proteins into the granule. Important questions to be solved are how the cis-regulatory sequence and structural elements on the RNA mediate the assembly of the full RNP in order to coordinate its selective transport, and how post-translational modifications of the **IDRs** mediate RNP remodeling in response to specific signals¹. Performing both CLIP and RNA-centric methods under dynamic states will be essential to resolve how specific RBPs are released, rebound or repositioned on RNAs in response to stimuli. Comparisons between localised mRNAs will tell whether they share a subset of core RBPs, and how these RBPs mediate mRNA recruitment to transport machineries and the translational apparatus. Finally, studies of RNA-RNAs in addition to protein-RNA and protein-protein contacts will be needed to fully disentangle the principles of RNP assembly¹⁷³.

Such understanding of RNP remodelling is of paramount importance as it underlies many aspects of cellular remodelling, including cellular polarity and movement, axon guidance, synaptic plasticity and memory formation. Moreover, deregulated RNP dynamics can lead to formation of aberrant condensates and aggregates in many neurologic diseases, such as amyotrophic lateral sclerosis and fragile X syndrome¹⁷⁴. Combining the RNA and protein centric methods in models of these diseases will be essential to understand how changes in RNP assembly contribute to the disease processes by affecting specific RNAs on their pathway of biogenesis, transport, translation and degradation.

[H1] Reproducibility and data deposition

[H2] Reproducibility and comparative value of CLIP data

To understand which features of the RNA drive the binding of an RBP in physiological conditions, how these interactions evolve and are remodeled, and how crosstalk of RBPs takes place on individual RNAs, comparisons of multiple datasets produced across conditions, cell types, species, and RBPs are necessary. Although data have been obtained by multiple CLIP variants for many RBPs, and in some cases also by complementary methods such as RIP and TRIBE,

only few studies have examined such data in complementary ways^{132,175,176}. For comparisons to be most informative, it is essential to distinguish the technical from biological sources of variation between CLIP experiments. Technical variation can have four primary causes: 1) differences in the conditions of crosslinking, stringencies of lysis and washing during the purification and quality control of the purified protein-RNA complexes (Box 1), and cDNA library preparation between protocols (Figure 2). 2) variations in the way the RBP is purified, such as use of different antibodies for IP of endogenous RBP or affinity purification of tagged RBP. 3) unintentional variations in the way the method has been implemented, such as subtle variations in the density of cultured cells, UV crosslinking and RNase fragmentation conditions. 4) stochastic variation in the capture of RNAs and identified binding sites between samples, especially when the RNAs have low expression and the efficiency of UV crosslinking is low. It is thus advisable that comparative analyses aiming to identify biologically-relevant changes in the endogenous RNA binding properties of RBPs are designed in such a way that data accuracy and the technical sources of variation can be addressed.

The most valuable indicator of CLIP data accuracy is its cross-validation with orthogonal information, such as the motif enrichment in peaks defined by various CLIP datasets^{37,177}, or the position-dependent enrichment of peaks around the regulated RNA elements, as shown by the RNA maps¹³². Binding motifs have been identified by *in vitro* methods for hundreds of RBPs, and regulated elements can be defined from increasingly available RNAseq data, obtained upon RBP knockout or knockdown^{108,178,179}. So far, integration of these data with CLIP has been qualitative; for example, enriched motifs have been identified from the large resources of PAR-CLIP and eCLIP data^{35,108,109}, and these are often similar to the motifs determined for the respective RBPs in *in vitro* studies. However, comparisons of motif enrichments in CLIP peaks obtained for the same RBP in various datasets have not been done. Thus, it will be important to develop approaches that can use orthogonal information to evaluate CLIP data accuracy on a large scale.

Although a necessary indicator of data quality, reproducibility across replicate CLIP experiments is less informative than the cross-validation with orthogonal data. This is because cross-contamination from a co-IPed RBP can be reproducible, as can technical biases of crosslinking, nuclease digestion and ligation. These reproducible biases can in fact distort the data, potentially boosting the significance of otherwise low-occupancy sites. Therefore, as more and more CLIP data for the same RBP is produced across labs and across variant methods, it will be essential to perform comparative benchmarking of these data and reconstruct comprehensive and accurate sets of binding sites. For instance, while the peak identification methods mentioned above can yield tens of thousands of peaks for some well characterized RBPs, it is informative to assess peak reproducibility for replicate samples within a lab, across labs and across CLIP variants⁴², and for samples that assess biological variation, such as samples obtained from different animals¹¹. A concern remains that reproducible peaks, just as peaks in general, are more likely to be located in relatively abundant RNAs. Peaks in less expressed RNAs may be less reproducible and therefore missed in the final results, although some false negative sites can be recovered with computational models trained on the CLIP data¹¹⁴. With the rapidly increasing number of available datasets and computational approaches, it will be possible to perform more benchmarking comparisons and thus gain insights into the experimental and computational steps that aid the specificity and sensitivity, and thus the reproducibility of data.

[H2] Data resources

Resources that provide CLIP data across studies are essential for compiling the RBP interaction data and enabling comparisons across data sets. The raw sequencing data are made available upon publication from general public repositories such as the Sequence Read Archive¹⁸⁰ or the European Nucleotide Archive (<http://www.ebi.ac.uk/ena>), which enforce the tracking of appropriate metadata. Alignments of reads are provided as .bam files that can be visualized with tools such as the Integrative Genomics Viewer (IGV, <http://software.broadinstitute.org/software/igv/>)¹⁸¹. Specialized databases such as doRINA (<https://dorina.mdc-berlin.de/>)¹⁸², ENCORI (previous known as starBase; <http://starbase.sysu.edu.cn/>)¹⁸³ and POSTAR2 (<http://lulab.life.tsinghua.edu.cn/postar>)¹⁸⁴ enable the exploration of processed CLIP peaks, along with additional information such as annotation and gene expression. doRINA also allows upload of user-provided binding site data for visualization in the context of the information held in the database. A tool called SEQing has also been developed to visualize Arabidopsis iCLIP binding sites¹⁸⁵, again in the context of gene expression data. Nevertheless, CLIP data integration can be challenging, as many CLIP variants use tailored design of barcodes and UMIs, which require customised analysis. Databases of RBP binding motifs have started to emerge as well. CISBP-RNA¹⁷⁸ summarizes data on *in vitro* RBP-RNA interactions, while ATTRACT contains curated data from a variety of sources¹⁸⁶, albeit without resolving discrepancies in the motifs that are inferred for the same protein from different types of experiments. Overall, it will be important for the RNA community that these resources remain well maintained and further integrated.

[H1] Limitations and optimizations

[H2] RBP-specific challenges in CLIP data analysis

RBPs can differ in many aspects that can influence data analysis and interpretation. Perhaps the clearest indicator of binding characteristics is the type of motif that is recognised. Some RBPs, such as Pumilo-family proteins, primarily bind relatively long, well-defined motifs, which overlap with sharp crosslinking peaks¹², while others recognise short (often only 2-4 nucleotides long) and degenerate motifs, which often occur in multivalent clusters to drive *in vivo* binding¹⁸⁷. Binding peaks for such RBPs can be dispersed over long clusters of motifs, as exemplified by RBPs binding to LINE-derived RNA elements that contain enriched motifs that are dispersed over hundreds of nucleotides¹⁸⁸. An even more extreme example are RBPs with limited sequence preferences, such as FUS or SUZ12, which show broad distribution of crosslinking across nascent transcripts^{100,189}. In such cases, technical biases such as uridine crosslinking preferences are more likely to contribute to peaks of crosslinking, and therefore such peaks need to be considered with caution. Thus, strategies to assign the binding sites from CLIP data ideally need to be adjusted to the binding characteristics of each RBP, but such approaches are still to be developed.

Many RBPs interact with large RNPs, and their RNA interactions are often dominated by one or a few abundant ncRNAs, such as snRNA for spliceosome and rRNA for the ribosome. Nevertheless, even such RBPs can have additional moonlighting functions outside of their primary

RNP, as has been seen for ribosomal proteins¹⁹⁰. Thus, one needs to be cautious not to automatically assign such secondary binding to background. Moreover, even though the standard IP conditions of CLIP are quite stringent, stable RNPs may not fully disassemble, and in such cases, the RBP partners generate considerable 'extrinsic background' to the resulting data. Such RBPs tend to bind to similar RNAs and perform shared functions, so in some cases it can be informative to design CLIP such that it simultaneously profiles the RNA interactome of many RBPs that are associated with specific stable RNPs; for example, Sm proteins are IPed in 'spliceosome iCLIP' to yield the RNA interactome of multiple RBPs associated with various snRNPs, thus revealing not just the direct binding of Sm proteins on snRNAs, but also the branch points and the sites of spliceosomal assembly on pre-mRNAs¹⁹¹.

[H2] Challenges of RNA-centric methods

[H3] RNA affinity capture methods

The development of RNA-centric methods that are based on RNA affinity capture has greatly expanded our knowledge on RBPs bound to specific RNAs. However, an inherent limitation of these methods is the potential loss of transient and compartment-specific interactions and possibility of co-purifying post-lysis false-positive interactions⁷⁶. Moreover, the addition of aptamer can change the secondary structure of the RNA and the corresponding protein binding pattern¹⁹². To address these issues, the post-lysis integrity of the RNP can be improved with formaldehyde or UV crosslinking, followed by either biotin-labelled antisense oligo RNA affinity purification (RAP)¹⁹³, peptide nucleic acid (PNA)-assisted affinity purification^{194,195}, or both biotin and then 2'-O-methylated antisense RNA mediated tandem RNA isolation (TRIP)¹⁹⁶.

[H3] Proximity-based methods

Proximity-based methods are highly complementary, as they can overcome the limitations listed for the affinity-based methods. However, they can contain limitations of their own, such as the need for sufficient available lysine or other electron-rich amino acids on the protein surface for efficient biotinylation. Moreover, the free enzyme that is in the process of searching for the targeted RNA can biotinylate nonspecific proteins. Such background biotinylation can to some extent be corrected when analysing the data in a cell-specific or tissue-specific way, and general contaminants can be diminished from the dataset by referring to the CRAPome database¹⁹⁷. Another issue could be a limited detection range (10-20nm). The proximity enzymes that are currently used differ mainly in their labeling range and substrates, and can be broadly grouped in peroxidases and biotin ligases (Table 1)¹⁹⁸. Biotin ligases convert biotin and ATP into biotinoyl-5'-adenylate (bioAMP) which diffuses around the activation site and covalently bonds to the nearby lysine residues¹⁹⁹.

The efficiency of proximity ligases depends on the redox environments and proximal nucleophile concentrations, which might explain why BioID and TurboID were found to be effective in every compartment when tagged with a nuclear localization sequence (NLS), mitochondrial targeting sequence (MTS) or ER-targeting sequences, whereas miniTurboID was more effective in open cytosolic environment rather than membrane-enclosed organelles²⁰⁰. miniTurboID can be used at

a lower temperature (20°C to 37°C) compared to BioID (37°C) and BioID2 (optimal is 50°C)^{200,201}. However, it is concerning that constitutive expression of TurboID in the absence of exogenous biotin leads to decreased size and viability in *Drosophila melanogaster*²⁰⁰ and even long incubation time (more than 6 hrs) or use of excess biotin (50 µM) may result in nonspecific biotinylation in the cell²⁰⁰. Deletion of the N-terminal region was found to decrease the stability of miniTurboID in *Caenorhabditis elegans*²⁰⁰. Recently, with the help of enzyme reconstruction algorithms and residue replacements on optimized biotin ligases, a new BirA enzyme, AirID (ancestral BirA for proximity-dependent biotin identification), has been developed²⁰². AirIDw was found to be less toxic compared to TurboID, and tamavidin2-Rev beads were used instead of streptavidin beads to release proteins efficiently in presence of free biotin.

[H2] Challenges and opportunities in identifying the RNA binding sites

To fully understand RNP assembly, it is important to define with high resolution the sites on RNAs that recruit specific RBPs, as well as the sites on RBPs that bind to RNAs. The field is still learning how to extract RNA interaction parameters from CLIP data as well as how to interpret the potential functions of these interactions. Defining the crosslinking peaks of high occupancy, as described earlier, is an important step, but such peaks should not be directly equated to functionally-relevant binding sites. For instance, many RBPs bind to broad regions that contain multiple occurrences of a motif, which rarely fully overlap with the peaks of high crosslinking, and it is the total residence time of the protein in such broader regions of the RNA that determines the functional outcome^{36,91,187}. Recently, femtosecond UV laser crosslinking followed by CLIP (KIN-CLIP) was shown capable of characterising the *in vivo* binding kinetics at individual sites and the functionality of binding site clusters⁹¹.

To come closer to the full binding site assignment, it is necessary to combine CLIP data with analysis of RNA sequences and structural motifs¹¹⁴. Further indication of the functional relevance of binding sites can be obtained by assessing their evolutionary conservation. However, many RNA sequences are not strongly constrained in evolution - for example, even though the length and arrangement of lncRNAs and introns is under considerable evolutionary constraint, most of their sequence shows weak conservation across species, and rapid accumulation of repetitive elements, indicating weak functional constraint²⁰³. Nevertheless, even largely neutrally evolving sequences can contain high-affinity binding sites that are under some selection, as demonstrated by the observation that many RBPs contain high-occupancy intronic binding sites within repetitive elements, where they often act to repress inclusion of cryptic exons²⁰⁴.

Finally, even the most optimal CLIP-defined binding peaks (i.e., highly specific data, with optimal background analysis, etc), or sites that are computationally predicted based on RNA sequence/structure features don't lead only to sites that are functional, i.e., have a regulatory impact on the RNA. To discern the sites that are likely functionally relevant, it is valuable to integrate CLIP with orthogonal transcriptomic data from RBP perturbation experiments, which is particularly informative if it leads to position-dependent regulatory principles (i.e. RNA maps)^{12,124,205,206}. Such integration identifies CLIP peaks that likely mediate the regulation of specific elements (alternative exons, etc), while it also distinguishes the RNAs detected by RNA-seq that are likely directly regulated by the RBP (i.e., they contain CLIP peaks at expected positions) from those that change upon RBP perturbation due to off-target effects, feedback loops via other RBPs, or other types of cellular compensation. Thus, the sensitivity and specificity of patterns observed by an RNA map can be used as a valuable measure of the quality CLIP and RNA-seq data that are being integrated¹⁴. Moreover, to understand binding sites that drive functions beyond RNA processing, additional types of orthogonal data sets can be integrated with CLIP, as has been exemplified by studies of RNA stability¹⁷⁶, translation^{88,149} and localization^{207,208}.

Finally, sequencing-based approaches can be integrated by insights from proteomics. In particular, the sites on RBPs that bind to specific RNA sites can be simultaneously defined through a combination of UV crosslinking, high-resolution mass spectrometry and a dedicated computational workflow to identify both the crosslinked peptides and RNA oligonucleotides - an approach that can be RNA-centric, or applied to the whole RBPome³⁴. Recently, several additional approaches were developed for high-throughput mapping of crosslinked peptides or amino acids within RBPs¹. With the ever-increasing capacity of these complementary methods to monitor specific functions of RBPs, integrative approaches are bound to become increasingly fruitful.

[H1] Outlook

In the decade and a half since the first CLIP studies, the method has undergone much development and the nature of the obtained data is much better understood. It is clear that there is no one-size-fits-all guideline for the design and analysis of CLIP experiments. It will be important to learn more about the pros and cons of the various experimental and computational approaches of CLIP variants through comparisons of the increasing amounts of available data. In the meantime, it is important to be aware of the steps that can be taken for quality control and optimization in order to tailor the experimental and computational steps according to the RBP that is studied, the input material, and the type of questions that are asked.

We expect many new applications of CLIP to be developed in coming years, with increasing integration of CLIP with data from methods based on enzymatic-tagging and RNA-centric approaches. These complementary methods have not yet been used in combination, but we hope that this Primer will encourage their integrative use. Cross-method comparisons will be valuable from the technical perspective, to better understand the advantages of each method, and correct for technical biases. Integration of data from CLIP that primarily detects direct protein-RNA

interactions with those that also detect RNA-proximal proteins will help to understand which proteins are recruited to RNAs primarily via direct recognition of specific RNA elements versus protein-protein interactions with other RBPs. Another valuable application will be to study specific RBPs in subcellular compartments with complementary methods to provide insights into the assembly properties of RBPs at organelles or **bimolecular condensates**²⁰⁹. For example, such methods could be applied to chloroplasts, which are unique to plants and rely heavily on post-transcriptional mechanisms for controlling the expression of their genome²¹⁰.

Important questions in RNP remodelling and combinatorial assembly can be unravelled when CLIP and complementary methods are used under comparative scenarios. For example, comparative CLIP of one RBP from cells lacking another RBP can reveal how individual RBPs compete for binding to overlapping sites¹²⁶, or how larger RNPs compete, such as the role of exon-junction complex in blocking access of splicing machinery to regions around exon-exon junctions in spliced RNAs¹⁵⁹. The competitive and combinatorial assembly principles can be further unravelled by "*in vitro* CLIP" experiments, in which recombinant RBPs with varying concentrations are incubated with long transcripts, followed by modelling and machine learning²¹¹. Moreover, CLIP can be performed with purified RNPs in specific states, such as for example to define helicase-RNA contacts in specific spliceosomal states by "purified spliceosome iCLIP" (psiCLIP)²¹². Finally, a long-term challenge will be to understand how RNA regulatory networks are remodelled on various timescales: from the timescales of cellular signal-response, development and aging, to mutation-driven changes in cancer or other diseases, and finally, the timescale of organismal evolution. Such questions have started to be addressed by CLIP studies across species or in response to disease mutations^{213,214}. A particularly important question will be to understand how variations in the IDRs of RBPs, which tend to evolve faster than structured domains, and are hotspots of disease-causing mutations and post-translational modifications¹, might affect the regulation of specific types of RNA binding sites.

Two emerging applications of transcriptomic techniques that we have not covered in this article are mapping of RNA structure and RNA modifications genome-wide, as the topic has been comprehensively covered elsewhere^{17,215-217}. Integration of protein-RNA interactions with information of RNA structure and RNA-RNA spatial interactions is already opening new doors to understanding the roles of RNA molecules in organizing RNP assembly^{17,50,217-219}. Recently, an RNA pull-down method was used to identify proteins bound to 186 RNA structures conserved across yeast species²²⁰. This approach enables a streamlined study of dozens of short RNA fragments to uncover RBPs that tend to bind similar RNA structures or other types of similar RNA motifs from a group of RNAs. It thus offers a valuable complement to the more global RNA interactome capture on the one hand, and the RNA-centric approaches on the other.

Over 100 RNA modifications have been described, most of which affect the assembly of protein-RNA complexes, therefore their transcriptomic understanding is essential orthogonal information to CLIP and other methods that interrogate protein-RNA interactions. Interestingly, mutations of certain methyltransferases can stabilize the covalently linked protein-RNA catalytic intermediates, thus enabling CLIP to be performed without the need for UV crosslinking, as has been done for m5C-miCLIP²²¹. Most methods to date have been developed for transcriptomic studies of m6A, the type of modification that is most common in mRNAs, and these include variants of CLIP, such

as m6A-miCLIP, which employ antibodies that recognise m6A-containing RNA²²². The success of such approaches critically depends on the quality of the antibodies recognizing the modification²²³. Therefore, similar to studies of protein-RNA interactions, integration of data from complementary methods will be valuable to gain a full picture of RNA modifications and their roles in RNP assembly^{216,224}.

From the computational angle, we expect the methods for site and motif identification to reach maturity, leading to high-quality databases of *in vivo* RBP binding motifs. As most of the computational methods work with uniquely mapping reads, improvements are foreseen in the quantification of sites located in repeat elements as well as at exon-exon boundaries or in splicing and polyadenylation isoforms. Ultimately, we can start wondering, what could we do if we had information on all the protein-RNA interaction sites? For example, we could construct a whole-cell model that includes them to comprehensively predict RNA fates, and their roles in cellular changes during development and disease? The path taken towards such an ultimate aim will require integration of protein-centric and RNA-centric methods to gain understanding not just of the full RNP assembled on each transcript, but also the spatial RNP dynamics as each transcript moves through the cell, and the temporal RNP dynamics as post-translational protein modifications and RNA methylation modulate the RNP. As such, RNPs will surely continue to teach us about the highly interconnected and ever-changing world of living cells.

Acknowledgements

We thank Flora Lee, Anob Chakrabarti and Reem Abouward for suggestions on the manuscript. This work was supported by the German Research Foundation (DFG) through grants STA653/13-1 and STA653/14-1 to D.S., KO5364/1-1 to T.K., the Intramural Research Program of the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health to M.H. and J. Ma., by the European Union's Horizon 2020 research and innovation programme (835300-RNPdynamics) to J.U. and J.Mu., Swiss National Science Foundation (310030_189063) to M.Z and by the Biozentrum Basel International Ph.D. Program Fellowships for Excellence to M.K. The Francis Crick Institute receives its core funding from Cancer Research UK (FC001110), the UK Medical Research Council (FC001110), and the Wellcome Trust (FC001110).

Competing interests

The authors declare no competing interests

Highlighted references:

Benhalevy, D., Anastasakis, D.G., and Hafner, M. (2018). Proximity-CLIP provides a snapshot of protein-occupied RNA elements in subcellular compartments. *Nat. Methods* 15, 1074–1082.

Subcellular compartment-specific proximity labelling is combined with CLIP to monitor RNA-protein interactions at specific locations in the cell.

Briese, M., Haberman, N., Sibley, C.R., Faraway, R., Elser, A.S., Chakrabarti, A.M., Wang, Z., König, J., Perera, D., Wickramasinghe, V.O., et al. (2019). A systems view of spliceosomal assembly and branchpoints with iCLIP. *Nat. Struct. Mol. Biol.* 26, 930–940.

Adaptation of CLIP to simultaneously profile the RNA interactome of many RBPs that are associated with stable RNPs; in this case, this determined the RNA interaction profiles of spliceosomal proteins.

Chakrabarti, A.M., Haberman, N., Praznik, A., Luscombe, N.M., and Ule, J. (2018b). Data Science Issues in Studying Protein–RNA Interactions with CLIP Technologies. *Annu. Rev. Biomed. Data Sci.* 1, 235–261.

Analysis of RNA splicing maps is presented as a way to assess the sensitivity and specificity of CLIP data, along with a thorough review of computational methods.

Feng, H. et al. Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites. *Mol. Cell* 74, 1189–1204.e6 (2019).

De novo motif discovery is performed on >100 RBPs using eCLIP data by joint modeling of sequence specificity and crosslink sites, and evaluation of motifs by allele imbalance.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr, Jungkamp, A.-C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141, 129–141.

Development of PAR-CLIP that enables identifications of crosslink sites with analysis of transitions.

Hwang, H.-W., Saito, Y., Park, C.Y., Blachère, N.E., Tajima, Y., Fak, J.J., Zucker-Scharff, I., and Darnell, R.B. (2017). cTag-PAPERCLIP Reveals Alternative Polyadenylation Promotes Cell-Type Specific Protein Diversity and Shifts Araf Isoforms with Microglia Activation. *Neuron* 95, 1334–1349.e5.

Development of a knockin mouse in which a GFP-tagged RBP is conditionally expressed in selected cell populations, enabling cell-type specific CLIP; in this case, GFP-PABP is used to map the 3' ends of mRNAs in excitatory and inhibitory neurons, astrocytes and microglia.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods* 8, 559–564.

This study evaluates how differences in cross-linking and ribonuclease digestion affect the sites obtained with HITS-CLIP and PAR-CLIP, both marked by specific crosslinking-induced mutations.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* 17, 909–915.

Development of iCLIP that enables amplification of truncated cDNAs and identifications of crosslink sites with analysis of truncations.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456, 464–469.

This study introduced the use high-throughput sequencing for CLIP, and thereby validated the RNA map of splicing regulation by Nova proteins.

Maticzka, D., Lange, S.J., Costa, F., and Backofen, R. (2014). GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.* 15, R17.

This study presented the first computational framework for modelling sequence- and structure-binding preferences of RBPs from CLIP data.

McMahon, A.C., Rahman, R., Jin, H., Shen, J.L., Fieldsend, A., Luo, W., and Rosbash, M. (2016). TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. *Cell* 165, 742–753.

This study establishes a method to identify RNA binding sites of RBPs through fusion with ADAR catalytic domain and analysis of RNA editing.

Meyer, K., Köster, T., Nolte, C., Weinholdt, C., Lewinski, M., Grosse, I., and Staiger, D. (2017). Adaptation of iCLIP to plants determines the binding landscape of the clock-regulated RNA-binding protein AtGRP7. *Genome Biol.* 18, 204.

The first plant iCLIP study identifies RNA-binding partners of an hnRNP-like protein in the reference plant *Arabidopsis thaliana*.

Mukherjee N, Wessels HH, Lebedeva S, Sajek M, Ghanbari M, Garzia A, Munteanu A, Yusuf D, Farazi T, Hoell JI, Akat KM, Akalin A, Tuschl T, Ohler U. (2019) Deciphering human ribonucleoprotein regulatory networks. *Nucleic Acids Res.* Jan 25;47(2):570-581.

This study produced 114 PAR-CLIP experiments for 64 RBPs in the HEK cell line, and presents a comparative analysis of these RBPs.

Munschauer, M. et al. The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature* 561, 132–136 (2018).

RAP-MS and CLIP maps are used in a complementary fashion to map the assembly of NORAD lncRNA into an RNP that links proteins involved in DNA replication or repair.

Schneider C, Kudla G, Wlotzka W, Tuck A, Tollervey D. (2012) Transcriptome-wide analysis of exosome targets. *Mol Cell.* 2012 Nov 9;48(3):422-33.

Development of split-CRAC, where an RBP undergoes in vitro cleavage during affinity purification, which allows separate identification of RNA sites crosslinked to the N- and C-terminal regions of the RBP.

Sutandy, F.X.R., Ebersberger, S., Huang, L., Busch, A., Bach, M., Kang, H.-S., Fallmann, J., Maticzka, D., Backofen, R., Stadler, P.F., et al. (2018). In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. *Genome Res.* 28, 699–713.

Development of “in vitro iCLIP”, in which recombinant RBPs are incubated with long transcripts followed by modeling and machine learning, to study how protein-RNA interactions are determined by cis-acting sequences and modulated by trans-acting RBPs.

Ule, J., Jensen, K., Mele, A., and Darnell, R.B. (2005). CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* 37, 376–386.

A detailed description of the CLIP protocol that establishes the workflow and explains RNase optimisation, SDS-PAGE purification conditions and cDNA library preparation that are conceptually followed by most later variants.

Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., et al. (2020a). A large-scale binding and functional map of human RNA-binding proteins. *Nature* 583, 711–719.

This study produced eCLIP experiments for 103 RBPs from HepG2 and 120 RBPs from K562 cell line, each in duplicates and with SMIinput controls, which are available as part of the ENCODE resource (<https://www.encodeproject.org/>), and presents a comparative analysis of these RBPs.

Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N.M., and Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* 152, 453–466.

This study demonstrates the quantitative capacity of CLIP to compare binding of an RBP between conditions - in this case, to demonstrate the displacement of U2AF2 by hnRNPC at cryptic splice sites within intronic Alu elements.

Zarnegar, B.J., Flynn, R.A., Shen, Y., Do, B.T., Chang, H.Y., and Khavari, P.A. (2016). irCLIP platform for efficient characterization of protein-RNA interactions. *Nat. Methods* 13, 489–492.

This study presents a nonisotopic method for the detection of protein–RNA complexes by using an infrared labeled adapter, which simplifies their visualisation after SDS-PAGE separation.

References:

1. Gebauer, F., Schwarzl, T., Valcárcel, J. & Hentze, M. W. RNA-binding proteins in human genetic disease. *Nat. Rev. Genet.* 1–14 (2020).
2. Lee, F. C. Y. & Ule, J. Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Mol. Cell* **69**, 354–369 (2018).
3. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8**, 479–490 (2007).
4. Corley, M., Burns, M. C. & Yeo, G. W. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Mol. Cell* **78**, 9–29 (2020).
5. Masliah, G., Barraud, P. & Allain, F. H.-T. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cell. Mol. Life Sci.* **70**, 1875–1895

- (2013).
6. Lerner, M. R. & Steitz, J. A. Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 5495–5499 (1979).
 7. Tenenbaum, S. A., Carson, C. C., Lager, P. J. & Keene, J. D. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 14085–14090 (2000).
 8. Niranjankumari, S., Lasda, E., Brazas, R. & Garcia-Blanco, M. A. Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo. *Methods* **26**, 182–190 (2002).
 9. Köster, T., Haas, M. & Staiger, D. The RIPper case: identification of RNA-binding protein targets by RNA immunoprecipitation. *Methods Mol. Biol.* **1158**, 107–121 (2014).
 10. Ule, J. *et al.* CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**, 1212–1215 (2003).
 11. Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469 (2008).
 12. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).
 13. König, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17**, 909–915 (2010).
 14. Chakrabarti, A. M., Haberman, N., Praznik, A., Luscombe, N. M. & Ule, J. Data Science Issues in Studying Protein–RNA Interactions with CLIP Technologies. *Annu. Rev. Biomed. Data Sci.* **1**, 235–261 (2018).
 15. McMahon, A. C. *et al.* TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. *Cell* **165**, 742–753 (2016).
 16. Benhalevy, D., Anastasakis, D. G. & Hafner, M. Proximity-CLIP provides a snapshot of

- protein-occupied RNA elements in subcellular compartments. *Nat. Methods* **15**, 1074–1082 (2018).
17. Lin, C. & Miles, W. O. Beyond CLIP: advances and opportunities to measure RBP-RNA and RNA-RNA interactions. *Nucleic Acids Res.* **47**, 5490–5501 (2019).
 18. Ramanathan, M., Porter, D. F. & Khavari, P. A. Methods to study RNA–protein interactions. *Nat. Methods* **16**, 225–234 (2019).
 19. Ule, J., Jensen, K., Mele, A. & Darnell, R. B. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* **37**, 376–386 (2005).
 20. Jungkamp, A.-C. *et al.* In vivo and transcriptome-wide identification of RNA binding protein target sites. *Mol. Cell* **44**, 828–840 (2011).
 21. Zhang, Y. *et al.* Integrative genome-wide analysis reveals HLP1, a novel RNA-binding protein, regulates plant flowering by targeting alternative polyadenylation. *Cell Res.* **25**, 864–876 (2015).
 22. Meyer, K. *et al.* Adaptation of iCLIP to plants determines the binding landscape of the clock-regulated RNA-binding protein AtGRP7. *Genome Biol.* **18**, 204 (2017).
 23. Moore, M. J. *et al.* Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat. Protoc.* **9**, 263–293 (2014).
 24. Kishore, S. *et al.* A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods* **8**, 559–564 (2011).
 25. Friedersdorf, M. B. & Keene, J. D. Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biol.* **15**, R2 (2014).
 26. König, J., Zarnack, K., Luscombe, N. M. & Ule, J. Protein–RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet.* **13**, 77–83 (2012).
 27. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).

28. Patton, R. D. *et al.* Chemical crosslinking enhances RNA immunoprecipitation for efficient identification of binding sites of proteins that photo-crosslink poorly with RNA. *RNA* **26**, 1216–1233 (2020).
29. Terao, K. & Ogata, K. Proteins of small subunits of rat liver ribosomes that interact with poly(U). II. Cross-links between poly(U) and ribosomal proteins in 40 S subunits induced by UV irradiation. *J. Biochem.* **86**, 605–617 (1979).
30. Fiser, I., Scheit, K. H. & Kuechler, E. Poly(4-thiouridylic acid) as messenger RNA and its application for photoaffinity labelling of the ribosomal mRNA binding site. *Eur. J. Biochem.* **74**, 447–456 (1977).
31. Fiser, I., Scheit, K. H., Stöffler, G. & Kuechler, E. Identification of protein S 1 at the messenger RNA binding site of the Escherichia coli ribosome. *Biochem. Biophys. Res. Commun.* **60**, 1112–1118 (1974).
32. Wagenmakers, A. J., Reinders, R. J. & van Venrooij, W. J. Cross-linking of mRNA to proteins by irradiation of intact cells with ultraviolet light. *Eur. J. Biochem.* **112**, 323–330 (1980).
33. Miller, R. L. & Plagemann, P. G. Effect of ultraviolet light on mengovirus: formation of uracil dimers, instability and degradation of capsid, and covalent linkage of protein to viral RNA. *J. Virol.* **13**, 729–739 (1974).
34. Kramer, K. *et al.* Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat. Methods* **11**, 1064–1070 (2014).
35. Feng, H. *et al.* Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites. *Mol. Cell* **74**, 1189–1204.e6 (2019).
36. Sugimoto, Y. *et al.* Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol.* **13**, R67 (2012).
37. Haberman, N. *et al.* Insights into the design and interpretation of iCLIP experiments. *Genome Biol.* **18**, 7 (2017).

38. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**, 858–862 (2001).
39. 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. Natl. Acad. Sci. U. S. A.* **106**, 9613–9618 (2009).
40. Zhang, C. & Darnell, R. B. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.* **29**, 607–614 (2011).
41. Zarnegar, B. J. *et al.* irCLIP platform for efficient characterization of protein-RNA interactions. *Nat. Methods* **13**, 489–492 (2016).
42. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514 (2016).
43. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
44. Blondal, T. *et al.* Isolation and characterization of a thermostable RNA ligase 1 from a *Thermus scotoductus* bacteriophage TS2126 with good single-stranded DNA ligation properties. *Nucleic Acids Res.* **33**, 135–142 (2005).
45. Buchbender, A. *et al.* Improved library preparation with the new iCLIP2 protocol. *Methods* **178**, 33–48 (2020).
46. Ascano, M., Hafner, M., Cekan, P., Gerstberger, S. & Tuschl, T. Identification of RNA-protein interaction networks using PAR-CLIP. *Wiley Interdiscip. Rev. RNA* **3**, 159–177 (2012).
47. Chi, S. W., Zang, J. B., Mele, A. & Darnell, R. B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460**, 479–486 (2009).
48. Helwak, A., Kudla, G., Dudnakova, T. & Tollervey, D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153**, 654–665 (2013).

49. Grosswendt, S. *et al.* Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. *Mol. Cell* **54**, 1042–1054 (2014).
50. Sugimoto, Y. *et al.* hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. *Nature* **519**, 491–494 (2015).
51. Corley, M. *et al.* Footprinting SHAPE-eCLIP Reveals Transcriptome-wide Hydrogen Bonds at RNA-Protein Interfaces. *Mol. Cell* **0**, (2020).
52. van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol.* **18**, 424–428 (2000).
53. Xu, W., Rahman, R. & Rosbash, M. Mechanistic implications of enhanced editing by a HyperTRIBE RNA-binding protein. *RNA* **24**, 173–182 (2018).
54. Brannan, K. *et al.* Robust single-cell discovery of RNA targets of RNA binding proteins and ribosomes. (2020).
55. Taliaferro, J. M., Wang, E. T. & Burge, C. B. Genomic analysis of RNA localization. *RNA Biol.* **11**, 1040–1050 (2014).
56. Adekunle, D. A. & Wang, E. T. Transcriptome-wide organization of subcellular microenvironments revealed by ATLAS-Seq. *Nucleic Acids Res.* **48**, 5859–5872 (2020).
57. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).
58. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
59. Wang, P. *et al.* Mapping spatial transcriptome with light-activated proximity-dependent RNA labeling. *Nat. Chem. Biol.* **15**, 1110–1119 (2019).
60. Li, Y., Aggarwal, M. B., Ke, K., Nguyen, K. & Spitale, R. C. Improved Analysis of RNA Localization by Spatially Restricted Oxidation of RNA-Protein Complexes. *Biochemistry* **57**, 1577–1581 (2018).
61. Li, Y., Aggarwal, M. B., Nguyen, K., Ke, K. & Spitale, R. C. Assaying RNA Localization in

- Situ with Spatially Restricted Nucleobase Oxidation. *ACS Chem. Biol.* **12**, 2709–2714 (2017).
62. Fazal, F. M. *et al.* Atlas of Subcellular RNA Localization Revealed by APEX-Seq. *Cell* **178**, 473–490.e26 (2019).
 63. Padrón, A., Iwasaki, S. & Ingolia, N. T. Proximity RNA Labeling by APEX-Seq Reveals the Organization of Translation Initiation Complexes and Repressive RNA Granules. *Mol. Cell* **75**, 875–887.e5 (2019).
 64. Kaewsapsak, P., Shechner, D. M., Mallard, W., Rinn, J. L. & Ting, A. Y. Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *Elife* **6**, (2017).
 65. Hung, V. *et al.* Spatially resolved proteomic mapping in living cells with the engineered peroxidase APEX2. *Nat. Protoc.* **11**, 456–475 (2016).
 66. Chen, C.-L. & Perrimon, N. Proximity-dependent labeling methods for proteomic profiling in living cells. *Wiley Interdiscip. Rev. Dev. Biol.* **6**, (2017).
 67. Choder, M. mRNA imprinting: Additional level in the regulation of gene expression. *Cell. Logist.* **1**, 37–40 (2011).
 68. Gemmill, D., D'souza, S., Meier-Stephenson, V. & Patel, T. R. Current approaches for RNA-labelling to identify RNA-binding proteins. *Biochem. Cell Biol.* **98**, 31–41 (2020).
 69. Slobodin, B. & Gerst, J. E. A novel mRNA affinity purification technique for the identification of interacting proteins and transcripts in ribonucleoprotein complexes. *RNA* **16**, 2277–2290 (2010).
 70. Hogg, J. R. & Collins, K. RNA-based affinity purification reveals 7SK RNPs with distinct composition and regulation. *RNA* **13**, 868–880 (2007).
 71. Leppek, K. & Stoecklin, G. An optimized streptavidin-binding RNA aptamer for purification of ribonucleoprotein complexes identifies novel ARE-binding proteins. *Nucleic Acids Res.* **42**, e13 (2014).

72. Lee, H. Y. *et al.* RNA-protein analysis using a conditional CRISPR nuclease. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 5416–5421 (2013).
73. Flather, D. *et al.* Generation of Recombinant Polioviruses Harboring RNA Affinity Tags in the 5' and 3' Noncoding Regions of Genomic RNAs. *Viruses* **8**, (2016).
74. Hartmuth, K. *et al.* Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 16719–16724 (2002).
75. Windbichler, N. & Schroeder, R. Isolation of specific RNA-binding proteins using the streptomycin-binding RNA aptamer. *Nature Protocols* vol. 1 637–640 (2006).
76. Mili, S. & Steitz, J. A. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA* **10**, 1692–1694 (2004).
77. Simon, M. D. *et al.* The genomic binding sites of a noncoding RNA. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 20497–20502 (2011).
78. McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232–236 (2015).
79. Munschauer, M. *et al.* The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature* **561**, 132–136 (2018).
80. Chu, C. *et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161**, 404–416 (2015).
81. Theil, K., Imami, K. & Rajewsky, N. Identification of proteins and miRNAs that specifically bind an mRNA in vivo. *Nat. Commun.* **10**, 4205 (2019).
82. Flynn, R. A. *et al.* Systematic discovery and functional interrogation of SARS-CoV-2 viral RNA-host protein interactions during infection. *bioRxiv* (2020)
doi:10.1101/2020.10.06.327445.
83. Schmidt, N., Lareau, C. A., Keshishian, H. & Melanson, R. A direct RNA-protein interaction atlas of the SARS-CoV-2 RNA in infected human cells. *BioRxiv* (2020).
84. Mukherjee, J. *et al.* β -Actin mRNA interactome mapping by proximity biotinylation. *Proc.*

- Natl. Acad. Sci. U. S. A.* **116**, 12863–12872 (2019).
85. Yi, W. *et al.* CRISPR-assisted detection of RNA–protein interactions in living cells. *Nat. Methods* **17**, 685–688 (2020).
 86. Han, Y. *et al.* Directed Evolution of Split APEX2 Peroxidase. *ACS Chem. Biol.* **14**, 619–635 (2019).
 87. Hafner, M. *et al.* RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* **17**, 1697–1712 (2011).
 88. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
 89. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
 90. De, S. & Gorospe, M. Bioinformatic tools for analysis of CLIP ribonucleoprotein data. *Wiley Interdiscip. Rev. RNA* **8**, (2017).
 91. Sharma, D., Zagore, L. L., Brister, M. M. & Ye, X. The kinetic landscape of an RNA binding protein in cells. *bioRxiv* (2020).
 92. Lee, C.-Y. S. *et al.* Recruitment of mRNAs to P granules by condensation with intrinsically-disordered proteins. *Elife* **9**, (2020).
 93. Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.* **6**, 1139–1152 (2014).
 94. Krakau, S., Richard, H. & Marsico, A. PureCLIP: capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biol.* **18**, 240 (2017).
 95. Drewe-Boss, P., Wessels, H.-H. & Ohler, U. omniCLIP: probabilistic identification of protein–RNA interactions from CLIP-seq data. *Genome Biol.* **19**, 183 (2018).
 96. Huppertz, I., Haberman, N. & Ule, J. ‘Read–through marking’ reveals differential nucleotide composition of read-through and truncated cDNAs in iCLIP. *Wellcome Open Research* vol.

- 3 77 (2018).
97. Hocq, R., Paternina, J., Alasseur, Q., Genovesio, A. & Le Hir, H. Monitored eCLIP: high accuracy mapping of RNA-protein interactions. *Nucleic Acids Res.* **46**, 11553–11565 (2018).
 98. Ameer, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* **18**, 1435–1440 (2011).
 99. Sibley, C. R. *et al.* Recursive splicing in long vertebrate genes. *Nature* **521**, 371–375 (2015).
 100. Rogelj, B. *et al.* Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. *Sci. Rep.* **2**, 603 (2012).
 101. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
 102. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–8 (2009).
 103. Siddharthan, R., Siggia, E. D. & van Nimwegen, E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.* **1**, e67 (2005).
 104. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
 105. Alipanahi, B., DeLong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
 106. Wright, J. E. *et al.* A quantitative RNA code for mRNA target selection by the germline fate determinant GLD-1. *EMBO J.* **30**, 533–545 (2011).
 107. Zhao, Y. & Stormo, G. D. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.* **29**, 480–483 (2011).

108. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
109. Mukherjee, N. *et al.* Deciphering human ribonucleoprotein regulatory networks. *Nucleic Acids Res.* **47**, 570–581 (2019).
110. Brümmer, A., Kishore, S., Subasic, D., Hengartner, M. & Zavolan, M. Modeling the binding specificity of the RNA-binding protein GLD-1 suggests a function of coding region-located sites in translational repression. *RNA* **19**, 1317–1326 (2013).
111. Liu, N. *et al.* N⁶-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature* **518**, 560–564 (2015).
112. Ray, D. *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.* **27**, 667–670 (2009).
113. Fukunaga, T. *et al.* CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. *Genome Biol.* **15**, R16 (2014).
114. Maticzka, D., Lange, S. J., Costa, F. & Backofen, R. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.* **15**, R17 (2014).
115. Bahrami-Samani, E., Penalva, L. O. F., Smith, A. D. & Uren, P. J. Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Res.* **43**, 95–103 (2015).
116. Pietrosanto, M., Mattei, E., Helmer-Citterich, M. & Ferrè, F. A novel method for the identification of conserved structural patterns in RNA: From small scale to high-throughput applications. *Nucleic Acids Res.* **44**, 8600–8609 (2016).
117. Paraskevopoulou, M. D., Karagkouni, D., Vlachos, I. S., Tastsoglou, S. & Hatzigeorgiou, A. G. microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions. *Nat. Commun.* **9**, 3601 (2018).
118. Livi, C. M., Klus, P., Delli Ponti, R. & Tartaglia, G. G. catRAPID signature: identification of ribonucleoproteins and RNA-binding regions. *Bioinformatics* **32**, 773–775 (2016).
119. Khorshid, M., Hausser, J., Zavolan, M. & van Nimwegen, E. A biophysical miRNA-mRNA

- interaction model infers canonical and noncanonical targets. *Nat. Methods* **10**, 253–255 (2013).
120. Breda, J., Rzepiela, A. J., Gumienny, R., van Nimwegen, E. & Zavolan, M. Quantifying the strength of miRNA-target interactions. *Methods* **85**, 90–99 (2015).
121. Stražar, M., Žitnik, M., Zupan, B., Ule, J. & Curk, T. Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics* **32**, 1527–1535 (2016).
122. Pan, X. & Shen, H.-B. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC Bioinformatics* **18**, 136 (2017).
123. Van Nostrand, E. L. *et al.* Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biol.* **21**, 90 (2020).
124. Ule, J. *et al.* An RNA map predicting Nova-dependent splicing regulation. *Nature* **444**, 580–586 (2006).
125. Gruber, A. J. *et al.* Discovery of physiological and cancer-related regulators of 3' UTR processing with KAPAC. *Genome Biol.* **19**, 44 (2018).
126. Zarnack, K. *et al.* Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152**, 453–466 (2013).
127. Wang, S. *et al.* Enhancement of LIN28B-induced hematopoietic reprogramming by IGF2BP3. *Genes Dev.* **33**, 1048–1068 (2019).
128. Lambert, N. *et al.* RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol. Cell* **54**, 887–900 (2014).
129. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
130. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990).
131. Ghanbari, M. & Ohler, U. Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Res.* **30**, 214–226 (2020).

132. Chakrabarti, A. M., Haberman, N., Praznik, A., Luscombe, N. M. & Ule, J. Data Science Issues in Studying Protein–RNA Interactions with CLIP Technologies. *Annu. Rev. Biomed. Data Sci.* **1**, 235–261 (2018).
133. Wang, Q. *et al.* The PSI-U1 snRNP interaction regulates male mating behavior in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 5269–5274 (2016).
134. Zisoulis, D. G. *et al.* Comprehensive discovery of endogenous Argonaute binding sites in *Caenorhabditis elegans*. *Nat. Struct. Mol. Biol.* **17**, 173–179 (2010).
135. Licatalosi, D. D. & Darnell, R. B. RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* **11**, 75–87 (2010).
136. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).
137. Gerstberger, S., Hafner, M., Ascano, M. & Tuschl, T. Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease. *Adv. Exp. Med. Biol.* **825**, 1–55 (2014).
138. Yamaji, M. *et al.* DND1 maintains germline stem cells via recruitment of the CCR4-NOT complex to target mRNAs. *Nature* **543**, 568–572 (2017).
139. Kim, K. K., Yang, Y., Zhu, J., Adelstein, R. S. & Kawamoto, S. Rbfox3 controls the biogenesis of a subset of microRNAs. *Nat. Struct. Mol. Biol.* **21**, 901–910 (2014).
140. Xu, Q. *et al.* Enhanced Crosslinking Immunoprecipitation (eCLIP) Method for Efficient Identification of Protein-bound RNA in Mouse Testis. *J. Vis. Exp.* (2019) doi:10.3791/59681.
141. Li, W., Jin, Y., Prazak, L., Hammell, M. & Dubnau, J. Transposable elements in TDP-43-mediated neurodegenerative disorders. *PLoS One* **7**, e44099 (2012).
142. Vourekas, A. *et al.* The RNA helicase MOV10L1 binds piRNA precursors to initiate piRNA processing. *Genes Dev.* **29**, 617–629 (2015).
143. Vourekas, A. *et al.* Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. *Nat. Struct. Mol. Biol.* **19**, 773–781 (2012).

144. Vourekas, A., Alexiou, P., Vrettos, N., Maragkakis, M. & Mourelatos, Z. Sequence-dependent but not sequence-specific piRNA adhesion traps mRNAs to the germ plasm. *Nature* **531**, 390–394 (2016).
145. Miller, M. R., Robinson, K. J., Cleary, M. D. & Doe, C. Q. TU-tagging: cell type-specific RNA isolation from intact complex tissues. *Nat. Methods* **6**, 439–441 (2009).
146. Ule, J., Hwang, H.-W. & Darnell, R. B. The Future of Cross-Linking and Immunoprecipitation (CLIP). *Cold Spring Harb. Perspect. Biol.* **10**, (2018).
147. Saito, Y. *et al.* Differential NOVA2-Mediated Splicing in Excitatory and Inhibitory Neurons Regulates Cortical Development and Cerebellar Function. *Neuron* **101**, 707–720.e5 (2019).
148. Hwang, H.-W. *et al.* cTag-PAPERCLIP Reveals Alternative Polyadenylation Promotes Cell-Type Specific Protein Diversity and Shifts Araf Isoforms with Microglia Activation. *Neuron* **95**, 1334–1349.e5 (2017).
149. Sawicka, K. *et al.* FMRP has a cell-type-specific role in CA1 pyramidal neurons to regulate autism-related transcripts and circadian memory. *Elife* **8**, (2019).
150. Köster, T., Reichel, M. & Staiger, D. CLIP and RNA interactome studies to unravel genome-wide RNA-protein interactions in vivo in *Arabidopsis thaliana*. *Methods* **178**, 63–71 (2020).
151. Schmal, C., Reimann, P. & Staiger, D. A circadian clock-regulated toggle switch explains AtGRP7 and AtGRP8 oscillations in *Arabidopsis thaliana*. *PLoS Comput. Biol.* **9**, e1002986 (2013).
152. Reichel, M. *et al.* In Planta Determination of the mRNA-Binding Proteome of *Arabidopsis* Etiolated Seedlings. *Plant Cell* **28**, 2435–2452 (2016).
153. Zhang, Z. *et al.* UV crosslinked mRNA-binding proteins captured from leaf mesophyll protoplasts. *Plant Methods* **12**, 42 (2016).
154. Marondedze, C., Thomas, L., Serrano, N. L., Lilley, K. S. & Gehring, C. The RNA-binding protein repertoire of *Arabidopsis thaliana*. *Sci. Rep.* **6**, 29766 (2016).
155. Bach-Pages, M. *et al.* Discovering the RNA-Binding Proteome of Plant Leaves with an

- Improved RNA Interactome Capture Method. *Biomolecules* **10**, (2020).
156. Köster, T., Marondedze, C., Meyer, K. & Staiger, D. RNA-Binding Proteins Revisited - The Emerging Arabidopsis mRNA Interactome. *Trends Plant Sci.* **22**, 512–526 (2017).
157. Beckmann, B. M. *et al.* The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat. Commun.* **6**, 10127 (2015).
158. Howard, J. M. *et al.* HNRNPA1 promotes recognition of splice site decoys by U2AF2 in vivo. *Genome Res.* **28**, 689–698 (2018).
159. Blazquez, L. *et al.* Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing. *Mol. Cell* **72**, 496–509.e9 (2018).
160. Tollervey, J. R. *et al.* Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci.* **14**, 452–458 (2011).
161. Yamazaki, T. *et al.* Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. *Mol. Cell* **70**, 1038–1053.e7 (2018).
162. Modic, M. *et al.* Cross-Regulation between TDP-43 and Paraspeckles Promotes Pluripotency-Differentiation Transition. *Molecular Cell* (2019)
doi:10.1016/j.molcel.2019.03.041.
163. Horos, R. *et al.* The Small Non-coding Vault RNA1-1 Acts as a Riboregulator of Autophagy. *Cell* **176**, 1054–1067.e12 (2019).
164. Holmqvist, E. *et al.* Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. *EMBO J.* **35**, 991–1011 (2016).
165. Gottwein, E. *et al.* Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines. *Cell Host Microbe* **10**, 515–526 (2011).
166. Gay, L. A., Sethuraman, S., Thomas, M., Turner, P. C. & Renne, R. Modified Cross-Linking, Ligation, and Sequencing of Hybrids (qCLASH) Identifies Kaposi's Sarcoma-Associated Herpesvirus MicroRNA Targets in Endothelial Cells. *J. Virol.* **92**, (2018).
167. Kutluay, S. B. *et al.* Global changes in the RNA binding specificity of HIV-1 gag regulate

- virion genesis. *Cell* **159**, 1096–1109 (2014).
168. Apolonia, L. *et al.* Promiscuous RNA binding ensures effective encapsidation of APOBEC3 proteins by HIV-1. *PLoS Pathog.* **11**, e1004609 (2015).
169. Flynn, R. A. *et al.* Dissecting noncoding and pathogen RNA–protein interactomes. *RNA* **21**, 135–143 (2015).
170. Banerjee, A. K. *et al.* SARS-CoV-2 Disrupts Splicing, Translation, and Protein Trafficking to Suppress Host Defenses. *Cell* (2020) doi:10.1016/j.cell.2020.10.004.
171. Nabeel-Shah, S. *et al.* SARS-CoV-2 Nucleocapsid protein attenuates stress granule formation and alters gene expression via direct interaction with host mRNAs. *Cold Spring Harbor Laboratory* 2020.10.23.342113 (2020) doi:10.1101/2020.10.23.342113.
172. Pandya-Jones, A. *et al.* A protein assembly mediates Xist localization and gene silencing. *Nature* **587**, 145–151 (2020).
173. Tauber, D., Tauber, G. & Parker, R. Mechanisms and Regulation of RNA Condensation in RNP Granule Formation. *Trends Biochem. Sci.* **45**, 764–778 (2020).
174. Formicola, N., Vijayakumar, J. & Besse, F. Neuronal ribonucleoprotein granules: Dynamic sensors of localized signals. *Traffic* **20**, 639–649 (2019).
175. Uren, P. J. *et al.* High-throughput analyses of hnRNP H1 dissects its multi-functional aspect. *RNA Biol.* **13**, 400–411 (2016).
176. Blackinton, J. G. & Keene, J. D. Functional coordination and HuR-mediated regulation of mRNA stability during T cell activation. *Nucleic Acids Res.* **44**, 426–436 (2016).
177. Kishore, S. *et al.* A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature Methods* vol. 8 559–564 (2011).
178. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
179. Jolma, A. *et al.* Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. *Genome Res.* **30**, 962–973 (2020).

180. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**, D13–21 (2008).
181. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
182. Blin, K. *et al.* DoRiNA 2.0—upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.* **43**, D160–D167 (2015).
183. Li, J.-H., Liu, S., Zhou, H., Qu, L.-H. & Yang, J.-H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* (2013) doi:10.1093/nar/gkt1248.
184. Zhu, Y. *et al.* POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Res.* **47**, D203–D211 (2019).
185. Lewinski, M., Bramkamp, Y., Köster, T. & Staiger, D. SEQing: web-based visualization of iCLIP and RNA-seq data in an interactive python framework. *BMC Bioinformatics* **21**, 113 (2020).
186. Giudice, G., Sánchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. ATtRACT—a database of RNA-binding proteins and associated motifs. *Database* **2016**, (2016).
187. Jankowsky, E. & Harris, M. E. Specificity and nonspecificity in RNA-protein interactions. *Nat. Rev. Mol. Cell Biol.* **16**, 533–544 (2015).
188. Attig, J. *et al.* Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing. *Cell* **174**, 1067–1081.e17 (2018).
189. Beltran, M. *et al.* The interaction of PRC2 with RNA or chromatin is mutually antagonistic. *Genome Res.* **26**, 896–907 (2016).
190. Warner, J. R. & McIntosh, K. B. How common are extraribosomal functions of ribosomal proteins? *Mol. Cell* **34**, 3–11 (2009).
191. Briese, M. *et al.* A systems view of spliceosomal assembly and branchpoints with iCLIP.

- Nat. Struct. Mol. Biol.* **26**, 930–940 (2019).
192. Cai, S. *et al.* Investigations on the interface of nucleic acid aptamers and binding targets. *Analyst* **143**, 5317–5338 (2018).
193. McHugh, C. A. & Guttman, M. RAP-MS: A Method to Identify Proteins that Interact Directly with a Specific RNA Molecule in Cells. *Methods Mol. Biol.* **1649**, 473–488 (2018).
194. Zeng, F. *et al.* A protocol for PAIR: PNA-assisted identification of RNA binding proteins in living cells. *Nat. Protoc.* **1**, 920–927 (2006).
195. Bell, T. J., Eiríksdóttir, E., Langel, U. & Eberwine, J. PAIR technology: exon-specific RNA-binding protein isolation in live cells. *Methods Mol. Biol.* **683**, 473–486 (2011).
196. Matia-González, A. M., Iadevaia, V. & Gerber, A. P. A versatile tandem RNA isolation procedure to capture in vivo formed mRNA-protein complexes. *Methods* **118-119**, 93–100 (2017).
197. Mellacheruvu, D. *et al.* The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* **10**, 730–736 (2013).
198. Trinkle-Mulcahy, L. Recent advances in proximity-based labeling methods for interactome mapping. *F1000Res.* **8**, (2019).
199. Cronan, J. E. Targeted and proximity-dependent promiscuous protein biotinylation by a mutant *Escherichia coli* biotin protein ligase. *J. Nutr. Biochem.* **16**, 416–418 (2005).
200. Branon, T. C. *et al.* Efficient proximity labeling in living cells and organisms with TurboID. *Nat. Biotechnol.* **36**, 880–887 (2018).
201. Kim, D. I. *et al.* An improved smaller biotin ligase for BioID proximity labeling. *Mol. Biol. Cell* **27**, 1188–1196 (2016).
202. Kido, K. *et al.* AirID, a novel proximity biotinylation enzyme, for analysis of protein–protein interactions. *Elife* **9**, e54983 (2020).
203. Kapusta, A. & Feschotte, C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet.* **30**, 439–452 (2014).

204. Attig, J. & Ule, J. Genomic Accumulation of Retrotransposons Was Facilitated by Repressive RNA-Binding Proteins: A Hypothesis. *Bioessays* **41**, e1800132 (2019).
205. Martí-Gómez, C., Lara-Pezzi, E. & Sánchez-Cabo, F. dSreg: a Bayesian model to integrate changes in splicing and RNA-binding protein activity. *Bioinformatics* **36**, 2134–2141 (2020).
206. Witten, J. T. & Ule, J. Understanding splicing regulation through RNA splicing maps. *Trends Genet.* **27**, 89–97 (2011).
207. Goering, R. *et al.* FMRP promotes RNA localization to neuronal projections through interactions between its RGG domain and G-quadruplex RNA sequences. *Elife* **9**, (2020).
208. Dermit, M. *et al.* Subcellular mRNA Localization Regulates Ribosome Biogenesis in Migrating Cells. *Dev. Cell* **55**, 298–313.e10 (2020).
209. Lyon, A. S., Peeples, W. B. & Rosen, M. K. A framework for understanding the functions of biomolecular condensates across scales. *Nat. Rev. Mol. Cell Biol.* (2020)
doi:10.1038/s41580-020-00303-z.
210. del Campo, E. M. Post-transcriptional control of chloroplast gene expression. *Gene Regul. Syst. Bio.* **3**, 31–47 (2009).
211. Sutandy, F. X. R. *et al.* In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. *Genome Res.* **28**, 699–713 (2018).
212. Strittmatter, L. M. *et al.* PsiCLIP reveals dynamic RNA binding by DEAH-box helicases before and after exon ligation. 2020.03.15.992701 (2020) doi:10.1101/2020.03.15.992701.
213. Porter, D. F. & Khavari, P. A. easyCLIP Quantifies RNA-Protein Interactions and Characterizes Recurrent PCBP1 Mutations in Cancer. 635888 (2019) doi:10.1101/635888.
214. Ule, J. & Blencowe, B. J. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Mol. Cell* **76**, 329–345 (2019).
215. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **169**, 1187–1200 (2017).
216. Capitanichik, C. A., Toolan-Kerr, P., Luscombe, N. M. & Ule, J. How do you identify m6A

- methylation in transcriptomes at high resolution? A comparison of recent datasets. *Front. Genet.* **11**, 398 (2020).
217. Lu, Z. & Chang, H. Y. Decoding the RNA structure. *Curr. Opin. Struct. Biol.* **36**, 142–148 (2016).
218. Cai, Z. *et al.* RIC-seq for global in situ profiling of RNA-RNA spatial interactions. *Nature* **582**, 432–437 (2020).
219. Foley, S. W. *et al.* A Global View of RNA-Protein Interactions Identifies Post-transcriptional Regulators of Root Hair Cell Fate. *Dev. Cell* **41**, 204–220.e5 (2017).
220. Casas-Vila, N., Sayols, S., Pérez-Martínez, L., Scheibe, M. & Butter, F. The RNA fold interactome of evolutionary conserved RNA structures in *S. cerevisiae*. *Nat. Commun.* **11**, 2789 (2020).
221. Hussain, S. *et al.* NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell Rep.* **4**, 255–261 (2013).
222. Linder, B. *et al.* Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods* **12**, 767–772 (2015).
223. Helm, M., Lyko, F. & Motorin, Y. Limited antibody specificity compromises epitranscriptomic analyses. *Nat. Commun.* **10**, 5669 (2019).
224. Tang, Y. *et al.* m6A-Atlas: a comprehensive knowledgebase for unraveling the N6-methyladenosine (m6A) epitranscriptome. *Nucleic Acids Res.* (2020)
doi:10.1093/nar/gkaa692.
225. Rees, J. S., Li, X., Perrett, S., Lilley, K. S. & Jackson, A. P. Selective Proteomic Proximity Labeling Assay Using Tyramide (SPPLAT): A Quantitative Method for the Proteomic Analysis of Localized Membrane-Bound Protein Clusters. *Curr. Protoc. Protein Sci.* **80**, 10.5:10.5.1 (2015).
226. Martell, J. D. *et al.* A split horseradish peroxidase for the detection of intercellular protein-protein interactions and sensitive visualization of synapses. *Nat. Biotechnol.* **34**, 774–780

- (2016).
227. Xue, M. *et al.* Optimizing the fragment complementation of APEX2 for detection of specific protein-protein interactions in live cells. *Sci. Rep.* **7**, 12039 (2017).
228. Roux, K. J., Kim, D. I. & Burke, B. BioID: A Screen for Protein-Protein Interactions. *Curr. Protoc. Protein Sci.* **74**, 4.3.1 (2013).
229. Schopp, I. M. *et al.* Split-BioID a conditional proteomics approach to monitor the composition of spatiotemporally defined protein complexes. *Nat. Commun.* **8**, 15690 (2017).
230. De Munter, S. *et al.* Split-BioID: a proximity biotinylation assay for dimerization-dependent protein interactions. *FEBS Lett.* **591**, 415–424 (2017).
231. Cho, K. F. *et al.* Split-TurboID enables contact-dependent proximity labeling in cells. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 12143–12154 (2020).
232. Ramanathan, M. *et al.* RNA-protein interaction detection in living cells. *Nat. Methods* **15**, 207–212 (2018).
233. Kucukural, A., Özadam, H., Singh, G., Moore, M. J. & Cenik, C. ASPeak: an abundance sensitive peak detection algorithm for RIP-Seq. *Bioinformatics* **29**, 2485–2486 (2013).
234. Golumbeanu, M., Mohammadi, P. & Beerewinkel, N. BMix: probabilistic modeling of occurring substitutions in PAR-CLIP data. *Bioinformatics* **32**, 976–983 (2016).
235. Zhang, Z. & Xing, Y. CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. *Nucleic Acids Res.* **45**, 9260–9271 (2017).
236. Park, S. *et al.* CLIPick: a sensitive peak caller for expression-based deconvolution of HITS-CLIP signals. *Nucleic Acids Res.* **46**, 11153–11168 (2018).
237. Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.* **20**, 1434–1442 (2013).
238. Shah, A., Qian, Y., Weyn-Vanhenyryck, S. M. & Zhang, C. CLIP Tool Kit (CTK): a flexible

- and robust pipeline to analyze CLIP sequencing data. *Bioinformatics* **33**, 566–567 (2017).
239. Wang, Z. *et al.* iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol.* **8**, e1000530 (2010).
240. Chen, B., Yun, J., Kim, M. S., Mendell, J. T. & Xie, Y. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome Biol.* **15**, R18 (2014).
241. Uren, P. J. *et al.* Site identification in high-throughput RNA–protein interaction data. *Bioinformatics* **28**, 3013–3020 (2012).
242. Tree, J. J., Granneman, S., McAteer, S. P., Tollervey, D. & Gally, D. L. Identification of bacteriophage-encoded anti-sRNAs in pathogenic *Escherichia coli*. *Mol. Cell* **55**, 199–213 (2014).
243. Comoglio, F., Sievers, C. & Paro, R. Sensitive and highly resolved identification of RNA–protein interaction sites in PAR-CLIP data. *BMC Bioinformatics* **16**, 32 (2015).
244. Palmer, L. E., Weiss, M. J. & Paralkar, V. R. YODEL: Peak calling software for HITS-CLIP data. *F1000Res.* **6**, 1138 (2017).
245. Huppertz, I. *et al.* iCLIP: protein-RNA interactions at nucleotide resolution. *Methods* **65**, 274–287 (2014).
246. Zhao, Y. *et al.* SpyCLIP: an easy-to-use and high-throughput compatible CLIP platform for the characterization of protein–RNA interactions with high accuracy. *Nucleic Acids Res.* **47**, e33–e33 (2019).
247. Schneider, C., Kudla, G., Wlotzka, W., Tuck, A. & Tollervey, D. Transcriptome-wide analysis of exosome targets. *Mol. Cell* **48**, 422–433 (2012).
248. Miniard, A. C., Middleton, L. M., Budiman, M. E., Gerber, C. A. & Driscoll, D. M. Nucleolin binds to a subset of selenoprotein mRNAs and regulates their expression. *Nucleic Acids Res.* **38**, 4807–4820 (2010).
249. Choudhury, N. R. *et al.* Tissue-specific control of brain-enriched miR-7 biogenesis. *Genes Dev.* **27**, 24–38 (2013).

250. Zielinski, J. *et al.* In vivo identification of ribonucleoprotein-RNA interactions. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 1557–1562 (2006).
251. Rogell, B. *et al.* Specific RNP capture with antisense LNA/DNA mixmers. *RNA* **23**, 1290–1302 (2017).
252. Sharma, S. Isolation of a sequence-specific RNA binding protein, polypyrimidine tract binding protein, using RNA affinity chromatography. *Methods Mol. Biol.* **488**, 1–8 (2008).
253. Tsai, B. P., Wang, X., Huang, L. & Waterman, M. L. Quantitative profiling of in vivo-assembled RNA-protein complexes using a novel integrated proteomic approach. *Mol. Cell. Proteomics* **10**, M110.007385 (2011).
254. Yoon, J.-H., Srikantan, S. & Gorospe, M. MS2-TRAP (MS2-tagged RNA affinity purification): tagging RNA to identify associated miRNAs. *Methods* **58**, 81–87 (2012).
255. Carey, J., Cameron, V., de Haseth, P. L. & Uhlenbeck, O. C. Sequence-specific interaction of R17 coat protein with its ribonucleic acid binding site. *Biochemistry* **22**, 2601–2610 (1983).
256. Lim, F., Downey, T. P. & Peabody, D. S. Translational repression and specific RNA binding by the coat protein of the Pseudomonas phage PP7. *J. Biol. Chem.* **276**, 22507–22513 (2001).
257. Deckert, J. *et al.* Protein composition and electron microscopy structure of affinity-purified human spliceosomal B complexes isolated under physiological conditions. *Mol. Cell. Biol.* **26**, 5528–5543 (2006).
258. Villaseñor, R. *et al.* ChromID identifies the protein interactome at chromatin marks. *Nat. Biotechnol.* **38**, 728–736 (2020).
259. Zhang, Z. *et al.* Capturing RNA–protein interaction via CRUIS. *Nucleic Acids Res.* **48**, e52–e52 (2020).
260. Han, S. *et al.* RNA-protein interaction mapping via MS2 or Cas13-based APEX targeting. 2020.02.27.968297 (2020) doi:10.1101/2020.02.27.968297.

261.Lin, X. & Lawrenson, K. In Vivo Analysis of RNA Proximity Proteomes Using RiboPro.

bioRxiv 2020.02.28.970442 (2020).

Figures and legends

Figure 1: Overview of the general CLIP workflow

A schematic overview of the core steps that are common to most variants of the CLIP protocol. The RNA is in light gray, cDNA in dark grey, and adapters/primers in blue. The RNA-binding proteins (RBP) are in blue or grey blobs, as well as the peptide remaining on the RNA after proteinase K treatment. The figure is adapted from Lee et al², with permission from Elsevier.

Figure 2: Overview of primary CLIP variants and TRIBE

A comparative schematic of methods is subdivided into three sections. Red: Includes all steps prior to immunoprecipitation (when applicable), including treatment of cultured cells and crosslinking. Blue: RNA manipulation. Green: cDNA preparation and sequencing. Methods that share similar crosslinking or RNA modification strategies are grouped by the extended vertical lines in the red section. Note that PAR-CLIP is predominantly performed using 4-thiouridine as photoreactive nucleoside, but 6-thioguanosine can also be used and results in a G-to-A transition. Note that CRAC, which closely resembles HITS-CLIP, uses protein tags that allow denaturing purification.

Figure 3: Overview of CLIP analysis.

(A) Main steps of sample preparation with associated sources of noise. RBP-RNA interactions are dynamic and therefore, the probability of an RBP to crosslink to a cognate site in an RNA at the time of experiment is affected by multiple factors: synergistic or antagonistic interactions between RBPs on the same RNA region, the interaction affinity (the residence time of the RBP being low on low-affinity sites and high on high-affinity sites), the availability of the RBP and the cognate site, due to time-dependent stochastic fluctuations in expression and localization. After crosslinking, cells are lysed and the RNAs fragmented. An RBP-specific antibody is used to immunoprecipitate the protein along with crosslinked RNA fragments. Cross-reactivity (to 'blue' protein) or lack of antibody binding can lead to false or undetected sites (grey box). The size distribution of fragments can impact the recovery of crosslinking sites. The crosslink constitutes a roadblock for reverse transcription, leading stochastically to different types of fragments: those that are accurately transcribed across the crosslink sites, those where reverse transcription stops at the crosslink site and those where mutations or deletions are introduced at the site of crosslink. iCLIP variants aim to capture the fragments that truncate at the crosslink position, while PAR-CLIP aims to capture fragments where readthrough occurs. (B) Main computational steps leading to the extraction of peaks from CLIP data. First, adapter sequences as well as PCR duplicates are removed (for clarity the parts that are removed at this step are shown as faded colors), and then the inserts are mapped to the genome or transcriptome. The central panel shows read density profiles in the region of the tubulin (TUBB) gene, the tracks corresponding to samples from K562 cells obtained in the ENCODE project: PUM2-eCLIP, PUM2-SMInput, and RNA-seq (figure from the IGV genome browser). Various approaches are used to distinguish peaks of high

RBP occupancy from background. Background models are constructed from regions neighboring the putative peaks in the CLIP sample itself, or from the same region as the peak in the SMInput or the RNA-seq samples (indicated by colored brackets). Peaks are defined as contiguous regions where the number of reads is significantly higher than expected based on the background models (indicated in the cartoon by the colored dashed lines that show the average coverage in different types of background, same color scheme as in the left panel). Some tools consider not only the read counts but also the number and pattern of crosslink-diagnostic mutations (red boxes in individual reads shown under the peak). (C) Peak analysis. Typically, peaks that are reproducibly identified in replicate experiments are extracted for further analyses. Here, the agreement between the peaks obtained in two replicates of PUM2 eCLIP is shown as a function of the number of top peaks selected from each replicate. Peaks are sorted by score, the top x peaks (x indicated by the x-axis) are extracted, and the proportion of overlapping peaks is shown on the y-axis. Two peaks are considered as overlapping if they share at least one nucleotide. Reproducible peaks can then be annotated with their location in different genomic regions, the types of RNAs in which they occur or the region of protein-coding RNAs (5' UTR, CDS, 3' UTR) in which they reside. The sequences of the most enriched peaks are also typically used to search for enriched sequence motifs that point to the sequence preference of the RBP. In this case, the motif identified from the top peaks is indeed known to be the recognition element of the Pumilio2 protein.

Figure 4: CLIP applications in model organisms.

Shown are representative applications of CLIP in cultured cells, intact tissue, several animal models, and transgenic plants. The red circle indicates nucleotides such as 4-thiouridine (³SU). Points to be considered for each experimental system as well as the unique advantages are presented.

Tables and boxes

Table 1: Characteristics of the commonly used proximity enzymes

Enzyme	Source (size)	Labeling range	Substrate (incubation time)
Peroxidase based (fixed cells)			

Horseradish Peroxidase (HRP) ²²⁵	Horseradish (44KDa)	200-300 nm	For fluorescent microscopy = Amplex Red, H ₂ O ₂ , For proximity labeling = Biotin phenol, H ₂ O ₂ , For electron microscopy : DAB, OsO ₄ , H ₂ O ₂ (5-10 mins)
split HRP (sHRP) ²²⁶	Horseradish (44KDa)	200-300 nm	For fluorescent microscopy = Amplex Red, H ₂ O ₂ , For proximity labeling = Biotin phenol, H ₂ O ₂ , For electron microscopy : DAB, OsO ₄ , H ₂ O ₂ (45 mins)
APEX ^{64,198}	Pea (Synthetic) (28kDa)	10-20 nm	Biotin-phenol (30-60min)
APEX ²²⁷	Soybean (Synthetic) 28kDa	10-20 nm	Biotin-phenol, Biotin- aniline, Biotin-naphthylamine (30-60min)
Split APEX ²⁸⁶	Soybean (Synthetic) 28kDa	10-20 nm	Biotin-phenol, Biotin- aniline, Biotin-naphthylamine (30-60min)
Biotin ligase (live cells)			
BioID ²²⁸	<i>E.coli</i> (37KDa)	10-15nm	Biotin (50 μM) (6-24 hrs)
Split BioID ^{229,230}	<i>E.coli</i>	10-15nm	Biotin (50 μM) (6-24 hrs)
BioID ²⁰¹	<i>A.aeolicus</i> (27kDa)	10-15nm	Biotin (3.2 μM) (6-24 hrs)
TurboID ²⁰⁰	<i>E.coli</i> (35kDa)	10-15nm	Biotin (10-60 min)
MiniTurboID ²⁰⁰	<i>E.coli</i> (28kDa)	10-15nm	Biotin (10-60 min)
Split TurboID ²³¹	<i>E.coli</i>	10-15nm	Biotin (10-60 min)

BASU ²³²	<i>B.subtilis</i> (28kDa)	10-15nm	Biotin (200 µM) (30 min - 18hrs)
AirID ²⁰²	Synthetic	--	Biotin (5 µM) (3hrs)

Table 2: Available peak detection software

Feature/ Method	Supported Protocols	Background	Model for peak calling	Uses crosslink- diagnostic events	Additional Features	Repository	Documentation Examples Comments
ASPeak ²³³	RIP-seq,H	external sample (RNA-seq or RIP-input)	NB (parametrised for each genomic interval)	none		https://sourceforge.net/projects/as-peak/	Available/Available
BMix ²³⁴	P	Substitutions other than crosslink- diagnostic	Mixture model for substitutions, sources of error modeled based on non-crosslink- induced mutations	substitution		https://github.com/cbg-ethz/BMix	Limited / Test case / No tutorial
CLAM ²³⁵	H,P,i,e RIP-seq	Resampled foreground reads within gene	Benjamini- Hochberg False Discovery Rate	none	Integrated pre- processing explicit use of multimappers	https://github.com/Xinglab/CLAM	Very detailed / Available / Tutorial with reproducible analyses
CLIPick ²³⁶	H	Coverage simulated based on gene expression data	Cubic spline interpolation	none		https://github.com/CLIPick/CLIPick-package	Very detailed / Available / Tutorial
CLIPper ²³⁷	e,H,P,i	Resampled foreground reads within transcript	FDR relative to resampled coverage per position, cubic spline interpolation	none		https://github.com/YeoLab/clipper	Sparse / None / Preselected genome assemblies

			to extract peaks, Poisson distribution to calculate enrichment p-value				
CLIP Tool Kit ²³⁸	H,P,i,e	Randomized diagnostic events across reads	Binomial distribution for diagnostic events	multiple	Integrated pre-processing	https://github.com/chaolinzhanglab/ctk	Documentation / Available / Tutorials for how to pre-process data from each CLIP variant
iCount ²³⁹	i	Resampled foreground reads within gene region	FDR relative to resampled coverage per region	truncation	Integrated pre-processing, kmer-finder, web-interface	https://github.com/tomazc/iCount	Detailed / Available / Tutorial on iCLIP data analysis
OmniCLIP ⁹⁵	H,P,i,e iCLAP, CRAC	External sample (RNA-seq, SMI)	NHMM (GLM for coverage profile, DMM diagnostic events)	multiple	Multiple inputs handled in one run	https://github.com/phillipppdre/omniCLIP	Limited / None / No tutorial
PIPE-CLIP ²⁴⁰	H,P,i,e	No	ZTNB for coverage, binomial for diagnostic events	multiple	Integrated pre-processing, integrated motif analysis (external installation required)	https://github.com/QBR/C/PIPE-CLIP	Limited / Available / Tutorial based on Galaxy, discontinued
Piranha ²⁴¹	H,P,i,e RIP-seq	Low coverage regions from foreground sample	ZTNB model (ZTNBR if covariates are provided)	none	Cross-sample analysis (differential binding detection)	https://github.com/smithlabcode/piranha	Available / None / No tutorial
PureCLIP ⁹⁴	i,e	Optional external sample (RNA-seq, SMI)	NHMM (LTG for coverage, ZTB for truncations)	truncation		https://github.com/skrakau/PureCLIP	Detailed / Available / Tutorial on how to pre-process data
pyCRAC ²⁴²	CRAC,H,P, i,e	Resampled foreground reads within gene	FDR relative to randomised distribution	multiple	Integrated pre-processing, supports multimappers	https://git.ed.ac.uk/sgrannem/pycrac	Very detailed / Examples / Tutorial with results and

					integrated motif analysis		visualizations
wavCluster ²⁴³	P	Substitutions other than crosslink-diagnostic	CWT of the coverage function	substitution	Integrated motif analysis	https://github.com/FedericoComoglio/wavCluster	Documentation available in Bioconductor
YODEL ²⁴⁴	H	No	Highest coverage within cluster of overlapping reads	none	Multiple inputs handled in one run	https://github.com/LancePalmerSt Jude/YODEL/	None available

i: iCLIP, e: eCLIP, P: PAR-CLIP, H:HITS-CLIP

CWT: continuous wavelet transform, DMM: dirichlet-multinomial mixture, EM: expectation-maximization, GLM: generalised linear model, LTG: left-truncated gamma distribution, NHMM: non-homogeneous markov model, ZTB: zero-truncated binomial distribution, ZTNB: zero-truncated negative binomial, ZTNBR: zero-truncated negative binomial regression;

Protocol/data for which the software was primarily developed are shown in bold-face.

Box 1: Purification of RBP-RNA complexes in CLIP

Most CLIP experiments are done using immunoprecipitation (IP) against intact endogenous RBPs under conditions aimed to remove other RBPs that interact with the RBP-of-interest, e.g. using denaturing detergents and high salt. An alternative approach, established first in yeast by the CRAC method, is to use affinity tags such as His-tag, FLAG-tag, SpyTag or others, which enable the use of fully denaturing conditions during purification, thereby maximizing stringency in order to fully dissociate even the most stable RNPs^{2,39,212,245,246}. Moreover, split-CRAC is performed using cleavable proteins with a tag on either end of the protein, which can reveal the distinct RNA binding roles of different domains in an RBP²⁴⁷. SDS-PAGE separation of the immunoprecipitated RBP-RNA complexes and transfer to nitrocellulose enables further purification by size-selection, as it fractionates RBPs of different molecular weight, and by reducing the amount of co-purified non-crosslinked RNAs that does not bind as well to nitrocellulose. Moreover, visualization of RBP-RNA complexes after SDS-PAGE separation and membrane transfer is used to determine appropriate conditions of RNase fragmentation and to optimise the various steps of purification with the use of negative controls in order to achieve maximal sensitivity and specificity of the purified RBP. Visualization also enables appropriate size-selection of the specific RBP

crosslinked to RNAs, according to guidelines that incorporate the size of adapter and RNA fragments¹⁹. Originally, radiolabelling was used for visualisation, whereas irCLIP circumvents this by introducing use of an adapter with infrared fluorescent label⁴¹. On the other hand, eCLIP omits the estimation of extrinsic background via visualisation, and instead excises a broad area up to ~75 kDa above where the unligated RBP is estimated to migrate based on its analysis via Western blot⁴².

Supplementary Table: Key methods to identify protein partners of a specific RNA

Method	Conditions / Description	Advantages and Limitations
1. RNA affinity-capture based methods to identify direct RNA binders		
By using RNA probes immobilized on beads ^{248,249}	In vitro transcribed or synthesized RNA baits - covalently linked to a solid support, incubation with the whole cellular lysates	These methods work best with shorter transcripts (<100 nt) such as pri/pre miRNAs, specific regulatory motifs.
By using modified antisense oligos	RNA antisense purification coupled with mass spectrometry (RAP-MS) ¹⁹³	Chemical modifications of the probe affect the secondary structure of RNA, resulting in structural rearrangements that interfere with complex formation.
	PNA (peptide nucleic acid analogues) probes ²⁵⁰	These hybrids increase the stability and affinity by significantly increasing the melting temperatures, stability and fast capturing, they are also resistant to proteases and nucleases. Due to UV crosslinking, the RNP structures remain intact.
	Antisense locked nucleic acid (LNA)/DNA oligonucleotides ²⁵¹ 20-mer probes with full complementarity to the target RNA sequence	
	Antisense oligos 5' or 3' end-modified by biotin or other means ²⁵² . Used in <i>C. elegans</i> in vIPR (in vivo Interactions by pulldown of RNA) ⁸¹ .	High-affinity binding can be captured via antiDIG antibody - rapid, specific, resistant to high salt concentration, heat, pH and proteolysis.

By using aptamers (see main text for advantages and limitations)	MS2 aptamers: MS2 tagged endogenous RNA and stable expression of MCP-HTBH tag (6x histidine clusters separated by a TEV cleavage sequence with an in vivo biotinylation site), streptavidin beads ²⁵³ , or MS2-TRAP method ²⁵⁴ .
	PP7 aptamers: 25 nt long stem-loop aptamer fused to the 5' / 3' end of RNA, affinity purification via PP7 binding coat protein (Kd ~ 1 nM) ^{255,256} .
	S1 and D8 aptamer : S1 (44-nt long) binds to streptavidin (Kd ~ 70 nM)) and D8 (33-nt long) binds to Sephadex (polysaccharide dextran B512) ⁷¹ .
	Tobramycin binding aptamer (40 nt) and streptomycin binding aptamer (46 nt) bind to tobramycin (Kd ~ 5 nM) and streptomycin (Kd ~ 1 μM) ^{74,257} .
	CRISPR/Csy4 aptamer : in vitro generated RNA transcripts with 16 nt hairpin (5 bp stem and 5 nt loop) binds (Kd = 50 pM) irreversibly with an inactive, biotinylated form of Csy4 endoribonuclease ⁷² .

2. Proximity based methods in live cells to identify direct and transient RNA binders		
RNA-protein interaction detection (RaPID) ^{232,258}	constitutive expression of BioID and BoxB, heterogenous λN labelled RNA, streptavidin pulldown	Transient and direct heterogeneous RNA interactors can be mapped, 16hrs long incubation time fails to identify dynamic interactions.
RNA BioID ⁸⁴	constitutive expression of BioID and MCP, endogenous MS2 labelled RNA, Biotin, streptavidin pulldown	Conditional variation of RNA behaviour can be mapped for the entire lifetime. Labelling time is at least 6hrs so dynamics cannot be mapped.
CRISPR-based RNA-United Interacting System (CRUIS) ²⁵⁹	dLwaCas13a (creating R474A and R1046A mutations in the LwaCas13a) fused with PafA, PupE, streptavidin beads	With an 19-aa linker (~7 nm) the labeling radius is 17 nm (~50) bases.. One sgRNA gives RNA site specific RBPome information. Due to the size, CRISPR-based targeting might affect the structure and the patterns of interacting protein to the RNA.
CRISPR-assisted RNA-protein interaction detection method (CARPID) ⁸⁵	RNA-targeting type VI-D CRISPR single effector dCasRx coexpressing BASU, two gRNA sequences spaced by a 30-nucleotide repeat to target lncRNA transcript, biotin, streptavidin beads	Targeting the same RNA with two different sgRNAs, reduces the background proteome. Due to the large size, CRISPR-based targeting might affect the structure of the targeted RNA, and the interacting proteins.
MS2 or Cas13-based APEX targeting ²⁶⁰	MS2 stem loop, MS2 coat protein-fused APEX2 (MCP-APEX2) or Cas13-APEX2 fusion (dCas13-APEX2) with sgRNA, biotin-phenol and H2O2, streptavidin beads	One sgRNA only gives RNA site specific information.
RiboPro (Ribonucleic acid proximity protein labelling) ²⁶¹	Catalytically dead Cas13 (dCas13) expressing APEX2 (dPspCas13b-Flag-APEX2-HA), sgRNA, biotin-phenol and H2O2, streptavidin beads	

[H1] Glossary

Intrinsically disordered region (IDR): A polypeptide region that doesn't form a defined three-dimensional structure in solution, but tends to contain multivalent, assembly-promoting segments, the functionality of which is heavily modulated by posttranslational modifications¹.

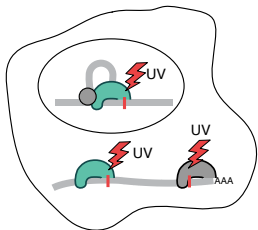
Biomolecular condensate is a membraneless assembly of proteins and/or nucleic acids, driven by multivalent interactions formed by protein domains, IDRs and/or nucleic acids²⁰⁹.

Positional weight matrix (PWM): representation of binding sites of nucleic acid-binding proteins, including RBPs. The matrix columns correspond to individual, contiguous positions in binding sites, while the rows correspond to the four possible nucleotides. The value in a given row and column gives the relative frequency with which the nucleotide specified by the row occurs at the position specified by the column in binding sites of the RBP.

Watson-Crick face: Part of the nucleobases that are involved in hydrogen bonding for canonical base-pairing.

Preparation of crosslinked cell lysate

1. Covalent protein-RNA crosslinking



2. Cell lysis

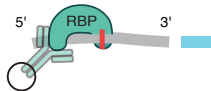
3. RNA fragmentation



Purification of specific crosslinked RNA fragments

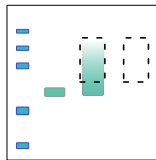
4. Immunoprecipitation (IP) or affinity purification of protein-RNA complexes

5. Ligation of adapter to fragmented RNA



6. Purification of protein-bound RNA by SDS-PAGE

RNase 
IP RBP control



Proteinase K digestion



cDNA library sequencing and analysis

7. Reverse transcription

RNA: 
5'  3'

Most cDNAs truncate at crosslink sites:

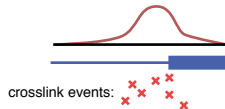
3' 

A readthrough cDNA:

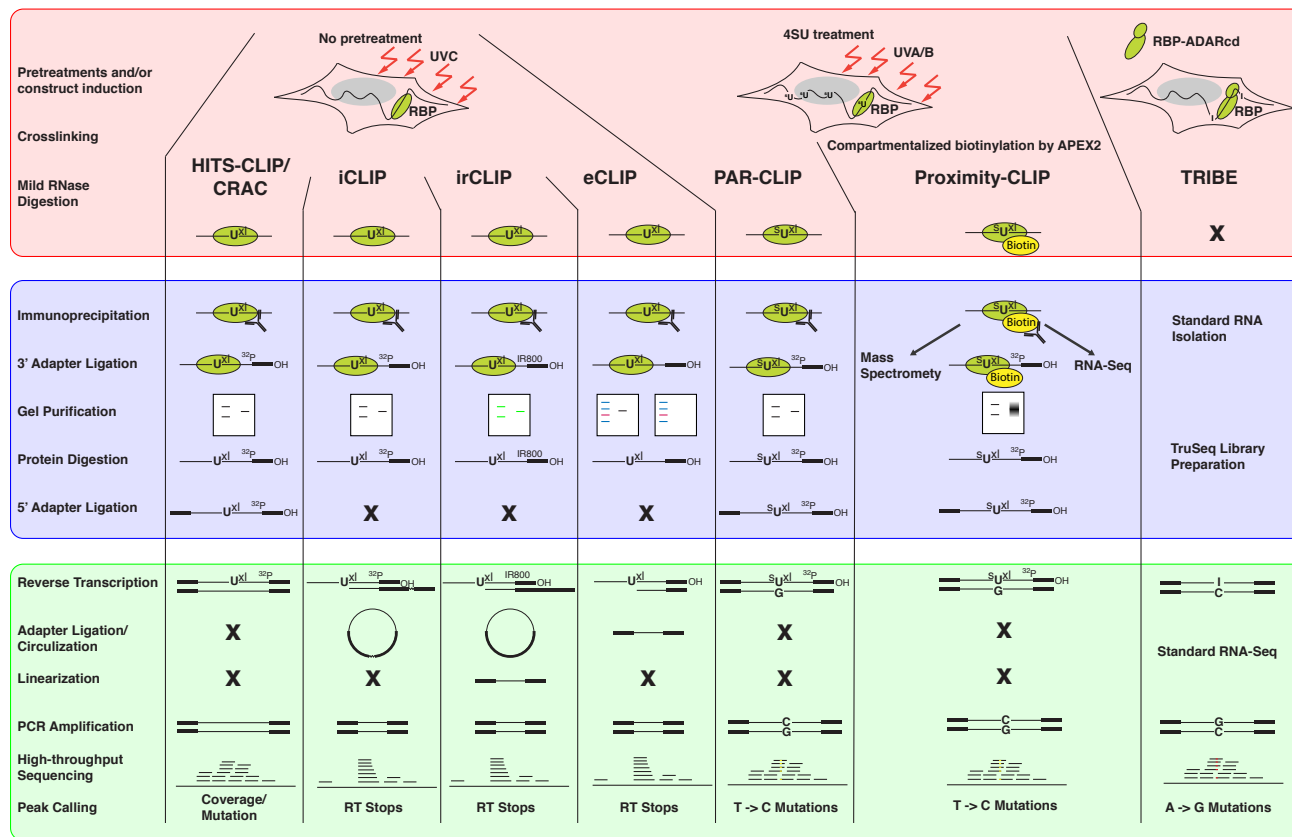
3' 

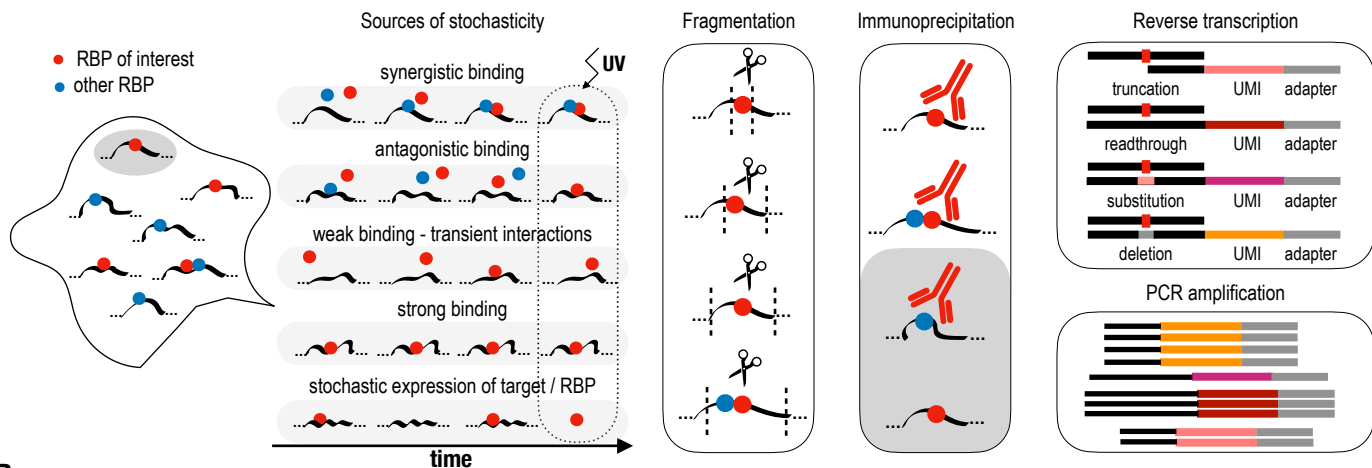
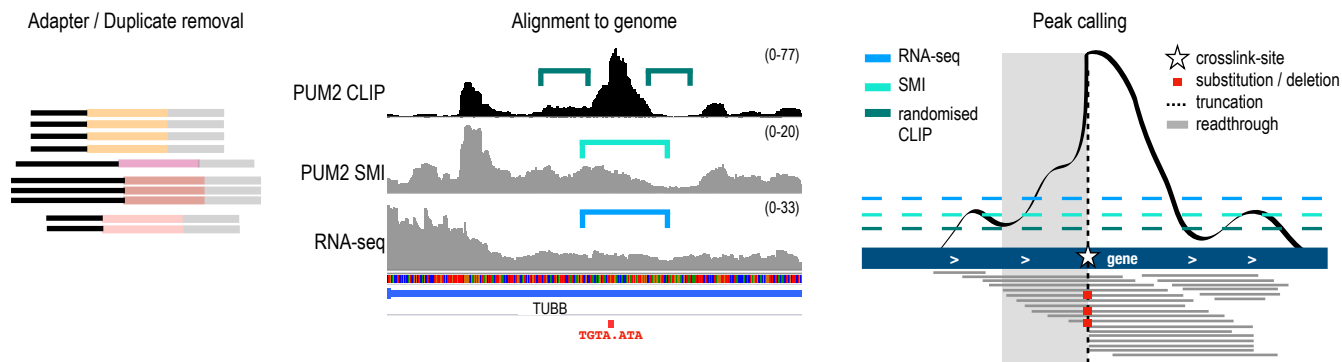
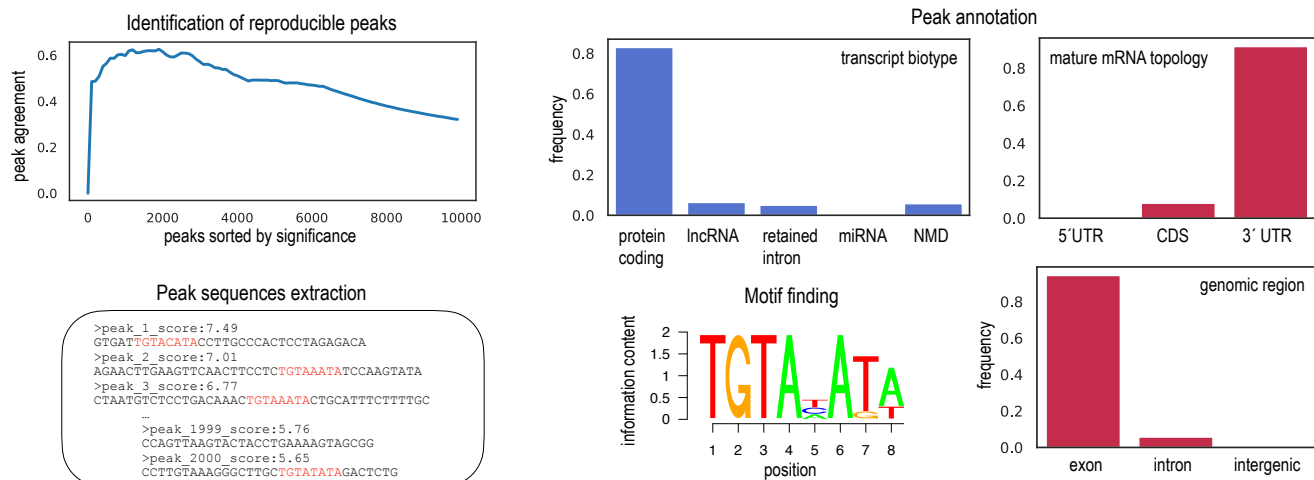
8. PCR and high-throughput sequencing

9. Bioinformatic determination of binding peaks



10. Integration with complementary data



A.**B.****C.**

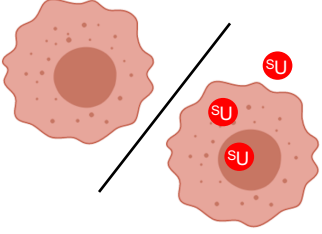



	Cell culture	Tissue	Animal models	Plants
				
Considerations	RBPs and RNAs expressed in a cell type specific manner are not recovered	Thick tissues require mechanical dissection	Thick tissues require mechanical dissection	Harsh denaturing required for plant cell walls
	Exogenous expression system	Heterogenous samples; multiple cell types can complicate data interpretation Degradation of RNP complexes Limitations to deliver modified nucleotides	Increased difficulty in generating mutant strains/tagging Difficulty in delivering modified nucleotides	Higher energy UV required for crosslinking in the presence of UV absorbing pigments and to reach inner cell layers Heterogenous samples; multiple cell types can complicate data interpretation
Unique Advantages	Monolayer allows for optimized UV crosslinking Ready uptake of nucleoside analogs Multiple RBPs can be compared using the same cell line/transcriptome background Rapid development of transgenic cell lines for CLIP/TRIBE	CLIP analysis in health and disease Identifies physiological interactions in the whole spectrum of cells and tissues	cTAG-CLIP for identifying cell type-specific binding patterns Detect cell-type specific regulatory events in less abundant cell types	Epitope tagged RBPs in loss-of-function mutants to mimic endogenous expression pattern

Figure 6 CLIP applications in model organisms