Sensitive detection and structural characterisation of UV-induced cross-links in protein-RNA complexes using CLIR-MS

Chris P. Sarnowski[1,2]; Anna Knörlein[3#]; Tebbe de Vries[4#]; Michael Götze[1,5]; Irene Beusch[4,6]; Ruedi Aebersold[1,7]; Frédéric H.-T. Allain[4]; Jonathan Hall[3]; Alexander Leitner[1,*]

[1] Institute of Molecular Systems Biology, Department of Biology, ETH Zürich, Zurich, Switzerland

[2] Systems Biology PhD Program, University of Zürich and ETH Zürich, Zurich, Switzerland

[3] Institute of Pharmaceutical Sciences, Department of Chemistry and Applied Biosciences, ETH Zürich, Zurich, Switzerland

[4] Institute of Biochemistry, Department of Biology, ETH Zürich, Zurich, Switzerland

[5] Current address: Department of Biology, Chemistry, Pharmacy, Institute of Chemistry and Biochemistry, Freie Universität Berlin, Berlin, Germany

[6] Current address: Department of Biochemistry and Biophysics, University of California, San Francisco (UCSF), San Francisco, CA, USA

[7] Faculty of Science, University of Zürich, Zurich, Switzerland

[#] These authors contributed equally.

* Corresponding author (Email: leitner@imsb.biol.ethz.ch).

Abstract

Cross-linking coupled with mass spectrometry is an increasingly popular methodology for elucidating structural information from biological complexes. Whilst protein-protein cross-linking workflows are widely used and well characterised, adoption of protein-RNA cross-linking workflows for structural studies is less widespread, and data produced from such experiments remains less well understood. The cross-linking of stable isotope labelled RNA coupled to mass spectrometry (CLIR-MS) workflow uses isotope labelled RNA to simultaneously confirm that peptides are cross-linked to RNA and aid cross-link localisation in an RNA sequence. For broader application of CLIR-MS as part of the structural analysis of ribonucleoproteins, the method must be sensitive, robust, and its reaction products need to be well characterised. We enhanced our previously published workflow, improving coverage and sensitivity. We used it to infer common properties of protein-RNA cross-links such as cross-linking distance, and to assess the impact of substitution of uracil with 4-thio-uracil in structural proteomics experiments. We profiled the compositional diversity of RNA-derived peptide modifications, and subsequently defined a more inclusive data analysis approach which more than doubles the number of cross-link spectrum matches compared with our past work. We defined distance restraints from these cross-links, and with the aid of visualisation software, demonstrated that on their own they provide sufficient information to localise an RNA chain to the correct position on the surface of a protein. We applied our enhanced workflow and understanding to characterise the binding interface of several protein-RNA complexes containing classical and uncommon RNA binding domains. The enhanced sensitivity and understanding demonstrated here underpin a wider adoption of protein-RNA cross-linking in structural biology.

Introduction

Proteins and ribonucleic acids (RNAs) form functional units that are fundamental for the survival of a cell.[1–4] Understanding the structural nature of vital protein-RNA assemblies contributes to comprehensive explanation of how such complexes function. The nature of RNA recognition by a set of common protein structure motifs, such as RNA recognition motifs[5] (RRMs), zinc fingers[6] (ZnFs), and DEAD box helicase domains[7], is generally well understood. However, structural studies of critical protein-RNA assemblies such as the ribosome[8,9] and spliceosome[10,11], as well as proteome-wide RNA-binding protein studies[12,13], reveal a plethora of proteins that bind RNA but lack a canonical RNA binding domain. Technologies that enable the study of the structural nature of such interactions therefore help understanding the functions of these newly discovered complexes. As with RNA, proteins also form functional complexes with other proteins. Cross-linking coupled with mass spectrometry (XL-MS) is a widely used technique in structural biology to describe how protein complexes are assembled, therefore helping understand their functions. The protein-protein XL-MS technique establishes spatial proximity between non-adjacent amino acids, thus providing information on which parts of two or more interacting proteins are in contact with one another, or even how a single protein is folded[14–17]. Despite the prevalence and importance of protein-RNA interactions, analogous XL-MS technology to study the structural and spatial nature of these interactions remains less mature than approaches for protein-protein interactions.

A number of features of a typical protein-protein XL-MS workflow[18] contribute to the practicality and prevalence of the technique. Firstly, protein samples are most commonly cross-linked using a chemical reagent with a known spacer length and amino acid specificity. These inform the distance represented by a cross-link, and aid localisation of the cross-link to precise residues. Additionally, cross-linked peptides may be enriched for using techniques such as size-exclusion chromatography (SEC)[19] or immobilised metal affinity chromatography (IMAC)[20]. Enrichment compensates to some extent low cross-linking reaction yields, thus increasing analytical coverage of cross-linked species. One of many specialised software tools[21] can then be used to interpret

resulting mass spectra, most commonly by comparison with an *in silico* digested reference database of cross-linked peptides. Within each peptide, the precise cross-linked amino acid can be identified thanks to predictable peptide fragmentation during tandem mass spectrometry (MS/MS) analysis, and amino acid specific chemistry. The result of such a workflow is a set of pairs of amino acids that must reside in close proximity with one another in the 3D structure of the folded protein(s), within a certain distance determined by the length of the cross-linker molecule. These point-to-point distances can be specified as restraints in structural models, either as the sole source of experimental data[22], or in combination with data from complementary structural techniques[23,24], and provide a relatively fast way of studying protein structures in the solution state.

For protein-RNA cross-linking data to be utilised in a similar fashion, precise sites of cross-link attachment should be identifiable on both the protein and the RNA at or close to single amino acid- or nucleotide-resolution, and the distance represented by a cross-link should be well characterised. A variety of experimental techniques have been developed that exploit UV cross-linking of proteins to nucleic acids[25], providing information about the cross-link at varying structural resolutions (from whole molecule to single residue). Whilst cross-linking and immunoprecipitation (CLIP)[26–30] and related techniques provide localisation of the binding site of a given protein on an RNA at high-resolution, the site of cross-linking on the protein remains localised to a specific domain at best, providing insufficient detail for use as a restraint for structural modelling. Alternatively, previously published MS based methods successfully employ UV cross-linking to localise RNA contact sites on proteins to single amino acid positions[31–33] but offer limited resolution on the RNA side of the interaction. Whilst such an approach can confirm the amino acids mediating interaction with RNA, it is of limited value in positioning the RNA sequence in relation to the protein, and cannot be used as a true distance restraint in a structural modelling pipeline.

Our previously reported cross-linking of isotope-labelled RNA (CLIR-MS) workflow[34] builds on previous MS-based RNA-cross-linking techniques[31,35.]. It additionally localises the RNA cross-

linking site, employing stable isotope labelling of selected RNA segments to achieve up-to single nucleotide resolution whilst simultaneously assigning the cross-link to a single amino acid on the interacting protein in a single experiment. Given the positional specificity afforded by this approach, a UV-induced protein-RNA cross-link from a CLIR-MS experiment with position-specific RNA labelling more closely represents an ideal distance restraint in an analogous fashion to above-mentioned applications of protein-protein cross-links.

Despite the promise shown by the CLIR-MS pipeline and related mass spectrometry-based techniques to study protein-RNA interactions, several challenges have remained towards routine application of the method. The RNA-derived products of the UV cross-linking reaction remain relatively poorly characterised, and little is known about the details of the reaction mechanism[36]. The spatial/geometric properties reflected in a cross-link in the context of a protein-RNA complex have not been systematically characterised, leaving previous interpretations of protein-RNA cross-linking distances reliant on the assumption that such cross-links are 'zero-distance'[29], or on inferences from putative mechanisms for such reactions proposed in older literature[37]. Systematic study of cross-linking distance would therefore aid more accurate interpretation of protein-RNA cross-links. Furthermore, sensitivity is a critical factor to consider when analysing protein-RNA cross-links with MS, because the yield of protein-nucleic acid UV cross-linking reactions is low. Often, less than 10% of starting material is converted to cross-linked complex using conventional conditions,[38] and only a very small fraction of this material – peptides at the RNA binding interface – is eventually analysed. Improvements to sample preparation, data acquisition, and incorporation of a full set of well-characterised reaction products could all increase overall sensitivity of the CLIR-MS technique.

In this work we address several of the inherent limitations of the original CLIR-MS workflow. Specifically, we describe enhancements to sample preparation, encompassing protein-RNA adduct enrichment and in mass spectrometric instrumentation and data acquisition which collectively increase the numbers of cross-links identified from a given protein-RNA sample.

Additionally, we take advantage of stable isotope labelling of RNA in CLIR-MS to more thoroughly characterise and validate the RNA-derived adducts to peptides, further enhancing the number of cross-links identified by the specialised data analysis approach in a CLIR-MS experiment. Together, these advances help overcome the sensitivity challenges. We then apply the optimised sample preparation and data analysis pipeline to a varied set of model RNA binding proteins (PTBP1, FOX1, and MBNL1) in complex with their cognate RNAs, and compare cross-link identifications with published structures. We systematically assess the distance over which UV-induced protein-RNA cross-links form, and characterise distinct behaviours of canonical uracil with 4-thio-uracil. Using these data, we show that CLIR-MS derived cross-links alone contain sufficient information to describe the occupancy space of an RNA relative to a protein in complex. Finally, we study the binding interface of a non-canonical RNA binding domain, the ubiquitin-like domain of the U2 snRNP protein SF3A1 to stem-loop 4 of the U1 snRNA using CLIR-MS restraints. The work highlights the value of protein-RNA cross-link data as an independent data type, in addition to its previously demonstrated value as part of an integrative modelling pipeline when combined with high-resolution data from NMR spectroscopy[34].

Results

Optimisation of sample preparation and data analysis

The CLIR-MS sample preparation protocol (**Fig. 1a**) comprises cross-linking of protein-RNA complexes by UV irradiation, digestion of complexes with nucleases and proteases, enrichment for cross-linked peptide-RNA adducts with titanium oxide metal oxide affinity chromatography (MOAC), sample clean-up using C_{18} solid phase extraction (SPE), and analysis by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS)[34]. To improve the sensitivity of the CLIR-MS method, we systematically optimised these steps to improve recovery and detection of protein-RNA adducts. We assumed that this would result in both increased total numbers of cross-link spectrum matches (XLSMs) during data analysis, as well as an increased number of unique peptide-RNA identifications from these XLSMs. A unique identification is here defined as a combination of a cross-linked amino acid position with a unique RNA sequence composition cross-linked to that amino acid position. Unless otherwise stated, results refer to changes in the total number of XLSMs from an experiment.

Changes to sample preparation prior to analysis

We selected the Polypyrimidine Tract Binding Protein 1 (PTBP1) in complex with the internal ribosome entry site (IRES) of the encephalomyocarditis virus (EMCV) as a model for sample preparation, because the protein contains multiple RRM domains, and the RNA is relatively long (88 nucleotides). Together, this results in a large number of amino acid cross-linking sites on the protein, comprising a variety of RNA species linked to them, as previously reported[34]. The complex therefore provides an adequate complexity for generalised sample preparation optimisation measures which may be transferred to other complexes. The previously published CLIR-MS sample preparation and analysis procedure[34] was used as a baseline for protocol optimisations.

We first tested the impact of miniaturising the C_{18} SPE step (**Fig. 1a** step 3) on the number of XLSMs made from a PTBP1-IRES CLIR-MS sample. Prior literature suggests that in conventional

proteomics experiments, stop-and-go extraction (Stage) tips[39] provide superior sample recovery for low sample amounts compared with larger C_{18} cartridges that were used in the original CLIR-MS protocol. Given the low expected sample amount after MOAC enrichment, the improvement in sample recovery may also be expected in a CLIR-MS experiment. We prepared the PTBP1-IRES CLIR-MS sample as described previously[34], and additionally with the same protocol but with SPE cartridges replaced with Stage tips. We then analysed both samples with LC-MS/MS using identical acquisition parameters and compared the number of XLSMs produced by an xQuest search[40] (**Fig. 1c**). When using Stage tips, the number of XLSMs identified from each replicate increased by approximately 33% (from 129 to 172, mean of both replicates) compared with larger SPE cartridges. This suggests that miniaturising the SPE step, and thus minimising the surface area of vessels in contact with samples, improves recovery of peptide-RNA adducts prior to LC-MS/MS analysis.

Changes to mass spectrometry data acquisition

Next, we investigated whether changes to MS acquisition parameters could increase the number of XLSMs provided by a CLIR-MS experiment (**Fig. 1a** step 4). Multiple activation methods for MS/MS peptide sequencing are available on modern instruments, some of which may provide superior fragmentation, or preservation of post-translational modifications (PTMs)[41]. Assuming that a peptide cross-linked to a short piece of RNA (one to four nucleotides) behaves like a peptide with a PTM during MS analysis, the choice of fragmentation method may therefore influence the number of identified XLSMs. To test this, the aliquots of the same PTBP1-IRES sample were injected multiple times and analysed with three different activation methods, (ion trap) collision induced dissociation (CID), electron transfer/higher-energy collision-induced dissociation (EThcD) and higher-energy collision-induced dissociation (HCD). The numbers of XLSMs returned in each case were compared (**Fig. 1d**). Injections measured with the EThcD acquisition method produced an average of just 32 identifications per injection (mean of both replicates), representing a decrease on the CID method used previously for CLIR-MS. HCD appeared to perform best, with

an average of 118 identifications per MS run (mean of both injections), representing an increase of 74% compared with CID (mean of 68 identifications per injection). Importantly, HCD also appeared better suited to detection of longer RNA adducts, returning a greater number of di- and trinucleotide containing XLSMs than CID or EThcD (**Fig. 1d**). These longer RNA adducts are essential in assigning RNA localisation, given the small number of building blocks (four nucleobases) and the fact that the short oligonucleotide is not directly sequenced in our MS method. Based on these observations, we selected HCD as the default fragmentation method for LC-MS/MS analysis for CLIR-MS.

Evaluating the robustness of the CLIR-MS protocol

To test whether these optimisations could be generalised, we compared our original and revised protocols on a different protein-RNA complex, the FOX1 RRM with its cognate RNA, the FOX Binding Element (5'-UGCAUGU-3', FBE). FOX1 regulates eukaryotic RNA splicing by binding the RNA sequence UGCAUG[42], and is structurally well characterised (PDB ID: 2ERR) using solution state NMR spectroscopy[43], meaning identified cross-links could be validated against a known structure. The number of XLSMs increased from less than 60 per replicate with the previous protocol to over 200 per replicate with the enhanced protocol (mean of both replicates, **Fig. 1e**), suggesting that optimisations made using the PTBP1-IRES complex have broad applicability. The identifications from the conditions with the greatest number of identified XLSMs for each complex are plotted in (**Fig. 1f and Fig. 1b**) for the FOX1-FBE and PTBP1-IRES complexes, respectively. The PTBP1 protein exhibits several sites that cross-link to RNA, owing to its relatively large size and multiple RRMs. Conversely, the FOX1 RRM exhibits only two major protein sites cross-linking to RNA, surrounding phenylalanine residues at positions 126 and 160 which are consistent with the published structure[43].

In summary, we demonstrate that miniaturisation of sample clean-up and optimisation of LC-MS/MS fragmentation conditions contributed to an increase in the number of XLSMs made in a CLIR-MS experiment and that these enhancements are not specific to a single complex.

Increasing the number of XLSMs with a deeper understanding of possible UV-XL products

Next, we attempted a more thorough characterisation of UV-induced RNA-derived peptide modifications found in a CLIR-MS sample, to investigate whether a broader set of reaction products could be routinely incorporated into the data search strategy (**Fig. 1a** step 5). Stable-isotope labelled RNA is primarily employed to localise cross-links to specific nucleotides in the RNA – those which are present in both light and heavy form in the sample result in a characteristic mass shift. Additionally, the presence of these mass-shifted species confirms that peptide modifications truly derive from RNA, rather than any potential side product of the UV cross-linking reaction. In the present work, two labelling schemes, either with *in vitro* transcribed RNA incorporating ^{13}C and ^{15}N, or chemical synthesis of RNA using nucleotides containing ^{13}C ribose, were used (further details in **Methods**). Many different RNA-derived products have been described in previous peptide-centric mass spectrometric analyses of UV-induced protein-RNA cross-links[31,33,35,44,45]. These suggest that a variety of neutral losses, including those corresponding to atomic compositions such as $-H_2$, $-H_2O$, $-HPO_3$, occur from the peptide-RNA adduct during cross-linking, sample preparation, or mass spectrometric analysis.

To establish a comprehensive set of detectable RNA-derived peptide modifications, the FOX1-FBE complex was selected. With only a single RRM, it yields a less complex sample with a smaller number of peptides cross-linked to RNA which may improve detection of low abundant RNA-derived products during mass spectrometric analysis. Two FOX1-FBE samples were prepared using the CLIR-MS protocol; one was subjected to 254 nm irradiation to induce cross-links, and the other was left unirradiated as a control. Both samples were analysed by LC-MS/MS, and the data subjected to an “open” modification search using MSFragger[46], to discover all possible mass additions. The detected peptide modifications were grouped into 0.1 Da mass bins, and the number of identifications within each mass bin compared between irradiated and non-irradiated samples (**Fig. 2a and 2b**) and annotated according to previous literature (**Tab. 1** and **Tab. 2**). Many RNA-derived modifications were observed in the UV irradiated condition, forming

clusters corresponding to expected masses of mono-, di-, tri-, and tetranucleotide RNA species cross-linked to peptides (grey shading, **Fig. 2a**). As expected, many of the most common peptide mass additions in the irradiated sample were present in both light and heavy isotopic forms, with a delta mass corresponding to that expected from the respective RNA attachment. Some abundant modifications that are not specific to the crosslinking reaction were present in both samples such as +258 Da, a modification commonly observed as a result of His-tagged *in vitro* expressed protein preparations, as used for this protein[47].

The putative list of loss products was then used to define a “closed” xQuest search (for the UV cross-linked samples only), expanding on the set of products used in the original protocol[34]. In this search approach, both the light and heavy isotope forms of each proposed RNA species must be detected in order to produce an XLSM, thereby validating putative RNA-derived modification types found in the open modification search. The product types that were subjected to confirmation using the xQuest search are shown in **Tab. 1**. Putative identifications from the open search that could not be validated are described in **Tab. 2**. Based on the confirmatory results of the xQuest search, we routinely incorporated the expanded product list of validated loss products from **Tab. 1** into the xQuest search parameters for a CLIR-MS experiment, thereby increasing the number of XLSMs obtained from a CLIR-MS data set compared with the original protocol[34]. A comparison of results from two xQuest searches of the same data (the UV cross-linked samples also used in **Fig. 2a**), one using the restricted set of modifications as previously published[34], and one using the expanded set presented here (validated in **Tab. 1**), is shown in (**Fig. 2c**). Not only did the number of XLSMs increase with the more inclusive search from 313 to 1177 (sum of both replicates), but the quantity of unique amino acid and RNA sequence combinations also increased from 71 to 118 (sum of both replicates). The additional cross-links detected fall around the same amino acid positions on the protein as those detected with the more restricted parameters, suggesting the additional identifications contribute to robustness of results. The expanded cross-link product set therefore builds on the sample preparation and data acquisition optimisations to further increase

the sensitivity of the technique and reduce sample amount requirements. Note that this expanded product list was used to search the data presented in earlier sections describing sample preparation enhancements.

Comparing CLIR-MS data with previously published structural models

Having optimised sample preparation and data analysis to achieve greater numbers of XLSMs from each CLIR-MS experiment, we then set out to better characterise the structural distance represented by a protein-RNA cross-link, to aid more faithful incorporation of restraints in structural models. We applied the technique to a broader set of protein-RNA complexes for which prior published structural models derived from established structural techniques exist, and compared CLIR-MS results for each complex with the respective published structure.

We selected three representative protein-RNA complexes, containing the proteins FOX1, MBNL1 and PTBP1, to provide coverage of different modes of RNA recognition, different protein sizes, and where the RNA sequence recognised in the published structural model consists of just a few nucleotides. The short RNA mimics the behaviour of a section of segmentally labelled RNA in a CLIR-MS experiment. PTBP1 has many diverse roles in RNA metabolism, including regulation of splicing activity[48] and translation initiation[49]. Published structures exist for each of its four RRMs in complex with a short polypyrimidine RNA sequence (PDB IDs: 2AD9, 2ADB, 2ADC), which were produced from solution-state NMR experiments[50].The muscleblind-like (MBNL) proteins also act as splicing regulators, controlling tissue specific alternative splicing by targeting CUG and CCUG RNA repeat sequences[51]. Unlike the PTB proteins, RNA binding is mediated by ZnF domains. An NMR-derived structure also exists for the ZnF 1-2 pair of MBNL1 in complex with a short RNA[52] (PDB ID: 5U9B). CLIR-MS data from the FOX1 protein in complex with the FBE is also included for comparison to an existing structure (PDB ID: 2ERR). Taken together, comparisons of CLIR-MS data from these complexes with their respective structures should facilitate inference of a

generalisable protein-RNA cross-linking distance. We exclusively selected solution-state NMR structures for comparison with CLIR-MS data (where cross-linking takes place in the solution state) to avoid introducing biases of crystallisation conditions into our distance measurements.

Comparison of CLIR-MS results with published structures of model complexes

We used full-length PTBP1 protein, an MBNL1 construct spanning positions 1-269, and the aforementioned FOX1 RRM construct. Short RNA sequences, corresponding to bound RNA sequences in the published structures for each complex, were synthesised. Schematics of each complex are overlaid in **Fig. 3a-c**. The optimised CLIR-MS protocol was applied to each of the complexes, and the cross-links identified are shown in **Fig. 3a-c**.

The optimised sample preparation protocol returned hundreds of cross-link identifications in each complex, further supporting the broad applicability of the CLIR-MS method to a diverse set of RNA binding proteins. In the case of the FOX1 RRM, the majority of cross-linked amino acid sites on the protein (**Fig. 3a**) fell in clusters around two phenylalanine residues (F126 and F160 respectively). These are well explained by the published structure, with the cross-linked amino acids located on the β-sheet surface of the protein expected to recognise the RNA. Furthermore, these aromatic residues are found π-π stacked with nucleotide bases in the structure, suggesting a close interaction between protein and RNA which may be particularly conducive to cross-linking, as noted elsewhere[53].

In the MBNL1 complex (**Fig. 3b**), CLIR-MS analysis suggests that two main groups of amino acids interact with the short RNA (5'-CGCUU-3'), around L69 in ZnF 2 of the 1-2 pair, and around F237 in ZnF 4 of the ZnF 3-4 pair, respectively. This matches published data from MBNL1 interacting with the intronic binding site in human cardiac troponin T pre-mRNA, which demonstrated that only ZnF 2 and ZnF 4 are involved in RNA binding[52]. As with the RRM-based complexes, the cross-linking sites in the ZnF 1-2 are well explained by the published NMR-derived structure of RNA interaction with this pair, falling on the position of the protein surface expected to mediate RNA interaction. Prior comparison of unbound ZnF 1-2 and ZnF 3-4 pairs suggests they are highly

conserved, both in terms of sequence and structure[52]. The highly cross-linked residues detected by CLIR-MS, L69 (ZnF 2) and F237 (ZnF 4) are also observed at equivalent positions within the tandem ZnFs, suggesting similar structural modes of RNA binding between the two ZnF pairs in this sample.

PTBP1 exhibited fewer amino acid positions cross-linked to RNA when in complex with the short RNA (5'-UCUCU-3', **Fig. 3c**), than when in complex with a longer RNA such as the IRES RNA (**Fig. 1b**). This may indicate that in the case of the PTBP1-IRES complex, some amino acid sites actively recognise RNA, whereas others make weak unspecific contacts with RNA in the context of a longer RNA bound across multiple RRMs, without contributing to a selective interaction. The latter interactions will be consequently unidentifiable when a shorter RNA is bound, such as the case in **Fig. 3**. As with the FOX1 complex, the majority of cross-linking sites are well explained by the published structures, falling on the expected exposed β-sheet faces of the proteins across all RRMs, or otherwise nearby on the surface of the protein. The exceptions were in RRM2, where some cross-linking sites were unexpectedly found on the opposite face of the protein than expected. This may originate from a non-specific RNA-protein interaction that may be particularly conducive to UV cross-linking.

In summary, the cross-linking sites identified by the optimised CLIR-MS pipeline for three different model protein-RNA complexes are numerous, and generally well explained by existing structural models produced using solution-state NMR spectroscopy.

Measuring the distance represented by a CLIR-MS protein-RNA cross-link

XL-MS data is frequently used to define distance restraints in structural modelling pipelines, where the distance represented by a cross-link depends on the reaction chemistry. UV-induced protein-RNA cross-links are often called "zero-length" cross-links, however in the absence of either a chemical cross-linking reagent or understanding of the chemical reaction mechanism of UV cross-linking, there is no clear consensus on the distance represented by cross-link formation. Empirical comparisons of CLIR-MS data with published structures therefore provide a strategy to understand

the distance represented by a UV-induced protein-RNA cross-link, which is vital for faithful use of CLIR-MS data in structural applications.

To undertake such a comparison, we filtered the cross-links identified from each complex for mononucleotide adducts, and for each cross-link measured the distance from Cα of the amino acid to N1 (pyrimidines) or N9 (purines) of the nearest matching nucleotide (backbone to backbone) in the respective published NMR-derived structural model ensembles. In structural proteomics, backbone-backbone distances are commonly used for modelling, in absence of known side-chain orientations. The cross-links used in each comparison were annotated on the respective published structures as shown in **Fig. 3d-f**. Furthermore, a control set of distances was generated from each of these structures, also from Cα to N1 or N9, but covering all theoretical pairwise combinations of nucleotides and amino acids present in each structural ensemble. The distributions of measured and control distances were plotted for each structure (**Fig. 3g-i**). In each case, experimentally detected cross-link distances form a distribution centred on a shorter distance than, and clearly separated from the theoretical control distances, demonstrating the specificity of the cross-links. The mean protein-RNA cross-linking distances for each sample were 9.7 Å, 10.9 Å and 12.1 Å for FOX1, PTBP1 and MBNL1 complexes respectively, with an upper limit of around 20 Å.

A small secondary distribution of cross-link distances greater than 20 Å was observed in the PTBP1 RRM comparison distribution; all of these values derive from cross-links in RRM2 which are not so well explained by the structure, as mentioned above, and may therefore be an artefact of comparison of CLIR-MS results from a full-length protein with isolated RRM models, rather than true cross-linking distances. Whilst observed distances were broadly consistent between the different complexes, the ZnF-mediated RNA binding of MBNL1 appeared to exhibit slightly longer distances than the RRM mediated binding of PTBP1 and FOX1. This could be explained by differing amino acid compositions of ZnFs and RRMs. The set of all cross-linked amino acid types found in the MBNL1 complex tend to have longer side chains, such as tyrosine, tryptophan, and phenylalanine (although L69 was identified as cross-linked in the most XLSMs). Whilst these were

also found cross-linked in the PTBP1 and FOX1 complexes, cross-links with amino acids bearing shorter side chains such as glycine, serine, threonine, and proline were also observed, which may explain the slight shift in distance distributions.

From these comparisons of CLIR-MS derived cross-links with prior structural models, we conclude that the cross-links detected in a CLIR-MS experiment are highly specific to their structural context, and represent a mean proximity of respective peptide and RNA backbones of around 10-12 Å.

The utility of CLIR-MS derived cross-links as an independent structural data type

Protein-protein XL-MS data are frequently employed as a standalone data type for low-resolution placement of proteins relative to one another in a complex[54,55]. This is possible because the cross-link is precisely localised to a single amino acid position on both peptides, and the distance represented by the cross-link is known thanks to well characterised reaction chemistry. Cross-links yielded by the CLIR-MS workflow may be precisely localised on both the protein (by peptide fragmentation and MS/MS) and RNA sequences (by selective isotope labelling and overlay of oligonucleotide adducts found at the same amino acid position) of the complex. Furthermore, the structural comparison described above revealed the distances represented by these protein-RNA cross-links. Taken together, precisely localised CLIR-MS cross-links should therefore contain sufficient structural information to tether an RNA to the correct position on the surface of a protein, providing a low-resolution description of how the two molecules interact. To evaluate this use case, we used DisVis[56,57] to visualise the accessible interaction space of RNA relative to its corresponding protein in a complex, as constrained by CLIR-MS cross-linking data.

We separated the protein and RNA chains of the published structural models shown in **Fig. 3d-f**, and collated a list comprising only mononucleotide cross-links that were identified for each complex in the CLIR-MS experiments shown in **Fig. 3a-c**. The RNA position associated with each mononucleotide was assumed to be the nearest nucleotide in the published structures, as measured in **Fig. 3g-i**. These cross-links were specified as restraints with distances from 0 to 12

Å, in line with the upper quartile of all measured distances (from Cα to N1 for pyrimidines or N9 for purines) observed in **Fig. 3g-i**. We then submitted the components to DisVis[56] for occupancy analysis, with protein as the fixed chain and RNA specified as the scanning chain. The outputs are shown in **Fig. 3j-l**, where grey shading represents spatial occupancy of the RNA chain relative to the protein (displayed as centre-of-mass of the RNA), given the specified cross-links. For all protein-RNA complexes tested, the compatible positioning of the RNA relative to the protein derived from the CLIR-MS cross-links closely resembled the placement of the RNA in published structural models of each of these complexes (**Fig. 3d-f**).. In each of these relatively small model complexes, RNA contact sites fall close together on the surface of the protein. Together with the short RNA sequences, it is here more challenging to precisely determine RNA orientation due to a large degree of rotational freedom for the RNA. Nonetheless, this proof of concept on well-studied complexes suggests that given a short linear RNA and a solved unbound protein structure, CLIR-MS derived cross-links alone contain sufficient information, when combined with our empirically derived cross-linking distance, to accurately identify the occupancy space of a linear, non-structured RNA relative to a protein in a complex.

Comparing cross-linking of 4-thio-uracil with uracil in a CLIR-MS experiment

Due to the low reaction yield of the protein-RNA cross-linking reaction, many experimental workflows that rely on UV cross-linking of protein to RNA substitute uracil for 4-thio-uracil (4SU) to increase the proportion of protein-RNA complex that is cross-linked[35,45,58]. Given the reliance of CLIR-MS on a UV cross-linking reaction between protein and RNA, substitution of uracil with 4SU could also be used here to increase the reaction yield, and hence detectability of cross-links by the pipeline. Production of RNA by solid phase synthesis facilitates position-specific incorporation of chemically modified nucleotides such as 4SU, meaning the cross-linking behaviour of 4SU at a specified nucleotide position can be evaluated. Three separate FOX1-FBE samples were prepared, each with one of the three uracil positions in the FBE RNA heptanucleotide replaced

with 4SU (schematics overlaid in **Fig. 4a**). Samples were irradiated with 365 nm UV light, ensuring that the cross-links formed resulted only from the substituted base, given that only 4SU reacts at this wavelength. Samples were then analysed using the optimised CLIR-MS workflow and identified cross-links are shown in (**Fig. 4a**). As expected, the numbers of XLSMs are relatively high, compared with a similar sample mass using natural nucleotides shown in previous figures, and especially so considering all cross-links derive from a single nucleotide position. Most cross-linking involved positions U1 and U7. According to the published structure, U1 exhibits some conformational heterogeneity in its binding, and U7 is not held rigidly in place by specific hydrogen bonds[43]. Position U5, which the published structure indicates is firmly held in place by hydrogen bonds to multiple amino acid residues[43], did not cross-link so strongly. The major cross-linked protein sites differed from those obtained in samples containing only natural nucleotides, with a loss of cross-links around amino acid position 160 when any of the uracil positions were replaced with 4SU, and a gain of cross-links surrounding N151. However, these amino acid positions are still close to RNA in the published structure. These results indicate that 4SU cross-linking activity may be distinct from that of natural uracil.

We then compared identified 4SU-derived cross-links with the published FOX1-FBE structure ensemble, with distances once again measured from Cα of the amino acid to N1 of 4SU. The distribution of observed distances and the control distance set containing all possible amino acid and nucleotide pairs in the complex are shown in **Fig. 4b**. The distribution of 4SU cross-links is less well resolved from the control set of distances compared with the natural uracil cross-links shown in **Fig. 3g**, with a median distance of around 18 Å for both the measured and control distances. This suggests that the distance represented by a 4SU-derived cross-link may be longer than for a natural nucleotide. However, definition of an upper bound distance based on these data is not appropriate, given the relatively small size of this complex and the relatively long median observed distance. From these data, we conclude that the structural meaning of a 4SU derived UV

cross-link may be distinct from that of a natural nucleobase, a factor that must be considered if using 4SU in structural applications.

Characterising a non-canonical protein-RNA interaction using only CLIR-MS restraints

The data shown so far demonstrate that the optimised CLIR-MS sample preparation and data analysis steps provide larger numbers of protein-RNA cross-link identifications than the original protocol, that CLIR-MS derived cross-links represent proximity of around 10-20 Å between protein and RNA backbones, and that a set of CLIR-MS derived cross-links provides sufficient information to describe the spatial arrangement of a protein and an RNA in complex. We then exploited this improved pipeline to study a non-canonical protein-RNA interaction type between the ubiquitin-like domain of U2 snRNP protein component SF3A1 and stem-loop 4 (SL4) of the U1 snRNA. The protein-RNA complexes of the cellular splicing machinery are essential for regulating gene expression. Interaction of SL4 with SF3A1 was previously observed during formation of pre-spliceosomal complexes[59]. Further characterisation of this interaction revealed that the ubiquitin-like (UBL) domain found near the C-terminus of SF3A1 mediates the interaction with U1 snRNA SL4[60]. As well as being functionally important for splicing, the SF3A1-SL4 interaction is also structurally significant, given that the UBL domain is not considered to be a canonical RNA-binding domain[1]. Published structures exist for the U1 snRNP and the SF3A1 protein components in their unbound states, but the structural basis of the SF3A1-UBL interaction with SL4 RNA has to date remained poorly understood. Using CLIR-MS, we aimed to generate a set of protein-RNA cross-links which identify the key amino acid residues mediating the RNA interaction, and which nucleotides in SL4 they interact with. With these cross-links, we aimed to describe the low-resolution spatial arrangement of the complex.

SF3A1-UBL protein and SL4 RNA were reconstituted *in vitro* with a 50:50 mixture of RNA with natural isotopic abundance and stable isotope labelled RNA, respectively; a schematic representation is shown overlaid in **Fig. 5a**. After UV cross-linking, samples were prepared using the optimised CLIR-MS workflow. Identified cross-links are shown in **Fig. 5a**. The results highlight two major cross-linking regions in the protein sequence responsible for recognition of the RNA, corresponding to clusters at Q715-K717 (in the β1-β2 loop, UniProt numbering) and at E760-F763 (around strands β3 and β4). For every cross-linked amino acid position detected, unique ribonucleotide compositions at that position were identified, and systematically overlaid. Results were plotted as a heat map, revealing probable RNA sequence positions with which the respective amino acid sites interact (**Fig. 5b**). The analysis revealed that the (G)UUCG(C) terminal loop, inferred from a published model containing SL4[61], cross-links with amino acids around the E760 protein site (highlighted with red box, **Fig. 5b**). However, based on this analysis alone, ambiguity remained as to which nucleotides cross-link with the amino acids around Q715.

To reduce the ambiguity, we conducted an occupancy analysis with DisVis, to establish a subset of mutually compatible cross-links. We used the highest scoring 5% of protein-RNA contact site position pairs from the heat map in **Fig. 5b** as distance restraints. Restraints were specified from Cα of amino acids to N1 (pyrimidines) or N9 (purines) of the nucleotide, with permitted distances from 0 Å to 12 Å (the upper quartile value of all observed distances in **Fig. 3g-i**). We used the free protein structure[62] (PDB ID: 1ZKH) as the fixed chain, and the structure of SL4 (subset from PDB ID: 6QX9) as the scanning chain[61]. The unbound structures are shown in **Fig. 5c**, with amino acid positions found most frequently cross-linked to RNA coloured in orange. A permitted occupation space of the RNA relative to the protein was calculated with DisVis based upon a subset of the ambiguous distance restraints (**Fig. 5d**). DisVis assigns a z-score to each specified restraint to determine which of the tested restraints were most frequently violated during the occupancy analysis. The cross-links with the most favourable (lowest) z-scores are located between the nucleotides at the top of the RNA stem loop structure, and amino acids E760-F763 in the protein,

consistent with the heat map analysis (**Fig. 5b**). Based on these analyses, the UUCG tetraloop contacts the surface of the protein near E760-F763. The other major cross-linked amino acid site suggests that the lower part of the stem loop is then tethered to the protein around Q715-K717. Unlike with the model complexes in **Fig. 3j-l**, the contact sites between protein and RNA are more spatially separated on the protein surface, meaning that there is directionality to the permitted occupancy space of the RNA. The CLIR-MS distance restraints are therefore in this case sufficient to describe both the position and the orientation of a rigid stem loop RNA structure in relation to a ubiquitin-like protein domain.

A high-resolution 3D structure of the interaction between SF3A1-UBL and SL4 of the U1 snRNA (**Fig. 5e**) was determined separately, using X-ray crystallography and validated using CLIR-MS, solution state NMR spectroscopy and functional assays (described separately[63]). The structure confirms that the position of the RNA relative to the protein as achieved using CLIR-MS restraints only is similar to the bona fide high resolution structure. The crystal structure also confirmed contacts of amino acid residues around E760-F763 (around strands β3 and β4) with the top of the stem loop of the RNA, with F763 stacking on cytosine of UUCG tetraloop and K765 forming a salt bridge to the phosphate backbone of U1-SL4 at the terminal loop. The crystal structure revealed an interaction between the C-terminal residues (RGGR motif) with the major groove of the RNA, however the CLIR-MS analysis, conducted using the same protein construct, yielded no cross-links in this region (around amino position 790). This is likely due to inherent incompatibility of the C-terminal sequence, RGGRKK, with trypsin digestion and analysis by LC-MS/MS. These positions remained undetectable even in LC-MS/MS analysis of non-cross-linked protein subjected to shotgun proteomics analysis. Importantly, cross-links detected by CLIR-MS (**Fig. 5a**) indicate a close proximity of the β1–β2 loop to the RNA. This interaction is less apparent in the crystal structure (**Fig. 5e**), but was in agreement with NMR chemical perturbations and by mutative functional assay, which demonstrate that K717 forms a salt bridge with U1-SL4[63].

Here, when used as an independent data type, CLIR-MS cross-links result in a low-resolution characterisation of a novel non-canonical protein-RNA interaction. Characterisations carried out using this methodology therefore represent a reliable starting point for more sophisticated, atomic-scale integrative structural modelling workflows[64].

Discussion

The enhancements to the CLIR-MS protocol presented here provided greater coverage of the cross-link species created in a sample after UV irradiation, taking the form of a distribution of RNA modifications over a set of consecutive amino acids, rather than at a single amino acid position. Since the first application of CLIR-MS, a newer generation of more sensitive mass spectrometers have become available which additionally contribute to increased sensitivity of the workflow. The increased density of cross-links now identified in each sample builds confidence in a detected protein-RNA interaction site, increasing the standalone value of CLIR-MS data. The insights gained here through application of the technique to well-studied complexes shed light on the properties of protein-RNA cross-links and their use in structural biology.

The variety of RNA-derived peptide modifications observed is rather striking. The analysis approach used here considers the biological information contained in a cross-link (i.e. protein proximity to a given nucleotide sequence) constant between neutral loss products (**Tab. 1**) with the same RNA sequence. Whilst the biological information contained in different neutral loss products with the same sequence composition is equal, this diversity may have implications for the analytical workflow. Despite the enrichment step, peptide-RNA adducts are often present in very low in abundance in the final LC-MS/MS sample, close to the limits of detection. If the variety of RNA-derived adducts is a result of sample preparation, further optimisations could be considered to reduce the number of adduct types produced. For example, different combinations of nucleases may leave distinct RNA adduct types; of the nucleases used here, RNases A[65] and T1[66] yield a 5’ hydroxy product, but leave a 2'-3' cyclic phosphate or a 2' or 3' phosphate attached (which may explain the observed -H_2O loss). However, benzonase leaves the phosphate attached[67], as noted previously[35]. Such technical stratification may unnecessarily reduce signal intensity. Alternative approaches using chemical cleavage of RNA in comparable experimental setups have been recently demonstrated[53,68], which may reduce the variety of RNA product types, but at the same time result in near-exclusively mononucleotide attachments to peptides. The reduced proportion of

polynucleotide adducts resulting from this approach may however make it more challenging to precisely assign the cross-linking site on an RNA sequence, because the data lack the required sequence context on the RNA side.

To be useful as distance restraints, protein-RNA cross-links should be of a well-defined length and localised both to a single amino acid on the peptide, and to a single nucleotide on the RNA. In experiments shown here, localisation on a peptide was achieved by MS/MS, like in a conventional MS-mediated proteomics experiment. For the RNA side, short stretches of segmentally labelled RNA used in a CLIR-MS experiment narrow the cross-linked ribonucleotides to those within the labelled sequence, as isotope pairing is a requirement to produce an identification. Some further analysis is however required to refine the position to a single nucleotide. As shown previously[34], this may be achieved by overlaying detected mono-, di-, and trinucleotide species. This approach was applied systematically in **Fig. 5b**, such that probable sites of RNA interaction are computed for every cross-linked amino acid position. The rich variety of polynucleotide RNA sequence compositions found linked to a particular peptide therefore together contain the information to localise the cross-link at up to single nucleotide resolution. We therefore consider them beneficial enough in structural studies to select nuclease digestion over alternative chemical RNA degradation approaches[68]. Remaining ambiguity (i.e. in the case of labelled segments with highly redundant sequences) may be technically overcome by shortening the segment of labelled RNA, with the maximum resolution being a single labelled nucleotide position. Selecting a set of nucleases which produce a uniform RNA product may however be a worthwhile enhancement, representing a practical compromise between reducing sample complexity to improve signal strength of low abundant species, whilst also maintaining the information content of polynucleotide RNA adducts.

Comparing published structures with protein-RNA cross-linking data from CLIR-MS experiments demonstrated UV-induced protein-RNA cross-links consistently form over a distance of 10-12Å (measured backbone to backbone), even between different types of RNA binding domain. A clear

understanding of this parameter is vital for structural interpretation, if cross-linking data is to be used to specify restraints that reflect true proximity. The distance measurements shown here appear to agree with the chemical structures of cross-linking products proposed in prior literature, which suggests a mechanism for UV-induced protein-RNA cross-linking[36,37]. The chemical mechanism of the reaction remains relatively poorly characterised, and further research in this area (see reference[53]) will enhance interpretation and utilisation of these data.

The consistency in unique structural information (i.e. combinations of amino acid sites with an attached RNA sequence) obtained over different irradiation energies suggests that varying this parameter does not introduce structural artefacts. Each of the complexes analysed here had the cross-linking energy optimised with gel electrophoresis-based assays for maximum yields of cross-linked complex with minimal UV-induced multimerisation. The results imply that this step may not be so critical in maximising the unique data produced by an experiment. The unique data obtained when the temperature at irradiation is varied also remains fairly consistent, again highlighting the robustness of the information obtained. Closer examination of this data however reveals more subtle trends. Although the CLIR-MS method is not currently designed for quantitative structural interpretation, the inverse correlations in numbers of identifications found at the most prominent RNA-contacting amino acids with temperature are noteworthy, despite the non-redundant structural information remaining broadly constant in both cases. The in-solution dynamics of a protein-RNA complex will likely vary over such a broad temperature range, and these results could suggest the suitability of UV cross-linking approaches, with further optimisations, for studying protein-RNA dynamics. Indeed, recent work using femtosecond laser-induced UV cross-linking in protein-RNA complexes demonstrates the potential of UV cross-linking for dynamics studies[69].

Incorporation of 4SU in place of uracil is conventionally accepted by the scientific community as a strategy for increasing the yield of a protein-RNA UV cross-linking experiment, under the assumption that structure or function of an RNA are not impacted[35,58,70]. The data we present here

may alter how such data is interpreted in a structural context. Longer cross-linking distances do not exclude the use of such data in structural modelling workflows, but may require distinct treatment (i.e. a longer distance restraint for 4SU-derived cross-links). Indeed, in the case of protein-protein cross-linking data, complementary approaches with differing specificity and cross-link distance both add value to computational pipelines that predict protein structures[71]. In the most extreme case, a particularly long cross-linking distance could be compared with proximity-tagging proteomics workflows such as BioID[72], which provide valuable biological information, even without specifying a precise interaction distance.

Recently published proteome-wide studies of RNA binding proteins captured by RNA pull-down also highlight distinct behaviours of uracil and 4SU, with each pulling down a different subset of the proteome[44]. Our observation that 4SU induced cross-links lead to distinct cross-linked amino acid positions of FOX1 compared with natural uracil is consistent with 4SU cross-linking capturing a distinct subset of protein-RNA binding interactions from natural bases in the proteome-wide study. The authors speculate that the differences may result from distinct lifetimes of the radical species generated when a natural nucleotide is irradiated compared with a 4SU. Indeed, more fundamental studies of sulphur-substituted nucleotides provide evidence of a longer-lived triplet state radical[73]. This may have the potential to react in more transient RNA-bound conformations in solution than a natural nucleotide, although further experimental work would be required to examine this hypothesis. An additional recent study suggests increased 4SU incorporation may impact splicing efficiency[74]. Taken together with the results presented here, these observations suggest that 4SU and natural uracil may have subtly distinct behaviours beyond the difference in reaction yields, which must be considered when interpreting results generated using 4SU cross-linking.

The DisVis analyses of protein-RNA cross-links clearly demonstrate the information content of CLIR-MS results. Whilst the RNA binding behaviour of many canonical RNA binding protein domains is well understood, a large array of novel RNA-binding protein domains are increasingly

observed by practitioners[1]. The structural characterisation of every novel RNA binding protein using established structural biology techniques will be an enormous undertaking for the scientific community, hence technical advances that accelerate the process are an attractive prospect. In the model complexes with prior structures studied here in **Fig. 3**, CLIR-MS derived restraints contain sufficient information to position the RNA on the correct RNA binding surface of the protein, even in the absence of other complementary data types. This represents an additional use case to the one shown previously[34], where an integrative modelling approach used multiple structural data sources to determine a final model. The CLIR-MS technique can therefore now be confidently applied for *de novo* low-resolution structural characterisations, providing an attractive pipeline for the study of protein-RNA binding sites in the solution state.

The model complexes studied in **Fig. 3**, all have short, single stranded and flexible RNAs. Furthermore, especially in the case of the MBNL1 complex, cross-linked amino acids tend to fall within a relatively small spatial cluster on the surface of each protein. In such cases, the RNA occupation spaces provided by DisVis reflect the remaining rotational degrees of freedom. In the case of the SF3A1-UBL interaction with U1 snRNA SL4 shown in **Fig. 5**, the RNA is instead formed into a rigid stem-loop structure. Furthermore, multiple clusters of cross-linked amino acids are spread more widely across the surface of the protein. This results in a narrower and more elongated occupation space. These different behaviours could indicate that CLIR-MS data is likely most successful as a standalone data type when applied to study the interaction of more rigid RNA structural features with a protein, and where multiple, spatially separated protein-RNA contact sites are identified by cross-linking.

Overall, the optimised CLIR-MS protocol and data analysis approach provide much greater numbers of identifications than the original protocol[34], improving the confidence in identified protein-RNA interaction sites detected. These identifications compare favourably with existing structures derived from other established structural techniques, when used to study complexes with well characterised structures. We used these comparisons to make general inferences about

the distances over which UV-induced protein-RNA cross-links form, and the robustness of protein-RNA cross-linking data. Our optimisations and observations guide the interpretation of protein-RNA cross-linking data, whilst demonstrating a new use case as an independent source of structural data. We therefore propose that CLIR-MS data is well suited to low-resolution binding interface characterisation for rigid complexes when considered in isolation, or as an additional complementary data type in more sophisticated integrative modelling pipelines[64] to achieve high-resolution, atomic-scale models. In cases of the latter, CLIR-MS data may prove particularly valuable when probing flexible protein regions where more established structural techniques relying on conformational homogeneity may struggle to provide coverage, as shown in our previous work[34]. The method may prove a useful tool to reliably study the emerging plethora of non-canonical RNA binding domains with the relative speed of an MS-based pipeline.

Methods

Protein expression and purification

PTBP1 and FOX1 RRM were prepared as described previously[34,43]. MBNL1 (amino acids 1-269 of MBNL140) was obtained in pGEX-6P1 (GE Healthcare). Plasmids were transformed into *E. coli* BL21 (DE3) codon+ (RIL) (Agilent Technologies) for protein expression. Cells were grown in K-MOPS minimal medium until $OD_{600\ nm}$ ~0.5, shifted from 37 °C to 20 °C, and induced at $OD_{600\ nm}$ 0.7-0.8. Expression was carried out for 22-24 h. After harvesting the cells by centrifugation (15 min, 6000 rpm, 4 °C, Sorvall SLC6000 fixed angle rotor), dry pellets were frozen at -20 °C. Cells were thawed and resuspended to ~0.25 g/mL in phosphate buffered saline (PBS; 140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4 pH 7.3) using 1x cOmplete™ EDTA-free (Roche) tablet per 2 L culture. The cell suspension was homogenised using a 100 µm H10Z cell and Microfluidizer (Microfluidics) operated at 15000 psi, over three cycles. The lysate was subsequently clarified by centrifugation (60 min, 17000 rpm, 4 °C, Sorvall SS-34 fixed angle rotor). The resulting supernatant was incubated for 4-5 h at 4 °C on glutathione sepharose 4B (GE Healthcare) (3 mL per 2 L expression). After this, all further purification steps were performed at RT. The supernatant-resin slurry was loaded onto gravity flow columns and washed with 2 bed volumes PBS, followed by 10 bed volumes PBS + 1 M NaCl. Protein was eluted step-wise using 50 mM Tris, 500 mM NaCl, 50 mM glutathione (reduced) pH 8 (adjusted with NaOH). Pooled eluate was dialysed overnight at 4 °C into 20 mM sodium phosphate (NaP), 25 mM NaCl, 10 µM $ZnSO_4$, 5 mM beta-mercaptoethanol pH 7. The GST-tag was cleaved by addition of HRV3C (1mg per 100 mg protein). Cleaved GST was separated by anion exchange chromatography (HiTrap Q HP 5 mL (GE Healthcare)). The equivalent of 1 L expression was injected per run, concentrated to 1 mL. To obtain non-degraded MBNL1Δ101 samples, flow through was again concentrated and subjected to gel-filtration chromatography, and buffer exchanged to 20 mM NaP, 50 mM NaCl, 10 µM $ZnSO_4$, 5 mM beta-mercaptoethanol pH 6. Protein was then concentrated, aliquoted, snap frozen in liquid nitrogen, and stored at -80 °C until use.

For SF3A1-UBL (amino acids 704-793), the protein sequence fused to an N-terminal GB1 solubility tag and a 6x TEV-cleavable His-tag was cloned into pET24b (Novagen). Plasmids were transformed into *E. coli* BL21 (DE3) codon+ (RIL) (Agilent Technologies) for protein expression. Cells were induced at $OD_{600\ nm}$ 0.6-0.8 with 1 mM isopropyl-β-d-thiogalactopyransoide (IPTG). Expression was carried out for 4 h at 37 °C. Cells were grown in LB-medium (DIFCOTM LB-Broth, Fisher Scientific) with chloramphenicol and kanamycin. After harvesting the cells by centrifugation (10 min, 5000 x *g*, 4 °C, Sorvall SLC6000 fixed angle rotor), pellets were resuspended in 20 mM Tris pH 8, 1 M NaCl (buffer A), 10 mM imidazole, with cOmplete™ EDTA-free protease inhibitor. Cell lysis was carried out with a microfluidizer (Microfluidics), and the lysate centrifuged for clarification (30 min, 5000 x *g*, 4 °C). Protein purification was carried out by Ni-affinity chromatography, either with Ni-NTA beads (QIAGEN), step-wise by gravity flow, or using an ÄKTA Prime purification system (Amersham Biosciences) equipped with 5 mL HisTrap column (GE Healthcare), with an imidazole gradient of buffer B (20 mM Tris pH 8, 0.25 M NaCl, 500 mM imidazole). The buffer of the fusion proteins was exchanged by dialysis to buffer C (20 mM Tris pH 8, 0.25 M NaCl, 2.5 mM β-mercaptoethanol). The fusion protein was then cleaved overnight at 4 °C, using 6x His tag TEV (purified in house). GB1-6His and the His-TEV protease were removed from the solution with Ni-NTA beads, and the solution incubated with RNaseOUT (Invitrogen) for 15 min. The protein was then purified by size exclusion chromatography, using a HiLoad 16/60 Superdex 75 pg (GE) in 10 mM sodium phosphate pH 6, 50 mM NaCl. Protein was then concentrated, aliquoted, snap frozen in liquid nitrogen, and stored at -80 °C until use.

Preparation of RNA

Multiple RNA isotope labelling strategies are available for CLIR-MS[75], and are employed in experiments presented here. $^{13}C^{15}N$ labelling results from transcription of RNA in isotopically labelled cell culture medium, as demonstrated previously[34]. Alternatively, chemically synthesised short RNAs are employed, where RNA is synthesised using ^{13}C ribonucleotides (also used elsewhere[53]). $^{13}C^{15}N$ *in vitro* transcribed RNA sequences, EMCV IRES RNA (sequence: 5'-

GGAUACUGGCCGAAGCCGCUUGGAAUAAGGCCGGUGUGCGUUUGUCUAUAUGUUAUUUU CCACCAUAUUGCCGUCUUUUGGCAAUGUG-3') and U1 snRNP SL4 RNA (sequence: 5'-GGGGACUGCGUUCGCGCUUUCCCC-3') were prepared as described previously[34]. For chemically synthesised RNAs, standard phosphoramidites were purchased from Thermo Fisher Scientific. ^{13}C ribose-labelled phosphoramidites were purchased from Pitsch Nucleic Acids. 4-thiouridine phosphoramidites were synthesised as described previously[76,77]. All other chemicals were obtained from Sigma-Aldrich, Fluorochem, TCI, and Fisher Scientific.

Synthesis of oligonucleotides for model complexes

All oligonucleotides used were synthesised on a 50 nmol scale with the MM12 synthesiser (Bio Automation Inc.) using 500 Å UnyLinker CPG (Controlled-pore glass, ChemGenes) with standard synthesis conditions. Coupling time for the phosphoramidites was 2 × 180 s. The RNA phosphoramidites were used as 0.08 M solutions in dry acetonitrile (ACN). The activator BTT (CarboSynth) was prepared as 0.24 M solution in dry ACN. 0.02 M I_2 solution in Tetrahydrofuran (THF)/Pyridine/water (70:20:10, w/v/v/v) was used as oxidising reagent. Capping reagent A was THF/lutidine/acetic anhydride (8:1:1) and capping reagent B was 16 % N-methylimidazole in THF. Detritylation was performed using 3 % dichloroacetic acid in dichloromethane.

For deprotection and cleavage from the solid support, the CPG was treated with gaseous methylamine for 1.5 h at 70 °C. For RNAs containing 4-thiouridine, the oligonucleotide was first incubated with 1 M DBU (1,8-diazabicyclo[5.4.0]undec-7-ene) in dry ACN (1 mL) for 3 h at RT and the CPG resin was washed with 5 mL in ACN. Afterwards the oligonucleotide was deprotected and cleaved from the solid support by using ammonia containing 50 mM NaSH (1 mL) at RT for 24 h. Desilylation for all RNA was carried out by treatment with a mixture of N-methyl-2-pyrrolidone (60 µL), triethylamine (30 µL), and triethylamine trihydrofluoride (40 µL) at 70 °C for 2 h. The reaction was quenched by adding trimethylethoxysilane (200 µL, 5 min, RT). Purification was carried out on an Agilent 1200 series preparative RP-HPLC using an XBridge OST C18 column (10 × 50 mm, 2.5

µm; Waters) at 65 °C with a flow rate of 5 mL/min, gradient 10–50 % B in 5 min (A= 0.1 M aqueous triethylamine/acetic acid, pH 8.0; B= 100 % ACN).

Fractions containing the DMT-protected product were collected, dried under vacuum, and treated with 40 % aqueous acetic acid for 15 min at RT to remove the DMT group. Samples were dried under vacuum and dissolved in 200 µL of water, and purified by RP-HPLC on an XBridge OST C18 column (10 × 50 mm, 2.5 µm; Waters) at 65 °C with a flow rate of 5 mL/min, gradient 2–20 % B in 6 min (A= 0.1 M aqueous triethylamine/acetic acid, pH 8.0; B= 100 % ACN).

Fractions containing the desired product were collected and dried under vacuum. Mass and purity were confirmed by LC–MS (Agilent 1200/6130 system) on an Acquity OST C18 column (2.1 × 50 mm; Waters).. The column oven was set to 65 °C, flow-rate: 0.3 mL/min, gradient 1–35 % B in 15 min (A= water containing 0.4 M hexafluoroisopropanol, 15 mM triethylamine; B= methanol). UV absorption of the final products was measured on a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific).

Cross-linking of protein-RNA complexes

5 nmol of purified protein-RNA complex was prepared per enrichment replicate, at concentration between 0.1 and 1.0 mg/mL, depending on the sample. RNA in the sample consisted of equimolar mixtures of unlabelled RNA and stable isotope labelled RNA (either ^{13}C only for chemically synthesised RNA or $^{13}C^{15}N$ for in vitro transcribed RNA). The sample was subjected to 254 nm irradiation in a UVP Ultraviolet Crosslinker (Ultraviolet Products), with the sample cooled on a metal plate, pre-cooled to -20 °C, throughout. 4-thiouracil containing samples were irradiated four times with 150 mJ/cm^2 at 365 nm in a Vilber Lourmat Bio-link BLX Crosslinker (Collegien). Each irradiation step was followed by a pause of 1 min to allow the sample to cool. Unless otherwise stated, cross-linking irradiation energy was optimised using SDS-PAGE analysis to maximise cross-linking yield of the protein-RNA heterodimer, whilst minimising UV-induced multimers and degradation products.

Digestion and enrichment

Samples were precipitated using 0.1 volumes 3 M sodium acetate (pH 5.2) and 3 volumes of ethanol precooled to -20 °C, and kept at -20 °C for at least 2 h. Pellets of precipitated complexes were collected by centrifugation (30 min, 13000 × *g*, 4 °C). Pellets were washed in 2 volumes of 80 % ethanol in water (v/v) at -20 °C. The centrifugation step was repeated, and pellets air dried for 10 min. 50 µL of 50 mM Tris-HCl (pH 7.9) with 4 M urea was used to resuspend the pellet, and the solution then diluted with 150 µL 50 mM Tris-HCl, pH 7.9. 5 µg and 5 U per mg of cross-linked sample, of RNases A (Roche Diagnostics) and T1 (Thermo Scientific) respectively, were added, and RNA digestion carried out for 2 h at 52 °C on a ThermoMixer (Eppendorf). After cooling on ice, 2 µL of 1 M $MgCl_2$, and 125 U of benzonase (Sigma Aldrich) per mg of cross-linked complex, was added to each sample. Further RNA digestion was then carried out for 1 h at 37 °C on a ThermoMixer. Sequencing grade trypsin (Promega) was added at a 24:1 protein:enzyme ratio (w/w). Samples were incubated overnight at 37 °C on a shaking incubator, then heated to 70 °C for 10 min to deactivate trypsin. After deactivation, samples were cleaned up by solid-phase extraction (SepPak 50 mg tC18 cartridges, Waters), and dried in a vacuum centrifuge.

Titanium dioxide metal oxide affinity chromatography (MOAC) was used to enrich protein-RNA crosslinks as described previously[34,78]. In brief, dried samples were resuspended in 100 µL MOAC loading buffer (water:ACN:trifluoroacetic acid (TFA), 50:50:0.1 (v/v/v) with 300 mg/mL lactic acid), and incubated on a ThermoMixer at 1200 rpm for 30 min with 5 mg of pre-equilibrated TiO_2 beads (10 µm Titansphere PhosTiO, GL Sciences). Beads were settled by centrifugation (1 min, 10000 × *g*, RT), and the supernatant carefully removed and discarded. 100 µL fresh MOAC loading buffer was added, and the sample incubated for a further 15 min. Centrifugation was repeated, the supernatant removed, and 100 µL MOAC washing buffer (water:ACN:TFA, 50:50:0.1 (v/v/v)) was added. After a further 15 min incubation, centrifugation was repeated and the supernatant discarded. Peptide-RNA adducts were then eluted from the beads with 50 µL MOAC elution buffer (50 mM ammonium phosphate, pH 10.5). Samples were incubated for 15 min, and beads again

settled by centrifugation. The supernatant was carefully collected, stored on ice, and elution repeated a second time and combined with the first eluate. Eluate solution was immediately acidified to pH 2-3 with TFA. Eluates were purified with C_{18} solid phase extraction using self-packed Stage tips. In brief, two layers of C_{18} membrane (Empore, 3M) packed in a 200 µL tip (MaxRecovery, Axygen) were washed with 80 µL 100% ACN with 0.1% formic acid (FA), 80% ACN with 0.1% FA in water, then equilibrated twice with 80 µL 5% ACN with 0.1% FA in water. Sample was applied to the membrane, and the membrane then washed 3 times with 80 µL 5% ACN with 0.1% FA in water. Purified peptide-RNA adducts were eluted from the membrane three times with 50 µL 50% ACN with 0.1% FA in water. LoBind tubes (Eppendorf) pre-washed with 50% ACN with 0.1% FA in water were used to collect purified peptide-RNA adducts. The sample was then dried in a vacuum centrifuge. For PTBP1-IRES samples cleaned up with C_{18} cartridges in **Fig. 1**, SepPak tC18 cartridges (Waters) were used for clean-up instead.

Analysis of samples with LC-MS/MS

Each dried sample was resuspended in 20 µL mobile phase A (described below), and 5 µL of each sample was injected for LC-MS/MS analysis. For MS method optimisation experiments in **Fig. 1**, resulting samples from multiple enrichment replicates were pooled, and 3 µL sample was injected to evaluate each acquisition method. For data shown in all figures except, LC-MS/MS analysis was performed using an Easy-nLC 1200 HPLC system (Thermo Fisher Scientific) coupled to an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific) equipped with a Nanoflex (Thermo Fisher Scientific) nanoflow electrospray source. Peptide-RNA adducts were separated using a PepMap RSLC column (250 mm × 75 um, 2 µm particle size, Thermo Fisher Scientific) with gradient of 6-40% mobile phase B (A= water:ACN:FA, 98:2:0.15 (v/v/v); B= water:ACN:FA acid, 20:80:0.15 (v/v/v)) with a flow rate of 300 nL/min over 60 min. Peptide-RNA adducts were separated on a PepMap RSLC column (150 mm × 75 um, 2 µm particle size, Thermo Fisher Scientific) using a gradient of 5-30% mobile phase B (A= water:ACN:FA, 98:2:0.15 (v/v/v); B =water:ACN:FA acid, 2:98:0.15 (v/v/v)) with a flow rate of 300 nL/min over 60 min.

The Orbitrap Fusion Lumos was used in data dependent acquisition mode, and the Orbitrap mass analyser used for precursor ion spectra acquisition, with resolution of 120000. Fragmentation was achieved with higher-energy collisional dissociation (HCD), using stepped collision energies of 21.85 %, 23 % and 24.15 %. For the MS method optimisation experiments shown in **Fig. 1**, alternative fragmentation methods (CID with normalised collision energy of 35 %; EThcD with supplemental activation at 25 %; stepped HCD collision energies from 23-31 %) were used. Precursor ions with charge states between +2 and +7 were selected with a quadrupole isolation window of 1.2 m/z and cycle time of 3 s, with a dynamic exclusion period of 30 s. Resultant fragment ions were detected in the ion trap at rapid resolution. For experiments in **Fig. 5**, fragment ions were detected in the Orbitrap using a resolution of 30000.

The Orbitrap Elite was operated in data dependent acquisition mode. The Orbitrap analyser was used for acquisition of precursor ion spectra with a resolution of 120000. Collision-induced dissociation (CID) with normalised collision energy of 35 % was used for fragmentation. The dynamic exclusion period was set to 30 s. Fragment ions were detected in the ion trap at normal resolution.

Data analysis with xQuest (light-heavy labelled species)

Data files produced by the mass spectrometer (Thermo Fisher .RAW format) were converted to centroided mzXML files using msconvert.exe (ProteoWizard msConvert v.3.0.9393c[79]). Files were then searched using xQuest (version 2.1.5, available at https://gitlab.ethz.ch/leitner_lab/xquest_xprophet)[40,80] against a database containing only the sequence of the target protein. xQuest was originally designed to analyse protein-protein XL-MS data, with workflows in which an equimolar mixture of light and heavy isotopes of a chemical cross-linking reagent have been used to covalently link peptides. During a CLIR-MS experiment, a light-heavy stable isotope labelled RNA segment cross-linked to a peptide behaves similarly to a monolink[17] (type 0 cross-link[81]) in peptide-peptide cross-linking nomenclature, thus enabling xQuest to also process such data.

All amino acid types were permitted as possible modification sites, and all possible RNA-derived adducts of 1-4 nucleotides in length, based on the RNA sequence of the respective complex and including all loss products, were considered possible modifications. For RNA produced by solid phase synthesis, a delta mass of 5.016774 Da per labelled nucleotide in the expected RNA modification was specified, to restrict identifications to those containing labelled RNA. For *in vitro* transcribed sequences, delta masses were defined according to expected $^{13}C^{15}N$ labelling patterns, described previously[34]. A +/- 15 ppm mass tolerance window and 60 s retention time tolerance was used for pairing of light-heavy species. Further parameters for xQuest searching (described previously[80]): Enzyme = trypsin, maximum missed cleavages = 2, MS1 mass tolerance = 10 ppm, MS2 mass tolerance = 0.2 Da for ion trap MS2 data or 10 ppm for Orbitrap MS2 data. Identifications with an ld.Score > 20 (according to the scoring scheme described previously[80]) were considered. FDR estimations may be less reliable when calculated using low numbers of peptide-spectrum matches, or when related ion species are present but not of interest in a dataset[82]. Given the numbers of spectral identifications observed in CLIR-MS protein-RNA cross-linking data sets are rather low compared with conventional proteomics experiments, the score-threshold was selected for enhanced stringency over an FDR calculation. Further processing was completed using custom Python 3.7.1 scripts. CLIR-MS plots shown here have amino acid numbering retrospectively adjusted from the FASTA file numbering to match prior structural models of each complex published in the PDB. Identifications were further refined for mass accuracy. Where multiple identifications were produced against the same spectrum, only the highest scoring identification was retained. Raw data files and xQuest search engine result files are accessible in PRIDE, described in ‘Data Availability’.

The sparse nature of metal oxide-enriched protein-RNA adduct samples means XLSMs made are often near the limits of detection by MS, and hence are vulnerable to fluctuations in instrument performance over time. Comparisons are therefore only made within batches where data acquisition took place at a similar time.

Data analysis with MSFragger (open modification search)

Thermo Fisher .RAW files were converted as above. Default parameters for an open search using MSFragger (v2.1) were loaded, and the following modified: modification range = 150-1400 Da, fragment mass tolerance = 0.2 Da, allowed missed cleavages = 2, minimum peptide length = 5, top peaks = 250, min_fragments_modelling = 3, min_matched_fragments = 5, allow_multiple_variable_mods_on_residue = 1. The search was executed using the FragPipe GUI (v11.0), and results outputted to a comma-separated value (csv) file. Identifications with an expect score of greater than 0.05 were removed. Remaining matches to decoy sequences were also removed. Mass additions to peptide were then aggregated in 0.1 Da mass bins, and RNA-derived mass additions to peptides were manually annotated according to existing literature where possible, to form a putative list of RNA-derived products. All putative products were subjected to validation using a light-heavy dependent xQuest search.

Comparison of CLIR-MS results with published structures

Protein-RNA cross-links identified from the xQuest search were filtered, retaining only identifications where a mononucleotide RNA was cross-linked to a peptide. A custom script in PyMOL (version 2.3.2, Schrödinger) was used to compare identified cross-links with published ensembles of models derived from NMR spectroscopy for each model complex; distances for cross-links were measured from the Cα atom of the amino acid position in the cross-link identification, to N1 (pyrimidines) or N9 (purines) of the closest nucleotide in the structure matching the nucleotide type observed in the cross-link identification. Distances were outputted to a csv file and distributions plotted using the Python package Plotly[83]. Identified protein-RNA cross-links were plotted on published structures using PyXlinkViewer[84], with code modified to plot nucleotide cross-links.

Visualisation of structural information content with DisVis

A restraint list of amino acid and nucleotide positions was prepared from protein-RNA cross-links identified by CLIR-MS. Amino acid positions were taken directly from the xQuest results. For complexes in **Fig. 3**, RNA positions were selected based on the closest matching nucleotide, as measured in published structures. For the U1 snRNA SL4 sequence in **Fig. 5**, RNA positions were derived by overlaying non-redundant RNA compositions detected at every cross-linked amino acid position with the full RNA sequence. For the heatmap plot in **Fig. 5b**, the score contribution of the cross-linked adduct at each nucleotide position in the total RNA sequence is normalised by the length of the RNA adduct. The 5% highest scoring positions in the heatmap (in arbitrary units according to the heatmap scale) were used to define RNA sequence positions for restraints. In both **Fig. 3** and **Fig. 5**, the restraint distances were defined between a minimum of 0 Å, and a maximum of 12 Å, with 12 Å corresponding to the upper quartile of all cross-linking distances measured against published structural ensembles in **Fig. 3g-i**. Restraints were specified from Cα of the amino acid position to N1 (pyrimidines) or N9 (purines) of the nucleotide position. For complexes in **Fig. 3**, published structures were downloaded from the PDB, and protein and RNA chains were exported as separate molecules. Protein structures were specified as fixed chains, and RNA structures as scanning chains. Individual protein and RNA chains submitted to the DisVis web server[56,57] (https://wenmr.science.uu.nl/disvis/), together with the restraints file in the required format. “Occupancy Analysis” was enabled, and “Complete Scanning” was selected. All other parameters were left at default values. Outputs from DisVis analysis were visualised with UCSF Chimera[85].

Author Contributions

A.L., F.H.-T.A., J.H. and R.A. conceived the study, and supervised C.P.S., A.K. and T.d.V. A.K. produced chemically synthesized RNA. T.d.V. produced double-isotope labelled RNA. A.K. and T.d.V. (FOX1), T.d.V. (SF3A1, PTBP1, MBNL1), I.B. (MBNL1) expressed and purified proteins for the study. A.K. and T.d.V. performed cross-linking of protein-RNA complexes. C.P.S. performed TiO_2 enrichment, C_{18} clean-up, and measured samples by mass spectrometry. C.P.S. and M.G. wrote analysis scripts and conducted data analysis using xQuest. C.P.S. performed DisVis analysis. C.P.S. and A.L. wrote the manuscript together. All authors contributed to manuscript revisions and approved the final manuscript.

Acknowledgements

We thank P. Picotti for access to laboratory and MS infrastructure. We thank C. von Schroetter for assistance in producing *in vitro* transcribed RNA. We thank the lab of C. Branlant for the MBNL1 plasmid. We thank N. de Souza for critical comments during preparation of the manuscript. Sources of funding for this work were: ETH Zürich (Research Grant ETH-24 16-2 to A.L., F.H.-T.A., J.H., and R.A.); Strategic Focus Area for the ETH Domain “Personalized Health and Related Technologies” (TechTransfer Project PHRT-503 to A.L. and F.H.-T.A.); European Research Council (Advanced Grant ERC-20140 AdG 670821 to R.A.); and National Center of Competence in Research, RNA & Disease (NCCR RNA & Disease, 51NF40-182880). Funding from the ETH Scientific Equipment program and the European Union Grant ULTRA-DD (FP7-JTI 115766) was used to purchase the mass spectrometers used in this work.

References

1. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* **19**, 327–341 (2018).

2. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).

3. 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).

4. Gebauer, F., Schwarzl, T., Valcárcel, J. & Hentze, M. W. RNA-binding proteins in human genetic disease. *Nat. Rev. Genet.* **22**, 185–198 (2021).

5. Cléry, A., Blatter, M. & Allain, F. H. T. RNA recognition motifs: boring? Not quite. *Curr. Opin. Struct. Biol.* **18**, 290–298 (2008).

6. Hall, T. M. T. Multiple modes of RNA recognition by zinc finger proteins. *Curr. Opin. Struct. Biol.* **15**, 367–373 (2005).

7. Linder, P. & Jankowsky, E. From unwinding to clamping - the DEAD box RNA helicase family. *Nat. Rev. Mol. Cell Biol.* **12**, 505–516 (2011).

8. Schluenzen, F. *et al.* Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell* **102**, 615–623 (2000).

9. Ramakrishnan, V. Ribosome structure and the mechanism of translation. *Cell* **108**, 557–572 (2002).

10. Matera, A. G. & Wang, Z. A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* **15**, 108–121 (2014).

11. Plaschka, C., Lin, P. C. & Nagai, K. Structure of a pre-catalytic spliceosome. *Nature* **546**, 617–621 (2017).

12. Tsvetanova, N. G., Klass, D. M., Salzman, J. & Brown, P. O. Proteome-Wide Search Reveals Unexpected RNA-Binding Proteins in Saccharomyces cerevisiae. *PLoS One* **5**, e12671 (2010).

13. Castello, A. *et al.* Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol. Cell* **63**, 696–710 (2016).

14. Leitner, A., Faini, M., Stengel, F. & Aebersold, R. Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends Biochem. Sci.* **41**, 20–32 (2016).

15. O'Reilly, F. J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nat. Struct. Mol. Biol.* **25**, 1000–1008 (2018).

16. Sinz, A. Cross-Linking/Mass Spectrometry for Studying Protein Structures and Protein–Protein Interactions: Where Are We Now and Where Should We Go from Here? *Angew. Chemie - Int. Ed.* **57**, 6390–6396 (2018).

17. Leitner, A. *et al.* Probing native protein structures by chemical cross-linking, mass

spectrometry, and bioinformatics. *Mol. Cell. Proteomics* **9**, 1634–1649 (2010).

18. Iacobucci, C. *et al.* First Community-Wide, Comparative Cross-Linking Mass Spectrometry Study. *Anal. Chem.* **91**, 6953–6961 (2019).

19. Leitner, A. *et al.* Expanding the Chemical Cross-Linking Toolbox by the Use of Multiple Proteases and Enrichment by Size Exclusion Chromatography. *Mol. Cell. Proteomics* **11**, M111.014126 (2012).

20. Steigenberger, B., Pieters, R. J., Heck, A. J. R. & Scheltema, R. A. PhoX: An IMAC-Enrichable Cross-Linking Reagent. *ACS Cent. Sci.* **5**, 1514–1522 (2019).

21. Yu, C. & Huang, L. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Anal. Chem.* **90**, 144–165 (2018).

22. Young, M. M. *et al.* High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5802–5806 (2000).

23. Koukos, P. I. & Bonvin, A. M. J. J. Integrative Modelling of Biomolecular Complexes. *J. Mol. Biol.* **432**, 2861–2881 (2020).

24. Rout, M. P. & Sali, A. Principles for Integrative Structural Biology Studies. *Cell* **177**, 1384–1403 (2019).

25. Budowsky, E. I., Axentyeva, M. S., Abdurashidova, G. G., Simukova, N. A. & Rubin, L. B. Induction of polynucleotide-protein cross-linkages by ultraviolet irradiation. Peculiarities of the high-intensity laser pulse irradiation. *Eur. J. Biochem.* **159**, 95–101 (1986).

26. Ule, J. *et al.* CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science* **302**, 1212–1215 (2003).

27. 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).

28. Ramanathan, M., Porter, D. F. & Khavari, P. A. Methods to study RNA–protein interactions. *Nat. Methods* **16**, 225–234 (2019).

29. Lee, F. C. Y. & Ule, J. Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Mol. Cell* **69**, 354–369 (2018).

30. 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).

31. 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).

32. Queiroz, R. M. L. L. *et al.* Comprehensive identification of RNA–protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat. Biotechnol.* **37**, 169–178 (2019).

33. Panhale, A. *et al.* CAPRI enables comparison of evolutionarily conserved RNA interacting regions. *Nat. Commun.* **10**, 2682 (2019).

34. Dorn, G. *et al.* Structural modeling of protein-RNA complexes using crosslinking of segmentally isotope-labeled RNA and MS/MS. *Nat. Methods* **14**, 487–490 (2017).

35. Kramer, K. *et al.* Mass-spectrometric analysis of proteins cross-linked to 4-thio-uracil- and 5-bromo-uracil-substituted RNA. *Int. J. Mass Spectrom.* **304**, 184–194 (2011).

36. Meisenheimer, K. M. & Koch, T. H. Photocross-linking of nucleic acids to associated proteins. *Crit. Rev. Biochem. Mol. Biol.* **32**, 101–140 (1997).

37. Shetlar, M. D., Carbone, J., Steady, E. & Hom, K. Photochemical Addition of Amino Acids and peptides to Polyuridylic Acid. *Photochem. Photobiol.* **39**, 141–144 (1984).

38. Darnell, R. B. HITS-CLIP: Panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip. Rev. RNA* **1**, 266–286 (2010).

39. Rappsilber, J., Ishihama, Y. & Mann, M. Stop And Go Extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**, 663–670 (2003).

40. Rinner, O. *et al.* Identification of cross-linked peptides from large sequence databases. *Nat. Methods* **5**, 315–318 (2008).

41. Frese, C. K. *et al.* Improved Peptide Identification by Targeted Fragmentation Using CID, HCD and ETD on an LTQ-Orbitrap Velos. *J. Proteome Res.* **10**, 2377–2388 (2011).

42. Zhang, C. *et al.* Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes Dev.* **22**, 2550–2563 (2008).

43. Auweter, S. D. *et al.* Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. *EMBO J.* **25**, 163–173 (2006).

44. Shchepachev, V. *et al.* Defining the RNA interactome by total RNA-associated protein purification. *Mol. Syst. Biol.* **15**, e8689 (2019).

45. Peil, L. *et al.* Identification of RNA-associated peptides, iRAP, defines precise sites of protein-RNA interaction. *bioRxiv,* https://doi.org/10.1101/456111 (2018).

46. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).

47. Geoghegan, K. F. *et al.* Spontaneous α-N-6-phosphogluconoylation of a 'His tag' in Escherichia coli: The cause of extra mass of 258 or 178 Da in fusion proteins. *Anal. Biochem.* **267**, 169–184 (1999).

48. Spellman, R. & Smith, C. W. J. Novel modes of splicing repression by PTB. *Trends Biochem. Sci.* **31**, 73–76 (2006).

49. Hellen, C. U. T. & Sarnow, P. Internal ribosome entry sites in eukaryotic mRNA molecules. *Genes Dev.* **15**, 1593–1612 (2001).

50. Oberstrass, F. C. *et al.* Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science* **309**, 2054–2057 (2005).

51. Pascual, M., Vicente, M., Monferrer, L. & Artero, R. The Muscleblind family of proteins: An emerging class of regulators of developmentally programmed alternative splicing. *Differentiation* **74**, 65–80 (2006).

52. Park, S. *et al.* Structural Basis for Interaction of the Tandem Zinc Finger Domains of Human

Muscleblind with Cognate RNA from Human Cardiac Troponin T. *Biochemistry* **56**, 4154–4168 (2017).

53. Knörlein, A. *et al.* Structural requirements for photo-induced RNA-protein cross-linking. *ChemRxiv,* https://doi.org/10.33774/chemrxiv-2021-05zhj (2021).

54. Chen, Z. A. *et al.* Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **29**, 717–726 (2010).

55. Klatt, F. *et al.* A precisely positioned MED12 activation helix stimulates CDK8 kinase activity. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 2894–2905 (2020).

56. Van Zundert, G. C. P. & Bonvin, A. M. J. J. DisVis: Quantifying and visualizing accessible interaction space of distance-restrained biomolecular complexes. *Bioinformatics* **31**, 3222–3224 (2015).

57. van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **428**, 720–725 (2016).

58. Hafner, M. *et al.* Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).

59. Sharma, S., Wongpalee, S. P., Vashisht, A., Wohlschlegel, J. A. & Black, D. L. Stem–loop 4 of U1 snRNA is essential for splicing and interacts with the U2 snRNP-specific SF3A1 protein during spliceosome assembly. *Genes Dev.* **28**, 2518–2531 (2014).

60. Martelly, W., Fellows, B., Senior, K., Marlowe, T. & Sharma, S. Identification of a noncanonical RNA binding domain in the U2 snRNP protein SF3A1. *RNA* **25**, 1509–1521 (2019).

61. Charenton, C., Wilkinson, M. E. & Nagai, K. Mechanism of 5’ splice site transfer for human spliceosome activation. *Science* **364**, 362–367 (2019).

62. Lukin, J.A., Dhe-Paganon, S., Guido, V., Lemak, A., Avvakumov, G.V., Xue, S., Newman, E.M., Mackenzie, F., Sundstrom, M., Edwards, A., Arrowsmith, C.H., S. G. C. (SGC). Solution structure of a human ubiquitin-like domain in SF3A1. *The Protein Data Bank* (PDB ID: 1ZKH) https://www.rcsb.org/structure/1ZKH doi:http://doi.org/10.2210/pdb1ZKH/pdb.

63. de Vries, T. *et al.* Sequence-specific RNA recognition by an RGG motif connects U1 and U2 snRNP for spliceosome assembly. *Proc. Natl. Acad. Sci.* **119**, e2114092119 (2022).

64. Schneidman-Duhovny, D. *et al.* A method for integrative structure determination of protein-protein complexes. *Bioinformatics* **28**, 3282–3289 (2012).

65. Findly, D., Herries, D. G., Mathias, A. P., Rabin, B. R. & Ross, C. A. The Active Site and Mechanism of Action of Bovine Pancreatic Ribonuclease. *Nature* **190**, 781–784 (1961).

66. Takahashi, K. The Structure and Function of Ribonuclease T1. *J. Biochem.* **67**, 833–839 (1970).

67. Nestle, M. & Roberts, W. K. An extracellular nuclease from Serratia marcescens. I. Purification and some properties of the enzyme. *J. Biol. Chem.* **244**, 5213–5218 (1969).

68. Bae, J. W., Kwon, S. C., Na, Y., Kim, V. N. & Kim, J.-S. S. Chemical RNA digestion enables robust RNA-binding site mapping at single amino acid resolution. *Nat. Struct. Mol. Biol.* **27**,

678–682 (2020).

69. Sharma, D. *et al.* The kinetic landscape of an RNA-binding protein in cells. *Nature* **591**, 152–156 (2021).

70. Castello, A. *et al.* System-wide identification of RNA-binding proteins by interactome capture. *Nat. Protoc.* **8**, 491–500 (2013).

71. Fajardo, J. E. *et al.* Assessment of chemical-crosslink-assisted protein structure modeling in CASP13. *Proteins Struct. Funct. Bioinforma.* **87**, 1283–1297 (2019).

72. Roux, K. J., Kim, D. I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J. Cell Biol.* **196**, 801–810 (2012).

73. Mai, S. *et al.* The origin of efficient triplet state population in sulfur-substituted nucleobases. *Nat. Commun.* **7**, 13077 (2016).

74. Altieri, J. & Hertel, K. J. The influence of 4-thiouridine labeling on pre-mRNA splicing outcomes. *bioRxiv* https://doi.org/10.1101/2021.09.03.458914 (2021).

75. Götze, M. *et al.* Single Nucleotide Resolution RNA–Protein Cross-Linking Mass Spectrometry: A Simple Extension of the CLIR-MS Workflow. *Anal. Chem.* **93**, 14626–14634 (2021).

76. Milecki, J., Nowak, J., Skalski, B. & Franzen, S. 5-Fluoro-4-thiouridine phosphoramidite: New synthon for introducing photoaffinity label into oligodeoxynucleotides. *Bioorganic Med. Chem.* **19**, 6098–6106 (2011).

77. Shah, K., Wu, H. & Rana, T. M. Synthesis of Uridine Phosphoramidite Analogs: Reagents for Site-Specific Incorporation of Photoreactive Sites into RNA Sequences. *Bioconjug. Chem.* **5**, 508–512 (1994).

78. Leitner, A. *et al.* Probing the phosphoproteome of HeLa cells using nanocast metal oxide microspheres for phosphopeptide enrichment. *Anal. Chem.* **82**, 2726–2733 (2010).

79. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).

80. Walzthoeni, T. *et al.* False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nat. Methods* **9**, 901–903 (2012).

81. Schilling, B., Row, R. H., Gibson, B. W., Guo, X. & Young, M. M. MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. *J. Am. Soc. Mass Spectrom.* **14**, 834–850 (2003).

82. Lin, A., Plubell, D. L., Keich, U. & Noble, W. S. Accurately Assigning Peptides to Spectra When only a Subset of Peptides Are Relevant. *J. Proteome Res.* **20**, 4153–4164 (2021).

83. Plotly Technologies Inc. Collaborative data science.

84. Schiffrin, B., Radford, S. E., Brockwell, D. J. & Calabrese, A. N. PyXlinkViewer: A flexible tool for visualization of protein chemical crosslinking data within the PyMOL molecular graphics system. *Protein Sci.* **29**, 1851–1857 (2020).

85. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and

analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

86. Panhale, A. *et al.* CAPRI enables comparison of evolutionarily conserved RNA interacting regions. *Nat. Commun.* **10**, 2682 (2019).

Figures

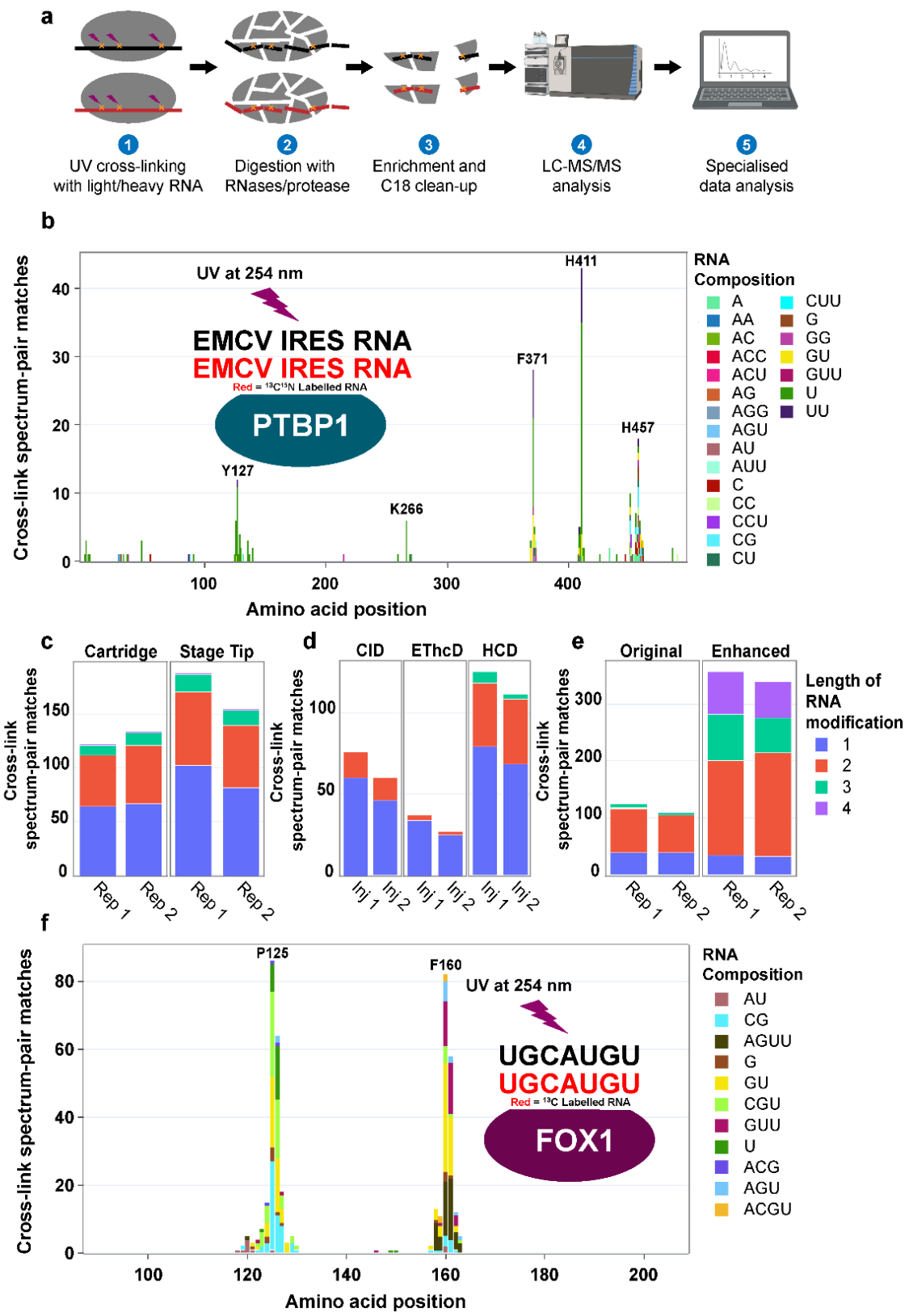

Figure 1: Optimisation of sample preparation and data acquisition for the CLIR-MS workflow.

a) Overview of the CLIR-MS workflow, as established in Ref. 34.

b) Structural information obtained from the PTBP1-IRES complex using the optimal experimental conditions in panels c) and d) respectively. Overlaid, schematic representations of the PTBP1 protein and EMCV IRES RNA in complex, used for data in panels b), c), and d).

c) Comparison of the number of XLSMs made from PTBP1-IRES samples (complex schematic shown in b) prepared using (conventional) cartridges and Stage tips for the final C_{18} clean-up step.

d) Comparison of analysis of a single PTBP1-IRES CLIR-MS sample (schematic shown in b) utilising different activation types for peptide fragmentation.

For c)-e), Rep = Replicate, Inj = Injection.

e) Numbers of identifications produced when the same FOX1-FBE sample (schematic shown in f) is prepared and analysed with and without the enhancements in method design. The optimisations from panels c and d transfer to other protein-RNA complexes.

f) Structural information obtained from the FOX1-FBE complex using the enhanced experimental conditions in panel e. Overlaid, schematic representation of the FOX1 protein and FBE RNA in complex, used for data in panels e and f.

In panels a, b, and f, heavy isotope RNA is represented in red, and light RNA in black (used in equimolar ratio for sample preparation).

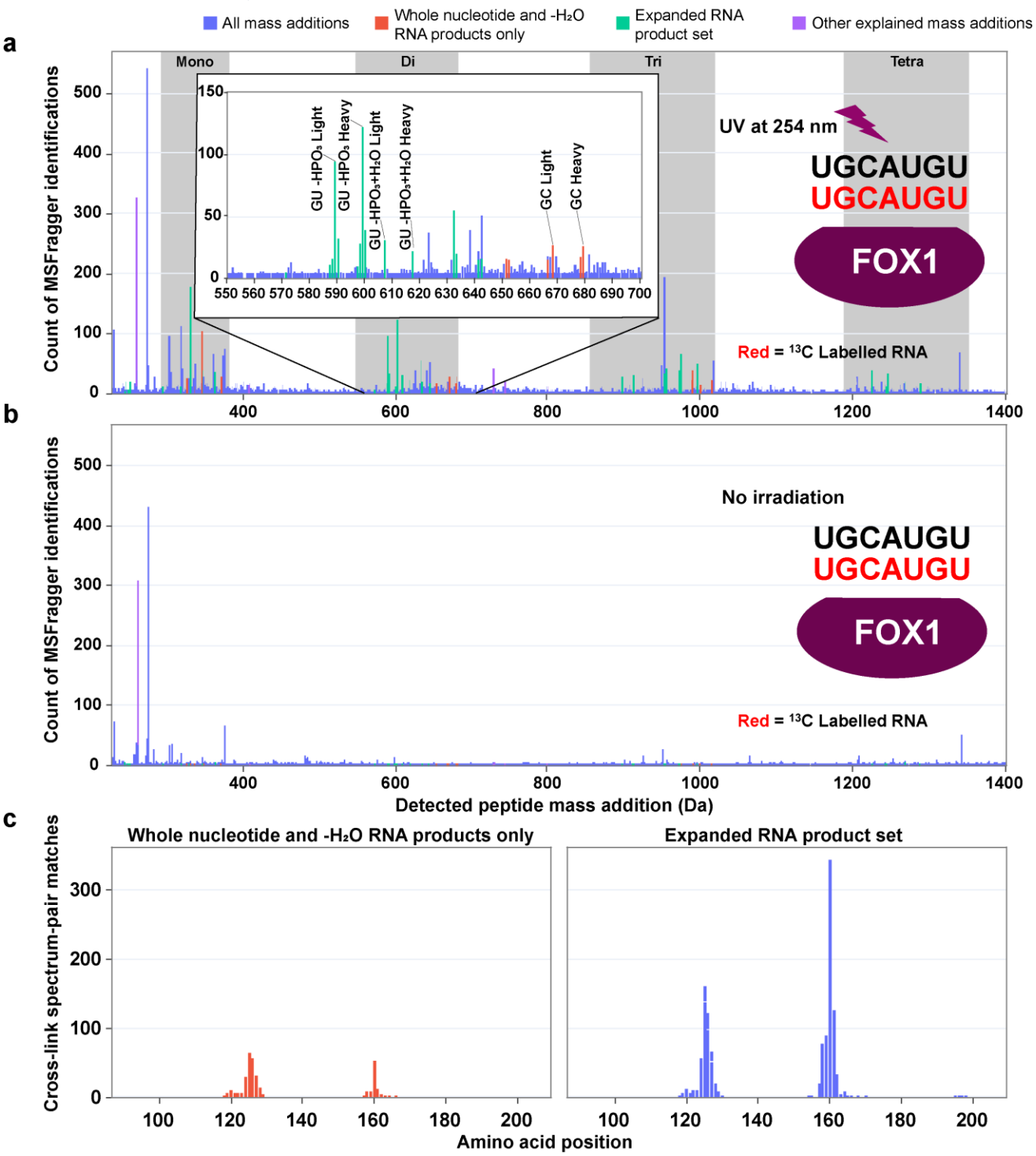

Figure 2: Optimisation of data analysis through better understanding of the cross-linking products.

a) and b) Open modification searches of the data produced from the a) irradiated and b) unirradiated complexes (data from FOX1-FBE). Any peptide mass additions between 150 and 1400 Da were considered by the search software. The number of identifications found corresponding to each 0.1 Da mass bin is shown. Regions corresponding to mono- (~300-400

Da), di- (~600-700 Da), tri- (~900-1000 Da) and tetranucleotide (~1200-1300 Da) adducts are highlighted in grey. The bars for each mass bin are coloured according to whether they have been used in the original CLIR-MS study[34] or in the present work. c) Comparison of closed xQuest searches of the same FOX1-FBE data from a) using the entire set of cross-linking products identified in this work (right) in comparison to the modifications specified previously (left)[34].

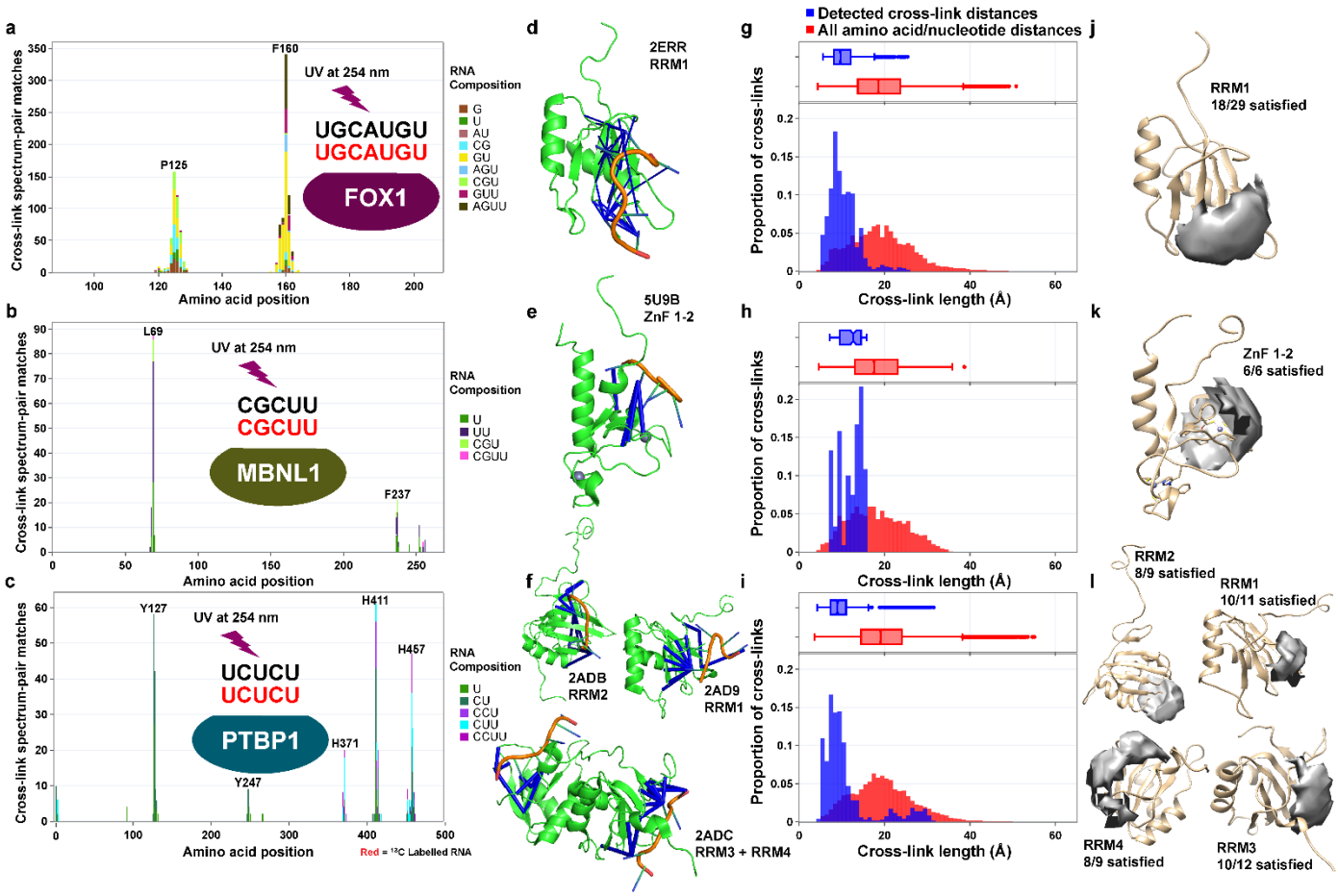

Figure 3: Application of CLIR-MS workflow to model complexes, and comparisons with published structures.

a)-c) CLIR-MS results from the complexes FOX1-FBE a), MBNL1-CGCUU b), and PTBP1-UCUCU c), with schematic representation of each protein-RNA complex overlaid

d)-f) Published structures from the PDB that correspond to the complexes illustrated in panels a)-c). Cross-links involving mononucleotide RNA adducts identified using CLIR-MS are superimposed on the structures in blue. Structures were visualised with PyMOL and the PyXlinkViewer plugin (The PyMOL Molecular Graphics System, Version 2.3.2 Schrödinger, LLC).

g)-i) Distribution of distances when cross-links from CLIR-MS are measured against the published structures in panels. d)-f), compared with all theoretically possible pairwise combinations of nucleotide and amino acid in each structure. Euclidean distances are measured from the Cα atom of the amino acid to the glycosidic nitrogen atom in the nucleotide (N1 for pyrimidines or N9 for purines). Boxes span Q1-Q3, with centre line representing the median. Whiskers represent upper

and lower fences of data points, or highest/lowest values where all values are within this range. g) Measured distances (blue), n=990; Control distances (red), n=18480. h) Measured distances, n=120; Control distances, n=9200. i) Measured distances, n=720; Control distances, n=51960. j)-l) Visualisation of the structural information contained in point-to-point distance restraints obtained from data in a)-c) using DisVis. The occupancy space shown relates to the number of satisfied restraints that restrict the solution space to ≤0.01% of the conformations sampled, except for MBNL1 (k, 0.39%) and PTBP1 RRM2 (l, 0.07%) where this threshold was too stringent to output any permitted occupancy space. Visualisations are produced using UCSF Chimera.

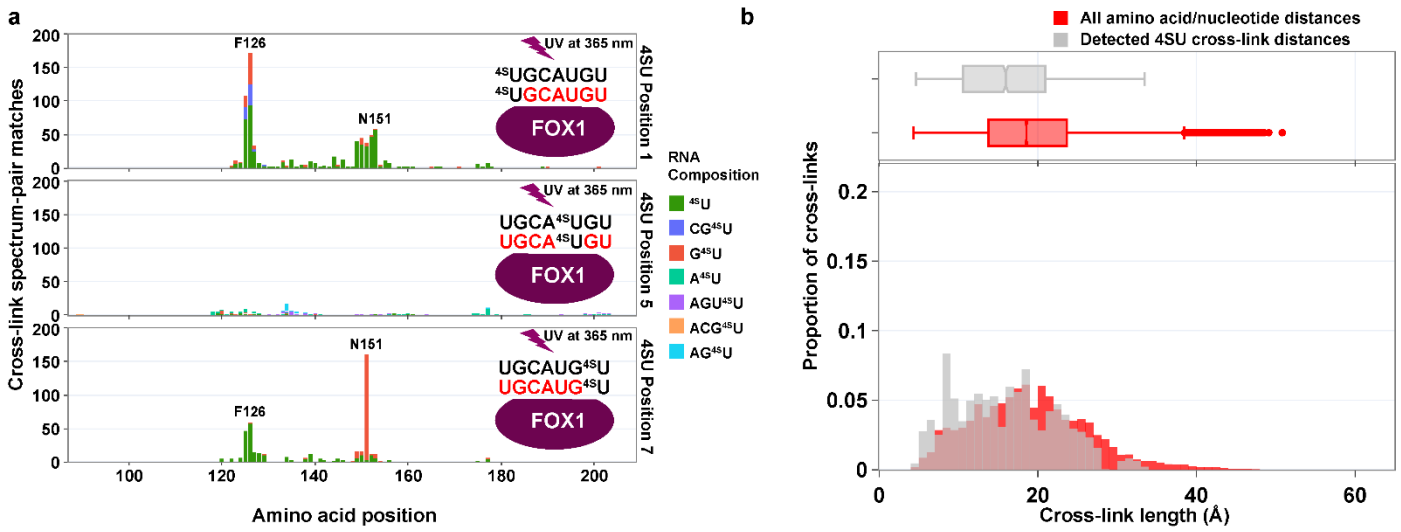

Figure 4: 4-thio-uracil leads to qualitative as well as quantitative changes to results compared with natural uracil.

a) CLIR-MS results for FOX1-FBE after replacing specific uracil positions with photoactive 4-thio-uracil.

b) Cross-link distances detected in the FOX1-FBE complex where 4-thio-uracil is used in place of natural uracil, as measured against the published structure for the protein-RNA complex. Boxes span Q1-Q3, with centre line representing the median. Whiskers represent upper and lower fences of data points, or highest/lowest values where all values are within this range. g) Measured distances (grey), n=2790; Control distances (red), n=18480.

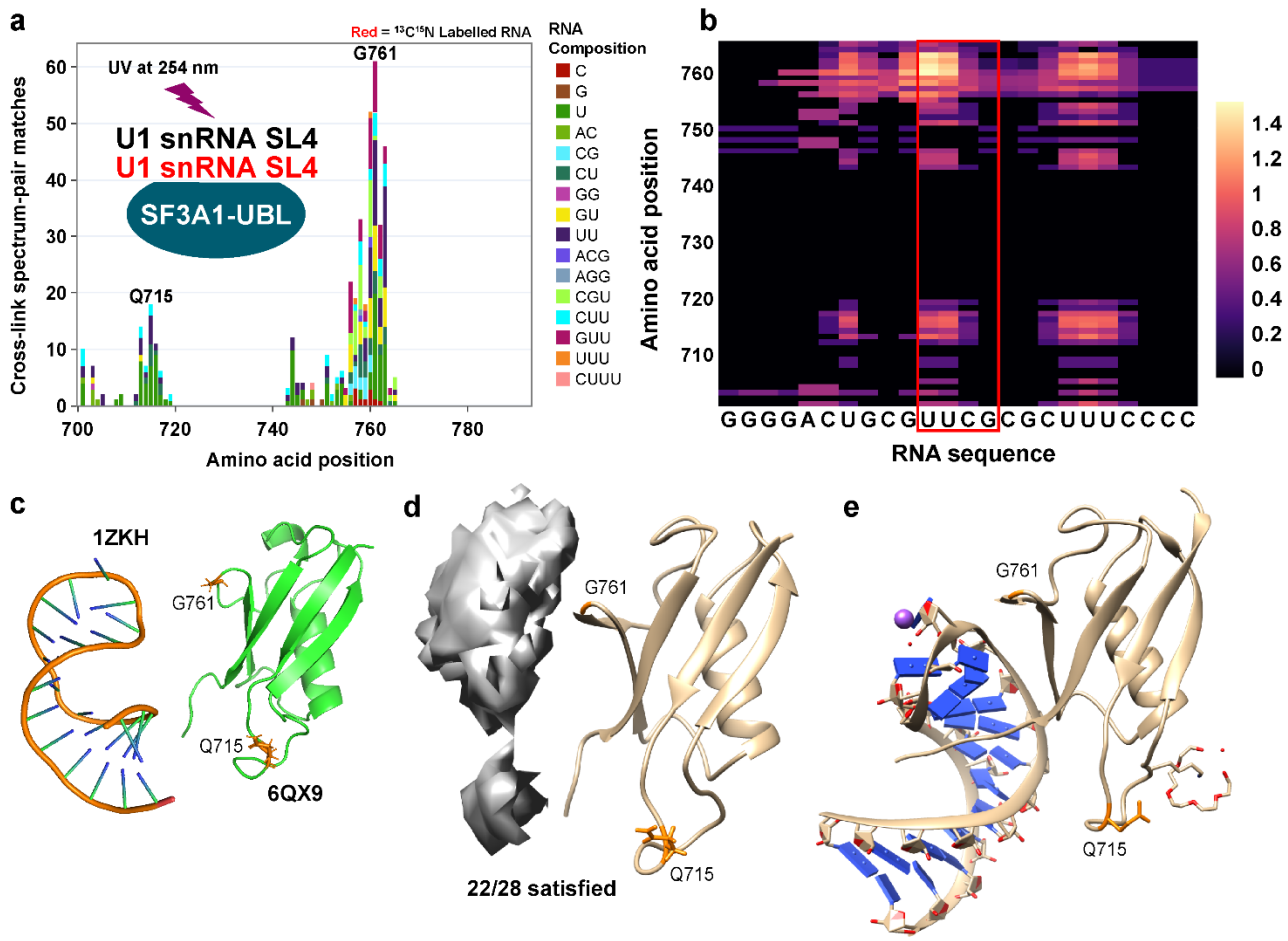

Figure 5: CLIR-MS derived cross-links describe a non-canonical protein-RNA interaction.

a) Redundant XLSMs from CLIR-MS analysis of SF3A1-UBL with U1 snRNA SL4. Overlaid, schematic representation of the protein-RNA complex used for this experiment.

b) Non-redundant amino acid position and RNA composition cross-link combinations, systematically overlaid for every XLSM to describe possible RNA interaction sites suggested by CLIR-MS data.

c) Previously published structural models of unbound SF3A1-UBL (PDB ID: 1ZKH) and the U1 snRNA SL4 (PDB ID: 6QX9). Highly cross-linked amino acids from (a) are coloured orange. Structures visualised with PyMOL (The PyMOL Molecular Graphics System, Version 2.3.2 Schrödinger, LLC).

d) Visualisation of the structural information contained in point-to-point distance restraints obtained from data in a) using the models in c) and DisVis. Occupancy space shown is for the number of

cross-links satisfied where the solution space is reduced to 0.01% of the total number of conformations sampled. Visualisation produced using UCSF Chimera.

e) Structure of the SF3A1-UBL interacting with the U1 snRNA SL4, determined using X-ray crystallography (described separately[63]). Visualisation produced using UCSF Chimera.

Tables

Table 1: Types of RNA-derived peptide adducts validated for use in further xQuest searches.

Losses from RNA-derived peptide adducts that were putatively detected in an open modification search of a CLIR-MS data set generated using the FOX1-FBE complex, and subsequently validated by xQuest search, specifying such losses in search parameters. Only adduct types that were validated by xQuest search in the FOX1-FBE complex were used for further xQuest searches.

Approximate Δm from whole mononucleotide (Da)	**Proposed Atomic Composition of Loss**	**Putative explanation**	**Detected in open search**	**Detected in FOX1-FBE xQuest validation search**	**Previous evidence in literature**
-2	H_2	Net loss of H_2	Yes	Yes	Kramer 2011[35]
-18	H_2O	Neutral loss of water (possibly from difference between 2' or 3' phosphate vs 2'-3' cyclic phosphate nucleotide terminus)	Yes	Yes	Panhale 2019[33], Dorn 2017[34], Kramer 2014[31]
-20	H_4O	Combined neutral loss of water and H_2	Yes	Yes	
-36	H_4O_2	Loss of 2x water	Yes	Yes	
-62	HPO_3+H_2O		Yes	Yes	

-64	HPO_2	Neutral loss from phosphate group, combined with oxidation (of a nearby amino acid)	Yes	Yes	
-80	HPO_3	Neutral loss from phosphate group	Yes	Yes	Panhale 2019[33], Kramer 2014[31], Shchepachev 2019[44]
-82	H_3PO_3	Combined neutral loss from phosphate group and of H_2	Yes	Yes	
-98	H_3PO_4	Combined neutral loss from phosphate group and of water	Yes	No	Panhale 2019[33], Kramer 2014[31]
-116	H_5PO_5	Combined neutral loss from phosphate group and 2x water	Yes	No	

Table 2: Putative RNA-derived peptide adduct types identified in open modification search not validated by xQuest search

Several other nucleotide-derived adducts are putatively identified in the open modification search of CLIR-MS data generated from the FOX1-FBE complex. However, these were not subjected to further validation with an xQuest search, nor are they routinely encoded as default loss products for future xQuest searches, owing to their practical incompatibility with the CLIR-MS approach, and the atoms carrying the isotope label in each nucleotide.

Free nucleobases other than guanine (i.e. a mononucleotide that has lost its phosphate and ribose components) are not identified, as in previous literature[33], as they have a mass below the minimum specified mass shift specified in the open search parameters used here (150 Da).

Mass addition to peptide (Da)	Putative explanation	Light form detected in open search	Heavy form detected in open search	Previous evidence in literature	Rationale for exclusion from further xQuest searches
151	Guanine nucleobase only (loss of sugar and phosphate)	Yes	N/A	Panhale 2019[86]	Nucleotides in FOX1-UGCAUGU, experiments are labelled with ^{13}C ribose. A nucleobase without ribose exhibits no shift.
712/727 (Light/ Heavy)	Former Uxx trinucleotide, light, with loss of 2 bases; residual backbones from 2 further nucleotides (including labelled ribose) remain	Yes	No	Kramer 2011[35], Panhale 2019[86]	Challenging to encode in parameters where ribose is labelled; similar information content to a mononucleotide without residual backbones
711/726 (Light/ Heavy)	Former Cxx trinucleotide, light, with loss of 2 bases; residual backbones from 2 further nucleotides (including labelled ribose) remain	Yes	No	Kramer 2011[35], Panhale 2019[86]	Challenging to encode in parameters where ribose is labelled; similar information content to a mononucleotide without residual backbones