Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins

Katharina Kramer[1,10], Timo Sachsenberg[2,3,10], Benedikt M Beckmann[4,9], Saadia Qamar[1], Kum-Loong Boon[5], Matthias W Hentze[4], Oliver Kohlbacher[2,3,6,7] & Henning Urlaub[1,8]

RNA-protein complexes play pivotal roles in many central biological processes. Although methods based on high-throughput sequencing have advanced our ability to identify the specific RNAs bound by a particular protein, there is a need for precise and systematic ways to identify RNA interaction sites on proteins. We have developed an experimental and computational workflow combining photo-induced cross-linking, high-resolution mass spectrometry and automated analysis of the resulting mass spectra for the identification of cross-linked peptides, cross-linking sites and the cross-linked RNA oligonucleotide moieties of such RNA-binding proteins. The workflow can be applied to any RNA-protein complex of interest or to whole proteomes. We applied the approach to human and yeast mRNA-protein complexes *in vitro* and *in vivo*, demonstrating its powerful utility by identifying 257 cross-linking sites on 124 distinct RNA-binding proteins. The open-source software pipeline developed for this purpose, RNP^{xl}, is available as part of the OpenMS project.

RNA molecules bind to proteins to form ribonucleoprotein complexes (RNPs). These are indispensable for the synthesis, stability, transport and activity of mRNAs[1] and noncoding RNAs[2,3]. RNA-binding proteins (RBPs) assume numerous functions in RNPs. RBPs can modulate or stabilize RNA structures, thereby making RNA catalytically active, for example, during pre-mRNA splicing[4]. RNA can also guide a catalytically active RBP to its destination; examples of this are microRNA– or long noncoding RNA–mediated translational control and epigenetic modulation[5,6]. RBPs are also involved in splicing and can recruit or repel other proteins, induce hydrolysis of RNA or protect RNA from degradation.

Consequently, much attention is devoted to the identification of proteins interacting directly with RNAs and the identification of the corresponding RNA sites. Classical pulldown experiments can provide evidence that a certain RNA and protein interact with one another, directly or indirectly. Studies using UV-induced cross-linking and, subsequently, mass spectrometry (MS) have identified proteins in direct contact with RNA[7–10]. Although the identification of RNAs bound by RBPs and their exact interaction sites benefits from high-throughput sequencing techniques after cross-linking, such as PAR-CLIP[11], so far no complementary approach has been described that allows for the identification of the corresponding cross-link sites within the proteins.

Here we report a methodology for purifying peptide–RNA oligonucleotide conjugates derived from *in vitro*– or *in vivo*–UV-irradiated RNP complexes that allows mass spectrometric sequencing of the cross-linked peptide and RNA moiety and the subsequent identification of the cross-linked peptides and RNA by automated database searching. We applied the strategy to three biological systems of human and yeast origin. Overall, we have identified 749 cross-links, which were mapped to 257 unique amino acids or protein regions in 124 different proteins. We demonstrated the unbiased nature of the approach by identifying proteins lacking classical RNA-binding motifs or annotated RNA-binding function.

RESULTS

Experimental workflow

We designed an integrated experimental and computational strategy for the automated identification of protein-RNA cross-linking sites, including the cross-linked amino acids and the cross-linked RNA moiety, i.e., the cross-linked nucleotides (**Fig. 1a**). The cross-linked RNA moiety consists of mono-, di- or trinucleotides, thus making it difficult to precisely assign which part of the RNA, or which RNA, is cross-linked (see Discussion). Isolated and UV-cross-linked RNPs and non-cross-linked controls are first hydrolyzed with endoproteinases and nucleases. The resulting

[1]Bioanalytical Mass Spectrometry Group, Department of Cellular Biochemistry, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany. [2]Center for Bioinformatics, University of Tübingen, Tübingen, Germany. [3]Department of Computer Science, University of Tübingen, Tübingen, Germany. [4]European Molecular Biology Laboratory, Heidelberg, Germany. [5]Department of Cellular Biochemistry, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany. [6]Quantitative Biology Center, University of Tübingen, Tübingen, Germany. [7]Faculty of Medicine, University of Tübingen, Tübingen, Germany. [8]Bioanalytics Research Group, Department of Clinical Chemistry, University Medical Center, Göttingen, Germany. [9]Present address: Molecular Infection Biology Research Group, Integrative Research Institute (IRI) for the Life Sciences, Humboldt University Berlin, Berlin, Germany. [10]These authors contributed equally to this work. Correspondence should be addressed to H.U. (henning.urlaub@mpibpc.mpg.de) or O.K. (oliver.kohlbacher@uni-tuebingen.de).

RECEIVED 13 JANUARY; ACCEPTED 21 JULY; PUBLISHED ONLINE 31 AUGUST 2014; DOI:10.1038/NMETH.3092

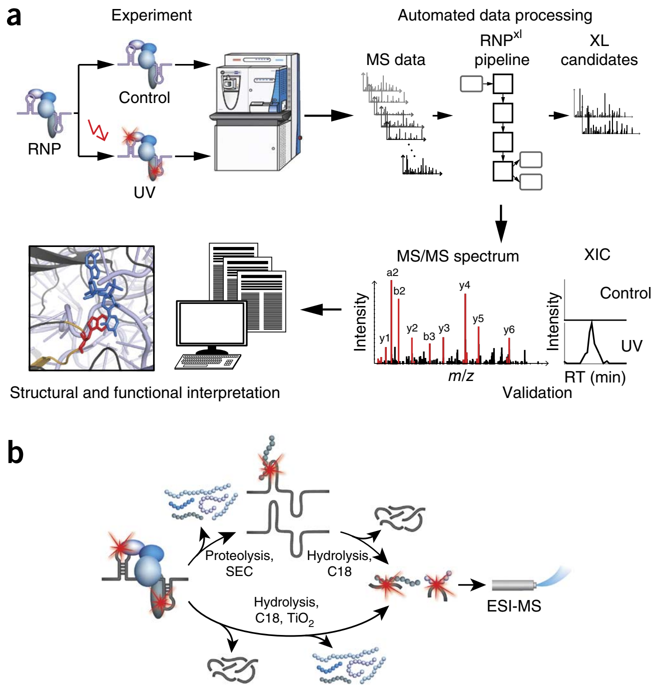

Figure 1 | Overview of the procedure and experimental workflow. (**a**) Schematic outline of the entire approach. UV-irradiated protein-RNA complexes are processed for LC-ESI-MS/MS analysis in parallel with a non-irradiated control. Mass spectra are subjected to a data analysis workflow that yields potential cross-linked peptides. Results are validated by comparison of extracted-ion chromatogram (XIC; RT, retention time) intensities in control versus UV-irradiated sample and evaluation of expected and observed fragmentation patterns in the MS/MS spectrum. Identified cross-links are compared to published RNA-binding functionality and structural data when available. (**b**) Isolation of cross-linked heteroconjugates. Isolated cross-linked protein-RNA complexes are enriched by size-exclusion chromatography (SEC) followed by hydrolysis and reversed-phase C18 chromatography (upper workflow), whereas complexes from *in vivo*–cross-linked whole cells are hydrolyzed and then enriched by C18 chromatography and TiO_2 solid-phase extraction (lower workflow). Enriched peptide–RNA oligonucleotide heteroconjugates from both experimental workflows are then directly analyzed by LC-ESI-MS/MS.

peptide–RNA oligonucleotide heteroconjugates are enriched (**Fig. 1b**) and analyzed by electrospray-ionization (ESI) MS on Orbitrap instruments in data-dependent acquisition mode. MS data are submitted to a dedicated data analysis workflow based on the OpenMS platform[12,13] (http://www.openms.de/). In contrast to conventional MS search engines, the computational workflow introduced here, named RNP^{xl}, removes all tandem mass spectra (MS/MS spectra) not corresponding to cross-linked species, calculates masses of possible peptide-RNA conjugates and searches MS/MS spectra against a sequence database with the OMSSA[14] algorithm. Fragment spectra assigned to peptide–RNA oligonucleotide cross-links are annotated in order to identify the cross-linked amino acids and nucleotides.

We applied this approach independently to three biological systems: (i) RNPs derived from incubation of a transcribed pre-mRNA tagged with an aptamer recognized by the MS2 bacteriophage coat protein and reconstituted *in vitro* with human nuclear extract, (ii) purified pre-mRNA or mRNA (hereafter '(pre-)mRNA') complexes from yeast, obtained by affinity capture of the nuclear cap-binding protein subunit 2 (Cbp20), and (iii) whole yeast cells metabolically labeled with 4-thiouridine (4SU) and cross-linked *in vivo*. UV-based protein-RNA cross-linking in the first two systems was carried out at 254 nm, the protein moiety was digested with trypsin under denaturing conditions and non-cross-linked peptides were removed by size-exclusion chromatography (SEC). The RNA pool was hydrolyzed with endonucleases, non-cross-linked oligonucleotides were removed by reversed-phase chromatography and remaining cross-linked peptide-oligonucleotide conjugates were subjected to liquid chromatography–tandem MS (LC-MS/MS) (**Fig. 1b**). In the whole yeast cells, UV irradiation was performed at 365 nm *in vivo*, cells were lysed and polyadenylated mRNA with its associated RBPs was recovered by oligo(dT) affinity selection. After digestion of the cross-linked protein and RNA moiety, non-cross-linked oligonucleotides were removed by reversed-phase chromatography, and cross-linked conjugates were enriched with a TiO_2 matrix before LC-MS/MS analysis (**Fig. 1b** and Online Methods).

In all experiments with unlabeled RNA, UV-irradiated and non-irradiated samples were treated in parallel (**Fig. 1a**) to reliably distinguish between cross-linked peptide–RNA oligonucleotides and false positives (for example, residual non-cross-linked species). The experimental workflow resulted in two raw data files per experiment (cross-linked and control, except for yeast 4SU-labeled RNPs; Online Methods), which were then analyzed computationally. A typical data set from a cross-linked sample yields 5,000–10,000 MS/MS spectra.

Mass spectrometry data analysis

In previous studies with isolated proteins, cross-linked peptide-oligonucleotides have been shown to yield information-rich fragments of the cross-linked peptide moiety upon MS-based sequencing. These fragments can be used to identify the cross-linked peptide[15]. Although, in principle, the resulting spectra could be compared to theoretical spectra in a conventional database search approach (for example, OMSSA[14] or MASCOT[16]), in the case of peptide-oligonucleotide cross-links, such a search is complicated because database search engines require defined modification masses. The cross-linked RNA moiety is not known and can encompass different combinations of nucleotides, including their derivatives (such as those that lose water or terminal phosphate groups). Moreover, despite the biochemical purification procedures, the overall number of MS/MS spectra is quite large compared to the number of spectra containing cross-link information.

We designed RNP^{xl} to reduce data to the relevant spectra, search these spectra against entire proteomes and annotate spectra of cross-linked peptide–RNA oligonucleotide conjugates (**Fig. 2a**, Online Methods, **Supplementary Software** and **Supplementary Note**; http://www.openms.de/RNPxl). Briefly, MS data are prepared for analysis (which involves conversion to mzML format, peak calling, and alignment of chromatographic retention times of data from UV- and non-irradiated samples; **Supplementary Fig. 1**). Application of the different filter algorithms reduced the originally acquired mass spectra successively (**Supplementary Fig. 2**). These filters include (i) searching against a target-decoy database to remove spectra corresponding to non-cross-linked peptides with a false discovery rate of peptide-to-spectrum matches (FDR) of 1%

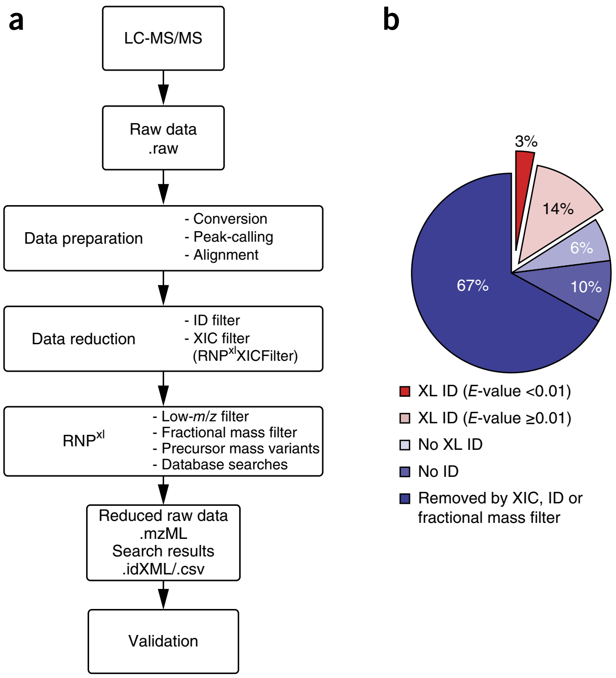

Figure 2 | Data analysis workflow. (**a**) Outline of the RNP^{xl} data analysis pipeline. Raw LC-ESI-MS/MS data are converted into an open mass spectrometry format (.mzML) and peak called; retention-time alignment between the UV sample and the non-irradiated control is performed. The overall amount of data is reduced by removing MS/MS spectra of confidently identified non-cross-linked peptides (ID filter), spectra appearing in both the UV-irradiated sample and control with comparable intensities (XIC filter with RNP^{xl}XIC), and spectra with small precursor masses (low-*m/z* filter) and residual short oligonucleotides (fractional mass filter). Finally, RNP^{xl} creates precursor mass variants, submits data into the search engine and summarizes the search results. (**b**) Results of the data analysis procedure for a single data set of yeast RNA-binding proteins. The XIC, ID and fractional mass filters excluded approximately two-thirds (67%) of the overall 9,728 fragment spectra. Additionally, 16% of the spectra were disregarded, as these did not yield any database search result for a cross-linked heteroconjugate. Of the remaining 17% potential cross-link candidates, 14% had a low score (*E*-value of ≥0.01), yielding a final list of 317 (3%) potential cross-link candidates for manual validation.

(the 'ID' filter); (ii) an extracted-ion chromatogram (XIC)-based filter to remove MS/MS spectra of precursors present in the UV- and non-irradiated control in a certain retention time window, and (iii) a fractional mass filter[15,17] to eliminate pure RNA oligonucleotides. Finally, a precursor mass variant list is generated, in which masses of all possible nucleotide combinations of putatively cross-linked oligonucleotides are subtracted from the experimental precursor mass. The resulting data are searched against the UniProt[18] database containing the respective proteome (human or yeast; **Supplementary Fig. 3**) using OMSSA[14].

In a single representative data set of UV-cross-linked RNPs from yeast (**Fig. 2b**), of an initial 9,728 spectra, two-thirds were removed by the ID, XIC and fractional mass filters. The ID filter removed 2,823 spectra (29%), the XIC filter removed 3,313 spectra (34%) and the fractional mass filter removed 335 spectra (3%) when applied in that order. Another 10% of the spectra remained completely unassigned, and 6% were excluded because they matched non-cross-linked peptides with medium or low confidence (FDR above 1%). 17% of the submitted MS/MS spectra represented potential peptide-RNA cross-links, and among these, 3% (317 spectra) exhibited an *E* value (expected number of random database hits with equal or better score) better than 0.01 and were kept as potential peptide-RNA cross-links (**Supplementary Data**).

Protein-RNA cross-links

From the human protein-RNA complexes, we identified 189 cross-links matching 60 tryptic peptides with their cross-linked nucleotides or oligonucleotides (hereafter '(oligo)nucleotides'). In 79% of the cases, the cross-linked nucleotide could be identified. In half of the 60 tryptic peptides, we identified the cross-linked amino acid. The cross-linked peptides were mapped to 35 different proteins. A large majority of the cross-linked peptides (54) lie in known RNA-binding motifs such as RNA-recognition motifs (RRMs) and K homology (KH) domains (**Fig. 3a**, **Supplementary Table 1**, **Supplementary Figs. 4** and **5** and **Supplementary Data** spectra H01–H60).

In the UV-irradiated yeast RNPs (isolated by affinity purification of Cbp20), we identified 184 peptide–RNA oligonucleotide cross-links, which mapped to 64 tryptic peptides with their cross-linked (oligo)nucleotides. In 89% of the cases, the cross-linked nucleotide could be identified (**Supplementary Table 2**, **Supplementary Fig. 6** and **Supplementary Data** spectra Y01–Y64). In 39 of the 64 tryptic peptides, the cross-linked amino acid could be pinpointed. The cross-links involved 49 different proteins, the majority derived from the ribosome (137 cross-links in 34 ribosomal proteins). Nonribosomal proteins included nucleolar proteins 3 and 13 (Npl3 and Nop13), polyadenylate-binding protein Pab1 and single-stranded nucleic acid–binding protein Sbp1 with cross-links located in the RRM (**Fig. 3b**). Of particular interest are three enzymes (adenosylhomocysteinase Sah1, alcohol dehydrogenase Adh1 and glyceraldehyde-3-phosphate dehydrogenase Tdh2) containing a Rossmann fold[19]. These and five additional proteins (enolase Eno1, inorganic pyrophosphatase Ipp1, peroxiredoxin Tsa1, phosphoglycerate kinase Pgk1 and pyruvate kinase Cdc19) were not known to be classical RBPs (**Fig. 3b**).

We obtained the broadest spectrum of proteins cross-linked to RNA after isolation of polyadenylated mRNA from UV-irradiated yeast cells that had incorporated 4SU into mRNA (**Supplementary Table 3**, **Supplementary Figs. 7** and **8** and **Supplementary Data** spectra Y65–Y104). We identified 376 cross-links, which corresponded to 133 unique cross-linking sites or regions in 57 different proteins. In all of the cases, the cross-linked nucleotide could be identified unambiguously. 161 cross-links were found in ribosomal and 215 in nonribosomal proteins including metabolic enzymes (peptidyl-prolyl *cis-trans* isomerase Cpr1 and phosphoglycerate kinase Pgk1), DNA-binding proteins (such as nonhistone chromosomal protein 6A and 6B (NHP6A and NHP6B), endonuclease PI-SceI (Vma1) and multiprotein bridging factor 1 Mbf1), a nucleotide-binding protein (elongation factor 1 alpha Tef1) and RBPs (**Fig. 3c**). Among the latter are RNA helicases (Dbp1 and Sub2), proteins containing RRMs (such as Pub1), KH motifs (heterogeneous nuclear RNP K-like protein 2 Hek2 and Scp160) and Pumilio repeats (Puf3), as well as proteins containing motifs that are not frequently associated with RNA binding, i.e., the coiled-coil domain (Bfr1), HTA-La-type RNA-binding domain (Sro9) and RBPs with not-yet well-defined motifs (Gag-p49 derived from *TY1A-LR4*).

Our data also reveal structural details of the protein-RNA interaction. Analyzing four cross-linked peptides, we confirmed that the cross-link locations agree with the known crystal structures of three protein-RNA complexes (**Fig. 4**). U2AF 65-kDa subunit (U2AF65) is an RNA-binding protein that contains three RRMs and interacts with the polypyrimidine tract on pre-mRNAs[20]. The three-dimensional (3D) structure of the first two RRMs in complex with poly(U) RNA has been described (PDB ID: 2YH1; ref. 20). The MS/MS spectra narrowed down the cross-linking sites to Leu261 or Phe262 and Phe199, respectively (**Fig. 4a**). Phe262 corresponds to the conserved aromatic residue in the RNP2 consensus sequence, and Phe199 to the second conserved aromatic residue in the RNP1 consensus sequence (**Supplementary Fig. 4**). The location of the cross-linked peptides agrees exactly

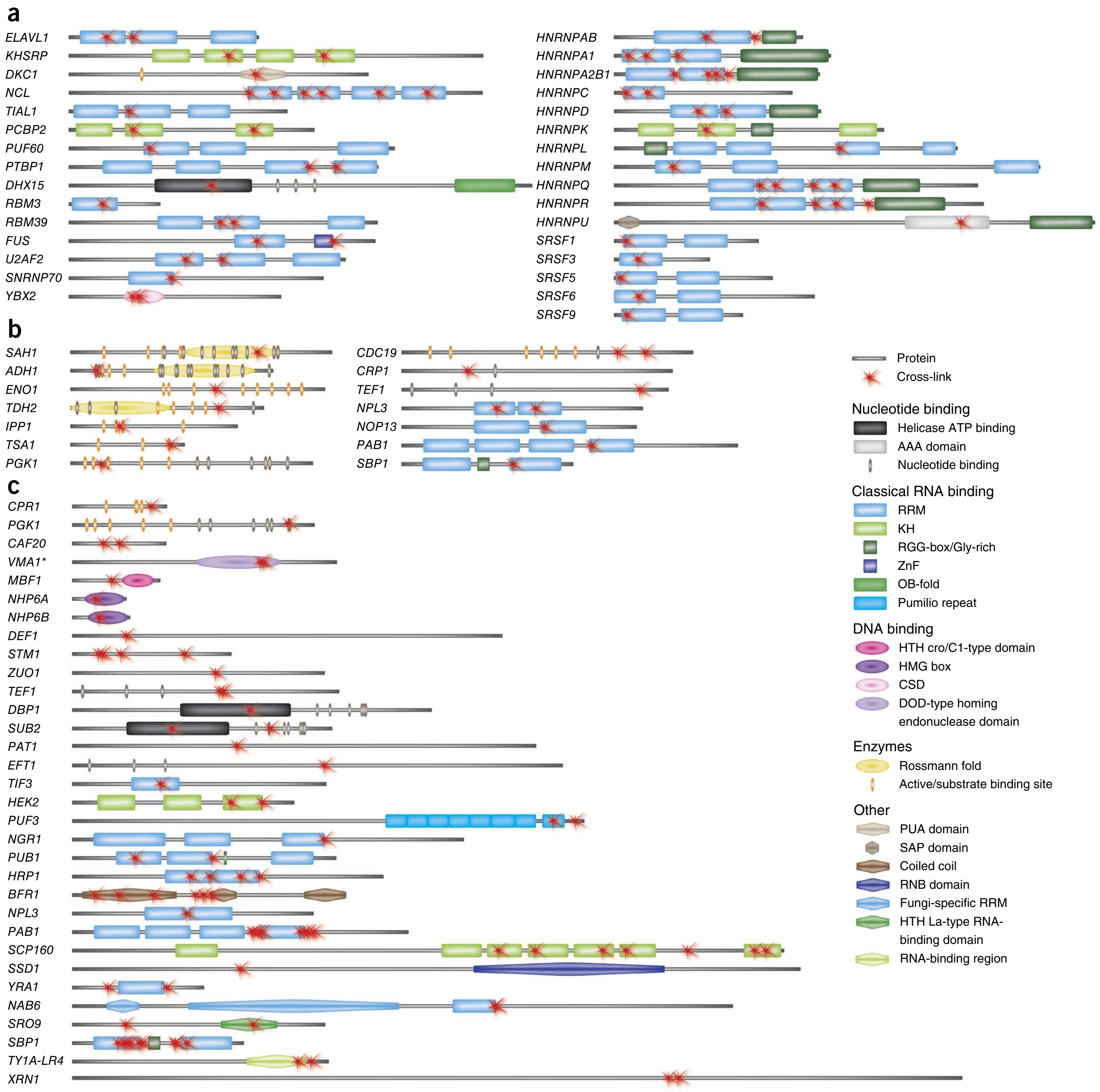

Figure 3 | Distribution of cross-linking sites in identified RNA-binding proteins with annotated domain structure. (**a**–**c**) Human proteins (**a**), yeast proteins isolated with TAP-tagged Cbp20 (**b**) and yeast proteins cross-linked to 4SU and isolated with oligo(dT) (**c**) are shown, represented by their corresponding gene names to avoid ambiguity. Ribosomal proteins are not included. Putative RNA- or DNA-binding domains, nucleotide-binding sites, and active/substrate binding sites are given if annotated in protein databases. All appearing domains and sites are listed on the lower right with their assigned symbols. Information is derived from the UniProt database and supplemented with details from the US National Center for Biotechnology Information (NCBI), Pfam, Superfamily and CATH (class, architecture, topology, homology) databases (**Supplementary Tables 1–3**). Yeast gene *VMA1* (asterisk) encodes V-type proton ATPase catalytic subunit A and endonuclease PI-SceI, but only the latter is shown because it contains the cross-linked regions.

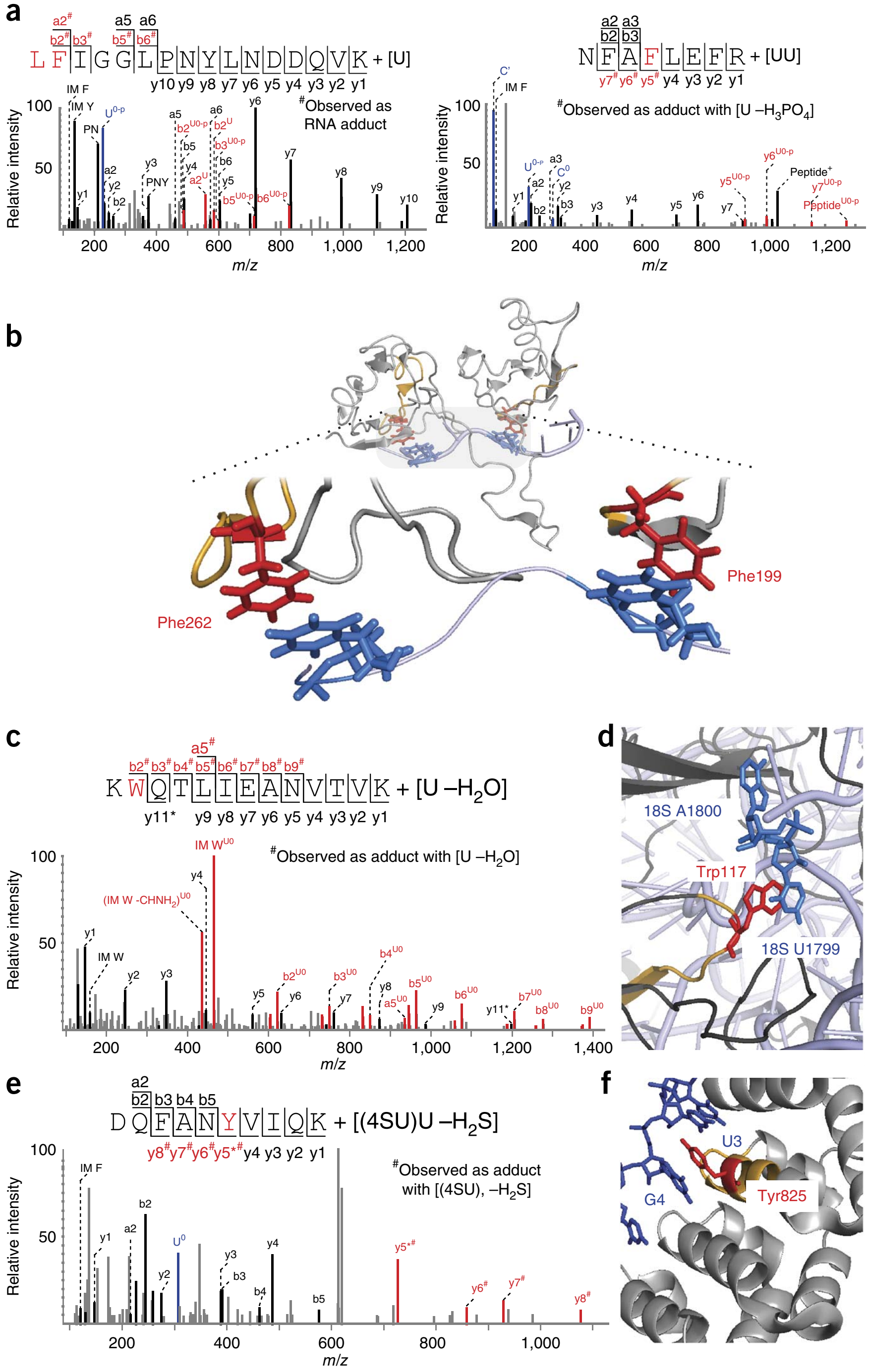

Figure 4 | MS/MS fragment spectra of cross-linked heteroconjugates and structural interpretation. (**a**,**c**,**e**) Manually validated MS/MS spectra. Regular peptide fragments are shown in black, RNA fragments in blue, and specific fragment ions derived from peptide-RNA cross-links in red ($U^{0\text{-p}}$, U $-H_3PO_4$; U^0, U $-H_2O$). (**b**,**d**,**f**) Structural interpretation of identified cross-links. Proteins are shown in gray, RNA (nucleotides) in blue, cross-linked peptides in orange, and amino acids in red. (**a**) Left, fragment spectrum of peptide LFIGGLPNYLNDDQVK from human splicing factor U2AF 65-kDa subunit (Leu261–Lys276) cross-linked to U via Leu261 or Phe262. Right, fragment spectrum of U2AF peptide NFAFLEFR (Asn196–Arg203) cross-linked to a UU dinucleotide through F199. (**b**) Phe262 and Phe199 are found in close spatial proximity to uracil in the structure of RRM1 and RRM2 of U2AF with a poly(U) oligonucleotide[20]. (**c**) MS/MS fragment spectrum of 40S ribosomal protein S1 peptide KWQTLIENANVTVK cross-linked to [U $-H_2O$] via Trp117. (**d**) In the structure of the yeast ribosome[21], S1 Trp117 is in close spatial proximity to U1799 of the 18S ribosomal RNA. (**e**) Fragment spectrum of peptide DQFANYVIQK from mRNA-binding protein Puf3 (Asp820–Lys829) cross-linked to [(4SU)U $-H_2S$]. The shift of the y series by a fragment of the cross-linked 4SU base identifies Tyr825 as the cross-linked amino acid. (**f**) Structure of the RNA-binding domain of Puf3 to a recognition sequence[22]. Tyr825 stacks between U3 and G4 of the cocrystallized oligonucleotide.

with the available structure of RRM1 and RRM2 of U2AF65 together with an 8-mer poly(U) (**Fig. 4b**). Yeast ribosomal protein S1 was found cross-linked to uridine through a tryptophan residue (**Fig. 4c**). Within the crystal structure of the yeast ribosome[21] (PDB IDs: 3U5F, 3U5G, 3U5H and 3U5I), this tryptophan residue (Trp117) stacks perfectly with U1799 of the 18S rRNA (**Fig. 4d**). In mRNA-binding protein Puf3, residue Tyr825 was found cross-linked (**Fig. 4e**). In the structure of the protein's RNA-binding domain to one of its recognition sequences (PDB ID: 3K49; ref. 22), this residue stacks between U3 and G4 of the cocrystallized oligonucleotide (**Fig. 4f**).

DISCUSSION

Our method combines UV cross-linking with MS and database search to identify RNA-protein contact sites at the peptide and amino acid as well as at the nucleotide level in the context—and against the background—of whole-cell extracts. Earlier studies established the specificity of combining UV cross-linking and MS in structural investigations of moderately complex protein-RNA complexes[15,23–27] and also through mutagenesis studies of identified cross-linked amino acids in single RNPs[28,29]. More complex samples pose considerably greater challenges to MS data analysis.

Peptides cross-linked to RNA oligonucleotides cannot be comprehensively identified by conventional database search engines, as this would limit the identification of cross-linked peptides to a very restricted set of possible modifications (**Supplementary Fig. 9**). Our purification procedure for peptide–RNA oligonucleotide cross-links used various endonucleases to hydrolyze intact RNA. Consequently, the length distribution, the composition and possible modifications of the cross-linked RNA oligonucleotides could not be predicted. The number of potential RNA adducts was too large to allow search strategies analogous to the identification of post-translational modifications with conventional database searches. A conventional, limited database search for cross-linked species might nonetheless still be used as an initial survey in order to roughly anticipate the number of cross-links. Even then, a conventional database search (for example, taking N-terminal uridinylation into account) only led to limited results, i.e., less than two-thirds of cross-linked species were identified compared to our dedicated RNP^{xl} workflow (**Supplementary Table 4**).

RNP^{xl} allows the generation of freely defined precursor variants containing modified and nonmodified nucleotides. In practice, the generation of precursor variants up to tetranucleotides will be sufficient for most analyses. Peptides cross-linked to longer RNA oligonucleotides are difficult to identify, as MS/MS spectra are dominated by RNA fragment ions, so that no (or very poor) sequence information on the cross-linked peptide can be obtained[30]. Nonetheless, cross-links to tri- or tetranucleotides can allow location of the cross-linking site on the RNA if the oligonucleotide composition identified matches a unique RNA sequence[25].

Our data analysis yielded several insights. (i) Nearly all amino acid residues except aspartic acid, asparagine, glutamic acid and glutamine were found to be cross-linked to nucleotides (**Supplementary Tables 1–3**). (ii) The cross-linked nucleotide was usually U or 4SU. (iii) Cross-linking to 4SU systematically showed a loss of H_2S in the precursor mass, and fragment ions from the peptide moiety often showed an adduct mass of 94 Da, corresponding to a uracil derivative that has lost H_2S (ref. 15). (iv) Several cross-links, particularly from yeast ribosomal proteins copurified with (pre-)mRNA by TAP-tagged Cbp20 protein, carried an additional mass of 151.9938 Da. Cross-links with an additional 152 Da were first reported in a cross-linking study of snurportin 1 to U1 small nuclear RNA[26] and involve almost exclusively cysteine-containing peptides. We note that this adduct mass is associated with the presence of dithiothreitol (DTT) in the sample. DTT may promote formation of cross-links between cysteine residues and RNA bases under UV irradiation at 254 nm (personal communication, U. Zaman, Max Planck Institute for Biophysical Chemistry).

Although we used different purification strategies for RNP isolation, the comparison of the cross-linked peptides derived from yeast RNP complexes irradiated at 254 nm (nonsubstituted RNA) and 365 nm (4SU-substituted RNA) demonstrates that the two cross-linking methods were complementary for numerous proteins, as has been observed before[7]. In Pgk1, Tef1, Npl3 and several ribosomal proteins, different regions of the proteins were identified as being cross-linked to RNA, whereas in Pab1, Sbp1 and other ribosomal proteins, the same amino acids or peptides have been identified as cross-linking to RNA (**Supplementary Table 5**).

Our results highlight the utility of a UV-based irradiation approach to identify direct contact sites between proteins and RNA, and we substantially extend the scope of previous studies in which entire proteins—but not the specific cross-linking site—were identified by MS[7–10]. The majority of cross-linking sites identified reside in known RNA-binding domains, such as RRMs and KH motifs (**Fig. 3** and **Supplementary Tables 1–3**). Among the 20 heterogeneous nuclear RNP proteins in the human database, we found 13 (more than 60%) to be cross-linked. In the 44 canonical RNA-binding motifs (RRM and KH motifs) of these proteins, we identified 25 distinct cross-linked peptides or amino acids; these account for more than 55% of the canonical RNA-binding motifs in these proteins. The cross-link sites also included less frequently described RNA-binding motifs such as Pumilio repeats and WD domains, as well as other sequence motifs such as the AAA domain, the RanBP-type zinc finger, the PUA domain, coiled-coil domains, the HMG box and the DOD-type homing endonuclease domain. Cross-linking sites within several metabolic enzymes underscore the enzymes' ability to interact with RNA[7]. The RNA cross-links occurred within the ATP- or substrate-binding sites, the Rossmann fold and other regions of these proteins.

The comparison of the cross-linking sites with available 3D structures demonstrates the structural specificity of the cross-linking approach while at the same time proving useful for predicting (novel) RNA-binding sites. In the case of DNA-binding proteins, the 3D structures of NHP6A in complex with dsDNA[31] and of domain I of the *Saccharomyces cerevisiae* homing endonuclease PI-SceI[32] both show amino acids that were found to be cross-linked to RNA (Arg40 and Tyr328) while being located in the DNA-binding domain, thus suggesting a dual function of these proteins. In the 3D structure of the 80S yeast ribosome[21], most ribosomal proteins—with their cross-linking sites—are found in close proximity to the 28S or 18S rRNA. The cross-linking sites identified in receptor for activated C kinase 1 (RACK1), and in the ribosome-associated proteins Stm1 (ref. 21) and Zuo1 (ref. 33), are not proximal to rRNA, suggesting alternative conformations of translating ribosomes on mRNA or involvement of these proteins in direct mRNA binding. In total, we have pinpointed 39 cross-linking sites in 26 different proteins (about one-third of yeast's 79 ribosomal proteins).

Although the quality of computational predictions of RNA-protein interactions has greatly improved in recent years, there are substantial differences between our experimental approach and such prediction methods. Computationally predicted RNA-binding sites describe potential interactions, which are not necessarily realized in a specific protein-RNA complex. Because our approach is not biased toward known RNA-binding domains, it provides a basis for improving computational predictions of RNA-binding motifs, as in metabolic enzymes and transcription factors, in proteins that contain more than one RNA interaction site[34], or in those proteins that form a composite RNA interaction site through protein-protein interactions[35]. Our method is complementary to the now well-established RNA-sequencing approaches that identify cross-linked RNAs and nucleotides (for example, PAR-CLIP[11]). In contrast to deep-sequencing approaches, cross-linked peptides cannot be amplified, and the identification of corresponding cross-linking sites within proteins relies on adequate enrichment procedures and the sensitivity of MS instruments. The number of cross-links we identified and our ability to identify cross-linked amino acids (for example, proline and glycine residues) that had not been previously identified reflect the recent advancements in MS instrumentation. The steady improvement of mass spectrometers can be expected to enable even more comprehensive identification of cross-linking sites with our method.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. PRIDE: mass spectrometry data have been deposited with the data set identifier PXD000513.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We dedicate this article to the memory of our colleague and friend Andreas Bertsch. The authors thank M. Raabe, U. Pleßmann and T. Conrad for technical assistance, U. Zaman for providing unpublished data, R. Hofele for helpful discussions, R. Lührmann for support and providing infrastructure, S. Klinge for help with the yeast ribosome structure in PyMol, and S. Aiche and C. Bielow for discussions on workflow implementation. This work was supported by a German Research Foundation (DFG) grant to H.U. (SFB860, INST 186/859-1), a Higher Education Commission, Pakistan (HEC)/German Academic Exchange Service (DAAD) stipend to S.Q., funding from the European Research Council (ERC) under the Union's Seventh Framework Programme (FP7/2007-2013)/ERC-2011-ADG_20110310 to M.W.H. and a Marie Curie Fellowship (FP7/2007-2013)/MC-IEF-301031 to B.M.B.

AUTHOR CONTRIBUTIONS

K.K., B.M.B., S.Q., K.-L.B., M.W.H. and H.U. designed biochemical experiments. K.-L.B. and K.K. designed and transformed the yeast strain. K.K. and B.M.B. carried out experiments for the yeast systems; K.K. analyzed the resulting data. S.Q. performed experiments in the human system; K.K. and S.Q. analyzed the resulting data. K.K., T.S., O.K. and H.U. designed data analysis strategy; T.S. implemented it. K.K. and T.S. tested the data analysis tools. K.K., T.S., B.M.B., M.W.H., O.K. and H.U. wrote the paper; all authors contributed comments throughout all stages of the manuscript. K.K., T.S. and S.Q. compiled the supplementary materials.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Glisovic, T., Bachorik, J.L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**, 1977–1986 (2008).
2. Matera, A.G., Terns, R.M. & Terns, M.P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.* **8**, 209–220 (2007).
3. Yates, L.A., Norbury, C.J. & Gilbert, R.J. The long and short of microRNA. *Cell* **153**, 516–519 (2013).
4. van der Feltz, C., Anthony, K., Brilot, A. & Pomeranz Krummel, D.A. Architecture of the spliceosome. *Biochemistry* **51**, 3321–3333 (2012).
5. Sabin, L.R., Delás, M.J. & Hannon, G.J. Dogma derailed: the many influences of RNA on the genome. *Mol. Cell* **49**, 783–794 (2013).
6. Mercer, T.R. & Mattick, J.S. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* **20**, 300–307 (2013).
7. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).
8. Baltz, A.G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46**, 674–690 (2012).
9. Mitchell, S.F., Jain, S., She, M. & Parker, R. Global analysis of yeast mRNPs. *Nat. Struct. Mol. Biol.* **20**, 127–133 (2013).
10. Klass, D.M. *et al.* Quantitative proteomic analysis reveals concurrent RNA-protein interactions and identifies new RNA-binding proteins in *Saccharomyces cerevisiae*. *Genome Res.* **23**, 1028–1038 (2013).
11. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).
12. Kohlbacher, O. *et al.* TOPP—the OpenMS proteomics pipeline. *Bioinformatics* **23**, e191–e197 (2007).
13. Sturm, M. *et al.* OpenMS—an open-source software framework for mass spectrometry. *BMC Bioinformatics* **9**, 163 (2008).
14. Geer, L.Y. *et al.* Open mass spectrometry search algorithm. *J. Proteome Res.* **3**, 958–964 (2004).
15. 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).
16. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
17. Pourshahian, S. & Limbach, P.A. Application of fractional mass for the identification of peptide-oligonucleotide cross-links by mass spectrometry. *J. Mass Spectrom.* **43**, 1081–1088 (2008).
18. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **40**, D71–D75 (2012).
19. Hentze, M.W. Enzymes as RNA-binding proteins: a role for (di)nucleotide-binding domains? *Trends Biochem. Sci.* **19**, 101–103 (1994).
20. Mackereth, C.D. *et al.* Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature* **475**, 408–411 (2011).
21. Ben-Shem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334**, 1524–1529 (2011).
22. Zhu, D., Stumpf, C.R., Krahn, J.M., Wickens, M. & Hall, T.M. A 5′ cytosine binding pocket in Puf3p specifies regulation of mitochondrial mRNAs. *Proc. Natl. Acad. Sci. USA* **106**, 20192–20197 (2009).
23. Urlaub, H., Thiede, B., Müller, E.C., Brimacombe, R. & Wittmann-Liebold, B. Identification and sequence analysis of contact sites between ribosomal proteins and rRNA in *Escherichia coli* 30 S subunits by a new approach using matrix-assisted laser desorption/ionization-mass spectrometry combined with N-terminal microsequencing. *J. Biol. Chem.* **272**, 14547–14555 (1997).
24. Kühn-Hölsken, E., Dybkov, O., Sander, B., Lührmann, R. & Urlaub, H. Improved identification of enriched peptide RNA cross-links from ribonucleoprotein particles (RNPs) by mass spectrometry. *Nucleic Acids Res.* **35**, e95 (2007).
25. Luo, X. *et al.* Structural and functional analysis of the *E. coli* NusB-S10 transcription antitermination complex. *Mol. Cell* **32**, 791–802 (2008).
26. Kühn-Hölsken, E. *et al.* Mapping the binding site of snurportin 1 on native U1 snRNP by cross-linking and mass spectrometry. *Nucleic Acids Res.* **38**, 5581–5593 (2010).
27. Mozaffari-Jovin, S. *et al.* The Prp8 RNase H-like domain inhibits Brr2-mediated U4/U6 snRNA unwinding by blocking Brr2 loading onto the U4 snRNA. *Genes Dev.* **26**, 2422–2434 (2012).
28. Ghalei, H., Hsiao, H.H., Urlaub, H., Wahl, M.C. & Watkins, N.J. A novel Nop5-sRNA interaction that is required for efficient archaeal box C/D sRNP formation. *RNA* **16**, 2341–2348 (2010).
29. Müller, M. *et al.* A cytoplasmic complex mediates specific mRNA recognition and localization in yeast. *PLoS Biol.* **9**, e1000611 (2011).
30. Schmidt, C., Kramer, K. & Urlaub, H. Investigation of protein-RNA interactions by mass spectrometry—techniques and applications. *J. Proteomics* **75**, 3478–3494 (2012).
31. Allain, F.H. *et al.* Solution structure of the HMG protein NHP6A and its interaction with DNA reveals the structural determinants for non-sequence-specific binding. *EMBO J.* **18**, 2563–2579 (1999).
32. Werner, E., Wende, W., Pingoud, A. & Heinemann, U. High resolution crystal structure of domain I of the *Saccharomyces cerevisiae* homing endonuclease PI-SceI. *Nucleic Acids Res.* **30**, 3962–3971 (2002).
33. Leidig, C. *et al.* Structural characterization of a eukaryotic chaperone—the ribosome-associated complex. *Nat. Struct. Mol. Biol.* **20**, 23–28 (2013).
34. Schmitzová, J. *et al.* Crystal structure of Cwc2 reveals a novel architecture of a multipartite RNA-binding protein. *EMBO J.* **31**, 2222–2234 (2012).
35. Urlaub, H., Raker, V.A., Kostka, S. & Lührmann, R. Sm protein-Sm site RNA interactions within the inner ring of the spliceosomal snRNP core structure. *EMBO J.* **20**, 187–196 (2001).

ONLINE METHODS

Assembly and isolation of human protein-RNA complexes. *In vitro*–transcribed MS2-tagged PM5 pre-mRNA[36] was prebound with MS2-MBP fusion protein[37]and incubated for 30 min at 0 °C with 10 ml HeLa nuclear extract[38]. Protein-RNA complexes were isolated by MS2 affinity selection with amylose beads (New England BioLabs) as previously reported[37].

Preparation of yeast (pre-)mRNPs by TAP tag purification. *Yeast strain.* Amplification of TAP tag[39] construct (*CBP2*–PreScission protease cleavage site–2 ProteinA) from pBS1539-Psc (*URA3*) plasmid was carried out under Phusion polymerase, with forward primer 5′-TCAGACCAGGTTTCGATGAAGAAAGAGAAGATGATAACTACGTACCTCAGTCCATGGAAAAGAGAAGAT-3′ and reverse primer 5′-TATATATATATCTGTGTGTAGAATCTTTCTCAGATATAAATTGATTGATTTACGACTCACTATAGGGCGA-3′. The PCR construct was transformed into yeast strain BJ2168. The positive yeast clone was confirmed by DNA sequencing and western blotting.

Yeast cell extract. Yeast cells (*MAT**a** gal2 leu2 pep4-3 prc1-407 prb1-1122 trp1 ura3-52*) were grown in YPD (1% yeast extract, 2% peptone, 2% glucose) substituted with 50 mg/l ampicillin and 10 mg/l tetracycline, pelleted at 4,500 r.p.m. (6,728 r.c.f.) for 10 min, suspended in 0.7 volumes (v/w) AGK buffer (10 mM HEPES, pH 7.5, 1.5 mM $MgCl_2$, 200 mM KCl, 10% glycerin) and flash frozen in liquid nitrogen. Cell beads were ground (ZM 200, Retsch). Cell debris were pelleted at 17,000 r.p.m. for 30 min, and, optionally, polysomes were separated in a second ultracentrifugation at 37,000 r.p.m. for 60 min.

TAP tag purification. Typically, 10 ml yeast cell extract (30–35 mg/ml protein) were incubated with 600 µl IgG beads suspension (IgG Sepharose 6 Fast Flow, GE Healthcare) for 2 h at 4 °C. The supernatant was eluted by gravity, and the beads were washed with 20 ml CBB buffer (calmodulin binding buffer; 25 mM Tris, pH 7.9, 150 mM NaCl, 1 mM $MgOAc_2$, 1 mM imidazole, 2 mM $CaCl_2$, 2 mM DTT). The complex was released by incubation with 12 µl PreScission (10 mg/ml) in 2 ml CBB and 1 µl rRNasin (Promega) overnight.

The second purification step was carried out with 400 µl calmodulin beads suspension (Calmodulin Affinity Resin, Agilent). The sample was incubated with beads for 1 h at 4 °C, washed with 20 ml CBB and eluted twice by incubation with 1 ml CEB (calmodulin elution buffer; 25 mM Tris, pH 7.9, 150 mM NaCl, 1 mM $MgOAc_2$, 1 mM imidazole, 25 mM EGTA, 0.02% NP-40, 2 mM DTT) for 5 min.

UV cross-linking (254 nm). The cross-linking apparatus built in-house was equipped with four 8-W lamps (254 nm; G8T5, Sankyo Denki). Cross-linking was done on ice in custom-made Petri dishes in which either 10 or 1 ml sample solution had a depth of 1 mm. The Petri dishes were placed on an ice-cold metal block at a distance of 1 cm under the lamps.

Cross-linking of yeast (pre-)mRNPs was typically carried out on the IgG eluate, in initial experiments also on cell extract and calmodulin eluate. For cross-linking of yeast cell extract, the extract was dialyzed against AGK without glycerin, as glycerin is a radical scavenger. The sample was irradiated for 2 min and subsequently ethanol precipitated. Cross-links were typically isolated by size-exclusion and C18 chromatography (see below).

Human protein-RNA complexes were cross-linked after elution from amylose beads. Samples were UV irradiated for 10 min, followed by ethanol precipitation. Cross-links were isolated by size-exclusion and C18 chromatography followed by TiO_2 solid-phase extractions (see below).

Preparation of *in vivo*–4SU-labeled yeast mRNPs. *Growth, in vivo labeling and cross-linking.* Cells (strain BY4141, *MAT**a** his3Δ0 leu2Δ0 met15Δ0 ura3Δ0*) were grown in YPD (1% yeast extract, 2% peptone, 2% glucose) to OD_{600} = 0.5, and RNA was *in vivo*–4SU-labeled as described in ref. 40. Cells were allowed to grow for another 3 h before cells were pelleted at 15,000*g* for 15 min. (m)RNPs were cross-linked by 365-nm UV light and were cooled on ice using an XL-1500 Spectro Linker (Spectronics Corporation) as described[40].

Purification of mRNPs. After pelleting of cells at 2,880*g* for 5 min, cells were resuspended in lysis buffer (20 mM Tris-HCl, pH 7.5, 500 mM LiCl, 0.5% LiDS, 1 mM EDTA, 5 mM DTT, 1× EDTA-free protease inhibitor cocktail (Roche)). Cells were lysed using acid-washed glass beads in a FastPrep device (MPI) using five pulses at 6 m/s for 60 s with 60 s of pausing in between. Lysates were then cleared by centrifugation at 9,400*g* for 2.5 min in a table-top centrifuge. Purification of mRNPs was performed using magnetic oligo(dT) beads (NEB) as described in refs. 7,41. Finally, cross-links were isolated by C18 chromatography and TiO_2 solid-phase extraction (see below).

Enrichment and isolation of cross-links. For identification of peptide–RNA oligonucleotide cross-links in MS analysis, the moiety of peptides and RNA oligonucleotides that are not cross-linked has to be removed. This step is important, as residual non-cross-linked peptides and RNA oligonucleotides will strongly interfere with the detection of peptide–RNA oligonucleotide cross-links in the mass spectrometer. Enrichment of cross-linked peptide–RNA oligonucleotides and the removal of non-cross-linked peptides using TiO_2 has been established previously[15,25]. Size-exclusion chromatography (SEC) was performed for nonsubstituted yeast and human RNPs. Application of SEC or TiO_2 solid-phase extraction for purification of peptide–RNA oligonucleotides depends on the nature of the cross-linked RNPs and their RNA moieties. In principle, SEC and TiO_2 enrichment are complementary, as both techniques remove non-cross-linked peptides from the sample. SEC is generally beneficial for complexes isolated under native conditions, which might still contain 'contaminating' non-cross-linked proteins that do not interact with RNA directly. 4SU-substituted yeast RNPs were isolated under stringent conditions with oligo(dT) and contained almost exclusively cross-linked proteins; size exclusion was omitted because the majority of proteins that are not cross-linked had been removed during this isolation. SEC is generally not applicable at all if peptides and RNA do not exhibit sufficient difference in size: for example, for cross-linking experiments of proteins bound to short oligonucleotides (approximately 5–40 nucleotides). C18 reversed-phase (RP) chromatography is absolutely essential in order to remove the non-cross-linked RNA oligonucleotides before MS analysis. Without this step, MS-based detection of peptide–RNA oligonucleotide heteroconjugates is severely hampered by residual non-cross-linked RNA oligonucleotides.

For SEC, the sample was digested with 1:50 (w/w) trypsin (sequencing-grade modified trypsin, Promega) overnight in the presence of 0.1% SDS and 50 mM Tris, pH 7.9. Protein amounts were determined by the Bradford assay. SEC was carried out on a SMART system (Pharmacia Biotech) equipped with a Superdex 200 column (PC 3.2/30, 2.4 ml, Amersham Biosciences) at a flow rate of 40 µl/min with 20 mM Tris, pH 7.5, 150 mM NaCl, 1.5 mM $MgCl_2$ as running buffer. Fractions of 100 µl were collected. Fractions showing a strong absorbance at 254 nm over that at 280 nm (see **Supplementary Fig. 10**) were pooled and ethanol precipitated.

Protein and RNA digestion as well as C18 desalting and TiO_2 enrichment was done according to established protocols[15,25].

For hydrolysis of proteins and RNA, precipitated samples from cross-linking or SEC were dissolved in the presence of 50 µl 4 M urea, 50 mM Tris, pH 7.9, 1.5 mM $MgCl_2$ and diluted with 150 µl 50 mM Tris, pH 7.9, 1.5 mM $MgCl_2$ to a final concentration of 1 M urea, 50 mM, Tris pH 7.9, 1.5 mM $MgCl_2$. RNA was hydrolyzed with 1 µl benzonase (25 U, Novagen) for 30 min at 37 °C and subsequently with 1 µl each of RNases A and T1 (both Ambion; RNase A 1 µg/µl, RNase T1 1 U/µl) for 60 min at 52 °C. Protein digestion was carried out with 1:20 (w/w) trypsin (sequencing-grade modified trypsin, Promega) at 37 °C overnight. Protein amounts were determined by the Bradford assay. For samples completely proteolyzed before SEC, 1 µg trypsin was added instead for hydrolysis of nucleases and residual longer protein fragments.

Desalting was done immediately after digestion for all samples. 10 µl acetonitrile (ACN) and 2 µl 10% formic acid (FA) (v/v) were added to the sample (volume 200 µl) before loading on C18 columns (in-house; C18 AQ 120 Å 5 µm, Dr. Maisch GmbH). Samples were washed with 120 µl 0.1% FA (v/v) and eluted stepwise with 120 µl 50% ACN (v/v), 0.1% FA and 60 µl 80% ACN (v/v), 0.1% FA (v/v). The combined eluate was dried in a centrifugal evaporator. Samples were either enriched further with TiO_2 or directly subjected to LC-MS/MS analysis.

For TiO_2 enrichment, samples were dissolved in 60 µl buffer A (200 mg/ml dihydroxy benzoic acid, 80% ACN, 5% trifluoroacetic acid (TFA)) and loaded on TiO_2 microcolumns (in-house; titansphere, 5 µm, GL Science). The sample was washed with buffer A (180 µl) and extensively with buffer B (80% ACN, 5% TFA; 300 µl) to remove any residual DHB and finally eluted with 120 µl ammonia. The eluate was dried in a centrifugal evaporator.

Mass spectrometry. Pellets from C18 desalting or TiO_2 enrichment were dissolved with 50% ACN, 0.1% FA and diluted to a final concentration of 10% ACN and 0.1% FA. Typically, each sample was injected twice on either the same or different instruments.

Most LC-MS/MS analyses were carried out with an LTQ Orbitrap Velos (Thermo) directly coupled to a nanoflow–liquid chromatography system (Agilent 1100 series). Samples were loaded on a C18 trapping column at a flow rate of 10 µl/min in 3% buffer B (buffer A: 0.1% FA; buffer B: 95% ACN, 0.1% FA) and washed for 5 min. A linear gradient of 3–36% buffer B (flow rate 300 nl/min) flushed the analytes onto the analytical column and separated them in an elution time of 37 min (60-min overall run time) or 97 min (120-min overall run time). Residual analytes were eluted by raising buffer B to 95% for 7.5 min. The columns were built in-house. (trapping column: inner diameter = 150 µm, length = 2 cm; analytical column: inner diameter = 75 µm, length = 15 cm; C18 material, see above). The instrument was run in data-dependent acquisition mode. MS1 was acquired from 350 to 1,600 *m*/*z* (or Thomson, Th) at a resolution setting of 30,000 FWHM (full-width at half-maximum). The ten most intense precursors were chosen for fragmentation by higher-energy collision-induced dissociation (HCD). For MS/MS, the following parameters were set: minimal signal required, 5,000; isolation width, 2 Th; normalized collision energy, 45; dynamic exclusion, 20 s; resolution setting, 7,500 FWHM.

Alternatively, LC-MS/MS experiments were conducted with a Q Exactive instrument (Thermo) coupled to an EASY-nLC II (Thermo). The sample was loaded on a C18 trapping column (inner diameter, 100 µm; length, 4 cm; C18 material, see above) and washed with 25 µl buffer A. A linear gradient of 4–36% buffer B within 92 min at a flow rate of 250 nl/min separated the analytes on the C18 analytical column (inner diameter, 50 µm; length, 10 cm; C18 AQ 120 Å 3 µm, Dr. Maisch GmbH). A final elution step at 95% buffer B for 8 min removed any residual analytes. The instrument was set to a TOP12 method in data-dependent acquisition mode. MS1 was recorded from 350 to 1600 *m*/*z* at a resolution setting of 70,000 FWHM. MS/MS fragmentation was done on the 12 most intense precursors with HCD fragmentation. MS/MS parameter were chosen as follows: minimal signal required, 10,000; isolation width, 2 Th; normalized collision energy, 30; dynamic exclusion, 20 s; resolution setting, 17,500 FWHM.

Data analysis with RNP^{xl}. The experimental workflow resulted in two raw data files per experiment (cross-linked and control, except for yeast 4SU-labeled RNPs), which were then subjected to the computational workflow (see below). For samples derived from yeast 4SU-labeled RNPs, no non-UV-irradiated control was performed. In initial MS analyses we confirmed that non-cross-linked proteins were completely removed through very stringent washing steps. Accordingly, the control MS data did not provide any additional benefit for the automated data analysis.

MS raw data in Thermo's .raw format was converted into the .mzML-format[42] with msconvert of the ProteoWizard software package[43] (http://proteowizard.sourceforge.net/).

All subsequent steps were performed with tools from the OpenMS software library (http://www.openms.de/RNPxl; see detailed instructions as well as links to sample pipelines and a test data set in the **Supplementary Note**).

Data in profile mode were peak called (centroided). The non-irradiated control was aligned relative to the UV-irradiated sample on the basis of high-intensity features to correct for retention time shifts[44].

Next, a search was performed against a target-decoy version of the UniProt yeast database, also including contaminant sequences as distributed with the MaxQuant[45] software package. Oxidation of methionine, carbamylation of lysines and N termini, and phosphorylation of tyrosine, serine and threonine were considered as variable modifications. All MS/MS spectra with a confident peptide-to-spectrum match (FDR <1%) were filtered from the sample's data set.

For all remaining MS/MS spectra, differential analysis was performed on precursor intensities calculated from extracted-ion chromatograms of UV-irradiated sample and non-irradiated

doi:10.1038/nmeth.3092

control. If the same precursor was observed in the control at a comparable intensity (fold change less than 2), the corresponding MS/MS spectrum was removed from the UV-sample data file.

The reduced sample data file was subjected to our novel data analysis tool, RNP^{xl}. Prior to generation of precursor variants, spectra were filtered if their precursor corresponded to a small RNA oligonucleotide according to fractional mass (M <1,750 Da and fractional mass <0.2) or if the precursor was too small for an identifiable cross-link (M <600 Da).

Precursor variants for unlabeled RNA were generated with all calculated masses of RNAs meeting the following criteria: maximum length of four nucleotides, at least one U in sequence, RNA modifications: none, $-H_2O$, $-HPO_3$, $-H_3PO_4$, $-H_2O + 152$, $-HPO_3 + 152$, $-H_3PO_4 + 152$. In addition, +152 was considered as a modification without an additional nucleotide. This lead to a maximum number of 281 precursor variants per MS/MS spectrum.

For 4SU-substituted RNA, the following parameters were chosen for precursor mass variant generation: maximum RNA length, four nucleotides; at least one 4SU in sequence; RNA modifications: none, $-H_2O$, $-H_2S$, $-HPO_3$, $-H_2S$ $-HPO_3$, $-H_3PO_4$. Alternatively, a mass of a post-translational modification was defined on all 20 amino acids, resembling 4SU cross-linked under loss of H_2S with a stable adduct of 94.0167 Da ($C_4H_2N_2O$) after fragmentation[15]: PTM mass, 306.0253 ($C_9H_{11}N_2O_8P$), neutral loss, 212.0086 Da ($C_5H_9O_7P$). For the corresponding searches, RNA masses for precursor mass variant generation were calculated with the following criteria: maximum length, three nucleotides; RNA modification, $-H_2O$. The resulting precursor mass variants were searched with OMSSA[14] in four separate searches allowing the described PTM on sets of five amino acids: K, F, H, R, Y; S, G, P, W, M; A, V, T, C, L; or I, N, D, Q, E. In these searches, only oxidation of methionine was considered as an additional variable modification.

Precursor mass variants with a mass-to-charge ratio below 250 m/z were disregarded for further processing. Parameters for OMSSA searches were chosen as follows: precursor mass tolerance, 10 p.p.m.; fragment mass tolerance, 0.01 Da; variable modifications (unless noted otherwise above), oxidation of methionines and carbamylation of N termini and lysines. Typically, all results with an OMSSA score better than 1×10^{-5} (native RNA) or 1×10^{-8} (4SU-labeled RNA) were considered for manual validation.

In general, the computational workflow runs on standard personal computers and is integrated into the OpenMS environment[12,13] (see **Supplementary Note** tutorial for details). The search results are reported in two formats: a tabular comma-separated values (CSV) file, which can be imported into spreadsheet applications (for example, Microsoft Excel), and an idXML file, used to annotate the raw data with the search results in mzML format in TOPPView[46] (a graphical MS data visualization tool that is part of OpenMS).

Validation of search results. First, correct assignment of monoisotopic peak and charge state was confirmed. For experiments with unlabeled RNA, extracted-ion chromatograms of non-irradiated control and UV-irradiated sample were compared with Xcalibur (Thermo Fisher Scientific). Cross-link candidates were rejected if the same precursor was observed in the control measurement with an extracted-ion chromatogram area of more than half of the area in the UV-irradiated sample.

The assignment of peptide fragments was verified by annotating the raw data in mzML format with the search result output of the RNP^{xl} tool in idXML format using TOPPView[46]. Unassigned peaks, especially those of high intensity, were annotated manually, either as internal peptide fragments, RNA marker ions, or peptide sequence ions shifted by the mass of the cross-linked RNA or fragments thereof. In the majority of fragment spectra, y- and/or b-type fragment-ion series were observed that unambiguously identified the cross-linked peptide sequence (see **Supplementary Data**). The fragment spectra of the putatively cross-linked species were omitted if (i) RNA marker ions did not match to the computed RNA composition, i.e., an adenosine predicted but not visible as a strong marker ion in the lower m/z regime; (ii) a large number of high intensity signals, especially in the higher m/z range was observed but could not be annotated or (iii) a peptide sequence tag was manually assigned that did not correspond to the computed candidate sequence and also could not be explained by an overlapping precursor and its corresponding fragment spectra. Special focus was on the following peptide and RNA ions: the a2/b2 ion pair usually well observable in HCD fragment spectra; high-intensity immonium ions; fragments resulting from cleavage N and C termini to proline, where the N-terminal ion should have high intensity and the C-terminal ion should be barely or not observable, and RNA marker ions of the nucleic acid bases when more than one nucleotide was cross-linked. Localization of the cross-linking site on the peptide was done by comparing automatically annotated regular peptide fragments and manually annotated signals corresponding to peptide fragments shifted by the mass of the cross-linked RNA or its fragments. Only if the last regular peptide fragment and the first shifted fragment clearly pointed to a single amino acid or if an immonium ion bound to RNA was observed was localization to a single amino acid possible. More details are given in the **Supplementary Note** tutorial. For peptide fragmentation characteristics with HCD and a collection of further literature about collision-induced fragmentation, see ref. 47.

Comparison to standard database search. Comparative database searches with OMSSA[14] and Mascot[16] were performed on the MS data set of cross-linked human RNPs. A novel post-translational modification was defined on arginine and lysine with a mass of 324.0359 Da ($C_9H_{13}N_2O_9P$) and a neutral loss with the same mass. At least one arginine or lysine is present in tryptic peptides (with the possible exception of the protein C terminus).

For OMSSA, data was processed as described above (conversion, peak picking). Next, the ID filter pipeline was modified by taking only the newly defined modification and oxidation of methionine into account as variable modifications. All other parameters remained unchanged. Consequently, results up to an FDR of 1% were reported and are listed in **Supplementary Table 4**.

For Mascot (version 2.3.02), data were converted into the text-based .msm file format with Raw2MSMS version 1.10 (ref. 48). Variable PTMs were the same as for OMSSA. Two missed cleavages were allowed for trypsin. Peptide tolerance was set to 10 p.p.m., MS/MS tolerance to 20 mmu, instrument to ESI-TRAP. The same version of the human UniProt database with the MaxQuant contaminant database (http://maxquant.org/downloads.htm) was used as for OMSSA but without reverse sequences. A Mascot

doi:10.1038/nmeth.3092

score of 29 (P <0.05, determined by Mascot) was applied for the comparison; the complete results are also listed in **Supplementary Table 4**.

Data deposition and software availability. The mass spectrometry data described in this work have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org/) via the PRIDE partner repository[49] with the data set identifier PXD000513 and can be viewed with the PRIDEInspector Tool. The software described above is available in source code for all major platforms as well as precompiled binaries for Windows and OS X as part of the OpenMS software suite at http://www.OpenMS.de/RNPxl/ (where future software updates can be found) and as **Supplementary Software**. The software is distributed as open-source software under a three-clause BSD license. Detailed documentation and a tutorial on the use of the software is available in the **Supplementary Note**.

36. Bessonov, S., Anokhina, M., Will, C., Urlaub, H. & Lührmann, R. Isolation of an active step I spliceosome and composition of its RNP core. *Nature* **452**, 846–850 (2008).
37. Deckert, J. *et al.* Protein composition and electron microscopy structure of affinity-purified human spliceosomal B complexes isolated under physiological conditions. *Mol. Cell. Biol.* **26**, 5528–5543 (2006).
38. Dignam, J.D., Lebovitz, R.M. & Roeder, R.G. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res.* **11**, 1475–1489 (1983).
39. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030–1032 (1999).
40. Creamer, T.J. *et al.* Transcriptome-wide binding sites for components of the *Saccharomyces cerevisiae* non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. *PLoS Genet.* **7**, e1002329 (2011).
41. Castello, A. *et al.* System-wide identification of RNA-binding proteins by interactome capture. *Nat. Protoc.* **8**, 491–500 (2013).
42. Martens, L. *et al.* mzML—a community standard for mass spectrometry data. *Mol. Cell. Proteomics* **10**, R110.000133 (2011).
43. Chambers, M.C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
44. Lange, E. *et al.* A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics* **23**, i273–i281 (2007).
45. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
46. Sturm, M. & Kohlbacher, O. TOPPView: an open-source viewer for mass spectrometry data. *J. Proteome Res.* **8**, 3760–3763 (2009).
47. Michalski, A., Neuhauser, N., Cox, J. & Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *J. Proteome Res.* **11**, 5479–5491 (2012).
48. Olsen, J.V. *et al.* Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4**, 2010–2021 (2005).
49. Vizcaíno, J.A. *et al.* The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41**, D1063–D1069 (2013).

doi:10.1038/nmeth.3092