Identification of peptide-RNA heteroconjugates by mass spectrometry

Dissertation

for the award of the degree

“Doctor rerum naturalium” (Dr. rer. nat.)

of the Georg-August-Universität Göttingen

within the doctoral program Molecular Biology
of the Georg-August University School of Science (GAUSS)

submitted by

Aleksandar Chernev

from Sofia, Bulgaria

Göttingen, 2020

Thesis Committee

Prof. Henning Urlaub
Bioanalytical Mass Spectrometry Group
Max Planck Institute for Biophysical Chemistry, Göttingen
Bioanalytics, Department of Clinical Chemistry
University Medical Centre
Georg-August-Universität, Göttingen

Prof. Markus Bohnsack
Department of Molecular Biology
University Medical Centre
Georg-August-Universität, Göttingen

Prof. Tim Beißbarth
Department of Medical Bioinformatics
University Medical Centre
Georg-August-Universität, Göttingen

Members of the Examination Board

Prof. Henning Urlaub
(Reviewer)
Bioanalytical Mass Spectrometry Group
Max Planck Institute for Biophysical Chemistry, Göttingen
Bioanalytics, Department of Clinical Chemistry,
University Medical Centre
Georg-August-Universität, Göttingen

Prof. Markus Bohnsack
(Reviewer)
Department of Molecular Biology
University Medical Centre
Georg-August-Universität, Göttingen

Further members of the Examination Board

Prof. Tim Beißbarth
Department of Medical Bioinformatics
University Medical Centre
Georg-August-Universität, Göttingen

Prof. Jörg Stülke
Department of General Microbiology
Institute for Microbiology and Genetics
Georg-August-Universität, Göttingen

Dr. Alexander Stein
Membrane Protein Biochemistry Group
Max Planck Institute for Biophysical Chemistry, Göttingen

Dr. Alex Faesen
Biochemistry of Signal Dynamics Group
Max Planck Institute for Biophysical Chemistry, Göttingen

Date of the oral examination: 15.09.2020

In times of trouble, one should laugh. Just laugh!

- Gintama

Table of Contents

List of Figures

List of Tables

Extended supplementary tables on attached CD:

SupplTableCD_1_Ecoli_XL_combined_FDR001.xlsx

SupplTableCD_2_HeLa_XL_combined_FDR001.xlsx

Abbreviations

ACN	Acetonitrile
AGC	Automatic gain control
ATP	Adenosine triphosphate
BCA	Bicinchoninic acid
CID	Collision-induced dissociation
CLIP-seq	Cross-linking immunoprecipitation-high-throughput sequencing
Da	Dalton (g/mol)
DC	Direct current
DHB	2,5-dihydroxybenzoic acid
DMSO	Dimethyl sulfoxide
DNA/RNA	Deoxy-/ribonucleic acid
e.g.	for example (exempli gratia)
EDTA	Ethylenediaminetetraacetic acid
ESI	Electrospray ionization
et al.	and others (*et alli*)
FA	Formic acid
FDR	False discovery rate
GdnHCL	Guanidine hydrochloride
HCD	C trap dissociation
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HFIP	Hexafluoroisopropanol
IMAC	Immobilized metal affinity chromatography
LB	Lysogeny broth
LC	Liquid chromatography
LDS	Lithium dodecyl sulfate
LTQ	Linear Trap Quadropole
m/z	mass-to-charge
MALDI	Matrix-assisted laser desorption ionization
MOPS	(3-(N-morpholino)propanesulfonic acid)
MS	Mass spectrometry
MS/MS or MS2	Tandem mass spectrometry
MWCO	Molecular weight cut-off
NCE	Normalized collision energy
NGS	Next-generation sequencing
OD	Optical density
OMSSA	Open Mass Spectrometry Search Algorithm
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate Buffered Saline
PCR	Polymerase chain reaction
PDB	Protein Data Bank
ppm	parts per million
PTM	Post-translational modification
PVDF	Polyvinylidene fluoride
Q Exactive HF	Q Exactive High Field
RBP	RNA-binding protein
RF	Radio frequency
RNAse	Ribonuclease
RP	Reversed-phase
rpm	revolutions per minute
RRM	RNA-recognition motif

SAX	Strong anion exchange chromatography
SDS	Sodium dodecyl sulfate
TEA	Triethylamine
TEAB	Triethylammonium bicarbonate
TFA	Trifluoroacetic acid
UHPLC	Ultra High Performance Liquid Chromatography
UV	ultra violet
v/v	volume/volume
w/v	weight/volume
w/w	weight/weight
XIC	extracted-ion chromatogram
XML	Extensible Markup Language

Summary

Proteins and nucleic acids are two of the major constituents of life and their interplay is at the center of most biological processes. A deep understanding of the interaction between these biomolecules is crucial for structural and functional elucidation of numerous cellular mechanisms. Cross-linking of proteins to nucleic acids by UV irradiation or chemical reagents enables the preservation of association information in covalent bonds that can be examined by a variety of analytical methods.

Mass spectrometry (MS) is particularly useful in detecting the proteins associated with RNA. In recent years, significant progress was made in identifying cross-linked peptide-RNA heteroconjugates, which provide direct evidence for the contact sites between proteins and RNA. However, the analysis of protein-RNA cross-links by MS remains a very challenging task due to the low yield of the cross-linking reactions and the laborious manual annotation of mass spectra that is required to validate the results.

In this work, a strategy for fully automated annotation was developed that substantially speeds up the manual analysis of cross-links. The different elements of peptide-RNA heteroconjugate fragment spectra were identified and categorized, allowing the development of comprehensive and fully descriptive scoring functions. The created scores are a prerequisite for the employment of a false discovery rate estimation and pave the path towards full automation of the cross-link analysis.

Furthermore, the cross-linking behavior of the four canonical ribonucleotides was examined in controlled experiments with model RNA-binding proteins Hsh49 and GAPDH. In addition to the previously described cross-links to uracil, UV-induced heteroconjugates formed with cytosine, guanine and adenine could be detected by mass spectrometry. All identified spectra were formed by generating a covalent bond between the nucleobase and various amino acids. The mass characteristics of the observed precursors and their fragmentation products were investigated in detail and are summarized for future utility in mass spectrometric analyses. As an alternative to UV irradiation, sulfite-mediated cross-linking was demonstrated to be useful in the identification of cytosine contacts with lysines.

The toolkit for cross-link enrichment was complemented with two novel workflows for purification of oligoribonucleotide heteroconjugates, based on silica-based purification and strong anion exchange chromatography. Both methods result in significant depletion of interfering non-cross-link species. The workflows could be successfully employed in the study of *in vivo* UV-generated cross-links of *E. coli*, as well as in investigation of the protein-RNA interactions in HeLa cytoplasmic extract. A large number of cross-link sites

could be detected, providing contact information for known RNA-binding proteins and identifying novel RNA interaction partners.

Finally, the ability of nanoelectrospray mass spectrometry to identify heteroconjugates with large RNA moieties was explored. A synthetic peptide-RNA standard was generated by click chemistry and used to determine appropriate chromatographic separation and electrospray ionization conditions. The RNA moiety could be successfully fragmented by collision-induced dissociation, providing comprehensive sequencing information. Thus, demonstrating the capability of nanoelectrospray mass spectrometry to acquire additional structural information from cross-linked samples by revealing the identity of the interacting RNA and the localization of the cross-link site on the nucleotide chain.

1. Introduction

1.1 Mass spectrometry

Mass spectrometry is a powerful analytical technique that allows the determination of the mass-to-charge ratio (*m/z*) of ionized molecules, through their electromagnetic properties. The mass of an analyte can be deduced from the distribution of naturally occurring heavy isotope elements. Each molecule is represented by a mixture of mass variants, including molecules formed by only “light” atoms (e.g. ^{1}H, ^{12}C, ^{14}N; monoisotopic mass) or “light” and “heavy” (e.g. ^{2}H, ^{13}C, ^{15}N) stable isotope atoms. The distribution of these variants creates multiple, equally spaced signals (isotope envelope) with predictable profile, based on the statistical prevalence of the heavy isotope atom. Of main significance for biological molecules are the ratio of heavy isotopes of carbon (1.10% 13C 13.0034 Da) and nitrogen (0.37% 15N 15.0001 Da) that determine isotopic distribution of 1 Da-spaced peaks to the monoisotopic peak. From the observed *m/z* value between the isotopic distribution can be deduced the charge state of the analyte and its mass.

A mass spectrometer generally consists of three elements – ion source, mass analyzer and detector. In the ion source, the sample is ionized to produce charged molecules in the gas phase that can be manipulated and directed through the application of electromagnetic fields into the mass analyzer. The mass analyzer separates the ions according to their mass-to-charge and directs them to the detector, where the amount of ions at specific *m/z* value can be determined.

1.1.1 Ionization of macrobiomolecules

The employment of mass spectrometry in the study of biological molecules was enabled by the discovery of soft ionization techniques that allow ionization of large biomolecules without causing their degradation or fragmentation. In the last decades, two methods have contributed significantly to the advances of the field – matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI).

1.1.1.1 MALDI-MS

In MALDI, the analytes from a sample are first embedded in a crystalline matrix, with strong optical absorbance that can efficiently absorb laser irradiation. Laser pulses are absorbed by the matrix, causing ablation and ionization of the analytes, predominantly into singly charged ions [1]. The introduction of the ions to the gas phase from solid matrix, makes MALDI particularly powerful in spatial analysis of molecules from embedded complex samples such as tissue sections [2]. The speed and sensitivity of the method make it also an especially useful tool for microbiological identification and detection of pathogens [3].

1.1.1.2 ESI-MS

Electrospray ionization relies on the generation of ions from solution by application of high voltage and heat [4]. The transfer of ions from liquid to gas phase allows the direct coupling of the ionization method with liquid chromatography separation techniques. Predominantly higher charge states of ions are obtained, enabling the more comprehensive structural investigation of biomolecules. These aspects have made ESI widely used in identification of analytes from complex mixtures and the method of choice for proteomic investigations.

In standard ESI-MS experiments, molecules eluting from a liquid chromatography setup are nebulized from a fine emitter, held at high electrostatic potential difference, forming small charged droplets. Through the application of heat, the solvent continuously evaporates, shrinking the droplets further, until the analyte is freed as an ion in the gas phase. The ionization of analytes is more efficient from solution with pH at which the molecules would be charged in solution (e.g. acidic buffers for measurements performed in positive mode). Two existing models describe the possible mechanism of ionization – the ion evaporation and charge residue models. The ion evaporation model postulates that solvent evaporates from the surface of a droplet until the surface field strength is strong enough to cause the emission of the charged analyte by field desorption [5]. The charge residue model suggests that highly charged droplets shrink until the Coulombic repulsion of the charges becomes higher than the surface tension (Rayleigh limit). At that point, the droplet undergoes a fission event, bursting into several smaller droplets. This process repeats until a droplet contains a single charged analyte [6]. It is not fully clear what is the contribution of the two models for the ionization of molecules with different physicochemical properties, however the existing consensus is that larger analytes (>1000 Da) ionize predominantly by the charge residue model, while smaller analytes can be emitted through the ion evaporation model [7]. The introduction of lower flow rates (nl/min) by nano-liquid chromatography (nano-LC) separation allows the generation of smaller initial droplets, increasing the ionization efficiency and sensitivity [8,9].

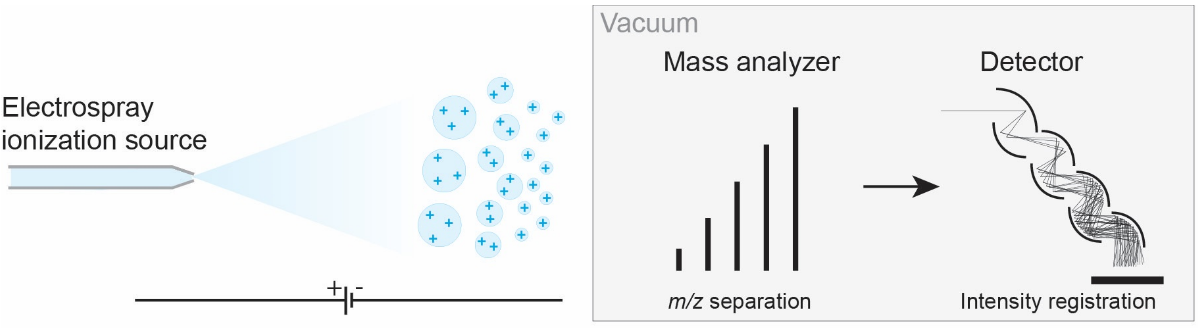


Figure 1.1 Components of a mass spectrometer

A mass spectrometer can be viewed as formed by 3 main components – ionization source, mass analyzer and detector. The ionization source produces gas phase ions, which are introduced to the vacuum of the instrument. The mass analyzer separates the ions according to their *m/z* value through the application of electromagnetic fields, leading them to the detector that registers the number of ions present in particular *m/z* value.

1.1.2 Mass analyzers

Various mass analyzers are currently present, possessing different characteristics, such as mass resolution, accuracy and range. The features, strengths and weaknesses of the mass analyzers used in this study are shortly discussed below.

1.1.2.1 Quadrupole

Quadrupole mass analyzers consist of four metal rods placed in parallel to each other and connected electrically in pairs. By applying radio frequency (RF) or direct current (DC) potential to the pairs, the mass analyzer can manipulate the trajectory of ions by creating oscillating electric fields. Application of only RF potential allows the transmission of all ions, while the combination RF and DC potential allows for the isolation of a particular *m/z* value. Quadrupoles and other multipoles (e.g. hexapoles) can act as beam-type collision cells by introduction of inert gas and application of increased RF-energy. As an analyzer, quadrupoles are very robust, inexpensive and can operate in fast duty cycles at lower vacuum levels. However, they suffer from relatively poor resolution and limited mass range [10].

1.1.2.2 Linear ion trap

Linear ion trap analyzers are formed from modified quadrupoles, where static electrical potential is applied to the end of the rods, trapping the ions. Thus, the confined ions can be accumulated over time, leading to higher sensitivity. Ejection of ions is performed by application of specific resonance RF potential for the particular *m/z* value of the analyte. Application of multiple frequencies allows the isolation of an analyte by expulsion of all other ions in the trap. Fragmentation is achieved by application of supplemental resonance excitation voltage for the specific *m/z*, which forces the corresponding analyte to collide with inert gas multiple times, slowly acquiring activation energy. The activated molecule decomposes to product ions that are stored in the trap and can be detected in a mass scan. Trapped product ions can be further selected and fragmented, providing several levels of structural information. Linear ion traps are characterized by very fast scan rate and high sensitivity, but suffer from low resolution [11].

1.1.2.3 Orbitrap

Orbitrap analyzers are ion traps, consisting of two components – two barrel-like (outer) electrodes and spindle-like (inner) electrode. Linear electric field is formed between the barrel-like electrodes and the central spindle electrode. When ions are injected tangentially into the mass analyzer, their inertia is balanced by the attraction force to the inner electrode. The ions take elliptical trajectories with harmonic axial oscillation frequency, which is proportional to the *m/z* value of the ion. The movement of the ions creates an image current in the outer electrodes that can be converted to a mass spectrum by Fourier transformation. Orbitraps have very high resolving power and accuracy but a slow scan rate and reduced sensitivity compared to a linear ion trap [12].

1.1.3 Fragmentation

Additional level of information for an analyte can be obtained by introducing energy to the (precursor) molecule, inducing its fragmentation and recording the mass spectrum of products (product or fragment spectrum). This process is termed tandem mass spectrometry (MS/MS or MS2). The characteristic signals in the fragment spectrum allow distinguishing isobaric compounds and identifying components of highly complex mixtures. Biological studies most commonly employ collision-induced dissociation (CID), where the precursor molecules are accelerated to increase their kinetic energy, followed by collision with inert gas (e.g. He, Ar, N_2) [13]. Two variant of CID were utilized in this work – ion trap CID generated in a linear ion trap and beam-type CID generated in a Higher-energy C-trap dissociation (HCD) cell. Ion trap CID is a slow-heating method, in which the precursor molecule is activated by multiple collisions with helium. Due to the electromagnetic properties of ion traps, the excitation of an analyte causes destabilization of the trajectory of all ions contained in the trap with *m/z* lower than ~30% *m/z* of the precursor, which leads to their loss in the product spectrum. In addition, the gradual activation grants sufficient time for molecular energy rearrangement, resulting in prevalent rupture of labile bonds. Post-translational modifications (PTM), such as phosphorylation or glycosylation are preferentially cleaved, often leading to inadequate fragmentation and inability to localize the modification site. During beam-type fragmentation, the precursor molecule is accelerated through a multipole collision chamber with inert gas (e.g. N_2). Fewer collision events are required for fragmentation to occur, generally preserving labile modifications. Unlike ion trap CID, where only the precursor is excited, in beam-type CID all ions are accelerated and could take part in a collision event. That may lead to further decomposition of product ions, but preserves the information of the low *m/z* parts of the product spectrum.

1.1.3.1 Peptide fragmentation

Shotgun proteomics studies that involve the identification of peptide mixtures are generally carried out in acidic solutions, which promote protonation of the molecules. Fragmentation of protonated peptides preferentially affects the peptide bond, producing a-, b- and y-ion series by charge directed reactions (Fig. 1.2). The prominent fragmentation pathways can be described with the “mobile proton” model [14]. Upon activation of the precursor ion by collision events with inert gas, protonation can occur at energetically less favored sites. Protonation of the amide nitrogen weakens the peptide bond and facilitates nucleophilic attack on the carbon atom of the amide bond, which can participate in a number of rearrangement (reviewed in [15]). One of the main reactions (b_x-y_z pathway) involves the oxygen of the neighboring N-terminal peptide bond, leading to the generation of b- and y-ions. Once formed, b-ions may further fragment producing lower b-ions ($b_x \rightarrow b_{x-1}$ pathway) or an a-ion ($b_x \rightarrow a_x$ pathway), especially in the case of beam-type CID, which creates characteristic intense a_2/b_2 pairs and shorter b-ion series. Trying to predict the exact probability of different fragmentation pathways to occur and the profile of a fragment spectrum is unfeasible due to the versatile chemical nature and possible spatial orientation of amino acid side chains. However, certain fragmentation trends can be expected. The presence of the imino acid proline creates high intensity y-ions (proline effect) by promoted cleavage of the neighboring N-terminal peptide bond [16]. High proton affinity side chains such as histidine also enhance the cleavage of the peptide bond N-terminal to the side chain (histidine effect) [14]. Similar effect can be observed for the positive amino acids, lysine and arginine [17]. Glutamine and asparagine promote C-terminal fragmentation of the peptide bond through nucleophilic attack by the side chain oxygen [18]. So do aspartate and glutamate with the involvement of a “locally mobile” carboxylic proton [19]. Excitation of protonated peptides can also result in neutral loss of small molecules, creating non-sequencing ions that provide little identification value. Water loss can occurs from the peptide C-terminus and the side chains of aspartate, glutamate, threonine and serine. Ammonia loss is observed from the side chains of asparagine, glutamine, lysine and arginine. Formation of internal ions can occur from intense y-ions and generally occurs in peptides containing proline or histidine. Internal immonium ions are predominantly generated from further fragmentation of ions by the a_x->a_{x-1}/I_x pathway [20].

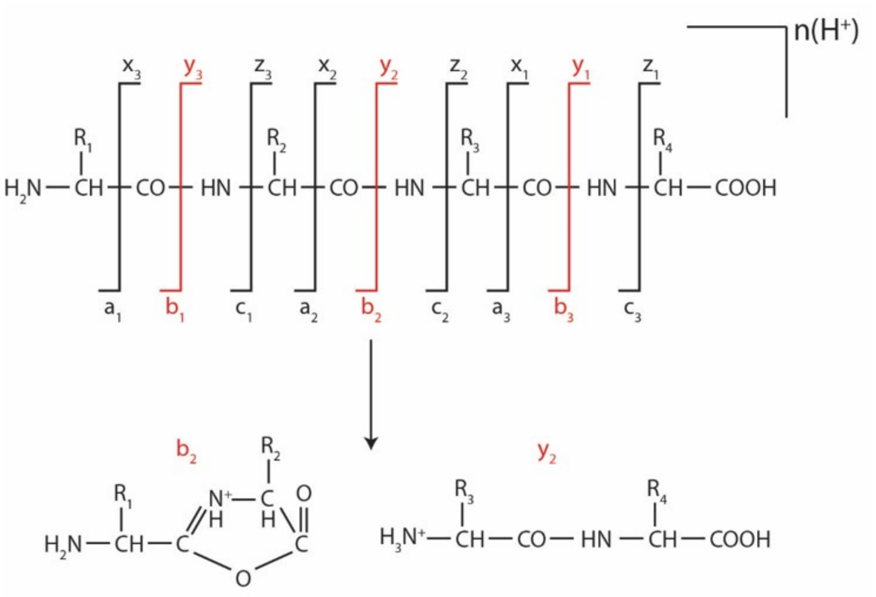

n(H+)
x3 y3 z3 x2 y2 z2 x1 y1 z1
R1 R2 R3 R4
H2N—CH—CO—HN—CH—CO—HN—CH—CO—HN—CH—COOH
a1 b1 c1 a2 b2 c2 a3 b3 c3
b2
y2
H2N—CH—C N+—C O H H C O
H3N+—CH—CO—HN—CH—COOH

Figure 1.2 Peptide fragmentation

Nomenclature of sequencing peptide ions generated by collision-induced dissociation. Fragments derived from the N-terminus can be presented as a-, b- and c-ions and C-terminal fragments are noted as x-, y- and z-ions [21]. Fragmentation from collision-induced dissociation occurs preferentially at the peptide bond, creating predominantly b- and y-ions.

1.1.3.2 Nucleic acid fragmentation

Nucleic acids are detected most efficiently in negative mode, in the presence of neutral or slightly basic buffers, owing to the negative charge of the phosphate groups of the oligonucleotide. Nomenclature of the fragmentation is analogous to the peptide fragmentation, with the generation of a-, b-, c- and d-ions from the 5'-end or w-, x-, y- and z-ion from the 3'-end (Fig. 1.3 A). The main pathway for fragmentation of DNA under CID conditions involves the loss of a base as initial step, either as a neutral or an ion, and successive cleavage of the 3' C-O bond of the deoxyribose, generating complementary w- and [a-B_n]-ions [22]. On the other hand, RNA produces mostly c- and y-ions, generated in reaction of the of 2'-hydroxyl hydrogen atom with the 5'-phosphate oxygen (Fig. 1.3 B) and to a lesser extent w- and [a-B_n]-ions [23].

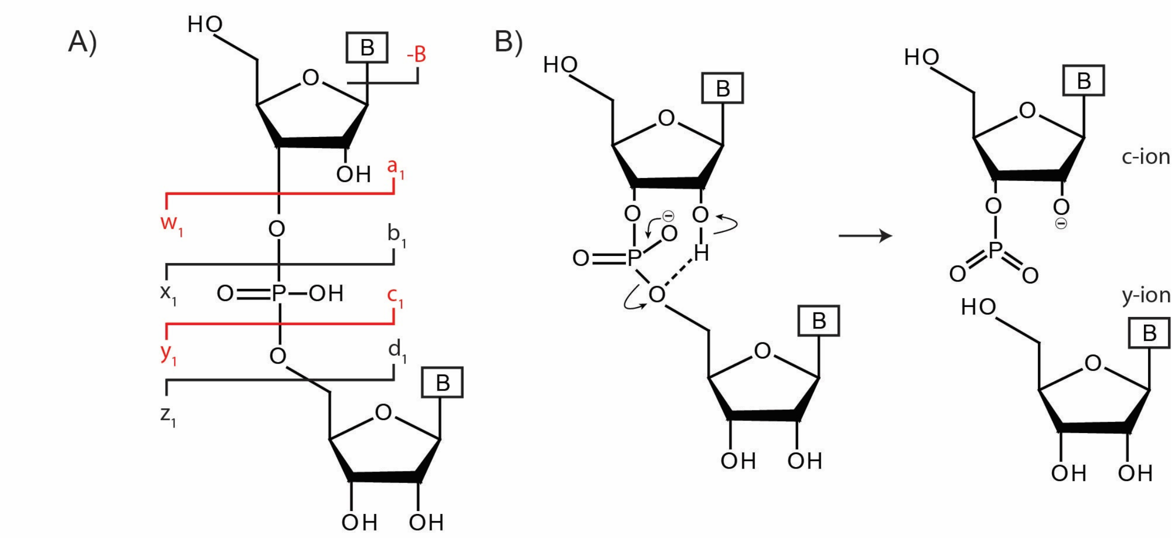


Figure 1.3 Fragmentation of RNA

A) Nomenclature of RNA fragments is analogous to peptide sequencing ions [24]. Ions generated from the 5'-end are noted as a-, b-, c- and d-ions, while 3'-end fragments are called w-, x-, y- and z-ions. Loss of the nucleobase is notated as -B_n, where B is substituted by the one letter code of the nucleobase and n – the position of the nucleobase in the sequence of the oligonucleotide. B) Proposed mechanism for formation of c- and y-ions in negative mode, redrawn according to [23].

1.1.4 Hybrid instruments

Two or more mass analyzers can be combined in the same instrument to exploit their individual strengths, making a hybrid instrument. Two types of hybrid mass spectrometers were used in this study – Q Exactive HF and Fusion (Lumos) Tribrid (Fig. 1.4). Q Exactive HF combines two mass analyzers – a quadrupole based mass filter and high-field Orbitrap. The Orbitrap is used to obtain high resolution and accuracy scans, while the quadrupole can effectively isolate ionized analytes for fragmentation. Beam-type CID fragmentation can be performed with nitrogen in a specialized multipole cell (HCD cell). Fusion (Lumos) Tribrid contains an additional mass analyzer – a linear ion trap. The presence of three mass analyzers allows exceptional versatility in fragmentation and scanning possibilities. Analytes can be fragmented in the ion routing multipole with nitrogen by beam-type CID or in the linear ion trap with helium by ion trap CID. The products can be analyzed by either a high resolution and accuracy scan in the Orbitrap or in faster and more sensitive linear ion trap scans. For the analysis of peptide-RNA heteroconjugates, beam-type fragmentation is preferred, as it preserves the information of low *m/z* RNA marker and other product ions. The fragment scans are commonly recorded in the Orbitrap to allow unambiguous identification of low intensity ions by the high accuracy provided by the mass analyzer.

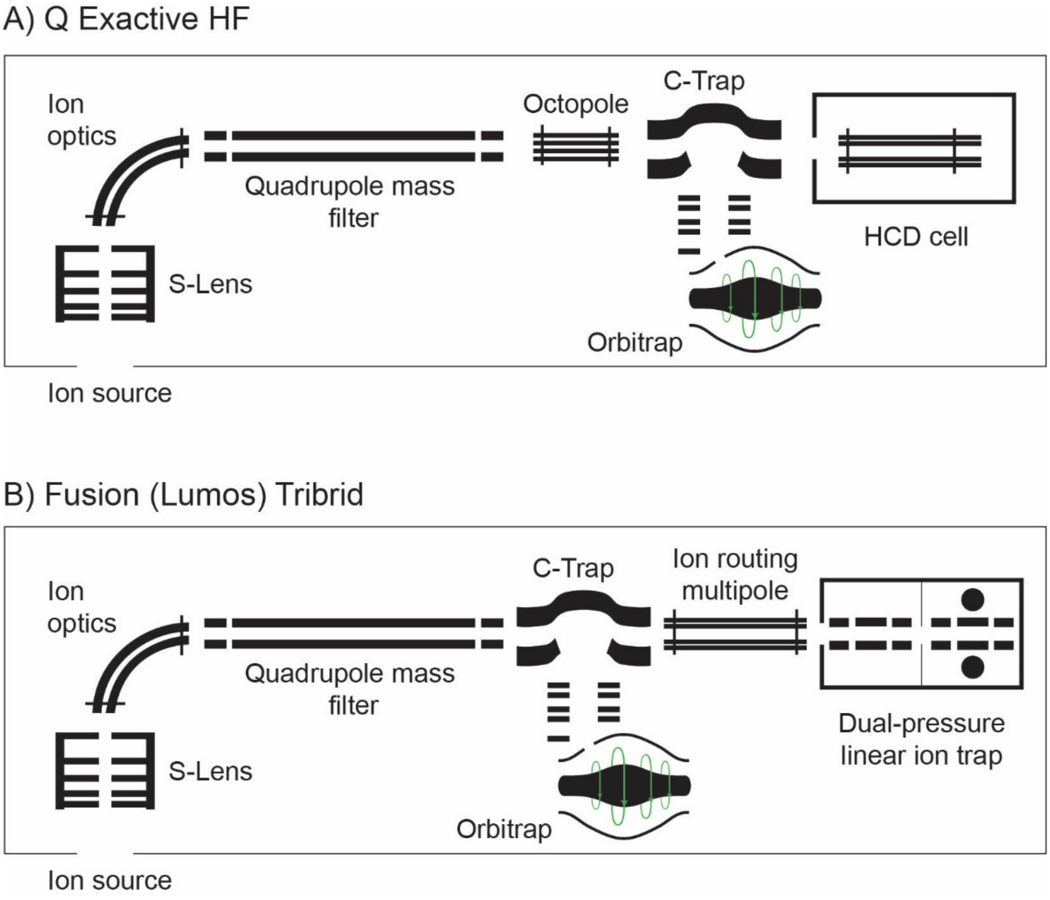


Figure 1.4 Hybrid mass spectrometers used in this study

Schematic representation of Orbitrap hybrid instruments. A) Q Exactive HF - Generated ions enter the vacuum of the instrument from the ion source and are captured and focused by the S-Lens. Next, the ions are transferred through the ion optics for accumulation in the C-trap. Collected ions for a precursor scan are injected to the Orbitrap, which creates a survey mass spectrum of the analytes in the sample. For generation of fragment spectra, precursors can be isolated in the quadrupole mass filter and fragmented in a HCD cell. The generated products are collected in the C-trap and injected for a product scan in the Orbitrap. B) Fusion (Lumos) Tribrid incorporates in addition Dual pressure linear ion trap that permits additional fragmentation and scan options. The first chamber of the trap has higher concentration of helium atoms, used to perform collisional cooling of arriving ions, reducing their initial energy and concentrating them in the center. Cooled ions are transferred to the low pressure chamber, where ion trap CID and mass spectrum scan can occur. The ion routing multipole allows the transfer back and forth between the different components of the instrument and serves as a HCD fragmentation cell.

1.1.5 Identification of peptides and proteins by mass spectrometry

The ability of mass spectrometry to deduce the exact mass of an analyte and characteristic fragmentation pattern make it an invaluable analytical tool for confirming the identity of a molecule. The development of soft ionization techniques in combination with the increased speed and sensitivity of mass spectrometers has enabled the employment of mass spectrometric analysis of highly complex mixtures such as whole cells, tissues and entire organisms. In a standard proteomic experiments, the proteins of the sample are extracted and digested with an endoproteinase. Commonly, trypsin has been utilized, owing to its robustness, high proteolytic activity and cleavage specificity after positive amino acids. Tryptic peptides naturally localize positive charges in both the C-terminal basic amino acid and the N-terminal amino group, increasing the probability of generating broad sequencing series from both peptide ends. Typically, the peptide mixture is separated by reversed-phase chromatography, so that a limited number of analytes are transferred at a certain time to the mass spectrometer. This has two effects: i) it allows for better ionization of the peptide that competes with other molecules for the charge of the solvent ii) reduces the number of signals the instrument should analyze at a particular time point, enabling more thorough investigation. Explorative experiments are generally performed in data dependent mode - the ionized peptides are first detected by the instrument in the survey (MS1) spectrum. The most intense signals are selected, isolated and fragmented, generating fragment (MS2) spectra. This acquisition cycle repeats throughout the entire elution time of the sample, collecting tens of thousands of spectra containing peptide sequencing information. Depending on the complexity of the sample, additional prefractionation steps can be performed, either on the protein or peptide level. The analysis of samples from higher eukaryotes (e.g. human samples) generally necessitates employment of several levels of orthogonal fractionation methods to achieve in-depth investigation (reviewed in [25]).

The generated spectral results files are processed by a search engine that reports the peptide identifications in the sample. First, the protein sequences are *in silico* digested and the expected *m/z* values of the generated peptides are calculated. The theoretical precursors are compared with the experimentally observed precursors. Whenever a match is detected, a theoretical fragment spectrum is produced and compared to the experimentally acquired fragment spectrum, calculating a similarity score. False discovery rate (FDR) estimation is generally performed by executing a search with reversed or randomized protein sequences, as an estimation of matching a peptide by chance. Different variants of search engines, scoring equations and false discovery rate strategies have been developed (reviewed in [26]).

1.2 Identification of RNA-binding proteins

From transcription to degradation, RNA is covered with RNA-binding proteins (RBPs), forming ribonucleoprotein particles. The interaction of many canonical RBPs is achieved through modular arrangements of characteristic protein domains, with notable examples including the RNA-recognition motif (RRM), the K-homology and the zinc finger domain [27]. Advances in structure determination techniques and the introduction of MS-based proteomics investigation of the RNA interactome have revealed a surprisingly large numbers of RBPs, many of which do not possess a known RNA-binding domain [28]. A wide array of unexpected interaction partners has emerged, including a plethora of metabolic enzymes and DNA-binding proteins [29,30]. The biological function of their association with RNA is still unclear for many of the identified proteins. While some proteins have been demonstrated to possess a moonlighting gene regulation functions, often acting on their own mRNA, the unpredictably great number of identified partners suggests that the association may serve an alternative purpose [31]. The concurrence of RNA-binding and catalytic regions in many of the identified proteins indicates a possible allosteric riboregulation mechanism exhibited by the RNA [32]. Alternative explanation for the higher affinity of a protein towards RNA may be found in the formation of higher-order assemblies, ensuring a specific cellular localization of the ribonucleoprotein particle [33]. Or the formation of the protein-RNA complex has no particular biological role, it is simply a transient association caused by the biophysical properties of the protein, favored by its high affinity towards nucleotide factors, phosphorylated metabolites or other structurally similar to RNA molecules. Additional work is required to elucidate the level of importance of the novel identifications, however, it is clear that a lot is still unknown of the interplay between proteins and RNA.

1.2.1 Methods for investigation of protein-RNA association

A variety of methods for studying the association of proteins and RNA are currently available. Possible interactions between proteins and RNA can be elucidated *in vitro* by reconstituting the ribonucleoprotein complex or *in vivo* by the utilization of cross-linking in combination with pull-down techniques. Classical biochemical methods, such as electrophoretic mobility shift assay, allow the detection of interaction between purified molecules [34]. The contacts of a protein-RNA complex can be further studied in detail by determining the molecular structure through X-ray crystallography, electron microscopy or nuclear magnetic resonance spectroscopy [35]. Alternatively, RNA pull-down and protein immunoprecipitation techniques enable the identification of RNPs formed in the cell by an RNA molecule or protein of interest. The combination with high-throughput explorative methods permits the comprehensive investigation of protein-RNA interactions *in vivo*. Novel RBPs can be identified by protein microarray assays or mass spectrometry based techniques. Similarly, RNAs associating with proteins are primarily identified and characterized by RNA sequencing methods. In the last years, a large number of cross-linking immunoprecipitation strategies (CLIP-seq family of methods) have emerged, providing a much deeper understanding of the dynamic nature of protein-RNA association (reviewed in [36]).

1.2.2 Photoreactivity of RNA

Cross-linking of proteins and RNA transforms the non-covalent spatial interaction into a newly formed covalent bond that allows the employment of biochemical analysis in denaturing conditions. The aromatic character of the nucleobases predisposes the absorption of light in the ultraviolet (UV) region and the generation of an electronic excited state. Commonly, irradiation with low pressure mercury lamps that emit UV light at 254 nm has been utilized. The excited RNA molecule can return to the ground state by emitting a photon (through fluorescence or phosphoresce) or through non-radiative decay pathways of internal energy conversion. Alternatively, the nucleobase can undergo a photochemical (light) reaction, leading to dissociation, structural rearrangement, generation of radical species or addition of another molecule.

The photoaddition of proteins to RNA has been widely utilized in the study of protein-RNA association, but the mechanism of the underlying reactions is not fully understood. The UV-induced cross-linking reaction is very inefficient, as the canonical nucleobases were selected chemically in prebiotic conditions characterized by prominent UV irradiation [37]. The excited states of the pyrimidine and purine building blocks of RNA have ultrashort lifetimes and are characterized with greater photostability than related organic molecules. In addition, photodamage reactions, formation of lesions (e.g. pyrimidine dimers) and self-cleavage pathways compete with the formation of protein-RNA cross-links [38]. The chemical versatility of the amino acid side chains and the unpredictable effects of the spatial protein conformation further hamper the direct investigation of the cross-linking mechanism in cellular systems. Thus, the bulk of information about protein-RNA cross-linking reactions was acquired by monitoring the photochemistry of simple model chemical compounds, such as nucleotides and amino acids.

Mass spectrometric analysis of protein-RNA cross-links have identified almost exclusively uracil as the cross-linked nucleotide [39]. Insights into the photoreactivity of uridine with amino acids can be obtained from the studies of Shetlar et al., who have performed extensive photoaddition experiments by UV irradiation of polyuridilyc acid with 19 common amino acids (excluding proline) [40]. The presence of a cross-link product was assessed by fluorescamine assay after depletion of the unreacted amino acids by gel column chromatography, allowing the detection of a primary amino group of the cross-linked amino acid. All 19 amino acids were found to be reactive with polyuridilyc acid. Highest reactivity was obtained by the sulfur containing (Cys and Met), aromatic (Trp, Phe and Tyr) and basic (Arg and Lys) amino acids. In addition, evidence of limited photoreactivity of proline can be found in experiments performed with ^{14}C labeled uracil [41]. Thus, all 20 amino acids could undergo a photochemical reaction with uracil. The localization results obtained by mass spectrometric studies in the recent years follow the observations of the fluorescamine assay [39,42]. While certain amino acids have shown substantially lower reactivity (e.g. Asp) and are seldom observed in mass spectrometric studies until now, the possibility of detecting prominent cross-links generated with them in the future cannot be fully dismissed [40,43]. The effects of energy transfer between neighboring nucleotides, as well as the spatial orientation and structure of protein-RNA complexes might modulate the reactivity of amino acids, leading to uncharacteristic efficiency of cross-linking in particular complexes. Therefore, the possibility to generate a covalent bond with any of the 20 amino acids should be considered.

The photochemical cross-linking reaction between proteins and RNA is predominantly thought to proceed by a radical mechanism [44]. The exact photochemical reaction is unclear for many of the reactive amino acids. Some insights into the cross-linking of uracil can be found in the studies performed by Varghese et al. [45] . After UV irradiation of a solution of cysteine and uracil, four cross-link products could be observed (Fig. 1.5). Cysteine addition products are formed by generation of covalent bond at either the 5th or 6th position of the pyrimidine ring. Further investigation have found that products I and II are predominantly formed in deaerated samples (under nitrogen) and are stable to heat and acid, but unstable in alkali solutions and photoreversible. On the other hand, products III and IV are predominantly formed in aerated solutions and are stable in all mentioned conditions [46].

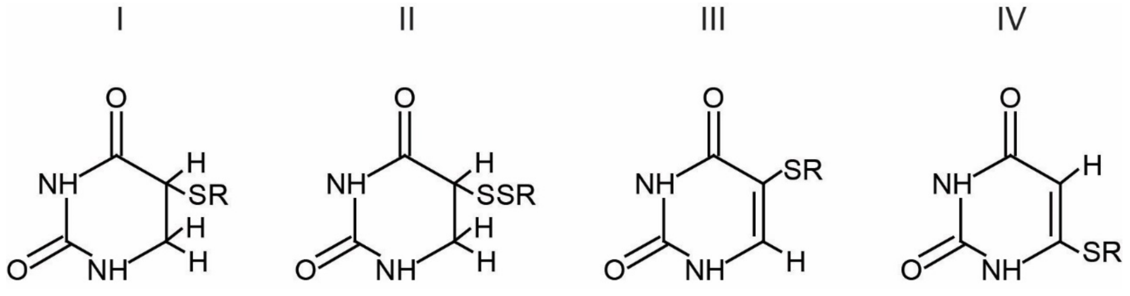


Figure 1.5 Photoaddition products of cysteine and uracil

Described cross-linked products of cysteine and uracil, as described by Varghese et al. [45].

1.3 Elucidation of protein-RNA interactions by UV-induced cross-linking and mass spectrometry

Mass spectrometry in combination with UV cross-linking approaches can be separated in three different categories, according to the resolution information of the interaction they provide: i) identifying RNA-associated proteins ii) spotting the peptides involved in the interaction iii) pinpointing the exact amino acid or subpeptide region that is cross-linked to the RNA.

Protein level workflows include protocols in which enrichment is achieved by purification of cross-linked RNA species, followed by digestion of the linked proteins and quantification against a non-irradiated control. Most commonly mRNA purification with oligo(dT)-based pull-downs have been utilized [47–49]. With this approach, the non-cross-linked peptides of the linked protein are identified in standard proteomics search, from which the number of significantly enriched proteins to the control sample can be determined. This strategy ensures high sensitivity, owing to the highly optimized instrumental and bioinformatics setup for linear peptides. Identification of multiple peptides from a protein gives accumulative confidence of the identification. In addition, protein-RNA complexes with covalent bonds formed at different residues of the protein contribute to a common, overlapping pool of linear peptides, leading to an increased signal readouts. A disadvantage of this approach is that no confident localization information can be obtained. The method also relies on the assumption that enrichment of proteins in the irradiated sample constitutes direct interaction to RNA. However, UV-irradiation can also cause protein-protein cross-linking, that can be highly efficient with certain proteins [50]. Therefore, protein interactors of the cross-linked protein might be also enriched due to a UV-generated covalent bond or very strong association that could not be disrupted by the denaturing buffers. On the other hand, extremely strong interaction partners of the RNA would be enriched in both irradiated and control sample, resulting in no significant difference and dismissal as a false negative.

Peptide level strategies are also based on identification of linear peptides that are neighboring the cross-linked peptide. Before enrichment of the RNA, the proteins are partially digested with endoproteinases LysC or ArgC. Purification is performed for both irradiated and control samples, followed by complete digestion with trypsin, releasing linear peptides that can be identified in a standard proteomics search engine [51–53]. Benefits of this strategy include localizing the cross-linked protein region. Ideally, the detected tryptic peptides would be adjacent to the cross-linked peptide, which could be deduced by *in silico* extension. The reliability of the localization is dependent on the efficiency of the initial enzyme digestion, as the presence of miscleavages could shift the localization window several peptides into either direction. Another limitation is the fact that if the LysC or ArgC fragments do not contain a tryptic peptide within, the cross-linked site is not detected. In comparison with the protein level approach, the number of detectable peptides per protein is substantially decreased, essentially exchanging some of the sensitivity and confidence of the identification for a more precise localization information.

The third type of strategy that is also explored in this study relies on the identification of peptide-RNA heteroconjugates, providing direct evidence of the interaction and in most cases pinpointing the exact amino acid involved in the interaction [39,43,53,54]. The highly specific information comes at the cost of limited sensitivity due to the low yield and suboptimal ionization of the cross-link species. Challenging bioinformatic analysis and labor-intensive manual validation are required by the combinatorial complexity of possible precursors and convoluted fragment spectra. Moreover, deep understanding of the observed adducts and collision induced behavior of the heteroconjugates are essential for an accurate assignment.

1.3.1 Identification of peptide-RNA heteroconjugates

Investigation of protein-RNA contacts by the detection of peptide-RNA cross-links is a challenging task due to the large number of different variants of cross-link products (Fig. 1.6). In order to detect the linkage sites, proteins and RNAs have to be hydrolyzed to a mixture of high abundant peptides and RNA fragments, accompanied by low abundant peptide-RNA heteroconjugates. Due to the limited accessibility of the cross-link sites, enzymatic digestion with RNAses often produces a mixture of cross-links with RNA moiety that typically ranges from one to four nucleotides in length, each of which can be represented by one of the four RNA bases. In addition, a significant number of mass modifications can be observed on each RNA moiety combination (e.g. $-H_2O$, $-HPO_3$, $+HPO_3$), further expanding the possible variants that should be considered [39,55]. Thus, for every peptide that can be generated during proteolytic digestion, a large number of precursor variants has to be calculated and matched to the acquisition data, greatly complicating the analysis.

Due to the physicochemical differences of peptides and RNA, peptide-RNA heteroconjugates create complex fragment spectra. Unlike post-translational modifications such as phosphorylation, every amino acid has the possibility to be linked with UV-generated RNA adduct. When subjected to collision-induced dissociation, the RNA moiety can produce multiple adducts [39,42]. Therefore, a number of possible adduct variants should be considered for every ion in the fragment spectrum, creating a combinatorial complexity that cannot be handled by standard search engines.

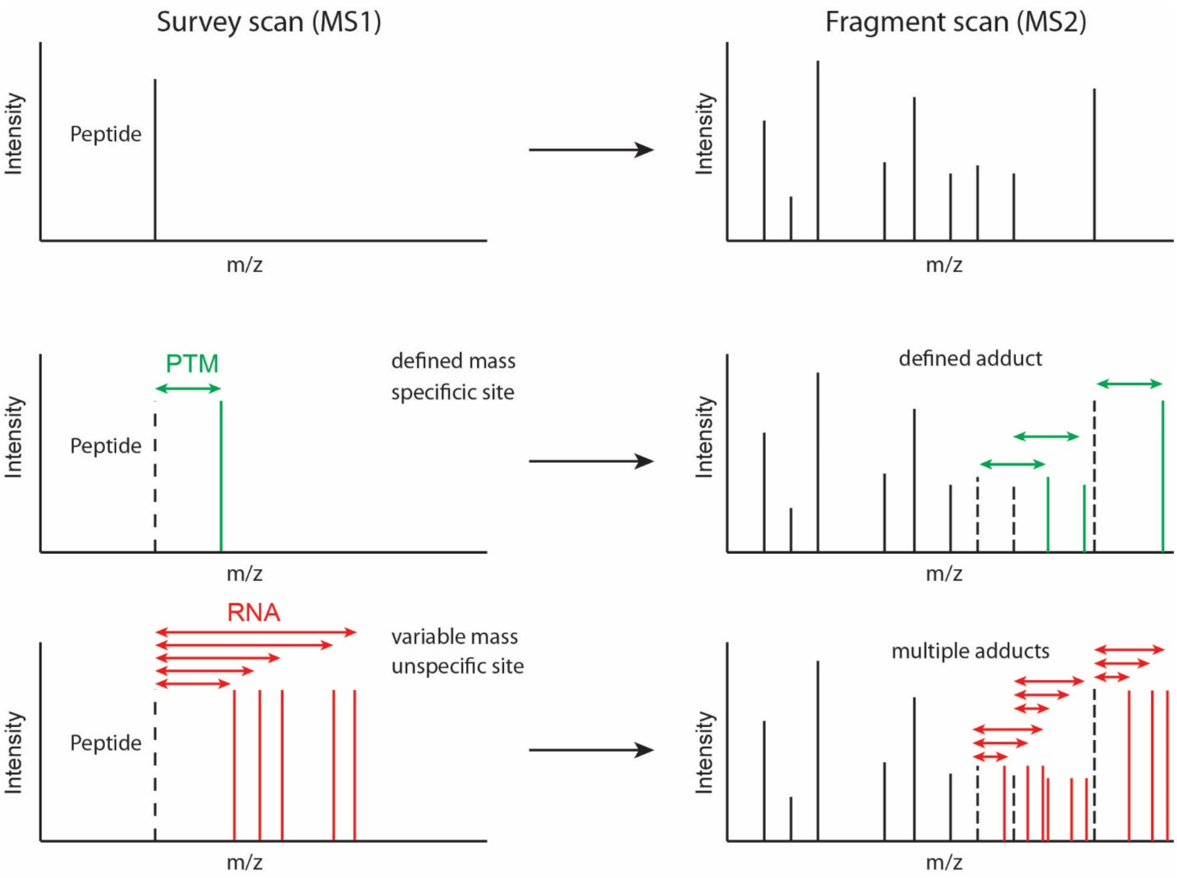


Figure 1.6 Identification of peptide-RNA heteroconjugates

Database search of peptides is achieved through matching the *m/z* value of the precursor and expected fragments (a-, b- and y-ions) to the experimentally acquired spectrum. When a post-translational modification (PTM) is considered, an additional variant of the peptide is calculated by adding the known mass of the modification at the precursor and fragment level. The amino acid specificity and limited fragmentation of modifications lead to only slight complication of the identification search. Identification of peptide-RNA heteroconjugates requires the generation of numerous precursor variants for every peptide. As cross-linking can happen at any position of the peptide and the RNA adduct is prompt to fragmentation down to several different products, multiple variants of every ion could be observed, greatly complicating the data analysis.

1.3.2 Biochemical enrichment of peptide-RNA heteroconjugates

The presence of negatively charged phosphate groups in the RNA moiety of cross-links leads to increased losses and lower efficiency of ionization in positive mode mass spectrometry [56]. In addition, the UV-induced cross-linking reaction has very low yield, creating a low number of covalent bonds between proteins and RNA [57]. The overwhelming abundance of non-cross-link peptides and RNA fragments, generated during the hydrolysis steps of sample preparations, can completely suppress the signal of the cross-link species. Therefore, effective depletion of the non-cross-linked species and enrichment of the peptide-RNA heteroconjugates is a prerequisite for the successful detection and identification of cross-link sites.

Identification of cross-links has been a long-standing interest of the Urlaub Research Group, leading to the development of several purification workflows [39]. Two effective strategies have emerged for the enrichment of cross-links from reconstituted protein-RNA complexes (Fig. 1.7 A). Depletion of non-cross-linked peptides can be achieved by utilizing the size difference between undigested RNA oligonucleotides and peptides [57,58]. The cross-linked protein-RNA complex is digested with endoproteinase (e.g. trypsin) and the RNA containing species are separated from the peptides by size exclusion chromatography in denaturing conditions. Next, the RNA is digested with RNAses and the non-cross-linked RNA fragments are depleted by C18 reversed-phase chromatography, leading to a sample enriched in peptide-RNA heteroconjugates. Alternatively, enrichment may be achieved on the basis of the phosphate groups present in cross-links [59,60]. The protein-RNA complexes are first hydrolyzed with endoproteinases and RNAses, followed by depletion of non-cross-linked RNA fragments by reversed-phase C18 chromatography. The enrichment of peptide-RNA heteroconjugates over non-cross-linked peptides is achieved by immobilized metal affinity chromatography (IMAC) or titanium dioxide enrichment (TiO_2), which can selectively bind phosphorylated organic molecules. Cross-linking investigations of highly complex samples (e.g. yeast cells) that contain a large number of phosphorylated molecules has typically utilized initial purification of the mRNAs and the associated proteins through oligo(dT) hybridization or affinity capture of cap-binding proteins [39]. Peptide-RNA heteroconjugates can be further isolated though size exclusion or C18/TiO_2 workflows.

1.3.3 Data analysis

Initially, processing of mass spectrometric data of UV-irradiated samples was done completely manually by calculating and matching possible cross-link variants, followed by annotation and validation of the fragment spectra. This limited the approach to very simple systems and identification of only high abundant cross-links. To address this problem, computer-aided strategies were developed that eventually lead to the establishment of the RNP^{xl} computational workflow for identification of peptide-RNA heteroconjugates based on the open-source OpenMS project [39,61]. The RNP^{xl} workflow consists of a series of bioinformatic pipelines that assist manual evaluation by reducing the amount of cross-link candidates that need to be considered to a smaller fraction of higher-probability hits (Fig. 1.7 B). To exploit the full power of the workflow, an UV cross-linked sample and non-irradiated control are processed together. The signals present in the control samples are used to remove non-cross-linked species by comparing the extracted-ion chromatogram (XIC filtering). A standard proteomics search is executed to remove fragment spectra that match linear peptides, reducing the data that needs to be considered for cross-link candidates (ID filtering). Finally, matching of cross-links is performed. Possible precursor variants are generated by removing expected RNA adducts from the experimentally observed precursors. In this way, the generation of vast lists of precursor variants from large databases is circumvented, greatly reducing the time of analysis. The subtraction precursor variants are employed in standard proteomics search with the Open Mass Spectrometry Search Algorithm (OMSSA) [62]. As an output, the RNP^{xl} workflow provides a list of cross-link candidates, as well as peptide fragment information. The utilization of the OpenMS framework allows the employment of annotation tools that facilitate subsequent annotation and manual validation of the cross-link candidates.

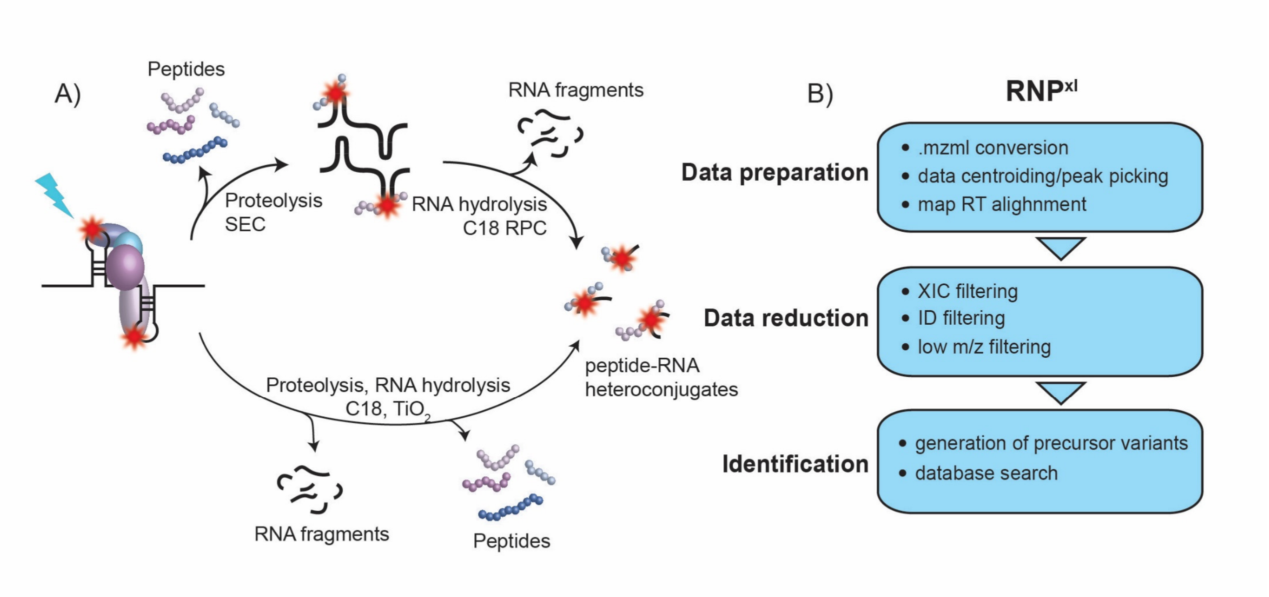


Figure 1.7 Established strategies for enrichment of cross-links and data analysis

A) The low yield of the cross-link reaction necessitates the employment of enrichments strategies to purify peptide-RNA heteroconjugates and deplete competing non-cross-linked species. Two commonly utilized strategies are presented [39]: i) The cross-linked protein-RNA complexes are digested with endoproteinases and the RNA species are separated from the non-cross-linked peptide by size exclusion chromatography. Subsequently, the RNA is hydrolyzed and non-cross-linked RNA fragments are removed by C18 reversed-phase chromatography. Ii) Alternatively, the protein-RNA complex is digested with proteinases and RNAses and non-cross-linked RNA fragments are removed by C18 reversed-phase chromatography. The heteroconjugates are enriched over non-cross-linked peptides by TiO_2 chromatography. B) Overview of the OpenMS RNP^{xl} data analysis workflow [39]. First, the acquisition data is prepared for downstream analysis by conversion to an open source format (.mzml). Signals are centroided and the retention time of the non-irradiated control and cross-linked samples are aligned. To facilitate faster analysis, spectra that do not lead to possible cross-links are removed by comparison with the control sample (XIC filtering), matching to linear peptides (ID filtering) or logical minimum size restrictions (low m/z filtering). The identification is achieved through generation of possible RNA adducts by subtraction from the experimentally observed precursors and standard proteomics search.

1.4 Objectives

UV-induced cross-linking in combination with mass spectrometry can provide valuable information about the contact sites between proteins and RNA. A variety of biochemical enrichment strategies and a dedicated bioinformatic workflow have been developed to tackle the challenges of identifying low abundant cross-link species that come in numerous mass variants. However, the analysis of peptide-RNA heteroconjugates still requires colossal amount of cumbersome manual annotation and expert validation that limits the wide use of the method.

The established RNP^{xl} data analysis workflow automates the identification of cross-link candidates and provides limited fragmentation data about the peptide moiety. This information is insufficient to make a sound judgement of the quality of an identification or to localize the site of the interaction. In order to confirm the validity of the heteroconjugate, the majority of the signals observed in the fragment spectrum have to be annotated and examined for consistency. A major aim of this study is to expand the role of the RNP^{xl} workflow to provide additional information about all types of fragments that can be observed by mass spectrometry. Thus, substituting the role of manual authentication and localization, ultimately leading to fully automated data analysis.

The amount of protein-RNA complexes analyzed by UV cross-linking and mass spectrometry has risen steadily in the last years, yet almost exclusively uracil-conjugated cross-links have been identified. Uracil is the only ribonucleobase for which we have a clear understanding of fragmentation behavior and expected mass adducts, limiting the type of protein-RNA contacts that can be surveyed with this approach. This is inconsistent with existing literature observations, where cross-linking to all four nucleobases is expected to occur to a certain extent. An additional objective of this study is to systematically investigate and describe the UV-induced RNA adducts that can be successfully analyzed by mass spectrometry.

Purification of cross-links derived from complex systems is very challenging, owing to the large numbers of metabolites and modified peptides that hamper the enrichment process. Substantial success has been achieved with methods targeting mRNA species, while the application of unbiased strategies for purification of total peptide-RNA heteroconjugates has provided unsatisfactory results. There is a need for universal enrichment method that enables the efficient depletion of peptides and metabolites derived from complex systems. This problem is addressed with the assessment of solid-phase RNA extraction techniques as possible alternative workflows.

Existing strategies for identification of peptide-RNA heteroconjugates were developed by adaptation of proteomics workflows. Therefore, predominantly insight about the peptide moiety is obtained. The final goal of this work is to explore the theoretical capabilities of nanoelectrospray mass spectrometry to deliver additional information about peptide-RNA heteroconjugates, with the focus set on elucidating the RNA moiety.

2. Materials and methods

2.1 Materials

2.1.1 Chemicals, solvents and reagents

2,5-dihydroxybenzoic acid (DHB)	Sigma-Aldrich (Germany)
2-mercaptoethanol	Roth (Germany)
4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)	Sigma-Aldrich (Germany)
Acetic acid	Merck (Germany)
Acetonitrile (ACN, LiChrosolv gradient grade)	Merck (Germany)
Ammonium hydroxide (NH4OH, 25% v/v)	Merck (Germany)
Bis(2-hydroxyethyl)amino-tris(hydroxymethyl)methane (bis-Tris)	Sigma-Aldrich (Germany)
Calcium chloride (CaCl2)	Merck (Germany)
Chloroform	Merck (Germany)
Coomassie Brilliant Blue G-250	Sigma-Aldrich (Germany)
Copper(I) bromide (CuBr)	Sigma-Aldrich (Germany)
Ethanol	Merck (Germany)
Ethylenediaminetetraacetic acid (EDTA)	Roth (Germany)
Formaldehyde (37% v/v)	Sigma-Aldrich (Germany)
Formic acid (FA)	Sigma-Aldrich (Germany)
Glucose	Merck (Germany)
Glycerol	Merck (Germany)
Guanidine hydrochloride (GdnHCL)	Sigma-Aldrich (Germany)
Hexafluoroisopropanol (HFIP)	Sigma-Aldrich (Germany)
Isopropanol	Merck (Germany)
Magnesium chloride (MgCl2)	Sigma-Aldrich (Germany)
Magnesium sulfate (MgSO4)	Merck (Germany)
Methanol (MeOH, LiChrosolv gradient grade)	Merck (Germany)
Ortho-phosphoric acid	Merck (Germany)
Silver nitrate (AgNO3)	Sigma-Aldrich (Germany)
Sodium carbonate (Na2CO3)	Merck (Germany)
Sodium chloride (NaCl)	Merck (Germany)
Sodium dodecyl sulfate (SDS)	Serva Electrophoresis (Germany)
Sodium hydrogen sulfite (39% w/v)	Merck (Darmstadt, Germany)
Sodium metabisulfite	Sigma-Aldrich (Germany)
Sodium thiosulfate (Na2S2O3)	Merck (Darmstadt, Germany)
Triethylamine (TEA)	Sigma-Aldrich (Germany)
Trifluoroacetic acid (TFA)	Roth (Germany)
Tris(benzyltriazolylmethyl)amine	Sigma-Aldrich (Germany)
Tris(hydroxymethyl)aminomethane (Tris)	Roth (Germany)

Uracil	Sigma-Aldrich (Germany)
Urea	Sigma-Aldrich (Germany)
Water (LiChrosolv gradient grade)	Merck (Germany)
Zinc chloride (ZnCl2)	Sigma-Aldrich (Germany)

2.1.2 Enzymes

Antarctic phosphatase 5 000 U/ml	New England Biolabs (Germany)
Benzonase 25 U/µl	Novagen, Merck (Germany)
DNAse I 6 U/µl	Zymo Research (Germany)
Lysozyme from chicken egg white ~100 000 U/mg	Sigma-Aldrich (Germany)
Nuclease P1 100 000 U/ml	New England Biolabs (Germany)
Pierce Universal nuclease 250 U/µl	Thermo Fischer Scientific (Germany)
RNAse A 1 mg/ml	Ambion, Applied Biosystems (Germany)
RNAse I 10 U/µl	Thermo Fischer Scientific (Germany)
RNAse T1 1000 U/µl	Ambion, Applied Biosystems (Germany)
Trypsin (sequencing grade)	Promega (USA)

2.1.3 Proteins, peptides and (oligo)nucleotides

Glyceraldehyde-3-phosphate Dehydrogenase from rabbit muscle (GAPDH)	Sigma-Aldrich (Germany)
Yeast protein Hsh49	Kindly provided by Alexander Wulf (Bioanalytical Mass Spectrometry, MPIbpc)
poly(U)$_{25}$-3'-biotin	Purimex (Germany)
poly(G)$_{25}$-3'-biotin	Purimex (Germany)
poly(C)$_{25}$-3'-biotin	Purimex (Germany)
poly(A)$_{25}$-3'-biotin	Purimex (Germany)
Uridine-5′-monophosphate	Sigma-Aldrich (Germany)
Uridine-^{15}N$_2$ 5′-monophosphate	Sigma-Aldrich (Germany)
Uridine-^{13}C$_9$,^{15}N$_2$ 5′-monophosphate	Sigma-Aldrich (Germany)
Guanosine-5′-monophosphate	Sigma-Aldrich (Germany)
Guanosine-^{15}N$_5$ 5′-monophosphate	Sigma-Aldrich (Germany)
Guanosine-^{13}C$_{10}$,^{15}N$_5$ 5′-monophosphate	Sigma-Aldrich (Germany)
Cytidine-5′-monophosphate	Sigma-Aldrich (Germany)
Cytidine-^{15}N$_3$ 5′-monophosphate	Sigma-Aldrich (Germany)
Cytidine-^{13}C$_9$,^{15}N$_3$ 5′-monophosphate	Sigma-Aldrich (Germany)
Adenosine-5′-monophosphate	Sigma-Aldrich (Germany)
Adenosine-^{15}N$_5$ 5′-monophosphate	Sigma-Aldrich (Germany)
Adenosine-^{13}C$_{10}$,^{15}N$_5$ 5′-monophosphate	Sigma-Aldrich (Germany)

5’-UAGACAU*UGCAGUCACAG-3’ *=(5-ethynyl-2’-deoxyuridine)	Baseclick (Germany)
ALYTFAEGF*K *= (4-azidophenylalanine)	Eurogentec (Belgium)
4-thiouridine	Carbosynth (United Kingdom)

2.1.4 Commercial kits and buffers

InstantBlue Protein Stain	Expedeon (United Kingdom)
RNeasy Maxi Kit (RPE, RLT and RW1 buffer)	Qiagen (Germany)
Invitrogen TRIzol Reagent	Thermo Fischer Scientific (Germany)
Direct-zol RNA Miniprep Plus (RNA PreWash and Wash buffer)	Zymo Research (Germany)
Biotin Chromogenic Detection Kit (Streptavidin-AP Conjugate, Washing Buffer, Blocking Solution, Detection Buffer, Substrate Solution)	Thermo Fischer Scientific
Antarctic Phosphatase Reaction Buffer	New England Biolabs (Germany)
NuPAGE MOPS SDS Buffer Kit (MOPS SDS Running Buffer, Sample Reducing Agent, LDS Sample Buffer)	Thermo Fischer Scientific (Germany)
Pierce BCA Protein Assay Kit	Thermo Fischer Scientific (Germany)
Triethylammonium bicarbonate buffer 1 M pH 8.5 (TEAB)	Sigma-Aldrich (Germany)
Triethylammonium acetate buffer 1 M pH 7 (TEAA)	Sigma-Aldrich (Germany)
Phosphate-Buffered Saline pH7.4 (PBS)	Thermo Fischer Scientific (Germany)

2.1.5 Commonly used buffers and solutions

LC-MS Loading Buffer	2% (v/v) ACN 0.05% (v/v) TFA
Colloidal Coomassie stain solution	20% (v/v) Methanol 0.08% (w/v) Coomassie Brilliant G-250 8% Ammonium sulfate 1.6% (v/v) Ortho-phosphoric acid
Sodium acetate 3 M pH 5.2	24.6 g sodium acetate in 100 ml water, pH adjusted with glacial acetic acid

SAX Separation buffer	6M Urea 50 mM bis-Tris pH6 400 mM NaCl
SAX Elution buffer	2 M NaCl 50 mM bis-Tris pH 6

2.1.6 Other consumables

Titansphere TiO2 Bulk 10 µm	GL Sciences (Japan)
Pierce Strong Anion Exchange Spin Columns (Mini, Maxi)	Thermo Fischer Scientific (Germany)
C18 Micro SpinColumns	Harvard Apparatus ( )
NuPAGE 4-12% Bis-Tris Protein Gels 1mm	Thermo Fischer Scientific (Germany)
Phase Lock Gel Tubes	Quantabio (USA)
Zeba Spin Desalting Columns (7K MWCO, 0.5 mL)	Thermo Fischer Scientific (Germany)
Amersham Hybond P PVDF membrane	Sigma-Aldrich (Germany)
Distal Coated SilicaTip Emitter	New Objective (USA)
Cell Culture Dish 60/15,145/20 mm	Greiner Bio-One (Austria)
Mono Q 5/50 GL SAX column	Sigma-Aldrich (Germany)
Sep-Pak Vac C18 columns 1cc	Waters (Germany)
Reprosil-Pur basic C18	Dr. Maisch (Germany)
Diamond Tower Pack tips	Gilson (Germany)
Falcon tubes	Greiner Bio-One (Austria)
Safe-Lock Tubes	Eppendorf (Germany)

2.1.7 *E. coli* strains and media components

E. coli XL10-Gold	Stratagene (USA) Kindly provided by Dr. Constantin Cretu (Macromolecular Crystallography, MPIbpc)
E. coli K-12 BW25113	Dharmacon (USA)
Keio Knockout pyrD [63]	
M9 Minimal Salts Base	Formedium (United Kingdom)
Casamino acids	Formedium (United Kingdom)
LB medium	MP Biomedicals (Germany)
LB-agar medium	MP Biomedicals (Germany)
Kanamycin	Roth (Germany)

2.1.8 Instruments and laboratory equipment

Heraeus Multifuge X3R	Thermo Fischer Scientific (Germany)
Heraeus Fresco 17 Microcentrifuge	Thermo Fischer Scientific (Germany))
Heraeus Pico 17 Microcentrifuge	Thermo Fischer Scientific (Germany))
Heraeus HERAsafe HS Safety Cabinet	Thermo Fischer Scientific (Germany))
UV Cross-linking apparatus build in-house 4x8W lamps 254 nm / 365 nm	Sankyo Denki (Japan)
Thermomixer comfort	Eppendorf (Germany)
Thermomixer C	Eppendorf (Germany)
Pharmacia Ultrospec 3000 pro	GE Healthcare (Germany)
NanoDrop 1000 Spectrophotometer	Thermo Fischer Scientific (Germany)
Savant SPD121P Speed Vac	Thermo Fischer Scientific (Germany)
Eppendorf Concentrator 5301	Eppendorf (Germany)
Vortex-Genie 2	Scientific Industries (USA)
Pharmacia Amersham EPS 350	GE Healthcare (Germany)
PowerPack 200	Bio-Rad Laboratories (Germany)
PerfectBlue 'Semi-Dry' Electro Blotter	Peqlab (Germany)
Sonorex Super RK 103 H	Bandelin (Germany)
XCell SureLock Mini-Cell	Thermo Fischer Scientific (Germany)
Orion 2-Star Benchtop pH meter	Thermo Fischer Scientific (Germany)
ÄKTAmicro	GE Healthcare (Munich, Germany)
Varioklav Classic 400	Thermo Fischer Scientific (Germany)
Sonifier cell disrupter S-450D	Emerson Electric (USA)
Dionex Ultimate 3000 UHPLC	Thermo Fischer Scientific (Germany)
Linear Ion Trap XL (LTQ XL)	Thermo Fischer Scientific (Germany)
Q Exactive HF	Thermo Fischer Scientific (Germany)
Orbitrap Fusion Tribrid	Thermo Fischer Scientific (Germany)
Orbitrap Fusion Lumos Tribrid	Thermo Fischer Scientific (Germany)

2.1.9 Software and online tools

MaxQuant 1.5.0.3	Max Planck Institute for Biochemistry (Germany)
KNIME Analytics Platform	KNIME (Switzerland)
OpenMS	University of Tübingen (Germany)
PyCharm	JetBrains (Czech Republic)
Python	Python Software Foundation (USA)
R Studio	R Studio (USA)
R	The R Foundation for Statistical Computing (Austria)
Xcalibur 4.1	Thermo Fischer Scientific (Germany)
Proteome Discoverer 2.1	Thermo Fischer Scientific (Germany)
RoboOligo	University of Cincinnati (USA)
UCSF Chimera 1.14	University of California (USA)
ChemSketch 2015	ACD Labs
Adobe Creative Suite 5	Adobe (California)
Microsoft Office 2016	Microsoft Corporation (USA)
ProteinProspector (http://prospector.ucsf.edu)	University of California (USA)
Mongo Oligo Mass Calculator (http://mods.rna.albany.edu/masspec/Mongo-Oligo)	University at Albany (USA)
STRING database 11 https://string-db.org/	STRING [64]
UniProt database https://www.uniprot.org/	UniProt Consortium (2017)
ProtParam tool https://web.expasy.org/protparam/	ExPASy Server [65]

2.2 Methods

2.2.1 Standard biochemical methods

2.2.1.1 Alcohol precipitation

Protein and RNA were precipitated by adding 3 volumes of ice-cold ethanol or 1 volume isopropanol and 1/10 volume of 3 M sodium acetate pH 5.2. Incubation was performed at -20 °C for 2 hours, followed by 16,000x*g* centrifugation at 4 °C for 30 minutes. The pellet was washed twice with 80% (v/v) ice-cold ethanol, centrifuged as described above and air-dried for 2 minutes, after which it was resuspended in the appropriate buffer for subsequent processing.

2.2.1.2 Purification of total RNA from bacterial cells

Total RNA purification was performed with Qiagen RNeasy Maxi kit according to manufacturer’s protocol or by using TRIzol extraction as explained below. For TRIzol extraction pelleted cells were combined with 1 ml of TRIzol reagent, incubated for 5 minutes at room temperature and centrifuged for 5 minutes at 12,000x*g*. The supernatant was transferred to Phase Lock Gel 2 ml tubes and 600 µl chloroform were added. Phase separation was achieved at 12,000x*g* for 20 minutes. The upper aqueous phase was transferred to a new tube, precipitated with 1 ml of isopropanol and centrifuged for 10 minutes at 12,000x*g* at 4 °C. The pellet was washed with 1 ml of 80% (v/v) ethanol and centrifuged for 5 minutes at 7,500x*g* at 4 °C. The RNA was air-dried for 5-10 minutes and resuspended in RNAse-free water. Concentration, yield and quality were estimated spectrophotometrically.

2.2.1.3 Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE)

Separation of proteins was performed with the NuPAGE system. The samples were supplemented with 1x Sample Reducing Agent, 1x LDS Sample Buffer and heated for 10 minutes at 70 °C. Separation was performed with MOPS SDS Running Buffer, on 4-12% Bis-Tris 1.0 mm gels for 50-60 minutes at 200 V.

Visualization of proteins after SDS-PAGE was performed with colloidal Coomassie or InstantBlue Protein Stain [66]. The gel was submerged in the staining solution and incubated overnight, while shaking. Destaining was achieved by washing the gel several times with deionized water.

2.2.1.4 Silver staining of polyacrylamide gels

The silver staining protocol was adapted from [67]. All incubation and washing steps were performed with at least 10 times the gel volume. The gels were fixed for at least 3 hours with 50% (v/v) methanol, 10% (v/v) acetic acid, followed by washing twice in 50% (v/v) ethanol and once with 30% (v/v) ethanol for about 20 minutes. The gel was sensitized with 0.8 mM $Na_2S_2O_3$ for 60 seconds and washed with deionized water in 3 steps for 20 seconds. Impregnation was performed with 2 g/l $AgNO_3$ in the presence of 0.026% (v/v) formaldehyde for sensitivity increase. The gels were washed 3 times with water for 20 seconds. The silver staining was developed with 60 g/l Na_2CO_3, 0.0185% (v/v) formaldehyde, 16 µM $Na_2S_2O_3$ for 2 to 10 minutes depending on the desired staining intensity. Next, the gels were washed twice for 2 minutes each with deionized water. The reaction was stopped by adding 50% (v/v) methanol, 10% (v/v) acetic acid.

2.2.1.5 North-western blot

The SDS gels were blotted onto a Amersham Hybond PVDF blotting membrane via semi-dry transfer for 90 minutes with 0.8 mA/cm^2 (transfer buffer; 20% (v/v) methanol, 50 mM MOPS, 50 mM Tris Base, 0.1% (w/v) SDS, 1 mM EDTA, pH 7.7). The blotted proteins were stained with colloidal Coomassie for 30 seconds. Destaining was performed with 50% (v/v) methanol, 10% (v/v) acetic acid. The membrane was reactivated for 20 second with methanol and developed with the Biotin Chromogenic Detection Kit. All steps were performed with gentle shaking. Blocking was completed for 30 minutes with Blocking Buffer, followed by 30 minutes incubation with the Streptavidin-AP Conjugate. Unbound conjugate was washed 3 times with Washing Buffer. The membrane was incubated for 10 minutes with Detection Buffer and the blot was developed in the dark for about 10 minutes with Substrate Solution. The reaction was stopped by rinsing the membrane with deionized water.

2.2.1.6 Estimation of protein and nucleic acid concentration

Protein concentration of complex mixtures (e.g. HeLa cytoplasmic extract) was determined with the use of the Pierce BCA Protein Assay kit, according to the manufacturer's protocol. Briefly – standard curve was prepared with BSA in the range of 125 ng/µl to 2 µg/µl. Dilution series of the protein sample were prepared to final volume of 100 µl in two replicates, supplemented with 2 ml of Working Reagent and incubated for 30 minutes at 37 °C. The absorbance at 562 nm was measured and concentration was estimated based on the reference standard curve.

The concentration of purified proteins was estimated using the absorbance at 280 nm. Extinction coefficients of the proteins were calculated using the ProtParam tool. Spectrophotometric measurements were performed in triplicates on NanoDrop.

Determination of the concentration and quality of purified RNA and DNA was performed spectrophotometrically by measuring the sample absorbance profile (assuming 40 ng-cm/µl for RNA and 50 ng-cm/µl for DNA).

For rough estimation of expected yield of total protein and RNA from whole cells the following presumptions were employed:

Table 2.1 Macromolecular components of *E. coli* and HeLa cells

Assumed for *E. coli* OD_{600} of 1.0 ≈ $8x10^8$ cells/ml

Component	**Amount per HeLa Cell**	**Amount per *E. coli* cell**
Total dry weight	400 pg	0.4 pg
Total DNA	15 pg	0.017 pg
Total RNA	30 pg	0.10 pg
Total protein	300 pg	0.2 pg
Cytoplasmic ribosomes	4×10^6	3×10^4
Cytoplasmic tRNA molecules	6×10^7	4×10^5
Cytoplasmic mRNA molecules	7×10^5	4×10^3

Values collected from the website of Thermo Fischer Scientific [68]

2.2.1.7 RNeasy silica-based purification

Isolation of long RNA-containing species (>200 nucleotides) was performed with Qiagen RNeasy Maxi kit. Centrifugation steps were performed at 3,000x*g*. The sample was mixed with an appropriate amount of RLT buffer supplemented with β-mercaptoethanol, according to manufacturer's protocol and loaded onto the silica column by centrifugation. Weak interactors were washed away with high concentration of chaotropic agents by addition of 15 ml RW1 buffer and centrifugation for 5 minutes. Salts were removed by two wash steps with 10 ml RPE buffer each and spinning 2 and 10 minutes to ensure removal of residual ethanol. Elution was performed in two steps by addition of 1.2 ml RNAse-free water, incubation for 1 minute at room temperature and collection for 3 minutes by centrifugation.

2.2.1.8 TRIzol-assisted silica-based purification

Enrichment of RNA species longer than 17 nucleotides was performed with Direct-zol RNA Miniprep Plus kit. Centrifugation steps were performed at 16,000x*g* for 30 seconds. Samples were mixed thoroughly with 3 volumes of TRIzol reagent and 4 volumes 100% ethanol. The mixture was loaded onto a silica column by centrifugation. For samples containing significant amounts of DNA, the column was washed with 400 µl RNA Wash

Buffer and DNA digestion was performed with 75 µl DNA Digestion Buffer and 5 µl DNAseI (6 U/µl) for 15 minutes at room temperature. The bound RNA was washed with 400 µl RNA PreWash buffer and 700 µl RNA Wash Buffer, followed by centrifugation for 2 minutes to ensure complete removal of the wash buffer. Elution was performed with 100-150 µl RNAse-free water.

2.2.1.9 Strong anion exchange chromatography (SAX) of peptides, proteins and RNA

To generate complex mixture of peptides, 400 µg of HeLa nuclear extract (kindly provided by the Department of Cellular Biochemistry, MPIbpc) was digested with 4 µg of trypsin overnight in the presence of 25 mM Tris-HCl pH 7.9 (100 µl final volume). Three technical replicates were performed. 25 µl of the digested mixture was supplemented with 375 µl 4 M Urea, 50 mM bis-Tris-HCl pH 6, 200 mM NaCl and loaded onto Pierce SAX (Q) spin column mini. Centrifugation steps were performed at 2,000x*g* for 5 minutes, with 400 µl buffer. The peptides were eluted stepwise with increasing concentration of NaCl (200, 400, 600, 800, 1000 mM) in 4 M Urea, 50 mM bis-Tris-HCl pH 6 buffers. The flow through and the 200 mM elution step were combined. Samples were desalted by C18 reversed-phase chromatography using Harvard Apparatus SpinColumns, dried under vacuum and resuspended in 50 µl 50% (v/v) ACN, 0.1% (v/v) FA by vortexing and sonication for 2 minutes. Additional 450 µl of 5% ACN, 0.1% FA were added and 5 µl was used for LC-MS analysis.

To test the behavior of intact proteins, HeLa nuclear extract (~100 µg) was combined with 400 µl 4 M Urea, 50 mM bis-Tris-HCl pH 6. The sample was loaded onto Pierce SAX (Q) spin column mini and eluted in stepwise manner with increasing NaCl concentration (200, 400, 600, 800, 1000, 2000 mM). The fractions were ethanol precipitated, resuspended in LDS sample buffer and analyzed by SDS-PAGE with Coomassie staining.

Roughly 75 µg of total RNA (~50 µl), purified from pyrD *E. coli* was supplemented with 375 µl 4 M Urea, 50 mM bis-Tris-HCl pH 6 and loaded onto Pierce SAX (Q) spin column mini. Purification and stepwise elution was performed as described above. The fractions were loaded and separated on an SDS-PAGE gel and fixed overnight in 50 % (v/v) methanol, 10 % (v/v) acetic acid, followed by silver staining.

2.2.1.10 Reversed-phase chromatography (C18)

C18 Reversed-phase chromatography was utilized for desalting peptide samples before LC-MS analysis and for depletion of polar small molecules, such as digested nucleotides. Depending on the amount of peptides and sample volume, the following option were used: Harvard Apparatus C18 SpinColumns, Sep-Pak C18 1 cc Vac Cartridge, in-house assembled columns (AQ 120 Å 5 µM, Dr. Maisch GmbH) [42] or StageTips [69]. The samples were adjusted to 5% (v/v) ACN and acidified with FA or TFA to 0.1% (v/v) final concentration. The columns were activated with methanol and equilibrated with elution (80% (v/v) ACN, 0.1% (v/v) FA or TFA) and washing buffer (5% (v/v) ACN, 0.1% (v/v) FA or TFA). The peptide sample was loaded onto the column, washed several times with washing buffer and eluted with elution buffer containing high amount of organic solvent.

2.2.1.11 Titanium dioxide enrichment of cross-linked peptides (TiO_2)

Titanium dioxide purification was used for the depletion of non-cross-linked peptides [39]. Spin columns were assembled as described in [42], with Titansphere TiO_2 Bulk 10 µm beads as chromatographic matrix. Depending on the starting material and the size of the column, subsequent washing steps were done with either 60 or 200 µl. The beads were equilibrated with washing Buffer B (80% (v/v) ACN, 5% (v/v) TFA) and Buffer A (200 mg/ml DHB or 5% (v/v) glycerol, 80% (v/v) ACN, 5% (v/v) TFA). The digested peptide mixture was resuspended in Buffer A and loaded onto the column. Unspecific interactors were removed with 3 washes of Buffer A and the competitor compound was eliminated with 3 washes of Buffer B. Elution was performed with 0.3 M NH4OH and the sample was dried in a centrifugal evaporator.

2.2.2 Cell culture and UV cross-linking

UV irradiation at 254 nm or 365 nm was performed with in-house build cross-linking apparatus as previously described in [42].

2.2.2.1 Bacterial cell culture and media

E. coli cells were cultured in Lysogeny broth (LB) medium and minimal supplemented M9 media. Media and solutions were sterilized by autoclaving at 121 °C for 15 min or filter-sterilized. The medium was supplemented with 25 µg/ml kanamycin prior to inoculation.

Cells to be irradiated at 254 nm were inoculated at a starting OD_{600} of 0.05 in 1 L LB medium and incubated at 37 °C and 160 rpm until OD_{600} of 0.6 was reached. Cells were harvested with centrifugation at 4,000x*g* at 25 °C. The pellet was resuspended in 48 ml of ice-cold PBS pH 7.4, diluted to OD_{600} of 10 and transferred to 145 mm petri dish. UV irradiation was carried out for 2 hours on ice in a 4 °C room with constant gentle shaking. The cells were collected by 40 minutes of centrifugation at 4 °C, 4,000x*g* and pellets were stored at -80 °C.

4-thiouridine incorporation was performed in supplemented M9 medium. 400 ml of medium, supplemented with 100 µM 4-thiouridine were inoculated with *E. coli* pyrD cells to initial OD_{600} of 0.1. The culture was incubated at 30 °C, 160 rpm until OD_{600} of 0.45 was reached. 4-thiouridine was added to a final concentration of 1 mM, followed by 2 hour incubation at 30 °C, 160 rpm. Afterwards, an additional 400 ml of M9 supplemented media with 1 mM 4-thiouridine was added and the cells were incubated for another hour. Harvesting was performed with 4,000x*g* for 40 minutes at 25 °C. The pellet was resuspended in ice-cold PBS pH 7.4 to OD_{600} of 10. The cells were irradiated as described above for 1 hour at 365 nm, followed by centrifugation for 40 minutes with 4,000x*g* at 4 °C.

Supplemented M9 medium

Component (stock solution)	End concentration
M9 salt solution (5x)	1x
Glucose (20 % w/v)	0.4 % (w/v)
CaCl2 (1M)	100 µM
MgSO4 (1M)	2 mM
Casamino acids (10 % w/v)	0.2 % (w/v)
Uracil (10mg/ml in DMSO)	10 µg/ml

Lysogeny broth medium (LB)

Component	End concentration
Tryptone	1 % (w/v)
Yeast extract	0.5 % (w/v)
NaCl	1 % (w/v)
Casamino acids (10%)	0.2% (w/v)
Uracil (10mg/ml in DMSO)	10 µg/ml

2.2.2.2 Disruption of bacterial cells

For disruption of *E. coli* cells, frozen pellets (~3.2x10^{11} cells / 254 nm; ~4x10^{11}cells / 4SU-365nm) were mixed with 9 mg of lysozyme in 300 µl PBS pH 7.4 and incubated for 5 minutes at room temperature. To the cells 3 ml of 8M Urea, 100 mM HEPES-NaOH pH 7.4, 20 mM EDTA were added. Cells were disrupted on ice by sonication (0.5/2s on/off, 150 pulses, 30% vibration amplitude) using a Sonifier cell disrupter. The lysate was diluted to 24 ml and 1M Urea final concentration with RNAse-free water.

2.2.2.3 Generation of HeLa cytoplasmic extract

Crude HeLa cytoplasmic extract was originally produced by the HeLa Bioreactor Facility of MPIbpc and kindly provided by Dr. Olexandr Dybkov (Department of Cellular Biochemistry, MPIbpc). Shortly - HeLa S3 cells were grown to density 6x10^6 cells/ml and pelleted at 1,300x*g* for 10 minutes at 4 °C. The cells were washed with PBS pH 7.4 and resuspended in 10 mM HEPES-KOH pH 7.6, 10 mM potassium acetate, 0.5 mM magnesium acetate. After 5 minutes swelling on ice the cells were homogenized with a glass douncer and nuclei were separated by centrifugation at 18,000x*g* for 5 minutes.

Cross-linking was performed for 10 minutes at 254 nm in 2 ml fractions of cytoplasmic extract placed in 6 cm wide petri dishes. For each condition, 1 ml of cross-linked extract was used.

2.2.3 Enrichment of peptide-RNA heteroconjugates from simple mixtures

2.2.3.1 Cross-linking of the mtRNAP/TEFM complex

Protein samples in complex with RNA/DNA scaffold [70] (5’ to 3’ DNA - CATGGGGTAACTAGTTCGACGCCAGACG; CGTCTGGCGTGATCACGACTACCCCATG and RNA - UGAUGGUAAUGCUCCUGUCGUGAUC) were kindly provided by Dr. Hauke Hillen (Department of Molecular biology, MPIbpc). Three types of protein-RNA/DNA complexes were analyzed - mtRNAp/TEFM complex (~ 2 nmol) and individual proteins mtRNAp (~ 1.93 nmol) and TEFM (~ 0.70 nmol). The complexes were split into two for irradiated and non-irradiated sample that were processed in parallel. Cross-linking was performed for 5 minutes at 254 nm. The samples were ethanol precipitated and the pellet was dissolved in 50 µl 50 mM Tris-HCl pH 7.9, 6 M Urea. An additional 250 µl 50 mM Tris-HCl pH 7.9 were added to lower the chaotropic agent concentration to levels appropriate for enzymatic digestion. Incubation steps were performed at 500 rpm in a thermomixer. To the mtRNAP/TEFM sample 2 µl of RNAse I was added and it was incubated at 37 °C for 2 hours. The mtRNAP and TEFM samples were digested with 2 µl RNAse T1 for 2 hours at 52 °C. All samples were adjusted to 5 mM $MgCl_2$, supplemented with 2 µl Benzonase and incubated for an additional 1 hour at 37 °C. Protease digestion was performed overnight with trypsin at 37 °C at a 1:20 (w/w) enzyme to protein ratio. The next day the samples were supplemented with additional trypsin (0.5 µg for mtRNAP and TEFM; 2 µg for mtRNAP/TEFM), as well as RNAse I, RNAse T1 and Benzonase (2 µl each) and incubated for 2 hours at 37 °C, followed by 1h incubation at 52 °C. Afterwards ACN and TFA were added to final concentration of 5% (v/v) and 1% (v/v), respectively. The samples were depleted of free nucleotides by reversed-phase C18 chromatography, dried and resuspended in 100 µl 200 mg/ml DHB, 80% (v/v) ACN, 5% (v/v) TFA. Standard TiO_2 enrichment protocol in the presence of DHB was carried out. The sample were dried under vacuum, peptides were resuspended in 25 µl 2% (v/v) ACN, 0.1% (v/v) FA by vortexing and sonication for 2 minutes. 5 µl were used for LC-MS analysis.

2.2.3.2 Cross-linking with homopolyribonucleotides

Rabbit GAPDH and Hsh49 were utilized for systemic analysis of cross-linking products with homopolyribonucleotides. Control and cross-link samples were processed in parallel. In each sample 30 µg of protein were mixed with equimolar amount of poly(U)$_{25}$-3'-biotin, poly(G)$_{25}$-3'-biotin, poly(C)$_{25}$-3'-biotin or poly(A)$_{25}$-3'-biotin oligonucleotides. Cross-linking was performed for 2 (Hsh49) and 5 (GAPDH) minutes at 254 nm. From the samples, 10 % (~3 µg protein amount) were removed for North-western blot analysis with Biotin Chromogenic Detection Kit, the rest was used for C18/TiO_2 purification. Urea was added to 1 M final concentration and RNAse digestion was performed for 2 hours at 37 °C. RNases were selected according to the sequence of the homopolyribonucleotide: poly(U) - 1 µl RNAse I and 1µl RNAse A; poly(G) – 1 µl RNA I and 1 µl RNAse T1; poly(C) - 1 µl RNAse I and 1ul RNAse A; poly(A) – 2 µl RNAse I. Next, overnight protease digestion was performed with trypsin at 1:20 (w/w) ratio at 37 °C. Digested nucleotides were depleted by C18 Reversed-phase chromatography (Harvard Apparatus C18 SpinColumn). TiO_2 enrichment was achieved with glycerol as competitor. The purified peptides were dried under vacuum and resuspended in 2 µl 50% (v/v) ACN, 0.1% (v/v) FA by vortexing and sonication for 1 minute. The sample was diluted with 15 µl 0.1% (v/v) FA and 5 µl were used for LC-MS analysis.

2.2.3.3 Cross-linking with individual nucleotide monophosphates

Cross-linking experiments were performed with rabbit GAPDH for all four ribonucleotides monophosphates as light and heavy isotope substituted versions (^{15}N and $^{15}N^{13}C$). To form a complex, 2 nmol of protein was mixed with 320 nmol of nucleotide (40 µl final volume in 30 mM Tris-HCl pH 7.9). The mixture was incubated on ice for 20 minutes and irradiated for 10 minutes at 254 nm. Additional 70 µl of 50 mM Tris-HCl pH 7.9 were added and the non-cross-linked nucleotides were depleted by Zeba Spin Column (7K MWCO). The desalted mixture was supplemented with 5 µl 1M Tris-HCl pH 7.9 and overnight trypsin digestion was performed (1:20 w/w, 37 °C). ACN and FA were added to a final concentration of 5% (v/v) and 0.1% (v/v), followed by C18 Reversed-phase chromatography (Harvard Apparatus C18 SpinColumns). Elution was performed with 80% (v/v) ACN, 0.5% (v/v) TFA, 5% (v/v) glycerol. The peptide mixture was incubated with 1 mg of Titansphere TiO_2 Bulk 10 µm beads for 10 minutes at room temperature and transferred onto a Harvard Apparatus C18 SpinColumn. Unspecific interactors were washed away twice with 80% (v/v) ACN, 0.5% (v/v) TFA, 5% (v/v) glycerol and 80% (v/v) ACN, 0.5% (v/v) TFA, and once with 8% (v/v) ACN, 0.05% (v/v) TFA. The cross-linked peptides were eluted from the TiO_2 beads to the underlying C18 matrix with 500 mM Na_2HPO_4 and 0.3 M NH_4OH. Desalting was accomplished with 5% (v/v) ACN, 0.1% (v/v) FA, followed by elution with 80% (v/v) ACN, 0.1% (v/v) FA. Peptides were dried under vacuum, resuspended in 25 µl LC-MS Loading Buffer and 5 µl were used for LC-MS analysis.

2.2.3.4 Protein-RNA cross-linking of rabbit GAPDH with isolated *E. coli* RNA

To form protein-RNA complex, 600 µg of GAPDH were incubated with 600 µg of *E. coli* derived total RNA for 30 minutes on ice. The complex was cross-linked for 10 minutes at 254nm. Urea was added to 1 M final concentration and proteins were digested with trypsin (1:10 w/w) for 2 hours at 37 °C. Purification of cross-linked peptides was performed with Zymo Direct-zol Miniprep Plus kit. After elution, urea was added to 1 M end concentration and the RNA was digested overnight with 1 µl RNAse I, 1 µl RNAse A and 0.25 µl RNAse T1 at 37 °C. The sample was desalted with Harvard Apparatus C18 SpinColumns, dried under vacuum, resuspended in 20 µl LC-MS Loading Buffer and 5 µl were used for LC-MS analysis.

2.2.3.5 Sulfite mediated cross-linking

Chemical cross-linking mediated by sulfite was performed with solutions of sodium hydrogen sulfite and freshly dissolved metabisulfite. Experiments were done in triplicates. For complex formation, 1 nmol of rabbit GAPDH was incubated with 1 nmol poly(UC)$_{12}$ in 100 mM HEPES-NaOH pH 7 for 30 minutes on ice. Cross-linking reagent was added to 50 mM end concentration and the reaction was incubated for 3 hours at 37 °C. The proteins were digested with trypsin in the presence of 1 M Gdn-HCl and enrichment of heteroconjugates was performed with Zymo Direct-zol Miniprep Plus kit. RNAse digestion was performed with 1 µl RNAse I and 1 µl RNAse A. In addition, 1 µl of Antarctic phosphatase was added and the sample was supplemented with 1/10 volume Antarctic Phosphatase Reaction Buffer (10x), followed by overnight incubation at 37 °C. Depletion of digested nucleotides was performed with Harvard Apparatus C18 SpinColumns. The enriched heteroconjugates were dried under vacuum and resuspended in 20 µl LC-MS Loading buffer, 5 µl were used for LC-MS analysis.

2.2.4 Isolation of heteroconjugates from *E. coli*

Purification of cross-links from *E. coli* cell lysate was performed by a two-step enrichment workflow based on silica and SAX enrichment. Sample pre-digestion was performed by addition of 200 µg trypsin to 24 ml of cell lysate and overnight incubation at 37 °C.

2.2.4.1 Silica-based enrichment of *E. coli* cross-links

Enrichment of RNA-containing molecules was performed with the Direct-zol RNA Miniprep and RNeasy Maxi kit by extracting respectively 2 and 5 ml of digested *E. coli* cell lysate as described above. Due to the low capacity of the Direct-zol kit, the sample was split among 10 miniprep spin columns. The aqueous eluate of the columns was adjusted to 1 M urea, 12.5 mM HEPES-NaOH pH 7.4 and protein digestion was performed with 20 µg trypsin for 2 hours at 37 °C. Afterwards, the samples were purified again in the same way and the eluate was adjusted to 1 M urea, 12.5 mM HEPES-NaOH pH 7.4, 1 mM $MgCl_2$. RNA digestion was performed overnight at 37 °C with 2 µl Universal Nuclease, 2 µl RNAse A, 1 µl RNAse T1 and 6 µl RNAse I. Non-cross-linked RNA fragments were removed by C18 reversed-phase chromatography with Sep-Pak C18 Cartridge, dried under vacuum and resuspended in 20 µl LC-MS Loading buffer, of which 5 µl were injected for LC-MS analysis.

2.2.4.2 SAX-based enrichment of *E. coli* cross-links

Strong anion exchange chromatography was performed with Pierce SAX (Q) spin column Maxi. Centrifugation steps were performed at 500x*g* for 5 minutes. 12 ml of digested *E. coli* lysate were mixed with 7 ml SAX Separation buffer. The sample was loaded onto a spin column, washed twice with 19 ml SAX Separation buffer and eluted in two steps of 5 ml SAX Elution buffer. The eluate was ethanol precipitated and resuspended in 2 ml DNAse I buffer (Zymo Research). DNA digestion was performed with 120 units DNAse I for 30 minutes at room temperature. Next, complete protein digestion was performed with 20 µg trypsin for 2 hour at 37 °C. To deplete the resulting linear peptides, the sample was mixed with 17 ml SAX Separation buffer and subjected to the same purification and precipitation procedure. The pellet was resuspended in 2 ml 1 M Urea, 5 mM Tris-HCl pH 7.9, 1 mM

$MgCl_2$ and RNA was digested overnight with 2 µl Universl Nuclease, 2 µl RNAse A, 1 µl RNAse T1 and 6 µl RNAse I at 37 °C. Non-cross-linked RNA fragments were depleted with Sep-Pak C18 Cartridge, dried under vacuum, resuspended in 25 µl LC-MS Loading Buffer and 5 µl were used for subsequent analysis.

2.2.5 Isolation of heteroconjugates from HeLa cytoplasm

2.2.5.1 C18/TiO_2 enrichment

The RNA in 1 ml HeLa cytoplasmic extract was digested by addition of 2 µl RNase I, 2 µl RNAse A and 2 µl RNAse T1 and incubation for 2 hours at 37 °C. The sample was supplemented with 93 µl 7 M GdnHCl and digested with additional 2 µl RNAse I, 2 µl RNAse A and 2 µl RNAse T1 for 1 hour at 37 °C. Tris-HCl pH 7.9 was added to 20 mM final concentration, followed by trypsin digestion (~1:200 w/w). Digested nucleotides were removed with C18 reversed-phase chromatography (SepPak C18 Cartlidge). Elution was performed with 80% (v/v) ACN, 5% (v/v) TFA, 5% (v/v) Glycerol. TiO_2 enrichments was carried out by 10 minutes incubation on a rotating wheel with TiO_2 beads (~10:1 w/w beads/peptides). Unspecific interactors were depleted by 3 washes with Buffer A and 3 washes Buffer B. The acidic content was decreased by 2 washes with 8% (v/v) ACN, 0.5% (v/v) TFA and the beads were transferred to a Harvard Apparatus C18 SpinColumns. The cross-linked peptides were eluted onto the C18 material by successive application of 500 mM Na_2HPO_3 and 0.3 M NH_4OH. Desalting was performed with 20 mM TEAB pH 8.5, followed by elution with 80% (v/v) ACN, 20 mM TEAB pH 8.5. To ensure optimal peptide digestion, 0.5 µg trypsin was added to the eluate and the sample was incubated for 2 hour at 37 °C, followed by Harvard Apparatus C18 reversed-phase chromatography and drying under vacuum. The pellet was resuspended in 25 µl LC-MS Loading Buffer and 5 µl were used for LC-MS analysis.

2.2.5.2 Silica-based purification

To 1 ml UV-irradiated HeLa cytoplasmic extract EDTA, HEPES-NaOH pH 7.4 and GdnHCl were added to final concentration of 5 mM, 20 mM and 1 M. Protein digestion was done overnight with trypsin in 1:200 (w/w) ratio at 37 °C. Depletion of peptides was performed with Direct-zol or RNeasy kit and the eluate was further digested with 5 µg of Trypsin for 2 hours at 37 °C, followed by second step of enrichment. The RNA was digested overnight with 4 µl RNAse I, 2 µl RNAse A, 2 µl Universal nuclease, 0.5 µl RNAse T1 and 2 µl Nuclease P1 in the presence of 50 mM Tris-HCl pH 7.9, 1 mM $MgCl_2$ and 1 mM $ZnCl_2$ at

37 °C. The next day, GdnHCl was added to 1 M end concentration and fresh mixture of RNAses was added for additional 1 h digestion at 37 °C. Digested nucleotides were removed by Harvard Apparatus C18 reversed-phase chromatography, the eluate was dried under vacuum, resuspended in 20 µl LC-MS Loading Buffer and 5 µl were used for LC-MS analysis.

2.2.5.3 SAX-based purification

To 1 ml of cytoplasmic extract were added EDTA, HEPES-NaOH pH 7.4 and Urea to end concentration of 5 mM, 20 mM and 1 M. Protein digestion was carried out overnight with trypsin in 1:200 (w/w) ratio at 37 °C. One volume of SAX Separation buffer was mixed with the sample and loaded onto Pierce SAX (Q) mini spin column. Linear peptides were removed by 2 washing steps with SAX Separation buffer, followed by two elution steps with SAX Elution buffer. The pH of the eluate was raised with HEPES-NaOH pH 7.4 and the DNA was digested with DNAse I (180 units) in the presence of 2 mM $MgCl_2$ and 0.5 mM $CaCl_2$ for 15 minutes at room temperature. Urea was added to 1 M end concentration and complete protein digestion was performed with 5 µg of trypsin for 2 hours at 37 °C. The generated linear peptides were depleted with a second purification step and the eluate was diluted to lower the salt concentration of 200 mM with RNAse-free water. RNA digestion was performed overnight in the presence of 50 mM Tris-HCl pH 7.9, 1 mM $MgCl_2$ and 1 mM $ZnCl_2$ with 4 µl RNAse I, 2 µl RNAse A, 2 µl Universal nuclease, 0.5 µl RNAse T1 and 2 µl Nuclease P1 at 37 °C. The next day, GdnHCl was added to 1 M end concentration and fresh mixture of RNAses were added for additional 1 hour digestion at 37 °C. Digested nucleotides were removed by Harvard Apparatus C18 reversed-phase chromatography, the eluate was dried under vacuum, resuspended in 20 µl LC-MS Loading Buffer and 5 µl were used for LC-MS analysis.

2.2.5.4 Sulfite-mediated cross-linking of HeLa cytoplasmic extract

To 1 ml of cytoplasmic extract were added 25 µl 2 M freshly dissolved metabisulfite, followed by incubation at 37 °C for 5 hours. The sample was precipitated with isopropanol and resuspended in 167 µl 6 M GdnHCl. To promote desulfonation of the nucleobases, 5 µl 1 M TEAB pH 8.5 were added, followed by 30 minutes incubation at 37 °C. The sample was diluted to 1 M GdnHCl and supplemented with EDTA and HEPES-NaOH pH 7.4 to 5 mM and 50 mM end concentration. Protein digestion was performed overnight with 33.75 µg trypsin in 1:400 (w/w) ratio at 37 °C. Enrichment of cross-links was performed with two step workflow based on the Direct-zol kit as described above.

2.2.6 Synthesis of peptide-RNA heteroconjugate standard

Standard peptide-oligonucleotide conjugates were synthetized by copper-catalyzed click reaction. Synthetic RNA oligonucleotide 5’-UAGACAU*UGCAGUCACAG-3’ that contained modified 5-ethynyl-2’-deoxyuridine base was mixed with synthetic peptide ALYTFAEGF*K containing 4-azidophenylalanine in molar ratio 1:5. The reaction mixture was incubated at 37 °C for 3 hours in the presence of 7.5 mM copper(I) bromide as catalyzer and 15mM tris(benzyltriazolylmethyl)amine as stabilization agent. The reaction product was monitored by SAX column (Mono Q 5/50) equipped on ÄKTAmicro and direct infusion mass spectrometric analysis in LTQ XL. Digestion of the heteroconjugate standard was performed with 1 µl RNAse T1 for 2 hours at 52 °C. The digest mixture was desalted by C18 Reversed-phase chromatography on self-assembled stage tips at pH 8.5. Washing was achieved with 100 mM TEAB pH 8.5 and elution with 100 mM TEAB pH 8.5, 80% (v/v) ACN.

2.2.7 LC-MS/MS in negative mode

2.2.7.1 Nano-liquid chromatography conditions

Chromatographic separation of heteroconjugates in negative mode was performed with ion pairing buffers containing 8.15 mM triethylamine adjusted with 200 mM hexafluoroisopropanol to pH 8.6. Gradient separation was formed by applying 5 to 44% mobile phase containing 80% (v/v) ACN for 13 minutes. Flowrate was set to 300 nl/min and column temperature was kept at 50 °C.

2.2.7.2 MS acquisition in negative mode

MS acquisition was achieved with Fusion Tribrid mass spectrometer. Ionization was performed with equipped distal coated silica emitter (i.d. 30 µm). Spray voltage was set to 1600 V and ion transfer tube temperature was kept at 275 °C. Orbitrap was used as the mass analyzer for survey scans. Precursor acquisition was achieved with resolution of 120,000 and AGC target was set to $2x10^5$. Charge states 4 through 12 were selected for fragmentation with 1.6 *m/z* quadrupole isolation window. Dynamic exclusion was set to 7 s. Fragmentation was performed with either beam-type collision induced dissociation (NCE

15%, 20%, 25% and 30%) or with ion trap-based collision induced dissociation (NCE 20%, 30% and 40%). The fragment spectra were measured in the Orbitrap mass analyzer with AGC target of 5x10^4 and resolution of 30,000. Maximum injection time for the product scan was set to 120 ms.

2.2.8 LC-MS/MS in positive mode

2.2.8.1 Nano-liquid chromatography conditions

Chromatographic separation prior mass spectrometric acquisition was performed using the Ultimate 3000 UHPLC. Peptides were first concentrated onto a trap column (Thermo PepMap 5 µm 100 Å C18 300 µm x 5mm or packed in-house ReproSil-Pur 1.9 µm 120 Å C18-AQ 100 µm x 30 mm). Separation was achieved with a linear gradient formed with mobile phase A (0.1% v/v FA) and mobile phase B (80% v/v ACN, 0.08% v/v FA) on an analytical column packed in-house (ReproSil-Pur 1.9 µm 120 Å C18-AQ 75 µm x 300 mm) with a constant flow rate of 300 nl/min. Gradient details for the presented samples in this study are shown in Table 2.2.

Table 2.2 Nano-liquid chromatography

Experiment	Gradient (%B)	Gradient duration	Column temperature
mtRNAp/TEFM complex	8-45%	43 min	50 °C
GAPDH/Hsh49 (polyribonucleotides)	8-46%	43 min	50 °C
GAPDH (mononucleotides)	8-42%	43 min	50 °C
HeLa nuclear extract SAX	5-46%	43 min	50 °C
GAPDH (*E. coli* RNA)	5-42%	43 min	50 °C
E. coli cells (I, II, III)	10-50%	165 min	50 °C
E. coli cells (IV-IX)	5-42%	163 min	50 °C
HeLa cytoplasm (I)	5-42%	163 min	50 °C
HeLa cytoplasm (II-VI)	5-42%	165 min	50 °C
GAPDH (poly(UC))	8-42%	43 min	50 °C

2.2.8.2 ESI-MS/MS analysis

Data acquisition of samples in positive mode was performed with the Orbitrap mass analyzer for both survey and product scans. The nano-LC system was directly coupled to electrospray source with 30 µm (i.d.) stainless steel emitter and eluting analytes were ionized by application of 2300-2400 V in the source. The temperature of the ion transfer tube was kept at 275 °C. The acquisition was performed in data dependent manner for analytes with charge states from 2 to 8. Fragment spectra were generated with beam-type collision-induced dissociation and recorded with high resolution (15,000 – 30,000) starting from 110 *m/z*. Detailed parameters of the mass spectrometric analysis are shown in Table 2.3.

Table 2.3 ESI-MS/MS acquisition in positive mode

Mass spectrometric scan parameters of the experiments presented in this study. AGC – Automatic gain control; IT – Injection time; NCE - Normalized collision energy.

Experiment	Instrument	TopN/ TopSpeed	Precursor scan (MS1) Resolution / AGC target / Range	Fragment scan (MS2) Resolution / AGC target / IT / NCE	Isolation window	Dynamic exclusion
mtRNAp/TEFM complex	Q Exactive HF	Top 30	60 000 / 10^6 / 350-1600 *m/z*	15 000 / 10^5 / 100 ms / 30	1.6 *m/z*	25s
GAPDH/Hsh49 (polyribonucleotides)	Fusion Tribrid	3s TopS	60 000 / 10^6 / 380-1580 *m/z*	15 000 / 10^5 / 128 ms /30	1.4 *m/z*	10s
GAPDH (mononucleotides)	Fusion Tribrid	3s TopS	120 000 / 10^6 / 350-1500 *m/z*	30 000 / 10^5 / 128 ms /30	1.6 *m/z*	10s
HeLa nuclear extract (SAX)	Fusion Lumos Tribrid	3s TopS	60 000 / 5x10^5 / 380-1580 *m/z*	15 000 / 5x10^4 / 128 ms /30	1.4 *m/z*	20s
GAPDH (*E. coli* RNA)	Fusion Tribrid	3s TopS	120 000 / 10^6 / 360-1500 *m/z*	30 000 / 10^5 / 250 ms /30	1.6 *m/z*	9s
E. coli cells (I, II, III)	Fusion Lumos Tribrid	3s TopS	60 000 / 5x10^5 / 380-1580 *m/z*	15 000 / 5x10^4 / 250 ms / 28	1.2 *m/z*	20s
E. coli cells (IV-IX)	Fusion Tribrid	3s TopS	120 000 / 10^6 / 360-1500 *m/z*	30 000 / 10^5 / 250 ms / 30	1.6 *m/z*	9s
HeLa cytoplasm (I)	Fusion Tribrid	3s TopS	120 000 / 10^6 / 360-1500 *m/z*	30 000 / 10^5 / 250 ms/ 30	1.6 *m/z*	9s
HeLa cytoplasm (II-VI)	Q Exactive HF	Top 15	120 000 / 10^6 / 380-1580 *m/z*	30 000 / 2x10^5 / 250 ms /30	1.4 *m/z*	10s
GAPDH (poly(UC))	Fusion Tribrid	3s TopS	120 000 / 10^6 / 350-1500 *m/z*	15 000 / 10^5 / 256 ms /30	1.6 *m/z*	10s

2.2.9 Data analysis

2.2.9.1 Protein databases

Analysis of reconstituted complexes and individual proteins was done using FASTA files containing the corresponding modified protein sequences, including purification tags and other inserts. Identification of *E. coli* and HeLa proteins was performed with *E. coli* K12 and human Swiss-Prot (UniProtKB) databases that were downloaded on 26.06.19, containing respectively 4,456 and 20,368 protein entries. For the purpose of FDR estimation, the FASTA files were concatenated with the list of common contaminants provided by MaxQuant (245 protein entries).

2.2.9.2 Peptide identification with MaxQuant

The acquisition .raw files were submitted to MaxQuant. Peptides were matched within 6 ppm precursor accuracy in the survey scan and fragments were matched within 20 ppm accuracy to the product scan. Oxidation on methionine and protein N-terminal acetylation were considered as variable modification. For identification of sulfite-induced cross-linking C-NH_3 and C-NH_3-HPO_3 (Suppl. Table 3.4) were added as possible variable modifications on lysines or protein N-terminal. Tryptic peptides up to two miscleavages with minimal length of 5 (GAPDH/poly(UC)) or 6 (HeLa nuclear extract SAX) amino acids were matched with the detected analytes. False discovery rate was set to 1% and estimated with reversed decoy sequence.

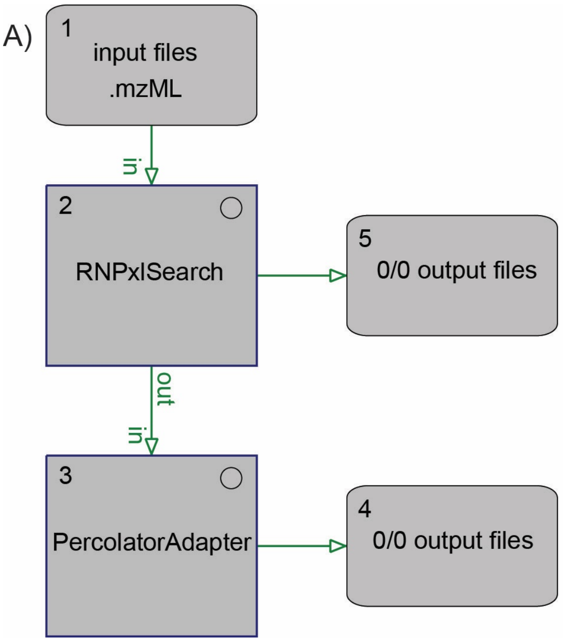
2.2.9.3 Identification of cross-links with the RNP^{xl} workflow

2.2.9.3.1 Data transformation

In order to identify cross-links with the RNP^{xl} workflow, first the acquisition .raw data was converted to compatible open XML-based format (.mzml) using Proteome Discoverer. Spectra with minimal precursor mass of 350 Da and maximum mass of 5000 Da were extracted and peaks were filtered with minimum signal to noise ratio (S/N) of 1.5. The selected signals were recorderd in an .mzml file that was used for precursor variation search with the RNPxlSearch engine.

2.2.9.3.2 Precursor variant search of complex samples

Precursor variant search was performed with the RNPxlSearch node of OpenMS (2.4.0 2019-04-11). Identification of peptides and cross-links from the .mzml file was done with mass tolerance of 6 ppm for the precursor and 20 ppm for fragments in the product scan. Charge states from 2 to 5 were considered, without isotope correction. Oxidation of methionine was included as variable modification for all samples. For workflows involving prolonged incubation at elevated temperature in urea, carbamylation of lysines and peptide N-terminals was included as well. When analyzing TiO_2 enrichment samples, phosphorylation of serine, threonine or tyrosine was added as a variable modification. A maximum of 2 modification per peptide sequence were allowed. Tryptic peptides with up to 3 miscleavages and a size from 5 to 30 amino acids were considered. Heteroconjugate detection was performed by generating precursor adduct variants with up to 3 nucleotide in length. All 4 nucleotides in their monophosphate form were selected as candidates for cross-linking. Plausible neutral loss adducts in the fragment spectrum and precursor modifications are shown in Fig. 2.1. FDR estimation was enabled by setting scoring to “slow” and decoys to “true”. The .idxml file from the RNPxlSearch node was further processed with the Percolator algorithm [71] to improve the rate of identifications, with default parameters.



B)

U	C9H13N2O9P
U-H2O	C9H11N2O8P
U-HPO3	C9H12N2O6
U-H3PO4	C9H10N2O5
U'	C4H4N2O2
U'-H2O	C4H2N2O
C3O	C3O

C	C9H14N3O8P
C-NH3	C9H11N2O8P
C-H2O	C9H12N3O7P
C-HPO3	C9H13N3O5
C-NH3-HPO3	C9H10N2O5
C-H3PO4	C9H11N3O4
C'	C4H5N3O
C'-NH3	C4H2N2O
C'-H2O	C4H3N3

G	C10H14N5O8P
G-H2O	C10H12N5O7P
G-NH3	C10H11N4O8P
G-HPO3	C10H13N5O5
G-NH3-HPO3	C10H10N4O5
G-H3PO4	C10H11N5O4
G'	C5H5N5O
G'-H2O	C5H3N5

A	C10H14N5O7P
A-H2O	C10H12N5O6P
A-NH3	C10H11N4O7P
A-HPO3	C10H13N5O4
A-H3PO4	C10H11N5O3
A-NH3-HPO3	C10H10N5O4
A'	C5H5N5
A'-NH3	C5H2N4

C)

U (-H2O, -H2O-HPO3, -HPO3)
G (-H2O, -H2O-HPO3, -HPO3, -NH3, -NH3-HPO3)
C (-H2O, -H2O-HPO3, -HPO3, -NH3, -NH3-HPO3)
A (-H2O, -H2O-HPO3, -HPO3, -NH3, -NH3-HPO3)

Figure 2.1 Overview of data analysis workflow for complex samples

A) Schematic representation of the precursor variant search workflow B) Possible neutral losses considered in the fragment spectrum C) Possible precursor modifications

2.2.9.3.3 Precursor variant search of isolated complexes and individual proteins

Detection of heteroconjugates was done essentially as described in the previous section, with the following modifications. OpenMS (2.4.0 2018-05-31) workflows were run in the KNIME Analytics Platform (version 3.4.2). Miscleavages were limited to 2 and no maximum size of the peptide was considered. In the experiments with homopolyribonucleotides and individual nucleotide monophosphates, cross-linking was limited to the corresponding type of nucleotide. The top 3 hits per spectrum were exported in the resulting .idxml file. For samples with heavy isotope labeled nucleotides, the lists of neutral fragment losses were modified accordingly and both heavy and light (e.g. -$^{15}NH_3$/-$^{14}NH_3$) losses were considered at the precursor and fragment level. Loss of 2H searches were executed separately, by modifying the respective fragment neutral losses and precursor adducts. DNA searches were considered analogously to RNA with the corresponding compositional changes for the deoxyribose moiety and thymine.

2.2.9.3.4 Manual validation of fragment spectra

Results in the .idxml format were loaded in the TOPPView tool of OpenMS. Decision on the individual quality of inspected spectra was based on a multitude of factors including: i) consistent beam-type fragmentation information, such as prominent y-ion series and strong a2/b2-ion pairs; ii) coherent intensity trends of peptide sequencing ions (e.g. proline effect), immonium ions and other effects of the amino acid side chains [72]; iii) presence of expected nucleotide marker ions; iv) overall matched intensity of the spectrum and number of unmatched signals; v) regular isotopic envelope profile; vi) expected tendencies of the observed neutral loss adducts; vii) annotation of internal fragments and other uncommon elements. Individual spectra were exported in vector image format (.svg) for generation of figures and the identification information was exported in tubular form (.csv) for filtering and supplementary table formation.

2.2.9.3.5 Automated annotation with custom python script

Initially, the automated annotation was performed with a custom python script based on the pyteomics package [73]. Information about the peptide sequence, matched adducts and their mass spectrometric parameters was extracted from RNP^{xl} identification information. Lists of possible shifted series and other ions were generated and matched to the spectral information of the .mzml files (further details described in section 3.1.6).

2.2.9.4 Data filtering and visualization

Simple processing of the table files (e.g. removal of contaminants, q-value filtering) was performed with Excel or R. Bar charts were generated with the ggplot2 R package [74]. Molecular graphics and structure analysis were done with UCSF Chimera [75]. Generation of spectral, workflow and structure figures was done with Illustrator.

2.2.9.5 Gene ontology analysis

Gene ontology enrichment analysis was performed with the STRING database. Decoy hits and contaminants were removed. Results were filtered to include unique proteins at 1% spectral FDR. Hits deriving form shared peptides that lead to more than one protein accession were reduced to the first accession of the list to avoid artificial inflation of the group count.

3. Results

3.1 Automated annotation of protein-RNA cross-link spectra

UV-induced cross-linking mass spectrometry is a powerful tool that can successfully identify RNA-binding proteins and determine the exact amino acid of the protein involved in the interaction. The established biochemical enrichment strategies rely on substantial depletion of the non-cross-linked species that are present in excess, allowing the detection of the extremely low abundant peptide-RNA heteroconjugate molecules. Similarly, the established MS data processing workflow depends on filtering out all spectra that can match to linear peptides or signals that are present in the non-irradiated control, allowing the rest of the spectra to be matched to protein-RNA candidates, based on the precursor mass and fragments series matching the corresponding linear peptide. Successively, those candidates need to be extensively manually validated and annotated, before the cross-link spectrum can be reported as a reliable hit (Suppl. Fig. 3.1).

With the improvement of sensitivity and speed of mass spectrometers, every single experimental file could harbor thousands of cross-link candidates that require manual validation. Analysis of so many spectra could take weeks for annotation and verification. This presents the major time and effort-consuming step of the identification workflow. To address this problem, a strategy for automated annotation was outlined that mimics the manual annotations a human expert performs. For this purpose, the different elements and characteristics of cross-links in conditions of collision-induced fragmentation were determined and are presented in more detail below.

Ions observed in fragment protein-RNA cross-link spectra fall into 3 categories: i) RNA fragment ions that contain only fragments/information from the cross-linked nucleic acid ii) peptide fragments that contain only parts of the involved peptide iii) mass “shifted” or “adduct” fragment ions that entail both RNA and peptide derived parts in their structure, named after the mass shifts observed on peptide fragments by the RNA adducts (Fig. 3.1).

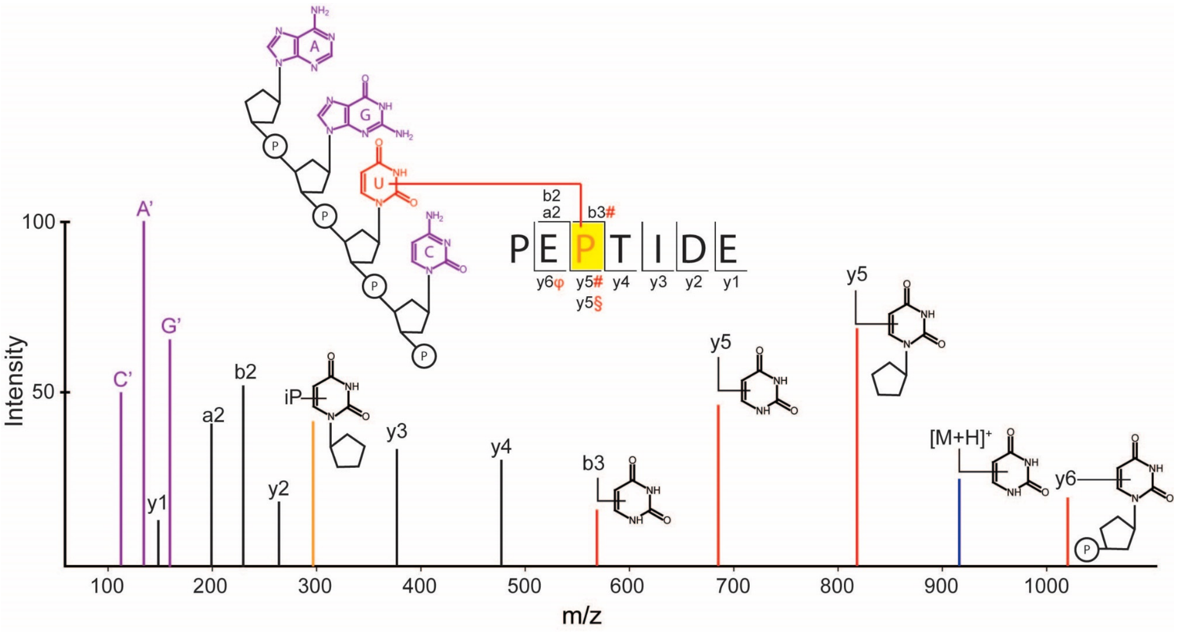


Figure 3.1 Schematic representation of a peptide-RNA heteroconjugate fragment spectrum

Illustrative annotated cross-link spectrum of peptide “PEPTIDE” and RNA oligonucleotide “AGUC”. Characteristic ions generated by collision-induced fragmentation are indicated and annotated: Nucleobase ions (purple) originating from the fragmentation of the non-cross-linked bases are indicated with the letter of the base and apostrophe. Standard peptide derived a-, b- and y-ions (black) are indicated with the respective letter and index of the ion position (annotations for single charges are omitted). Shifted ions (red) with a neutral loss adduct of the cross-linked uridine nucleotide are visualized as conjugated with a schematic representation of the RNA adduct. Peptide precursors (blue) are indicated as M and immonium ion (orange) are marked with a prefix “i” in front of the respective amino acid letter code.

3.1.1 RNA marker ions

At the energy levels usually used in beam-type fragmentation to obtain a meaningful peptide sequencing (~30 NCE), the more labile RNA nucleotides that are not covalently bound to the peptide are almost completely shattered, mainly due to the fragmentation of the N-glycosidic bonds. Therefore, RNA fragment ions of the non-cross-linked nucleotides are mostly represented by strong nucleobases marker ions and more rarely as low intensity intact nucleotide/nucleoside ions. The cross-linked nucleotide in contrast gives low intensity or no base marker ions, partially because it would require the occurrence of two fragmentation events to free the base ion, one at the N-glycosidic bond and one at the UV generated peptide-RNA covalent bond. The empirically observed tendency of the nucleobases ion intensities in cross-link spectra (A’>G’>C’>>U’) is in accordance with their reported proton affinities [76]. Adenine, guanine and cytosine give rise to intense marker ions and uracil base ions are often absent or of low intensity (Fig. 3.2). The high intensity of adenine, guanine and cytosine nucleobase marker ions can lead to suppression of the other signals in the spectrum, an effect that becomes more prominent with the increasing of the RNA adduct length (Suppl. Fig. 3.2). The presence of marker ions does not explicitly indicate a cross-link hit, as they are often observed also in non-cross-link spectra, when significant amount of RNA is present in the sample. Nevertheless, the consistent presence of the respective marker ions and their effects is important quality control used for validation. Unexplainable absence of expected base ions (e.g. A’, G’, C’) is a strong indicator of a false positive. Marker ions can also provide information for the correct assignment of the right RNA adduct. For example - the difference of an oxygen atom between adenine and guanine results in the same mass as the commonly observed methionine oxidation modification. In such cases, marker ions are often the only way for accurately assign the correct RNA adduct.

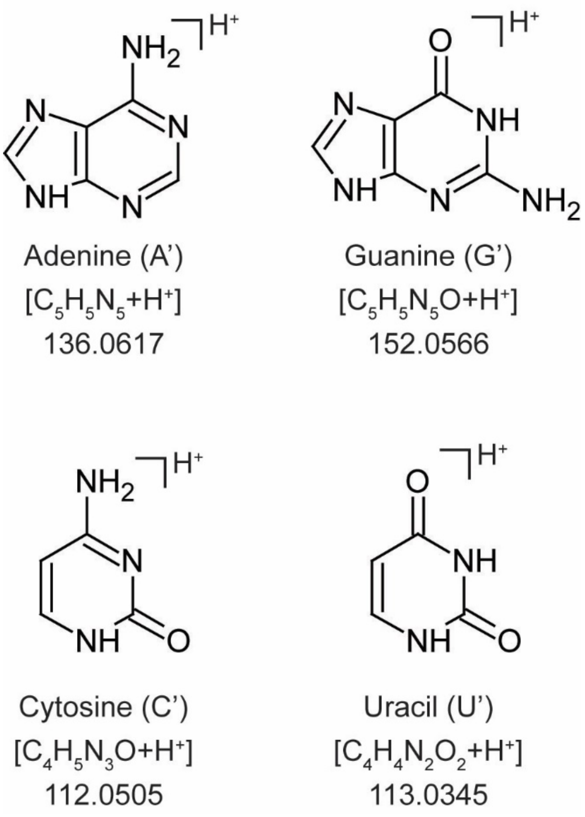


Figure 3.2 RNA base marker ions

Commonly observed marker ions of the 4 canonical RNA bases, derived from the non-cross-linked nucleotides by cleavage of the N-glycosidic bond between the nucleobase and ribose.

3.1.2 Peptide-derived ions

Ions from the peptide moiety follow the usual beam-type fragmentation, creating prominent y-ion series and less prominent a- and b-ion series. They are mostly created from the non-cross-linked C-terminus for y-ions and N-terminus for the a-, b-series, but can also encompass the cross-link position in events of total neutral loss of the RNA adduct. Therefore, localization of the cross-link position based on the non-shifted peptide ions is often inaccurate. Additionally, the generated ions are undistinguishable from linear peptide fragments and identification solely based on them can easily lead to a false positive.

3.1.3 Shifted ion series

Shifted ion series are formed from the a-, b- and y-ion peptide series shifted with (neutral loss) adducts of the cross-linked nucleotide. They provide the strongest (cumulative) evidence of the validity of the heteroconjugate spectrum and the localization of the cross-linked amino acid. At the same time, due to the large number of possible neutral losses, the shifted ions are also the major cause of the combinatorial complexity observed in fragment cross-link spectra. Generally, the shifted ions with adducts resulting from the fragmentation of the N-glycosidic bond of the linked nucleotide are the most commonly observed species.

3.1.4 Shifted (intact peptide) precursor ions

In standard proteomics research, the identification of neutral losses on the precursor does not provide substantial information for the identity of the hit and the information of such ions are not considered when scoring the spectrum. However, characteristic neutral losses (e.g. N-glycosidic bond fragmentation) can be a strong evidence for the validity of a cross-link hit, even though it provides no localization information where that cross-link occurred in the sequence of the peptide.

3.1.5 Shifted immonium ions

Amino acids that create prominent immonium marker ions (e.g. iY, iF, iH) or other amino acid fragments (e.g. lysine fragments K' 129.10/84.08), when cross-linked, often produce shifted immonium ions with RNA adducts. The additional mass added by the RNA adducts sometimes also stabilizes non-traditionally intensive immonium ions (e.g. iC) that can be detected in the spectrum. Whenever present, shifted immonium ions provide strong and precise localization information as well as important additional verification information.

3.1.6 Automated annotation

In order to automate the calculations of the different elements, all known uridine adducts (Table 3.1) and all amino acid immonium/fragment ions (Suppl. Table 3.1) were noted. At first, simple calculator scripts were created in R and used to generate a-, b- and y-ion shifted series and all possible shifted immonium ions. Later, the functionality was expanded into a python-based script that allowed identification of the shifted ion series, immonium ions, precursor shifted ions and their characteristics (e.g. mass accuracy, relative intensity) from a spectrum (Fig.3.3).

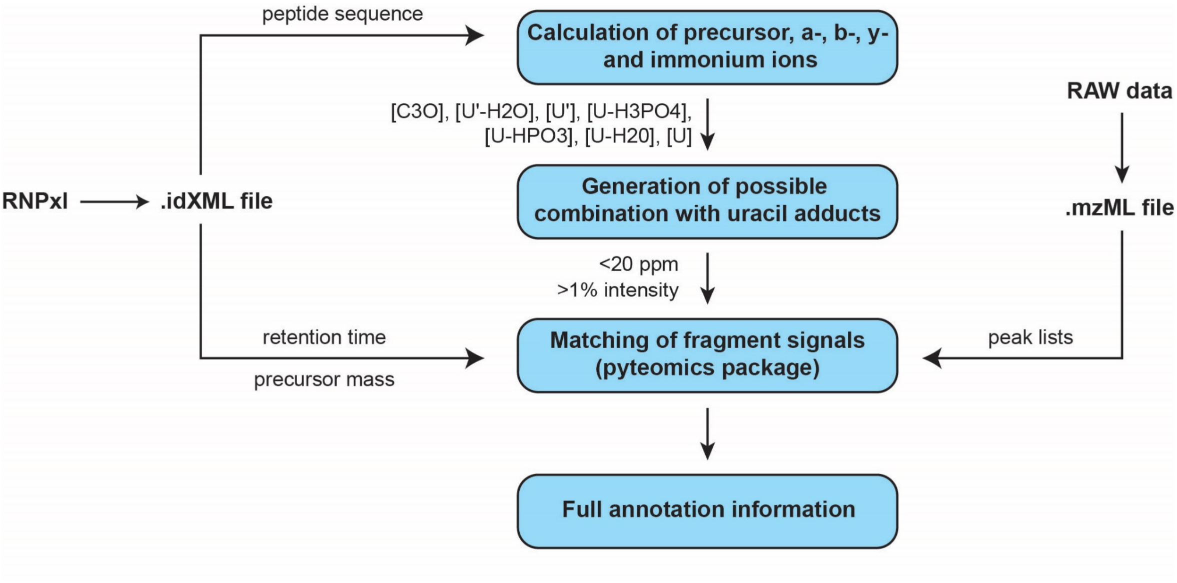

peptide sequence
Calculation of precursor, a-, b-, y- and immonium ions
RAW data
[C3O], [U'-H2O], [U'], [U-H3PO4], [U-HPO3], [U-H20], [U]
RNPxl
.idXML file
Generation of possible combination with uracil adducts
.mzML file
<20 ppm
>1% intensity
retention time
precursor mass
Matching of fragment signals (pyteomics package)
peak lists
Full annotation information

Figure 3.3 Schematic representation of automated annotation of peptide-RNA cross-links

The identification information contained in the .idXML results file from the RNP^{xl} workflow are used to extract the peptide sequence information, retention time and precursor mass of the cross-link candidates. Common collision-induced dissociation products are calculated and their combinations with neutral loss products of uridine monophosphate are generated. The possible fragment adducts are matched to the corresponding MS/MS spectrum and all annotation fitting certain criteria are reported in tubular form or spectral view in a dedicated graphical user interface.

In addition, further quality information, such as the presence of expected RNA marker ions was supplemented. With the functionality of the pyteomics package [73], the entire spectrum files could be iterated with the identification candidates reported from the RNP^{xl} tool. Functional annotation of ion species is reported in tabular text form for the entire file that allows global overview or can be loaded in a simple graphical user interface and recalled for the view of individual spectrum (Suppl. Fig. 3.3 A, B). Thus, this approach allows simultaneous employment with the TOPPAS tool of OpenMS and greatly reduces the time required for manual verification of cross-link candidate spectra. The script was expanded with additional functionality that enables the iteration of all spectra with custom submitted masses, useful for mining of entire datasets when investigating new possible shifts (e.g. putative DNA cross-links adducts). Logical restrictions on the uridine fragmentation were set up (Suppl. Fig. 3.3 C) to limit the report of nonsensical shifted ions.

Table 3.1 Established uridine neutral loss adducts commonly observed in protein-RNA cross-link spectra

Abbreviation	Formula	Monoisotopic mass
U	C9H13N2O9P	324.0359
U-H2O	C9H11N2O8P	306.0253
U-HPO3	C9H12N2O6	244.0695
U-H3PO4	C9H10N2O5	226.059
U'	C4H4N2O2	112.0273
U'-H2O	C4H2N2O	94.01671
C3O	C3O	51.99492

This strategy greatly increased the speed of manual validation and was incorporated in the successive versions of the RNP^{xl} tool for OpenMS [77] and Proteome Discoverer™ by Johannes Veit and Dr. Timo Sachsenberg (Applied Bioinformatics Group, University of Tübingen) [55]. In consequence, complete automated annotation and visualization of shifted and non-shifted species could be performed in successive versions of the TOPPAS tool in the OpenMS suite and Proteome Discoverer (Suppl. Fig. 3.4).

3.1.7 Software improvements

In addition to the visual inspection improvements, Dr. Timo Sachsenberg performed considerable improvements on the RNP^{xl} workflow. A dedicated search engine (RNPxlSearch) was introduced to replace the OMSSA search engine. RNPxlSearch provides significant speed improvements (up to two orders of magnitude) and generation of automated cross-link localization score according to the observed shifted ions [78]. In consequence to these improvements and the implementation of the automated annotation, the time and effort needed for analysis of protein-RNA cross-links was reduced dramatically. However, the identification of cross-link spectra is still based only on the precursor mass and the peptide generated fragments, while all other elements of the spectrum are assigned post-identification. Therefore, the search engine has a high chance of misassigning or dismissing hits with prominent shifted ion series or intense RNA marker ions, due to the less extensive peptide fragments in the spectrum. In this way, short peptides and heteroconjugates cross-linked close to the C-terminus are disproportionally penalized in the scoring. Additionally, because the generated score is only partially descriptive of the cross-link spectrum quality, there is no sensible way to generate a false discovery rate estimation and filtering, necessitating manual validation of all spectra.

In order to tackle this problem, a strategy for generation of a combined score taking into account all elements of the cross-link spectrum was designed. For this purpose, 590 manually validated spectra were extracted from experiments of the mtRNAP/TEFM complex (Table 3.2) and data generated in previous years by Dr. Kundan Sharma (Bioanalytical Mass Spectrometry Group, MPIbpc), exemplary for various RNA adducts, sequences, cross-linked amino acids and spectral quality. To generate a combined score, Dr. Timo Sachsenberg created subscores incorporating the information of different spectrum elements and performed training on the extracted exemplary spectra and a reference yeast dataset [39]. Further cycles of manual assessment of the resulting score, retraining and optimization were performed. While the software is still in development, a working version (OpenMS 2.4.0 from 2019-04-11) was used for FDR evaluation and filtering of highly complex datasets in the following chapters.

3.2 Systematic nucleotide cross-link evaluation

In the mass-spectrometric studies performed until now in the Urlaub Research Group, the reported cross-linked nucleotide is almost exclusively uridine. Guanosine was only detected couple of times in low quality and singular spectra that were of unconvincing low quality [39,79]. An interesting observation is that lysine cross-links were often detected with RNA adduct U-H2O (C9H1N2O8P 306.0256 Da), which upon fragmentation generated shifted ions that are adducted with U'-H2O (C4H2N2O 94.067 Da). During literature review, a paper involving transamination reaction between cytosine and primary amines mediated by UV irradiation was noticed, that could be an explanation of the observed phenomenon [80]. Indeed, cytosine with ammonia net loss has the same exact mass and chemical composition as uracil with water net loss (C-NH3 = C9H1N2O8P = U-H2O 306.0256 Da).

To test whether the observed lysine cross-links that were presumed to be formed with uracil, are actually a transamination product of cytosine and to verify the low quality guanosine cross-links, a set of experiments for systematic cross-linking analysis of the 4 canonical RNA nucleobases was initiated. Two model RNA binding proteins – spliceosomal factor Hsh49 (part of the SF3b complex) and glyceraldehyde-3-phoshphate dehydrogenase (GAPDH) were utilized. Both proteins were previously demonstrated to efficiently bind RNA in our laboratory (data not shown). Hsh49 contains two RRM domains and GAPDH has a Rossmann fold domain. The model proteins were cross-linked with equimolar amount of poly(U)$_{25}$, poly(C)$_{25}$, poly(G)$_{25}$ and poly(A) $_{25}$ RNA oligonucleotides labeled on the 3'-end with biotin. The resulting products were analyzed by North-western blot and subjected to

C18/TiO$_2$ enrichments workflow, followed by LC-MS analysis (Suppl. Table 3.3). Evidence of higher order products corresponding to the molecular weight of Hsh49 + RNA (~27 kDa + ~8 kDa) and 2xHsh49+RNA can be detected with all four homopolyribonucleotides. Although small amounts of Hsh49 seem to undergo protein-protein cross-linking to itself during UV irradiation, the band of 2xHsh49+RNA is probably primarily a result of the length of the synthetic oligonucleotide that can easily accommodate binding of several proteins simultaneously. In the case of the second model protein GAPDH, the formation of distinct cross-link products is less apparent. GAPDH usually forms a homotetramer that is known to very efficiently generate protein-protein cross-link products under UV light [50]. After irradiation, the majority of GAPDH formed aggregates that did not enter the gel, which most likely is responsible for the observed differences between Hsh49 and GAPDH (Fig.3.4 A, B).

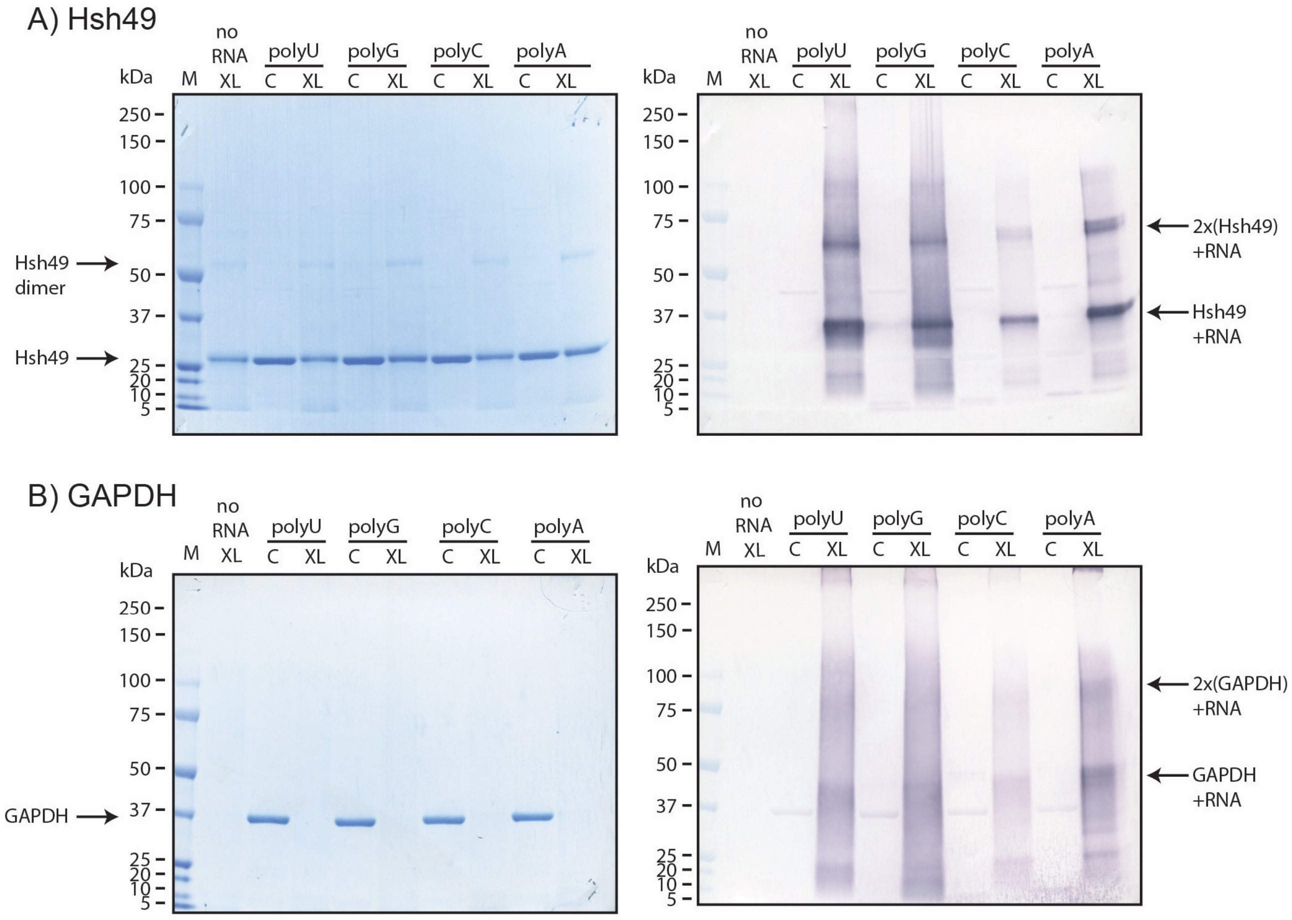


Figure 3.4 Model proteins Hsh49 and GAPDH can form cross-links to the four homopolyribonucleotides

Hsh49 (A) or GAPDH (B) were cross-linked with poly(U)$_{25}$-3'-biotin, poly(G)$_{25}$-3'-biotin, poly(C)$_{25}$-3'-biotin or poly(A)$_{25}$-3'-biotin synthetic oligonucleotides by irradiation at 254 nm. In parallel, a reaction with only protein and no RNA was irradiated under the same conditions. The samples were separated on SDS-PAGE and plotted onto a PVDF membrane. The immobilized proteins were visualized with Coomassie stain (left panel). The membrane was destained, blocked and developed with the Biotin Chromogenic Kit (right panel).

In additional experiments, GAPDH was cross-linked with the four nucleoside monophosphates and their heavy isotope ^{15}N and $^{15}N/^{13}C$ analogs to validate suspected cross-links and fragment shifts. The protein was mixed with great molar excess (160 fold) of free nucleotides to promote contact and irradiated at 254 nm. The large number of unreacted nucleotides was depleted by size exclusion spin columns, so that they would not hinder successive enrichment steps and to avoid the generation of possible non-UV-induced adducts by reactions such as spontaneous glycation [81]. The formed heteroconjugate products were purified by C18/TiO_2 workflow and analyzed by LC-MS (Suppl. Table 3.4).

3.2.1 Uracil cross-links

As expected, a substantial number of cross-link spectra were identified with poly(U) (Suppl. Table 3.3) for both model proteins. All seven commonly observed fragment shifts of uridine could be identified (Fig. 3.3). All fragment adducts could also be validated in the experiments with heavy labeled uridine monophosphate analogs to be of RNA origin.

Additionally, cross-links with only the uracil base (U' C4H4N2O2 112.0273 Da) as RNA precursor adduct could be detected. These, were probably generated during the harsh conditions of the enrichment process or during ionization (Suppl. Fig. 3.5). A significant number of cross-links to U-H2O could be identified in both types of experiments. Several of the spectra identified lysine as the cross-linked amino acid, indicating that at least part of the reported cross-links in the past were indeed formed with uracil.

Further searches to detect not fully additive adducts such as the products described by Varghese [45] were executed. A lesser number of uracil cross-links with -2H (U-2H, C9H11N2O9P, 322.0202 Da) net loss could be identified, as well as few hits leading to U-H2O-2H (C9H9N2O8P 304.0097 Da). The two hydrogen atom deficit can be observed in all collision-induced neutral losses of the precursor adduct in the fragment spectrum. However, the spectral number and precursor intensity of the detected -2H heteroconjugates was considerably lower than the hits corresponding to fully additive and water loss heteroconjugates (Fig. 3.6).

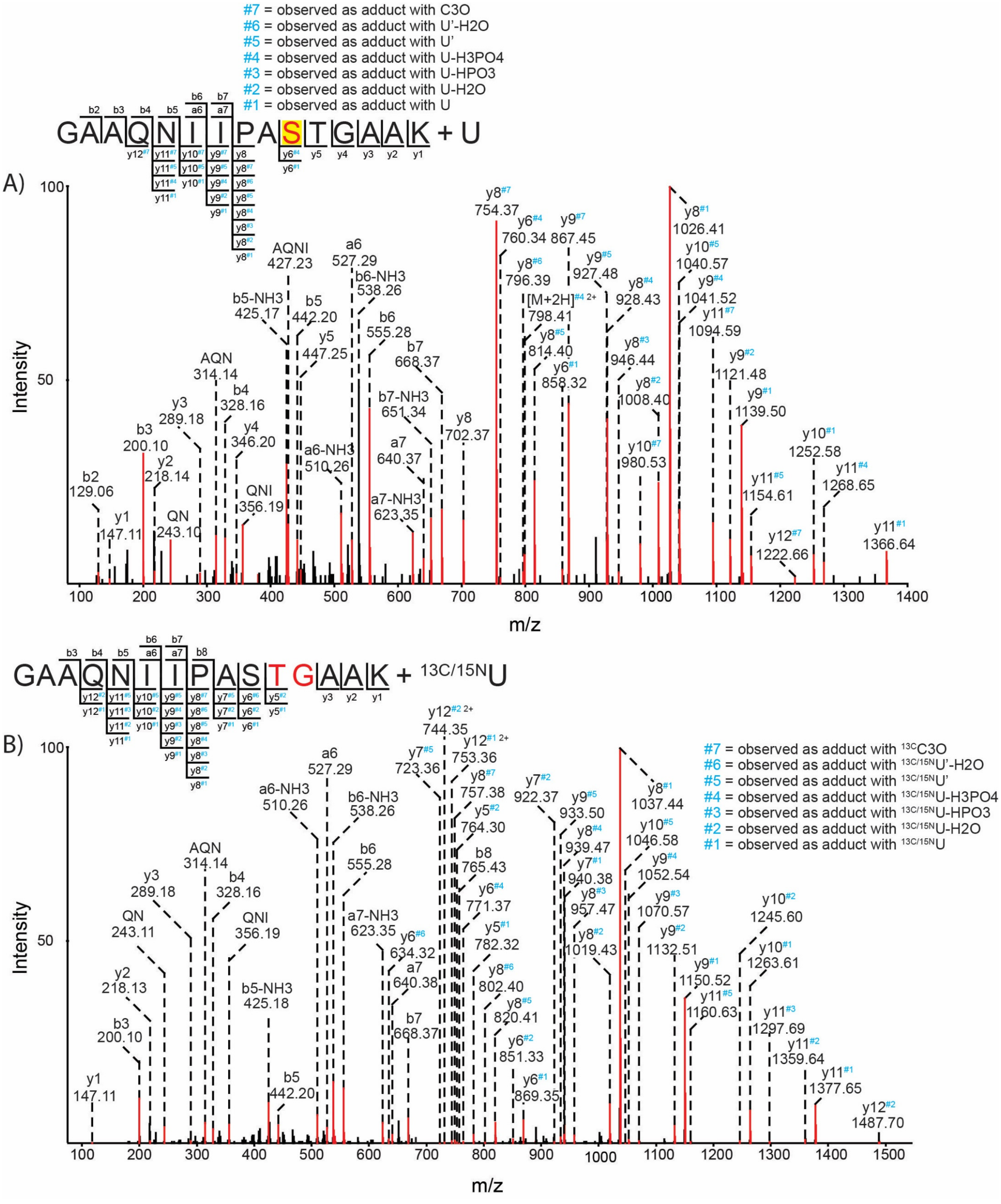


Figure 3.5 Exemplary cross-link spectrum to uracil

A) Fragment spectrum of the peptide $_{199}$GAAQNIIPASTGAAK$_{213}$ with uridine monophosphate generated by irradiation of GAPDH with poly(U). The shifted y-ion series localize the cross-link at $_{208}$S. All seven known uridine neutral losses can be identified in the spectrum. The peptide and sequencing ions are indicated above the spectrum. Identified peaks are highlighted in red and annotated on top. Shifted ions are indicated with a superscript of the respective adduct. Legend of the observed adducts can be found on top of the peptide sequence. B) Fragment spectrum of the peptide $_{199}$GAAQNIIPASTGAAK$_{213}$ with uridine monophosphate, generated by irradiation of GAPDH with heavy isotope labeled (^{13}C/^{15}N) uridine monophosphate. All seven known uridine neutral losses can be observed to be shifted with the expected mass addition by the heavy isotopes, thus confirming the correct assignment as shifted peptide ions.

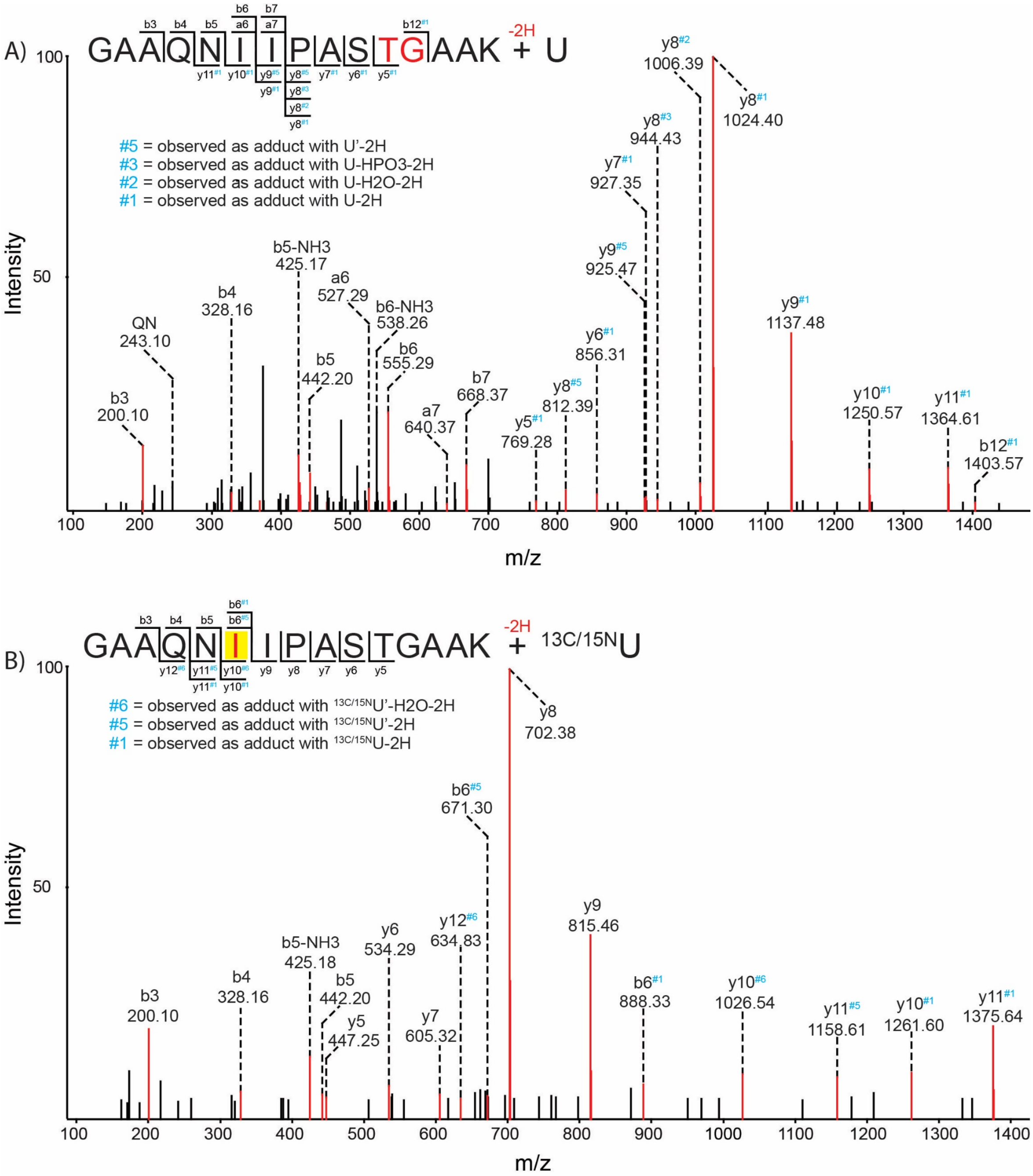


Figure 3.6 New type of uracil adduct detected (-2H net loss)

Fragment spectrum of the peptide $_{199}$GAAQNIIPASTGAAK$_{213}$ with uridine monophosphate, generated by irradiation of GAPDH with uridine monophosphate and its heavy isotope (^{13}C/^{15}N) labeled analog. All detected shifts have a -2H net loss. A) Cross-link spectrum to unlabeled uridine monophosphate. B) Cross-link spectrum to heavy isotope labeled (^{13}C/^{15}N) uridine monophosphate, confirming the identity of the observed shifted fragments.

3.2.2 Cytosine cross-links

Searches for cytosine cross-links identified a substantial number of cross-linked peptides, all found with an RNA adduct corresponding to cytosine with ammonia net loss and mostly cross-linked to lysines (Suppl. Table 3.2, 3.3, Fig. 3.7). These results suggest that many of the reported in the past lysine cross-links to U-H2O were actually formed by cytosine. Experiments with heavy labeled cytidine monophosphates revealed that the ammonia loss occurred with a heavy labeled nitrogen, thus demonstrating that it originates from the 4-NH2 amino group of the nucleobase (Suppl. Fig. 3.6). A single -NH3-2H loss cross-link peptide could also be identified. No difference could be observed between the fragmentation patterns of C-NH3 and U-H2O cross-links. Therefore, when analyzing protein interactions with unlabeled or native RNA, it is not possible to distinguish by mass spectrometry which of the two pyrimidine nucleotides generated the 306.0256 Da adduct and took part in the interaction.

UV irradiation has been reported to promote deamination of cytosine to uracil [82]. Such initial transformation, followed by cross-linking reaction may result in misassignment of the adduct and erroneous conclusions about the originating position in the RNA. To assess the contribution of this effect, in the context of mass spectrometric analysis of cross-links, additional searches for uracil heteroconjugates were performed within the cytosine datasets. Few spectra of fully additive U (C9H13N2O9P 324.0359 Da) cross-link spectra could be identified.

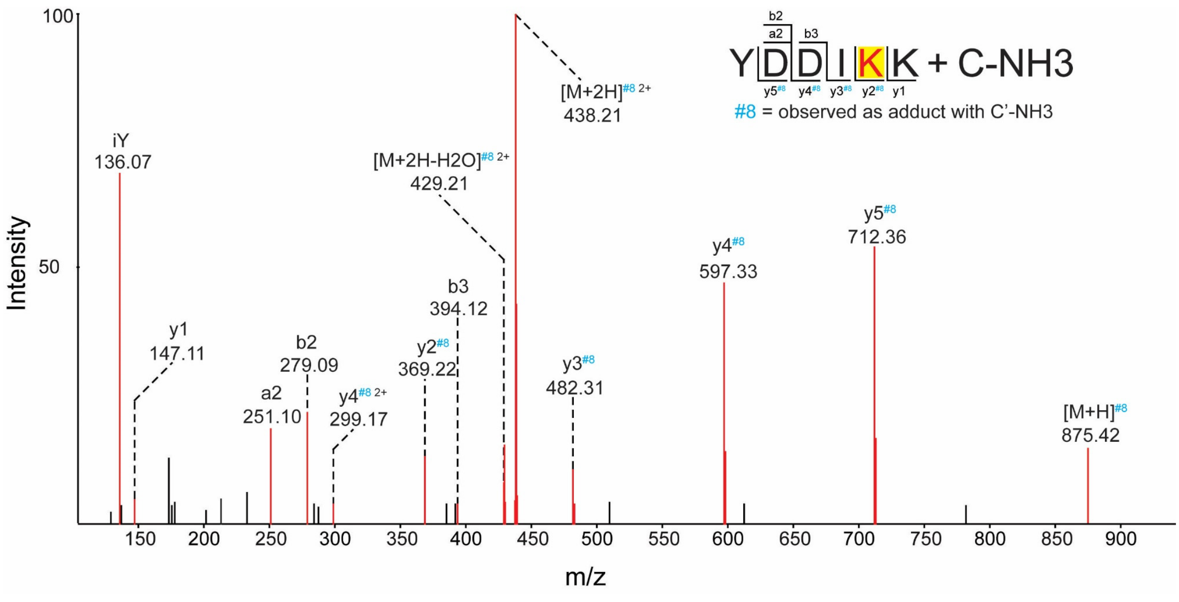


Figure 3.7 Exemplary cross-link spectrum to cytosine

Fragment spectrum of peptide $_{253}$YDDIKK$_{258}$ cross-linked to cytidine monophosphate with NH3 net loss, generated by irradiation of GAPDH with poly(C). The shifted y-ion series confidently identify $_{257}$K as the cross-linked amino acid. Characteristic for cytosine cross-link spectra – only single neutral loss (C'-NH3) is commonly observed, creating high intensity shifts and conclusive localization of the cross-link site.

3.2.3 Guanine cross-links

Several guanine cross-links could be identified, but at considerably lower numbers than the heteroconjugates observed with pyrimidine nucleotides (Suppl. Table 3.3, 3.4, Fig. 3.8). Guanine adducts were observed as fully additive nucleotides (G C10H14N5O8P 363.0580 Da), as well as ammonia (G-NH3 C10H11N4O8P 346.0315) and water loss modifications (G-H2O C10H12N5O7P 345.0474 Da). Additionally, guanine base could be identified as a precursor adduct (G’ C5H5N5O 151.0494 Da), as well as cross-links with -2H net loss (G-2H C10H12N5O8P 361.0423 Da; Suppl. Fig. 3.7).

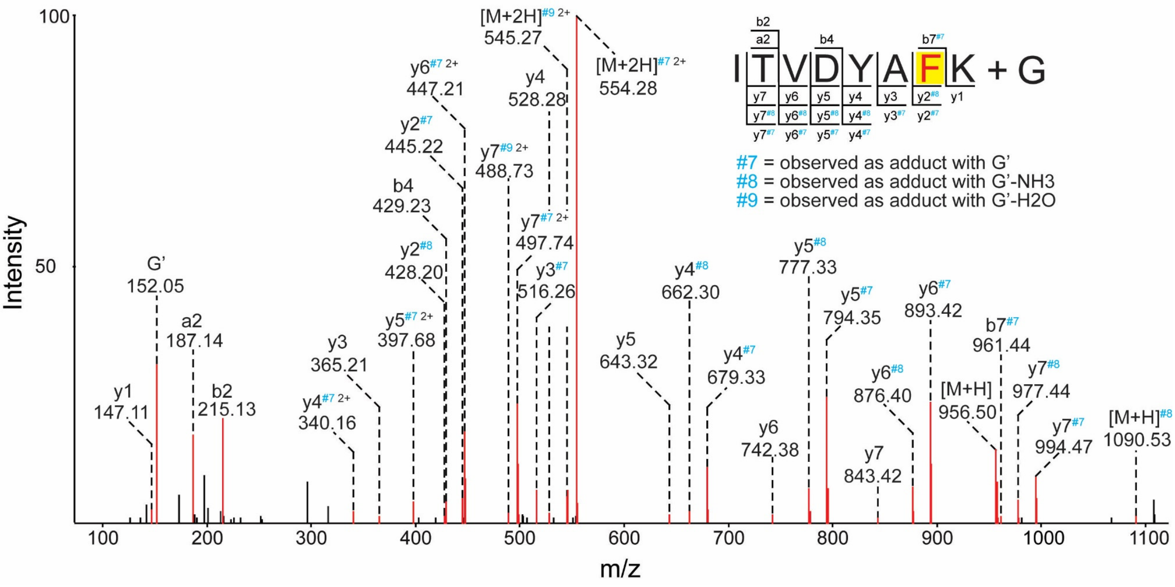


Figure 3.8 Exemplary cross-link spectrum to guanine

Fragment spectrum of peptide $_{178}$ITVDYAFK$_{185}$ with guanidine monophosphate, generated by irradiation of HSH49 with poly(G). The shifted b- and y-ion series localize $_{184}$F as the cross-linked amino acid. Several different neutral losses can be identified, all generated by fragmentation of the N-glycosidic bond. The guanine marker ion (G’ 152.05 *m/z*) is suppressed as the nucleobase is involved in the formation of the protein-RNA covalent bond.

3.2.4 Adenine cross-links

No adenine cross-links could be identified in the poly(A) or single nucleotide analog experiments. Purines are less stable than pyrimidines as they are prone to N-glycosidic cleavage at low pH [83]. Therefore, it is possible that the harsh acidic conditions of the TiO_2 enrichment resulted in loss of adenine cross-linked peptides. In order to address this hypothesis, an alternative experiment was set up. Large quantities of *E. coli* extracted RNA were cross-linked with GAPDH and purified by TRIzol/silica based workflow. The resulting mass spectrometric files were searched against all 4 nucleotides. Few unambiguous cross-link spectra of adenosine monophosphate with ammonia net loss could be identified (Fig. 3.9).

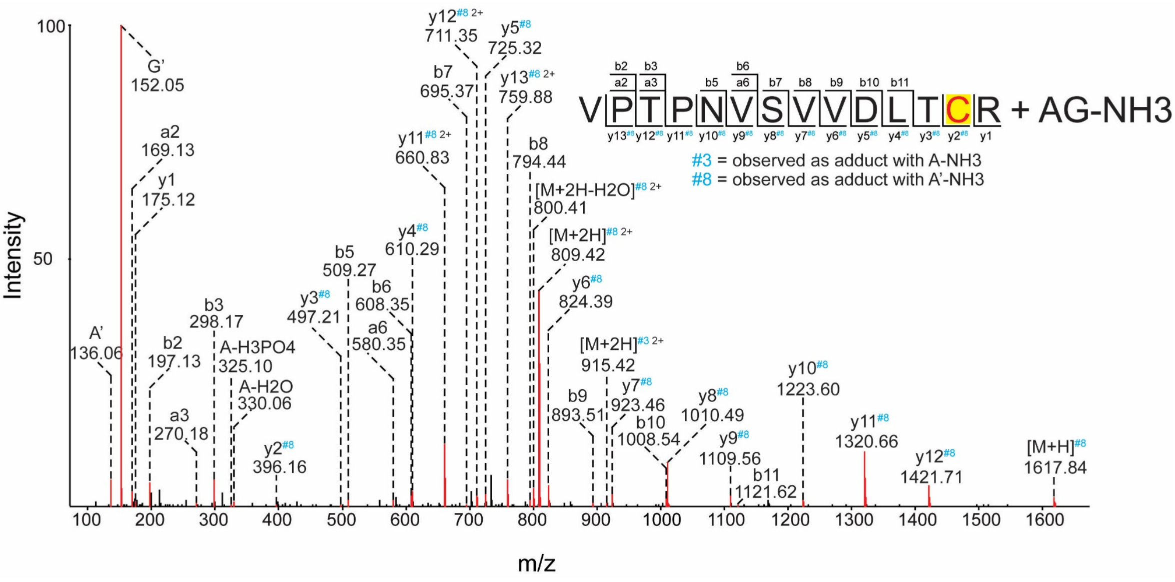


Figure 3.9 Exemplary cross-link spectrum to adenine

Fragment spectrum of peptide $_{233}$VPTPNVSVVDLTCR$_{246}$ with AG-NH3, generated by irradiation of GAPDH with *E. coli* isolated RNA. The shifted y-ion series localize $_{245}$C as the cross-linked amino acid. Analogous to the cytosine cross-links, a single neutral loss (A'-NH3) is predominantly observed.

3.2.5 Pyrimidine bases can cross-link in DNA

During an investigation of the mtRNAP/TEFM elongation complex with RNA/DNA scaffold [84] (sample kindly provided by Hauke Hillen, Department of Molecular Biology, MPIbpc) by UV irradiation and mass spectrometry, high confidence spectra of protein-DNA cross-links were identified (Fig. 3.10, Table 3.2). Both cytosine and thymine could be identified as cross-linked with a covalent bond formed at the pyrimidine base. Cytosine cross-links were formed with lysines through -NH3 net loss and had identical fragmentation behavior to the RNA bases. The thymine heteroconjugate was fully additive. This was the first instance of observed clear spectral evidence of DNA cross-link mediated by nucleobase in our laboratory. Later, Dr. Alexandra Stützer (Bioanalytical Mass Spectrometry Group, MPIbpc), with the use of a chromatin model system, demonstrated that such cross-links can be readily identified by mass-spectrometry after UV irradiation of protein-DNA complexes.

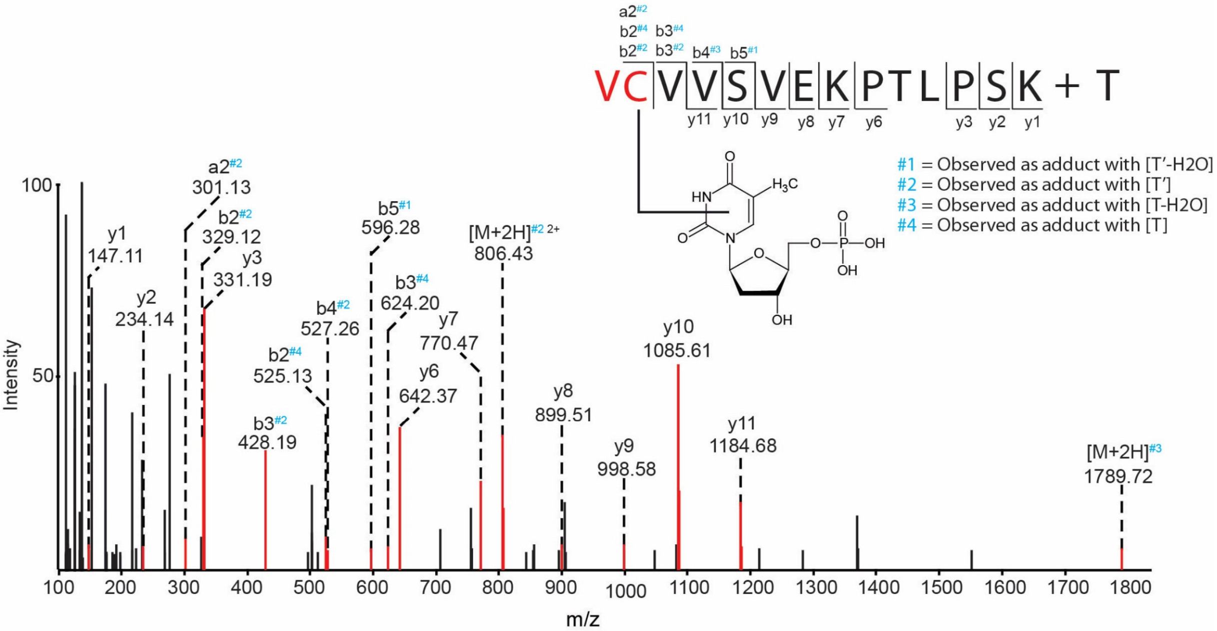


Figure 3.10 Exemplary DNA cross-link spectrum to the nucleobase thymine

Fragment spectrum of peptide $_{412}$VCVVSVEKPTLPSK$_{425}$ with thymine monophosphate, generated by irradiation of mtRNAP with RNA/DNA scaffold. The shifted a-, b-ion series localize $_{412}$VC$_{413}$ as the cross-linked region.

Table 3.2 Protein-RNA/DNA cross-links identified in the human mtRNAP/TEFM complex

Protein (UniProt ID)	**Peptide**	**RNA/DNA adduct(s)**	**Cross-link localization**	**Sample**
mtRNAP (O00411)	$_{202}$LSLDVEQAPSGQHSQAQLSGQQQR$_{225}$	AU	-	mtRNAP/TEFM
	$_{1090}$QIGGGIQSITYTHNGDISR$_{1108}$	GU	-	mtRNAP/TEFM
	$_{412}$VCVVSVEKPTLPSK$_{425}$	U, GU, ACU-HPO3, AGU	$_{413}$C	mtRNAP/TEFM
TEFM (Q96QE5)	$_{43}$ITPNVTFCDENAK$_{55}$	G', U	$_{50}$C	mtRNAP/TEFM
	$_{43}$ITPNVTFCDENAKEPENALDK$_{63}$	U	$_{50}$CDENA$_{54}$	mtRNAP/TEFM
	$_{42}$KITPNVTFCDENAK$_{55}$	U, GU	$_{50}$C	mtRNAP/TEFM
	$_{153}$KLLKPDIER$_{161}$	UC-NH3 or UU-H2O	$_{153}$KL$_{154}$	mtRNAP/TEFM
mtRNAP (O00411)	$_{564}$EQPWPLPVQMELGK$_{577}$	U	$_{573}$MEL$_{575}$	mtRNAP
	$_{1090}$QIGGGIQSITYTHNGDISR$_{1108}$	U, GU	$_{1100}$Y	mtRNAP
	$_{412}$VCVVSVEKPTLPSK$_{425}$	U, CU, GU, AU, AAU, AGU, dT	$_{50}$C	mtRNAP
	$_{602}$LVPVLYHVYSFR$_{613}$	GU	$_{607}$Y	mtRNAP
TEFM (Q96QE5)	$_{43}$ITPNVTFCDENAK$_{55}$	U, GU	$_{50}$C	TEFM
	$_{43}$ITPNVTFCDENAKEPENALDK$_{63}$	U, GU, CU, ACU-HPO3	$_{50}$C	TEFM
	$_{153}$KLLKPDIER$_{161}$	UC-NH3 or UU-H2O, ACU-NH3-HPO3 or AUU-H3PO4, dTC-NH3, dAC-NH3, dATC-NH3	$_{153}$K	TEFM
	$_{42}$KITPNVTFCDENAKEPENALDK$_{63}$	GU	-	TEFM
	$_{154}$LLKPDIER$_{161}$	AUC-NH3-HPO3 or AUU-H3PO4, dCTG-NH3-HPO3, dATC-NH3-HPO3, dTC-NH3	$_{156}$K	TEFM

3.3 Sulfite-mediated cross-linking of proteins to RNA

Similarly to UV irradiation, sulfonation has been described to promote the transamination cross-linking of cytosine to the ε-amino group of lysine [80]. Such a specific reaction would circumvent the complexity difficulties in the analysis of protein-RNA interactions observed with UV light. Therefore, the bisulfite reaction was further investigated for its usefulness for structural studies of protein-RNA interactions by mass spectrometry.

Sulfite-mediated derivatization of cytosine with primary amines is well documented and an efficient chemical reaction (Fig.3.11) [85]. However, the available protocols require high concentrations of the reactants and incubation at low pH and high temperatures that are not suitable for obtaining native-like structural information of protein-RNA complexes. Therefore, a set of experiments was performed to evaluate if the reaction can proceed under physiological conditions and at lower concentrations with adequate efficiency.

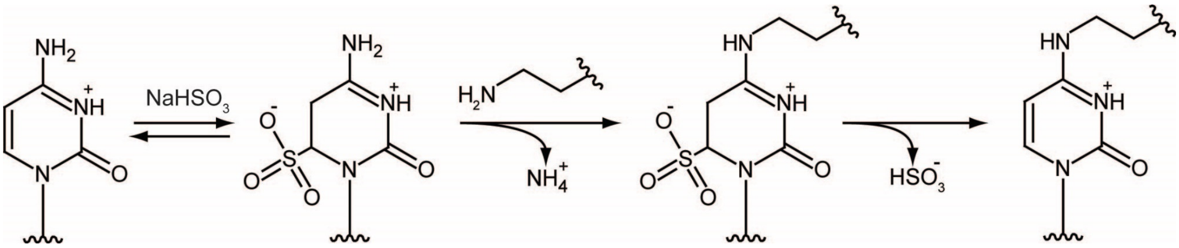


Figure 3.11 Sulfite-mediated transamination of cytosine with primary amines

Schematic representation of sulfite-mediated conjugation of cytosine with lysine side chain. Reaction mechanism is redrawn according to [85].

Sulfite-mediated cross-linking could be successfully used to identify cross-links between the model protein GAPDH and poly(UC)$_{12}$ RNA oligonucleotide (Table 3.3) with both sodium bisulfite solution and sodium metabisulfite. The reaction seems to be highly specific to lysines, but it was less efficient in the tested conditions than utilizing UV irradiation (data not shown). The specificity of the reaction and the limited neutral loss fragments observed allow identification to be carried with standard mass spectrometry search engines, such as MaxQuant [86] as a simple modification FDR controlled search (Suppl. Table 3.4), without the need for manual validation. Therefore, sulfite-mediated cross-linking is an interesting and highly specific method for the analysis of protein-RNA cross-linking by mass spectrometry, however, under the investigated conditions it was inferior to standard UV-mediated cross-linking.

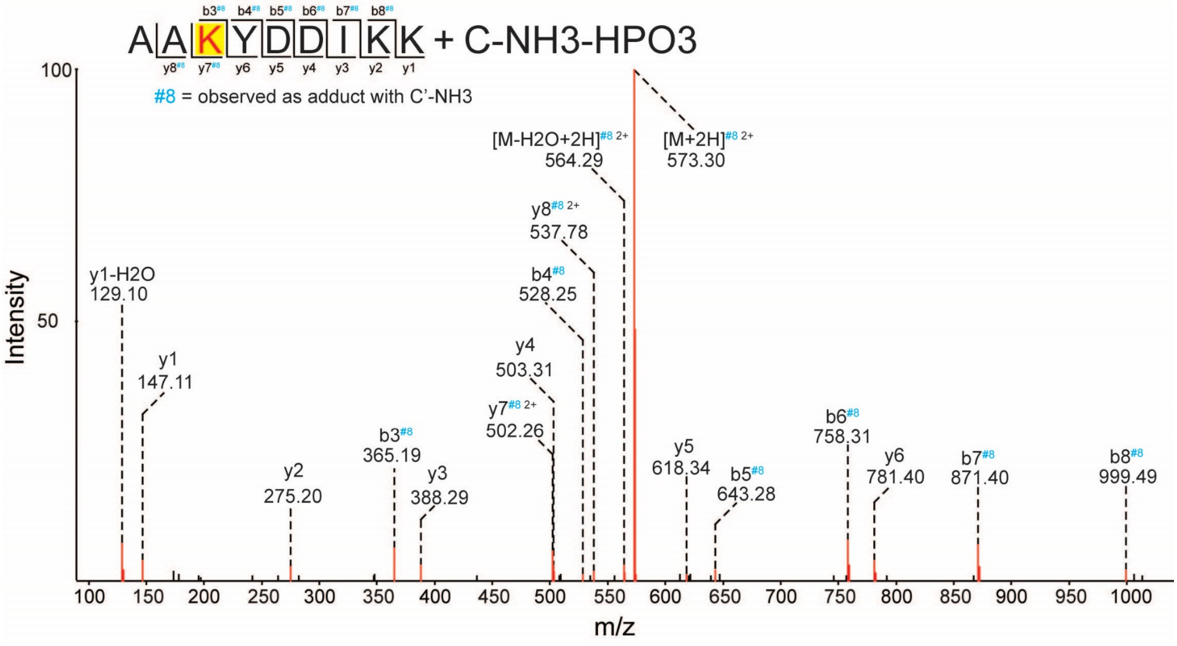


Figure 3.12 Exemplary sulfite-mediated cross-link spectrum

Fragment spectrum of peptide $_{250}$AAKYDDIKK$_{258}$ with C-NH3-HPO3, generated by bisulfite cross-linking of GAPDH with poly(UC). The shifted b- and y-ion series localize $_{252}$K as the cross-linked amino acid. The fragmentation behavior is completely analogous to UV generated cytosine cross-links.

Table 3.3 Identified sulfite-induced cross-links

GAPDH and poly(UC) were cross-linked in the presence of 50 mM bisulfite solution or freshly dissolved metabisulfite.

Protein (UniProt ID)	**Peptide**	**RNA adduct(s)**	**Cross-link localization**	**Crosslinker**
GAPDH (P46406)	$_{250}$AAKYDDIKK$_{258}$	C-NH3-HPO3	$_{252}$K	bisulfite
	$_{60}$AENGKLVINGK$_{70}$	C-NH3-HPO3	$_{63}$GK$_{64}$	bisulfite
	$_{214}$AVGKVIPELNGK$_{225}$	C-NH3-HPO3	$_{217}$K	bisulfite
	$_{196}$DGRGAAQNIIPASTGAAKAVGK$_{217}$	C-NH3-HPO3	$_{213}$K	bisulfite
	$_{199}$GAAQNIIPASTGAAKAVGK$_{217}$	C-NH3-HPO3	$_{213}$K	bisulfite
	$_{253}$YDDIKKVVK$_{261}$	C-NH3-HPO3	$_{257}$K	bisulfite
	$_{250}$AAKYDDIKK$_{258}$	C-NH3-HPO3	$_{252}$K	metabisulfite
	$_{60}$AENGKLVINGK$_{70}$	C-NH3-HPO3	$_{63}$GK$_{64}$	metabisulfite
	$_{214}$AVGKVIPELNGK$_{225}$	C-NH3-HPO3	$_{217}$K	metabisulfite
	$_{196}$DGRGAAQNIIPASTGAAKAVGK$_{217}$	C-NH3-HPO3	$_{212}$AK$_{213}$	metabisulfite
	$_{199}$GAAQNIIPASTGAAKAVGK$_{217}$	C-NH3-HPO3	$_{213}$K	metabisulfite
	$_{253}$YDDIKKVVK$_{261}$	C-NH3-HPO3	$_{257}$K	metabisulfite

3.4 Purification of cross-links from complex sources

Due to the low generation efficiency of protein-RNA cross-links, biochemical enrichment is pivotal for their identification. The established reversed-phase C18/TiO_2 purification strategy relies on the enrichment of cross-link heteroconjugates based on their phosphate group. It is suitable for analysis of low complexity protein mixtures and isolated or reconstituted complexes, but it is not suitable for analysis of cross-link experiment of entire cells, where a large amount of phosphopeptides and other metabolites with phosphate groups compete with the cross-linked heteroconjugates. In such cases, a different strategy utilizing size exclusion chromatography that takes advantage of the different sizes of the RNA molecule and digested peptides or specific pull-downs (e.g. oligo(dT) hybridization) can be used to deplete the overwhelming amount of non-cross-linked peptides. Although fairly successful [39], these workflows show some disadvantages. Size exclusion chromatography is often not efficient in separating peptides from RNA and specific pull-down strategies limit the amount of starting material that can be used and can only enrich for certain RNA species. Therefore, the mass spectrometric investigation of complex samples can benefit from the development of a new biochemical enrichment method that would allow effective depletion of non-cross-linked peptides and can be used for the analysis of all RNA species in the cell.

The amount of purified cross-linked heteroconjugates and the presence of non-cross-linked peptides are the major limiting factors in detection of cross-links in a sample. Therefore, a closer examination was taken at classical solid phase extraction strategies for RNA purification that allow substantial depletion of peptide species, namely silica-based purification and strong anion exchange (SAX) chromatography [87,88]. When utilizing silica-based purification, the negatively charged phosphate groups of nucleic acids bind the negatively charged silanol groups through a cation salt bridge (Fig. 3.13) in the presence of chaotropic salts or alcohols, whereas proteins have low affinity to silica and can be washed away. The bound nucleic acids are then eluted with low ionic strength solutions. Similarly, when using strong anion exchange separation, the phosphate groups of nucleic acids form a strong electrostatic interaction with the immobilized, positively charged groups of the matrix (e.g. quaternary ammonium) and can be differentially eluted from proteins that have a weaker interaction by using a high ionic strength or change of the buffer pH.

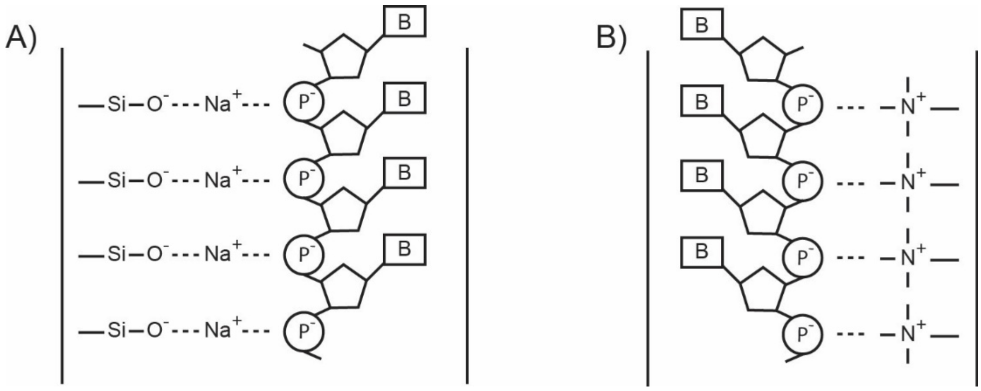


Figure 3.13 Solid-phase extraction of RNA

Oligonucleotides can be enriched on the basis of the phosphodiester backbone by cooperative binding with solid-phase matrixes. A) Silica-based purification relies on the interaction of the negatively charged silanol groups with the phosphate groups through a cation salt bridge B) Strong anion exchange binding is mediated by strong electrostatic interactions of the immobilized positively charged groups with the negatively charged phosphate backbone.

In order to determine the appropriate conditions for optimal partition of peptides and nucleic acids during strong anion exchange chromatography, a set of preliminary experiments was performed. Membrane-based spin columns with immobilized quaternary ammonium functional groups were used as a strong anion exchange matrix. To characterize the peptide elution behavior, a trypsin digested peptide mixture derived from HeLa nuclear extract was subjected to a stepwise elution strong anion exchange chromatography and analyzed by LC-MS and MaxQuant (Fig. 3.14 A). In parallel, an undigested protein mixture was also fractionated and analyzed by SDS-PAGE (Fig. 3.14 B). Most peptides and intact proteins were removed from the column with 400 mM NaCl at pH 6. When total RNA derived from *E. coli* was examined, it eluted from the column at salt concentrations higher than 600 mM at pH 6 (Fig. 3.14 C). Even though proteins and RNA could be differentially eluted, the behavior of RNA molecules covered with cross-linked proteins might be substantially altered, making it difficult to separate it from the non-cross-linked proteins. Thus, it became apparent that initial digestion of proteins to smaller polypeptide stretches is warranted before separation by strong anion exchange chromatography. To account for any possible sequestering effects that the cross-linked peptides might exert, the lowest salt concentration that allows substantial peptide depletion (400 mM NaCl, pH 6) was selected as a separation buffer.

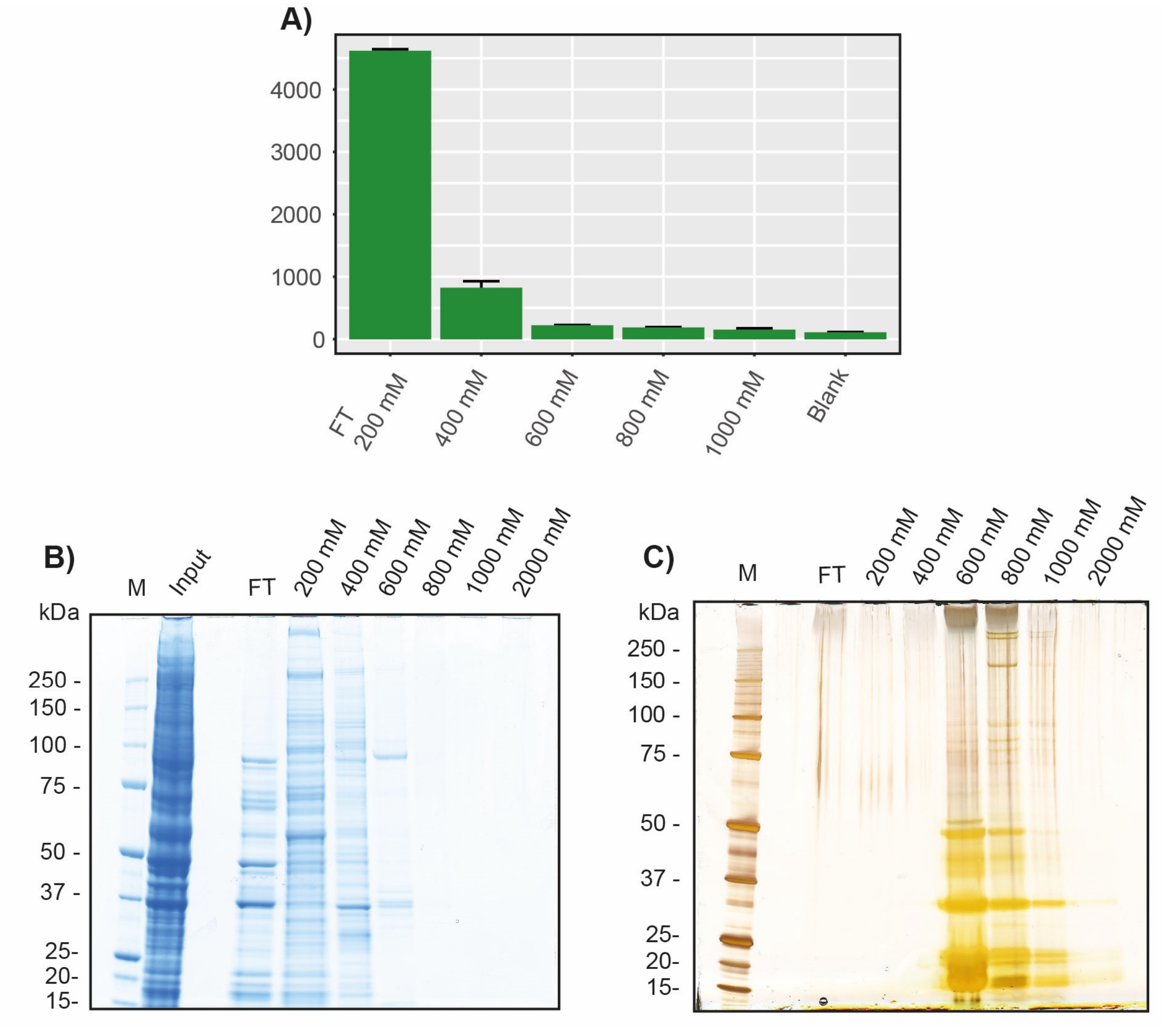


Figure 3.14 Characterization of elution conditions of peptides, proteins and RNA subjected to strong anion exchange chromatography

A) Number of identified peptides eluted at different salt concentrations from strong anion exchange column based on three replicates. Blank count based on blank samples after each replicate series for estimation of carry-over. B) SDS-PAGE analysis of differentially eluted proteins from HeLa nuclear extract that were separated by strong anion exchange chromatography C) Silver staining of *E.coli* total RNA subjected to strong anion exchange chromatography

Silica enrichment protocols have been developed for purification of non-cross-linked RNA species [88,89]. It is not clear if the spatial hindrance of the attached cross-linked proteins or damage from the UV irradiation might make them ineffective when utilized for enrichment of cross-linked RNA. To investigate that possibility, a feasibility study was conducted with simple complexes that confirmed silica enrichment could be successfully used for the identification of cross-linked heteroconjugates. Although both purification of intact cross-linked proteins and digested peptides could be demonstrated, the yield of purified RNA and cross-link numbers observed with intact proteins was diminished (data not shown). Therefore, a protein digestion step was implemented before purification for the silica-based workflows as well. Most promising results of this preliminary testing were achieved with protocols based on the Qiagen RNEasy Maxi kit, optimized for purification of RNA species longer than 200 nucleotides and the Zymo Research Direct-zol RNA Miniprep Plus kit, optimized for isolation of RNA species longer than 17 nucleotides (based on manufacturer's product description).

3.4.1 Investigation of RNA-binding proteins in *Escherichia coli*

The incorporation of automated annotation and the development of improved scoring substantially shortens the identification timeline of peptide-RNA contacts derived from simple mixtures, allowing workflow completion in the matter of days. In addition to that, the combination of the faster dedicated RNPxlSearch search engine and the consistent sensitivity developments of the instrumentation might allow extending the existing workflow to global studies of RNA binding proteins derived from highly complex mixtures, such as whole cells. In order to address this question, silica and SAX-based workflows were utilized to identify RNA-binding proteins in *Escherichia coli* and localize the exact position where the contact with the nucleic acid takes place on the protein sequence.

Incorporation of photoreactive analogs, such as thionucleotides, is a well-established solution to improve the low yield of nucleotide cross-linking with UV at 254 nm. This strategy has been investigated with eukaryotic cells such as the yeast *Saccharomyces cerevisiae* [90] and mammalian cells [91], but is not well characterized with bacterial cells such as *Escherichia coli*.

In initial experiments employing 4-thiouracil incorporation and irradiation at 365 nm in *E. coli* XL Gold, no protein-RNA cross-linking heteroconjugate products could be identified (data not shown). Utilizing mutant auxotroph strain K12 pyrD (Fig. 3.15) that is unable to synthetize uracil with complex media such as Lysogeny broth (LB) also had an unsatisfying outcome – only few proteins could be identified after irradiation at 365 nm, represented by highly abundant proteins (such as ribosome proteins) and proteins in natural contact with 4thio-uracil RNA (e.g. tRNA sulfurtransferase) (Exemplary file – I 4SU LB). Therefore, incorporation was tested in completely synthetic media (M9) supplemented with controlled amounts of uracil and 4-thiouridine based on a protocol adapted from [92]. These conditions gave rise to a substantial number of protein-RNA heteroconjugates identifications (Exemplary file: III 4SU M9) and were used for the following experiments. Subsequently, optimization of the irradiation time for upscaling to larger numbers of cells was performed.

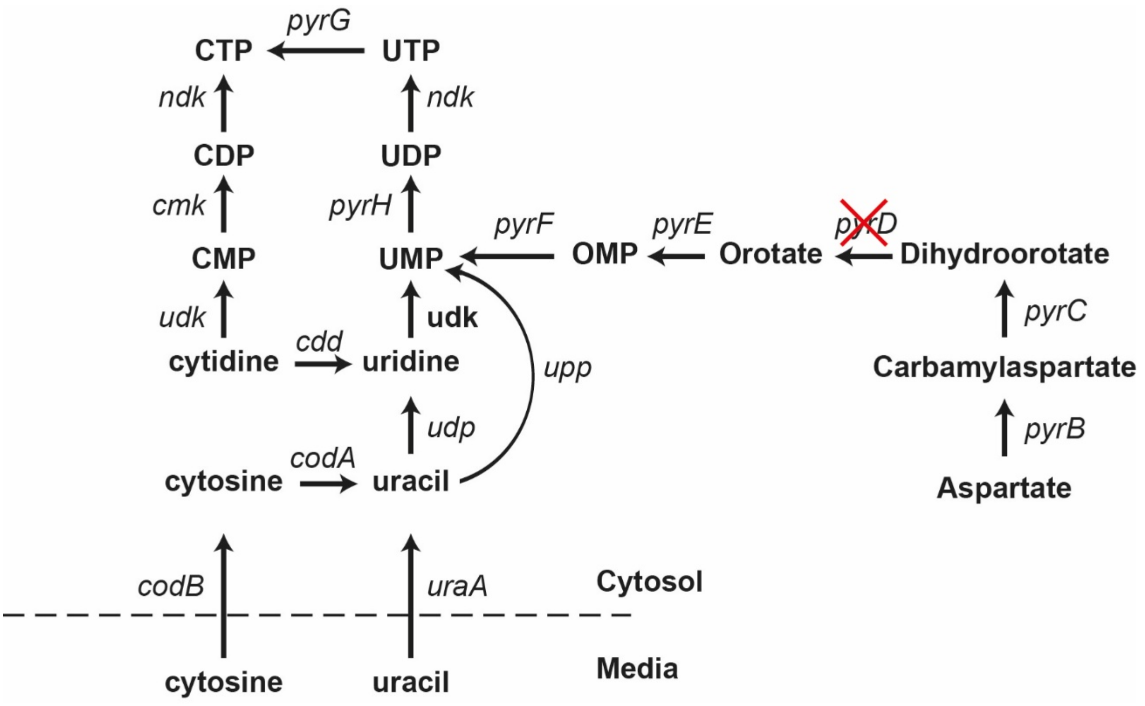


Figure 3.15 Overview of pyrimidine nucleotide biosynthetic and salvage pathways.

Gene names are used to indicate involved enzymes. Pathways redrawn as described in [93].

Identified cross-linking products between proteins and 4-thiouridine mostly involve a net loss of H2S [94]. Thus, the adduct formed has a molecular composition of C9H1N2O8P (306.0256 Da) that is equivalent to U-H2O and C-NH3 adducts. Moreover, the fragmentation behavior of such heteroconjugates is completely identical, making them undistinguishable. Therefore, it is not possible to determine by mass spectrometry, which of these pyrimidine bases had formed the cross-link. For ease of presentation in following tables, the 306.0256 Da adduct is notated as “N”. In spectrum figure representations the most likely cross-linked nucleotide for the selected conditions is presented – for experiments with 4SU incorporation and irradiation at 365 nm – 4SU, for irradiation at 254 nm – U-H2O or C-NH3.

Under the identified conditions, purification protocols utilizing silica-based purification and strong anion exchange chromatography were established. Due to the low efficiency of cross-linking a large amount of starting material is required. This poses practical difficulties as cross-linking protocols call for high amounts of trypsin to be used (e.g. 1:20 w/w) in order to ensure efficient digestion of the less accessible cross-linked complexes down to peptides that can be easily ionized and identified [42]. Simply upscaling would result in impractical and expensive protocols. To circumvent the need for enormous amount of enzyme, it was decided to modify the upscaled enrichment workflows into a two-step purification protocol. This way, a sample could be initially digested with low amount of protease (e.g. 1:200 w/w), allowing the removal of proteins from the cross-linked RNA species with reasonable expenditure of enzyme required. RNA and heteroconjugates could be purified, depleting most of peptides. Thus, a second digestion step could be performed where a substantially higher ratio of trypsin to peptides can ensure cleavage around the challenging cross-linked sites. Next, repurification of the RNA and heteroconjugates would assist in maximal removal of peptides. Finally RNAse digestion, desalting and LC-MS analysis would follow (Fig. 3.16).

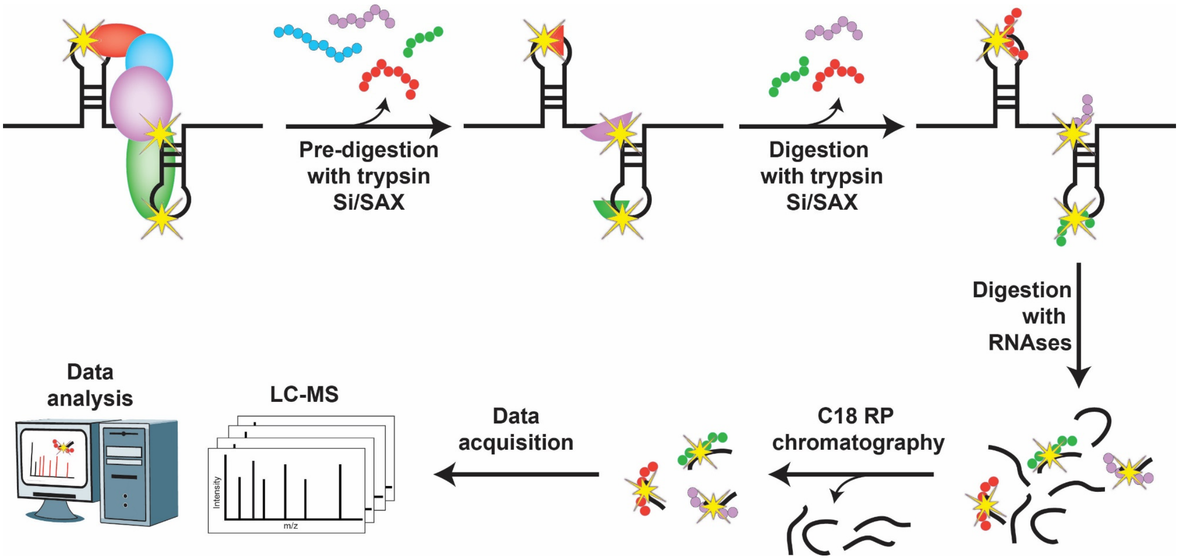


Figure 3.16 Schematic representation of two-step purification silica and strong anion exchange workflows

Taking into account the above mentioned considerations, both purification strategies were employed to analyze the RNA-binding proteins of 4SU-labeled and unlabeled *Escherichia coli* cells. The combined result files (exemplary files I-IX) gave rise to an enormous number of peptide-RNA heteroconjugate candidate spectra (99,132), making it impossible to employ standard, thorough manual validation strategy. For that reason spectral FDR calculation based on the working combined score from RNPxlSearch and rescoring from Percolator was applied individually for each file. The filtered spectral results at 1% FDR can be found as supplementary table on the attached CD. An overview of the identified spectral, peptide and protein counts, as well as the cross-linked nucleotides according to the automatic annotation, can be seen in Figure 3.17. In total 34,282 cross-link spectra could be identified, leading to 1,377 unique peptide sequences and 468 proteins from the *E. coli* proteome. Both silica-based and strong anion exchange-based workflows resulted in considerable depletion of non-cross-linked peptides.

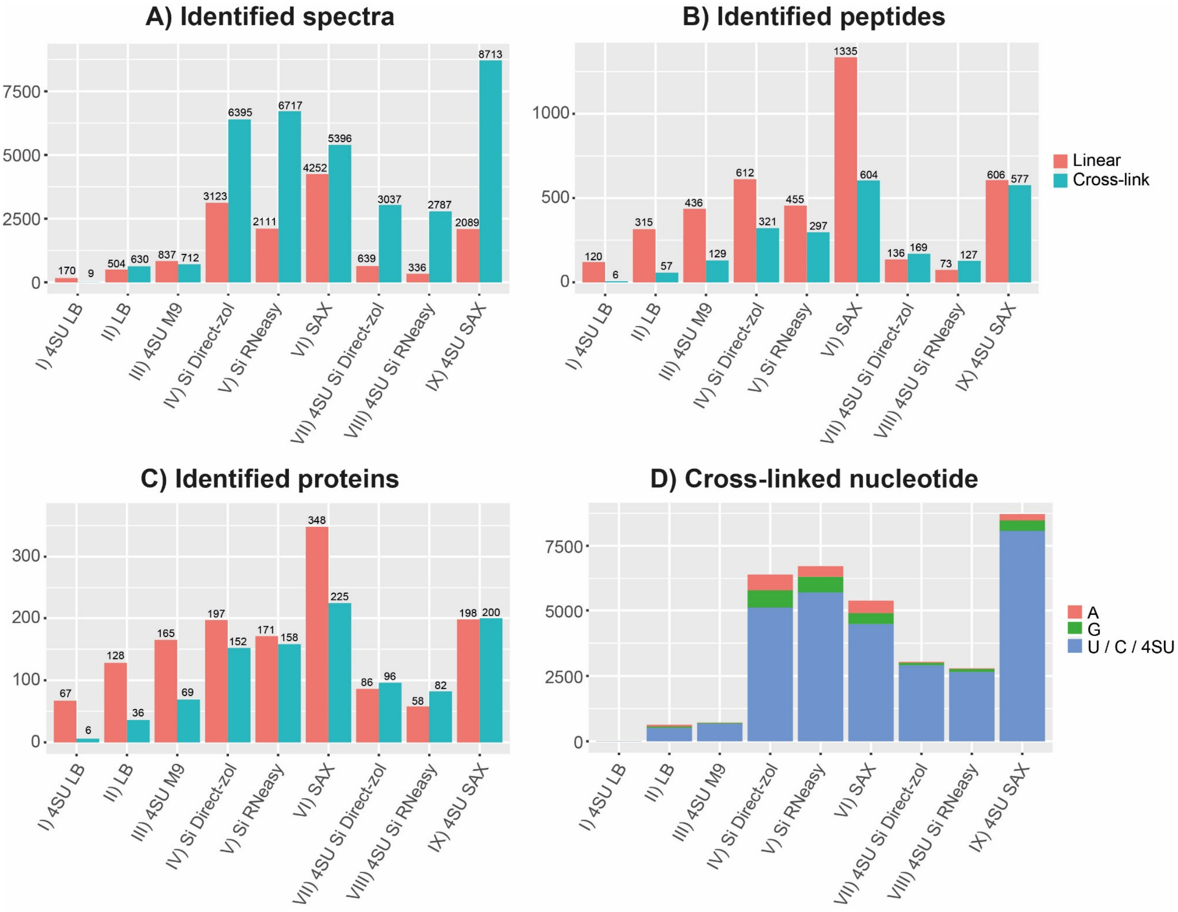


Figure 3.17 Overview of identification results from *E. coli* samples

A) Number of identified spectra at 1% FDR that lead to identification of linear peptides or peptide-RNA cross-links B) Peptide sequence hits derived at 1% spectral FDR. Modified (e.g. methionine oxidation) and unmodified peptides were counted as a single sequence. C) Identified proteins at 1% spectral FDR without any calculations of protein interference. D) Cross-linked nucleotide as reported by the RNPxlSearch localization score.

All 1% FDR-filtered identifications from presented files (I-IX) were combined, hits matching possible contaminants and linear peptides were removed, and gene ontology enrichment analysis was performed with the STRING database [64]. Functional enrichment in molecular functions are presented in Table 3.4. As shown, 107 out of 179 RNA-binding proteins were identified and the results were substantially enriched on RNA-binding proteins (FDR 3.67E-36) with thorough representation of ribosome constituent proteins (53 of 57 FDR 6.08E-23) and tRNA-binding proteins (29 of 39 FDR 6.81E-11). In addition to that, 13 of 18 mRNA-binding proteins in the database could be detected (8.74E-05). Protein details and full GO enrichment results can be found as supplementary tables on the attached CD.

Table 3.4 Gene ontology enrichment analysis of protein-RNA cross-links identified in *E. coli*.

Unique protein hits at 1% spectral FDR were submitted into the STRING database. Shared peptides that lead to more than one protein accession were limited to only the first accession hit from the list to avoid artificial inflation of the group count.

GO-term	Description	Count in gene set	False discovery rate
GO:0003723	RNA binding	107 of 179	3.61E-36
GO:0003735	structural constituent of ribosome	53 of 57	6.08E-23
GO:0005198	structural molecule activity	56 of 73	1.14E-21
GO:0097159	organic cyclic compound binding	251 of 1377	2.44E-17
GO:1901363	heterocyclic compound binding	251 of 1377	2.44E-17
GO:0019843	rRNA binding	43 of 55	8.66E-17
GO:0003676	nucleic acid binding	157 of 684	2.1E-16
GO:0005488	binding	328 of 2121	6.78E-16
GO:0140098	catalytic activity, acting on RNA	59 of 138	4.58E-14
GO:0140101	catalytic activity, acting on a tRNA	35 of 59	4.14E-11
GO:0000049	tRNA binding	29 of 39	6.81E-11
GO:0004812	aminoacyl-tRNA ligase activity	21 of 26	3.07E-08
GO:0036094	small molecule binding	137 of 792	4.13E-06
GO:0017076	purine nucleotide binding	95 of 493	1.08E-05
GO:0035639	purine ribonucleoside triphosphate binding	93 of 482	1.33E-05
GO:0032555	purine ribonucleotide binding	94 of 492	1.58E-05
GO:0000166	nucleotide binding	118 of 680	3.27E-05
GO:0043168	anion binding	122 of 712	3.27E-05
GO:0003729	mRNA binding	13 of 18	8.74E-05
GO:0032553	ribonucleotide binding	95 of 525	9.05E-05
GO:0016874	ligase activity	31 of 102	9.74E-05
GO:0030554	adenyl nucleotide binding	83 of 440	9.85E-05
GO:0005524	ATP binding	82 of 435	0.00011
GO:0008144	drug binding	95 of 530	0.00011
GO:0097367	carbohydrate derivative binding	97 of 548	0.00013
GO:0045182	translation regulator activity	12 of 23	0.0017
GO:0044877	protein-containing complex binding	16 of 43	0.0027
GO:0090079	translation regulator activity, nucleic acid binding	11 of 21	0.003
GO:0043021	ribonucleoprotein complex binding	15 of 40	0.0039
GO:0043167	ion binding	182 of 1312	0.0053
GO:0002161	aminoacyl-tRNA editing activity	7 of 9	0.0085
GO:0048027	mRNA 5'-UTR binding	6 of 6	0.009
GO:0070180	large ribosomal subunit rRNA binding	6 of 6	0.009

Manually browsing through the resulting files showed a considerable number of spectra that would have been approved under manual validation conditions but could not reach the 1% FDR cut off of the combined RNPxlSearch scoring. Additionally, some examples of spectra that passed the FDR filtering, but would be considered of poor quality by manual validation criteria could be detected. Since the false discovery rate filtering was performed only on spectral level, with majority of spectral hits leading to few high abundant known RNA-binding proteins, the false discovery rate at peptide and protein level might be substantially different. Manual validation has been the standard used for years in the Urlaub Research Group for evaluation of cross-linked spectra and utilized to reject hits of lower quality with the intent to minimize the chance of reporting a false-positive hit. Due to the large number of spectra, a comprehensive manual evaluation would require substantial amount of time and could not be performed on the entirety of the dataset. In order to compare how the established manual validation strategy would evaluate the automatically selected spectra, a closer look was taken into a subset of cross-linked proteins that constitute the *E. coli* RNA polymerase complex. All cross-link spectra at 1% FDR that lead to identification of a cross-link site were extracted and manually validated. The majority of identified spectra (224 out of 243) and identified peptides (29 out of 35) would also fit manual validation criteria. Identified cross-link sites that were also manually validated are mapped onto crystal structure of the RNA polymerase complex and are visualized in Figure 3.18. As shown, most cross-links fall around the transcription bubble, in close proximity to the expected nascent RNA chain.

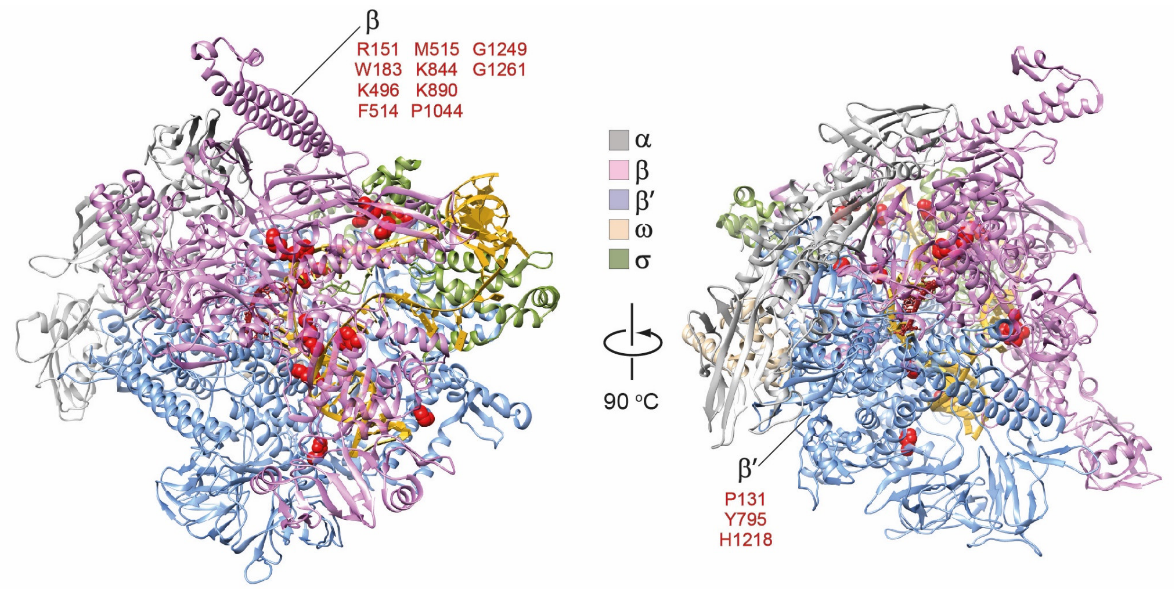


Figure 3.18 Identified cross-link sites in *E. coli* RNA polymerase

Manually confirmed hits in which the cross-linked amino acid could be localized were mapped onto crystal structure [PDB ID 5IPM] [95]. DNA is colored goldenrod, RNA is colored brown. Proteins are differentially colored. Cross-linked residues are represented as red spheres according to the atom’s Van der Waals radii.

In the next step were addressed proteins that have not been described as RNA-binding or their interaction with nucleic acid was not extensively studied. The majority of these proteins were presented only by a single cross-linked peptide. Whereas, almost all identified spectra of the RNA polymerase complex were of high quality and the majority of the cross-linked peptides could be manually verified, the quality of the cross-linked spectra of the examined proteins varied substantially. In total, 64 proteins were examined and 42 could be validated and are shown in Suppl. Table 3.5. Within the confirmed hits that were not previously described as RNA-binding proteins in the UniProt database, seven enzymes involved in the lipid biosynthetic process could be identified (LPXD, FABA, FABG, FABI, FABZ, PSS, ACP). To no surprise, the list also contained five glycolytic enzymes, including Glyceraldehyde-3-phosphate dehydrogenase A, Fructose-bisphosphate aldolase class 2 and Phosphoglycerate kinase (G3P1, ALF, DLDH, ODP2, PGK). Three cross-linked peptides could be confirmed for G3P1 – $_{161}$VINDNFGIIEGLM(Ox)TTVHATTATQK$_{184}$, $_{185}$TVDGPSHKDWR$_{195}$ and $_{322}$VLDLIAHISK$_{331}$. When aligned with the sequence of rabbit muscle GAPDH by Clustal Omega [96], the two localized sites with amino acid resolution - $_{192}$K and $_{331}$K of G3P1 fall in close proximity with $_{196}$DGRG$_{199}$ and $_{329}$M sites identified between GAPDH and poly(U) (Suppl. Table 3.2).

In addition, peptides from five DNA-binding proteins such as transcription factors Met repressor (Fig. 3.19 A) and Probable transcriptional regulatory protein YebC, as well as proteins involved in DNA damage response - Exodeoxyribonuclease III, protein RecA and DNA-binding protein HU-alpha could be identified as cross-linked to RNA adducts (METJ, YEBC, EX3, RECA, DBHA). Other interesting examples involved in protein transport, are the ribosome associated Trigger factor (Fig. 3.19 B) and Protein translocase subunit SecY (TIG, SECY). Additionally, several suspected or known but less extensively structurally studied RNPs could be confirmed and the cross-link site localized (YHBY, YFIF, TSAB, TRUD, RNR, DUSB).

Examples of confirmed hits that have a resolved structure in complex with RNA are shown in Fig. 3.20 (RNB, NUSA/NUSB, MNMA). Cross-linked sites or regions of Exoribonuclease 2 (RNB) could be identified in 4 peptides - $_{32}$G[F]GFLEVDAQK$_{42}$,$_{43}$SY[F]IPPPQMK$_{52}$, $_{69}$[ERESA]EPEELVEPFLTR$_{85}$, $_{580}$LVDNGAIA[F]IPAPFLHAVR$_{598}$. When mapped onto a crystal structure [PDB ID 2IX1], $_{588}$F falls in close proximity to the resolved poly(A) RNA oligonucleotide and $_{33}$F, $_{45}$F and $_{81}$ERESA$_{85}$ are localized in close proximity to the expected binding path of a longer nucleotide chain (Fig. 3.20 A). Transcription termination proteins NUSA and NUSB were presented with respectively two peptides – $_{4}$EILAVVEAVSNE[K]ALPR$_{20}$, $_{132}$EHEGEIITGVV[K]K$_{144}$ for NUSA and one peptide $_{96}$SDV[PYK]VAINEAIELAK$_{112}$ for NUSB. If mapped onto EM structure [PDB ID 5MS0], the localized $_{96}$PYK$_{112}$ region is in close proximity with the resolved structure of the nascent RNA (Fig. 3.20 B). In the case of the tRNA-specific 2-thiouridylase Mnma, a peptide $_{150}$DQSY[F]LYTLSHEQIAQSLFPVGELEKPQVR$_{179}$ could be identified. When mapped onto

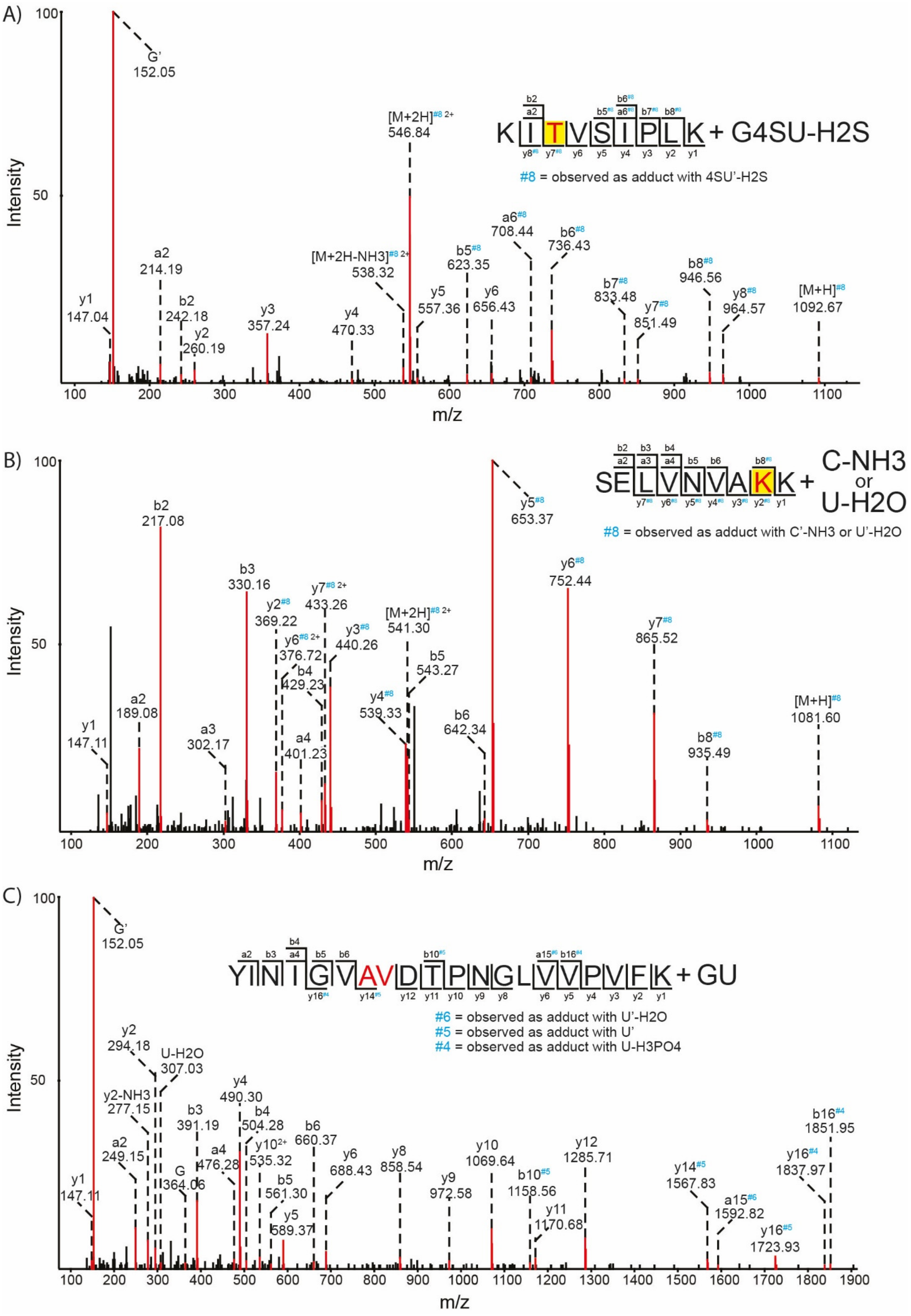


Figure 3.19 Exemplary cross-link spectra of novel RBPs identified in *E. coli*.

A) Fragment spectrum of Met repressor cross-linked peptide $_{24}$KI[T]VSIPLK$_{32}$ with RNA adduct G4SU-H2S. The validity of the identification is supported by almost complete sequencing series of y-ions, nearly complete series of b-ions and observed proline effect. In addition a strong suppressive G' marker ion and prominent precursor peak shifted with 94 Da nucleobase adduct produced after cleavage of the glycolytic bond can be observed.

The shifted y7 and y8 ions localize $_{26}$T as the cross-linked amino acid B) Trigger factor cross-linked peptide $_{30}$SELVNVA[K]K$_{38}$ to C-NH3 or U-H2O RNA adduct. The identity of the hit is supported by almost complete y- and b- sequencing ions. The shifted y2 and b8 ions localize the cross-linked residue at $_{37}$K. C) ODP2 cross-linked peptide $_{493}$YINIGVAVDTPNGLVVPVFK$_{512}$ with GU. The spectrum shows substantial number of identified y- and b- sequencing ions, coherent proline effect and expected high intensity G' marker ion. The shifted y14 ion localizes the cross-linked site at either $_{499}$A or $_{500}$V.

crystal structure [PDB ID 2DER], the cross-linked residue $_{154}$F is located in the vicinity of nucleotides $_{34}$U and $_{35}$U of tRNA-Glu (Fig. 3.20 C). Overall, considering also a number of cases not presented here, available protein-RNA structures are highly consistent with the identified cross-linking results.

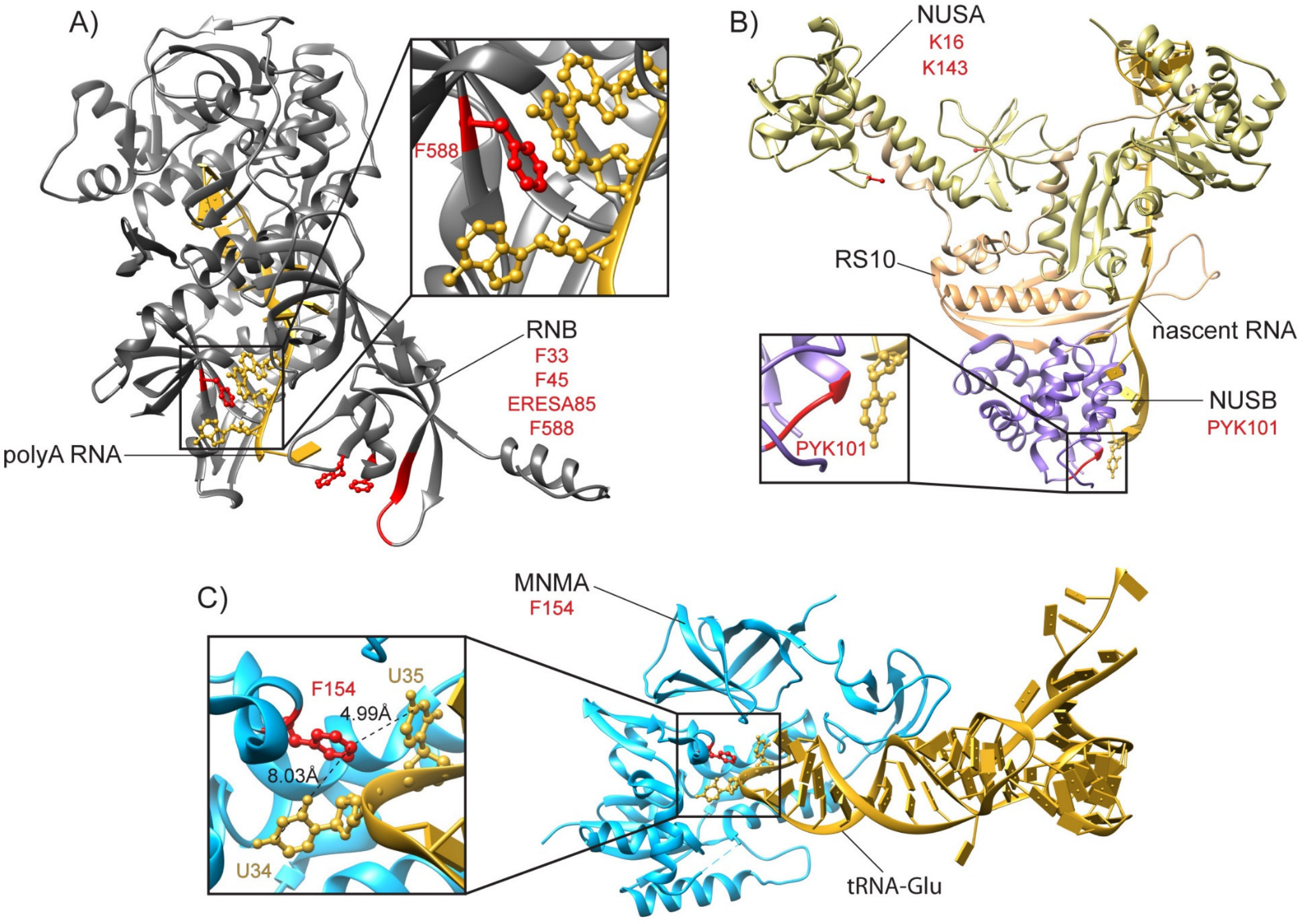


Figure 3.20 Identified heteroconjugates of *E. coli* proteins are in good agreement with available structures.

Cross-linked sites and regions of confirmed peptides were mapped onto available crystal and EM structures. Cross-linked residues are marked in red. A) Identified contacts between Exoribonuclease 2 and RNA are mapped onto X-ray crystallographic structure [PDB ID 2IX1] [97] B) Cross-links detected in Transcription termination proteins NUSA and NUSB are illustrated on lambda-based antitermination complex [PDB ID 5MS0] [98]. Some proteins are omitted for clarity. C) The Identified cross-link site of tRNA-specific 2-thiouridylase Mnma is mapped onto available crystal structure [PDB ID 2DER] [99].

3.4.2 Investigation of RNA-binding proteins in cytoplasmic extract of HeLa cells

Next, the established workflows were tested on a highly complex human sample – cytoplasmic extract from HeLa cells. First, a feasibility experiment with silica-based purification was performed under standard irradiation conditions – 10 minutes / 254 nm (Exemplary file: I). A substantial number of cross-link sites could be observed under those conditions. Subsequently, UV-generated cross-links from cytoplasmic extract were analyzed with all 4 available workflows: standard C18/TiO_2 (Exemplary file: II), two silica-based purification protocols (Direct-zol, III and RNeasy, V) and strong anion exchange based workflow (SAX, IV). Additionally, a sulfite-mediated workflow based on the Direct-zol purification was employed in parallel (VI).

An overview of the identification results from the HeLa samples at 1% spectral FDR can be seen in Fig. 3.21. Extended tables with the identification results are available on the attached CD. In total 11,294 cross-link spectra, 857 unique peptide sequences and 488 proteins from the human proteome could be identified. Similarly to the *E. coli* samples, the strong anion exchange and silica workflows depleted the majority of non-cross-linked peptides. As expected, the standard C18/TiO_2 protocol did not lead to substantial enrichment of cross-links over non-cross-linked peptides. Additionally, browsing through the identification results revealed a quality difference on the spectral level in comparison with the other workflows. The majority of fragment spectra contained a considerable amount of noise, most likely due to the increased interference of the co-eluting linear peptides. Similar to the *E. coli* samples, the cross-links were identified predominantly to pyrimidine bases. The sulfite-mediated cross-linking gave rise to only few cross-link hits with simple fragmentation behavior, mostly represented by the most abundant RNA-binding proteins, such as ribosomal constituents. Although considerable amounts of enzyme and prolonged incubation were employed to digest down the RNA, the majority of detected precursor adducts were of length of two or three nucleotides. Most likely the access to cross-linked RNA regions is heavily hindered and additional optimization of the RNA digestion step is needed.

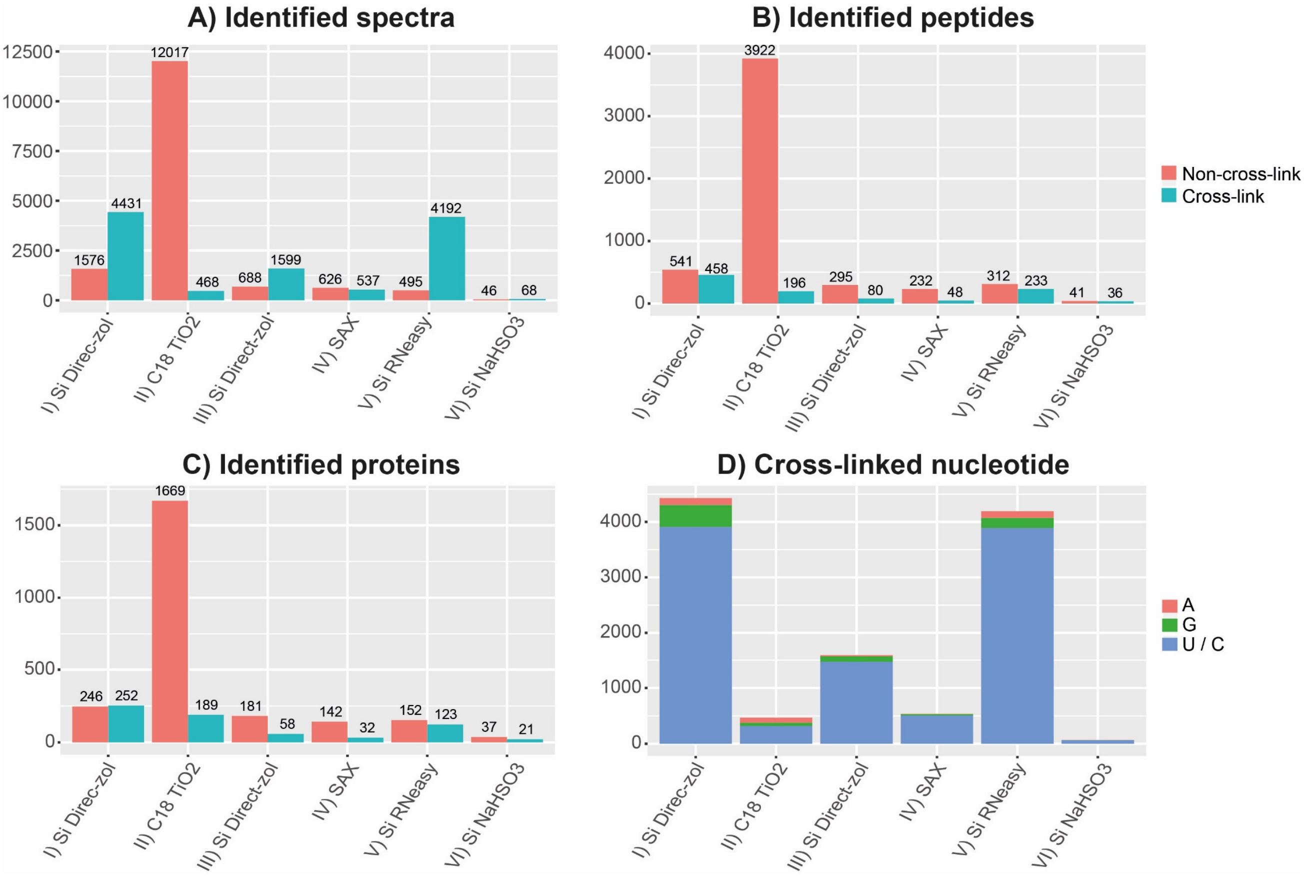


Figure 3.21 Overview of Identification results from HeLa cytoplasm

A) Number of identified spectra at 1% FDR that lead to identification of linear peptides or peptide-RNA cross-links B) Peptide sequence hits derived at 1% spectral FDR. Modified (e.g. methionine oxidation) and unmodified peptides were counted as a single sequence C) Identified proteins at 1% spectral FDR without any protein interference calculation. D) Cross-linked nucleotide as reported by the RNPxlSearch localization score.

A considerable number of the identified protein hits has been identified as part of the mRNA interactome (138) or candidate RBPs (44) by a proteomics study performed by Castello et al. [49]. Their resurgent detection confirms the validity of the suggested direct interaction and adds localization information of the cross-link site that is not obtainable in proteomics-based results.

The combined protein results at 1% spectral FDR were submitted for gene ontology enrichment analysis with STRING (Table 3.5). Significant enrichment of known RNA-binding proteins could be observed (128 out of 850 FDR 3.93E-59). Substantial enrichment examples include structural constituent of the ribosome (53 of 146 FDR 8.44E-39), mRNA-binding proteins (44 of 198 9.98E-25), tRNA-binding proteins (10 of 56 FDR 4.99E-05), snRNA-binding proteins (9 of 38 2.55E-05) and U3 snoRNA-binding proteins (3 of 5 FDR 7.7E-04).

Table 3.5 Gene ontology enrichment analysis of protein-RNA cross-links identified in HeLa cytoplasmic extract

Unique protein hits at 1% spectral FDR were submitted into the STRING database. Shared peptides that lead to more than one protein accession were limited to only the first accession hit from the list to avoid artificial inflation of the group count.

GO-term	*Description*	*Count in gene set*	*False discovery rate*
GO:0003723	RNA binding	128 of 850	3.93E-59
GO:0003735	structural constituent of ribosome	53 of 146	8.44E-39
GO:0003676	nucleic acid binding	180 of 3332	1.75E-26
GO:0005198	structural molecule activity	75 of 679	3.38E-25
GO:0003729	mRNA binding	44 of 198	9.98E-25
GO:1901363	heterocyclic compound binding	233 of 5305	1.93E-24
GO:0097159	organic cyclic compound binding	233 of 5382	1.34E-23
GO:0045182	translation regulator activity	25 of 124	6.31E-13
GO:0019843	rRNA binding	18 of 60	1.20E-11
GO:0003730	mRNA 3'-UTR binding	18 of 63	2.18E-11
GO:0090079	translation regulator activity, nucleic acid binding	21 of 99	2.99E-11
GO:0005488	binding	348 of 11878	2.52E-10
GO:0008135	translation factor activity, RNA binding	18 of 84	1.07E-09
GO:0140098	catalytic activity, acting on RNA	32 of 345	1.25E-08
GO:0043021	ribonucleoprotein complex binding	19 of 116	1.49E-08
GO:0003727	single-stranded RNA binding	16 of 80	2.87E-08
GO:0005524	ATP binding	74 of 1462	3.10E-08
GO:0035639	purine ribonucleoside triphosphate binding	85 of 1794	3.10E-08
GO:0030554	adenyl nucleotide binding	76 of 1524	3.13E-08
GO:0032559	adenyl ribonucleotide binding	75 of 1514	5.22E-08
GO:0017076	purine nucleotide binding	86 of 1865	6.89E-08
GO:0019899	enzyme binding	96 of 2197	8.77E-08
GO:0032555	purine ribonucleotide binding	85 of 1853	1.02E-07
GO:0008144	drug binding	78 of 1710	6.14E-07
GO:0000166	nucleotide binding	90 of 2097	6.17E-07
GO:0048027	mRNA 5'-UTR binding	9 of 24	1.16E-06
GO:0097367	carbohydrate derivative binding	91 of 2163	1.16E-06
GO:0003725	double-stranded RNA binding	13 of 70	1.40E-06
GO:0036094	small molecule binding	96 of 2460	1.25E-05
GO:0044877	protein-containing complex binding	49 of 968	1.87E-05
GO:0003743	translation initiation factor activity	10 of 50	2.53E-05
GO:0017069	snRNA binding	9 of 38	2.55E-05
GO:0016462	pyrophosphatase activity	43 of 819	3.54E-05
GO:0140101	catalytic activity, acting on a tRNA	14 of 115	3.74E-05
GO:0005515	protein binding	203 of 6605	4.34E-05
GO:0000049	tRNA binding	10 of 56	4.99E-05
GO:0017111	nucleoside-triphosphatase activity	41 of 778	4.99E-05
GO:0043168	anion binding	100 of 2696	4.99E-05
GO:0030621	U4 snRNA binding	5 of 6	5.14E-05

GO:0051082	unfolded protein binding	13 of 106	6.71E-05
GO:0008187	poly-pyrimidine tract binding	7 of 23	7.82E-05
GO:0003746	translation elongation factor activity	6 of 16	0.00014
GO:0017091	AU-rich element binding	7 of 26	0.00015
GO:0019900	kinase binding	36 of 678	0.00015
GO:0003697	single-stranded DNA binding	12 of 99	0.00016
GO:0008143	poly(A) binding	6 of 17	0.00017
GO:0019901	protein kinase binding	33 of 599	0.00017
GO:0004812	aminoacyl-tRNA ligase activity	8 of 43	0.00032
GO:0004386	helicase activity	14 of 147	0.00033
GO:0016887	ATPase activity	24 of 392	0.00055
GO:0036002	pre-mRNA binding	7 of 36	0.00077
GO:0042162	telomeric DNA binding	7 of 36	0.00077
GO:0043022	ribosome binding	8 of 50	0.00077
GO:0043024	ribosomal small subunit binding	5 of 14	0.0008
GO:0070181	small ribosomal subunit rRNA binding	4 of 8	0.0017
GO:0008266	poly(U) RNA binding	5 of 19	0.0025
GO:0031369	translation initiation factor binding	6 of 31	0.0025
GO:0030622	U4atac snRNA binding	3 of 3	0.0031
GO:0017070	U6 snRNA binding	4 of 11	0.0041
GO:0008026	ATP-dependent helicase activity	9 of 90	0.0059
GO:0035925	mRNA 3'-UTR AU-rich region binding	4 of 13	0.0065
GO:0042802	identical protein binding	63 of 1754	0.0069
GO:0034511	U3 snoRNA binding	3 of 5	0.0077
GO:0008092	cytoskeletal protein binding	37 of 882	0.0078

In the case of the small ribosomal subunit hits, a thorough inspection was performed of the identified cross-link sites. The heteroconjugate spectra leading to those protein identifications were manually evaluated. In total 20 constituent proteins of the small subunit could be confirmed. The majority of the spectral identifications (1,790 out of 2,058) and peptide identifications (81 of 95) could be validated. The confirmed hits are presented in Suppl. Table 3.6 and localized amino acid sites are visualized in Fig. 3.22. Similar to the observations with the RNA polymerase complex of *E. coli*, the cross-linked identifications based on the RPNxlSearch are in good agreement with the determination of manual evaluation. When considering highly abundant RNA-protein complexes the result from manual validation and FDR estimation are almost interchangeable.

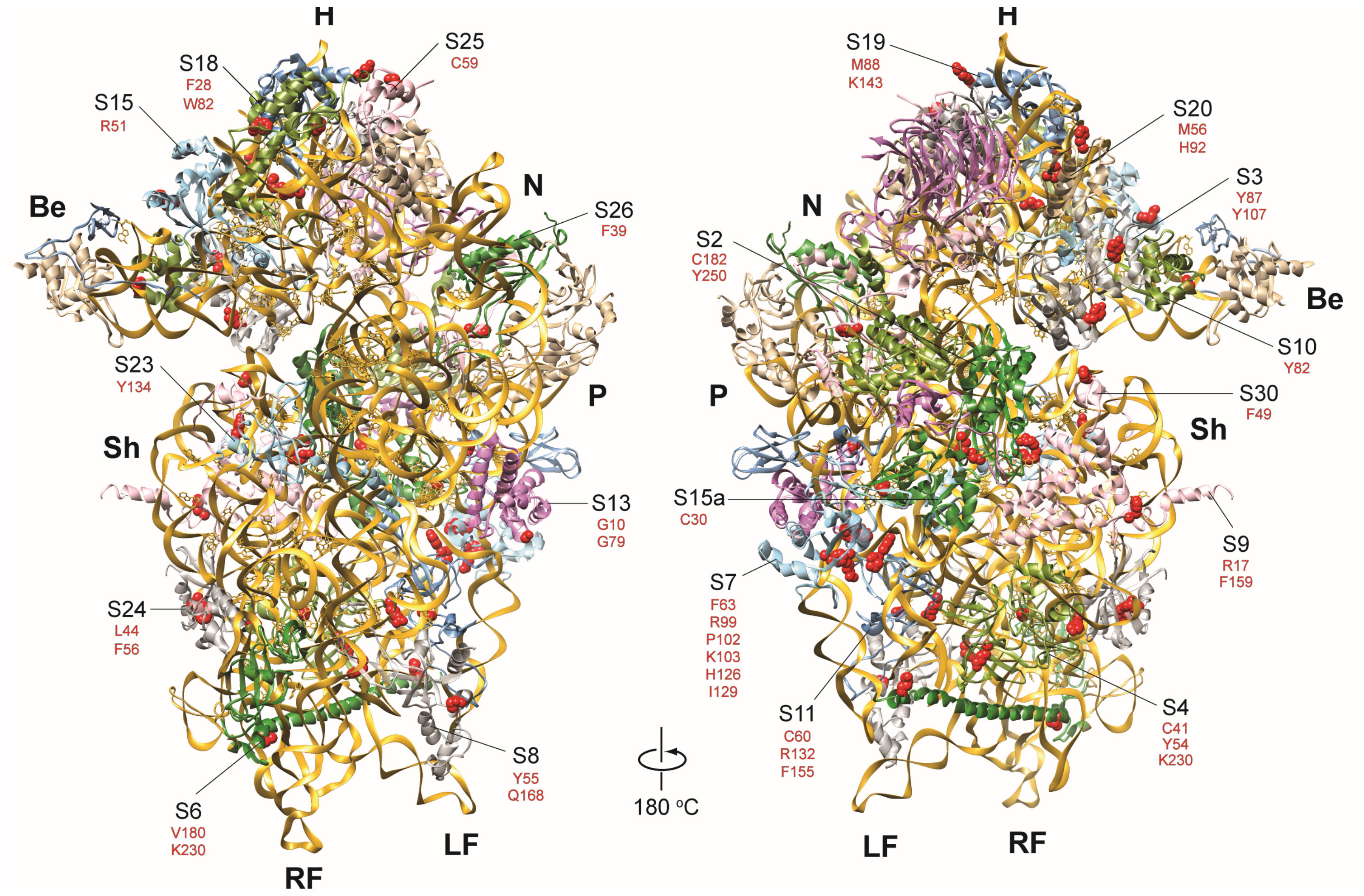


Figure 3.22 RNA contacts identified in the human small ribosomal subunit

Cross-link sites that could be localized to amino acid resolution were mapped onto cryo-EM structure [PDB ID 6QZP] [100]. RNA is colored in goldenrod, identified residues are marked in red and proteins are differentially colored. H-Head; Be-Beak; Sh-Shoulder; RF-Right foot; LF-Left foot; P-Platform, N-Neck

Next, a closer look was taken into novel and less studied protein hits. 63 proteins represented by 74 cross-linked peptides could be manually validated (Suppl. Table 3.6). For 35 of the protein hits, some evidence of interaction with RNA could be found in the UniProt database, the majority reported by quantitative mass spectrometry experiments [47,48].

Examples of protein hits not previously reported as RNA-binding in the UniProt database include: EKC/KEOPS complex subunit LAGE3 (LAGE3) involved in the formation of threonylcarbamoyl groups on adenosine in tRNA; Oxysterol-binding protein-related protein 6 (OSBL6) that regulates the transport of cholesterol; Dual specificity mitogen-activated protein kinase kinase 3 (MP2K3) that takes part in MAP signaling (Fig. 3.23 A); Catenin delta-1 (CTND1) which regulates the cell adhesion of C-, E- and N-cadherins; Cytoskeleton-associated protein 5 (CKAP5) and Echinoderm microtubule-associated protein-like 4 (EMAL4) that interact with microtubules; protein LTV1 homolog (LTV1) inferred to be involved in ribosome biogenesis (Fig. 3.23 B); Obg-like ATPase 1 (OLA1) involved in regulation of global protein phosphorylation in cancer cells [101]; YY1-associated protein 1 (YYAP1) that has transcription coregulation activity; Actin and Tubulin variants.

The list of manually confirmed heteroconjugates hits includes three enzymes involved in glycolysis, including Alpha-enolase and Fructose-bisphosphate aldolase (ENOA, ALDOA/ALDOC, GNPI2) and 4 structural constituents of the cytoskeleton, among which Actin and Desmoplakin (ACTB, DESP, TBB5, TBB6). A considerable number of cadherin-binding proteins was present (CTNND1, CKAP5, MAP4, SND1, EMAL4, OLA1, PDLI5, 1433S, EF1D, TRI25, ZCCHV). In addition to that, five proteins binding damaged DNA could be identified, including the XRCC5/XRCC6 dimer involved in non-homologous end joining (XRCC5, XRCC6, APEX1, APTX, HMGB1). Furthermore, four other proteins involved in DNA repair could be detected, including DNA-dependent protein kinase catalytic subunit and Ubiquitin-conjugating enzyme E2 N (PRKDC, UBE2N, TRIM25, PRP19).

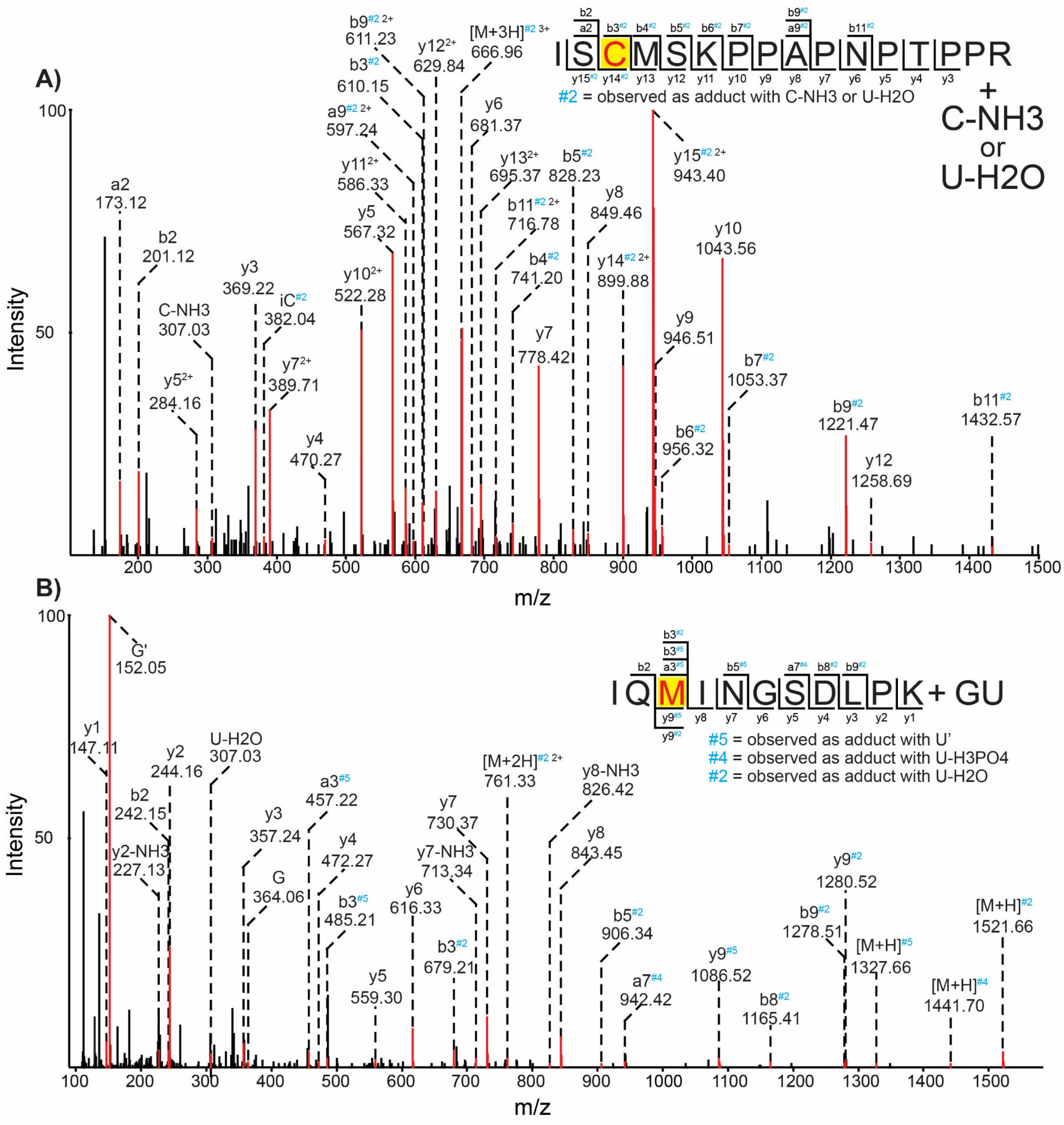


Figure 3.23 Examples of novel RNA-binding proteins identified in HeLa cells

A) MS/MS spectrum of MP2K3 cross-linked peptide $_{27}$IS[C]MSKPPAPNPTPPR$_{42}$ to pyrimidine derived adduct. The identification is supported by extensive a-, b- and y-ions and consistently observed proline effect. The cross-link site is localized at the $_{29}$C residue by b- and y-ions. B) MS/MS spectrum of LTV1 cross-linked peptide $_{402}$IQ[M]INGSDLPK$_{412}$ with GU. The assignment of the identification is backed up by broad a-, b- and y-ion series, as well as prominent proline effect in y2. $_{404}$M could be assigned as the cross-linked amino acid by shifted y9 and a3/b3 ions.

An important consideration when reviewing results obtained by the C18/TiO$_2$ workflow is the indiscriminative principle of enrichment based on the presence of a phosphate group. The first step of the protocol involves the digestion of RNA to nucleotides by RNAses. Therefore, both long RNA species and nucleotide binding proteins would be enriched and could be identified, making it impossible in some cases to distinguish the origin of the adduct. For example several ATP-binding proteins cross-linked to adenosine could be identified in the C18/TiO$_2$ protocol (RFC2, NUBP1). Interestingly, high quality spectra of purine water loss adducts A-H2O ($C_{10}H_{12}N_5O_6P$ 329.0525 Da) and G-H2O ($C_{10}H_{12}N_5O_7P$ 345.0474 Da) could be identified as a precursor adducts to a peptide of Endoplasmic reticulum chaperone BiP (Fig. 3.24 A).

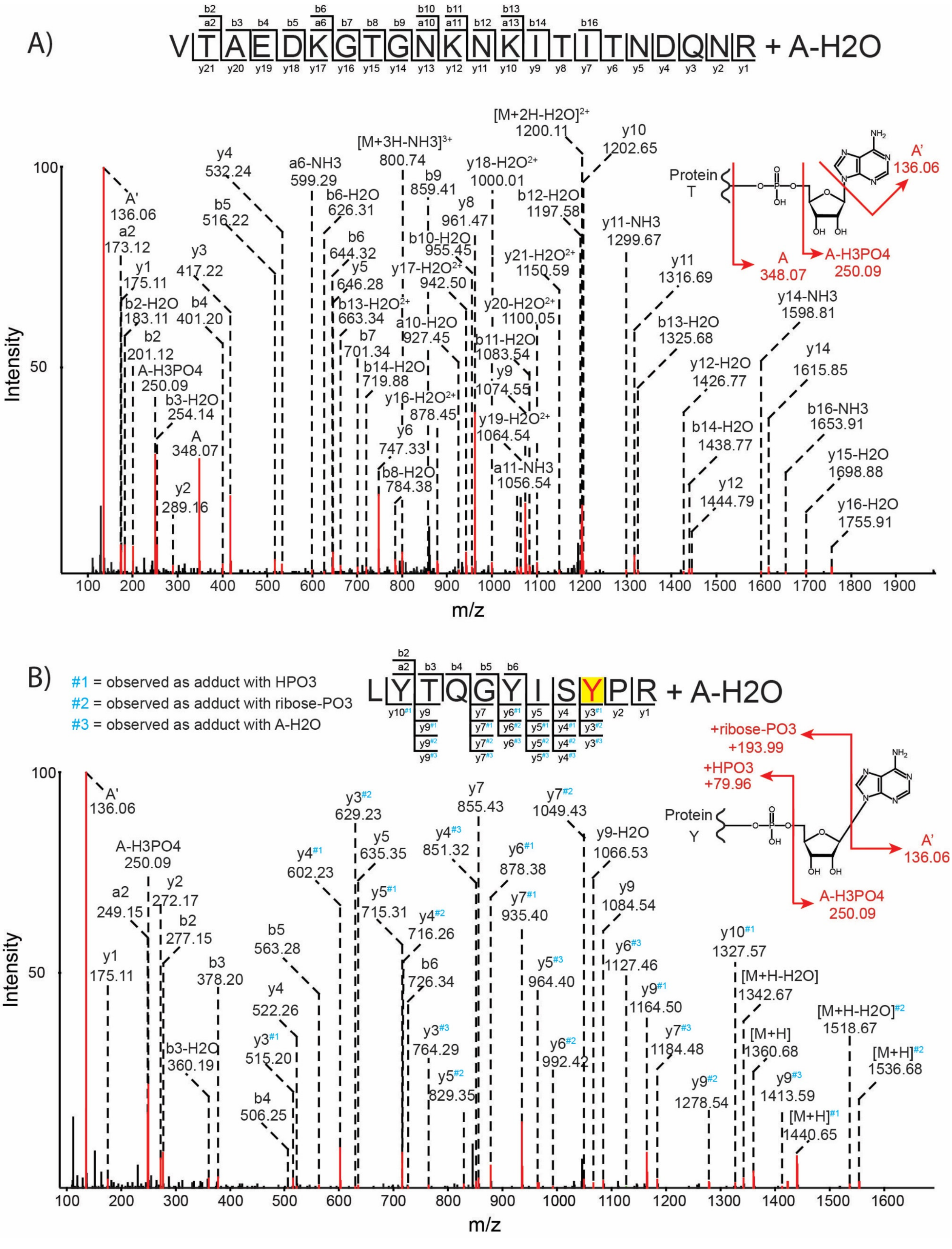


Figure 3.24 Identification of AMPylation by mass spectrometry

A) Fragment spectrum of AMPylated peptide $_{511}$VTAEDKGTGNKNKITITNDQNR$_{532}$ of Endoplasmic reticulum chaperone BiP. The assignment is supported by the presence of complete sequencing y-ion series and extensive a- and b-ion series. Strong signals of the adenine nucleobase (136.06) and the nucleoside monophosphate (348.07) and H3PO4 (250.09) net losses can be observed. B) Fragment spectrum of peptide

LYTQGYIS[Y]PR from DNA topoisomerase 3 AMPylated on tyrosine. The localization of the modification can be confirmed by prominent y-ion series shifted with neutral loss adducts. Strong marker ions of adenine nucleobase and nucleoside can be detected at 136.06 and 250.09 *m/z*.

Adenine does not contain an oxygen in the nucleobase, so the water loss must have occurred from the ribose-phosphate or peptide moiety of the heteroconjugate. In the fragment spectrum, strong marker ions of the nucleobase (A’), nucleoside (A-H3PO4), as well as the full nucleotide (A) could be identified. The cross-link with G-H2O behaved similarly with instense marker ions, inconsistent with previous observations of the same precursor adduct in the model protein systems. Signals of a cross-linked nucleobase are generally not of high intensity, as they require two fragmentation events to form. This implies the covalent bond was not formed at the nucleobase. The presence of a fully additive A marker ion from an A-H2O precursor adduct indicates that the water loss involved the peptide moiety. The identified peptide $_{511}$VTAEDKGTGNKNKITITNDQNR$_{532}$ contains $_{518}$T that has been demonstrated to be AMPylated, as part of a regulatory mechanism of the chaperone [102]. Therefore, the detected heteroconjugates were not a result of an UV-induced cross-linking reaction, but a post-translational modification (AMPylation and GMPylation) enriched by the C18/TiO_2.

A puzzling observation is the detection of an AMPylated peptide of DNA topoisomerase 3 by a silica-based purification of a sulfite-induced cross-links (Fig. 3.24 B). Analytes with single phosphate groups should not be enriched by the mechanism of this workflow, which is effective only for RNA species longer than 17 nucleotides. Topoisomerase 3β has been shown to bind RNA and is the major topoisomerase for mRNAs [103]. Most likely the protein was enriched as part of a cross-link complex with long RNA and happened to bear an AMPylation modification. Therefore, cross-links with single nucleotide adducts with water loss from any workflow should be carefully examined to determine if they were created by a UV-induced covalent bond involving the nucleobase or by enzymatic formation of phosphodiester bond with the phosphate group.

Another interesting finding is the presence of C-H2O (C9H12N3O7P 305.0413 Da) adducts in the dataset. In the reconstituted complexes cross-linking experiments, cytosine heteroconjugates were only observed with NH3 net loss. Several high-quality spectra of cross-links with C-H2O adducts were identified. In few of these cases, cysteine could be identified as the cross-linked residue. Exemplary fragment spectra of the newly observed adduct are shown in Fig. 3.25. Unlike the AMPylation spectra, the characteristic fragmentation pattern of a nucleotide can be seen – the glycosidic bond is preferentially cleaved and the most prominently observed adduct is the nucleobase with net loss of H2O. This indicates that the pyrimidine base is involved in the formation of the cross-link covalent bond. A search for the observed adduct mass was submitted to the Unimod database [104]. No other possible modification that would explain the added mass could be found.

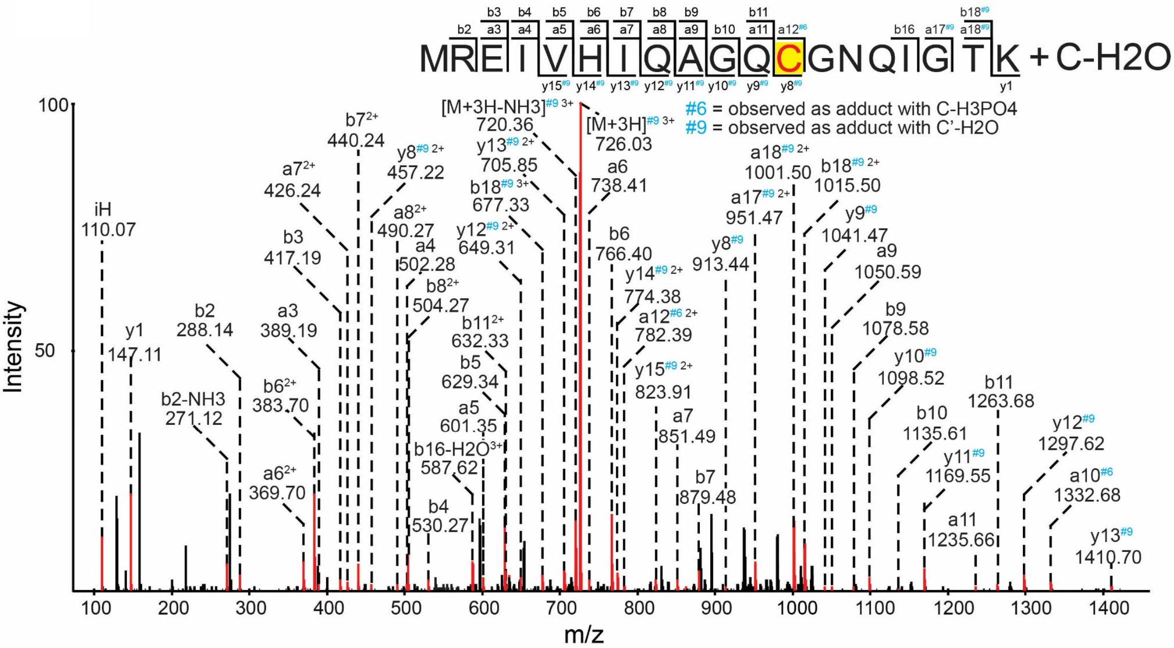


Figure 3.25 Exemplary spectra of cytosine cross-links with H2O net loss.

Fragment spectrum of Tubulin beta-6 chain peptide $_1$MREIVHIQAGQCGNQIGTK$_{19}$ with C-H2O. The identification is supported by extensive a-, b-, and y-ion series. Cysteine is the most likely cross-linked amino acid judging by the y8-y15 and a12 ions shifted with cytosine adduct.

Protein hits with an available structure with RNA are presented in Fig. 3.26 (YBOX1, SRP09, SRP14). A single cross-link site of Y-box-binding protein 1 could be identified in two peptide forms - $_{78}$NDTKEDV[F]VHQTAIK and $_{78}$NDTKEDV[F]VHQTAIKK$_{93}$. When mapped onto crystal structure [PDB ID 5YTT], $_{85}$F falls in close proximity with the co-crystalized RNA oligonucleotide. Both proteins comprising the signal recognition particle *Alu* element heterodimer could be identified and manually confirmed to be cross-linked to RNA. SRP09 was presented by two peptides leading to the same cross-linked amino acid residue – $_{42}$VTDDLV[C]LVYK$_{52}$ and $_{42}$VTDDLV[C]LVYKTDQAQDVK$_{60}$. For SRP14 a single peptide was detected - $_{22}$TSGSV[Y]ITLK$_{31}$. The localized sites $_{27}$Y and $_{48}$C fall into the RNA interface of the heterodimer when mapped onto crystal structure [PDB ID 4UYJ].

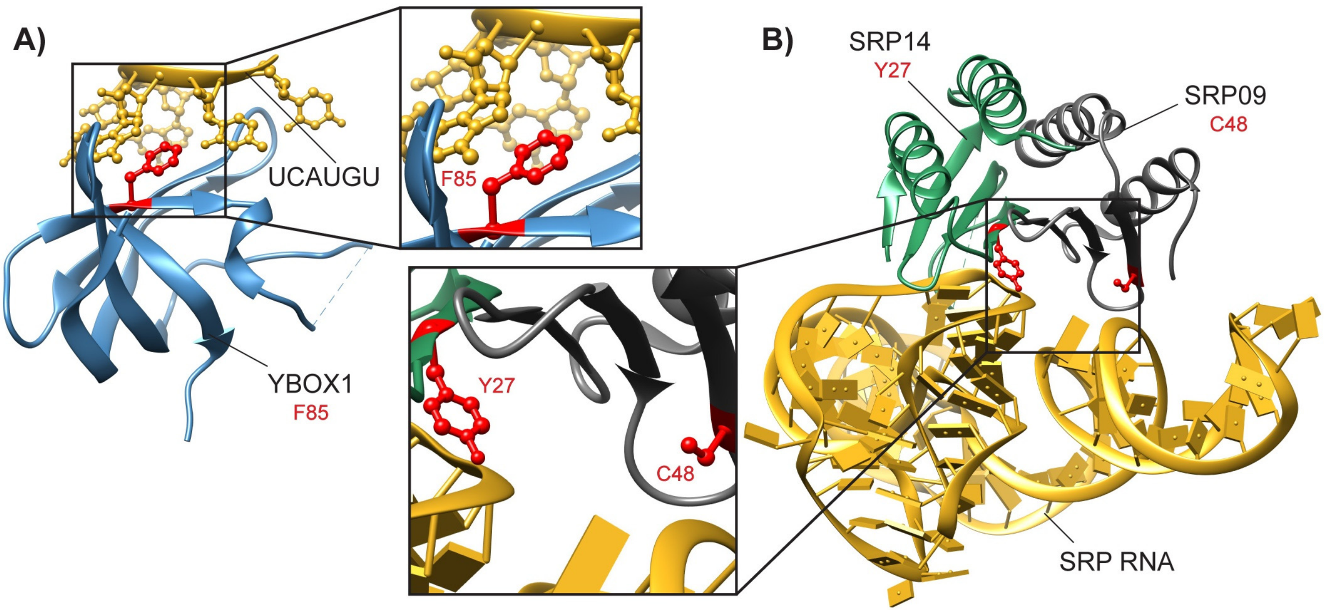


Figure 3.26 Identified heteroconjugates of HeLa proteins are in good agreement with available structures

Cross-linked sites of confirmed peptides were mapped onto available crystal and EM structures. Cross-linked residues are indicated in red. A) The identified contact between Y-box-binding protein 1 and RNA is mapped onto crystal structure [PDB ID 5YTT] [105]. B) Cross-links detected in the Signal recognition particle 9kDa and 14 kDa heterodimer are illustrated on crystal structure [PDB ID 4UYJ] [106].

Unfortunately, due to time and other constraints, no replicate experiments could be performed for the *E. coli* cells and HeLa cytoplasmic extracts in the frame of this work. Therefore, it is not possible to confidently compare the efficiency of the different workflows. Nevertheless, it is clear that both silica-based and strong anion exchange purification lead to considerable depletion of linear peptides and identification of a substantial number of cross-link sites. Both workflows enriched RNAs indiscriminately of their size and type, making them a valuable tool for the identification of various RNA-binding proteins.

3.5 Ionization and fragmentation of heteroconjugates with large RNA moiety

With the presented improvements in the analysis of peptide-RNA heteroconjugates, the determination of the cross-link sites on proteins by mass spectrometry has been dramatically improved. However, the sequence information obtained from the identification and annotation of cross-link spectra is highly peptide-centric. The current ESI-MS workflows limit the RNA moiety to one or a few nucleotides. Detection of species with larger RNA moieties is not possible due to the negative charge on the phosphate groups and the strong retention of nucleic acids on reversed-phase C18 material under acidic conditions (data not shown), that are in turn incompatible with the positive mode LC-MS analysis of peptides. In addition, the labile RNA is completely shattered at the energy levels required to obtain meaningful sequence information for peptides, leading to strong nucleobase marker ions signals and suppression of the other ions in the spectrum. Therefore, only compositional, but no sequencing information of the RNA adduct can be obtained, making it impossible to determine the identity of the interacting RNA or the position of that interaction on the nucleotide chain. Hence, development of a workflow for ionization of heteroconjugates with larger moiety and sequencing of the nucleotide chain was needed to complement the information obtained by the existing workflows.

In order to test the feasibility of such a workflow, a synthetic peptide-RNA standard was generated by copper-catalyzed click reaction between RNA oligonucleotide containing 5'-ethynyl-2'-deoxyuridine and synthetic peptide bearing 4-azidophenylalanined (Fig. 3.27).

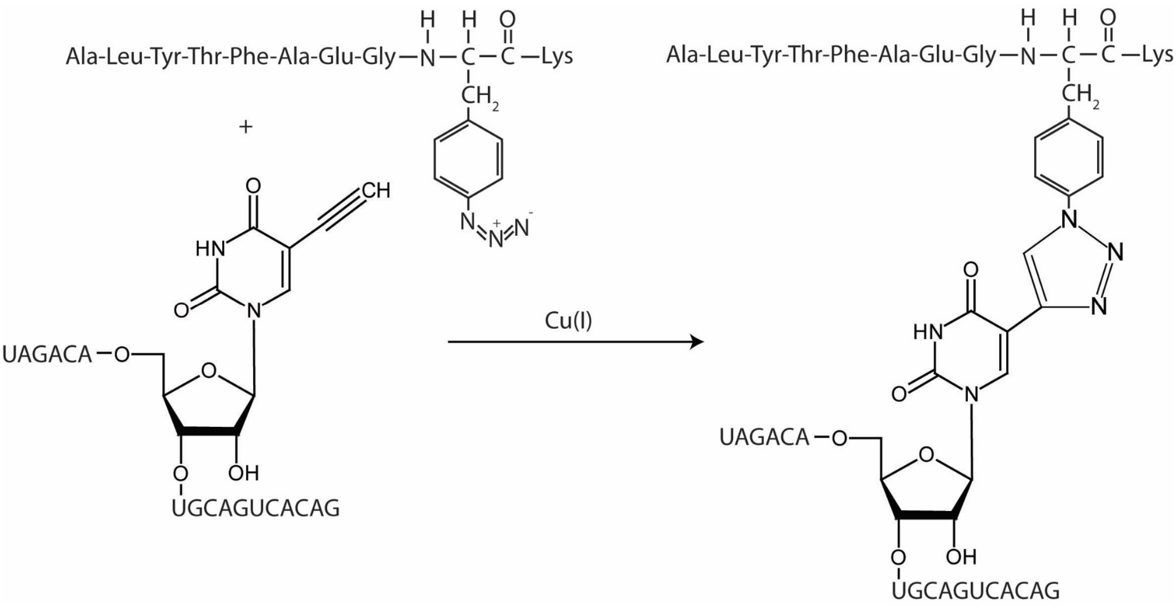

Ala-Leu-Tyr-Thr-Phe-Ala-Glu-Gly
Lys
+
UAGACA
UGCAGUCACAG
Cu(I)
Ala-Leu-Tyr-Thr-Phe-Ala-Glu-Gly
Lys
UAGACA
UGCAGUCACAG

Figure 3.27 Generation of synthetic peptide-RNA heteroconjugate

RNA oligonucleotide (5'-UAG ACA UUG CAG UCA CAG-3') containing 5'-ethynyl-2'-deoxyuridine and synthetic peptide (ALYTFAEGFK) bearing 4-azidophenylalanine were incubated in the presence copper(I)bromide.

Experiments for selection of the best ion pairing system and ionization source parameters were performed. Three ion pairing buffers were evaluated – TEAA pH 7, TEAB pH 8 and HFIP/TEA pH 8.6 buffer systems with ACN as mobile phase. The highest sensitivity was achieved with the HFIP/TEA ion pairing system (data not shown). Steel emitters could interact with the phosphodiester backbones, so distal coated silica emitters were selected instead. The HFIP/TEA system had high incidence of clogging for narrow emitters, thus, combination with needle of at least 30 µm inner diameter was required for successful continuous analysis of samples. Under these conditions, the standard peptide-RNA could be successfully chromatographically resolved and ionized in nanoelectrospray conditions. The heteroconjugate could be readily detected and fragmented. The RNA portion generating intense [a-B]-, c-, w- and y-ions that confidently revealed the sequence of the first six nucleotides from both the 5' and 3' ends (Fig. 3.28).

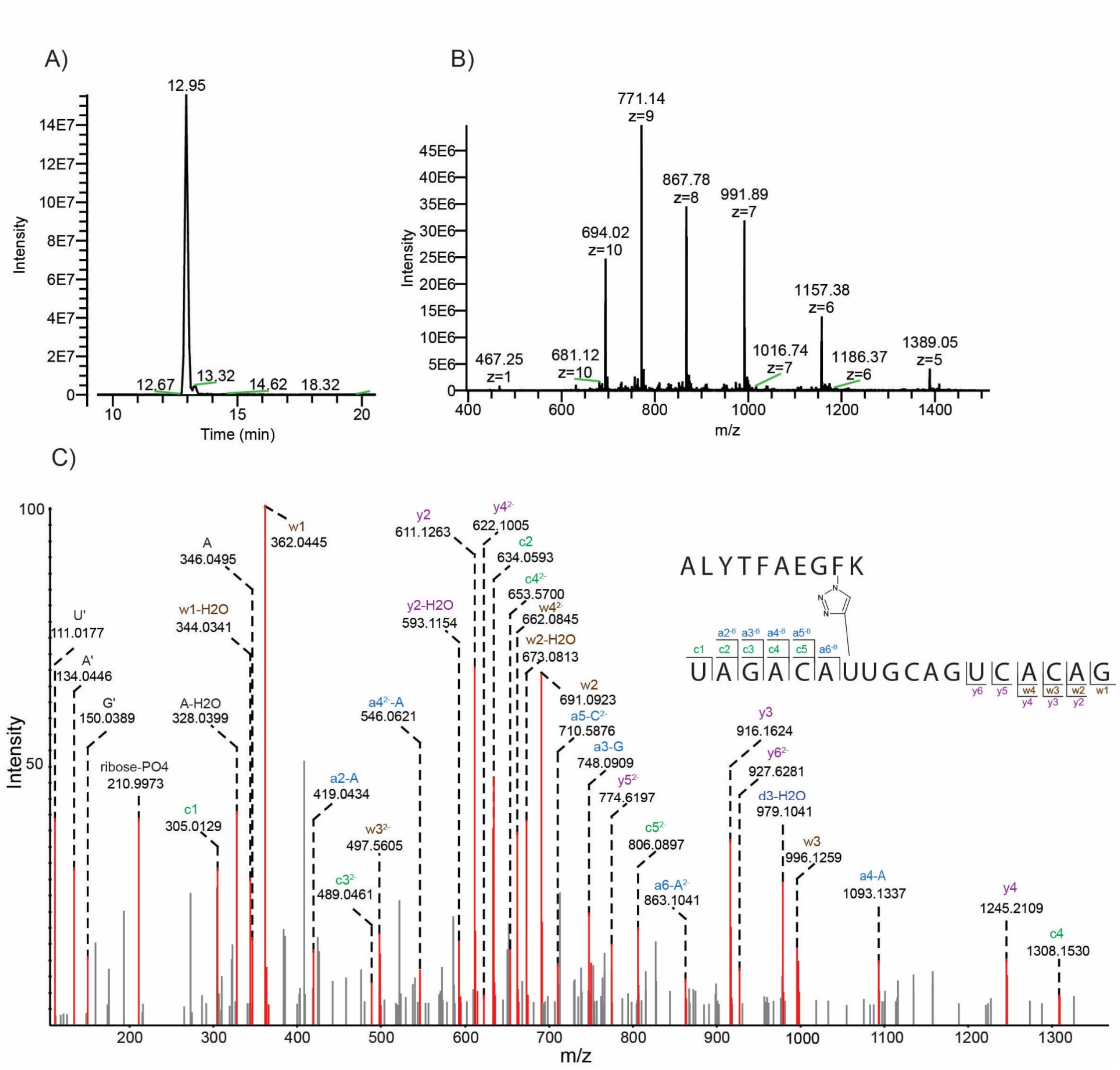


Figure 3.28 Sequencing of the long RNA moiety of heteroconjugates

Synthetic peptide-oligonucleotide standard was analyzed by nano-LC-ESI-MS/MS in negative mode. A) Elution profile of the heteroconjugates. B) Exemplary survey scan. C) Exemplary HCD fragment spectrum. The a-B-, c-, w- and y-ion series are annotated and marked in red.

This experiment confirmed that meaningful sequence information and identification of large RNA moieties of heteroconjugates is feasible. The 18-mer RNA moiety was digested with RNAse T1 to a smaller 6-nucleotide-long RNA fragment and analyzed under the same conditions. The small T1-derived oligonucleotide could be completely sequenced from both 5' and 3' ends, providing a highly confident localization information of the cross-link site by sequencing c-, w-, y-ions shifted with the peptide mass (Fig. 3.29). A peak corresponding to the cross-linked peptide with an uracil adduct can be seen at 1321.57 *m/z*. Isolation of the signal, and further fragmentation in an MS3 spectrum did not provide helpful peptide sequencing information (data not shown). These results demonstrate that the established nano-LC-MS/MS workflow can be successfully used in localizing the exact cross-link position on the nucleotide chain of T1 fragments up to 10 nucleotides in length.

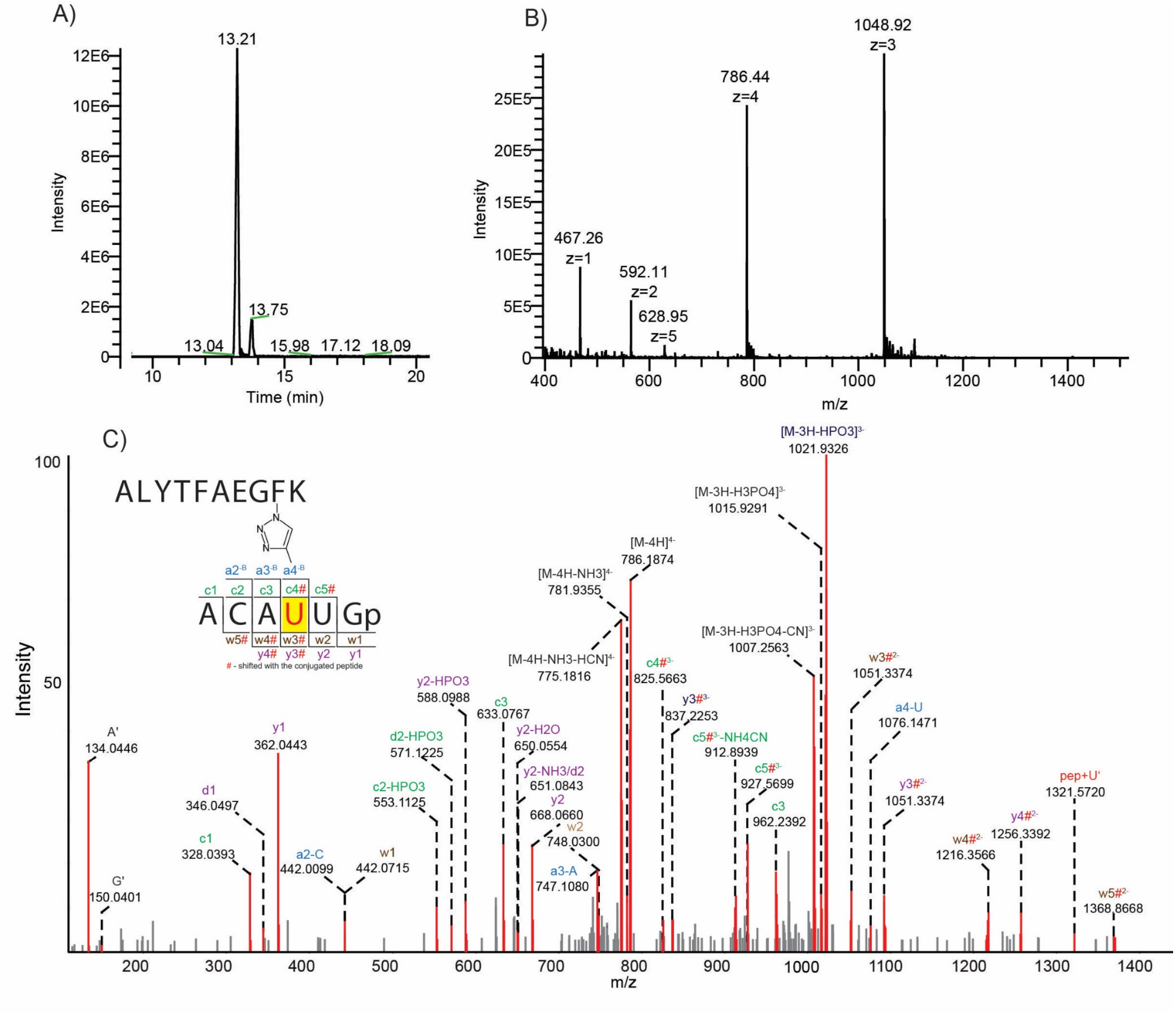


Figure 3.29 Localization of the cross-link site on T1 RNA fragments

RNAse T1 digest of peptide-oligonucleotide standard was analyzed by nano-LC-ESI-MS/MS in negative mode. A) Elution profile of the heteroconjugate. B) Exemplary survey scan. C) Exemplary beam-type collision-induced fragment spectrum. The [a-B]-, c-, w- and y-ion series are annotated and marked in red. Shifted ions with the mass of the entire peptide are indicated with #.

4. Discussion

4.1 Types of observed protein-RNA cross-links, fragment adducts and spectral characteristics

In recent years, mass spectrometric analysis of UV cross-linked protein-RNA complexes has emerged as a powerful tool for identification of novel RNA-binding proteins and elucidation of the amino acids involved in the interaction. The analysis of protein-RNA heteroconjugates is a particularly challenging task due to the low abundance of cross-linked species, the combinatorial complexity of possible adducts and the convoluted collision-induced fragmentation behavior. A prerequisite for identification of cross-links, is a clear understanding of which nucleotides are involved in the formation of the covalent bond and possible modifications they might harbor. Further insight into the fragmentation behavior of protein-RNA heteroconjugates is crucial for confident and accurate identification of cross-links, especially in complex samples.

In the first part of the results, the observations obtained from datasets of protein-RNA cross-link spectra acquired during the last decade in the Urlaub Research Group were summarized and the major elements of the fragment spectra were determined and categorized. With this information at hand, automated annotation of heteroconjugate spectra could be performed, greatly reducing the time and effort required for manual verification. In addition, the automated annotation allowed the screening of large numbers of fragment spectra for new adducts.

In order to gain new insights into the potential of mass spectrometry as a technology to elucidate protein-RNA interactions, a number of controlled experiments with model proteins and synthetic ribo(oligo)nucleotides were performed. In the following paragraphs, the fragmentation behavior of the different RNA nucleotide adducts that could be identified by mass spectrometry experiments are discussed in more detail.

4.1.1 Cross-links to pyrimidine bases

4.1.1.1 Cross-links to uracil

Uracil is the most reactive nucleotide observed in UV-induced protein-RNA cross-linking [40,107]. Previously performed mass spectrometric studies reported almost exclusively uracil-linked heteroconjugates [39]. The few exceptions were anecdotal in nature and lead to either questionable quality of the fragment spectra or lack of shifted ion information. The majority of identified cross-links to uracil were fully additive, implying that the identified products were formed by an addition cross-link reaction across the 5, 6-double bond or by reactions involving opening of the pyrimidine ring, such as the ones described in [45,108].

In general, fully additive uracil derived cross-links (U C9H13N2O9P 324.0359 Da) give rise to shifted ions with seven commonly observed adducts (Table 3.1). During fragmentation, the ion current is split among several of these neutral losses and often leads to low intensity shifted series. In addition, spectra of fully additive adducts have been observed to be susceptible to complete neutral loss of the adduct by breakage of the UV-generated covalent bond. This type of fragmentation event creates peptide fragments that surpass the cross-linking position, often complicating or preventing accurate localization of the cross-link site.

To a lesser extent, historically also -H2O net loss adducts (U-H2O C9H11N2O8P 306.053) of uracil have been reported. The water loss was presumed to account for the formation of cyclic phosphate on the ribose of the nucleotides, as a result of the enzymatic digestion or unspecific common loss in the gas phase during ionization [61]. Inconsistent with this presumption, the fragment spectrum of these cross-links contained almost exclusively shifted ions with what is described as the U'-H2O adduct (C4H2N2O 94.067 Da) and no noticeable examples of fully additive uracil base adducts (U' C4H4N2O2 112.0273 Da). This would strongly suggest the water loss occurred at the nucleobase or peptide moiety, most likely during the cross-linking reaction and not at the ribose-phosphate backbone due to the generation of cyclical phosphate. The mass of the reported adduct would also fit to C–NH3 that has not been considered in the previous mass spectrometry based studies. Such ammonia loss adducts were also previously described in the literature to occur in a transamination reaction of cytosine with lysine [80]. To shed some clarity into the situation, systematic cross-linking experiments with homopolyribonucleotides and isotopically labeled nucleotide monophosphate with model proteins were performed. Water loss adducts could be identified from both poly(U) and isotopically labeled nucleotide experiments. The identified cross-linked amino acid profile in these experiments resembled that generally observed of uracil, consisting of mostly aromatic and nucleophilic amino acids, including lysine.

Interestingly, a new type of uracil adduct that had a deficit of two hydrogen atoms from the fully additive chemical composition could be identified (U-2H C9H11N2O9P 322.0202 Da). The -2H loss was observed in all neutral loss adducts in the fragment spectra, indicating that the loss occurred from the nucleobase or peptide moiety, rather than the sugar phosphate backbone. The mass of the adduct fits to the products III or IV described by Varghese et al. [45]. The number of these cross-links, as both peptide sequences and spectral count, were substantially lower than fully additive and water loss adducts. The observed precursor of the -2H heteroconjugates were also of considerably lower intensity. In addition, the cross-link sites of the new adduct commonly coincide with the sites identified with the more prominent precursor adducts. Therefore, the identification of these cross-links generally did not contribute new contact information of the interaction between RNA and proteins. Unlike the -H2O net loss adducts, that are a subset of the fully additive uracil adducts, the -2H net loss is present in all fragmentation products of the nucleotide. There is no straightforward way to incorporate these adducts in the RNP^{xl} workflow concurrently with the common uracil adducts, but requires the employment of a separate search. As a consequence, inclusion of this minor adduct in global proteome-wide searches would substantially complicate standard investigative experiments and would be counterproductive. Nevertheless, the -2H heteroconjugates may offer an additional level of verification for many cross-link sites when analyzing simple protein complexes or focusing on a particular protein of interest. The combination of -2H net loss with the mass of commonly observed oxidation modification or the oxygen difference between guanine and adenine can easily lead to a misassignment on the precursor level to -H2O net loss and should be kept in mind when performing manual validation.

A puzzling result is the identification of a few high-quality spectra of heteroconjugates with U-H2O-2H precursor adducts, as well as a single hit with a C-NH3-2H adduct (Suppl. Table 3.2, 3.3). Upon collision-induced dissociation of these precursors, the fragment spectra contain high intensity, shifted ion series with a 92.00 Da mass shift that would fit to the U'-H2O-2H (C4N2O) adduct and displayed analogous behavior to the precursors with U-H2O/C-NH3 adducts. The similarity in the MS/MS spectra would suggest that the leaving water group involves the nucleobase oxygen and presumably originates by covalent bond formation at the 4th position of the pyrimidine ring, without affecting the 5,6-double bond. In this the case, the additional deficit of 2H atoms must result from a double bond or cyclical structure formation somewhere else in the involved nucleobase or amino acid residues by an unknown mechanism.

Adducts with -2H net loss generally eluted later than their fully additive counterpart during C18 reversed-phase chromatographic separation. That behavior is expected with respect to the presumed formation mechanism involving restoration of the 5,6-double bond and the aromatic character of the pyrimidine base. Therefore, although possible, it is unlikely to observe co-elution, overlapping and distortion of the isotope envelopes of -2H and fully additive adducts.

4.1.1.2 Cross-links to cytosine

A substantial number of cross-links could be identified within the experiment with poly(C) and the isotopically labeled cytidine monophosphate ribonucleotides. All of the observed precursor adducts had an ammonia loss, predominantly presented by C-NH3 RNA adducts (C9H11N2O8P 306.053 Da). They are indistinguishable by mass and fragmentation behavior from the U-H2O adduct observed with uracil, generating mostly C'-NH3 neutral loss adducts in the fragments spectrum (C4H2N2O 94.0167 Da). The great majority of the detected heteroconjugates are linked to the amino acid lysine. This represents a difference with the U-H2O adducts observed in the uridine experiments that showed a range of cross-linked amino acids similar to the fully additive adduct, without clear amino acid preference to a single residue. Therefore, some of the cross-links observed and reported in the past as U-H2O adducts to lysine were actually formed with C-NH3 or a co-eluting mixture of their isobaric products. In the ^{15}N-labeled cytidine monophosphate samples, the observed nominal masses of the 306 and 94 Da adducts (C-NH3, C'-NH3) in the unlabeled nucleotide experiments are increased by 2 Da to 308/96 Da that correspond to ^{15N}C-^{15}NH3 and ^{15N}C'-^{15}NH3. Similarly, in the ^{13}C^{15}N samples, an increase to 317 and 100 Da is observed, that would fit the chemical compositions of ^{13C15N}C-^{15}NH3 and ^{13C15N}C'-^{15}NH3. Therefore, the ammonia leaving group contains a labeled nitrogen. This clearly confirms that the ammonia loss originates from the 4-NH2 position of the pyrimidine and not from the lysine residue of the peptide moiety.

Interestingly, few high-quality spectra of cross-links with RNA adduct with mass 305.04 Da could be identified in the samples of *E. coli* cells and HeLa cytoplasm. This mass corresponds to a cytidine monophosphate with water loss (C-H2O). Upon fragmentation, a neutral loss adduct of 93.03 Da could be observed (Figure 3.25). This would match the mass of a shift C'-H2O (C4H1N2O), indicating that the water loss did not occur from the sugar moiety. No confident hits with this shift have been identified in the experiments with the model proteins or noted in the previous studies. Moreover, there is no heavy isotope confirmation for the origin and validity of the observed adduct. At the same time, the observed predominant fragmentation of the N-glycosidic bond strongly supports that it is indeed an adduct of RNA origin, involving a covalent bond between the peptide and the cytosine base. It is possible that this adduct is the result of a particular photochemical mechanism or space orientation of the amino acid residues in the cross-linked protein that was not present in the simple model systems and mixtures analyzed until now. Additional investigations are required to understand the nature of this type of adduct, as well as its usefulness for the elucidations of protein-RNA contacts.

Besides UV irradiation, cytosine cross-links could be successfully generated using either bisulfite solution or freshly dissolved metabisulfite at physiological pH and temperature. All observed heteroconjugates were formed by a transamination reaction of lysine with cytosine and are undistinguishable by fragmentation behavior from the UV-generated adduct (Fig. 3.9). The number of hits and their intensity, observed under the tested conditions, is well below the yield generated by UV irradiation.

The reaction between pyrimidine bases and sulfite was first investigated by Shapiro et al. [109], who demonstrated that uracil and cytosine can react with $NaHSO_3$ through chemical reaction addition to the 5,6-double bond. The saturation reaction with cytosine is more efficient at acidic pH and preferentially involves the protonated pyrimidine bases. If a primary amine group is in proximity to a sulfonated cytosine, a transamination reaction occurs. If no such reagent is present, deamination of cytosine to uracil by water is observed [110]. Substitutions at the 5th position of cytosine undergo through the deamination reaction much more slowly. This fact has been utilized in the mapping of 5-methylcytosines in DNA by genomic sequencing. Upon bisulfite addition, sulfonated cytosines can readily deaminate to uracils, while 5-methylcytosines would remain unaltered. After the reaction is completed, desulfonation can be promoted by high pH [111] and the bisulfite converted DNA is amplified by PCR, where the deaminated cytosines would be reaplaced by thymines [112].

The transamination reaction, whenever a suitable amine is present, has been found to be promoted by neutral pH [109]. Turchinsky et al. could demonstrate protein-RNA cross-linking mediated by bisulfite and later could identify an N4-substituted product of cytosine with lysine [80,110]. The same product could be identified by both irradiation with UV light at 254 nm and bisulfite conversion of the MS2 bacteriophage [80]. The generation of the transamination product due to UV irradiation is linked to the efficient formation of photohydrate, saturating the 5,6-double bond, similarly to the action of bisulfite, leading to the activation of an electrophilic center at the 4th position of cytosine [113].

The bisulfite-mediated cross-linking of lysine and cytosine has several advantages over UV irradiation of protein-RNA samples. The reaction is highly specific and produces adducts with a single prominent neutral loss generated by the fragmentation of the N-glycosidic bond. If the RNA adduct can be reduced to a single nucleotide during the sample preparation procedure, these adducts can be regarded as a simple modification in proteomics search engines, such as MaxQuant. This allows for the utilization of highly optimized and straightforward algorithms resulting in robust FDR controlled results that do not necessitate further manual verification. In addition, the complexity of the analytes in the sample is dramatically reduced, omitting a number of known and obscure products that stem from UV irradiation induced radical mechanisms. Drawbacks of bisulfite-mediated cross-linking include the low efficiency of cross-linking to cytosine and lysine under the milder physiological conditions used in this study. Furthermore, bisulfite is not membrane permeable, which exclude its use for *in vivo* applications. The high specificity also limits the use of bisulfite to protein-RNA contacts where lysines are in appropriate spatial positions to cytosine bases. Nevertheless, the bisulfite cross-linking reaction has unutilized potential and can be further optimized to become a highly specific tool for analysis of protein-RNA contacts.

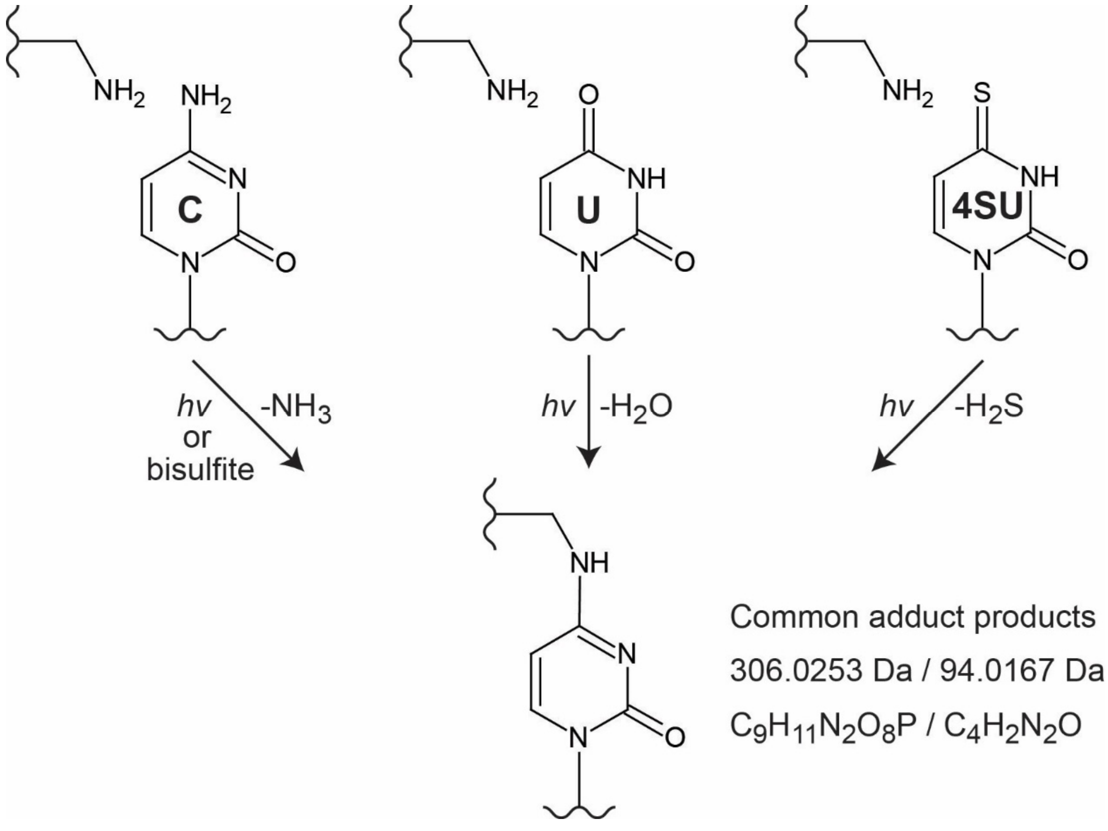


Figure 4.1 UV promoted cross-linking of RNA pyrimidine bases can lead to the formation of the same photoproduct

Cytosine, uracil and 4-thiouracil can form identical 4-substituted adducts after the loss of NH3, H2O or H2S. These are usually presented as additional mass of 306.0253 Da to the peptides in the precursor scan and characteristic high intensity neutral loss series adduct of 94.0167 Da in the fragment spectrum.

While the sulfonation of cytosine and subsequent reactions are well documented, little is known of the sulfonation effects on the other nucleobases [114]. The identical composition and fragmentation behavior of the UV-generated U-H2O adducts suggest that it is the same product as generated by cytosine transamination. Most likely, it is formed analogously to cytosine by a dark reaction after a photohydration event at the 5,6-double bond of uracil. Similarly to cytosine, uracil can also efficiently react with bisulfite [115], forming a sulfonated product. Therefore, it would be interesting to investigate whether sulfonation of the uracil base by bisulfite would also promote cross-linking reaction with proteins.

Noteworthy, unlike uracil, no fully additive adduct of cytosine could be detected. It has been proposed that addition of an amino acid across the 5,6-double bond, analogously to photohydration, would result in unstable adduct that would readily deaminate to uracil in aqueous solution [116]. Indeed, a couple of spectra of fully additive U adducts could be detected in the poly(C) experiments. It is not possible to exclude that the transformation to uracil occurred before the cross-linking reaction or that a contamination of the synthetic oligonucleotide was present. In any case, the occurrence of this phenomenon seems to be negligible compared to the large number of C-NH3 cross-links observed.

An interesting observation is the detection of precursor adducts comprising of only the uracil nucleobase (Suppl. Table. 3.2). The precursor adducts is likely degraded after the enrichment process or during ionization of the sample, leading to a loss of the ribose-phosphate moiety. This creates a simple heteroconjugate that behaves like a peptide with a small modification. Consistent conversion of all precursor adducts to a simple nucleobase modification would promote higher sensitivity and simple fragment spectrum that can be analyzed with any proteomics search engine. Preliminary attempts to induce deglycosylation of peptide-RNA heteroconjugates with in-source fragmentation or incubation with an organic acid and heating were unsuccessful and resulted in overall degradation of the sample (data not shown).

An interesting photoreaction that would lead to full addition of pyrimidine bases has been reported to involve opening of the pyrimidine ring. Cytosine and uracil participate in ring opening reactions with alkylamines that upon heating or acidification lead to N1 alkyl-substituted pyrimidines [108]. The ring opening reaction of uridine has been demonstrated to involve photoaddition of water to 5,6-double bond, followed by a dark exchange reaction with primary amines. Heating of this product produces the parent nucleosides [117]. Similarly, the irradiation of thymidine with primary amines at 0 °C results in efficient formation of open ring adducts, that slowly convert to N-substituted thymine under loss of deoxyribose. The formation of a photoexchange pyrimidine adduct with the ε-amino group lysine has been also confirmed in chromatin samples from calf thymus [118]. The described thermal reaction results in freeing the cross-linked proteins with lysine residues modified by a thymine base (Fig. 4.2). Such adducts would not have been detected by current workflows that rely on purification after extended incubation at room or higher temperature. Therefore, such a reaction could have untapped potential for studying thymine to lysine contacts in DNA as a simple modification. In the case of RNA, uridine has been reported to form only transient cross-links that are reversed upon heating [117]. At the same time, the photoexchange formation of N1-alkylated uracil bases with primary amines might indicate that an analogous reaction is also possible for uracil bases in RNA. Further investigations are required to identify the possibility of using such adducts to study protein-RNA/DNA contacts.

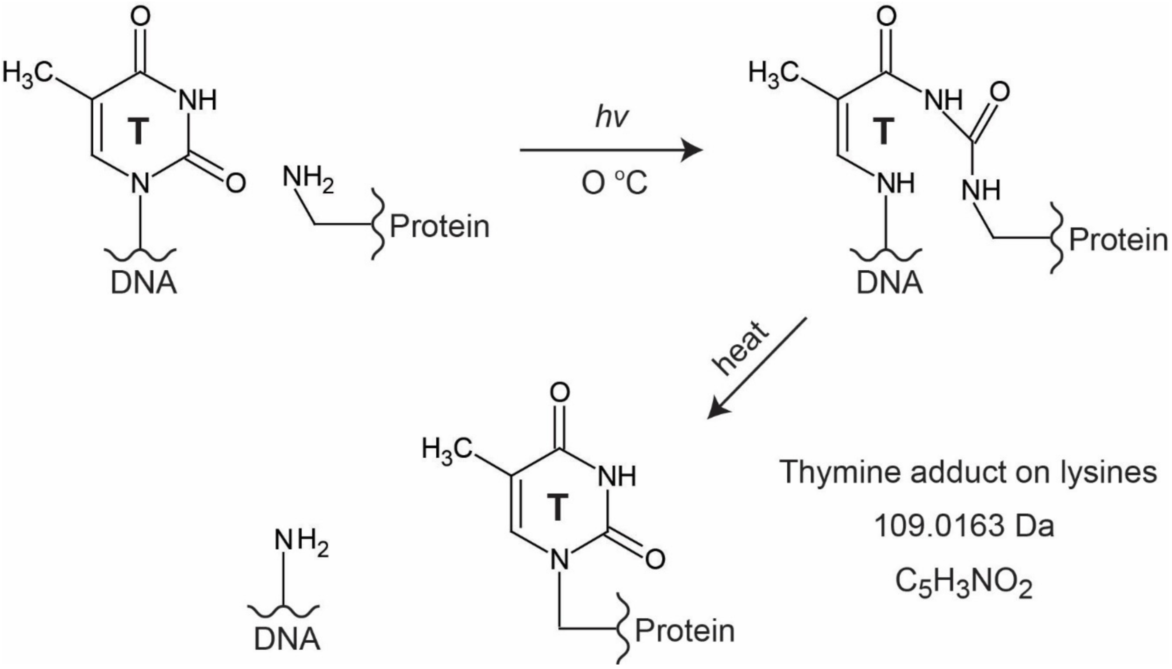


Figure 4.2 Photoexchange reaction of thymine and lysine

Irradiation of protein-DNA complexes at low temperatures leads to a ring opening reaction of thymine with lysine, possibly through a hydrate intermediate. Heating the reaction induces pyrimidine modification on the lysines of the protein. Reaction drawn as described by [119].

4.1.1.3 Cross-links to 4-thiouracil

The incorporation of photoreactive nucleotide analogs can provide significant increase in detectable protein-RNA cross-links *in vitro* and *in vivo*. The introduction of chromophores also allows for excitation at longer wavelengths, which leads to reduction of the UV-induced damage and unwanted photochemical reactions in the system. Commonly used chromophores include halopyrimidines (e.g. 5-bromouracil, 5-iodocytosine), azide-labeled nucleotides and thioribonucleotides (e.g. 4-thiouracil, 6-thiogunanine) [120]. Substitution of the ketone group oxygen with sulfur in thioribonucleotides ensures similar base pairing properties and generates minimal steric distortions, as the difference of the Van der Waals radii between the two elements is only 0.45 Å. In aqueous solutions, the uracil analog 4-thiouracil is predominantly in 2-keto-thione form, shifting the absorption peak to the near UV range (λ ≈ 330) [121]. In this study, 4-thiouridine has been utilized to label *E. coli* cells *in vivo*, resulting in random incorporation of the 4-thiouracil base in all RNA species of the organism. The near UV-induced (365 nm) photoreaction between proteins and 4-thiouracil labeled RNA was studied previously in our laboratory and the primary observed peptide photoproduct was generated under loss of H2S from the thiol group on the 4th position of the pyrimidine (4SU-H2S C9H11N2O8P 306.053 Da) [94]. Fragmentation of the precursor creates intense amino acid series shifted by 94 Da that correspond to fragmentation of the N-glycosidic bond (C4H2N2O 94.0167 Da). The behavior of the observed cross-linking product is identical to the observed adducts generated from uracil and cytosine by water and ammonia loss, respectively (Fig. 4.1). Therefore, similarly to bisulfite-induced cross-linking, if RNA digestion can be optimized down to single nucleotides, it is possible to identify 4-thiouracil generated cross-links with the same simple modification and proteomics-based search engine. However, unlike bisulfite promoted cross-linking, 4-thiouridine reacts with a multitude of amino acids [39] that would complicate the search and localization of the cross-link by such a strategy. A noticeable disadvantage of using a 4-thiouridine based strategy is the need for preceding incorporation, which limits the method to cell cultures and synthetic oligonucleotide systems. In addition, little is known about the physiological effect of the labeling and the toxicity it exerts on the cells, which may cause considerable changes in the protein-RNA interactome and detection of protein-RNA contacts uncharacteristic for unstressed systems.

4.1.1.4 Cross-links to DNA bases

The main focus of this study is the detection of protein-RNA contacts through the identification of peptide-RNA heteroconjugates. All characterized UV-induced cross-links with RNA were formed by reactions involving the nucleobases. Therefore, similar reactions can be expected to occur with the pyrimidine and purine DNA nucleobases. Several high-quality cross-link spectra of peptide-DNA heteroconjugates could be identified during investigation of the mtRNAP/TEFM elongation complex assembled with an RNA/DNA scaffold (Table 3.2). Most of them involved transamination reactions of cytosine with lysine as observed within RNA, generating a deoxyribose version of the commonly observed precursor adduct (dC-NH3 C9H11N2O7P 290.0304 Da). In addition, a fully additive cross-link of thymine could be identified, cross-linked to the $_{412}$VC$_{413}$ region of mtRNAP, likely linked to the side chain of cysteine. The major neutral loss product of the heteroconjugate in the MS/MS spectra is thymine base (T’ C5H6N2O2 126.0429 Da). The photoaddition of cysteine to polythymidylic acid was first demonstrated by Smith and Meun [122]. Later several cysteine products with thymine, in great part analogous to the products observed with uracil, could be isolated and characterized [123,124]. The identified fully additive cross-link could correspond to products observed by Varghese [124], involving a photoaddition to the 5,6-double bond.

4.1.2 Cross-links to purine bases

The number of detected cross-links with purine bases was substantially lower in comparison to pyrimidine bases. Several effects may contribute to this result: i) lower reactivity of the purine bases ii) unstable products that decompose in the process of biochemical enrichment and ionization iii) generation of unexpected mass adducts that were not included in the search.

Some insights into the efficiency of photoaddition of different nucleic bases with amino acids can be extracted from Shetlar et al. [107]: Homopolyribonucleotides of cytosine, guanine and adenine were irradiated in the presence of 19 amino acids (excluding proline), similarly to the investigation of polyuridilyc acid. The detection of photoaddition was accomplished by a fluorescamine assay, i.e. by reaction with the primary amine groups of the amino acids. The results indeed indicate a lower photoaddition reactivity of the purine nucleotides in comparison to the pyrimidine polyribonucleotides, and especially to uracil. Partial explanation for the observed reduced reactivity of purines might be found in their simple photodeactivation mechanism, leading directly to a ground state. On the other hand, pyrimidines possess much richer photodynamics that may lead to trapping in local energy minima, increasing the time spent in excited state for up to several picoseconds [125].

Purine nucleotides are less stable than pyrimidines under physiological conditions and more readily undergo spontaneous rupture of the N-glycosidic bond, that is believed to be a result of increased susceptibility to acid catalyzed deglycosylation [126]. In the standard C18/TiO_2 enrichment workflow, with which the majority of the heteroconjugates were detected, the cross-linked species are subjected to high concentrations of strong organic acid (5% v/v TFA) that may lead to deglycosilation and the inability to enrich the purine cross-linked peptides. In addition, the protein and RNA digestion steps require lengthy incubations at elevated temperatures. Therefore, utilization of TiO_2 protocols with lower acid concentrations or alternative workflows may lead to more frequent detection of purine-based heteroconjugates.

4.1.2.1 Cross-links to guanine

Despite the inherent instability of purines, a number of cross-links could be identified in the experiments with model proteins with poly(G) and the isotopically labeled guanosine monophosphate analogs. Precursor adducts of fully additive (G), ammonia net loss (G-NH3), water net loss (G-H2O), 2H net loss (G-2H) and guanine base (G’) could be identified. In several of the detected heteroconjugates, the cross-linked amino acid was identified as phenylalanine, tyrosine or lysine (Suppl. Table 3.2) that are in good agreement with the obtained photoreactivity results with single amino acids [107]. In the case of fully additive cross-links, a multitude of neutral losses could be identified in the fragment spectrum, with guanine nucleobase adducts (G’) being the most prominent adduct. This type of profile strongly resembles the observation within the MS/MS spectra of fully additive uracil adducts. In cases where -H2O or -NH3 loss were observed in the precursor level, the fragment spectrum contained almost exclusively shifts of the nucleobase with the corresponding loss (G’-H2O, G’-NH3), analogously to the behavior of U-H2O/C-NH3 adducts. One exception is a G-H2O adduct identified within a peptide of the endoplasmic reticulum chaperone BiP. However, judging by the fragmentation spectrum profile, this finding most likely corresponds to a GMPylation post translational modification and not to a UV-induced cross-linked adduct.

Little is known of the photochemical products of guanine and proteins. Steinmaus et al. investigated the photochemistry of the purine bases with isopropanol and ethanol as threonine and serine analogs and could demonstrate that the predominant product is substitution of a hydrogen atom at C8 position of the purines with the alcohol [127]. In a photo-cross-linking study, Rohrbach and Bodley could demonstrate the photoaddition of guanine to a reactive cysteine residue in elongation factor G [128]. Xu et al. could demonstrate addition of lysine to C5 and C8 position of guanine [129]. These reactions can explain the generation of some of the observed G-2H adducts (Suppl. Table 3.2), however, in the literature there is no previous report or proposed mechanism for the detected fully additive and water/ammonia loss cross-links.

4.1.2.2 Cross-links to adenine

No adenine cross-links could be identified with poly(A) or adenosine monophosphate nucleotide analogs in C18/TiO_2 based experiments. Employing an alternative experiment involving cross-linking GAPDH with *E. coli* derived RNA and silica purification based workflow resulted in the identification of several high-quality spectra of adenine cross-links. The precursor adducts were formed with A-NH3 and upon fragmentation formed prominent neutral loss series with A’-NH3. Investigation of HeLa cytoplasm, revealed a fully additive version of adenosine monophosphate (Suppl. Table 3.6). In addition, phosphodiester linked heteroconjugates (A-H2O), resulting from enzymatic AMPylation were detected (e.g. Endoplasmic reticulum chaperone BiP, DNA topoisomerase 3).

4.2 Identification of heteroconjugates with large RNA moieties

Mass spectrometry is the method of choice for identification of proteins and obtaining sequencing information for peptides. When it comes to the analysis of nucleic acids, it is less adept in competing with the sensitivity and scope provided by next generation sequencing (NGS) techniques. Employment of current mass spectrometry workflows is only advantageous for identification of modified bases in RNA and DNA or nucleic acid based therapeutics [130]. Alike, mass spectrometry-based workflows have been utilized to analyze protein-RNA heteroconjugates in a peptide-centric manner, identifying the regions of the protein contacting the nucleic acid. On the other hand, NGS-based workflows were established to provide information in RNA-centric metric manner, elucidating where these interactions occur on the ribonucleotide chain.

The ability of mass spectrometry to simultaneously provide information about the peptide and RNA moiety of cross-links has been explored previously. Urlaub et al. demonstrated that MALDI-MS could be utilized to identify the contact sites of cross-linked *E. coli* 30S ribosomal complexes by sequencing the oligoribonucleotide moiety of the heteroconjugates [131,132]. The strategy was further developed into a MALDI-MS workflow that could provide both peptide and RNA sequencing information, elucidating the exact contact sites in spliceosomal protein-RNA complexes [133,134].

The use of mass spectrometry for identification of cross-linked species has naturally shifted towards electrospray ionization techniques that allow the coupling to liquid chromatography and the thorough analysis of complex mixtures. In this way, the high-throughput examination of large number of analytes present in the cross-linking sample was enabled, simplifying the acquisition process and ultimately leading to the detection of lower abundant cross-links. However, the current LC-ESI-MS workflows for analysis of cross-linked protein-RNA complexes involve extensive digestion of the ribonucleotide chain, that is necessary to obtain exhaustive sequencing information of the involved peptide. In this way, the acquired positional information of where cross-links occur within the protein sequence is often precise enough to pinpoint the cross-linked amino acid. Identification of cross-linked amino acids at this resolution in peptide-RNA heteroconjugates with larger RNA moieties is incompatible with the established protocols that utilize positive mode ESI-MS. Increasing the number of nucleotides in the RNA adduct reduces the ionization efficiency due to the negative charges of the sugar phosphate backbone. Additionally, the energies utilized for efficient sequencing of the peptide moiety readily fragment the N-glycosidic bond of the nucleotides and lead to very strong nucleobase marker ions that suppress the other signals in the spectrum (Suppl. Fig. 3.2). As a result, the spectra provide only compositional, but no sequencing information for the involved RNA moiety. The standard use of C18 reversed-phase chromatography for analysis in positive mode is also incompatible with the analysis of larger RNA moieties due to the high affinity of the ribonucleotide chains to the matrix under acidic conditions. These factors point towards negative mode for the analysis of heteroconjugates with large RNA moiety, which was successfully employed in this study

An 18 nucleotide long peptide-RNA standard was successfully synthesized and utilized to evaluate appropriate ionization conditions and ion pairing systems. The standard heteroconjugate could be detected and fragmented, providing sequencing information of the 6 nucleotides from both the 5' and 3' ends (Fig. 3.28). Reduction of the 18-mer to 6-mer with RNAse T1 allowed for complete sequencing of the RNA moiety and localization of the cross-linked site (Fig. 3.29).

The presented results confirm the feasibility of the LC-ESI-MS approach to localize the cross-link site on the nucleotide chain. This provides complementary information to the standard protocols and allows matching the positional cross-link information of both interaction partners. Advantageous to next generation sequencing techniques, such an approach provides direct identification evidence of the contact site for both the peptide and RNA moiety of the heteroconjugate. At the same time, unlike the tremendously sensitive NGS methods, mass spectrometry workflows usually require femtomole amounts of analytes in order to generate a high quality fragment spectrum. This may present a serious challenge for the inefficient UV cross-linking reaction.

Currently, no software tool can match peptide-RNA heteroconjugates by RNA sequencing fragmentation. Identification of cross-links from simple systems could be aided by submission of the expected cross-linked peptide as a post-transcriptional modification in dedicated software, such as RoboOligo [135]. The development of a dedicated workflow reciprocal to RNP^{xl} would be needed to perform analysis in a systematic way as well as investigation of complex systems.

4.3 Data analysis of cross-links

The RNP^{xl} workflow and recent optimizations by the introduction of the dedicated RNPxlSearch search engine and automated annotation allow for fast exploration of a multitude of possible RNA adducts and their corresponding fragmentation mechanism over highly complex datasets. The incorporation of the identified spectral trends and elements into different scoring functions allows the calculation of an inclusive combined score and the employment of an FDR estimation. Such a calculation is greatly needed, as the number of heteroconjugate candidate spectra increases drastically within complex samples, analyzed with modern state-of-the-art instruments, and makes individual, manual verification impossible. Manual validation criteria have been established through analysis of low complexity samples of substantial amount and are traditionally quite strict, resulting in high confidence results at the cost of increased false negative hits. It would be extremely challenging for a human expert to assess intuitively the search space explosion that comes with proteome wide searches. At the same time, a simple equation of subscores can greatly overestimate the importance of certain elements within inconsistent spectra.

Indeed, browsing through FDR filtered results, a number of very low quality spectra could be identified. The majority of these examples are remarkably noisy, in which the algorithm could match a great number of sequencing ions in low intensity signals, resulting in high score. Alternatively, many spectra did not qualify at 1% FDR, although they presented high quality exhaustive sequencing information. The identification provided by 1 % spectral FDR filtering was compared with manual validation in the *E. coli* and HeLa samples by analyzing selected complexes of well-known RNA-binding proteins, as well as a number of less abundant proteins or not known to interact with RNA, presented mostly by single hits. The majority of the spectra in the RNA polymerase complex and small ribosomal subunit were of high quality and could be confirmed manually. Most of the spectra that did not qualify to manual validation criteria were of borderline passing quality. Similar observations were acquired when analyzing simple complex mixtures – the FDR controlled results almost completely overlap with the decision of manual validation. The situation for the rest of the evaluated spectra varied considerably – their quality ranged from high to very low. Partial explanation for these findings can be found in the low starting amount of heteroconjugates that are generated in complex mixtures and the heavy competition they encounter. The combination of these limiting factors would lead to low intensity signals, close to the limit of detection of the instrument. Naturally, fragment spectra from such analytes would generate poor sequencing information that cannot satisfy the stringent manual validation criteria. Alternatively, these hits are false positive identifications of different analytes that happen to match the precursor mass and limited amount of the fragment peaks. Without dedicated heavy isotope or spike-in experiments, it is impossible to evaluate the true false discovery rate at different levels and determine the contribution of each of those two alternatives.

The FDR filtering is spectral-based and the majority of high quality spectra lead to the same few peptides in high abundant protein-RNA complexes, such as ribosome and cold-shock proteins. At the same time, a lower fraction of the spectral identifications leads to a large section of the peptide and protein assignments. Therefore, even if the false discovery rate is controlled at 1% for the spectral level, the low number of hits that comprise this 1% could lead to much higher percentage of peptide and protein false positives. Standard and robust calculations of peptide and protein level FDR could not be employed, due to the relatively low number of assignments. A dedicated strategy is required to combine the information from the various RNA adducts of each site in a sensible manner. Therefore, the current FDR estimation provides appropriate results to achieve a global overview of a sample, but should be treated with caution when it comes to individual identification sites, especially on novel RNA-binding proteins. A way to approach this issue is the presented two level verification strategy, combining FDR filtering with selective manual validation. First, the massive numbers of candidate spectra are filtered automatically by a simple decoy-based strategy, followed by manual validation of individual proteins of interest. Thus, decreasing the probability of reporting a false positive hit due to score or human bias. In the current state, automated analysis of simple protein-RNA complexes is feasible, but the combinatorial complexity of proteome-wide search space makes it impossible to confidently determine the protein-RNA interactions of cellular systems solely on the basis of the RNPxlSearch FDR estimation. The performance of the algorithm in complex systems appears to be even more reliant on the quality of the sample and the depletion of non-cross-linked species. Thus, improved isolation of peptide-RNA heteroconjugates by the more discerning purification workflows directly decreases the probability of reporting false positive hits. Additional improvements in the combined scoring would be required to achieve the final goal of data analysis workflow that is completely automated and provides reliable results irrelevant of the sample quality.

In recent years, several alternative strategies for data analysis of peptide-RNA heteroconjugates have emerged. Open database search engines, such as MSFragger allow the identification of cross-links on the basis of the observed peptide fragments and post-identification mapping of the adduct mass [54,136]. This workflow shares many of the disadvantages of the original RNP^{xl} tool, requiring substantial amount of post-identification manual verification and high probability of missing spectral identifications with limited number of peptide fragments. Alternatively, adaptation of existing commercial software PEAKS allows the combination of de novo sequencing that provides peptide sequence tags and matching of multiple modifications per peptide [53]. This strategy allows the identification of shifted and peptide ions, but due to the inability of the software to consider neutral losses, can provide only partial annotation of the observed fragment adducts. Successful analysis of datasets obtained from 4-thiouracil incorporation in yeast cells was achieved with the use of the XiSearch library for peptide-based mass spectrometry [43]. By specifying the cross-linked nucleotide with the use of photoreactive analogs, the combinatorial precursor problem is significantly simplified. This also limits the number of fragments to be considered in the MS2 spectrum down to the fragments generated by 4-thiouridine dissociation. Adaptation of the XiSearch engine allowed the matching of multiple neutral loss fragments from 4-thiouridine and identification of cross-links with RNA moiety up to three nucleotides.

4.4 Purification of peptide-RNA heteroconjugates from complex mixtures

Identification of peptide-RNA heteroconjugates is largely dependent on efficient biochemical enrichments that leads to depletion of the overwhelming number of non-cross-linked species. Due to the peptide-centric nature of the analysis and the thorough digestion of the RNA moiety, the reduction of the pool of linear peptides is crucial for obtaining successful results. Investigation of RNA-binding proteins has been typically performed through specific pull-down techniques targeting a single RNA species, such as olido(dT) capture of mRNAs [39,48]. Although efficient, these methods are limited to a particular type of RNA molecule and are not applicable to all organisms. In this study, two additional unbiased methods for purification of peptide-RNA heteroconjugates are presented – silica-based and SAX-based enrichment workflows. Both methods have been demonstrated to effectively enrich RNA-peptide species, depleting the bulk of linear peptides. Generally, the resulting fragment spectra are of relative high quality, with low probability of precursor interference. In comparison, the results obtained with standard C18/TiO_2 workflow enrich for a large number of linear peptides. Particularly, the enriched phosphopeptides have a high chance of fitting to the precursor mass of a cross-link, due to the high oxygen content of the phosphate group in both analytes. The bulk number of isolated linear peptides also leads to prominent precursor interference and markedly noisier spectra. Both silica-based and SAX-based workflows rely on the interaction of the matrix with the phosphate backbone and are therefore less likely to purify heteroconjugates generated with single ribonucleotides. In addition, they can be easily adapted to retrieve both, protein or peptide level information, in parallel to the analysis of heteroconjugates, that provides an additional level of verification.

In the span of this study, several papers dedicated to the purification of peptide-RNA heteroconjugates were published. Asencio et al. developed a solid-phase, silica-based extraction protocol for cross-linked nucleic acid protein complexes. Subsequently, silica-based purification in combination with TiO_2 was employed for the determination of peptide-RNA heteroconjugates by mass spectrometry [43]. The combination of two different purification strategies could provide enhanced depletion of the observed unspecific interactors. Such a strategy may be especially beneficial in situations where high amount of starting material is used to compensate the losses associated with additional purification steps. Alternative method for isolation of cross-links by filter-based purification that relies on the size difference of RNA and protein digests has been developed by Panhale et al. [53]. Cross-linked protein-RNA complexes could also be purified by a phenol-toluol extraction, on the basis of their physicochemical properties [137]. In addition, protocol for complete chemical digestion of the cross-linked RNA moiety to single nucleotides with the use of hydrofluoric acid has emerged [54]. These developments provide a wide range of tools that can be employed, combined and further improved to efficiently isolate cross-link species.

4.5 Conclusions and future perspectives

The identification of peptide-RNA cross-links can provide valuable structural information about the interactions of proteins and RNA. Detection of the heteroconjugate species is a challenging task, hampered by the low yield of the cross-linking reaction, combinatorial complexity of possible RNA adducts and convoluted fragmentation behavior. In this study, the possible heteroconjugates of the 4 canonical RNA nucleotides with proteins were explored and their fragmentation profile was characterized. The discovered mass adducts can be directly utilized for detection, identification and verification of peptide-RNA cross-links from any system. The acquired understanding of the generated products during collision-induced dissociation allowed for categorizing them into distinct spectral elements and the development of a strategy for full automated annotation, greatly alleviating the effort required for manual validation. The assignment of all features of a fragment spectrum was used for the development of a descriptive score that can be used for FDR estimation in a simple target-decoy strategy. As an alternative to UV-irradiation, bisulfite can be utilized for highly specific lysine to cytosine cross-linking reaction. Furthermore, two efficient workflows for enrichment of peptide-RNA heteroconjugates from various RNA species were established. Both silica-based and SAX-based purification could be employed to investigate RNA-protein interactions in *E. coli* cells and Hela cytoplasm, providing exhaustive depletion of non-cross-linked species and a high number of detected cross-link sites. The interaction sites of a large number of known RNA-binding proteins could be revealed. Additionally, a considerable number of novel RNA-binding proteins was identified, including enzymes involved in the lipid biosynthesis of *E. coli* and cytoskeleton associated proteins in Hela cells.

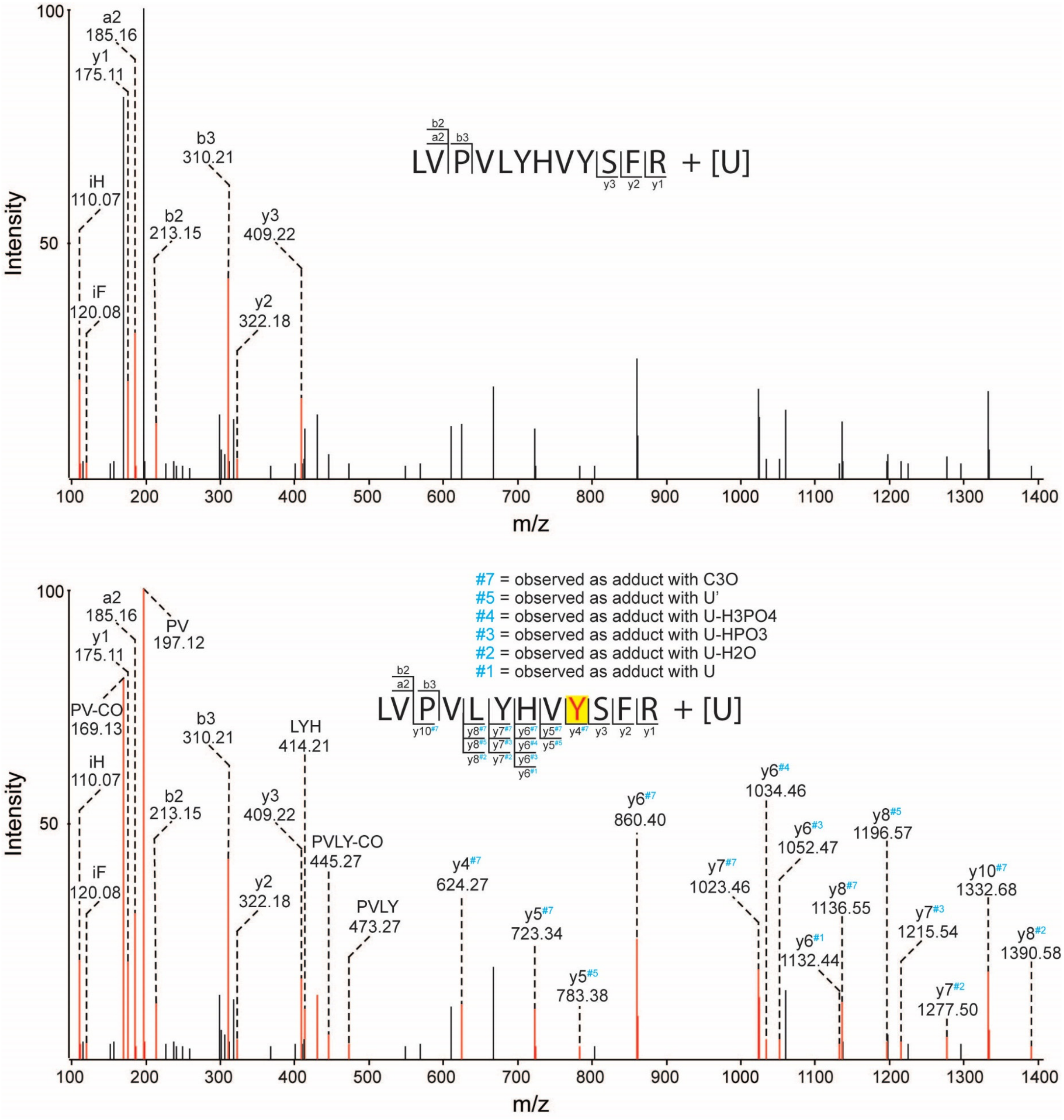
A great deal of effort has been exerted the last years in the field of protein-RNA investigation by mass spectrometry and especially in the detection of peptide-RNA heteroconjugates. This has led to substantial improvements in the speed, amount and confidence of the obtained information. At the same time, the generated output cannot compete with the depth and volume of data produced by other structural mass spectrometry investigations, such as protein-protein cross-linking workflows. Without a doubt, the continuous improvement of instrumentation and optimization of biochemical enrichment workflows will lead to more comprehensive explorations in the future. However, in order to release the full potential of mass spectrometry in the study of protein-RNA interactions, two key aspects should be addressed: Reducing the combinatorial complexity of the problem and improving the absolute yield of the cross-linking reaction. Optimization leading to complete digestion of the RNA moiety to single nucleotide level would allow for the use of standard proteomics search engines, substantial reduction of the search space and converging the split signal intensity of multiple precursor adducts into a single one. Combinations of specific cross-linking reactions, as the transamination mediated by bisulfite or photoaddition of reactive analogs, can promote the decrease of unknown products in the sample and thus its complexity. Moreover, development of a chemical or enzymatic method for degradation of the RNA adduct down to the cross-linked nucleobase would propel the search towards the simplicity and sensitivity of a post-translational modification searches, providing the depth of ubiquitination or phosphorylation investigations. Increase in the cross-linking yield may come from the discovery of more efficient chemical cross-linking methods or exploration of enhanced photochemical reactions, such as the employment of two photon excitation or the assessment of photosensitizing compounds. Breakthrough improvements in the abovementioned directions can lead to the ultimate goal – a simple and sensitive assay with fully automated and reliable data analysis that can be directly employed in any laboratory with access to a mass spectrometer.

Appendix

Supplementary Table 3.1 Immonium ions and amino acid fragments

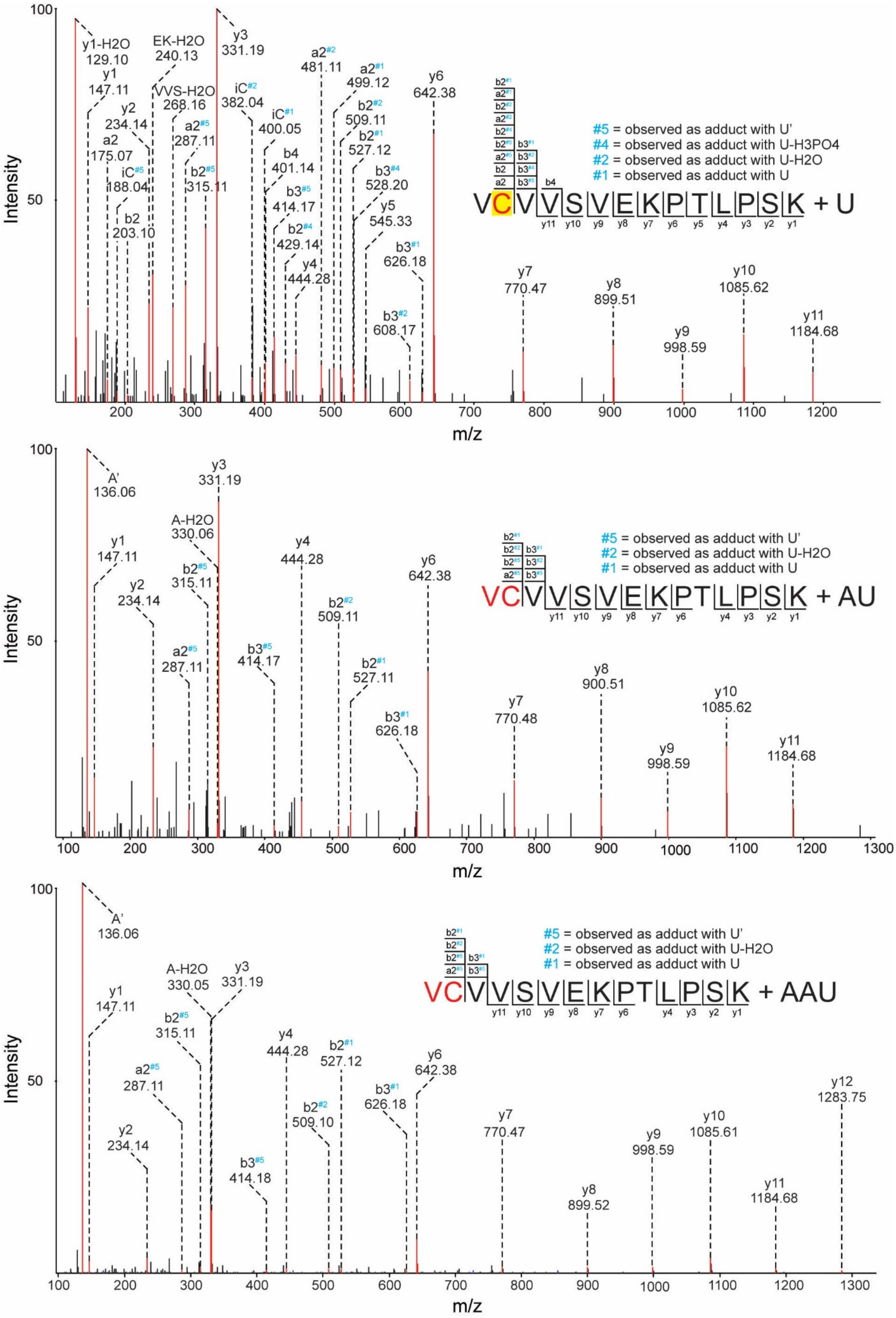
Amino acid	Formula	Monoisotopic mass
Lysine iK (K_1, K_2)	C5H13N2 (C6H13N2O, C5H10N)	101.1079 (129.1027, 84.0813)
Glycine iG	CH4N	30.0344
Alanine iA	C2H6N	44.0500
Serine iS	C2H6NO	60.0449
Proline iP	C4H8N	70.0657
Valine iV	C4H10N	72.0813
Threonine iT	C3H8NO	74.0606
Cysteine iC	C2H6NS	76.0221

Amino acid	Formula	Monoisotopic mass
Iso\leucine iI\iL	C5H12N	86.0970
Asparagine iN	C3H7N2O	87.0558
Aspartate iD	C3H6NO2	88.0399
Glutamine iQ	C4H9N2O	101.0715
Glutamate iE	C4H8NO2	102.0555
Methionine iM	C4H10NS	104.0534
Histidine iH	C5H8N3	110.0718
Phenylalanine iF	C8H10N	120.0813
Arginine iR	C5H13N4	129.1140
Tyrosine iY	C8H10NO	136.0762
Tryptophan iW	C10H11N2	159.0922



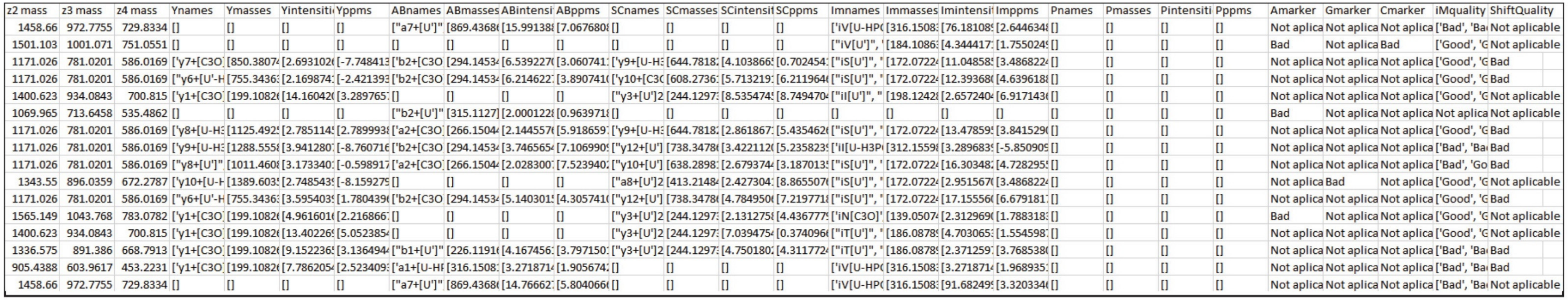
Supplementary Figure 3.1 Manual annotation of protein-RNA heteroconjugate spectra

The execution of the RNP^{xl} workflow results in a list of protein-RNA cross-link candidates that need to be manually annotated and validated. The search engine annotates only the peptide derived fragments (top panel). Shifted a-, b-, and y-ion series, precursor shifted ions and immonium ions need to be manually calculated, matched and assigned in order to achieve a fully annotated spectrum (bottom panel).

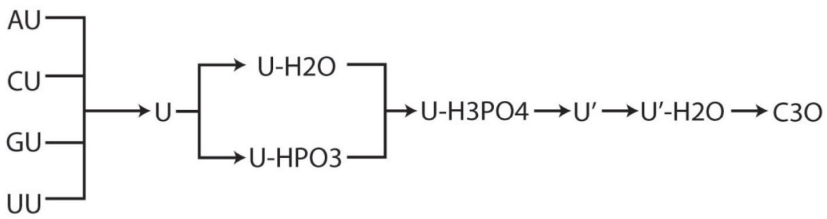


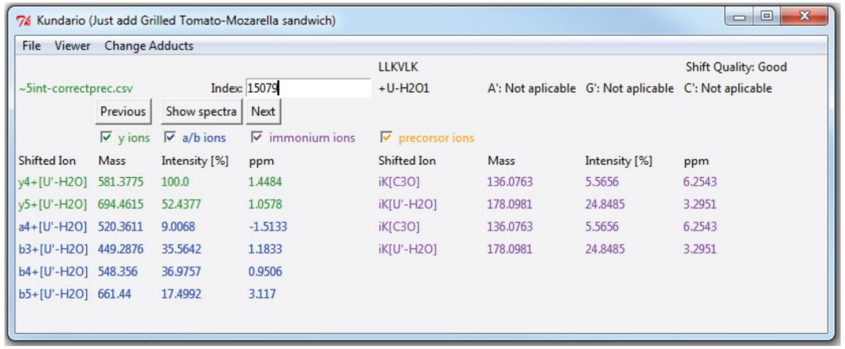
Supplementary Figure 3.2 Nucleobase marker ion suppression

The non-cross-linked nucleotides give rise to highly intense nucleobase marker ions that can suppress the other signals in the spectrum and pose one of the limitation of analyzing protein-RNA heteroconjugates with large RNA moieties



z2 mass	z3 mass	z4 mass	Ynames	Ymasses	Yintensiti	Yppms	ABnames	ABmasses	ABintensi	ABppms	SCnames	SCmasses	SCintensit	SCppms	Imnames	Immasses	Imintensit	Imppms	Pnames	Pmasses	Pintensiti	Pppms	Amarker	Gmarker	Cmarker	iMquality	ShiftQuality
1458.66	972.7755	729.8334	[]	[]	[]	[]	["a7+[U']"	[869.4368(	[15.99138{	[7.067680{	[]	[]	[]	[]	['iV[U-HP(	[316.1508:	[76.18108!	[2.644634{	[]	[]	[]	[]	Not aplica	Not aplica	Not aplica	['Bad', 'Ba	Not aplicable
1501.103	1001.071	751.0551	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	["iV[U']", '	[184.1086:	[4.344417:	[1.755024!	[]	[]	[]	[]	Bad	Not aplica	Bad	['Good', 'G	Not aplicable
1171.026	781.0201	586.0169	['y7+[C3O]	[850.3807‹	[2.693102(	[-7.748413	['b2+[C3O	[294.1453‹	[6.539227(	[3.060741:	['y9+[U-H3	[644.7818:	[4.103866!	[0.702454:	["iS[U']", '	[172.0722‹	[11.04858!	[3.486822‹	[]	[]	[]	[]	Not aplica	Not aplica	Not aplica	['Good', 'G	Bad
1171.026	781.0201	586.0169	["y6+[U'-H	[755.3436:	[2.169874:	[-2.421393	['b2+[C3O	[294.1453‹	[6.214622:	[3.890741(	['y10+[C3(	[608.2736:	[5.713219:	[6.211964(	["iS[U']", '	[172.0722‹	[12.39368(	[4.639618{	[]	[]	[]	[]	Not aplica	Not aplica	Not aplica	['Good', 'G	Bad
1400.623	934.0843	700.815	['y1+[C3O]	[199.1082(	[14.16042(	[3.289765:	[]	[]	[]	[]	["y3+[U']2	[244.1297:	[8.535474!	[8.749470‹	["iI[U']", "	[198.1242{	[2.657240‹	[6.917143(	[]	[]	[]	[]	Not aplica	Not aplica	Not aplica	['Good', 'G	Not aplicable
1069.965	713.6458	535.4862	[]	[]	[]	[]	["b2+[U']"	[315.1127]	[2.000122{	[0.963971{	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	Bad	Not aplica	Not aplica	Not aplica	Not aplicable
1171.026	781.0201	586.0169	['y8+[U-H3	[1125.492!	[2.785114!	[2.789993{	['a2+[C3O]	[266.1504‹	[2.144557(	[5.918659:	['y9+[U-H3	[644.7818:	[2.861867:	[5.435462(	["iS[U']", '	[172.0722‹	[13.47859!	[3.841529(	[]	[]	[]	[]	Not aplica	Not aplica	Not aplica	['Good', 'G	Bad
1171.026	781.0201	586.0169	['y9+[U-H3	[1288.555{	[3.941280:	[-8.760716	['b2+[C3O	[294.1453‹	[3.746565‹	[7.106990!	["y12+[U']	[738.3478(	[3.422112(	[5.235823!	['iI[U-H3P(	[312.1559{	[3.289683!	[-5.850909	[]	[]	[]	[]	Not aplica	Not aplica	Not aplica	['Bad', 'Ba	Bad
1171.026	781.0201	586.0169	["y8+[U']"	[1011.460{	[3.173340:	[-0.598917	['a2+[C3O]	[266.1504‹	[2.028300:	[7.523940:	["y10+[U']	[638.2898:	[2.679374‹	[3.187013!	["iS[U']", '	[172.0722‹	[16.30348:	[4.728295!	[]	[]	[]	[]	Not aplica	Not aplica	Not aplica	['Bad', 'Go	Bad
1343.55	896.0359	672.2787	['y10+[U-H	[1389.603!	[2.748543!	[-8.159279	[]	[]	[]	[]	["a8+[U']2	[413.2148‹	[2.427304:	[8.865507(	["iS[U']", '	[172.0722‹	[2.951567(	[3.486822‹	[]	[]	[]	[]	Not aplica	Bad	Not aplica	['Good', 'G	Not aplicable
1171.026	781.0201	586.0169	["y6+[U'-H	[755.3436:	[3.595403!	[1.780439(	['b2+[C3O	[294.1453‹	[5.140301!	[4.305741(	["y12+[U']	[738.3478(	[4.784950(	[7.219771{	["iS[U']", '	[172.0722‹	[17.15556(	[6.679181:	[]	[]	[]	[]	Not aplica	Not aplica	Not aplica	['Good', 'G	Bad
1565.149	1043.768	783.0782	['y1+[C3O]	[199.1082(	[4.961601(	[2.216866:	[]	[]	[]	[]	["y3+[U']2	[244.1297:	[2.131275{	[4.436777!	['iN[C3O]'	[139.0507‹	[2.312969(	[1.788318:	[]	[]	[]	[]	Bad	Not aplica	Not aplica	['Good', 'G	Not aplicable
1400.623	934.0843	700.815	['y1+[C3O]	[199.1082(	[13.40226!	[5.052385‹	[]	[]	[]	[]	["y3+[U']2	[244.1297:	[7.039475‹	[0.374096(	["iT[U']", '	[186.0878!	[4.703065:	[1.554598:	[]	[]	[]	[]	Not aplica	Not aplica	Not aplica	['Good', 'G	Not aplicable
1336.575	891.386	668.7913	['y1+[C3O]	[199.1082(	[9.152236!	[3.136494‹	["b1+[U']"	[226.1191(	[4.167456:	[3.797150:	["y3+[U']2	[244.1297:	[4.750180:	[4.311772‹	["iT[U']", '	[186.0878!	[2.371259:	[3.768538(	[]	[]	[]	[]	Not aplica	Not aplica	Not aplica	['Bad', 'Ba	Bad
905.4388	603.9617	453.2231	['y1+[C3O]	[199.1082(	[7.786205‹	[2.523409:	['a1+[U-HF	[316.1508:	[3.271871‹	[1.905674:	[]	[]	[]	[]	['iV[U-HP(	[316.1508:	[3.271871‹	[1.968935:	[]	[]	[]	[]	Not aplica	Not aplica	Not aplica	['Bad', 'Ba	Bad
1458.66	972.7755	729.8334	[]	[]	[]	[]	["a7+[U']"	[869.4368(	[14.76662:	[5.804066(	[]	[]	[]	[]	['iV[U-HP(	[316.1508:	[91.68249!	[3.320334(	[]	[]	[]	[]	Not aplica	Not aplica	Not aplica	['Bad', 'Ba	Not aplicable





~5int-correctprec.csv

Index: 15079

LLKVLK

+U-H2O1

A': Not aplicable G': Not aplicable C': Not aplicable

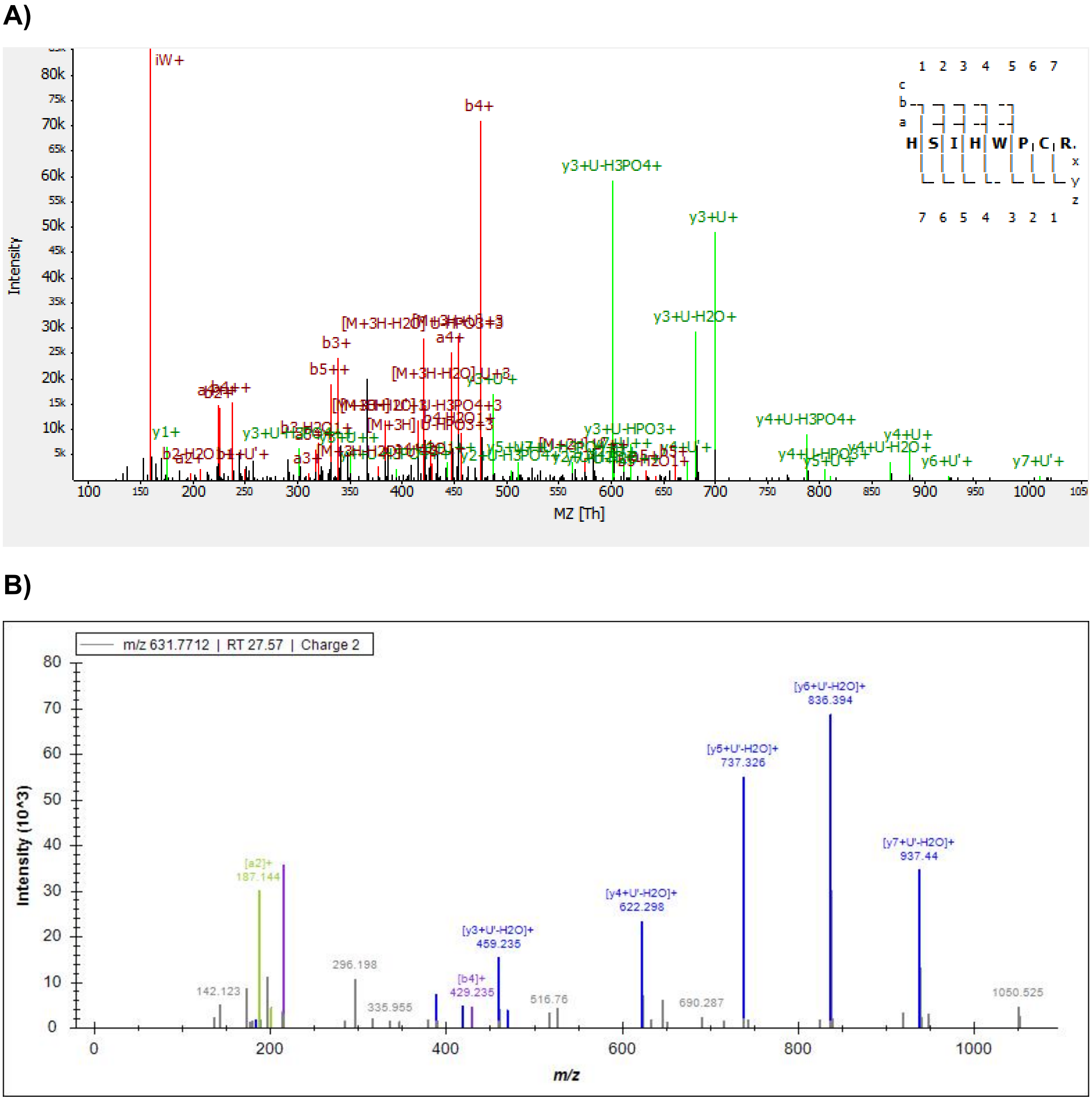
Shift Quality: Good

☑ y ions ☑ a/b ions ☑ immonium ions ☑ precorsor ions

Shifted Ion	Mass	Intensity [%]	ppm	Shifted Ion	Mass	Intensity [%]	ppm
y4+[U'-H2O]	581.3775	100.0	1.4484	iK[C3O]	136.0763	5.5656	6.2543
y5+[U'-H2O]	694.4615	52.4377	1.0578	iK[U'-H2O]	178.0981	24.8485	3.2951
a4+[U'-H2O]	520.3611	9.0068	-1.5133	iK[C3O]	136.0763	5.5656	6.2543
b3+[U'-H2O]	449.2876	35.5642	1.1833	iK[U'-H2O]	178.0981	24.8485	3.2951
b4+[U'-H2O]	548.356	36.9757	0.9506				
b5+[U'-H2O]	661.44	17.4992	3.117				

Supplementary Figure 3.3 Automated annotation output – python script

A) Example of tubular output of the python automated annotation. Additional columns are added to the RNP^{xl} output text files, that have the information of matched shifted ions, and their characteristics – mass accuracy, intensity, logical restrictions B) Logical restriction of possible downstream fragmentation of uracil based cross-links C) Example of a spectral view of the graphical user interface



Supplementary Figure 3.4 Examples of fully automated annotation of cross-link spectra

Output generated by current version of TOPPAS 2.4.0 (A) and Proteome Discoverer 2.1 (B)

Supplementary Table 3.2 Cross-links identified between GAPDH and HSH49 with poly(U), poly(G), poly(C) and poly(A)

Cross-linked amino acid localization is based on an exemplary fragment spectrum and may differ in other fragment spectra.

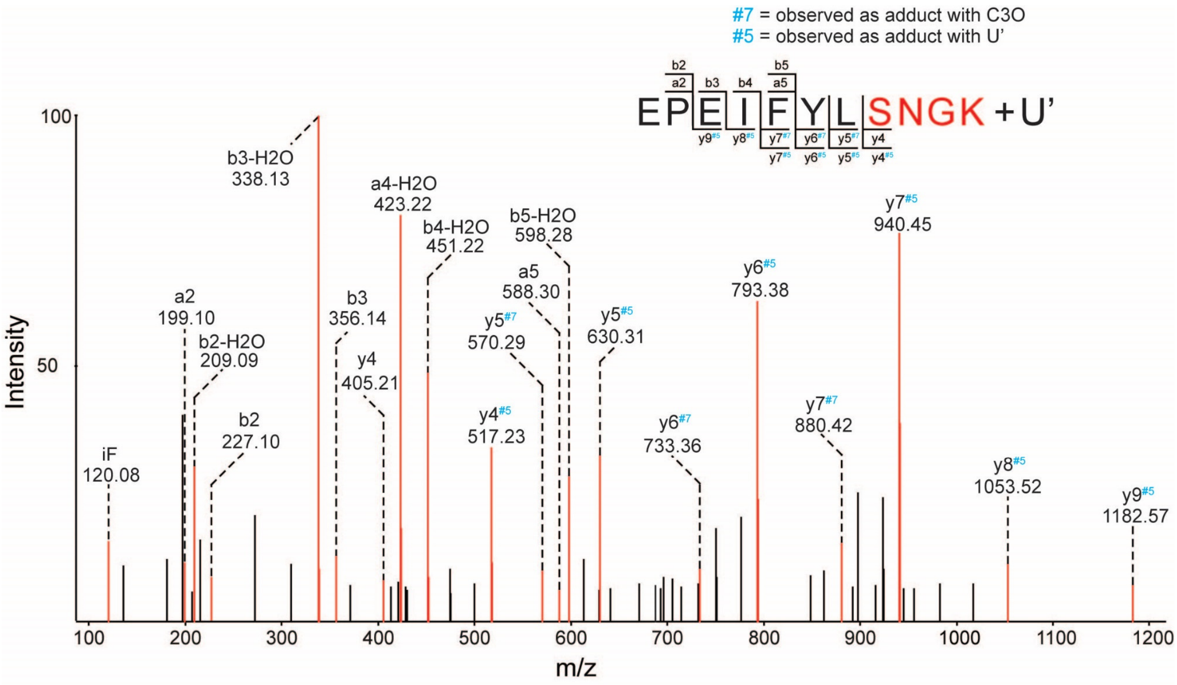
Protein (UniProt ID)	**Peptide**	**RNA adduct(s)**	**Cross-link localization**
GAPDH (P46406)	$_{60}$AENGKLVINGK$_{70}$	U-H2O	$_{65}$LV$_{66}$
	$_{71}$AITIFQER$_{78}$	U-H2O, U	$_{74}$IF$_{75}$
	$_{71}$AITIFQERDPANIK$_{84}$	U’-H2O, U-H2O, U, UU-H2O, UU	$_{75}$FQERD$_{79}$
	$_{214}$AVGKVIPELNGK$_{225}$	U’, U, UU	$_{217}$K
	$_{196}$DGRGAAQNIIPASTGAAK$_{213}$	U’, U-H2O, U, UU-H2O, UU, UUU	$_{196}$DGRG$_{199}$
	$_{199}$GAAQNIIPASTGAAK$_{213}$	U’, U-H2O-2H, U-H2O, U-2H, U, UU-H2O, UU	$_{205}$I
	$_{199}$GAAQNIIPASTGAAKAVGK$_{217}$	U’	$_{213}$K
	$_{12}$IGRLVTR$_{18}$	U, UU	$_{13}$GR$_{14}$
	$_{308}$LISWYDNEFGYSNR$_{321}$	U	$_{318}$Y
	$_{226}$LTGMAFR$_{232}$ $_{226}$LTGM(Ox)AFR$_{232}$	U’-H2O, U’, U-H2O-2H, U-H2O, U-2H, U, UU	$_{229}$M
	$_{15}$LVTRAAFNSGK$_{25}$	U-H2O, U, UU	$_{17}$TR$_{18}$
	$_{262}$QASEGPLK$_{269}$	U-H2O	$_{267}$PL$_{268}$
	$_{4}$VGVNGFGR$_{11}$	U’-H2O, U’, U-H2O, U-2H, U	$_{8}$G
	$_{4}$VGVNGFGRIGR$_{14}$	U’, U-H2O, U, UU-H2O, UU	$_{10}$GR$_{11}$
	$_{117}$VIISAPSADAPMFVMGVNHEK$_{137}$ $_{117}$VIISAPSADAPM(Ox)FVMGVNHEK$_{137}$ $_{117}$VIISAPSADAPM(Ox)FVM(Ox)GVNHEK$_{137}$	U-H2O, U	-
	$_{218}$VIPELNGK$_{225}$	U-H2O, UU-H2O	$_{220}$P
	$_{233}$VPTPNVSVVDLTCR$_{246}$	U	$_{244}$TC$_{245}$
	$_{322}$VVDLMVHMASKE$_{333}$ $_{322}$VVDLMVHM(Ox)ASKE$_{333}$	U	$_{329}$M
	$_{250}$AAKYDDIK$_{257}$	C-NH3	$_{252}$K
	$_{250}$AAKYDDIKK$_{258}$	C-NH3	$_{252}$K
	$_{60}$AENGKLVINGK$_{70}$	C-NH3	$_{64}$K
	$_{106}$AGAHLKGGAK$_{115}$	C-NH3	$_{111}$K
	$_{214}$AVGKVIPELNGK$_{225}$	C-NH3	$_{217}$K
	$_{54}$FHGTVKAENGK$_{64}$	C-NH3	$_{59}$K
	$_{199}$GAAQNIIPASTGAAK$_{213}$	C-NH3	$_{199}$GA$_{200}$
	$_{199}$GAAQNIIPASTGAAKAVGK$_{217}$	C-NH3	$_{213}$K
	$_{258}$KVVKQASEGPLK$_{269}$	C-NH3	$_{261}$K
	$_{226}$LTGM(Ox)AFR$_{232}$	C-NH3, C-NH3-2H	$_{231}$FR$_{232}$
	$_{185}$TVDGPSGKLWR$_{195}$	C-NH3	$_{192}$K
	$_{4}$VGVNGFGR$_{11}$	C-NH3	$_{8}$GFGR$_{11}$
	$_{233}$VPTPNVSVVDLTCR$_{246}$	C-NH3	$_{243}$LTCR$_{246}$
	$_{322}$VVDLM(Ox)VHM(Ox)ASKE$_{333}$	C-NH3	$_{332}$K
	$_{322}$VVDLM(Ox)VHMASKE$_{333}$	C-NH3	$_{332}$K
	$_{259}$VVKQASEGPLK$_{269}$	C-NH3	$_{261}$K
	$_{253}$YDDIKK$_{258}$	C-NH3	$_{257}$K
	$_{144}$IVSNASCTTNCLAPLAK$_{160}$	G	-
	$_{233}$VPTPNVSVVDLTCR$_{246}$	G’	-
HSH49 (Q99181)	$_{161}$ADLAIK$_{166}$	U-H2O	$_{163}$L
	$_{150}$CAYVYFEDFEK$_{160}$	U	$_{150}$CA$_{151}$
	$_{112}$DMILPIAK$_{119}$ $_{112}$DM(Ox)ILPIAK$_{119}$	U-H2O-2H, U-H2O, U, UU-H2O, UU	$_{114}$I
	$_{137}$EPEIFYLSNGK$_{147}$	U’, U-H2O-2H, U-H2O, U, UU	$_{142}$Y
	$_{131}$FGKLIR$_{136}$	U, UU	$_{132}$GKL$_{134}$
	$_{127}$IFNKFGK$_{133}$	U-H2O, U	$_{130}$K
	$_{127}$IFNKFGKLIR$_{136}$	U	$_{131}$FGK$_{133}$
	$_{38}$IKYPK$_{42}$	U’, U, U-H2O, UU-H2O, UU	$_{40}$Y
	$_{69}$IMNNTVR$_{75}$	U-H2O, UU-H2O, UU	$_{69}$IM$_{70}$
	$_{178}$ITVDYAFK$_{185}$	U’-H2O,U-H2O, U-2H, U, UU-H2O, UU	$_{184}$F
	$_{178}$ITVDYAFKENGK$_{189}$	U’, U-H2O, U-2H, U, UU	$_{184}$F
	$_{80}$LIKVR$_{84}$	U-H2O, U, UU	$_{82}$K
	$_{134}$LIREPEIFYLSNGK$_{147}$	U-H2O, U	$_{142}$Y
	$_{201}$LLNKEALK$_{208}$	U-H2O	$_{204}$KE$_{205}$
	$_{76}$LYDRLIK$_{82}$	U-H2O, U, UU	$_{79}$R
	$_{114}$NLADSIDSDQLVK$_{126}$	U-H2O-2H, U-H2O, U	$_{125}$V
	$_{85}$QVTNSTGTTNLPSNISK$_{101}$	U-H2O-2H, U-H2O, U, UU	$_{96}$PS$_{97}$
	$_{85}$QVTNSTGTTNLPSNISKDMILPIAK$_{109}$	U	-
	$_{167}$SLNNQLVANNR$_{177}$	U-H2O, U	$_{172}$L
	$_{83}$VRQVTNSTGTTNLPSNISK$_{101}$	U-H2O, U, UU-H2O, UU	$_{83}$VR$_{84}$
	$_{194}$YGDDVDRLLNK$_{204}$	U-H2O	$_{200}$RLL$_{202}$
	$_{161}$ADLAIKSLNNQLVANNR$_{177}$	C-NH3	$_{166}$K
	$_{205}$EALKHNMLK$_{213}$ $_{205}$EALKHNM(Ox)LK$_{213}$	C-NH3	$_{208}$K
	$_{137}$EPEIFYLSNGKLK$_{149}$	C-NH3	$_{145}$NGKL$_{148}$
	$_{131}$FGKLIR$_{136}$	C-NH3	$_{133}$K
	$_{190}$GNAKYGDDVDR$_{200}$	C-NH3	$_{193}$K
	$_{209}$HNMLK$_{213}$	C-NH3	$_{213}$K
	$_{127}$IFNKFGK$_{133}$	C-NH3	$_{130}$K
	$_{38}$IKYPK$_{42}$	C-NH3	$_{39}$K

	$_{38}$IKYPKDK$_{44}$	C-NH3	$_{42}$KD$_{43}$
	$_{178}$ITVDYAFK$_{185}$	C-NH3	$_{184}$FK$_{185}$
	$_{178}$ITVDYAFKENGK$_{189}$	C-NH3	$_{185}$K
	$_{80}$LIKVR$_{84}$	C-NH3	$_{82}$K
	$_{201}$LLNKEALK$_{208}$	C-NH3	$_{204}$K
	$_{38}$IKYPK$_{42}$	G-NH3, G-2H	$_{40}$Y
	$_{178}$ITVDYAFK$_{185}$	G-NH3, G-2H, G, GG	$_{184}$F
	$_{178}$ITVDYAFKENGK$_{189}$	G-NH3, G-2H, G, GG-NH3, GG	$_{184}$F
	$_{80}$LIKVR$_{84}$	G-2H	$_{82}$K
	$_{83}$VRQVTNSTGTTNLPSNISK$_{101}$	G-H2O	$_{83}$VR$_{84}$

Supplementary Table 3.3 Cross-linked identified between GAPDH and nucleotide monophosphate analogs

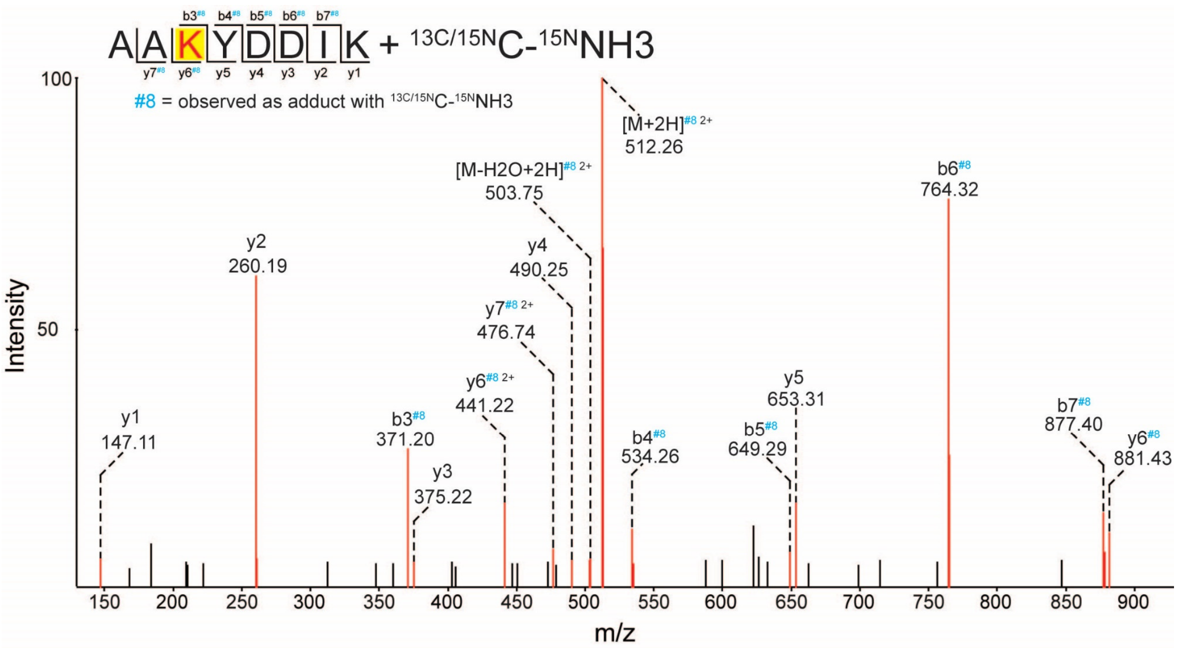
Cross-linked amino acid localization is based on an exemplary fragment spectrum and may differ in other fragment spectra.

Protein (UniProt ID)	**Peptide**	**RNA adduct(s)**	**Cross-link localization**	**Cross-linked nucleotide**
GAPDH (P46406)	$_{71}$AITIFQER$_{78}$	U	$_{75}$F	U
	$_{196}$DGRGAAQNIIPASTGAAK$_{213}$	U	$_{196}$DGRGAA$_{201}$	U
	$_{54}$FHGTVK$_{59}$	U-2H	$_{55}$H	U
	$_{199}$GAAQNIIPASTGAAK$_{213}$	U, U-2H	$_{210}$GAAK$_{213}$	
	$_{144}$IVSNASCTTNCLAPLAK$_{160}$	U-H2O, U-2H	$_{154}$C	U
	$_{226}$LTGMAFR$_{232}$ $_{226}$LTGM(Ox)AFR$_{232}$	U-H2O, U, U-2H	$_{231}$F	U
	$_{4}$VGVNGFGR$_{11}$	U	$_{8}$GF$_{9}$	U
	$_{4}$VGVNGFGRIGR$_{14}$	U	$_{8}$GFGR$_{11}$	U
	$_{233}$VPTPNVSVVDLTCR$_{246}$	U	$_{244}$TC$_{245}$	U
	$_{322}$VVDLM(Ox)VHM(Ox)ASK$_{332}$	U	$_{328}$HM$_{329}$	U
	$_{71}$AITIFQER$_{78}$	^{15N}U, ^{15N}U-2H	$_{75}$F	^{15N}U
	$_{71}$AITIFQERDPANIK$_{84}$	^{15N}U	$_{75}$F	^{15N}U
	$_{196}$DGRGAAQNIIPASTGAAK$_{213}$	^{15N}U	$_{196}$DGRGA$_{200}$	^{15N}U
	$_{54}$FHGTVK$_{59}$	^{15N}U, ^{15N}U-2H	$_{55}$H	^{15N}U
	$_{199}$GAAQNIIPASTGAAK$_{213}$	^{15N}U, ^{15N}U-H2O, ^{15N}U-2H	$_{208}$S	^{15N}U
	$_{199}$GAAQNIIPASTGAAKAVGK$_{217}$	^{15N}U	$_{213}$K	^{15N}U
	$_{144}$IVSNASCTTNCLAPLAK$_{160}$	^{15N}U, ^{15N}U-2H-H2O, ^{15N}U-2H	$_{150}$CTTNC$_{154}$	^{15N}U
	$_{308}$LISWYDNEFGYSNR$_{321}$	^{15N}U, ^{15N}U-2H	$_{316}$F	^{15N}U
	$_{226}$LTGMAFR$_{232}$ $_{226}$LTGM(Ox)AFR$_{232}$	^{15N}U, ^{15N}U-H2O, ^{15N}U-2H	$_{231}$F	^{15N}U
	$_{4}$VGVNGFGR$_{11}$	^{15N}U	$_{9}$F	^{15N}U
	$_{4}$VGVNGFGRIGR$_{14}$	^{15N}U	$_{10}$GR$_{11}$	^{15N}U
	$_{117}$VIISAPSADAPMFVMGVNHEK$_{137}$ $_{117}$VIISAPSADAPM(Ox)FVMGVNHEK$_{137}$ $_{117}$VIISAPSADAPM(Ox)FVM(Ox)GVNHEK$_{137}$	^{15N}U	$_{122}$P	^{15N}U
	$_{218}$VIPELNGK$_{225}$	^{15N}U	$_{220}$PELNGK$_{225}$	^{15N}U
	$_{233}$VPTPNVSVVDLTCR$_{246}$	^{15N}U, ^{15N}U-2H	$_{245}$C	^{15N}U
	$_{322}$VVDLM(Ox)VHM(Ox)ASK$_{332}$	^{15N}U	$_{328}$H	^{15N}U
	$_{71}$AITIFQER$_{78}$	$^{15N/13C}$U	$_{75}$F	$^{15N/13C}$U
	$_{196}$DGRGAAQNIIPASTGAAK$_{213}$	$^{15N/13C}$U	$_{196}$DGRGA$_{200}$	$^{15N/13C}$U
	$_{199}$GAAQNIIPASTGAAK$_{213}$	$^{15N/13C}$U-H2O, $^{15N/13C}$U, $^{15N/13C}$U-2H	$_{210}$G	$^{15N/13C}$U
	$_{226}$LTGMAFR$_{232}$ $_{226}$LTGM(Ox)AFR$_{232}$	$^{15N/13C}$U-H2O, $^{15N/13C}$U, $^{15N/13C}$U-2H	$_{231}$F	$^{15N/13C}$U
	$_{4}$VGVNGFGR$_{11}$	$^{15N/13C}$U, $^{15N/13C}$U-2H	$_{9}$F	$^{15N/13C}$U
	$_{322}$VVDLMVHM(Ox)ASK$_{332}$ $_{322}$VVDLM(Ox)VHM(Ox)ASK$_{332}$	$^{15N/13C}$U	$_{328}$H	$^{15N/13C}$U
	$_{144}$IVSNASCTTNCLAPLAK$_{160}$	$^{15N/13C}$U -2H	SCTTNCL	$^{15N/13C}$U
	$_{250}$AAKYDDIK$_{257}$	C-NH3	$_{252}$K	C
	$_{250}$AAKYDDIKK$_{258}$	C-NH3	$_{252}$K	C
	$_{214}$AVGKVIPELNGK$_{225}$	C-NH3	$_{216}$GK$_{217}$	C
	$_{54}$FHGTVK$_{59}$	C-NH3	$_{54}$F	C
	$_{54}$FHGTVKAENGK$_{64}$	C-NH3	$_{59}$K	C
	$_{199}$GAAQNIIPASTGAAK$_{213}$	C-NH3	$_{199}$G	C
	$_{199}$GAAQNIIPASTGAAKAVGK$_{217}$	C-NH3	$_{213}$K	C
	$_{262}$QASEGPLK$_{269}$	C-NH3	$_{262}$Q	C
	$_{185}$TVDGPSGKLWR$_{195}$	C-NH3	$_{191}$GK$_{192}$	C
	$_{259}$VVKQASEGPLK$_{269}$	C-NH3	$_{261}$K	C
	$_{253}$YDDIKK$_{258}$	C-NH3	$_{257}$KK$_{258}$	C
	$_{250}$AAKYDDIK$_{257}$	^{15N}C-^{15N}NH3	$_{252}$K	^{15N}C
	$_{250}$AAKYDDIKK$_{258}$	^{15N}C-^{15N}NH3	$_{252}$K	^{15N}C
	$_{60}$AENGKLVINGK$_{70}$	^{15N}C-^{15N}NH3	$_{63}$GK$_{64}$	^{15N}C
	$_{214}$AVGKVIPELNGK$_{225}$	^{15N}C-^{15N}NH3	$_{216}$GK$_{217}$	^{15N}C
	$_{54}$FHGTVK$_{59}$	^{15N}C-^{15N}NH3	$_{54}$F	^{15N}C
	$_{199}$GAAQNIIPASTGAAK$_{213}$	^{15N}C-^{15N}NH3	$_{199}$G	^{15N}C
	$_{262}$QASEGPLK$_{269}$	^{15N}C-^{15N}NH3	$_{262}$Q	^{15N}C
	$_{259}$VVKQASEGPLK$_{269}$	^{15N}C-^{15N}NH3	$_{261}$K	^{15N}C
	$_{253}$YDDIKK$_{258}$	^{15N}C-^{15N}NH3	$_{257}$KK$_{258}$	^{15N}C
	$_{250}$AAKYDDIK$_{257}$	$^{15N/13C}$C-^{15N}NH3	$_{252}$K	$^{15N/13C}$C
	$_{60}$AENGKLVINGK$_{70}$	$^{15N/13C}$C-^{15N}NH3	$_{63}$GK$_{64}$	$^{15N/13C}$C
	$_{214}$AVGKVIPELNGK$_{225}$	$^{15N/13C}$C-^{15N}NH3	$_{217}$K	$^{15N/13C}$C
	$_{54}$FHGTVK$_{59}$	$^{15N/13C}$C-^{15N}NH3	$_{54}$F	$^{15N/13C}$C
	$_{54}$FHGTVKAENGK$_{64}$	$^{15N/13C}$C-^{15N}NH3	$_{59}$K	$^{15N/13C}$C
	$_{199}$GAAQNIIPASTGAAK$_{213}$	$^{15N/13C}$C-^{15N}NH3	$_{199}$G	$^{15N/13C}$C
	$_{262}$QASEGPLK$_{269}$	$^{15N/13C}$C-^{15N}NH3	$_{262}$Q	$^{15N/13C}$C
	$_{185}$TVDGPSGKLWR$_{195}$	$^{15N/13C}$C-^{15N}NH3	$_{192}$K	$^{15N/13C}$C
	$_{259}$VVKQASEGPLK$_{269}$	$^{15N/13C}$C-^{15N}NH3	$_{261}$K	$^{15N/13C}$C
	$_{233}$VPTPNVSVVDLTCR$_{246}$	G’	$_{243}$LTCR$_{246}$	G
	$_{250}$AAKYDDIKK$_{258}$	^{15N}G-2H	$_{253}$YD$_{254}$	^{15N}G



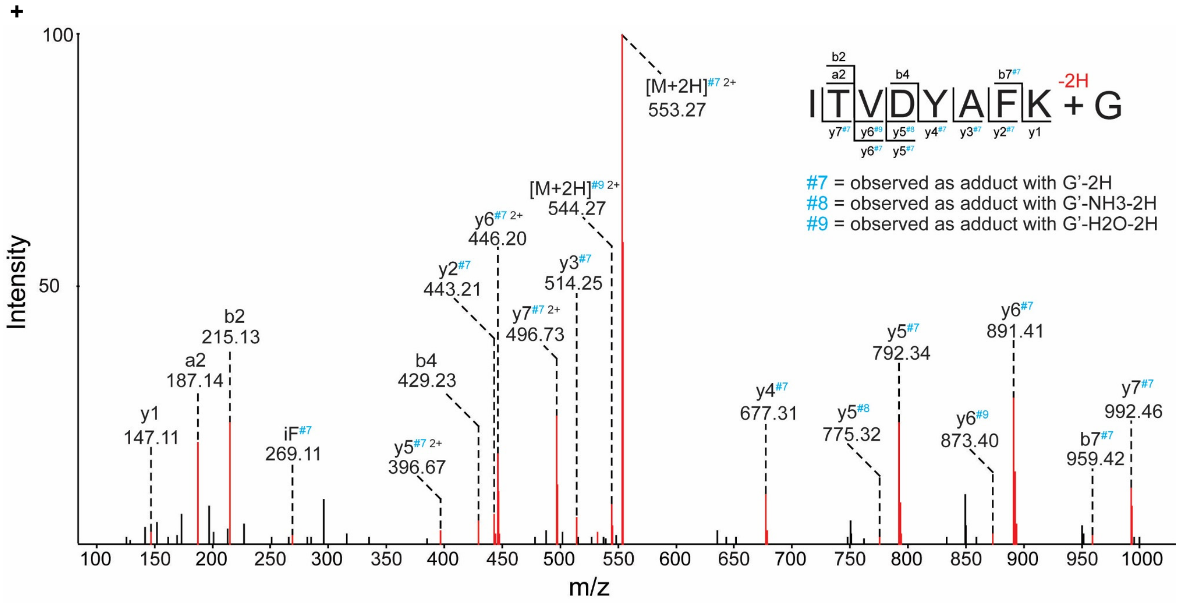
Supplementary Figure 3.5 Exemplary cross-link spectrum with the uracil nucleobase

Fragment spectrum of peptide $_{137}$EPEIFYLSNGK$_{147}$ with uracil, generated by irradiation of Hsh49 with poly(U). The identification is supported by prominent y-ion and b-ion series. The shifted y4-y9 ions localize the cross-link site to the $_{144}$SNGK$_{147}$ region.



Supplementary Figure 3.6 The 4-NH2 group of cytosine is lost during cross-linking

Fragment spectrum of GAPDH peptide $_{250}$AAKYDDIK$_{257}$ cross-linked to heavy labeled (^{13}C/^{15}N) cytidine monophosphate with ^{15N}NH3 net loss. The deficit of an ^{15}N ammonia group, indicates that the loss occurred from the 4-NH2 group of the pyrimidine nucleobase. The identification is supported by complete y-ion sequencing series and extensive b-ion series. The shifted ion series localize $_{253}$K as the cross-linked amino acid.



Supplementary Figure 3.7 Exemplary spectrum of guanosine cross-link with -2H net loss

Fragment spectrum of peptide $_{178}$ITVDYAFK$_{185}$ with guanosine monophosphate with 2H net loss, generated by irradiation of HSH49 with poly(G). The shifted b- and y-ion series localize $_{184}$F as the cross-linked amino acid. Several different neutral losses can be identified, all generated by fragmentation of the N-glycosidic bond and 2H deficit.

Supplementary Table 3.4 Identification of cytosine cross-links with a standard proteomics search

Parameters used to for identification of sulfite-mediated cross-links by MaxQuant.

Modification	**C-NH3**
Description	CMP with NH3 loss
Composition	C9H11O8N2P
Specificity	K, protein N-terminal
Neutral loss	Ribose-P (C5H9O7P)

Modification	**C-NH3-HPO3**
Description	CMP with NH3 and HPO3 loss
Composition	C9H10O5N2
Specificity	K, protein N-terminal
Neutral loss	Ribose (C5H8O4)

Supplementary Table 3.5 Identified and manually validated cross-link sites in *E. coli*

The sample origin is noted in parenthesis. Ambiguous pyrimidine generated 306 Da adduct is noted as “N”. Cross-linked amino acid localization is based on an exemplary fragment spectrum and may differ in other fragment spectra.

Protein (UniProt ID)	**Peptide**	**RNA Adducts**
DNA-directed RNA polymerase subunit beta (P0A8V2)	$_{842}$DT[K]LGPEEITADIPNVGEAALSK$_{864}$ $_{842}$DTK(Carbamyl)LGPEEITADIPNVGEAALSK$_{864}$	N(VI), N-HPO3(II), UN(IV)
	$_{504}$EFFGSSQLSQ[F]MDQNNPLSEITHK$_{527}$ $_{504}$ (Carbamyl)EFFGSSQLSQF[M]DQNNPLSEITHK$_{527}$	AN(VI), CN(IX), GN (IX), N(VI), UA-NH3(VI), UG-NH3(IX)
	$_{181}$GS[W]LDFEFDPK$_{191}$	AAG-NH3(IV,VI), AAG-NH3-HPO3(VI)
	$_{530}$ISALG[PGGLTRE]R$_{542}$	CN-HPO3(I)
	$_{1036}$ITQGDDLA[P]GVLK$_{1048}$	AN(VI)
	$_{479}$LSLGDLDTLMPQDMINA[K]PISAAVK$_{503}$	AN(VI), N(VI), UU(VI)
	$_{1247}$ST[G]SYSLVTQQPLGGK$_{1262}$	AGN(VII), AUN(VII), CN(IX), CG-H2O(VII), GN(VII,IX), N(VI), UN (VII,IX)
	$_{1247}$STGSYSLVTQQPLG[G]KAQFGGQR$_{1269}$ $_{1247}$(Carbamyl)STGSYSLVTQQPL[GG]KAQFGGQR$_{1269}$	ACN(IX), AUN(IX), AU-HPO3(I), CN(IX), CN-HPO3(I), GN(IX), UN(IX), UN-HPO3(I)
	$_{55}$SVFPIQ[SY]SGNSELQYVSYR$_{74}$	UN(IX), UN-HPO3(I), UG-NH3(IX), UU(VI)
	$_{144}$VIVSQLH[R]SPGVFFDSDK$_{161}$	CN(IX)
	$_{887}$VTP[K]GETQLTPEEK$_{900}$	N(VI)
DNA-directed RNA polymerase subunit beta' (P0A8T7)	$_{315}$AIT[GSNKR]PLK$_{325}$	AAN(VI)
	$_{203}$EELNETN[SET]KR$_{214}$	CN(VII)
	$_{418}$EHPVLL[NRAP]TLHR$_{431}$ $_{418}$(Carbamyl)EHPVLL[NRAP]TLHR$_{431}$	AN(VI), AN-HPO3(IX), CN(IX)
	$_{1151}$EPAILAEISGIVS[FG]KETK$_{1170}$	CN(IX)
	$_{1207}$GDVISDGPEAP[H]DILR$_{1222}$	CN(IX)
	$_{124}$IGLLLDM[P]LR$_{133}$	GN(IX), UN(IX)
	$_{124}$IGLLLDM[PL]RDIER$_{137}$	CN(IX)
	$_{40}$KPET[INY]RTFKPER$_{53}$	CN(IX), GN(IX), AGN(IX)
	$_{1175}$LVITPVDGSDPYEEMI[PK]$_{1192}$	CN(IX)
	$_{1068}$TAG[GK]DLRPALK$_{1079}$	AC-NH3(VI)
	$_{790}$TANSG[Y]LTR$_{798}$	CN(2)
	$_{1141}$VADLFEARRPKEPAILAEISGIVSFGK$_{1167}$	AAA-NH3(VI), CU(VI)
	$_{347}$VDYS[GR]SVITVGPYLR$_{362}$	CN(IX), CG-NH3(IX), UN(IX)
RNA polymerase sigma factor RpoD (P00579)	$_{104}$[EM]GTVELLTR$_{113}$ $_{104}$[EM](Oxidation)GTVELLTR$_{113}$	N(IV,VI,II), U(VI), UN(IV), U-HPO3(II), UU(IV)
	$_{424}$GYKF[S]TYATWWIR$_{436}$	N(VI)
	$_{452}$IPVHM(Oxidation)IETINK$_{462}$	CG-H2O(VI)
	$_{100}$MYMREM(Oxidation)GTVELLTR$_{113}$ $_{100}$M(Oxidation)YM(Oxidation)REMGTVELLTR$_{113}$ $_{100}$M(Oxidation)YMREM(Oxidation)GTVELLTR$_{113}$	U(II), N(II)
RNA polymerase sigma factor RpoD / RpoS (P00579 / P13445)	$_{427/142}$[FS]TYATWWIR$_{436/151}$	N(VI)
UDP-3-O-(3-hydroxymyristoyl)glucosamine N-acyltransferase LPXD_ECOLI (P21645)	$_{71}$SAALVVKN[PY]LTYAR$_{85}$	CN-HPO3(III)
Carbamoyl-phosphate synthase small chain CARA_ECOLI (P0A6F1)	$_{331}$SL[F]DGTLQGIHR$_{342}$	UN-HPO3(III), AGN-HPO3(III,IX), AGN(IX)

Probable transcriptional regulatory protein YebC YEBC_ECOLI (P0A8A0)	$_{109}$TVAEVR[H]AFSK$_{119}$	CN-HPO3(III)
3-hydroxydecanoyl-[acyl-carrier-protein] dehydratase FABA_ECOLI (P0A6Q3)	$_{163}$VGLFQDTSA[F]$_{172}$	CN-HPO3(III), N(VIII,VII), CN(VIII,VII), GN(VIII,VII), UN(VII)
Shikimate kinase 1 AROK_ECOLI (P0A6D7)	$_{117}$DK[KRPL]LHVETPPR$_{130}$	CN(IX), GN(IX)
Protein YhgF YHGF_ECOLI (P46837)	$_{328}$ATMGLDP[GL]RTGVK$_{341}$	N(VI)
RNA-binding protein YhbY YHBY_ECOLI (P0AGK4)	$_{1}$MNLS[T]KQK$_{8}$	ACN(IX)
Modulator of FtsH protease HflK HFLK_ECOLI (P0ABC7)	$_{52}$LGG[LGG]GK$_{59}$	CN(IX,VIII,VII), CCN(IX,VII), UN(VIII,VII)
Trigger factor TIG_ECOLI (P0A850)	$_{38}$KV[R]IDGFR$_{45}$	CN(IX), GN(IX), UN(IX)
	$_{46}$KGKV[P]MNIVAQR$_{57}$	CN(IX), GN(IX)
	$_{47}$GKV[P]MNIVAQR$_{57}$	CN(IX), GN(IX)
	$_{30}$SELVNVA[K]K$_{38}$	N(VI)
Protein translocase subunit SecY SECY_ECOLI (P0AGA2)	$_{256}$RV[Y]AAQSTHLPLK$_{268}$	CN(IX), GN(IX), AGN(IX), N(VI)
D-alanyl-D-alanine carboxypeptidase DacA DACA_ECOLI (P0AEB2)	$_{390}$IID[Y]IK$_{395}$	CN(IX), UN(VIII,VII), GN(VIII,VII), CUN(VII), AGN(VII)
Dual-specificity RNA methyltransferase RlmN RLMN_ECOLI (P36979)	$_{333}$VL[M]SYGFTTIVR$_{344}$	N(IX)
Cysteine synthase A CYSK_ECOLI (P0ABK5)	$_{106}$ALGANLVLTEG[AK]GMK$_{121}$	CN(IX), ACN(IX)
CDP-diacylglycerol--serine O-phosphatidyltransferase PSS_ECOLI (P23830)	$_{335}$LQ[Y]YVNTDQLVVR$_{347}$	GN(VIII)
Met repressor METJ_ECOLI (P0A8U6)	$_{24}$KI[T]VSIPLK$_{32}$	GN(IX)
Guanylate kinase KGUA_ECOLI (P60546)	$_{69}$DAFLEHAEV[F]GNYYGTSR$_{86}$	GN(IX)
Enoyl-[acyl-carrier-protein] reductase [NADH] FabI FABI_ECOLI (P0AEK4)	$_{184}$VNAISA[GPIR]TLAASGIK$_{201}$	GN(IX)
Exodeoxyribonuclease III EX3_ECOLI (P09030)	$_{210}$FSWFD[Y]RSK$_{218}$	GN(IX)
3-hydroxyacyl-[acyl-carrier-protein] dehydratase FabZ FABZ_ECOLI (P0A6Q6)	$_{101}$FKRPVVPGDQMIMEVT[F]EK$_{119}$	N(IX)
	$_{103}$RPVVPGDQMIMEVT[F]EK$_{119}$	N(IX)
Cell division protein ZapD ZAPD_ECOLI (P36680)	$_{203}$LNLSLDSQLYPQI[SG]HK$_{219}$	CN(IX), GN(IX)
Chemotaxis protein CheY CHEY_ECOLI (P0AE67)	$_{92}$KENIIAAAQAGA[S]GYVVKPFTAATLEEK$_{119}$	UN(IX), CN(IX), GN(IX), AGN(IX)
	$_{93}$ENIIAAAQAGAS[G]YVVKPFTAATLEEK$_{119}$ $_{93}$(Carb)ENIIAAAQAGAS[G]YVVKPFTAATLEEK$_{119}$	CN(IX), UN(IX), ACN(IX), GN(IX)
Protein RecA RECA_ECOLI (P0A7G6)	$_{200}$IGVMFGNPETTTGGNALK[F]YASVR$_{223}$	CN(IX), GN(IX)
Phosphocarrier protein HPr PTHP_ECOLI (P0AA04)	$_{46}$SL[F]KLQTLGLTQGTVVTISAEGEDEQK$_{72}$	GN(IX), AU(IX)
tRNA-specific 2-thiouridylase MnmA MNMA_ECOLI (P25745)	$_{150}$DQSY[F]LYTLSHEQIAQSLFPVGELEKPQVR$_{179}$	CUN(IX)
Glyceraldehyde-3-phosphate dehydrogenase A G3P1_ECOLI (P0A9B2)	$_{185}$TVDGPSH[K]DWR$_{195}$	N(VI)
	$_{322}$VLDLIAHIS[K]$_{331}$	N(VI)
	$_{161}$VINDNFGIIEGLM(Oxidation)TTVHATTATQK$_{184}$	G-NH3-HPO3(VI)
Uncharacterized tRNA/rRNA methyltransferase YfiF YFIF_ECOLI (P0AGJ5)	$_{152}$KA[YH]VVDEAELTK$_{164}$	U(VI), N(VI)
	$_{93}$[SF]IDPEVLR$_{101}$	GN(VII)
tRNA threonylcarbamoyladenosine biosynthesis protein TsaB TSAB_ECOLI (P76256)	$_{64}$GPG[S]FTGVR$_{72}$	N(VI)
	$_{119}$[M]GEVYWAEYQR$_{129}$	N(VI)
Dihydrolipoyl dehydrogenase DLDH_ECOLI (P0A9P0)	$_{399}$LIFD[K]ESHR$_{407}$	N(VI)
Fructose-bisphosphate aldolase class 2 ALF_ECOLI (P0AB71)	$_{306}$ANEA[Y]LQGQLGNPK$_{319}$	N(VI)
Transcription termination/antitermination protein NusA NUSA_ECOLI (P0AFF6)	$_{132}$EHEGEIITGVV[K]K$_{144}$	N(VI)
	$_{4}$EILAVVEAVSNE[K]ALPR$_{20}$	N(VI)
Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex ODO2_ECOLI (P0AFG6)	$_{197}$NSTAMLTTFNEVNMKPI[M]DLR$_{217}$	N(VI)
Acyl carrier protein ACP_ECOLI (P0A6A8)	$_{10}$[K]IIGEQLGVK$_{19}$	N(VI)
DNA-binding protein HU-alpha DBHA_ECOLI (P0ACF0)	$_{71}$IAAANVPAFVS[G]KALK$_{86}$	N(VI)
	$_{1}$[M]NKTQLIDVIAEK$_{13}$	GN(V,IX)
Transcription antitermination protein NusB NUSB_ECOLI (P0A780)	$_{96}$SDV[PYK]VAINEAIELAK$_{112}$	CU(VI)
	$_{392}$VDFS[K]FGEIEEVELGR$_{407}$	N(VI)

Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex ODP2_ECOLI (P06959)	$_{493}$YINIGV[AV]DTPNGLVVPVFK$_{512}$	GU(VI)
Phosphoglycerate kinase PGK_ECOLI (P0A799)	$_{356}$ISYISTGGGAF[L]EFVEGK$_{373}$	N(VI), AN(VI)
3-oxoacyl-[acyl-carrier-protein] reductase FabG FABG_ECOL (P0AEK2)	$_{98}$[MKDEE]WNDIIETNLSSVFR$_{116}$ $_{98}$[M(Ox)KDEE]WNDIIETNLSSVFR$_{116}$	N(VI)
Transaldolase B TALB_ECOLI (P0A870)	$_{242}$LTIAPALL[K]ELAESEGAIER$_{261}$	N(VI)
tRNA pseudouridine synthase D TRUD_ECOLI (Q57261)	$_{68}$IHAREVS[F]AGQK$_{81}$	CN(IX), CU(VI)
	$_{120}$KLR[L]GALK$_{127}$	CN(IX)
	$_{121}$LR[L]GALK$_{127}$	CN(IX)
	$_{22}$ANPEDFVVVEDLG[F]EPDGEGEHILVR$_{47}$	AUA-NH3(IX), AN(VI),AAN(VI)
	$_{231}$GS[W]FVATTEELAELQR$_{246}$	N(IX,VI), CN(IX), CU(IX)
Ribonuclease R RNR_ECOLI (P21499)	$_{105}$KDDL[Y]LSSEQMK$_{116}$	UN(IX,VII), GN(IX), N(VI,IV), UGN(VII)
	$_{756}$QV[G]KKVNFEPDSAFR$_{770}$	CG-H2O(IX)
	$_{663}$LDDLFIDGLVHVSSLDND[Y]YR$_{683}$	CN(IX), N(IX), AN(IX), GN(IX), UN(IX)
Exoribonuclease 2 RNB_ECOLI (P30850)	$_{43}$SY[F]IPPPQMK$_{52}$	CN(IX, VII), GN(IX,VIII,VII), N(Vi,V,IV), AGN(VII), UN(VII)
	$_{60}$IIAVIHSEKERESAEPEELVEPFLTR$_{85}$	N(IX)
	$_{69}$[ERESA]EPEELVEPFLTR$_{85}$	UN(IX)
	$_{580}$LVDNGAIA[F]IPAPFLHAVR$_{598}$	N(VI), CN(IX), UN(IX), GN(IX), ACN(IX), AGN(IX)
	$_{32}$G[F]GFLEVDAQK$_{42}$	N(VI,V,IV)
tRNA-dihydrouridine synthase B DUSB_ECOLI (P0ABT5)	$_{85}$INVESGAQIIDINMG[C]PAK$_{103}$ $_{85}$INVESGAQIIDINM(Ox)G[C]PAK$_{103}$	N(III,IX,VI), U(IX), U-HPO3(III)

Supplementary Table 3.6 Identified and manually validated cross-link sites in HeLa cytoplasmic extract

The sample origin is noted in parenthesis. Ambiguous pyrimidine generated 306 Da adduct is noted as "N". Cross-linked amino acid localization is based on an exemplary fragment spectrum and may differ in other fragment spectra.

Protein (UniProt ID)	Peptide	RNA Adducts
40S ribosomal protein S30 (P62861)	$_{42}$FVNVVPT[F]GKK$_{52}$	N(I), CU(I)
	$_{42}$FVNVVPTFGK$_{51}$	N(I)
40S ribosomal protein S10 (P46783)	$_{81}$D[Y]LHLPPEIVPATLR$_{95}$	N(I), U(I)
40S ribosomal protein S10/Putative 40S ribosomal protein S10-like (P46783/ Q9NQ39)	$_{25/25}$KDV[HM]PKHPELADK$_{38/38}$	N(I)
40S ribosomal protein S11 (P62280)	$_{153}$KQ[F]QKF$_{158}$	AN(I)
	$_{36}$[YY]KNIGLGFK$_{35}$	N(I), AN(I)
	$_{36}$[YY]KNIGLGFKTPK$_{38}$	ACN(V),ACU(V)
	$_{49}$EAIEGTYIDKK[C]PFTGNVSIR$_{69}$	N(I)
	$_{119}$DVQIGDIVTVGEC[R]PLSK$_{136}$	N(I), AN(I)
40S ribosomal protein S13 (P62277)	$_{10}$[G]LSQSALPYR$_{19}$	N(I)
	$_{79}$[G]LAPDLPEDLYHLIK$_{93}$	N(I)
40S ribosomal protein S15 (P62841)	$_{51}$[R]KQHSLLK$_{58}$	AN(I), AU(I),AAU(I), AAN(I), AAU-HPO3(V)
	$_{48}$GLR[R]KQHSLLK$_{58}$	AAU(I,V), AU-HPO3(V), AAU-HPO3(V), AGU-HPO3(V), AU(V), AGU(V)
	$_{51}$[R]KQHSLLKR$_{59}$	AAU-HPO3(V)
40S ribosomal protein S15a (P62244)	$_{23}$RQVLIR[PC]SK$_{32}$	N(I)
	$_{24}$QVLIRP[C]SK$_{32}$	N(I)
40S ribosomal protein S18 (P62269)	$_{24}$RKIA[F]AITAIK$_{34}$	GG(I)
	$_{25}$KIA[F]AITAIK$_{34}$	GG(I)
	$_{26}$IAFAITAIK$_{34}$	GG(I)
	$_{76}$QYKIPD[W]FLNR$_{86}$	CG(V)
40S ribosomal protein S19 (P39019)	$_{134}$IAGQVAAAN[K]K$_{144}$	N(I)
	$_{63}$HLY[LR]GGAGVGSMTK$_{77}$	UU(I), UN(I),
	$_{134}$IAGQVAAAN[K]KH$_{145}$	CGN(III)
	$_{85}$NGV[M]PSHFSR$_{94}$	G-H2O(III), G-NH3(III)
40S ribosomal protein S2 (P15880)	$_{174}$IGKPHTVP[C]K$_{183}$	N(I)
	$_{174}$IGKPHTVP[C]KVTGR$_{187}$	N(I)
	$_{228}$[GC]TATLGNFAK$_{238}$	UU(I)
	$_{247}$TYS[Y]LTPDLWK$_{257}$	UU-HPO3(I,III,IV,V), U(I,III,IV,V), N(I,III,IV,V), CG-HPO3(I), UU(I,II,III,IV,V), GG-H2O(I), UN(I,III,IV,V), AA-H2O(I), UUN(I), UUU(I, II, IV,V), U-HPO3(III,V), CUU-HPO3(III,V), GUU-HPO3(III), UUU-HPO3(III,IV,V), UG-NH3(III,V), GUU(III), CUU(III), CUN(V)
	$_{247}$TYS[Y]LTPDLWKETVFTK$_{263}$	UU-HPO3(V), U(V), N(V), CU-HPO3(V), UN(V), UU(I,V), UUU-HPO3(V)
	$_{247}$TYS[Y]LTPDLWKETVFTKSPYQEFTDHLVK$_{275}$	UU(V)
	$_{239}$ATFDAISKTY[SY]LTPDLWK$_{257}$	N(V), U(V), UU(V), UU-HPO3(V)
	$_{239}$ATFDAISKTYSYLTPDLWKETVFTK$_{263}$	UN(V), UU(V)
40S ribosomal protein S20 (P60866)	$_{50}$VKGPVR[M]PTK$_{59}$	N(I), AN(I), AU(I), ACU(III), AGU-HPO3(V)
	$_{50}$VKGPVR[M]PTKTLR$_{62}$	AU-HPO3(V), ACU-HPO3(V), ACU(V), AN(V), ACN(V), AGU-HPO3(V), AGU(V)
	$_{52}$GPVR[M]PTK$_{59}$	AN(I)
	$_{52}$GPVR[M]PTKTLR$_{62}$	ACU(III,V), AU(III), AU-HPO3(V)
	$_{84}$IHKRLIDLHSPSEIVK$_{99}$	ACU(V)
	$_{88}$LIDL[H]SPSEIVK$_{99}$	N(I),AN(I), AU(I)
	$_{47}$NLKVKGPVR[M]PTK$_{59}$	ACU(V), AGU(V), AN(V), AU-HPO3(V), ACU-HPO3(V), AGU-HPO3(V)
40S ribosomal protein S23	$_{125}$VANVSLLAL[Y]K$_{135}$	U(I)

(P62266)	$_{81}$ITAFVPNDGCLNFIEENDEVLVAGFGR$_{107}$	C-H2O(II)
40S ribosomal protein S24 (P62847)	$_{21}$[KQ]M(Oxidation)VIDVLHPGK$_{32}$	GCN(Vi)
	$_{44}$[L]AKM(Oxidation)YKTTPDVIFVFGFR$_{61}$	UU(V), CGU-HPO3(V), GU(V), CGU(V)
40S ribosomal protein S24 (P62847)	$_{47}$[MY]KTTPDVIFVFGFR$_{61}$	N(I), CN(V), CGU(V), CGU-HPO3(V), CU(V), CU-HPO3(V), GU-HPO3(V)
	$_{47}$M(Oxidation)YKTTPDVIFVFGFRTHFGGGK$_{68}$	CUU(V)
	$_{62}$THFGGGKTTGFGMIYDSLDYAK$_{83}$ $_{62}$THFGGGKTTGFGM(Oxidation)IYDSLDYAK$_{83}$	UN(I), AGG(V), CCU-HPO3(IV), CU(V), CUN(III,IV,V), CUU(III,IV.V), CUU-HPO3(IV,V), UU(V), UUU(III)
	$_{62}$THFGGGKTTGFGMIYDSLDYAKK$_{84}$ $_{62}$THFGGGKTTGFGM(Oxidation)IYDSLDYAKK$_{84}$	UU(V), UN(V), CUU(III,V), CUU-HPO3(V), CUN(V), CCU(V)
	$_{62}$THFGGGKTTGFGM(Oxidation)IYDSLDYAKKNEPK$_{88}$	CUU(V)
	$_{50}$TTPDVI[F]VFGFRTHFGGGK$_{68}$	U(V), UU(V), UN(V), CU(V), CUU(V), CUU-HPO3(V), CUN(V), GUN(V), CCN(V)
40S ribosomal protein S25 (P62851)	$_{53}$ATYDKL[C]KEVPNYK$_{66}$	N(I)
40S ribosomal protein S26 / **Putative 40S ribosomal protein S26-like 1** (P62854/Q5JNZ5)	$_{101/101}$FRPAGAAPRPPPKPM$_{115/115}$	N(I), UN(I)
	$_{35/35}$AIKK[F]VIR$_{42/42}$	AUU(III)
	$_{39/39}$FV[IR]NIVEAAAVR$_{51/51}$	GU(I), AGU(I), GGU(I)
40S ribosomal protein S3 (P23396)	$_{77}$FGFPEGSVEL[Y]AEK$_{90}$	N(I)
	$_{46}$TEIIIL[ATR]TQNVLGEK$_{62}$	CU(I)
	$_{107}$[Y]KLLGGLAVR$_{116}$	UG-NH3(I), GG-NH3(I), CG-NH3(I), AUG-NH3(I), AG-NH3(I), AGG-NH3(I), ACG-NH3(I)
40S ribosomal protein S4, X/Y isoform 1/2 (P22090/P62701/Q8TD47)	$_{52/52/52}$LK[Y]ALTGDEVKK$_{63/63/63}$	N(I)
40S ribosomal protein S4, X isoform (P62701)	$_{40}$E[C]LPLIIFLR$_{49}$	N(I)
	$_{222}$LSNIFVIG[K]GNKPWISLPR$_{240}$	GN(Vi)
40S ribosomal protein S6 (P62753)	$_{222}$[EK]RQEQIAK$_{230}$	GN(VI)
	$_{222}$EKRQEQIA[K]R$_{231}$	GN(Vi)
	$_{176}$IQRL[V]TPR$_{183}$	UU(I), CUU(III)
	$_{224}$RQEQIA[K]R$_{231}$	GN(VI), GN-HPO3(Vi)
40S ribosomal protein S7 (P62081)	$_{59}$AIII[F]VPVPQLK$_{70}$	N(I), UU(I)
	$_{100}$IL[P]KPTR$_{106}$	UN(I)
	$_{100}$ILP[K]PTRK$_{107}$	UN(I)
	$_{99}$[R]ILPKPTR$_{106}$	GU(I)
	$_{100}$RILPKPTRK$_{107}$	CUN(V), CUU(V)
	$_{119}$SRTLTAV[H]DAILEDLVFPSEIVGK$_{142}$	N(I,IV), U(I), UU(I,V), CU(V), CN(V), AA-H2O(I), UU-HPO3(IV,V), UN(I,V), AGG(V), CUN(V), CUU(I,V), CUU-HPO3(IV), GUU(I,IV)
	$_{119}$SRTLTAV[H]DAILEDLVFPSEIVGKR$_{143}$	N(V), UN(V), UU(I,V), CU(V), CN(V), CC(V), UU-HPO3(IV,V), CCN(V), CCU(V), CUN(V), CUU(V), GUU-HPO3(V)
	$_{121}$TLTAVHDA[I]LEDLVFPSEIVGK$_{142}$	N(I), UU(I), AA-H2O(I), GG-H2O(I), UN(I), UG-NH3(I), CUN(V), CUU(IV,V), UUU(V)
	$_{121}$TLTAVHDAILEDLVFPSEIVGKR$_{143}$	U(V), UN(V), CUU(V), CUN(V)
40S ribosomal protein S8 (P62241)	$_{50}$GGNKK[Y]RALR$_{59}$	GG-NH3(V)
	$_{158}$ISSLLEEQFQ[Q]GK$_{170}$	N(I), CU(I)
	$_{99}$[NC]IVLIDSTPYR$_{110}$	N(I)
	$_{111}$[QW]YESHYALPLGR$_{123}$	N(I)
40S ribosomal protein S9 (P46781)	$_{156}$HID[F]SLR$_{162}$	N(I)
	$_{11}$KTYVTP[R]RPFEK$_{22}$	AU(V), ACU(V)
	$_{12}$TYVTPR[R]PFEK$_{22}$	AU(III,V), ACU(V)
Gamma-interferon-inducible protein 16 IF16_HUMAN (Q16666)	$_{752}$SVIHS[H]IK$_{759}$	N(I)
Nucleolar protein 16 NOP16_HUMAN (Q9Y3C1)	$_{34}$IE[C]SHIR$_{40}$	N(I)
Alpha-enolase ENOA_HUMAN (P06733)	$_{427}$NFRNPLAK$_{434}$	G-H2O(I)
Centrosomal protein of 104 kDa CE104_HUMAN (O60308)	$_{701}$ALQGQLA[AL]K$_{710}$	G-NH3(I)
Zinc finger CCCH-type antiviral protein 1 ZCCHV_HUMAN (Q7Z2W4)	$_{271}$[SC]TPSPDQISHR$_{282}$	N(I)

EKC/KEOPS complex subunit LAGE3 LAGE3_HUMAN (Q14657)	$_{19}$GGH[SCRGGV]DTAAAPAGGAPPAHAPGPGR$_{47}$	N(I)
X-ray repair cross-complementing protein 5 XRCC5_HUMAN (P13010)	$_{243}$HSIHWP[C]R$_{250}$	N(I), U(I)
	$_{243}$HSIHW[PC]RLTIGSNLSIR$_{260}$	GU(V), CG(V)
Oxysterol-binding protein-related protein 6 OSBL6_HUMAN (Q9BZF3)	$_{763}$LT[F]VK$_{767}$	CN(I), CU(I), GU(I)
Dual specificity mitogen-activated protein kinase kinase 3 MP2K3_HUMAN (P46734)	$_{27}$IS[C]MSKPPAPNPTPPR$_{42}$ $_{27}$IS[C]M(Oxidation)SKPPAPNPTPPR$_{42}$	N(I)
High mobility group protein B1 HMGB1_HUMAN (P09429)	$_{97}$RPPSAFFLF[C]SEYRPK$_{112}$	N(I)
Ankyrin repeat domain-containing protein 17/ **Ankyrin repeat and KH domain-containing protein 1** ANR17/ANKH1_HUMAN (O75179/Q8IWZ3)	$_{1308/1280}$GADVNAP[P]VPSSR$_{1320/1292}$	UU(I)
Probable tRNA pseudouridine synthase 1 TRUB1_HUMAN (Q8WWH5)	$_{149}$[Y]TAIGELGK$_{157}$	CU(I), N(I), CUU(III)
Kelch domain-containing protein 4 KLDC4_HUMAN (Q8TBB5)	$_{362}$KEEPEGGSR[PACGGA]GTQGPVQLVK$_{386}$	N(I)
Catenin delta-1 CTND1_HUMAN (O60716)	$_{600}$YQEAAPNVANNTGPHAAS[C]FGAK$_{622}$	N(I)
Cytoskeleton-associated protein 5 CKAP5_HUMAN (Q14008)	$_{1103}$FQPASAPAED[C]ISSSTEPKPDPK$_{1125}$	N(I)
Microtubule-associated protein 4 MAP4_HUMAN (P27816)	$_{634}$[KC]SLPAEEDSVLEK$_{647}$	N(I)
Aprataxin APTX_HUMAN (Q7Z2E3)	$_{141}$DAAQEAEAGTGLEPGSNSGQ[C]SVPLKK$_{167}$	N(I)
Signal recognition particle 14 kDa protein SRP14_HUMAN (P37108)	$_{22}$TSGSV[Y]ITLK$_{31}$	N(I)
Survival of motor neuron-related-splicing factor 30 SPF30_HUMAN (O75940)	$_{209}$VGVGT[C]GIADKPMTQYQDTSK$_{229}$	N(I)
Protein LTV1 homolog LTV1_HUMAN (Q96GA3)	$_{402}$IQ[M]INGSDLPK$_{412}$	N(I), GU(III)
Methionine aminopeptidase 2 MAP2_HUMAN (P50579)	$_{386}$NFDVG[H]VPIR$_{395}$	N(I)
Staphylococcal nuclease domain-containing protein 1 SND1_HUMAN (Q7KZF4)	$_{415}$VNVTVDYIRPASPATETV[PAF]SER$_{438}$	N(I)
	$_{26}$MVLSG[C]AIIVR$_{36}$	N(I)
E3 ubiquitin/ISG15 ligase TRIM25 TRI25_HUMAN (Q14258)	$_{316}$GISTKPV[Y]IPEVELNHK$_{332}$	N(I), U(I), CU(I,V), UN(I), UU(I), GU(I,V), CN(V), GU-HPO3(V), AN(I)
	$_{314}$LRGISTKPV[Y]IPEVELNHK$_{332}$	N(I), CN(V), CU(V), GU(V)
	$_{470}$VALS[ECY]TVASVAEMPQNYRPHPQR$_{494}$	N(I)
Glutamate-rich WD repeat-containing protein 1 GRWD1_HUMAN (Q9BQ67)	$_{9}$[RTCETGE]PMEAESGDTSSEGPAQVYLPGR$_{37}$	N(I)
	$_{10}$[TCE]TGEPMEAESGDTSSEGPAQVYLPGR$_{37}$	N(I)
DNA-dependent protein kinase catalytic subunit PRKDC_HUMAN (P78527)	$_{15}$LQETLSAAD[RCGA]ALAGHQLIR$_{36}$	N(I)
Echinoderm microtubule-associated protein-like 4 EMAL4_HUMAN (Q9HC35)	$_{76}$AVIPMS[C]ITNGSGANR$_{91}$	N(I)
2'-5'-oligoadenylate synthase 3 OAS3_HUMAN (Q9Y6K5)	$_{646}$QD[C]FNMAQGFR$_{656}$	N(I)
Ras GTPase-activating protein-binding protein 2 G3BP2_HUMAN (Q9UN86)	$_{327}$YPDSHQL[F]VGNLPHDIDENELK$_{348}$	N(I)
	$_{324}$IIRYPDSHQL[F]VGNLPHDIDENELK$_{348}$	N(I)
	$_{371}$LPNFG[F]VVFDDSEPVQR$_{387}$	N(I)
FLYWCH family member 2 FWCH2_HUMAN (Q96CP2)	$_{60}$[KGVHCV]MSLGVPGPATLAK$_{78}$	N(I)
	$_{61}$[GVHCV]MSLGVPGPATLAK$_{78}$	N(I)
Obg-like ATPase 1 OLA1_HUMAN (Q9NTK5)	$_{341}$GFI[M]AEVMK$_{349}$	N(I)
	$_{36}$STFFNVLTNSQASAENFPFCTIDPNESR$_{63}$	C-H2O(II)
Uncharacterized protein C7orf50 CG050_HUMAN (Q9BRJ6)	$_{97}$SGAELALDYL[C]R$_{108}$	N(I)
	$_{155}$ARELTVQ[K]AEALMR$_{168}$	AGU-HPO3(V), ACG-NH3(V)
Actin ACTB/ACTA/ACTG/ACTH/ACTC/ACTS/ ACTBL_HUMAN (P60709/P62736/P63261/P63267/P68032/P68133/Q562R1)	$_{327}$[I]KIIAPPER$_{335}$	GN(I)
Signal recognition particle 9 kDa protein SRP09_HUMAN (P49458)	$_{42}$VTDDLV[C]LVYKTDQAQDVK$_{60}$	N(I)
	$_{42}$VTDDLV[C]LVYK$_{52}$	N(I,V), U(I), AN(I)
X-ray repair cross-complementing protein 6 XRCC6_HUMAN (P12956)	$_{400}$YTPRRNIPPYFVALVPQEEELDDQK$_{424}$	CU(I), GU(I,V)
	$_{75}$IISS[DRD]LLAVVFYGTEK$_{92}$	CU(I)
Far upstream element-binding protein 3 FUBP3_HUMAN (Q96I24)	$_{344}$GDWSVGAPGGVQEITYTVPADK[C]GLVIGK$_{372}$	N(I,V), U(I,V), AN(I,V), UU(I,V), UU-HPO3(V), AUU-HPO3(V), AU(V), CUU(V), ACN-HPO3(V), UUU-HPO3(V), ACU(V), AUU(V), AUN(V)
Helicase-like transcription factor HLTF_HUMAN (Q14527)	$_{434}$VIEDVAFA[C]ALTSSVPTTK$_{452}$	N(I)
YY1-associated protein 1 YYAP1_HUMAN (Q9H869)	$_{587}$[CIK]PAPVIHPASVIFTVPATTVK$_{610}$	N(I)
DNA-(apurinic or apyrimidinic site) lyase APEX1_HUMAN (P27695)	$_{99}$[CSENKL]PAELQELPGLSHQYWSAPSDK$_{125}$	N(I)
Ubiquitin-conjugating enzyme E2 N UBE2N_HUMAN (P61088)	$_{54}$LELFLPEEY[PMA]APK$_{68}$	N(I)
Tubulin TBB5/TBB4B/TBB3/TBB2A/TBB2B_HUMAN	$_{104}$[GHY]TEGAELVDSVLDVVR$_{121}$	N(I)

(P07437/P68371/Q13509/Q13885/Q9BVA1)		
Endoplasmic reticulum chaperone BiP BIP_HUMAN (P11021)	$_{511}$VTAEDKGTGNKNKITITNDQNR$_{532}$	A-H2O(II), G-H2O(II)
Zinc finger Ran-binding domain-containing protein 2 ZRAB2_HUMAN (O95218)	$_{138}$AVGPASILKEVEDKESEGEEEDEDEDLSK$_{166}$	G-H3PO4(II)
Fructose-bisphosphate aldolase A/C ALDOA/ALDOC_HUMAN (P04075/P09972)	$_{174/174}$YASICQQNGIVPIVEPEILPDGDHDLKR$_{201/201}$	C-H2O(II)
Desmoplakin DESP_HUMAN (P15924)	$_{104}$SRELDECFAQANDQMEILDSLIR$_{126}$	C-H2O(II)
Y-box-binding protein 1/Y-box-binding protein 3 YBOX1/YBOX3_HUMAN (P67809/ P16989)	$_{78/110}$NDTKEDV[F]VHQTAIKK$_{93/125}$	CU(I,II)
Y-box-binding protein 1/ Y-box-binding protein 2/ Y-box-binding protein 3 YBOX1/YBOX2/YBOX3 (P67809/Q9Y2T7/P16989)	$_{78/114/110}$NDTKEDV[F]VHQTAIK$_{92/127/124}$	CU(I,II)
Annexin A7 ANXA7_HUMAN (P20073)	$_{322}$LLVSMCQGNRDENQSINHQMAQEDAQR$_{348}$	C-H2O(II)
Elongation factor 1-delta EF1D_HUMAN (P29692)	$_{242}$KLQIQCVVEDDKVGTDLLEEEITK$_{265}$	C-H2O(II)
14-3-3 protein sigma 1433S_HUMAN (P31947)	$_{88}$VETELQGV[C]DTVLGLLDSHLIKEAGDAESR$_{117}$	C-H2O((II)
Heat shock 70 kDa protein 4 HSP74_HUMAN (P34932)	$_{276}$LMSANASDLPLSIE[C]FMNDVDVSGTMNR$_{303}$	C-H2O(II)
Replication factor C subunit 2 RFC2_HUMAN (P35250)	$_{1}$MEVEAVCGGAGEVEAQDSDPAPAFSK$_{26}$	A(II)
Cytosolic Fe-S cluster assembly factor NUBP1 NUBP1_HUMAN (P53384)	$_{1}$M(Oxidation)EEVPHDCPGADSAQAGR$_{18}$	A(II)
Heterogeneous nuclear ribonucleoprotein U HNRPU_HUMAN (Q00839)	$_{638}$GNFTLPEVAECFDEITYVELQKEEAQK$_{664}$	C-H2O(II)
Chromobox protein homolog 3 CBX3_HUMAN (Q13185)	$_{53}$GFTDADNTWEPEENLD[C]PELIEAFLNSQK$_{81}$	C-H2O(II)
ATP-binding cassette sub-family F member 1 ABCF1_HUMAN (Q8NE71)	$_{353}$ALSIPPNIDVLL[C]EQEVVADETPAVQAVLR$_{382}$	C-H2O(II)
Glucosamine-6-phosphate isomerase 2 GNPI2_HUMAN (Q8TDQ7)	$_{96}$HIDIDPNNAHILDGNAADLQ[AEC]DAFENK$_{124}$	C-H2O(II)
NudC domain-containing protein 2 NUDC2_HUMAN (Q8WVJ2)	$_{95}$DAAN[CWTSL]LESEYAADPWVQDQMQR$_{120}$	C-H2O(II)
PDZ and LIM domain protein 5 PDLI5_HUMAN (Q96HC4)	$_{199}$TAVNVPRQPTVTSVCSETSQELAEGQR$_{225}$	C-H2O(II)
SRSF protein kinase 1 SRPK1_HUMAN (Q96SB4)	$_{346}$DTEGGAAEINCNGVIEVINYTQNSNNETLR$_{375}$	C-H2O(II,III)
Tubulin beta-6 chain TBB6_HUMAN (Q9BUF5)	$_{1}$M(Oxidation)REIVHIQAGQCGNQIGTK$_{19}$ $_{1}$MREIVHIQAGQ[CGN]QIGTK$_{19}$	C-H2O(II)
Pre-mRNA-processing factor 19 PRP19_HUMAN (Q9UMS4)	$_{207}$YRQVASHVGLHSASIPGILALDLCPSDTNK$_{236}$	C-H2O(II)
Phosphatidylethanolamine-binding protein 1 PEBP1_HUMAN (P30086)	$_{81}$YREWHHFLVVNMK$_{93}$ $_{81}$YREWHHFLV[V]NM(Oxidation)K$_{93}$	CN(V), CU(V), GU(V)
DNA topoisomerase 3-alpha/beta-1 TOP3A/TOP3B_HUMAN (Q13472/O95985)	$_{354/328}$LYTQGYIS[Y]PR$_{364/338}$	A-H2O(VI)

Bibliography

1. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., and Matsuo, T. (1988). Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Commun. Mass Spectrom. *2*, 151–153.

2. Ryan, D.J., Spraggins, J.M., and Caprioli, R.M. (2019). Protein identification strategies in MALDI imaging mass spectrometry: a brief review. Curr. Opin. Chem. Biol. *48*, 64–72.

3. Singhal, N., Kumar, M., Kanaujia, P.K., and Virdi, J.S. (2015). MALDI-TOF mass spectrometry: An emerging technology for microbial identification and diagnosis. Front. Microbiol. *6*, 1–16.

4. Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., and Whitehouse, C.M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. Science (80-.). *246*, 64–71.

5. Thomson, B.A., and Iribarne, J. V. (1979). Field induced ion evaporation from liquid surfaces at atmospheric pressure. J. Chem. Phys. *71*, 4451–4463.

6. Dole, M., Hines, R.L., Mack, L.L., Mobley, R.C., Ferguson, L.D., and Alice, M.B. (1968). Gas phase macroions. Macromolecules *1*, 96–97.

7. Wilm, M. (2011). Principles of electrospray ionization. Mol. Cell. Proteomics *10*, M111.009407.

8. Karas, M., Bahr, U., and Dülcks, T. (2000). Nano-electrospray ionization mass spectrometry: Addressing analytical problems beyond routine. Fresenius. J. Anal. Chem. *366*, 669–676.

9. Wilm, M., and Mann, M. (1996). Analytical properties of the nanoelectrospray ion source. Anal. Chem. *68*, 1–8.

10. Mirzaei, H., and Carrasco, M. (2016). Modern Proteomics – Sample Preparation, Analysis and Practical Applications. Mod. Proteomics - Sample Prep. Anal. Pract. Appl. *919*, 43–62. Available at: http://link.springer.com/10.1007/978-3-319-41448-5 [Accessed June 10, 2020].

11. Douglas, D.J., Frank, A.J., and Mao, D. (2005). Linear ion traps in mass spectrometry. Mass Spectrom. Rev. *24*, 1–29.

12. Zubarev, R.A., and Makarov, A. (2013). Orbitrap mass spectrometry. Anal. Chem. *85*, 5288–5296.

13. Johnson, A.R., and Carlson, E.E. (2015). Collision-Induced Dissociation Mass Spectrometry: A Powerful Tool for Natural Product Structure Elucidation. Anal. Chem. *87*, 10668–10678.

14. Wysocki, V.H., Tsaprailis, G., Smith, L.L., and Breci, L.A. (2000). Mobile and localized protons: A framework for understanding peptide dissociation. J. Mass Spectrom. *35*, 1399–1406.

15. Paizs, B., and Suhai, S. (2005). Fragmentation pathways of protonated peptides. Mass Spectrom. Rev. *24*, 508–48. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15389847 [Accessed July 14, 2014].

16. Breci, L.A., Tabb, D.L., Yates, J.R., and Wysocki, V.H. (2003). Cleavage N-terminal to proline: Analysis of a database of peptide tandem mass spectra. Anal. Chem. *75*, 1963–1971.

17. Farrugia, J.M., O'Hair, R.A.J., and Reid, G.E. (2001). Do all b2 ions have oxazolone structures? Multistage mass spectrometry and ab initio studies on protonated N-acyl amino acid methyl ester model systems. Int. J. Mass Spectrom. *210–211*, 71–87.

18. Harrison, A.G. (2003). Fragmentation reactions of protonated peptides containing glutamine or glutamic acid. J. Mass Spectrom. *38*, 174–187.

19. Rožman, M. (2007). Aspartic Acid Side Chain Effect-Experimental and Theoretical Insight. J. Am. Soc. Mass Spectrom. *18*, 121–127.

20. Ambihapathy, K., Yalcin, T., Leung, H.W., and Harrison, A.G. (1997). Pathways to immonium ions in the fragmentation of protonated peptides. J. Mass Spectrom. *32*, 209–215.

21. Roepstorff, P., and Fohlman, J. (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed. Mass Spectrom. *11*, 601.

22. Wu, J., and McLuckey, S.A. (2004). Gas-phase fragmentation of oligonucleotide ions. Int. J. Mass Spectrom. *237*, 197–241.

23. Schürch, S., Bernal-Méndez, E., and Leumann, C.J. (2002). Electrospray tandem mass spectrometry of mixed-sequence RNA/DNA oligonucleotides. J. Am. Soc. Mass Spectrom. *13*, 936–945.

24. Mcluckey, S.A., Van Berkel, G.J., and Glish, G.L. (1992). Tandem Mass Spectrometry of Small, Multiply Charged Oligonucleotides. J. Am. Soc. Mass Spectrom. *3*, 60–70.

25. Steen, H., and Mann, M. (2004). The ABC's (and XYZ's) of peptide sequencing. Nat. Rev. Mol. Cell Biol. *5*, 699–711. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15340378.

26. Nesvizhskii, A.I. (2007). Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol. Biol. *367*, 87–119.

27. Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins : modular design for efficient function. *8*, 479–490.

28. Hentze, M.W., Castello, A., Schwarzl, T., and Preiss, T. (2018). REVIEWS A brave new world of RNA-binding proteins. Nat. Publ. Gr. *19*, 327–341. Available at: http://dx.doi.org/10.1038/nrm.2017.130.

29. Castello, A., Hentze, M.W., and Preiss, T. (2015). Metabolic Enzymes Enjoying New Partnerships as RNA-Binding Proteins. Trends Endocrinol. Metab. *26*, 746–757.

30. He, C., Sidoli, S., Warneford-Thomson, R., Tatomer, D.C., Wilusz, J.E., Garcia, B.A., and Bonasio, R. (2016). High-Resolution Mapping of RNA-Binding Regions in the Nuclear Proteome of Embryonic Stem Cells. Mol. Cell *64*, 416–430.

31. Matia-González, A.M., Laing, E.E., and Gerber, A.P. (2015). Conserved mRNA-binding proteomes in eukaryotic organisms. Nat. Struct. Mol. Biol. *22*, 1027–1033.

32. Liao, Y., Castello, A., Fischer, B., Leicht, S., Föehr, S., Frese, C.K., Ragan, C., Kurscheid, S., Pagler, E., Yang, H., *et al.* (2016). The Cardiomyocyte RNA-Binding Proteome: Links to Intermediary Metabolism and Heart Disease. Cell Rep. *16*, 1456–1469.

33. Boeynaems, S., Alberti, S., Fawzi, N.L., Mittag, T., Polymenidou, M., Rousseau, F., Schymkowitz, J., Shorter, J., Wolozin, B., Van Den Bosch, L., *et al.* (2018). Protein Phase Separation: A New Phase in Cell Biology. Trends Cell Biol. *28*, 420–435.

34. Hellman, L.M., and Fried, M.G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nat. Protoc. *2*, 1849–1861.

35. Jones, S. (2016). Protein–RNA interactions: structural biology and computational modeling techniques. Biophys. Rev. *8*, 359–367.

36. Ramanathan, M., Porter, D.F., and Khavari, P.A. (2019). Methods to study RNA–protein interactions. Nat. Methods *16*, 225–234. Available at: http://dx.doi.org/10.1038/s41592-019-0330-1.

37. Beckstead, A.A., Zhang, Y., Vries, S. De, and Kohler, B. (2016). Life in the light : nucleic acid photoproperties as a. 24228–24238.

38. Wurtmann, E.J., and Wolin, S.L. (2009). RNA under attack: Cellular handling of RNA damage RNA under attack: Cellular handling of RNA damage E. J. Wurtmann et.al. Crit. Rev. Biochem. Mol. Biol. *44*, 34–49.

39. Kramer, K., Sachsenberg, T., Beckmann, B.M., Qamar, S., Boon, K.-L., Hentze, M.W., Kohlbacher, O., and Urlaub, H. (2014). Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. Nat. Methods *11*, 1064–1070. Available at: http://dx.doi.org/10.1038/nmeth.3092.

40. Shetlar, M.D., Carbone, J., Steady, E., and Hom, K. (1984). Photochemical addition of amino acids and peptides to polyuridylic acid. Photochem. Photobiol. *39*, 141–144.

41. Smith, K.C. (1969). Photochemical addition of amino acids to 14C-uracil. Biochem. Biophys. Res. Commun. *34*, 354–357.

42. Sharma, K., Hrle, A., Kramer, K., Sachsenberg, T., Staals, R.H.J., Randau, L., Marchfelder, A., van der Oost, J., Kohlbacher, O., Conti, E., *et al.* (2015). Analysis of protein-RNA interactions in CRISPR proteins and effector complexes by UV-induced cross-linking and mass spectrometry. Methods *89*, 138–148.

43. Shchepachev, V., Bresson, S., Spanos, C., Petfalski, E., Fischer, L., Rappsilber, J., and Tollervey, D. (2019). Defining the RNA interactome by total RNA -associated protein purification . Mol. Syst. Biol. *15*, 1–23.

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

45. Varghese, A.J. (1974). Photoaddition products of uracil and cysteine. BBA Sect. Nucleic Acids Protein Synth. *374*, 109–114.

46. Varghese, A.J. (1976). Photochemical Addition of Amino Acids and Related Compounds to Nucleic Acid Constituents. In Aging, Carcinogenesis, and Radiation Biology (Springer US), pp. 207–223.

47. Baltz, A.G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., *et al.* (2012). The mRNA-Bound Proteome and Its Global

Occupancy Profile on Protein-Coding Transcripts. Mol. Cell *46*, 674–690.

48. Castello, A., Horos, R., Strein, C., Fischer, B., Eichelbaum, K., Steinmetz, L.M., Krijgsveld, J., and Hentze, M.W. (2013). System-wide identification of RNA-binding proteins by interactome capture. Nat. Protoc. *8*, 491–500. Available at: http://www.nature.com/doifinder/10.1038/nprot.2013.020.

49. Kwon, S.C., Yi, H., Eichelbaum, K., Föhr, S., Fischer, B., You, K.T., Castello, A., Krijgsveld, J., Hentze, M.W., and Kim, V.N. (2013). The RNA-binding protein repertoire of embryonic stem cells. Nat. Struct. Mol. Biol. *20*, 1122–1130.

50. Itri, F., Monti, D.M., Della Ventura, B., Vinciguerra, R., Chino, M., Gesuele, F., Lombardi, A., Velotta, R., Altucci, C., Birolo, L., *et al.* (2016). Femtosecond UV-laser pulses to unveil protein-protein interactions in living cells. Cell. Mol. Life Sci. *73*, 637–648.

51. Castello, A., Frese, C.K., Fischer, B., Järvelin, A.I., Horos, R., Alleaume, A.M., Foehr, S., Curk, T., Krijgsveld, J., and Hentze, M.W. (2017). Identification of RNA-binding domains of RNA-binding proteins in cultured cells on a system-wide scale with RBDmap. Nat. Protoc. *12*, 2447–2464. Available at: http://dx.doi.org/10.1038/nprot.2017.106.

52. Mullari, M., Lyon, D., Jensen, L.J., and Nielsen, M.L. (2017). Specifying RNA-Binding Regions in Proteins by Peptide Cross-Linking and Affinity Purification. J. Proteome Res. *16*, 2762–2772.

53. Panhale, A., Richter, F.M., Ramírez, F., Shvedunova, M., Manke, T., Mittler, G., and Akhtar, A. (2019). CAPRI enables comparison of evolutionarily conserved RNA interacting regions. Nat. Commun. *10*. Available at: http://dx.doi.org/10.1038/s41467-019-10585-3.

54. Bae, J.W., Kwon, S.C., Na, Y., Kim, V.N., and Kim, J.S. (2020). Chemical RNA digestion enables robust RNA-binding site mapping at single amino acid resolution. Nat. Struct. Mol. Biol. *27*, 678–682.

55. Veit, J., Sachsenberg, T., Chernev, A., Aicheler, F., Urlaub, H., and Kohlbacher, O. (2016). LFQProfiler and RNP^{xl}: Open-Source Tools for Label-Free Quantification and Protein-RNA Cross-Linking Integrated into Proteome Discoverer. J. Proteome Res. *15*.

56. Liu, S., Zhang, C., Campbell, J.L., Zhang, H., Yeung, K.K.C., Han, V.K.M., and Lajoie, G.A. (2005). Formation of phosphopeptide-metal ion complexes in liquid chromatography/electrospray mass spectrometry and their influence on phosphopeptide detection. Rapid Commun. Mass Spectrom. *19*, 2747–2756.

57. Urlaub, H., Kruft, V., Bischof, O., Müller, E.C., and Wittmann-Liebold, B. (1995). Protein-rRNA binding features and their structural and functional implications in ribosomes as determined by cross-linking studies. EMBO J. *14*, 4578–4588.

58. Kramer, K., Sachsenberg, T., Beckmann, B.M., Qamar, S., Boon, K.L., Hentze, M.W., Kohlbacher, O., and Urlaub, H. (2014). Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. Nat. Methods *11*, 1064–1070.

59. Kühn-Hölsken, E., Dybkov, O., Sander, B., Lührmann, R., and Urlaub, H. (2007). Improved identification of enriched peptide-RNA cross-links from ribonucleoprotein particles (RNPs) by mass spectrometry. Nucleic Acids Res. *35*.

60. Richter, F.M., Hsiao, H.H., Plessmann, U., and Urlaub, H. (2009). Enrichment of protein-RNA crosslinks from crude UV-irradiated mixtures for MS analysis by on-line chromatography using titanium dioxide columns. Biopolymers *91*, 297–309.

61. Kramer, K. (2013). Investigation of protein – RNA interactions by UV cross-linking and mass spectrometry : methodological improvements toward in vivo applications Dissertation. PhD Thesis Univ. Göttingen.

62. Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., and Bryant, S.H. (2004). Open mass spectrometry search algorithm. J. Proteome Res. *3*, 958–964.

63. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. Mol. Syst. Biol. *2*.

64. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., *et al.* (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.

65. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., and Bairoch, A. (2005). The Proteomics Protocols Handbook - Chapter 52: Protein Identification and Analysis Tools on the ExPASy Server.

66. Neuhoff, V., Arold, N., Taube, D., and Ehrhardt, W. (1988). Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity

using Coomassie Brilliant Blue G-250 and R-250. Electrophoresis *9*, 255–262.

67. Blum, H., Beier, H., and Gross, H.J. (1987). Improved silver staining of plant proteins, RNA and DNA in polyacrylamide gels. Electrophoresis *8*, 93–99.

68. Macromolecular Components of E. coli and HeLa Cells | Thermo Fisher Scientific - DE Available at: https://www.thermofisher.com/de/en/home/references/ambion-tech-support/rna-tools-and-calculators/macromolecular-components-of-e.html [Accessed July 6, 2020].

69. Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat. Protoc. *2*, 1896–1906.

70. Hillen, H.S., Parshin, A.V., Agaronyan, K., Morozov, Y.I., Graber, J.J., Chernev, A., Schwinghammer, K., Urlaub, H., Anikin, M., Cramer, P., *et al.* (2017). Mechanism of Transcription Anti-termination in Human Mitochondria. Cell *171*.

71. Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., and MacCoss, M.J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods *4*, 923–925.

72. Kapp, E.A., Schütz, F., Reid, G.E., Eddes, J.S., Moritz, R.L., O'Hair, R.A.J., Speed, T.P., and Simpson, R.J. (2003). Mining a Tandem Mass Spectrometry Database to Determine the Trends and Global Factors Influencing Peptide Fragmentation. Anal. Chem. *75*, 6251–6264.

73. Goloborodko, A.A., Levitsky, L.I., Ivanov, M. V., and Gorshkov, M. V. (2013). Pyteomics - A python framework for exploratory data analysis and rapid software prototyping in proteomics. J. Am. Soc. Mass Spectrom. *24*, 301–304.

74. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York Available at: http://link.springer.com/10.1007/978-0-387-98141-3.

75. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

76. Green-Church, K.B., and Limbach, P.A. (2000). Mononucleotide gas-phase proton affinities as determined by the kinetic method. J. Am. Soc. Mass Spectrom. *11*, 24–32.

77. Röst, H.L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.C., Gutenbrunner, P., Kenar, E., *et al.* (2016). OpenMS: A flexible open-source software platform for mass spectrometry data analysis. Nat. Methods *13*, 741–748.

78. Sachsenberg, T. (2017). Computational Methods for Mass Spectrometry-based Study of Protein-RNA or Protein-DNA Complexes and Quantitative Metaproteomics. PhD Thesis Univ. Tübingen.

79. Sharma, K. (2015). Investigation of prokaryotic immune defense system with quantitative and structural mass spectrometry. PhD Thesis Univ. Göttingen.

80. Budowsky, E.I., Simukova, N.A., Turchinsky, M.F., Boni, I. V, and Skoblov, Y.M. (1976). Induced formation of covalent bonds between nucleoprotein components. V. UV or bisulfite induced polynucleotide-protein crosslinkage in bacteriophage MS2. Nucleic Acids Res. *3*, 261–276.

81. Wei, Y., Chen, L., Chen, J., Ge, L., and He, R.Q. (2009). Rapid glycation with D-ribose induces globular amyloid-like aggregations of BSA with high cytotoxicity to SH-SY5Y cells. BMC Cell Biol. *10*, 10.

82. Privat, E., and Sowers, L.C. (1996). Photochemical deamination and demethylation of 5-methylcytosine. Chem. Res. Toxicol. *9*, 745–750.

83. Shabarova, Z., and Bogdanov, A. (1994). Advanced Organic Chemistry of Nucleic Acids (Wiley) Available at: https://onlinelibrary.wiley.com/doi/book/10.1002/9783527615933 [Accessed July 6, 2020].

84. Hillen, H.S., Parshin, A. V., Agaronyan, K., Morozov, Y.I., Graber, J.J., Chernev, A., Schwinghammer, K., Urlaub, H., Anikin, M., Cramer, P., *et al.* (2017). Mechanism of Transcription Anti-termination in Human Mitochondria. Cell *171*, 1082-1093.e13. Available at: https://doi.org/10.1016/j.cell.2017.09.035.

85. Hermanson, G.T. (2013). Nucleic Acid and Oligonucleotide Modification and Conjugation. Bioconjugate Tech., 959–987. Available at: http://linkinghub.elsevier.com/retrieve/pii/B9780123822390000236.

86. Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

87. Koubek, J., Lin, K.F., Chen, Y.R., Cheng, R.P., and Huang, J.J.T. (2013). Strong anion-exchange fast performance liquid chromatography as a versatile tool for preparation and purification of RNA produced by in vitro transcription. RNA *19*, 1449–1459.

88. Koo, K., Foegeding, P.M., and Swaisgood, H.E. (1998). Isolation of RNA and DNA fragments using diatomaceous earth. Biotechnol. Tech. *12*, 549–552.

89. Salvo-Chirnside, E., Kane, S., and Kerr, L.E. (2011). Protocol: High throughput silica-based purification

of RNA from Arabidopsis seedlings in a 96-well format. Plant Methods *7*, 40.

90. Beckmann, B.M. (2017). RNA interactome capture in yeast. Methods *118–119*, 82–92. Available at: http://dx.doi.org/10.1016/j.ymeth.2016.12.008.

91. Duffy, E.E., Schofield, J.A., and Simon, M.D. (2019). Gaining insight into transcriptome-wide RNA population dynamics through the chemistry of 4-thiouridine. Wiley Interdiscip. Rev. RNA *10*, 1–18.

92. Bezerra, R., and Favre, A. (1990). In vivo incorporation of the intrinsic photolabel 4-thiouridine into Escherichia coli RNAs. Biochem. Biophys. Res. Commun. *166*, 29–37.

93. Turnbough, C.L., and Switzer, R.L. (2008). Regulation of Pyrimidine Biosynthetic Gene Expression in Bacteria: Repression without Repressors. Microbiol. Mol. Biol. Rev. *72*, 266–300.

94. Kramer, K., Hummel, P., Hsiao, H.H., Luo, X., Wahl, M., and Urlaub, H. (2011). Mass-spectrometric analysis of proteins cross-linked to 4-thio-uracil- and 5-bromo-uracil-substituted RNA. Int. J. Mass Spectrom. *304*, 184–194.

95. Liu, B., Zuo, Y., and Steitz, T.A. (2016). Structures of E. coli σS-transcription initiation complexes provide new insights into polymerase mechanism. Proc. Natl. Acad. Sci. U. S. A. *113*, 4051–4056.

96. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., *et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*.

97. Frazão, C., McVey, C.E., Amblar, M., Barbas, A., Vonrhein, C., Arraiano, C.M., and Carrondo, M.A. (2006). Unravelling the dynamics of RNA degradation by ribonuclease II and its RNA-bound complex. Nature *443*, 110–114.

98. Said, N., Krupp, F., Anedchenko, E., Santos, K.F., Dybkov, O., Huang, Y.H., Lee, C.T., Loll, B., Behrmann, E., Bürger, J., *et al.* (2017). Structural basis for λN-dependent processive transcription antitermination. Nat. Microbiol. *2*, 1–13.

99. Numata, T., Ikeuchi, Y., Fukai, S., Suzuki, T., and Nureki, O. (2006). Snapshots of tRNA sulphuration via an adenylated intermediate. Nature *442*, 419–424.

100. Natchiar, S.K., Myasnikov, A.G., Kratzat, H., Hazemann, I., and Klaholz, B.P. (2017). Visualization of chemical modifications in the human 80S ribosome structure. Nature *551*, 472–477. Available at: http://dx.doi.org/10.1038/nature24482.

101. Xu, D., Song, R., Wang, G., Jeyabal, P.V.S., Weiskoff, A.M., Ding, K., and Shi, Z.Z. (2016). Obg-like ATPase 1 regulates global protein serine/threonine phosphorylation in cancer cells by suppressing the GSK3β- inhibitor 2-PP1 positive feedback loop. Oncotarget *7*, 3427–3439.

102. Preissler, S., Rato, C., Chen, R., Antrobus, R., Ding, S., Fearnley, I.M., and Ron, D. (2015). AMPylation matches BiP activity to client protein load in the endoplasmic reticulum. Elife *4*, 1–33.

103. Ahmad, M., Shen, W., Li, W., Xue, Y., Zou, S., Xu, D., and Wang, W. (2017). Topoisomerase 3ß is the major topoisomerase for mRNAs and linked to neurodevelopment and mental dysfunction. Nucleic Acids Res. *45*, 2704–2713.

104. Creasy, D.M., and Cottrell, J.S. (2004). Unimod: Protein modifications for mass spectrometry. Proteomics *4*, 1534–1536.

105. Yang, X.J., Zhu, H., Mu, S.R., Wei, W.J., Yuan, X., Wang, M., Liu, Y., Hui, J., and Ying Huang, X. (2019). Crystal structure of a Y-box binding protein 1 (YB-1)–RNA complex reveals key features and residues interacting with RNA. J. Biol. Chem. *294*, 10998–11010.

106. Bousset, L., Mary, C., Brooks, M.A., Scherrer, A., Strub, K., and Cusack, S. (2014). Crystal structure of a signal recognition particle Alu domain in the elongation arrest conformation. Rna *20*, 1955–1962.

107. Shetlar, M.D., Hom, K., Carbone, J., Moy, D., Steady, E., and Watanabe, M. (1984). Photochemical addition of amino acids and peptides to homopolyribonucleotides of the major DNA bases. Photochem. Photobiol. *39*, 135–140.

108. Hom, K., Strahan, G., and Shetlar, M.D. (2000). Ring opening photoreactions of cytosine and uracil with ethylamine. Photochem. Photobiol. *71*, 243–253.

109. Shapiro, R., and Weisgras, J.M. (1970). Bisulfite-catalyzed transamination of cytosine and cytidine. Biochem. Biophys. Res. Commun. *40*, 839–843.

110. Shapiro, R. (1976). Addition of Amino Acids and Related Substances to Nucleic Acids by Nucleophilic Catalysis. Aging, Carcinog. Radiat. Biol., 225–242.

111. Piperi, C., Farmaki, E., Vlastos, F., Papavassiliou, A.G., and Martinet, N. (2008). DNA methylation signature analysis: How easy is it to perform? J. Biomol. Tech. *19*, 281–284.

112. Clark, S.J., Smallwood, S.A., Lee, H.J., Krueger, F., Reik, W., and Kelsey, G. (2017). Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). Nat. Protoc. *12*, 534–547.

113. Johns, H.E., LeBlanc, J.C., and Freeman, K.B. (1965). Reversal and deamination rates of the main ultraviolet photoproduct of cytidylic acid. J. Mol. Biol. *13*, 849–861.

114. Marfey, P., and Kantesaria, P. (1975). The effect of sodium bisulfite on the melting profiles of nucleic acids and on their respective nucleosides. Physiol. Chem. Phys. *7*, 167–175.

115. Hayatsu, H., Wataya, Y., Kai, K., and Iida, S. (1970). Reaction of Sodium Bisulfite with Uracil, Cytosine, and Their Derivatives. Biochemistry *9*, 2858–2865.

116. Shetlar, M.D. (1980). Cross-Linking of Proteins to Nucleic Acids by Ultraviolet Light. In Photochemical and Photobiological Reviews (Springer US), pp. 105–197.

117. Shetlar, M.D., Hom, K., and Venditto, V.J. (2013). Photohydrate-mediated reactions of uridine, 2′-deoxyuridine and 2′-deoxycytidine with amines at near neutral pH. Photochem. Photobiol. *89*, 869–877.

118. Saito, I., and Matsuura, T. (1985). Chemical Aspects of UV-Induced Cross-Linking of Proteins to Nucleic Acids. Photoreactions with Lysine and Tryptophan. Acc. Chem. Res. *18*, 134–141.

119. Saito, I., and Matsuura, T. (1985). Chemical Aspects of UV-Induced Cross-Linking of Proteins to Nucleic Acids. Photoreactions with Lysine and Tryptophan. Acc. Chem. Res. *18*, 134–141.

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

121. Favre, A., Saintomé, C., Fourrey, J.L., Clivio, P., and Laugâa, P. (1998). Thionucleobases as intrinsic photoaffinity probes of nucleic acid structure and nucleic acid-protein interactions. J. Photochem. Photobiol. B Biol. *42*, 109–124.

122. Smith, K.C., and Meun, D.H.C. (1968). Kinetics of the Photochemical Addition of [35S] Cysteine to Polynucleotides and Nucleic Acids. Biochemistry *7*, 1033–1037.

123. Smith, K.C. ed. (1980). Photochemical and Photobiological Reviews (Boston, MA: Springer US) Available at: http://link.springer.com/10.1007/978-1-4684-3641-9 [Accessed July 8, 2020].

124. Varghese, A.J. (1973). Properties of Photoaddition Products of Thymine and Cysteine. Biochemistry *12*, 2725–2730.

125. Barbatti, M., Aquino, A.J.A., Szymczak, J.J., Nachtigallová, D., Hobza, P., and Lischka, H. (2010). Relaxation mechanisms of UV-photoexcited DNA and RNA nucleobases. Proc. Natl. Acad. Sci. U. S. A. *107*, 21453–21458.

126. Rios, A.C., Yua, H.T., and Tor, Y. (2015). Hydrolytic fitness of N-glycosyl bonds: Comparing the deglycosylation kinetics of modified, alternative, and native nucleosides. J. Phys. Org. Chem. *28*, 173–180.

127. Steinmaits, H., Rosenthal, I., and Elad, D. (1971). Light- and γ-Ray-Induced Reactions of Purines and Purine Nucleosides with Alcohols. J. Org. Chem. *36*, 3594–3598.

128. Rohrbach, M.S., and Bodley, J.W. (1977). Photo-cross-linking of guanine nucleotides to the nucleotide binding site of elongation factor G. Arch. Biochem. Biophys. *183*, 340–346.

129. Xu, X., Muller, J.G., Ye, Y., and Burrows, C.J. (2008). DNA-protein cross-links between guanine and lysine depend on the mechanism of oxidation for formation of C5 vs C8 guanosine adducts. J. Am. Chem. Soc. *130*, 703–709.

130. Fabris, D. (2010). A Role for the MS Analysis of Nucleic Acids in the Post-Genomics Age. J. Am. Soc. Mass Spectrom. *21*, 1–13.

131. Urlaub, H., Thiede, B., Müller, E.C., Brimacombe, R., and Wittmann-Liebold, B. (1997). 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.

132. Urlaub, H., Thiede, B., Müller, E.C., and Wittmann-Liebold, B. (1997). Contact sites of peptide-oligoribonucleotide cross-links identified by a combination of peptide and nucleotide sequencing with MALDI MS. J. Protein Chem. *16*, 375–383.

133. Kühn-Hölsken, E., Lenz, C., Sander, B., Lührmann, R., and Urlaub, H. (2005). Complete MALDI-ToF MS analysis of cross-linked peptide-RNA oligonucleotides derived from nonlabeled UV-irradiated ribonucleoprotein particles. RNA *11*, 1915–1930.

134. Urlaub, H., Hartmuth, K., Kostka, S., Grelle, G., and Lührmann, R. (2000). A general approach for identification of RNA-protein cross-linking sites within native human spliceosomal small nuclear

ribonucleoproteins (snRNPs): Analysis of RNA-protein contacts in native U1 and U4/U6.U5 snRNPs. J. Biol. Chem. *275*, 41458–41468.

135. Sample, P.J., Gaston, K.W., Alfonzo, J.D., and Limbach, P. a (2015). RoboOligo: software for mass spectrometry data to support manual and de novo sequencing of post-transcriptionally modified ribonucleic acids. Nucleic Acids Res. *43*, e64. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25820423.

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

137. Urdaneta, E.C., Vieira-Vieira, C.H., Hick, T., Wessels, H.H., Figini, D., Moschall, R., Medenbach, J., Ohler, U., Granneman, S., Selbach, M., *et al.* (2019). Purification of cross-linked RNA-protein complexes by phenol-toluol extraction. Nat. Commun. *10*, 1–17. Available at: http://dx.doi.org/10.1038/s41467-019-08942-3.

Acknowledgements

First I would like to thank Prof. Henning Urlaub for giving me the opportunity to work on this project and for the continuous scientific and personal support throughout the years.

I am thankful to Prof. Markus Bohnsack and Prof. Tim Beißbarth for their time and suggestions as part of my thesis advisory committee. I would like to also thank Prof. Jörg Stülke, Dr. Alexander Stein and Dr. Alex Faesen for taking part in my examination board.

Sincere thanks to Dr. Alexandra Stützer and Alexander Wulf for proof-reading, as well as for the numerous discussions and suggestions in the last years. I appreciate all the support provided in the beginning of my PhD from Dr. Kundan Sharma and Dr. Uzma Zaman.

I am immensely grateful to Dr. Hauke Hillen, Dr. Goran Kokic, Dr. Katharina Hofmann, Dr. Jana Schmitzova, Dr. Leyla El Ayoubi, Dr. Constantin Cretu and Dr. Maria Tauber for providing numerous protein-RNA/DNA samples, on which I could test all my wild ideas and that helped me figure out all the ways that don't work.

Many thanks to our collaborators in the Kohlbacher lab in Tübingen for the development and support of the OpenMS suite and the RNP^{xl} tool, especially to Dr. Timo Sachsenberg, Johannes Veit and Eugen Netz for the improvements of the search engine and annotation.

I am heavily indebted to Monika Raabe and Uwe Pleßmann for their help and guidance with all instrument-related questions. I am grateful to Prof. Claudia Höbartner for her help in generating the peptide-DNA/RNA heteroconjugate standards.

I would like to thank Dr. Olexandr Dybkov and Dr. Kuan-Ting Pan for always finding the time and patience to answer all my annoying questions.

I am grateful for all the support I received throughout the years from the MolBio coordination office by Dr. Steffen Burkhardt and Kerstin Grüniger. I would like to also thank Jennifer Reinhold, Dr. Christian Kordowski and Sibyla Valkova for their help in the final stretch.

I am deeply grateful and would like to thank all current and previous members of the Bioanalytical Mass Spectrometry Lab for all the things they taught me and the everyday discussions that expanded my scientific understanding and critical thinking.

This work was partly supported by the Ph.D. program "Molecular Biology" (IMPRS) at the Georg August University Göttingen and the SFB860 (Sonderforschungberiech 860).

Moreover, I truly appreciate all the stuff completely unrelated to science. I thank Kundan for knowing what one should say in the face of *La Chupacabra*, Andreas for finding everything interesting, while preventing forest fires and Sunit for punching me on regular basis, squeezing in an extra kick or two, here and there, whenever I felt down. I am grateful for Wulfie's miracles and aura reading capabilities, for the bizarre google searches that the conversations with Alexandra prompted and for the highly inappropriate TV show taste of Luisa. I greatly appreciate Iwan's climber physiognomy and him always calling me out on my bullshit. I thank Fanni for always being upstairs and hung(a)ry. I am grateful for the politically incorrect Fridays with Jasmin, all the mango Schiffy never brought and the exquisite sketch taste of Andreia that almost rivals mine. I would like to thank Sasha for all the puns and Goran for not having an unsarcastic bone in his body.

I am deeply thankful to Чипето and Fernando for making me laugh every day from far away.

Finally, I would like to thank my family for their love and support, and for always being there for me.

Благодаря ви!