Molecular Cell
Resource

The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts

Alexander G. Baltz,[1,3] Mathias Munschauer,[1,3] Björn Schwanhäusser,[1] Alexandra Vasile,[1] Yasuhiro Murakawa,[1] Markus Schueler,[1] Noah Youngs,[2] Duncan Penfold-Brown,[2] Kevin Drew,[2] Miha Milek,[1] Emanuel Wyler,[1] Richard Bonneau,[2] Matthias Selbach,[1] Christoph Dieterich,[1] and Markus Landthaler[1,*]
[1]Max Delbrück Center for Molecular Medicine, Berlin Institute for Medical Systems Biology, 13125 Berlin, Germany
[2]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10003, USA
[3]These authors contributed equally to this work
*Correspondence: markus.landthaler@mdc-berlin.de
DOI 10.1016/j.molcel.2012.05.021

SUMMARY

Protein-RNA interactions are fundamental to core biological processes, such as mRNA splicing, localization, degradation, and translation. We developed a photoreactive nucleotide-enhanced UV crosslinking and oligo(dT) purification approach to identify the mRNA-bound proteome using quantitative proteomics and to display the protein occupancy on mRNA transcripts by next-generation sequencing. Application to a human embryonic kidney cell line identified close to 800 proteins. To our knowledge, nearly one-third were not previously annotated as RNA binding, and about 15% were not predictable by computational methods to interact with RNA. Protein occupancy profiling provides a transcriptome-wide catalog of potential *cis*-regulatory regions on mammalian mRNAs and showed that large stretches in 3′ UTRs can be contacted by the mRNA-bound proteome, with numerous putative binding sites in regions harboring disease-associated nucleotide polymorphisms. Our observations indicate the presence of a large number of mRNA binders with diverse molecular functions participating in combinatorial posttranscriptional gene-expression networks.

INTRODUCTION

During and immediately after transcription, nascent messenger RNAs (mRNAs) associate with proteins to form messenger ribonucleoprotein (mRNP) complexes that mediate and regulate most aspects of mRNA metabolism and function. Throughout their life cycle, mRNP complexes consist of a dynamically changing repertoire of proteins that define processing, cellular localization, and the decay and translation rate of mRNAs.

The mammalian genome has been predicted to encode about 600 RNA-binding proteins (de Lima Morais et al., 2011), based on the presence of one or more catalytic or noncatalytic domains that can interact with RNA. However, several proteins implicated in other cellular processes exhibit RNA-binding activity despite the absence of recognizable RNA-binding domains. Among them, cytosolic aconitase (also known as iron-regulatory protein 1) posttranscriptionally regulates specific target mRNAs depending on cellular iron levels (Kennedy et al., 1992). This and other examples of RNA-binding activity of unexpected proteins highlight the need to thoroughly catalog the cellular repertoire of RNA-binding proteins in order to define the system that regulates the posttranscriptional fate of mRNAs.

More than 30 years ago, initial attempts were made to isolate and analyze the poly(A)$^+$RNA-bound proteome by oligo(dT) sepharose chromatography. Purifications of mRNPs from in vitro UV-irradiated polysomal fractions (Greenberg, 1979), from UV-irradiated intact cells (Wagenmakers et al., 1980), and from untreated cells (Lindberg and Sundquist, 1974) revealed the association of a specific set of proteins with mRNA. Later on, similar methods were applied to characterize hnRNP particles and to identify the mRNA polyadenylate-binding protein (Adam et al., 1986; Choi and Dreyfuss, 1984). Recently, screening approaches and oligo(dT) purification procedures were used to provide a catalog of yeast RNA-binding proteins (Scherrer et al., 2010; Tsvetanova et al., 2010). However, methods for comprehensive identification of mammalian mRNA-binding proteins have remained elusive.

Another prerequisite for our understanding of the function of RNA-interacting proteins is a systematic identification of their binding sites and the definition of their RNA targets. Current genomic approaches use UV crosslinking and immunoprecipitation (CLIP) of mRNA-protein complexes in combination with next-generation sequencing to identify RNA binding sites (König et al., 2010; Licatalosi et al., 2008). One recently developed method, PAR-CLIP, employs the photoreactive thionucleosides, 4-thiouridine and 6-thioguanosine, to increase the crosslinking efficiency between protein and RNA and to provide near nucleotide resolution of the RNA-binding site (Hafner et al., 2010).

Here, we use a photoreactive nucleoside-enhanced UV crosslinking and oligo(dT) affinity purification approach to identify the mRNA-bound proteome and to globally map the sites of protein-mRNA interactions in mammalian cells. Using high-resolution quantitative mass spectrometry we uncovered close to 800 proteins in oligo(dT)-purified protein-mRNA complexes. A large number of these proteins were neither previously annotated

nor could be predicted by state of the art computational methods to interact with RNA. We validated the RNA binding function of 19 of these mRNA-interacting proteins. Using PAR-CLIP-seq, we identified mRNA binding preferences and distinct binding patterns for five of these proteins. Furthermore, protein occupancy profiling on mRNA by next-generation sequencing of protein-crosslinked RNA fragments, provided a transcriptome-wide view of the interaction sites of the mRNA-bound proteome and revealed widespread binding of proteins to 3′ untranslated regions (3′ UTRs) of mRNAs.

RESULTS

Optimization of mRNP Oligo(dT) Affinity Purification

To characterize the protein-mRNA interactome, we sought to improve existing methods to identify the protein content of oligo(dT) affinity-purified mRNP complexes and to determine the mRNA regions contacted by the mRNA-bound proteome (Figure 1A). A key feature of our approach is the use of photoreactive nucleoside analogs, 4-thiouridine (4SU) and 6-thioguanosine (6SG), to metabolically label cellular RNA without detectable incorporation into DNA (Figure S1A available online) (Favre et al., 1993). Both 4SU and 6SG are readily taken up by cultured mammalian cells and dramatically enhance the crosslinking efficiency of proteins to RNA by UV 365 nm irradiation (Hafner et al., 2010). Photocrosslinking of living cells stabilizes mRNP complexes and facilitates their isolation by oligo(dT) affinity purification (Setyono and Greenberg, 1981; Wagenmakers et al., 1980). Protein-denaturing conditions during the purification ensure a stringent isolation of proteins in direct contact with mRNA through covalent bonds and thus enable the identification of the mRNA-interacting proteins by mass spectrometry (Figure 1A). Moreover, 4SU-labeled RNA, crosslinked to proteins, can readily be identified by characteristic T to C (T-C) transitions in complementary DNA (cDNA) (Hafner et al., 2010), providing a way to globally identify the RNA segments that are crosslinked by the mRNA-bound proteome (Figure 1A).

We initially tested this approach by purifying protein-mRNA complexes using oligo(dT) beads from extracts of UV-irradiated and nonirradiated intact human embryonic kidney (HEK) 293 cells after growth in medium supplemented with or without 4SU and 6SG. Resolving the RNase-treated eluate by SDS-PAGE revealed that the combination of metabolic labeling of RNA with photoreactive nucleosides and irradiation at UV 365 nm allowed a high recovery of proteins (Figure 1B). We further examined the amount of mRNA obtained in precipitates from extracts of crosslinked and noncrosslinked cells. A qRT-PCR analysis showed that comparable amounts of *GAPDH* mRNA were precipitated (Figure S1B), suggesting that labeling of RNA and UV irradiation had only a minor effect on the mRNA pull-down efficacy.

As expected when probing the oligo(dT) precipitate for the presence of known RNA-binding proteins by western analysis, we were able to detect the heterogeneous nuclear ribonucleoprotein K (HNRNPK) (Figure S1C). However, the Argonaute protein, AGO2/EIF2C2, was not detectable after a single oligo(dT) purification (Figure S1C), likely due the insufficient precipitation of mRNAs and/or incomplete capture of mRNAs with shortened poly(A) tails, such as microRNA/AGO-targeted mRNAs. Thus, we measured the degree of depletion of *GAPDH* mRNA after one oligo(dT) precipitation. The *GAPDH* transcript is abundant and targeted by AGO proteins (Hafner et al., 2010; Kishore et al., 2011). Figure 1C shows that only about 70% of this transcript was depleted in the supernatant when compared to input RNA. Three additional consecutive pull-downs from the same extract reduced the amount of *GAPDH* mRNA in the supernatant to about 5% (Figure 1C). A western analysis of the pooled eluates of four oligo(dT) purifications validated the presence of AGO2 protein (Figure 1D) as well as the RNA-binding protein QUAKING, indicating that multiple consecutive oligo(dT) precipitations are required to precipitate crosslinked AGO protein. Known DNA-binding proteins were not detected in the pooled precipitates (Figure S1D).

Characterization of the Oligo(dT)-Purified RNA

To obtain a more detailed picture of the RNA present in the pooled precipitates of four consecutive oligo(dT) purifications, we constructed a cDNA library by random priming of 4SU- and 6SG-labeled RNA derived from irradiated and nonirradiated cells. Digital gene expression analysis of the cDNA library of nonirradiated cells, labeled with 4SU and 6SG, revealed that about 88% of the sequence reads mapped to mRNA and 8% ribosomal RNA (rRNA) genes, whereas in RNA precipitates obtained from UV-irradiated cells the rRNA content increased to 36%, likely reflecting crosslinking of ribosomes to mRNA transcripts (Figures 1E and S1E). In contrast a standard mRNA purification procedure, involving only a single oligo(dT) precipitation, of untreated cells resulted in 96% mRNA and 2% rRNA (Figures 1E and S1E).

Furthermore a comparison of the different RNA libraries showed that the abundance of mRNAs obtained by a single oligo(dT) purification from untreated cells and metabolically labeled transcripts derived from noncrosslinked and UV-crosslinked cells correlated well (Pearson correlation coefficient of 0.87 and 0.82, respectively; Figure 1F), indicating that the oligo(dT)-precipitated mRNA closely reflected the cellular mRNA pool. To monitor the incorporation of photoreactive nucleotides into mRNA, we isolated 4SU- and 6SG-labeled RNA from the oligo(dT) precipitate of noncrosslinked cells by biotinylation and streptavidin purification (Dölken et al., 2008). The abundance of the thionucleotide-containing RNA was in good agreement with cellular mRNA (Pearson correlation coefficient of 0.90), suggesting efficient and unbiased metabolic labeling of transcripts (Figure 1F). In summary, we concluded that at least four consecutive oligo(dT) purifications are required to efficiently purify cellular mRNA-protein complexes, while a significant enrichment of mRNA over other classes of RNA can still be observed in the resulting precipitates.

Identification of mRNA-Bound Proteins by Quantitative Mass Spectrometry

To identify proteins crosslinked to mRNAs, we performed oligo(dT) purifications, as described above, and precipitates were analyzed by SILAC-based quantitative mass spectrometry (Mann, 2006). For this purpose, cells were grown in medium

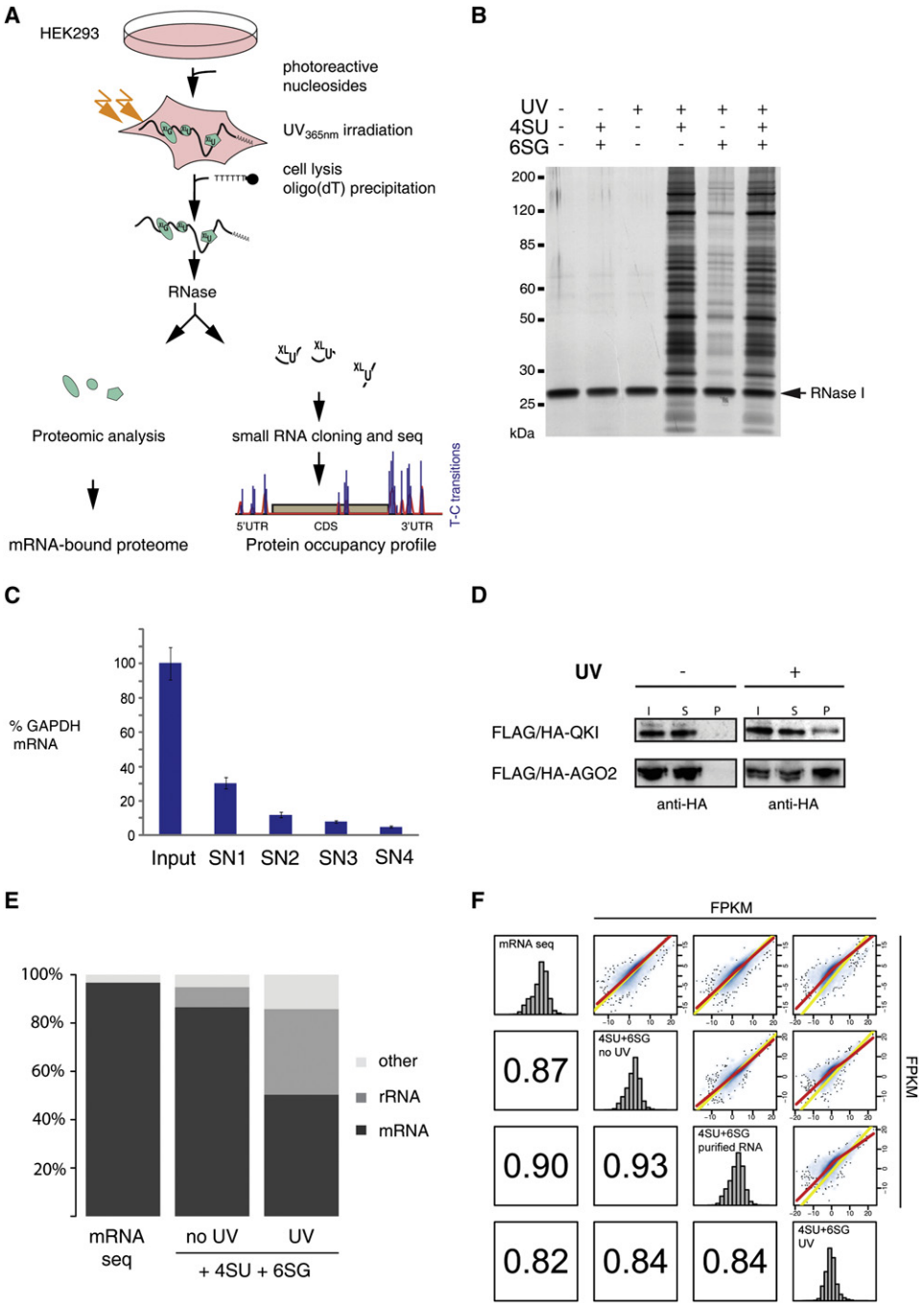


Figure 1. Analysis of Oligo(dT)-Purified mRNPs

(A) Illustration of the experimental setup to identify the mRNA-bound proteome and its occupancy profile on RNA. Transcripts were labeled with photoreactive nucleosides and proteins were crosslinked to RNA by 365 nm UV irradiation. mRNP complexes were isolated after cell lysis by oligo(dT) precipitation under denaturing conditions. For the identification of the mRNA-bound proteome, mRNPs were eluted from the beads, nuclease treated, and analyzed by quantitative mass spectrometry. For identification of the protein binding pattern on RNA, mRNPs were RNase I treated, followed by proteinase K digest to remove RNA-bound proteins. RNA molecules were converted into a cDNA library and next-generation sequenced.

(B) SDS-PAGE analysis of proteins crosslinked to polyadenylated RNA. HEK293 cells were grown in medium supplemented with 4SU and/or 6SG and UV irradiated at 365 nm. Cells were lysed using denaturing conditions, and protein-mRNA complexes were isolated by oligo(dT) precipitation. Protein-RNA complexes were eluted from oligo(dT) beads, treated with RNase I, separated on a SDS gradient gel, and visualized by silver staining.

(C) *GAPDH* mRNA depletion. qRT-PCR analysis of *GAPDH* mRNA in supernatants (SN1 to SN4) after each round of oligo(dT) bead precipitation (four in total) compared to *GAPDH* mRNA in extract before precipitations (input) are shown as percent of input. The error bars display the calculated maximum and minimum expression levels that represent the standard error of the mean expression level with a 95% confidence interval.

supplemented with "light" or "heavy" stable isotope labeled amino acids to compare the protein abundance in oligo(dT) precipitates of crosslinked cells to that of noncrosslinked cells (Figure S2A). We performed two independent experiments (L1 and L2) in which the "light" labeled cells were UV irradiated and proteins in the oligo(dT) precipitations were compared to the precipitate of noncrosslinked "heavy" labeled cells. In a single "label swap" experiment (H1), the "heavy" labeled cells were crosslinked and the recovered proteins were compared to those of "light" labeled noncrosslinked cells.

Proteins in pooled oligo(dT) precipitates were separated by SDS-PAGE followed by in-gel digest. Peptides were analyzed by nanoflow high-performance liquid chromatography (HPLC) and online mass spectrometry on a high-resolution instrument. In total, 300 hr of data acquisition for all three experiments resulted in 1,383,856 MS/MS spectra. Processing the data with MaxQuant (Cox and Mann, 2008), we identified 1,326 proteins and observed a significant overlap between experiments. Seven hundred ninety proteins were identified in all three proteomic analyses, and 561 of those were quantified with at least three observed SILAC-peptide ratios in each experiment (Figure 2A). To further examine the reproducibility we compared log2 SILAC ratios from biological replicates L1 and L2 (Figure 2B). Seven hundred seventy-eight out of 800 proteins identified in both experiments were specifically enriched in the precipitates of UV-crosslinked cells relative to the nonirradiated control cells (SILAC log2 fold changes < 0 in both cases). Hence, 97% of all identified proteins showed specific enrichment. In addition we observed no correlation between the fold enrichment and the cellular protein abundance (Figure S2B), suggesting that the degree of enrichment was independent of the number of protein molecules present in the cell.

Next, we plotted the log2 SILAC ratios from both biological replicates against the label swap experiment. As expected, most SILAC ratios were inverted by the label swap (Figures 2C and 2D). The proteins with low SILAC ratios in both the biological replicates and the label swap experiment were largely contaminants such as trypsin, LysC, and keratins. We therefore excluded 135 proteins with negative log2 SILAC ratios in the label swap experiment. Among the excluded proteins were six known RNA-binding proteins: the small SNRPE, PDCD11, ELAVL3, RBM16, PA2G4, and RBPMS. Requiring an enrichment of at least 3-fold (determined by SILAC ratios) in at least one of three analyses, we excluded 54 proteins, which did not meet this stringent enrichment criteria, thereby ending up with 889 proteins. In order to facilitate further analysis, we retained only the longest isoforms of the identified proteins, resulting in a nonredundant list of 797 proteins (Table S1).

We further subdivided the 797 proteins into three classes (Table S1). Class I included 505 proteins (63%), which were enriched more than 3-fold in all three proteomic analyses. One hundred ninety proteins (24%) showed a 3-fold enrichment in two experiments (class II), and 102 proteins (13%) in only one experiment (class III).

Overview of Identified mRNA-Interacting Proteins

We first classified the identified proteins into functional categories based on gene annotation. As expected, ribosomal proteins, RNA helicases, translation factors, and RNA-binding proteins were most frequent, making up close to 70% of the identified proteins (Figure 3A). The low numbers of highly expressed cellular proteins such as metabolic enzymes, histones, and heat-shock proteins, suggested that the oligo(dT) purification was specific. The median relative abundance of RNA-binding proteins, translations factors, and ribosomal proteins was slightly higher but comparable to that of other functional groups of proteins (Figure 3B).

Confirming the method, we discovered RNA-interacting proteins present in complexes that influence surveillance and translation of spliced mRNAs. We detected all proteins, RBM8A/Y14, MAGOH, EIF4A3, and CASC3/BTZ, making up the core of the exon junction complex (Bono et al., 2006). However, we identified an order of magnitude more molecules of EIF4A3 and CASC3 (Table S1), which were previously shown to directly interact with RNA, compared to MAGOH and the RNA-recognition motif-containing protein RBM8A. We observed EIF4A1, EIF4B, EIF4E, EIF4G1, and EIF4H, all of which are present in the translation initiation complex (Jackson et al., 2010). Additionally, we identified three of the four Argonaute proteins expressed in HEK293 cells. On the other hand, the identified mRNA binders only partially overlapped with sets of proteins found in nuclear RNA-containing structures. Ninety-nine out of 172 proteins detected in spliceosomal complexes (Bessonov et al., 2008) were observed to interact with mRNA (Figure 3C). Two hundred forty-one identified mRNA interactors were also found in the nucleolus proteome (Andersen et al., 2005) (Figure 3C).

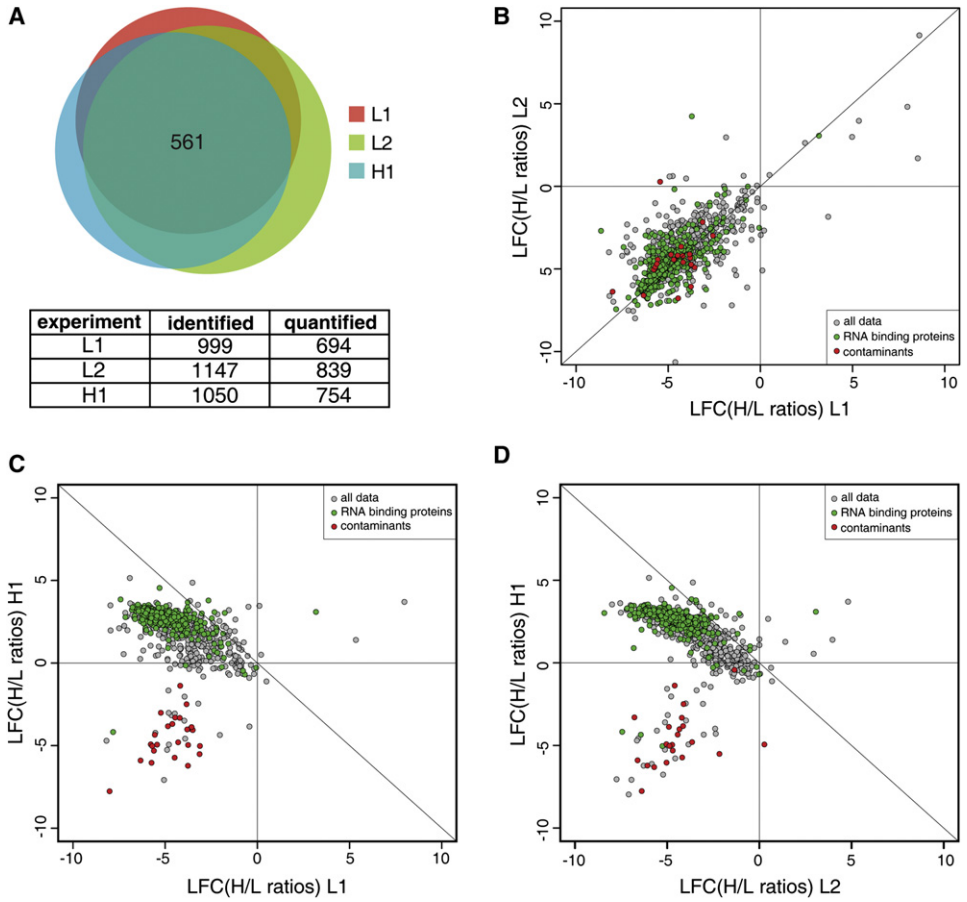
In addition to the expected mRNA-interacting proteins, we identified 245 proteins (Table S1), which have not been previously annotated as RNA binding (Figure 3A, candidate mRNA binders). Eighty percent of these proteins were detected in at least two out of three proteomic analyses, and about 47%

(D) Western analysis of FLAG/HA-tagged RNA-binding proteins QUAKING (QKI) and ARGONAUTE 2 (AGO2/EIF2C2) in input extract (I), supernatant after precipitation (S), and oligo(dT)-purified material (P) of UV-crosslinked and noncrosslinked cells.

(E) Read count distribution over different RNA types. mRNA was purified either from total TRIzol-extracted RNA by a single oligo(dT) precipitation (mRNA seq), or by four rounds of oligo(dT) precipitation from cellular extract of UV-irradiated and nonirradiated cells (4SU+6SG UV and 4SU+6SG no UV, respectively). Crosslinked proteins were removed by Proteinase K digest prior to RNA analysis by next-generation sequencing of recovered RNA. The read count distribution over different RNA classes (mRNA, rRNA, and other) was inferred by multiplication of the FPKM values with the respective length of the longest transcript of a given gene.

(F) Pair-wise correlation between RNA abundance expressed as log2 FPKM of RNA described in (E). For assessment of the incorporation of photoreactive nucleoside into RNA, the 4SU- and 6SG-containing RNA was purified from oligo(dT)-precipitated RNA of noncrosslinked cells by biotinylation and streptavidin pull-down (4SU+6SG purified RNA) and analyzed by next-generation sequencing. The diagonal is shown as yellow line for of each pairwise comparison, whereas a LOESS regression line is shown in red.

See also Figure S1.



experiment	identified	quantified
L1	999	694
L2	1147	839
H1	1050	754

Figure 2. Identification of the mRNA-Bound Proteome by Quantitative Mass Spectrometry

(A) Summary of proteomic experiments. In two replicates, the proteomic composition of oligo(dT) precipitates was analyzed for "light" labeled crosslinked cells (experiments L1 and L2) and one experiment for "heavy" labeled crosslinked cells (H1). The overlap of identified proteins in different experiments is shown in the Venn diagram. The table indicates the number of identified and quantified proteins, as determined by SILAC ratios of proteins in each experiment.

(B) Comparison of the log2 fold changes (LFC) of "heavy" to "light" SILAC ratios (H/L) of proteins quantified in biological replicates L1 and L2. Previously known RNA-binding proteins are indicated in green, and known contaminants are in red.

(C and D) As in (B), proteins were quantified in L1 (C) or L2 (D) plotted against proteins quantified in label swap experiment H1.

See also Figure S2 and Table S1.

were observed in all three pull-downs. We applied an adaptation of a multiple association network integration algorithm (Mostafavi et al., 2008) to predict proteins with RNA-binding function, using gene ontology (GO) data, InterPro and Pfam domain data, gene coexpression, protein-protein interaction, and structural similarity data (Drew et al., 2011). This algorithm demonstrated strong predictive power, as evidenced by the precision-recall values for RNA binding (see Table S2, part A) and by previous field-wide tests of function prediction algorithms (Peña-Castillo et al., 2008). The most informative data types for predicting RNA-binding were determined to be GO process, GO localization, and 3D protein structure similarity, followed to a lesser degree by protein-protein interaction data, gene coexpression, and InterPro domains. A full description of the algorithm and benchmarking results appear in the Supplemental Experimental Procedures.

After applying the algorithm to the 245 candidate mRNA-interacting proteins detected by our assay, 112 proteins (48 belonging to class I) could not be predicted as RNA binders (Table S1) (even at a very low precision level $\geq$ 20%, and when using the function prediction algorithm in a manner that minimizes false negative predictions at the expense of false positive predictions). This finding strongly suggests that our experiments uncovered RNA-interacting proteins that use distinct or highly divergent RNA-binding domains and occupy discrete regions of the known protein association networks. Some of our discoveries include proteins that are functionally annotated as transcription factors (JUN, NXF1), and putative protein kinases (FASTKD1, FASTKD2, FASTKD5). Additionally, several proteins encoded by open reading frames (C1orf35, C16orf80, C11orf31, C9orf114, C19orf47) were observed to be RNA binding.

Overrepresentation of Nucleic Acid Binding Domains

Next, we classified the identified proteins based on their three-dimensional structure and amino acid sequence. For the

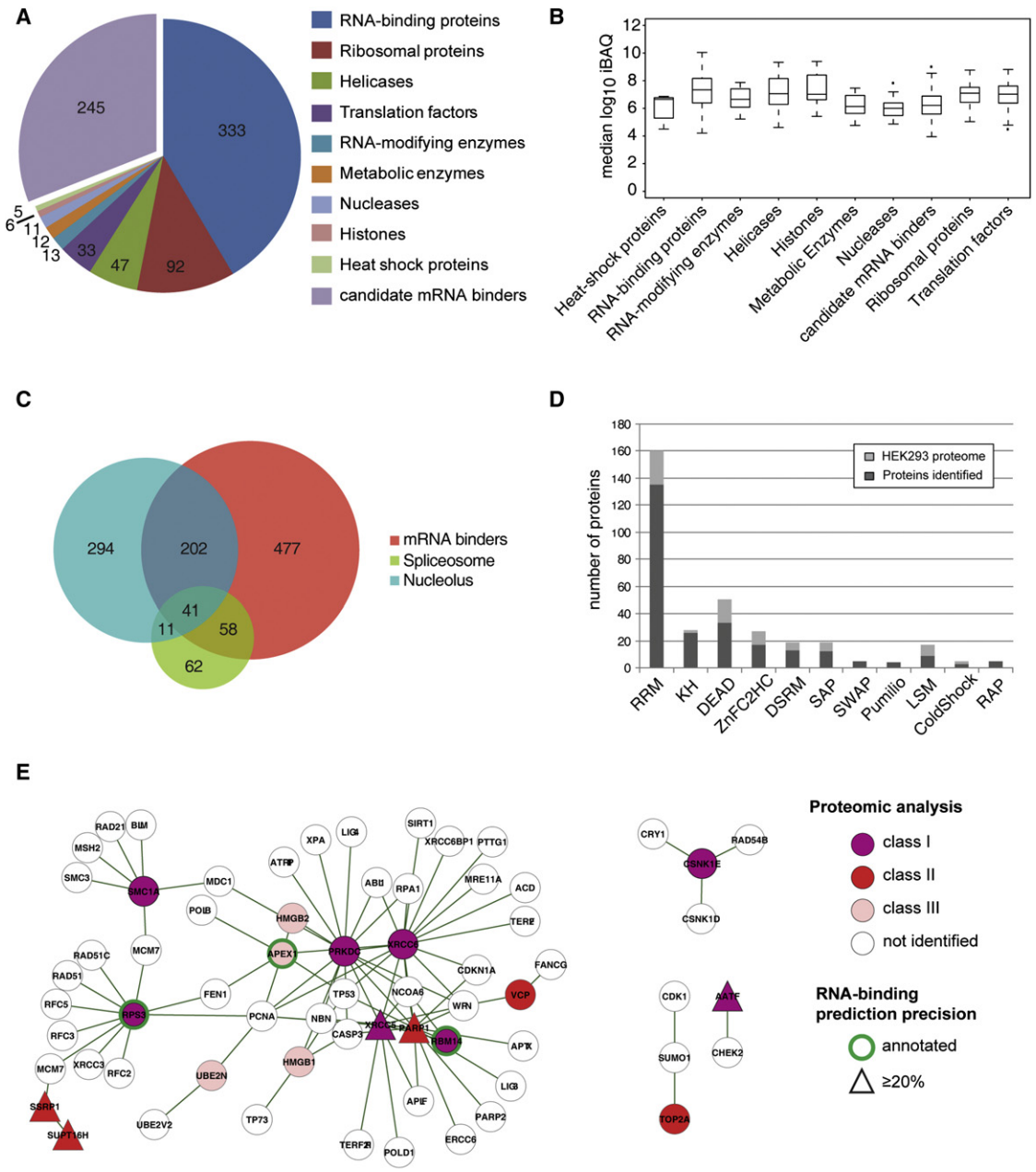


Figure 3. Overview of Identified mRNA-Interacting Proteins

(A) Number of identified proteins belonging to different functional categories.

(B) Median relative number of protein molecules belonging to different functional categories as determined, shown as box plots. Protein amounts were calculated as the sum of all peptide peak intensities divided by the number of theoretically observable tryptic peptides (Schwanhäusser et al., 2011). The median is shown as horizontal line, and the surrounding box defines the upper and lower quartile. The sample range is defined by the whiskers, while dots indicate potential outliers.

(C) Overlap of identified mRNA binders with proteins present in the spliceosome and the nucleolus.

(D) Number of identified proteins with specific RNA-binding domains (dark gray) was compared to the respective number of RNA-binding domain-containing proteins in HEK293 proteome (light gray).

(E) mRNA-bound proteins and their first neighbors based on PPI were analyzed for overrepresented Gene Ontology terms. The members of the cluster enriched for DNA damage response are depicted as nodes in the PPI network. Only proteins with direct interactions are shown. The node color represents the number of times the proteins were identified in our analysis (purple, detected in all three experiments; red, detected in two experiments; pink, detected in one experiment), and the edges indicate direct PPIs between the cluster members. The node shape indicates the level of prediction. Annotated RNA-binding proteins are marked by green node border color.

See also Figure S3 and Tables S1 and S3.

structural classification, we first queried the set of mRNA-interacting proteins against the Protein Folding Project database (Drew et al., 2011). This database provided SCOP superfamily classifications derived from sequence similarity (psi-blast), fold recognition, and Rosetta de novo structure prediction for the identified RNA-interacting proteins. An enrichment analysis of superfamilies showed an overrepresentation of folds associated with single and double-stranded RNA (dsRNA)-binding function (RNA-binding domain, eukaryotic type KH domain, and dsRNA-binding domain-like), helicases (P-loop containing nucleoside triphosphate hydrolases) and nucleases (Pin domain-like) with a corrected p value ≤ 0.05 (Table S3, part A). Interestingly, we also found two structural superfamilies significantly enriched that are associated with DNA binding ("winged helix" DNA-binding domain, and AlbA-like) suggesting that these DNA-binding folds could also interact with RNA. The "winged helix" DNA-binding domain is present in a number of RNA helicases. The AlbA-like fold was found in POP7 and in C9orf23. Notably, the AlbA-like superfamily had been previously suggested to be involved in RNA binding (Aravind et al., 2003).

To obtain an additional perspective of the mRNA-bound proteome associated structures, we performed Pfam and InterPro domain enrichment analysis using the identified proteins. As expected, most of the significantly enriched domains were various RNA-interaction motifs (Table S3, part B). Besides the commonly recognized RNA-binding domains, we found an overrepresentation of several domains with putative RNA-binding activity (Table S3, part B). Among these were the SWAP/SURP domain and the RAP domain, for which an RNA binding activity was suggested based on sequence comparisons (Denhez and Lafyatis, 1994; Lee and Hong, 2004). The RAP domain was also significantly enriched among the 112 identified RNA interactors that were not predicted to be RNA binding (Table S3).

Finally, to estimate the depth of the mRNA-bound proteome we covered in our oligo(dT) precipitations, we compared the number of identified proteins encoding at least one specific RNA-binding domain to the number of respective RNA-binding domain containing proteins observed in the deep HEK293 proteome (Geiger et al., 2012). Figure 3D shows that the majority of RNA-binding domain-containing proteins expressed in HEK293 were identified by our analyses.

mRNA-Bound Proteome Connects Posttranscriptional Regulation to DNA-Related Processes

In order to systematically examine the connectivity of the identified mRNA binders and their potential relationship to non-mRNA related biological processes, we generated a network based on protein-protein interaction (PPI) (Figure S3A). When comparing the PPI network of mRNA binders to a random network of equal size, we observed a higher average clustering coefficient, indicating the presence of highly interconnected protein clusters within the network (Figure S3A). Because these clusters are indicative of functional modules mediating the regulation of complex biological processes, we analyzed the set of mRNA binders and their first neighbors, based on PPI, for an enrichment of GO terms linked to biological processes (Ashburner et al., 2000). As expected, the most significantly overrepresented GO terms were mRNA splicing, localization, processing, and translation (Figure S3B). In addition we observed an overrepresentation for DNA-related processes, namely "response to DNA damage," "DNA-dependent transcription," and "DNA duplex unwinding" (Figure S3B).

Figure 3E shows the PPI sub-network for members linked to the term "response to DNA damage." Central to this network are XRCC6/Ku70, XRCC5/Ku80, and the DNA-activated protein kinase (PRKDC). These proteins were identified in all three proteomic analyses (Table S1). Besides their role in DNA double-strand break repair and recombination, the proteins have been shown to interact with RNA structures, such as the RNA-stem loop region in yeast telomerase TLC1 and the RNA-component of human telomerase (Ting et al., 2005). In addition, XRCC6 had been suggested to bind internal ribosomal entry site (IRES) elements and is likely involved in the regulation of IRES-mediated mRNA translation (Silvera et al., 2006).

Validation of RNA-Binding Function of Several Candidate mRNA Interactors

To validate the RNA-binding activity of a subset of the identified proteins, we applied PAR-CLIP. In brief, HEK293 cell lines, stably expressing epitope-tagged mRNA binders, were grown in the presence of 4SU and UV irradiated. Immunopurified (IPed) and RNase-treated protein-RNA complexes were radiolabeled with T4 polynucleotide kinase, separated by SDS-PAGE, and blotted onto a membrane. The radiolabeled protein-RNA complexes were visualized by phosphorimaging, while protein precipitation was monitored by western analysis. As positive controls, we used five known RNA-binding proteins: CAPRIN1 (Shiina et al., 2005), HNRNPD (Knapinska et al., 2011), HNRNPR (Hassfeld et al., 1998), HNRPNU (Kiledjian and Dreyfuss, 1992), as well as MYEF2, which is a transcriptional repressor (Haas et al., 1995) with an RNA recognition motif (RRM) domain. As expected, the tagged proteins, IPed from UV-irradiated cells, efficiently crosslinked to RNA, when compared to proteins that were IPed from nonirradiated cells (Figure 4A). In contrast, we were unable to detect radiolabeled protein-RNA complexes in immunoprecipitations of phosphoglycerate kinase 1 (PGK1) and lactate dehydrogenase A (LDHA) (Figure 4B), two metabolic enzymes that were not identified in our proteomic analyses as potential RNA binders (Table S1).

We generated HEK293 cell lines stably expressing 29 putative mRNA-interacting proteins as epitope-tagged versions. Twenty-one proteins could be IPed and were used in the PAR-CLIP assay (Table S4). We tested the RNA-binding activity of 18 candidates belonging to class I and three members of class II, BTF3, C16orf80, and PRDX1 (Figure 4C). For all proteins, except BZW1 and C16orf80, we observed an increased radioactive signal in immunoprecipitations of irradiated cells, indicating that these proteins were crosslinked to RNA, and thus are likely in close contact or directly binding to RNA.

Interestingly, several of the crosslinked proteins possess enzymatic activities: ALKBH5 (2-oxoglutarate oxygenase), C22orf28 (RNA ligase), CSNK1E (kinase), MKRN2 (ubiquitin ligase), PRDX1 (peroxidase), and USP10 (ubiquitin thioesterase). Furthermore several of the identified RNA-binding proteins have been implicated in transcriptional regulation either by inhibition

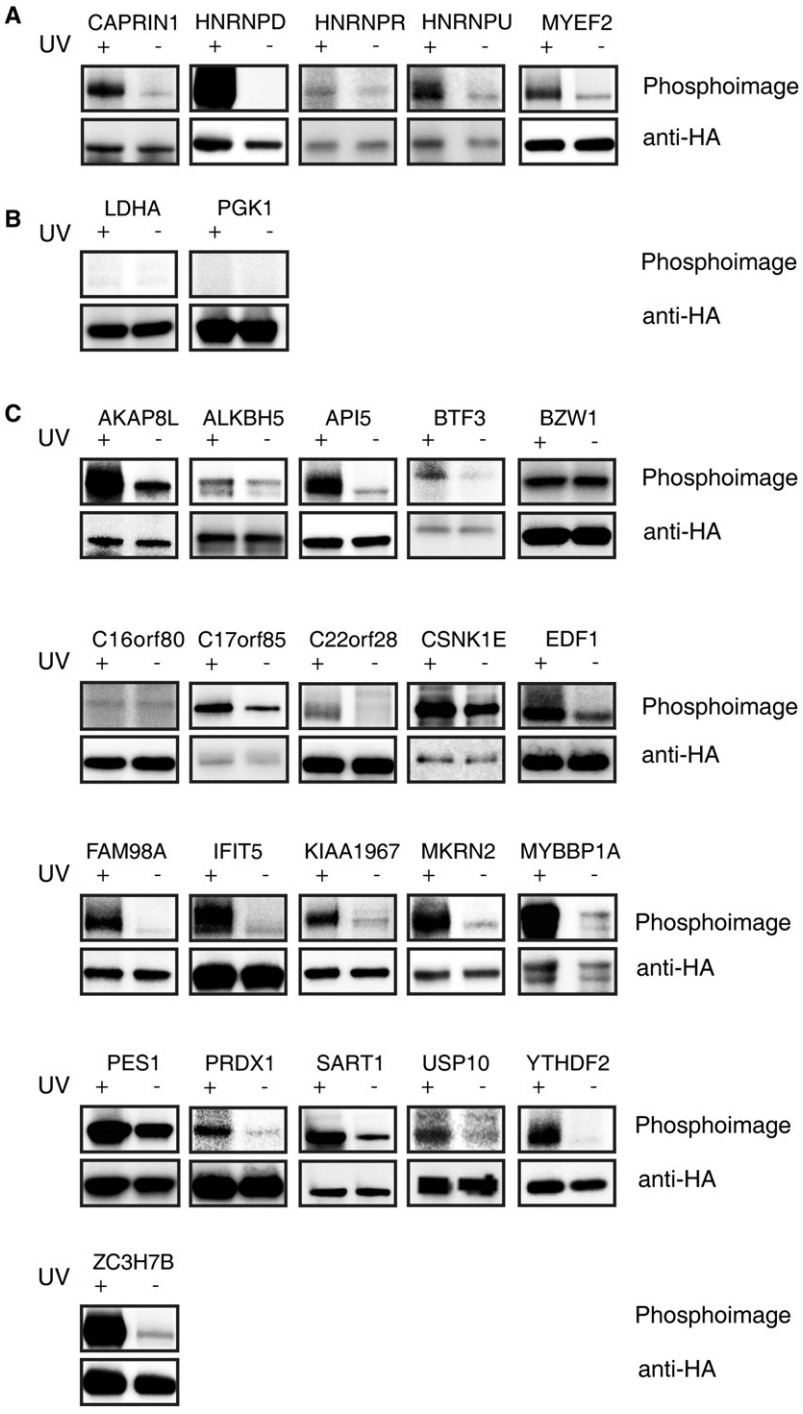


Figure 4. Validation of RNA-Binding Activity of Candidate mRNA Binders

RNA-binding activity of candidate mRNA binders was determined by PAR-CLIP. Protein-RNA complexes were separated by SDS-PAGE and blotted onto nitrocellulose membrane. Western analysis with an anti-HA antibody confirmed the correct size and equal loading of the IPed protein. Phosphorimaging indicated efficient radioactive labeling of covalently bound nucleic acid in the mRNP complex. The assay was performed at least twice for each protein. Representative results are shown.

(A) CAPRIN1, HNRNPD, HNRNPR, HNRNPU, and MYEF2 served as positive controls.

(B) Metabolic enzymes LDHA and PGK1, both not detected in oligo(dT) precipitations, served as negative controls.

(C) Results of PAR-CLIP assay for 21 putative mRNA binders are shown.

The radioactive signal in noncrosslinked immunoprecipitates is likely due to the presence of protein kinases and/or background crosslinking.

See also Figure S4 and Table S4.

Identification of RNA-Binding Sites of Several Candidate mRNA Interactors

To confirm that a representative subset of the identified RNA interactors is binding mRNA transcripts, we applied PAR-CLIP in combination with next-generation sequencing (Hafner et al., 2010). In PAR-CLIP experiments, crosslinking of 4SU-labeled RNA to proteins leads to specific T-C changes in cDNA sequences, marking the protein binding site on the target RNA (Hafner et al., 2010).

We performed duplicate experiments for five proteins: ALKBH5, C22orf28, C17orf85, and ZC3H7B, as well as the known RNA-binding protein CAPRIN1 (Table S5). Diagnostic T-C changes in aligned reads demonstrated efficient RNA-protein crosslinking (Figure S5A). All PAR-CLIP sequencing data (Table S5) were analyzed with a computational analysis pipeline (Lebedeva et al., 2011). The binding sites are available at http://dorina.mdc-berlin.de/cgi-bin/hgTracks/ (Anders et al., 2012). The mRNA targets for the respective proteins are listed in Table S5.

Analyses of PAR-CLIP-seq data confirmed that the five tested proteins all bind predominately mRNA. We used RNA immunoprecipitation (RIP) coupled to semiquantitative RT-PCR to confirm the interactions of these proteins with some of their top mRNA targets (Figure S5B). Although all proteins displayed a preference for mRNA, the distribution of binding sites on protein coding transcripts differed. The binding sites of CAPRIN1 were equally distributed over coding sequences (CDS) and 3′ UTR regions. CAPRIN1 localizes to stress granules in proliferating cells and was suggested to have a role in mRNA transport and local translational control (Shiina et al., 2005). In addition our data indicated that ZC3H7B has a binding

of histone deacetylases (KIAA1967) or by acting as transcription factor (BTF3, MYBBP1A, and EDF1). Since EDF1 harbors a prokaryotic-type helix-turn-helix motif, suggesting that this protein may function in DNA binding, we further examined the nature of the crosslinked nucleic acid. When we incubated the immunoprecipitate with RNase I, but not with DNase I, the radioactive signal of the ribonuclease-treated complex was reduced, indicating that EDF1 was crosslinked to RNA (Figure S4).

preference for 3′ UTRs, but can also interact with sequences in introns and CDSs. ZC3H7B was previously shown to form a ternary complex with the translation initiation factor EIF4G and the rotavirus nonstructural protein NSP3 in virus infected cells (Vitour et al., 2004).

The majority of binding sites of ALKBH5 and C22orf28 were identified in CDSs after normalization of the number of sequence clusters to the overall length of the different transcript regions (Figures 5A and S5C). ALKBH5 is a 2-oxoglutarate-dependent oxygenase and a direct target of hypoxia-inducible factor 1α (HIF-1α) (Thalhammer et al., 2011). In contrast to C22orf28, ALKBH5 binding sites were preferentially distributed to distal 5′ regions of CDSs (Figure 5B).

C22orf28 is the essential subunit of a human transfer RNA (tRNA) splicing ligase complex (Popow et al., 2011). A closer inspection of the C22orf28 target transcripts revealed that the ligase contacts the X-box binding protein 1 (*XBP1*) mRNA. This interaction could be confirmed by RIP-RT/PCR (Figure S5B). Interestingly, C22orf28 RNA binding sites in *XBP1* are flanking an intron (Figure 5C), which is removed by endoplasmic reticulum stress-induced unconventional cytoplasmic splicing (Yoshida et al., 2001), suggesting C22orf28 is the ligase in this enzyme-mediated splicing event.

Protein Occupancy Profiling Provides Catalog of Protein-mRNA Contact Sites

Present day CLIP data only provides insight into the transcriptome-wide RNA binding sites of close to 30 mammalian RNA interactors (Milek et al., 2012), less than 5% of the 800 mRNA binders identified in this study, leaving the majority of *cis*-regulatory mRNA elements contacted by these proteins intangible. Therefore, we set out to globally identify the RNA regions that interact with the mRNA-bound proteome by assessing the transcriptome-wide T-C transition profile in cDNA sequences derived from 4SU-labeled RNA crosslinked to all mRNA binders. The crosslinked 4SU residues indicate the RNA contact sites of RNA-interacting proteins and thus should enable us to globally profile the protein occupancy on the mRNA transcriptome.

We generated protein occupancy cDNA libraries for two biological replicates. In brief, we crosslinked 4SU-labeled cells and purified protein-mRNA complexes using oligo(dT) beads. The precipitate was treated with RNase I to reduce the protein-crosslinked RNA fragments to a length of about 30–60 nt. To remove noncrosslinked RNA, protein-RNA complexes were precipitated with ammonium sulfate and blotted onto nitrocellulose. The RNA was recovered by Proteinase K treatment, ligated to cloning adapters, and reverse transcribed. The resulting cDNA libraries were PCR-amplified and next-generation sequenced (Table S6).

When mapping the sequence reads to the human reference genome, we observed diagnostic T-C changes (Figures 6A and 6B) for both profiling libraries, indicative for crosslinking of 4SU-containing RNA to proteins (Hafner et al., 2010). The majority of the sequence reads mapped to mRNA sequences (86% and 81%; Figures 6C and 6D), confirming that the bulk of oligo(dT)-precipitated transcripts were derived from protein-coding genes and therefore the purified proteins predominately bound to mRNA. A comparison of a transcriptome-wide sequence-normalized read count indicated that the proteins preferentially bound exons over introns (Figure S6).

To assess the reproducibility of our assay, we computed rank correlation coefficients for all transcripts using a sliding window approach to compare sequence coverage over entire transcripts. Figure 6E shows the density distribution of rank correlation coefficients for corresponding transcripts in both experiments (median 0.712) compared to the correlation of randomly selected unrelated transcripts (median 0.015). Next we compared the median coverage over entire transcripts (median of all windows for each transcript) between replicate experiments (Figure 6F) and obtained a rank correlation coefficient of 0.984, suggesting a high degree of similarity between replicate experiments, both in coverage signal for individual transcript regions and overall transcript sequence coverage.

We further analyzed the reproducibility of the occurrence of T-C changes at specific positions and found high agreement between the two profiles (e.g., about 80% of the T-C positions with at least 5 nucleotide changes in one replicate showed at least two transitions in the other experiment [Figure 6G]). Finally, we correlated the absolute number of T-C changes at corresponding transcript positions, considering only sites with at least two transitions in one of the profiles, resulting in a high Pearson correlation coefficient of 0.862 (Figure 6H).

We generated a consensus occupancy profile by using the mean number of T-C changes at positions with at least two T-C changes in each of the two libraries. The transcriptome-wide occupancy profile is available at http://dorina.mdc-berlin.de/cgi-bin/hgTracks/ (Anders et al., 2012). Figure 6I shows the consensus T-C transition profile and mean sequence coverage of reads mapping to the genomic region encoding *EEF2*. As expected, T-C changes and sequence coverage were higher in exonic compared to intronic sequences.

Zooming into the 3′ UTR of *EEF2* (Figures 7A and S7A) as well as the 3′ UTRs of *CBX3* (Figure 7B) and *TP53* (Figure 7C) we observed distinct T-C transition profiles indicating regions of protein binding. Intriguingly, three distinct regions with T-C changes in the *TP53* 3′ UTR overlap with previously determined RNA-binding sites, identified either by deletion studies and/or PAR-CLIP experiments (Figure 7C).

To assess whether the occupancy profile indeed reflects binding patterns of RNA interactors, we compared the T-C transition probability around miRNA binding sites in AGO PAR-CLIPs and the occupancy profile. In both cases we observed an increased probability of T-C changes upstream of miRNA binding sites (Figures 7D and 7E), suggesting that the occupancy profile recapitulates the T-C transition pattern of AGO PAR-CLIPs even in the context of other RNA binders. Furthermore, we observed T-C changes in 76% of 32163 AGO binding sites, suggesting that the occupancy profiles encloses the majority of contact sites of this protein.

To estimate the general distribution of protein binding to different transcript regions, we averaged the relative density of T-C changes in distinct exonic sequences. While protein binding to 3′ UTRs was equally distributed, binding in 5′ UTRs and CDS showed a preference for 3′ regions (Figure 7F).

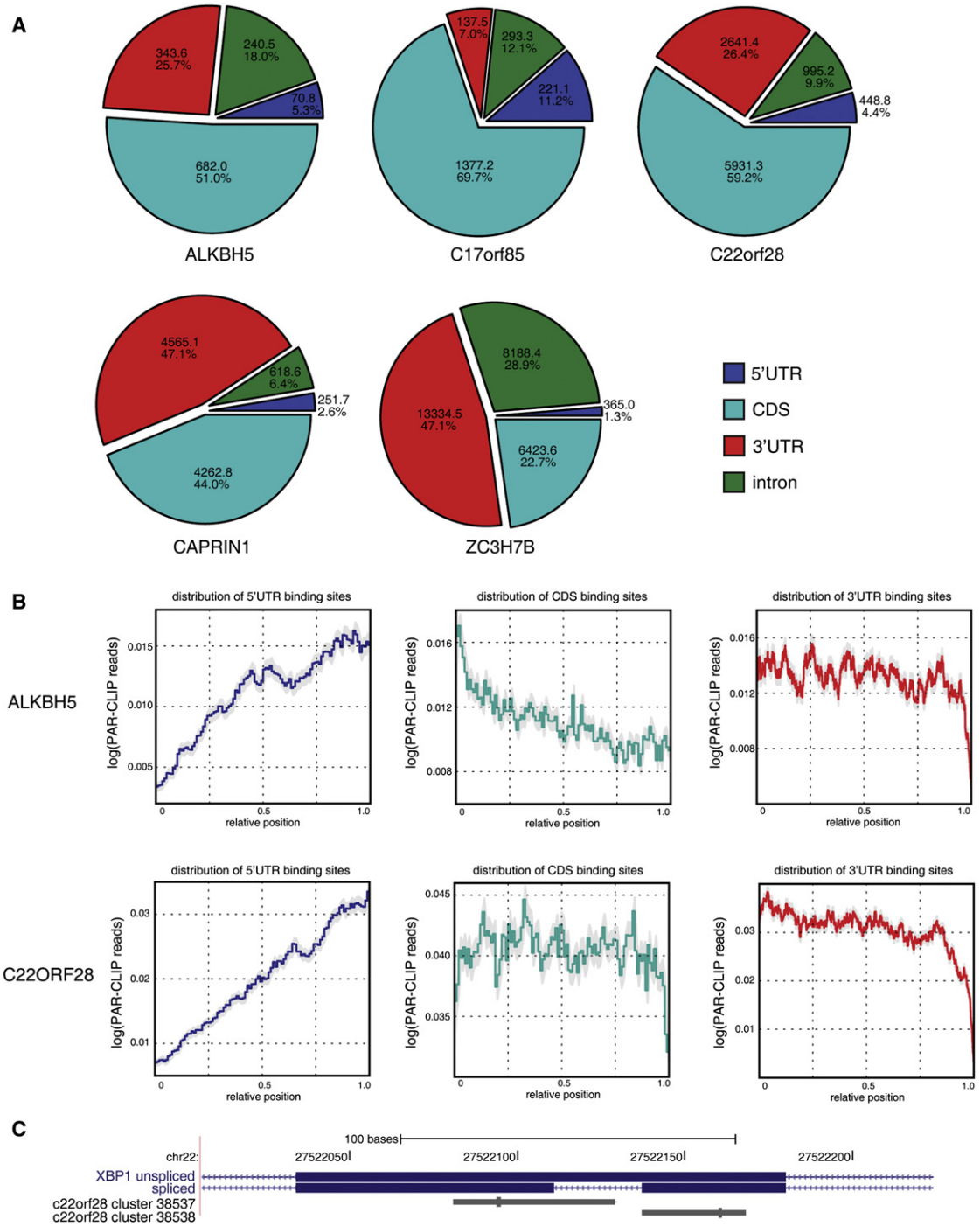


Figure 5. PAR-CLIP Analysis of Candidate RNA-Binding Proteins

(A) Distribution of mRNA binding sites based on PAR-CLIP sequence clusters for the indicated proteins are shown. The absolute number and percentage distribution of sequence clusters in different transcript regions are indicated.

(B) PAR-CLIP sequence coverage along transcript regions is shown for ALKBH5 and C22orf28.

(C) Genome browser view of spliced and unspliced *XBP1* transcript isoforms. Putative C22orf28 binding sites flanking the *XBP1* intron are indicated in dark gray. See also Figure S5 and Table S5.

Since we were unable to differentiate whether RNA fragments mapping to mRNA coding sequences were crosslinked to RNA-binding proteins or to translating ribosomes, we further focused our analysis on 3′ UTR sequences. The occupancy profiles indicated that extensive regions within 3′ UTRs can be bound by proteins. A transcriptome-wide analysis of 3′ UTRs showed

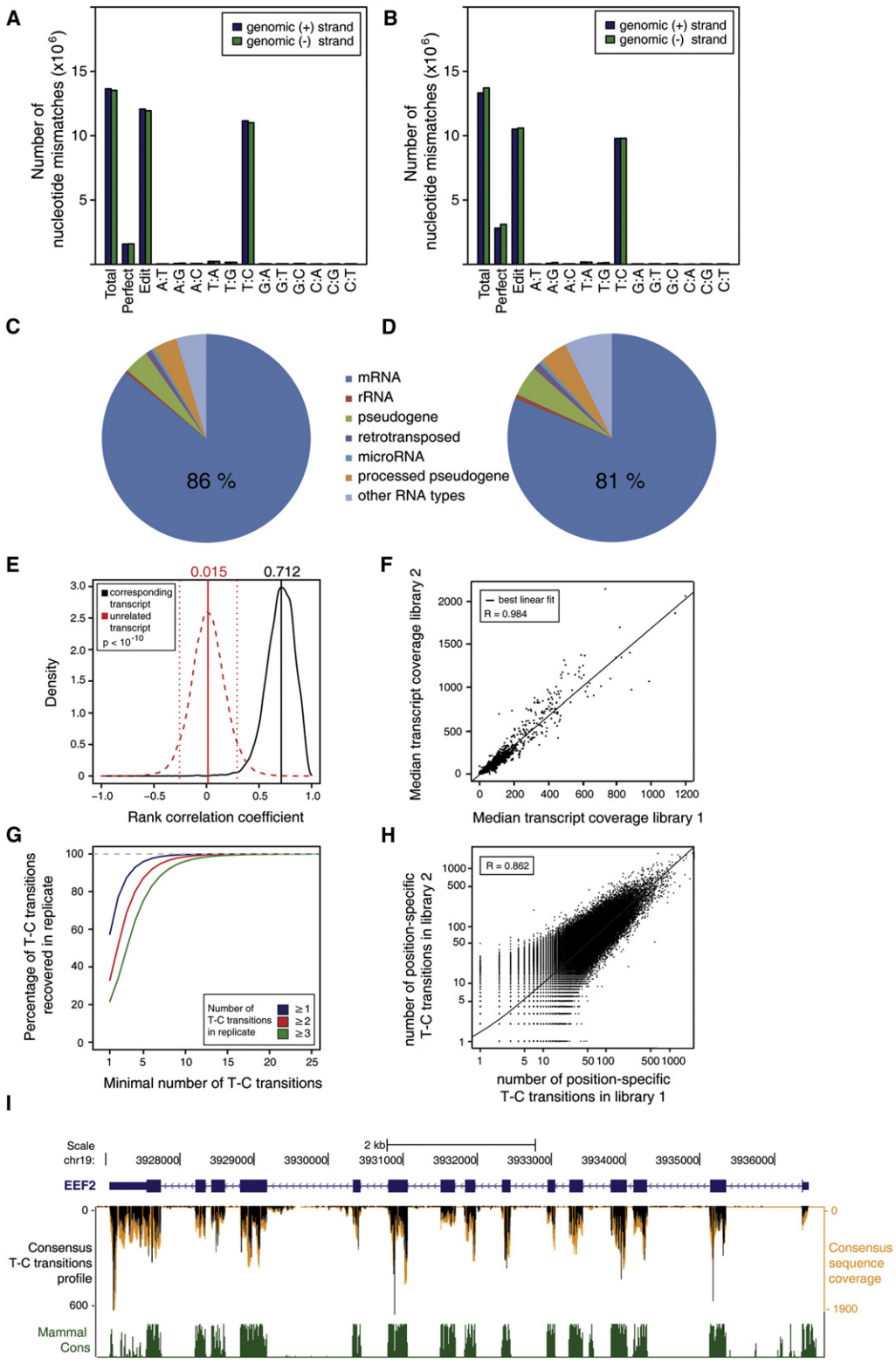


Figure 6. Comparison of Two Independently Generated Protein Occupancy Profiles

(A and B) Specific mismatches in aligned sequence reads demonstrate efficient protein-RNA crosslinking. The frequency of nucleotide mismatches in occupancy profiling reads aligned to human genome is shown for library 1 (A) and 2 (B). T-C mismatches are the signature of efficient crosslinking of 4SU-labeled RNA to protein.

that 28% of uridines were converted to cytidines (Figure S7B), arguing for widespread protein contacts in this transcript region during the life cycle of polyadenylated mRNAs.

Assuming that the minimal RNA binding region of a protein is at least 3 nucleotides centered around a crosslinked uridine, we analyzed the evolutionary conservation of such contact sites across 44 vertebrate species and observed a significantly elevated PhyloP conservation score (Pollard et al., 2010) (Figure 7G), suggesting that the crosslinked regions are of functional importance. Next we extended our analysis by examining the density of single nucleotide polymorphisms (SNPs) in minimal RNA binding regions centered around positions with T-C changes. Crosslinked regions showed a significantly lower SNP frequency compared to noncrosslinked control regions (T-C = 0.004106, non-T-C = 0.005663, binominal test: p value $< 2.2 \times 10^{-16}$), suggesting that sites with T-C changes are under stronger negative selection in humans further supporting their functional relevance.

Putative RNA *Cis*-Regulatory Elements with Trait/Disease-Associated Polymorphisms

SNPs occurring in binding sites of RNA-interacting proteins could be a contributing factor to *cis*-modulation of gene expression by changing the affinity of a regulatory protein to untranslated RNA regions. We examined trait/disease-associated SNPs (TASs), obtained from a listing of genome-wide association studies (Hindorff et al., 2009), for their presence in potential RNA binding sites. We focused on TASs within 10 nt around crosslinking sites. In total, we identified 28 TASs within potential protein binding sites in introns and 3′ UTRs of mRNAs as well as intergenic regions (Table S7). As shown in Figures S7C and S7D, rs9299 and rs8321 are TASs that are located in the 3′ UTRs of *HOXB5* and *ZNRD1*, respectively. rs9299 has been reported to be linked to childhood obesity (Bradfield et al., 2012), while rs8321 was described to be associated with AIDS progression (Limou et al., 2009).

DISCUSSION

Maturation, localization, decay and translational regulation of mRNAs involve the formation of complexes of RNA-interacting proteins with their target transcripts (Martin and Ephrussi, 2009; Moore and Proudfoot, 2009). Here, we present an approach to characterize the protein-mRNA interactome of a human cell line, based on in vivo UV crosslinking of proteins to mRNA followed by oligo(dT) affinity purification.

Using quantitative proteomics, we identified close to 800 proteins, which were isolated based on their ability to crosslink to thionucleotide-labeled polyadenylated RNA. Sequencing of RNA in the oligo(dT) precipitate and RNA crosslinked to the copurified proteins showed that the majority of transcripts were derived from protein-coding genes. Close to 90% of the identified mRNA binders were observed in at least two purifications of protein-mRNA complexes of crosslinked cells. As expected a majority of the mRNA binders were proteins previously described to interact with RNA based on their functional annotation as RNA-binding proteins, helicases, nucleases, and RNA-modifying enzymes. Two hundred forty-five proteins had previously not been shown to interact with RNA nor have recognizable RNA interaction domains, indicating the need for experimental methods to discover RNA binders.

A subset of the identified mRNA binders has been implicated in diseases. Specific genotypic variations in 59 out of the 797 observed proteins lead to monogenic disorders (Table S1), including neurogenerative diseases (Alzheimer's, Fragile X syndrome, amytrophic lateral sclerosis, and spinocerebellar ataxias), muscular diseases (myotonic dystrophy and Emery-Dreyfuss muscular dystrophy), and cancers (Ewing sarcoma and chondrosarcoma). Thirteen of the disease-associated proteins have, to our knowledge, previously not been shown to interact with RNA (Table S1), suggesting modulation of RNA metabolism in the respective diseases.

The mRNA-interacting proteins also provide interesting insights into how posttranscriptional regulation is connected to other cellular pathways and regulatory mechanisms. In particular, transcription seems to be tightly coupled to the subsequent RNA metabolism. Several proteins, for which we confirmed their RNA-binding activity, were shown to function in transcriptional regulation. KIAA1967/DBC1, also known as Deleted in Breast cancer 1, was initially identified as an inhibitor of the histone acetyltransferase SIRT1 (Kim et al., 2008). Recently, KIAA1967 was shown to be present in a complex that integrates alternative mRNA splicing with RNA polymerase II transcript elongation (Close et al., 2012). Another identified RNA binder is the Myb-binding protein 1a (MYBBP1A). MYBBP1A interacts with and regulates the activity of several transcription factors, including c-Myb (Favier and Gonda, 1994) and NFκB (Owen et al., 2007). Likewise EDF1, also identified as RNA-binding, interacts with the basic leucine zipper proteins, ATF1, c-Jun, and c-Fos, and acts as transcriptional coactivator (Kabe et al., 1999). It is presently unknown by what mechanism these proteins modulate

(C and D) Distribution of mapped reads to different RNA types for library 1 (C) and 2 (D).
(E) Density of transcript-wise rank correlation coefficients based on sequence coverage of the two libraries between corresponding (black) and unrelated (red) transcripts. Solid lines indicate medians, and dashed lines indicate the 5% and 95% quantiles, respectively.
(F) Scatterplot of median transcript-coverage values of the two libraries. The solid line represents the best linear fit. The rank correlation coefficient based on all pair-wise comparisons is indicated.
(G) Reproducibility of individual T-C transition sites. The reproducibility was measured as the percentage of sites with a minimal number of T-C transitions, which also showed a certain number of transitions (≥ 1, ≥ 2, ≥ 3) in the replicate experiment.
(H) Scatterplot of absolute numbers of position-specific T-C transition events for all T positions inside transcripts, which showed at least two transitions in one of the two replicates. The solid line indicates the best linear fit. The Pearson correlation coefficient is indicated.
(I) Browser view of genomic region encoding *EEF2* gene. Consensus T-C transition profile (in black number of T-C transitions) and sequence coverage (orange) are indicated. T-C transitions are diagnostic for crosslinking sites. Phastcon conservation of placental mammals is shown in green.
See also Figure S6 and Table S6.

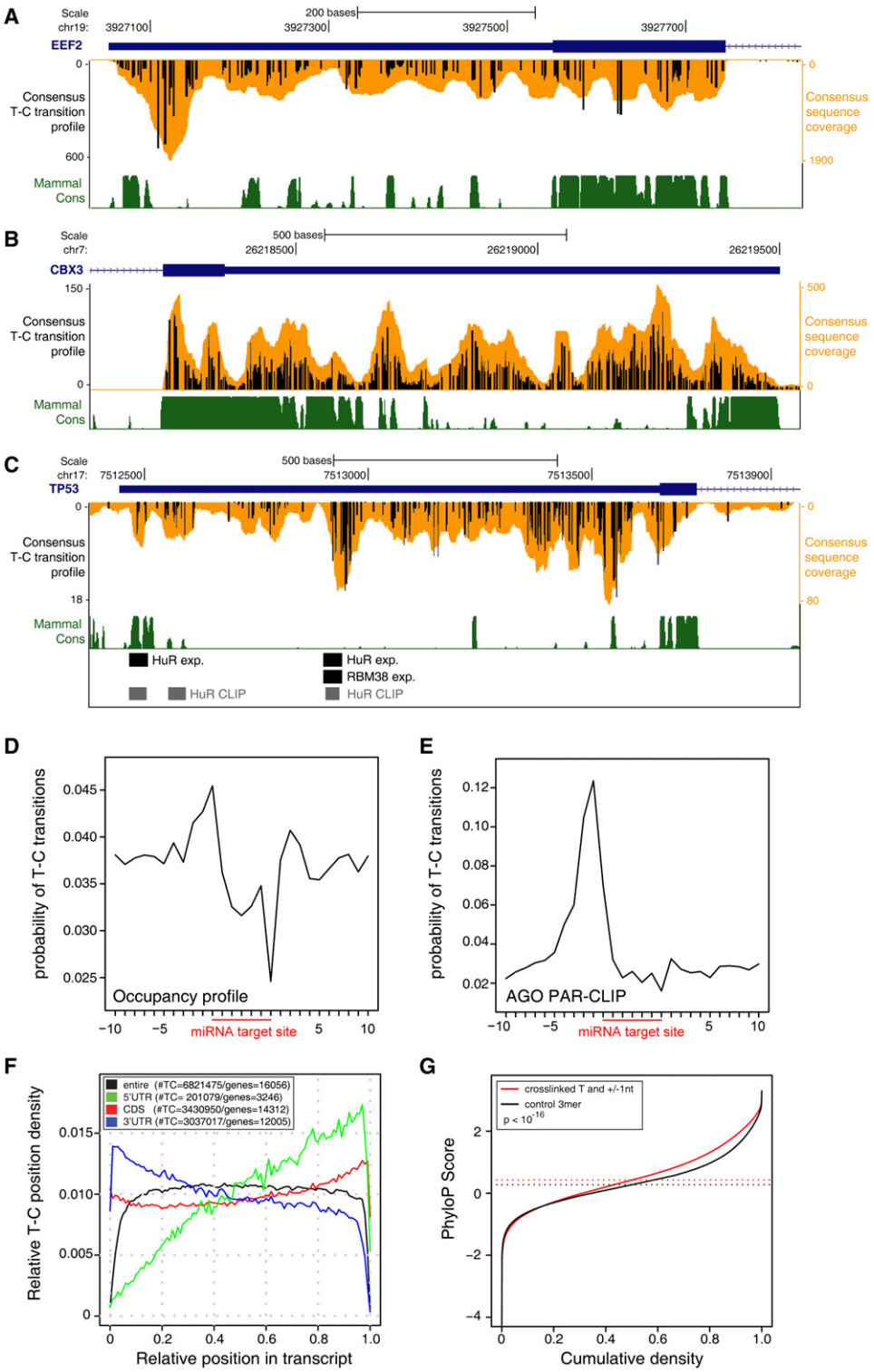


Figure 7. Protein Occupancy Profiling on mRNA Provides a Global Map of Potential *Cis*-Regulatory Transcript Regions

(A–C) Browser view of genomic regions encoding 3′ UTRs of *EEF2* (A), *CBX3* (B), and *TP53* (C) genes. Consensus T-C transition profile (in black) and sequence coverage (orange) are indicated. Phastcon conservation of placental mammals is shown in green. The binding sites of HuR/ELAVL1 in the *TP53* 3′ UTR that were determined by PAR-CLIP (Kishore et al., 2011) are shown in gray, and HuR/ELAVL1 (Mazan-Mamczarz et al., 2003; Zou et al., 2006) and RBM38 (Zhang et al., 2011) binding sites determined by other experimental approaches are indicated in black.

(D and E) Probability of T-C transitions around miRNA binding sites in occupancy profile and AGO PAR-CLIPs (Hafner et al., 2010).

transcription and whether the RNA binding function is required for this activity.

Recent studies identifying RNA-binding proteins in yeast revealed a large number of cytoplasmic proteins with catalytic activities, many of them acting in metabolism (Scherrer et al., 2010; Tsvetanova et al., 2010). In contrast, our study revealed only eleven metabolic enzymes among the 797 experimentally determined RNA-interacting proteins. However, we discovered a number of nonmetabolic enzymes. Among them were C22orf28 and ALKBH5, two proteins that possess catalytic activities. Our findings suggest that C22orf28 is the elusive RNA ligase involved in the cytoplasmic nuclease-mediated splicing of the *XBP1* mRNA (Uemura et al., 2009). On the other hand ALKBH5, found only in vertebrates, possibly functions in oxidative RNA demethylation, since it shows similarity to the *Escherichia coli* DNA methylation repair enzyme AlkB and possesses a 2-oxoglutarate oxygenase activity (Thalhammer et al., 2011). Interestingly, our set of mRNA binders also included other RNA-modifying enzymes, among them the methyltransferase NSUN2. Despite its narrow substrate range, catalyzing a 5-methylcytosine modification on tRNAs, NSUN2 might have a broader role in mRNA modification as evidenced by a recent finding of widespread occurrence of 5-methylcytosine in human mRNA (Squires et al., 2012). The identification of ALKBH5, NSUN2, and several other potential RNA-modifying enzymes suggests that base modifications in mRNA might be more prevalent than anticipated.

Complementing the identification of the mRNA-bound proteome, we were able to determine the mRNA regions that can crosslink to proteins in HEK293 cells. One of the most interesting outcomes was that, during the life cycle of an mRNA molecule, widespread regions of the 3′ UTRs provide contact sites for RNA-interacting proteins. Up to 28% of all uridines present in 3′ UTRs showed diagnostic T-C changes in the protein occupancy profiling sequence reads. This number is reasonably high, considering observations that typically only one of a few uridines in RNA binding sites, when substituted by 4SU, crosslinks to proteins (Hafner et al., 2010). The evolutionary conservation of crosslinked sites suggested that the identified protein-bound RNA segments are of functional importance. In the future, a central task will be to overlap occupied regions with evolutionary constrained sequences and RNA candidate structures (Lindblad-Toh et al., 2011) as well as with RNA interaction data of individual proteins, to identify specific RNA regulatory elements and their structural contexts.

Our results support the view that transcripts are generally bound and regulated by multiple RNA-interacting proteins (Keene, 2007). The combinatorial assembly of *cis*-regulatory factors, which takes place in a spatial and time-resolved manner, determines the fate of an mRNA molecule. Untranslated regions of protein-coding transcripts seem to provide ample sequence elements for proteins to bind and to function in the regulation of mRNA biogenesis, localization, decay, and translation. Until now, comprehensive high-resolution mapping of protein-RNA interactions using different CLIP approaches lead to the discovery of protein-RNA interaction sites that control distinct posttranscriptional processes. However, these studies focused on the binding specificity and function of single RNA-binding proteins (Hafner et al., 2010; König et al., 2010; Ule et al., 2003). Conversely, our dual approach should allow monitoring of quantitative changes in the protein-mRNA interactome, by analyzing differential protein binding to mRNA and its occupancy on specific RNA regions, in response to intra- and extracellular signals in an unbiased manner.

Additionally, the protein occupancy profile narrows the genomic sequence search space for *cis*-regulatory elements in untranslated mRNA regions. The identification of occupied mRNA sites will be valuable for the examination of rapidly emerging data on genetic variation between individuals. Some polymorphic variations within a population possibly contribute to complex traits and diseases by impacting posttranscriptional and/or translational regulation of gene expression. In addition, we envision that the knowledge of RNA-binding sites will broaden the use of antisense molecules to modulate gene expression in experimental as well as therapeutic applications (Kole et al., 2012).

In summary, the identification of the mRNA-bound proteome and its occupancy profile on protein-coding transcripts offers a systems-wide view on the protein-mRNA interactome, describing its components and the RNA sites of interactions. Application of this approach in the future will greatly contribute to a better understanding of cellular functions of mRNP complexes with the goal to elucidate the posttranscriptional regulatory code that defines growth, differentiation, and disease.

EXPERIMENTAL PROCEDURES

Identification of mRNA-Interacting Proteins

HEK293 cells were grown in medium supplemented with 4-thiouridine and 6-thioguanosine. Living cells, grown in light SILAC medium, were irradiated with 365 nm UV light whereas the control cells, grown in heavy SILAC medium were not UV crosslinked. In the label swap experiment, the cells grown in heavy SILAC medium were crosslinked and the cells grown in light SILAC medium were used as control. After crosslinking, cells were harvested and lysed in lysis/binding buffer. Oligo(dT) beads were added to cell extract and incubated at room temperature. Beads were extensively washed in lysis/binding buffer containing 1% lithium dodecyl sulfate. Protein-mRNA complexes were eluted from beads in elution buffer. For mass spectrometry the RNA was removed by nuclease treatment. Digested protein samples were prepared for mass spectrometry analysis as described in the Supplemental Experimental Procedures.

Validation of RNA-Binding Activity

Cells, stably expressing His/FLAG/HA-tagged proteins, were labeled with 100 μM 4SU, UV irradiated, and lysed in NP-40 lysis buffer. 4SU-labeled nonirradiated cells were used as control. Immunoprecipitation was carried

(F) Relative density of T-C positions along the entire transcript and different regions (5′ UTR, CDS, 3′ UTR) are shown.
(G) Comparison of PhyloP score of 3-mer sequences centered around crosslinked T (red) to random noncrosslinked 3-mers (black) is shown. The p value indicates the significance of the difference of the PhyloP score distribution between crosslinked and control regions as given by a two-sample Kolmogorov-Smirnov test.
See also Figure S7 and Table S7.

out with anti-FLAG magnetic beads. Bound protein-RNA complexes were 5′ end labeled with T4 Polynucleotide Kinase and γ-^{32}P-ATP. The crosslinked mRNP complexes were resolved on a gradient gel, and the corresponding protein-RNA complexes were analyzed by phosphorimaging and western blotting.

PAR-CLIP

PAR-CLIP protocol was performed as described (Hafner et al., 2010). In contrast to the published cloning procedure the 3′ ligation was performed with barcoded 3′ adapters. The PAR-CLIP cDNA sequencing data was analyzed using the PAR-CLIP analysis pipeline (Lebedeva et al., 2011).

Protein Occupancy Profiling on mRNA

HEK293 cells were grown in medium supplemented with 200 μM 4SU 16 hr prior to crosslinking. Harvested cells were resuspended of lysis/binding buffer. Oligo(dT) precipitation was performed as described earlier. Eluted protein-RNA complexes were RNase I treated, followed by ammonium sulfate precipitation. Precipitate was separated by SDS-PAGE and transferred onto a nitrocellulose membrane. RNA was extracted from membrane by proteinase K treatment and phenol/chloroform extraction. Recovered RNA was dephosphorylated using CIP. After dephosphorylation RNA was phenol/chloroform extracted, ethanol precipitated and 5′ end labeled with T4 PNK. Radiolabeled RNA was extracted and subsequent small RNA cloning and adaptor ligations were performed as described previously (Hafner et al., 2010).

More-detailed description of the methods is provided in the Supplemental Experimental Procedures

ACCESSION NUMBERS

The sequencing data have been deposited in the GEO database under the accession number GSE38157.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and seven tables and can be found with this article online at doi:10.1016/j.molcel.2012.05.021.

ACKNOWLEDGMENTS

We would like to express our gratitude to Nikolaus Rajewsky (MDC) and members of his lab for sharing the PAR-CLIP computational analysis pipeline for this study as well as Claudia Langnick and Mirjam Feldkamp from the Wei Chen lab (MDC) for sequencing. We thank the volunteers participating in the Human Proteome Folding Project on IBM's World Community Grid. As part of the Berlin Institute for Medical Systems Biology at the MDC, the research group of M.L. is funded by the Federal Ministry for Education and Research (BMBF) and the Senate of Berlin, Berlin, Germany. A.B. and M.M. are funded by the MDC/NYU PhD program.

Received: March 20, 2012
Revised: May 14, 2012
Accepted: May 17, 2012
Published online: June 8, 2012

REFERENCES

Adam, S.A., Nakagawa, T., Swanson, M.S., Woodruff, T.K., and Dreyfuss, G. (1986). mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. Mol. Cell. Biol. *6*, 2932–2943.

Anders, G., Mackowiak, S.D., Jens, M., Maaskola, J., Kuntzagk, A., Rajewsky, N., Landthaler, M., and Dieterich, C. (2012). doRiNA: a database of RNA interactions in post-transcriptional regulation. Nucleic Acids Res. *40* (Database issue), D180–D186.

Andersen, J.S., Lam, Y.W., Leung, A.K., Ong, S.E., Lyon, C.E., Lamond, A.I., and Mann, M. (2005). Nucleolar proteome dynamics. Nature *433*, 77–83.

Aravind, L., Iyer, L.M., and Anantharaman, V. (2003). The two faces of Alba: the evolutionary connection between proteins participating in chromatin structure and RNA metabolism. Genome Biol. *4*, R64.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al; The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Bessonov, S., Anokhina, M., Will, C.L., Urlaub, H., and Lührmann, R. (2008). Isolation of an active step I spliceosome and composition of its RNP core. Nature *452*, 846–850.

Bono, F., Ebert, J., Lorentzen, E., and Conti, E. (2006). The crystal structure of the exon junction complex reveals how it maintains a stable grip on mRNA. Cell *126*, 713–725.

Bradfield, J.P., Taal, H.R., Timpson, N.J., Scherag, A., Lecoeur, C., Warrington, N.M., Hypponen, E., Holst, C., Valcarcel, B., Thiering, E., et al; the Early Growth Genetics (EGG) Consortium. (2012). A genome-wide association meta-analysis identifies new childhood obesity loci. Nat Genet., in press. Published online April 8, 2012.

Choi, Y.D., and Dreyfuss, G. (1984). Isolation of the heterogeneous nuclear RNA-ribonucleoprotein complex (hnRNP): a unique supramolecular assembly. Proc. Natl. Acad. Sci. USA *81*, 7471–7475.

Close, P., East, P., Dirac-Svejstrup, A.B., Hartmann, H., Heron, M., Maslen, S., Chariot, A., Söding, J., Skehel, M., and Svejstrup, J.Q. (2012). DBIRD complex integrates alternative mRNA splicing with RNA polymerase II transcript elongation. Nature *484*, 386–389.

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.

de Lima Morais, D.A., Fang, H., Rackham, O.J., Wilson, D., Pethica, R., Chothia, C., and Gough, J. (2011). SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res. *39* (Database issue), D427–D434.

Denhez, F., and Lafyatis, R. (1994). Conservation of regulated alternative splicing and identification of functional domains in vertebrate homologs to the Drosophila splicing regulator, suppressor-of-white-apricot. J. Biol. Chem. *269*, 16170–16179.

Dölken, L., Ruzsics, Z., Rädle, B., Friedel, C.C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P., and Koszinowski, U.H. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. RNA *14*, 1959–1972.

Drew, K., Winters, P., Butterfoss, G.L., Berstis, V., Uplinger, K., Armstrong, J., Riffle, M., Schweighofer, E., Bovermann, B., Goodlett, D.R., et al. (2011). The Proteome Folding Project: proteome-scale prediction of structure and function. Genome Res. *21*, 1981–1994.

Favier, D., and Gonda, T.J. (1994). Detection of proteins that bind to the leucine zipper motif of c-Myb. Oncogene *9*, 305–311.

Favre, A., Moreno, G., Salet, C., and Vinzens, F. (1993). 4-Thiouridine incorporation into the RNA of monkey kidney cells (CV-1) triggers near-UV light long-term inhibition of DNA, RNA and protein synthesis. Photochem. Photobiol. *58*, 689–694.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol. Cell. Proteomics *11*, M111, 014050.

Greenberg, J.R. (1979). Ultraviolet light-induced crosslinking of mRNA to proteins. Nucleic Acids Res. *6*, 715–732.

Haas, S., Steplewski, A., Siracusa, L.D., Amini, S., and Khalili, K. (1995). Identification of a sequence-specific single-stranded DNA binding protein that suppresses transcription of the mouse myelin basic protein gene. J. Biol. Chem. *270*, 12503–12510.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M., et al. (2010).

Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129–141.

Hassfeld, W., Chan, E.K., Mathison, D.A., Portman, D., Dreyfuss, G., Steiner, G., and Tan, E.M. (1998). Molecular definition of heterogeneous nuclear ribonucleoprotein R (hnRNP R) using autoimmune antibody: immunological relationship with hnRNP P. Nucleic Acids Res. *26*, 439–445.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.

Jackson, R.J., Hellen, C.U., and Pestova, T.V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. Nat. Rev. Mol. Cell Biol. *11*, 113–127.

Kabe, Y., Goto, M., Shima, D., Imai, T., Wada, T., Morohashi, K., Shirakawa, M., Hirose, S., and Handa, H. (1999). The role of human MBF1 as a transcriptional coactivator. J. Biol. Chem. *274*, 34196–34202.

Keene, J.D. (2007). RNA regulons: coordination of post-transcriptional events. Nat. Rev. Genet. *8*, 533–543.

Kennedy, M.C., Mende-Mueller, L., Blondin, G.A., and Beinert, H. (1992). Purification and characterization of cytosolic aconitase from beef liver and its relationship to the iron-responsive element binding protein. Proc. Natl. Acad. Sci. USA *89*, 11730–11734.

Kiledjian, M., and Dreyfuss, G. (1992). Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. EMBO J. *11*, 2655–2664.

Kim, J.E., Chen, J., and Lou, Z. (2008). DBC1 is a negative regulator of SIRT1. Nature *451*, 583–586.

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

Knapinska, A.M., Gratacós, F.M., Krause, C.D., Hernandez, K., Jensen, A.G., Bradley, J.J., Wu, X., Pestka, S., and Brewer, G. (2011). Chaperone Hsp27 modulates AUF1 proteolysis and AU-rich element-mediated mRNA degradation. Mol. Cell. Biol. *31*, 1419–1431.

Kole, R., Krainer, A.R., and Altman, S. (2012). RNA therapeutics: beyond RNA interference and antisense oligonucleotides. Nat. Rev. Drug Discov. *11*, 125–140.

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

Lebedeva, S., Jens, M., Theil, K., Schwanhäusser, B., Selbach, M., Landthaler, M., and Rajewsky, N. (2011). Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. Mol. Cell *43*, 340–352.

Lee, I., and Hong, W. (2004). RAP—a putative RNA-binding domain. Trends Biochem. Sci. *29*, 567–570.

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

Limou, S., Le Clerc, S., Coulonges, C., Carpentier, W., Dina, C., Delaneau, O., Labib, T., Taing, L., Sladek, R., Deveau, C., et al; ANRS Genomic Group. (2009). Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). J. Infect. Dis. *199*, 419–426.

Lindberg, U., and Sundquist, B. (1974). Isolation of messenger ribonucleoproteins from mammalian cells. J. Mol. Biol. *86*, 451–468.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al; Broad Institute Sequencing Platform and Whole Genome Assembly Team; Baylor College of Medicine Human Genome Sequencing Center Sequencing Team; Genome Institute at Washington University. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482.

Mann, M. (2006). Functional and quantitative proteomics using SILAC. Nat. Rev. Mol. Cell Biol. *7*, 952–958.

Martin, K.C., and Ephrussi, A. (2009). mRNA localization: gene expression in the spatial dimension. Cell *136*, 719–730.

Mazan-Mamczarz, K., Galbán, S., López de Silanes, I., Martindale, J.L., Atasoy, U., Keene, J.D., and Gorospe, M. (2003). RNA-binding protein HuR enhances p53 translation in response to ultraviolet light irradiation. Proc. Natl. Acad. Sci. USA *100*, 8354–8359.

Milek, M., Wyler, E., and Landthaler, M. (2012). Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing. Semin. Cell Dev. Biol. *23*, 206–212.

Moore, M.J., and Proudfoot, N.J. (2009). Pre-mRNA processing reaches back to transcription and ahead to translation. Cell *136*, 688–700.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. *9* (*Suppl 1*), S4.

Owen, H.R., Elser, M., Cheung, E., Gersbach, M., Kraus, W.L., and Hottiger, M.O. (2007). MYBBP1a is a novel repressor of NF-kappaB. J. Mol. Biol. *366*, 725–736.

Peña-Castillo, L., Tasan, M., Myers, C.L., Lee, H., Joshi, T., Zhang, C., Guan, Y., Leone, M., Pagnani, A., Kim, W.K., et al. (2008). A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biol. *9* (*Suppl 1*), S2.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Popow, J., Englert, M., Weitzer, S., Schleiffer, A., Mierzwa, B., Mechtler, K., Trowitzsch, S., Will, C.L., Lührmann, R., Söll, D., and Martinez, J. (2011). HSPC117 is the essential subunit of a human tRNA splicing ligase complex. Science *331*, 760–764.

Scherrer, T., Mittal, N., Janga, S.C., and Gerber, A.P. (2010). A screen for RNA-binding proteins in yeast indicates dual functions for many enzymes. PLoS ONE *5*, e15499.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Setyono, B., and Greenberg, J.R. (1981). Proteins associated with poly(A) and other regions of mRNA and hnRNA molecules as investigated by crosslinking. Cell *24*, 775–783.

Shiina, N., Shinkura, K., and Tokunaga, M. (2005). A novel RNA-binding protein in neuronal RNA granules: regulatory machinery for local translation. J. Neurosci. *25*, 4420–4434.

Silvera, D., Koloteva-Levine, N., Burma, S., and Elroy-Stein, O. (2006). Effect of Ku proteins on IRES-mediated translation. Biol. Cell *98*, 353–361.

Squires, J.E., Patel, H.R., Nousch, M., Sibbritt, T., Humphreys, D.T., Parker, B.J., Suter, C.M., and Preiss, T. (2012). Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Res., in press. Published online February 16, 2012. 10.1093/nar/gks144.

Thalhammer, A., Bencokova, Z., Poole, R., Loenarz, C., Adam, J., O'Flaherty, L., Schödel, J., Mole, D., Giaslakiotis, K., Schofield, C.J., et al. (2011). Human AlkB homologue 5 is a nuclear 2-oxoglutarate dependent oxygenase and a direct target of hypoxia-inducible factor 1α (HIF-1α). PLoS ONE *6*, e16210.

Ting, N.S., Yu, Y., Pohorelic, B., Lees-Miller, S.P., and Beattie, T.L. (2005). Human Ku70/80 interacts directly with hTR, the RNA component of human telomerase. Nucleic Acids Res. *33*, 2090–2098.

Tsvetanova, N.G., Klass, D.M., Salzman, J., and Brown, P.O. (2010). Proteome-wide search reveals unexpected RNA-binding proteins in Saccharomyces cerevisiae. PLoS ONE *5*, e12671.

Uemura, A., Oku, M., Mori, K., and Yoshida, H. (2009). Unconventional splicing of XBP1 mRNA occurs in the cytoplasm during the mammalian unfolded protein response. J. Cell Sci. *122*, 2877–2886.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212–1215.

Vitour, D., Lindenbaum, P., Vende, P., Becker, M.M., and Poncet, D. (2004). RoXaN, a novel cellular protein containing TPR, LD, and zinc finger motifs, forms a ternary complex with eukaryotic initiation factor 4G and rotavirus NSP3. J. Virol. *78*, 3851–3862.

Wagenmakers, A.J., Reinders, R.J., and van Venrooij, W.J. (1980). Cross-linking of mRNA to proteins by irradiation of intact cells with ultraviolet light. Eur. J. Biochem. *112*, 323–330.

Yoshida, H., Matsui, T., Yamamoto, A., Okada, T., and Mori, K. (2001). XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. Cell *107*, 881–891.

Zhang, J., Cho, S.J., Shu, L., Yan, W., Guerrero, T., Kent, M., Skorupski, K., Chen, H., and Chen, X. (2011). Translational repression of p53 by RNPC1, a p53 target overexpressed in lymphomas. Genes Dev. *25*, 1528–1543.

Zou, T., Mazan-Mamczarz, K., Rao, J.N., Liu, L., Marasa, B.S., Zhang, A.H., Xiao, L., Pullmann, R., Gorospe, M., and Wang, J.Y. (2006). Polyamine depletion increases cytoplasmic levels of RNA-binding protein HuR leading to stabilization of nucleophosmin and p53 mRNAs. J. Biol. Chem. *281*, 19387–19394.