Resource

Cell

Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins

Alfredo Castello,[1,4] Bernd Fischer,[1,4] Katrin Eichelbaum,[1] Rastislav Horos,[1] Benedikt M. Beckmann,[1] Claudia Strein,[1] Norman E. Davey,[1] David T. Humphreys,[2] Thomas Preiss,[2,3] Lars M. Steinmetz,[1] Jeroen Krijgsveld,[1,*] and Matthias W. Hentze[1,2,*]

[1]European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, Heidelberg 69117, Germany

[2]Molecular Genetics Division, Victor Chang Cardiac Research Institute, Sydney NSW 2010, Australia

[3]Genome Biology Department, The John Curtin School of Medical Research, The Australian National University, Building 131, Garran Road, Acton ACT 0200, Australia

[4]These authors contributed equally to this work

*Correspondence: jeroen.krijgsveld@embl.de (J.K.), hentze@embl.de (M.W.H.)

DOI 10.1016/j.cell.2012.04.031

SUMMARY

RNA-binding proteins (RBPs) determine RNA fate from synthesis to decay. Employing two complementary protocols for covalent UV crosslinking of RBPs to RNA, we describe a systematic, unbiased, and comprehensive approach, termed "interactome capture," to define the mRNA interactome of proliferating human HeLa cells. We identify 860 proteins that qualify as RBPs by biochemical and statistical criteria, adding more than 300 RBPs to those previously known and shedding light on RBPs in disease, RNA-binding enzymes of intermediary metabolism, RNA-binding kinases, and RNA-binding architectures. Unexpectedly, we find that many proteins of the HeLa mRNA interactome are highly intrinsically disordered and enriched in short repetitive amino acid motifs. Interactome capture is broadly applicable to study mRNA interactome composition and dynamics in varied biological settings.

INTRODUCTION

RNA biology is orchestrated by the interplay of RNAs with RNA-binding proteins (RBPs) within dynamic ribonucleoproteins (RNPs) (Glisovic et al., 2008). Both the RBP repertoire and RBP activities of cells respond to a multitude of biological cues and environmental stimuli. Against this background, it is unsurprising that numerous diseases have been linked to defects in RBP expression and function, including neuropathies, muscular atrophies, metabolic disorders, and cancer (Cooper et al., 2009; Darnell, 2010; Lukong et al., 2008).

Intensive efforts have been undertaken to better understand RBPs, and much of our current knowledge of RNA-protein interactions has been accumulated stepwise for more than 20 years. Many RBPs interact with messenger RNAs (mRNAs) via a limited set of modular RNA-binding domains (RBDs), including the RNA recognition motif (RRM), heterogeneous nuclear RNP K-homology domain (KH), zinc fingers (Znf), etc. (Lunde et al., 2007). These motifs have informed in silico algorithms to identify other proteins harboring similar signatures as putative additional RBPs (Anantharaman et al., 2002). However, numerous noncanonical RBDs have been reported (Lee and Hong, 2004; Niessing et al., 2004; Rammelt et al., 2011; Zalfa et al., 2005), reflecting limitations in the scope of computational predictions. More recently, systematic experimental protocols for the identification and characterization of RBPs have been developed. Two studies using protein microarrays and RNA probes identified about 200 RBPs from budding yeast, including several novel candidates (Scherrer et al., 2010; Tsvetanova et al., 2010). In an alternative in vitro approach, stable isotope labeling by amino acids in cell culture (SILAC) and mass spectrometry (MS) were used to identify the association of polypeptides with immobilized RNA probes (Butter et al., 2009). The most abundant proteins captured by this assay matched bona fide RBPs that are known to bind to the respective RNA elements. However, this approach does not discriminate direct RNA-protein interactions from indirect protein-protein interactions with RBPs; moreover, bona fide RBPs cannot be distinguished from nonphysiological RNA binding. Thus, comprehensive in vivo mRNA interactomes have remained elusive.

To covalently couple proteins directly bound to RNA in vivo, UV light of 254 nm can be used to crosslink the naturally photoreactive nucleotide bases, especially pyrimidines, and specific amino acids (Phe, Trp, Tyr, Cys, and Lys) (Brimacombe et al., 1988; Hockensmith et al., 1986). Recently, photoactivatable-ribonucleoside-enhanced crosslinking (PAR-CL) has been popularized. The photoactivatable nucleotide 4-thiouridine (4SU) is taken up by cultured cells and incorporated into nascent RNAs, and efficient crosslinking is induced by 365 nm UV light irradiation (Hafner et al., 2010). UV crosslinking requires direct contact ("zero" distance) between protein and RNA and does not promote protein-protein crosslinking (Greenberg, 1979; Pashev et al., 1991; Suchanek et al., 2005). Both conventional

UV crosslinking (cCL) and PAR-CL have been used for the determination of RNAs bound by particular RBPs (Hafner et al., 2010; Licatalosi et al., 2008).

Because the crosslinking chemistries of cCL and PAR-CL are distinct (Greenberg, 1979; Wetzel and Söll, 1977), we used both techniques in parallel to determine "all" RBPs bound to polyadenylated RNA in HeLa cells, advancing work that started with heterogeneous nuclear ribonucleoproteins (hnRNPs) in the 1980s (Dreyfuss et al., 1984). We show that the in vivo HeLa mRNA interactome includes hundreds of proteins that were previously unknown to bind RNA, and we discuss resulting insights into RNA biology.

RESULTS AND DISCUSSION

In Vivo Capture of HeLa RBPs

To determine the repertoire of proteins that directly bind to mRNAs in living HeLa cells, we "froze" protein-mRNA interactions by covalent UV crosslinking. Taking advantage of the complementary crosslinking chemistries of cCL (254 nm) and PAR-CL (4SU/365 nm), we employed both techniques in parallel. RBPs covalently bound to polyadenylated RNAs in vivo are captured on oligo(dT) magnetic beads following cell lysis. Unlike strategies based on antibodies or protein tags, nucleic acid hybridization allows the use of highly stringent biochemical conditions to minimize contaminations, including 500 mM lithium chloride and lithium dodecyl sulfate (LiDS; 0.5%). Following stringent washes, proteins are released by RNase treatment and are identified using MS (Figure 1A).

With this protocol, which we term "interactome capture," mRNAs are enriched over 18S ribosomal RNA (rRNA), accompanied by a substantial decrease in total RNA levels after oligo(dT) pull-down (Figures 1B and S1A available online). The β-actin, glyceraldehyde 3-phosphate dehydrogenase (GAPDH), and thymidylate synthase (TS) mRNAs are efficiently isolated (recovering 25%–70% of the starting material) following both cCL or PAR-CL. Enrichment of mRNAs over rRNAs was independently confirmed by using a Bioanalyzer Chip (Figure S1B). DNA does not copurify because no PCR amplification occurred when samples were RNase treated before the oligo(dT) pull-down (Figure 1B) or when reverse transcriptase (RT) was omitted from the RT reaction during complementary DNA (cDNA) preparation (data not shown). RNA isolated by oligo(dT) purification was also analyzed by next-generation sequencing. As expected, mRNA was the predominant RNA population, followed by a residual pool of rRNA and mitochondrial rRNA (Figure 1C). Other RNAs were of low abundance or were not detected.

We next analyzed the proteins isolated by interactome capture. Gel electrophoresis, combined with silver staining, reveals complex protein patterns from either of the two UV-crosslinking methods, whereas control reactions from nonirradiated cells or mock pull-downs with control beads lacking oligo(dT) were remarkably clean (Figures 1D and S1C). Both the cCL and PAR-CL protocols recover similar RBP patterns with some notable differences. Importantly, the patterns of isolated proteins differ profoundly from the whole HeLa proteome, indicating that interactome capture can successfully select against abundant cellular proteins. Used as a positive control, the polypyrimidine tract-binding protein 1 (PTBP1), a well-known RBP, was strongly enriched by both cCL and PAR-CL and was undetectable in the negative control samples (Figure 1E). Likewise, cytosine-uracil-guanine (CUG) triplet repeat RNA-binding protein 1 (CELF1) was isolated by both crosslinking methods; however, cCL captured this protein more effectively, exemplifying an RBP that is favored by one crosslinking chemistry compared to the other. Most importantly, negative controls for the abundant α-tubulin, β-actin, and DNA-binding histones H3 and H4 confirm the high selectivity and specificity of the protocol.

Proteomic Determination of the HeLa mRNA Interactome

Following release by RNase treatment, proteins were cleaved into peptides with trypsin. To maximize protein identification, sample complexity was reduced by peptide fractionation using isoelectric focusing. The resulting fractions were analyzed by high-resolution nano-LC-MS/MS. By combining the data from cCL and PAR-CL, we identified 1,651 proteins in the UV-crosslinked samples, whereas only 434 proteins were identified in controls, including 335 proteins that were also found in the collective set of proteins identified in the UV-crosslinking samples (Figure 2A and Table S1). Therefore, 1,316 proteins were exclusively identified by interactome capture. The overlap of proteins identified by cCL and PAR-CL was approximately two-thirds (64%) (Figure 2B). Although both UV-crosslinking protocols yield comparably high numbers of proteins, 24% of the identified proteins were found exclusively in cCL samples, compared with 12% for PAR-CL; these data correlate well with the protein patterns shown in Figures 1D and S1C. From a total of 4,797 proteins detected in the HeLa whole-cell lysate (Table S1), 1,361 were also present in the crosslinked samples after oligo(dT) pull-down (Table S1), whereas 290 were exclusively found in samples after interactome capture (Figure S2A).

To apply statistical data analysis, protein enrichment in crosslinked samples over controls was assessed by two label-free quantification methods that use different information available from tandem mass spectrometry. The spectral count method estimates differential protein abundance by comparing the number of peptide identifications for each protein. Taking the natural variation between biological replicates into account, the bioconductor package DESeq (Anders and Huber, 2010) provides a statistical test for assessment of differential abundance of count data.

The ion count method was applied as a second quantification approach. Ion chromatograms for each peptide were extracted and used to quantify the relative amount of peptide ions between one crosslinking and one negative control experiment. Taking biological variance into account, a moderated t test implemented in the software limma (Smyth, 2004) was used to detect enriched proteins.

We determined significant enrichment of spectral counts and ion counts for a large number of proteins (Figures 2C, 2D, S2B, and S2C). In addition, biological replicates that were analyzed by both statistical methods showed a strong correlation, even for the comparison between PAR-CL and cCL (Figures 2D, S2B, and S2C). The number of significantly enriched proteins

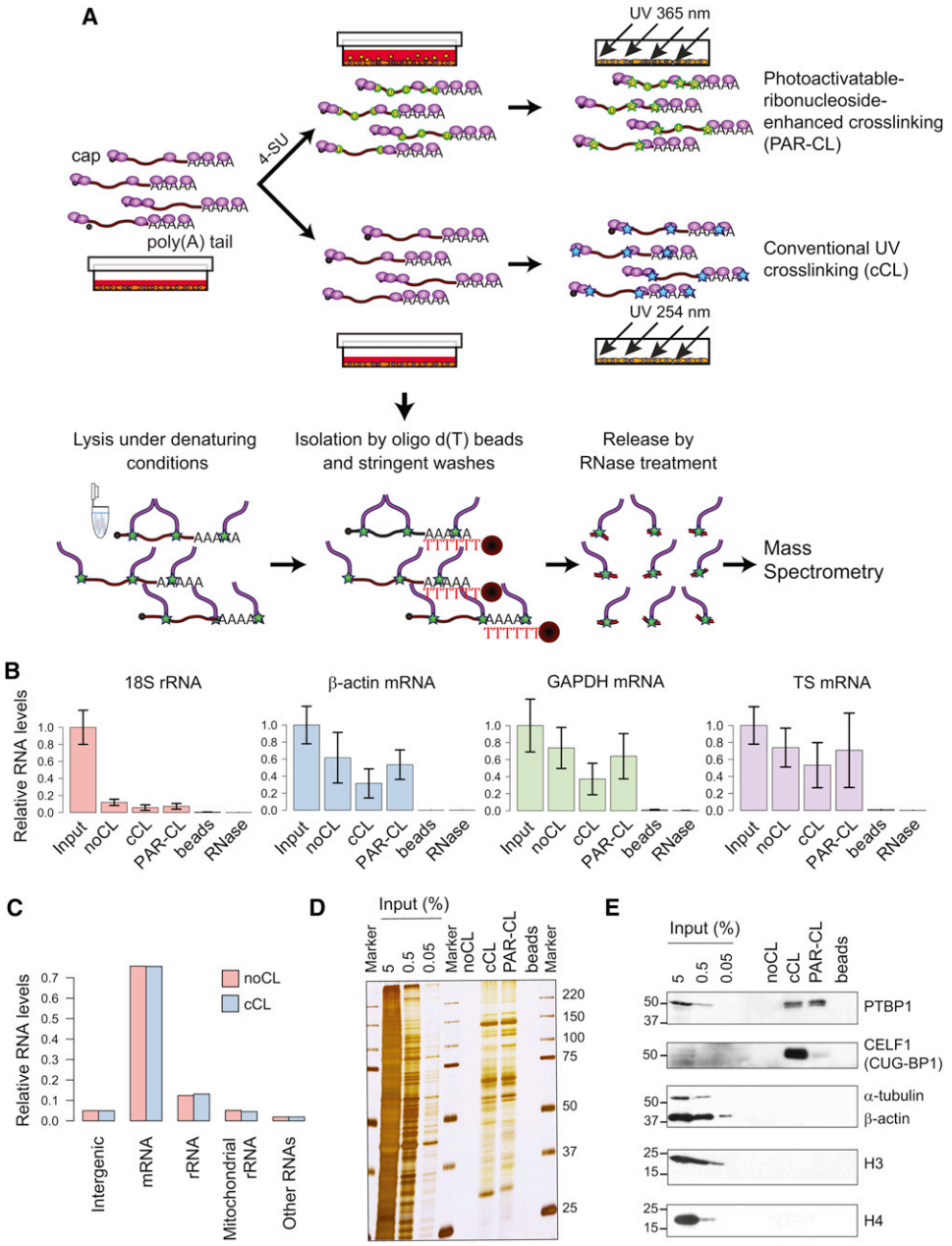

Figure 1. In Vivo Capture of HeLa RBPs

(A) mRNA-protein interactions are preserved by employing either UV cCL or PAR-CL protocols on proliferating HeLa cells. mRNA-protein complexes are isolated by pull-down with oligo(dT) magnetic beads and stringently washed, and then bound proteins are eluted with RNase and identified by MS.

(B) After applying either cCL or PAR-CL, poly(A)$^+$ RNAs were selected as in (A). As controls, beads lacking oligo(dT) (beads), RNase T1- and A-treated lysates (RNase), or nonirradiated cells (noCL) were used. Levels of 18S rRNA, β-actin, GAPDH, and TS mRNAs in samples were monitored by RT-qPCR. SDs were calculated from four biological replicates.

(C) RNAs isolated following cCL or noCL protocols were analyzed by sequencing, and the relative amounts of different RNAs are plotted.

(D–E) Samples were digested with RNases, and released proteins were analyzed by silver staining (D) and western blotting against PTBP1, CUG-BP, α-tubulin, β-actin, and histones (H)3 and H4 (E).

See also Figure S1.

was 493 from the spectral count method and 797 from the more sensitive ion count method (false discovery rate 0.01 in both cases) (Figure 2E). Combining the two analyses, 860 proteins were enriched after UV crosslinking by at least one of the two quantification methods. Because these 860 proteins qualify as RBPs according to stringent biochemical and statistical criteria,

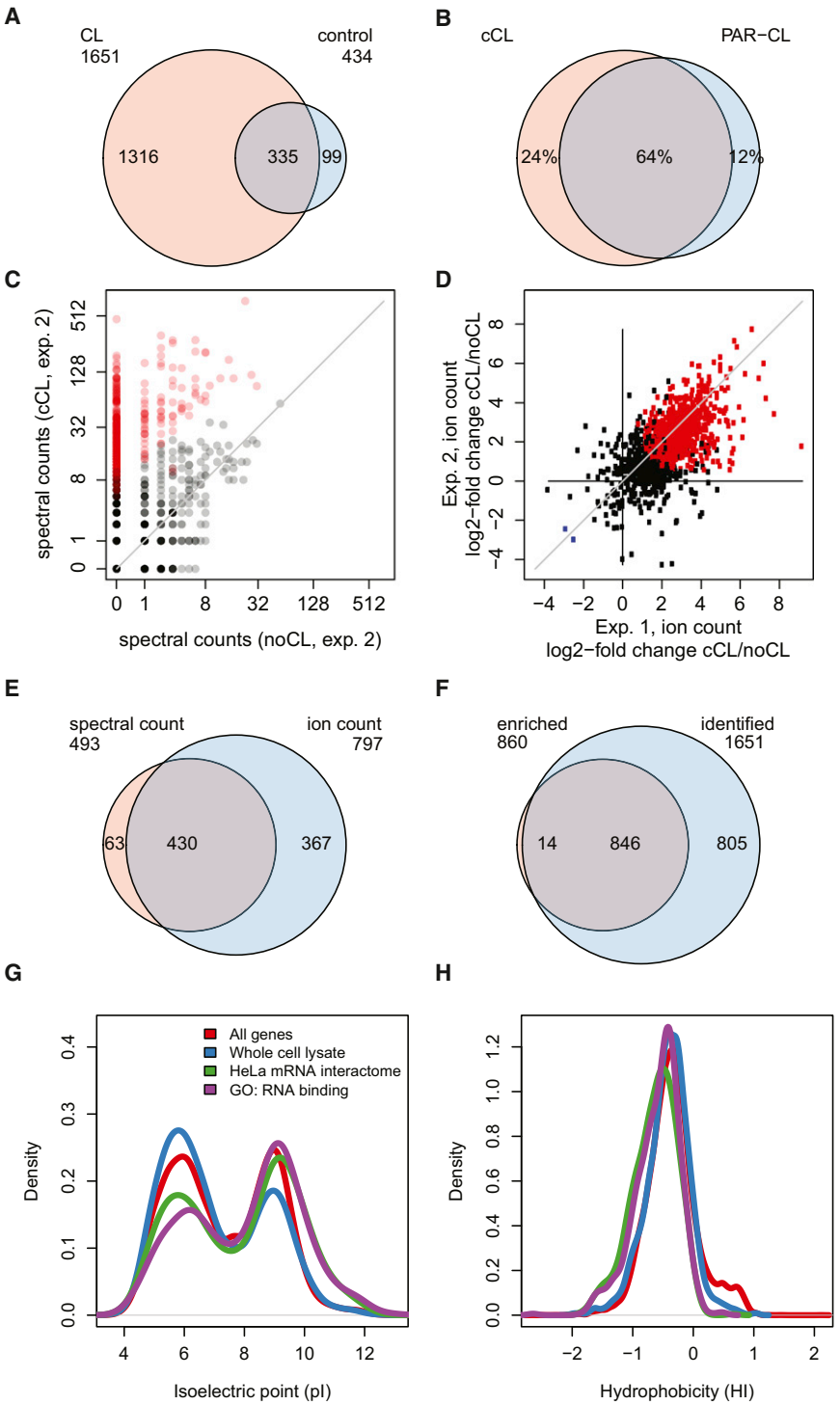

Figure 2. Proteomic Analysis of HeLa mRBPs

Poly(A)$^+$ RNAs were isolated as in Figure 1A. Three experimental (two cCL and one PAR-CL) and three control (two noCL and one 4SU noCL) biological replicates were processed by MS.

(A) Venn diagram comparing the number of proteins identified in the three crosslinking (CL) experiments (pink) or in their respective controls (blue).

(B) Percentage of proteins identified in one cCL (pink) or in one PAR-CL experiment (blue).

(C) Scatter plot of spectral counts comparing a cCL experiment to a control experiment. Each dot represents one protein. Axes depict the number of unique peptide identifications. Proteins in red are significantly enriched according to the DESeq method.

(D) Scatter plot comparing the differential ion counts of two biological replicates. Axes show the log2-fold change in ion counts between cCL and noCL. Proteins significantly enriched according to the ion-count method in cCL or control experiments are depicted by red or blue dots, respectively.

(E and F) Venn diagrams comparing the number of proteins (E) quantitatively enriched by the spectral count method (pink) or by the ion-count method (blue) or (F) quantitatively enriched in crosslinking experiments compared to controls (HeLa mRNA interactome, pink); total number of proteins identified in crosslinking experiments (blue).

(G) Density of the calculated isoelectric points (pI) of all human proteins (red), HeLa whole-cell lysate (blue), HeLa mRNA interactome (green), and proteins annotated as RNA binding (purple).

(H) Density of hydrophobicity for the same protein groups as in (G).

See also Figures S2 and S3 and Tables S1 and S3.

we refer to them as the "HeLa mRNA interactome" (Figure 2F). Note that 14 of these proteins are listed as "enriched," but not as "identified," because their corresponding peptides were not identified in crosslinked samples with false discovery rate (FDR) <0.01; however, they reached this identification threshold by taking into consideration data from the control experiments. Although the quantitative value extracted for these peptides is significantly larger in crosslinked samples than in controls (FDR < 0.01), qualifying their inclusion in the mRNA interactome, these 14 proteins should nonetheless be considered against this background and are indicated in Table S1 (red font).

Earlier analyses of complex proteomes, for example from *C. elegans* or *D. melanogaster*, noticed a technical bias of MS regarding protein abundance, isoelectric point (pI), hydrophobicity, and protein size (Brunner et al., 2007; Schrimpf et al., 2009). Compared to proteins predicted from the human genome, basic, hydrophobic, and low abundance proteins are underrepresented in the HeLa whole-cell lysate (Figures 2G, 2H, and S2D; red versus blue line); however, protein size did not substantially affect protein identification (Figure S2E). In contrast to the whole-cell lysate,

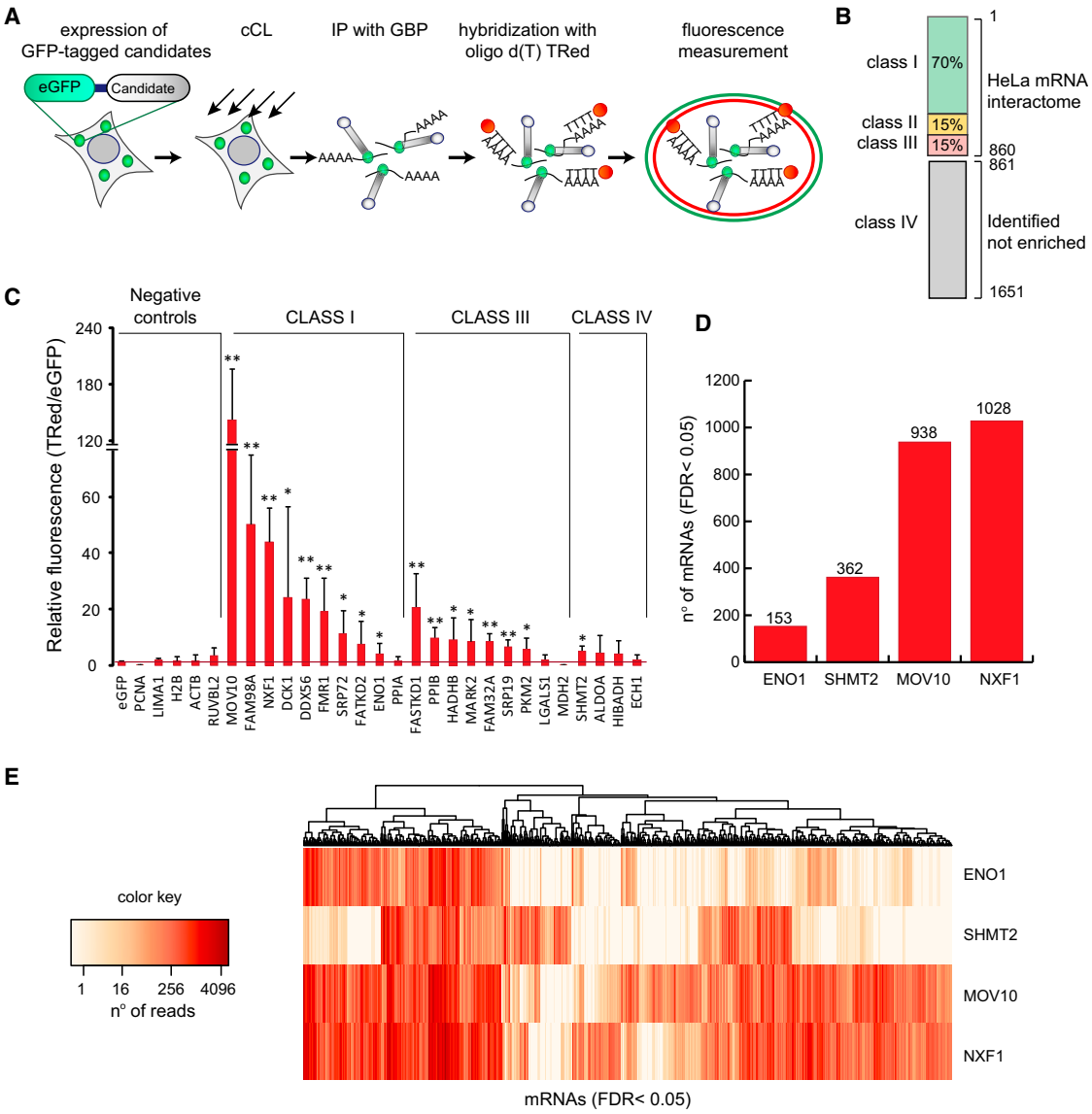

Figure 3. Experimental Validation of the HeLa mRNA Interactome

(A) Scheme of the dual fluorescence validation method.

(B) Classification of the identified proteins.

(C) Relative TRed/EGFP fluorescence ratios from controls and candidate EGFP/YFP-tagged proteins after normalization to the ratio of unfused EGFP. Error bars represent SDs from nine independent IPs (three biological replicates). *$p < 0.05$ and **$p < 0.01$ after t test.

(D and E) RNAs crosslinked to and coimmunoprecipitated with YFP/EGFP-tagged MOV10, NXF1, ENO1, and SHMT2 were analyzed by sequencing. (D) Number of genes significantly enriched ($p < 0.05$) over control samples (RNAs coimmunoprecipitated with EGFP). (E) Heat map showing the mRNAs bound to each protein ($p < 0.05$).

See also Figure S3 and Tables S2 and S7.

basic proteins were more prevalent in the HeLa mRNA interactome than acidic ones, as seen for proteins annotated by the gene ontology (GO) term "RNA-binding" (Figure 2G, green versus purple line). Moreover, these latter protein sets showed similar densities for hydrophobicity and mRNA abundance (Figures 2H and S2D). Therefore, the HeLa mRNA interactome displays the expected chemical and biological features.

Experimental Validation of the HeLa mRNA Interactome

For validation, we developed a fluorescence-based quantitative method to monitor mRNA-protein interactions. We generated "Tet-on" HeLa cell lines stably expressing enhanced green fluorescent protein (EGFP)/yellow fluorescent protein (YFP)-tagged proteins (23 candidates and 6 negative controls) (Figure 3A). Following Tet induction and UV crosslinking, EGFP/YFP chimeric proteins were immunoprecipitated (IP) with a high

affinity and specificity single-chain antibody from *Lama paca* (Rothbauer et al., 2008). Immunoprecipitates were stringently washed, and the presence of crosslinked polyadenylated RNAs was revealed by hybridization of Texas red (TRed)-labeled oligo(dT). Thus, the fluorescence ratio of TRed (RNA) over EGFP (protein expression) serves as a quantitative measure of poly(A) RNA binding.

All 1,651 identified proteins were ranked according to their spectral and ion count scores. For the 860 proteins of the interactome, the top 70%, next 15%, and bottom 15% were assigned to classes I–III, respectively (Figure 3B). The remaining identified proteins were considered as class IV. Candidate RBPs from classes I, III, and IV were selected, including underrepresented categories such as kinases and intermediary metabolism enzymes (see below).

All negative controls, including three DNA-binding proteins (RUVBL1, PCNA, and H2B), showed TRed/EGFP ratios close to background (unfused EGFP) (Figure 3C). Conversely, nine out of ten class I candidates display significantly higher relative fluorescence values (Figure 3C). Seven out of nine proteins from class III and one out of four from class IV were also validated by this assay. Notably, the number of validated proteins in each class correlates well with the MS quantification data. Some of the nonvalidated candidates could represent false negatives because the EGFP/YFP tag may interfere with RNA binding of some RBPs.

For an independent test of RNA binding and to obtain insights into the spectrum of bound RNAs, we identified RNAs crosslinked to GFP/YFP-fused MOV10, NXF1, ENO1, SHMT2, or EGFP alone following GFP/YFP immunoprecipitation by next-generation sequencing. After cDNA library preparation, primer ligation, and amplification, equal amounts of DNA were subjected to Sequencing by Oligonucleotide Ligation and Detection (SOLiD); this normalization procedure overestimates RNA binding by the negative control EGFP because a far greater number of cell equivalents were used. As shown in Figures 3D and S3A and Table S2, a large number of mRNAs are significantly enriched in immunoprecipitations of RBP candidates compared to the EGFP control. Nevertheless, a small set of highly abundant HeLa mRNAs was prevalent in EGFP samples (Figure S3B); these contaminants likely passed the detection threshold due to the overrepresentation of this sample in the sequencing runs. MOV10 and NXF1 display broad RNA binding, whereas the enzymes ENO1 and SHMT2 bind specific and distinct subsets of RNAs (Figure 3E). Evidently, even the class IV candidate SHMT2 is validated by both assays, confirming that this class harbors additional bona fide RBPs.

Technical Aspects of the Interactome Capture Protocol

To differentiate bona fide RBPs from nonspecific binders, we applied stringent biochemical and statistical criteria. This choice minimizes false positives but comes at the price of false negatives. Their number is difficult to estimate, especially because we presently do not know how many of the class IV proteins represent physiological RBPs. For example, IRP1 (ACO1), the regulatory RBP of cellular iron homeostasis, failed to be identified in the crosslinked samples, although it is detected in the HeLa whole-cell lysate (Table S1). Such a false-negative result could originate from the lack of IRP1 binding to its target mRNAs due to an iron-replete state of the cultured cells (Hentze et al., 2010) and/or from inefficient crosslinking when bound to its targets. Generalizing this limitation, our approach will fail to detect physiological RBPs when: (1) they are not expressed in HeLa cells, (2) they do not bind polyadenylated RNAs, (3) their RNA-binding activity is inhibited in proliferating HeLa cells, or (4) bound RBPs fail to be crosslinked by both cCL and PAR-CL. However, most of the RRM-containing proteins (136/151, see below), all of the hnRNPs (18/18), and almost all of the RNA helicases (19/23) detected in the HeLa whole-cell lysate are also found in the HeLa mRNA interactome, suggesting that it represents a reasonably comprehensive atlas of the HeLa cell mRNA-binding proteins.

In theory, both UV-crosslinking protocols should select for proteins that directly bind to RNA and discriminate against those that associate indirectly as subunits of larger RNA-binding complexes without directly contacting the RNA because the UV-crosslinking protocols do not mediate protein-protein crosslinking (Greenberg, 1979; Pashev et al., 1991; Suchanek et al., 2005) and because the purification conditions (0.5 M LiCl; 0.5% LiDS) will dissociate most noncovalent protein-protein interactions. The core exon junction complex represents a high-affinity heterotetramer composed of eIF4AIII (EIF4A3), Y14 (RBM8A), MAGOH, and Barentz (CASC3, BTZ), whose cocrystal structure with RNA is known (Figure 4A) (Bono et al., 2006). Consistent with the structural information and supporting the selectivity of interactome capture, we find eIF4AIII and CASC3 to be components of the mRNA interactome, whereas Y14 and MAGOH are absent (Table S1). Although we consider the mRNA interactome as being validated as a complex data set, each individual member of it should be considered as a high-probability RBP, recommended for individual validation by researchers planning to explore these proteins' functions in RNA biology in greater depth.

cCL versus PAR-CL

The two-pronged approach with cCL and PAR-CL offers advantages over the use of a single method because the majority of RBPs of the interactome are independently confirmed by a second protocol. Whereas most of the proteins are similarly captured by the two techniques, for a few dozen proteins, cCL or PAR-CL showed superior performance compared to the other, providing technically useful information (Figures S3C and S3D and Table S3). For example, CELF1 is more efficiently captured by cCL than PAR-CL, in agreement with Figure 1E (Figure S3C); the converse applies to the Y-box-binding protein 1 (YBX1).

PAR-CL has recently been popularized as being more efficient than cCL in protein-RNA crosslinking (Hafner et al., 2010). About 12% of the interactome was captured solely by PAR-CL (Figure 2B), but twice as many RBPs (24% of the interactome) were identified only by cCL and could have been missed if PAR-CL had been used alone.

Known and Previously Unknown RBPs

To benchmark the HeLa mRNA interactome against known RBPs, we carried out a gene set enrichment analysis assessing functional and structural properties using gene ontology. As

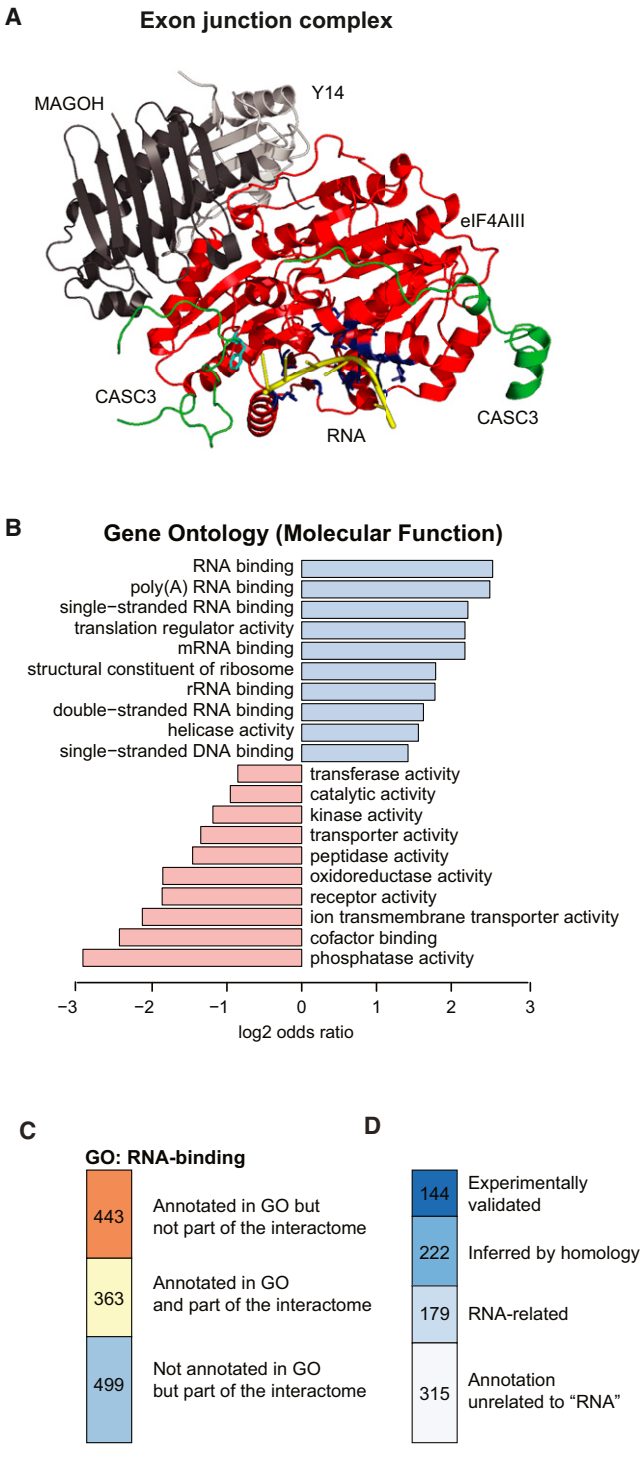

Figure 4. Analysis of the HeLa mRNA Interactome

(A) Ribbon diagram of the crystal structure of the core exon junction complex consisting of eIF4AIII, Magoh, Y14 (residues 66–174), and CASC3 (residues 137–286), associated with U_{15} RNA at 2.2 Å resolution (PDB 2J0Q) (Bono et al., 2006). EIF4A3 (eIF4AIII, red) and CASC3 (green) contact the RNA directly (yellow), which is in contrast to Y14 (light gray) and Magoh (dark gray). Amino acids from eIF4AIII and CASC3 in contact with the RNA are shown in dark blue and cyan, respectively.

(B) Ten of the most significant over- (blue) and underrepresented (pink) molecular function GO terms of the mRNA interactome.

(C) Comparison of the mRNA interactome with the GO term "RNA binding."

(D) Number of experimentally validated RBPs, RBPs inferred by homology, RBPs with the GO annotation "RNA related," or RBPs without RNA-related annotation in the mRNA interactome.

See also Figure S4 and Table S4.

predicted, RNA-binding annotations far exceed DNA binding in the HeLa mRNA interactome (Figure S4A), with RNA binding itself being the most enriched GO term, followed by more defined RNA-binding activities such as mRNA binding (Figure 4B and Table S4). In addition, other RNA biology-related functions and processes are highly represented, e.g., protein synthesis and RNA metabolism (Figures S4B–S4E). Kinases, phosphatases, receptors, transporters, proteins involved in mitosis, DNA synthesis, and intermediary metabolism are statistically underrepresented (Figures 4B and S4B–S4E and Table S4); some RBPs from these underrepresented categories will be discussed in greater detail below.

To estimate the number of "previously unknown" RBPs, we assembled a catalog of experimentally validated RBPs and compared it with the HeLa mRNA interactome and the GO annotation "RNA binding" in ENSEMBL (Figures 4C and 4D). Because some well-known RBPs are not annotated as RNA binding in public databases, we further removed proteins with GO annotations related to RNA (e.g., RNA metabolism). Even after this stringent counterselection, the HeLa mRNA interactome adds 315 high-probability RBPs to those identified in the past decades (Figure 4D). In addition, the HeLa mRNA interactome provides direct experimental support for RNA binding of a large number of proteins (222) that, in spite of being annotated in GO as RNA binders, had only been inferred to represent RBPs by homology.

Insights into Modes of RNA Binding

Globular Domains

About half of the mRNA interactome proteins harbor known RBDs, and as a consequence, several classical (e.g., RRM, KH, and DEAD box helicase) and nonclassical (e.g., LSM and YTH) RBDs are statistically overrepresented (Figures 5A–5C). Dual-specificity domain families with sparse evidence for RNA binding were also present in our data set (Figure 5C). For example, the SAF-A/B, Acinus, and PIAS (SAP) domain (Figures S5A and S5B) is commonly associated with DNA binding; however, in the exonuclease ERI1, it has been shown to interact with the 3′ end stem loop of histone mRNA (Yang et al., 2006) (Y. Cheng and D.J. Patel, personal communication; PDB 1ZBH) (Figure S5C). Our data support a broader role of SAP domains in RNA binding because most of the SAP-domain-containing proteins identified in the HeLa whole-cell lysate are also found in the HeLa mRNA interactome (12/14), and eight of these do not harbor a canonical RBD.

Another example is tryptophan-aspartic acid 40 (WD40), which consists of repeats of a 31–60 residue-conserved motif (WD40 motif) that forms β-propeller structures known as WD domains (Figure S5D). This protein architecture generates an excellent platform for the evolution of diverse binding specificities (mostly protein binding), and the domain family has

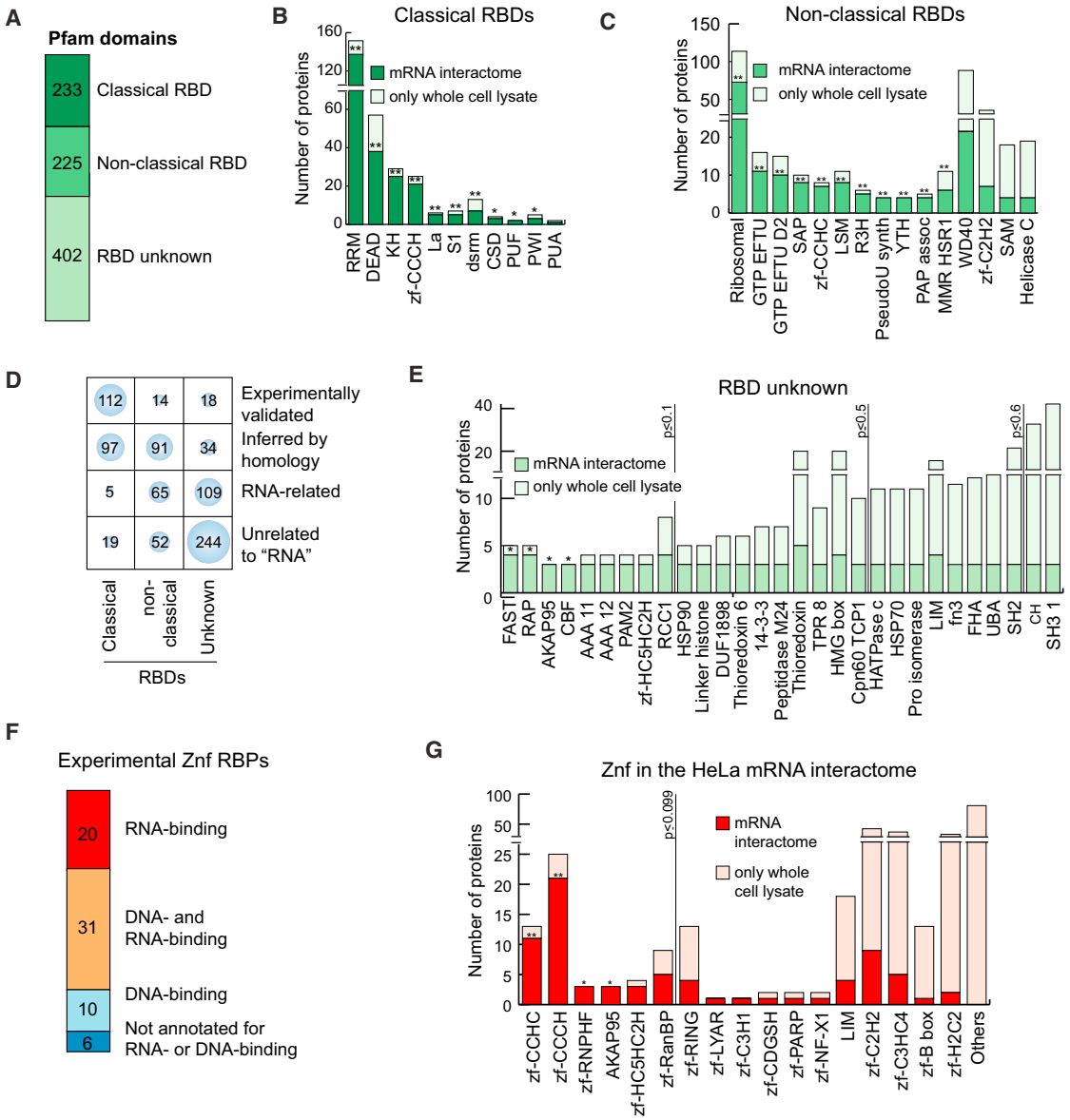

Figure 5. Globular Domains in HeLa mRNA Interactome Proteins

(A) Number of proteins harboring classical, nonclassical, or unknown RBDs in the mRNA interactome. For the purpose of this figure, we considered the RBDs listed in Lunde et al., (2007) as classical and protein domains that have been experimentally shown to bind RNA in at least one example as nonclassical.

(B) Number of proteins annotated with each classical domain in the mRNA interactome (dark) or only identified in HeLa whole-cell lysate (light).

(C) Number of proteins annotated with each nonclassical domain. Only domains with four hits or more are shown. Proteins containing both classical and nonclassical RBDs are listed in (B).

(D) Balloon plot cross-referencing functional (GO) and structural (Pfam domains) annotations of the proteins in the HeLa mRNA interactome.

(E) Distribution of Pfam domains in the proteins of the HeLa mRNA interactome without known RBD. Only Pfam domains with at least three hits are shown.

(F) Comparison of different Znf proteins of the mRNA interactome with the GO terms RNA binding and DNA binding.

(G) Occurrence of Znf motifs within the HeLa mRNA interactome (red) and HeLa whole-cell lysate (pink).

See also Figure S5 and Table S5.

expanded significantly in higher eukaryotes (Stirnimann et al., 2010). Interestingly, the WD domain of Gemin5 displays RNA-binding activity (Lau et al., 2009), suggesting that WD domains can interact, at least in some instances, with RNA. In agreement, 23 WD domain-containing proteins are found to be associated with poly(A)$^+$ RNAs in HeLa cells, none of which harbor classical RBDs. The physicochemical properties of these putative mRNA-binding WD domains differ from WD domains of proteins that are not present in the HeLa mRNA interactome, being enriched for most of the amino acids typically found at protein-RNA

interfaces (especially basic amino acids) (Lunde et al., 2007) (Figure S5E). Homology modeling of the WD domain of UTP15 revealed clusters of basic amino acids at the surface of the β-propeller that may serve as a platform for docking RNA (Figure S5F).

Orphan proteins without known RNA-binding motifs constitute half of the HeLa mRNA interactome, and most of these also lack RNA-binding or RNA-related GO annotations (Figures 5A and 5D). We searched for domains or motifs that are enriched among these proteins, which could represent RBDs. Two domains co-occur in all members of the poorly characterized *f*as-*a*ctivated *s*erine/*t*hreonine (FAST) kinase family: the FAST kinase domain and the RNA-binding domain abundant in Apicomplexans (RAP) (Figures 5E and S5G). The RAP domain is a putative RBD without supporting experimental evidence (Lee and Hong, 2004). Homology modeling of the RAP domain revealed an endonuclease-like fold that generates an interface rich in basic and aromatic residues that might be involved in RNA binding (Figures S5H and S5I). We identified all (six) human RAP-domain containing proteins in crosslinked samples, including four in the HeLa interactome (Table S5); two of these (FASTKD2 and FASTKD1) were validated independently as RBPs (Figure 3C). Therefore, FAST kinases represent a family of directly RNA-binding kinases.

Znf are classical nucleotide-binding domains that are subclassified by the order of the zinc-contacting amino acids (Lunde et al., 2007). We found 69 Znf-containing proteins within the mRNA interactome, many of which were previously uncharacterized as possessing RNA-binding activity (Figure 5F). CCCH, CCHC, and RNPHF Znf motifs are well known to bind RNA and, expectedly, are enriched in the mRNA interactome. AKAP95 and HC5HC2H Znf subtypes, previously thought to bind exclusively DNA, are also overrepresented in our data set (Figure 5G), suggesting that they also represent bona fide RBDs. The remaining Znf domain classes occurred more sporadically.

Seven peptidyl-prolyl *cis-trans* isomerases (PPI) are found in the interactome (Table S5). PPIs play regulatory roles in spliceosome and ribonucleoprotein dynamics by interconverting *cis* and *trans* conformations of proline isomers (Mesa et al., 2008). PPIE and PPIL4 contain one RRM (grouped with the proteins harboring classical RBDs in Figures 5A and 5B) (Mi et al., 1996; Zeng et al., 2001). However, five additional PPIs lacking known RBDs are also present within the mRNA interactome (Figure 5E and Table S5), and we validated PPIB as an RBP (Figure 3C). PPIG and PPIA contribute to ribonucleoprotein dynamics (Mesa et al., 2008; Pan et al., 2008), the latter being essential for hepatitis C virus (HCV) replication (Foster et al., 2011). The presence of several PPIs in the mRNA interactome suggests that this protein family plays broader roles in RNA biology than previously anticipated.

Repetitive Disordered Motifs

Large portions of the human proteome are intrinsically disordered, natively lacking stable three-dimensional structure. Disordered regions are frequently endowed with high functional density containing multiple interaction interfaces and may be involved in regulatory functions, including facilitation of RNA folding as RNA chaperones (Dyson and Wright, 2005; Tompa and Csermely, 2004). Proteins within the mRNA interactome are highly enriched in intrinsically disordered regions compared to the human proteome or HeLa whole-cell lysate ($p = 2.2 \times 10^{-16}$) (Figure 6A). However, the physicochemical properties of these unstructured segments of RBPs differ from comparable regions of whole-cell lysate, with a prevalence of glycine (G), arginine (R), and lysine (K) residues (Figure 6B). Unexpectedly, tyrosine (Y) is also enriched in these segments (especially in proteins containing classical RBDs), although the ''order-promoting'' aromatic residues are depleted in disordered regions of the human proteome (Figure S6A) (Radivojac et al., 2007). Amino acids that are enriched in the unstructured regions of the mRNA interactome are also commonly found in globular RBDs (Lunde et al., 2007); conversely, acidic amino acids, which are usually of low abundance in those interfaces, are underrepresented (Figure S6A). Another striking property of disordered segments in RBPs is that low complexity and repetitive amino acid sequences are overrepresented compared to similar regions within the human proteome or HeLa whole-cell lysate ($p = 2.2 \times 10^{-16}$) (Figures 6C and 6D). These features apply to RBPs lacking known RBDs and RNA-related GO annotations (disorder, $p = 3.7 \times 10^{-6}$; complexity, $p = 4.25 \times 10^{-5}$; repetitive sequences: $p = 2.6 \times 10^{-9}$) (Figures 6A–6D).

Several repetitive sequences in unstructured regions of RBPs form recognizable patterns shared between evolutionarily unrelated proteins of the mRNA interactome (Figures 6E, S6B, and S6C). Arginine co-occurs preferentially with serine (S) (Figure 6E), reflecting the regulatory importance of arginine-serine (RS) dipeptides, particularly in the serine-arginine (SR) protein family (Twyffels et al., 2011). Arginine also combines with glycine, forming the arginine-glycine-glycine (RGG) box RNA-binding motif (Figure 6E), which binds a guanine-rich *sc1* RNA sequence in fragile X mental retardation protein 1 (FMR1) with nanomolar affinity (Phan et al., 2011). The FMR1 segment $R_{533}GGGGR_{538}$ recognizes *sc1* RNA by shape complementarity and intermolecular hydrogen-bonding interactions with the Watson-Crick bases G31 and G7 (Phan et al., 2011). RGG boxes vary in the length and number of repeated units, and they are often found in mRNA interactome proteins in combination with classical or nonclassical RBDs or other repetitive motifs as well as in proteins lacking known RNA-binding architectures (Figures 6F, 6G, and S6D). This suggests that RGG boxes are broadly used platforms for RNA binding, which could contribute cooperatively to the modular design of RBPs by increasing the affinity and the specificity of the protein-RNA interaction. In some instances, glycine also combines with tyrosine, forming tyrosine-glycine-glycine (YGG) boxes (Figure 6E). The function of this motif is unknown; nevertheless, we find it frequently in combination with RBDs or RGG boxes (Figures 6F, 6G, and S6D). YGG boxes could employ a similar mechanism of RNA binding as RGG boxes by using the tyrosine side chain to interact with RNA bases by stacking or hydrogen bonding.

Basic disordered tails are often used by transcription factors to bind DNA (Vuzman and Levy, 2012). In this regard, lysine-rich segments are also found in mRNA interactome proteins, and they are especially abundant among the previously unknown RBPs (Figures 6E–6G). In some cases, poly(K) motifs coincide with experimentally validated nuclear localization signals (NLS); however, they are frequently longer than the classical NLS

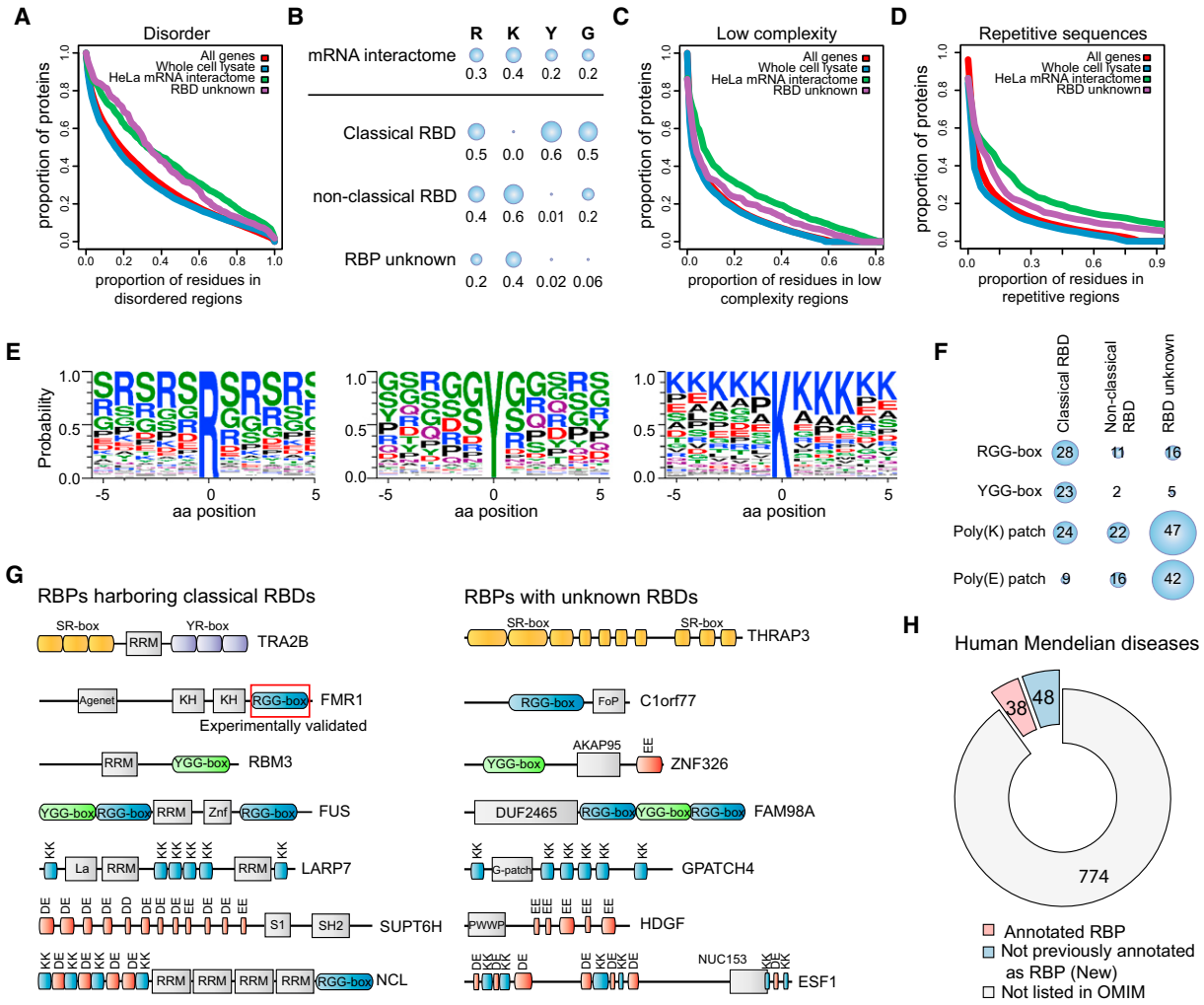

Figure 6. Repetitive Motifs in HeLa mRNA Interactome Proteins

(A) Distribution of calculated disorder regions of all human proteins (red), HeLa whole-cell lysate (blue), mRNA interactome (green), and proteins lacking known RBDs (purple).

(B) Enrichment of amino acids in disordered regions of the mRNA interactome.

(C and D) Distribution of calculated low-complexity regions (C) and repetitive dipeptide sequences (D) for the same protein groups as in (A).

(E) Sequence logos of amino acids around repetitive residues. A position weight matrix is computed from all 11-mer sequences around all residues in repetitive regions. Sequence logos are shown for the central amino acids R, Y, or K. The height of the letters is proportional to the probability of amino acid occurrence at each position.

(F) Occurrence of disordered repetitive motifs in mRNA interactome proteins.

(G) Schematic representation of repetitive motif distribution in proteins containing classical RBDs or lacking known RBDs.

(H) Number of proteins of the mRNA interactome listed in the OMIM database: proteins annotated in GO as RNA-binding (red), proteins not annotated as RNA binding (blue).

See also Figure S6 and Table S6.

definition and form patches with nonrandom distribution (Figure 6G). Hypothetically, poly(K) patches could establish electrostatic interactions with the phosphate backbone of RNA in analogy with the basic tails in DNA-binding proteins (DBPs) (Vuzman and Levy, 2012). Length and net charge of basic tails in homeodomain transcription factors influence their DNA-binding properties (Vuzman et al., 2010; Vuzman and Levy, 2010). Poly(K) patches in RBPs could follow similar principles for binding affinity and specificity. Alternatively, poly(K) tracts could be involved in interactions with acidic protein patches, which we also observe in HeLa RBPs (Figures 6F, 6G, S6C, and S6D), as occurs with K-rich histone tails (McBryant et al., 2010).

The presence of repetitive motifs within disordered regions and their conservation in nonhomologous RBPs point toward an emerging role of such intrinsically disordered domains in RNA biology.

Insights into Mendelian Disease

Eighty-six proteins of the mRNA interactome are listed in the Online Mendelian Inheritance in Man (OMIM) database as being

associated with human Mendelian disease (ENSEMBL 63). Most of these were previously unknown to be RBPs (Figure 6H and Table S6). Disturbances of RNA metabolism can now be explored for these 48 proteins to further understand their roles in the respective human disorders. In some cases, the same syndromes are caused by alterations of both known and previously unknown RBPs (Table S6). For instance, non-insulin-dependent diabetes mellitus can be caused by mutations in the well-known RBP IGF2BP2 and also by mutations in the interactome protein PTPN1 (also called PTP1B). PTPN1 is a phosphatase, one of the most underrepresented functions in the mRNA interactome (Figure 4B); it has also been implicated in cancer (Lessard et al., 2010).

Similarly, a FASTKD2 mutation generating a premature stop codon was identified in patients with infantile mitochondrial encephalomyopathy associated with cytochrome *c* oxidase deficiency (mitochondrial complex IV deficiency in OMIM), an infrequent developmental disease with severe symptoms (Ghezzi et al., 2008). This mutation generates a truncated protein lacking part of the FAST kinase and the whole RAP domain with decreased susceptibility to apoptotic stimuli (Figure S5G). Thus, the role of FASTKD2 as an RBP (validated in Figure 3C) in apoptosis and infantile mitochondrial encephalomyopathy associated with cytochrome *c* oxidase deficiency calls for further exploration.

"Moonlighting" Enzymes and REM Networks

Cytosolic aconitase is an enzyme that plays a key physiological role as an iron-regulated mRNA-binding protein (iron regulatory protein 1/IRP1) (Hentze and Argos, 1991; Rouault et al., 1991). Other enzymes of intermediary metabolism have been implicated in "moonlighting" as RNA-binding proteins, although the evidence supporting RNA binding in vivo is limited (Cieśla, 2006; Hentze, 1994). Using the "reactome" annotation (Joshi-Tope et al., 2005), the HeLa mRNA interactome harbors 17 enzymes of intermediary metabolism, and the extended class IV list increases this count to 46 (Table 1). In part, this list confirms earlier experiments (Cieśla, 2006; Elzinga et al., 1993, 2000; Kiri and Goldspink, 2002; Liu et al., 2001; Nagy and Rigby, 1995; Nakagawa et al., 1995; Pioli et al., 2002; Shetty et al., 2004), and it also identifies metabolic enzymes not previously known as RBPs. We validated four of these as RBPs by the dual fluorescence assay (Figure 3C); ENO1 and SHMT2 were also validated by sequencing of associated RNAs (Figures 3D and 3E).

The HeLa cell RNA-binding enzymes cover much of the landscape of intermediary metabolism, including carbohydrate, amino acid, lipid, and nucleotide metabolism, and they appear not to cluster into particular pathways. If functionally relevant, as proposed by the REM (*R*NA, *e*nzyme, and *m*etabolite) network hypothesis (Hentze and Preiss, 2010), these proteins could broadly connect intermediary metabolism with RNA biology and posttranscriptional gene regulation.

Oxidoreductase, transferase, and kinase are prevalent catalytic activities among these enzymes. Six of the RNA-binding enzymes in the mRNA interactome and, additionally, 12 in the identification set use NAD^+, $NADP^+$, NADH, NADPH, FAD, or $FADH_2$ as cofactors via the dinucleotide-binding (Rossmann) fold. The Rossmann fold constitutes an RBD for GAPDH and LDH (Nagy and Rigby, 1995; Pioli et al., 2002), but Rossmann-fold-containing proteins are underrepresented in the HeLa interactome overall (Figure S4E). Therefore, this domain does not appear to suffice for RNA binding unless the (metabolic) state of proliferating HeLa cells is incompatible with RNA binding by the other Rossmann-fold-containing proteins.

Finally, five of the metabolic enzymes in the interactome and an additional five in the identification data set share their ability to simultaneously bind ATP and an anionic substrate such as succinate, L-aspartate, or pyruvate. The role of this property for RNA binding also deserves further exploration.

Outlook

The mRNA interactome capture methodology was developed here to generate a comprehensive atlas of mRNA (strictly: poly(A) RNA)-binding proteins of a living cell. In spite of their limitations, we chose HeLa cells for their economy and ease of handling as well as the wealth of available tools and information. We believe that this work offers an informative snapshot of RNA biology. Interactome capture can now be adapted to study the mRNA interactomes of other cells and organisms. The approach can also be applied to investigate changes in interactome composition as a function of different biological conditions such as metabolic changes, differences in cell growth/the cell cycle, forms of stress (hypoxia, oxidative stress, nutrient deprivation, etc.), developmental and differentiation stages, or the response to drugs. Applied to query such biological contexts, mRNA interactomes and their responses could offer unprecedented insights into biological states, complementing analyses of transcriptomes and proteomes.

EXPERIMENTAL PROCEDURES

In Vivo Isolation of HeLa RBPs

HeLa cells were grown overnight in the presence (PAR-CL) or absence (cCL) of 4-thiouridine. Cells were irradiated with UV light at 254 nm (for cCL) or 365 nm (for PAR-CL), harvested, and lysed. Poly(A)$^+$ mRNAs and crosslinked proteins were captured with oligo(dT)$_{25}$ magnetic beads (NE Biolabs) as described in the Supplemental Information.

Mass Spectrometry, Protein Identification, and Quantification

Proteins were processed following standard protocols, and the resulting peptides were fractionated and analyzed on a nano-HPLC system (Proxeon) or nano-Acquity UPLC system (Waters) coupled directly to an LTQ Orbitrap Velos (Thermo Fisher Scientific). A detailed description of the sample preparation, protein identification, and quantification can be found in the Supplemental Information.

GFP-Based Method for Detection of mRNA-Protein Interactions

HeLa cells expressing N- or C-terminally EGFP/YFP-tagged proteins (Table S7) were induced with tetracycline, irradiated with UV light, and lysed. GFP-binding protein (GBP; GFP agarose trap, Chromotek)-immunoprecipitated mRNAs were detected using an oligo(dT)$_{25}$ probe fused to TRed dye (Sigma). RNAs coimmunoprecipitated with GFP/YFP-tagged proteins were identified by RNASeq. Detailed protocols can be found in the Supplemental Information.

ACCESSION NUMBERS

The data associated with this manuscript are accessible from the ProteomeCommons.org Tranche (https://proteomecommons.org/dataset.jsp?i =Sy2f3AM%2BCJtz81p4Vibcy44KiiM2cAvgjP8YM%2FXraQjyL1WMyR41UOEJk6iM3Z8hNYVD2YZ1TGaEo4NlWYDA1gKl3SAAAAAAAABMmw%3D%3D and http://www.ebi.ac.uk/arrayexpress [Accession E-MTAB-869]). The

Table 1. RNA-Binding Metabolic Enzymes

Protein	Class	Reactome	Pathway	Cofactor
ALDH18A1	mRNA interactome	13	amino acids and derivatives	dinucleotide
PKM2	mRNA interactome	474	carbohydrates	nucleotide and anionic substrate
ENO1	mRNA interactome	474	carbohydrates	
LTA4H	mRNA interactome	22258, 15369	lipids and lipoproteins; prostanoid metabolism	
ALDH6A1	mRNA interactome	13	amino acids and derivatives	dinucleotide
DUT	mRNA interactome	1698, 957	nucleotides; pyrimidine metabolism	
ASS1	mRNA interactome	13	amino acids and derivatives	nucleotide and anionic substrate
TXN	mRNA interactome	1698	nucleotides	
HADHB	mRNA interactome	22279	fatty acids; triacylglycerol and ketone body	dinucleotide
MDH2	mRNA interactome	1046	pyruvate and TCA	dinucleotide
ADK	mRNA interactome	1698, 522	nucleotides; purine	nucleotide and anionic substrate
FDPS	mRNA interactome	22258	lipids and lipoproteins	
SUCLG1	mRNA interactome	1046	pyruvate and TCA cycle	nucleotide and anionic substrate
FASN	mRNA interactome	22279, 11193	fatty acids, ketone, vitamins, and cofactors	dinucleotide
NQO1	mRNA interactome	13	amino acids and derivatives	dinucleotide
P4HB	mRNA interactome	22258	lipids and lipoproteins	
NME1	mRNA interactome	1698	nucleotides	nucleotide and anionic substrate
SHMT2	candidate RBP		amino acid and folate	
AK2	candidate RBP	1698	nucleotides	nucleotide and anionic substrate
CPS1	candidate RBP	13	amino acids and derivatives	nucleotide and anionic substrate
SDHA	candidate RBP	1046	pyruvate and TCA cycle	dinucleotide
CAD	candidate RBP	1698, 957	nucleotides; pyrimidine	
AKR1B1	candidate RBP	22258, 11057	lipids and lipoproteins; steroid hormones, and vitamins A and D	dinucleotide
BLVRB	candidate RBP	9431	porphyrins	dinucleotide
DLD	candidate RBP	13, 1046, 2071	amino acids and pyruvate; TCA cycle	dinucleotide
MTHFD1	candidate RBP	11238, 11167, 11193	vitamins and cofactors; folate and pterines	dinucleotide
GMPR2	candidate RBP	1698, 522	nucleotides; purine	dinucleotide
PGK1	candidate RBP	723, 474	carbohydrates: glucose	nucleotide and anionic substrate
DECR1	candidate RBP	22279, 22258	fatty acid, triacylglycerol, and ketone body	dinucleotide
ENO3	candidate RBP	723, 474	carbohydrates; glucose	
MVK	candidate RBP	22258	lipids and lipoproteins	nucleotide and anionic substrate
GAPDH	candidate RBP	723, 474	carbohydrates; glucose	dinucleotide
TPI1	candidate RBP	723, 474	carbohydrates; glucose	
LDHB	candidate RBP	1046, 2071	pyruvate and TCA cycle	dinucleotide
KYNU	candidate RBP	13	amino acids and derivatives	
DHCR24	candidate RBP	22258	lipids and lipoproteins	dinucleotide
CAT	candidate RBP	1698, 522	nucleotides; purine	
ACLY	candidate RBP	1505, 22279, 22258	energy integration; fatty acid, triacylglicerol; lipids and lipoproteins	nucleotide and anionic substrate
IDH1	candidate RBP	1046, 22258, 16957	lipids and lipoproteins; pyruvate and TCA cycle	dinucleotide
HADH	candidate RBP	22258, 22279	lipids and lipoproteins; fatty acid, triacylglycerol, and ketone body	dinucleotide
ALDOA	candidate RBP	723, 474	carbohydrates; glucose	

Table 1. *Continued*

Protein	Class	Reactome	Pathway	Cofactor
ALAS2	candidate RBP	9431	porphyrins	
TKT	candidate RBP	1505, 474	energy integration; carbohydrates	
PGAM1	candidate RBP	723, 474	carbohydrates; glucose	
GATM	candidate RBP	13, 813	amino acids and derivatives; creatine	
IDH2	candidate RBP	1046	pyruvate and TCA cycle	

R/Bioconductor data package mRNAinteractomeHeLa contains the R-scripts used for the analysis in this manuscript (http://www.bioconductor.org). Distribution of disordered and repetitive regions in HeLa RBPs can be found in http://www.embl.de/mRNAinteractome.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, six figures, and seven tables and can be found with this article online at doi:10.1016/j.cell.2012.04.031.

ACKNOWLEDGMENTS

We thank Drs. Toby Gibson, Teresa Carlomagno, Wolfgang Huber, and Maria Moreno (EMBL) for expert advice and helpful discussions, and we also thank Dr. Markus Landthaler (MCD, Berlin) for generously sharing his expertise on PAR-CL. We are grateful to Drs. Jan Ellenberg, Rainer Pepperkok (EMBL), Stefan Pusch, and Andreas von Deimling (Universitätsklinikum Heidelberg) for plasmids; Dr. Iain Mattaj (EMBL) for the rabbit GFP antibody; and Dr. Matthias Gromeier (Duke University Medical Center, Durham, USA) for the HeLa Flip-In TRex cell line. We acknowledge Alexis Perez and the EMBL Flow Cytometry Core Facility for FACS experiments and EMBL Gene Core Facility for support throughout this work. This work was supported by grants 514904, 573731, and 632757 from the National Health and Medical Research Council and was also supported by DP0878224 from the Australian Research Council awarded to T.P. A.C. is the beneficiary of a Marie Curie postdoctoral fellowship (FP7).

Received: November 21, 2011
Revised: March 8, 2012
Accepted: April 24, 2012
Published online: May 31, 2012

REFERENCES

Anantharaman, V., Koonin, E.V., and Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res. *30*, 1427–1464.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

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.

Brimacombe, R., Stiege, W., Kyriatsoulis, A., and Maly, P. (1988). Intra-RNA and RNA-protein cross-linking techniques in Escherichia coli ribosomes. Methods Enzymol. *164*, 287–309.

Brunner, E., Ahrens, C.H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E.W., Panse, C., de Lichtenberg, U., Rinner, O., et al. (2007). A high-quality catalog of the Drosophila melanogaster proteome. Nat. Biotechnol. *25*, 576–583.

Butter, F., Scheibe, M., Mörl, M., and Mann, M. (2009). Unbiased RNA-protein interaction screen by quantitative proteomics. Proc. Natl. Acad. Sci. USA *106*, 10626–10631.

Cieśla, J. (2006). Metabolic enzymes that bind RNA: yet another level of cellular regulatory network? Acta Biochim. Pol. *53*, 11–32.

Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell *136*, 777–793.

Darnell, R.B. (2010). RNA regulation in neurologic disease and cancer. Cancer Res. Treat. *42*, 125–129.

Dreyfuss, G., Choi, Y.D., and Adam, S.A. (1984). Characterization of heterogeneous nuclear RNA-protein complexes in vivo with monoclonal antibodies. Mol. Cell. Biol. *4*, 1104–1114.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. *6*, 197–208.

Elzinga, S.D., Bednarz, A.L., van Oosterum, K., Dekker, P.J., and Grivell, L.A. (1993). Yeast mitochondrial NAD(+)-dependent isocitrate dehydrogenase is an RNA-binding protein. Nucleic Acids Res. *21*, 5328–5331.

Elzinga, S.D., van Oosterum, K., Maat, C., Grivell, L.A., and van der Spek, H. (2000). Isolation and RNA-binding analysis of NAD+ -isocitrate dehydrogenases from Kluyveromyces lactis and Schizosaccharomyces pombe. Curr. Genet. *38*, 87–94.

Foster, T.L., Gallay, P., Stonehouse, N.J., and Harris, M. (2011). Cyclophilin A interacts with domain II of hepatitis C virus NS5A and stimulates RNA binding in an isomerase-dependent manner. J. Virol. *85*, 7460–7464.

Ghezzi, D., Saada, A., D'Adamo, P., Fernandez-Vizarra, E., Gasparini, P., Tiranti, V., Elpeleg, O., and Zeviani, M. (2008). FASTKD2 nonsense mutation in an infantile mitochondrial encephalomyopathy associated with cytochrome c oxidase deficiency. Am. J. Hum. Genet. *83*, 415–423.

Glisovic, T., Bachorik, J.L., Yong, J., and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett. *582*, 1977–1986.

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

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.

Hentze, M.W. (1994). Enzymes as RNA-binding proteins: a role for (di)nucleotide-binding domains? Trends Biochem. Sci. *19*, 101–103.

Hentze, M.W., and Argos, P. (1991). Homology between IRE-BP, a regulatory RNA-binding protein, aconitase, and isopropylmalate isomerase. Nucleic Acids Res. *19*, 1739–1740.

Hentze, M.W., and Preiss, T. (2010). The REM phase of gene regulation. Trends Biochem. Sci. *35*, 423–426.

Hentze, M.W., Muckenthaler, M.U., Galy, B., and Camaschella, C. (2010). Two to tango: regulation of Mammalian iron metabolism. Cell *142*, 24–38.

Hockensmith, J.W., Kubasek, W.L., Vorachek, W.R., and von Hippel, P.H. (1986). Laser cross-linking of nucleic acids to proteins. Methodology and first applications to the phage T4 DNA replication system. J. Biol. Chem. *261*, 3512–3518.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., et al. (2005). Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. *33*(Database issue), D428–D432.

Kiri, A., and Goldspink, G. (2002). RNA-protein interactions of the 3′ untranslated regions of myosin heavy chain transcripts. J. Muscle Res. Cell Motil. *23*, 119–129.

Lau, C.K., Bachorik, J.L., and Dreyfuss, G. (2009). Gemin5-snRNA interaction reveals an RNA binding function for WD repeat domains. Nat. Struct. Mol. Biol. *16*, 486–491.

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

Lessard, L., Stuible, M., and Tremblay, M.L. (2010). The two faces of PTP1B in cancer. Biochim. Biophys. Acta *1804*, 613–619.

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.

Liu, X., Szebenyi, D.M., Anguera, M.C., Thiel, D.J., and Stover, P.J. (2001). Lack of catalytic activity of a murine mRNA cytoplasmic serine hydroxymethyltransferase splice variant: evidence against alternative splicing as a regulatory mechanism. Biochemistry *40*, 4932–4939.

Lukong, K.E., Chang, K.W., Khandjian, E.W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends Genet. *24*, 416–425.

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

McBryant, S.J., Lu, X., and Hansen, J.C. (2010). Multifunctionality of the linker histones: an emerging role for protein-protein interactions. Cell Res. *20*, 519–528.

Mesa, A., Somarelli, J.A., and Herrera, R.J. (2008). Spliceosomal immunophilins. FEBS Lett. *582*, 2345–2351.

Mi, H., Kops, O., Zimmermann, E., Jäschke, A., and Tropschug, M. (1996). A nuclear RNA-binding cyclophilin in human T cells. FEBS Lett. *398*, 201–205.

Nagy, E., and Rigby, W.F. (1995). Glyceraldehyde-3-phosphate dehydrogenase selectively binds AU-rich RNA in the NAD(+)-binding region (Rossmann fold). J. Biol. Chem. *270*, 2755–2763.

Nakagawa, J., Waldner, H., Meyer-Monard, S., Hofsteenge, J., Jenö, P., and Moroni, C. (1995). AUH, a gene encoding an AU-specific RNA binding protein with intrinsic enoyl-CoA hydratase activity. Proc. Natl. Acad. Sci. USA *92*, 2051–2055.

Niessing, D., Hüttelmaier, S., Zenklusen, D., Singer, R.H., and Burley, S.K. (2004). She2p is a novel RNA binding protein with a basic helical hairpin motif. Cell *119*, 491–502.

Pan, H., Luo, C., Li, R., Qiao, A., Zhang, L., Mines, M., Nyanda, A.M., Zhang, J., and Fan, G.H. (2008). Cyclophilin A is required for CXCR4-mediated nuclear export of heterogeneous nuclear ribonucleoprotein A2, activation and nuclear translocation of ERK1/2, and chemotactic cell migration. J. Biol. Chem. *283*, 623–637.

Pashev, I.G., Dimitrov, S.I., and Angelov, D. (1991). Crosslinking proteins to nucleic acids by ultraviolet laser irradiation. Trends Biochem. Sci. *16*, 323–326.

Phan, A.T., Kuryavyi, V., Darnell, J.C., Serganov, A., Majumdar, A., Ilin, S., Raslin, T., Polonskaia, A., Chen, C., Clain, D., et al. (2011). Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction. Nat. Struct. Mol. Biol. *18*, 796–804.

Pioli, P.A., Hamilton, B.J., Connolly, J.E., Brewer, G., and Rigby, W.F. (2002). Lactate dehydrogenase is an AU-rich element-binding protein that directly interacts with AUF1. J. Biol. Chem. *277*, 35738–35745.

Radivojac, P., Iakoucheva, L.M., Oldfield, C.J., Obradovic, Z., Uversky, V.N., and Dunker, A.K. (2007). Intrinsic disorder and functional proteomics. Biophys. J. *92*, 1439–1456.

Rammelt, C., Bilen, B., Zavolan, M., and Keller, W. (2011). PAPD5, a noncanonical poly(A) polymerase with an unusual RNA-binding motif. RNA *17*, 1737–1746.

Rothbauer, U., Zolghadr, K., Muyldermans, S., Schepers, A., Cardoso, M.C., and Leonhardt, H. (2008). A versatile nanotrap for biochemical and functional studies with fluorescent fusion proteins. Mol. Cell. Proteomics *7*, 282–289.

Rouault, T.A., Stout, C.D., Kaptain, S., Harford, J.B., and Klausner, R.D. (1991). Structural relationship between an iron-regulated RNA-binding protein (IRE-BP) and aconitase: functional implications. Cell *64*, 881–883.

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.

Schrimpf, S.P., Weiss, M., Reiter, L., Ahrens, C.H., Jovanovic, M., Malmström, J., Brunner, E., Mohanty, S., Lercher, M.J., Hunziker, P.E., et al. (2009). Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. PLoS Biol. *7*, e48.

Shetty, S., Muniyappa, H., Halady, P.K., and Idell, S. (2004). Regulation of urokinase receptor expression by phosphoglycerate kinase. Am. J. Respir. Cell Mol. Biol. *31*, 100–106.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. *3*, Article3.

Stirnimann, C.U., Petsalaki, E., Russell, R.B., and Müller, C.W. (2010). WD40 proteins propel cellular networks. Trends Biochem. Sci. *35*, 565–574.

Suchanek, M., Radzikowska, A., and Thiele, C. (2005). Photo-leucine and photo-methionine allow identification of protein-protein interactions in living cells. Nat. Methods *2*, 261–267.

Tompa, P., and Csermely, P. (2004). The role of structural disorder in the function of RNA and protein chaperones. FASEB J. *18*, 1169–1175.

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.

Twyffels, L., Gueydan, C., and Kruys, V. (2011). Shuttling SR proteins: more than splicing factors. FEBS J. *278*, 3246–3255.

Vuzman, D., and Levy, Y. (2010). DNA search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail. Proc. Natl. Acad. Sci. USA *107*, 21004–21009.

Vuzman, D., and Levy, Y. (2012). Intrinsically disordered regions as affinity tuners in protein-DNA interactions. Mol. Biosyst. *8*, 47–57.

Vuzman, D., Azia, A., and Levy, Y. (2010). Searching DNA via a "Monkey Bar" mechanism: the significance of disordered tails. J. Mol. Biol. *396*, 674–684.

Wetzel, R., and Söll, D. (1977). Analogs of methionyl-tRNA synthetase substrates containing photolabile groups. Nucleic Acids Res. *4*, 1681–1694.

Yang, X.C., Purdy, M., Marzluff, W.F., and Dominski, Z. (2006). Characterization of 3'hExo, a 3′ exonuclease specifically interacting with the 3′ end of histone mRNA. J. Biol. Chem. *281*, 30447–30454.

Zalfa, F., Adinolfi, S., Napoli, I., Kühn-Hölsken, E., Urlaub, H., Achsel, T., Pastore, A., and Bagni, C. (2005). Fragile X mental retardation protein (FMRP) binds specifically to the brain cytoplasmic RNAs BC1/BC200 via a novel RNA-binding motif. J. Biol. Chem. *280*, 33403–33410.

Zeng, L., Zhou, Z., Xu, J., Zhao, W., Wang, W., Huang, Y., Cheng, C., Xu, M., Xie, Y., and Mao, Y. (2001). Molecular cloning, structure and expression of a novel nuclear RNA-binding cyclophilin-like gene (PPIL4) from human fetal brain. Cytogenet. Cell Genet. *95*, 43–47.

EXTENDED EXPERIMENTAL PROCEDURES

In Vivo Isolation of HeLa RBPs

HeLa cells were grown overnight on 15 × 500 cm^2 dishes (per condition) in DMEM medium supplemented with 10% fetal calf serum. In the cases of PAR-CL and its control experiment, medium also contained 100 μM 4-thiouridine (4SU). After PBS wash, 80%–90% confluent cell dishes were placed on ice and irradiated with 0.15 J/cm^2 UV light at 254nm (for cCL) or 365nm (for PAR-CL) as previously described (Hafner et al., 2010; Ule et al., 2003). Cells were harvested and lysed in a buffer containing 500 mM LiCl and 0.5% LiDS, and homogenized using a narrow gauge needle (0.4 mm diameter). Poly(A)$^+$ mRNAs and crosslinked proteins were captured with oligo(dT)$_{25}$ magnetic beads (NE Biolabs). Subsequently, oligo(dT)$_{25}$ beads were washed with buffers containing decreasing concentrations of LiCl and LiDS. RNAs and crosslinked proteins were eluted with 20 mM Tris HCl, pH 7.5. For RNA analysis, samples were digested with proteinase K, and RNA was isolated with RNeasy kit (QIAGEN). For protein analysis, samples were treated with RNase T1 and RNase A (Sigma), and released proteins were analyzed by western blotting, silver staining or MS.

Sample Preparation for Mass Spectrometry

Samples were supplemented with 100 mM DTT and concentrated using Amicon Ultra Centrifugal Filters (0.5 ml, 3 kDa cut off) (Millipore). The following steps including alkylation, buffer exchange and digestion were performed according to standard protocols. Briefly, 200 μl 8 M urea in 0.1 M Tris/HCl, pH 8.5, were added and concentrated. After addition of 100 μl iodoacetamide the samples were mixed at 600 rpm for 1 min and incubated without agitation for 5 min, followed by concentration of the sample. The buffer was exchanged by adding 100 μl 8 M urea in 0.1 M Tris/HCl, pH 8.0, followed by concentration of the sample for three successive rounds. After addition of 0.5 μg of endoproteinase Lys-C in 40 μl 8 M urea in 0.1 M Tris/HCl, pH 8.0, and mixing at 600 rpm for 1 min, the filter units were incubated at room temperature overnight. 120 μl 50 mM NH_4HCO_3 with 0.5 μg trypsin were added and incubated at room temperature for 4 hr. After centrifugation of the filter units to collect the peptides, 50 μl of 0.5 M NaCl were added followed by centrifugation. Samples were acidified with CF_3COOH and desalted using Sep-Pak® cartridges (Vac 1cc (50 mg) tC18) as described elsewhere (Villén and Gygi, 2008). The peptide samples were fractionated into 12 fractions on an Agilent 3100 OFFGEL Fractionator (settings as described in the manual) using Immobiline™ DryStrips (ph 3-10 NL, 13 cm, GE Healthcare). Dried samples were resuspended in 360 μl H_2O and diluted into 1.44 ml 1.25 x IEF stock solution (6% glycerol, 2% Ampholytes pH 3-10 (1:50)). The IEF stripes were rehydrated for 30 min before adding 150 μl of diluted sample to each vial. Focusing was performed at a constant current of 50 mA with a maximum voltage of 8000 V. After reaching 20 kVh the samples were collected, acidified with CF_3COOH and desalted using StageTips (Rappsilber et al., 2007). The Stage Tips were equilibrated with 20 μl Methanol, 20 μl 50% Acetonitril and 0.1% Formic acid and 40 μl 0.1% CF_3COOH. After loading the sample the tips were washed with 40 μl 0.1% Formic acid. Following elution with 40 μl 50% acetonitrile, 0.1% Formic acid the peptide samples were dried in a speed vacuum centrifuge. The samples were diluted in 10 μl 4% acetonitrile, 0.1% formic acid.

LC-MS/MS

Peptides were separated using the nanoAcquity UPLC system (Waters) fitted with a trapping (nanoAcquity Symmetry C_{18}, 5μm, 180 μm x 20 mm) and an analytical column (nanoAcquity BEH C_{18}, 1.7μm, 75μm x 200 mm) or the Proxeon EasyNanoLC system (Thermo Fisher) fitted with a trapping column (self-packed Hydro-RP C_{18} (Phenomenex), 100 μm x 2.5 cm, 4 μm) and an analytical column (self-packed Reprosil C_{18} (Dr. Maisch) 75 μm x 15 cm, 3 μm, 100 Å). The outlet of the analytical column was coupled directly to an LTQ Orbitrap Velos (Thermo Fisher Scientific) using the Proxeon nanospray source. Solvent A was water, 0.1% formic acid and solvent B was acetonitrile, 0.1% formic acid. The samples (5 μL) were loaded with a constant flow of solvent A at 15 μL/min onto the trapping column when using the nanoAcquity UPLC system. Trapping time was 1 min. Peptides were eluted via the analytical column at a constant flow of 0.3 μL/min. During the elution step, the percentage of solvent B increased in a linear fashion from 3% to 40% B in 15 min for the noncrosslinked control samples or from 3% to 40% in 45 min for the crosslinked samples. On the Proxeon EasyNanoLC system the samples were loaded with a constant pressure (250 bar) of solvent A with a total volume of 20 μl onto the trapping column. Peptides were eluted via the analytical column at a constant flow of 0.3 μL/min. During the elution step, the percentage of solvent B increased in a linear fashion from 5% to 25% B in 40 min followed by an increase from 25% to 80% in 5 min. The peptides were introduced into the mass spectrometer via a Pico-Tip Emitter 360 μm OD x 20 μm ID; 10 μm tip (New Objective) and a spray voltage of 2.1 kV was applied. The capillary temperature was set at 200°C. Full scan MS spectra with mass range 300-1700 *m/z* were acquired in profile mode in the FT with a resolution of 30,000. The filling time was set at a maximum of 300 ms with limitation of 10^6 ions. The most intense ions (up to 15) from the full scan MS were selected for sequencing in the LTQ. Normalized collision energy of 40% was used, and the fragmentation was performed after accumulation of 3 × 10^4 ions or after filling time of 100 ms for each precursor ion (whichever occurred first). MS/MS data were acquired in centroid mode. Only multiply charged (2+, 3+) precursor ions were selected for MS/MS. The dynamic exclusion list was restricted to 500 entries with maximum retention periods of 30 s and a relative mass window of 10 ppm. To improve the mass accuracy, a lock mass correction using a background ion (*m/z* 445.12003) was applied.

Protein Identification

The peak lists were generated using the transproteomics pipeline. The peak lists were searched against the human part of UniProtKB (downloaded 13.04.2011) containing 35,346 sequences supplemented with reversed decoy sequences as well as a list of most

common contaminations using the Mascot search engine (version 2.2.03, Matrix Science, London, UK). The peptide mass tolerance was set to 15 ppm and the MS/MS mass tolerance to 0.5 Da. One missed cleavage as well as peptide charges of +1, +2 and +3 were allowed. A fixed modification carbamidomethyl (C) was used and the oxidation on methionine as a variable modification.

False identification rate and identification error probabilities of peptides were estimated by a non-parametric density estimation. Density estimation was performed separately for peptides of different length. Only the set of peptides with false identification rates of 0.01 was used for further analysis. Peptides were grouped into protein groups, which were defined as gene models (ENSEMBL 63). Each protein group is uniquely identified by its ENSEMBL gene ID. Only peptides uniquely mapping to one protein group were used for further analysis. The identification error probability of a protein group was computed as the product of the identification error probabilities of the single peptides assuming independence of the identification events. To remove dependent identification events, only the lowest identification error probability was used, if a peptide was identified multiple times. The false identification rate of protein groups was estimated by non-parametric density estimation using the protein identification error probability as a score. Protein groups are accepted such that the resulting false identification rate is less than 0.01.

Protein Quantification

Differential protein abundance was assessed by two methods using independent information available from LC/MS/MS. The spectral count method seeks for a statistical enrichment of the number of peptide identifications in the CL samples compared to control samples. The software package DESeq (Anders and Huber, 2010) provides a statistical test for count data taking into account the natural biological variation between samples.

The m/z was recalibrated for each LC/MS/MS run individually by fitting a linear function using the identified peptides. The retention time of all LC/MS/MS runs were aligned by multiple alignment (Fischer et al., 2006). Loess was used as a regression method. The retention time was predicted for all identified peptides. Ion counts were used as an estimate of peptide abundance. The ion counts for each peptide were estimated by integrating the intensity in the MS spectra over ± 80 s and ± 6 ppm of the predicted peptide position in the LC/MS/MS run. For one pair of crosslinking and control sample, the differential peptide abundance was estimated by the log2-peptide ratio. This resulted in differential peptide abundance measures for three biological experiments. The differential abundance of a protein group was estimated as the trimmed mean over all differential peptide abundances where 20% of the data are trimmed on both sides. Only protein groups with at least two quantification events were used for differential testing. Taking into account the biological variation between replicates, the differential protein abundance was tested against the null hypothesis that the difference is zero by a moderated t test. Since local variance estimates for three samples are unstable, the moderated t test gains in power over a standard t test by combining the local variance estimated for each protein with a global variance estimated over all proteins by an empirical Bayesian method (Lönnstedt and Speed, 2002; Smyth, 2004). p values were corrected for multiple testing by the method of Benjamini-Hochberg. Protein groups with an adjusted p value of max. 0.01 were reported in the mRNA-interactome.

Protein Identification Bias and Gene Set Enrichment Analysis

Four sets of proteins were compared to each other: all human proteins annotated in ENSEMBL (version 63), proteins identified in the HeLa whole-cell lysate, the HeLa mRNA-interactome and proteins annotated by the GO-term "RNA-binding" in ENSEMBL (version 63). Four values were computed for each protein: the pI implemented in the trans proteomic pipeline; the mean of the amino acid hydrophobicity index; the mean normalized mRNA level over 16 arrays of HeLa cells extracted from the ArrayExpress atlas (ArrayExpress accession E-MTAB-62); and the length of proteins in terms of the number of amino acids.

Gene Ontology, Interpro, Reactome pathways, and Superfamily annotation were downloaded from ENSEMBL (version 63) and Reactome (version 36). Enrichment of categories was tested for the mRNA interactome compared to the background of proteins identified in the whole-cell lysate. p values were calculated by Fisher's exact test. p values were corrected for multiple testing by the method of Benjamini-Hochberg.

Comparison between cCL and PAR-CL

To compare the efficiency of the two crosslinking protocols, peptide abundance ratios between cCL and PAR-CL samples are computed for two experiments and summarized for each protein as above. Protein ratios of the two samples are tested for difference from zero by a moderated t test (limma). p values are corrected for multiple testing (Benjamini-Hochberg). Proteins with a false discovery rate smaller than 0.2 are colored in red (enriched in the cCL) or blue (enriched in the PAR-CL).

Fluorescence-Based Method for Detection of mRNA-Protein Interactions

One 100 mm dish of HeLa cells expressing different N- or C-terminally EGFP/YFP tagged proteins at 40%–50% of confluence was induced for 16 hr with 1 μg/ml tetracycline. Cells were then irradiated with UV254 light (see above), resuspended in NP40 lysis buffer and homogenized as above. Lysates were diluted in 500 mM NaCl, 0.05% SDS, 50 mM pH7.5 Tris-HCl buffer and incubated with GFP-binding protein (GBP) coupled to agarose beads (GFP agarose trap, Chromotek) (Rothbauer et al., 2008). GBP beads were washed with decreasing concentrations of NaCl and SDS, blocked with buffer containing E. Coli tRNA and 150 mM LiCl, and then incubated in hybridization buffer containing 500 mM LiCl, 0.05% LiDS and 40nM oligo$(dT)_{25}$ probe fused to Texas Red (TRed) dye (Sigma). Excess of oligo $(dT)_{25}$ TRed was removed by washing with decreasing concentrations of LiCl. GBP beads

were resuspended in 200 mM LiCl and 20 mM Tris HCl, pH 7.5 buffer and transferred to an opaque black 96 well plate. Fluorescence was measured in TECAN Safire II microplate reader.

Analysis of RBP-Bound RNAs by Deep Sequencing

For high-throughput sequencing by crosslinking and GBP immunoprecipitation (HITS-GBP CLIP), three 100 mm dishes of HeLa cells at 80% confluence were irradiated with UV254 light and GBP immunoprecipitates generated as above. Samples were eluted with 200 mM glycyl glycine, pH 2.5, and immediately neutralized with 1M trizma base, pH 10.8. Eluted RNA-protein complexes were digested with proteinase K and released RNAs were further isolated using the RNeasy kit (Quiagen).

All next-generation sequencing libraries were obtained using the Total RNA-Seq kit following the manufacturer's protocol (Applied Biosystems). Briefly, RNA was digested with RNaseIII for 3 min (GBP HITS-CLIP) or 10 min (RNA pull down with oligo(dT)) at 37°C. Adapters were then hybridized (10 min at 65°C), and ligated (overnight at 16°C) before generating cDNA by reverse transcription (30 min at 37°C). cDNA of about 150 - 250 nt size was excised from a 10% urea PAGE gel (Invitrogen Novex) and the slice was directly loaded into a PCR reaction that was amplified for 15 (RNA from oligo(dT) pull downs) or 18 (GBP HITS-CLIP) cycles. Libraries were multiplexed at a final concentration of 0.5pM for emulsion PCR before being sequenced on either a v4 or 5500 XL SOLiD sequencer (Applied Biosystems). Sequence tags were mapped with Applied Biosystems ''Lifescope'' software using default settings. Identified mRNAs were evaluated for differences in read counts between the candidate RBP and the EGFP negative control immunoprecipitation samples. p values were computed using a negative binomial model (bioconductor package DESeq). The dispersion was calculated as the maximum of the per-gene dispersion estimate and the median over all genes. p values were corrected for multiple testing by the method of Benjamini-Hochberg.

Analysis of Protein Domains

Classical and nonclassical RBDs were selected from all Pfam domains by manual annotation. First, all classical Pfam RBDs are tested for enrichment in the HeLa mRNA interactome compared to the whole-cell lysate by Fisher's exact test. In a second step, the non-classical RNA-binding domains are tested for enrichment using only the proteins that do not contain any of the classical RBD. In a final step, all domains present in at least three proteins of the mRNA interactome are tested for enrichment using only proteins that do not contain any known RNA-binding domain. p values were corrected for multiple testing by the method of Benjamini-Hochberg.

Disordered, Low-Complexity, and Repetitive Regions

For all positions in a protein a score for intrinsic disorder is computed using IUPred (Dosztányi et al., 2005). Amino acid residues with a score larger than 0.4 are called disordered. For each protein the ratio of disordered residues is estimated.

The complexity is computed for each amino acid residue as the Shannon entropy of the amino acid distribution within a window of 10 amino acid residues before and after the position. The entropy of the whole protein database is ~4.1 bits. We consider amino acid residues with an entropy smaller than 3 bits as low complexity. The proportion of low complexity residues within the set of disordered residues is computed for each protein.

The probability of occurrence in the disordered regions of the whole-cell lysate is computed for each pair of amino acids neighboring in the protein sequence (and with one or two wild-card amino acids in between). Overrepresentation of an amino acid pair in the disordered regions is tested by a binomial test. p values are corrected for multiple testing by the method of Benjamini-Hochberg. Di-mers with a p value smaller than 0.01 are regarded as repetitive within disordered regions.

For each of the three measures a distribution is plotted for the proteins in the HeLa mRNA interactome, proteins detected in the whole-cell lysate, and all known human proteins. Furthermore a distribution is plotted for the set of proteins that do neither contain a known RNA-binding domain, nor a RNA-related GO annotation (RBD unknown). The difference of the distributions of the mRNA interactome, proteins lacking known RBD and the whole-cell lysate was tested by Kolmogorov-Smirnov test.

To report amino acid di-mers, we restricted ourselves to those di-mers that are overrepresented in the mRNA interactome. We counted for each di-mer, how often it was called ''repetitive'' in the mRNA interactome and in the whole-cell lysate and tested for enrichment by Fischer's exact test. Only patterns with a corrected p value (Benjamini-Hochberg) smaller than 0.05 are listed in Figure S6. The size of the symbols is proportional to the odds-ratio.

For amino acid logos, amino acid composition is reported only for disordered regions. All values in Figures 6 and S6 are log2 fold changes between the regarded set of proteins and the whole-cell lysate, thus the whole-cell lysate has log2-fold change of 0 for all properties. All subsequences in a window of five amino acids before and after a repetitive residue are listed and sorted by their central amino acid. For each central amino acid, a sequence logo is plotted in Figures 6 and S6. The height of the symbols is proportional to the distribution of the amino acids at this position in the mRNA interactome.

Establishment of Stable HeLa Cell Lines

Chimeric cDNAs encoding the different EGFP- or YFP-tagged proteins were amplified by PCR from already established EGFP/YFP-containing plasmid libraries. Alternatively, a HeLa cDNA library and EGFP plasmid were used as templates for fusion PCR. Resulting chimeric cDNAs were cloned into pCDNA5/FRT/TO (Invitrogen). HeLa TRex Flip-in cell lines (Table S7) were established as described in the manufacturer's protocol (Flip In TRex, Invitrogen).

Western Blotting and Silver Staining

Proteins co-isolated by oligo(dT) pull down were analyzed by silver staining, according to standard protocols, and by western blotting using antibodies against CUG-BP (Santa Cruz Biotechnology), PTBP1, α-tubulin, β-actin (all from Sigma) and Histone 3 and 4 (Abcam) following the manufacturers' recommendations. EGFP/YFP-tagged proteins were detected with a rabbit antibody or a rat monoclonal GFP antibody (Chromotek).

Real-Time PCR

Oligo(dT)-isolated and proteinase K-digested RNAs as well as total RNA from whole-cell lysate, were purified using RNeasy kit (Quiagen), visualized using an RNA 6000 Pico Bioanalyzer Chip (Agilent technologies), and then retrotranscribed into cDNA (SuperScript II, Invitrogen) following the manufacturer's recommendations. Reverse-transcriptase quantitative PCR (RT-qPCR) was performed by SYBR green (Applied biosystems) with specific primers against 18S rRNA (all from 5′ to 3′; forward: GAAACTGC GAATGGCTCATTAAA, reverse: CACAGTTATCCAAGTGGGAGAGG), *GAPDH* (f: GTGGAGATTGTTGCCATCAACGA, r: CCCATTC TCGGCCTTGACTGT), β-actin (f: CGCGAGAAGATGACCCAGAT, r: TCACCGGAGTCCATCACGAT) and thymidylate synthase (f: GGCAGAATACAGAGATATGGAATCAGA, r: TCGTCAGGGTTGGTTTTGATG) mRNAs in a 7500 Real time PCR system (Applied Biosystems), and normalized against a Renilla luciferase spike-in control mRNA, added during cell lysis (f: GAATTTGCAGCATA TCTTGAACCAT, r: GGATTTCACGAGGCCATGATAA).

SUPPLEMENTAL REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Dosztányi, Z., Csizmók, V., Tompa, P., and Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol. *347*, 827–839.

Fischer, B., Grossmann, J., Roth, V., Gruissem, W., Baginsky, S., and Buhmann, J.M. (2006). Semi-supervised LC/MS alignment for differential proteomics. Bioinformatics *22*, e132–e140.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.C., Munschauer, M., et al. (2010). PAR-CliP–a method to identify transcriptome-wide the binding sites of RNA binding proteins. J. Vis. Exp. *41*, e2034.

Kelley, L.A., and Sternberg, M.J. (2009). Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protoc. *4*, 363–371.

Lönnstedt, I., and Speed, T. (2002). Replicated microarray data. Statist. Sinica *12*, 31–46.

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.

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.

Villén, J., and Gygi, S.P. (2008). The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. Nat. Protoc. *3*, 1630–1638.

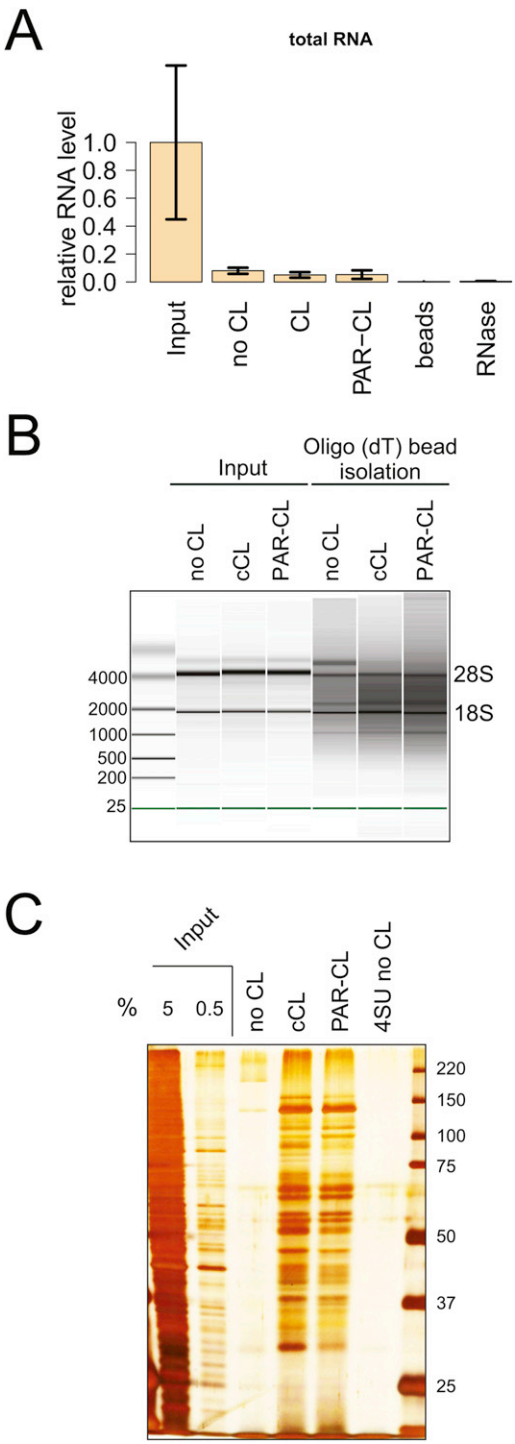

Figure S1. Interactome Capture: RNA and Protein Quality Control, Related to Figure 1

After applying cCL, PAR-CL, or control protocols, RNA and crosslinked proteins were isolated with oligo(dT) beads.

(A) RNA isolated by oligo(dT) pull-down was measured by using a nanodrop device. Error bars represent SDs from four independent experiments.

(B) Proteins were digested with proteinase K and the released RNAs were analyzed using a 6000 Pico bioanalyzer.

(C) In parallel, RNAs were digested with RNase T1 and A, and released proteins were analyzed by silver staining of a SDS-polyacrylamide gel.

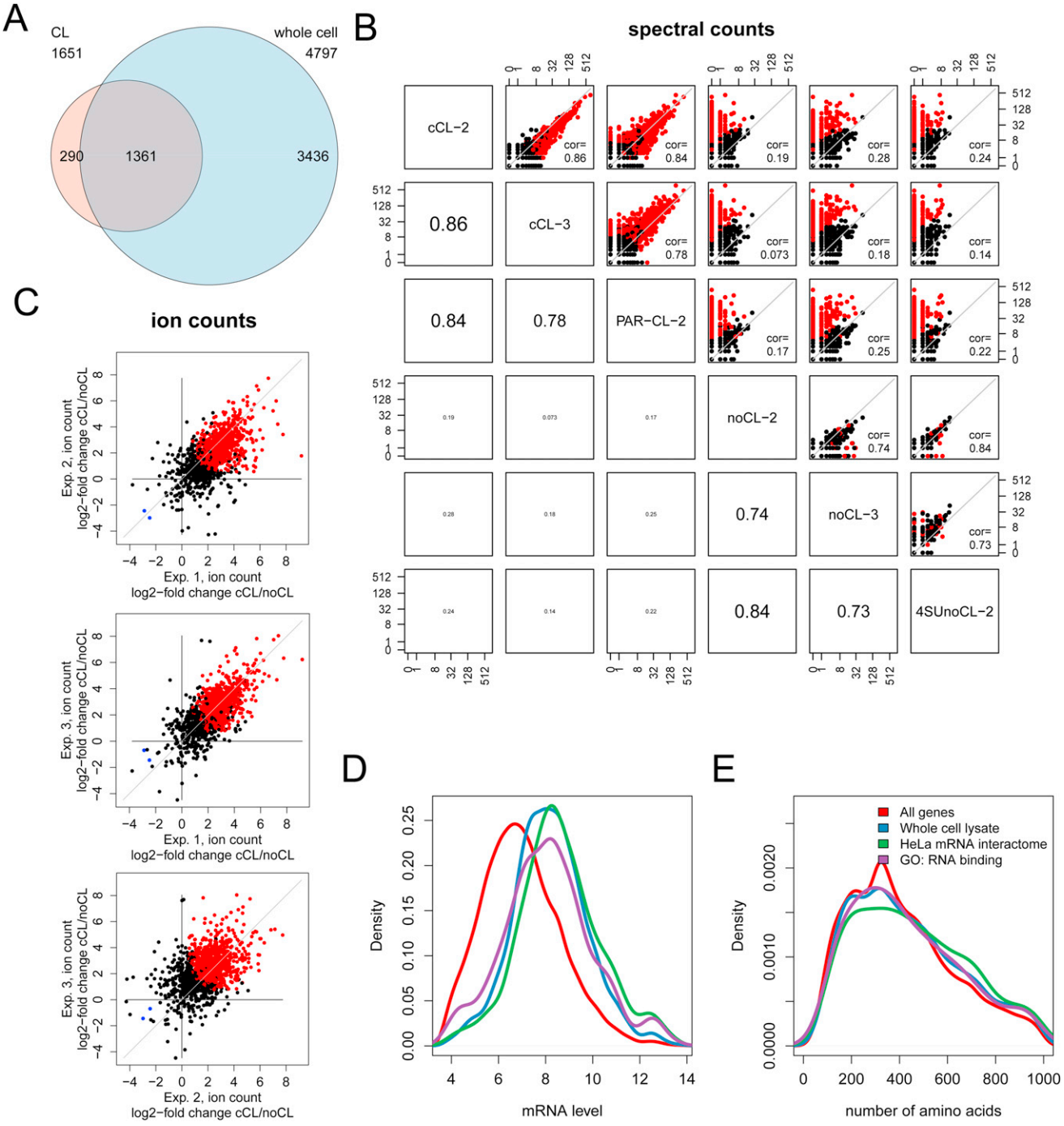

Figure S2. Mass Spectrometry Data Analysis, Related to Figure 2

(A) Venn diagram comparing the number of identified proteins in the three crosslinking (CL) experiments (pink) or in HeLa whole-cell lysate (blue).

(B) Scatter plots of spectral counts comparing three CL experiments to three control experiments. Each dot represents one protein. The axes depict the number of unique peptide identifications. Proteins in red are significantly enriched according to the spectral count method.

(C) Scatter plots of differential ion-counts of three CL experiments compared to controls. Each dot represents one protein. Axes depict the log2-fold change in ion-counts between the CL experiment and the control experiment. Proteins significantly enriched according to the ion-count method in CL or control experiments are depicted by red or blue dots, respectively.

(D) Density of mRNA levels of all human proteins (red), HeLa whole-cell lysate (blue), HeLa mRNA interactome (green), and proteins annotated by GO term "RNA-binding" (purple).

(E) Density of the number of amino acids (protein length) for the same protein groups as in (D).

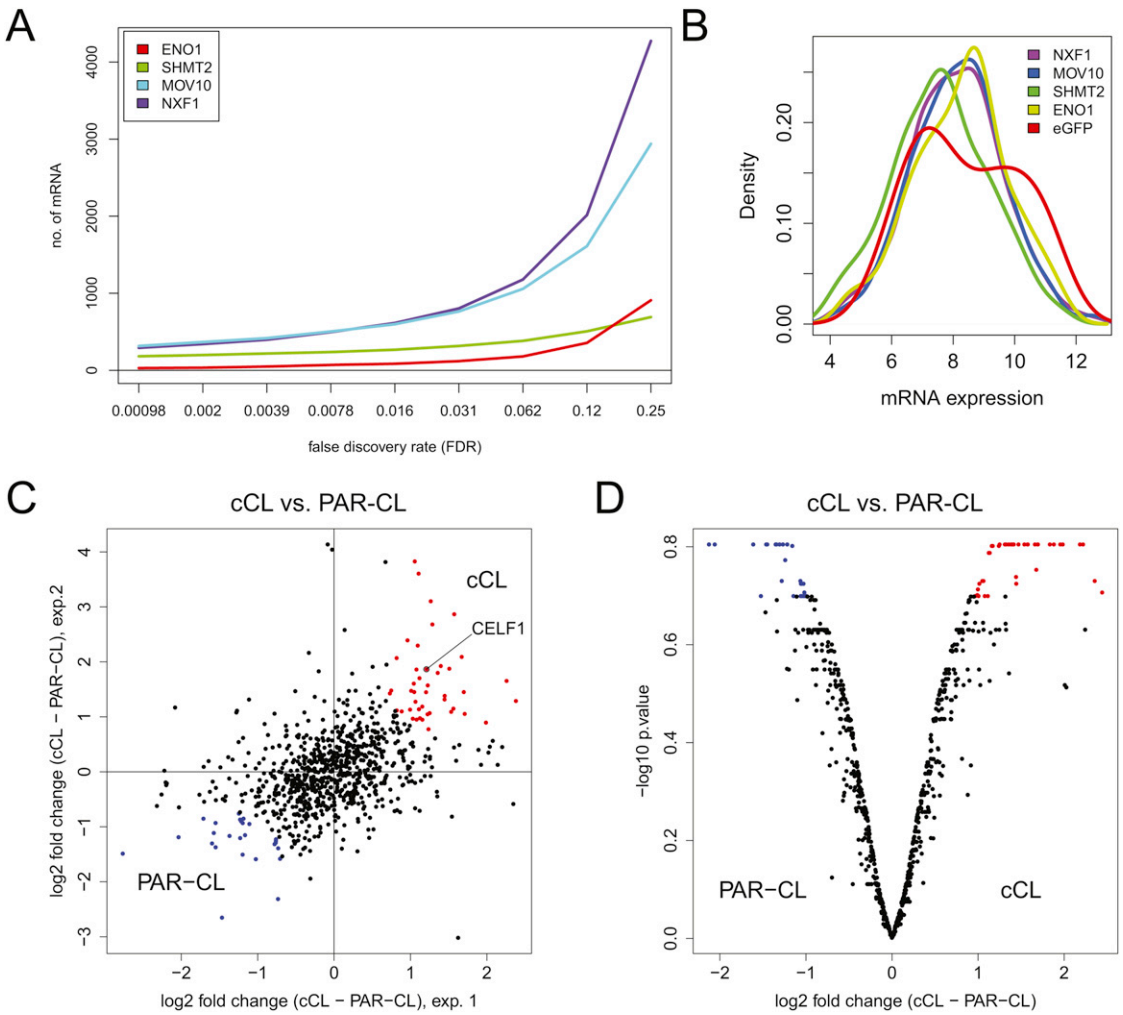

Figure S3. CLIPseq Quality Control and Quantitative Comparison of cCL and PAR-CL Data Sets, Related to Figures 2 and 3

(A) Number of specific mRNA transcripts called from four independent CLIPseq experiments for different FDR cut-offs.

(B) Density of mRNA levels in HeLa cells of the genes called by the CLIPseq experiments at $p < 0.05$.

(C and D) Scatter plot and Volcano plot comparing two cCL and two PAR-CL experiments. Each dot represents one protein. The axis depicts the log2-fold change in ion-counts between the cCL experiment and the PAR-CL. Proteins significantly enriched according to the ion-count method in cCL or PAR-CL with 20% FDR are depicted in red or blue, respectively. Scatter plot shows a correlation of 0.43. CELF1 (CUG-BP) that was found to be more efficiently crosslinked to RNAs by cCL (Figure 1E) is indicated in the scatter plot.

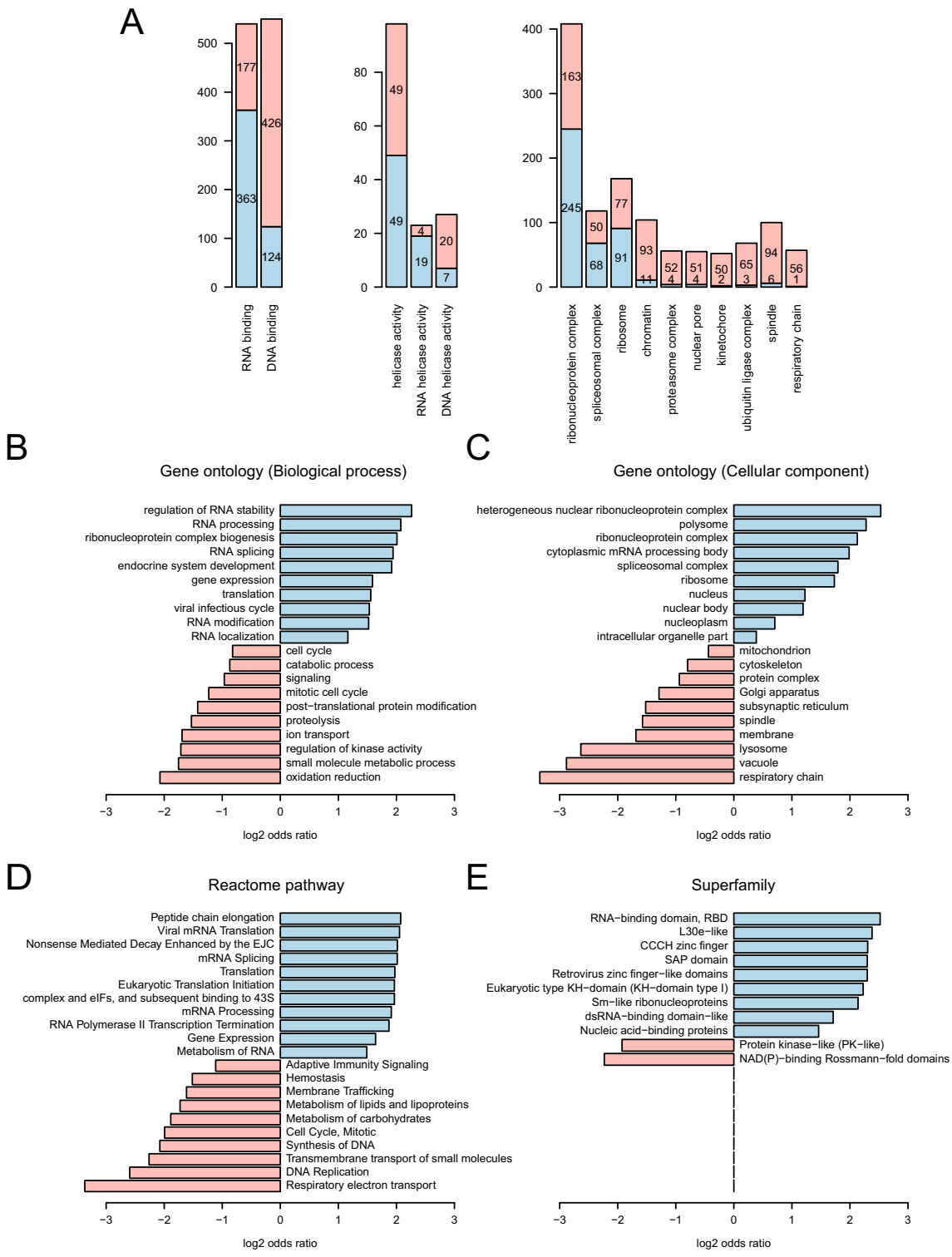

Figure S4. Gene Ontology Analysis of the HeLa mRNA Interactome, Related to Figure 4

(A) "RNA-binding" and "DNA-binding" or "helicase," "RNA-helicase" and "DNA-helicase" molecular function "GO terms" present in the HeLa mRNA interactome (blue) or absent from the HeLa mRNA interactome but identified in the whole-cell lysate (pink). Number of proteins annotated as a component of different cellular macrocomplexes ("cellular compartment" GO terms) present in the HeLa mRNA interactome (blue) or identified exclusively in the whole-cell lysate (pink). (B–E) Ten of the most significantly over- (blue) and under-represented (pink) "biological process" GO terms, (C) "cellular compartment" GO-terms, (D) reactome pathways, and (E) superfamilies in the HeLa mRNA interactome.

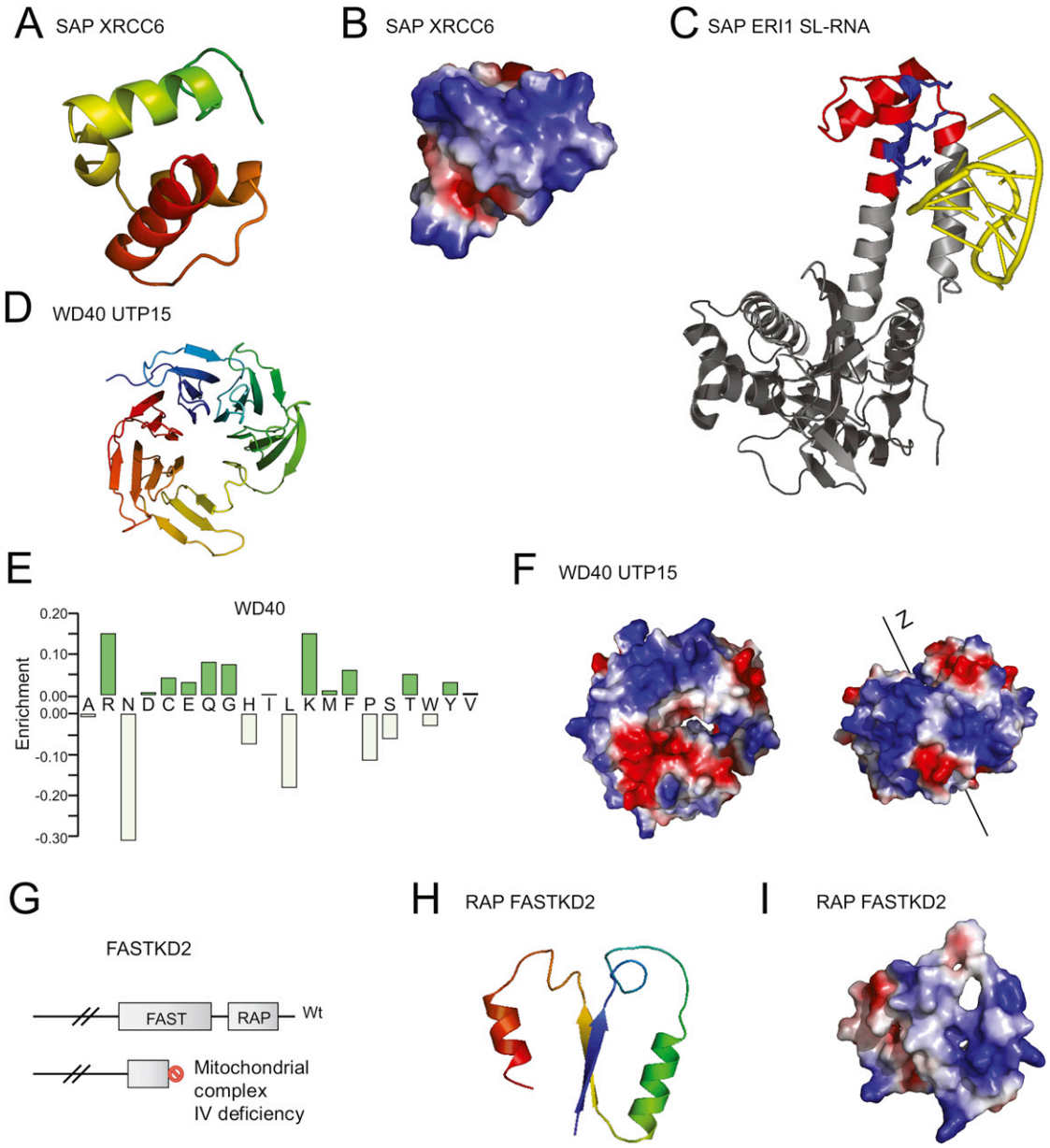

Figure S5. Structural Analysis of SAP, WD40, and RAP Domains, Related to Figures 5 and 6

(A) Ribbon diagram representation of the crystal structure of the SAP domain of XRCC6 (PDB 1JEQ).

(B) Surface charge of the XRCC6 SAP domain. Positive and negative charge patches are colored in blue and red, respectively.

(C) Ribbon diagram representation of the crystal structure of the 3′-5′ exonuclease ERI1 bound to RNA (PDB 1ZBH). RNA is shown in yellow, the SAP domain in red and its amino acids contacting the RNA in blue.

(D) Homology modeling of the WD domain of UTP15 using Phyre2 (Kelley and Sternberg, 2009) and represented as a ribbon diagram. UTP15 was modeled based on the WD domain of WDR5 (100% confidence, 24% identity).

(E) Amino acid enrichment of the WD domains in the HeLa RBPs compared to WD domains in the HeLa whole-cell lysate.

(F) Surface charge of the UTP15 WD domain model. Colors as in (B).

(G) Scheme of the protein domain organization of FASTKD2 and its truncated version found in Mitochondrial complex IV deficiency.

(H) Homology modeling of FASTKD2 RAP domain using Phyre2 and represented as a Ribbon diagram. The FASTKD2 RAP domain was modeled based on the crystal structure of a putative endonuclease-like protein 2 from *Neisseria gonorrhoeae* (91% confidence, 23% identity).

(I) Surface charge of the FASTKD2 RAP domain model. Colors as in (B).

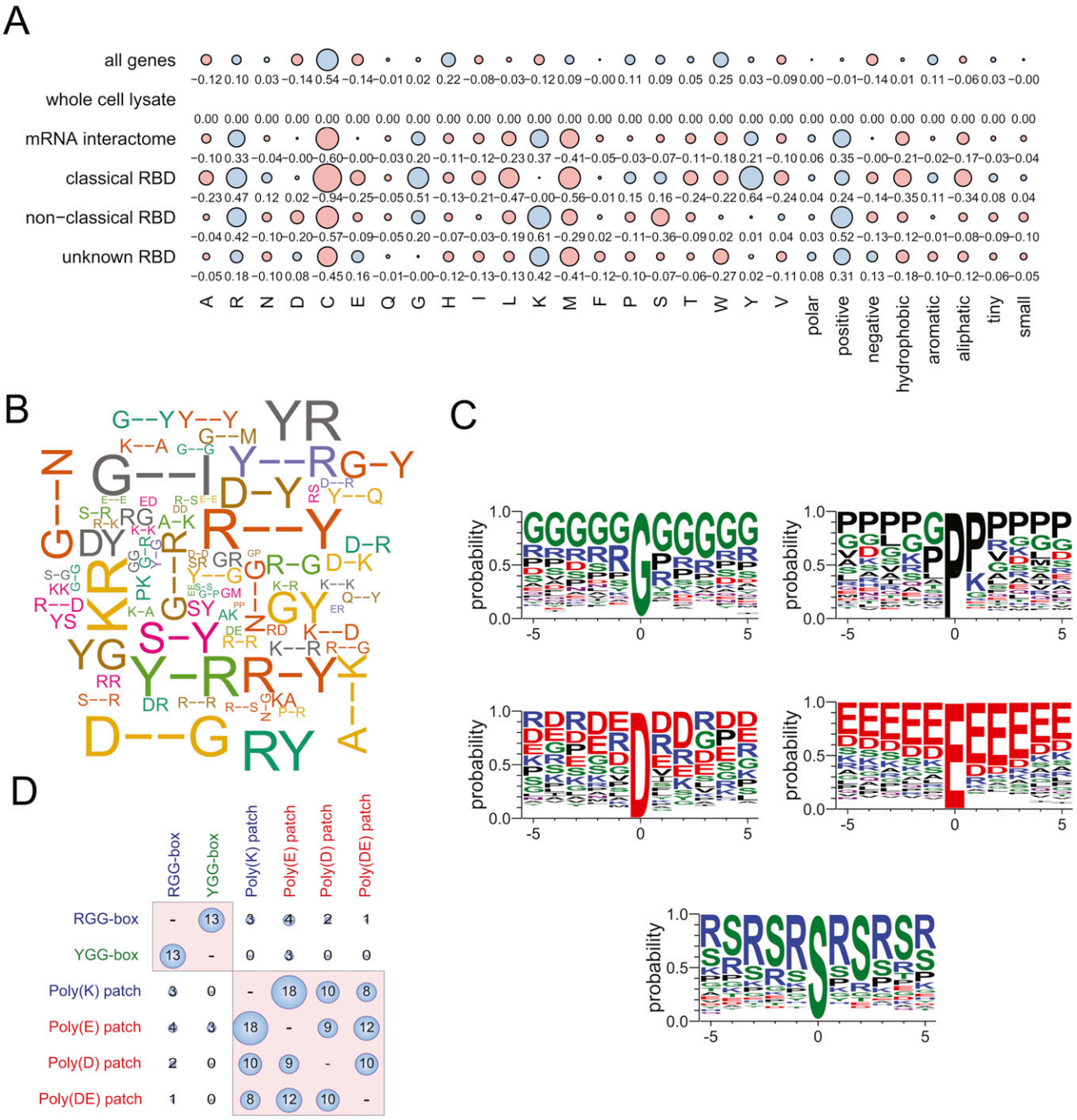

Figure S6. Identification of Repetitive Motifs in the Disordered Regions of the HeLa mRNA Interactome Proteins, Related to Figure 6

(A) Amino acid enrichment in disordered regions of HeLa RBPs compared to similar segments in HeLa whole-cell lysate proteins. Blue balloons indicate overrepresentation whereas pink balloons indicate underrepresentation.

(B) Repetitive sequence patterns that occur significantly more often in the mRNA interactome than in whole-cell lysate proteins. The size of the patterns symbolizes the odds-ratio of occurrence, comparing the mRNA interactome with the whole-cell lysate.

(C) Sequence logos for repetitive regions. 11-mers around all repetitive residues were extracted from the HeLa mRNA interactome and grouped by their central residue. The height of the letters is proportional to the probability of occurrence of amino acids at each position.

(D) Co-occurrence of repetitive motifs in HeLa mRNA interactome proteins. Balloon plot showing the number of cases in which two different repetitive motifs are found together in the same protein.