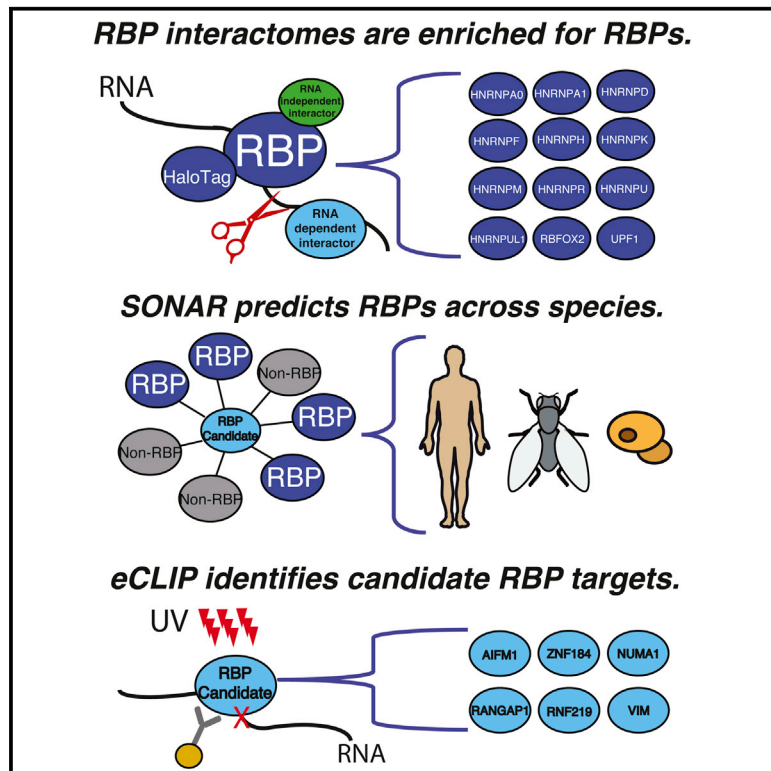# SONAR Discovers RNA-Binding Proteins from Analysis of Large-Scale Protein-Protein Interactomes

## Graphical Abstract



## Highlights

- Quantitative proteomics show that RNA-dependent interactors are frequently RBPs

- SONAR identifies thousands of unannotated candidate RBPs in yeast, fly, and human

- SONAR candidate RBPs are enriched for proteins with zinc finger domains

- Enhanced CLIP identifies transcriptome-wide targets of SONAR candidate RBPs

## Authors

Kristopher W. Brannan, Wenhao Jin, Stephanie C. Huelga, ..., Marie K. Schwinn, Danette L. Daniels, Gene W. Yeo

## Correspondence

geneyeo@ucsd.edu

## In Brief

The repertoire of RNA-binding proteins (RBPs) is incomplete and is expanding to include many non-canonical RNA associated proteins. Brannan et al. describe the interaction network for 12 nuclear and cytoplasmic RBPs and develop an RBP classifier termed SONAR, which identifies previously unannotated RNA-protein associations utilizing existing protein-protein interaction networks.

## Accession Numbers

GSE86035
PXD003999
PXD004000

Molecular Cell

# Article

CellPress

# SONAR Discovers RNA-Binding Proteins from Analysis of Large-Scale Protein-Protein Interactomes

Kristopher W. Brannan,[1,7] Wenhao Jin,[2,7] Stephanie C. Huelga,[1,7] Charles A.S. Banks,[3] Joshua M. Gilmore,[3] Laurence Florens,[3] Michael P. Washburn,[3,4] Eric L. Van Nostrand,[1] Gabriel A. Pratt,[1] Marie K. Schwinn,[5] Danette L. Daniels,[5] and Gene W. Yeo[1,2,6,8,*]

[1]Department of Cellular and Molecular Medicine, Stem Cell Program and Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA 92093, USA
[2]Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore, 117597 Singapore, Singapore
[3]Stowers Institute for Medical Research, 1000 East 50th Street, Kansas City, MO 64110, USA
[4]Department of Pathology and Laboratory Medicine, University of Kansas Medical Center, 3901 Rainbow Boulevard, Kansas City, KS 66160, USA
[5]Promega Corporation, 2800 Woods Hollow Road, Madison, WI 53711, USA
[6]Molecular Engineering Laboratory, A*STAR, 138673 Singapore, Singapore
[7]Co-first author
[8]Lead Contact
*Correspondence: geneyeo@ucsd.edu
http://dx.doi.org/10.1016/j.molcel.2016.09.003

## SUMMARY

RNA metabolism is controlled by an expanding, yet incomplete, catalog of RNA-binding proteins (RBPs), many of which lack characterized RNA binding domains. Approaches to expand the RBP repertoire to discover non-canonical RBPs are currently needed. Here, HaloTag fusion pull down of 12 nuclear and cytoplasmic RBPs followed by quantitative mass spectrometry (MS) demonstrates that proteins interacting with multiple RBPs in an RNA-dependent manner are enriched for RBPs. This motivated SONAR, a computational approach that predicts RNA binding activity by analyzing large-scale affinity precipitation-MS protein-protein interactomes. Without relying on sequence or structure information, SONAR identifies 1,923 human, 489 fly, and 745 yeast RBPs, including over 100 human candidate RBPs that contain zinc finger domains. Enhanced CLIP confirms RNA binding activity and identifies transcriptome-wide RNA binding sites for SONAR-predicted RBPs, revealing unexpected RNA binding activity for disease-relevant proteins and DNA binding proteins.

## INTRODUCTION

RNA-binding proteins (RBPs) act in both the nucleus and cytoplasm during every step of the RNA life cycle to exert precise and responsive control of gene expression. Given the number and expression levels of eukaryotic RBPs, and the fact that many are required for viability, it is unsurprising that evidence demonstrating their importance in development and disease emerges daily (Castello et al., 2013; Lukong et al., 2008).

The catalog of RBPs is diverse and has grown to include metabolic enzymes, cell-cycle regulators, and many other factors not previously associated with RNA biology (Beckmann et al., 2015; Conrad et al., 2016). As this diverse RBP catalog grows, complex regulatory networks are emerging that rely on highly coordinated and concurrent target association by numerous RBPs. With the advent of high-throughput techniques to accurately and robustly identify RBP target binding sites (Van Nostrand et al., 2016) and protein interaction partners, it has become possible to elucidate the complete list and the precise molecular functions of RBPs in these emerging RBP-RNA networks in a scalable manner.

Initial efforts to define the complete repertoire of RBPs largely focused on proteins with known or predicted RNA binding domains (RBDs). Attempts to predict RBPs from primary sequence and protein structure have yielded only moderate success (Si et al., 2015), particularly since more than a third of RBPs have no prior RNA binding related homology or annotation (Castello et al., 2013). To expand the catalog of RBPs beyond ones with sequence or structure homology, several experimental approaches have been undertaken. In budding yeast *Saccharomyces cerevisiae*, protein microarrays and affinity purification followed by mass spectrometry identified a significant number of previously unannotated RBPs, including unconventional RNA interactions for a number of known enzymes (Klass et al., 2013; Scherrer et al., 2010; Tsvetanova et al., 2010). More recently, UV cross-linking of proteins to RNA followed by oligo(dT) capture and mass spectrometry (referred to as interactome capture) have been used to identify proteins which bind directly to mRNA. This general approach has identified nearly 900 RBPs in human HeLa cells, 729 RBPs in human HuH-7 cells, 382 nuclear RBPs in human K562 cells, 555 RBPs in mouse embryonic stem cells, 678 RBPs in yeast, and 523 in *Drosophila* embryos (Beckmann et al., 2015; Castello et al., 2012; Conrad et al., 2016; Kwon et al., 2013; Matia-González et al., 2015; Sysoev et al., 2016; Wessels et al., 2016). However, a major limitation of these studies is that the use of oligo(dT) beads to isolate proteins bound to polyadenylated mature mRNAs is biased against

the identification of proteins that interact with premature, unspliced, and unpolyadenylated RNA.

As many RBPs interact co-transcriptionally to influence splicing, 5′ and 3′ end formation, and nuclear localization, for both unpolyadenylated pre-mRNAs and non-coding RNAs, it is important to adopt approaches to extend the RBP catalog to include RBPs not identified by polyadenylated mRNA interactome-capture techniques. One attractive approach is to examine protein-protein interaction (PPI) networks for both nuclear and cytoplasmic RBPs and to discriminate between RNA-mediated and direct protein-protein interactions. At a small scale, quantitative mass spectrometry-based proteomics, in combination with RNase treatments, has been successfully utilized to identify RBPs that are concurrently bound to the same RNA substrates (Flury et al., 2014; Klass et al., 2013).

Here, we propose an RNase-coupled proteomics strategy to identify proteins with RNA binding activity by prioritizing proteins that interact with multiple known (or annotated) RBPs in an RNA-dependent manner. To evaluate this concept, we first focused on RBPs localized in both the nucleus and the cytoplasm, such as the heterogeneous nuclear ribonucleoparticle proteins (HNRNPs), for which RNA binding roles are well documented and transcriptome-wide targets have been characterized (Huelga et al., 2012; Lee et al., 2015; Van Nostrand et al., 2016). We used HaloTag fusion (HT-RBP) pull-downs followed by quantitative mass spectrometry, in both untreated and RNase treated conditions, to characterize the interactomes of ten of the major HNRNPs, as well as the splicing factor (RBFOX2), and the nonsense-mediated decay (NMD) pathway regulator (UPF1). With this approach, we demonstrate that interactors associating with multiple HT-RBPs are highly enriched for annotated RBPs as well as candidate RBPs.

To generalize RBP discovery from PPI networks, we developed a classification algorithm *S*upport *V*ector *M*achine *o*btained from *n*eighborhood *a*ssociated *R*BPs (termed SONAR, available at https://github.com/YeoLab/SONAR) to predict if a given protein of interest is a candidate RBP. Enhanced CLIP (Van Nostrand et al., 2016) performed on a subset of SONAR predictions and HT-RBP interactors corroborated that they are bona-fide RBPs, demonstrating that SONAR successfully discovers uncharacterized candidate RBPs by leveraging protein-protein interactomes.

## RESULTS

### HaloTag Fusion Pull Down of 12 RBPs Followed by Quantitative Proteomics Identifies RNA-Dependent Interactors

We performed a small-scale proteomic study with and without RNase treatment to identify proteins that interact with known RBPs in an RNA-dependent manner. Specifically, we performed an unbiased quantitative proteomic study to systematically and comprehensively identify proteins in complex with 12 canonical RBPs (ten hnRNP proteins, the splicing factor RBFOX2, and nonsense-mediated decay factor UPF1) using HT technology. The rapid binding kinetics of the HaloLink Sepharose-based resin decreases false positives by minimizing resin exposure time to lysate, and covalent capturing allows for transient interactors to be purified by prohibiting diffusion of the primary capture target

from the resin (Daniels et al., 2012; Deplus et al., 2013). Plasmids encoding a full-length open reading frame (ORF) of each RBP fused to a HT-RBP were transfected in HEK293T cells in biological replicates. Confocal microscopy performed on cells expressing these HT-RBPs labeled with fluorescent TMR HaloTag ligand revealed nuclear and cytoplasmic localization (Figure S1). As these proteins are known to interact with RNA substrates in both the nucleus and cytoplasm, we expect that our fusion pull-down approach will identify interacting proteins that associate with both nascent and mature mRNAs.

Cells were lysed 24 hr post-transfection and half of the lysates were left untreated, while the remaining half were subjected to stringent RNase digestion (Figure 1A). Briefly, protein complexes were covalently captured on HaloLink resin, subjected to washing to remove non-specific interactions, and the remaining protein interactors were eluted in urea and analyzed by silver staining (Figure S2A). Following multidimensional protein identification technology (MudPIT) mass spectrometry of eluted interactors, normalized spectral abundance factors (NSAF) values were calculated (Figure 1B), and these values were used to group replicates by hierarchical clustering (Figure S2B). Outlier experiments were removed and the remaining replicate experiments were averaged together. Within each experiment, the distribution of $\log_2$ enrichment scores compared to the control experiment was computed for each protein interactor (HNRNPF is presented as an example in Figure 1B and the others in Figure S2C). The set of significantly enriched protein interactors within each experiment was defined using a *Z* score. Specifically, protein interactors were significantly enriched for interacting with a given HT-RBP if they had a score greater than 1.5 SDs ($\sigma$) from the mean ($\mu$) of the positive values in the distribution (red dashed line, Figure 1B). Interactor enrichment scores were computed for both the RNase-treated and untreated conditions and a total of 3,853 significantly enriched interactors were detected (Table S1). For each HT-RBP, the enrichment scores for protein interactors in the untreated sample were compared to the RNase-treated sample (Figure S2D). Expectedly, the HT-RBP was often the most enriched protein precipitated in the experiments (red dots, Figure S2D). Analysis of all enriched interactions for all HT-RBP baits revealed that our approach detects approximately similar numbers of RNA-independent and RNA-dependent interacting proteins. Indeed, nearly half of all significant protein interactions with HT-RBPs are lost with RNase treatment (Figure 1C).

We found that RNase treatment depleted known RNA-dependent interactors, such as the exon-junction complex proteins (Figures 1D and 1E), which displayed enriched interactions in the untreated sample and indicates RNase-sensitive (RNA-dependent) interactions with HT-HNRNPs. In contrast, we observe known RNA-independent interactions between HT-UPF1 and the other UPF components of the EJC complex, which confirms that our experimental conditions and computational approach reliably distinguishes known RNA-independent versus RNA-dependent protein-protein interactions. Similarly, many of the interactions between HT-RBP baits and the components of the SF3B splicing complex were RNA-dependent (Figure 1D).

We performed a Gene Ontology analysis on the RNA independent interactors for each HT-RBP (Table S2). Terms associated
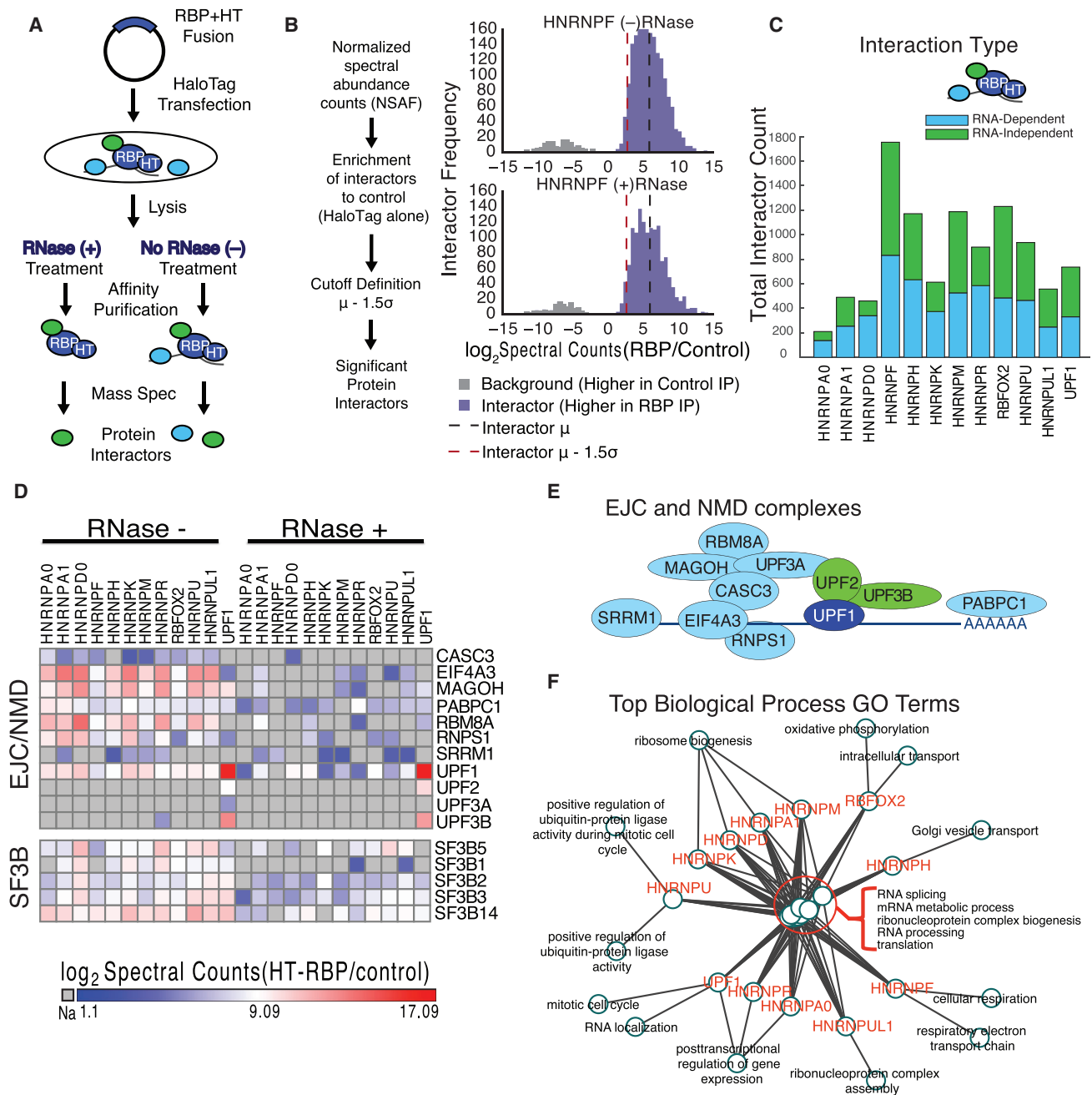
**Figure 1. Identification of Enriched RNA-Binding Protein-Protein Interactors for HT-RBPs**

(A) HaloTag fusion pull-down and MS experimental procedure. RBP-HaloTag fusion protein constructs are transfected into HEK293T cells in replicate, the cells are lysed, and half the lysate is treated with RNase. The affinity purified products are subjected to LC/MS/MS to identify protein interactors.

(B) Analysis flow chart for post processing of MS data. NSAF enrichment score distribution compared to control (HaloTag alone) for hnRNPF pull-down. The grey data points (enrichment < 0) represent the background. For enrichment score higher in hnRNPF experiment (blue), a mean (μ, black dashed line) and SD (σ) is computed. Significant interactions have an enrichment score greater than 1.5 times the SD (1.5σ, red dashed line).

(C) Number of enriched RNA-dependent and RNA-independent interactions for all HT-RBP baits.

(D) Heatmap of specific interactions displaying log$_2$ fold enrichment over control for all HT-RBP baits. The interactions are grouped into NMD/EJC complexes and SF3B complexes.

(E) Exon-junction (EJC) and NMD factors. The green indicates RNA-independent interactors, and the light blue indicates RNA-dependent interactors for HT-UPF1 (dark blue).

(F) Gene ontology characterization of RNA-independent HT-RBP interactors. The shared and unique gene ontology terms displayed as interaction network, where the red text nodes are HT-RBP baits, and the central highlighted nodes are enriched terms shared by all baits (see also Table S2).
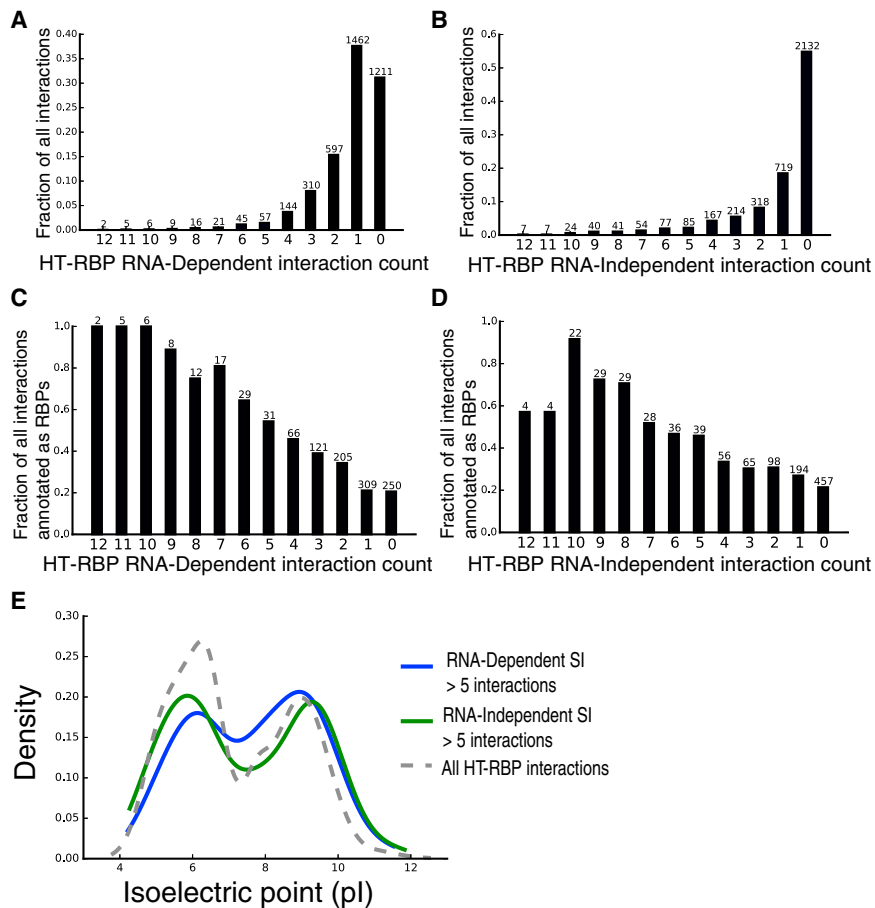
CellPress



**Figure 2. Super Interactors Are Enriched for Known and Candidate RBPs**

(A) Bar chart displays the fraction of all RNA-dependent interacting proteins that come up in 1 (unique interactor) and 2 to 12 HaloTag-RBP experiments (shared interactor). The number of interactors is given at the top of each bar.

(B) Bar chart displays the fraction of all RNA-independent interacting proteins that come up in 1 (unique interactor) and 2 to 12 HaloTag-RBP experiments (shared interactor). The number of interactors is given at the top of each bar.

(C) Bar chart displays the fraction of unique (1 HT-RBP) and shared (2–12 HT-RBPs) RNA-dependent interacting proteins that are RBPs. The number of RBP interactors is given at the top of each bar.

(D) Bar chart displays the fraction of unique (1 HT-RBP) and shared (2–12 HT-RBPs) RNA-independent interacting proteins that are RBPs. The number of RBP interactors is given at the top of each bar.

(E) Density of the calculated isoelectric points (pI) of RNA-dependent super interacting proteins (blue line) and RNA-independent super interacting proteins (green line) compared to all proteins in the HT-RBP interaction set (gray dashed line).

with RNA processing are among the most statistically significantly enriched Gene Ontology (GO) terms identified for HT-RBP RNA-independent interactions (Figure 1F). We also observed significantly enriched ontology terms unique to particular HNRNPs. For example, HNRNPU uniquely interacts in an RNA-independent manner with E3 ubiquitin-protein ligases and proteasome subunits, HNRNPH uniquely interacts with factors involved in Golgi vesicle transport, and HNRNPF and RBFOX2 interact with unique subsets of nuclear encoded mitochondrial proteins, including respiratory electron transport chain components (Figure 1F). In summary, HaloTag fusion pull downs of canonical RBPs successfully identified thousands of RNA-dependent and -independent interactors in human cells.

## RNA-Dependent Interactors that Interact with Multiple RBPs Are Frequently RBPs

We reasoned that the more HT-RBPs a protein interacts with, the more likely it is an RBP itself, particularly if the interactions are RNA-dependent. In fact, for both RNA-dependent and RNA-independent interactors, we observed a strong correlation between the number of HT-RBP baits that a given protein interacts with and the likelihood that the protein is itself an annotated RBP. Nearly half of all RNA-dependent and RNA-independent interactors associate with a single HT-RBP bait and a decreasing number of proteins interact with an increasing numbers of baits (Fig-

ures 2A and 2B). However, as the number of interacting HT-RBPs increases, the fraction of annotated RBPs represented in the set of interactors also increases, such that ~60% of interactors that interact with five or more HT-RBPs are annotated RBPs (Figures 2C and 2D). This effect is even stronger for RNA-dependent interactions. 100% of RNA-dependent interactors associating with more than ten HT-RBPs are known RBPs (Figure 2C), compared to ~70% of RNA-independent interactors (Figure 2D). Based on these analyses, we define proteins that interact with five or more HT-RBPs "super interactors" (SI). The ~12- to 20-fold enrichment of annotated RBPs within the SI proteins is notable, as known RBPs currently comprise ~5%–8% of all annotated protein-coding genes. Higher isoelectric points are characteristic of annotated RBPs (Castello et al., 2012), and when we compared the distribution of isoelectric points for RNA-dependent and RNA-independent SI proteins, we see higher isoelectric points for RNA-dependent SI proteins (Figure 2E), further supporting an enrichment in RNA binding activity in this group of proteins. Our findings support our hypothesis that RNA-dependent interactors of multiple RBPs are enriched for known or candidate RBPs.

## Protein-Protein Interaction Networks Allow Global Prediction of RBPs

Based on our observations that the more HT-RBPs a given protein interacts with, the more likely it is an RBP, we developed a classification algorithm SONAR to calculate an RBP classification score (RCS) for protein baits of interest from large-scale PPI data sets. First, we developed a list of 1,787
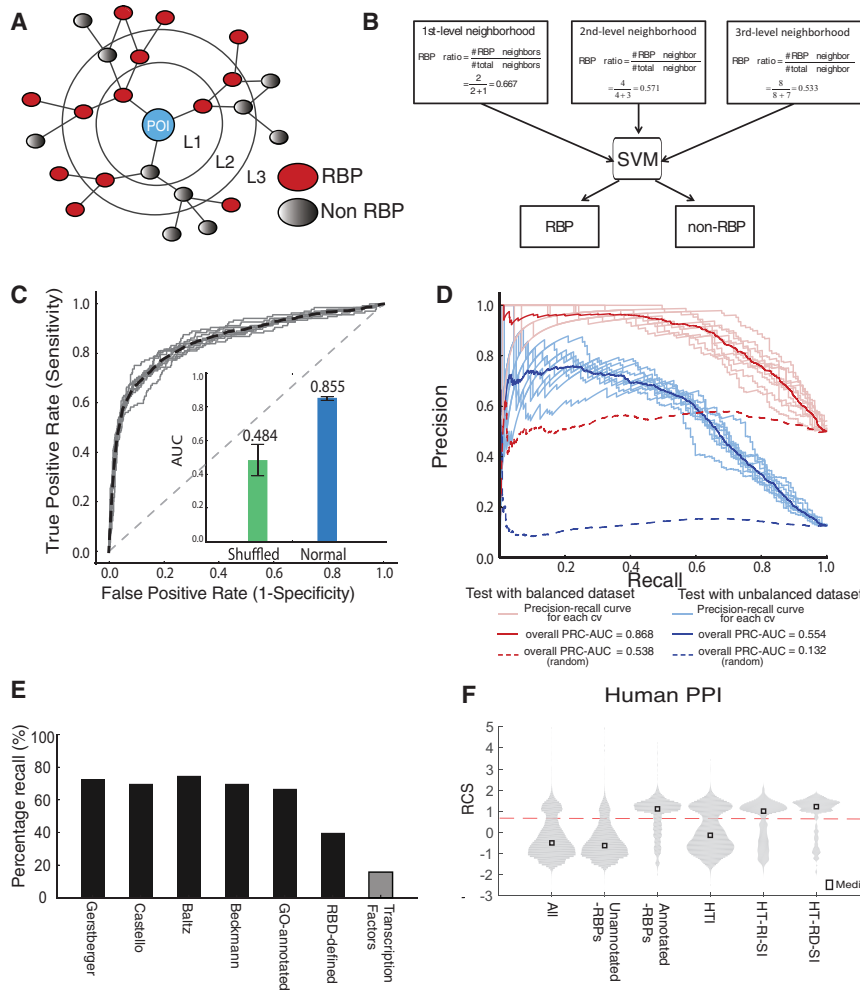
**Figure 3. SONAR RBP Classification Approach**

(A) Diagram of neighborhood classification strategy. Protein of interest (POI) from a given interactome data set with its depicted neighborhood with interactions at different levels is shown. The first level interactions are direct interactions with POI, the second level interactions are interactions with first level neighbors, and third level interactions are interactions with second level neighbors.

(B) Determination of RCS as described in the Experimental Procedures section.

(C) ROC-AUC analysis of classifier performance for human proteins from BioPlex network. The data are represented as mean ± SEM.

(D) PRC-AUC analysis of classifier performance for human proteins from BioPlex network.

(E) Percent recall for SONAR trained on BioPlex PPI network for six RBP lists depicting different RBP annotations and percent of annotated transcription factors (TF) predicted as candidate RBPs.

(F) Violin plots of RCS distributions for all human BioPlex interactors, non-RBP interactors within BioPlex, annotated RBP interactors within BioPlex, all HT interactors, HT-RI-SI, and HT-RD-SI within Bioplex. The median of the distributions are denoted with a square box (see also Table S4).

the receiver operating curve (ROC-AUC) and under the precision recall curve (PRC-AUC) illustrate the high sensitivity (~0.7) and specificity (~0.9) of the classifier (Figures 3C and 3D). We obtained different definitions of RBPs based on interactome capture (Castello et al., 2012), Baltz (Baltz et al., 2012), Beckmann (Beckmann et al., 2015), manual curation (Gene Ontology or GO-annotated), sequence composition alone (RBD-defined), and some combination of experimentally defined and manual curation (Gerstberger et al., 2014) (Figures 3E and S3B). Importantly, the performance of SONAR, as measured by the percentage recall, is robust to different definitions of RBPs (Figures 3E and S3B). Curiously, the RBD-defined set consisting of the largest set (2,551) of proteins predicted to associate with RNA based solely on sequence composition appears to have a low fraction of SONAR-predicted RBPs, unlike the other defined lists. This strongly suggests that sequence composition alone may yield a high false positive rate when annotating RBPs. We also observed that 30% of known transcription factors are predicted by SONAR to have RNA-association activity (Figure 3E).

The RCS value for each held-out protein in the BioPlex data was determined as the mean classifier score (or output) over ten iterations of 10-fold cross-validation (Table S4). The RCS distribution for annotated RBPs (that were not used in training of SONAR) was statistically significantly higher (p < 10$^{-16}$, Kolgomorov-Smirnov two-tailed test) compared to non-annotated RBPs, as well as the full set of proteins in the BioPlex data. In

human annotated RBPs by combining the list of 1,072 by Sundararaman et al. (2016) and 1,542 by Gerstberger et al. (2014) (Table S3). Next, we generated "neighborhoods" for all proteins in the human BioPlex PPI data set, which consists of dozens of thousands of interactions (Huttlin et al., 2015). Proteins (RBP and non-RBPs) are represented as nodes in these neighborhoods with edges connecting two proteins if they interact (Figure 3A). Interestingly, known RBPs in the BioPlex data set exhibit statistically significant differences from non-RBPs in terms of network properties. RBPs tend to have higher degree centrality (the number of edges incident upon a node; p < 10$^{-15}$) and higher closeness centrality (distance from the node to all others, p < 10$^{-11}$). RBPs and non-RBPs have approximately similar betweenness centrality (i.e., the number of times the node acts as a bridge along the shortest path between two other nodes, p < 0.002) compared to non-RBPs (Figure S3A). Within each neighborhood, the fractions of annotated RBPs were extracted as features for the SVM classifier (Figure 3B). The classifier was trained using annotated RBPs as positively labeled examples and other (presumably non-RBP) proteins as negatively labeled examples. Performance was evaluated using 10-fold cross-validation. The areas under
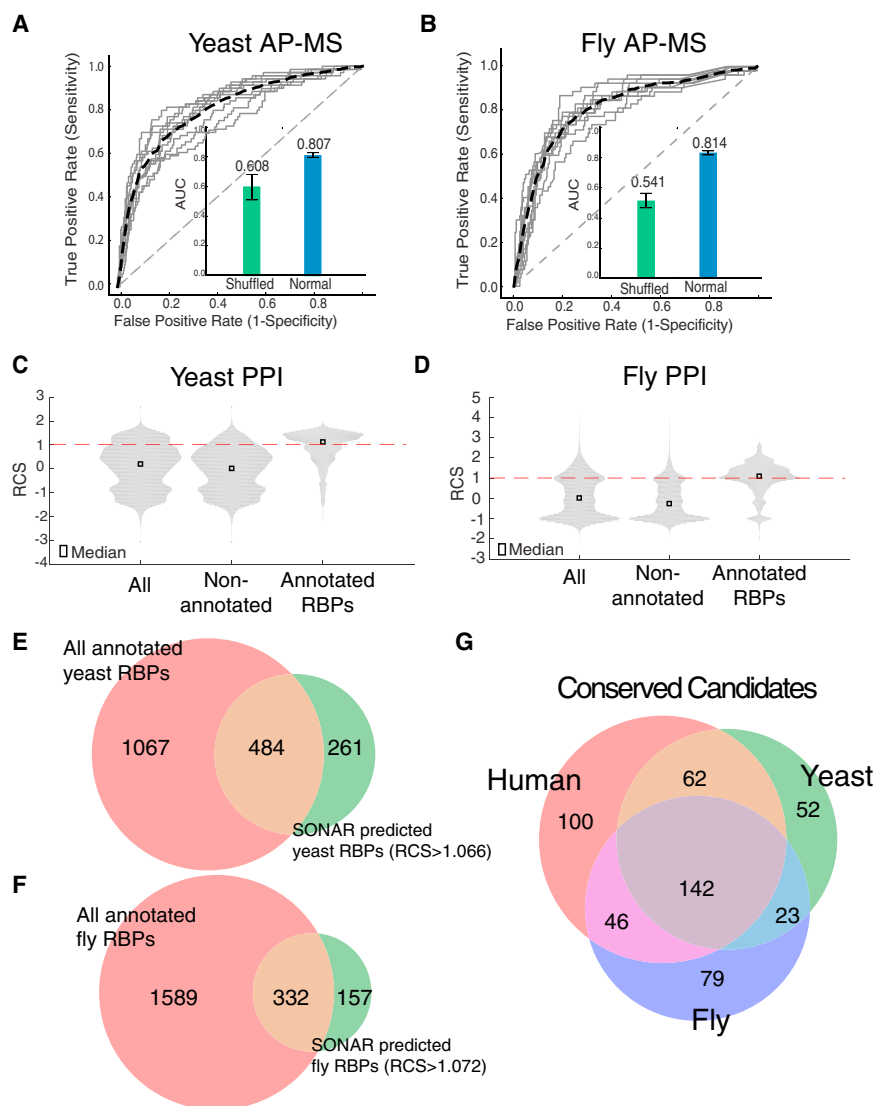
**CellPress**



Figure 4. SONAR RBP Classification Scores Predict Thousands of RBPs Using PPI Networks from Multiple Species

(A) ROC-AUC analysis of classifier performance for yeast (*Saccharomyces cerevisiae*) proteins from BioGRID network. The data are represented as mean ± SEM.

(B) ROC-AUC analysis of classifier performance for fly (*Drosophila melanogaster*) proteins from BioGRID network. The data are represented as mean ± SEM.

(C) Violin plots of RCS distributions for all yeast BioGRID interactors, non-RBP interactors, and annotated RBP interactors within yeast BioGRID interactors.

(D) Violin plots of RCS distributions for all fly BioGRID interactors, non-RBP interactors, and annotated RBP interactors within BioGRID interactors.

(E) Venn diagram showing overlap between all annotated yeast RBPs and SONAR predicted yeast RBPs (RCS > 1.066, threshold for false positive rate 0.1).

(F) Venn diagram showing overlap between all annotated fly RBPs and SONAR predicted fly RBPs (at RCS > 1.072, threshold for false positive rate 0.1).

(G) Venn diagram showing overlap between conserved high RCS scoring predicted RBPs in human (light red), in yeast (green), and in fly (purple).

candidate RBPs exhibited higher iso-electric points (p $< 10^{-6}$, Figure S4A), higher proportions of residues in disordered regions (p $< 10^{-5}$, Figure S4B), and higher amino acid compositions reflective of RNA binding (p $< 10^{-5}$, Figure S4C). These features were also similar to annotated RBPs (red lines, Figure S4). We did not detect any appreciable differences in low-complexity regions (Figure S4D) or total protein size (Figure S4E). In summary, SONAR leverages publically available PPI networks to predict proteins that have attributes of RBPs.

## SONAR Identifies Hundreds of Previously Unannotated RBPs across Multiple Species

The flexibility of SONAR was next evaluated on available yeast (*Saccharomyces cerevisiae*) and fly (*Drosophila melanogaster*) protein-protein interactome data sets. Training and testing for the classifiers followed the same procedure as for the human data set. Interestingly, we found that SONAR performed very well (AUC > 0.8) yeast (Figure 4A) and fly (Figure 4B) PPI data sets generated by affinity precipitation followed by mass spectrometry (AP-MS), similar to BioPlex. Performance metrics are provided for human, yeast, and fly in Figure S3C (chosen to achieve a false discovery rate of 10%). However, SONAR performs weakly using PPI data sets from yeast two-hybrid (Y2H) methodology (Figure S3D). As Y2H relies on bait and prey
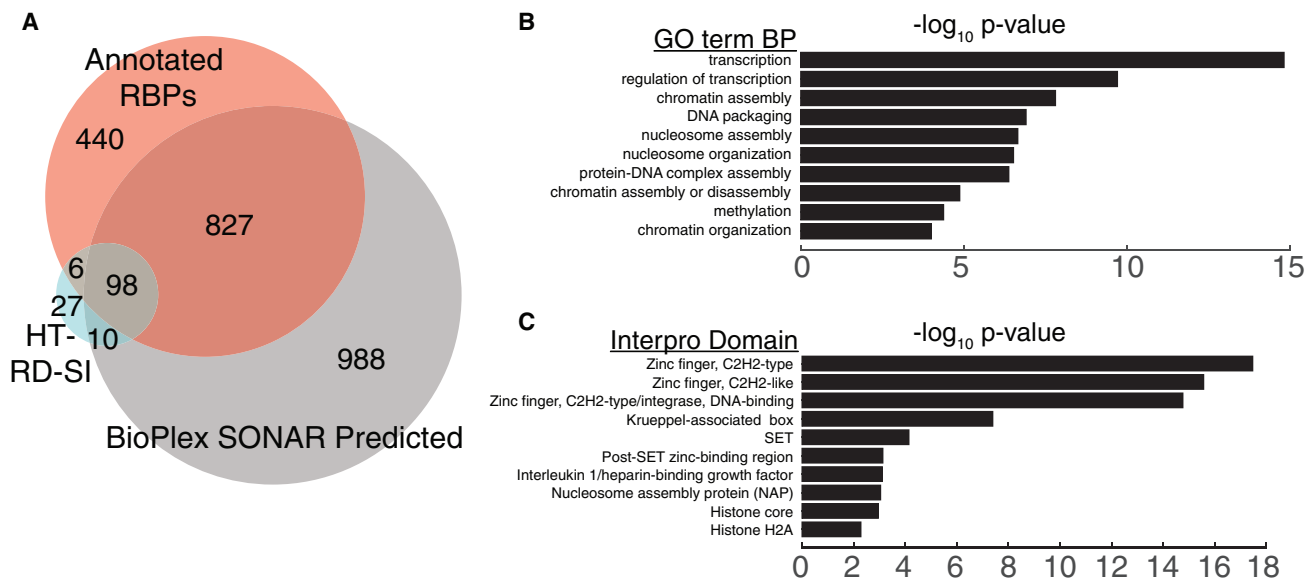
fact, ~70% of annotated RBPs had an RCS of greater than our user-defined SONAR threshold of 0.79, whereas only 10% of all unannotated RBPs scored above threshold (Figure 3F). Interestingly, 30% of the complete set of HaloTag interactors (HTI) that were also present in the BioPlex data set scored above 0.79 (Figure 3F). The complete set of HaloTag-RBP-super interactor proteins had a score distribution similar to that for annotated RBPs. Pertinently, SI proteins that were extracted from the RNA-dependent (HT-RD-SI) interactions had significantly higher RCS values than SI proteins obtained from the RNA-independent (HT-RI-SI) interactors (Figure 3F).

Next, we evaluated attributes that reflect biophysical properties thought to be associated with known RBPs (Castello et al., 2012; Lunde et al., 2007). We compared annotated RBPs and candidate proteins previously unannotated as RBPs, but predicted by SONAR to be RNA-associated (RCS > 0.79) (blue lines, Figure S4). In comparison to all proteins in the BioPlex network (dashed lines, Figure S4A), we found that SONAR-predicted

**Figure 5. SONAR Predicts Human Candidate RBPs Enriched for Proteins with Zinc Finger and DNA Binding Domains**
(A) Venn diagram showing overlap between all HT-RBP RNA-dependent SI proteins contained in the BioPlex network (light blue), all annotated RBPs contained in the BioPlex network (red), and all BioPlex SONAR predicted RBPs (at RCS > 0.79; gray).
(B) Bar graph displaying -$\log_{10}$ p values for GO Biological Process (BP) terms enriched in the set of RBP candidates (RCS > 0.79 and not previously annotated as RBPs) compared to all interactors within the BioPlex data set.
(C) Bar graph displaying -$\log_{10}$ p values for InterPro protein domains enriched in the set of RBP candidates (RCS > 0.79 and not previously annotated as RBPs) compared to all interactors within the BioPlex data set.

proteins to interact, bringing together the activation domain and binding domain of a transcription factor to activate a reporter gene, we expect that when these bait and preys are RBPs, they are not directly bound to RNA molecules. This result supports the concept that AP-MS experiments identifies endogenous RBP interactors, as they are more likely to be in their normal cellular context of binding RNA. As with RCS distributions derived from the human BioPlex network, the yeast (Figure 4C) and fly (Figure 4D) RCS distributions for annotated RBPs was statistically significantly higher (p < $10^{-16}$ by Kolgomorov-Smirnov two-tailed test) compared to non-annotated RBPs, as well as the full set of proteins in the BioGRID data.

SONAR candidate RBPs overlap with 31% of annotated yeast (Figure 4E) and 17% of annotated fly RBPs (Figure 4F). Here, we defined annotated yeast and fly RBPs using the union of human orthologs and proteins identified from interactome-capture studies (Figures S5A and S5B for yeast and S5C for fly). Performance for SONAR was also robust when trained using RBP definitions solely derived from interactome-capture or only from computationally defined orthologs (Figure S5E). Importantly, 350 human, 290 fly, and 279 yeast proteins that were predicted by SONAR as candidate RBPs had assignable orthologs, with 142 SONAR predicted candidate RBPs conserved across all three species (Figure 4G). These conserved candidate RBPs are involved in protein folding, chromosome organization, and nuclear import and include HT-RBP SI proteins (Table S4). In all, these results demonstrate that SONAR identifies previously unannotated and conserved RBPs across multiple species without explicit sequence and structure homology information.

## Candidate RBPs Are Enriched for Proteins with Zinc Finger and DNA Binding Domains

The set of BioPlex proteins that SONAR predicts as RBPs (RCS > 0.79) overlaps with ~70% of the list of annotated RBPs contained in BioPlex and ~80% of HT-RD SI proteins from the HaloTag PPI network (Figure 5A). There are 998 SONAR candidate RBPs that have no previous annotation as RBPs, ten of which are HT-RD-SI RBP candidates (Figure 5A). We performed GO analysis to determine what categories of proteins are enriched in SONAR predicted RBPs aside from the expected RNA processing categories. Intriguingly, for previously unannotated SONAR predicted RBPs, we found significant enrichment for GO terms involved in transcription, chromatin modification, chromosome organization, DNA binding, and nucleosome assembly (Figure 5B; Table S5), and protein domains contained by BioPlex candidate RBPs are significantly enriched for DNA binding domains, with zinc finger binding domains being the most enriched (Figure 5C; Table S5). We also found that SONAR predicted RBPs overlapped with 27% of the newly identified non-canonical RBPs termed "enigmRBPs" (Beckmann et al., 2015), as well as ~68% of recently reported DNA-RBPs (Conrad et al., 2016) (Figure S5D). In fact, the concordance with the latter study is also consistent with our recent data generated as part of the ENCODE consortium. We identified ten previously annotated DNA binding proteins that SONAR predicted as RBPs (RCS > 0.79), which interact directly with endogenous RNA by eCLIP (Table S5; publicly available data sets at https://www. encodeproject.org). We believe that the large set of SONAR predicted RBP candidates (998) would significantly expand the

existing repertoire of RNA-associated proteins upon further experimental confirmation.

### CLIP and eCLIP Validate RNA Binding Activity and Discover RNA Targets of Candidate RBPs

We next sought to determine SONAR's ability to predict RNA binding activity in vivo for a selection of HT-RD-SI proteins with SONAR RCS values between −1.3 and +1.7. We transiently transfected HEK293T cells with plasmids expressing V5-tagged fusions of a selection of candidate RBPs. Cells were UV irradiated and RBP-RNA complexes were immunoprecipitated using commercially available V5 antibody and T4 PNK radiolabeled, separated on an SDS-PAGE gel, and transferred to a nitrocellulose membrane for western blot and phosphor-image analysis. Negative controls were either non-transfected or expressed V5-tagged GFP, while positive controls expressed the V5-tagged RBPs HNRNPM and HNRNPUL1. We found that seven out of eight V5-tagged RBPs tested efficiently bound radiolabeled RNA in a UV-cross-linking-dependent manner (Figure S6A; Supplemental Information). In contrast, we did not detect radiolabeled RNA immunoprecipitated from cells expressing V5-tagged GFP in either condition or from non-transfected cells. This assay validates SONAR predicted (RCS > 0.79) at 100%, but we note here that 3 HT-RD-SI proteins (AIFM1, VIM, and SMU1) with low SONAR scores also detectably bind RNA above negative controls by this assay, indicating a false negative rate for SONAR (Figure S6A).

Next, we determined the transcriptome-wide binding sites of a subset of these RBP candidates and included the zinc-finger protein of unknown function ZNF184 using enhanced CLIP. Briefly, HEK293T were subjected to UV-mediated cross-linking, lysis, and treatment with limiting amount of RNase, followed by immunoprecipitation (IP) of protein-RNA complexes using commercially available antibodies that interrogate the endogenous proteins. RNA fragments protected from RNase digestion by RBP occupancy were subjected to 3′ RNA linker ligation, reverse-transcription, and 3′ DNA linker ligation to generate eCLIP libraries for high-throughput Illumina sequencing. The improved efficiency of eCLIP enabled the generation of a size-matched input (SMInput) library for each biological sample, in which 2% of the pre-immunoprecipitation sample was subjected to identical library generation steps including ribonuclear protein complex size selection on nitrocellulose membranes. In total, 42 eCLIP (including SMInput) libraries were sequenced to ∼seven million reads each, of which ∼30%–70% mapped uniquely to the human genome (Table S6).

We demonstrate that these candidate RBPs bind RNAs in different genic regions (Figure 6A) and recognize different binding motifs (Figure 6B). The RAN-GTPase activating protein RANGAP1 showed preferential enrichment in 5′ UTRs and intron-less genes (Figures 6A and 6C). RANGAP1 localizes at the nuclear periphery and is important for import of cargo through nuclear pore complexes. Disrupted nuclear import caused by RANGAP1 inhibition through interaction with C9orf72 hexanucleotide repeat expansion RNA (HREs) is involved in ALS pathology (Zhang et al., 2015). Interestingly, we find that the motif (CGGCGG) enriched by RANGAP1 eCLIP is similar to the $G_4C_2$ pattern responsible for HRE G-quadra-

plexes (Figure 6B). The mitotic spindle organizing protein NUMA1 preferentially bound to intronic regions including its own pre-mRNA (Figures 6A and 6D). The ring-type zinc finger protein RNF219, previously implicated in Alzheimer's disease, preferentially bound 3′ UTRs (Figures 6A and 6E). The zinc finger containing DNA binding protein ZNF184 had enriched clusters over all types of transcript regions, including introns such as the distal intron of the CENPM gene (Figures 6A and 6F). In all, SONAR-predicted RBPs indeed interact with hundreds to thousands of enriched binding sites in the human transcriptome.

CLIP cluster discovery was performed using the CLIPper (Lovci et al., 2013) algorithm. SI proteins that had low RCS values, such as AIFM and VIM, had fewer significantly enriched clusters (Figure S6C). Predicted RBPs with high RCS values such as NUMA1, RANGAP1, RNF219, and ZNF184 had far more enriched clusters and bound transcripts in specific regions (Figures 6A and S6C). Interestingly, SONAR RCS values appear positively correlated with the fraction of interactions with HT-RBPs that are RNA-dependent ($R^2$ = 0.72) (Figure S6B) and with the number of identified eCLIP clusters ($R^2$ = 0.42) (Figure S6C). Overall, these findings demonstrate that the SONAR algorithm is effective at predicting bona-fide RBPs.

### DISCUSSION

The increasing importance of RBPs in development and disease has accelerated the need to comprehensively, rapidly, and accurately identify new components of ribonuclear particles in a variety of organisms. Previous computational approaches to define the complete catalog of RBPs rely heavily on the presence of RNA binding domains (Baltz et al., 2012; Gerstberger et al., 2014). These approaches lack sensitivity given the growing number of interactome-capture studies that identify RBPs that lack characterized RBDs (Beckmann et al., 2015; Conrad et al., 2016). However, interactome-capture studies of polyA-selected mRNPs may not identify proteins interacting with nascent or pre-mRNA and interactome capture followed by mass spectrometry is a technically demanding technique, precluding its adoption for the vast majority of researchers.

In this study, we have developed a simple yet powerful computational approach to discover RBPs based on the realization that without RNAase treatment, pull down of RNA-associated proteins enriches for interactors that are directly bound to RNA. Therefore, the higher the number of interacting RBPs, the more likely a previously unannotated protein is itself an RBP. As a pilot study, we have independently identified RBP-interactomes for a diverse set of RBPs, namely a splicing factor (RBFOX2), a decay regulator (UPF1), and ten major HNRNP proteins using a HaloTag purification approach followed by mass spectrometry analysis. Together, these 12 proteins are localized in the nucleus and cytoplasm and are ideal for identifying known and previously unannotated RBPs bound to both nuclear and cytoplasmic RNA substrates. By comparing RNase-treated and RNase-untreated conditions, we were able to uncover thousands of RNA-dependent and RNA-independent interactions. Interactors include many unexpected proteins involved in diverse biological processes including oxidative phosphorylation, proteasome function, energy production, mitochondrial
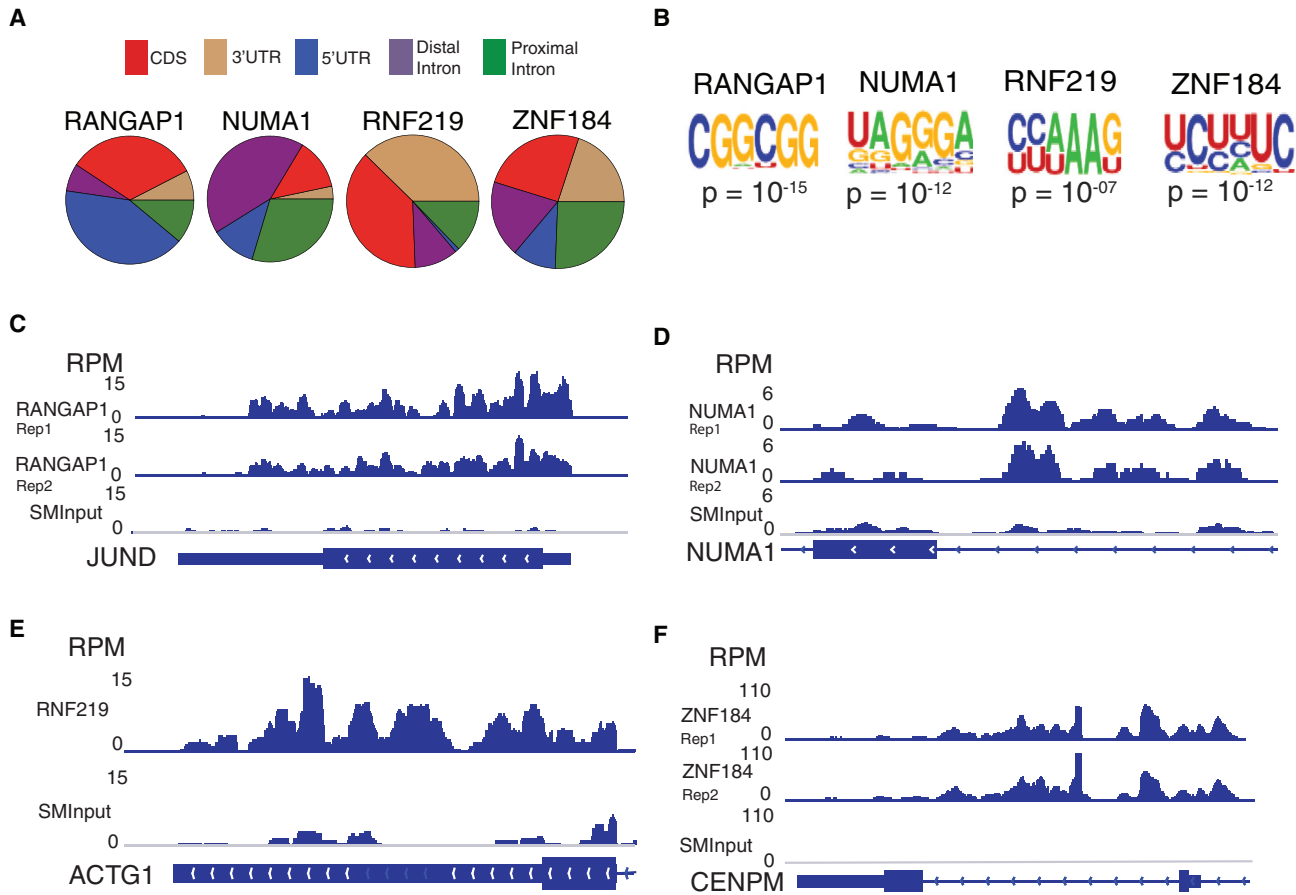
**Figure 6. Enhanced CLIP Validation of Candidate RBPs Predicted by HT-RBP Interactome and SONAR Classification**

(A) Distributions across transcript regions for peaks enriched >8-fold over size-matched input ($-\log_{10}$ p > 5) from eCLIP experiments for four RBP candidates.

(B) Motifs called and p values for input normalized peaks described in Figure 5D.

(C) Genome browser track view of RANGAP1 eCLIP data in reads per million (RPM) showing enrichment above input on the intronless JUND gene locus.

(D) Genome browser track view of NUMA1 eCLIP data in reads per million (RPM) showing enrichment above input on a NUMA1 intron.

(E) Genome browser track view of RNF219 eCLIP data in reads per million (RPM) showing enrichment above input on the ACTG1 3′ UTR region.

(F) Genome browser track view of ZNF184 eCLIP data in reads per million (RPM) showing enrichment above input on the CENPM distal intron.

organization, and Golgi vesicle transport. Importantly, this resource confirms that proteins that interact with multiple HT-RBPs are enriched for RBPs.

As there are many large-scale PPI data sets available that are not subject to RNase treatment, we developed an algorithm termed SONAR that leverages the neighborhood of RBP interactors in PPI networks to identify unannotated RBPs. SONAR does not rely on sequence or structural homology or modeling which historically prevented the identification of proteins that have RNA binding activity through yet unknown mechanisms. Instead, SONAR relies on experimentally determined interaction information from thousands of independently generated affinity purification experiments, which in aggregate is more stringent than replicates of a few to a dozen precipitation experiments. Furthermore, as we demonstrated, SONAR can be readily applied to PPI data sets from multiple organisms, which led to our identification of previously unannotated, evolutionarily conserved RBPs. Interestingly, many of these candidate RBPs have zinc finger binding

domains and are involved in processes such as transcription, chromosome organization, and chromatin modification, in support of recent predictions that there are likely RNA binding roles for DNA binding proteins (Conrad et al., 2016). While it is possible that this observation is the result of direct protein-protein interactions between DNA binding protein baits and nuclear RBPs, our results also suggest the possibility that many DNA binding proteins likely also interact with RNA during transcription or RNA processing. The second possibility is consistent with our own eCLIP studies performed as part of the ENCODE consortium, which show that ten high SONAR-scoring proteins previously reported to have DNA binding and transcription factor activity can also interact directly with endogenous RNA targets.

Lastly, we validated SONAR predictions of RNA binding activity by eCLIP for a set of RBP candidates. Surprisingly, for six candidate eCLIP validations, higher SONAR scores correlated with higher numbers of eCLIP-identified binding sites, with VIM and AIFM libraries having the fewest peaks above size-matched

input controls, and NUMA1, RANGAP1, and ZNF184 having many more clusters enriched above size-matched input controls. In retrospect, this interesting observation is consistent with our features utilized by SONAR. It is reasonable to imagine that the higher the fraction of known RBPs within the neighborhoods of a candidate RBP, the higher the number of different RNA molecules bound by various RBPs (hence more binding sites) that the protein should interact with.

Among our validated examples of previously unannotated RNA-associated proteins, the ring-type zinc finger protein RNF219 preferentially bound 3′ UTRs, which is intriguing in light of its reported protein-protein interactions with components of the deadenylation machinery (Hein et al., 2015; Li et al., 2014), and it will be interesting in the future to determine whether RNF219 affects stability of its RNA targets. The mitotic spindle organizing protein NUMA1 preferentially bound introns, and considering the strong RNA independent interaction of NUMA1 with UPF1 and RBFOX2 baits, it is possible that NUMA1 plays a role in alternative splicing or nonsense mediated decay of its targets. The RAN-GTPase activating protein RANGAP1 preferential bound many 5′ UTRs. Disruption of RANGAP1 mediated mRNA export and localization could be a possible mechanism underlying C9orf72 HRE related ALS pathology (Zhang et al., 2015), and in future experiments, it is important to determine if C9orf72 HRE disruption of RANGAP1 RNA binding may lead to mRNA export/import inhibition and altered stability or translation for RANGAP1 mRNA targets. The DNA binding protein ZNF184 bound transcripts in introns, UTRs, and coding sequences. Given the recent finding that mRNA fate can be determined by promoter regions alone (Zid and O'Shea, 2014), one interesting explanation of this phenomenon may be that RNA binding activity of promoter specific transcription factors and chromatin remodelers influences transcript localization, and the group of transcription factors predicted to bind RNA in this study could serve as valuable candidates for testing this hypothesis.

In conclusion, the SONAR algorithm is a powerfully simple and effective method for the de novo identification of candidate RBPs. As with many methods, the strength of SONAR's approach, which lies on its ability to leverage large-scale PPI data sets, is also a potential weakness. First, SONAR relies on the coverage of these PPI networks, meaning not all proteins have been subjected to affinity precipitation followed by mass spectrometry. Second, SONAR relies on the quality of the independent data sets within these PPI networks. Finally, SONAR requires definitions of known or annotated RBPs to generate its scores for uncharacterized proteins. Fortunately, as more interactome-capture data sets are generated and more SONAR-predicted RBPs are validated, we can build an increasingly rigorous data set of known RBPs. Also, efforts to improve the quality and coverage of PPI data sets are underway. Nevertheless, in this study, the RBP interactomes presented here represent a valuable first step in building a comprehensive resource for the identification of RNA-dependent and RNA-independent interactors of RBPs. Also, future studies verifying the large set of classifier predicted RBPs would expand the list of RBPs to include many unexpected classes of proteins involved in diverse biological processes.

## EXPERIMENTAL PROCEDURES

### HaloTag Affinity Purification

HaloTag affinity purification was performed as follows, with and without RNase treatment in replicate (2 μl RNase A; A797C 4 mg/mL for each 12 million cell pellet lysate). Protein complexes were covalently captured on HaloLink resin pre-equilibrated in resin wash buffer (Tris-buffered saline [TBS] and 0.05% IGEPAL CA-640; Sigma) for 15 min at 22°C with rotation. Resin was then washed five times with wash buffer to remove non-specific interactions and protein interactors were eluted in 8 M urea. Additional control experiments with HaloTag only were also performed with and without RNase digestion in replicate. Affinity purified complexes were than analyzed by MudPIT as previously described (Florens and Washburn, 2006).

### Mass Spectrometry Data Analysis

Spectral counting was performed and NSAF determined (dNSAF values) as previously described (Zhang et al., 2010). False positives/contaminants were removed by comparing to control experiments. Any potential protein interactor with a dNSAF value that was not 2× greater than its value in any of the control corresponding experiments was assigned a 0 value.

For equivalent comparison to the RBP RNA targets, the lists of protein-protein interactors were filtered for those targets that are expressed at a detectable RNA level in HEK293T cell RNA-seq data (Huelga et al., 2012). Replicates were clustered, outlier data sets (UPF1_RNASE_2, M_RNASE_1, H_RNASE_1, D0_1) were removed, and the remaining replicates were averaged together. An enrichment score was calculated for each protein interactor by calculating a $\log_2$ ratio of the dNSAF value in the RBP purification experiment compared to the control experiment:

$$\text{Enrichment Score} = \log_2[(\text{dNSAF RBP} + 0.000001)/(\text{dNSAF Control} + 0.000001)]$$

A pseudocount of 0.000001 representing the lowest measurable dNSAF value was added within the enrichment score to prevent losing targets with zero values in the control experiment (divide by zero errors). The final set of protein-protein interactors was determined using all protein interactors with enrichment greater than the $\mu-1.5\sigma$ in each data set.

RNA-dependent interactions were defined as: ((no_RNase enrichment score) − (RNase enrichment score)) / (no_RNase enrichment score) > 0.3. Post lysis interactions were filtered and defined as: ((no_RNase enrichment score) − (RNase enrichment score)) / ((no_RNase enrichment score) + (RNase enrichment score)) < −0.2. All other interactions were defined as RNA independent.

### PPI Network Construction and Neighborhood Features Generation

All the protein-protein interaction networks were represented as an undirected and unweighted graph G = (V, E). Each vertex v ∈ V refers to a protein and each edge $(v, u) \in$ E displays an interaction between two proteins, $v$ and $u$. The definition of level-k is described in Sela et al. (2012): briefly, $u$ is $v$'s level-k neighbor when there exists a simple path of length $k$ from $v$ to $u$. It should be noted that there may be more than one simple path between $v$ and $u$, thus $u$ can be $v$'s level-$k_1$ and level-$k_2$ ($k_1 \neq k_2$) simultaneously. Additionally, if there are $n$ (n > 1) paths of $k$ from $v$ to $u$, the protein $u$ will be counted $k$ times in the feature calculation of level-k neighbors.

In this study, we only consider the given protein's first three levels of neighbors (shown in Figure 3A), and an example of the feature generation process is displayed in Figure 3B. With the three features (calculations described in Supplemental Information), we obtain a feature vector for each protein of the PPI network. We constructed a human PPI network using interaction data from the BioPlex project (Huttlin et al., 2015). For each protein in this network, a PPI feature was generated with the approach above. PPI networks for two other model organisms (Saccharomyces cerevisiae and Drosophila melanogaster) were constructed with the interaction data downloaded from BioGRID 3.4 database (Stark et al., 2006). For each species, two PPI networks were assembled with the data from AP-MS and yeast-two-hybrid experiments, respectively. The 3-level PPI features were generated using the same calculation above for each protein in the PPI networks.

**Cell**Press

## Classifier Building and Evaluation

Classifiers were trained to recognize RBPs using the SVM classification model with a radial basis function (RBF) kernel as implemented within the scikit-learn package in Python (Pedregosa et al., 2011). We divided the proteins in each PPI network into two categories, RBPs and non-RBPs, based on the RBP annotation lists as described in the main text. Each protein is represented by a feature vector and a class label (i.e., RBP or non-RBP). We trained the classifiers and evaluated their performance using ROC-AUC and PRC-AUC analysis with 10-fold cross-validation. The closer the classifier's ROC curve to the upper left corner, the better the classifier's performance in both sensitivity and specificity. To train our classifiers with a balanced data set, we oversampled RBP samples in the training step. During testing, each unseen protein obtained a classification score from a trained classifier. We generated classification scores for every protein in the PPI network during the cross-validation training/testing process and repeated ten times to get an average score for each protein.

## eCLIP

eCLIP was performed as previously described (Van Nostrand et al., 2016). Briefly, for each experiment, 20 million HEK293T cells were UV cross-linked (254 nM, 400 mJ/cm2) and lysed on ice and lysates were sheared with RNase I (Ambion). RBP-RNA complexes were inmmunoprecipitated with antibodies (see Supplemental Information) specific to candidates of interest (Protein G sheep anti-rabbit Dynabeads) and immunoprecipitated material was stringently washed. Dephosphorylation with FastAP (Thermo Fisher) and T4 PNK (NEB) are followed by on-bead ligation of barcoded RNA adapters to the 3′ end (T4 RNA Ligase, NEB). RNA-protein complexes are run on standard protein gels and transferred to nitrocellulose membranes, and the region 75 kDa (∼150 nt of RNA) above the protein molecular weight is excised and proteinase K (NEB) treated to isolate RNA. Reverse transcription is carried out using Affinityscript (Agilent), followed by treatment with ExoSAP-IT (Affymetrix) to remove excess oligonucleotides. A DNA randomer adaptor is then ligated to the cDNA fragment 3′ end (T4 RNA Ligase, NEB). After cleanup (Dynabeads MyOne Silane, Thermo Fisher), samples were first subjected to qPCR to determine final PCR cycle number, and then PCR amplified (Q5, NEB) and size selected via agarose gel electrophoresis. Samples were sequenced on the Illumina HiSeq 2500 platform as two Paired End 50 bp (for N5) or 55 bp (for N10) reads.

## ACCESSION NUMBERS

The accession number for the eCLIP data reported in this paper is GSE86035. The complete MudPIT data sets can be obtained from ftp://massive.ucsd.edu/ using the following MassIVE accession numbers for the MOCK and Halotag RBP data sets with password (KBGWY39029): MSV000079668 and MSV000079669, as well as at ProteomeXchange: PXD003999 and PXD004000.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2016.09.003.

## AUTHOR CONTRIBUTIONS

Conceptualization, K.W.B., W.J., S.C.H., D.L.D., and G.W.Y.; Investigation, K.W.B., W.J., S.C.H., D.L.D., M.P.W., C.A.S.B., J.M.G., L.F., and M.K.S.; Validation, K.W.B., D.L.D., and M.P.W.; Formal Analysis, K.W.B., W.J., S.C.H., E.L.V.N., M.P.W., and G.A.P.; Writing - Original Draft, K.W.B., W.J., S.C.H., and G.W.Y.; Writing - Review & Editing, K.W.B., W.J., S.C.H., D.L.D., M.P.W., and G.W.Y.; Funding Acquisition, G.W.Y.; and Supervision, G.W.Y.

## ACKNOWLEDGMENTS

## REFERENCES

Baltz, A.G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., et al. (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol. Cell 46, 674–690.

Beckmann, B.M., Horos, R., Fischer, B., Castello, A., Eichelbaum, K., Alleaume, A.M., Schwarzl, T., Curk, T., Foehr, S., Huber, W., et al. (2015). The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. Nat. Commun. 6, 10127.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., et al. (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell 149, 1393–1406.

Castello, A., Fischer, B., Hentze, M.W., and Preiss, T. (2013). RNA-binding proteins in Mendelian disease. Trends Genet. 29, 318–327.

Conrad, T., Albrecht, A.S., de Melo Costa, V.R., Sauer, S., Meierhofer, D., and Ørom, U.A. (2016). Serial interactome capture of the human cell nucleus. Nat. Commun. 7, 11212.

Daniels, D.L., Méndez, J., Mosley, A.L., Ramisetty, S.R., Murphy, N., Benink, H., Wood, K.V., Urh, M., and Washburn, M.P. (2012). Examining the complexity of human RNA polymerase complexes using HaloTag technology coupled to label free quantitative proteomics. J. Proteome Res. 11, 564–575.

Deplus, R., Delatte, B., Schwinn, M.K., Defrance, M., Méndez, J., Murphy, N., Dawson, M.A., Volkmar, M., Putmans, P., Calonne, E., et al. (2013). TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. EMBO J. 32, 645–655.

Florens, L., and Washburn, M.P. (2006). Proteomic analysis by multidimensional protein identification technology. Methods Mol. Biol. 328, 159–175.

Flury, V., Restuccia, U., Bachi, A., and Muhlemann, O. (2014). Characterization of phosphorylation- and RNA-dependent UPF1 interactors by quantitative proteomics. J. Proteome Res. 13, 3038–3053.

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

Hein, M.Y., Hubner, N.C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I.A., Weisswange, I., Mansfeld, J., Buchholz, F., et al. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell 163, 712–723.

Huelga, S.C., Vu, A.Q., Arnold, J.D., Liang, T.Y., Liu, P.P., Yan, B.Y., Donohue, J.P., Shiue, L., Hoon, S., Brenner, S., et al. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep. 1, 167–178.

Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., et al. (2015). The BioPlex network: A systematic exploration of the human interactome. Cell 162, 425–440.

Klass, D.M., Scheibe, M., Butter, F., Hogan, G.J., Mann, M., and Brown, P.O. (2013). Quantitative proteomic analysis reveals concurrent RNA-protein

interactions and identifies new RNA-binding proteins in Saccharomyces cerevisiae. Genome Res. 23, 1028–1038.

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

Lee, S.R., Pratt, G.A., Martinez, F.J., Yeo, G.W., and Lykke-Andersen, J. (2015). Target discrimination in nonsense-mediated mRNA decay requires Upf1 ATPase activity. Mol. Cell 59, 413–425.

Li, S., Wang, L., Fu, B., Berman, M.A., Diallo, A., and Dorf, M.E. (2014). TRIM65 regulates microRNA activity by ubiquitination of TNRC6. Proc. Natl. Acad. Sci. USA 111, 6970–6975.

Lovci, M.T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T.Y., Stark, T.J., Gehman, L.T., Hoon, S., et al. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat. Struct. Mol. Biol. 20, 1434–1442.

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.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

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.

Sela, D., Chen, L., Martin-Brown, S., Washburn, M.P., Florens, L., Conaway, J.W., and Conaway, R.C. (2012). Endoplasmic reticulum stress-responsive transcription factor ATF6α directs recruitment of the mediator of RNA polymerase II transcription and multiple histone acetyltransferase complexes. J. Biol. Chem. 287, 23035–23045.

Si, J., Cui, J., Cheng, J., and Wu, R. (2015). Computational prediction of RNA-binding proteins and binding sites. Int. J. Mol. Sci. 16, 26303–26317.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34, D535–D539.

Sundararaman, B., Zhan, L., Blue, S.M., Stanton, R., Elkins, K., Olson, S., Wei, X., Van Nostrand, E.L., Pratt, G.A., Huelga, S.C., et al. (2016). Resources for the comprehensive discovery of functional RNA elements. Mol. Cell 61, 903–913.

Sysoev, V.O., Fischer, B., Frese, C.K., Gupta, I., Krijgsveld, J., Hentze, M.W., Castello, A., and Ephrussi, A. (2016). Global changes of the RNA-bound proteome during the maternal-to-zygotic transition in Drosophila. Nat. Commun. 7, 12128.

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, 5.

Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods 13, 508–514.

Wessels, H.H., Imami, K., Baltz, A.G., Kolinski, M., Beldovskaya, A., Selbach, M., Small, S., Ohler, U., and Landthaler, M. (2016). The mRNA-bound proteome of the early fly embryo. Genome Res. 26, 1000–1009.

Zhang, Y., Wen, Z., Washburn, M.P., and Florens, L. (2010). Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. Anal. Chem. 82, 2272–2281.

Zhang, K., Donnelly, C.J., Haeusler, A.R., Grima, J.C., Machamer, J.B., Steinwald, P., Daley, E.L., Miller, S.J., Cunningham, K.M., Vidensky, S., et al. (2015). The C9orf72 repeat expansion disrupts nucleocytoplasmic transport. Nature 525, 56–61.

Zid, B.M., and O'Shea, E.K. (2014). Promoter sequences direct cytoplasmic localization and translation of mRNAs during starvation in yeast. Nature 514, 117–121.