Systematic discovery of endogenous human ribonucleoprotein complexes

Anna L. Mallam[1,2,3,§,*], Wisath Sae-Lee[1,2,3], Jeffrey M. Schaub[1,2,3], Fan Tu[1,2,3], Anna Battenhouse[1,2,3], Yu Jin Jang[1], Jonghwan Kim[1], Ilya J. Finkelstein[1,2,3], Edward M. Marcotte[1,2,3,§], Kevin Drew[1,2,3,§,*]

[1] Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA
[2] Center for Systems and Synthetic Biology, University of Texas at Austin, Austin, TX 78712, USA
[3] Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, TX 78712, USA
[§] Correspondence: amallam@utexas.edu (A.L.M.), marcotte@icmb.utexas.edu (E.M.M.), kdrew@utexas.edu (K.D.)

* These authors contributed equally to this work

Short title: A resource of human ribonucleoprotein complexes

Summary: An exploration of human protein complexes in the presence and absence of RNA reveals endogenous ribonucleoprotein complexes

Abstract

Ribonucleoprotein (RNP) complexes are important for many cellular functions but their prevalence has not been systematically investigated. We developed a proteome-wide fractionation-mass-spectrometry strategy called differential fractionation (DIF-FRAC) to discover RNP complexes by their sensitivity to RNase A treatment. Applying this to human cells reveals a set of 115 highly-stable endogenous RNPs, and a further 1,428 protein complexes whose subunits associate with RNA, thus indicating over 20% of all complexes are RNPs. We show RNP complexes either dissociate, change composition, or form stable protein-only complexes upon RNase A treatment, uncovering the biochemical role of RNA in complex formation. We combine these data into a resource, rna.MAP (rna.proteincomplexes.org), which demonstrates that well-studied complexes such as replication factor C (RFC) and centralspindlin exist as RNP complexes, providing new insight into their cellular functions. We apply our method to red blood cells and mouse embryonic stem cells to demonstrate its ability to identify cell-type specific roles for RNP complexes in diverse systems. Thus the methodology has the potential uncover RNP complexes in different human tissues, disease states and throughout all domains of life.

Keywords: ribonucleoprotein complex, RNP, RNA binding protein, RBP, proteomics, DIF-FRAC, protein complexes, biochemical fractionation, mass spectrometry, interactome

Introduction

Large macromolecular complexes are crucial to many essential biochemical functions and their full characterization is necessary for a complete understanding of the cell. A worldwide effort is underway to systematically identify complexes using high throughput mass spectrometry techniques across many cell types, tissues and species (Havugimana et al., 2012; Hein et al., 2015; Wan et al., 2015; Drew et al., 2017; Huttlin et al., 2017) but these techniques currently only consider the protein subunits of complexes ignoring other constituent biomolecules. Ribonucleoprotein (RNP) complexes, a specific subclass of complexes consisting of RNA and protein (Castello et al., 2013; Gerstberger et al., 2014; Hentze et al., 2018), are particularly important to study due to their indispensable role in cellular functions such as translation (ribosome), splicing (spliceosome), and RNA degradation (exosome), as well as their critical role in human diseases including amyotrophic lateral sclerosis (ALS) (Scotter et al., 2015), spinocerebellar ataxia (Yue et al., 2001), and autism (Voineagu et al., 2011). Unfortunately, we currently lack a full account of all RNP complexes in the cell inhibiting our understanding of vital biological processes. Recent advances in methodology have identified many new RNA-associated proteins, highlighting the importance of protein-RNA interactions throughout the proteome (Baltz et al., 2012; Castello et al., 2012; Brannan et al., 2016; Castello et al., 2016; He et al., 2016; Treiber et al., 2017; Bao et al., 2018; Huang et al., 2018; Queiroz et al., 2018; Trendel et al., 2018) yet none of these methods are capable of directly exploring the prevalence of protein-RNA interactions in the context of macromolecular complexes. Moreover, these techniques rely on crosslinking, modified-nucleotide-incorporation, the use of specific RNA baits, and/or poly(A) RNA-capture, all of which biases their identifications.

To address these limitations and discover the pervasiveness of RNA in macromolecular complexes, we developed an unbiased strategy to systematically discover endogenous RNP complexes. Our method, 'differential fractionation for interaction analysis' (DIF-FRAC), measures the sensitivity of protein complexes to RNase A treatment using native size-exclusion chromatography followed by mass spectrometry. DIF-FRAC is based on a high throughput co-fractionation mass spectrometry (CF-MS) approach that has been applied to a diverse set of tissues and cells types en route to generating human and metazoan protein complex maps (Havugimana et al., 2012; Wan et al., 2015; Drew et al., 2017). DIF-FRAC builds upon CF-MS by comparing chromatographic separations of cellular lysate under control and RNA degrading

conditions (Figure 1A). DIF-FRAC then discovers RNP complexes by identifying concurrent shifts of known protein complex subunits upon RNA degradation.

Using DIF-FRAC we discover a set of 115 highly stable RNP complexes and further generate a system-wide resource of 1,428 protein complexes where the majority of subunits bind RNA, representing 20 % of known human protein complexes. In depth analysis of DIF-FRAC data further shows unprecedented characterization of RNP complexes providing distinct roles for RNA in protein complexes including complex compositions that are RNA-dependent, identification of RNA as peripheral to complex formation, and discerning RNA as a structural component responsible for the stability of the complex.

A distinct advantage of the DIF-FRAC method is its lack of reliance on crosslinking, nucleotide incorporation, genetic manipulation or poly(A) RNA capture efficiency and therefore can be used to investigate a wide variety of cell types, tissues and species. We apply DIF-FRAC to mouse embryonic stem cells (mESCs) and human erythrocytes (red blood cells; RBCs) to show the method is highly adaptable and can be extended to discover RNP complexes in diverse samples, tissue types, and species. Finally, we provide our resource, rna.MAP, to the community as a fully searchable web database at rna.proteincomplexes.org.

Results and Discussion

Differential fractionation (DIF-FRAC) identifies RNP complexes

The DIF-FRAC strategy detects RNP complexes by identifying changes in the elution of a protein complex's subunits upon degradation of RNA (Figure 1). We applied DIF-FRAC to human HEK 293T cell lysate using size-exclusion chromatography (SEC) to separate the cellular proteins in a control and an RNase A-treated sample into 50 fractions (Figure 1A). Upon degradation of RNA, we see a change in the bulk chromatography absorbance signal consistent with higher-molecular-weight species (>1000 kDa) becoming lower-molecular-weight species (Figure 1B). The loss of absorbance signal in the high-molecular-weight region in the absence of RNA suggests this peak corresponds to RNA and RNP complexes. The distribution of cellular RNA in these fractions measured using RNA-seq confirmed we are accessing a diverse RNA landscape of mRNAs, small RNAs, and lncRNAs (Figure S1). As a negative control, we applied DIF-FRAC to human erythrocytes, which have substantially lower amounts of RNA due to the loss of their nucleus and ribosomes upon maturation (Keerthivasan et al., 2011). For this reason,

we expect limited change in the proteome's elution profiles upon RNase A treatment. Accordingly, the absorbance chromatography signal of erythrocyte lysate shows negligible difference in a DIF-FRAC experiment (Figure 1C). Together these data establish that DIF-FRAC is capable of identifying bulk changes to the RNA-bound proteome.

Mass spectrometry was used to identify and quantify the resident proteins in each fraction for both the control and RNase A-treated chromatographic separations, resulting in 8,946 protein identifications. Using the these abundance measurements, we compared elution profiles (i.e. abundance change over chromatographic separation) between the control and RNase A treated experiments for each protein. A difference in a protein's elution profile between experiments is indicative of a protein-RNA interaction. For example, the known RNA helicase DDX21 shows a substantial change in its elution profile upon RNase A treatment (Figure 1D), consistent with DDX21 association with RNA (Calo et al., 2015). Alternatively, the glucose synthesis enzyme PGM1, a protein not known to associate with RNA, shows no shift in agreement with it not binding RNA (Figure 1E).

We can further examine these elution profile differences in context of physically associated proteins to identify RNP complexes. For example, subunits of the spliceosome, a known RNP complex, show elution profiles that coelute in the control but shift markedly upon RNA degradation (Figure 1F). In contrast, subunits of the non-RNA-associated hexameric MCM complex (M_r ~550 kDa) (Figure 1G), as well as the 8-subunit COP9 signalosome (M_r ~500 kDa) (Figure S2A), co-elute regardless of RNase A treatment, consistent with the complexes not interacting with RNA. Thus, DIF-FRAC exhibits a robust signal that can be used to differentiate between non-RNA-associated complexes and RNP complexes.

Systematic identification of RNP complexes

In order to systematically identify RNP complexes in a DIF-FRAC experiment, we developed a computational framework to identify statistically significant changes in a protein's elution behavior. The DIF-FRAC score evaluates the degree to which two chromatographic separations differ (Figure 1H). Briefly, the DIF-FRAC score is a normalized Manhattan-distance between a protein's control and RNase A treated elution profiles (see Methods). To identify significant changes, we calculated P-values by comparing each protein's DIF-FRAC score to an abundance-controlled background distribution of DIF-FRAC scores from non-RNA-associated proteins

(Figure S3A, see Methods for full description). We evaluated the score's performance on a curated set of known RNA-associated proteins and see strong correspondence between precision and high-ranking proteins (Figure 1I, Figure S3B). DIF-FRAC identifies 1012 proteins with significant elution profile differences in HEK 293T cells with a false discovery rate (FDR) cutoff of 5% (Table S1). To evaluate correspondence to known RNA-associated proteins, this set of statistically significant hits was compared to RNA-associated proteins identified using alternative methods indicating that DIF-FRAC is highly accurate (Figure S4A-K). Importantly, DIF-FRAC identifies 196 novel human RNA-associated proteins, indicating we are uncovering previously undiscovered RNA biology (Table S2, Figure S4L). DIF-FRAC identified RNA-associated proteins that are also strongly enriched in RNA binding domains annotated by Interpro (Finn et al., 2017)(Figure S4M).

We identified RNP complexes by searching for known protein complexes that are sensitive to RNase A treatment. We first compiled a global set of 6,271 nonredundant annotated protein complexes from hu.MAP, a resource of protein complexes identified from high throughput mass spectrometry data (Drew et al., 2017), and CORUM, a manually curated repository of experimentally characterized protein complexes from mammalian organisms (Ruepp et al., 2010). We then analyzed DIF-FRAC elution profiles for the protein subunits of each complex in this global set to determine its stability under our separation conditions and its sensitivity to RNase A treatment. In this way, we identify 441 complexes where the majority of subunits have significant DIF-FRAC scores, indicating an RNA component (Table 1 and Table S3). Within this set, we further detect 115 highly-stable RNP complexes whose subunits co-elute in the DIF-FRAC control experiment (i.e. form a stable complex) as well as have high DIF-FRAC scores for the majority of their subunits (see Methods) (Figure 1J). We call these complexes 'RNP Select' (Figure 2 and Table S3)). Finally, to ascertain the total number of annotated protein complexes that likely function with an RNA component, we evaluate all known evidence of RNA-protein interactions in addition to DIF-FRAC evidence and identify 1,428 complexes that contain mostly RNA-associated proteins (see Methods). This suggests that greater than 20% of known protein complexes function as RNP complexes (Table 1 and Table S3).

Notable examples of previously uncharacterized RNP complexes include the human activating signal cointegrator-1 (hASC-1) complex, whose loss of function causes spinal

muscular atrophy (Knierim et al., 2016), tripeptidyl-peptidase II, a large protease linked to cancer, and the SPATA5-SPATA5L1 complex, an uncharacterized complex linked to epilepsy, hearing loss, and mental retardation syndrome (Tanaka et al., 2015) (Table 1). We provide the complete set of RNP complexes as a fully searchable web database, rna.MAP, at rna.proteincomplexes.org. This represents the first detailed resource of human RNP complexes giving researchers testable hypotheses about RNP complexes.

Replication factor C is an RNP ComplexA detailed look at the RNP Select complexes reveals the replication factor C (RFC) complex exists as a stable RNP complex (Figure 3). During replication and DNA damage repair, the RFC complex is responsible for loading PCNA, a DNA polymerase processing factor, onto DNA. Our DIF-FRAC data identify two previously observed variants of the RFC complex, RFC1-5 and RFC2-5 (Figure 3A), where RFC1-5 appears to be the dominant variant as well as the RNA-associated form (Figure 3B). Consistent with the RFC complex interacting with RNA, the homologous clamp loader in *E.coli*, γ complex, is shown to load the DNA clamp onto RNA-primed template DNA (Yao and O'Donnell, 2012). Eukaryotic RFC has also been shown to be capable of loading PCNA onto synthetic RNA-primed DNA (Yuzhakov et al., 1999). In light of this, we tested whether purified RFC complex from *S. cerevisiae* could directly bind different species of nucleic acids. We observe that it not only binds dsDNA and DNA-RNA hybrids, but also dsRNA with surprisingly tight binding constants in the nanomolar range (Figure 3C and Figure S5). These data validate the use of DIF-FRAC to identify RNP complexes, and suggest a previously uncharacterized role for dsRNA in RFC complex function.

Classification of RNP Complexes

RNA performs a variety of roles in macromolecular complexes. For example, it can bind as a substrate, function as an integral structural component, or act as a regulator of a complex's composition. Mirroring these roles, DIF-FRAC data reveals that RNP complexes can remain fully intact (Figure 4A), become destabilized (Figure 4B), or adopt different compositions (Figure 4C) upon RNA degradation. We therefore categorize stable RNP complexes into to three groups which we term 'apo-stable' (the protein complex is stable without RNA), 'structural' (RNA is essential for the RNP complex structure and/or subunit solubility), and 'compositional' (RNA promotes different stable combinations of protein-complex subunits perhaps in a

regulatory role) (Figure 4 and Table 1). Specific examples of apo-stable RNP complexes include the exosome, RNase P, and the multi-synthetase complex (Figure 4A). Elution profiles show that in the absence of RNA, subunits in these RNP complexes co-elute as a smaller, non-RNA-associated complex consistent with a molecular weight of the apo form. Available atomic structures of the exosome with and without RNA support the concept that RNA is peripheral to the stability of the complex (Gerlach et al., 2018; Weick et al., 2018).

Structural RNP complexes include the 40S and 60S ribosomal subcomplexes (Figure 4B and Figure S6). Upon degradation of RNA, the abundance of ribosomal protein subunits markedly decreases, suggesting the ribosome breaks apart and subunits become insoluble. This result is consistent with solved structures of the ribosome (Anger et al., 2013), demonstrating the centrality of rRNAs to the overall complex architecture (Figure 4B). Interesting exceptions to this behavior are the elution profiles for RPLP0, RPLP1 and RPLP2. These proteins co-elute in the RNase A treated sample, suggesting RNA does not mediate their interaction. In agreement with this observation, the atomic structure of the human ribosome confirms that interactions between RPLP0, RPLP1 and RPLP2 are protein mediated (Figure 4B). This example demonstrates how DIF-FRAC data can not only identify RNA-protein mediated interactions, but also contain structural information about RNP subcomplexes.

We finally observe RNA acting as a nonexclusive subunit of RNP complexes, where two different protein complexes occur in the presence and absence of RNA. We deem these compositional RNPs (Figure 4C). For example, the WCRF (Williams syndrome transcription factor-related chromatin remodeling factor) complex, NuRD (Nucleosome Remodeling Deacetylase) complex and Cohesin complex are reported to together form a chromatin-remodeling supercomplex (CORUM ID: 282). We observe the WCRF and NuRD complexes co-eluting in the control experiment forming a 12-subunit complex that shifts its elution upon RNA degradation. Interestingly, we observe the supercomplex (WCRF, NuRD and Cohesin) eluting as a ~17-subunit complex in the RNA degradation condition. This composition change provides an explanation for why several NuRD-containing complexes are observed experimentally (Xue et al., 1998; Hakimi et al., 2002), as they likely represent both RNP complexes and non-RNA-associated complexes.

We also find an uncharacterized compositional RNP complex containing the cell growth regulators DRG1 and ZC3H15 (DRFP1) (Ishikawa et al., 2005) that are implicated in lung

cancer (Lu et al., 2016). ZC3H15 stabilizes DRG1 and prevents degradation possibly by preventing poly-ubiquitination (Ishikawa et al., 2005). Our result suggests RNA is involved in ZC3H15's role in stabilizing DRG1. We observe a shift to a non-RNA-associated complex containing DRG1-ZC3H15-LRRC41 in the absence of RNA (Figure 4C). LRRC41 is a probable substrate recognition component of E3 ubiquitin ligase complex (Kamura et al., 2004). Our data suggests that in the absence of RNA, LRRC41 binds the DRG1-ZC3H15 complex and recruits ubiquitin machinery. A further example of a compositional RNP complex is the transcription factor (TF)IIIC-TOP1-SUB1 complex, which is involved in RNA polymerase III pre-initiation complex (PIC) assembly (Male et al., 2015). DIF-FRAC shows this 7-subunit complex changes composition to the five-subunit TFIIIC upon RNA degradation (Figure 4C), offering further insights into the mechanism of TFIIIC-dependent PIC formation.

Finally, we identify the chromatin remodeling BRG/hBRM associated factors (BAF; the mammalian SWI/SNF complex; SWI/SNF-A) and polybromo-associated BAF (PBAF; SWI/SNF-B) complexes as compositional RNP complexes, which are some of the most frequently mutated complexes in cancer (Hodges et al., 2016; Tang et al., 2017) (Figure S7). BAF and PBAF complexes share a set of common core subunits, but also each have signature subunits that are related to their respective functions. Elution profiles show these core subunits co-elute with PBAF-only subunits in the control, but shift to co-elute with BAF-only subunits upon RNA degradation (Figure S7). These data suggest BAF exists as a non-RNA-associated complex while PBAF functions as an RNP complex, consistent with its known role in transcription and supporting a previously described RNA-binding model where lncRNAs interact with SWI/SNF complexes in cancer (Tang et al., 2017). Together these examples describe the various physical relationships between RNA and macromolecular protein complexes.

Evaluating RNPs in multiple proteomes

To fully account for the prevalence of RNP complexes we require robust methods that can be applied to many different cell types, tissues and species. Human erythrocytes are particularly intriguing to search for RNP complexes because they lack major sources of bulk RNA. Specifically, they discard their nucleus during maturation and contain little or no RNA based on their lack of stain with RNA-binding dyes (Lee et al., 1986). Consistent with this, they display virtually no DIF-FRAC signal on a global scale; no change in the bulk SEC absorbance signal is

seen upon RNA degradation (Figure 1C) and we identify substantially fewer RNA-associated proteins (111) compared to HEK 293T (Figure 5A). This suggests the DIF-FRAC method is robust to discriminate between cell types with varying utility of RNP complexes. Of these 111 RNA-associated proteins, however, 63 of them have been previously annotated to associate with RNA in high throughput experiments (Table S1). Interestingly, while we see consistent behavior of the non-RNA-associated COP9 signalosome complex in erythrocytes as compared to HEK 293T (Figure S2A,B), our data indicate the molecular chaperone CCT (also called TRiC) complex (Spiess et al., 2004) has a significant DIF-FRAC score in erythrocytes compared to HEK 293T (Figure S8A,B) pointing to CCT as a novel RNP complex. Consistent with this result, the CCT subunit CCT5 has been shown to bind RNA nonspecifically to the 5'UTR of p53 mRNA (Takagi et al., 2005), and the CCT complex has a role in localizing scaRNAs to Cajal bodies (Freund et al., 2014). In total, these data suggest RNA plays a limited yet important role in erythrocytes.

Recently there has been greater appreciation for the role of RNA in embryonic development specifically in the targeted localization of maternal RNA and the importance of lncRNAs for cell differentiation and stem cell potency. We therefore applied DIF-FRAC to mouse embryonic stem cells to establish the prevalence of RNP complexes in defining the pluripotent state. We identified 1,165 significant RNA-associated proteins in mESCs (Figure 5A, Table S1), including 466 that are novel, representing a 35% increase in the number of annotated mouse RNA-associated proteins (Figure 5B). Identified RNA-associated proteins are reproducible between species as we see strong overlap between equivalent proteins in human and mouse (Figure 5C). Additionally, the behavior of complexes is reproducible including the non-RNA-associated COP9 signalosome (Figure S2C) and the RFC complex, which specifically behaves as a compositional RNP complex in both species (Figure 3A).

Among the identified mESC RNA-associated proteins are members of the centralspindlin complex, a heterotetramer consisting of Racgap1 and Kif23 and involved in cytokinesis (Figure 5D) (Yuce et al., 2005; White and Glotzer, 2012). Previously unknown to contain an RNA component, we identify Racgap1, Kif23 and the centralspindlin interaction partner Ect2 as significantly sensitive to RNase A treatment in mESC. In agreement with this mESC result, we observe this behavior in human cells showing conservation across species (Figure 5E). Our results suggest a physical interaction between the centralspindlin complex and RNA and

corroborates previous studies that report Kif23 (ZEN-4 in *C.elegans*) as a positive regulator of RNP granules in germ line (Wood et al., 2016), as well as localization of several RNA species to the midbody during cytokinesis (Clemson et al., 1996; Lecuyer et al., 2007; Zheng et al., 2010). The ability to apply DIF-FRAC to diverse systems will allow identification of conserved RNA-associated proteins and RNP complexes in different human tissues, disease states, and throughout all domains of life.

Characterization of individual RNA-associated proteins

Although our resource focuses primarily on RNP complexes, the DIF-FRAC method is capable of characterizing individual RNA-associated proteins as well. Inspection of elution profiles for individual proteins reveals at least four distinct DIF-FRAC signals (Figure 6). Specifically, the signals manifest as elution-profile changes with RNase A treatment that show either (1) an apparent decrease in molecular weight of the RNA-associated protein consistent with the degradation of an RNA component (Figure 6A); (2) a decrease in abundance, suggesting the RNA-associated protein becomes insoluble or is degraded (Figure 6B); (3) an apparent increase in molecular weight, suggesting the RNA-associated protein forms a higher-order species or aggregate (Figure 6C); or (4) an increase in abundance, indicative of the RNA-associated protein becoming more soluble (Figure 6D). A full analysis of all identified RNA-associated proteins shows a majority (796) decrease in molecular weight, while 216 RNA-associated proteins increase in size (Figure S9). Aside from RNA acting as an interaction partner to RNA-associated proteins, RNA has been shown to regulate the oligomerization state of proteins both positively (Huthoff et al., 2009; Bleicheret and Baserga, 2010; Xie et al., 2018) and negatively (Yoshida et al., 2004). Our data suggests that while the majority of RNA-associated proteins form higher-order assemblies with RNA, the oligomerization of ~20 % is potentially inhibited by RNA. Alternatively, RNA has also been shown to alter the solubility state of proteins (Maharana et al., 2018). We observe an increase in abundance for 535 proteins upon RNase A treatment, a decrease in abundance for 470 proteins, and no change in abundance for only 7 proteins. This suggests RNA impacts the solubility for most RNA-associated proteins and may function to tune protein availability in the cell.

Looking specifically at individual proteins, we find BANF1, a chromatin organizer protein, appears insoluble under our experimental conditions without RNA (Figure 6B).

Interestingly, the BANF1 mutation Ala12-Thr12 causes Hutchinson-Gilford progeria syndrome, a severe and debilitating aging disease, by a reduction in protein levels (Puente et al., 2011). One hypothesis is that this reduction is caused by disruption of the RNA-BANF1 interaction, leading to insolubility and degradation. Furthermore, RNA has also been shown to solubilize proteins linked to pathological aggregates (Maharana et al., 2018). Our data identifies a number of CREC family members (CALU, RCN1, RCN2, and SDF4; Figure 6C and Table S1) as RNA-associated proteins that increase in molecular weight upon RNA degradation. The CREC family is a group of multiple EF-hand, low-affinity calcium-binding proteins with links to amyloidosis (Vorum et al., 2000). DIF-FRAC demonstrates a dependence of RNA on the oligomerization state of CALU, which could play a role in the formation of amyloid deposits similar to that observed for prion-like RNA-associated proteins (Maharana et al., 2018). Based on these examples and the many disease links to DIF-FRAC identified RNP complexes (Table 1), we anticipate our data will generate testable RNA-related hypotheses about disease-related states.

Conclusion

Here, we report the design, development, and application of a robust fractionation-based strategy to determine RNP complexes on a proteome-wide scale. We successfully used DIF-FRAC to identify 115 stable RNP complexes throughout the human interactome, and applied it to multiple tissue types and species. To our knowledge, this is the first experimental study to investigate native human ribonucleoprotein complexes system wide. Combining this with previous data, we generate a resource of the RNA-bound human proteome and demonstrate that upwards of 20% of protein complexes contain an RNA component, highlighting the prevalence of RNP complexes in the cellular milieu. Together our results provide a valuable tool for researchers to investigate the role of RNPs in protein function and disease.

The DIF-FRAC methodology offers important advances over previous techniques that examine RNA-protein interactions. Specifically, interactions are probed proteome-wide in a native, whole lysate sample using a strategy that is not reliant on labelling or cross-linking efficiency. We show DIF-FRAC can be applied effectively to multiple cell types and organisms, and has the potential to provide information on protein-RNA interactions in disease states. Furthermore, DIF-FRAC is a broadly applicable framework that can be extended to examine other large-scale proteomic changes in a system of interest.

We also introduce three classifications of RNP complexes (apo-stable, structural, and compositional) that provide a useful framework to organize the roles of RNAs in macromolecular complexes. Additionally, DIF-FRAC provides information on the biochemical characteristics (i.e. molecular weight, solubility) of RNP complexes in the presence and absence of RNA that offer clues to disease pathophysiology. We anticipate this technique to be a powerful tool to uncover the molecular mechanisms of RNA related diseases. Overall, the DIF-FRAC method described and demonstrated here charters new territories in the cellular landscape of RNA-protein interactions. We have utilized DIF-FRAC to provide the first system-wide resource of human RNPs, providing a broadly applicable tool for studying cellular interactions and responses in multiple cell types and states.

Acknowledgements

We would like to thank Matt Davis for helpful comments, Claire D. McWhite for plotting software and the Texas Advanced Computing Center for high-performance computing. This work was supported by grants from the NIH (K99 HD092613 to K.D., R35 GM122480, DP1 GM106408, R21 GM119021, R01 DK110520, and R01 HD085901 to E.M.M., R01 GM112722 to J.K. and R01 GM120554 to I.J.F.); NSF (IOS-1237975 to E.M.M., and 1453358 to I.J.F.); Army Research Office (W911NF-12-1-0390 to E.M.M.); and Robert Welch Foundation (F1515 to EMM and F1808 to I.J.F.). Data collection by the UT Austin Proteomics Facility was supported by CPRIT RP110782 to Maria Person.

Figure Legends:

Figure 1: Differential fractionation (DIF-FRAC) identifies RNP complexes. (A) The DIF-FRAC workflow requires two equivalent cell culture lysates for a control and an RNase A treated sample. Lysate is separated into fractions by size exclusion chromatography, and proteins in each fraction are identified by mass spectrometry to determine individual protein elution profiles proteome-wide for each condition. An elution shift of a protein upon RNase A treatment is indicative of an RNA-protein interaction. Elution shifts are cross-referenced with known protein complexes to identify RNP complexes (B) Separations of HEK293T lysate under control (black) and RNase A treated (red) conditions monitored by bulk SEC chromatography absorbance profiles at A_{280} show loss of high molecular weight signal upon treatment. (C) Separations of erythrocyte lysate under control (black) and RNase A treated (red) conditions monitored by bulk SEC chromatography absorbance profiles at A_{280} show overlapping absorbance signal. (D) RNA binding protein elution profile for nucleolar RNA helicase 2 (DDX21) (Abundance = count of unique peptide spectral matches). The elution profile shows sensitivity to RNase A treatment. (E) Elution profile for phosphoglucomutase (PGM1) which does not interact with RNA is not sensitive to RNase A treatment. (F) Elution profiles for subunits of the spliceosome RNP complex show co-elution of complex in control and a shift in elution upon RNase A treatment. (G) Elution profile for the non-RNA-associated MCM complex shows no detectable elution shift. In (B)-(G) dashed lines correspond to the elution volumes of molecular weight standards thyroglobulin (M_r = 669 kDa), apoferritin (M_r = 443 kDa), albumin (M_r = 66 kDa), and carbonic anhydrase (M_r = 29 kDa). Molecular weight labels on subsequent plots are removed for clarity. (H) A DIF-FRAC score is calculated for each protein from the absolute value of the differences in the elution profiles between control and RNase A treated samples, and then summed and normalized. See also Figure S3A. (I) DIF-FRAC P-value calculated on HEK 293T data shows strong ability to discriminate known RNA-associated proteins from other proteins. See also Figure S3B. (J) DIF-FRAC identifies 115 stable 'RNP Select' complexes and 441 complexes where the majority of subunits were RNA-associated. 1428 total complexes were identified where the majority of subunits were RNA-associated determined by a combination of DIF-FRAC and other published datasets.

Figure 2: DIF-FRAC reveals a map of stable RNP complexes. Our data demonstrate at least 115 annotated protein complexes exist as stable RNP complexes we call RNP Select (top). RNP Select complexes are defined as complexes whose protein subunits co-elute in the control DIF-FRAC sample (> 0.75 average correlation coefficient) and > 50% of subunits have a DIF-FRAC P-value > 0.5. Colors indicate different classes of RNP complexes: apo-stable (red), structural (orange), and compositional (green). (1-8) Elution profiles for RNP complexes in RNP Select. Abundance represents count of unique peptide spectral matches for each protein.

Figure 3: Replication factor C is an RNP Complex (A) Elution profiles in both human and mouse demonstrates RFC1-5 forms an RNP complex (blue/yellow highlight). A smaller subcomplex of RFC2-5 (green highlight) becomes the dominant form upon RNA degradation. (B) A cartoon to show the RNA-dependence of annotated complexes RFC1-5 (blue) and RFC2-5 (green) as determined by DIF-FRAC. RNA is shown in grey. (C) Electromorphic mobility shift assays (EMSA) of various concentrations of purified *S. cerevisiae* RFC mixed with 1 nM ^{32}P-labeled oligonucleotides. Representative gels showing RFC binds dsDNA, DNA/RNA hybrid and dsRNA substrates. RFC-nucleic acid complexes were separated on 10% native gels. Binding constants are in the nanomolar range (Figure S5).

Figure 4: DIF-FRAC identifies 3 classes of RNP complexes. (A) Elution profiles of the exosome (top, CORUM 7443), RNase P (middle, CORUM 123), and the multi-synthetase complex (bottom, CORUM 3040) demonstrate they are 'apo-stable' RNP complexes that bind RNA and remain intact in the absence of RNA. Blue shading represents RNA bound complex. Red shading represents RNA unbound complex. (B) Elution profiles of the 60S ribosomal subunit (CORUM 308) show it destabilizes upon RNA degradation and is a 'structural' RNP complex. DIF-FRAC elution data show the ribosomal subunits RPLP0, RPLP1 and RPLP2 (orange) remain as a subcomplex upon RNA degradation, consistent with their position in the solved ribosome structure (bottom, PDB 4V6X). (C) Examples of 'compositional' RNP complexes. Green shading represents RNA unbound complex. (Top) Elution profiles of NuRD-WCRF-Cohesin (CORUM 282) and NuRD-WCRF suggest that the different forms are promoted by RNA association. (Middle) DIF-FRAC suggests that the Drg1-ZC3H15-LRRC41 complex (hu.MAP 2767) forms only in the absence of RNA. (Bottom) The TFIIIC containing-TOP1-

SUB1 complex (CORUM 1106) loses two subunits upon RNA degradation. In (A)-(C) dashed lines correspond molecular weight standards described in Figure 1.

Figure 5: DIF-FRAC identifies RNP complexes across cell types and species. (A) DIF-FRAC identifies RNA-associated proteins in mESC (mouse embryonic stem cells; 1165), HEK 293T cells (1012), and human erythrocytes (111). (B) Venn diagram of overlap between previously published large-scale RNA-protein interaction studies, literature annotated RNA-associated proteins, and DIF-FRAC identified RNA-associated proteins in mESC. (C) RNA-associated human-mouse orthologues are identified reproducibly in DIF-FRAC experiments. Elution profiles of the centralspindlin complex for (D) mESC and (E) HEK 293T demonstrate centralspindlin is an RNP complex in both species. Yellow and blue shading represent RNA bound complex in mESC and HEK 293T respectively.

Figure 6: Four distinct DIF-FRAC signals for RNA-associated proteins. Examples of elution profiles for proteins that (A) decrease in size, (B) decrease in abundance (less soluble), (C) increase in size and (D) increase in abundance (more soluble) upon RNA degradation.

Figure S1: DIF-FRAC accesses a diverse RNA landscape. Box plots show the RNA abundance of mRNA, lncRNA, small RNA, other ncRNA, and pseudogenes in control fractions 16-23 of HEK 293T cell lysate (TPM = Transcripts Per Million). Boxes indicate median (inner joint), first quartile (left) and third quartile (right). Lines indicate 1.5 interquartile range. Dots indicate outliers.

Figure S2: Non-RNA-associated complexes are insensitive to RNase A treatment. DIF-FRAC elution profiles show subunits of the non-RNA-associated COP9 signalosome complex (M_r ~500 kDa (Oron et al., 2002)) in control (black) and RNase A treated (red) for (A) HEK293T lysate, (B) erythrocytes and (C) mESC do not shift upon RNase A treatment. Abundance represents count of unique peptide spectral matches. Vertical dotted lines represent protein standards described in Figure 1.

Figure S3: DIF-FRAC Score accurately discriminates between RNA-associated proteins and non-binders.
(A) Workflow to calculate abundance corrected P-values for each protein's DIF-FRAC score. (B) Precision recall analysis shows the DIF-FRAC Score recalls a substantial number of known RNA-associated proteins in HEK 293T cells. (C) High DIF-FRAC P-values have high precision in recovering known RNA-associated proteins in mouse embryonic stem cells. (D) Precision recall analysis shows the DIF-FRAC Score recalls a substantial number of known RNA-associated proteins in mESC.

Figure S4: DIF-FRAC RNA-associated proteins show substantial overlap with other high-throughput studies. (A-K) Venn diagrams show overlap of DIF-FRAC RNA-associated proteins from HEK 293T cells with 11 high-throughput RNA association studies (Bao et al., 2018; Hentze et al., 2018; Huang et al., 2018; Queiroz et al., 2018; Trendel et al., 2018). (L) Venn diagram shows overlap of DIF-FRAC RNA-associated proteins from HEK 293T cells and combined set of high throughput RNA association studies. (M) Enrichment of RNA binding structural motifs in DIF-FRAC-identified RNA-associated proteins from HEK 293T cells.

Figure S5: Affinity of nucleic acid for the *S. cerevisiae* RFC complex. Binding curves from electromorphic mobility shift assays (EMSA) of various concentrations of purified *S. cerevisiae* RFC mixed with 1 nM ^{32}P-labeled oligonucleotides. Data was fit to a hyperbolic equation (solid line). The calculated k_D + 95% CI is 1.9 ± 0.5 nM for dsDNA (black), 7.5 ± 4.9 nM for DNA/RNA hybrid (red), and 25 ± 11 nM for dsRNA (blue). Error bars denote standard deviation.

Figure S6: DIF-FRAC analysis of the 40S ribosomal subunit. Elution profiles of the 40S ribosomal subunits demonstrate it is destabilized upon RNA degradation ('structural' RNP complex).

Figure S7: DIF-FRAC analysis of human BAF and PBAF complexes. Elution profiles of annotated human BAF and PBAF complexes demonstrate core subunits common to both complexes coelute in both control and RNase A treated samples, but at different molecular

weights. PBAF-only subunits (light grey) coelute with core subunits only in the control sample, while BAF-only subunits (dark grey) coelute with the core subunits as a lower molecular weight complex only when RNA is degraded. Together, these elution profiles suggest that PBAF is an RNP complex, but the BAF complex does not associate with RNA.

Figure S8: DIF-FRAC analysis of the CCT complex. Elution profiles show (A) HEK 293T CCT complex is insensitive to RNase A treatment while (B) erythrocyte CCT complex is sensitive to RNase A treatment suggesting it behaves as an RNP complex in red blood cells.

Figure S9: Analysis of DIF-FRAC shift types of RNA-associated proteins. Upon RNase A treatment we observe different types of changes to protein elution profiles. Each point in the graph represents one RNA-associated protein. Molecular weight shift is the weighted average difference between control and RNase A treated profiles, where a negative value (left side of graph) represents lower molecular weight elution upon treatment and positive value (right side of graph) represents gain in molecular weight (see Methods for calculation). Abundance change is the normalized change in abundance upon RNase A treatment. A positive value (top of graph) represents gain in solubility and a negative value (bottom of graph) represents loss in solubility. Examples from Figure 6 are annotated.

Materials and Methods

Experimental Design

Human cell culture and extract preparation

HEK293T cells (ATCC CRL3216) cultured in DMEM (Gibco) supplemented with 10 % (v/v) FBS (Life Technologies) were continually split over 7 days to give four 10-cm dishes of adherent cells. For the control fractionation sample, two 10-cm dishes of cells were harvested at 80-100 % confluence without trypsin by washing in ice cold phosphate buffered saline (PBS) pH 7.2 (0.75 mL; Gibco) and placed on ice. Cells (approximately 0.1 g wet weight) were lysed on ice (5 min) by resuspension in Pierce IP Lysis Buffer (0.8 mL; 25 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% NP-40 and 5% glycerol; Thermo Fisher) containing 1x protease inhibitor cocktail III (Calbiochem). The resulting lysate was clarified (17,000g, 10 min, 4°C) and left at room temperature (30 min). The sample was filtered (Ultrafree-MC filter unit (Millipore);

12,000g, 2 min, 4°C) to remove insoluble aggregates. RNase A treated samples were prepared on the same day in an identical manner, except RNase A (8 μL, 80 μg, Thermo Fisher, catalogue #EN0531) was added after lysate clarification and the sample left at room temperature (30 min) before filtration.

Mouse embryonic stem cell culture

Gelatin adapted mouse J1 ES cells (ATCC® SCRC-1010™) were cultured in Dulbecco's Modified Eagle's Medium (DMEM, Life Technologies) containing 18% fetal bovine serum (FBS, Gemini), 50 U/mL of penicillin/streptomycin with 2 mM L-glutamine (Life Technologies), 0.1 mM non-essential amino acid (Life Technologies), 1% nucleosides (Sigma-Aldrich), 0.1 mM β-mercaptoethanol (Sigma-Aldrich), and 1,000 U/mL recombinant leukemia inhibitory factor (LIF, Chemicon). ES cells were plated on 15-cm dishes coated with 0.1% gelatin and incubated at 37°C and 5% CO_2. Cells were passaged every 2 days. Lysis and RNase A treatment were done as described in the HEK 293T protocol.

Erythrocyte cell preparation

Leukocyte-reduced red blood cells (RBCs) were obtained from an anonymous donor and purchased from Gulf Coast Regional Blood Center (Houston, Texas). The RBCs used in this experiment were kept at 4°C for 54 days before lysis to ensure reticulocytes mature into RBCs. Prior to cell lysis, RBCs were washed with ice cold PBS (pH 7.4, Gibco) for 3 times at 600 g for 15 min at 4°C. RBCs were then lysed in hypotonic solution (5 mM Tris-HCl, pH 7.4) containing protease and phosphatase inhibitors (cOmplete, EDTA-free Protease Inhibitor Cocktail, Roche and PhosSTOP, Roche) with a ratio of 1 volume packed RBC: 5 volumes hypotonic solution. Hemolysate (soluble fraction of RBC lysate) was collected by centrifuging white ghosts (membrane fraction of RBC lysate) at 21,000 g for 40 mins at 4°C. Hemolysate was collected and stored at -80°C until further use. On the day of experiment, hemolysate was thawed and treated with Hemoglobind (Biotech Support Group) in order to remove hemoglobin from hemolysate. A total of 4-5 mg of total proteins were split into control and RNase A treated samples. The RNase sample was treated with RNase A as described in the protocol of RNase A treatment of lysate from HEK293T cells. Both samples were filtered (Ultrafree-MC filter unit (Millipore); 12,000 g, 2 min, 4°C) to remove insoluble aggregates prior to fractionation.

Biochemical fractionation using native size-exclusion chromatography

All lysates were subject to size exclusion chromatography (SEC) using an Agilent 1100 HPLC system (Agilent Technologies, ON, Canada) with a multi-phase chromatography protocol as previously described (Havugimana et al., 2012). Soluble protein (1.25 mg, 250 μL) was applied to a BioSep-SEC-s4000 gel filtration column (Phenomenex) equilibrated in PBS, pH 7.2 (HEK 293T and mESC lysate) or pH 7.4 (erythrocytes) at a flow rate of 0.5 mL min^{-1}. Fractions were collected every 0.375 mL. The elution volume of molecular weight standards (thyroglobulin (M_r = 669 kDa); apoferritin (M_r = 443 kDa); albumin (M_r = 66 kDa); and carbonic anhydrase (M_r = 29 kDa); Sigma) was additionally measured to calibrate the column (Figure 1B).

Mass spectrometry

Fractions were filter concentrated to 50 μL, denatured and reduced in 50 % 2,2,2-trifluoroethanol (TFE) and 5 mM tris(2-carboxyethyl)phosphine (TCEP) at 55 °C for 45 minutes, and alkylated in the dark with iodoacetamide (55 mM, 30 min, RT). Samples were diluted to 5 % TFE in 50 mM Tris-HCl, pH 8.0, 2 mM $CaCl_2$, and digested with trypsin (1:50; proteomics grade; 5 h; 37 °C). Digestion was quenched (1 % formic acid), and the sample volume reduced to ~100 μL by speed vacuum centrifugation. The sample was washed on a HyperSep C18 SpinTip (Thermo Fisher), eluted, reduced to near dryness by speed vacuum centrifugation, and resuspended in 5 % acetonitrile/ 0.1 % formic acid for analysis by LC-MS/MS.

Peptides were separated on a 75 μM x 25 cm Acclaim PepMap100 C-18 column (Thermo) using a 3-45 % acetonitrile gradient over 60 min and analyzed online by nanoelectrospray-ionization tandem mass spectrometry on an Orbitrap Fusion or Orbitrap Fusion Lumos Tribrid (Thermo Scientific). Data-dependent acquisition was activated, with parent ion (MS1) scans collected at high resolution (120,000). Ions with charge 1 were selected for collision-induced dissociation fragmentation spectrum acquisition (MS2) in the ion trap, using a Top Speed acquisition time of 3-s. Dynamic exclusion was activated, with a 60-s exclusion time for ions selected more than once. MS from HEK 293T cells was acquired in the UT Austin Proteomics Facility.

Construction and sequencing of RNA-seq libraries of DIF-FRAC samples

Fractions from a biological replicate SEC separation corresponding to higher molecular weight species (approximately >1.5 MDa; fractions 16-23 in Figure 1B) were analyzed by total RNA sequencing. Total RNA was isolated from each fraction (0.375 mL) by addition of Trizol (1.125 mL; Thermo Fisher) and the sample (1.4 mL) was transferred to a Phasemaker tube (Thermo Fisher). Total RNA was extracted following the protocol supplied by the manufacturer and further cleaned up using a RNeasy MinElute Cleanup Kit (Qiagen). RNA integrity number (RIN) was measured using an Agilent Bioanalyzer and samples were ribo-depleted using a using a RiboZero Gold (Human/Mouse/Rat) kit (Illumina) to remove rRNAs. RNA libraries were prepared for sequencing according to vendor protocols using NEBNext R Small RNA Library Prep Set for Illumina R (Multiplex Compatible), Cat #E7330L, according to the protocol described by Podnar et al. (Podnar et al., 2014). RNA was fragmented using elevated temperature in carefully controlled buffer conditions to yield average fragment sizes of 200 nucleotides. These fragments were directionally ligated to 5′ and 3′ adaptors so that sequence orientation is preserved throughout sequencing. Reverse transcription and PCR were performed to complete the DNA sequencing libraries, which were sequenced using an Illumina NextSeq 500 instrument (75-nt single reads) at the Genomic Sequencing and Analysis Facility at the University of Texas at Austin.

***S. cerevisiae* RFC purification:**

RFC was purified as previously described (Finkelstein et al., 2003; Kim et al., 2017). Briefly, full-length *S. cerevisiae* RFC was expressed in BL21(DE3) ArcticExpress (Agilent) *E. coli* co-transformed with pLant2b-RFC-AE (pIF117) and pET11-RFC-BCD (pIF116). RFC was subsequently purified by SP and Q (GE Healthcare) ion exchange chromatography. Protein concentration was determined by comparison to a BSA titration curve using Coomassie-stained SDS-PAGE.

Electrophoretic Mobility Shift Assay (EMSA):

Oligonucleotide constructs were based on an earlier description (Kobayashi et al., 2006). Each of the four nucleic acid substrates were radiolabeled with [γ-^{32}P]-ATP using T4 Polynucleotide Kinase (NEB). Free nucleotide was removed using G-25 MicroSpin columns (GE Healthcare).

Oligonucleotides were subsequently heated to 75°C and slowly cooled to room temperature to allow proper annealing. 1 nM oligonucleotide and various concentrations of RFC (0 to 256 nM) were incubated for 15 minutes at room temperature in a buffer containing 25 mM Tris-HCl [pH 7.5], 50 mM NaCl, 2 mM $MgCl_2$, 2 mM DTT, and 0.1 mg/mL BSA. Reactions were quenched with 6x loading dye (10 mM Tris-HCl [pH 7.6], 60% glycerol, 60 mM EDTA, 0.15% [w/v] Orange G) and subsequently separated by native acrylamide gel electrophoresis. Gels were dried on Zeta-Probe Membrane (Bio-Rad) at 80°C for two hours. Bands were visualized by a Typhoon FLA 7000 phosphorimager (GE Healthcare). Binding was quantified using FIJI (Schindelin et al., 2012). Subsequent data were fit to a hyperbolic equation to determine the k_D for oligonucleotide binding.

Oligonucleotides used:

Name	Sequence
dsDNA	5' - CTC GAG GTC GTC ATC GAC CTC GAG ATC A – 3'
DNA/RNA	5' - rCrUrC rGrArG rGrUrC rGTC ATC GAC CTC GAG ATC A – 3'
dsRNA	5' - rCrUrC rGrArG rGrUrC rGrUrC rArUrC rGrArC rCrUrC rGrArG rArUrC rA – 3'

Calculated k_D from fitting to hyperbolic equation (Bound = $(v*[E])/(k_D+[E])$), where "[E]" is the concentration of the enzyme, and "v" and "k_D" are solved by linear regression.

Bioinformatic analysis

Protein identification

Prior to protein identification, human and mouse proteomes were downloaded from UniProt website (Apweiler et al., 2004). Raw formatted mass spectrometry files were first converted to mzXML file format using MSConvert (http://proteowizard.sourceforge.net/tools.shtml) and then processed using MSGF+ (Kim and Pevzner, 2014), X! TANDEM (Craig and Beavis, 2004) and Comet (Lingner et al., 2011) peptide search engines with default settings. MSBlender (Kwon et al., 2011) was used for integration of peptide identifications and subsequent mapping to protein identifications. A false discovery rate of 1% was used for peptide identification. Protein elution profiles were assembled using unique peptide spectral matches for each protein across all fractions collected.

DIF-FRAC score and P-value significance calculation

In order to determine the significance of a protein's sensitivity to RNase A treatment, we compare the protein's control elution profile to its RNase A treated elution profile as schematized in Figures 2A and S3A. Specifically, we first calculate the L1-norm of the two elution profiles (**equation 1**).

$$D_p = \sum_{i=1}^{N} \mid X_{p,i} - Y_{p,i} \mid \quad \textbf{(1)}$$

Where N represents the total number of fractions collected and p represents an individual protein. X and Y represent abundance matrices of control and experiment (RNase A treated) respectively. We next normalize D_p by the total abundance seen for protein p in both the control and experiment (**equation 2**).

$$D_p^{norm} = \frac{D_p^2}{\sum_{i=1}^{N} X_{p,i} + \sum_{i=1}^{N} Y_{p,i}} \quad \textbf{(2)}$$

We observed D^{norm} is biased by high abundance proteins and we therefore evaluate significance of a protein's sensitivity to RNase A treatment by comparing to a background of proteins with similar abundance. Specifically, we create a distribution of D^{norm} from proteins in a window surrounding protein p and have not been annotated as RNA-associated proteins in the literature (**equation 3**). See Figure S3A for schematic.

$$W_p = D_{p+s}^{norm}, ..., D_p^{norm}, ..., D_{p-s}^{norm} \quad \textbf{(3)}$$

where s is a window size of 100 and unannotated RNA-associated proteins are in order of abundance.

We posit that the proteins in distribution, W_p, is a mixture of unannotated RNA-associated proteins as well as non-RNA-associated proteins. In order to evaluate significance of a protein's D^{norm} being greater than what is expected by non-RNA binders, we model the distribution W_p using a two component gaussian mixture model (GMM). To ensure an accurate model fit we

evaluate our GMM fit using three criteria (equation 4). First, we calculate the Baysian Information Criterion (BIC) for both the two component GMM and a one component GMM and ensure the two component GMM has a lower BIC (**equation 4a**). Second, we ensure the component with the lowest mean μ (i.e. non-RNA-associated component) has the largest weight (**equation 4b**). Finally, we ensure the largest component weight is greater than a given weight threshold t_{weight} (**equation 4c**). t_{weight} can be estimated by the expected fraction of non-RNA binders in the proteome. In practice we set t_{weight} to be between 0.6 and 0.75.

$$BIC_{2-component} \leq BIC_{1-component} \quad \textbf{(4a)}$$

$$\arg\min_{\mu} = \arg\max_{weight} \quad \textbf{(4b)}$$

$$\max_{weight} \geq t_{weight} \quad \textbf{(4c)}$$

If all three criteria are passed the lowest mean component of the two component GMM is used, otherwise the one component is used (**equation 5**).

$$GMM_p = \begin{cases} GMM_{2-component}, & \text{if criteria are met (equation 4)} \\ GMM_{1-component}, & \text{otherwise} \end{cases} \quad \textbf{(5)}$$

We next calculate the Z-score of protein p's D^{norm} score relative to the non-RNA-associated component of GMM_p (**equation 6**).

$$Z_p = \frac{D_p^{norm} - \mu_{GMM_p}}{\sigma_{GMM_p}} \quad \textbf{(6)}$$

where μ_{GMM_p} and σ_{GMM_p} are the mean and standard deviation of the first component of GMM_p respectively.

Finally, we calculate a P-value of Z_p using the normal distribution survival function and then false discovery correct P-values across all proteins using the Benjamini/Hochberg correction (Benjamini and Hochberg, 1995). RNA-associated proteins were considered significant at a 0.05 FDR corrected P-value.

RNA-associated annotations, overlap comparisons and score performance analysis Low throughput RNA binding annotations were defined as proteins with Gene Ontology (Ashburner et al., 2000) "RNA binding" annotations limited to those with evidence codes: EXP, IDA, IPI, IMP, IGI, IEP, TAS, NAS, or IC. In addition, proteins with "ribonucleoprotein" in their UniProt keywords were also included. High throughput RNA association annotations were primarily collected from Hentze *et al.* 2018 supplemental table S2 (Hentze et al., 2018). In addition, we gathered more recent high throughput datasets from (Bao et al., 2018; Huang et al., 2018; Queiroz et al., 2018; Trendel et al., 2018).

To estimate the coverage of identified RNA-associated proteins by the DIF-FRAC method independent of cell type and machine setup, the Venn diagrams in Figures 5B and S4L report only proteins with mean abundance > 10, where mean abundance is the average peptide spectral matches identified in the control and RNase A treated HEK 293T cells. To compare directly the RNA-associated proteins identified in the high throughput sets to the DIF-FRAC method, Venn diagrams in Figure S4A-K report all proteins.

To calculate Precision vs Neg Ln P-value plots (Figure 1I and Figure S3C), we first added a pseudocount (+1e-308) to DIF-FRAC P-values and then applied -1*ln(P-value) where ln is the natural log. Precision is defined as TP/AP, where TP (true positives) is defined as proteins annotated as either high throughput or low throughput RNA binding (see above) and a Neg Ln P-value greater than a given value. AP (all predictions) is defined as any protein with a Neg Ln P-value greater than a given value. To calculate Precision vs Recall plots (Figures S3B and S3D), precision is defined above and recall is defined as TP/AKP where TP is true positives and AKP (all known positives) is defined as proteins annotated as either high throughput or low throughput RNA binding.

Classification of DIF-FRAC elution profiles

To calculate the amount a protein shifts upon RNase A treatment, we calculate the average fraction a protein is observed weighted by the PSMs observed in each fraction. The difference between the weighted average of the treated and untreated elution profiles provides the total shift amount. A protein's shift in elution from a high molecular weight to a low molecular weight results in a negative shift value whereas a shift from low molecular weight to high molecular weight corresponds to a positive value.

To calculate the amount a protein's abundance changes upon RNase A treatment, we calculate the difference of a protein's total PSMs observed in the untreated and treated samples. We further normalize this value by dividing by the sum of the total PSMs from both samples. This results in a value between 1.0 and -1.0 where a positive value corresponds to an increase in abundance upon RNase A treatment and a negative value corresponds to a decrease in abundance upon RNase A treatment.

Assembly of RNP complexes

We define the global set of RNP complexes by first creating a combined non-redundant set of CORUM (Ruepp et al., 2010) and hu.MAP (Drew et al., 2017) complexes (Jaccard coefficient < 1.0). For every complex in this global set we tested if > 50% of the protein subunits were 1) identified as an RNA-associated protein by DIF-FRAC (P-value > 0.05), 2) annotated by high throughput methods or 3) annotated by low throughput methods (see above for description of annotations). RNP Select complexes are defined as complexes whose protein subunits co-elute in the DIF-FRAC control sample (> 0.75 average Pearson correlation coefficient among subunits) and > 50% of subunits have a DIF-FRAC P-value > 0.5.

RNA-Seq Analysis

After performing quality control on the sequencing fastq files using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/), 3' adapter contamination was removed using Cutadapt (v1.10) (Martin, 2011). Alignment of the 8 RNA fraction datasets was then performed with the Hisat2 transcriptome-aware aligner (v2.1.0) (Kim et al., 2015), against a Hisat2 reference index built using GRCh38/hg38 primary assembly genome fasta from Gencode (v27, Ensembl release 90) (Harrow et al., 2012) annotated with the corresponding v27 GTF

(General Transfer Format) annotations. The Hisat suite Stringtie program (v1.3.3b) (Pertea et al., 2016) was used to quantify gene-level expression from the alignment files. TPM (Transcripts Per Million), a sequencing-depth-normalized estimate of reads mapping to the gene, was used for further analysis.

Data Deposition

Data will be deposited in Pride upon acceptance.

Code Repository

Source code is freely available on GitHub: https://github.com/marcottelab/diffrac

References

Anger, A.M., Armache, J.P., Berninghausen, O., Habeck, M., Subklewe, M., Wilson, D.N., and Beckmann, R. (2013). Structures of the human and Drosophila 80S ribosome. Nature *497*, 80-85.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M.*, et al.* (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. *32*, D115-119.

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

Baltz, A.G., Munschauer, M., Schwanhausser, 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.

Bao, X., Guo, X., Yin, M., Tariq, M., Lai, Y., Kanwal, S., Zhou, J., Li, N., Lv, Y., Pulido-Quetglas, C.*, et al.* (2018). Capturing the interactome of newly transcribed RNA. Nat. Methods *15*, 213-220.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J. Roy Stat. Soc. B Met. *57*, 289-300.

Bladen, C.L., Udayakumar, D., Takeda, Y., and Dynan, W.S. (2005). Identification of the polypyrimidine tract binding protein-associated splicing factor p54(nrb) complex as a candidate DNA double-strand break rejoining factor. J. Biol. Chem. *280*, 5205-5210.

Bleicheret, F., and Baserga, S.J. (2010). Ribonucleoprotein multimers and their functions. Crit. Rev. Biochem. Mol. *45*, 331-350.

Bochar, D.A., Savard, J., Wang, W., Lafleur, D.W., Moore, P., Cote, J., and Shiekhattar, R. (2000). A family of chromatin remodeling factors related to Williams syndrome transcription factor. Proc. Natl. Acad. Sci. U.S.A. *97*, 1038-1043.

Brannan, K.W., Jin, W., Huelga, S.C., Banks, C.A., Gilmore, J.M., Florens, L., Washburn, M.P., Van Nostrand, E.L., Pratt, G.A., Schwinn, M.K., *et al.* (2016). SONAR Discovers RNA-Binding Proteins from Analysis of Large-Scale Protein-Protein Interactomes. Mol. Cell. *64*, 282-293.

Brown, D.A., Di Cerbo, V., Feldmann, A., Ahn, J., Ito, S., Blackledge, N.P., Nakayama, M., McClellan, M., Dimitrova, E., Turberfield, A.H., *et al.* (2017). The SET1 Complex Selects Actively Transcribed Target Genes via Multivalent Interaction with CpG Island Chromatin. Cell Rep. *20*, 2313-2327.

Calo, E., Flynn, R.A., Martin, L., Spitale, R.C., Chang, H.Y., and Wysocka, J. (2015). RNA helicase DDX21 coordinates transcription and ribosomal RNA processing. Nature *518*, 249-253.

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., Frese, C.K., Horos, R., Alleaume, A.M., Foehr, S., Curk, T., Krijgsveld, J., and Hentze, M.W. (2016). Comprehensive Identification of RNA-Binding Domains in Human Cells. Mol. Cell. *63*, 696-710.

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

Clemson, C.M., McNeil, J.A., Willard, H.F., and Lawrence, J.B. (1996). XIST RNA paints the inactive X chromosome at interphase: Evidence for a novel RNA involved in nuclear chromosome structure. Journal of Cell Biology *132*, 259-275.

Craig, R., and Beavis, R.C. (2004). TANDEM: matching proteins with tandem mass spectra. Bioinformatics *20*, 1466-1467.

Davis, T.L., Walker, J.R., Campagna-Slater, V., Finerty, P.J., Paramanathan, R., Bernstein, G., MacKenzie, F., Tempel, W., Ouyang, H., Lee, W.H., *et al.* (2010). Structural and biochemical characterization of the human cyclophilin family of peptidyl-prolyl isomerases. PLoS Biol. *8*, e1000439.

Drew, K., Lee, C., Huizar, R.L., Tu, F., Borgeson, B., McWhite, C.D., Ma, Y., Wallingford, J.B., and Marcotte, E.M. (2017). Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. Mol. Syst. Biol. *13*, 932.

Dyer, M.A., Qadeer, Z.A., Valle-Garcia, D., and Bernstein, E. (2017). ATRX and DAXX: Mechanisms and Mutations. Cold Spring Harb. Perspect Med. *7*, a026567.

Efimov, A., Kharitonov, A., Efimova, N., Loncarek, J., Miller, P.M., Andreyeva, N., Gleeson, P., Galjart, N., Maia, A.R., McLeod, I.X., *et al.* (2007). Asymmetric CLASP-dependent nucleation of noncentrosomal microtubules at the trans-Golgi network. Dev. Cell *12*, 917-930.

Fan, J., and Pavletich, N.P. (2012). Structure and conformational change of a replication protein A heterotrimer bound to ssDNA. Genes Dev. *26*, 2337-2347.

Finkelstein, J., Antony, E., Hingorani, M.M., and O'Donnell, M. (2003). Overproduction and analysis of eukaryotic multiprotein complexes in *Escherichia coli* using a dual-vector strategy. Anal. Biochem. *319*, 78-87.

Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H.Y., Dosztanyi, Z., El-Gebali, S., Fraser, M.*, et al.* (2017). InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. *45*, D190-D199.

Freund, A., Zhong, F.L., Venteicher, A.S., Meng, Z., Veenstra, T.D., Frydman, J., and Artandi, S.E. (2014). Proteostatic control of telomerase function through TRiC-mediated folding of TCAB1. Cell *159*, 1389-1403.

Gerlach, P., Schuller, J.M., Bonneau, F., Basquin, J., Reichelt, P., Falk, S., and Conti, E. (2018). Distinct and evolutionary conserved structural features of the human nuclear exosome complex. Elife *7*, e38686.

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

Hakimi, M.A., Bochar, D.A., Schmiesing, J.A., Dong, Y.S., Barak, O.G., Speicher, D.W., Yokomori, K., and Shiekhattar, R. (2002). A chromatin remodelling complex that loads cohesin onto human chromosomes. Nature *418*, 994-998.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S.*, et al.* (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760-1774.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S.*, et al.* (2012). A census of human soluble protein complexes. Cell *150*, 1068-1081.

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

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.

Hentze, M.W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. Nat. Rev. Mol. Cell Biol. *19*, 327-341.

Higgs, M.R., Reynolds, J.J., Winczura, A., Blackford, A.N., Borel, V., Miller, E.S., Zlatanou, A., Nieminuszczy, J., Ryan, E.L., Davies, N.J.*, et al.* (2015). BOD1L Is Required to Suppress Deleterious Resection of Stressed Replication Forks. Mol. Cell *59*, 462-477.

Hodges, C., Kirkland, J.G., and Crabtree, G.R. (2016). The Many Roles of BAF (mSWI/SNF) and PBAF Complexes in Cancer. Cold Spring Harb Perspect Med *6*, a026930.

Huang, R.B., Han, M.T., Meng, L.Y., and Chen, X. (2018). Transcriptome-wide discovery of coding and noncoding RNA-binding proteins. Proc. Natl. Acad. Sci. U.S.A. *115*, E3879-E3887.

Huthoff, H., Autore, F., Gallois-Montbrun, S., Fraternali, F., and Malim, M.H. (2009). RNA-Dependent Oligomerization of APOBEC3G Is Required for Restriction of HIV-1. Plos Pathog *5*, e1000330.

Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P., Parzen, H.*, et al.* (2017). Architecture of the human interactome defines protein communities and disease networks. Nature *545*, 505-509.

Ishikawa, K., Azuma, S., Ikawa, S., Semba, K., and Inoue, J. (2005). Identification of DRG family regulatory proteins (DFRPs): specific regulation of DRG1 and DRG2. Genes Cells *10*, 139-150.

Ishikawa, K., Akiyama, T., Ito, K., Semba, K., and Inoue, J. (2009). Independent stabilizations of polysomal Drg1/Dfrp1 complex and non-polysomal Drg2/Dfrp2 complex in mammalian cells. Biochem. Biophys. Res. Commun. *390*, 552-556.

Jung, D.J., Sung, H.S., Goo, Y.W., Lee, H.M., Park, O.K., Jung, S.Y., Lim, J., Kim, H.J., Lee, S.K., Kim, T.S.*, et al.* (2002). Novel transcription coactivator complex containing activating signal cointegrator 1. Mol. Cell Biol. *22*, 5203-5211.

Kamura, T., Maenaka, K., Kotoshiba, S., Matsumoto, M., Kohda, D., Conaway, R.C., Conaway, J.W., and Nakayama, K.I. (2004). VHL-box and SOCS-box domains determine binding specificity for Cul2-Rbx1 and Cul5-Rbx2 modules of ubiquitin ligases. Genes Dev. *18*, 3055-3065.

Keerthivasan, G., Wickrema, A., and Crispino, J.D. (2011). Erythroblast enucleation. Stem Cells Int *2011*, 139851.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357-360.

Kim, S., and Pevzner, P.A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. *5*, 5277.

Kim, Y., de la Torre, A., Leal, A.A., and Finkelstein, I.J. (2017). Efficient modification of lambda-DNA substrates for single-molecule studies. Sci Rep-Uk *7*, 2071.

Knierim, E., Hirata, H., Wolf, N.I., Morales-Gonzalez, S., Schottmann, G., Tanaka, Y., Rudnik-Schoneborn, S., Orgeur, M., Zerres, K., Vogt, S.*, et al.* (2016). Mutations in Subunits of the Activating Signal Cointegrator 1 Complex Are Associated with Prenatal Spinal Muscular Atrophy and Congenital Bone Fractures. Am. J. Hum. Genet. *98*, 473-489.

Kobayashi, M., Figaroa, F., Meeuwenoord, N., Jansen, L.E.T., and Siegal, G. (2006). Characterization of the DNA binding and structural properties of the BRCT region of human replication factor C p140 subunit. J. Biol. Chem. *281*, 4308-4317.

Kwon, T., Choi, H., Vogel, C., Nesvizhskii, A.I., and Marcotte, E.M. (2011). MSblender: A probabilistic approach for integrating peptide identifications from multiple database search engines. J. Proteome Res. *10*, 2949-2958.

Lecuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T.R., Tomancak, P., and Krause, H.M. (2007). Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. Cell *131*, 174-187.

Lee, L.G., Chen, C.H., and Chiu, L.A. (1986). Thiazole orange: a new dye for reticulocyte analysis. Cytometry *7*, 508-517.

Lingner, T., Asshauer, K.P., Schreiber, F., and Meinicke, P. (2011). CoMet--a web server for comparative functional profiling of metagenomes. Nucleic Acids Res. *39*, W518-523.

Logan, C.V., Szabadkai, G., Sharpe, J.A., Parry, D.A., Torelli, S., Childs, A.M., Kriek, M., Phadke, R., Johnson, C.A., Roberts, N.Y.*, et al.* (2014). Loss-of-function mutations in MICU1 cause a brain and muscle disorder linked to primary alterations in mitochondrial calcium signaling. Nat. Genet. *46*, 188-193.

Lu, L., Lv, Y., Dong, J., Hu, S., and Peng, R. (2016). DRG1 is a potential oncogene in lung adenocarcinoma and promotes tumor progression via spindle checkpoint signaling regulation. Oncotarget *7*, 72795-72806.

Machwe, A., Lozada, E., Wold, M.S., Li, G.M., and Orren, D.K. (2011). Molecular cooperation between the Werner syndrome protein and replication protein A in relation to replication fork blockage. J. Biol. Chem. *286*, 3497-3508.

Maharana, S., Wang, J., Papadopoulos, D.K., Richter, D., Pozniakovsky, A., Poser, I., Bickle, M., Rizk, S., Guillen-Boixet, J., Franzmann, T.*, et al.* (2018). RNA buffers the phase separation behavior of prion-like RNA binding proteins. Science *360,* 918-921.

Male, G., von Appen, A., Glatt, S., Taylor, N.M., Cristovao, M., Groetsch, H., Beck, M., and Muller, C.W. (2015). Architecture of TFIIIC and its role in RNA polymerase III pre-initiation complex assembly. Nat. Commun. *6*, 7387.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10-12.

Mircsof, D., Langouet, M., Rio, M., Moutton, S., Siquier-Pernet, K., Bole-Feysot, C., Cagnard, N., Nitschke, P., Gaspar, L., Znidaric, M.*, et al.* (2015). Mutations in NONO lead to syndromic intellectual disability and inhibitory synaptic defects. Nat. Neurosci. *18*, 1731-1736.

Okuda-Ashitaka, E., Minami, T., Tsubouchi, S., Kiyonari, H., Iwamatsu, A., Noda, T., Handa, H., and Ito, S. (2012). Identification of NIPSNAP1 as a Nocistatin-interacting Protein Involving Pain Transmission. J. Biol. Chem. *287*, 10403-10413.

Oron, E., Mannervik, M., Rencus, S., Harari-Steinberg, O., Neuman-Silberberg, S., Segal, D., and Chamovitz, D.A. (2002). COP9 signalosome subunits 4 and 5 regulate multiple pleiotropic pathways in *Drosophila melanogaster*. Development *129*, 4399-4409.

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. *11*, 1650-1667.

Podnar, J., Deiderick, H., Huerta, G., and Hunicke-Smith, S. (2014). Next-Generation Sequencing RNA-Seq Library Construction. Curr. Protoc. Mol. Biol. *106*, 4 21 21-19.

Puente, X.S., Quesada, V., Osorio, F.G., Cabanillas, R., Cadinanos, J., Fraile, J.M., Ordonez, G.R., Puente, D.A., Gutierrez-Fernandez, A., Fanjul-Fernandez, M.*, et al.* (2011). Exome sequencing and functional analysis identifies BANF1 mutation as the cause of a hereditary progeroid syndrome. Am. J. Hum. Genet. *88*, 650-656.

Queiroz, R.M.L., Smith, T., Villanueva, E., Monti, M., Pizzinga, M., Marti-Solano, M., Mirea, D.-M., Ramakrishna, M., Harvey, R.F., Dezi, V.*, et al.* (2018). Unbiased dynamic characterization of RNA-protein interactions by OOPS. bioRxiv *333336*.

Rockel, B., Kopec, K.O., Lupas, A.N., and Baumeister, W. (2012). Structure and function of tripeptidyl peptidase II, a giant cytosolic protease. Biochim. Biophys. Acta *1824*, 237-245.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res. *38*, D497-501.

Schaffler, K., Schulz, K., Hirmer, A., Wiesner, J., Grimm, M., Sickmann, A., and Fischer, U. (2010). A stimulatory role for the La-related protein 4B in translation. RNA *16*, 1488-1499.

Schellhaus, A.K., Moreno-Andres, D., Chugh, M., Yokoyama, H., Moschopoulou, A., De, S., Bono, F., Hipp, K., Schaffer, E., and Antonin, W. (2017). Developmentally Regulated GTP binding protein 1 (DRG1) controls microtubule dynamics. Sci. Rep. *7*, 9996.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B.*, et al.* (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676-682.

Schonegge, A.M., Villa, E., Forster, F., Hegerl, R., Peters, J., Baumeister, W., and Rockel, B. (2012). The structure of human tripeptidyl peptidase II as determined by a hybrid approach. Structure *20*, 593-603.

Scotter, E.L., Chen, H.J., and Shaw, C.E. (2015). TDP-43 Proteinopathy and ALS: Insights into Disease Mechanisms and Therapeutic Targets. Neurotherapeutics *12*, 352-363.

Spagnolo, L., Rivera-Calzada, A., Pearl, L.H., and Llorca, O. (2006). Three-dimensional structure of the human DNA-PKcs/Ku70/Ku80 complex assembled on DNA and its implications for DNA DSB repair. Mol Cell *22*, 511-519.

Spiess, C., Meyer, A.S., Reissmann, S., and Frydman, J. (2004). Mechanism of the eukaryotic chaperonin: protein folding in the chamber of secrets. Trends Cell Biol. *14*, 598-604.

Suzuki, T., Behnam, M., Ronasian, F., Salehi, M., Shiina, M., Koshimizu, E., Fujita, A., Sekiguchi, F., Miyatake, S., Mizuguchi, T.*, et al.* (2018). A homozygous NOP14 variant is likely to cause recurrent pregnancy loss. J. Hum. Genet. *63*, 425-430.

Takagi, M., Absalon, M.J., McLure, K.G., and Kastan, M.B. (2005). Regulation of p53 translation and induction after DNA damage by ribosomal protein L26 and nucleolin. Cell *123*, 49-63.

Tanaka, A.J., Cho, M.T., Millan, F., Juusola, J., Retterer, K., Joshi, C., Niyazov, D., Garnica, A., Gratz, E., Deardorff, M.*, et al.* (2015). Mutations in SPATA5 Are Associated with Microcephaly, Intellectual Disability, Seizures, and Hearing Loss. Am. J. Hum. Genet. *97*, 457-464.

Tang, Y., Wang, J., Lian, Y., Fan, C., Zhang, P., Wu, Y., Li, X., Xiong, F., Li, X., Li, G.*, et al.* (2017). Linking long non-coding RNAs and SWI/SNF complexes to chromatin remodeling in cancer. Mol. Cancer *16*, 42.

The UniProt, C. (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res *45*, D158-D169.

Treiber, T., Treiber, N., Plessmann, U., Harlander, S., Daiss, J.L., Eichner, N., Lehmann, G., Schall, K., Urlaub, H., and Meister, G. (2017). A Compendium of RNA-Binding Proteins that Regulate MicroRNA Biogenesis. Mol. Cell *66*, 270-284 e213.

Trendel, J., Schwarzl, T., Prakash, A., Bateman, A., Hentze, M.W., and Krijgsveld, J. (2018). The Human RNA-Binding Proteome and Its Dynamics During Arsenite-Induced Translational Arrest. bioRxiv *329995*.

Vedadi, M., Blazer, L., Eram, M.S., Barsyte-Lovejoy, D., Arrowsmith, C.H., and Hajian, T. (2017). Targeting human SET1/MLL family of proteins. Protein Sci. *26*, 662-676.

Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J., and Geschwind, D.H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature *474*, 380-384.

Vorum, H., Jacobsen, C., and Honore, B. (2000). Calumenin interacts with serum amyloid P component. FEBS Lett. *465*, 129-134.

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., Kwan, J., Bezginov, A.*, et al.* (2015). Panorama of ancient metazoan macromolecular complexes. Nature *525*, 339-344.

Wang, T., Song, W., Chen, Y., Chen, R.B., Liu, Z., Wu, L.C., Li, M.C., Yang, J., Wang, L., Liu, J.H.*, et al.* (2016). Flightless I Homolog Represses Prostate Cancer Progression through Targeting Androgen Receptor Signaling. Clin. Cancer Res. *22*, 1531-1544.

Weick, E.M., Puno, M.R., Januszyk, K., Zinder, J.C., DiMattia, M.A., and Lima, C.D. (2018). Helicase-Dependent RNA Decay Illuminated by a Cryo-EM Structure of a Human Nuclear RNA Exosome-MTR4 Complex. Cell *173*, 1663-1677 e1621.

White, E.A., and Glotzer, M. (2012). Centralspindlin: At the Heart of Cytokinesis. Cytoskeleton *69*, 882-892.

Wilson, S.A., Brown, E.C., Kingsman, A.J., and Kingsman, S.M. (1998). TRIP: a novel double stranded RNA binding protein which interacts with the leucine rich repeat of Flightless I. Nucleic Acids Res. *26*, 3460-3467.

Wood, M.P., Hollis, A., Severance, A.L., Karrick, M.L., and Schisa, J.A. (2016). RNAi Screen Identifies Novel Regulators of RNP Granules in the *Caenorhabditis elegans* Germ Line. G3 (Bethesda) *6*, 2643-2654.

Xie, Q.Y., Chen, S.W., Tian, R.Y., Huang, X., Deng, R.L., Xue, B.B., Qin, Y.W., Xu, Y., Wang, J.J., Guo, M.M.*, et al.* (2018). Long Noncoding RNA ITPRIP-1 Positively Regulates the Innate Immune Response through Promotion of Oligomerization and Activation of MDA5. J. Virol. *92*, e00507-18.

Xue, Y.T., Wong, J.M., Moreno, G.T., Young, M.K., Cote, J., and Wang, W.D. (1998). NURD, a novel complex with both ATP-dependent chromatin-remodeling and histone deacetylase activities. Mol. Cell *2*, 851-861.

Yamamoto, S., Ogasawara, N., Yamamoto, K., Uemura, C., Takaya, Y., Shiraishi, T., Sato, T., Hashimoto, S., Tsutsumi, H., Takano, K.*, et al.* (2017). Mitochondrial proteins NIP-SNAP-1 and-2 are a target for the immunomodulatory activity of clarithromycin, which involves NF-kappa B-mediated cytokine production. Biochem. Bioph. Res. Co. *483*, 911-916.

Yang, R.Q., Gaidamakov, S.A., Xie, J.W., Lee, J., Martino, L., Kozlov, G., Crawford, A.K., Russo, A.N., Conte, M.R., Gehring, K.*, et al.* (2011). La-Related Protein 4 Binds Poly(A), Interacts with the Poly(A)-Binding Protein MLLE Domain via a Variant PAM2w Motif, and Can Promote mRNA Stability. Mol. Cell. Biol. *31*, 542-556.

Yao, N.Y., and O'Donnell, M. (2012). The RFC clamp loader: structure and function. Subcell Biochem *62*, 259-279.

Yoshida, Y., Izumi, H., Torigoe, T., Ishiguchi, H., Yoshida, T., Itoh, H., and Kohno, K. (2004). Binding of RNA to p53 regulates its oligomerization and DNA-binding activity. Oncogene *23*, 4371-4379.

Yuce, O., Piekny, A., and Glotzer, M. (2005). An ECT2-centralspindlin complex regulates the localization and function of RhoA. J. Cell Biol. *170*, 571-582.

Yue, S., Serra, H.G., Zoghbi, H.Y., and Orr, H.T. (2001). The spinocerebellar ataxia type 1 protein, ataxin-1, has RNA-binding activity that is inversely affected by the length of its polyglutamine tract. Hum. Mol. Genet. *10*, 25-30.

Yuzhakov, A., Kelman, Z., and O'Donnell, M. (1999). Trading places on DNA--a three-point switch underlies primer handoff from primase to the replicative DNA polymerase. Cell *96*, 153-163.

Zaghlool, A., Halvardson, J., Zhao, J.J., Etemadikhah, M., Kalushkova, A., Konska, K., Jernberg-Wiklund, H., Thuresson, A.C., and Feuk, L. (2016). A Role for the Chromatin-Remodeling Factor BAZ1A in Neurodevelopment. Hum. Mutat. *37*, 964-975.

Zemp, I., Wild, T., O'Donohue, M.F., Wandrey, F., Widmann, B., Gleizes, P.E., and Kutay, U. (2009). Distinct cytoplasmic maturation steps of 40S ribosomal subunit precursors require hRio2. J. Cell Biol. *185*, 1167-1180.

Zhao, L.Y., Liu, J., Sidhu, G.S., Niu, Y., Liu, Y., Wang, R., and Liao, D. (2004). Negative regulation of p53 functions by Daxx and the involvement of MDM2. J. Biol. Chem. *279*, 50566-50579.

Zheng, R., Shen, Z., Tripathi, V., Xuan, Z., Freier, S.M., Bennett, C.F., Prasanth, S.G., and Prasanth, K.V. (2010). Polypurine-repeat-containing RNAs: a novel class of long non-coding RNA in mammalian cells. J. Cell Sci. *123*, 3734-3744.

Figure 1

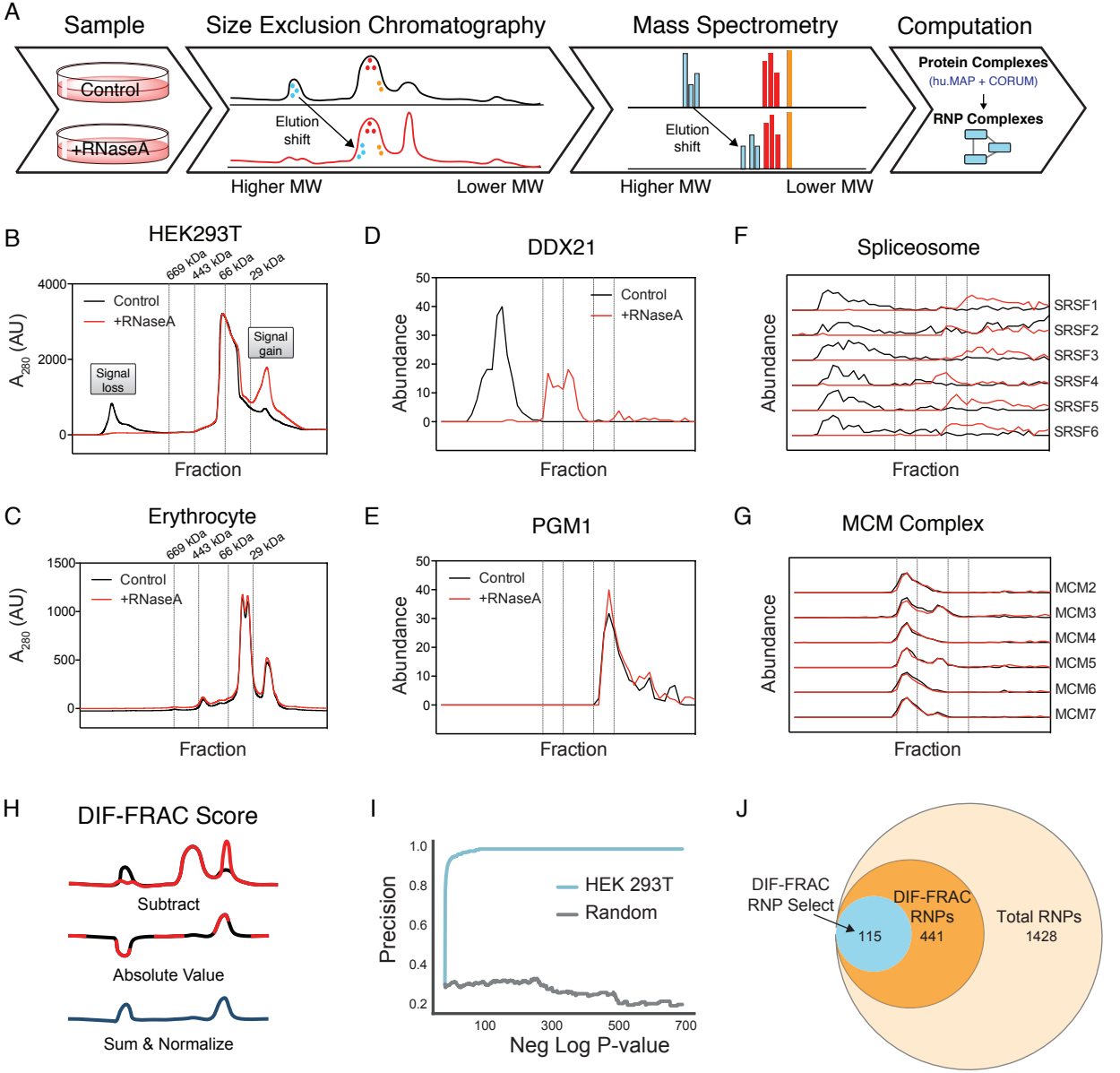

A
Sample
Size Exclusion Chromatography
Mass Spectrometry
Computation
Control
+RNaseA
Elution shift
Higher MW
Lower MW
Protein Complexes
(hu.MAP + CORUM)
RNP Complexes
B
HEK293T
669 kDa
443 kDa
66 kDa
29 kDa
Control
+RNaseA
Signal loss
Signal gain
A280 (AU)
Fraction
C
Erythrocyte
D
DDX21
Abundance
E
PGM1
F
Spliceosome
SRSF1
SRSF2
SRSF3
SRSF4
SRSF5
SRSF6
G
MCM Complex
MCM2
MCM3
MCM4
MCM5
MCM6
MCM7
H
DIF-FRAC Score
Subtract
Absolute Value
Sum & Normalize
I
Precision
HEK 293T
Random
Neg Log P-value
J
DIF-FRAC RNP Select
115
DIF-FRAC RNPs
441
Total RNPs
1428

Figure 2

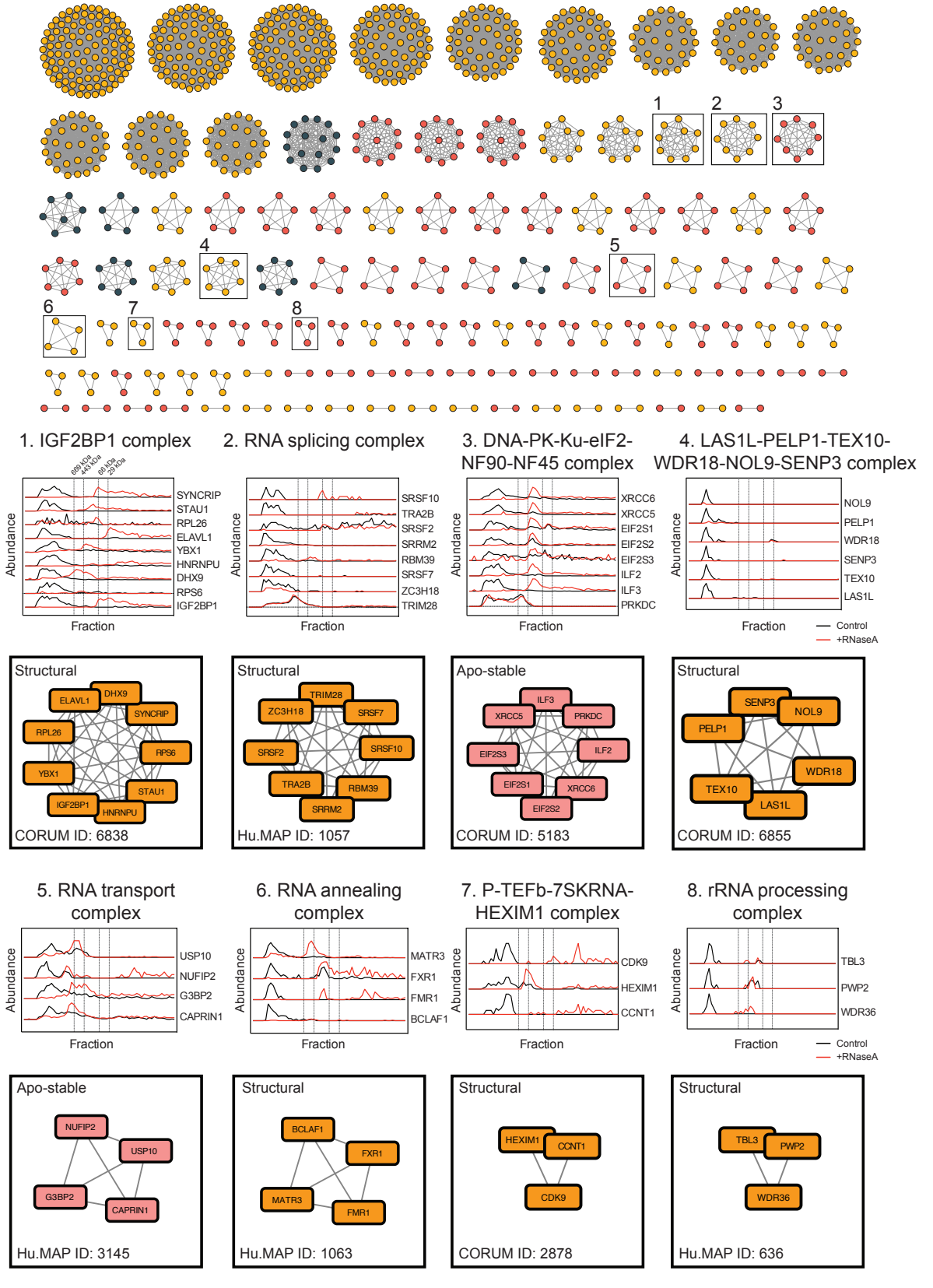

1
2
3
4
5
6
7
8
1. IGF2BP1 complex
669 kDa
443 kDa
66 kDa
29 kDa
SYNCRIP
STAU1
RPL26
ELAVL1
YBX1
HNRNPU
DHX9
RPS6
IGF2BP1
Abundance
Fraction
2. RNA splicing complex
SRSF10
TRA2B
SRSF2
SRRM2
RBM39
SRSF7
ZC3H18
TRIM28
3. DNA-PK-Ku-eIF2-NF90-NF45 complex
XRCC6
XRCC5
EIF2S1
EIF2S2
EIF2S3
ILF2
ILF3
PRKDC
4. LAS1L-PELP1-TEX10-WDR18-NOL9-SENP3 complex
NOL9
PELP1
WDR18
SENP3
TEX10
LAS1L
Control
+RNaseA
Structural
CORUM ID: 6838
Structural
Hu.MAP ID: 1057
Apo-stable
CORUM ID: 5183
Structural
CORUM ID: 6855
5. RNA transport complex
USP10
NUFIP2
G3BP2
CAPRIN1
6. RNA annealing complex
MATR3
FXR1
FMR1
BCLAF1
7. P-TEFb-7SKRNA-HEXIM1 complex
CDK9
HEXIM1
CCNT1
8. rRNA processing complex
TBL3
PWP2
WDR36
Apo-stable
Hu.MAP ID: 3145
Structural
Hu.MAP ID: 1063
Structural
CORUM ID: 2878
Structural
Hu.MAP ID: 636

Figure 3

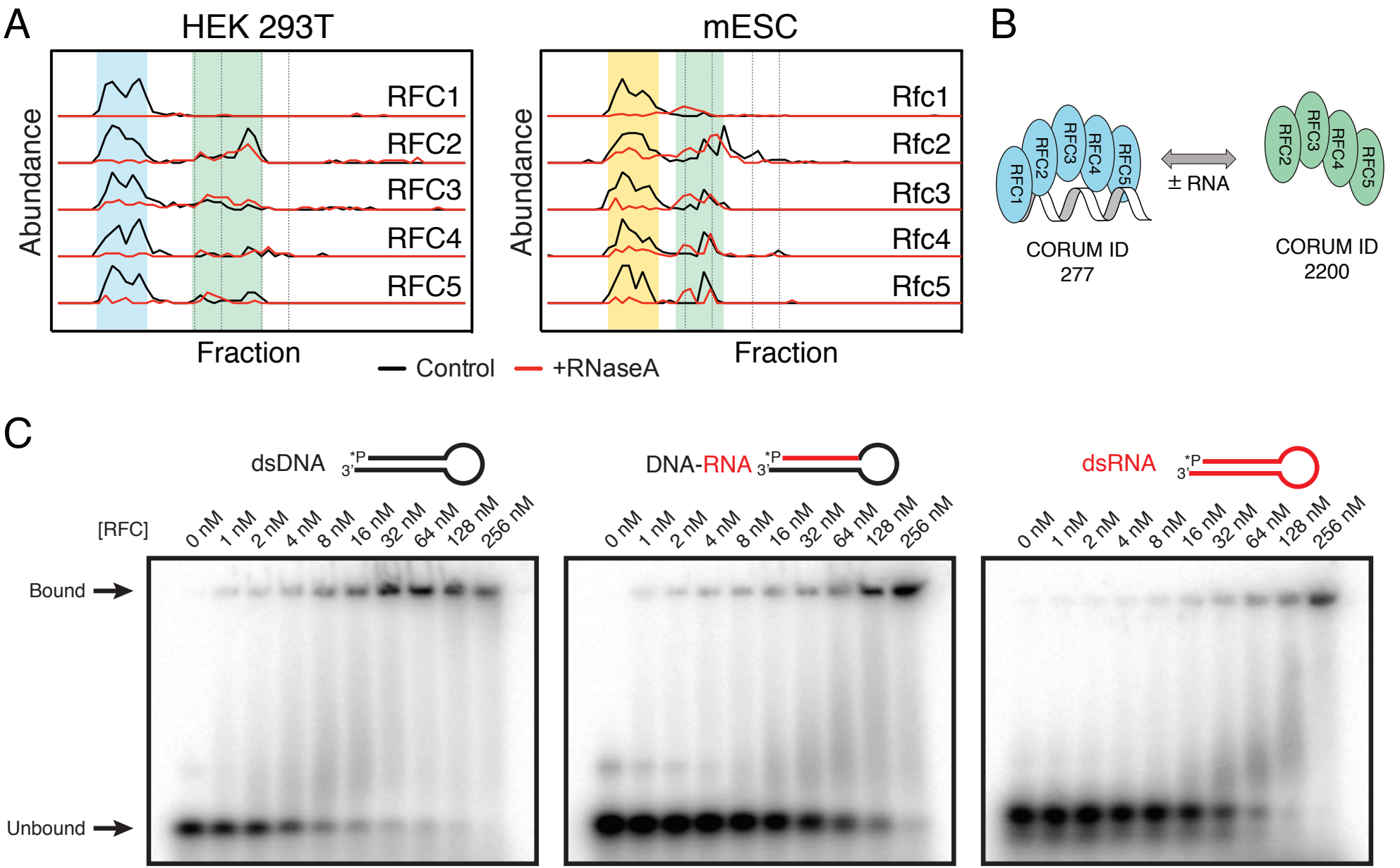

A
HEK 293T
mESC
Abundance
RFC1
RFC2
RFC3
RFC4
RFC5
Rfc1
Rfc2
Rfc3
Rfc4
Rfc5
Fraction
Control
+RNaseA
B
RFC1
RFC2
RFC3
RFC4
RFC5
± RNA
CORUM ID
277
CORUM ID
2200
C
dsDNA
DNA-RNA
dsRNA
*P
3'
[RFC]
0 nM
1 nM
2 nM
4 nM
8 nM
16 nM
32 nM
64 nM
128 nM
256 nM
Bound
Unbound

Figure 4

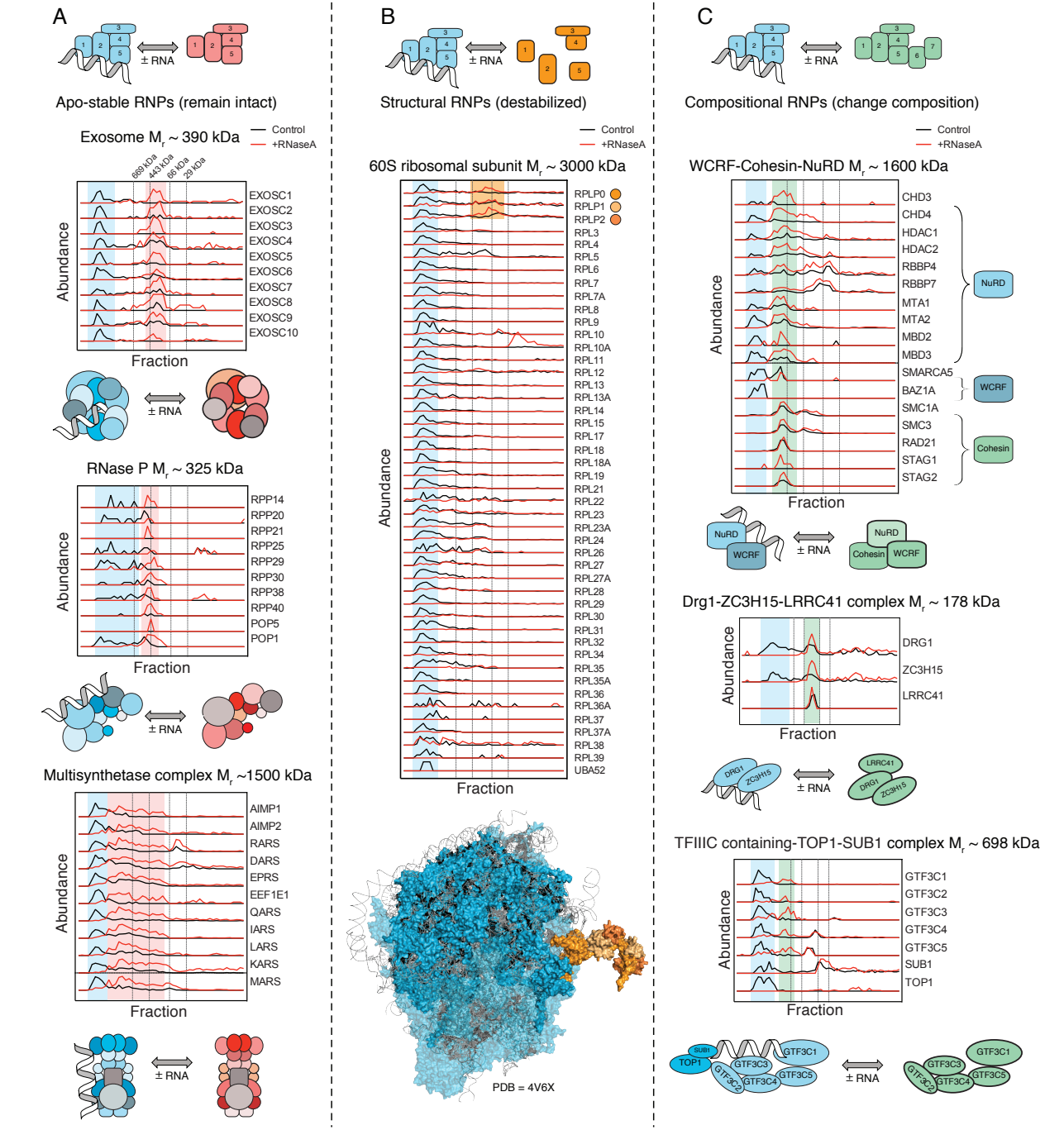

A
± RNA
Apo-stable RNPs (remain intact)
Exosome M_r ~ 390 kDa
Control
+RNaseA
689 kDa
443 kDa
66 kDa
29 kDa
Abundance
EXOSC1
EXOSC2
EXOSC3
EXOSC4
EXOSC5
EXOSC6
EXOSC7
EXOSC8
EXOSC9
EXOSC10
Fraction
± RNA
RNase P M_r ~ 325 kDa
Abundance
RPP14
RPP20
RPP21
RPP25
RPP29
RPP30
RPP38
RPP40
POP5
POP1
Fraction
± RNA
Multisynthetase complex M_r ~1500 kDa
Abundance
AIMP1
AIMP2
RARS
DARS
EPRS
EEF1E1
QARS
IARS
LARS
KARS
MARS
Fraction
± RNA
B
± RNA
Structural RNPs (destabilized)
Control
+RNaseA
60S ribosomal subunit M_r ~ 3000 kDa
Abundance
RPLP0
RPLP1
RPLP2
RPL3
RPL4
RPL5
RPL6
RPL7
RPL7A
RPL8
RPL9
RPL10
RPL10A
RPL11
RPL12
RPL13
RPL13A
RPL14
RPL15
RPL17
RPL18
RPL18A
RPL19
RPL21
RPL22
RPL23
RPL23A
RPL24
RPL26
RPL27
RPL27A
RPL28
RPL29
RPL30
RPL31
RPL32
RPL34
RPL35
RPL35A
RPL36
RPL36A
RPL37
RPL37A
RPL38
RPL39
UBA52
Fraction
PDB = 4V6X
C
± RNA
Compositional RNPs (change composition)
Control
+RNaseA
WCRF-Cohesin-NuRD M_r ~ 1600 kDa
Abundance
CHD3
CHD4
HDAC1
HDAC2
RBBP4
RBBP7
MTA1
MTA2
MBD2
MBD3
NuRD
SMARCA5
BAZ1A
SMC1A
WCRF
SMC3
RAD21
STAG1
STAG2
Cohesin
Fraction
NuRD
WCRF
± RNA
NuRD
Cohesin
WCRF
Drg1-ZC3H15-LRRC41 complex M_r ~ 178 kDa
Abundance
DRG1
ZC3H15
LRRC41
Fraction
DRG1
ZC3H15
± RNA
LRRC41
DRG1
ZC3H15
TFIIIC containing-TOP1-SUB1 complex M_r ~ 698 kDa
Abundance
GTF3C1
GTF3C2
GTF3C3
GTF3C4
GTF3C5
SUB1
TOP1
Fraction
SUB1
TOP1
GTF3C3
GTF3C1
GTF3C2
GTF3C4
GTF3C5
± RNA
GTF3C1
GTF3C3
GTF3C2
GTF3C4
GTF3C5

Figure 5

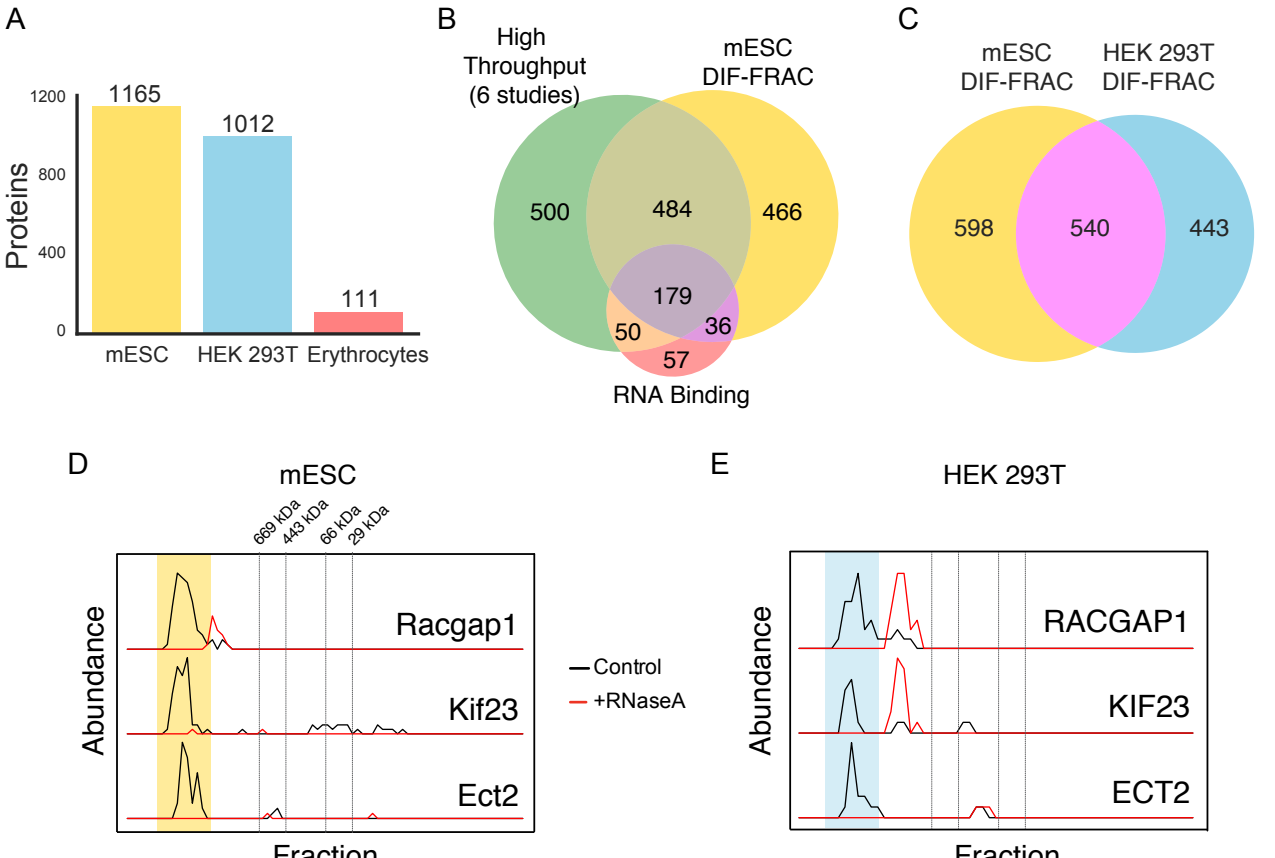

A
Proteins
1200
800
400
0
1165
1012
111
mESC
HEK 293T
Erythrocytes
B
High Throughput (6 studies)
mESC DIF-FRAC
500
484
466
179
50
36
57
RNA Binding
C
mESC DIF-FRAC
HEK 293T DIF-FRAC
598
540
443
D
mESC
669 kDa
443 kDa
66 kDa
29 kDa
Abundance
Racgap1
Kif23
Ect2
Control
+RNaseA
Fraction
E
HEK 293T
Abundance
RACGAP1
KIF23
ECT2
Fraction

Figure 6

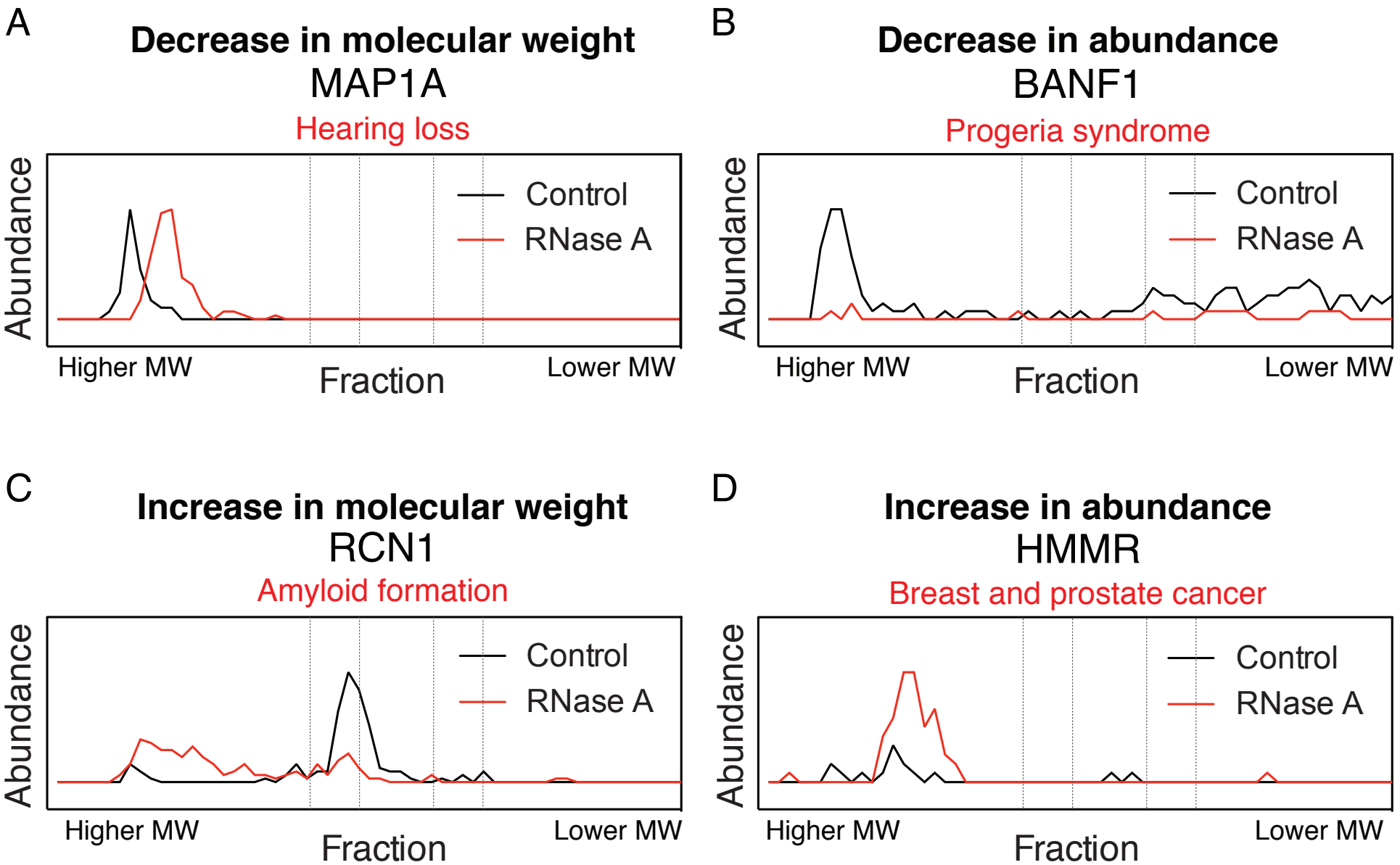
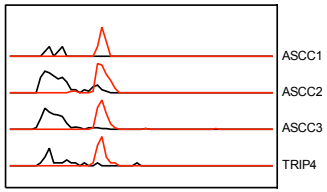
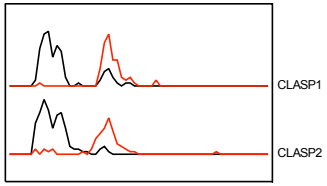
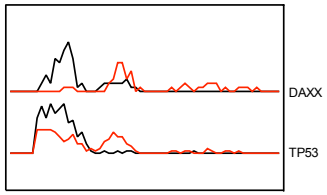
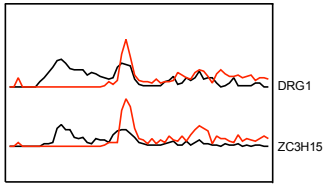
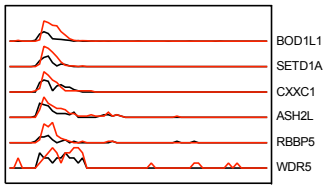
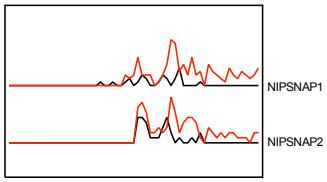
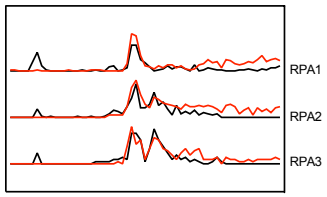
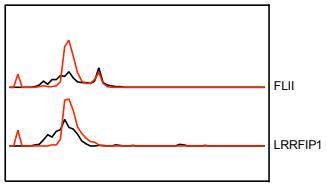
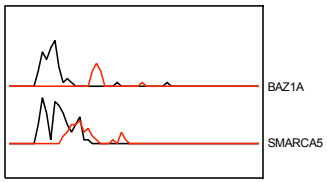
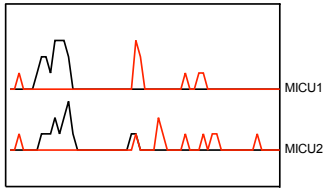
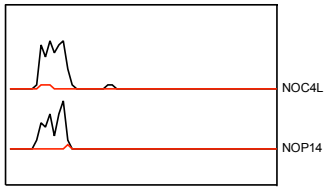
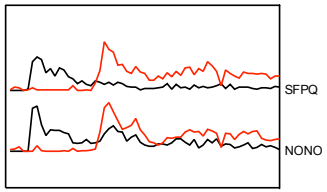
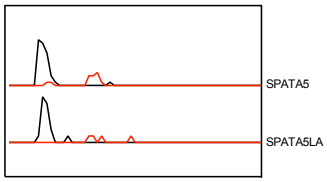
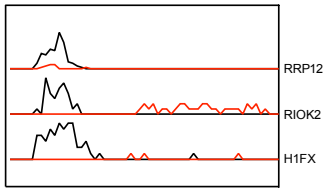
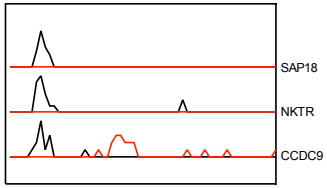
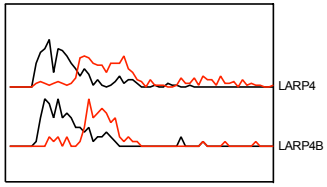
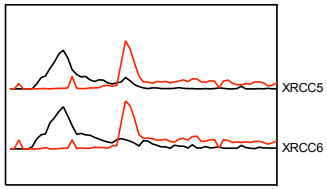
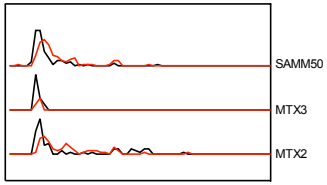
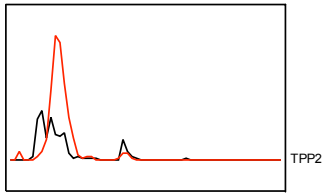

A
Decrease in molecular weight
MAP1A
Hearing loss
Abundance
Control
RNase A
Higher MW
Fraction
Lower MW
B
Decrease in abundance
BANF1
Progeria syndrome
Abundance
Control
RNase A
Higher MW
Fraction
Lower MW
C
Increase in molecular weight
RCN1
Amyloid formation
Abundance
Control
RNase A
Higher MW
Fraction
Lower MW
D
Increase in abundance
HMMR
Breast and prostate cancer
Abundance
Control
RNase A
Higher MW
Fraction
Lower MW

Table 1: Stable RNPs identified by DIF-FRAC

Gene Names*	Complex Name	Function	Soluble without RNA? $	Disease links	CORUM /Hu.Map #	RNA.MAP id	DIF-FRAC plot	RNP class †	Refs.
ASCC1 ASCC2 ASCC3 TRIP4	Human activating signal cointegrator 1 (hASC-1) complex	Transcription coactivator Cellular signaling	Yes	Spinal muscular atrophy	Yes/Yes	4984		Apo-stable	(Jung et al., 2002) (Knierim et al., 2016)
CLASP1 CLASP2	N/A	Microtubule binding Microtubule dynamics	Yes	N/A	No/Yes	3807		Apo-stable	(Efimov et al., 2007)
DAXX TP53	DAXX-TP53 complex	Transcription repression	Yes	Pancreatic neuroendocrine tumors Glioblastoma multiforme Adrenocortical tumors	Yes/No	4518		Apo-stable	(Zhao et al., 2004) (Dyer et al., 2017)
DRG1 ZC3H15	Drg1/Dfrp1 complex	Microtubule binding Microtubule polymerase GTPase	Yes	Lung adenocarcinoma	No/No	4096		Apo-stable	(Lu et al., 2016) (Schellhaus et al., 2017) (Ishikawa et al., 2009)

BOD1L1 SETD1A CXXC1* ASH2L RBBP5 WDR5	SET1A/SET1B complexes	Histone methyltransferase Transcription regulation	Yes	Fanconi anaemia Mixed lineage leukemia	Yes/Yes	2005, 3307		Apo-stable	(Vedadi et al., 2017) (Higgs et al., 2015) (Brown et al., 2017)
NIPSNAP1 NIPSNAP2	N/A	Vesicular transport	No	Inflammatory pain	No/Yes	5822		Apo-stable	(Okuda-Ashitaka et al., 2012) (Yamamoto et al., 2017)
RPA1 RPA2 RPA3*	Replication protein A complex	Single-stranded DNA binding DNA metabolism	Yes	Werner Syndrome	Yes/Yes	3204		Apo-stable	(Machwe et al., 2011) (Fan and Pavletich, 2012)
FLII LRRFIP1	FLII-LRRFIP1 complex	Transcriptional activation Actin binding	Yes	Prostate cancer	No/Yes	3626		Apo-stable	(Wilson et al., 1998) (Wang et al., 2016)
BAZ1A SMARCA5	WCRF complex	Chromatin remodeling	No	Intellectual disability	Yes/No	2105		Compositional	(Bochar et al., 2000) (Zaghlool et al., 2016)

MICU1* MICU2*	MICU1-MICU2 heterodimer	Calcium ion transport	Yes	Myopathy with extrapyramidal signs	Yes/Yes	4318		Structural	(Logan et al., 2014)
NOC4L NOP14	N/A	Ribosome processing and biogenesis	No	Recurrent pregnancy loss	No/Yes	4220		Structural	(Suzuki et al., 2018)
SFPQ NONO	PSF-p54(nrb) complex	Spicing factor DNA recombination	Yes	Intellectual disability	Yes/Yes	327		Apo-stable	(Bladen et al., 2005) (Mircsof et al., 2015)
SPATA5 SPATA5L1*	N/A	Spermatogenesis	No	Epilepsy, hearing loss, and mental retardation syndrome	No/Yes	5745		Structural	(Tanaka et al., 2015)
RRP12 RIOK2 H1FX	N/A	rRNA processing	No	N/A	No/Yes	5795		Structural	(Zemp et al., 2009)

SAP18 NKTR CCDC9	N/A	SAP18 is involved in RNA processing and splicing NKTR is involved in protein peptidyl-prolyl isomerization	No	N/A	No/Yes	6027		Structural	(Davis et al., 2010)
LARP4 LARP4B	N/A	Translation regulation	Yes	N/A	No/Yes	3327		Apo-stable	(Yang et al., 2011) (Schaffler et al., 2010)
XRCC5 XRCC6	Ku antigen complex	DNA damage and repair	Yes	Systemic lupus erythematosus	Yes/No	2930		Apo-stable	(Spagnolo et al., 2006)
SAMM50* MTX3* MTX2	N/A	Protein transport	Yes	N/A	No/Yes	1450		Apo-stable	
TPP2	Tripeptidyl-peptidase II	Serine protease	Yes	Muscle wasting Obesity Cancer	Yes/No			Apo-stable	(Schonegge et al., 2012) (Rockel et al., 2012)

* RNA-associated proteins newly identified by DIF-FRAC are shown in red. Those with an asterisk are above the 5% FDR cutoff. RNA-associated proteins in black have been previously annotated or identified by high-throughput methods (See Table S1).
# Evidence for all or some protein complex subunits interacting in CORUM or Hu.Map

$ Insolubility in the absence of RNA is inferred by an increase in apparent molecular weight of the complex upon RNA digestion, or a complete disappearance of signal. This is consistent with the RNP being solubilized by RNA, as suggested by Maharana et al (Maharana et al., 2018)
† RNP classes apo-stable, structural, or compositional as described in Fig. 4.

Figure S1

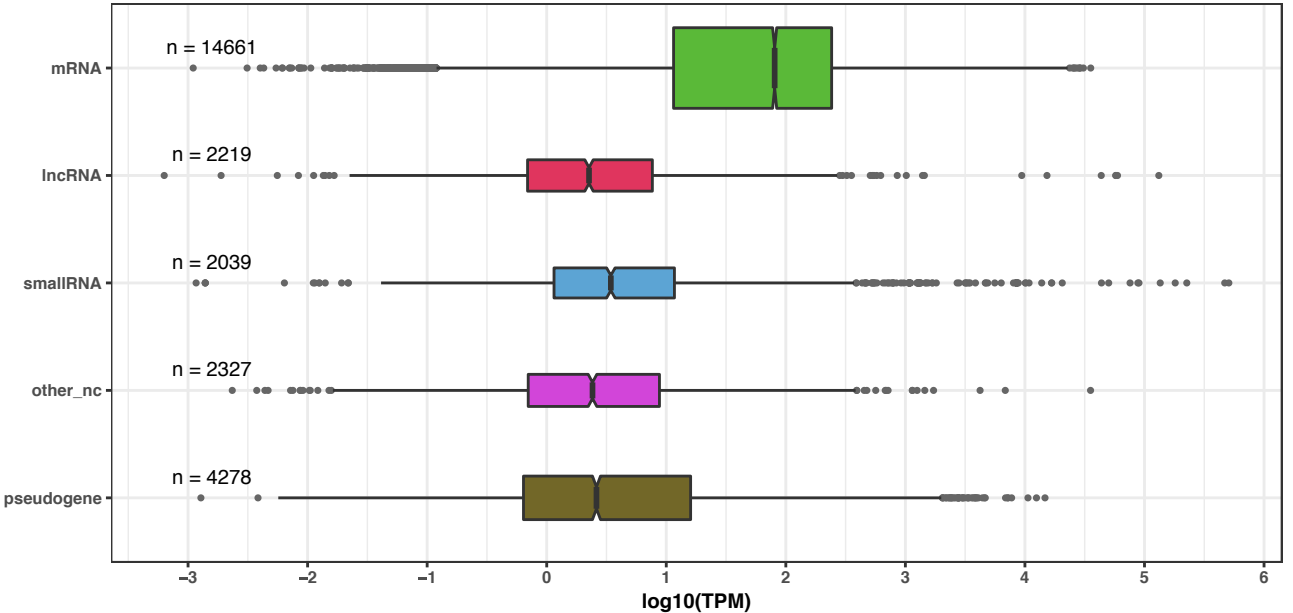

n = 14661
mRNA
n = 2219
lncRNA
n = 2039
smallRNA
n = 2327
other_nc
n = 4278
pseudogene
−3
−2
−1
0
1
2
3
4
5
6
log10(TPM)

Figure S2

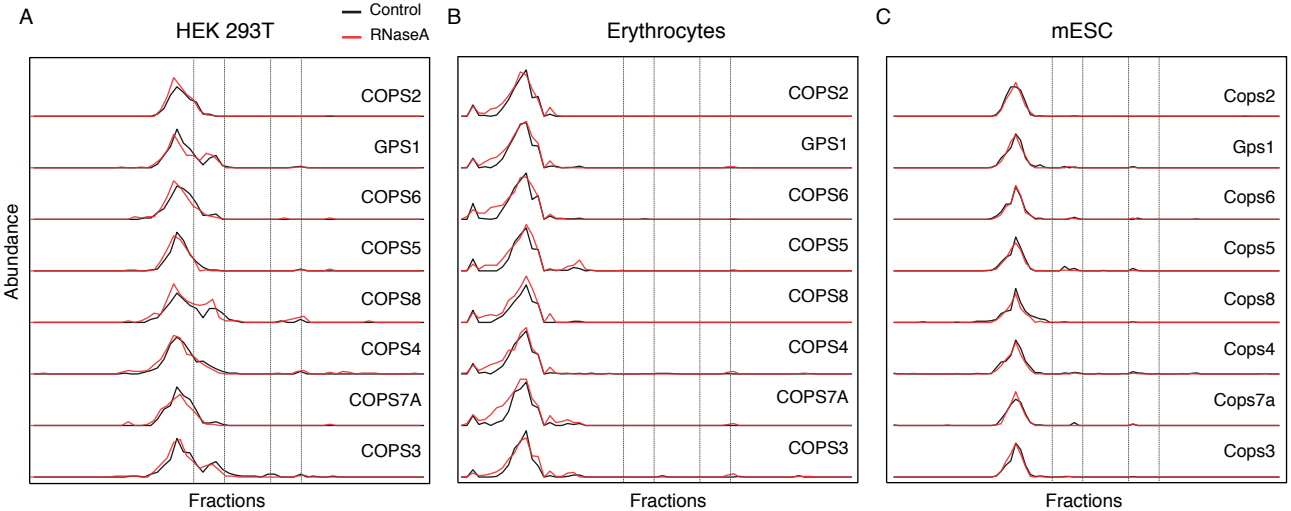
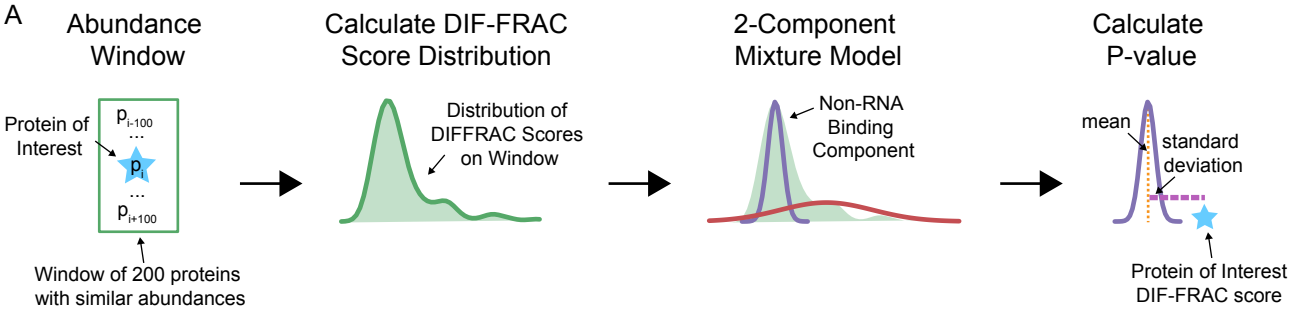

A
HEK 293T
Control
RNaseA
B
Erythrocytes
C
mESC
Abundance
COPS2
GPS1
COPS6
COPS5
COPS8
COPS4
COPS7A
COPS3
Cops2
Gps1
Cops6
Cops5
Cops8
Cops4
Cops7a
Cops3
Fractions
Fractions
Fractions

Figure S3


A
Abundance Window
Calculate DIF-FRAC Score Distribution
2-Component Mixture Model
Calculate P-value
Protein of Interest
p_{i-100}
p_i
p_{i+100}
Window of 200 proteins with similar abundances
Distribution of DIFFRAC Scores on Window
Non-RNA Binding Component
mean
standard deviation
Protein of Interest DIF-FRAC score

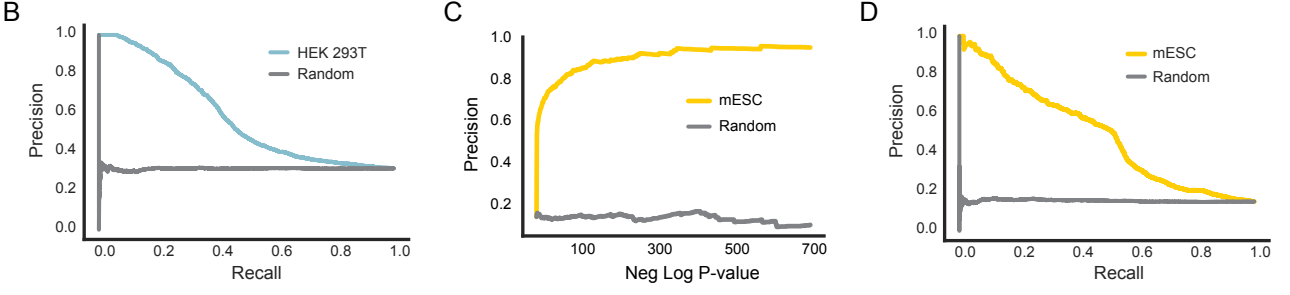

B
Precision
Recall
HEK 293T
Random
C
Precision
Neg Log P-value
mESC
Random
D
Precision
Recall
mESC
Random

Figure S4

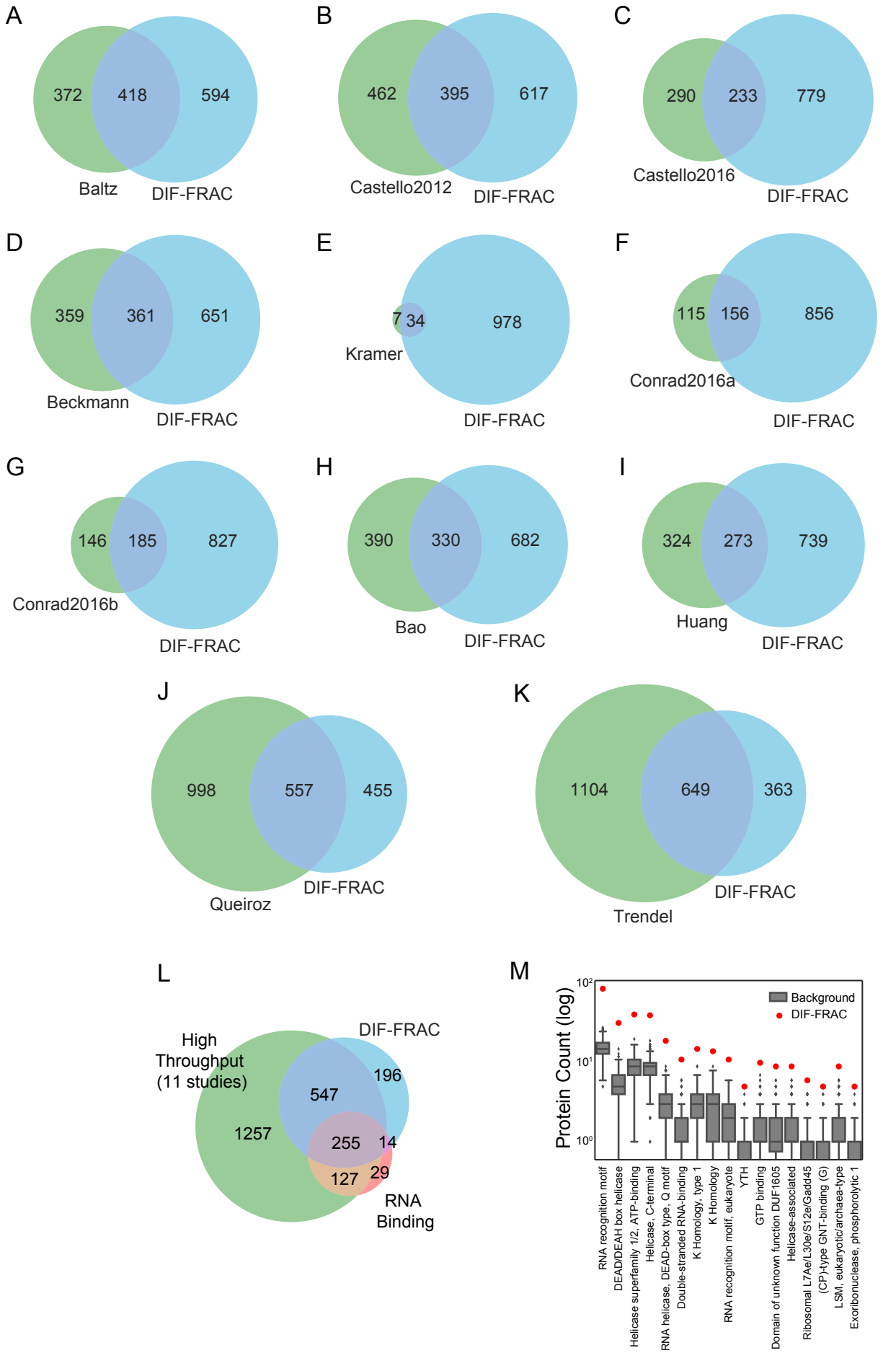

A
372
418
594
Baltz
DIF-FRAC
B
462
395
617
Castello2012
DIF-FRAC
C
290
233
779
Castello2016
DIF-FRAC
D
359
361
651
Beckmann
DIF-FRAC
E
7
34
978
Kramer
DIF-FRAC
F
115
156
856
Conrad2016a
DIF-FRAC
G
146
185
827
Conrad2016b
DIF-FRAC
H
390
330
682
Bao
DIF-FRAC
I
324
273
739
Huang
DIF-FRAC
J
998
557
455
Queiroz
DIF-FRAC
K
1104
649
363
Trendel
DIF-FRAC
L
High Throughput (11 studies)
DIF-FRAC
1257
547
196
255
14
127
29
RNA Binding
M
Protein Count (log)
Background
DIF-FRAC
RNA recognition motif
DEAD/DEAH box helicase
Helicase superfamily 1/2, ATP-binding
Helicase, C-terminal
RNA helicase, DEAD-box type, Q motif
Double-stranded RNA-binding
K Homology, type 1
K Homology
RNA recognition motif, eukaryote
YTH
GTP binding
Domain of unknown function DUF1605
Helicase-associated
Ribosomal L7Ae/L30e/S12e/Gadd45
(CP)-type GNT-binding (G)
LSM, eukaryotic/archaea-type
Exoribonuclease, phosphorolytic 1

Figure S5

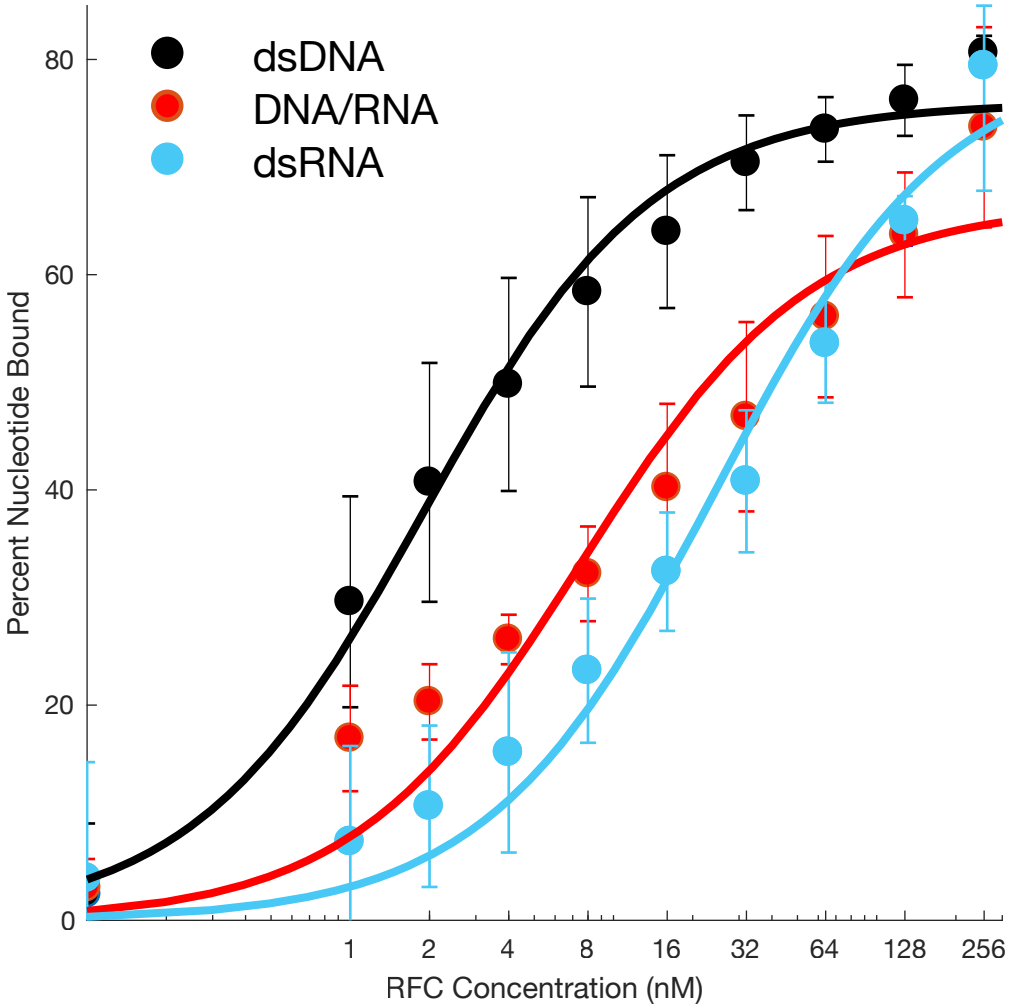

dsDNA
DNA/RNA
dsRNA
Percent Nucleotide Bound
0
20
40
60
80
1
2
4
8
16
32
64
128
256
RFC Concentration (nM)

Figure S6

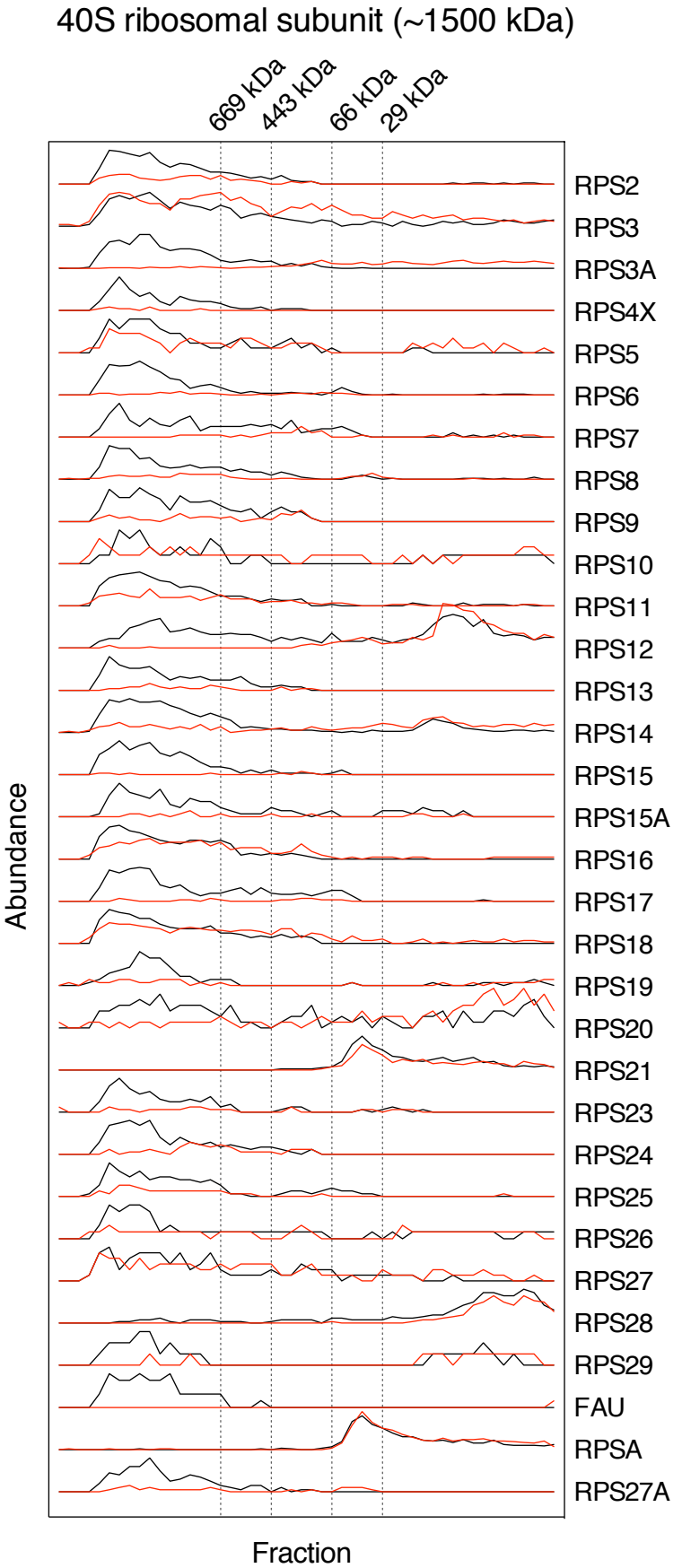

40S ribosomal subunit (~1500 kDa)
669 kDa
443 kDa
66 kDa
29 kDa
RPS2
RPS3
RPS3A
RPS4X
RPS5
RPS6
RPS7
RPS8
RPS9
RPS10
RPS11
RPS12
RPS13
RPS14
RPS15
RPS15A
RPS16
RPS17
RPS18
RPS19
RPS20
RPS21
RPS23
RPS24
RPS25
RPS26
RPS27
RPS28
RPS29
FAU
RPSA
RPS27A
Abundance
Fraction

Figure S7

BAF and PBAF complexes M ~ 2000 kDa

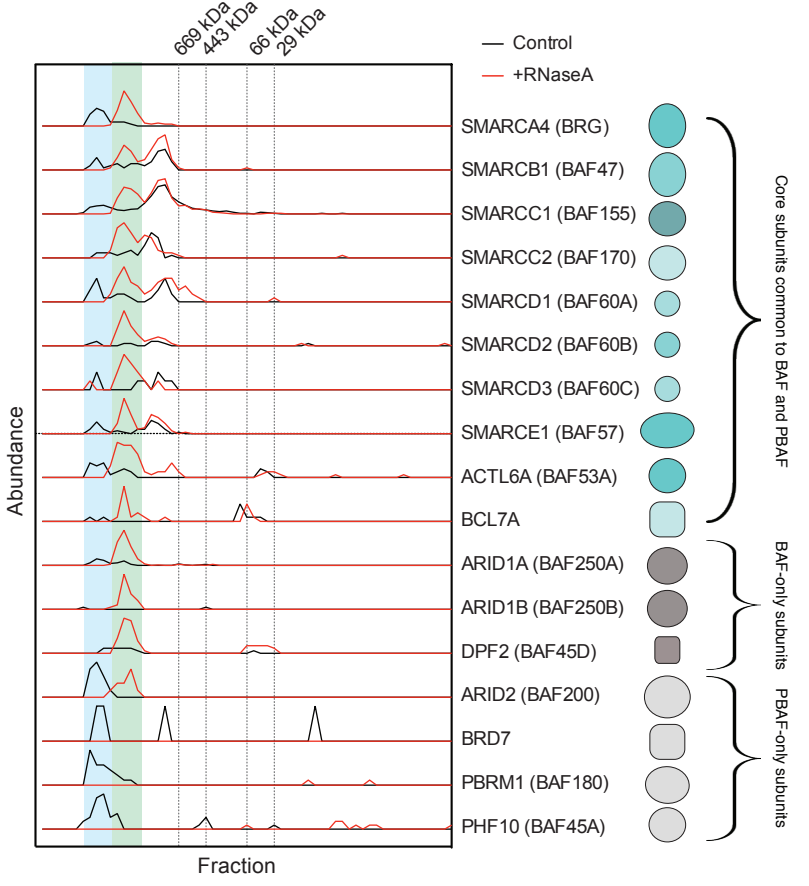


Figure S8

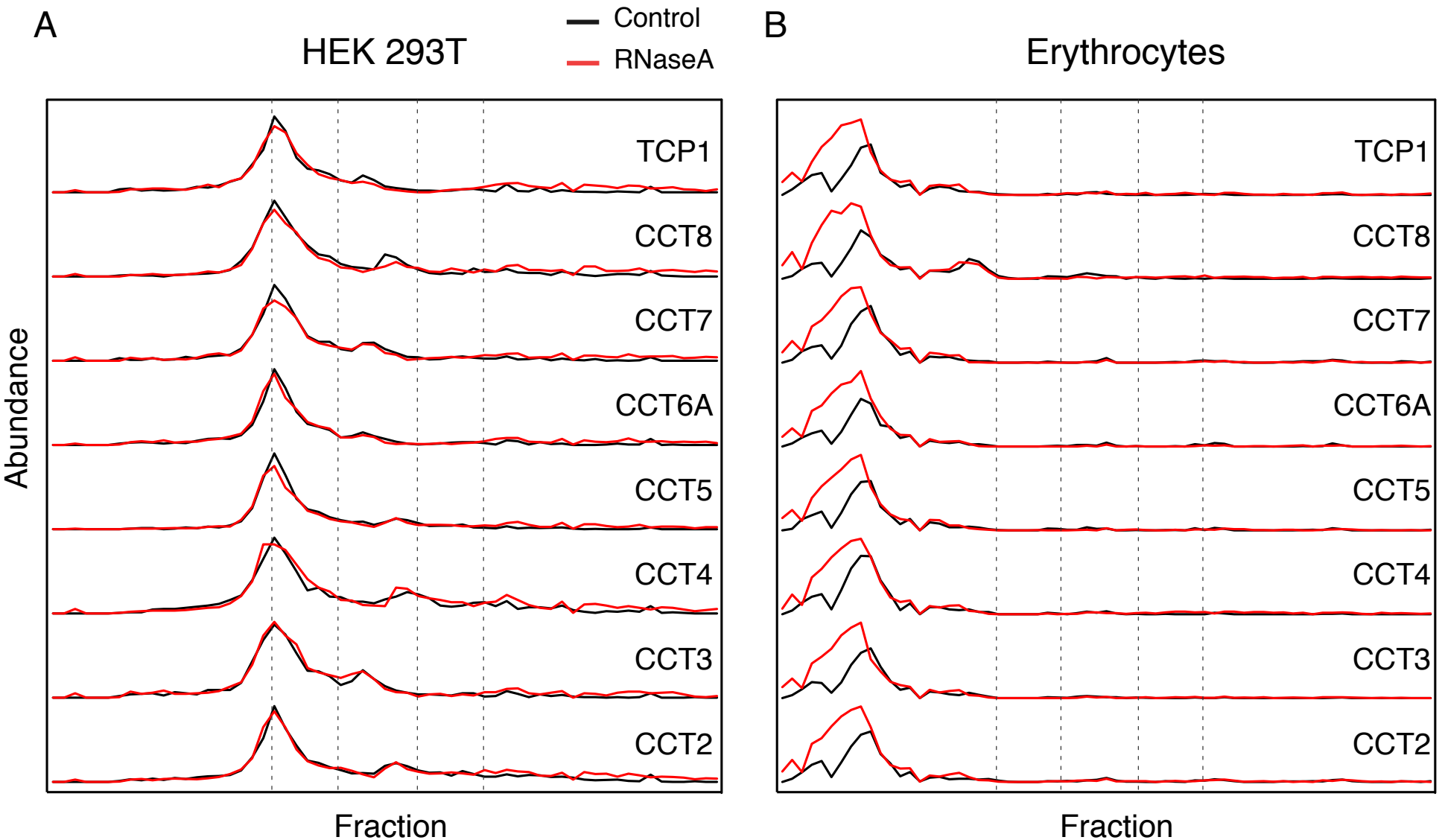
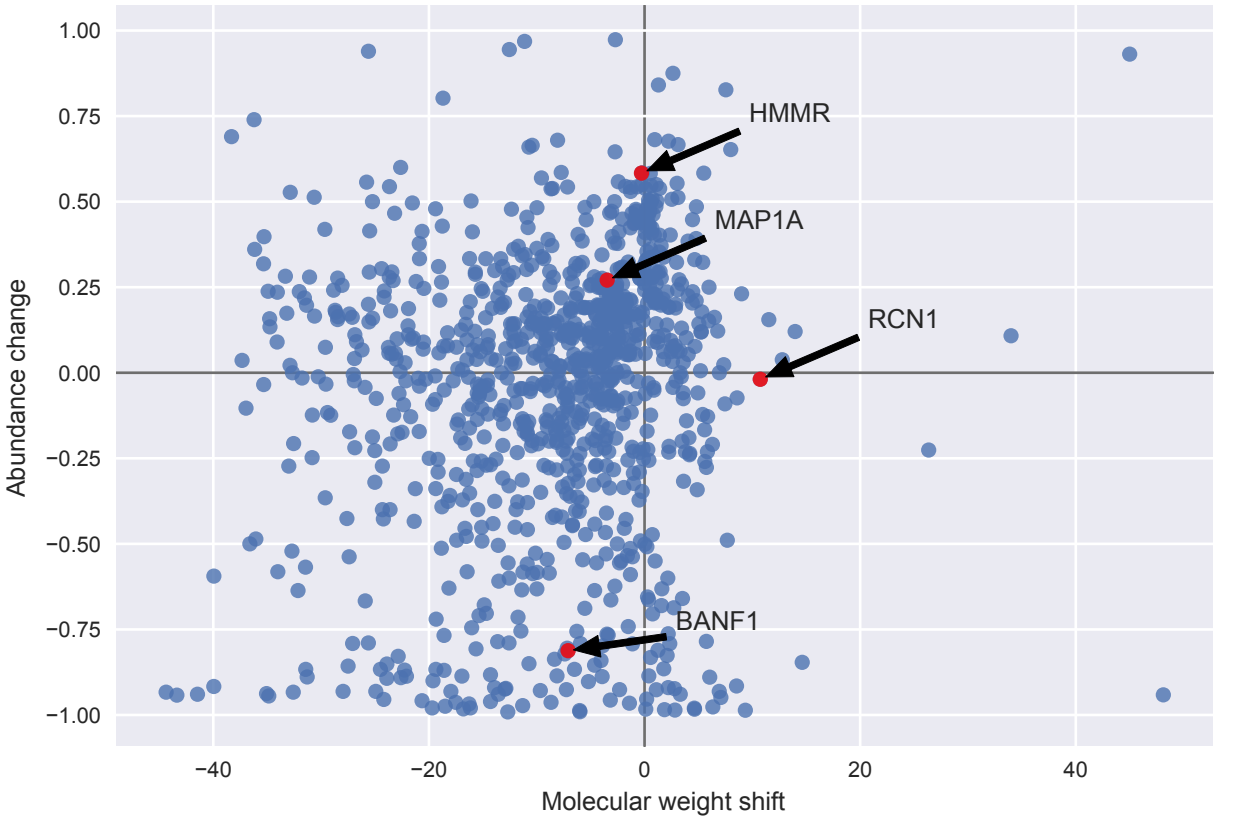

A
HEK 293T
Control
RNaseA
B
Erythrocytes
Abundance
TCP1
CCT8
CCT7
CCT6A
CCT5
CCT4
CCT3
CCT2
Fraction
Fraction

Figure S9


HMMR
MAP1A
RCN1
BANF1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Abundance change
−40
−20
0
20
40
Molecular weight shift