nature communications

Article

https://doi.org/10.1038/s41467-024-52765-w

# Capture of RNA-binding proteins across mouse tissues using HARD-AP

Received: 26 November 2023

Accepted: 20 September 2024

Published online: 28 September 2024

Check for updates

Yijia Ren[1,10], Hongyu Liao[1,10], Jun Yan[2,10], Hongyu Lu[1,10], Xiaowei Mao[3,4,5,10], Chuan Wang[1], Yi-fei Li[1], Yu Liu[6], Chong Chen[7], Lu Chen[1], Xiangfeng Wang[2], Kai-Yu Zhou[1], Han-Min Liu[1], Yi Liu[8], Yi-Min Hua[1] ✉, Lin Yu[1] ✉ & Zhihong Xue[1,9] ✉

RNA-binding proteins (RBPs) modulate all aspects of RNA metabolism, but a comprehensive picture of RBP expression across tissues is lacking. Here, we describe our development of the method we call HARD-AP that robustly retrieves RBPs and tightly associated RNA regulatory complexes from cultured cells and fresh tissues. We successfully use HARD-AP to establish a comprehensive atlas of RBPs across mouse primary organs. We then systematically map RNA-binding sites of these RBPs using machine learning-based modeling. Notably, the modeling reveals that the LIM domain as an RNA-binding domain in many RBPs. We validate the LIM-domain-only protein Csrp1 as a tissue-dependent RNA binding protein. Taken together, HARD-AP is a powerful approach that can be used to identify RBPomes from any type of sample, allowing comprehensive and physiologically relevant networks of RNA-protein interactions.

RNA-binding proteins (RBPs) associate with RNAs into dynamic ribonucleoproteins that modulate all aspects of RNA metabolism including transcription, translation, splicing, modification, intracellular trafficking, and decay[1,2]. Classically, RBPs are categorized based on their canonical RNA-binding domains (RBDs). An early study annotated ~400 mammalian RBPs that harbor 799 individual RBDs[3]. Recent high-throughput approaches have increased the number of recognized RBPs into the four-digit range and have revealed that numerous metabolism-related proteins, especially enzymes, associate with RNAs[4–6]. Given that previous studies were limited to cultured cell lines or primary cells, we are lacking a comprehensive picture of RBPs under physiological states across different tissues that is necessary for an understanding of the physiological connections between metabolism and RNA function.

Current approaches to identify RBPs mainly rely on the cross-linking of RNA-protein complexes followed by capture of the complexes through the polyadenylated tail of the RNA[7], incorporation of modified nucleotides to allow affinity enrichment[8,9], or organic phase-assisted separation of crosslinked RNA-protein complexes[10,11]. The polyA-based capture methods do not work on prokaryotic RNAs and or the eukaryotic species that lack polyA tails, which make up at least 95% of transcribed RNAs[7]. The methods based on the incorporation of

[1]Key Laboratory of Birth Defects and Related Disease of Women and Children of MOE, Department of Pediatrics, West China Second University Hospital, State Key Laboratory of Biotherapy, Sichuan University, Chengdu, Sichuan 610041, China. [2]National Maize Improvement Center, Frontiers Science Center for Molecular Design Breeding, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100094, China. [3]Sichuan Provincial Key Laboratory for Human Disease Gene Study and the Center for Medical Genetics, Department of Laboratory Medicine, Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, Sichuan 610072, China. [4]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China. [5]Shimmer Center, Tianfu Jiangxi Laboratory, Chengdu, Sichuan 641419, China. [6]State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China. [7]Department of Urology, Institute of Urology, State Key Laboratory of Biotherapy and Cancer Center, Sichuan University, Chengdu, Sichuan 610041, China. [8]Department of Physiology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. [9]Development and Related Diseases of Women and Children Key Laboratory of Sichuan Province, Chengdu, Sichuan 610041, China. [10]These authors contributed equally: Yijia Ren, Hongyu Liao, Jun Yan, Hongyu Lu, Xiaowei Mao. ✉e-mail: drhuayimin@scu.edu.cn; yu_scu@scu.edu.cn; xuez@scu.edu.cn

modified nucleotides can potentially address this limitation but cannot be used in tissue samples and may introduce potential biases due to variability in metabolism of different RNAs[8,9]. The methods based on organic phase separation need the UV-induced RNA-protein cross-linking to distinguish RBPs and non-RPBs, whereas the efficiency of UV-induced cross-linking is limited, and even lower for tissue samples[12–15]. In addition, many bona fide RBPs and proteins with post-translational modifications, especially glycosylated proteins, were reported to be trapped in the organic interphase independently of UV cross-linking[10,11]. An alternative way to recover RNA-binding complexes with high efficiency is needed.

To overcome these limitations, we engineered an RNA-binding protein that we call HARD (for high-affinity RNA-binding domain), which has high sequence-independent affinity for RNAs. We used immobilized HARD to develop a capture strategy that we call HARD-AP (HARD-mediated Affinity Purification) to isolate RBPs and tightly associated RNA regulatory complexes, which allow a comprehensive atlas of RBPs and RNA regulatory complexes from any cell or tissue samples with high specificity and sensitivity.

## Results

### Design of the HARD protein and its RNA binding activity

To design a protein that can tightly bind RNAs sequence independently, we analyzed structures of protein-RNA complexes available in the RCSB PDB database[16]. As starting points for our design, we selected the oligonucleotide/oligosaccharide-binding fold domain of *Sulfolobus solfataricus* single-stranded DNA-binding protein (SSB OB)[17–19] and part of the C-terminal region of the open reading frame 1 protein (ORF1p C-1/3) from mouse[20,21]. An individual RBD normally binds target RNAs with micromolar affinity[22]. Notably, both the SSB OB and ORF1p C-1/3 domain have nanomolar affinity for single-stranded RNAs[19,21]. The SSB OB domain binds as a monomer to the phosphate backbone of single-stranded RNA through a positively charged groove with a footprint of five bases (Supplementary Fig. 1a)[18,19]. The structure of the human protein has been solved[23] (Supplementary Fig. 1b). Mouse ORF1p C-1/3 has high sequence identity to human ORF1p (Supplementary Fig. 1c). The predicted structure of the mouse protein has a deeper and wider positively charged cleft for binding of the phosphate backbone of RNA than does the human ORF1p (Supplementary Fig. 1d). We designed the HARD protein by linking the mouse ORF1p C-1/3 domain to the SSB OB domain using three repeats of Gly-Gly-Gly-Gly-Ser-Ala as a linker (Fig. 1a).

To examine the RNA-binding activity of HARD protein, we expressed and purified the recombinant HARD protein fused with EGFP from *E. coli* (Fig. 1b). As a control, we expressed and purified EGFP in the same manner. The HARD protein adopts a stable monomeric structure as shown by gel filtration analysis (Supplementary Fig. 1e). We evaluated binding of the HARD protein to various nucleic acids using isothermal titration calorimetry (ITC). The HARD protein bound to 8nt single-stranded RNA, double-stranded RNA, single-stranded DNA, and double-stranded DNA with around 1 μM affinities (Fig. 1c and Supplementary Fig. 1f–h).

To enable use of the HARD protein to isolate RNAs, we conjugated the recombinant EGFP-HARD to N-hydroxysuccinimide-activated agarose beads. EGFP-conjugated beads were also prepared. Beads were incubated with the total RNA purified from HEK293 cells and bead-bound RNA was quantified after stringent washing. Quantitative RT-PCR (qRT-PCR) analysis of RNAs bound to the beads showed that ribosomal RNAs (rRNAs), polyA mRNAs, non-polyA mRNAs, long non-coding RNAs (lncRNAs), and small nuclear RNAs (snRNAs) were all efficiently captured by HARD beads, but negligible amounts were captured by EGFP beads (Fig. 1d, Supplementary Data 1). Different from the results of the isothermal titration calorimetry assay, the HARD beads captured heat-denatured single-stranded genomic DNA but not double-stranded genomic DNA (Supplementary Fig. 1i), suggesting that the HARD protein does not bind long double-stranded DNA.

To further characterize the RNA-binding preference of the HARD protein, we performed RNA-seq after removal of ribosomal RNAs to compare the distributions of RNAs captured by the HARD beads and input RNAs (Supplementary Data 2). Pearson correlation analysis of normalized RNA levels showed that the HARD-AP samples were highly correlated with each other and with the input samples (Pearson correlation $r = 0.97 \sim 0.98$) (Fig. 1e). Differential analysis showed that only few RNAs were significantly enriched in the input samples compared to the HARD-AP samples (Fig. 1f). In addition, the HARD-AP samples and input samples had highly similar distributions in terms of RNA abundance, RNA biotypes, RNA genome localization, and GC content (Fig. 1g-h and Supplementary Fig. 1j-l). For gene loci examined, normalized RNA-seq signals in the HARD-AP and input samples were highly similar (Fig. 1i). Taken together, these experiments demonstrated that the HARD-AP efficiently captures RNAs in a sequence and length-independent manner.

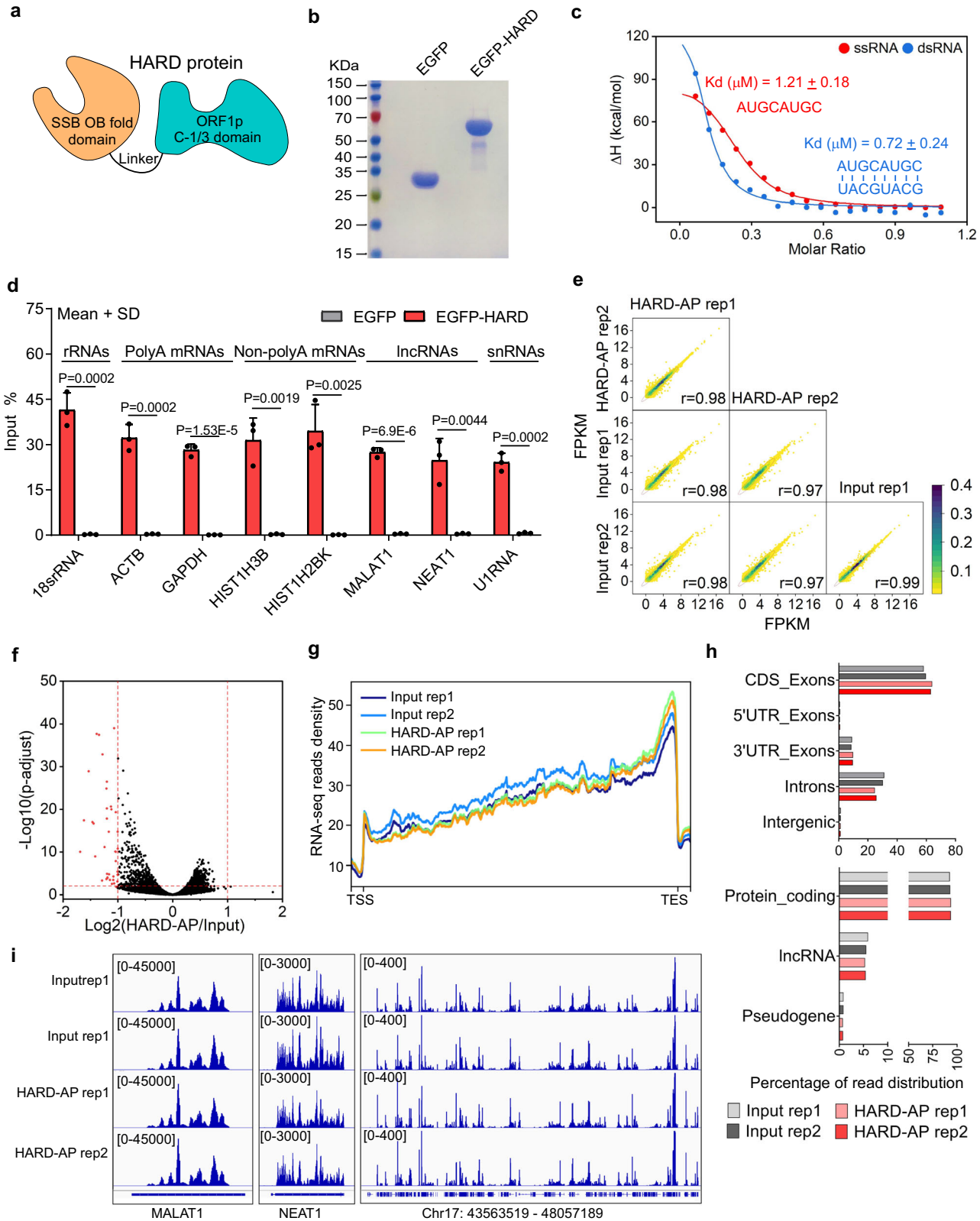
### Capture of RBPs from HEK293 cells using HARD-AP

UV light is widely used to covalently crosslink RNA to protein; this method does not crosslink protein to protein or DNA to protein[7,24]. However, UV-induced RNA-protein cross-linking has low efficiency and is hard to penetrate the tissue samples[12–15]. Thus, to retrieve RBPs and tightly associated RNA regulatory complexes with increased efficiency, we combined the UV cross-linking and high salt wash conditions, where the UV cross-linking helps preserve transient and weak RNA-protein interactions. The cellular ionic strength is comparable to 110–130 mM KCl[25,26]. We used the high salt wash buffer containing 500 mM NaCl, more than three times of the cellular ionic strength, to remove the non-specific contaminants in the HARD-AP, which is stringent enough to isolate specific RNA-protein complexes[27–30]. To be noted, HARD-AP will recover both the RBPs that directly bind to RNAs and the proteins that tightly associate with these direct RBPs, all of which are referred to as HARD RBPs in this study. It is well known that RNA-binding proteins cooperate with other proteins or form specific complexes (e.g. spliceosome, RNA exosome, etc.) to regulate the fate of RNAs in many cases[1,2]. It is therefore important to recover these tightly associated indirect RBPs in order to understand their regulatory mechanisms.

To capture RBPs from HEK293 cells, we first treated live cells with 254 nm UV light at a dose of 400 mJ/cm$^2$, and then lysed the cells and isolated RNA-protein complexes using HARD-AP (Fig. 2a). To prevent contamination from DNA-binding proteins, lysates were treated with DNase prior to capture of RNAs using the HARD beads. Silver staining of the eluted samples showed that the HARD beads isolated a substantial amount of protein from UV-treated and non-UV-treated samples but that EGFP beads did not bind protein (Fig. 2b). The similar silver staining patterns of UV-treated and non-UV-treated samples indicate the limited contribution of UV treatment, in line with the low efficiency of UV-induced RNA-protein cross-linking. Notably, little protein was detected when the samples were treated with RNase A prior to capture on the HARD beads, suggesting that the proteins captured by the HARD beads bind to RNAs directly or indirectly.

A western blotting analysis was performed to examine the levels of well-known RBPs and non-RBPs in different elutes. Known RBPs PTBP1, PSPC1, and NONO and non-canonical RBP GAPDH were detected in the HARD-AP elute but not the RNase A-treated sample (Fig. 2c). DNA-binding proteins Histone H3 and DNMT1 were not detected in the HARD-AP eluate (Fig. 2c). In the absence of UV cross-linking, the non-canonical RBP GAPDH was removed from the HARD beads, while the bona fide RBPs (PTBP1, NONO) were retained with significantly lower signals (Fig. 2c), suggesting that the wash condition was stringent to remove the non-specific contaminants.

To characterize proteins in the elutes, we cleaved the isolated proteins into peptides with trypsin and analyzed them by liquid




chromatography-tandem mass spectrometry (LC-MS/MS). We prepared three biologically independent replicates HARD-AP samples (HARD), EGFP-AP control samples (EGFP), RNase A-treated samples incubated with HARD beads (RNase), and non-UV crosslinking incubated with HARD beads (Non-UV). The data from replicates were pooled, and a search against the UniProt database using Proteome Discoverer was employed to identify peptides. From these samples, 25,208 unique peptides and 4,754 quantifiable proteins were identified (Supplementary Fig. 2a and Supplementary Data 3). The distribution of peptide size, charge, peptide number per protein, coverage, and protein mass met the requirement of quality control (Supplementary Fig. 2b–e and Supplementary Data 3). The data from replicates were highly correlated (Pearson correlation $r = 0.99 \sim 1$ for HARD, 0.99 ~ 1 for Non-UV, 0.75 ~ 0.89 for EGFP, 0.96 ~ 0.97 for RNase) (Supplementary

**Fig. 1 | The RNA-binding activity of HARD protein. a** Schematic diagram of HARD protein. HARD protein is composed of an SSB OB fold domain, a LINE-1 ORF1p C-1/3 domain and a flexible linker. **b** The Coomassie blue stained SDS-PAGE gel showing the purified recombinant EGFP-HARD and EGFP. This experiment was repeated once with similar results. **c** Isothermal titration calorimetry assay showing the binding of EGFP-HARD to single-stranded RNA (ssRNA) and double-stranded RNA (dsRNA). Shown are normalized data with the best fits (solid lines). Raw data are shown in Supplementary Fig. 1f. **d** Percentage of input RNA species isolated using HARD beads and EGFP beads as determined using qRT-PCR. Data are means ± SD; three independent biological samples were used for the analysis ($n = 3$); significance was determined using the two-tailed Student's t-test. **e** Scatter plot showing the correlation of normalized RNA-seq signals (FPKM) of HARD beads-enriched RNAs (HARD-AP) and input RNAs ($n = 60,623$). Two independent biological samples are labeled as rep1 and rep2. The colors scale indicates dot density. The Pearson correlation coefficients are given. **f** Volcano plot showing distributions of RNAs differentially detected in HARD-AP and Input samples ($n = 60,623$). Significantly differentially detected RNAs are labeled with red dots. The significance (p) was determined using the two-tailed Student's t-test and further adjusted using the Benjamini-Hochberg correction for multiple testing (p-adjust). **g** Metagene representation of RNA-seq signals of HARD-AP and Input samples in bodies of genes ($n = 60,623$). **h** Percentage of indicated RNA species and RNA-seq read distributions along genes in HARD-AP and input samples. **i** Genome browser tracks showing normalized RNA-seq signals along indicated genomic loci in HARD-AP and input samples. Source data for (b, e-f) are provided as a Source Data file.

Fig. 2f), and the HARD-AP and Non-UV replicates had small relative standard deviation (Median = 0.11) (Supplementary Fig. 2g), demonstrating that the HARD-AP protocol is reproducible.

To call positive targets isolated by HARD-AP, we referred to the methods (RICK, CARIC, Interactome capture[7–9]) that also utilize the affinity purification procedure and set the filtering criteria as follows: to be identified in at least two out of three HARD replicates; at least two unique peptides were identified; at least three-fold higher signal in the HARD-AP samples than in the EGFP replicates, with a p-adjust using the Benjamini-Hochberg correction for multiple testing of <0.01. By these criteria, the RBPome of HEK293 cells consists of 2202 proteins (Fig. 2d and Supplementary Data 3). When we replaced the EGFP samples with the RNase-treated samples to filter RBPs using the same criteria as above, we obtained 1575 RBPs, 1481 out of which overlaps the RBPome using the EGFP samples as the control (Fig. 2e), further suggesting that HARD beads capture proteins through their interactions with RNAs directly or indirectly. In addition, we successfully isolated 2426 RBPs from non-UV-treated HEK293 cells (Fig. 2d), which were filtered through the same criteria as the UV-crosslinked samples. Notably, 1719 RBPs were shared between the crosslinked and non-crosslinked samples (Fig. 2e), accounting for ~80% of RBPs derived from the UV-crosslinked condition. The data from two conditions showed a high Pearson correlation (Pearson correlation $r = 0.91$-$0.93$) (Supplementary Fig. 2f). We further compared the signal intensities of the shared RPBs between two conditions and found that 1233 (~72%) proteins did not exhibit significant difference between two conditions, and 271 (~22%) proteins showed significantly higher signals in crosslinked samples than in non-crosslinked samples. These data suggest the limited contribution of UV treatment and also suggest that HARD-AP can robustly isolate RBPs independent of UV-crosslinking. Yet, it also needs to be noticed that the inability of the HARD protein to tolerate denaturing conditions might lead to the presence of certain non-specific binding proteins that can endure the high salt washing conditions.

## Characterization of RBPomes in HEK293 cells captured by HARD-AP

Different methods have been used to characterize the RBPome of HEK293 cells. These methods include polyA-based capture methods (PAR-CLIP, pCLIP, comparative RIC, pCLAP, CAPRI), organic phase-separation-based methods (XRNAX, OOPS)[7,10,11,31–33]. Combined, these methods have identified 2,719 HEK293 proteins as RBPs (Fig. 2f, Supplementary Data 4). Based on 23 reported studies, there are 4,506 previously identified human RBPs from 8 cell lines (Supplementary Data 4). These 4,506 RBPs were all isolated through denatured procedures and thus all of them bind RNA directly. 1,685 out of 2,202 (~77%) RBPs identified by HARD-AP overlap with these 4,506 RBPs (Fig. 2f), suggesting that at least 77% of HARD-AP RBPs in HEK293 cells directly bind RNAs. Importantly, 517 known RBPs were retrieved by HARD-AP but not by any other RBPome analysis method from HEK293 cells (Fig. 2f).

As expected, there was over-representation of RNA-related gene ontology (GO) terms, including RNA binding, nucleotide binding, nucleoside phosphate binding, RNA metabolism, RNA processing, nucleobase-containing compound metabolism, and gene expression, in the HARD-AP RBPome (Fig. 2g).

Next, we examined performance of HARD-AP on retrieving the RNA-processing complexes Spliceosome, Integrator and RNA exosome[34–36]. For the highly abundant Spliceosome complex, HARD-AP captured a fraction of its subunits similar to other methods (Fig. 2h). The Integrator complex is Pol II-associated RNA-processing complex with relative low abundance in cells, 10 subunits of which directly bind RNAs[34]. HARD-AP captured 8 out of 14 subunits of the Integrator complex, whereas other methods retrieved none, one, or two (Fig. 2h, Supplementary Fig. 2h). It was known that all 11 subunits of the RNA exosome directly bind RNA[35]. Notably, 9 subunits of the RNA exosome were isolated by the HARD-AP, whereas most of other method captured no more than four (Fig. 2h, Supplementary Fig. 2i). Furthermore, proteins of the Mediator complex were exclusively identified by the HARD-AP (19 of 32 subunits), and the 26S proteasome complex were retrieved by HARD-AP and pCLAP (Fig. 2h). The Mediator complex associates with non-coding RNAs during the chromatin looping process to enhance transcription and interacts with newly transcribed RNAs during RNA polymerase II pausing to form dynamic transcriptional condensates[37–39]. The 26S proteasome complex reportedly act as endoribonucleases to degrade cellular RNAs[40,41]. All above suggested that HARD-AP could efficiently recover highly abundant or relative lowly abundant RNA regulatory complexes.

The HARD-AP-derived RBPs have relatively more acidic isoelectric point and higher hydrophobicities proteins than RBPs identified by other methods (Fig. 2i). We projected the HARD-AP-derived RBPs onto the human subcellular proteome database[42] and found that proteins that localize to the cytoplasm, nucleus, mitochondria, nucleoli, vesicles, and plasma membrane were identified (Supplementary Fig. 2j).

## Validation of HARD RBPs using the protein microarray

The HuProt Human proteome microarray contains over 21,000 GST-purified unique recombinant human proteins in yeast, including >81% of the canonically expressed proteins as defined by the Human Protein Atlas[43,44]. The proteins in the HuProt protein microarray are folded in their native conformation. The protein microarray has been successfully used for interrogating the direct RNA-protein interactions[45–48], such as TINCR and STAU1, SNORD50A/B and K-Ras, Bvht and CNBP. Thus, the protein microarray can be an alternative tool to independently validate the RNA-binding activities of proteins.

As shown in Fig. 3a, we generated a pool of Cy5-labeled RNAs by mixing the fragmented total RNAs and in vitro transcribed RNAs. To amplify the non-ribosomal RNAs (non-rRNAs), we generated the cDNA by reverse transcription using the total RNAs of HEK293 cells from which the ribosomal RNAs were first removed as the template; we then amplified the double-stranded templates of the T7 in vitro transcription by PCR using random primers; the RNAs were finally produced by the T7 in vitro transcription (Fig. 3a and Supplementary Fig. 3a, b). We randomly labeled the Cy5 dye to the fragmented total RNAs of HEK293 cells and the in vitro transcribed RNAs above (Supplementary

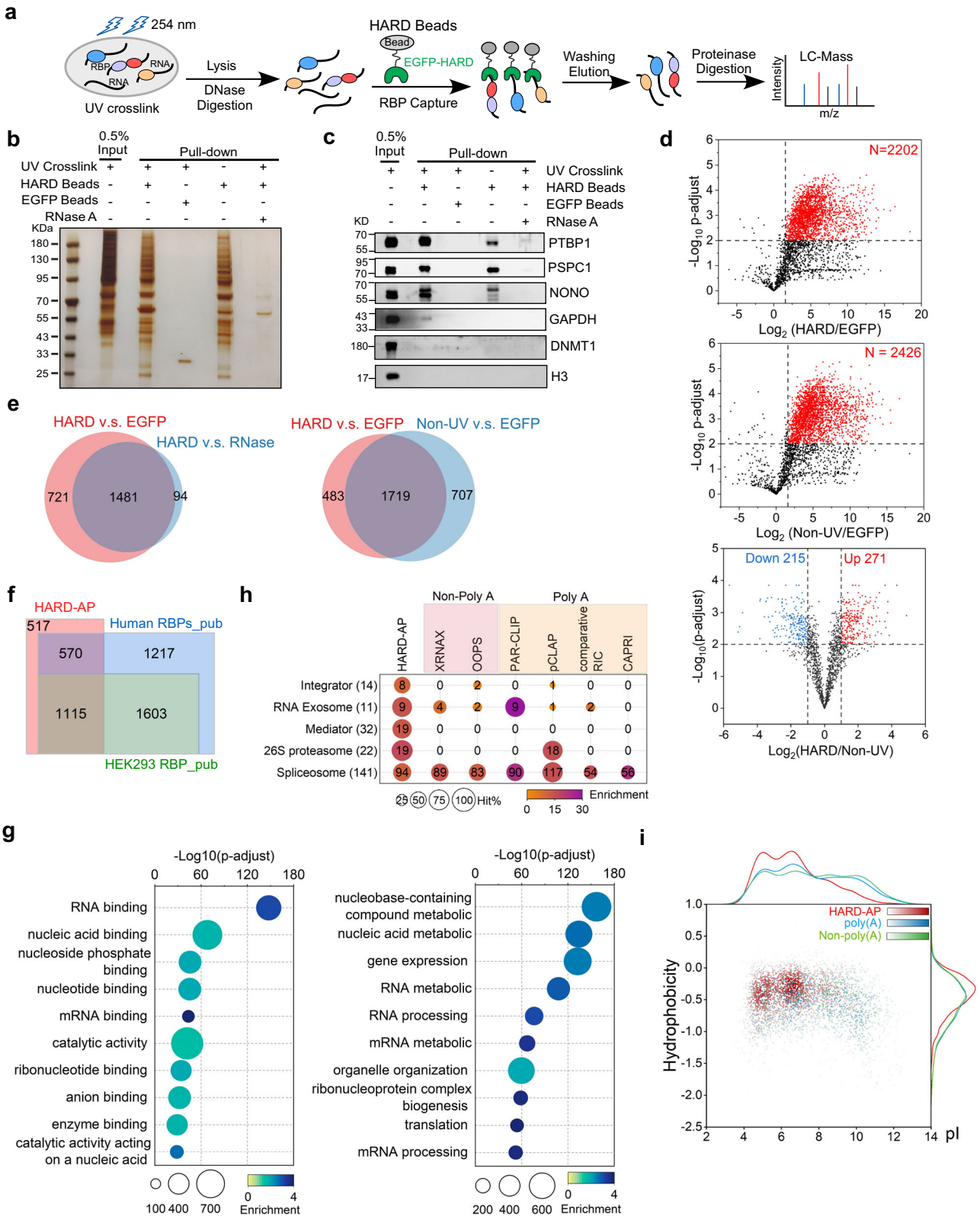



Fig. 3c, d). The calculated Cy5 labeling efficiency was 1 dye per 110nt for the fragmented total RNAs and 1 dye per 170nt for the in vitro transcribed RNAs. The pooled RNAs covered different types of RNAs with a high complexity, particularly non-rRNAs.

To investigate the RNA-protein interactions, we incubated the pooled Cy5-labeled RNAs with two independent protein microarrays and quantified the fluorescence signals on the microarrays after several washing steps. We used the fold change (FC) and signal-noise-ratio (SNR) to evaluate the RNA-binding activities of the proteins. The FC is defined as: Foreground Signal (F635) / Local Background Signal (B635); SNR is defined as: (F635 -B635)/ B635SD, where B635SD is the standard deviation of B635. The FC distributions of two protein microarray replicates had small relative standard deviation (Median = 0.10) (Supplementary Fig. 3e) and were highly correlated (Pearson

**Fig. 2 | Capture of RBPs from HEK293 cells using HARD-AP. a** Schematic of the HARD-AP procedure and LC-MS/MS analysis. **b** The silver stained SDS-PAGE gel showing protein precipitated from HEK293 cell lysates under indicated conditions. This experiment was repeated once with similar results. **c** Western blot analysis for indicated proteins in precipitates obtained under indicated conditions. This experiment was repeated once with similar results. **d** Volcano plot of distributions of proteins captured by HARD beads compared to EGFP beads in UV-crosslinked samples (top, $n = 3,457$), proteins captured by HARD beads in samples without UV treatment compared to EGFP beads (middle, $n = 3594$), and proteins captured by HARD beads in UV-crosslinked samples compared to non-UV samples (bottom, $n = 1719$). The fold changes were calculated from means of the ion intensities of three independent biological samples. The significance (p) was determined using the two-tailed Student's t-test and further adjusted using the Benjamini-Hochberg correction for multiple testing (p-adjust). **e** Left: Venn diagram comparing HARD-AP-derived RBPs using the EGFP-AP control sample (EGFP) and sample treated with RNase (RNase) as the negative control respectively; Right: Venn diagram comparing HARD-AP-derived RBPs in the UV-crosslinked samples and non-UV samples. **f** Venn diagram comparing RBPs isolated by HARD-AP to published RBPs. **g** Top GO terms over-represent in RBPs of HEK293 cells identified using HARD-AP. The GO enrichment analysis used the two-sided Fisher's exact test with the p-value adjusted using the Bonferroni correction for multiple testing. **h** Matrix bubble plot showing the comparison of Integrator, RNA Exosome, Mediator, 26S proteasome, and Spliceosome complexes captured by indicated methods from HEK293 cells. Hit% is the percentage of subunits of each complex captured by indicated method. The number of subunits captured by indicated methods is labeled on each bubble. The color scale indicates the enrichment. **i** Scatter plot of distributions of hydrophobicities vs. isoelectric points (pIs) of RBPs identified by HARD-AP (red), polyA-based methods (blue), and non-polyA-based methods (green) in HEK293 cells. Color scales indicate densities. Density plots outside axes illustrate distributions. Source data for (b-c, d-h) are provided as a Source Data file.

correlation $r = 0.90$) (Fig. 3b), demonstrating that the protein microarray assay is reproducible. On the protein microarray, the buffer, BSA, GST, IgA and IgG at different concentrations were used as negative controls, and the Alexa 647 labeled IgG was used as the positive control. As shown in Fig. 3c, d, the FC and SNR distributions of HARD RBPs ($n = 1447$) and HARD-AP specific RBPs ($n = 643$) in the HEK293 cells were almost all (~94%) significantly higher than the negative controls and were also very similar to that of the GO RBPs ($n = 1365$). GO analysis of RBPs with the FC > 1 showed the significant enrichment for RNA-related terms such as RNA binding, nucleotide binding, nucleobase-containing compound metabolism, RNA metabolism, etc. (Fig. 3e), providing a good validation of the protein microarray for the detection of protein-RNA interactions.

In addition, the well-known RBPs PTBP1, HNRNPK, RBM39, RBM19 and GAPDH showed highly specific Cy5 signals, while DNA-binding proteins histone H1, H3 and DNMT1 were not detected on the protein microarray (Fig. 3f). These data suggested the reliability of the protein microarray assay in measuring RNA-protein interactions. Importantly, a large fraction of the subunits of the RNA exosome (7 out of 8 subunits available on the array), integrator (6 out of 8 subunits available on the array), mediator (21 out of 26 subunits available on the array) and 26S proteasome (16 out of 18 subunits available on the array) showed direct RNA-binding activities with the signal significantly higher than the negative controls (FC > 1) (Supplementary Data 5). The representative subunits of these complexes on the protein microarray were shown in Fig. 3f, g. Furthermore, we compared the RNA-binding activities of the HARD RBPs and the published RBPs identified by the polyA-based methods and non-polyA-based methods on the protein microarray assays, and found similar FC and SNR distributions among them (Supplementary Fig. 3f). Thus, we used the high-throughput protein microarray assay as an alternative tool to systematically validate the direct RNA-binding activities of proteins captured by HARD-AP.

## Identification of RBPs in mice using HARD-AP

Due to technical limitations, previous studies have mainly focused on identification of RBPs in cultured cell lines (Supplementary Data 4). To obtain a comprehensive picture of the RBPome under physiological conditions, we applied the HARD-AP to characterize RBPs in organs of adult mice (brain, heart, lung, liver, and kidney) and mouse embryonic stem cells (mESCs) within 30 passagens. The mESCs were treated directly with a dose of 400 mJ/cm2 UV at 254 nm. To overcome the poor penetration of UV light through tissue samples and preserve transient and/or week RNA-protein interactions, we optimized the UV crosslinking conditions according to the previous study[49]. The freshly isolated organs were first frozen and ground into powder using liquid nitrogen, which was then cross-linked with a dose of 500 mJ/cm2 UV at 254 nm. We thus followed the same HARD-AP protocol as for HEK293 cells. As expected, the HARD-AP beads isolated substantial amounts of protein, whereas very little protein was captured by EGFP beads (Supplementary Fig. 4a).

LC-MS/MS analyses led to identification of 51,498 unique peptides and 7,618 proteins in the HARD-AP samples (Supplementary Fig. 4b). The distributions of peptide size, charge, peptide numbers per protein, coverage, and protein mass satisfied the quality control standard (Supplementary Fig. 4c–f and Supplementary Data 5). All HARD biological replicates were highly correlated (Pearson correlation $r = 0.99 \sim 1$) with small relatively standard deviations (Median = 0.09 ~ 0.20) (Supplementary Fig. 4g, h). Manifold approximation and projection (UMAP) analysis showed that the proteins identified using HARD and EGFP beads distributed into two different clusters, and there were sub-clusters within the HARD proteins, suggesting organ-specific RBPomes (Supplementary Fig. 4i). We obtained high-confidence RBPomes of 2891 proteins for mESCs, 3888 proteins for brain, 3235 proteins for heart, 3600 proteins for lung, 3246 proteins for liver, and 3575 proteins for kidney (Fig. 4a, Supplementary Data 6). Comparing to the 2202 RBPs identified in HEK293 cells, many more proteins appear to interact with RNAs under physiological conditions than in cultured cell lines with high passage numbers.

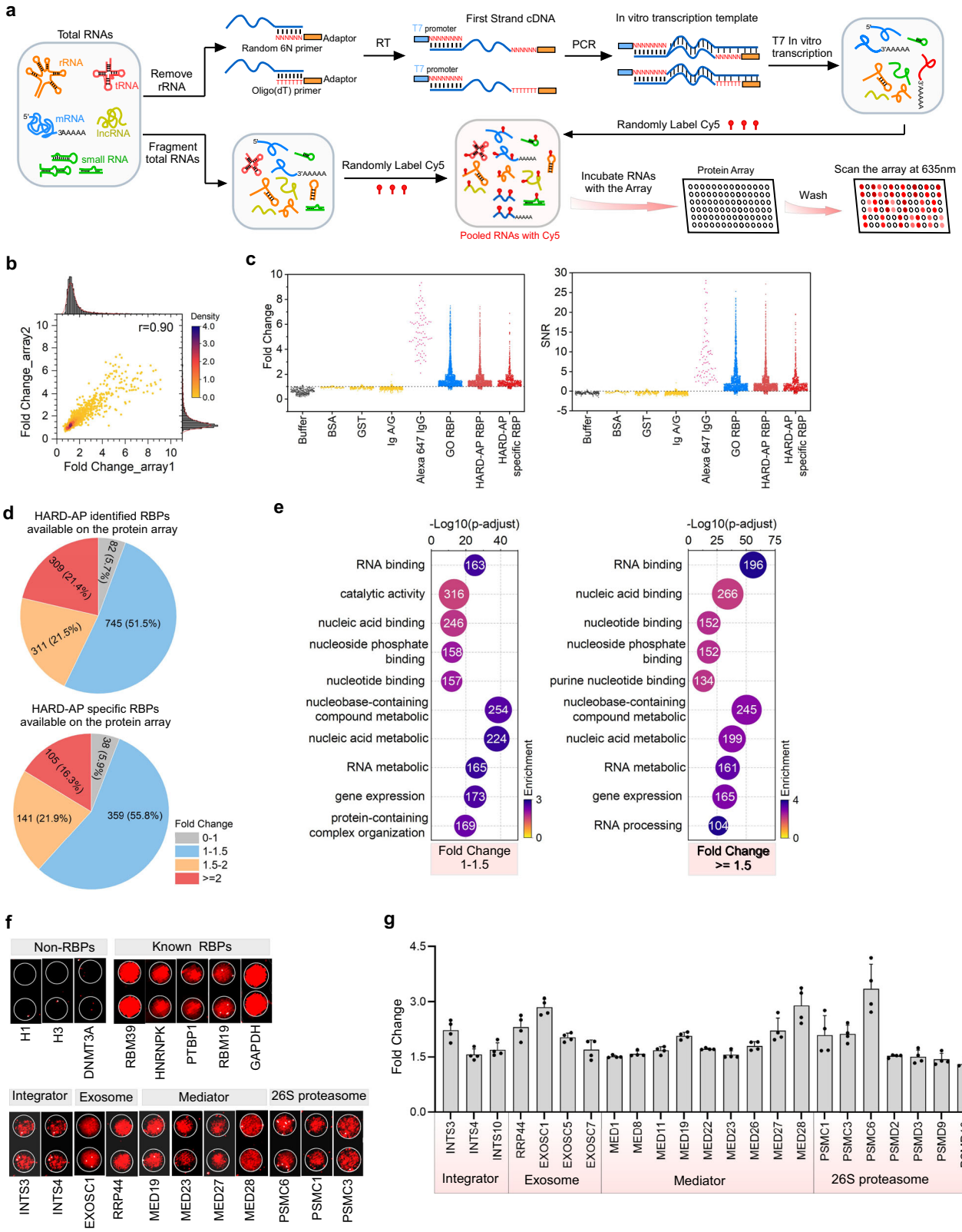
In addition, we also isolated RBPs from the non-crosslinked brain and kidney using HARD beads. We then performed the mass spectra database search together with the previous HARD beads-derived proteins from UV-crosslinked samples (HARD) and EGFP beads-derived proteins (EGFP). Following the same filtering criteria as the UV-treated sample, we identified 3045 and 3321 RPBs in the UV-treated and non-UV-treated brain samples, respectively; we also isolated 2785 and 3156 RPBs in the UV-treated and non-UV-treated kidney samples, respectively (Supplementary Fig. S5a, Supplementary Data 6). Among them, ~90% (2693 in brain, 2443 in kidney) RBPs in brain or kidney samples were shared between the crosslinked and non-crosslinked samples, ~70% of which did not show significant difference between two conditions (Supplementary Fig. S5b, c). The performance of HARD-AP on the non-crosslinked tissue samples is consistent with that on the non-crosslinked cell samples (Fig. 2d, e). These data further demonstrate the capacity of HARD-AP independent of crosslinking treatment and the limited contribution of UV treatment on tissue samples.

## Comparison of the RBPomes across mouse organs and cells

For HARD-AP-derived mouse RBPomes, 873 proteins were recovered in all samples test; 1405 proteins were present in all organs; 1700 ~ 1950 proteins were identified in the mESC and one organ (Fig. 4b, c), suggesting tissue-specific distributions of RBPomes. We obtained the mouse RBPome of 6746 RBPs, containing 4282 new RBPs, through combining HARD RBPs, GO annotated RBPs, and reported RBPs.

The HARD-AP method efficiently captured proteins from RNA-processing complexes Spliceosome (48–68 of 136 depending on organ), Integrator (5–11 of 14 depending on organ), RNA exosome

https://doi.org/10.1038/s41467-024-52765-w




(4–10 of 11 depending on organ), Mediator (6–15 of 32 depending on organ), and 26S proteasome (20 of 22); most of these proteins were not detected by other methods (Supplementary Fig. 6a). These data reflect the consistency and sensitivity of the HARD-AP method in recovering RNA-interacting complexes even in tissue samples.

GO terms analyses of the HARD-AP-derived mouse RBPomes had over-representation of RNA-related processes including RNA binding, nucleotide binding, ribonucleotide binding, nucleoside phosphate binding (Fig. 4d). In addition, the enriched GO terms include identical protein binding, enzyme binding, hydrolase activity, transferase activity, oxidoreductase activity, et al. (Fig. 4d). The identical protein binding term suggests that many of the identified RBPs form

**Fig. 3 | Validation of HARD RBPs using the protein microarray. a** Schematic of generating a pool of Cy5-labeled RNAs and hybridizing this pool with the human protein microarray. **b** Scatter plot showing the correlation of the fold change (Cy5 foreground signals over local background) of the HARD-AP identified RBPs ($n = 1447$) in HEK293 cells. The fold change was calculated from the average of four independent protein spots on the two independent protein arrays (the same for **c**). The Pearson correlation coefficient is given. **c** Distributions of the fold change and signal-noise-ratio (SNR) of indicated sets of proteins on the protein microarray (GO RBP $n = 1365$; HARD-AP RBP $n = 1447$; HARD-AP specific RBP $n = 643$). The buffer ($n = 320$), BSA ($n = 80$), GST ($n = 320$), Ig A/G ($n = 720$) at different concentrations were used as negative controls, and the Alexa 647 labeled IgG ($n = 80$) was used as the positive control. **d** Pie chart showing the distribution of fold change (Cy5 signals over local background) of the HARD-AP identified RBPs ($n = 1447$) and HARD-AP specific RBPs (n = 643) in HEK293 cells available on the protein microarray. **e** Top GO terms (molecular function and biological process) over-represent in HARD-AP RBPs with FC = 1-1.5 ($n = 745$) or FC ≥ 1.5 ($n = 620$) on two independent protein microarrays. The GO enrichment analysis used the two-sided Fisher's exact test with the *p*-value adjusted using the Bonferroni correction for multiple testing. The number of proteins is labeled on each bubble. The color scale indicates the enrichment. **f** Images of Cy5-RNA incubation signal of selected proteins on the protein microarray. **g** Fold change of subunits of the Integrator, Exosome, Mediator, and 26S proteasome complex. Subunits with fold change over 1.5 were selected for plotting. Data are from four protein spots on two independent protein arrays and shown as the Mean ± SD. Source data for (**b**–**d**, **g**) are provided as a Source Data file.

homodimers; dimerization is known to facilitate specific RNA recognition and improve RNA-binding affinity[50,51]. For example, the dimerization of Nova1 creates two recognition sites for RNA binding and enhances affinity for RNA[50]. Recently, numerous metabolic enzymes were reported to have RNA-binding activity in living cells[6]. These RNA-protein interactions function in feedback loops important for regulation of gene expression and/or in the control of enzymatic functions. Notably, the HARD-AP-derived RBPome each includes 1,410–1,988 proteins with enzyme binding or enzymic activity (hydrolases, transferases, kinases and oxidoreductase), most not previously reported to have RNA binding activity (Fig. 4d). The new RBPs recovered by HARD-AP in mouse are strongly enriched for the GO terms of nucleotide/ribonucleotide binding (764 proteins), various enzymatic activities (1568 proteins), and enzyme binding (735 proteins). These data suggest that many metabolic enzymes moonlight as RBPs under physiological conditions. Based on GO cellular component enrichments, the HARD-AP-derived RBPs are localized to the nucleus, cytosol, organelle lumen, and nucleoplasm (Supplementary Fig. 6b). Compared to non-RBPs in mouse, the HARD-AP-derived RBPs have relatively more acidic isoelectric points and lower hydrophobicities (Supplementary Fig. 6c).

Next, we mapped the orthologs of all known human RBPs and mouse RBPs identified using HARD-AP[52]. Notably, 3,048 HARD-AP-derived mouse RBPs are orthologous to known human RBPs and 3,334 human RBPs are also orthologous to the mouse HARD-AP RBPs (Fig. 4e). Furthermore, we analyzed the RNA binding activities of the human orthologs of the mouse HARD-AP RBPs on the protein microarray and found that their FC and SNR distributions were similar to that of GO RBPs (Supplementary Fig. 3g), providing a good validation of the RNA-binding activities of the mouse HARD-AP RBPs. All these orthogonal analyses above suggest that there are indeed a large number of proteins with RNA-binding activity in both mouse and human.

## Organ-enriched RBPs

To understand the organ distributions of RBPs, we performed the hierarchical clustering analysis of the HARD-AP-derived RBPs using their normalized ion intensities of mass spectrometry data, revealing organ-enriched clusters of RBPs abundance (Fig. 4f). We defined the enrichment as: standardized LC-MS/MS ion intensity ≥1. The organ-enriched clusters of RBPs showed significant tissue specificities (Supplementary Fig. 6d) and were also significantly enriched in GO terms associated with organ-specific physiological functions (Fig. 4g). For mESC, the enriched cluster is enriched in RNA-related functions (Fig. 4g). The top enriched terms for brain are vesicle-mediated transport, modulation of chemical synaptic transmission, regulation of trans-synaptic signaling, nervous system development, all related to brain physiological functions (Fig. 4g). For the heart, the top enriched terms were aerobic respiration, oxidative phosphorylation, and electron transport chain. For lung, enriched GO terms are cytoskeleton organization and actin filament-based process. For liver, the top enriched terms are oxidoreductase activity and various terms related to metabolic processes (oxoacid, carboxylic acid, amino acid and lipid). For kidney, the top enriched GO terms are oxidoreductase, vesicle-mediated transport and various terms related to metabolic processes (oxoacid, carboxylic acid, amino). There results indicate that the organ-enriched RBPs are tightly linked to organ-specific physiological functions. Since the protein levels of these proteins are high in the organs where they function, we examined the correlation between HARD-AP enrichment and their endogenous protein levels. As shown in Supplementary Fig. 6e, the enrichment of HARD-AP RBPs showed a week/trivial correlation with their endogenous protein levels in all organs tested (Pearson correlation $r = 0.34$ for brain, 0.33 for heart, −0.08 for lung, 0.27 for liver, −0.17 for kidney). After correcting the HARD-AP abundance with their endogenous protein levels, the hierarchical clustering analysis of the HARD-AP-derived RBPs also uncovered organ-enriched clusters as well (Supplementary Fig. 6f). Different from Fig. 4g, these clusters were strongly enriched for the GO terms of RNA binding, RNA metabolism, RNA processing, nucleobase-containing compound metabolism, nucleotide binding in brain, heart, lung and liver (Supplementary Fig. 6g). All data above further suggest the specificity of HARD-AP in capturing RBPs of tissues.

To confirm our results, we next used HARD-AP and western blot analyses to confirm that several identified proteins, which had not previously been reported to be associated with RNA, were indeed organ-dependent RBPs (Fig. 4h–j). Despite similar protein levels in brain and lung, both Bcr and Prkar1a showed brain-dependent enrichment in the HARD-AP sample. Similarly, Mylk3 was captured by the HARD beads from the heart lysate but not from the kidney lysate. Interestingly, in addition to the canonical Bcr with a molecular weight of 140–160 KD, we found a Bcr variant around 100 KD in all the organs examined (Fig. 4i). Following UV crosslinking, the levels of the canonical Bcr significantly decreased, and most of the Bcr shifted to ~100 KD in both brain and lung samples, which displayed distinct RNA-binding activities between the two types of samples (Fig. 4j). Furthermore, the human orthologs of Bcr, Prkar1a and Mylk3 also showed specific direct RNA-binding activities on the protein microarray (Fig. 4k).

All data above support the specificity of HARD-AP and the organ-dependent RNA-binding activities of these three proteins.

## Mapping of RNA-binding sites within RBPs using machine learning

Recently, machine learning, especially deep learning, has been shown to be able to accurately predict three-dimensional structures of proteins and interactions between biomolecules[53–56]. Thus, we systematically characterized RNA-biding sites (RBS) within RBPs in human and mouse using the protein structure-based deep-learning software GraphBind[56] to analyze globular domains and the protein sequence-based machine-learning software flDPnn[57] to analyze intrinsically disordered regions (IDRs) (Fig. 5a and Supplementary Data 7–8).

GraphBind extracts local patterns of structural and biophysicochemical features from protein structures to construct the embedded rules for RNA-binding residue prediction[56]. High-

https://doi.org/10.1038/s41467-024-52765-w

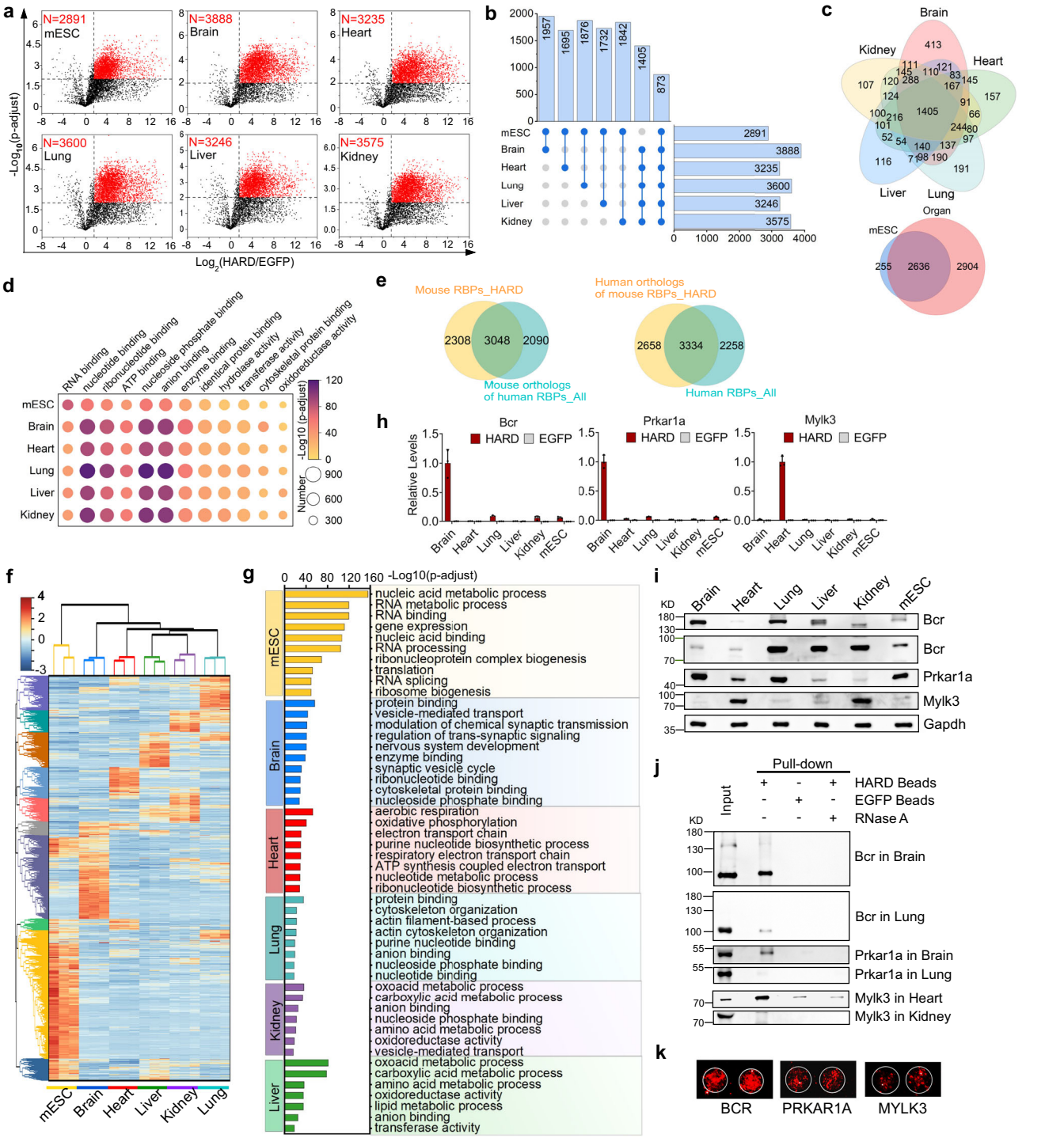
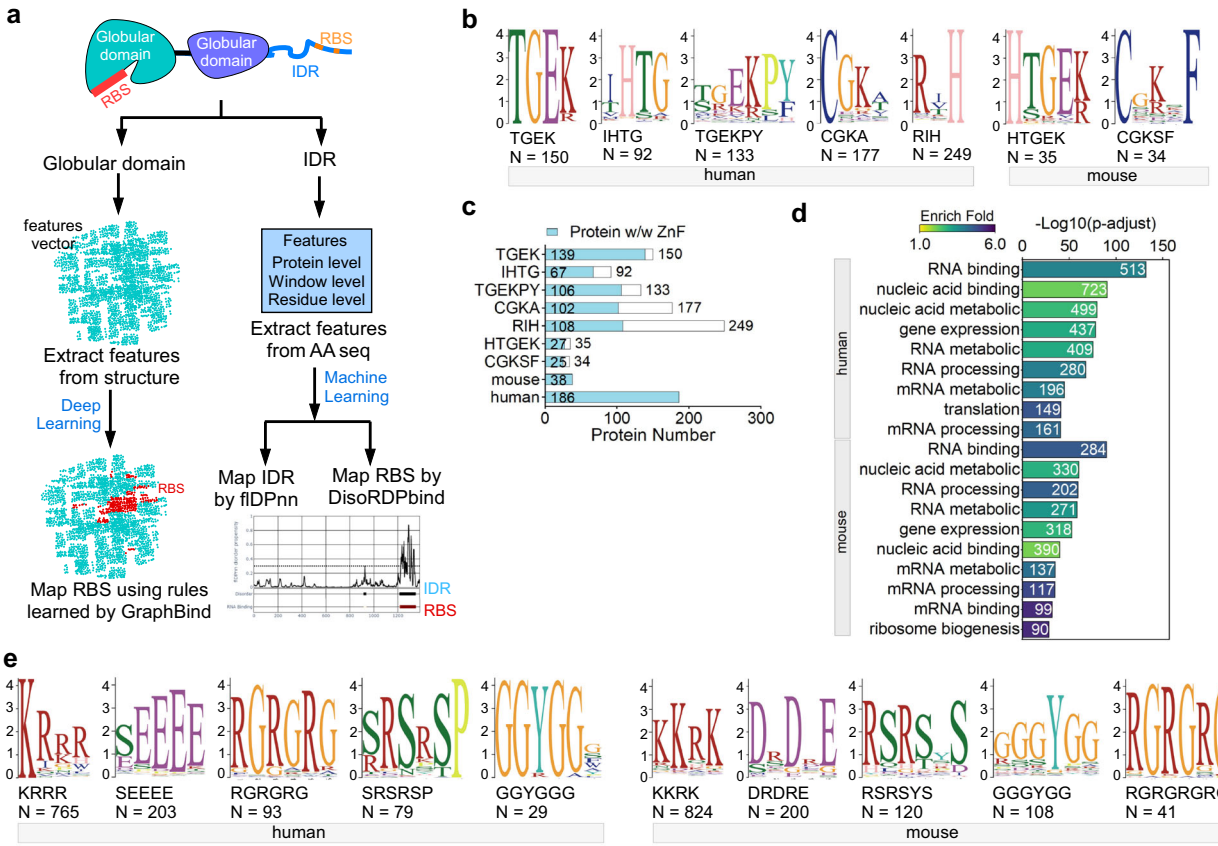


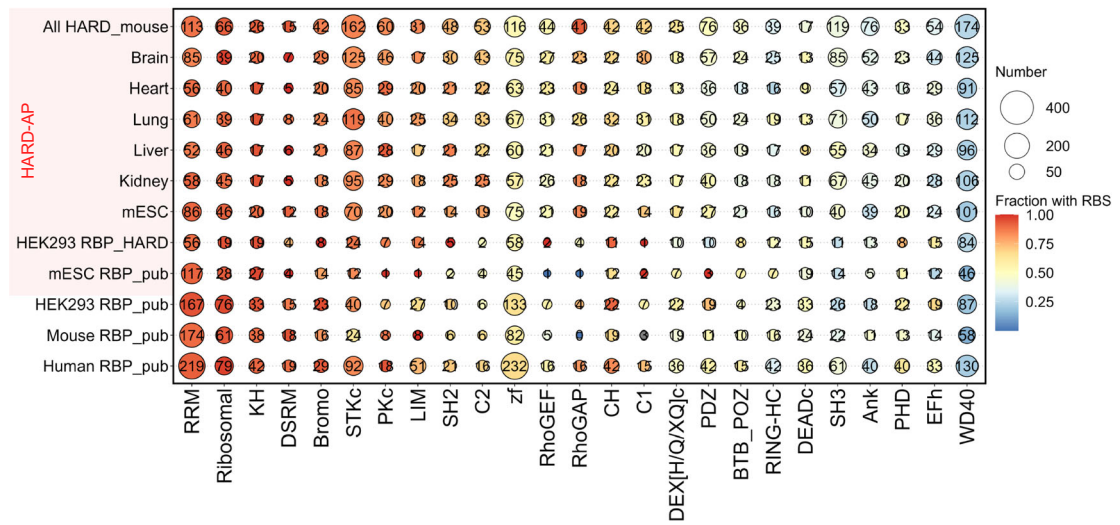

**Fig. 4 | The identification of RBPs in mouse cells and tissues using HARD-AP.** **a** Volcano plot showing distributions of proteins captured by HARD beads compared to EGFP beads in indicated samples. The fold changes were calculated from means of the ion intensities of three independent biological samples. The significance (p) was determined using the two-tailed Student's t-test and further adjusted using the Benjamini-Hochberg correction for multiple testing (p-adjust). **b** UpSet plot comparing RBPs identified in mESC and different mouse organs by HARD-AP. **c** Venn diagram comparing RBPs identified in different mouse samples by HARD-AP. **d** Matrix bubble plot showing enrichments of molecular function GO terms of in RBPs of indicated samples. The GO enrichment analysis used the two-sided Fisher's exact test with the p-value adjusted using the Bonferroni correction for multiple testing. **e** Venn diagram comparing the mouse RBPome identified using HARD-AP (mouse RBPs_HARD) and all human RBPome (human RBPs_All) to their indicated orthologs. **f** Heatmap of the hierarchical clustering analysis using normalized ion intensities from indicated samples. **g** Top GO terms over-represent in tissue- and cell-enriched RBPs identified by HARD-AP. The GO enrichment analysis used the two-sided Fisher's exact test with the *p*-value adjusted using the Bonferroni correction for multiple testing. **h** Relative levels of indicated proteins in different samples, which are calculated from the ion intensities of three independent biological samples. Data are means ± SD. HARD and EGFP represent proteins isolated by the HARD beads and EGFP beads respectively. **i** Western blot analysis showing the endogenous protein levels of Bcr, Prkar1a and Mylk3 in different mouse organs and mESC. This experiment was repeated once with similar results. **j** Western blot analysis for indicated proteins in indicated organ lysates after capture under indicated conditions. This experiment was repeated once with similar results. **k** Images of Cy5-RNA incubation signal of the human orthologs of Bcr, Prkar1a and Mylk3 on the protein microarray. Source data for (**a**–**c**) are provided as a Source Data file.

https://doi.org/10.1038/s41467-024-52765-w




f Numbers of HARD-AP-derived proteins harboring canonical RBDs

| Combined mouse | Brain | Heart | Kidney | Liver | Lung | mESC | HEK293 |
|---|---|---|---|---|---|---|---|
| 874 | 656 | 525 | 568 | 557 | 574 | 631 | 534 |




confidence protein structures of the globular domains of RBPs were obtained from the AlphaFold Protein Structure Database[58]. The RBS were enriched in positively charged residues such as arginine and lysine (25.1% for human RBPs, 24.7% for mouse RBPs) and hydroxylic residues such as serine (19.8 % for both human and mouse RBPs) (Supplementary Fig. 7a). Significant consensus patterns were not detected with the exception of the classical C2H2 zinc finger domain (zf-C2H2) of human and mouse RBPs (Fig. 5b, c and Supplementary Fig. 7b). This suggests that that RBS cannot be identified based solely on protein primary sequence.

Some IDRs have been reported to directly engage in interactions with RNA in human RBPs[4]. We, thus, applied machine-learning-based method flDPnn[57] to comprehensively examine distributions of IDRs in human and mouse RBPs (Supplementary Data 7-8). flDPnn has been

**Fig. 5 | Mapping of RNA-binding sites within RBPs. a** Schematic of strategies used to map RBSs using machine learning. **b** Sequence logos of enriched motifs of RNA-binding sites identified in globular domains of human and mouse RBPs. Number under each logo represents the number of RBPs harboring the motif. The Sequence logos were generated by MEME suite. **c** Bar plot showing the numbers of RBPs with indicated motifs divided into those with and without zinc finger (ZnF) domains. **d** Top ten GO terms in human and mouse RBPs containing RNA-binding IDRs. The GO enrichment analysis used the two-sided Fisher's exact test with the *p*-value adjusted using the Bonferroni correction for multiple testing. **e** Sequence logos of motifs enriched in IDRs of human and mouse RBPs. Number under each logo represents the number of RBPs harboring the motif. The Sequence logos were generated by MEME suite. **f** Numbers of RBPs identified by HARD-AP with canonical RBDs. **g** Matrix bubble plot showing enrichments of previously described domains (Conserved Domain Database) in the RBPs identified using HARD-AP as well as previously described RBPs. Color scales represent the fraction of domains harboring predicted RBS. Source data for (**f**, **g**) are provided as a Source Data file.

shown to accurately predict disorder and disorder functions including RNA binding[54,57]. Predicted RNA-binding IDRs were identified in 1,410 human RBPs and 1,252 mouse RBPs. Notably, for RBPs harboring RNA-binding IDRs, there is an over-representation of RNA-related GO terms including RNA/mRNA binding, nucleic acid binding, nucleic acid metabolism, gene expression, RNA metabolism, etc. (Fig. 5d). In addition, RNA-binding IDRs harbor a high proportion of small and polar and/or charged amino acids (27% for both human and mouse RBPs) (Supplementary Fig. 7a).

A sequence motif enrichment analysis of IDRs in identified RBPs yielded several significant consensus motifs: KRRR, KKRK, SEEEE, DRDRE, RG and SR repeats, and GGYGG (Fig. 5e and Supplementary Fig. 7b, c). The poly(K/R) motifs KRRR and KKRK are the most abundant of these motifs (765 in human and 824 in mouse). The poly(K/R) patch has been proposed to establish electrostatic interactions with the phosphate backbone of RNA in a manner analogous to the basic tails in DNA-binding proteins[7]. The RG and SR repeats were previously reported to bind RNAs and play roles in regulating transcription, pre-mRNA splicing, and mRNA translation[4,59,60].

Canonical RBPs are defined as proteins containing at least one of 799 known canonical RNA-binding domains (RBDs) (Supplementary Data 7-8). The HARD-AP identified 874 canonical RBPs in mouse, with from 441 to 656 in each tissue (Fig. 5f, Supplementary Data 7). To further understand the protein domains that recognize RNAs for the RBPome, we analyzed the distribution of domains within RBPs and then mapped the predicted RBS onto these domains. The HARD-AP-derived RBPomes strongly enrich previously described RNA interaction domains, a high proportion of which harbor predicted RBS, such as the RNA recognition motif (RRM) (85-100% with RBS), the ribosomal domain (90-98% with RBS), the K homology domain (KH) (90-100% with RBS), the double-stranded RNA-binding motif (DSRM) (85-100% with RBS), the zinc finger domain (49-92% with RBS), and the bromo domain (79-100% with RBS) (Fig. 5g), suggesting the reliability of the machine learning methods we used. Notably, the proportion of RBPs without predicted RBS is very close between published RBPs (10.9% for human, 17.9% for mouse) and those identified by HARD-AP (12.7% for human, 14.9% for mouse), indicating that HARD-AP did not capture significantly more non-specific associations and indirect RNA interactions (Supplementary Fig. 7d).

Notably, the catalytic domain of serine/threonine-specific kinase (STKc) or tyrosine-specific kinase (PKc) is strongly enriched in HARD-AP-identified RBPs in HEK293 cells and mouse samples; over 80% of these two domains harbor the predicted RBS (Fig. 5g). Several kinases, including CDK1, CDK9, and FAST family kinases FASTKD1 and FASTK2, were previously shown to directly bind to RNAs[8,61,62]. The STKc and PKc are conserved catalytic cores that are bilobal with a deep cleft between the lobes, and nucleotide-binding motifs line both sides of the cleft[63]. The predicted RBSs within these kinase domains have a high percentage of basic (21.7%) and hydrophilic amino acids (16.0%) (Supplementary Fig. 7e). RNA could compete with ATP to modulate kinase activity as is the case for CDK9 (Supplementary Fig. 7f), which is inhibited by binding to the 7SK small nuclear RNA[62].

## The conserved LIM domain as an RNA-binding domain

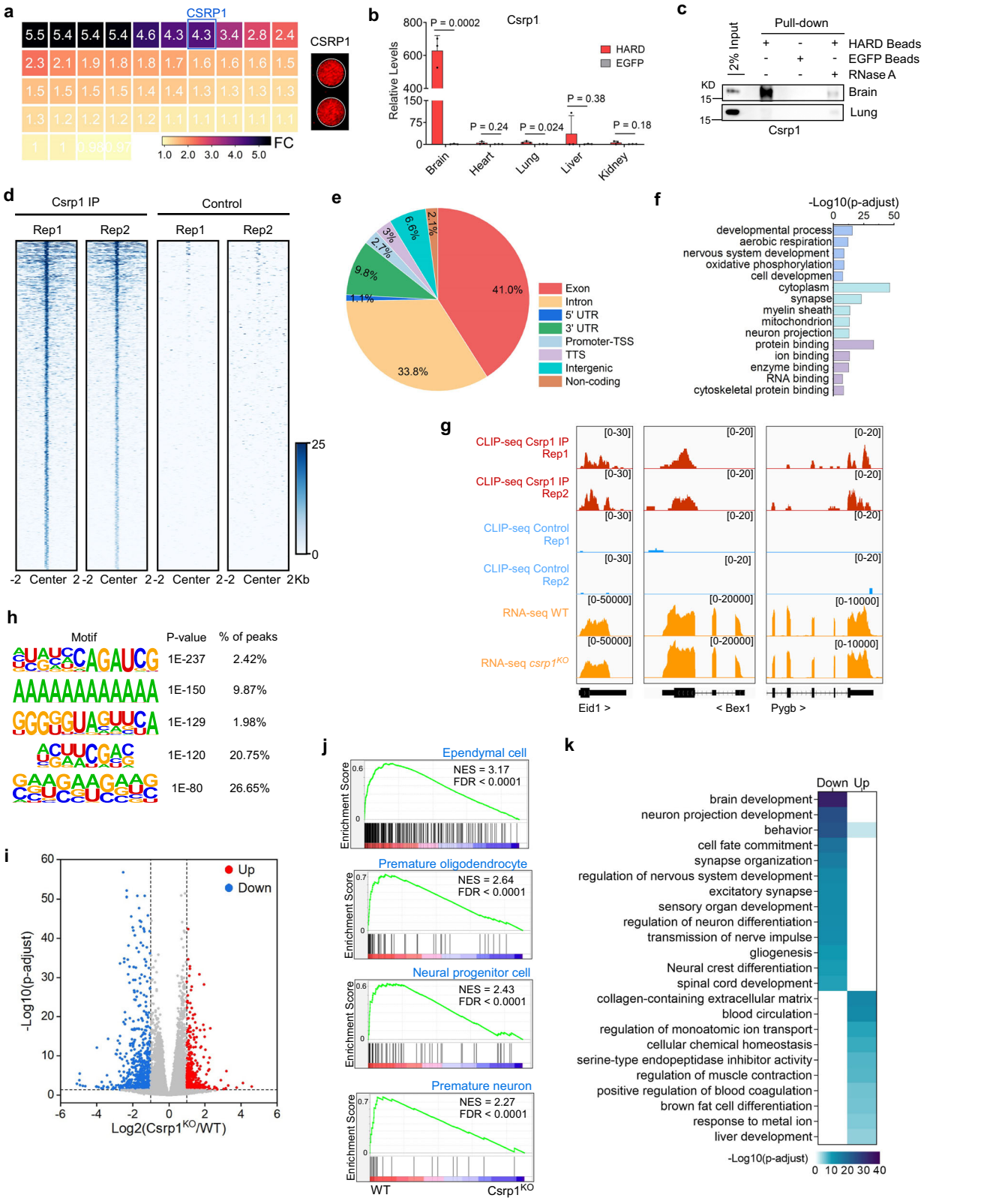
The conserved LIM domain is found in proteins involved in many processes including cell-fate determination, neuronal pathfinding, and tumorigenesis[64,65]. The LIM domain serves as a modular protein-binding interface, but it was not known as an RNA binding domain. In our HARD-AP RBPome, we identified dozens of LIM-containing proteins and found that 70-86% of the LIM proteins isolated by HARD-AP harbor the predicted RBS (Fig. 5g), suggesting the RNA-binding potential of the LIM domain. Notably, 24 LIM proteins showed specific direct RNA-binding activities with a signal of FC ≥ 1.5 on the protein microarray (Fig. 6a and Supplementary Fig. 8a), providing a good validation of the RNA-binding capacities of the LIM proteins and the reliability of the machine learning modeling methods we used.

The mouse protein Csrp1 protein has two LIM domains separated by an unstructured region (Supplementary Fig. 8b) and plays roles in neural development, lung fibrosis, smooth muscle development and cytoskeleton organization[66–69]. The human CSRP1 showed high RNA-binding activity on the protein microarray (Fig. 6a). In spite of the high protein levels of Csrp1 in lung and brain, Csrp1 was captured by HARD-AP in brain lysate but not in the lung lysate (Fig. 6b, c and Supplementary Fig. 8c), suggesting that the RNA-binding capability of Csrp1 is tissue/cell type-dependent. Importantly, the Csrp1 signal in the assay was dramatically diminished after RNase A treatment, suggesting that HARD beads captured Csrp1 through an interaction with RNA.

In the absence of differentiation inhibitor LIF, mESCs cultured in suspension spontaneously differentiate into three-dimensional aggregates called embryoid bodies (EBs) that recapitulate many aspects of early embryogenesis and induces various types of cells of three germ lineages[70]. Csrp1 was found to be highly expressed in embryoid bodies formed at day 9 (Supplementary Fig. 8d). To investigate the direct RNA-binding activity of Csrp1, we generated the mESCs that constitutively expressed V5-tagged Csrp1 (Supplementary Fig. 8e) and examined the RNAs bound to Csrp1-V5 in embryoid bodies using CLIP-seq as previously described[71]. Embryoid bodies that did not express V5-Csrp1 were used as the control. Two independent CLIP-seq experiments displayed high correlation; 10,028 specific peaks were identified by comparing to the controls (Fig. 6d and Supplementary Fig. 8f, g). Most of the Csrp1-binding were found to be on protein-coding RNAs (86.5%): exons (41.0%), introns (33.8%), and 3' untranslated regions (9.8%) (Fig. 6e and Supplementary Fig. 8h). GO analysis revealed that Csrp1-bound RNAs were expressed from genes significantly associated with cell development, cellular respiration, nervous system development, synapse, neuron projection, and so on (Fig. 6f). Examples of mRNAs bound by Csrp1 include those encoding neural differentiation regulators Eid1 and Bex1 and brain glycogen phosphorylase Pygb (Fig. 6g). A sequence motif enrichment analysis of the RNAs bound by Csrp1 uncovered multiple significant consensus motifs (Fig. 6h). These motifs comprise over 60% of Csrp1-binding peaks, suggesting that Csrp1 recognizes specific sequences of RNAs.

To understand the biological functions of Csrp1, we generated mESCs lacking Csrp1 through CRSPR/Cas9-mediated genome editing[72]. The loss of Csrp1 protein was confirmed by western blot (Supplementary Fig. 8i). Next, we carried out RNA-seq to compare the transcriptomes of wild-type mESCs and embryoid bodies and those that lack Csrp1. The absence of Csrp1 only caused few numbers of genes significantly expressed in $Csrp1^{KO}$ and WT mESCs and also did not significantly affect expression levels of genes of core transcriptional regulatory circuitry in mESCs[73] (Supplementary Fig. 8j-k). For the embryoid bodies derived from the mESCs, however, there were

https://doi.org/10.1038/s41467-024-52765-w




extensive differences in the transcriptomes of wild-type and those that lack Csrp1 (Fig. 6i). Gene Set Enrichment Analysis (GSEA) showed that the expression of genes regulating differentiation of neural lineage cells, including ependymal cell, premature oligodendrocyte, neural progenitor cell, and premature neuron, was significantly impaired when Csrp1 was deleted (Fig. 6j). GO analysis showed that the genes down-regulated in embryoid bodies lacking Csrp1 were significantly enriched in neural system-related processes (Fig. 6k). Together, these data suggest that Csrp1 is a tissue-dependent RNA binding protein required for normal neural lineage differentiation.

## Discussion

In this study, we engineered a RNA-binding protein HARD that has high affinity for various types of RNAs independent of their sequences or

**Fig. 6 | RNA-binding activity of the LIM domain. a** Left: Color matrix showing LIM domain protein on the protein microarray. Color scales represent the FC (Cy5 foreground signals over local background). Right: Image of Cy5-RNA incubation signal of the human CSRP1 on the protein microarray. **b** Relative levels of Csrp1 isolated from indicated mouse lysate using HARD beads and EGFP beads. Means ± SD are plotted; three independent biological samples were used for the analysis ($n = 3$). The $p$-values were calculated using a two-tailed Student's $t$-test. **c** Western blot analysis showing levels of Csrp1 isolated from the mouse brain and lung lysate using HARD beads and EGFP beads. This experiment was repeated once with similar results. **d** Heatmap of high-confidence CLIP-seq signals ± 2 kb around the center of peaks ($n = 10{,}028$). Csrp1 IP: EBs expressing Csrp1-V5; Control: WT EBs. Two biologically independent replicates are shown. **e** Distribution of Csrp1 CLIP-seq signals ($n = 10{,}028$) in different gene features. The average of the two biologically independent replicates is shown. **f** The top five GO terms associated with genes identified by CLIP-seq. The GO enrichment analysis used the two-sided Fisher's exact test with the $p$-value adjusted using the Bonferroni correction for multiple testing. **g** Representative genome browser tracks showing normalized CLIP-seq and RNA-seq signals. **h** The top five enriched Csrp1-binding motifs on all target RNAs (target sequences = 6071, background sequences = 39,667). Motif enrichment significance was computed by HOMER (binomial test without adjustment). **i** Volcano plot showing distributions of genes differentially expressed in $Csrp1^{KO}$ and WT embryoid bodies ($n = 22{,}625$). The significance (p) was determined using the two-tailed Student's t-test and further adjusted using the Benjamini-Hochberg correction for multiple testing (p-adjust). **j** Gene Set Enrichment Analysis of genes down-regulated in $Csrp1^{KO}$ embryoid bodies compared to WT embryoid bodies. NES, normalized enrichment score. **k** The most enriched GO terms in significantly expressed genes in the $Csrp1^{KO}$ embryoid bodies. The GO enrichment analysis used the two-sided Fisher's exact test with the $p$-value adjusted using the Bonferroni correction for multiple testing. Source data for (b-c) are provided as a Source Data file.

length. Using HARD, we developed a method that can robustly retrieve all RNA species and RBPs in cells and tissue samples. This allowed us to systematically characterize RBPs across major mouse tissues. These results identified 3985 new mouse RBPs and 4282 of their human homologs. Due to the limited efficiency of UV cross-linking, we combined the UV cross-linking and high salt wash condition to recover the RNA-regulatory complexes with high efficiency which include some subunits indirectly interacting with RNAs in the complexes. We performed the purification under high ionic strength conditions of 500 mM NaCl, which is more than three times of the cellular ionic strength and is stringent enough to isolate specific RNA-protein complexes with limited non-specific contanminants[27–30].

Thus, HARD-AP will recover both the RBPs that bind directly to RNAs and the proteins that tightly associate with these direct RBPs. However, given that the HARD protein fails to maintain its RNA-binding activity under the denaturing conditions, HARD-AP may contaminate potential non-specific binding proteins that could tolerate the high salt washing conditions. In the XRNAX study, Hentze et al. showed that partially digested RNA-protein covalent complexes can be efficiently recovered by the silica column under denaturing conditions, where the principle of silica matrix purification is based on the high affinity of the negatively charged backbone of nucleic acid towards the positively charged silica matrix. Based on the XRNAX study, we have successfully developed the tandem purification protocol combining the denaturing silica purification and HARD-AP in the Supplementary Fig. 9. In this protocol, we first performed the trysin/LysC partial digestion and isolated the UV crosslinked RNA-protein complexes through silica matrix column from the cell lysate under harsh denaturing conditions. Next, the RNA-protein complexes were eluted and further purified by the HARD beads. As showed in the Supplementary Fig. 9, the silver staining of the eluted samples showed that the HARD beads isolated a substantial amount of protein from the UV-treated sample while a negligible amount of proteins were isolated by HARD beads from non-UV-treated cell lysate. Additionally, little protein was detected when the samples were treated the RNase A prior to capture on the HARD beads. This protocol could be an easy alternative way to study the RNA-protein interactions under denaturing conditions with enhanced specificity, which is able to remove contaminants caused by negatively charged post-translational modifications or particular acidic sites within proteins. However, similarly, the efficiency of this tandem purification protocol largely depends on the efficiency of UV crosslinking. Notably, we also demonstrated the capacity of HARD-AP in robustly purifying RBPs from cell or tissue samples independent of crosslinking treatment.

We used the machine learning modeling methods to systematically map the RBS of all published and newly identified RBPs. These methods well mapped the known RBDs, and notably, help us to realize the conserved LIM domain as a RNA-binding domain. The RNA-binding activities of 24 LIM proteins were well validated by the RNA-protein interaction assay using the protein microarray. Furthermore, we demonstrated that the LIM-domain-only protein Csrp1 binds to RNA in neural cells to regulate neural lineage differentiation, highlighting the importance of studying RBP under physiological situations to uncover their biological functions and mechanisms.

We discovered that the organ-dependent RBPs were significantly associated with known physiological processes, suggesting the importance of RBPs in defining tissue-specific functions. In addition, the organ-derived RBPs are significantly enriched in metabolic enzymes such as hydrolase activity, transferase activity, kinase, and oxidoreductase activity in addition to RNA-related processes. These results suggest that the physiological environment may require enzymes to participate in extensive networks of protein-RNA interactions to achieve their physiological roles. Thus, many proteins are tissue/cell type-specific RBPs rather than canonical RBPs such as Csrp1, Bcr, Prkar1a and Mylk3 we tested. Their moonlighting RBP function can contribute to regulate gene regulation, cellular localization and enzymatic activitygg. We expect that the data reported here provide comprehensive and physiologically relevant tissue-specific networks of RNA-protein interactions and will serve a foundation for future studies of RBP functions and mechanisms. In addition, the HARD-based RNA purification methods can be a power tool to examine RNA-protein interaction in all cell types and tissues.

Furthermore, the HARD protein can be further engineered to create new research or therapeutic applications, such as intracellular RNA delivery tool by fusing HRAD with cell-penetrating peptide and as RNA modifiers by fusing it with RNA modifying enzymatic domains.

## Methods

### Expression and purification of the HARD protein

EGFP or EGFP-HARD was cloned into pET28a vector backbone with an additional 10xHis-tag on the N-terminus. The EGFP-HARD or EGFP alone were overexpressed in *E. coli* BL21(DE3) cells. One liter of cell culture was grown at 37 °C for overnight in LB medium with 50 μg/ml kanamycin until the optical density at 600 nm ($OD_{600}$) reached 0.8. IPTG was added to the final concentration of 0.2 mM and the culture grown at 16 °C for 24 h (hrs). Cells were collected and resuspended in Buffer A (1xPBS, 1 M NaCl and 10 mM imidazole). Cells were then lysed by sonication, and centrifuged at 4 °C for 30 min at 18,407 g. Solubilized proteins in the supernatant were purified using Ni-NTA resin (Cube biotech) and eluted with Buffer A with extra 500 mM imidazole. The eluted proteins were concentrated to remove imidazole and RNase through 30-KDa cut-off Amicon Ultra centrifugal filter unit (Millipore). Finally, the purity of the EGFP-HARD and EGFP protein were analyzed by SDS-PAGE and Coomassie Blue staining. Protein concentration was determined by OD280.

### Preparation of HARD/EGFP beads

Purified EGFP-HARD/EGFP proteins were changed to phosphate buffer (100 mM phosphate buffer pH7.0, 150 mM NaCl) through GM1250

desalting resin. EGFP-HARD/EGFP proteins were conjugated onto NHS-activated resin following manufacturer's instruction. Briefly, proteins were reacted with NHS resin for 1 hr at room temperature in phosphate buffer. Next, the resin was incubated in blocking buffer (1xPBS pH7.5, 1 M ethanolamine) for 2hrs at room temperature, and then washed with buffer (100 mM phosphate buffer, 500 mM NaCl). Finally, the resin was preserved in buffer (50 mM Tris-HCl pH7.4, 150 mM NaCl, 10% Glycerol) and stored at 2–8 °C for long terms.

### Plasmid preparation

The piggyBac 5′ and 3′ inverted repeats were synthesized and cloned into pUC57 to obtain pUC57.piggyBac. The EF1a promoter (from Addgene# #26777), eSpCas9(1.1) (from Addgene #71814), and IRES-NeoR-WPRE cassette (from Addgene#50917) were amplified from indicated plasmids and ligated in order to pUC57.piggyBac to get the PiggyBac_EF1a-eSpCas9-IRES-NeoR-WPRE using Gibson assembly master mix. The mouse Csrp1 ORF fused with V5 tag on the 3′ end was synthesized by Shanghai Sangon Biotech. The EF1a promoter (from Addgene# #26777) and Csrp1-V5 were ligated in order to pUC57.piggyBac to get the PiggyBac_EF1a-Csrp1-V5 using Gibson assembly master mix. The PGK promoter-PuroR-SV40 polyA cassette and two copies of U6 promoter-guide RNA scaffold (one copy with two Bbs I sites and the other with two Bsa I sites for inserting CRISPR targeting sequence) were synthesized and ligated in order to pUC57 to obtain pUC57_sgRNAduo-Puro using Gibson assembly master mix. Two targeting sequences of mouse Csrp1 were cloned into pUC57_sgRNAduo-Puro separately via Bbs I and Bsa I sites to obtain Csrp1_sgRNA1&2.

### Antibodies used in the study

Anti-DNMT1(Sino Biological, cat#100780-T10, 1:1000), anti-PTBP1(Sino Biological, cat#101043-T46, 1:1000), anti-Histone H3(Sino Biological, cat#100005-MM01, 1:10,000), anti-PSPC1(Proteintech, cat#16714-1-AP, 1:1000), anti-NONO(Proteintech, cat#11058-1-AP, 1:1000), anti-Bcr(Proteintech, cat#22585-1-AP, 1:1000), anti-Prkar1-a(Proteintech, cat#20358-1-AP, 1:1000), anti-Mylk3(Proteintech, cat#21527-1-AP, 1:1000), anti-Csrp1(ABclonal, cat#A19842, 1:1000), anti-V5 tag(Sino Biological, cat#100378-T36, 1:2000), anti-Gapdh(-Proteintech, cat#60004-1-Ig, 1:10000), anti-β-tubulin (Proteintech, cat#10068-1-AP, 1:10,000), Goat anti-mouse IgG (H + L) HRP (Sino Biological, cat#SSA007, 1:1000), Goat anti-rabbit IgG (H + L) HRP (Sino Biological, cat#SSA004, 1:1000).

### Cell culture

HEK293 cells were maintained in the medium (Dulbecco's Modified Eagle's Medium, 10% NBS, 1 × non-essential amino acid solution, 1 × GlutaMAX, 1 mM sodium pyruvate, 0.1 mM β-mercaptoethanol, 100 U/ml penicillin, 100 mg/ml streptomycin) under the condition of 37 °C and 5% $CO_2$. mESC Cell line V6.5 was from Laurie Boyer lab of MIT. V6.5 cells were maintained in the medium (DMEM, 10% FBS, 2,000U/ml LIF, 1 × non-essential amino acid solution, 1 × GlutaMAX, 1 mM sodium pyruvate, 0.1 mM β-mercaptoethanol, 100U/ml penicillin, 100 mg/ml streptomycin) under the condition of 37 °C and 5% $CO_2$.

### Embryoid body differentiation

mESCs were pre-plated to remove feeders and diluted to 100,000 cells/ml in standard ESCs medium lacking LIF. 2.5 ml diluted cells were plated on ultra-low-attachment 6-wells plate to induce aggregation. The medium was changed every other day. Ascorbic acid was added to a final concentration of 50 μg/ml from day 2 to day 9. EBs were collected at day 9 for RNA-seq and CLIP-seq.

### Generation of Csrp1$^{KO}$ mESC cell line and Csrp1-V5 over-expressing mESC cell line

The PiggyBac_EF1a-eSpCas9-IRES-NeoR-WPRE plasmid and Supper PiggyBac Transposase plasmid (Beijing Zoman Biotech) were transfected into V6.5 mESCs with the Lonza Nucleofector 2b using the mouse ES cell nucleofector Kit; these transfected cells were treated with G418 (400 μg/ml) for 5 days to obtain mESC cell line constitutively expressing eSpCas9 (Cas9 mESCs), which was considered as wildtype cells in the case of comparing gene expression with Csrp1$^{KO}$ cells. The Csrp1 targeting sgRNAs plasmid (Csrp1_sgRNA1&2) was transfected into Cas9 mESCs using Lonza Nucleofector as above. Two days later, puromycin was added to the medium with a final concentration of 1 μg/ml for 2 days' treatment. The puromycin-resistant cells were largely diluted and grown for 3 days. The clones were picked up individually under microscope and screened for Csrp1$^{KO}$ mESCs by western blot analysis using anti-Csrp1 antibody. The PiggyBac_EF1a-Csrp1-V5 plasmid and Supper PiggyBac Transposase plasmid were transfected into V6.5 mESCs using Lonza Nucleofector as above; these transfected cells were treated with puromycin (1 μg/ml) for 3 days to obtain Csrp1-V5 over-expressing mESCs.

### Mice

In this research, primary organs were isolated from 8 to 9 week-old mice (Mus musculus, C57BL/6 J). This mouse strain was originally acquired from GemPharmatech and housed in the Laboratory Animals facility at Sichuan University. The environmental conditions for their care included a temperature range of 18 − 22 °C, 50 − 60% humidity, and a 12 h light/dark cycle.

### Isolation of total RNAs and genomic DNAs

Cells were lysed in the Trizol regent. Total RNAs were isolated following manufacturer's instruction. For genomic DNAs, cells were lysed in the buffer (10 mM Tris-HCl pH 7.5, 10 mM EDTA, 10 mM NaCl, 0.5% N-Lauroylsarcosine sodium salt, 400 μg/ml Proteinase K) and then incubated at 55 °C overnight. The lysate was then precipitated by adding equal volume of isopropanol. The DNAs pellet was washed with 70% ethanol and dissolved in H2O.

### Nucleic acid-binding activity analysis of the HARD protein

A 400 μl binding assay was set up as below: 1x PBS, 15 μl HARD/EGFP beads, 10 μg HEK293 total RNAs/genomic DNAs/heat-denatured genomic DNAs, 200 U murine RNase inhibitor (mRI), 4 U DNase I (omit for DNAs-related assays). Genomic DNAs dissolved in H2O were heated for 5 min at 95 °C and immediately put on ice to get single-stranded DNAs (ssDNAs, heat-denatured genomic DNAs). Meanwhile, 10 μg of the same nucleic acids were diluted with ddH2O to 100 μl as input. The assay mixture was incubated at RT for 2 h with rotation. Next, beads were washed three times with 1 ml wash buffer (20 mM Tris pH 7.4, 500 mM NaCl, 0.1 % Tween-20) and 1 ml wash buffer (20 mM Tris pH 7.4, 50 mM NaCl, 0.1% Tween-20), alternately. Nucleic acids were eluted through incubating beads in 100 μl digestion buffer (1xPBS + 0.5% SDS) supplied with 1.6U Proteinase K (NEB) at 55 °C for 30 min. 5 μl 10% SDS and 1.6U Proteinase K were directly added to the input samples which were then following the same treatments. Next, each sample was supplied with 100 μl H2O, 100 μl 1-Bromo-3-chloropropane (BCP), and 100 μl phenol, mixed well by vertexing for 30 s and centrifuged at 15,871 g for 15 min at 4 °C. The supernatant was transferred to a new tube, mixed with 200 μl BCP, vortexed for 30 s and centrifuged at 15,871 g for 15 min at 4 °C. The supernatant was transferred to a new tube and supplied with 20 μl 5 M NaCl, 2 0 μl 3 M sodium acetate (NaOAc) pH5.2, and 2 μl glycogen (5 mg/ml, ThermoFisher). The mixture was then precipitated by adding 2.5 x volumes of 100% ethanol and centrifuged at 15,871 g for 15 min at 4 °C. The pellets were washed with 80% ethanol twice, and then dissolved in 23 μl H2O. After measuring the concentration, 1 μg RNAs were taken for 20 μl reverse transcription reaction (RT), and 1 μl RT products was used for 10 μl qPCR reaction. 1 μl dissolved DNAs were used for 10 μl qPCR reaction. qPCR was calculated by the ΔΔCT method and Percent Input method. All reactions were performed in duplicate. Oligonucleotide sequences are provided in Supplementary Data 1.

### Isothermal titration calorimetry

The binding assay was performed using a PEAQ ITC (Malvern Panalytical, UK) at 25 °C. The concentration of HARD protein was adjusted to 35 μM. The HARD protein was purified as described above. The concentrated protein was passed over the gel filtration and displayed as the monomer. The peak fractions from the gel filtration were collected and diluted to 35 μM with the gel filtration buffer (20 mM HEPES pH 7.0, 150 mM NaCl). The nucleic acid including ssRNA (AUGCAUGC), ssDNA (ATGCATGC), dsRNA (AUGCAUGC) or dsDNA (ATGCATGC) was separately dissolved in gel filtration buffer and then diluted to a final concentration of 200 μM (except dsDNA was 190 uM) with gel filtration buffer. The nucleic acid was injected 19 times (0.4 μl for injection 1 and 2 μl for injections 2–19) with 120 s intervals between injections. The titration data were analyzed using a one-site binding model, and the first injection was removed. The titration of nucleic acid into the buffer was deducted. The binding affinity (Kd) is presented as the Mean ± SD.

### RNA sequencing (RNA-seq)

Ribo-Zero RNA-seq: RNA-seq libraries were prepared from 2 μg of HEK293 input RNAs or HARD beads-bound RNAs. We first removed ribosomal RNAs (rRNAs) using Epicentre Ribo-Zero rRNA Removal Kit (Human). The libraries were generated using Illumina Stranded Total RNA Prep kit, purified by AMPure XP beads (Beckman), and quantified using the Agilent high sensitivity DNA assay on a Bioanalyzer 2100 system (Agilent). The libraries were finally sequenced on NovaSeq 6000 platform (Illumina).

mRNA RNA-seq: Libraries were prepared from 1 μg total RNAs. We purified the mRNA using VAHTS mRNA capture beads (Vazyme). The libraries were generated using VAHTS Unisersal V8 RNA-seq Library Prep Kit for Illuminia (Vazyme), and evaluated using Qsep with S2 Cartridge. The libraries were finally sequenced on NovaSeq 6000 platform (Illumina).

Bioinformatic analysis: The pair-end (PE) sequencing reads were first analyzed for quality control using FastQC (v0.11.9) (Babraham Bioinformatics), filtered and trimmed off adapters using Trim Galore (v0.6.7) (Babraham Bioinformatics). Ribosomal RNAs (rRNAs) were removed from trimmed reads using SortMeRNA[74]. Reads without rRNAs were mapped to hg38 human genome using STAR[75] (v2.7.1a) with default settings. RNA levels of each gene and biotypes of RNAs were quantified using FeatureCounts[76] (v2.0.1) and then normalized by the FPKM method. The bigwig files of aligned reads were generated by deeptools[77] (v2.0) and visualized using IGV[78] genome browser. Metagene was generated using deeptools[79].

### Isolation of RBPs through HARD-AP

For HEK293 cells or mESC cells, cultured cells with 80% confluence were first washed three times with ice-cold 1xPBS, and then immediately treated for 400 mJ/cm$^2$ at 254 nm wavelength using the Analytik Jena UV Crosslinker. Five primary organs (brain, heart, lung, liver, and kidney) were dissected from 8–9 weeks old mouse (C57BL/6), and then were washed three times with ice-cold 1xPBS to remove residual blood. After drying with gauze quickly, organ tissues were completely frozen with liquid nitrogen, and ground into powder under liquid nitrogen in a ceramic grinder. The ground powder was next transferred to a stainless-steel dish which was pre-cold with dry ice, and immediately cross-linked with a dosage of 500 mJ/cm$^2$ at 254 nm using the Analytik Jena UV Crosslinker as previously described[49].

UV-treated cells cultured in 10 cm dish (80% confluence) were lysed in 1 ml lysis buffer (1× PBS, 5 mM $MgCl_2$, 0.5 mM $CaCl_2$, 2000 U mRI (omit in RNase A-treated negative control samples), 1x cOmplete Proteinase inhibitor (Sigma)). Lysed cells were solubilized by sonication using ultrasonic disruptor with a 2 mm probe. The sonication program is as below: 5 s on, 25 s off at the power of 20 W. Sonicated cell lysate was centrifuged at 20,000 g for 5 min at 4 °C. 50 μl cell lysate was saved as the input. Cell lysate was supplied with 10ul DNase I and incubated at 37 °C for 1 h to clean up DNAs. For RNase A-treated control samples, extra 1 mg RNase A (Sigma) was added to the lysate besides DNase I, and cell lysate was incubated at 37 °C for 24hrs to completely clean up RNAs. HARD/EGFP beads were equilibrated with the lysis buffer, and then 0.5 ml beads were incubated with the cell lysate for 2 hrs at RT with rotation. Next, beads were washed three times with 1 ml wash buffer (20 mM Tris pH 7.4, 500 mM NaCl, 0.1 % Tween-20) and 1 ml wash buffer (20 mM Tris pH 7.4, 50 mM NaCl, 0.1% Tween-20), alternately. Finally, RBPs were eluted by 1.2 ml 8 M Urea solution for 5 min at RT. For tissue samples, 100 μl UV-treated tissue powder was lysed in 1 ml lysis buffer. Protein lysate was treated as above and incubated with 1 ml HARD/EGFP beads for binding. The left procedures were the same as above except 2.4 ml 8 M Urea solution for elution. 80% eluted proteins were used for LC-MS/MS analysis, 12.5% for silver staining analysis, and 4% for western blot.

For silica-HARD-AP tandem purification, 1 mg cell lysate was diluted in 0.5 ml TDB buffer (Tris-HCl pH7.4, 0.1% SDS) and partially digested by 200 ng Trypsin/LysC (Promega#V5071) for 30 min at 37 °C. The reaction was stopped by adding 3.5 ml Zymo Quick-RNA Midiprep Kit (Zymo#R1056) ZR RNA buffer and heated form 15 min at 60 °C. The cooled down lysate was mixed with 4.5 ml 100% ethanol and loaded on to the silica column in the kit. Do not discard the flow-through but save it for multiple purification. The column was then washed with 400 μl DNX buffer (50% ethanol, 40% Zymo ZR RNA buffer), 400 μl DNY buffer (2 M guanidinium chloride, 60% isopropanol) and 400 μl RPE buffer (80% ethanol, 100 mM NaCl, 10 mM Tris-HCl pH7.4). The column was eluted using 250 μl H2O. The saved flow-through was reloaded onto the column and washed with the same procedures as above. We repeated the purification four times for each sample and combined all elute in one tube (~1 ml). The elute was mixed with 100 μl 10x PBS buffer, loaded onto the HARD/EGFP beads and incubated for 2hrs with rotation at room temperature. Next, beads were washed three times with 1 ml wash buffer (20 mM Tris pH 7.4, 500 mM NaCl, 0.1 % Tween-20) and 1 ml wash buffer (20 mM Tris pH 7.4, 50 mM NaCl, 0.1% Tween-20), alternately. Finally, RBPs were eluted by 1.2 ml 8 M Urea solution for 5 min at RT.

### Western Blot and silver staining analysis of SDS-PAGE gel

20% Eluted proteins were mixed with 1/4 volume 100% trichloroacetic acid (TCA) and incubated for overnight at −20 °C. Precipitated proteins were collected by centrifuging at 15,000 g for 15 min °at 4 °C. The pellet was washed by cold acetone and dissolved in 40 μl 1xPBS, which was then treated with 2 ug RNase A for 1 hr at 37 °C to remove conjugated RNAs. 25 μl was used for silver staining analysis and 8 ul for western blot.

For western blot, proteins were separated by SDS-PAGE and electro-transferred to the 0.45 μm PVDF membrane. The membrane was blotted under 5% milk prepared by nonfat-dried milk and PBST (1x PBS + 0.1% Tween) and washed by PBST. The primary antibodies were incubated with membranes for overnight at 4 °C, and the HRP-conjugated secondary antibodies were incubated for 30 min at RT. For silver staining, proteins were separated on ExpressPlus PAGE Gel 4–20% (Genscript), which was stained using PAGE Gel Silver Staining Kit following manufacture's protocol.

### Liquid chromatography-tandem mass spectrometry (LC-MS/MS)

There are three independent biological replicates for all samples. EGFP samples were used as the control for the samples of HEK293 cells and five mouse organs under the conditions of UV treatment or non-UV treatment. We searched the database in five groups: HEK293 HARD, HEK293 EGFP, HEK293 RNase for the group 1; mESCs, mouse five organs for the group 2; HEK293 HARD, HEK293 EGFP, HEK293 RNase, HEK293 non-UV crosslinked for the group 3; Brain HARD, Brain EGFP,

Brain non-UV crosslinked for the group 4; Kidney HARD, Kidney EGFP, Kidney non-UV crosslinked for the group 5.

Trypsin Digestion: The protein solution was reduced with 5 mM dithiothreitol for 30 min at 56 °C and alkylated with 11 mM iodoacetamide for 15 min at RT in darkness. The alkylated samples were transferred to ultrafiltration tubes for FASP digestion. The samples were firstly replaced with 8 M urea for 3 times at 12000 g at room temperature for 20 min, and then replaced with 100 mM TEAB for 3 times. Trypsin was added at 1:50 trypsin-to-protein mass ratio for digestion overnight. The peptide was recovered by centrifugation at 12000 g for 10 min at RT and repeated for two times. Finally, the combined peptides were desalted by C18 SPE column.

LC-MS/MS Analysis: The tryptic peptides were dissolved in solvent A (0.1% formic acid, 2% acetonitrile/ in water), directly loaded onto a home-made reversed-phase analytical column (25 cm length, 75 μm i.d.). Peptides were separated with a gradient from 4% to 20% solvent B (0.1% formic acid in 90% acetonitrile) over 96 min, 20% to 32% in 18 min and climbing to 80% in 3 min then holding at 80% for the last 3 min, all at a constant flowrate of 500 nL/min on an EASY-nLC 1200 UPLC system (ThermoFisher). The separated peptides were analyzed in Exploris 480TM (ThermoFisher) with a nano-electrospray ion source. The electrospray voltage applied was 2.3 kV and the compensation voltages was −70 V. The full MS scan resolution was set to 60,000 for a scan range of 400–1200 m/z. Up to 15 most abundant precursors were then selected for further MS/MS analyses with 25 s dynamic exclusion. The HCD fragmentation was performed at a normalized collision energy (NCE) of 27%. The fragments were detected in the Orbitrap at a resolution of 30,000. Fixed first mass was set as 110 m/z. Automatic gain control (AGC) target was set at 75%, with an intensity threshold of 1E4 ions/s and MS2 maximum injection time was set as 100 ms.

Database Search: The resulting MS/MS data were processed using Proteome Discoverer search engine (v2.4.1.15). Tandem mass spectra were searched against Mus_musculus_10090_SP_20210721.fasta database (17089 entries) and Homo_sapiens_9606_SP_20200509.fasta database (20366 entries) concatenated with reverse decoy database. Trypsin (Full) was specified as cleavage enzyme allowing up to 2 missing cleavages. The mass tolerance for precursor ions was set as 10 ppm in first search and the mass tolerance for fragment ions was set as 0.02 Da. Carbamidomethyl on Cys was specified as fixed modification, and oxidation on Met and acetylation on protein N-term, were specified as variable modifications. FDR was adjusted to <1% and minimum score for modified peptides was set >40. Minimum peptide length was set at 6. All the other parameters in Proteome Discoverer were set to default values.

Protein quantification: Given that HARD and EGFP samples exhibit large difference in the complexity, the intensity of ion peaks quantified by Proteome Discoverer was used for protein quantification. This method displayed great performance in quantification yield, dynamic range, and reproducibility[80]. We performed the imputation of missing values using SampMin method[81] if a protein is observed at least twice in three independent samples, where SampMin method replaces all missing values in a sample with the minimum intensity value of that sample. To define the pool of RBPs captured by HARD-AP, we applied the following criteria: first, proteins were identified with two or more unique peptides in at least two out of three independent HARD beads samples; second, for quantified intensity, we selected proteins with at least threefold more signals in HARD beads samples than in EGFP beads samples and at the same time displayed a p-adjust using the Benjamini-Hochberg correction for multiple testing of <0.01.

### Normalization and hierarchical clustering of proteins captured by HARD-AP

Hierarchical clustering analysis: The heatmap of the hierarchical clustering analysis was generated using R package pheatmap with the following setting: clustering_distance_rows = euclidean, clustering_method = complete. The ion intensities of the mass spectrometry for the hierarchical clustering analysis were standardized by the R function scale().

Correcting the HARD-AP abundance with their endogenous protein levels: The mass spectrometry data of endogenous proteins in each organ were collected from the study by Guo T. et al.[82]. In this study, the proteomes of 41 mouse organs/tissues were quantitatively measured using the mouse of the same strain (C57BL/6) and age. We performed the imputation of missing values using SampMin method[81]. The ion intensities of the mass spectrometry of these samples were standardized by the R function scale() as well. We corrected the abundance of proteins isolated by HARD-AP by $m - n$, where $m$ = standardized ion intensities of proteins in HARD-AP, $n$ = standardized ion intensities of corresponding endogenous proteins. The corrected levels of HARD-AP RBPs were directly used for the hierarchical clustering analysis as above.

### Visualization of protein structures

Crystal structures or predicted 3D structures by Alphafold are visualized by PyMOL[83]. Electrostatic potential mapped onto the molecular surface of proteins were calculated by the tool in PyMOL as well.

### Gene Ontology (GO) Enrichment analysis and Gene Set Enrichment analysis (GSEA)

GO enrichment analysis was performed using the GO Consortium web interface (http://geneontology.org/) and UniProt identifier as input[84,85]. For all GO enrichment analyses, Fisher's exact test was used, with the $p$-value adjusted using the Bonferroni correction for multiple testing. GSEA was performed using GSEA software[86] (http://www.gsea-msigdb.org/gsea/) and the Molecular Signatures Database (MsigDB)[87]. All plots of GO enrichment were generated using Origin. The GSEA plots were generated by the GSEA software.

### Analysis of RBPs in hydrophobicity, isoelectric point and orthology

Proteins were computed using R package 'peptides' with the scales 'Kyte-Doolittle' for hydrophobicity, 'EMBOSS' for isoelectric point. Ortholog analysis was performed using bioDBnet web interface and UniProt identifier as input[52].

### Mapping of RNA-binding sites (RBS) within RBPs

As described in the study of Ying et al.[56], we installed the hierarchical graph neural networks-based deep learning predictor GraphBind on our local server and selected 495 non-redundant RNA-binding protein chains[56] (Supplementary Data 8) to train the predictor. The 3D structures of all RBPs predicted by AlphaFold were downloaded from AlphFold Protein Structure Database[58] and used as the input of GraphBind. IDRs and RBS within IDRs were mapped by sequence-based machine learning predictor flDPnn[57]. IDR regions harboring at least four consecutive resides with RNA-binding score over 0.5 were defined as RBS. The protein sequences were used as input.

### CLIP-seq

Library preparation and sequencing: We performed the CLIP-seq following the eCLIP-seq protocol[71] except several modifications as below. The V6.5 mESCs over-expressing V5-tagged Csrp1 and wildtype V6.5 mESCs were differentiated into embryoid bodies as described above separately. The embryoid bodies (EBs) of 9 days were resuspended in the PBS buffer and treated with UV crosslinking (254 nm, 400 mJ/cm2) using a Stratalinker. Crosslinked EBs were resuspended in 1 ml iCLIP lysis buffer (50 mM Tris pH 7.4, 100 mM NaCl, 1% Igepal CA630, 0.1% SDS, 0.5% sodium deoxycholate, RNase inhibitor, and protease inhibitor cocktail) and then solubilized for 1 min via sonication with a Covaris S220 instrument using following parameters: PIP 140 W, Duty factor 5%, CPB 200. The lysates were limited digested with 0.4 μl RNase

I for 5 min at 37 °C and then centrifuged for 15 min with 15,000 g. The 100 μl protein G magnetic beads were incubated with 10 μg anti-V5 tag antibody for 30 min at 4 C to prepare anti-V5 magnetic beads, and then added to the cell lysates and incubated for 2 h at 4 °C. Next, the beads were washed as below: two times with 1 ml high salt wash buffer (50 mM Tris pH 7.4, 1 M NaCl, 1% Igepal CA630, 0.1% SDS, 0.5% sodium deoxycholate); two times with 1 ml wash buffer (20 mM Tris pH7.4, 10 mM MgCl2, 0.2% Tween-20); two times with 1 ml 1x DNase buffer (10 mM Tris pH 7.4, 2.5 mM MgCl2, 0.5 mM CaCl2). Beads were resuspended in 100ul 1x DNase buffer supplemented with 4 U DNase I and incubated for 20 mins at 37 °C. Beads were then washed two times with 1 ml PK buffer (20 mM Tris pH 7.4, 50 mM NaCl, 1 mM EDTA). Beads were resuspended in 260ul PK buffer and supplemented with 40 μl proteinase K (NEB). Beads were incubated for 2 hrs at 37 °C and then mixed well with 200 μl phenol and 200 μl 1-Bromo-3-chloropropane (BCP), which was then centrifuged at 15,871 g for 15 min at 4 °C. The supernatant was transferred to a new tube, mixed with 200 μl BCP, vortexed for 30 s and centrifuged at 15,871 g for 15 min at 4 °C. The supernatant was transferred to a new tube and supplied with 30 μl 3 M sodium acetate (NaOAc) pH 5.2, and 2 μl glycogen (5 mg/ml, ThermoFisher). The mixture was then precipitated by adding 2.5x volumes of 100% ethanol and centrifuged at 15,871 g for 15 min at 4 °C. The pellets were washed with 80% ethanol twice, and then dissolved in 15 μl H2O. The resultant RNAs were first ligated with 3′ barcoded (NNNNNNNNNNNNNNN) RNA adapter and then 5′ RNA adapter. cDNAs were synthesized with M-MLV reverse transcriptase (Vazyme) and amplified by VAHTS HiFi amplification mix (Vazyme). The PCR products were cleaned using DNA clean beads (Vazyme) and separated on agarose gel. The library was purified using MinElute Gel Extraction Kit (Qiagen) and evaluated using Qsep with S2 Cartridge. The libraries were finally sequenced on NovaSeq 6000 platform (Illumina).

Bioinformatic analysis: We performed data processing following the eCLIP-seq pipeline (https://github.com/YeoLab/eCLIP)[71]. The sequencing reads were first analyzed for quality control using FastQC (v0.11.9) (Babraham Bioinformatics), and then unique molecular barcodes were extracted by umi_tools (v1.1.2). Reads were trimmed off adapters and filtered <18 bp using Cutadapt (v4.1) and aligned to the UCSC mm10 genome using STAR software (2.7.10a)[75]. High-quality mapping reads were extracted by setting parameter samtools -q to 40. Duplicate reads were removed using umi_tools dedup. Only uniquely mapping and de-duplicated reads (quality score > 40) were retained. To create CLIP-seq coverage plots, scale factors were calculated by ChIPseqSpikeInFree software(v1.2.4)[88], and the reads coverage were normalized by setting scaleFactor parameter and reformatted in the bigWig file format using deeptools (v3.5.1)[77]. The bigwig files of aligned reads were visualized using IGV[78] genome browser. CLIP-seq peaks were called using Clipper software (v2.1.2) with default parameters[71]. Significantly differentially binding sites between sample groups were identified using DiffBind software (v3.4.11) with the significance cut-off $q$-value ≤ 0.05 and fold change (FC) ≥ 2. Next, we use bedtools (v2.25.0) to identify overlaps between significantly differentially binding sites from Diffbind and original peak sites from Clipper to retrieve original peak sites that were used subsequently as input file of HOMER software. Motif finding for the CLIP-seq peaks of Csrp1 was performed with HOMER findMotifs program (−rna). Peak annotation was performed with HOMER annotatePeaks program (mm10).

### Protein microarray processing and analysis

Preparation of the pool of Cy5-labeled RNAs: The total RNAs of HEK293 cells were isolated using the Trizol regent following manufacturer's instruction and treated with DNase I to remove the contaminated genomic DNAs, which were then extracted by the 1-Bromo-3-chloropropane and precipitated by 2 volumes of 100% ethanol. To get fragmented total RNAs, 1.125 μg total RNAs was dissolved in 45 μl 1x Fragmentation buffer (100 mM Tris-HCl pH 8, 2 mM MgCl2) and heated for 6 mins at 94 °C. 32 tubes of fragmented products above were collected and precipitated by 2 volumes of 100% ethanol. The fragmented total RNAs were measured by the NanoDrop and analyzed by the Qsep Bio-Fragment analyzer. For in vitro transcribed RNAs, we first removed the rRNAs of 1 μg total RNAs of HEK293 cells using the Ribo-MagOff rRNA Depletion Kit (Human/Mouse/Rat) (Vazyme N420), which was then reverse transcribed into single-stranded cDNA using the M-MLV reverse transcriptase (Vazyme R021) with equal amount of RT6N (CGTGTGCTCTTCCGATCNNNNNN) and RT23T (CGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTTTTTTTTT) primers; T7 promoter was incorporated into the in vitro templates by PCR using the primer (TAATACGACTCACTATAGGGNNNNNNNN); the double-stranded templates of the T7 in vitro transcription were amplified by PCR using primers (Foward: TAATACGACTCACTATAGGG; Reverse: CGTGTGCTCTTCCGATC); the RNAs were finally produced by the T7 High Yield RNA Transcription kit (Vazyme TR101) using 1 μg template, and finally extracted by the 1-Bromo-3-chloropropane and precipitated by 2 volumes of 100% ethanol. RNAs were labeled using Label IT Nucleic Acid Cy5 Labeling Kit (Mirus MIR3700). The materials were reconstituted based on instructions of the Kit. We optimized the labeling procedure from the original manufacturers' protocol. For fragmented total RNAs, 25 μg RNAs in 25 μl H2O was mixed with 20 μl Label IT reagent, 20 μl Labeling buffer A, 135 μl H2O to obtain a final volume of 200 μl, and incubated for 1 h at 37 °C; for in vitro transcribed RNAs, 80 μg RNAs in 80 μl H2O was mixed with 40 μl Label IT reagent, 30 μl Labeling buffer A, 150 μl H2O to obtain a final volume of 300 μl, and incubated for 1 h at 37 °C. Each sample was then supplemented with 10 μg glycogen (Thermofisher), 1/10 volume of 5 M NaCl and 3 volume of 100% ethanol. After precipitating for at least 1 h at −20 °C, RNAs were washed with 80% ethanol and re-suspended in 30 μl H2O. The labeled RNAs could be stored at −80 °C, or proceed directly with the microarray hybridization. RNA labeling density was evaluated using NanoDrop. The efficacy of Cy5 dye incorporation was calculated as RNA Base:Dye ratio using following formulas[89].

$$\text{Base : Dye ratio} = (A_{base} * e_{dye})/(A_{dye} * e_{base})$$

$$A_{base} = A_{260} - (A_{dye} * C.F._{\cdot 260})$$

where

$A_{dye}$ − absorbance at excitation wavelength,Cy5(649nm)

$e_{dye}$ − extinction coefficient : $250000 M^{-1}cm^{-1}$(Cy5)

Abase − RNA base absorbance : $A_{260} - 0.05A_{dye}$(Cy5)

$e_{base}$ − RNA extinction coefficient : $8250 M^{-1}cm^{-1}$

$A_{260}$ − absorbance of nucleic acid at260nm

$C.F._{\cdot 260}$ − correction factor at260nm : 0.05(Cy5)

The RNA labeling Base/Dye labeling ratio in this work is presented in the Supplementary Fig. 3d.

Protein microarray hybridization and analysis: HuProt Human Protein Microarray v4.0 (CDI laboratories) was used. We performed the hybridization following the manufacturers' protocol with several modifications. Each microarray with the barcode facing up was incubated in 4.5 ml blocking buffer (40 mM HEPES pH8.0, 150 mM NaCl,

2 mM MgCl2, 0.5% BSA (w/v), 10 µg/ml salmon sperm DNA solution (Thermofisher)) for 1 h at room temperature with gentle agitation. During the blocking step, we mixed 12.5 µg labeled fragmented total RNAs in 15 µl H2O, 80 µg labeled in vitro transcribed RNAs in 30 µl H2O, 10 µl folding buffer (200 mM HEPES pH 7.4, 1 M NaCl) and 45 µl H2O, which was heated at 65 °C for 10 min and then cooled down at room temperature for 20 min. After completion of the blocking step, the pooled RNAs were added to 3 ml binding buffer (40 mM HEPES pH 7.4, 150 mM NaCl, 2 mM MgCl2, 0.01% Igepal CA-630, 5% glycerol, 0.2% BSA, 10 µg/ml salmon sperm DNA solution, 200U/ml mRI) and replace the blocking buffer. Microarray slides were incubated in the dark for 1 h with gentle agitation at 25 °C, and washed 3 times with 5 ml Binding buffer for 5 min each. After another 3 times washing with 5 ml washing buffer (40 mM HEPES pH 7.4, 150 mM NaCl, 2 mM MgCl2, 0.02% tween-20) for 5 min each, microarray slides were briefly dipped into a 50 ml conical tube filled with room temperature distilled water three times to remove salt, and immediately spin down in the slide holder or 50 ml conical tube at 200 g for 2 min at room temperature. The dry slide was scanned at 635 nm (Cy5) using a GenePix 4000B Microarray scanner (Molecular Devices) immediately after or at least within 2 h of the completion of the incubation. We used the fold change (FC) and signal-noise-ratio (SNR) to evaluate the RNA-binding activity of the proteins. The FC is defined as: Foreground Signal (F635) / Local Background Signal (B635); SNR is defined as: (F635 -B635)/ B635SD, where B635SD is the standard deviation of B635. The foreground signal at 635 nm, local background signal at 635 nm and SNR were quantified and calculated by the GenePix software from the scanned images.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The RNA-seq and CLIP-seq data have been deposited in the Gene Expression Omnibus under accession code GSE214173 and GSE226214. The mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium under accession code PXD037105, and PXD053382[90]. The protein microarray data have been deposited in the BioStudies[91] database under accession code S-BSST1172. Source data are provided with this paper.

## References


1. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).
2. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* **19**, 327–341 (2018).
3. Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.* **39**, D301–D308 (2011).
4. Castello, A. et al. Comprehensive identification of RNA-binding domains in human cells. *Mol. Cell* **63**, 696–710 (2016).
5. Gebauer, F., Schwarzl, T., Valcarcel, J. & Hentze, M. W. RNA-binding proteins in human genetic disease. *Nat. Rev. Genet.* **22**, 185–198 (2021).
6. Castello, A., Hentze, M. W. & Preiss, T. Metabolic enzymes enjoying new partnerships as RNA-binding proteins. *Trends Endocrinol. Metab. TEM* **26**, 746–757 (2015).
7. Castello, A. et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).
8. Bao, X. et al. Capturing the interactome of newly transcribed RNA. *Nat. Methods* **15**, 213–220 (2018).
9. Huang, R., Han, M., Meng, L. & Chen, X. Transcriptome-wide discovery of coding and noncoding RNA-binding proteins. *Proc. Natl Acad. Sci. USA* **115**, E3879–E3887 (2018).
10. Trendel, J. et al. The human RNA-binding proteome and its dynamics during translational arrest. *Cell* **176**, 391–403 (2018).
11. Queiroz, R. M. L. et al. Comprehensive identification of RNA-protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat. Biotechnol.* **37**, 169–178 (2019).
12. Fecko, C. J. et al. Comparison of femtosecond laser and continuous wave UV sources for protein-nucleic acid crosslinking. *Photochem. Photobiol.* **83**, 1394–1404 (2007).
13. Darnell, R. B. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip. Rev. RNA* **1**, 266–286 (2010).
14. Shetlar, M. D., Carbone, J., Steady, E. & Hom, K. Photochemical addition of amino acids and peptides to polyuridylic acid. *Photochem. Photobiol.* **39**, 141–144 (1984).
15. Paradiso, P. R., Nakashima, Y. & Konigsberg, W. Photochemical cross-linking of protein. nucleic acid complexes. The attachment of the fd gene 5 protein to fd DNA. *J. Biol. Chem.* **254**, 4739–4744 (1979).
16. Burley, S. K. et al. RCSB protein data bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **49**, D437–D451 (2021).
17. Wadsworth, R. I. & White, M. F. Identification and properties of the crenarchaeal single-stranded DNA binding protein from Sulfolobus solfataricus. *Nucleic Acids Res.* **29**, 914–920 (2001).
18. Kerr, I. D. et al. Insights into ssDNA recognition by the OB fold from a structural and thermodynamic study of Sulfolobus SSB protein. *EMBO J.* **22**, 2561–2570 (2003).
19. Morten, M. J. et al. High-affinity RNA binding by a hyperthermophilic single-stranded DNA-binding protein. *Extremophiles* **21**, 369–379 (2017).
20. Martin, S. L., Li, J. & Weisz, J. A. Deletion analysis defines distinct functional domains for protein-protein and nucleic acid interactions in the ORF1 protein of mouse LINE-1. *J. Mol. Biol.* **304**, 11–20 (2000).
21. Kolosha, V. O. & Martin, S. L. High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). *J. Biol. Chem.* **278**, 8112–8117 (2003).
22. Helder, S., Blythe, A. J., Bond, C. S. & Mackay, J. P. Determinants of affinity and specificity in RNA-binding proteins. *Curr. Opin. Struct. Biol.* **38**, 83–91 (2016).
23. Khazina, E. et al. Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nat. Struct. Mol. Biol.* **18**, 1006–1014 (2011).
24. Greenberg, J. R. Ultraviolet light-induced crosslinking of mRNA to proteins. *Nucleic Acids Res.* **6**, 715–732 (1979).
25. Liu, B., Poolman, B. & Boersma, A. J. Ionic strength sensing in living cells. *ACS Chem. Biol.* **12**, 2510–2514 (2017).
26. Maguire, M. E. & Cowan, J. A. Magnesium chemistry and biochemistry. *Biometals* **15**, 203–210 (2002).
27. Tsai, B. P., Wang, X., Huang, L. & Waterman, M. L. Quantitative profiling of in vivo-assembled RNA-protein complexes using a novel integrated proteomic approach. *Mol. Cell Proteom.* **10**, M110 007385 (2011).
28. Hacisuleyman, E. et al. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat. Struct. Mol. Biol.* **21**, 198–206 (2014).
29. Leppek, K. & Stoecklin, G. An optimized streptavidin-binding RNA aptamer for purification of ribonucleoprotein complexes identifies novel ARE-binding proteins. *Nucleic Acids Res.* **42**, e13 (2014).
30. Matia-Gonzalez, A. M., Iadevaia, V. & Gerber, A. P. A versatile tandem RNA isolation procedure to capture in vivo formed mRNA-protein complexes. *Methods* **118-119**, 93–100 (2017).
31. Garcia-Moreno, M. et al. System-wide profiling of RNA-binding proteins uncovers key regulators of virus infection. *Mol. Cell* **74**, 196–211 e111 (2019).

32. Mullari, M., Lyon, D., Jensen, L. J. & Nielsen, M. L. Specifying RNA-binding regions in proteins by peptide cross-linking and affinity purification. *J. Proteome Res.* **16**, 2762–2772 (2017).
33. Panhale, A. et al. CAPRI enables comparison of evolutionarily conserved RNA interacting regions. *Nat. Commun.* **10**, 2682 (2019).
34. Welsh, S. A. & Gardini, A. Genomic regulation of transcription and RNA processing by the multitasking Integrator complex. *Nat. Rev. Mol. Cell Biol.* **24**, 204–220 (2022).
35. Houseley, J., LaCava, J. & Tollervey, D. RNA-quality control by the exosome. *Nat. Rev. Mol. Cell Biol.* **7**, 529–539 (2006).
36. Black, C. S. et al. Spliceosome assembly and regulation: insights from analysis of highly reduced spliceosomes. *RNA* **29**, 531–550 (2023).
37. Henninger, J. E. et al. RNA-mediated feedback control of transcriptional condensates. *Cell* **184**, 207–225.e224 (2021).
38. Lai, F. et al. Activating RNAs associate with mediator to enhance chromatin architecture and transcription. *Nature* **494**, 497–501 (2013).
39. Li, W. et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516–520 (2013).
40. Kulichkova, V. A. et al. 26S proteasome exhibits endoribonuclease activity controlled by extra-cellular stimuli. *Cell Cycle* **9**, 840–849 (2010).
41. Jarrousse, A. S., Petit, F., Kreutzer-Schmid, C., Gaedigk, R. & Schmid, H. P. Possible involvement of proteasomes (prosomes) in AUUUA-mediated mRNA decay. *J. Biol. Chem.* **274**, 5925–5930 (1999).
42. Thul, P. J. et al. A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).
43. Venkataraman, A. et al. A toolbox of immunoprecipitation-grade monoclonal antibodies to human transcription factors. *Nat. Methods* **15**, 330–338 (2018).
44. Hu, S. et al. Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. *Cell* **139**, 610–622 (2009).
45. Xue, Z. et al. A G-rich motif in the lncRNA braveheart interacts with a zinc-finger transcription factor to specify the cardiovascular lineage. *Mol. Cell* **64**, 37–50 (2016).
46. Kretz, M. et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493**, 231–235 (2013).
47. Ramanathan, M., Porter, D. F. & Khavari, P. A. Methods to study RNA–protein interactions. *Nat. Methods* **16**, 225–234 (2019).
48. Siprashvili, Z. et al. The noncoding RNAs SNORD50A and SNORD50B bind K-Ras and are recurrently deleted in human cancer. *Nat Genet* **48**, 53–58 (2016).
49. Moore, M. J. et al. Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat. Protocols* **9**, 263–293 (2014).
50. Ramos, A. et al. Role of dimerization in KH/RNA complexes: the example of Nova KH3. *Biochemistry* **41**, 4193–4201 (2002).
51. Varani, L. et al. The NMR structure of the 38 kDa U1A protein - PIE RNA complex reveals the basis of cooperativity in regulation of polyadenylation by human U1A protein. *Nat. Struct. Biol.* **7**, 329–335 (2000).
52. Mudunuri, U., Che, A., Yi, M. & Stephens, R. M. bioDBnet: the biological database network. *Bioinformatics* **25**, 555–556 (2009).
53. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
54. Necci, M., Piovesan, D., Predictors, C., DisProt, C. & Tosatto, S. C. E. Critical assessment of protein intrinsic disorder prediction. *Nat. Methods* **18**, 472–481 (2021).
55. Peng, Z. & Kurgan, L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.* **43**, e121 (2015).
56. Xia, Y., Xia, C. Q., Pan, X. & Shen, H. B. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic Acids Res.* **49**, e51 (2021).
57. Hu, G. et al. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat. Commun.* **12**, 4438 (2021).
58. Varadi, M. et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
59. Jarvelin, A. I., Noerenberg, M., Davis, I. & Castello, A. The new (dis) order in RNA regulation. *Cell Commun. Signal. CCS* **14**, 9 (2016).
60. Thandapani, P., O'Connor, T. R., Bailey, T. L. & Richard, S. Defining the RGG/RG motif. *Mol. Cell* **50**, 613–623 (2013).
61. Popow, J. et al. FASTKD2 is an RNA-binding protein required for mitochondrial RNA processing and translation. *RNA* **21**, 1873–1884 (2015).
62. Nguyen, V. T., Kiss, T., Michels, A. A. & Bensaude, O. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature* **414**, 322–325 (2001).
63. Knighton, D. R. et al. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 407–414 (1991).
64. Matthews, J. M., Lester, K., Joseph, S. & Curtis, D. J. LIM-domain-only proteins in cancer. *Nature reviews. Cancer* **13**, 111–122 (2013).
65. Yasuoka, Y. & Taira, M. LIM homeodomain proteins and associated partners: then and now. *Curr. Topics Dev. Biol.* **145**, 113–166 (2021).
66. Ma, L., Greenwood, J. A. & Schachner, M. CRP1, a protein localized in filopodia of growth cones, is involved in dendritic growth. *J. Neurosci.* **31**, 16781–16791 (2011).
67. Järvinen, P. M. et al. Cysteine-rich protein 1 is regulated by transforming growth factor-β1 and expressed in lung fibrosis. *J. Cell. Physiol.* **227**, 2605–2612 (2012).
68. Chang, D. F. et al. Cysteine-rich LIM-only proteins CRP1 and CRP2 are potent smooth muscle differentiation cofactors. *Dev. Cell* **4**, 107–118 (2003).
69. Tran, T. C., Singleton, C., Fraley, T. S. & Greenwood, J. A. Cysteine-rich protein 1 (CRP1) regulates actin filament bundling. *BMC Cell Biol.* **6**, 45 (2005).
70. Simunovic, M. & Brivanlou, A. H. Embryoids, organoids and gastruloids: new approaches to understanding embryogenesis. *Development* **144**, 976–985 (2017).
71. Van Nostrand, E. L. et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514 (2016).
72. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
73. Young, R. A. Control of the embryonic stem cell state. *Cell* **144**, 940–954 (2011).
74. Kopylova, E., Noe, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
75. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
76. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
77. Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
78. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
79. Ramirez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).

80. Palomba, A. et al. Comparative evaluation of MaxQuant and proteome discoverer MS1-based protein quantification tools. *J. Proteome Res.* **20**, 3497–3507 (2021).
81. Liu, M. & Dongre, A. Proper imputation of missing values in proteomics datasets for differential expression analysis. *Brief. Bioinform.* **22**, bbaa112 (2021).
82. Lu, T. et al. Tissue-characteristic expression of mouse proteome. *Mol. Cell. Proteom.* **21**, 100408 (2022).
83. Schrodinger, L. L. C. *The PyMOL Molecular Graphics System*, Version 1.8 https://www.sciepub.com/reference/159710 (2015).
84. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* **25**, 25–29 (2000).
85. Gene Ontology, C. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325-D334 (2021).
86. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
87. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
88. Jin, H. et al. ChIPseqSpikeInFree: a ChIP-seq normalization approach to reveal global changes in histone modifications without spike-in. *Bioinformatics* **36**, 1270–1272 (2020).
89. Siprashvili, Z. et al. Identification of proteins binding coding and non-coding human RNAs using protein microarrays. *BMC Genom.* **13**, 633 (2012).
90. Perez-Riverol, Y. et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552 (2022).
91. Sarkans, U. et al. The BioStudies database-one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.* **46**, D1266–D1270 (2018).

## Acknowledgements

Z.X. is supported by the National Key Research and Development Program of China, Stem Cell and Translational Research (2018YFA0109200) and National Natural Science Foundation of China (General Program No. 31970600). L.Y. is supported by the Postdoctoral Science Foundation of China (2018M643467). Y. L. is supported by National Institutes of Health (R35 GM118118) and the Welch Foundation (I-1560). Y.H. is supported by the National Natural Science Foundation of China (General Program No. 82070324). X.M. is supported by the National Natural Science Foundation of China (T2322002), and Beijing Nova Program (Z211100002121040). We thank Fan Lai (Yunnan University, China) and Matthias W. Hentze (European Molecular Biology Laboratory) for insightful discussions and for critical evaluation of the manuscript. We thank Xiang Wang, Dong Deng, Zhenhua Shao, Haohao Dong, and Shiqian Qi, all of Sichuan University, for technical support. We thank PTM Biolabs Inc. for technical supports in LC-MS/MS and Wayen Biotech. for technical supports in protein microarray.

## Author contributions

Y.R.: study design, experimental work, data interpretation, and writing of the manuscript. H.Liao: study design, experimental work, data interpretation, and writing of the manuscript. J.Y.: deep learning modeling, bioinformatic evaluation, data interpretation, and manuscript review. H.Lu: study design, experimental work, data interpretation. X.M.: bioinformatic evaluation, data interpretation, statistical review, and manuscript review. C.W.: manuscript review. Y.F.L.: manuscript review. Y.L.: data review, manuscript review. C.C.: data review, manuscript review. L.C.: data review, manuscript review. X.W.: manuscript review. K.Y.Z.: manuscript review. H.M.L.: manuscript review. Y.L.: data interpretation and writing and review of the manuscript. Y.M.H.: study design, data interpretation, manuscript review. L.Y.: study design, experimental work, data interpretation, writing and review of the manuscript. Z.X.: study design, data interpretation, bioinformatic evaluation, writing and review of the manuscript, supervision of the work.

## Competing interests

Z.X., L.Y., Y.R., and H.L. are inventors on a patent (ZL202210541444.0) covering the HARD protein and its application. This patent has been authorized by the National Intellectual Property Administration, PRC. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52765-w.

**Correspondence** and requests for materials should be addressed to Yi-Min Hua, Lin Yu or Zhihong Xue.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.