RBPWorld for exploring functions and disease associations of RNA-binding proteins across species

Jian-You Liao [1,2,*,†], **Bing Yang** [1,†], **Chuan-Ping Shi**[1,†], **Wei-Xi Deng**[1], **Jin-Si Deng**[1],
Mei-Feng Cen[1], **Bing-Qi Zheng**[1], **Zi-Ling Zhan**[1], **Qiao-Ling Liang**[1], **Ji-En Wang**[1], **Shuang Tao**[1],
Daning Lu[1], **Maojin Liang**[3,*], **Yu-Chan Zhang**[4,*] and **Dong Yin** [1,*]

[1]Department of Medical Research Center, Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Guangdong-Hong Kong Joint Laboratory for RNA Medicine, Sun Yat-Sen Memorial Hospital, Sun Yat-sen University, 107 Yan Jiang West Road, Guangzhou, Guangdong 510120, China
[2]Department of Precision Medicine Center, Shenshan Central Hospital, Sun Yat-Sen Memorial Hospital, Sun Yat-sen University, 1 Heng Er Road, Dongyong Town, Shanwei, Guangdong 516621, China
[3]Department of Otolaryngology, Sun Yat-Sen Memorial Hospital, Institute of Hearing and Speech-Language Sciences, Sun Yat-sen University, 107 Yan Jiang West Road, Guangzhou, Guangdong 510120, China
[4]Department of Life Science, Guangdong Provincial Key Laboratory of Plant Resources, State Key Laboratory for Biocontrol, School of Life Science, Sun Yat-Sen University, No.135 Xingang Xi Lu, Haizhu District, Guangzhou, Guangdong 510275, China
*To whom correspondence should be addressed. Tel: +86 18 922 182 515; Email: yind3@mail.sysu.edu.cn
Correspondence may also be addressed to Yu-Chan Zhang. Tel: +86 13760816112; Email: zhyuchan@mail.sysu.edu.cn
Correspondence may also be addressed to Maojin Liang. Tel: +86 13760816112; Email: liangmj3@mail.sysu.edu.cn
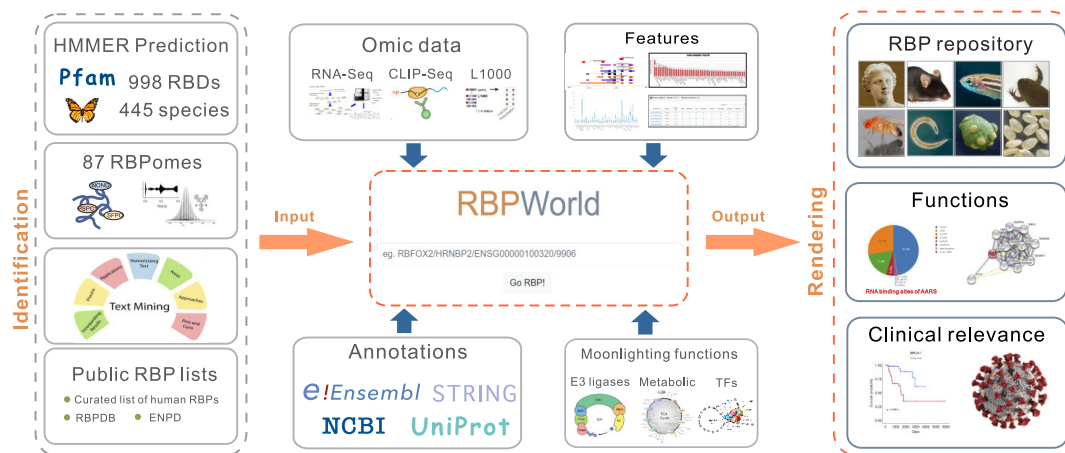Correspondence may also be addressed to Jian-You Liao. Tel: +86 1358054805; Email: liaojy3@mail.sysu.edu.cn
†The first three authors should be regarded as Joint First Authors.

Abstract

RNA-binding proteins (RBPs) play key roles in a wide range of physiological and pathological processes. To facilitate the investigation of RBP functions and disease associations, we updated the EuRBPDB and renamed it as RBPWorld (http://research.gzsys.org.cn/rbpworld/#/home). Leveraging 998 RNA-binding domains (RBDs) and 87 RNA-binding Proteome (RBPome) datasets, we successfully identified 1 393 413 RBPs from 445 species, including 3030 human RBPs (hRBPs). RBPWorld includes primary RNA targets of diverse hRBPs, as well as potential downstream regulatory pathways and alternative splicing patterns governed by various hRBPs. These insights were derived from analyses of 1515 crosslinking immunoprecipitation-seq datasets and 616 RNA-seq datasets from cells with hRBP gene knockdown or knockout. Furthermore, we systematically identified 929 RBPs with multi-functions, including acting as metabolic enzymes and transcription factors. RBPWorld includes 838 disease-associated hRBPs and 970 hRBPs that interact with 12 disease-causing RNA viruses. This provision allows users to explore the regulatory roles of hRBPs within the context of diseases. Finally, we developed an intuitive interface for RBPWorld, facilitating users easily access all the included data. We believe that RBPWorld will be a valuable resource in advancing our understanding of the biological roles of RBPs across different species.

Graphical abstract

Introduction

RNA-binding proteins (RBPs) are conserved and essential regulators of various cellular processes (1,2). They play a crucial role in the regulation of RNA metabolism and collaborate with RNAs to regulate specific cellular processes, such as signaling transduction (3), transcription (4) and protein synthesis (5). Despite the important role of RBPs in diverse pathological and physiological processes, the functions of most RBPs remain largely unclear. The rapid accumulation of genome-wide transcriptomic and proteomic datasets in public databases offers an unprecedented opportunity to enhance our understanding of the functional roles of RBPs in a wide range of cellular processes (6).

Several databases have been developed to help the investigation of RNA binding sites or motifs in RBPs. These includes Gene Ontology (GO) database (7), RNA-Binding Protein DataBase (RBPDB) (8), ATtRACT (9), StarBase (10) and POSTAR3 (11), all of which contribute significantly to the functional investigation of RBPs. In our previous studies, we constructed EuRBPDB (12), which significantly expands the repertoire of RBPs by including 315 222 RBPs identified from 162 species. EuRBPDB also incorporates 837 crosslinking immunoprecipitation (CLIP) sequence (CLIP-seq) datasets (13,14), allowing users to analyze the RNA targets of RBPs. Despite the advancements made by our database and others, there is still a significant lack of key information required for exploring RBP functions. For example, although many RBPs have been reported, the RNA targets for the majority of them are still missing within these databases. Additionally, RNA sequencing (RNA-seq) data from cells with knockdown or knockout (KD/KO) of RBP genes, referred to as bulkPerturb-seq, which are rapidly accumulating in public databases like The Gene Expression Omnibus (GEO) (15), remain to be fully integrated and utilized for comprehensive functional analysis of RBPs.

RBPs have emerged as key regulators in human diseases and hold promise as novel therapeutic targets (16). Mutations in RBPs can impact their expression, disrupt their interactions with cofactors and RNA targets (17,18), or interfere with other functions beyond RNA-binding (19,20), all of which can lead to severe human diseases. To date, three databases have been constructed to investigate the connections between RBPs and limited number of diseases. EuRBPDB and RBPDB (8) offer features to explore the association between RBPs and various human cancers. READDB (21) permits exploration of associations between a subset of RBPs and a limited number of human diseases. The absence of a comprehensive database for disease associations of RBPs hinders the study of RBP–disease associations.

To advance the functional and disease association studies of RBPs across various species, we have systematically updated our previously constructed RBP database, EuRBPDB. In this study, we first refined our RBP identification method, which result in the identification of 1 393 413 high-reliability RBPs from 445 species, including 3030 human RBPs (hRBPs). Next, we conducted a systematic analysis to identify the primary RNA targets of 1616 hRBPs, potential downstream pathways regulated by RBPs and the multi-functional properties of 929 hRBPs. Furthermore, we systematically identified 838 disease-associated RBPs and 1111 potential drugs targeting to 234 hRBPs. To enhance the user experience, we have redesigned the user interface (UI) of the RBP database to offer a more intuitive, user-friendly and mobile-compatible experience. We

have named this new version of the RBP database 'RBPWorld' to reflect its aim of including the comprehensive annotations of RBPs from a wide range of species.

By utilizing RBPWorld, users can interactively explore comprehensive RBP catalogs from 445 species and investigate the function and disease associations of RBPs across species. This platform will greatly facilitate the investigation and understanding of the biological role of RBPs across different species.

Materials and methods

RBP identification and annotation

The following key words were used to performed search in Pfam database (22) for RNA-binding domains (RBDs) and in the: 'RNA binding', 'RNA-binding', 'bind to.*RNA', 'bind to RNA', 'bind to nucleic acids', 'bind to DNA or RNA', 'bind to double-stranded RNA', 'bind to G-quartet sequences in.* RNA', 'bind to double-stranded RNA', 'bind to G-quartet sequences in.* RNA', 'bind to.* in.* RNA', 'bind to ribosomal RNA', 'bind *RNA', 'RNA processing' and 'bind.* RNA'. The HMM profiles of RBDs were extracted from Pfam database (release 35). All protein sequences of 445 species were downloaded from Ensembl database (release 110 for animal and release 57 for plant) (23).

The identification of RBPs in RBPWorld is based on the following evidence: (i) Using the HMMER package (v3.2.1) (24), we conducted a comprehensive search across all eukaryotic protein sequences against the RBD HMM profiles obtained from Pfam and EuRBPDB database. Proteins with *E*-value <0.0001 were considered as bona fide RBPs harboring RBD. (ii) We conducted a comprehensive search of gene annotations in the Gene Ontology (GO) database (7) using the aforementioned keywords related to RBDs. This approach allowed us to systematically identify genes annotated with functions indicative of RNA-binding activities, thereby aiding in the detection of potential RBPs. (iii) RBPome datasets were manually collected from published works (25–29). Then RBPs identified by at least seven distinct RBPome technologies were considered to be RBPs of high confidence. (iv) To further refine our list of RBPs, we integrated evidence from multiple sources: evidence for the identification of other RBPs was sourced from CLIP-seq datasets, annotations by Gerstberger *et al*. (1), GO (7) and the RBPDB (8) database.

The basic information, orthologs, paralogs, GO and phenotype annotations of RBPs were obtained from NCBI (30), GeneCards (31) and Ensembl databases (23). The protein–protein interaction information was parsed from STRING database (32). The pathway annotation was obtained from KEGG database (33). Expression data were obtained from GTEx (34).

RBP families

In RBPWorld, we employed two approaches to define RBP families. The first approach involved grouping RBPs based on shared RBDs (35). RBPs sharing the same family RBD (as defined by the Pfam database) based on sequence similarity were grouped into a family and named after the RBD. When a RBP harbored multiple types of RBDs, it was organized into distinct families based on those RBD types. RBPs lacking known RBDs were designated as members of the non-canonical RBP family. The second approach used paralogues analysis based on protein sequence, where RBP with iden-

tity ≥20 was clustered into a paralogous family ([1]). To cluster redundant RBPs, we first sort the families in descending order based on the number of members. Subsequently, we iteratively examine each RBP within these families. If an RBP is found in a subsequent family, it is removed from that family to ensure uniqueness. This dual strategy ensures a comprehensive and robust classification of RBP families, facilitating more accurate functional predictions and comparative analyses.

Annotation of multi-functional RBPs and condensate-associated RBPs

RBPWorld defines six categories of multi-functional RBPs (mfRBPs). These categories are annotated using resources as following: (i) DNA and RBP were retrieved from three sources: (a) AnimalTFDB4 ([35]), (b) Xiao *et al.* ([36]), (c) Van Nostrand *et al.* ([37]); (ii) RNA-binding E3 ligases were retrieved from IUUCD2.0 ([38]) and UbiBrowser2.0 ([39]) database; (iii) RNA-binding deubiquitinating enzymes (DUBs) were integrated from DUBase database; (iv) RNA-binding kinases annotations are retrieved from UniProt database ([40]); (v) RNA-binding transmembrane proteins prediction was performed using TMHMM-2.0 ([41]); (vi) Metabolic enzymes RBPs were downloaded from Reactome database ([42]). RBPWorld also systematically identified hRBPs localized to membrane-less condensates by Human Protein Atlas ([43]) and The RNA granule databases ([44]) into five categories: (i) nucleoli, (ii) nucleoli rim, (iii) fibrillar center, (iv) p-bodies and (v) stress granules (SGs).

Identification of primary RNAs targets for RBPs

CLIP-seq datasets were retrieved from ENCODE database ([45]) and StarBase ([10]). Peak and bam files of each dataset were downloaded. We used intersectBed of bedtools package (v2.27.1) ([46]) to annotate each peak. In annotating primary RNA targets using CLIP-seq data, RNA types with a read fraction <10% of the total read were excluded. If the read fraction of an RNA type exceeded 70%, we considered that the RBP specifically binds to this RNA type. If all RNA types bound by an RBP have peak fraction between 10% and 70%, these RNA types all considered as primary targets of the RBP. The primary RNA targets of RBPs were also obtained from the annotations performed by Gerstberger *et al.* ([1]), as well as the annotation information from the GO database. We integrated all primary RNA target information for each RBP. If an RBP possesses three or more primary RNA targets, we categorize as diverse.

Disease-associated RBPs and RBP-targeting drugs

Evidence of association between RBPs and diseases, classified into nine major categories, are sourced from nine distinct databases: (i) Open Targets platform ([47]), (ii) Clinvar ([48]), (iii) Gene Burden ([49]), (iv) Genomics england pane ([50]), (v) Gene2Phenotype ([51]), (vi) UniProt: UniProt literature ([40]), (vii) UniProt: Uniprot variants ([40]), (viii) Clingen ([52]) and (ix) Orphanet. RBPs bearing any Mendelian and somatic mutation evidence were classified as disease-associated RBPs. Drugs that affect to RBPs were retrieved from ChEMBL database ([53]). Identical small molecule drugs or inhibitors are consolidated into a single record, displaying the highest clinical trial phase achieved.

Analysis of RNA-seq datasets

BulkPerturb-seq datasets were acquired from the Sequence Read Archive (SRA) database ([54]) and ENCODE database ([37]). The curated RNA-seq fastq files were aligned to the human genome (GRCh38) using STAR (v2.6.1d) ([55]), and gene expression levels for each dataset was quantified using RSEM ([56]). Standardized the quantification results, using TPM exclusively for data presentation and analysis. Differential gene expression analysis was performed using R DESeq2 package with default parameters ([57]), while differential alternative splicing (DAS) were investigated using rMATS ([58]) with default parameters. The visualization of AS events was performed using Sashimi plots software ([59]).

Results

Refined RBP identification method identified 1 393 413 RBPs across species

In EuRBPDB study, RBPs were identified through a search based on RBDs, along with the utilization of RNA-binding Proteome (RBPome) datasets and manual curation from literature sources. To enhance our RBP identification procedure, we made the following improvements: ([i]) updated the RBD catalog by systematically interrogating the latest version of the Pfam database. As a result, our RBD catalog expands from 791 to 998. ([ii]) Incorporated 87 additional RBPome datasets to identify RBPs that might have been missed by domain annotation, as certain RBPs lack a Pfam-defined RBD. ([iii]) Included nine RNA interactome datasets obtained through probe-sets based RBP identification technologies like Comprehensive identification of RBPs by mass spectrometry (ChIRP-MS) ([60]) or RNA antisense purification coupled with mass spectrometry (RAP-MS) ([61]), and ([iv]) included RBPs from other RBP database, including GO ([7]) and RBPDB ([12]).

As a result of our refined RBP identification procedures, we have identified 3030 RBPs in humans and 1 393 413 RBPs in 445 species (Figure [1]A and B). Our analysis of the correlation between genomic complexity and the number of RBPs revealed a positive relationship, where species with higher ploidy, more chromosomes, a greater number of protein-coding genes and longer genomic length tending to have higher number of RBPs (Figure [1]C and D). Notably, the correlation between RBP count and the number of protein-coding genes was the strongest in both kingdoms, suggesting that the higher number of protein-coding genes may lead to a higher frequency of RBP generation. Interestingly, RBPs exhibit a significantly higher correlation with genomic length in plants than in animals, consistent with the observation that plant genomes tend to show a stronger correlation with protein-coding gene numbers than those in animals ([62]).

The RBP number in RBPWorld surpasses the count of RBPs included in other RBP database (Figure [1]E). Despite the large number of RBPs, we have identified across species, the majority of them exhibit conservation. Our conservation analysis revealed that the majority of hRBPs are conserved, with orthologs of 60.17% hRBPs found in over 200 animal species (Figure [1]F). Notably, only 32.35% of hRBPs identified through domain-search method were detectable using RBPome approaches, and 522 hRBPs that were detected by at least seven different RBPome technologies did not possess Pfam-defined RBDs (Figure [1]G). These findings underscore
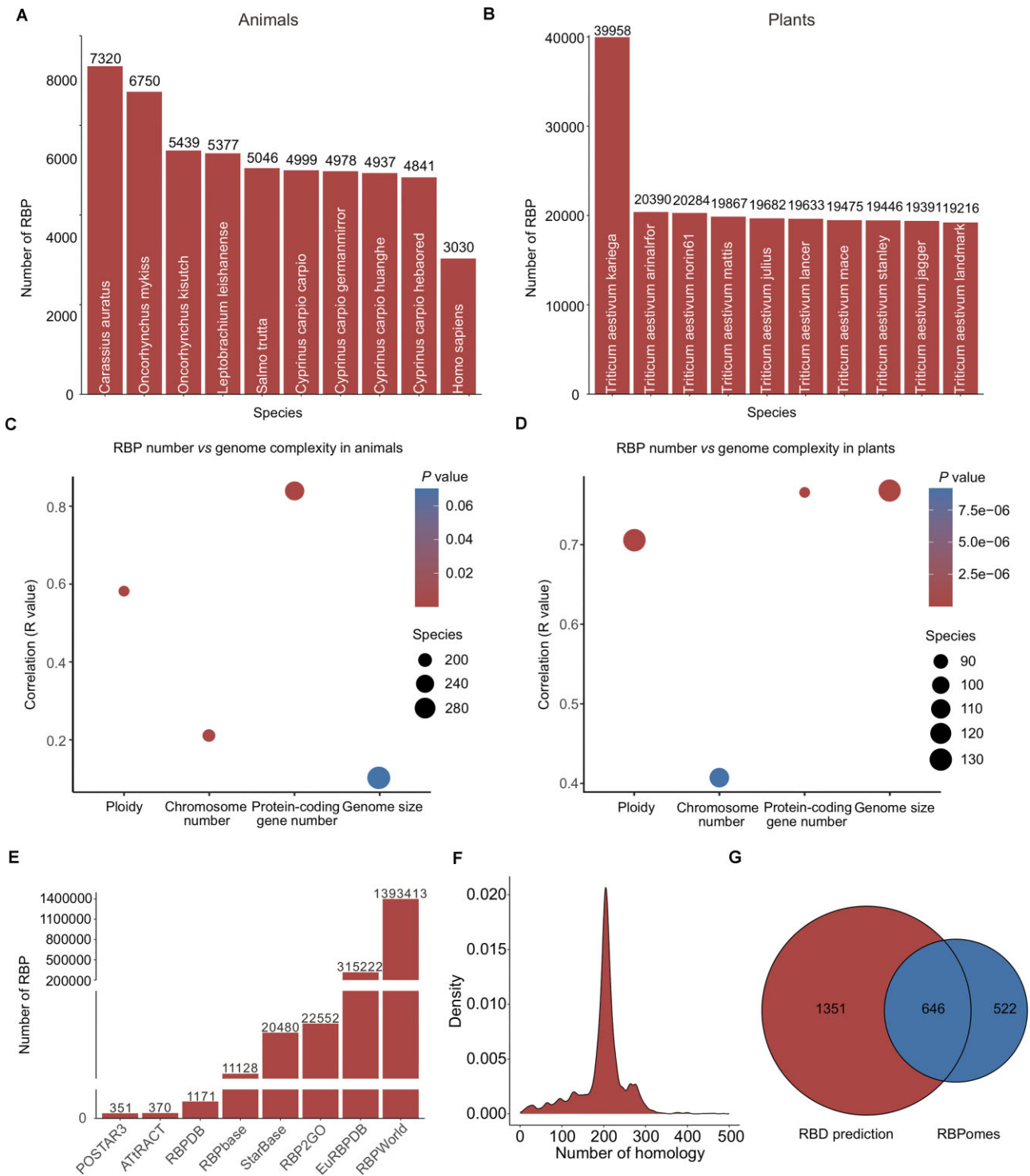
Figure 1. Refined RBP identification procedures enable comprehensive and high-confidence integration of RBP data across 445 species. (**A** and **B**) Bar charts illustrating the number of RBPs for the Top 9 species of animals (A) and Top 10 plants (B) collected in RBPWorld. (**C** and **D**) Bubble plots showing correlation between genomic complexity (chromosome ploidy, number of chromosomes, number of protein-coding gene and genome size) of animal species (C) or plants species (D) and the corresponding number of RBPs. Statistics were calculated as Pearson correlation coefficient. (**E**) Comparison of the number of RBPs collected by RBPWorld with those in other databases. (**F**) Density statistics of the number of hRBP homologous genes in all animals. Horizontal coordinates represent the number of homologous genes, while vertical coordinates indicate the density distribution. The total area under the curve is 1. (**G**) Venn plot showing distribution of RBPs identified through RBPome technology or RBD prediction. RBP, RNA-binding protein; RBPome, RNA-binding proteome.

the necessary of integrating domain-search-based method and RBPome techniques to identify the complete set of RBPs.

RNAs associated with diverse RBPs

To elucidate the functions of RBPs, we focused on identifying the primary RNA targets of various hRBPs. By integrating data from the GO database (7), the study by Gerstberger *et al*. (1) and available CLIP-seq datasets, we identified primary RNA targets for 1616 RBPs. When analyzing the primary RNA target of each RBP using CLIP-seq data, if the peak fraction of an RNA type exceeds 70%, we consider it the sole primary RNA target for that RBP. If all RNA types bound by an RBP have peak fraction between 10% and 70%, these RNA types all considered as primary targets of the RBP. RNA types with a total peak fraction <10% are ignored in our analysis. Diverse RNA is defined as binding to more than two different types of RNA (Supplementary Table S1). The majority of these RBPs are canonical RBPs possessing Pfam-defined RBDs. Subsequently, we categorized RBPs based on their interacting RNAs into different categories, including messenger RNA (mRNA)-binding, ribosomal RNA (rRNA)-binding, transfer RNA (tRNA)-binding, small nuclear RNA (snRNA)-binding and small nucleolar RNA (snoRNA)-binding. Interestingly, among the RBPs with identified primary RNA targets, 1365 were found to bind RNAs involved in protein synthesis, such as mRNAs, tRNAs and rRNAs (Figure 2A).

The RBP family was defined based on RBPs sharing the same RBDs (35). In total, we identified 849 RBP families across all 445 species (Figure 2B and C). Among these, humans host 735 RBP families. Interestingly, plants possess a notably smaller number of RBP families compared to animals (Figure 2B). In particular, human shows the highest number of RBPs with classical RBDs, such as RRM_1, zf_met and MMR_HSR1 (Figure 2C). Out of the identified RBP families in humans, 660 of them have been associated with specific target RNAs. Notably, RBPs in a family tend to interact with same types of primary RNA targets (Figure 2D). To define redundant properties of RBPs, we compared the evolutionary characteristics of RBPs with those of TFs, which constitute the other main group of gene regulatory factors. We used the phylogenetic homology classification as already defined by the Ensembl Compara database (23) and further grouped together paralogues with even closer homology. Most RBP paralogues share 20–70% sequence identity and, by this criterion, are identified to RBP families (Supplementary Table S2). By contrast, the 1704 human TFs, which diverged more recently than the RBPs, formed only 554 protein families by the homology criteria above.

RNA viruses like Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) rely on the host cell's RBPs for replication (63). These RBPs are crucial in regulating virus infections. Therefore, extensive efforts have been made to identify the proteins that interact with RNA viruses using probe-based RNA interactome techniques (64). Consequently, a substantial amount of RNA virus interactome data has been accumulated. In this study, we systematically identified hRBPs that interacting with RNA virus (vRBPs) by analyzing RNA interactome datasets and database annotations. Overall, we identified 970 vRBPs that interact with 12 disease-causing RNA virus, including SARS-CoV-2, Zika and Dengue RNA viruses. These vRBPs comprise 619 canonical RBPs and 351 non-canonical RBPs, deriving from 403 RBP families. No-

tably, they primarily bind to RNAs involved in protein synthesis, such as mRNAs, tRNAs and rRNAs *in vivo* (Figure 2E). In addition, vRBPs are enriched in distinct pathways and complexes specific to different viruses (Figure 2F).

Diverse 'moonlighting' functions of hRBPs

Several hRBPs exhibit additional functions beyond their role in RNA binding. These functions include acting as metabolic enzymes (65), transmembrane proteins (41), DNA- and RNA-binding proteins (DRBPs) (36,66), E3 ligases (E3s) or DUBs (38,39). Recognizing the importance of uncovering the 'moonlighting' functions of hRBPs in understanding their overall functionality, we conducted a systematical identification of these additional roles using annotations from Reactome (42), TMHMM-2.0 (41), AnimalTFDB4 (66), Uniprot (67), DUBase, iUUCD2.0 (38) and UbiBrowser2.0 (39) databases. Our analysis revealed diverse 'moonlighting' functions of hRBPs. Although the primary functions of some hRBPs may not be RNA-binding (68), we have designated their additional roles as 'moonlighting' functions for the sake of clarity in description. We identified 336 RNA-binding metabolic enzymes, 217 RNA-binding E3 ubiquitin ligases, 12 RNA-binding deubiquitinating enzymes, 66 RBP kinases, 300 DRBPs and 147 transmembrane RBPs. These mfRBPs account for the 30.6% of all hRBPs (Figure 3A), with the majority (94.42%) characterized by possessing only a single 'moonlighting' function (Figure 3B). Notably, we identified a total of 929 mfRBPs, some of which possess more than two moonlighting functions. Among these, 151 are targeted by ChEMBL drugs. Interestingly, certain mfRBPs, like DRBPs, tend to possess RBDs, with ∼78.6% of them featuring these domains. On the other hand, specific mfRBPs lean toward lacking known RBDs, for example, 68.3% of RNA-binding kinases did not encompass known RBDs (Figure 3C). Interestingly, these non-canonical RNA-binding kinases are involved in a broad spectrum of regulatory pathways, in contrast to canonical RNA-binding kinases, which predominantly regulate the apoptosis pathway (Supplementary Figure S1). Notably, all types of mfRBPs are involved in the regulation of a wide range of pathways (Figure 3D), rather than being specific to one or a few specific pathways. These results highlight the broad functional importance of mfRBPs in cellular processes.

In addition to exploring 'moonlighting' functions, we also systematically identified hRBPs localized to membrane-less condensates, such as nucleoli (69), P-bodies (70) and SGs (71). This analysis provides valuable insights into the regulatory roles of hRBPs in the formation and function of membrane-less condensates within cells. By incorporating the information from Human Protein Atlas (43) and the RNA granule databases (44), we identified 155 P-body RBPs, 614 SG RBPs, 371 dense fibrillar component (DFC) RBPs, 64 granular component (GC) RBPs and 88 fibrillar center (FC) RBPs. These condensate-associated RBPs (cRBPs) are account for 35.41% of all hRBPs (Figure 3A), with the majority of RBPs (20.60%) present in more than two condensates (Figure 3B). Notably, cRBPs tend to possess known RBDs (Figure 3C). Furthermore, cRBPs are involved in the regulation of a wide range of pathways (Figure 3D), suggesting their involvement in the regulation of diverse cellular processes.

Interestingly, mfRBPs have a preference for binding mRNA, reminiscent of specific mfRBPs like glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (72), heterogeneous
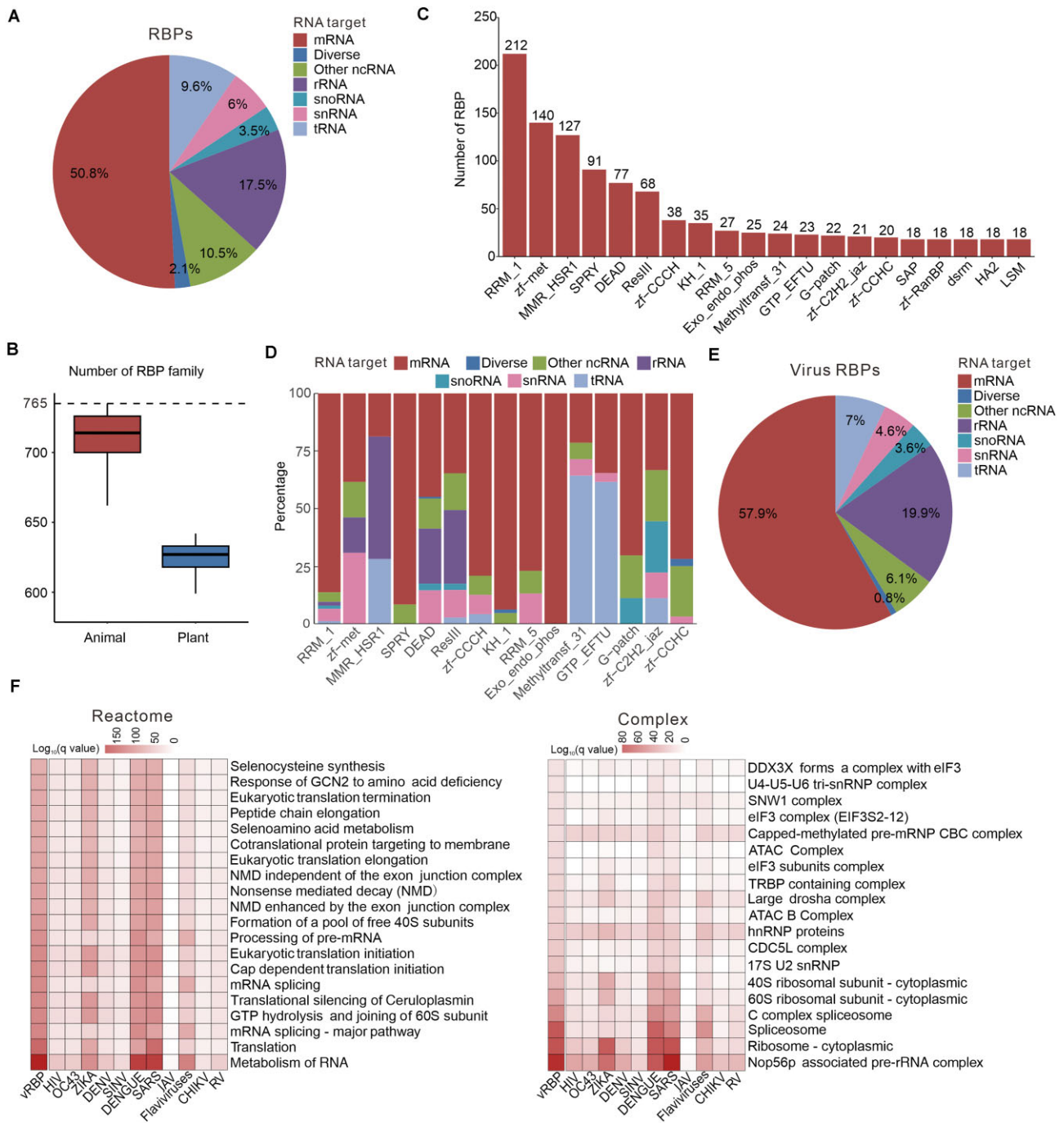
Figure 2. Exploring primary RNA targets of hRBPs and evolutionary conservation of RBP families. (**A**) A pie chart showing the distribution of primary RNA target types bound by RBPs that tend to bind specific RNA targets. (**B**) Statistics of RBP family numbers in different species. (**C**) Bar plot showing the number of RBPs with same RBD (RBP family) from RBPWorld with top 20 members. (**D**) Stacked bar chart showing the proportion of primary RNA targets of the top 15 RBP family members. (**E**) A pie chart showing the distribution of primary RNA target types bound by vRBPs that tend to bind specific RNA targets. (**F**) Heatmaps showing the Reactome pathway enrichment analysis result (left) and complex enrichment analysis result (right) of vRBPs. *q*-value of different pathway/complex were -log₁₀ transformed. RBD, RNA-binding domain; vRBP, hRBPs that interacting with RNA virus.

nuclear ribonucleoprotein L (HnRNP-L) (73) and L1td1 (LINE-1 type transposase domain-containing 1) (74), which have been found to regulate their own abundance by binding to their mRNA and controlling its stability, establishing a negative feedback regulatory loop. In contrast, different cRBPs exhibit distinct RNA binding preferences, potentially hinting at their specific functional roles. For example, DFC, a

membrane-less condensate responsible for rRNA processing (75), tends to associate with RBPs that bind rRNAs and snoRNAs (Figure 3E). These RNAs play crucial roles in the process of rRNA processing (76). Overall, our systematic identification of mfRBPs and cRBPs provides valuable insights into the intricate regulatory functions that hRBPs fulfill in various cellular processes.
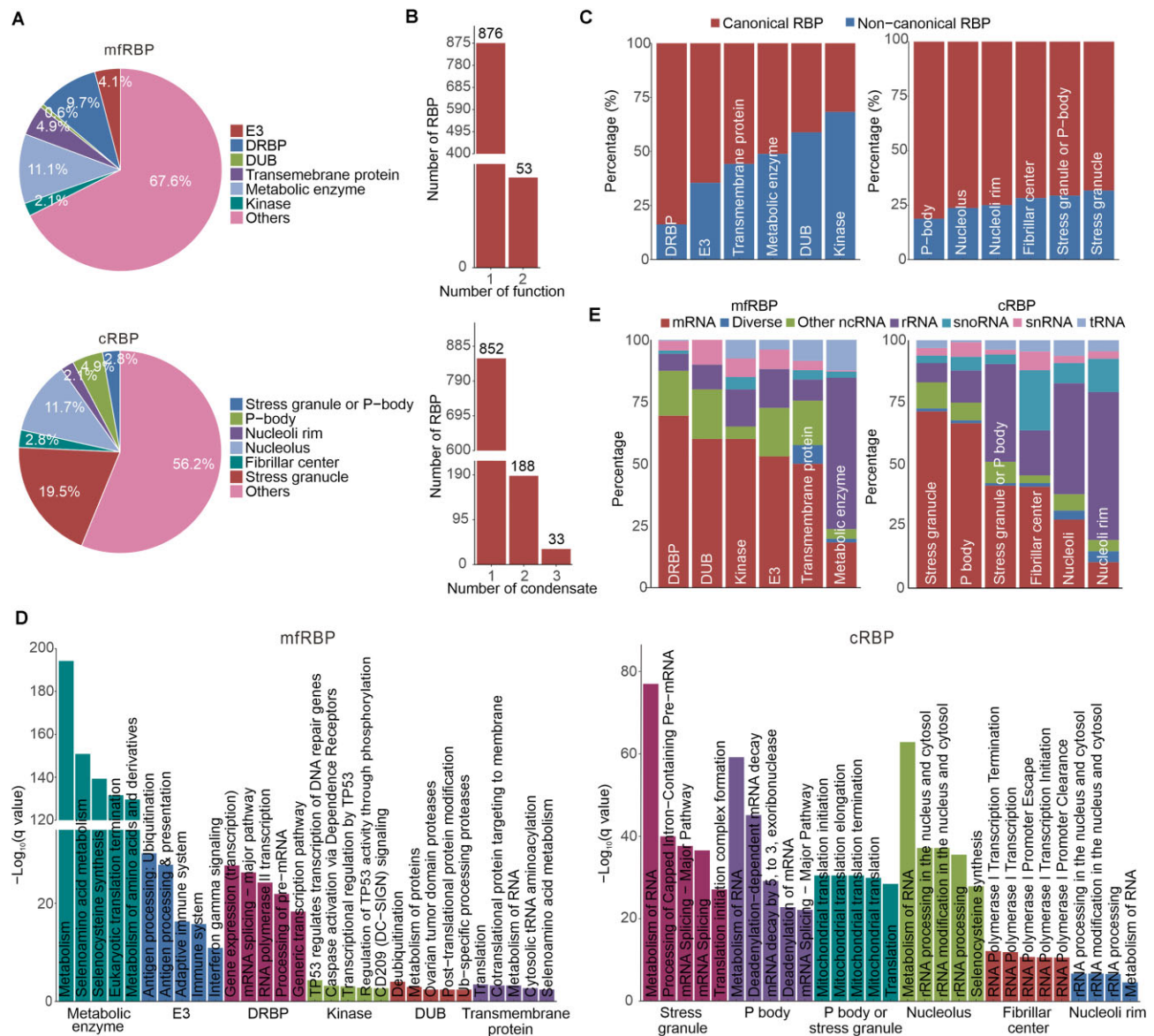
Figure 3. hRBPs perform diverse and multifaceted roles in cellular processes. (**A**) Pie charts showing the distribution of RBPs with multi-functional roles beyond binding RNA (up) and those capable of forming diverse types of condensates (down). (**B**) Bar charts showing the distribution of mfRBPs with only one function alone or two (up), as well as the distribution of cRBPs with varying numbers of condensate sublocations (down). (**C**) Distribution of hRBPs with or without an RBD (canonic or non-canonic) in mfRBPs (left) and cRBPs (right). (**D**) The top five pathways from the Reactome pathway enrichment analysis for different classification of mfRBPs (left) and cRBPs (right). *q*-value of different pathway/complex were -log$_{10}$ transformed. (**E**) The composition of different kinds of mfRBPs and cRBPs binding to specific RNA types. hRBP, human RBP; mfRBP, multi-functional RBP; cRBP, condensate-associated RBP; E3, RNA-binding E3 ligases; DRBP, DNA- and RNA-binding protein; DUB, RNA-binding deubiquitinating enzymes; Transmembrane protein, RNA-binding transmembrane protein; Metabolic enzyme, RNA-binding metabolic enzyme; Kinase, RNA-binding kinase.

Exploring RBP functions using bulkPerturb-seq datasets

To gain insights into the downstream regulatory networks of RBPs, we manually collected 616 bulkPerturb-seq datasets about RBP-coding genes derived from 62 different cell lines. Notably, the perturbation of these RBPs led to an average alteration in the expression of 877 genes. Interestingly, we found that the perturbation of tRNA, snoRNA and mRNA interacting RBPs resulted in the most pronounced changes in the cell transcriptome (Figure 4A). This observation aligns with their important roles in supporting cell growth.

Several RBPs play crucial roles in regulating AS of mR-NAs, giving rise to distinct protein isoforms with diverse func-

tions. To investigate the extent of each RBP's involvement in mRNA AS regulation, we detected DAS events occurring in cells undergoing the perturbation of various RBPs. In total, we identified 434 015 DAS events spanning across 613 datasets. On average, RBP perturbation triggered an occurrence of 708 DAS events. Notably, perturbations of certain RBPs lead to a large number of DAS events, suggesting their potential key role in regulating AS. We categorized those RBPs that induced over 2000 DAS events upon perturbation as hyper-DAS RBPs. Our analysis identified 29 hyper-DAS RBPs, most of them (71.43%) are mRNA binding RBPs, including several well-known AS regulators, such as CPEB1 (77), CELF2 (78) and RBM10 (79) (Figure 4B and C; Supplementary Figure S2).
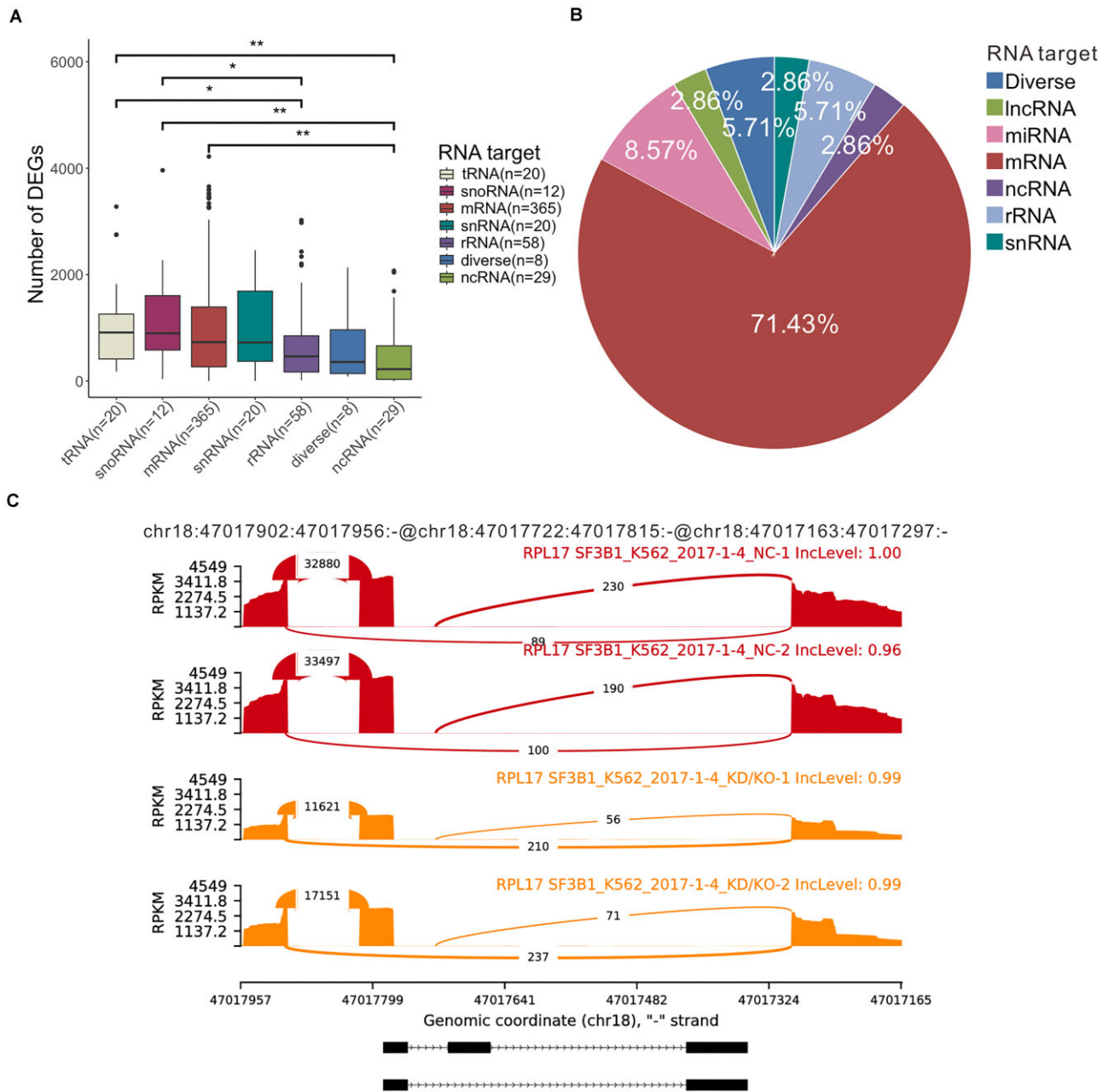
Figure 4. Impact of perturbation of genes encoding RBPs on gene expression profiles and AS events. (**A**) Boxplot exhibiting the number of differentially expressed genes (DEGs) upon 512 perturbed RBPs binding to specific RNA types. (**B**) Pie chart showing the distribution of primary RNA targets for RBPs with over 2000 DAS events, also known as hyper-DAS RBPs, following perturbations of genes encoding for RBPs. (**C**) Sashimi plot illustrating a significantly altered AS event involving exon skipping upon RBP knockdown, with a *P*-value of 0.00201. The IncLevelDifference value is −0.432. *$P <$ 0.05, **$P < 0.01$, ***$P < 0.001$.

Our results highlight the significant value of bulkPerturb-seq datasets. These datasets provide valuable insights into the functional roles and regulatory mechanisms of RBPs.

Mendelian and somatic diseases association of RBPs

To examine the regulatory role of hRBPs in human diseases, we assessed the disease associations of each RBP by utilizing disease association data extracted from nine resources gathering for target-disease evidence. Our analysis identified a total of 838 disease-associated hRBPs that were found to be mutated in various diseases, including 480 canonical

RBPs and 358 non-canonical RBPs (Figure 5A). The number of canonical RBPs and non-canonical RBPs that associated with Mendelian diseases shows a comparable distribution (Supplementary Figure S3A). Notably, non-canonical RBPs exhibits a tendency for association with somatic diseases, wherein the top 10 most relevant somatic diseases exclusively comprise various types of cancer (Figure 5B and C; Supplementary Figure 3B and C). Interestingly, we found that several hRBP subtypes show distinct preferences for either Mendelian or somatic diseases. For instance, tRNA interacting hRBPs displayed a tendency for Mendelian disease, whereas RNA-binding kinases showed a propensity for somatic diseases (Figure 5B). These findings imply that diverse hRBP
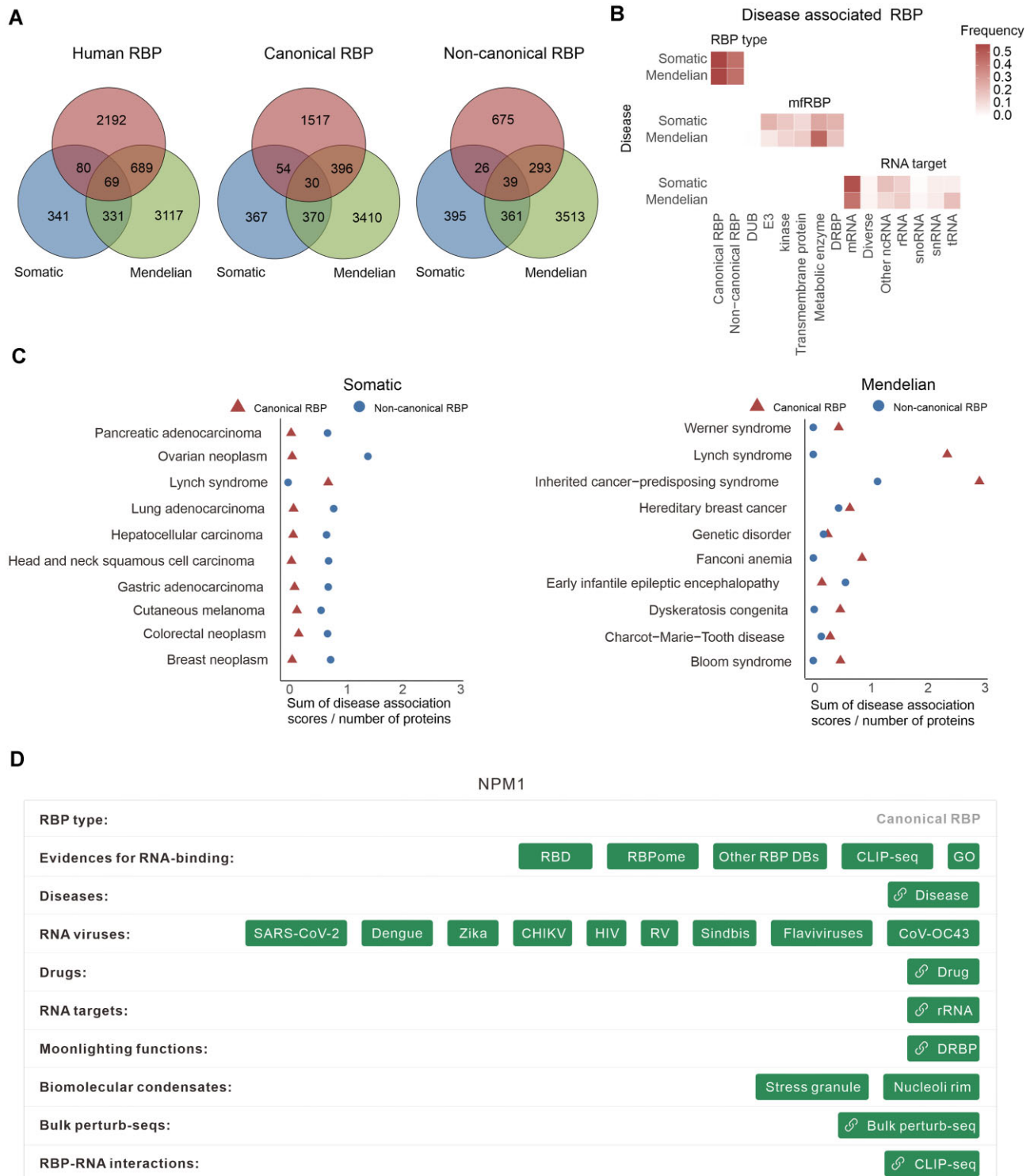
Figure 5. hRBPs involved in Mendelian and somatic genetic diseases. (**A**) The overlap of genes encoding canonical RBPs (left), non-canonical RBPs (middle) and hRBPs (right) with somatic mutant genes causing disease and genes involved in Mendelian genetic disorders. Disease-associated hRBPs were annotated with Mendelian and somatic disease associations extracted from the Open Targets platform. (**B**) Heatmap depicting the compositional differences between sets of RBPs associated with Mendelian mutation diseases and those linked to somatic mutation disorders, highlighting differences in RBP types, moonlighting functions and primary RNA targets. (**C**) Therapeutic areas of disease-associated RBPs. Disease mutations with an association score >0.2 from Open Targets were summed for selected therapeutic areas for RBPs. The accumulated association scores were normalized for the amounts of protein in each category. The top 10 therapeutic areas are shown for each disease category, highlighting the highest association scores for Mendelian and somatic mutation diseases. Additional information on methods is available at http://www.hentze.embl.de/public/hRBPdiseases. (**D**) 'Summary' section screenshot of NPM1 gene from RBPWorld.

subtypes may play distinct role in the development of Mendelian and somatic cell genetic diseases.

The RBPWorld platform serves as a valuable tool for the functions and disease association study of RBPs

To enable users to effectively utilize the deep exploration on RBP categories across species generated in this study, as well as the investigation of conservation, functions, disease associations and targeting drugs of RBPs, we have systematically updated the original RBP database EURBPDB. The database has been renamed to RBPWorld (http://research.gzsys.org.cn/rbpworld/), a title that better reflects its scope and utility. With RBPWorld, users can conveniently access comprehensive information about specific RBPs by inputting the gene symbol/ID in the top right corner of all pages. This input instantly retrieves all relevant information on a single RBP information page. The page includes detailed annotations, conservative properties, functions, RBP targeting drugs and disease associations. Additionally, RBPWorld offers convenient browsing and downloading options for all available data, providing users with a seamless experience. These features will greatly promote our understanding of the roles played by RBPs in various physiological and pathological processes.

Next, we illustrate the utility of RBPWorld with a specific example: Nucleophosmin 1 (NPM1), an extensively studied gene and the most frequently mutated gene in adult acute myeloid leukemia (AML) (80). NPM1 is an abundant nucleolar protein associated with ribosome maturation and export, centrosome replication, and response to stressful stimuli (81). By entering 'NPM1' on the homepage or in the upper right-hand corner of webpage, users can explore the RBP properties, functions and disease associations of NPM1 utilizing RBPWorld (Figure 5D). An RBD, nucleoplasmin, was found in NPM1 and other two members of nucleoplasmin family: nucleoplasmin 2 (NPM2) and nucleoplasmin 3 (NPM3). Proteins with nucleoplasmin domains were found functionally associated with ribosome functions (82). Evidence from RBPomes and CLIP-seq also suggested NPM1 as an RBP with high confidence. Leveraging RBPWorld, we found that NPM1 was conserved in five species, providing additional information on one more specie compared to EURBPDB. We focused on exploring the expression and function of NPM1 in human. Other than its well-known function in rRNA processing and ribosomal subunit assembly via binding to rRNA, we identified moonlighting function of NPM1: binding to DNA. Several studies have reported that the C-terminal domain of NPM1 binds with high affinity to G-quadruplex DNA (83). This binding effect is directly related to the cytoplasmic delocalization of NPM1, which is impaired in protein variants associated with AML. Another important characteristic of NPM1 was the ability to form membrane-less condensates in the nucleolus. Condensates section showed NPM1 forms SG and nucleoli rim, which were critical to protect rRNA and other nucleic acids from stressful conditions (84). Bulkperturb-seq dataset analysis showed that NPM1 knockdown affected gene expression levels and DAS, which mainly pertain to ribosome biosynthesis of cells and were dysregulated, highlighting the role of RBPWorld in identifying RBP-interacting genes. We also found that NPM1 is associated with infections by various viruses. It has been reported that B23/NPM1 affects the overall replication of viruses such as Human Immunodeficiency Virus (HIV), Hepatitis B Virus (HBV), Hepatitis C Virus (HCV), Hepatitis D Virus (HDV), and Human Papillomavirus (HPV) by playing functional roles at different stages of viral replication, including nuclear importation, viral genome transcription, assembly and final particle formation (85). Beyond viral infection diseases, RBPWorld also presented NPM1–disease associations generated from gene–disease association databases such as Open Targets Genetics and Clinvar. For investigators interested in NPM1 targeting, we also present two drugs that have entered Phase IV clinical trialsceritinib and crizotinibto facilitate their studies.

Overall, the updated version of RBPWorld emphasizes deeper multi-dimensional exploration of functional and disease association information across species of RBPs, providing researchers with barrier-free access to valuable bioinformatics data. Most of the above results were directly extracted from RBPWorld web interface, demonstrating its efficiency in deeply elucidate RBP functions and roles in diseases.

Discussion and conclusions

In this study, we conducted a comprehensive analysis to identify highly reliable RBPs from various species using our refined RBP identification methods. Our analysis yielded a total of 1 393 413 RBPs, which we categorized into 849 families. Furthermore, we investigated the conservation, functions, disease associations and targeting drugs of these RBPs. To enhance accessibility and facilitate further research, we developed a comprehensive platform, named RBPWorld (http://research.gzsys.org.cn/rbpworld/#/home), which enables users to explore the conservation, functions and disease associations across different species. RBPWorld serves as a valuable resource for researchers and provides insights into the critical roles of RBPs in various biological contexts.

Our findings revealed that the majority of RBPs exhibit high conservation across species, with all of them being conserved in more than half of the species included in our study. This underscores the fundamental roles that RBPs play in various biological processes. Moreover, we discovered that the functions of RBPs are more intricate than previously assumed. First, about one-third of the RBPs exhibit binding affinity toward multiple RNAs ($\geq$2). Second, 30.66% of RBPs possess additional functions beyond RNA-binding, such as DNA binding, metabolic enzymatic activity and E3 ligase function. Lastly, perturbing the expression of RBP genes can have far-reaching effect on a wide range of biological processes. Given their complex and pivotal regulatory roles, dysregulation of RBPs is anticipated to contribute to the development of severe diseases. Consistently, we identified strong associations between RBPs and various human genetic diseases, with nearly one-third of RBPs implicated in these conditions. Moreover, we observed significant alterations in the expression of over half of the RBPs in at least three cancer types we investigated. These results emphasize the potential relevance of RBPs as therapeutic targets and diagnostic markers in various diseases including cancers.

BulkPerturb-seq is a widely use and effective strategy for studying gene function. In our analysis, we manually collected 616 RBP bulkPerturb-seq datasets derived from 62 cell lines, systematically uncovering alternative splicing patterns and potential downstream pathways regulated by diverse RBPs. As the RBP field progresses rapidly, there is a growing accumula-

tion of bulkPerturb-seq datasets for RBPs. We are committed to continuously gathering and integrating these datasets into the RBPWorld platform, which will contribute to the emergence of novel insights into RBP functions. The RBPWorld platform also provides users with convenient and effective access to these valuable datasets for their studies on the functions and disease associations of RBPs. This facilitates the reuse and exploration of these datasets, fostering further discoveries and advancements in our understanding of RBPs.

Comparing to other currently available RBP databases (8,12,86), RBPWorld stands out for its extensive high-reliability RBP catalog, comprehensive functional annotations and extensive disease-related information. With its user-friendly web interface, RBPWorld serves as a powerful platform for decoding the RBP functions and their associations with diseases.

Data availability

RBPWorld platform is freely available at http://research.gzsys.org.cn/rbpworld/#/home.

Supplementary data

Supplementary Data are available at NAR Online.

Conflict of interest statement

None declared.

References

1. Gerstberger,S., Hafner,M. and Tuschl,T. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.
2. Castello,A., Fischer,B., Eichelbaum,K., Horos,R., Beckmann,B.M., Strein,C., Davey,N.E., Humphreys,D.T., Preiss,T., Steinmetz,L.M., *et al.* (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, **149**, 1393–1406.
3. Chen,J., Fang,X., Zhong,P., Song,Z. and Hu,X. (2019) N6-methyladenosine modifications: interactions with novel RNA-binding proteins and roles in signal transduction. *RNA Biol.*, **16**, 991–1000.
4. Steinmetz,E.J., Conrad,N.K., Brow,D.A. and Corden,J.L. (2001) RNA-binding protein Nrd1 directs poly (A)-independent 3'-end formation of RNA polymerase II transcripts. *Nature*, **413**, 327–331.
5. Song,P., Yang,F., Jin,H. and Wang,X. (2021) The regulation of protein translation and its implications for cancer. *Signal Transd. Target. Ther.*, **6**, 68.
6. ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
7. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still going strong. *Nucleic Acids Res.*, **47**, D330–D338.
8. Cook,K.B., Kazan,H., Zuberi,K., Morris,Q. and Hughes,T.R. (2011) RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.*, **39**, D301–D308.
9. Giudice,G., Sánchez-Cabo,F., Torroja,C. and Lara-Pezzi,E. (2016) ATtRACT-a database of RNA-binding proteins and associated motifs. *Database*, **2016**, baw035.
10. Li,J.H., Liu,S., Zhou,H., Qu,L.H. and Yang,J.H. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.*, **42**, D92–D97.
11. Zhao,W., Zhang,S., Zhu,Y., Xi,X., Bao,P., Ma,Z., Kapral,T.H., Chen,S., Zagrovic,B., Yang,Y.T., *et al.* (2022) POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res.*, **50**, D287–D294.
12. Liao,J.Y., Yang,B., Zhang,Y.C., Wang,X.J., Ye,Y., Peng,J.W., Yang,Z.Z., He,J.H., Zhang,Y., Hu,K., *et al.* (2020) EuRBPDB: a comprehensive resource for annotation, functional and oncological investigation of eukaryotic RNA binding proteins (RBPs). *Nucleic Acids Res.*, **48**, D307–D313.
13. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr., Jungkamp,A.C., Munschauer,M., *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
14. Van Nostrand,E.L., Pratt,G.A., Shishkin,A.A., Gelboin-Burkhart,C., Fang,M.Y., Sundararaman,B., Blue,S.M., Nguyen,T.B., Surka,C., Elkins,K., *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.
15. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M., *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
16. Gebauer,F., Schwarzl,T., Valcárcel,J. and Hentze,M.W. (2021) RNA-binding proteins in human genetic disease. *Nat. Rev. Genet.*, **22**, 185–198.
17. Wang,X., Liu,R., Zhu,W., Chu,H., Yu,H., Wei,P., Wu,X., Zhu,H., Gao,H., Liang,J., *et al.* (2019) UDP-glucose accelerates SNAI1 mRNA decay and impairs lung cancer metastasis. *Nature*, **571**, 127–131.
18. Hallegger,M., Chakrabarti,A.M., Lee,F.C.Y., Lee,B.L., Amalietti,A.G., Odeh,H.M., Copley,K.E., Rubien,J.D., Portz,B., Kuret,K., *et al.* (2021) TDP-43 condensation properties specify its RNA-binding and regulatory repertoire. *Cell*, **184**, 4680–4696.
19. Pavitt,G.D. (2018) Regulation of translation initiation factor eIF2B at the hub of the integrated stress response. *Wiley Interdiscipl. Rev. RNA*, **9**, e1491.
20. Wang,E., Lu,S.X., Pastore,A., Chen,X., Imig,J., Chun-Wei Lee,S., Hockemeyer,K., Ghebrechristos,Y.E., Yoshimi,A., Inoue,D., *et al.* (2019) Targeting an RNA-binding protein network in acute myeloid leukemia. *Cancer Cell*, **35**, 369–384.
21. Hashemikhabir,S., Neelamraju,Y. and Janga,S.C. (2015) Database of RNA binding protein expression and disease dynamics (READ DB). *Database*, **2015**, bav072.
22. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
23. Harrison,P.W., Amode,M.R., Austine-Orimoloye,O., Azov,A.G., Barba,M., Barnes,I., Becker,A., Bennett,R., Berry,A., Bhai,J., *et al.* (2024) Ensembl 2024. *Nucleic Acids Res.*, **52**, D891–D899.
24. Marchin,M., Kelly,P.T. and Fang,J. (2005) Tracker: continuous HMMER and BLAST searching. *Bioinformatics*, **21**, 388–389.
25. Backlund,M., Stein,F., Rettel,M., Schwarzl,T., Perez-Perri,J.I., Brosig,A., Zhou,Y., Neu-Yilik,G., Hentze,M.W. and Kulozik,A.E.

(2020) Plasticity of nuclear and cytoplasmic stress responses of RNA-binding proteins. *Nucleic Acids Res.*, **48**, 4725–4740.

26. Flynn,R.A., Belk,J.A., Qi,Y., Yasumoto,Y., Wei,J., Alfajaro,M.M., Shi,Q., Mumbach,M.R., Limaye,A., DeWeirdt,P.C., *et al.* (2021) Discovery and functional interrogation of SARS-CoV-2 RNA-host protein interactions. *Cell*, **184**, 2394–2411.

27. Gandhi,M., Groß,M., Holler,J.M., Coggins,S.A., Patil,N., Leupold,J.H., Munschauer,M., Schenone,M., Hartigan,C.R., Allgayer,H., *et al.* (2020) The lncRNA lincNMR regulates nucleotide metabolism via a YBX1 - RRM2 axis in cancer. *Nat. Commun.*, **11**, 3214.

28. Azman,M.S., Alard,E.L., Dodel,M., Capraro,F., Faraway,R., Dermit,M., Fan,W., Chakraborty,A., Ule,J. and Mardakheh,F.K. (2023) An ERK1/2-driven RNA-binding switch in nucleolin drives ribosome biogenesis and pancreatic tumorigenesis downstream of RAS oncogene. *EMBO J.*, **42**, e110902.

29. Mestre-Farràs,N., Guerrero,S., Bley,N., Rivero,E., Coll,O., Borràs,E., Sabidó,E., Indacochea,A., Casillas-Serra,C., Järvelin,A.I., *et al.* (2022) Melanoma RBPome identification reveals PDIA6 as an unconventional RNA-binding protein involved in metastasis. *Nucleic Acids Res.*, **50**, 8207–8225.

30. Sayers,E.W., Beck,J., Bolton,E.E., Bourexis,D., Brister,J.R., Canese,K., Comeau,D.C., Funk,K., Kim,S., Klimke,W., *et al.* (2021) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **49**, D10–D17.

31. Safran,M., Dalah,I., Alexander,J., Rosen,N., Iny Stein,T., Shmoish,M., Nativ,N., Bahir,I., Doniger,T., Krug,H., *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database*, **2010**, baq020.

32. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P., *et al.* (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.

33. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

34. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

35. Shen,W.K., Chen,S.Y., Gan,Z.Q., Zhang,Y.Z., Yue,T., Chen,M.M., Xue,Y., Hu,H. and Guo,A.Y. (2023) AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res.*, **51**, D39–D45.

36. Xiao,R., Chen,J.Y., Liang,Z., Luo,D., Chen,G., Lu,Z.J., Chen,Y., Zhou,B., Li,H., Du,X., *et al.* (2019) Pervasive chromatin-RNA binding protein interactions enable RNA-based regulation of transcription. *Cell*, **178**, 107–121.

37. Van Nostrand,E.L., Freese,P., Pratt,G.A., Wang,X., Wei,X., Xiao,R., Blue,S.M., Chen,J.Y., Cody,N.A.L., Dominguez,D., *et al.* (2020) A large-scale binding and functional map of human RNA-binding proteins. *Nature*, **583**, 711–719.

38. Zhou,J., Xu,Y., Lin,S., Guo,Y., Deng,W., Zhang,Y., Guo,A. and Xue,Y. (2018) iUUCD 2.0: an update with rich annotations for ubiquitin and ubiquitin-like conjugations. *Nucleic Acids Res.*, **46**, D447–D453.

39. Wang,X., Li,Y., He,M., Kong,X., Jiang,P., Liu,X., Diao,L., Zhang,X., Li,H., Ling,X., *et al.* (2022) UbiBrowser 2.0: a comprehensive resource for proteome-wide known and predicted ubiquitin ligase/deubiquitinase-substrate interactions in eukaryotic species. *Nucleic Acids Res.*, **50**, D719–D728.

40. UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.

41. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

42. Milacic,M., Beavers,D., Conley,P., Gong,C., Gillespie,M., Griss,J., Haw,R., Jassal,B., Matthews,L., May,B., *et al.* (2024) The

43. Uhlén,M., Fagerberg,L., Hallström,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,Å., Kampf,C., Sjöstedt,E., Asplund,A., *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.

44. Youn,J.Y., Dyakov,B.J.A., Zhang,J., Knight,J.D.R., Vernon,R.M., Forman-Kay,J.D. and Gingras,A.C. (2019) Properties of stress granule and P-body proteomes. *Mol. Cell*, **76**, 286–294.

45. Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A., Kaul,R., *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.

46. Greenwald,W.W., Li,H., Smith,E.N., Benaglio,P., Nariai,N. and Frazer,K.A. (2017) Pgltools: a genomic arithmetic tool suite for manipulation of Hi-C peak and other chromatin interaction data. *BMC Bioinf.*, **18**, 207.

47. Ochoa,D., Hercules,A., Carmona,M., Suveges,D., Gonzalez-Uriarte,A., Malangone,C., Miranda,A., Fumis,L., Carvalho-Silva,D., Spitzer,M., *et al.* (2021) Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res.*, **49**, D1302–D1310.

48. Landrum,M.J., Chitipiralla,S., Brown,G.R., Chen,C., Gu,B., Hart,J., Hoffman,D., Jang,W., Kaur,K., Liu,C., *et al.* (2020) ClinVar: improvements to accessing data. *Nucleic Acids Res.*, **48**, D835–D844.

49. Wang,Q., Dhindsa,R.S., Carss,K., Harper,A.R., Nag,A., Tachmazidou,I., Vitsios,D., Deevi,S.V.V., Mackay,A., Muthas,D., *et al.* (2021) Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature*, **597**, 527–532.

50. Martin,A.R., Williams,E., Foulger,R.E., Leigh,S., Daugherty,L.C., Niblock,O., Leong,I.U.S., Smith,K.R., Gerasimenko,O., Haraldsdottir,E., *et al.* (2019) PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.*, **51**, 1560–1565.

51. Thormann,A., Halachev,M., McLaren,W., Moore,D.J., Svinti,V., Campbell,A., Kerr,S.M., Tischkowitz,M., Hunt,S.E., Dunlop,M.G., *et al.* (2019) Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat. Commun.*, **10**, 2373.

52. Rehm,H.L., Berg,J.S., Brooks,L.D., Bustamante,C.D., Evans,J.P., Landrum,M.J., Ledbetter,D.H., Maglott,D.R., Martin,C.L., Nussbaum,R.L., *et al.* (2015) ClinGen–the clinical genome resource. *N. Engl. J. Med.*, **372**, 2235–2242.

53. Gaulton,A., Hersey,A., Nowotka,M., Bento,A.P., Chambers,J., Mendez,D., Mutowo,P., Atkinson,F., Bellis,L.J., Cibrián-Uhalte,E., *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res.*, **45**, D945–D954.

54. Katz,K., Shutov,O., Lapoint,R., Kimelman,M., Brister,J.R. and O'Sullivan,C. (2022) The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.*, **50**, D387–D390.

55. Sinyor,M., Schaffer,A. and Levitt,A. (2010) The sequenced treatment alternatives to relieve depression (STAR*D) trial: a review. *Can. J. Psychiatry*, **55**, 126–135.

56. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf.*, **12**, 323.

57. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

58. Shen,S., Park,J.W., Lu,Z.X., Lin,L., Henry,M.D., Wu,Y.N., Zhou,Q. and Xing,Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl Acad. Sci. U.S.A.*, **111**, E5593–E5601.

59. Katz,Y., Wang,E.T., Airoldi,E.M. and Burge,C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**, 1009–1015.

60. Chu,C. and Chang,H.Y. (2018) ChIRP-MS: rNA-directed proteomic discovery. *Methods Mol. Biol.*, **1861**, 37–45.

Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res.*, **52**, D672–D678.

61. McHugh,C.A. and Guttman,M. (2018) RAP-MS: a method to identify proteins that interact directly with a specific RNA molecule in cells. *Methods Mol. Biol.*, **1649**, 473–488.

62. Michael,T.P. and VanBuren,R. (2020) Building near-complete plant genomes. *Curr. Opin. Plant Biol.*, **54**, 26–33.

63. Sun,L., Li,P., Ju,X., Rao,J., Huang,W., Ren,L., Zhang,S., Xiong,T., Xu,K., Zhou,X., *et al.* (2021) In vivo structural characterization of the SARS-CoV-2 RNA genome identifies host proteins vulnerable to repurposed drugs. *Cell*, **184**, 1865–1883.

64. Iselin,L., Palmalux,N., Kamel,W., Simmonds,P., Mohammed,S. and Castello,A. (2022) Uncovering viral RNA-host cell interactions on a proteome-wide scale. *Trends Biochem. Sci.*, **47**, 23–38.

65. Wishart,D.S., Guo,A., Oler,E., Wang,F., Anjum,A., Peters,H., Dizon,R., Sayeeda,Z., Tian,S., Lee,B.L., *et al.* (2022) HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.*, **50**, D622–D631.

66. Cui,F., Li,S., Zhang,Z., Sui,M., Cao,C., El-Latif Hesham,A. and Zou,Q. (2022) DeepMC-iNABP: deep learning for multiclass identification and classification of nucleic acid-binding proteins. *Comput. Struct. Biotechnol. J.*, **20**, 2020–2028.

67. Pratt,H.E., Andrews,G.R., Phalke,N., Purcaro,M.J., van der Velde,A., Moore,J.E. and Weng,Z. (2022) Factorbook: an updated catalog of transcription factor motifs and candidate regulatory motif sites. *Nucleic Acids Res.*, **50**, D141–D149.

68. Huppertz,I., Perez-Perri,J.I., Mantas,P., Sekaran,T., Schwarzl,T., Russo,F., Ferring-Appel,D., Koskova,Z., Dimitrova-Paternoga,L., Kafkia,E., *et al.* (2022) Riboregulation of Enolase 1 activity controls glycolysis and embryonic stem cell differentiation. *Mol. Cell*, **82**, 2666–2680.

69. Scott,D.D., Trahan,C., Zindy,P.J., Aguilar,L.C., Delubac,M.Y., Van Nostrand,E.L., Adivarahan,S., Wei,K.E., Yeo,G.W., Zenklusen,D., *et al.* (2017) Nol12 is a multifunctional RNA binding protein at the nexus of RNA and DNA metabolism. *Nucleic Acids Res.*, **45**, 12509–12528.

70. Cui,Y.H., Xiao,L., Rao,J.N., Zou,T., Liu,L., Chen,Y., Turner,D.J., Gorospe,M. and Wang,J.Y. (2012) miR-503 represses CUG-binding protein 1 translation by recruiting CUGBP1 mRNA to processing bodies. *Mol. Biol. Cell*, **23**, 151–162.

71. Lee,E.K. (2012) Post-translational modifications of RNA-binding proteins and their roles in RNA granules. *Curr. Protein Pept. Sci.*, **13**, 331–336.

72. Seidler,N.W. (2013) Functional diversity. *Adv. Exp. Med. Biol.*, **985**, 103–147.

73. Lu,J., Zhong,C., Luo,J., Shu,F., Lv,D., Liu,Z., Tan,X., Wang,S., Wu,K., Yang,T., *et al.* (2021) HnRNP-L-regulated circCSPP1/miR-520h/EGR1 axis modulates autophagy and promotes progression in prostate cancer. *Mol. Ther. Nucleic Acids*, **26**, 927–944.

74. Emani,M.R., Närvä,E., Stubb,A., Chakroborty,D., Viitala,M., Rokka,A., Rahkonen,N., Moulder,R., Denessiouk,K., Trokovic,R., *et al.* (2015) The L1TD1 protein interactome reveals the importance of post-transcriptional regulation in human pluripotency. *Stem Cell Rep.*, **4**, 519–528.

75. Yao,R.W., Xu,G., Wang,Y., Shan,L., Luan,P.F., Wang,Y., Wu,M., Yang,L.Z., Xing,Y.H., Yang,L., *et al.* (2019) Nascent Pre-rRNA sorting via phase separation drives the assembly of dense fibrillar components in the human nucleolus. *Mol. Cell*, **76**, 767–783.

76. Peculis,B.A. (2001) snoRNA nuclear import and potential for cotranscriptional function in pre-rRNA processing. *RNA*, **7**, 207–219.

77. Bava,F.A., Eliscovich,C., Ferreira,P.G., Miñana,B., Ben-Dov,C., Guigó,R., Valcárcel,J. and Méndez,R. (2013) CPEB1 coordinates alternative 3'-UTR formation with translational regulation. *Nature*, **495**, 121–125.

78. Gazzara,M.R., Mallory,M.J., Roytenberg,R., Lindberg,J.P., Jha,A., Lynch,K.W. and Barash,Y. (2017) Ancient antagonism between CELF and RBFOX families tunes mRNA splicing outcomes. *Genome Res.*, **27**, 1360–1370.

79. Inoue,A., Yamamoto,N., Kimura,M., Nishio,K., Yamane,H. and Nakajima,K. (2014) RBM10 regulates alternative splicing. *FEBS Lett.*, **588**, 942–947.

80. Brunetti,L., Gundry,M.C., Sorcini,D., Guzman,A.G., Huang,Y.H., Ramabadran,R., Gionfriddo,I., Mezzasoma,F., Milano,F., Nabet,B., *et al.* (2018) Mutant NPM1 maintains the leukemic state through HOX expression. *Cancer Cell*, **34**, 499–512.

81. Yang,K., Yang,J. and Yi,J. (2018) Nucleolar stress: hallmarks, sensing mechanism and diseases. *Cell Stress*, **2**, 125–140.

82. Frehlick,L.J., Eirín-López,J.M. and Ausió,J. (2007) New insights into the nucleophosmin/nucleoplasmin family of nuclear chaperones. *BioEssays*, **29**, 49–59.

83. Chiarella,S., De Cola,A., Scaglione,G.L., Carletti,E., Graziano,V., Barcaroli,D., Lo Sterzo,C., Di Matteo,A., Di Ilio,C., Falini,B., *et al.* (2013) Nucleophosmin mutations alter its nucleolar localization by impairing G-quadruplex binding at ribosomal DNA. *Nucleic Acids Res.*, **41**, 3228–3239.

84. Yang,K., Wang,M., Zhao,Y., Sun,X., Yang,Y., Li,X., Zhou,A., Chu,H., Zhou,H., Xu,J., *et al.* (2016) A redox mechanism underlying nucleolar stress sensing by nucleophosmin. *Nat. Commun.*, **7**, 13599.

85. Lobaina,Y. and Perera,Y. (2019) Implication of B23/NPM1 in viral infections, potential uses of B23/NPM1 inhibitors as antiviral therapy. *Infect. Dis. Drug Targets*, **19**, 2–16.

86. Caudron-Herger,M., Jansen,R.E., Wassmer,E. and Diederichs,S. (2021) RBP2GO: a comprehensive pan-species database on RNA-binding proteins, their interactions and functions. *Nucleic Acids Res.*, **49**, D425–D436.