Refining the pool of RNA-binding domains advances the classification and prediction of RNA-binding proteins

Elsa Wassmer[1], Gergely Koppàny[1], Sven Diederichs[1,2,*] and Maïwen Caudron-Herger[3,*]

[1] Division of RNA Biology & Cancer, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

[2] Division of Cancer Research, Department of Thoracic Surgery, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, German Cancer Consortium (DKTK) - Partner Site Freiburg, 79106 Freiburg, Germany

[3] Research Group RNA-Protein Complexes & Cell Proliferation, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

* To whom correspondence should be addressed.

MCH: Tel: +49 6221 42 43 83; Email: m.caudron@dkfz-heidelberg.de

SD: Email: s.diederichs@dkfz-heidelberg.de

ORCID Maïwen Caudron-Herger: 0000-0002-7030-8153

ORCID Sven Diederichs: 0000-0001-7901-4752

ORCID Elsa Wassmer: 0000-0002-3595-3106

Keywords

RNA-binding proteins, RBP2GO database, RNA-binding domains, RNA-related protein families, intrinsically disordered regions

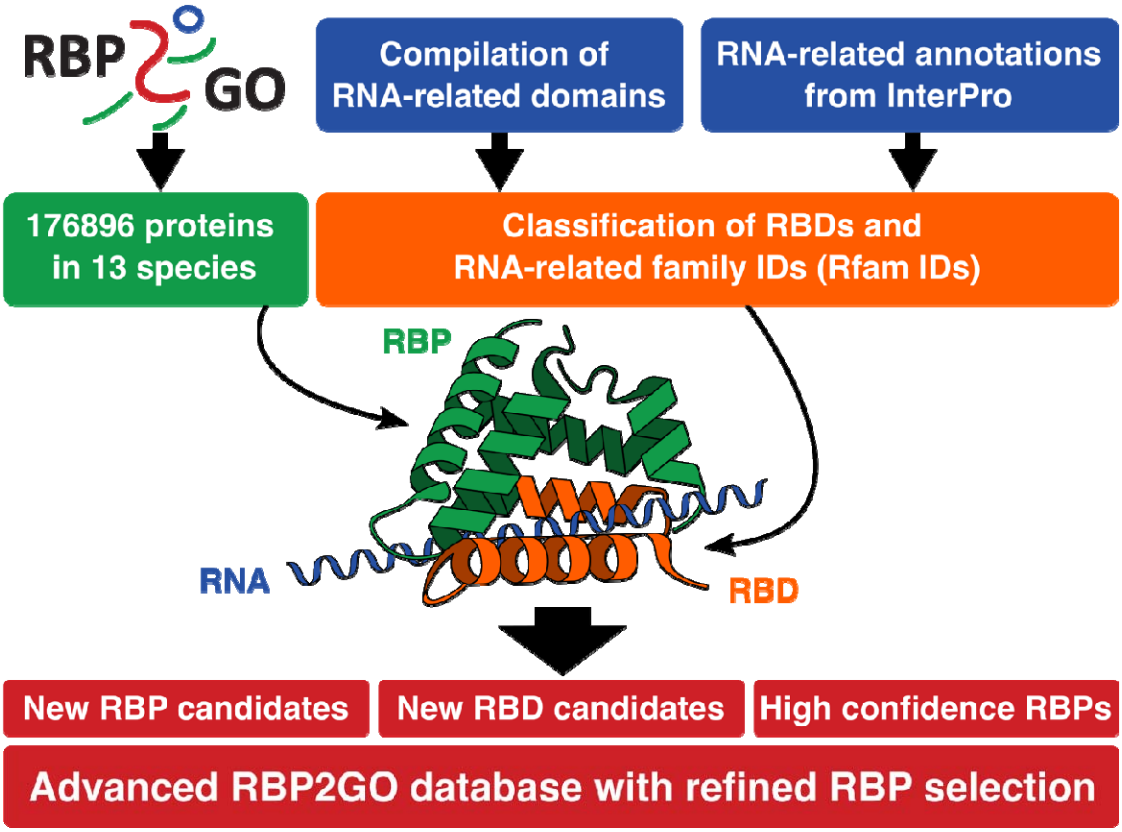
Key Points

- Comprehensive analysis of RNA-related protein domains and families enriched in RNA-binding proteins (RBPs)
- Pan-species prediction of new RBPs, and prediction and validation of new RNA-binding domains
- Online resource with complete dataset including high-confidence human RBPs according to a new scoring system

ABSTRACT

From transcription to decay, RNA-binding proteins (RBPs) influence RNA metabolism. Using the RBP2GO database that combines proteome-wide RBP screens from 13 species, we investigated the RNA-binding features of 176896 proteins. By compiling published lists of RNA-binding domains (RBDs) and RNA-related protein family (Rfam) IDs with lists from the InterPro database, we analyzed the distribution of the RBDs and Rfam IDs in RBPs and non-RBPs to select RBDs and Rfam IDs that were enriched in RBPs. We also explored proteins for their content of intrinsically disordered region (IDRs) and found a strong positive correlation between IDRs and RBDs. Our bioinformatic analysis indicated that RBDs/Rfam IDs were strong indicators of the RNA-binding potential of proteins and helped predicting new RBP candidates, especially in species with few available proteome-wide RBP identification studies. By further analyzing RBPs with no known RBD, we also predicted 15 new RBDs that were experimentally validated by RNA-bound peptides. Finally, we created the new RBP2GO composite score by combining the RBP2GO score with new quality factors linked to RBDs and Rfam IDs. Based on the RBP2GO composite score, we compiled a list of 2018 high-confidence human RBPs. The knowledge collected here was integrated into the RBP2GO database at https://RBP2GO.dkfz.de.

GRAPHICAL ABSTRACT



INTRODUCTION

RNA and RNA-binding proteins (RBPs) have emerged as central players in a large number of key cellular processes (1) with significant implications for our understanding of severe diseases such as neurological disorders or cancer (2-6). Therefore, it is fundamental to comprehensively examine RNA-protein complexes and their molecular functions.

Over the last decade, a variety of orthogonal techniques were developed to systematically identify RBPs based on proteome-wide analyses. The study of the mRNA-bound proteome (7-12) combines the UV-crosslinking of RNA molecules and their bound proteins with the use of oligo-dT beads to isolate the RNA-protein complexes and identify the bound proteins via mass spectrometry. The technique was further improved to identify the RNA-binding regions within the RBPs (13) and was applied on different human cells such as HEK293, HeLa and K562 (7,8,12), as well as other model organisms such as *Mus musculus* (10,14,15). Other strategies were explored to circumvent the sampling bias due to the use of oligo-dT beads, notably the isolation of crosslinked RNA-RBP complexes by organic phase separation (16-18), or the separation of RNA-dependent complexes on sucrose gradients upon RNase treatment (19-21).

These massive efforts resulted in the generation of an ever-growing number of datasets. In particular for model species such as *Homo sapiens*, six datasets were available in 2018 (13) as compared to 43 in 2020 (22), whereas only three datasets had been published for *Danio rerio* (23). Furthermore, previous work showed the poor overlap between the different datasets available for *Homo sapiens,* with many proteins being detected only once as RBP (22), highlighting the need for tools that could quantify the relevance of RBP candidates. A number of prediction algorithms were developed, mostly based on the protein sequence (24). Alternatively, the SONAR study predicted RNA-binding properties based on protein-protein interaction, as RBPs often interact with other RBPs (19,25). To facilitate the efficient screening of the remarkably growing RBP resources, we recently compiled all the available datasets into the comprehensive pan-species RBP2GO database (https://rbp2go.dkfz.de). RBP2GO encompasses 103 datasets and 22554 RBP candidates, along with their interactions and their functions (22). The database also displays the RBP2GO score for each protein, which reflects the probability for an RBP candidate to be a true RBP (22). However, due to a lack of data for some species, this score is only available for 9 out of the 13 featured species. Furthermore, it does not take into account data on the functional domains of the proteins.

A number of canonical RNA-binding domains (RBDs) such as the RNA-recognition motif (RRM) or the K-homology (KH) domain (26,27) have been catalogued (28) and further exploited to identify new RBPs (29). However, recent studies revealed that non-canonical RBDs and intrinsically disordered regions (IDRs) were also enriched in RBP candidates (8,30). Moreover, IDR-containing proteins such as NF-kappa-B-activating protein (NKAP) can bind to RNA in the absence of RBD via their IDRs (31). The potential of IDRs for binding to RNA was further confirmed in proteome-wide studies that directly identified peptides bound to RNA in *Homo sapiens* and *Drosophila melanogaster* (13,32). Currently, many of the newly identified RBP candidates are awaiting further experimental validation. Assessing

the specificity of the rapidly growing list of RBP candidates urgently needs further criteria to evaluate the likelihood of each protein to be a true RBP.

In this study, we addressed this issue by performing a pan-species analysis of the RNA-binding characteristics of the proteins included in the RBP2GO database. To this aim, we collected lists of RBDs and RNA-related protein families (Rfam) from published studies (8,13,29) and from the InterPro database (33). We examined their distribution in RBPs as compared to non-RBPs in order to compile a list of selected RBDs and Rfam identifiers (Rfam IDs) that were enriched in RBPs. Using data from the MobiDB database (34,35), we also analyzed the IDR content and its distribution in proteins with respect to their RBD content and found a strong correlation between the two features. Our bioinformatic approach indicates that the presence of a selected RBD / Rfam ID can be used to evaluate the RNA-binding potential of proteins as well as to predict new RBP candidates, in particular in species with very few available proteome-wide studies. In addition, we analyzed the RBPs lacking selected RBDs and predicted 15 new RBDs, that we validated based on studies of RNA-bound peptides. Finally, by calculating a new RBP2GO composite score that combines the former RBP2GO score (19) together with new quality factors characterizing the selected RBDs and Rfam IDs, we computed a list of 2018 high-confidence human RBPs. The results of our analysis were implemented into the RBP2GO database, along with new search options to browse 1) RBDs and Rfam IDs via a given InterPro ID and 2) proteins, with respect to their RNA-binding features. We anticipate that this additional information will considerably help refining our classification of RBP candidates and foster new developments in the field.

MATERIAL AND METHODS

Compilation of a list of RNA-binding domain candidates

Based on a literature search, the lists of RBDs from three studies were downloaded from the respective supplementary information (8,13,29). The domain IDs were converted into InterPro IDs if referenced from another database such as Pfam (36) and finally fused together with the list of IDs from InterPro. Outdated IDs were updated or discarded if no corresponding InterPro ID could be found. Altogether, 808 IDs were retrieved from the published datasets.

Based on a database search, we first downloaded the list of the InterPro IDs related to “RNA-binding” from the InterPro website (https://www.ebi.ac.uk/interpro/, InterPro v.88 released on March 10th, 2022) (33). Second, since the InterPro database provides a Gene Ontology (GO) annotation that is associated to the InterPro IDs, we downloaded all the IDs annotated with the GO term “RNA-binding” (GO:0003723). Third, we downloaded from the InterPro website all the domains that overlapped with the “RNA binding domain superfamily” (IPR035979). Taken together, these three datasets amounted to 2251 unique InterPro IDs.

The fusion of the IDs from the InterPro database and the published datasets resulted in a list of 2712 unique RNA-related IDs (Supplementary Table S1). These RNA-related InterPro IDs were further filtered for the “Domain” and “Repeat” types of IDs (1289 IDs, Figure 1A).

Selection of the RNA-binding domains

A three-step selection procedure was applied to the set of RBD candidates based on the hit ratios between RBPs (RBP candidates from proteome-wide studies) and non-RBPs in the species reported in the RBP2GO dataset (https://rbp2go.dkfz.de) (22). First, we selected the InterPro IDs of the RBD candidates that were enriched (hits in RBPs > hits in non-RBPs) in *Homo sapiens* (Hs) (Selection 1, Figure 1A) as the number of studies available for this species was by far the largest as compared to the other species (Supplementary Figure S1). We updated the UniProt IDs of the proteins according to the UniProt 2022_01 release from February 23rd, 2022 (37). The data for the obsolete UniProt IDs were either deleted when the ID was fused to another one already existing in the database, or duplicated if the ID was split between several new UniProt IDs. The updated sizes of the proteomes are reported in Supplementary Table S2. In a second step, for RBDs not present in Hs (Selection 2, Figure 1A), we selected the InterPro IDs of the RBD candidates that were enriched in *Mus Musculus* (Mm), *Saccharomyces cerevisiae* (Sc) and *Drosophila melanogaster* (Dm) in at least half of the species where the RBD was present. Finally, for RBDs present in neither of the four species Hs, Mm, Sc, Dm (Selection 3, Figure 1A), we selected the RBDs that were enriched in at least half of the remaining nine species where the RBD was present, *i.e.*, *Arabidopsis thaliana* (At), *Caenorhabditis elegans* (Ce), *Plasmodium falciparum* (Pf), *Escherichia coli* (Ec), *Danio rerio* (Dr), *Trypanosoma brucei* (Tb), *Salmonella Typhimurium* (ST), *Leishmania mexicana* (Lm) and *Leishmania donovani* (Ld). This three-step selection procedure led to the selection of 977 RBDs (Supplementary Table S3). The proteins of the species represented in the RBP2GO database were updated with the information on the selected RBDs (Supplementary Table S4). The data available in the Supplementary Table S4 was used to produce the graphs displayed in Figures 1B and 2A-D.

Protein expression levels in HeLa cells

The expression data displayed in the Figure 2E and Supplementary Figure S2 was downloaded for analysis from the EBI atlas website (https://www.ebi.ac.uk/gxa/experiments/E-PROT-19/Results) (38). The expression values can be found in the Supplementary Table S5.

Retrieval of the InterPro domain coordinates and compilation of the number and content fraction of RBDs per protein

The coordinates of each InterPro ID of the selected RBDs in the proteins were downloaded from the InterPro database (33) (https://ftp.ebi.ac.uk/pub/databases/interpro/releases/88.0/, released on March 10th, 2022).

To compute the RBD content fraction per protein, we merged separately all the overlapping RBDs present in a protein. Next, we used the protein length previously retrieved from UniProt (37) (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2022_01/, released on February 23rd, 2022) to calculate the fraction of the sequence annotated as RBD for each protein.

To compute the number of RBDs present in each protein, we merged the overlapping coordinates for the InterPro ID in R using the "bed_merge()" function from the "valr" package (39). We then merged, for each protein, the coordinates of the InterPro IDs overlapping by more than 10 amino acids using the same function. The number of coordinate pairs (start, end) generated was then calculated to obtain the number of RBDs per protein which were independent in the sense that they were not overlapping by more than 10 amino acids.

The Supplementary Table S4 was updated with this information and used to produce the Figure 3 and Supplementary Figure S3.

Compilation of a list of RNA-related family IDs and selection of the IDs enriched in RBP candidates.

In addition to the domains and repeats InterPro ID types (see above), we computed a list of RNA-related InterPro family IDs or Rfam IDs. The "RNA binding domain superfamily" (IPR035979) overlaps with several other domains. Accordingly, the list of 2712 RNA-related InterPro IDs (Supplementary Table S1) was filtered to keep only 1028 family IDs. The same procedure as the one used for the selection of the RBDs was applied to select for the Rfam IDs that were enriched in RBPs. This resulted in a final list of 672 Rfam IDs (Supplementary Table S6). The Supplementary Table S4 was updated accordingly with this information.

The results of these analyses were used to produce Figure 4 and Supplementary Figures S4 and S5.

Retrieval of the MobiDB-lite IDR content fraction

Among the various algorithms that were available for the prediction of IDRs (34), we selected the MobiDB-lite prediction as it is more conservative, through the combination of 10 different prediction tools (34,35). To retrieve the IDR content fraction of the proteins, we downloaded the results of the MobiDB-lite algorithm from the MobiDB database (https://mobidb.bio.unipd.it, Version 5.0.0) (35,40) for each of the 13 RBP2GO species. The Supplementary Table S4 was accordingly updated with this information (IDR content fraction).

The results of these analyses were used to produce Figure 5 and Supplementary Figure S6 and S7.

GO enrichment analysis for non-RBPs containing an RBD

For each species, a GO-enrichment analysis was performed in the non-RBPs containing one of the selected RBDs against the proteome with the PANTHER 17.0 classification system (http://pantherdb.org, release 2021_03) that was accessed via the "rbioapi" package in R (41). The results for the different species were then merged in a single table (Supplementary Table S7). The GO terms were filtered to keep the terms with an FDR <0.05. For each GO term, the number of species in which it was significantly enriched and its mean fold enrichment in these species were calculated. For the heatmaps (Figure 6), only the GO terms significantly enriched (FDR < 0.05) in more than four species out of 13, and with a mean fold enrichment > 5, were selected. The heatmaps were created with the geom_tile() function from the ggplot2 package (42), and the GO terms were sorted by decreasing mean fold enrichment.

Identification of new RBDs using the human RBP candidates lacking canonical RBDs

In this part of the analysis, we only included the human proteins as it is the species covered by the highest number of datasets. Based on their RBP2GO score, we selected the top 20% of the RBPs that did not contain any of the selected RBDs (935 "high score" proteins of interest, Supplementary Table S4) that we compared to three reference datasets: 1) the whole proteome (20741 proteins), 2) the non-RBPs without RBD (14411 proteins) and 3) the bottom 20% RBPs lacking RBDs based on their RBP2GO score (989 "low score" proteins, Supplementary Table S4). We extracted the InterPro IDs present in the "high score" dataset and filtered out the IDs that were present in the initial RBD candidate list. Next, we selected for the IDs of the type "Domain" or "Repeat", which resulted in a list of 637 IDs (Supplementary Table S8). For each InterPro ID (domain/repeat), we calculated the proportion of proteins containing the domain in the dataset of interest as well as in the three reference datasets. Next, for each domain, we computed its fold enrichment in the dataset of interest as compared to each of the reference datasets. A *P*-value was also generated as a result of a Fisher's exact test. The *P*-values were further adjusted for multiple testing by the FDR method (43). To identify new RBDs, we applied the following criteria: we selected the domains that were 1) significantly enriched against at least one reference dataset (>2-fold enrichment and adjusted *P*-value <0.05), 2) present in at least 0.5% of the proteins of the "high score" dataset (*i.e.* in at least five proteins) and 3) presenting an RBP vs. non-RBP hits ratio greater 1. Altogether, 15 domains satisfied these three criteria (Supplementary Table S8).

The results of these analyses were used to produce the panels A to E in Figure 7.

Validation of the newly discovered RBDs using RNA-binding peptides

Several studies featured in RBP2GO used approaches enabling the identification of the RNA-binding region of the detected RBP candidates (13,14,16,32,44-46). Using these datasets, we computed a new dataset combining the peptide information with the UniProt ID of the corresponding proteins and with the coordinates of the peptides in the proteins (Supplementary Table S9). Hereby, we only considered the human protein datasets. For each of the 624 proteins containing only the new RBDs (without containing any of the previously selected RBDs), we computed the proportion of peptides overlapping at least 50% with the new RBD relative to the total number of peptides found in the protein. This number was normalized to the fraction of the protein sequence that was covered by the domain. We then summed this number for all the domains corresponding to one InterPro ID over all proteins containing at least one peptide, and normalized by the total number of occurrences of this domain in the proteins containing at least one peptide. The same procedure was repeated for the selected RBDs and the non-selected domains (Figure 1A) using groups of proteins containing only selected RBDs or only non-selected domains respectively. To maintain the same selection criteria as for the new RBDs, we only took into account the selected and non-selected RBDs that were present in at least five RBPs (Supplementary Table S9). The results of these three analyses were used to produce Figure 7F.

Computation of the RBP2GO composite score

To reflect the importance of the presence of RBDs or the relation to an Rfam ID for an RBP candidate, we created a new "RBP2GO composite score" that integrates the knowledge on RBDs and Rfam IDs in addition to the information already provided by the RBP2GO score. This new score comprises three components. Component 1 represents one half of the score and is constituted by the listing count of the protein (*i.e.,* in how many datasets the protein is detected as RBP candidate). Component 1 is normalized to the total number of available datasets for the species and amounts to a maximum of 50 (*e.g.,* if the protein is found in all RBP datasets for the respective species). Component 2 represents one quarter of the score and is the average listing count for the top 10 interactors of the protein (according to the STRING database (STRING-DB.org). Component 2 is normalized and amounts to a maximum of 25. The last quarter is constituted by component 3, a score pertaining to the presence of RBDs or Rfam IDs that are associated to the protein of interest.

To define the component 3, a quality factor was attributed to each RBD / Rfam ID. To this aim, we took into consideration the ratio of hits in RBPs versus hits in non-RBPs in human (Hs) as a surrogate measure for the likelihood of the RBD or Rfam ID being truly associated with RBPs. If the ID was not found in Hs, we considered the ratio in mouse (Mm); if it was not found in Mm, we considered the one in yeast (Sc), and so on following the order of the species by decreasing number of studies (Supplementary Figure S1). For each ID, this ratio was used to attribute a semi-quantitative balancing quality factor: ratio <2 = quality factor of 1; ratio ≥2 = quality factor of 2; ratio ≥4 = quality factor of 3;

ratio ≥8 = quality factor of 4 and ratio ≥16 = quality factor of 5. For infinite ratios (ID only present in RBPs), the number of RBPs with this ID was in addition taken into account: for a number of RBPs ≤2 = quality factor of 2; number of RBPs from 2 to 4 = quality factor of 3; number of RBPs from 4 to 8 = quality factor of 4 and number of RBPs >8 = quality factor of 5. The attribution rules are summarized in the Supplementary Table S10. Finally, we summed the quality factors for all RBDs and RNA-related family IDs present for one protein, and limited this number to 25.

In total, this RBP2GO composite score ranges from 0 to 100, one half of it being computed based on experimental data and the other half based on its interactors and RBD / Rfam IDs.

The Pearson correlation coefficients between the three components of the composite score and the RBP2GO score were computed using the cor() function of the stats package in R. The results are displayed in Supplementary Figure S8.

The RBP2GO composite score was analyzed in the four groups of proteins, *i.e.*, non-RBPs without RBD, non-RBPs with RBD, RBPs without RBD and RBPs with RBD to produce the panel B in Figure 8. Supplementary Table S4 was updated accordingly.

Selection of high-confidence RBPs in Hs

We took advantage of the RBP2GO composite score to compute a list of selected high-confidence human RBPs. This list encompassed all human proteins, whether they were detected as RBPs in proteome-wide screens or not, that had a composite score superior or equal to 10, with at least two out of the three components of the score larger than zero. This resulted in a list of 2018 proteins, among which 1977 were already detected as RBPs in screens of the RBP2GO database (Supplementary Table S11).

Update of the RBP2GO database

The whole dataset compiled in this analysis is available as Supplementary Table S12. In order to make the data more accessible and searchable, this new dataset was included in a user-friendly and intuitive manner into the RBP2GO database. The “RBD_Status” column as well as the number of RBDs and the RBD content fractions were updated to take into account the 15 new RBDs (Supplementary Table S8) in addition to the selected RBDs and Rfam IDs. The new data, including the lists of RBDs and Rfam IDs can be directly downloaded from the download section of the database (https://rbp2go.dkfz.de). It is now also possible via the advanced search option from each species to search for proteins containing a specific InterPro ID, or containing an RBD and/or an Rfam ID. A new section with information about the selection procedures applied to the RBDs and Rfam IDs was also created under the sidebar menu “search domains”, which include a specific search module for InterPro IDs.

Linear regressions

For all linear regressions in this study, the regression lines were generated using the geom_smooth() function from the ggplot2 package (42), with the “lm” method. The correlation coefficients in the figures were computed using the “stat_cor()” function from the ggpubr package (47), except for the Supplementary Figure S8, for which the R base function cor() was used.

RESULTS

Proteome-wide RBP screens allow refining the pool of RNA-binding domains

Prior to the increase in the number of proteome-wide methods dedicated to the identification of RBPs and the resulting knowledge gained regarding RNA-protein interactions, proteins were thought to interact with RNA essentially via well-defined globular RBDs. Some proteins harbor a single RBD while others contain several RBDs, combined in a modular manner, which illustrates the great variability of RBD-RNA interactions (24,26).

Here, we combined and filtered RBD lists from previous works (8,13,29) together with data from the InterPro database (33) to compile a list of in total 2712 InterPro IDs corresponding to classical and non-classical RBDs (8) as well as domains with descriptions and GO terms related to RNA binding (Supplementary Table S1). By filtering for the “domain” and “repeat” types of InterPro IDs, we computed a starting list of 1289 RBD candidates (Figure 1A). Next, we applied a three-step selection procedure (as described in the Material and Methods section) that was based on the ratio between the presence of the domains in RBP candidates (experimentally detected as RBPs) as compared to non-RBPs in the 13 species included in the RBP2GO database in the order of the number of available datasets for the respective species (Supplementary Figure S1): *Homo sapiens* (Hs), *Mus musculus* (Mm), *Saccharomyces cerevisiae* (Sc), *Drosophila melanogaster* (Dm), *Arabidopsis thaliana* (At), *Caenorhabditis elegans* (Ce), *Plasmodium falciparum* (Pf), *Escherichia coli* (Ec), *Danio rerio* (Dr), *Trypanosoma brucei* (Tb), *Salmonella Typhimurium* (ST), *Leishmania mexicana* (Lm) and *Leishmania donovani* (Ld) (22). This analysis resulted in a set of 977 selected RBDs, further referred to as RBDs in the following (Supplementary Table S3).

We evaluated this selection of RBD candidates using the RBP2GO score of the proteins containing the selected RBDs in species with a higher number of proteome-wide RBP datasets such as Hs, Mm, Sc and Dm (Supplementary Figure 1). We confirmed that these proteins had a significantly higher RBP2GO score than the proteins with unselected or no RBDs, which corroborated our selection procedure (Figure 1B).

The proportion of RBD-containing proteins between RBPs and non-RBPs differs across species

Next, we analyzed the proportion of the selected RBDs in the proteins of the RBP2GO database that were classified as RBP or non-RBP candidates, further referred to as RBPs and non-RBPs. While some well-studied species such as human and mouse (Supplementary Figure S1) harbored more than 65% of the RBDs only in the RBPs and less than 10% only in the non-RBPs (Figure 2A and Supplementary Table S13), this seemed to be inverted in less studied species. In zebrafish (Dr) for example, 72% of the RBDs were found in non-RBPs only (Figure 2A and Supplementary Table S13).

This disparity between well-studied species (Hs, Mm, Sc, Dm) and less-studied species was also reflected in the proportion of RBPs as compared to RBD-containing proteins in the proteomes. The Hs proteome contained for example 3.5-fold less RBD-containing proteins than RBPs, while Dr and Tb inversely depicted more than 4.5-fold more RBD-containing proteins than RBPs (Supplementary Table S14). However, in the group of less-studied species, there were some exceptions such as Ec, which had 5-fold more RBPs (representing more than 25% of its proteome) than RBD-containing proteins (Figure 2B).

Interestingly, in the well-studied species, while the proportion of RBPs in the proteome was higher than for most of the less-studied species, only a small proportion of the RBPs contained an RBD (<10%, Figure 2C, dark orange group of proteins and Supplementary Table S14). Most RBPs did not contain an RBD, which strongly indicates that RBPs employ alternative ways in addition to already known RBDs to interact with RNA molecules.

The four most abundant RBDs in the RBPs of our dataset appeared all to be canonical RBDs (26,28), namely, the RNA recognition motif domain (RRM, IPR000504, found in 1700 proteins), Helicase superfamily 1/2 ATP-binding domain (IPR014001, found in 1079 proteins), Helicase C-terminal (IPR001650, found in 1060 proteins) and DEAD/DEAH box helicase domain (IPR011545, found in 701 proteins, Supplementary Table S4). Here again, there was a clear discrepancy between the well-studied species and less-studied species with regard to the proportion of proteins that both contained these domains and were experimentally detected as RBPs in proteome-wide screens. In the well-studied species, more than 50% of the proteins harboring at least one of these four canonical RBDs were classified as RBPs (Figure 2D and Supplementary Table S4). For many less-studied species, the percentage of RBPs dropped down to less than 15% of the proteins. Thus, the proteomes of the less-studied species may contain a number of yet undiscovered RBPs, including proteins containing selected RBDs.

In humans, 231 RBD-containing proteins had not been detected yet in the published RBP screens (Supplementary Table S14). We hypothesized that the expression levels of the proteins might impact their potential of being detected in the mass spectrometry-based studies. Therefore, we analyzed the protein expression levels (38) of the four different groups: non-RBPs without RBD, non-RBPs with

RBD, RBPs without RBDs and RBPs with RBDs. Both non-RBP groups had significantly lower expression levels (Figure 2E), which might potentially explain why the non-RBPs with RBDs may not have been detected as RBPs so far. Similarly, RBP2GO scores, which reflect the probability for a protein to be an RBP (22), positively correlated with the protein expression levels (Supplementary Figure S2).

In summary, a substantial number of non-RBPs with RBD, further referred as “RBP aspirants”, had not been detected in proteome-wide screens so far - especially in the less-studied species. Furthermore, the high proportion of RBPs without RBD, especially in the well-studied species, raises the question of how these proteins interact with RNA.

Proteins with RBDs appear as stronger RBP candidates

Taking advantage of the RBP2GO score (22), we evaluated its distribution in the different species and found that RBD-containing proteins had a significantly higher RBP2GO score in all species as compared to proteins without RBD (Figure 3A and Supplementary Figure S3A). This was confirmed when the non-RBPs and RBPs were split into two groups, depending on their RBD content. Both non-RBPs and RBPs, with RBD, had significantly higher RBP2GO scores than non-RBPs and RBPs lacking RBD (Figure 3B and Supplementary Figure S3B). For Dr and Tb, despite the significant difference, the proteins without RBD did not separate from the RBD-containing proteins as well as for the other species (Supplementary Figure S3A). This was, however, linked to the low number of proteins detected as RBPs so far in these two species (Figure 2B , Supplementary Figure S3B and Supplementary Table S14).

To obtain a more detailed view of the interplay between the presence of RBDs in proteins and the RBP2GO score, we evaluated the proportion of RBD-containing proteins for each interval of the RBP2GO score. We found a strong positive correlation between the proportion of RBD-containing proteins and the RBP2GO score up to a score of 50 (Figures 3C and Supplementary Figure S3C). Above 50, almost all proteins contained an RBD. This result supports the assumption that RBDs can be used to assess the RNA-binding potential of a candidate RBP, but also to identify new RBP candidates in the non-RBPs.

Exploiting detailed information from the InterPro database (33), we compiled the number of RBDs per protein, as well as the fraction of the protein sequence which they covered. We analyzed this data in RBPs versus non-RBPs and observed that RBPs had a higher RBD content fraction than non-RBPs, a difference that was significant for all species except Pf, Ec and Tc (Figure 3D and Supplementary Figure S3D). Moreover, in human proteins, the RBP2GO score increased with the number of RBDs per protein, whereas this was not visible for the non-RBPs (Figure 3E and Supplementary Table S14).

Taken together, these results highlight the strong positive correlation between the RBP2GO score and the number of RBDs in RBPs. This underlines the potential of RBDs for the evaluation of the

RNA-binding ability of proteins, including non-RBPs with RBDs, that might not be properly detected if lowly expressed.

The RBP-enriched Rfam IDs are also strong indicators of RNA-binding capacity

In addition to the information relative to the protein domains and repeats, the InterPro database provides annotation for protein families. In the initial lists of RNA binding-related InterPro IDs, 1028 IDs corresponded to “family” IDs (Supplementary Table S1). We subjected these IDs to the selection procedure, which we previously applied to the “domain” and “repeat” IDs (Figure 1A and Supplementary Figure S4A). We obtained a list of 672 selected RNA-related “family” IDs, referred to as Rfam IDs (Supplementary Tables S6 and Supplementary Figure S4A). The distribution of the Rfam IDs across RBPs and non-RBPs in the 13 RBP2GO species exhibited a similar pattern as for the RBDs (Figures 2A and 4A). The well-studied species displayed a large proportion of the Rfam IDs in RBPs only. For example, 98.2% of the Rfam IDs in Hs and more than 65% in Mm were found in RBPs. In less-studied species such as Tb, less than 7% of the Rfam IDs were related to RBPs only (Figure 4A and Supplementary Table S15). In Dr, 78.6% of the Rfam IDs were only present in non-RBPs. These observations confirmed the importance of a good study coverage for proper classification of proteins into RBPs or non-RBPs. Strikingly, the proportion of human RBPs containing RBDs and/or associated to Rfam IDs was only about 30%, leaving 70% of the RBPs without RNA-related InterPro ID (Supplementary Figure S4B and Supplementary Table S14).

Similar to the RBDs, the proteins annotated with an Rfam ID had a significantly higher RBP2GO score than the ones without Rfam IDs (Figure 4B and Supplementary Figure S5A). This was also verified when separating the proteins into non-RBPs and RBPs. Proteins annotated with an Rfam ID in both groups had a significantly higher RBP2GO score, except for the human non-RBPs, for which the difference was not significant mainly due to the small number of non-RBPs with Rfam ID (Figure 4C and Supplementary Figure S5B).

Similar to the RBD-containing proteins, calculating the proportion of Rfam ID-containing proteins for each interval of the RBP2GO score revealed a strong positive correlation between the RBP2GO score and the proportion of proteins associated with Rfam IDs (Figure 4D and Supplementary Figure S5C). For the well-studied species and within the proteins with high RBP2GO scores (>50), there were more proteins containing RBDs than proteins associated to Rfam IDs (Figures 3C and 4D). However, proteins with Rfam ID annotations seemed to have higher median RBP2GO score than proteins with RBDs only, *e.g.* in Hs, Mm, Dm, At, Ce and Ec (Figure 4E and Supplementary Figure S5D). In addition, RBPs annotated with both, RBDs and Rfam IDs, had a significantly higher RBP2GO score than the other RBP groups in Hs, Mm, Dm, Ce, and Ec, indicating a cumulative effect of the dual annotation (Figure 4E and Supplementary Figure S5D). The same was also observed for the non-RBP groups in Mm and Dm (Figure 4E).

Altogether, these results demonstrate that the selected Rfam IDs represent important protein features that can improve the evaluation of the RNA-binding potential of RBP candidates. While annotations with either RBDs or Rfam IDs seemed equivalent, having both annotations was linked to a higher RBP2GO score in several species, hence a greater likelihood for the protein to harbor physiologically relevant RNA-binding abilities.

Disordered regions are enriched in proteins with RBDs

While RBDs and Rfam IDs were both strong indicators of the RNA-binding potential of proteins, our analysis revealed a large group of RBP candidates (*e.g.* 70% in human) that were lacking such features (Supplementary Figure S4B, light orange group and Supplementary Table S14). In the past decades, IDRs emerged as potential RNA-binding regions and some IDRs were experimentally detected as directly interacting with RNA in proteome-wide screens (13,32). IDRs, unlike globular domains, do not fold into a specific three-dimensional structure, but rather remain lose and confer flexibility to proteins (48). Their structure can stabilize upon binding to interaction partners (49). They are implied in RNA binding for example via specific repeats such as the arginine and serine (RS) repeats in the NF-kappa-B-activating protein (NKAP), which plays an important role in splicing (30,31). Disordered regions that contain repeats of arginine and glycine, referred to as RGG-box, can also bind directly to RNA. Such IDR repeats are for example found in the fragile X mental retardation protein (FMRP/FMR1) (30,50).

To analyze the distribution of IDRs in the proteomes of the RBP2GO species, we used the data computed by the MobiDB-lite algorithm as it is more conservative, through the combination of 10 different prediction tools (35,40). The analysis of the distribution of the disordered regions in the proteins revealed that RBD-containing proteins had a significantly higher IDR fraction than proteins lacking RBDs, regardless of their status as RBP or non-RBP, in most species but especially in the well-studied ones (Figure 5A and Supplementary Figure S6). This is consistent with previous findings reporting the enrichment of disordered regions in RBPs (13,51). Surprisingly, in most species, the RBPs without RBD did not exhibit a higher disordered fraction than the RBD-containing non-RBPs, indicating that their RNA-binding ability was probably independent of the IDR content (Figure 5A and Supplementary Figure S6).

The combined analysis of the RBP2GO score of the RBPs with respect to the RBD and the IDR content revealed that the presence of IDRs in addition to RBDs only had a minor positive contribution to the increase of the RBP2GO score (Figure 5B and Supplementary Figure S7A). The same was confirmed for most of the species with respect to the RBD-containing non-RBPs. The minor contribution of the IDR to the RBP2GO score of the proteins was further validated by the weak linear correlation between the disordered fraction and the RBP2GO score for both the RBPs and the non-RBPs (Figure 5C and Supplementary Figure S7B).

According to our results, the IDRs appeared enriched in proteins (both RBPs and non-RBPs) that already contained RBDs. Therefore, the presence of IDRs in proteins seems largely redundant with the presence of RBDs and does not represent a strong independent indicator of the RNA-binding potential of proteins.

New RBP candidates can be identified from non-RBPs with RBDs

When splitting the 13 studied species into well-studied and less-studied species according to the number of available datasets (Supplementary Figure S1), the four well-studied species (Hs, Mm, Sc and Dm) summed up together 1465 non-RBPs with RBDs out of 5496 total proteins with RBDs (26.6%). For the nine less-studied species, there were 6273 non-RBPs with RBDs out of 9308 proteins with RBDs (67.3%) (Figure 6A and Supplementary Table S14). Adding the non-RBPs with RBDs to the collection of RBP candidates would increase the set of RBPs by 75% in the less-covered species as compared to only 10% in the well-studied species.

In order to evaluate the potential of the non-RBPs with RBDs as new RBP aspirants, we performed a Gene Ontology (GO) (52) enrichment analysis on the non-RBPs with RBDs in the eight species for which data was available from the PANTHER resource (53,54). We kept the terms that were enriched in at least half of the species and sorted them by the mean fold enrichment (Figure 6B and 6C). In the enriched GO terms for molecular function and biological processes, many terms were related to nucleic acid binding and metabolism such as “RNA polymerase activity” (as depicted in green in Figure 6B and 6C). In agreement with these findings, DNA-binding domains such as the high-mobility group (HMG)-box had already been associated to RNA-binding (13,55). Inversely, heterogeneous ribonucleoproteins and double-stranded RNA-binding proteins could also directly interact with DNA (56,57), illustrating that many proteins binding to DNA can also bind to RNA and *vice versa* (12,58). GO terms related to histones and chromatin were enriched, as well, but were less represented (orange, Figure 6B and 6C). RBPs can interact with chromatin and act in gene regulation, as for example the RNA-binding protein 25 (RBM25) which enhanced the transcription-dependent activities of the YY1 transcription factor (59-61). Metabolic activities also appeared to be enriched in the non-RBPs with RBDs (blue, Figure 6B and 6C), in accordance with recent studies showing that metabolic enzymes such as the iron regulatory protein 1 (IRP1), the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) or the glycolytic enzyme Enolase 1 (ENO1) were also interacting with RNA (62-65). Altogether, our GO analysis of the non-RBPs with RBDs strongly indicated that this group was enriched in molecular functions and biological processes reasonably linked to RNA such that this group may contain further so far experimentally unrecognized RBPs, referred to as RBP aspirants.

To further validate these findings, we screened the literature and found evidence for the RNA-binding properties of seven RBD-containing non-RBPs, which contained different types of RBDs (Supplementary Table S16). Individual studies experimentally confirmed the binding of these proteins to RNA (66-71), in support of our assumption that RBP aspirants (non-RBPs with RBDs) are strong

RBP candidates. Furthermore, six out of these seven proteins exhibited disordered regions, highlighting the frequent association of IDRs and RBDs (Supplementary Table S16) as suggested by the enrichment of IDRs in RBPs with canonical RBDs (72) and further confirmed in our pan-species analysis of RBPs and non-RBPs (Figure 5A).

We demonstrated here that a substantial number of RBD-containing non-RBPs, herein called RBP aspirants, are probably RBPs that had not been detected in proteome-wide screens so far. Using our selected list of RBDs as indicator of their RBP potential allowed us to identify 7738 new RBP aspirants, with more than 80% of them in less-studied species.

New RBDs can be predicted in RBPs without known RBDs and high RBP2GO score

Since a large proportion of RBP candidates did not contain any of the selected RBDs, we hypothesized that they may contain so far undetected RBDs. To test this hypothesis, we performed an enrichment analysis on the “domain” and “repeat” InterPro IDs present in the top 20% of the human RBPs lacking RBDs based on their RBP2GO score. This group of 935 proteins was compared to three different reference sets: the bottom 20% of the human RBPs based on their RBP2GO score, the non-RBPs with no RBD and the proteins of the whole human proteome (Figure 7A and Supplementary Table S4). 15 domains were significantly enriched in the top 20% RBPs without RBD as compared to at least one reference datasets. Within these domains, we found domains related to chaperones and importins, along with domains related to the binding of cytoskeletal proteins (Figure 7B and Supplementary Table S8). In particular, the WD40 repeats domain (IPR001680) does not exhibit a role in a specific cellular process, but is found in RBPs such as the WD-repeat containing protein 43 (WDR43) (73). The histidine kinase/HSP90-like ATPase domain (IPR003594) is present in the HSP90A protein which binds to RNA and is implicated in the loading of small RNAs into the RISC complexes (74). These findings could corroborate the RNA-binding ability of the 15 newly identified RBD candidates.

To validate the 15 domains as RBDs, we quantified the proportion of RBPs and non-RBPs containing these domains. In the well-studied species, there were only a small number of RBPs with new RBDs as compared to RBPs with selected RBDs (Figure 7C). In all four well-studied species, the proteins with new RBD candidates had a significantly higher RBP2GO score than RBPs without RBD. However, the RBP2GO score of the former group was still significantly lower than that of the proteins with selected RBDs (Figure 7D). In all four species, proteins containing only new RBD candidates were significantly enriched in RBPs as compared to non-RBPs (Figure 7E). These results supported the hypothesis that the 15 new RBD candidates can be true RBDs.

To further validate the newly identified RBD candidates, we took advantage of RNA-binding peptide datasets available from human proteome-wide RBP studies (13,16,32,44). We calculated the proportion of these peptides overlapping the newly discovered RBDs, normalized it and compared it to the normalized proportion of peptides overlapping with previously selected RBDs on the one hand

or with the RBDs that were not selected on the other hand (see Material and Methods for further information). The median proportion of peptides for the newly discovered RBDs appeared significantly higher than for the non-selected ones and not significantly different as compared to the selected RBDs (Figure 7F). For example, RNA-binding peptides were identified only in the Armadillo repeat (IPR000225), one of the newly identified RBDs, for three different proteins from RBP screens (Figure 7G).

In summary, we found 15 new RBD candidates in the human RBPs of our dataset that did not contain any RBD so far. We validated them using both the RBP2GO score of the related proteins as well as published experimental data. Consequently, we included them into the list of RBDs that can now be explored on the updated RBP2GO database (https://rbp2go.dkfz.de).

A new RBP2GO composite score supports defining a high-confidence human RBP list

In order to integrate the knowledge gained from the analysis of the RBDs and Rfam IDs, we computed a new RBP2GO composite score for each protein. This score combines three components that are based on experimental data directly related to the protein (component 1) and the top 10 interacting proteins according to the STRING database (component 2), as well as data related to RBDs and Rfam IDs (component 3, Figure 8A and Supplementary Table S10). Refer to the Material and Methods section for a detailed description. As expected, per definition, the first two components of the RBP2GO composite score strongly correlated with each other and with the RBP2GO score, but the correlation was decreased with the component 3 indicating that it added another layer of information (Supplementary Figure S8). The RBP2GO composite score allowed a better discrimination between proteins with and without RBDs (Figure 8B, Supplementary Table S12), providing a solid support for the evaluation of RBP candidates. Moreover, while the previous RBP2GO score was only available for 9 out of 13 RBP2GO species due to a lack of study coverage (22), the component 3 of the RBP2GO composite score can be provided for all species. Finally, we took advantage of the RBP2GO composite score to define a high-confidence list of human RBPs that contains 2018 proteins with an RBP2GO composite score >10 and with at least two components > 0 (Supplementary Table S11).

An advanced version of the RBP2GO database includes detailed knowledge on RBDs and Rfam IDs

Our analysis demonstrated that RBDs and Rfam IDs were strong indicators of the RNA-binding ability of the related proteins. Therefore, we expanded the RBP2GO database (https://rbp2go.dkfz.de) to allow the users to browse the information related to RBDs and Rfam IDs in an intuitive manner as well as to use these as search or filter criteria. Therefore, we added a dedicated section which contains a search tool to display the RNA-binding status of InterPro IDs to the sidebar (Figure 8C). We also

implemented a tab panel in the species-specific “Protein search” to provide the available information pertaining to the InterPro ID annotations as well as the RNA-binding properties of each domain, the presence of disordered regions in the proteins and the RBP2GO composite score (Figure 8D). Protein-specific links to the InterPro and MobiDB databases are provided to simplify access to complementary data if needed. Finally, we generated several options in the “Advanced search” for all the species. The users can navigate through the proteins of the database using a list of specific InterPro IDs and can specifically select proteins that contain RBDs and/or Rfam IDs.

DISCUSSION

RBPs are important interaction partners for RNA transcripts from the sites of RNA production until their degradation. It is evident that RNA-protein interactions are relevant in numerous key cellular processes. Thus, defects in RBPs are often associated with severe disease phenotypes such as muscular dystrophies, neurological disorders or cancers (3,75). Consequently, it is essential to investigate the composition of ribonucleoprotein complexes in order to understand the molecular mechanisms and functions governed by RNA-protein interactions.

In this study, we present a detailed analysis of RNA-binding features in RBP candidates from 13 species including datasets generated in both, eukaryotes and prokaryotes. Based on experimental knowledge on RBP candidates and non-RBPs, we compiled a list of selected RBDs and Rfam IDs. Using the previously defined RBP2GO score (22), we show that the presence of RBDs in proteins or their association with Rfam IDs particularly highlights their RNA-binding potential. The number of RBDs per RBP also shows a positive correlation with the RBP2GO score in human RBPs, which is supported by the modularity of canonical RBDs, with several of them often working together within the same protein to more efficiently bind to the target RNAs (26,76).

The distribution of the RBDs between RBPs and non-RBPs in the different species, as well as the higher proportion of non-RBPs with RBDs in the less-studied species, confirms the importance of a good study coverage for the proper classification of the proteins into RBPs and non-RBPs. One notable exception to this statement is *Escherichia coli*, which had only 48 non-RBPs with RBD despite a low study coverage with only three RBP screens (16,77,78). This unique situation for Ec may be explained by a significantly lower number of RBD-containing proteins as compared to other species. In the Ec dataset, there are only 247 RBD-containing proteins in total as compared to 1938 proteins in the Dr dataset with also three RBP screens. The difference may to be due to the size of the respective proteomes (4522 proteins for Ec versus 25801 for Dr). However, almost 80% of the RBD-containing proteins in Ec have been identified (Supplementary Table S17), as compared to only 11% in Dr (Supplementary Table S4). The high RBP identification rate in Ec seems to be due to one study (77), which detected 1107 RBP candidates in Ec, using a method based on silica beads purification. These RBP candidates cover 98% of the RBD-containing RBPs (185 out of 189) and 78% of the RBD-containing proteins in general (185 out of 237). The same study identified 1448 RBP candidates

in Sc, that cover 73% of the RBD-containing RBPs (438 out of 604) in this species (Supplementary Table S17). The RBP candidates from this study, that do not contain RBDs, will nonetheless need further experimental validation.

Using the list of selected RBDs, we extended the pool of RBP candidates by identifying the non-RBPs that contained selected RBDs in all the species of this study. This added 1465 RBP aspirants to the well-studied species and 6273 RBP aspirants to the less-studied species. These were validated both by a GO term enrichment analysis and a literature search. We noticed that the RBP aspirants had a significantly higher RBP2GO score than the non-RBPs lacking RNA-related feature. As non-RBPs, per definition, the only contribution to their RBP2GO score is made by their interaction partners (22). Hence, RBP aspirants seem to interact more with RBPs than the non-RBPs lacking any RNA-related annotation. This is in agreement with the previous findings reporting that proteins interacting with RBPs have a high probability to be RBPs themselves (19,25). Finally, we observed that the human non-RBPs were significantly less expressed than the RBPs, including the RBP aspirants. A lower expression level can contribute to the difficulty to capture such proteins in large mass spectrometry-based studies, which can explain why some human RBD-containing proteins had not been detected as RBPs yet, despite more than 40 published datasets (22).

Our analysis of IDRs demonstrated that, among RBPs, the RBD-containing proteins were enriched in IDRs as compared to the proteins lacking RBDs, which is in agreement with previous studies (72,79,80). However, our data also showed an enrichment in IDRs for non-RBPs containing RBDs as compared to the ones lacking RBDs. Since IDRs confer flexibility to a protein, they can link different domains, such as RBDs as documented for the RNA-binding protein FUS between its RRMs (81). IDRs can also collaborate with canonical RBDs for the recognition of RNA and participate in the specificity of RBPs (82). Based on our current analysis, it seems that IDRs alone are not sufficient to explain the RNA-binding properties of a large number of RBPs lacking RBDs. However, specific types of disordered regions such as arginine/serine (RS) repeats or RGG-boxes can bind to RNA (30,64,83). Presently, the MobiDB-lite algorithm, which we used, provides IDR subtypes based on the enrichment in specific amino acids and charge, but not on specific repeats (35,84). Thus, future options to further investigate the impact of IDRs in RNA-protein interactions could be to include a detailed analysis of the different repeat subtypes. Altogether, the question of how RBPs lacking RBDs are interacting with RNA remains unsolved and of great interest for future investigation, especially for proteins with so far unexpected affinity for RNA and functionally “riboregulated” (65).

Next, we also took advantage of our comprehensive dataset to identify 15 new RBDs within RBPs lacking selected RBDs but with high RBP2GO scores. Among them, some were found in proteins already known to bind to RNA, such as the histidine kinase/HSP90-like ATPase domain (IPR003594) present in HSP90A (74). The new RBDs were validated using experimental datasets describing RNA-bound protein peptides (13,16,32,44), demonstrating the utility of computational meta-analyses as orthogonal approach besides experimental studies.

Last but not least, we generated the RBP2GO composite score, which integrates both the experimental and sequence-based RNA-related features of the proteins. We applied the RBP2GO composite score to compute a high-confidence list of 2018 RBPs, thereby answering the need for a comprehensive catalog of human RBPs (Supplementary Table S11), which will help selecting candidates for further validation and characterization. In the advanced version of the RBP2GO database, the component of the RBP2GO composite score that is based on the quality factors of the related InterPro IDs was implemented also for species currently lacking an RBP2GO score. We believe that this additional information will facilitate the selection of RBP candidates for the species with currently low study coverage.

In summary, this study first provides a pan-species list of selected RBDs and Rfam IDs enriched in RBPs and then exploits this resource (i) to refine the list of RBPs from proteome-wide studies with higher confidence, (ii) to predict new RBP aspirants from the group of RBD-containing non-RBPs and (iii) to identify and validate new RBDs from the large group of RBPs lacking selected RBDs. The updated resource is freely available online (https://rbp2go.dkfz.de) and can be explored in an interactive way. We anticipate that this analysis and the compilation of the information in a unifying platform will stimulate our understanding of RBPs, their functions in key cellular processes and their implications in human diseases.

DATA AVAILABILITY

RBP2GO is available at https://rbp2go.dkfz.de.

The scripts for the main analysis, the production of the figures and the RBP2GO Shiny App are directly available on the RBP2GO database under the DOWNLOAD menu from the sidebar.

SUPPLEMENTARY DATA

Supplementary data comprises the following items:

Supplementary Figure S1, related to Figure 1. Plot of the number of published proteome-wide RBP screens per species. The vertical dashed line arbitrarily separates the well-studied species (on the left) from the less-studied species (on the right, in the gray area).

Supplementary Figure S2, related to Figure 2. Scatterplot of the expression of human proteins in HeLa cells versus their RBP2GO score. R represents the Pearson's correlation coefficient. The data used to generate this figure can be found in the Supplementary Table S5.

Supplementary Figure S3, related to Figure 3 Analysis of the distribution of the RBDs in the proteins of the RBP2GO database species (extended to all species). (**A**) Boxplots representing the distribution of the RBP2GO score in proteins without RBD (no RBD, light grey) and with RBD (RBD, dark grey). (**B**) Same as in (**A**), but separated into non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light orange) and RBPs with RBD (dark orange). (**C**) Proportion of RBD-containing proteins for each interval of the RBP2GO score. The red line represents a linear regression (up to a RBP2GO score of 50). R represents the Pearson's correlation coefficient. (**D**) Boxplots representing the distribution of the number of RBDs per protein in non-RBPs (blue) and RBPs (orange). *, **, *** and **** correspond to *P*-values < 0.05, 0.01, 0.001 and 0.0001, respectively, ns corresponds to non-significant, as resulting from a Wilcoxon rank sum test.

Supplementary Figure S4, related to Figure 4. Selection the Rfam IDs. (**A**) Flow chart showing the selection process of the Rfam ID candidates and depicting the three-step selection procedure based on the ratios between RBPs and non-RBPs in the species available in the RBP2GO database (22). The starting lists of Rfam ID candidates (InterPro type = Family) and the final list of selected Rfam IDs are found in Supplementary Tables S1 and S6, respectively. (**B**) Proportion in % of non-RBPs without RBD nor Rfam ID (light blue), non-RBPs with RBD but without Rfam ID (dark blue), non-RBPs without RBD but with Rfam ID (light green), non-RBPs with RBD and Rfam ID (dark green), RBPs without RBD nor Rfam ID (light orange), RBPs with RBD but without Rfam ID (dark orange), RBPs without RBD but with Rfam ID (light purple) and RBPs with RBD and Rfam ID (dark purple) in the different species.

Supplementary Figure S5, related to Figure 4. Analysis of the selected Rfam IDs in the proteins of the RBP2GO database species (extended to all species). (**A**) Boxplots representing the distribution of the RBP2GO score in proteins without Rfam ID (no Rfam ID, light grey) and with Rfam ID (Rfam ID, dark grey). (**B**) Same as in (**A**), but separated into non-RBPs without Rfam ID (light green), non-RBPs with Rfam ID (dark green), RBPs without Rfam ID (light purple) and RBPs with Rfam ID (dark purple). (**C**) Proportion of Rfam ID-containing proteins (in %) for each interval of the RBP2GO score. The red line represents a linear regression (up to a RBP2GO score of 50). R represents the Pearson's correlation coefficient. (**D**) Same as in (**A**), but separated between non-RBPs without RBD nor Rfam ID (light blue), non-RBPs with RBD but without Rfam ID (dark blue), non-RBPs without RBD but with Rfam ID (light green), non-RBPs with RBD and with Rfam ID (dark green), RBPs without RBD nor Rfam ID (light orange), RBPs with RBD but without Rfam ID (dark orange), RBPs without RBD but with Rfam ID (light purple) and RBPs with RBD and with Rfam ID (dark purple). *, **, *** and **** correspond to *P*-values < 0.05, 0.01, 0.001 and 0.0001, respectively, ns corresponds to non-significant, as resulting from a Wilcoxon rank sum test.

Supplementary Figure S6, related to Figure 5. Analysis of the disordered fraction in the proteins of the RBP2GO database species, part 1 (extended to all species). Boxplots representing the disordered fraction in proteins classified in the following four groups: non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light orange) and RBPs with RBD (dark orange). The proteins containing no disordered region were also included. The numbers given in grey are the number of proteins in each group. *, **, *** and **** correspond to *P*-values < 0.05, 0.01, 0.001 and 0.0001, respectively, ns corresponds to non-significant, as resulting from a Wilcoxon rank sum test.

Supplementary Figure S7, related to Figure 5. Analysis of the disordered fraction in the proteins of the RBP2GO database species, part 2 (extended to all species). (**A**) Boxplots representing the distribution of the RBP2GO score in eight groups of proteins: non-RBPs without RBD nor IDR (light green), non-RBPs without RBD but with IDR (light blue), non-RBPs with RBD but without IDR (dark green), non-RBPs with RBD and IDR (dark blue), RBPs without RBD nor IDR (light orange), RBPs without RBD but with IDR (light red), RBPs with RBD but without IDR (dark orange) and RBPs with RBD and IDR (dark red). *, **, *** and **** correspond to *P*-values < 0.05, 0.01, 0.001 and 0.0001, respectively, ns corresponds to non-significant, as resulting from a Wilcoxon rank sum test. (**B**) Boxplots representing the distribution of the RBP2GO score for non-RBPs (blue) and RBPs (orange) in function of the disordered fraction of the proteins. R represents the Pearson's correlation coefficient.

Supplementary Figure S8, related to Figure 8. Pearson's correlation coefficients between the three components of the RBP2GO composite score and the RBP2GO score. The coefficients are depicted from blue (low) to red (high).

Supplementary Table S1, related to Figure 1. List of the candidate RNA-related InterPro IDs based on literature and database catalogs.

Supplementary Table S2, related to Figure 1. Size of the proteomes of the RBP2GO-listed species.

Supplementary Table S3, related to Figure 1. List of selected RBDs.

Supplementary Table S4, related to Figures 2 to 5. List of all proteins of the RBP2GO database with information on RBDs, RNA-related family IDs and disorder.

Supplementary Table S5, related to Supplementary Figure S2. Expression of human proteins in HeLa cells, downloaded from (https://www.ebi.ac.uk/gxa/experiments/E-PROT-19/Results).

Supplementary Table S6, related to Supplementary Figure S4. List of selected RNA-related family IDs (Rfam IDs).

Supplementary Table S7, related to Figure 6. Results of the GO term enrichment analysis in the non-RBPs with RBD. If the GO terms are not related to one of the species, the fold-enrichment and the FDR are empty.

Supplementary Table S8, related to Figure 7. List of newly discovered RBDs and their enrichment (three last columns TRUE).

Supplementary Table S9, related to Figure 7. List of peptides and domains used for the validation of the new RBDs.

Supplementary Table S10, related to Figure 8. Attribution rules of quality factors to RBDs, Rfam IDs, and list of the IDs with their quality factors.

Supplementary Table S11, related to Figure 8. List of high-confidence human RBPs.

Supplementary Table S12, related to Figure 8B. List of the updated proteins of the RBP2GO database with the updated RBDs, RBP2GO score and RBP2GO composite score.

Supplementary Table S13, related to Figure 2A. Distribution of RBDs in the RBP2GO species into the three protein categories: RBPs only, non-RBPs only, and both (RBPs and non-RBPs).

Supplementary Table S14, related to Figure 3 to 5 and Supplementary Figures S3 to S7. Left (columns A:P): distribution of RBDs and Rfam IDs in the RBP2GO proteins. Right (columns T:AI): distribution of the RBDs and IDRs in the RBP2GO proteins.

Supplementary Table S15, related to Figure 4. Distribution of Rfam IDs in the RBP2GO species into the three protein categories: RBPs only, non-RBPs only, and both (RBPs and non-RBPs).

Supplementary Table S16, related to Figure 6. List of non-RBPs with RBD that were validated as RBPs by literature search.

Supplementary Table S17, related to the Discussion. List of RBPs detected in the Ec (top table) and Sc (bottom table) proteome-wide studies. The table also include the total number of RBPs detected in Ec and Sc, as well as the total number of RBD-containing proteins (highlighted in yellow).

AUTHOR CONTRIBUTIONS

M.C.-H. conceived the study and M.C.-H. and S.D. designed and supervised the study and the analysis. E.W. and G.K. performed the analysis of the RBDs. E.W. performed the analysis of the Rfam IDs, the new RBPs, the new RBDs and their validation and generated the high-confidence list of human RBPs. E.W. produced the figures, the supplementary tables and implemented the update of the RBP2GO database. E.W. wrote the manuscript, S.D. and M.C.-H. edited the manuscript and all authors have seen and approved the final manuscript.

ACKNOWLEDGEMENT

The authors sincerely acknowledge the IT core facility of the German Cancer Research Center for their precious support and help in deploying the new version of the RBP2GO database. The authors also deeply thank the lab members, collaborators and users for the feedback on the RBP2GO database, figures and manuscript.

FUNDING

Research on RNA–protein complexes in our lab is supported by the German Cancer Aid [70113919 to S.D.]; Baden-Württemberg Stiftung [BWF ISF2019-027 to M.C.-H.]. Funding for open access charge: DKFZ Core Funding.

Conflict of interest statement. S.D. is co-owner of siTOOLs Biotech, Martinsried, Germany, without relation to this work. The other authors disclose no conflicts of interest. This study is part of the Ph.D. thesis of E.W.

REFERENCES

1. Kishore, S., Luber, S. and Zavolan, M. (2010) Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Brief Funct Genomics*, **9**, 391-404.
2. Castello, A., Fischer, B., Hentze, M.W. and Preiss, T. (2013) RNA-binding proteins in Mendelian disease. *Trends Genet*, **29**, 318-327.
3. Gebauer, F., Schwarzl, T., Valcarcel, J. and Hentze, M.W. (2021) RNA-binding proteins in human genetic disease. *Nat Rev Genet*, **22**, 185-198.
4. 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.
5. Klingenberg, M., Gross, M., Goyal, A., Polycarpou-Schwarz, M., Miersch, T., Ernst, A.S., Leupold, J., Patil, N., Warnken, U., Allgayer, H. *et al.* (2018) The Long Noncoding RNA Cancer Susceptibility 9 and RNA Binding Protein Heterogeneous Nuclear Ribonucleoprotein L Form a Complex and Coregulate Genes Linked to AKT Signaling. *Hepatology*, **68**, 1817-1832.
6. Wurth, L. and Gebauer, F. (2015) RNA-binding proteins, multifaceted translational regulators in cancer. *Biochim Biophys Acta*, **1849**, 881-886.
7. Baltz, A.G., Munschauer, M., Schwanhausser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M. *et al.* (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell*, **46**, 674-690.
8. 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.
9. Beckmann, B.M., Horos, R., Fischer, B., Castello, A., Eichelbaum, K., Alleaume, A.M., Schwarzl, T., Curk, T., Foehr, S., Huber, W. *et al.* (2015) The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat Commun*, **6**, 10127.

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

11. Perez-Perri, J.I., Rogell, B., Schwarzl, T., Stein, F., Zhou, Y., Rettel, M., Brosig, A. and Hentze, M.W. (2018) Discovery of RNA-binding proteins and characterization of their dynamic responses by enhanced RNA interactome capture. *Nat Commun*, **9**, 4408.

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

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

14. Liao, Y., Castello, A., Fischer, B., Leicht, S., Foehr, S., Frese, C.K., Ragan, C., Kurscheid, S., Pagler, E., Yang, H. *et al.* (2016) The Cardiomyocyte RNA-Binding Proteome: Links to Intermediary Metabolism and Heart Disease. *Cell Rep*, **16**, 1456-1469.

15. Liepelt, A., Naarmann-de Vries, I.S., Simons, N., Eichelbaum, K., Fohr, S., Archer, S.K., Castello, A., Usadel, B., Krijgsveld, J., Preiss, T. *et al.* (2016) Identification of RNA-binding Proteins in Macrophages by Interactome Capture. *Mol Cell Proteomics*, **15**, 2699-2714.

16. Queiroz, R.M.L., Smith, T., Villanueva, E., Marti-Solano, M., Monti, M., Pizzinga, M., Mirea, D.M., Ramakrishna, M., Harvey, R.F., Dezi, V. *et al.* (2019) Comprehensive identification of RNA-protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat Biotechnol*, **37**, 169-178.

17. Trendel, J., Schwarzl, T., Horos, R., Prakash, A., Bateman, A., Hentze, M.W. and Krijgsveld, J. (2019) The Human RNA-Binding Proteome and Its Dynamics during Translational Arrest. *Cell*, **176**, 391-403 e319.

18. Urdaneta, E.C., Vieira-Vieira, C.H., Hick, T., Wessels, H.H., Figini, D., Moschall, R., Medenbach, J., Ohler, U., Granneman, S., Selbach, M. *et al.* (2019) Purification of cross-linked RNA-protein complexes by phenol-toluol extraction. *Nat Commun*, **10**, 990.

19. Caudron-Herger, M., Rusin, S.F., Adamo, M.E., Seiler, J., Schmid, V.K., Barreau, E., Kettenbach, A.N. and Diederichs, S. (2019) R-DeeP: Proteome-wide and Quantitative Identification of RNA-Dependent Proteins by Density Gradient Ultracentrifugation. *Mol Cell*, **75**, 184-199 e110.

20. Caudron-Herger, M., Wassmer, E., Nasa, I., Schultz, A.S., Seiler, J., Kettenbach, A.N. and Diederichs, S. (2020) Identification, quantification and bioinformatic analysis of RNA-dependent proteins by RNase treatment and density gradient ultracentrifugation using R-DeeP. *Nat Protoc*, **15**, 1338-1370.

21. Rajagopal, V., Loubal, A.S., Engel, N., Wassmer, E., Seiler, J., Schilling, O., Caudron-Herger, M. and Diederichs, S. (2022) Proteome-Wide Identification of RNA-Dependent Proteins in Lung Cancer Cells. *Cancers (Basel)*, **14**, 6109.

22. 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.

23. Despic, V., Dejung, M., Gu, M., Krishnan, J., Zhang, J., Herzel, L., Straube, K., Gerstein, M.B., Butter, F. and Neugebauer, K.M. (2017) Dynamic RNA-protein interactions underlie the zebrafish maternal-to-zygotic transition. *Genome Res*, **27**, 1184-1194.

24. Corley, M. (2020) How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Molecular Cell*, **78**, 9-29.

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

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

27. Maris, C., Dominguez, C. and Allain, F.H. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J*, **272**, 2118-2131.

28. 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**, D301-308.

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

30. Jarvelin, A.I., Noerenberg, M., Davis, I. and Castello, A. (2016) The new (dis)order in RNA regulation. *Cell Commun Signal*, **14**, 9.

31. Burgute, B.D., Peche, V.S., Steckelberg, A.L., Glockner, G., Gassen, B., Gehring, N.H. and Noegel, A.A. (2014) NKAP is a novel RS-related protein that interacts with RNA and RNA binding proteins. *Nucleic Acids Res*, **42**, 3177-3193.

32. Panhale, A., Richter, F.M., Ramirez, F., Shvedunova, M., Manke, T., Mittler, G. and Akhtar, A. (2019) CAPRI enables comparison of evolutionarily conserved RNA interacting regions. *Nat Commun*, **10**, 2682.

33. Blum, M., Chang, H.Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S. *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*, **49**, D344-D354.

34. Necci, M., Piovesan, D., Predictors, C., DisProt, C. and Tosatto, S.C.E. (2021) Critical assessment of protein intrinsic disorder prediction. *Nat Methods*, **18**, 472-481.

35. Necci, M., Piovesan, D., Clementel, D., Dosztanyi, Z. and Tosatto, S.C.E. (2021) MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavors in proteins. *Bioinformatics*, **36**, 5533-5534.

36. 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.

37. UniProt, C. (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*, **51**, D523-D531.

38. Bekker-Jensen, D.B., Kelstrup, C.D., Batth, T.S., Larsen, S.C., Haldrup, C., Bramsen, J.B., Sorensen, K.D., Hoyer, S., Orntoft, T.F., Andersen, C.L. *et al.* (2017) An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst*, **4**, 587-599 e584.

39. Riemondy, K.A., Sheridan, R.M., Gillen, A., Yu, Y., Bennett, C.G. and Hesselberth, J.R. (2017) valr: Reproducible genome interval analysis in R. *F1000Res*, **6**, 1025.

40. Piovesan, D., Necci, M., Escobedo, N., Monzon, A.M., Hatos, A., Micetic, I., Quaglia, F., Paladin, L., Ramasamy, P., Dosztanyi, Z. *et al.* (2021) MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res*, **49**, D361-D367.

41. Rezwani, M., Pourfathollah, A.A. and Noorbakhsh, F. (2022) rbioapi: user-friendly R interface to biologic web services' API. *Bioinformatics*, **38**, 2952-2953.

42. Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

43. Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289-300.

44. Mullari, M., Lyon, D., Jensen, L.J. and Nielsen, M.L. (2017) Specifying RNA-Binding Regions in Proteins by Peptide Cross-Linking and Affinity Purification. *J Proteome Res*, **16**, 2762-2772.

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

46. Reichel, M., Liao, Y., Rettel, M., Ragan, C., Evers, M., Alleaume, A.M., Horos, R., Hentze, M.W., Preiss, T. and Millar, A.A. (2016) In Planta Determination of the mRNA-Binding Proteome of Arabidopsis Etiolated Seedlings. *Plant Cell*, **28**, 2435-2452.

47. Kassambara, A. (2023) ggpubr: 'ggplot2' Based Publication Ready Plots.

48. Dunker, A.K., Babu, M.M., Barbar, E., Blackledge, M., Bondos, S.E., Dosztanyi, Z., Dyson, H.J., Forman-Kay, J., Fuxreiter, M., Gsponer, J. *et al.* (2013) What's in a name? Why these

proteins are intrinsically disordered: Why these proteins are intrinsically disordered. *Intrinsically Disord Proteins*, **1**, e24157.

49. Bhattacherjee, A. and Wallin, S. (2012) Coupled folding-binding in a hydrophobic/polar protein model: impact of synergistic folding and disordered flanks. *Biophys J*, **102**, 569-578.

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

51. Wang, K., Li, L., Fu, L., Yuan, Y., Dai, H., Zhu, T., Zhou, Y. and Yuan, F. (2019) Integrated Bioinformatics Analysis the Function of RNA Binding Proteins (RBPs) and Their Prognostic Value in Breast Cancer. *Front Pharmacol*, **10**, 140.

52. Gene Ontology, C. (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res*, **49**, D325-D334.

53. Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X. and Thomas, P.D. (2019) Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc*, **14**, 703-721.

54. Thomas, P.D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L.P. and Mi, H. (2022) PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci*, **31**, 8-22.

55. Hamilton, D.J., Hein, A.E., Holmes, Z.E., Wuttke, D.S. and Batey, R.T. (2022) The DNA-Binding High-Mobility Group Box Domain of Sox Family Proteins Directly Interacts with RNA In Vitro. *Biochemistry*, e1778.

56. 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.

57. Fierro-Monti, I. and Mathews, M.B. (2000) Proteins binding to duplexed RNA: one motif, multiple functions. *Trends Biochem Sci*, **25**, 241-246.

58. D, G.H., Kelley, D.R., Tenen, D., Bernstein, B. and Rinn, J.L. (2016) Widespread RNA binding by chromatin-associated proteins. *Genome Biol*, **17**, 28.

59. 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 e118.

60. Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172-177.

61. Chen, J.Y., Lim, D.H. and Fu, X.D. (2019) Mechanistic Dissection of RNA-Binding Proteins in Regulated Gene Expression at Chromatin Levels. *Cold Spring Harb Symp Quant Biol*, **84**, 55-66.

62. Castello, A., Hentze, M.W. and Preiss, T. (2015) Metabolic Enzymes Enjoying New Partnerships as RNA-Binding Proteins. *Trends Endocrinol Metab*, **26**, 746-757.

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

64. Balcerak, A., Trebinska-Stryjewska, A., Konopinski, R., Wakula, M. and Grzybowska, E.A. (2019) RNA-protein interactions: disorder, moonlighting and junk contribute to eukaryotic complexity. *Open Biol*, **9**, 190096.

65. 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 e2611.

66. Zeng, M., Sun, H., Chen, S., Wang, X., Yang, Y., Liu, Y., Tao, D., Yang, Z., Zhang, S. and Ma, Y. (2008) Identification of target messenger RNA substrates for mouse RBMY. *Mol Hum Reprod*, **14**, 331-336.

67. Gonatopoulos-Pournatzis, T., Wu, M., Braunschweig, U., Roth, J., Han, H., Best, A.J., Raj, B., Aregger, M., O'Hanlon, D., Ellis, J.D. *et al.* (2018) Genome-wide CRISPR-Cas9 Interrogation of Splicing Networks Reveals a Mechanism for Recognition of Autism-Misregulated Neuronal Microexons. *Mol Cell*, **72**, 510-524 e512.

68. Shen, H., Zheng, X., Luecke, S. and Green, M.R. (2010) The U2AF35-related protein Urp contacts the 39 splice site to promote U12-type intron splicing and the second step of U2-type intron splicing. *Genes & Dev.*, **24**, 2389-2394.

69. Dittmar, K.A., Jiang, P., Park, J.W., Amirikian, K., Wan, J., Shen, S., Xing, Y. and Carstens, R.P. (2012) Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. *Mol Cell Biol*, **32**, 1468-1482.

70. Morgan, M., Much, C., DiGiacomo, M., Azzi, C., Ivanova, I., Vitsios, D.M., Pistolic, J., Collier, P., Moreira, P.N., Benes, V. *et al.* (2017) mRNA 3' uridylation and poly(A) tail length sculpt the mammalian maternal transcriptome. *Nature*, **548**, 347-351.

71. Ansel, K.M., Pastor, W.A., Rath, N., Lapan, A.D., Glasmacher, E., Wolf, C., Smith, L.C., Papadopoulou, N., Lamperti, E.D., Tahiliani, M. *et al.* (2008) Mouse Eri1 interacts with the ribosome and catalyzes 5.8S rRNA processing. *Nat Struct Mol Biol*, **15**, 523-530.

72. Neelamraju, Y., Hashemikhabir, S. and Janga, S.C. (2015) The human RBPome: from genes and proteins to human disease. *J Proteomics*, **127**, 61-70.

73. Jain, B.P. and Pandey, S. (2018) WD40 Repeat Proteins: Signalling Scaffold with Diverse Functions. *Protein J*, **37**, 391-406.

74. Iwasaki, S., Kobayashi, M., Yoda, M., Sakaguchi, Y., Katsuma, S., Suzuki, T. and Tomari, Y. (2010) Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small RNA duplexes. *Mol Cell*, **39**, 292-299.

75. Pereira, B., Billaud, M. and Almeida, R. (2017) RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends Cancer*, **3**, 506-528.

76. Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y. and Yokoyama, S. (1999) Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. *Nature*, **398**, 579-585.

77. Shchepachev, V., Bresson, S., Spanos, C., Petfalski, E., Fischer, L., Rappsilber, J. and Tollervey, D. (2019) Defining the RNA interactome by total RNA-associated protein purification. *Mol Syst Biol*, **15**, e8689.

78. The Gene Ontology, C. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*, **47**, D330-D338.

79. Wang, C., Uversky, V.N. and Kurgan, L. (2016) Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics*, **16**, 1486-1498.

80. Ho, W.L. and Huang, J.R. (2022) The return of the rings: Evolutionary convergence of aromatic residues in the intrinsically disordered regions of RNA-binding proteins for liquid-liquid phase separation. *Protein Sci*, **31**, e4317.

81. Loughlin, F.E., Lukavsky, P.J., Kazeeva, T., Reber, S., Hock, E.M., Colombo, M., Von Schroetter, C., Pauli, P., Clery, A., Muhlemann, O. *et al.* (2019) The Solution Structure of FUS Bound to RNA Reveals a Bipartite Mode of RNA Recognition with Both Sequence and Shape Specificity. *Mol Cell*, **73**, 490-504 e496.

82. Varadi, M., Zsolyomi, F., Guharoy, M. and Tompa, P. (2015) Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PLoS One*, **10**, e0139731.

83. Zeke, A., Schad, E., Horvath, T., Abukhairan, R., Szabo, B. and Tantos, A. (2022) Deep structural insights into RNA-binding disordered protein regions. *Wiley Interdiscip Rev RNA*, **13**, e1714.

84. Das, R.K. and Pappu, R.V. (2013) Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci U S A*, **110**, 13392-13397.

FIGURE LEGENDS

Figure 1. Selection of the RNA-binding domains. (**A**) Flow-chart showing the selection process of the RBD candidates and depicting the three-step selection procedure based on the ratios between RBPs and non-RBPs in the species available in the RBP2GO database (22). The starting lists of RBD candidates and the final list of selected RBDs are found in Supplementary Table S1 and Supplementary Table S3, respectively. (**B**) Boxplots depicting the distribution of the RBP2GO score of the proteins from the groups with no candidate RBDs (grey, no RBD), unselected candidate RBDs (orange) and selected RBDs (green) in Hs (*Homo sapiens*), Mm (*Mus musculus*), Sc (*Saccharomyces cerevisiae*) and Dm (*Drosophila melanogaster*). The numbers below the boxplots indicate the size of the groups. *, ** and *** correspond to *P*-values < 0.05, 0.01 and 0.001, respectively, as resulting from a Wilcoxon rank sum test.

Figure 2. Distribution of the selected RBDs in the proteins of the RBP2GO database. (**A**) Number of selected RBDs present in each species and grouped according to the presence of the RBD either only in RBPs (orange), in both RBPs and non-RBPs (gray) or only in non-RBPs (blue). The dashed line represents the separation between the species with most (top) and fewer studieavailable. (**B**) Percentage of the proteome of the respective species comprised by RBPs (orange) or RBD-containing proteins (grey). The numbers at the end of the bars indicate the size of the respective proteomes in total. (**C**) Proportion in % of non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light orange) and RBPs with RBD (dark orange) in the different proteomes. (**D**) Fraction of RBPs among all proteins containing the indicated RBDs in the listed species. The percentages are depicted from black (high) to light grey (low). n.e. = non-existent, *i.e.,* the given domain is absent from the species. (**E**) Violin plots representing the protein expression level in HeLa cells according to a mass spectrometry experiment (38) in the four protein groups: non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light orange) and RBPs with RBD (dark orange). *** correspond to *P*-values < 0.001, ns corresponds to non-significant, as resulting from a Wilcoxon rank sum test.

Figure 3. Correlation of the RBDs in the proteins of the well-studied species of the RBP2GO database with the RBP2GO score. (**A**) Boxplots representing the distribution of the RBP2GO score in proteins without RBD (light grey) and with RBD (dark grey). (**B**) Same as in (**A**), but separated into non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light orange) and RBPs with RBD (dark orange). (**C**) Proportion of RBD-containing proteins for each interval of the RBP2GO score. The red line represents a linear regression (up to an RBP2GO score of 50). R represents the Pearson's correlation coefficient. (**D**) Boxplots representing the RBD content fraction per protein in non-RBPs with RBD (blue) and RBPs with RBD (orange). (**E**) Boxplots representing the distribution of the RBP2GO score per number of RBDs in non-RBPs (blue) and in RBPs (orange) in Hs. *** corresponds to *P*-values < 0.001 as resulting from a Wilcoxon rank sum test.

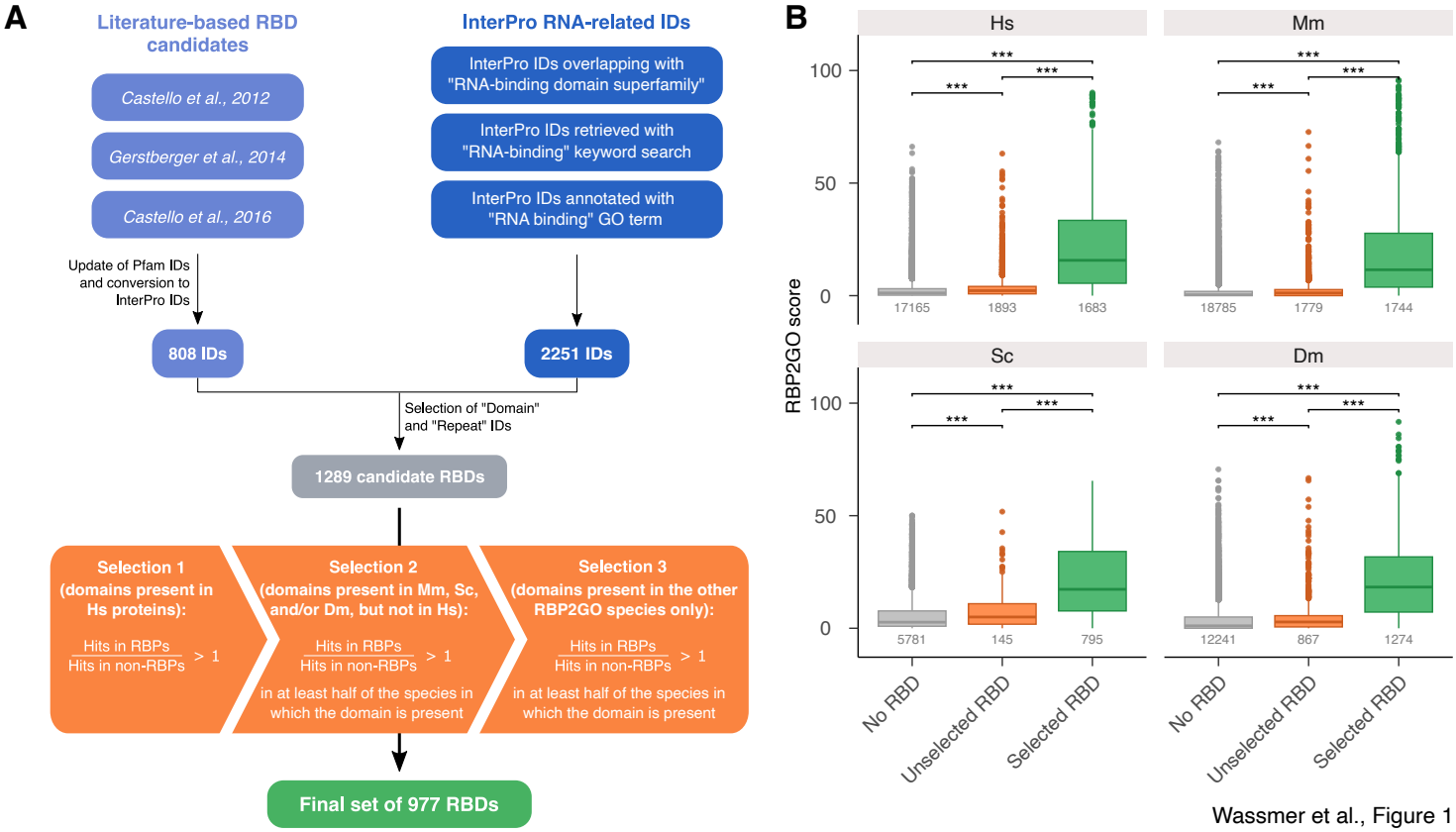
Figure 4. Distribution of the selected Rfam IDs in the proteins of the RBP2GO database. (**A**) Number of selected Rfam IDs present in each species and grouped according to the presence of the Rfam ID either only in RBPs (orange), in both RBPs and non-RBPs (gray) or only in non-RBPs (blue). (**B**) Boxplots representing the distribution of the RBP2GO score in proteins without Rfam ID (light grey) and with Rfam ID (dark grey). (**C**) Same as in (**B**), but separated into non-RBPs without Rfam ID (light green), non-RBPs with Rfam ID (dark green), RBPs without Rfam ID (light purple) and RBPs with Rfam ID (dark purple). (**D**) Proportion of Rfam ID-associated proteins for each interval of the RBP2GO score. The red line represents a linear regression (up to a score of 50). R represents the Pearson's correlation coefficient. (**E**) Same as in (**B**), but separated between non-RBPs without RBD nor Rfam ID (light blue), non-RBPs with RBD but without Rfam ID (dark blue), non-RBPs without RBD but with Rfam ID (light green), non-RBPs with RBD and Rfam ID (dark green), RBPs without RBD nor Rfam ID (light orange), RBPs with RBD but without Rfam ID (dark orange), RBPs without RBD but with Rfam ID (light purple) and RBPs with RBD and Rfam ID (dark purple). *, ** and *** correspond to *P*-values < 0.05, 0.01 and 0.001, respectively, ns corresponds to non-significant, as resulting from a Wilcoxon rank sum test.

Figure 5. Analysis of the disordered fraction in the human and murine proteins of the RBP2GO database. (**A**) Boxplots showing the distribution of the disordered fraction in the four groups of proteins: non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light orange) and RBPs with RBD (dark orange). The proteins containing no disordered region were also included. The numbers given in grey are the number of proteins in each group. (**B**) Boxplots representing the distribution of RBP2GO score in eight groups of proteins: non-RBPs without RBD nor IDR (light green), non-RBPs without RBD but with IDR (light blue), non-RBPs with RBD but without IDR (dark green), non-RBPs with RBD and IDR (dark blue), RBPs without RBD nor IDR (light orange), RBPs without RBD but with IDR (light red), RBPs with RBD but without IDR (dark orange) and RBPs with RBD and IDR (dark red). (**C**) Boxplots representing the distribution of the RBP2GO score for non-RBPs (blue) and RBPs (red) in function of the disordered fraction of the proteins. R represents the Pearson's correlation coefficient. *, ** and *** correspond to *P*-values < 0.05, 0.01 and 0.001, respectively, ns corresponds to non-significant, as resulting from a Wilcoxon rank sum test.

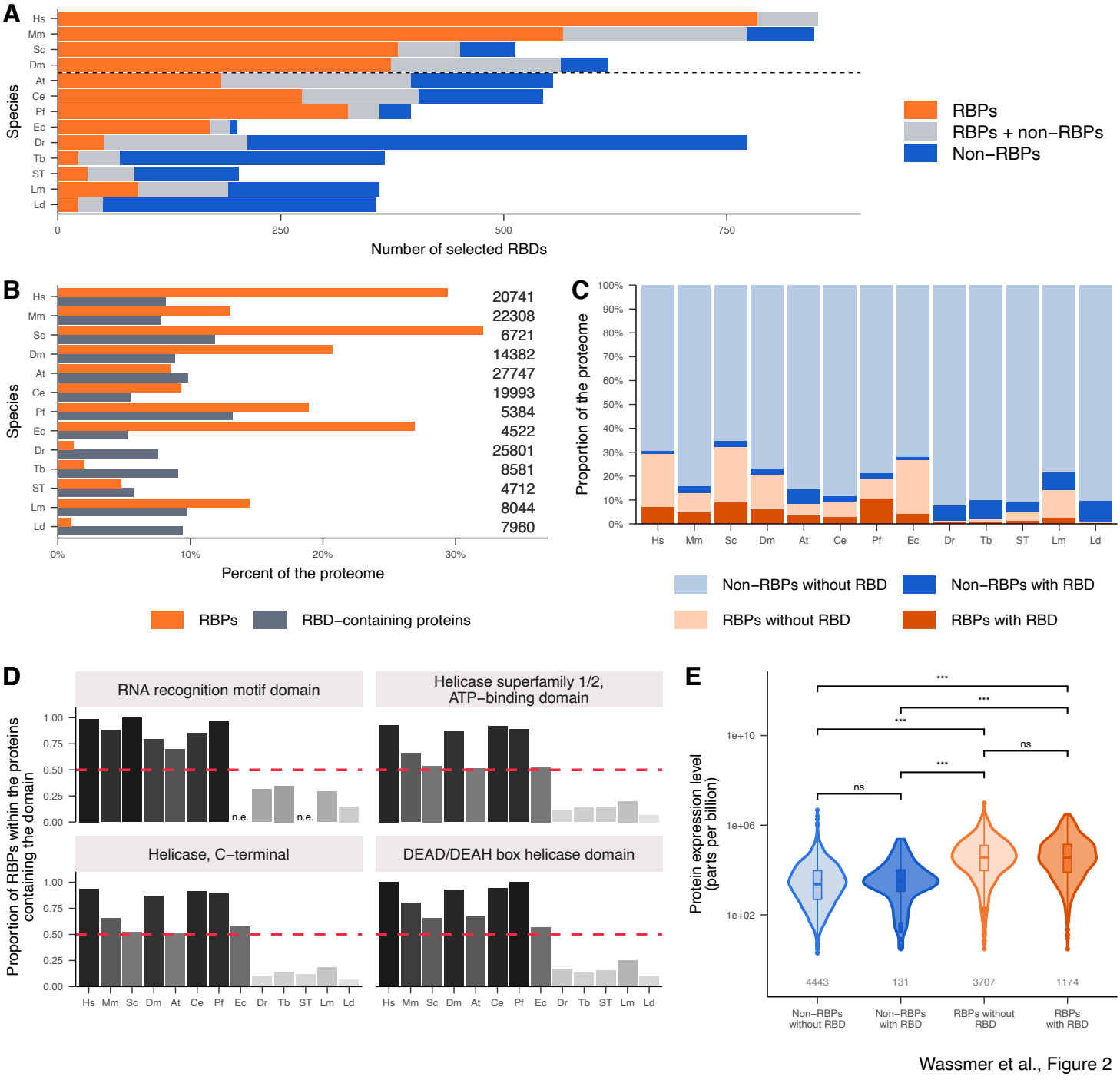
Figure 6. Discovery of candidates for RBP aspirants using the RBD content of the non-RBP proteins. (**A**) Number of RBPs and RBD-containing non-RBPs (numbers indicated in blue for this group) for species with a high (≥9, top bar) or a low (<9, bottom bar) number of RBP screen studies. (**B**) Heatmap showing the most enriched GO terms for molecular function in the RBD-containing non-RBPs as compared to the proteome of each species. The GO terms were classified by mean fold enrichment, from top to bottom. The terms are related among other molecular functions to metabolism (blue), to DNA and RNA (green), and to chromatin regulation (orange). (**C**) Same as in (**B**) for the most enriched GO terms for biological processes.

Figure 7. Discovery of new RBD candidates in RBPs lacking RBDs. (**A**) Volcano plots showing the enrichment and its adjusted *P*-values for each of the InterPro domains present in the RBPs without RBD within the top 20% of RBP2GO scores, with an RBP/non-RBP ratio > 1, as compared to three different control groups (RBPs without RBD within the bottom 20% RBP2GO scores; non-RBPs without RBD; whole proteome). The *P*-values were calculated using a Fisher's exact test and were adjusted with the FDR method. (**B**) Enrichment and *P*-values for the 15 InterPro IDs selected as candidate RBDs, see also Supplementary Table S8. Some InterPro IDs are related to chaperones (orange), cytoskeletal-protein binding (green), or importins (blue). (**C**) Proportion of proteins without RBD (grey), with selected RBDs (dark green) or new RBD candidates (light green) in non-RBPs or RBPs. (**D**) Boxplots showing the distribution of the RBP2GO score for the proteins without RBD (grey), containing a new RBD candidate (light green), or a selected RBD (dark green). *** corresponds to *P*-values < 0.001 as resulting from a Wilcoxon rank sum test. The numbers given in grey represent the size of the protein groups. (**E**) Fold enrichment of RBPs over non-RBPs across all proteins containing one of the new RBD candidates. The *P*-values were calculated using a Fisher's exact test. (**F**) Boxplots showing the distribution of the log2 of the normalized mean proportion of the RNA-binding peptides overlapping the domain + $1x10^{-2}$. The domains were categorized as not selected RBDs (grey, see Figure 1), new RBD candidates (light green) and selected RBDs (dark green, see Figure 1). Only domain present in at least 5 proteins were taken into account (see Material and Methods section for more details). The number above each box represents the median values and the number under each box represents the size of the group. The *P*-values were calculated using a Wilcoxon rank sum test. (**G**) Schematic representations of three proteins (grey) with Armadillo repeats (blue) and the overlapping RNA-binding peptides (red).

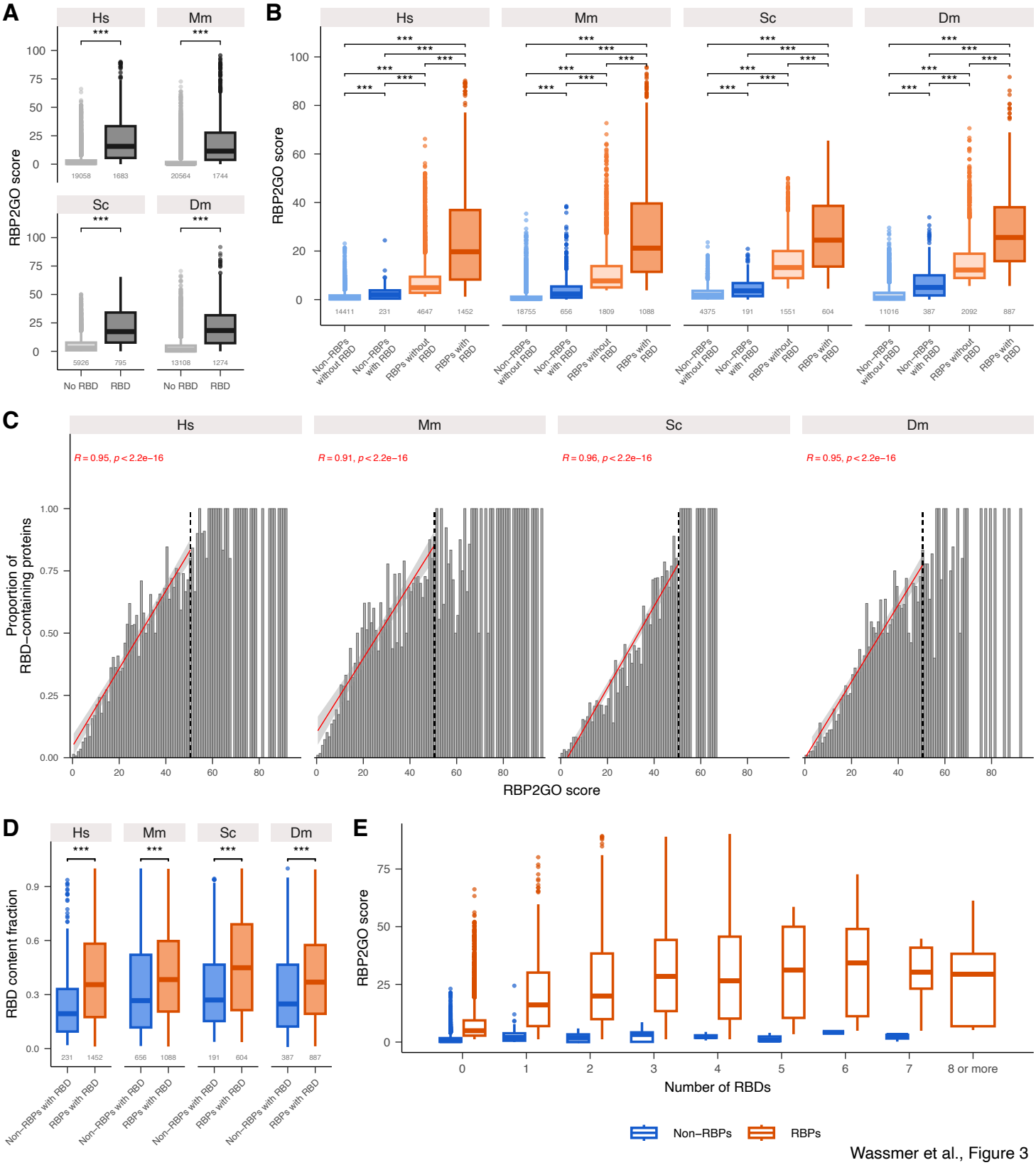
Figure 8. Development of the RBP2GO composite score and update of the RBP2GO database. (**A**) Schematic representation of the compilation of the RBP2GO composite score. The quality factors as well as their attribution rules are explained in the Supplementary Table S10. The maximum values are 50 for component 1 and 25 for each component 2 and 3. The RBP2GO composite score takes values between 0 and 100. (**B**) Violin plots representing the distribution of the RBP2GO score (orange) and the RBP2GO composite score (green) for non-RBPs without RBD nor Rfam ID, for non-RBPs with RBD and/or Rfam ID, for RBPs without RBD nor Rfam ID and for RBP with RBD and/or Rfam ID. (**C**) Screenshot of the sidebar of the advanced RBP2GO database (https://rbp2go.dkfz.de) showing the added menu for domains and families in orange. (**D**) Screenshot of the new "Domain information" tab added to the protein search section of each species in the RBP2GO database. This new tab displays all information pertaining to the RBDs and Rfam IDs of the selected protein, as well as information on its disorder content. It includes links to the InterPro (33) and MobiDB (40) databases. For each InterPro ID, a table indicates the type of the ID as "Domain", "Repeat" or "Family" and if it is linked to "RNA-binding". (**E**) Screenshot of the additional search function dedicated to the protein domains (found under the SEARCH DOMAINS of the sidebar as shown in **C**).



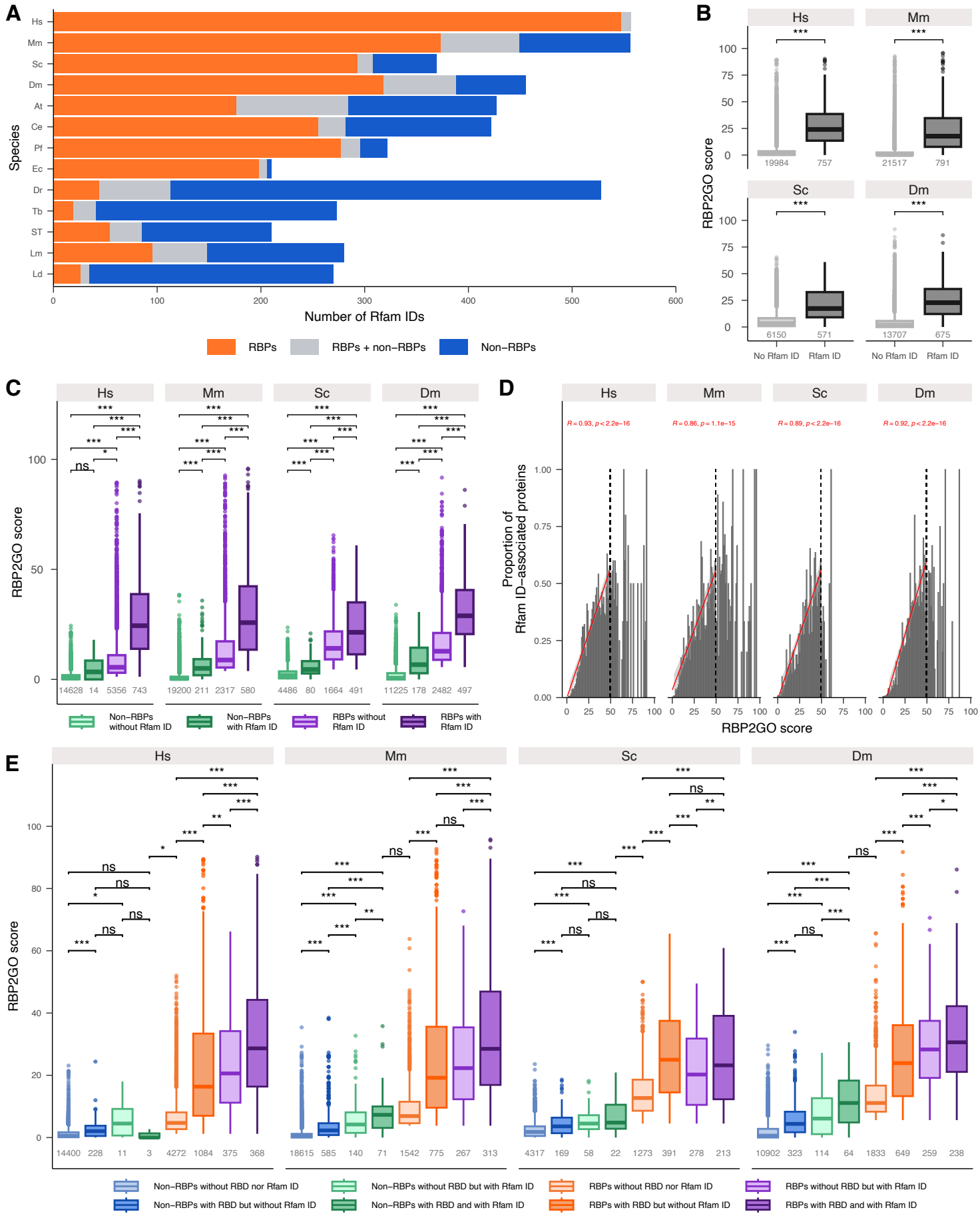
Wassmer et al., Figure 1



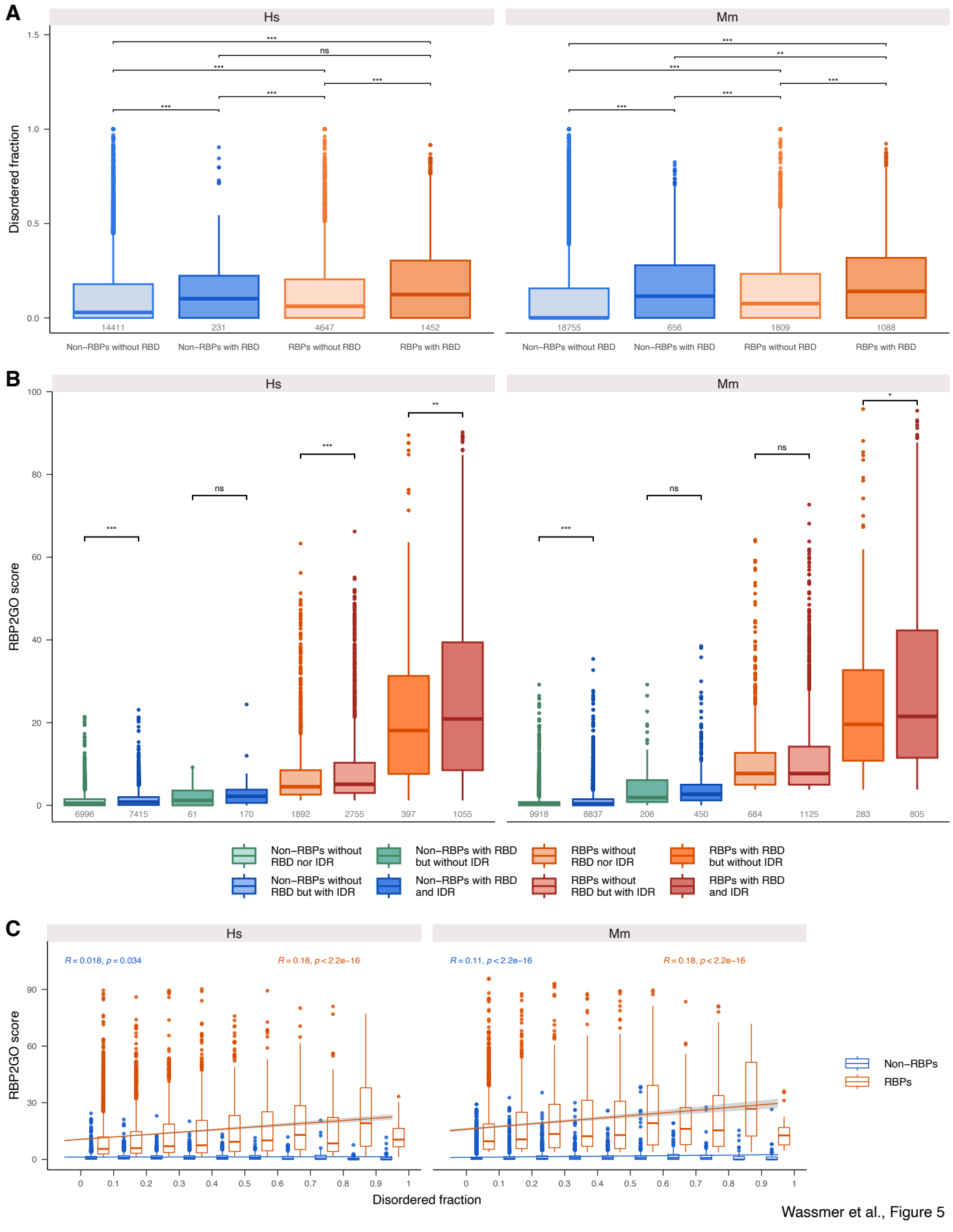
Wassmer et al., Figure 2



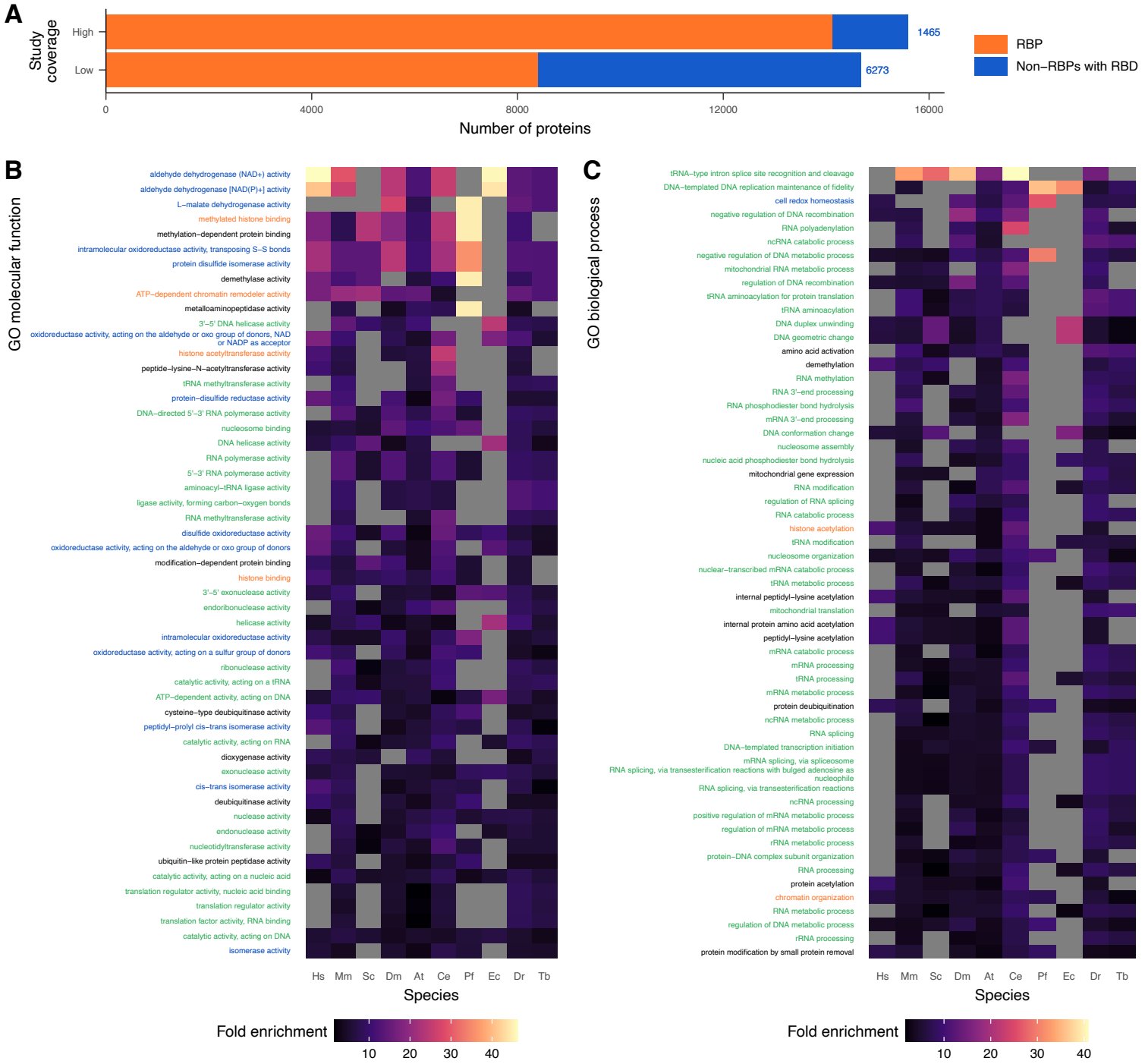
Wassmer et al., Figure 3



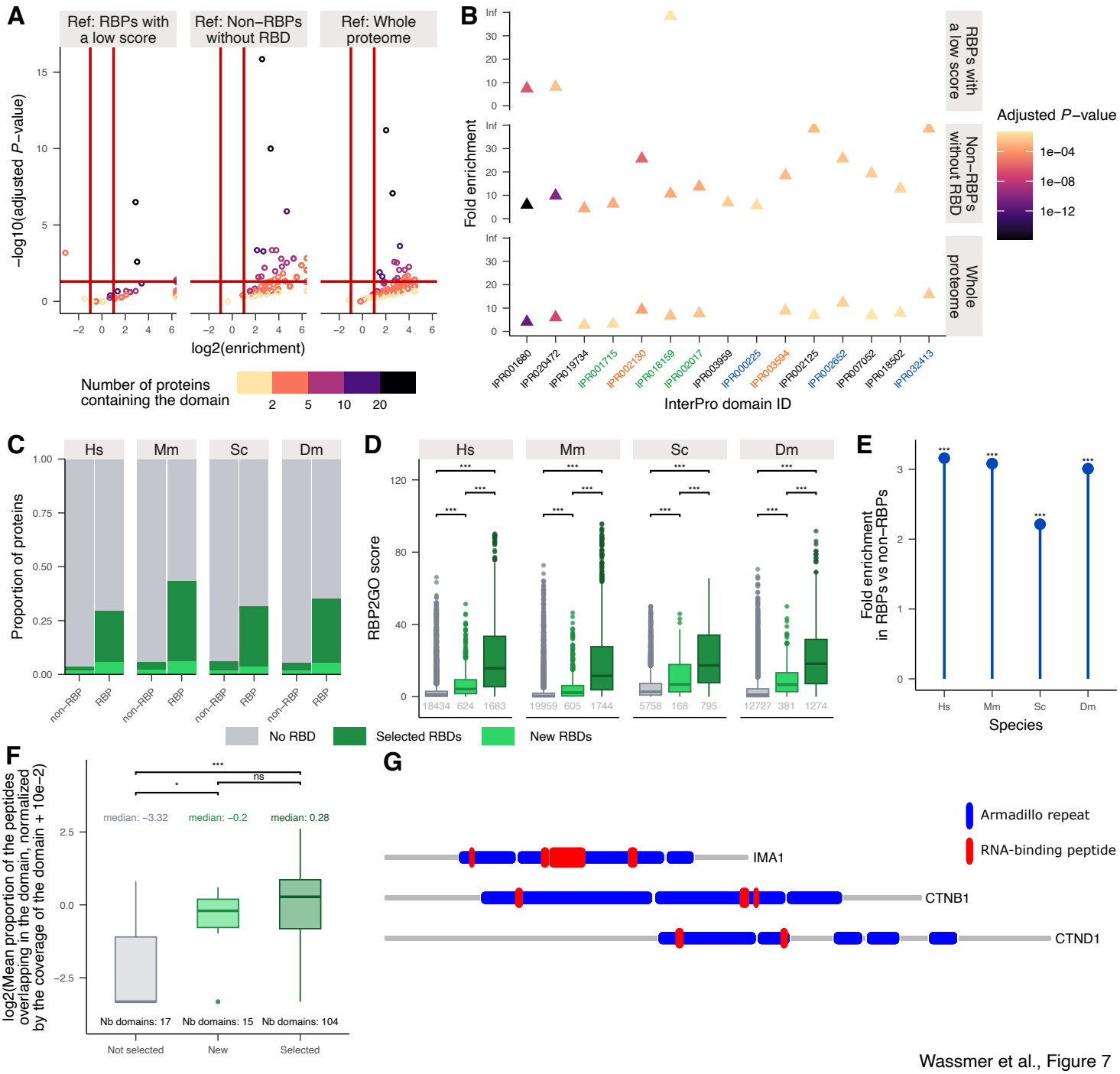
Wassmer et al., Figure 4



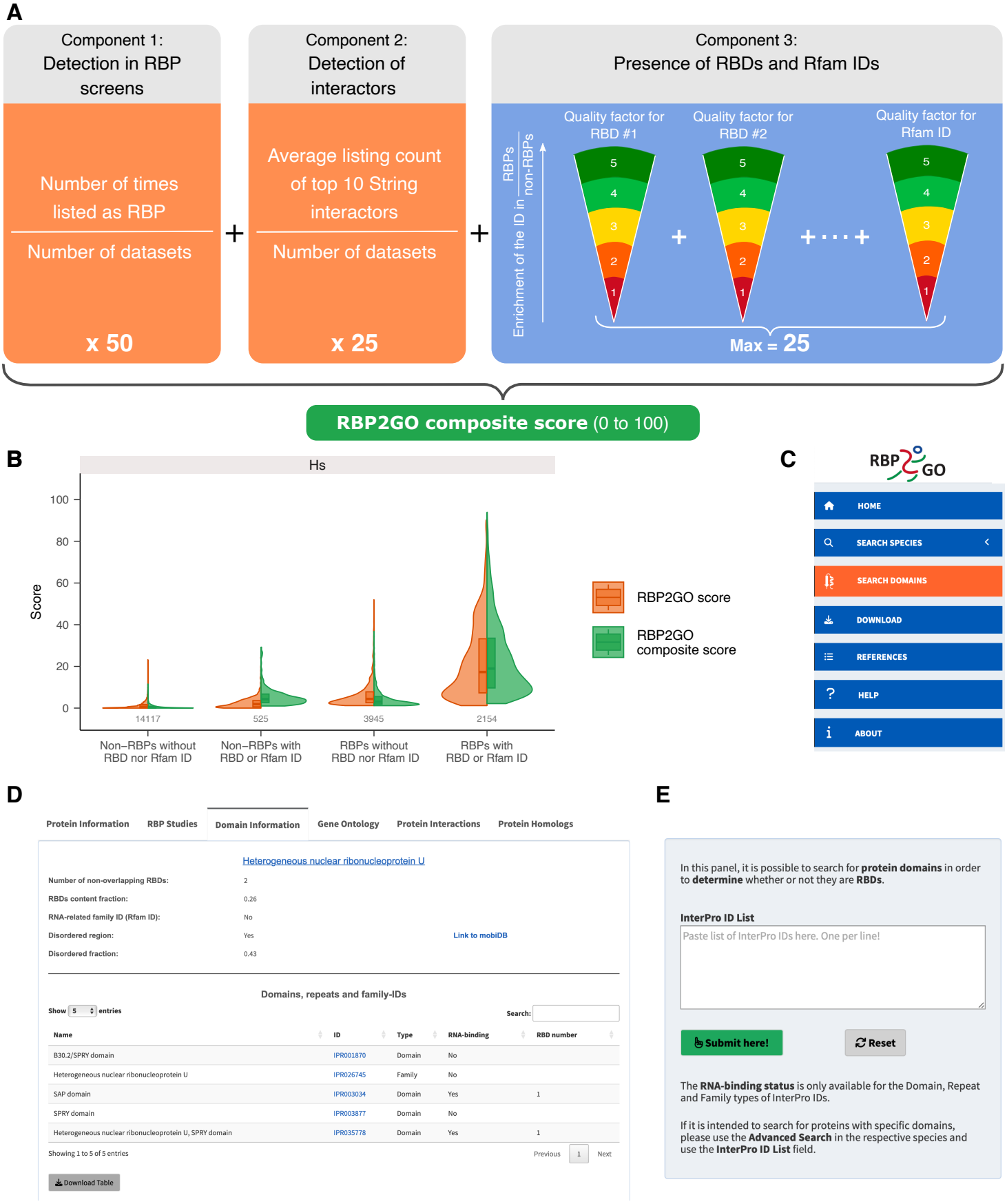
Wassmer et al., Figure 5



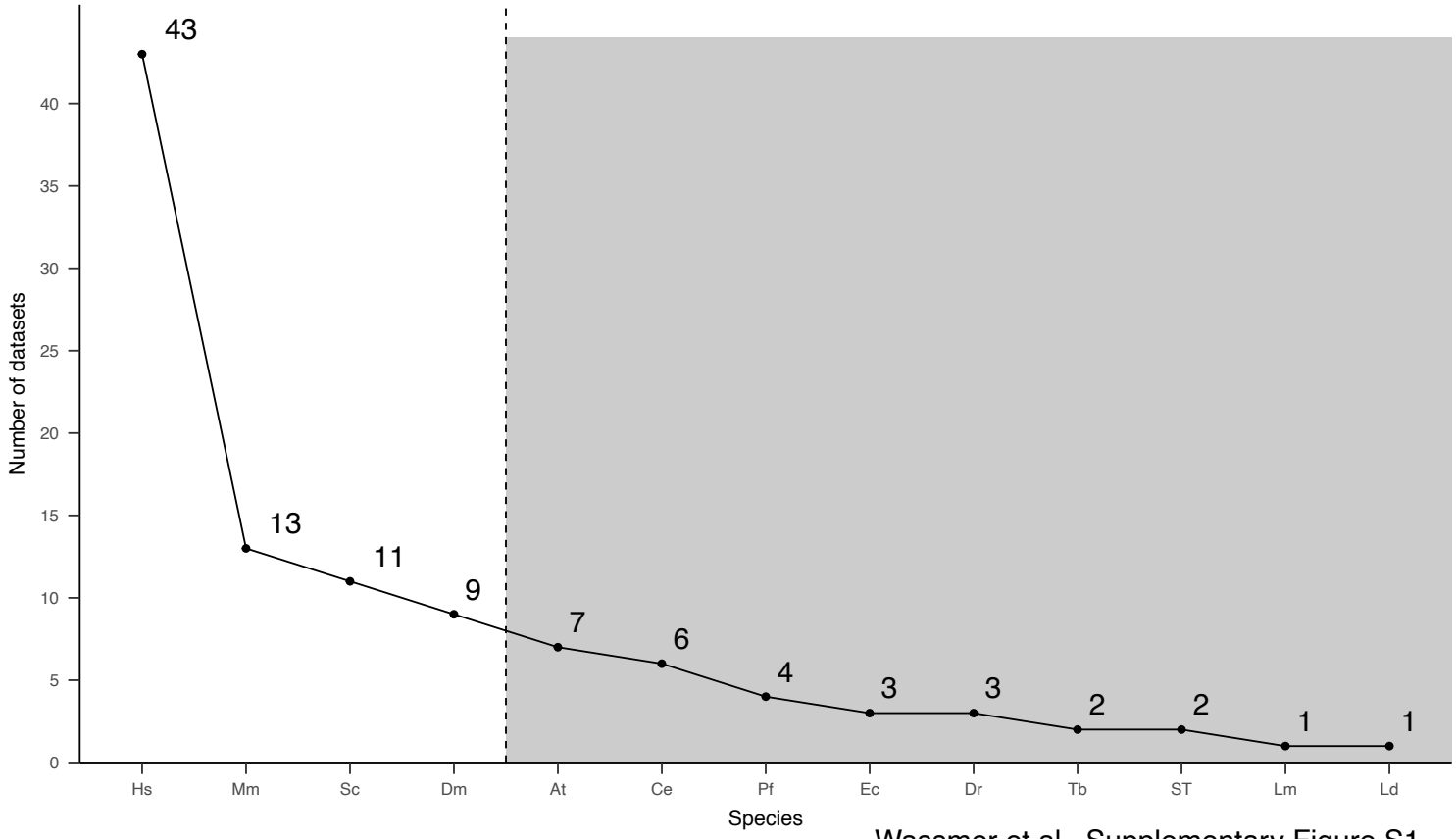
Wassmer et al., Figure 6



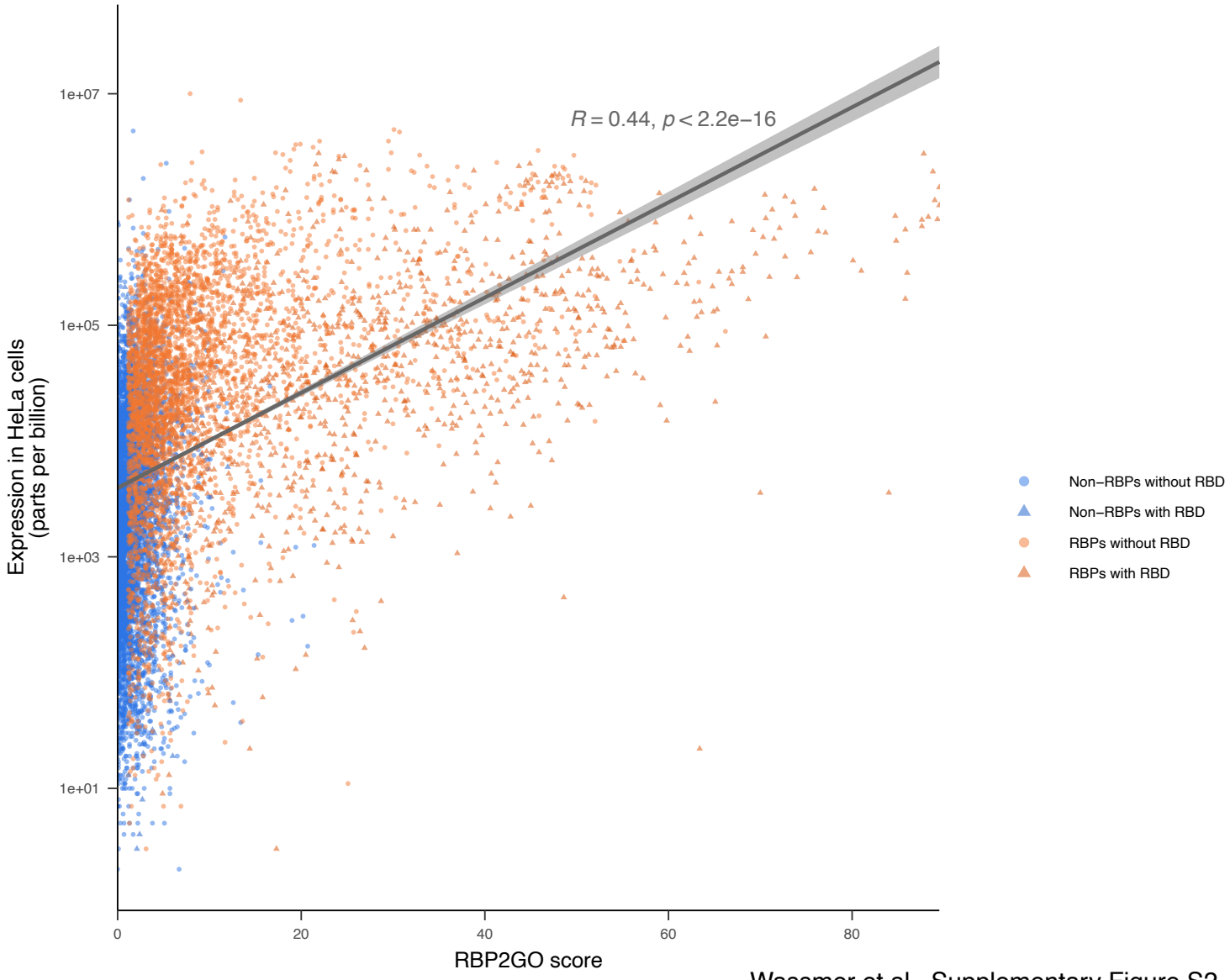
Wassmer et al., Figure 7



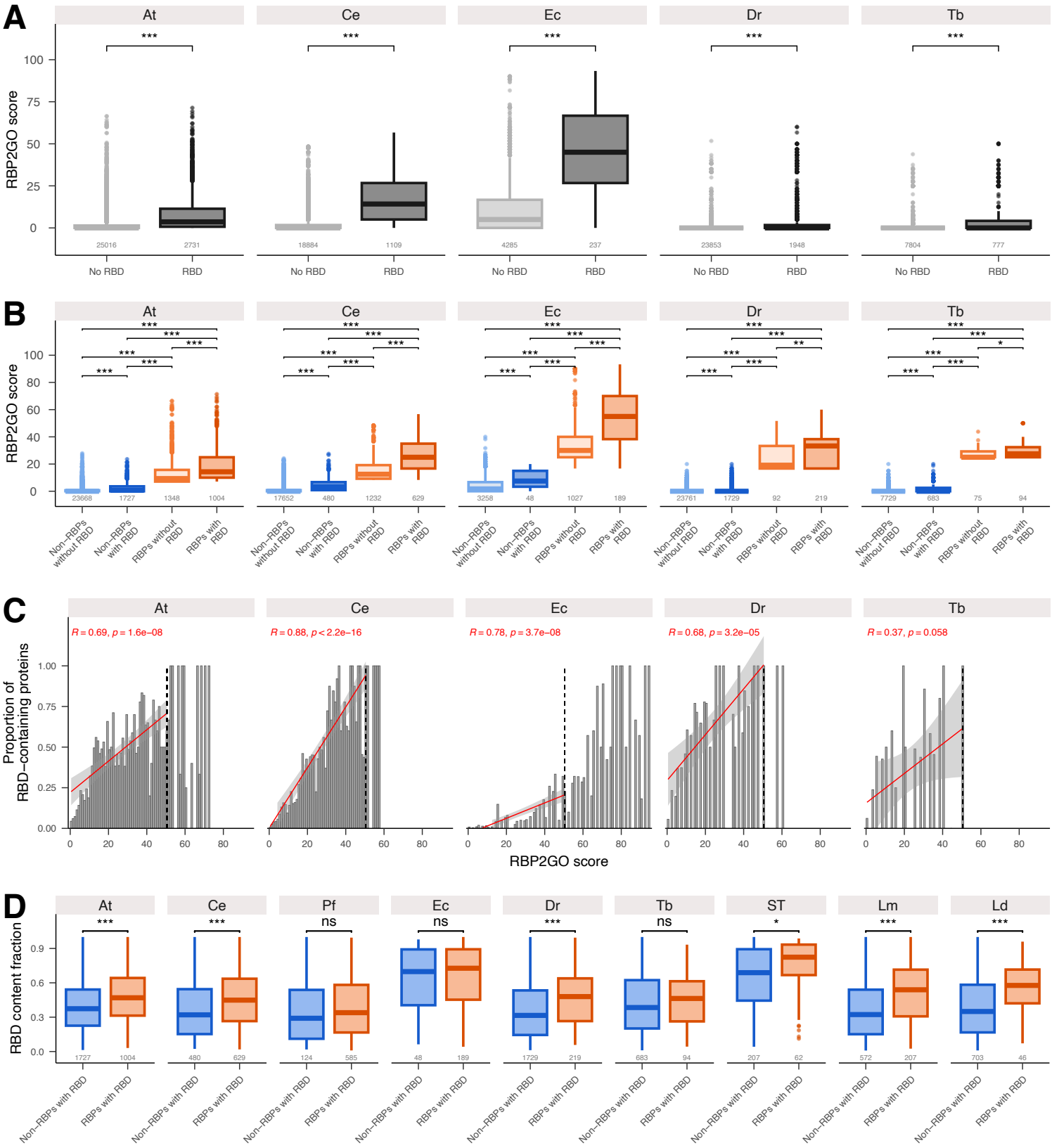
Wassmer et al., Figure 8



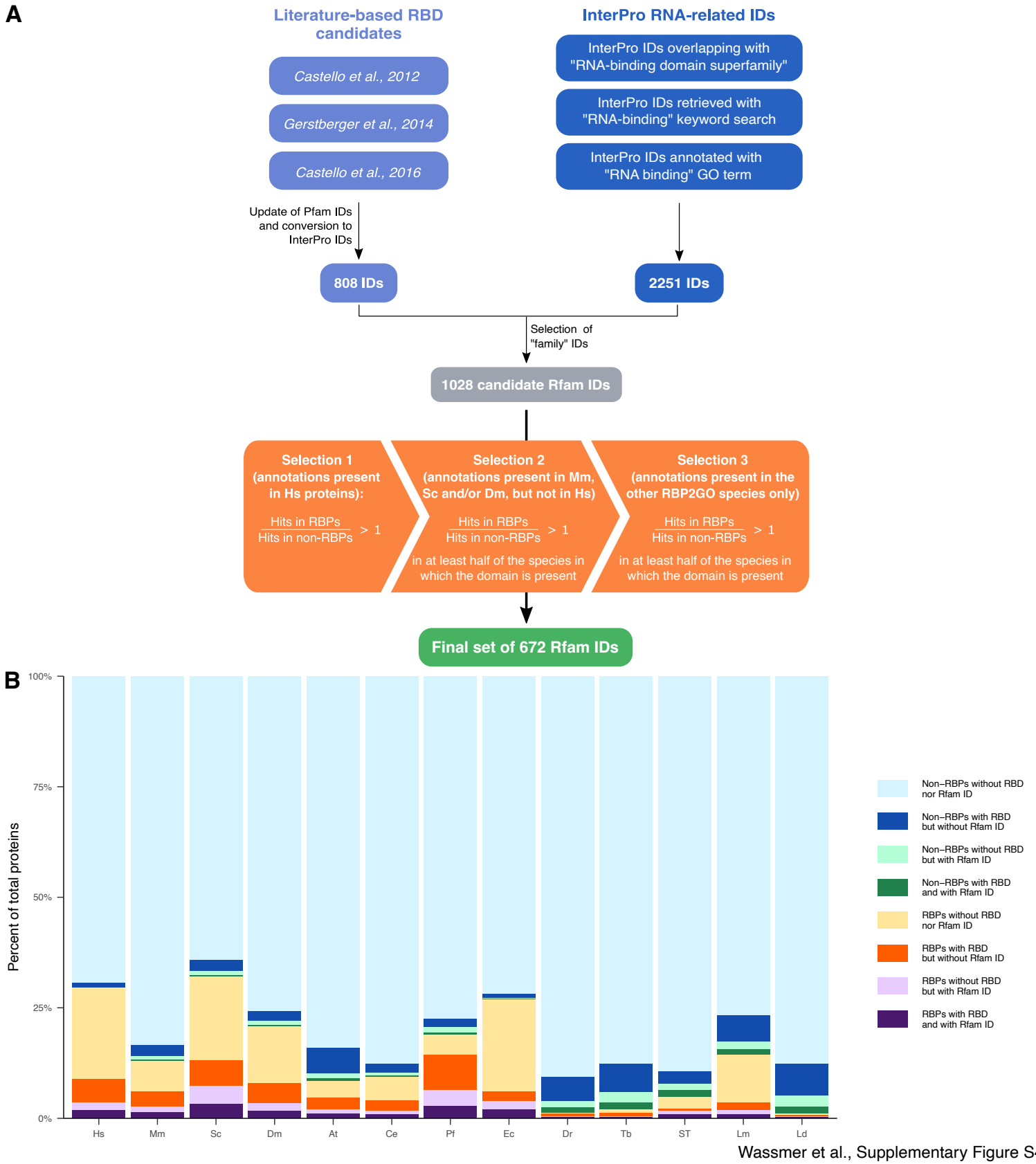
Wassmer et al., Supplementary Figure S1



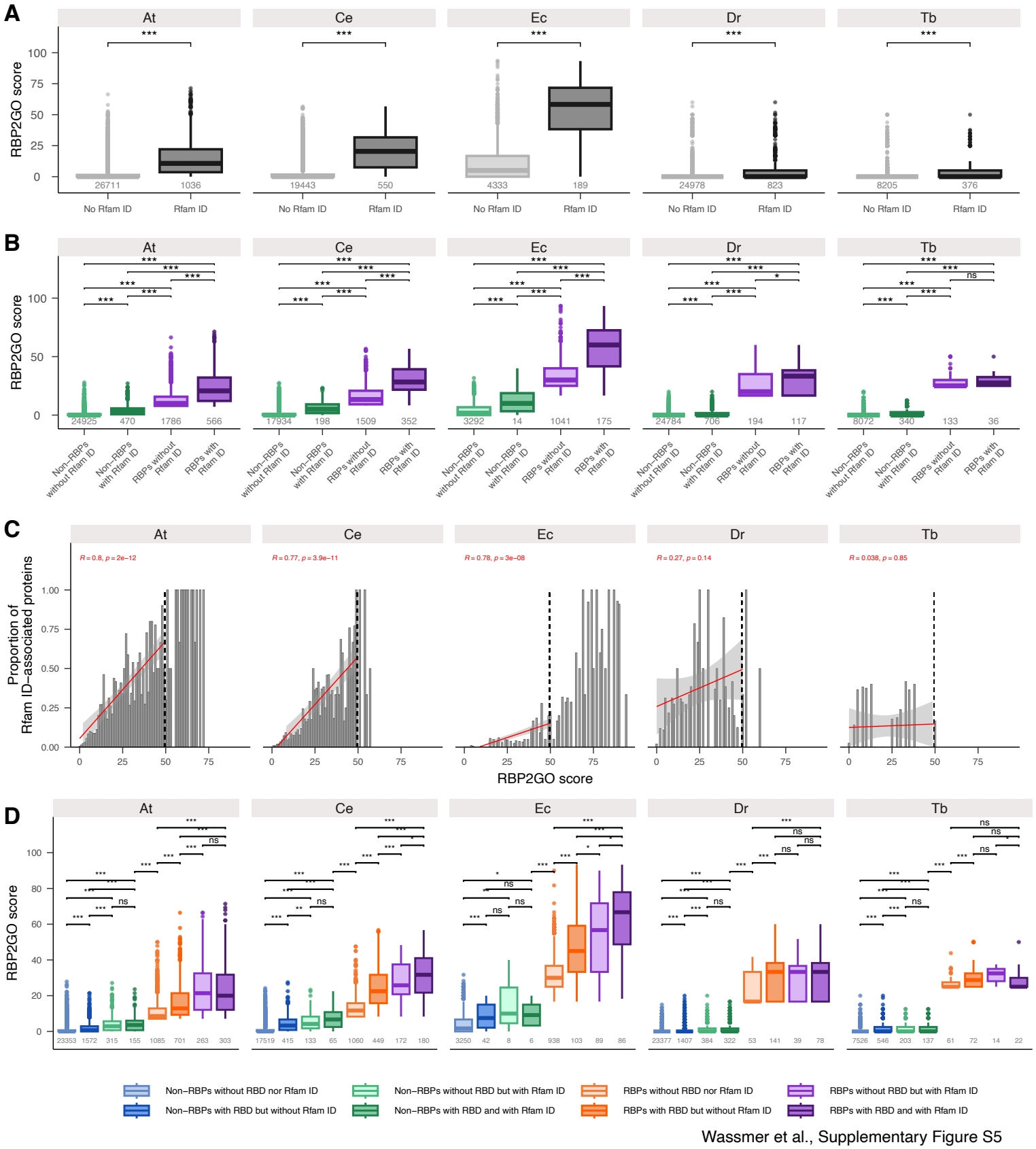
Wassmer et al., Supplementary Figure S2



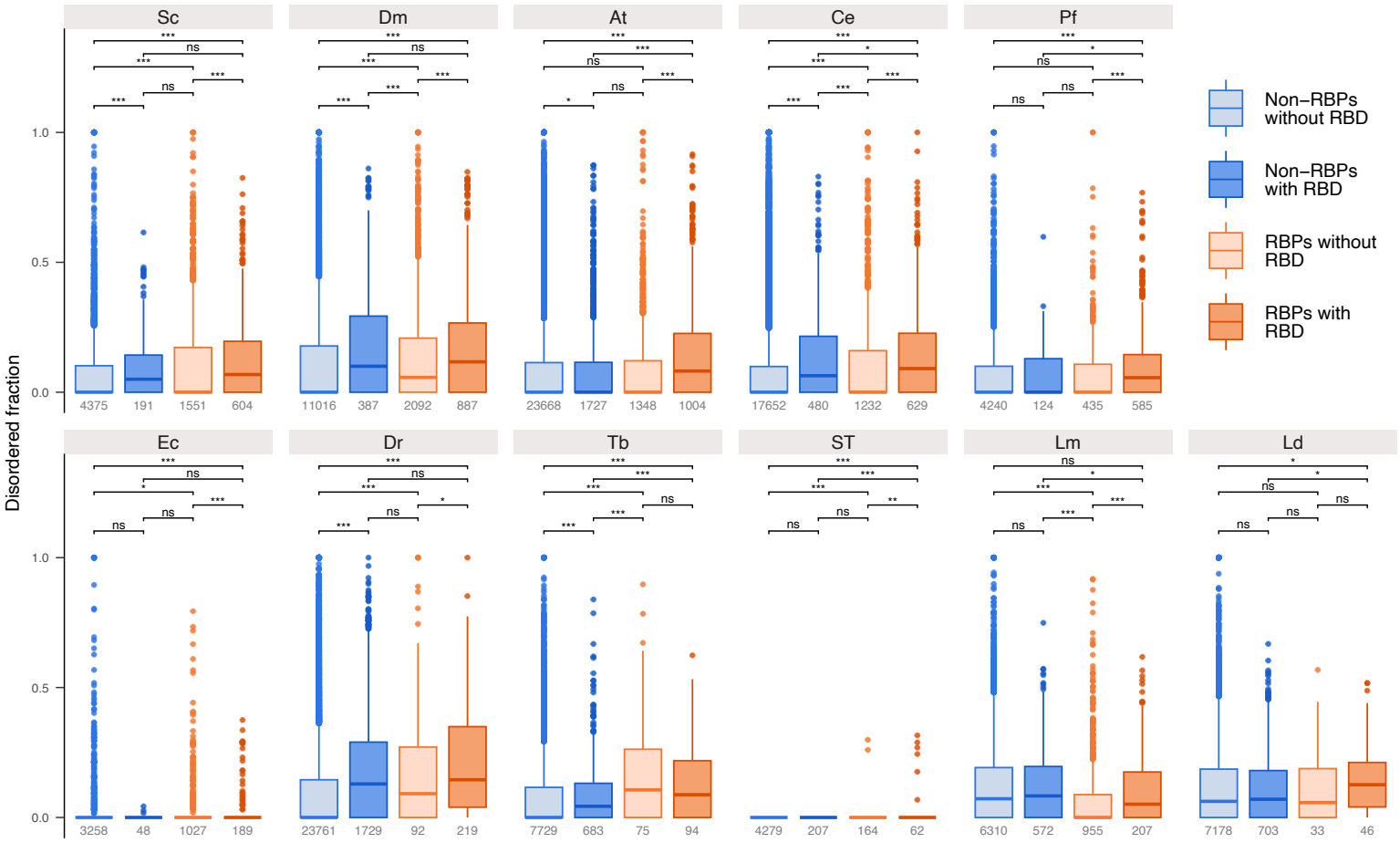
Wassmer et al., Supplementary Figure S3



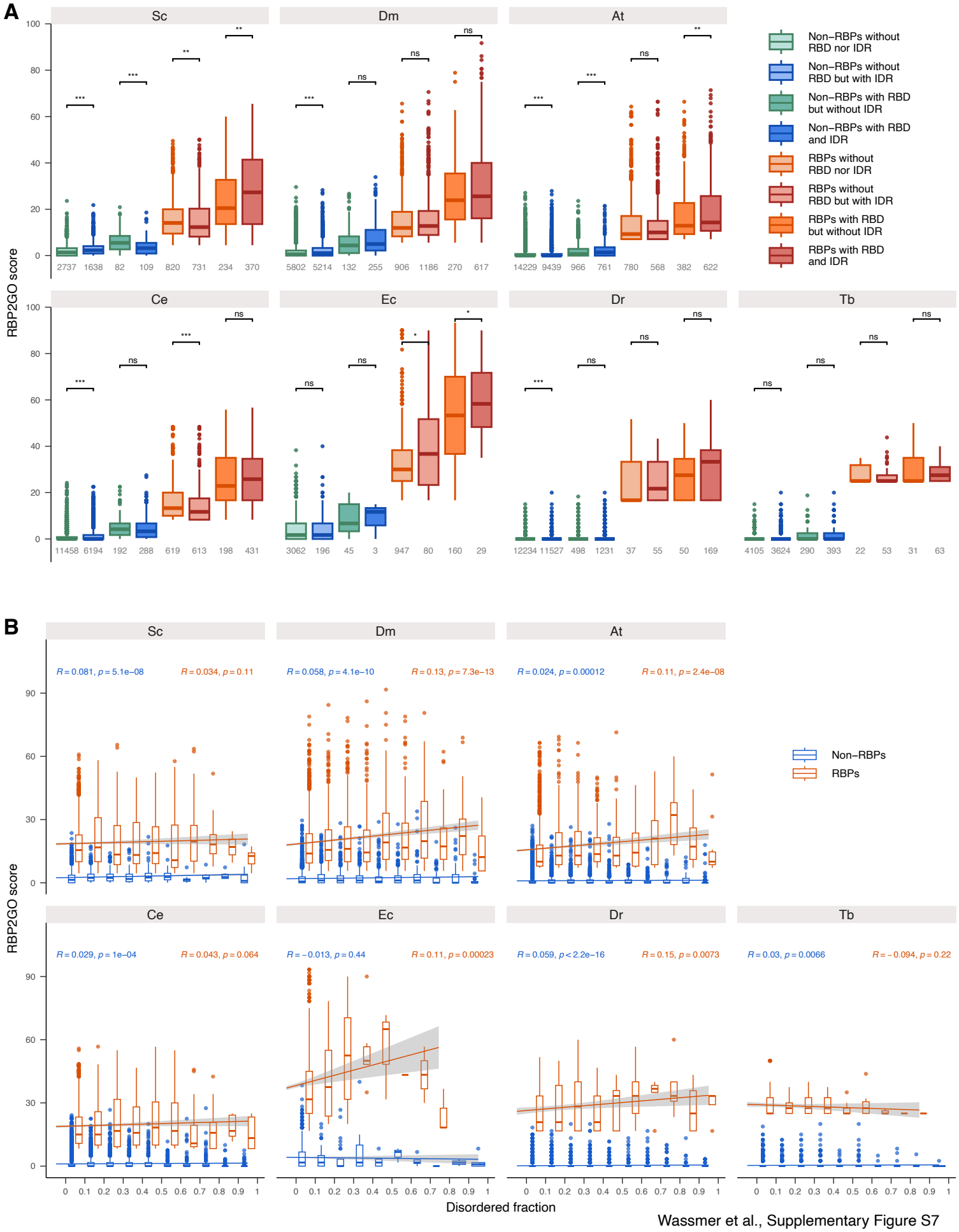
Wassmer et al., Supplementary Figure S4



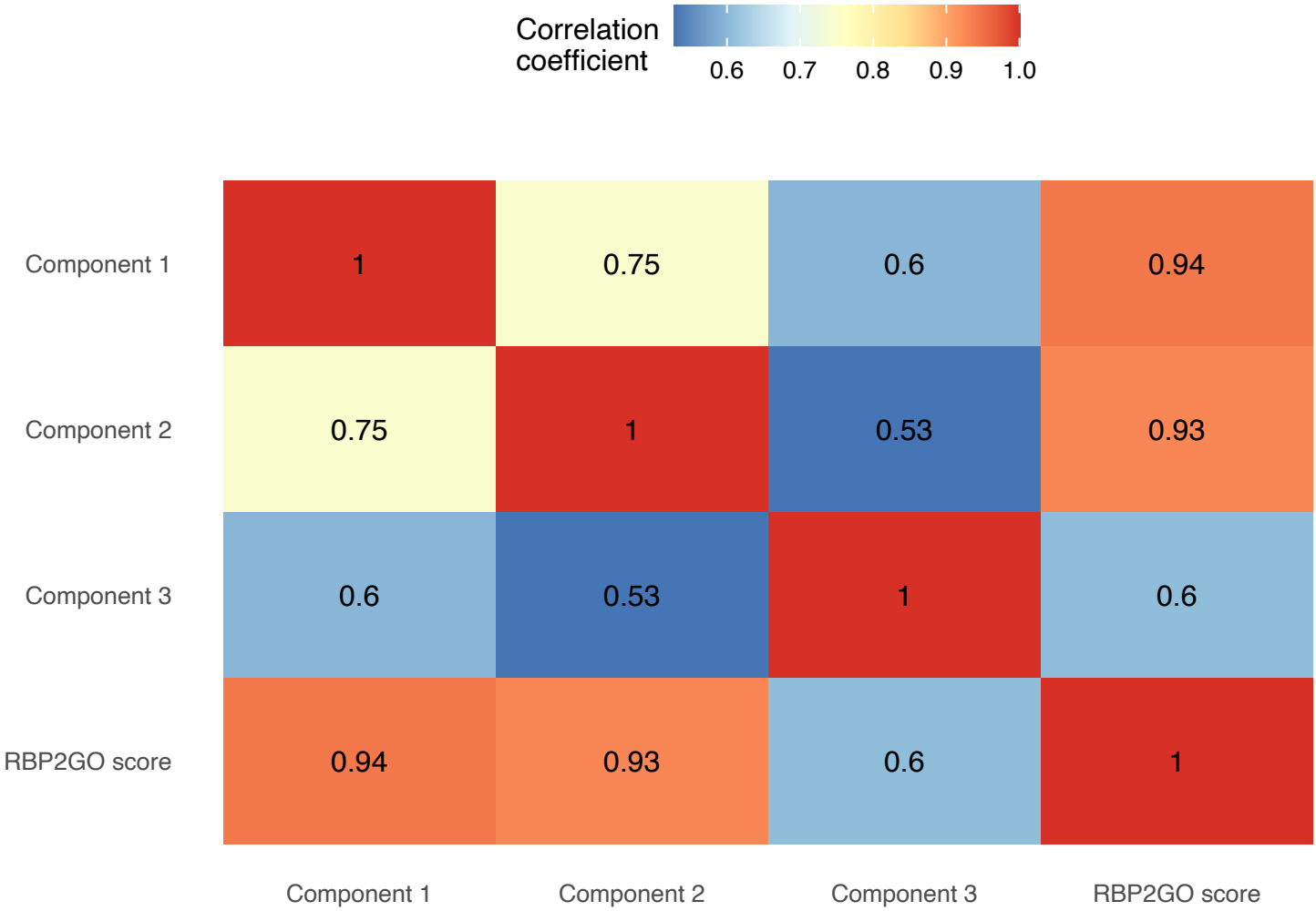
Wassmer et al., Supplementary Figure S5



Wassmer et al., Supplementary Figure S6



Wassmer et al., Supplementary Figure S7



Wassmer et al., Supplementary Figure S8