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

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

[1] Research Group "RNA-Protein Complexes & Cell Proliferation", 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, and German Cancer Consortium (DKTK), partner site Freiburg, a partnership between DKFZ and University Medical Center Freiburg, 79106 Freiburg, Germany

[†] These authors contributed equally to this work

* 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

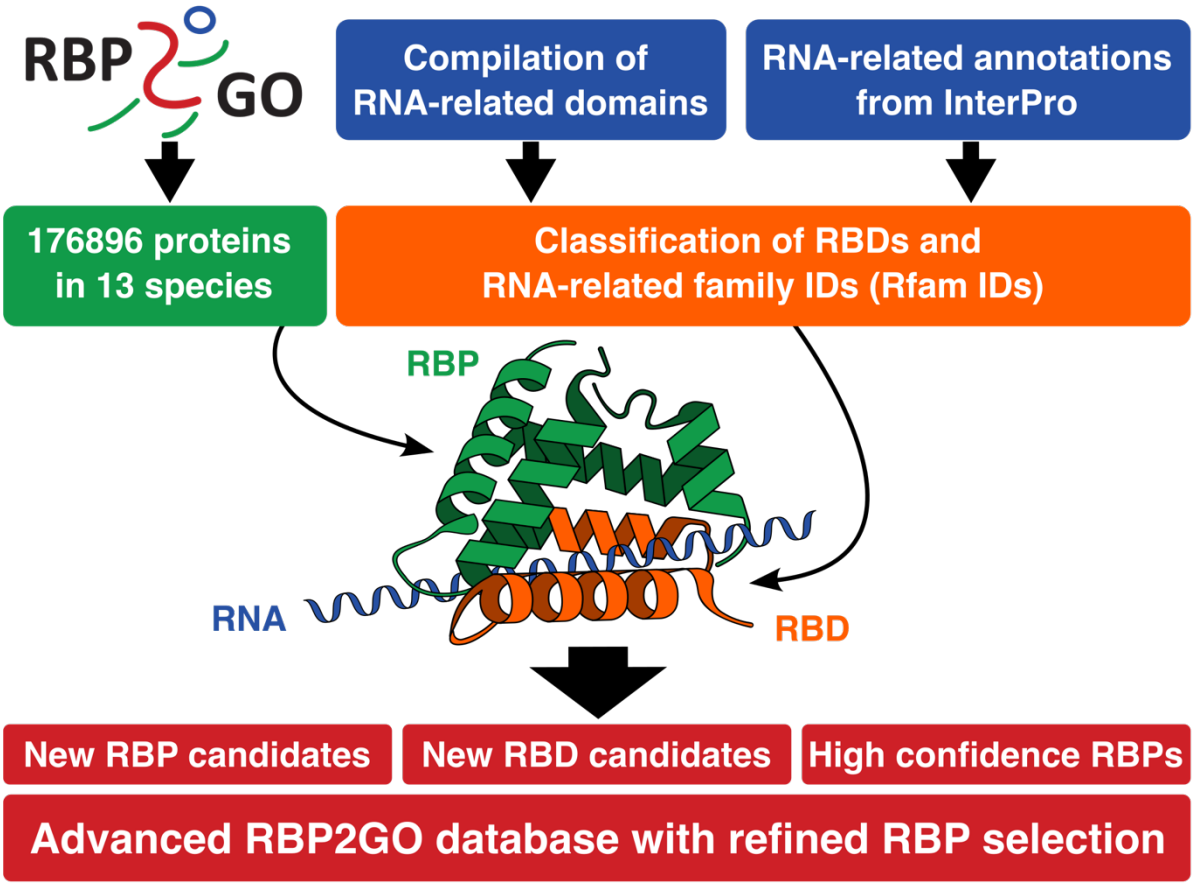
Key Points

- 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 in intrinsically disordered regions (IDRs) and low complexity regions (LCRs). We found a strong positive correlation between IDRs and RBDs and a co-occurrence of specific LCRs. 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 less investigated species. By further analyzing RBPs without RBD, we predicted new RBDs that were validated by RNA-bound peptides. Finally, we created the 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-2-Beta.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, 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

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

84 In this study, we addressed this issue by performing a pan-species analysis of the RNA-binding
85 characteristics of the proteins included in the RBP2GO database. To this aim, we collected lists of RBDs
86 and RNA-related protein families (Rfam) from published studies (8,13,29) and from the InterPro
87 database (33). We examined their distribution in RBP candidates as compared to non-RBPs in order to
88 compile a list of selected RBDs and Rfam identifiers (Rfam IDs) that were enriched in RBP candidates.
89 Using data from the MobiDB database (34), we also analyzed the IDR and low complexity region (LCR)
90 content and their distribution in proteins with respect to their RBD content and found a strong correlation
91 between the IDR/LCR and RBD content. Our bioinformatic approach indicates that the presence of a
92 selected RBD / Rfam ID can be used to evaluate the RNA-binding potential of proteins as well as to
93 predict new RBP candidates, in particular in species with very few available proteome-wide studies. In
94 addition, we analyzed the RBP candidates lacking selected RBDs and identified 15 RBDs that were not
95 in the starting collection of RBDs, that we further validated based on studies of RNA-bound peptides.
96 Finally, by calculating a new RBP2GO composite score that combines the former RBP2GO score (19)
97 together with new quality factors characterizing the selected RBDs and Rfam IDs, we computed a list
98 of 2018 high-confidence human RBPs. The results of our analysis were implemented into the RBP2GO
99 database, along with new search options to browse 1) RBDs and Rfam IDs via a given InterPro ID and
100 2) proteins, with respect to their RNA-binding features. We anticipate that this additional information will
101 considerably help refining our classification of RBP candidates and foster new developments in the
102 field.

103 MATERIAL AND METHODS

104 Compilation of a list of RNA-binding domain candidates

105 Based on a literature search, the lists of RBDs from three studies were downloaded from the respective
106 supplementary information (8,13,29). The domain IDs were converted into InterPro IDs if referenced
107 from another database such as Pfam (35) and finally fused together with the list of IDs from InterPro.
108 Outdated IDs were updated or discarded if no corresponding InterPro ID could be found. Altogether,
109 808 IDs were retrieved from the published datasets.

110 Based on a database search, we first downloaded the list of the InterPro IDs related to “RNA-binding”
111 from the InterPro website (https://www.ebi.ac.uk/interpro/, InterPro v.88 released on March 10th, 2022)
112 (33). Second, since the InterPro database provides a Gene Ontology (GO) annotation that is associated
113 to the InterPro IDs, we downloaded all the IDs annotated with the GO term “RNA-binding”
114 (GO:0003723). Third, we downloaded from the InterPro website all the domains that overlapped with

115 the “RNA binding domain superfamily” (IPR035979). Taken together, these three datasets amounted
116 to 2251 unique InterPro IDs.

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

120 **Selection of the RNA-binding domains**

121 A three-step selection procedure was applied to the set of RBD candidates based on the hit ratios
122 between RBPs (RBP candidates from proteome-wide studies, i.e., detected in at least one study) and
123 non-RBPs in the species reported in the RBP2GO dataset (https://rbp2go.dkfz.de) (22). First, we
124 selected the InterPro IDs of the RBD candidates that were enriched (hits in RBPs > hits in non-RBPs)
125 in *Homo sapiens* (Hs) (Selection 1, Figure 1A) as the number of studies available for this species was
126 by far the largest as compared to the other species (Supplementary Figure S1). We updated the UniProt
127 IDs of the proteins according to the UniProt release 2022_01 from February 23rd, 2022 (36). The data
128 for the obsolete UniProt IDs were either deleted when the ID was fused to another one already existing
129 in the database, or duplicated if the ID was split between several new UniProt IDs. The updated sizes
130 of the proteomes are reported in Supplementary Table S2. In a second step, for RBDs not present in
131 Hs (Selection 2, Figure 1A), we selected the InterPro IDs of the RBD candidates that were enriched in
132 *Mus Musculus* (Mm), *Saccharomyces cerevisiae* (Sc) and *Drosophila melanogaster* (Dm) in at least
133 half of the species where the RBD was present. Finally, for RBDs present in neither of the four species
134 Hs, Mm, Sc, Dm (Selection 3, Figure 1A), we selected the RBDs that were enriched in at least half of
135 the remaining nine species where the RBD was present, *i.e.*, *Arabidopsis thaliana* (At), *Caenorhabditis*
136 *elegans* (Ce), *Plasmodium falciparum* (Pf), *Escherichia coli* (Ec), *Danio rerio* (Dr), *Trypanosoma brucei*
137 (Tb), *Salmonella Typhimurium* (ST), *Leishmania mexicana* (Lm) and *Leishmania donovani* (Ld). This
138 three-step selection procedure led to the selection of 977 RBDs (Supplementary Table S3). The
139 proteins of the species represented in the RBP2GO database were updated with the information on the
140 selected RBDs (Supplementary Table S4). The data available in the Supplementary Table S4 was used
141 to produce the graphs displayed in Figures 1B and 2A-D.

142 **Protein expression levels in HeLa cells**

143 The expression data displayed in the Figure 2E and Supplementary Figure S2 was downloaded for
144 analysis from the EBI atlas website (https://www.ebi.ac.uk/gxa/experiments/E-PROT-19/Results) (37).
145 The expression values can be found in the Supplementary Table S5.

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

148 The coordinates of each InterPro ID of the selected RBDs in the proteins were downloaded from the
149 InterPro database (33) (https://ftp.ebi.ac.uk/pub/databases/interpro/releases/88.0/, released on March
150 10th, 2022).

151 To compute the RBD content fraction per protein, we merged separately all the overlapping RBDs
152 present in a protein. Next, we used the protein length previously retrieved from UniProt (36)
153 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2022_01/, released on
154 February 23rd, 2022) to calculate the fraction of the sequence annotated as RBD for each protein.

155 To compute the number of RBDs present in each protein, we merged the overlapping coordinates for
156 the InterPro ID in R using the "bed_merge()" function from the "valr" package (38). We then merged, for
157 each protein, the coordinates of the InterPro IDs overlapping by more than 10 amino acids using the
158 same function. The number of coordinate pairs (start, end) generated was then calculated to obtain the
159 number of RBDs per protein which were independent in the sense that they were not overlapping by
160 more than 10 amino acids.

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

163 **Compilation of a list of RNA-related family IDs and selection of the IDs enriched in RBP**
164 **candidates.**
165 In addition to the domains and repeats InterPro ID types (see above), we computed a list of RNA-related
166 InterPro family IDs or Rfam IDs. The "RNA binding domain superfamily" (IPR035979) overlaps with
167 several other domains. Accordingly, the list of 2712 RNA-related InterPro IDs (Supplementary
168 Table S1) was filtered to keep only 1028 family IDs. The same procedure as the one used for the
169 selection of the RBDs was applied to select for the Rfam IDs that were enriched in RBP candidates.
170 For more simplicity in the text, RBP candidates (including RBPs and non-validated RBPs) will be
171 referred to as RBPs. This resulted in a final list of 672 Rfam IDs (Supplementary Table S6). The
172 Supplementary Table S4 was updated accordingly with this information.

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

174 **Retrieval of the MobiDB-lite IDR content fraction**

175 Among the various algorithms that were available for the prediction of IDRs (39), we selected the
176 MobiDB-lite prediction as it is more conservative, through the combination of 10 different prediction
177 tools (34,39). To retrieve the IDR content fraction of the proteins, we downloaded the results of the

178 MobiDB-lite algorithm from the MobiDB database (https://mobidb.bio.unipd.it, Version 5.0.0) (34,40) for
179 each of the 13 RBP2GO species. The Supplementary Table S4 was accordingly updated with this
180 information (IDR content fraction).

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

182 **Linear regressions**

183 For all linear regressions in this study, the regression lines were generated using the geom_smooth()
184 function from the ggplot2 package (41), with the "lm" method. The correlation coefficients in the figures
185 were computed using the R base function "cor()" or the "stat_cor()" function from the ggpubr package
186 (42).

187 **Retrieval of the LCR frequency and amino acid compositional bias in the human proteins**

188 LCRs and sequences of the proteins were obtained directly from the UniProt database using the
189 UniProt.ws package (36): UniProt release 2023_05 from November 8th, 2023. These were used to
190 compute the LCR proportion within the proteins of the four main groups (non-RBPs without RBD, non-
191 RBPs with RBD, RBPs without RBD and RBPs with RBD) as depicted in Figure 5B. Amino acids were
192 grouped in categories (- charge: negatively charged; + charge: positively charged; Aromatic, Uncharged
193 non-polar and uncharged polar). The frequencies of specific motifs such as FGG, RGG and YGG were
194 calculated based on the protein sequence and according to the four main groups of proteins.

195 The results of these analyses were used to produce the panels in Supplementary Figure S8.

196 **GO enrichment analysis for non-RBPs containing an RBD**

197 For each species, a GO-enrichment analysis was performed in the non-RBPs containing one of the
198 selected RBDs against the proteome with the PANTHER 17.0 classification system
199 (http://pantherdb.org, release 2021_03) that was accessed via the "rbioapi" package in R (43). The
200 results for the different species were then merged in a single table (Supplementary Table S7). The GO
201 terms were filtered to keep the terms with an FDR <0.05. For each GO term, the number of species in
202 which it was significantly enriched and its mean fold enrichment in these species were calculated. For
203 the heatmaps (Figure 6), only the GO terms significantly enriched (FDR < 0.05) in more than four
204 species out of 13, and with a mean fold enrichment > 5, were selected. The heatmaps were created
205 with the geom_tile() function from the ggplot2 package (41), and the GO terms were sorted by
206 decreasing mean fold enrichment (Figure 6).

207 **Identification of new RBDs using the human RBPs lacking canonical RBDs**

208 In this part of the analysis, we only included the human proteins as it is the species covered by the
209 highest number of datasets. Based on their RBP2GO score, we selected the top 20% of the RBPs that
210 did not contain any of the selected RBDs (935 "high score" proteins of interest, Supplementary
211 Table S4) that we compared to three reference datasets: 1) the whole proteome (20741 proteins), 2)
212 the non-RBPs without RBD (14411 proteins) and 3) the bottom 20% RBPs lacking RBDs based on their
213 RBP2GO score (989 "low score" proteins, Supplementary Table S4). We extracted the InterPro IDs
214 present in the "high score" dataset and filtered out the IDs that were present in the initial RBD candidate
215 list. Next, we selected for the IDs of the type "Domain" or "Repeat", which resulted in a list of 637 IDs
216 (Supplementary Table S8). For each InterPro ID (domain/repeat), we calculated the proportion of
217 proteins containing the domain in the dataset of interest as well as in the three reference datasets. Next,
218 for each domain, we computed its fold enrichment in the dataset of interest as compared to each of the
219 reference datasets. A *P*-value was also generated as a result of a Fisher's exact test. The *P*-values
220 were further adjusted for multiple testing by the FDR method (44). To identify new RBDs, we applied
221 the following criteria: we selected the domains that were 1) significantly enriched against at least one
222 reference dataset (>2-fold enrichment and adjusted *P*-value <0.05), 2) present in at least 0.5% of the
223 proteins of the "high score" dataset (*i.e.*, in at least five proteins) and 3) presenting an RBP vs. non-
224 RBP hits ratio greater 1. Altogether, 15 domains satisfied these three criteria (Supplementary Table S8).

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

226 **Validation of the newly discovered RBDs using RNA-binding peptides**

227 Several studies featured in RBP2GO used approaches enabling the identification of the RNA-binding
228 region of the detected RBPs (13,14,16,32,45-47). Using these datasets, we computed a new dataset
229 combining the peptide information with the UniProt ID of the corresponding proteins and with the
230 coordinates of the peptides in the proteins (Supplementary Table S9). Hereby, we only considered the
231 human protein datasets. For each of the 624 proteins containing only the new RBDs (without containing
232 any of the previously selected RBDs), we computed the proportion of peptides overlapping at least 50%
233 with the new RBD relative to the total number of peptides found in the protein. This number was
234 normalized to the fraction of the protein sequence that was covered by the domain. We then summed
235 this number for all the domains corresponding to one InterPro ID over all proteins containing at least
236 one peptide, and normalized by the total number of occurrences of this domain in the proteins containing
237 at least one peptide. The same procedure was repeated for the selected RBDs and the non-selected
238 domains (Figure 1A) using groups of proteins containing only selected RBDs or only non-selected
239 domains respectively. To maintain the same selection criteria as for the new RBDs, we only took into
240 account the selected and non-selected RBDs that were present in at least five RBPs (Supplementary
241 Table S9). The results of these three analyses were used to produce Figure 7F-G.

242 **Computation of the RBP2GO composite score**

243 To reflect the importance of the presence of RBDs or the relation to an Rfam ID for an RBP, we created
244 a new "RBP2GO composite score" that integrates the knowledge on RBDs and Rfam IDs in addition to
245 the information already provided by the RBP2GO score. This new score comprises three components.
246 Component 1 is based on the listing count of the proteins (i.e., in how many datasets the protein is
247 detected as RBP). Component 2 is based on the average listing count for the top 10 interactors of the
248 protein, according to the STRING database (STRING-DB.org) (48). Component 3 provides a score
249 pertaining to the presence of RBDs or Rfam IDs that are associated to the protein of interest.

250 To define the component 3, a quality factor was attributed to each RBD / Rfam ID. To this aim, we took
251 into consideration the ratio of hits in RBPs versus hits in non-RBPs in human (Hs) as a surrogate
252 measure for the likelihood of the RBD or Rfam ID being truly associated with RBPs. If the ID was not
253 found in Hs, we considered the ratio in mouse (Mm); if it was not found in Mm, we considered the one
254 in yeast (Sc), and so on following the order of the species by decreasing number of studies
255 (Supplementary Figure S1). For each ID, this ratio was used to attribute a semi-quantitative balancing
256 quality factor: ratio <2 = quality factor of 1; ratio ≥2 = quality factor of 2; ratio ≥4 = quality factor of 3;
257 ratio ≥8 = quality factor of 4 and ratio ≥16 = quality factor of 5. For infinite ratios (ID only present in
258 RBPs), the number of RBPs with this ID was in addition taken into account: for a number of RBPs ≤2 =
259 quality factor of 2; number of RBPs from 2 to 4 = quality factor of 3; number of RBPs from 4 to 8 =
260 quality factor of 4 and number of RBPs >8 = quality factor of 5. The attribution rules are summarized in
261 the Supplementary Table S10. Finally, we summed the quality factors for all RBDs and RNA-related
262 family IDs present for one protein, and limited this number to 25.

263 In order to weight properly the contribution of each score, we applied a Random Forest algorithm in R,
264 using three independent datasets of RBPs and non-RBPs. Dataset 1: based on the proteins of the
265 RBP2GO database (22); dataset 2: based on the HydRA classification (49); dataset 3: based on
266 RBPbase superset (3). As a result of the Random Forest analysis and the calculated importance of the
267 components, while the RBP2GO composite score overall ranges from 0 to 100, component 1 can
268 amount to a maximum of 50, while component 2 and 3 can equally amount to a maximum of 25 (Figure 8
269 and Supplementary Table S11).

270 The Pearson correlation coefficients between the three components of the composite score and the
271 RBP2GO score were computed using the cor() function of the stats package in R. The results are
272 displayed in Supplementary Figure S9.

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

276 As an application of the RBP2GO composite score, the various proteome-wide strategies were grouped
277 according to their underlying methods, and the identified RBPs were analyzed, according to the three

278 components and the RBP2GO composite score itself (Supplementary Figure S10 and Supplementary
279 Table S12).

280 **Selection of high-confidence RBPs in *Homo sapiens* (Hs) and comparison with HydRA and**
281 **RBPbase RBPs**

282 We took advantage of the RBP2GO composite score to compute a list of selected high-confidence
283 human RBPs. This list encompassed all human proteins, whether they were detected as RBPs in
284 proteome-wide screens or not, that had a composite score superior or equal to 10, with at least two out
285 of the three components of the score larger than zero. This resulted in a list of 2018 proteins, among
286 which 1977 were already detected as RBPs in screens of the RBP2GO database (Supplementary
287 Table S13, Supplementary Figure S11). We compared the high-confidence RBP assignment to other
288 approaches, such as HydRA and RBPbase. The distributions of the RBP2GO composite scores and
289 HydRA scores were calculated for proteins with both scores available (Supplementary Figure S12).

290 **Update of the RBP2GO database**

291 The whole dataset compiled in this analysis is available as Supplementary Table S14. In order to make
292 the data more accessible and searchable, this new dataset was included in a user-friendly and intuitive
293 manner into the RBP2GO database. The "RBD_Status" column as well as the number of RBDs and the
294 RBD content fractions were updated to take into account the identified RBDs (Supplementary Table S8)
295 in addition to the selected RBDs and Rfam IDs. The new data, including the lists of RBDs and Rfam
296 IDs can be directly downloaded from the download section of the database (https://RBP2GO-2-
297 Beta.dkfz.de). It is now also possible via the advanced search option from each species to search for
298 proteins containing a specific InterPro ID, or containing an RBD and/or an Rfam ID. A new section with
299 information about the selection procedures applied to the RBDs and Rfam IDs was also created under
300 the sidebar menu "search domains", which include a specific search module for InterPro IDs.

301 For easy access to the analyze and the datasets, the whole script is available from the download section
302 of the RBP2GO database from the sidebar menu.

303 **RESULTS**

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

305 Prior to the increase in the number of proteome-wide methods dedicated to the identification of RBPs
306 and the resulting knowledge gained regarding RNA-protein interactions, RBPs were thought to interact
307 with RNA essentially via well-defined globular RBDs. Some proteins harbor a single RBD while others

308 contain several RBDs, combined in a modular manner, which illustrates the great variability of RBD-
309 RNA interactions (24,26).

310 Based on the datasets compiled in our recently published RBP2GO database (22), RBP candidates
311 were defined as proteins that were at least detected in one RNA-interactome dataset (Listing Count >=
312 1 in the RBP2GO database). RBP candidates are thus not necessarily validated experimentally. For
313 more simplicity, RBP candidates are named RBPs in the following. Non-RBPs were defined as the
314 proteins that had never been identified in the proteome-wide experimental studies (Listing Count = 0).
315 For each protein, we also defined an RBP2GO score, that reflects two aspects: first, we reasoned that
316 proteins that are independently listed in multiple datasets must have higher probabilities of being true
317 RBPs. Second, based on the finding of the SONAR study (25) and our previous analysis of the protein–
318 protein interactions within the CORUM database (19), proteins interacting with multiple RBPs are
319 frequently RBPs themselves. Accordingly, we computed and provided for each protein two separate
320 indicators of its RBP propensity: the count of the protein itself being listed as RBP (Listing Count) and
321 the average listing count of the up to ten interaction partners with the highest STRING interaction
322 database scores (AVG10 Int Listing Count). These two components were combined with an equal
323 weight in the RBP2GO score, which was then normalized to the number of datasets in the respective
324 species. As a result, the RBP2GO score can range from 0 to 100.

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

339 We evaluated this selection of RBD candidates using the RBP2GO score of the proteins containing the
340 selected RBDs in species with a higher number of proteome-wide RBP datasets such as *Homo sapiens*,
341 *Mus musculus*, *Saccharomyces cerevisiae*, *Drosophila melanogaster* (Supplementary Figure 1),
342 hereafter also referred to as the "well-studied species". We confirmed that these proteins had a
343 significantly higher RBP2GO score than the proteins with unselected or no RBDs, which corroborated
344 our selection procedure (Figure 1B).

345 **The proportion of RBD-containing proteins in proteomes differs across species**

346 Next, we analyzed the proportion of the selected RBDs in the proteins of the RBP2GO database that
347 were classified as RBP (Listing Count >= 1) or non-RBP (Listing Count = 0). Accordingly, the protein
348 can be grouped into RBPs with RBD (RBPs containing at least one selected RBD), RBPs without RBD,
349 non-RBPs with RBD and non-RBPs without RBD. While some well-studied species such as human and
350 mouse (Supplementary Figure S1) harbored more than 65% of the RBDs only in the RBPs (Figure 2A
351 and Supplementary Table S15), this seemed to be inverted in less studied species. In zebrafish (*Danio*
352 *rerio*), 72% of the RBDs were found only in non-RBPs (Figure 2A and Supplementary Table S15).

353 This disparity between well-studied species and less-studied species was also reflected in the
354 proportion of RBPs and RBD-containing proteins in the proteomes. The *Homo sapiens* proteome
355 contained 3.5-fold less RBD-containing proteins than RBPs, while *Danio rerio* and *Trypanosoma brucei*
356 inversely depicted >4.5-fold more RBD-containing proteins than RBPs (Supplementary Table S16).
357 However, in the group of less-studied species, *Escherichia coli*, which had 5-fold more RBPs than RBD-
358 containing proteins was an exception (Figure 2B).

359 Interestingly, in the well-studied species, while the proportion of RBPs in the proteome was higher than
360 for most of the less-studied species, only a small proportion of the RBPs contained an RBD (<10% of
361 the proteome, Figure 2C, dark orange group of proteins and Supplementary Table S16). Most RBPs
362 did not contain an RBD, which strongly indicates that RBPs employ alternative ways in addition to
363 already known RBDs to interact with RNA molecules.

364 The four most abundant RBDs in the RBPs of our dataset appeared all to be canonical RBDs (26,28),
365 namely, the RNA recognition motif domain (RRM, IPR000504, found in 1700 proteins), Helicase
366 superfamily 1/2 ATP-binding domain (IPR014001, found in 1079 proteins), Helicase C-terminal
367 (IPR001650, found in 1060 proteins) and DEAD/DEAH box helicase domain (IPR011545, found in 701
368 proteins, Supplementary Table S4). Here again, there was a clear discrepancy between the well-
369 studied species and less-studied species with regard to the experimental identification of proteins
370 containing such domains as RBPs. In the well-studied species, more than 50% of the proteins harboring
371 at least one of these four canonical RBDs were classified as RBPs (Figure 2D and Supplementary
372 Table S4). For many less-studied species, the percentage of RBPs dropped down to less than 15% of
373 the proteins. Thus, the proteomes of the less-studied species may contain a number of yet undiscovered
374 RBPs, including proteins containing selected RBDs (Supplementary Table S3).

375 In humans, 231 RBD-containing proteins had not been detected yet in the published RBP screens
376 (Supplementary Table S16). We hypothesized that the expression levels of the proteins might impact
377 their potential of being detected in the mass spectrometry-based studies. Therefore, we analyzed the
378 protein expression levels (37) of the four different groups: non-RBPs without RBD, non-RBPs with RBD,
379 RBPs without RBD and RBPs with RBD. Both non-RBP groups had significantly lower expression levels
380 (Figure 2E), which might potentially explain why the non-RBPs with RBD had not been detected as

381 RBPs yet. Similarly, RBP2GO scores, which reflect the probability for a protein to be an RBP (22),
382 positively correlated with the protein expression levels (Supplementary Figure S2).

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

387 **Proteins with RBD appear as stronger RBPs**

388 Taking advantage of the RBP2GO score (22), we evaluated its distribution in the different species and
389 found that RBD-containing proteins had a significantly higher RBP2GO score in all species as compared
390 to proteins without RBD (Figure 3A and Supplementary Figure S3A). This was confirmed when the non-
391 RBPs and RBPs were split into two groups, depending on their RBD content. In each group respectively,
392 the RBD-containing proteins had significantly higher RBP2GO scores, with the exception of *Danio rerio*
393 and *Trypanosoma brucei* (Figure 3B and Supplementary Figure S3B). This was, however, linked to the
394 low number of proteins detected as RBPs so far in these two species (Figure 2B and Supplementary
395 Table S16).

396 To obtain a more detailed view of the interplay between the presence of RBDs in proteins and the
397 RBP2GO score, we evaluated the proportion of RBD-containing proteins for each interval of the
398 RBP2GO score. We found a strong positive correlation between the proportion of RBD-containing
399 proteins and the RBP2GO score up to a score of 50 (Supplementary Figure S3C). Above 50, for the
400 four well-studied species and *Caenorhabditis elegans*, >80% of the proteins contained an RBD
401 (Supplementary Figure S3D and Supplementary Table S17). This result supports the assumption that
402 RBDs can be used to assess the RNA-binding potential of an RBP, but also to identify new RBP
403 candidates in the non-RBPs.

404 Exploiting detailed information from the InterPro database (33), we compiled the number of RBDs per
405 protein, as well as the fraction of the protein sequence they covered. We analyzed this data in RBPs
406 versus non-RBPs and observed that RBPs had a higher RBD content fraction than non-RBPs, a
407 difference that was significant for all species except *Plasmodium falciparum*, *Escherichia coli* and
408 *Trypanosoma brucei* (Figure 3C and Supplementary Figure S3E). Moreover, in human proteins, the
409 RBP2GO score increased with the number of RBDs per protein, whereas this was not visible for the
410 non-RBPs (Figure 3D).

411 Taken together, these results highlight the strong positive correlation between the RBP2GO score and
412 the number of RBDs in RBPs. This underlines the potential of RBDs for the evaluation of the RNA-
413 binding ability of proteins, including non-RBPs with RBD, that might not be properly detected if lowly
414 expressed.

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

416 In addition to the information relative to the protein domains and repeats, the InterPro database provides
417 annotation for protein families. In the initial lists of RNA binding-related InterPro IDs, 1028 IDs
418 corresponded to "family" IDs (Supplementary Table S1). We subjected these IDs to the selection
419 procedure, which we previously applied to the "domain" and "repeat" IDs (Figure 1A and Supplementary
420 Figure S4A). We obtained a list of 672 selected RNA-related "family" IDs, referred to as Rfam IDs
421 (Supplementary Tables S6 and Supplementary Figure S4A). The distribution of the Rfam IDs across
422 RBPs and non-RBPs in the 13 RBP2GO species exhibited a similar pattern as for the RBDs (Figures
423 2A and 4A). The well-studied species displayed a large proportion of the Rfam IDs in RBPs only. 98.2%
424 of the Rfam IDs in *Homo sapiens* and more than 65% in *Mus musculus* were found in RBPs. In less-
425 studied species such as *Trypanosoma brucei*, less than 7% of the Rfam IDs were related to RBPs only
426 (Figure 4A and Supplementary Table S18). These observations confirmed the importance of a good
427 study coverage for proper classification of proteins into RBPs or non-RBPs. Strikingly, 70% of the
428 human RBPs were not linked to RNA-related InterPro ID (Supplementary Figure S4B).

429 The proteins annotated with an Rfam ID had a significantly higher RBP2GO score than the ones without
430 Rfam ID (Figure 4B and Supplementary Figure S5A). This was also verified when splitting the proteins
431 into non-RBPs and RBPs (Supplementary Figure S5B-C), except for the human non-RBPs, for which
432 the difference was not significant mainly due to the small number of non-RBPs with Rfam ID.

433 There was also a strong positive correlation between the RBP2GO score and the proportion of proteins
434 associated with Rfam ID (Supplementary Figure S5D). For the well-studied species and within the
435 proteins with RBP2GO scores >50, proteins with RBD were more abundant than proteins with Rfam ID
436 (Supplementary Figure S3A and S5D). However, proteins with Rfam ID annotations seemed to have
437 higher median RBP2GO scores than proteins with RBD only (Figure 4C and Supplementary
438 Figure S5F). In addition, in five out of nine species, RBPs annotated with both, RBDs and Rfam IDs,
439 had a significantly higher RBP2GO score than the other RBP groups, indicating a cumulative effect of
440 the dual annotation (Figure 4C and Supplementary Figure S5F).

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

446 **Disordered regions are enriched in proteins with RBD**

447 While RBDs and Rfam IDs were both strong indicators of the RNA-binding potential of proteins, our
448 analysis revealed a large group of RBPs (*e.g.,* 70% in human) that were lacking such features
449 (Supplementary Figure S4B and Supplementary Table S16). In the past decades, protein regions

450 among those regions referred to as IDRs were experimentally detected as directly interacting with RNA
451 (13,32). IDRs are defined according to their structural properties. Unlike globular domains, IDRs do not
452 fold into a specific three-dimensional structure, but rather remain lose and confer flexibility to proteins
453 (50). Their structure can stabilize upon binding to interaction partners (51). Some IDRs are implicated
454 in RNA binding for example via specific repeats such as the arginine and serine (RS) repeats in the NF-
455 kappa-B-activating protein (NKAP), which plays an important role in splicing (30,31). Also, IDRs that
456 contain repeats of arginine and glycine, referred to as RGG-box, can also bind directly to RNA. Such
457 IDR repeats are for example found in the fragile X mental retardation protein (FMRP/FMR1) (30,52).
458 This motivated us to analyze the IDR content of the proteins.

459 To do so, we retrieved the data computed by the MobiDB-lite algorithm (34,40). The analysis of the IDR
460 content of the proteins revealed that RBD-containing proteins had a significantly higher IDR fraction
461 than proteins lacking RBDs, regardless of their status as RBP or non-RBP, in most species but
462 especially in the well-studied species (Supplementary Figure S6). This is consistent with previous
463 findings reporting the enrichment of IDRs in RBPs (13,53). In most species, the RBPs without RBD did
464 not exhibit a higher disordered fraction than the RBD-containing non-RBPs, indicating that their RNA-
465 binding ability was probably independent of the IDR content (Supplementary Figure S6).

466 The combined analysis of the RBP2GO score of the RBPs with respect to the RBD and the IDR content
467 revealed that the presence of IDRs in addition to RBDs only had a minor positive contribution to the
468 increase of the RBP2GO score (Figure 5A and Supplementary Figure S7A). The minor contribution of
469 the IDR to the RBP2GO score of the proteins was further validated by the weak linear correlation
470 between the disordered fraction and the RBP2GO score for both the RBPs and the non-RBPs
471 (Supplementary Figure S7B).

472 According to our results, the IDRs appeared enriched in proteins that already contained RBDs (both
473 RBPs and non-RBPs). Therefore, the presence of IDRs in proteins seems largely redundant with the
474 presence of RBDs and does not represent a strong independent indicator of the RNA-binding potential
475 of proteins that would otherwise explain, for example, how RBPs lacking RBDs could bind RNA.

476 IDRs frequently contain repeats, i.e., the underlying sequences depict amino acid content bias. Such
477 sequences are also referred to as low complexity sequences or low complexity regions (LCRs). Since
478 specific LCRs can bind to RNA (31,52), we also investigated the LCR content of the human RBPs and
479 non-RBPs. We found that charged amino acids were slightly enriched in RBPs (Supplementary
480 Figure S8A). More specifically, we observed that RGG and YGG repeats were enriched in proteins with
481 RBD, suggesting a co-occurrence of LCRs with RBD (Supplementary Figure S8B), similar to the co-
482 occurrence of IDRs with RBD. In support of this finding, human proteins with RBD (both non-RBPs and
483 RBPs) had a significantly higher proportion of LCRs than the proteins lacking RBDs (Figure 5B).

484 **New RBP candidates can be identified from non-RBPs with RBD**

485 When splitting the 13 studied species into well-studied and less-studied species according to the
486 number of available datasets (Supplementary Figure S1), the four well-studied species summed up
487 together 1465 non-RBPs with RBD out of 5496 proteins with RBD (26.6%). For the nine less-studied
488 species, there were 6273 non-RBPs with RBD out of 9308 proteins with RBD (67.3%) (Figure 6A and
489 Supplementary Table S16). Adding the non-RBPs with RBD to the collection of RBPs would increase
490 the set of RBPs by 75% in the less-covered species as compared to only 10% in the well-studied
491 species.

492 In order to evaluate the potential of the non-RBPs with RBD as new RBP aspirants, we performed a
493 Gene Ontology (GO) (54) enrichment analysis on the non-RBPs with RBD in the eight species for which
494 data was available from the PANTHER resource (55,56). We report the terms that were enriched in at
495 least half of the species, sorted by the mean fold enrichment (Figure 6B-C). In the enriched GO terms
496 for molecular function and biological processes, many terms were related to nucleic acid binding and
497 metabolism such as "RNA polymerase activity" (as depicted in green in Figure 6B-C). In agreement
498 with these findings, several high-mobility group (HMG) chromosomal proteins were already listed as
499 RBPs in the RBP2GO database (22) and experimentally validated, i.e., HMGN1 (19). Also, DNA-binding
500 domains such as the high-mobility group (HMG)-box had already been associated to RNA-binding
501 (13,57). Inversely, heterogeneous ribonucleoproteins and double-stranded RNA-binding proteins could
502 also directly interact with DNA (58,59), illustrating that many proteins binding to DNA can also bind to
503 RNA and *vice versa* (12,60). GO terms related to histones and chromatin were enriched, as well, but
504 were less represented (orange, Figure 6B-C). RBPs can interact with chromatin and act in gene
505 regulation, as for example the RNA-binding protein 25 (RBM25) which enhanced the transcription-
506 dependent activities of the YY1 transcription factor (61-63). Metabolic activities also appeared to be
507 enriched in the non-RBPs with RBD (blue, Figure 6B-C). This is also in agreement with the identification
508 of several metabolic enzymes in RBPome studies and their further validation such as the iron regulatory
509 protein 1 (IRP1), the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and the glycolytic enzyme
510 Enolase 1 (ENO1)(64-67). Altogether, our GO analysis of the non-RBPs with RBD strongly indicated
511 that this group was enriched in molecular functions and biological processes reasonably linked to RNA
512 such that this group may contain further so far experimentally unrecognized RBPs, referred to as RBP
513 aspirants.

514 To further validate these findings, we found evidence in the literature for the RNA-binding properties of
515 seven RBD-containing non-RBPs, which contained different types of RBDs (Supplementary
516 Table S19). Individual studies experimentally confirmed the binding of these proteins to RNA (68-73),
517 in support of our assumption that RBP aspirants (non-RBPs with RBD) are indeed RBPs. Furthermore,
518 six out of these seven proteins exhibited IDRs, in agreement with the frequent association of IDRs and
519 RBDs (Supplementary Table S16) as suggested by the enrichment of IDRs in RBPs with canonical
520 RBDs (74) and confirmed in our analysis above (Figure 5A).

521 We showed here that a substantial number of non-RBPs with RBD, herein called RBP aspirants, are
522 probably RBPs that had not been detected in proteome-wide screens so far. Using our selected list of

523 RBDs as indicator of their RBP potential allowed us to identify 7738 new RBP aspirants, with more than
524 80% of them in less-studied species.

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

526 Since a large proportion of RBPs did not contain any of the selected RBDs, we hypothesized that they
527 may contain so far undetected RBDs. To test this hypothesis, we performed an enrichment analysis of
528 the "domain" and "repeat" InterPro IDs present in the top 20% of the human RBPs without RBD based
529 on their RBP2GO score. This group of 935 proteins was compared to three different reference sets: the
530 bottom 20% of the human RBPs based on their RBP2GO score, the non-RBPs without RBD and the
531 human proteome (Figure 7A and Supplementary Table S4). 15 domains were significantly enriched as
532 compared to at least one of the reference datasets. Within these domains, we found domains related
533 to chaperones and importins, along with domains related to the binding of cytoskeletal proteins
534 (Figure 7B and Supplementary Table S8). In particular, the WD40 repeat (IPR001680) does not exhibit
535 a role in a specific cellular process, but is found in RBPs such as the WD-repeat containing protein 43
536 (WDR43) (75). The histidine kinase/HSP90-like ATPase domain (IPR003594) is present in the HSP90A
537 protein which binds to RNA and is implicated in the loading of small RNAs into the RISC complexes
538 (76). These findings could corroborate the RNA-binding ability of the 15 identified RBD candidates. In
539 addition, a literature search based on these 15 RBD candidates revealed that the WD40 repeat
540 (IPR001680) indeed is directly binding to RNA within the protein Gem-associated protein 5 or gemin5
541 (77), another evidence in support of the utility of our pipeline to identify RBDs among domains that were
542 not yet annotated as such in the InterPro database. For the other 14 domains, we did not find RNA-
543 binding evidence in the literature.

544 To validate the domains as RBDs, we quantified the proportion of proteins containing these domains.
545 In the well-studied species, only a small proportion of RBPs harbored new RBD candidates as
546 compared to RBPs with RBD (Figure 7C). In those species, the proteins with new RBD candidates had
547 a significantly higher RBP2GO score than RBPs without RBD. However, their RBP2GO score was still
548 significantly lower than that of the proteins with known RBDs (Figure 7D). Also, proteins containing only
549 new RBD candidates were significantly enriched in RBPs as compared to non-RBPs (Figure 7E). These
550 results supported the hypothesis that the new RBD candidates can be true RBDs.

551 To further validate the newly identified RBD candidates, we took advantage of RNA-binding peptide
552 datasets available from human proteome-wide RBP studies (13,16,32,45). We calculated and
553 normalized the proportion of these peptides overlapping the newly discovered RBDs, as compared to
554 the overlap with RBD and InterPro domains that were not selected. The median proportion of
555 overlapping peptides for the newly discovered RBDs appeared significantly higher than for the non-
556 selected domains and not significantly different as compared to the selected RBDs (Figure 7F). For
557 example, RNA-binding peptides were mapped to the Armadillo repeat (IPR000225), one of the newly
558 identified RBDs (Figure 7G).

559 In summary, our pipeline identified 15 RBDs, of which 14 had not been previously linked to RNA. We
560 validated those new RBDs using both the RBP2GO score of the related proteins as well as published
561 RNA-binding peptide datasets. Consequently, we included them into the list of RBDs that can now be
562 explored on the updated RBP2GO database (https://RBP2GO-2-Beta.dkfz.de).

563 **A new RBP2GO composite score helps defining a high-confidence human RBP list**

564 In order to integrate the knowledge gained from the analysis of the RBDs and Rfam IDs, we computed
565 a new RBP2GO composite score for each protein. This score combines three components that are
566 based on experimental data directly related to the protein (component 1) and the top 10 interacting
567 proteins according to the STRING database (component 2), as well as data related to RBDs and Rfam
568 IDs (component 3, Figure 8A and Supplementary Table S10). Refer to the Material and Methods
569 section for a detailed description. As expected, per definition, the first two components of the RBP2GO
570 composite score strongly correlated with each other and with the RBP2GO score, but the correlation
571 was decreased with the component 3 indicating that it added another layer of information
572 (Supplementary Figure S9). The RBP2GO composite score allowed a better discrimination between
573 proteins with and without RBD (Figure 8B, Supplementary Table S14), providing a solid support for the
574 evaluation of RBPs. Moreover, while the previous RBP2GO score was only available for 9 out of 13
575 RBP2GO species due to a lack of study coverage (22), the component 3 of the RBP2GO composite
576 score is now provided for all species.

577 The RBP2GO composite score can be used to learn about the different proteome-wide strategies used
578 to identify RBPs with regards to the three components of the score. As depicted in Supplementary
579 Figure S10, the various methods depict different spectrum of components and RBP2GO composite
580 score. The distributions of the three components and the RBP2GO composite score present
581 differences, from strategies with higher scores for all components, to strategies scoring better for
582 components 1 and 2 than for component 3.

583 Finally, we took advantage of the RBP2GO composite score to define a high-confidence list of human
584 RBPs that contains 2018 proteins with an RBP2GO composite score >10 and with at least two
585 components >0 (Supplementary Table S13). As a validation of both the RBP2GO composite score and
586 the high-confidence list, we found that human ribosomal proteins had a significantly higher RBP2GO
587 composite score than all other proteins (Supplementary Figure S11) and were as expected all included
588 in the high-confidence list of human RBPs (Supplementary Table S13). We also compared our high-
589 confidence RBP assignment to the RBP assignment from HydRa (49) and the RBPbase superset (3).
590 The intersection of the three datasets contains 1796 proteins, which represent 89% of the RBP2GO
591 high-confidence human RBPs (Supplementary Figure S12A). We also found that the high-confidence
592 human RBPs from our analysis had both a significantly higher RBP2GO composite score and HydRA
593 score (Supplementary Figure S12B-C), confirming the quality of this list as a core catalog of human
594 RBPs.

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

597 Our analysis demonstrated that RBDs and Rfam IDs were strong indicators of the RNA-binding ability
598 of the related proteins. Therefore, we expanded the RBP2GO database (https://RBP2GO-2-
599 Beta.dkfz.de) to allow the users to browse the information related to RBDs and Rfam IDs in an intuitive
600 manner as well as to use these as search or filter criteria. Therefore, we added a dedicated section
601 which contains a search tool to display the RNA-binding status of InterPro IDs to the sidebar (Figure 8C-
602 D). We also implemented a tab panel in the species-specific "Protein search" to provide the available
603 information pertaining to the InterPro ID annotations as well as the RNA-binding properties of each
604 domain, the presence of disordered regions in the proteins and the RBP2GO composite score
605 (Figure 8E). Protein-specific links to the InterPro and MobiDB databases are provided to simplify access
606 to complementary data if needed. Finally, we generated several options in the "Advanced search" for
607 all the species. The users can navigate through the proteins of the database using a list of specific
608 InterPro IDs and can specifically select proteins that contain RBDs and/or Rfam IDs.

609 **DISCUSSION**

610 RBPs are important interaction partners for RNA transcripts from the sites of RNA production until their
611 degradation. It is evident that RNA-protein interactions are relevant in numerous key cellular processes.
612 Thus, defects in RBPs are often associated with severe disease phenotypes such as muscular
613 dystrophies, neurological disorders or cancers (3,78). Consequently, it is essential to investigate the
614 composition of ribonucleoprotein complexes in order to understand the molecular mechanisms and
615 functions governed by RNA-protein interactions.

616 In this study, we present a detailed analysis of RNA-binding features in RBPs from 13 species including
617 datasets generated in both, eukaryotes and prokaryotes. Based on experimental knowledge on RBPs
618 and non-RBPs, we compiled a list of selected RBDs and Rfam IDs. Using the previously defined
619 RBP2GO score (22), we show that the presence of RBDs in proteins or their association with Rfam ID
620 particularly highlights their RNA-binding potential. The number of RBDs per RBP also shows a positive
621 correlation with the RBP2GO score in human RBPs, which is supported by the modularity of canonical
622 RBDs, with several of them often working together within the same protein to more efficiently bind to
623 the target RNAs (26,79).

624 The distribution of the RBDs between RBPs and non-RBPs in the different species, as well as the higher
625 proportion of non-RBPs with RBD in the less-studied species, confirms the importance of a good study
626 coverage for the proper classification of the proteins into RBPs and non-RBPs. One notable exception
627 to this statement is *Escherichia coli*, which had only 48 non-RBPs with RBD despite a low study
628 coverage with only three RBP screens (16,80,81). This unique situation for *Escherichia coli* may be

629 explained by a significantly lower number of RBD-containing proteins as compared to other species. In
630 the *Escherichia coli* dataset, there are only 247 RBD-containing proteins in total as compared to 1938
631 proteins in the *Danio rerio* dataset with also three RBP screens. The difference may be due to the size
632 of the respective proteomes (4522 proteins for *Escherichia coli* versus 25801 for *Danio rerio*). However,
633 almost 80% of the RBD-containing proteins in *Escherichia coli* have been identified (Supplementary
634 Table S20), as compared to only 11% in *Dani rerio* (Supplementary Table S4). The high RBP
635 identification rate in *Escherichia coli* seems to be due to one study (80), which detected 1107 RBPs
636 using a method based on silica beads purification. These RBPs cover 98% of the RBD-containing RBPs
637 (185 out of 189) and 78% of the RBD-containing proteins in general (185 out of 237). The same study
638 identified 1448 RBPs in *Saccharomyces cerevisiae*, that cover 73% of the RBD-containing RBPs (438
639 out of 604) in this species (Supplementary Table S20). The RBPs from this study, that do not contain
640 RBDs, will nonetheless need further experimental validation.

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

654 Our analysis of IDRs demonstrated that, among RBPs, the RBPs with RBD were enriched in IDRs as
655 compared to the proteins lacking RBDs, which is in agreement with previous studies (74,82,83).
656 However, our data also showed an enrichment in IDRs for non-RBPs with RBD as compared to the
657 RBPs without RBD. Since IDRs confer flexibility to a protein, they can link different domains, such as
658 RBDs as documented for the RNA-binding protein FUS between its RRMs (84). IDRs can also
659 collaborate with canonical RBDs for the recognition of RNA and participate in the specificity of RBPs
660 (85). Based on our current analysis, it seems that IDRs cannot explain the RNA-binding properties of a
661 large number of RBPs lacking RBDs. However, specific types of disordered regions containing LCRs
662 such as arginine/serine (RS) repeats or RGG-boxes can bind to RNA (30,66,86). Such LCRs seem to
663 correlate as well with the presence of RBDs in the proteins. Presently, the MobiDB-lite algorithm, which
664 we used, provides IDR subtypes based on the enrichment in specific amino acids and charge, but not
665 on specific repeats (34,87). Thus, future options to further investigate the impact of IDRs and LCRs in
666 RNA-protein interactions could be to include a detailed analysis of the different repeat subtypes.
667 Altogether, the question of how RBPs without RBD are interacting with RNA remains unclear and of

668 great interest for future investigations, especially for proteins with so far unexpected affinity for RNA
669 and functionally "riboregulated" (67).

670 Next, we also took advantage of our comprehensive dataset to identify new RBDs within RBPs without
671 RBD but with high RBP2GO scores. Among them, some were found in proteins already known to bind
672 to RNA, such as the histidine kinase/HSP90-like ATPase domain (IPR003594) present in HSP90A (76).
673 The new RBDs were validated using experimental datasets describing RNA-bound protein peptides
674 (13,16,32,45), demonstrating the utility of computational meta-analyses as orthogonal approach
675 besides experimental studies.

676 Last but not least, we generated the new RBP2GO composite score, which integrates both the previous
677 experimental analysis (components 1 and 2) and the new sequence-based RNA-related analysis
678 (component 3). This new score now allows refining the pool of existing RBPs and identifying new RBPs,
679 which are two independent results. We also applied the RBP2GO composite score to compute a high-
680 confidence list of 2018 RBPs, thereby answering the need for a comprehensive catalog of human RBPs
681 (Supplementary Table S13), which will help selecting candidates for further validation and
682 characterization. In the advanced version of the RBP2GO database, the component of the RBP2GO
683 composite score that is based on the quality factors of the related InterPro IDs was implemented also
684 for species currently lacking an RBP2GO score. We believe that this additional information will facilitate
685 the selection of RBPs for the species with currently low study coverage.

686 In summary, this study first provides a pan-species list of selected RBDs and Rfam IDs enriched in
687 RBPs and then exploits this resource (i) to refine the list of RBPs from proteome-wide studies with
688 higher confidence, (ii) to predict new RBP aspirants from the group of RBD-containing non-RBPs and
689 (iii) to identify and validate new RBDs from the large group of RBPs lacking selected RBDs. The updated
690 resource is freely available online (https://RBP2GO-2-Beta.dkfz.de) and can be explored in an
691 interactive way. We anticipate that this analysis and the compilation of the information in a unifying
692 platform will stimulate our understanding of RBPs, their functions in key cellular processes and their
693 implications in human diseases.

694 **DATA AVAILABILITY**

695 RBP2GO 2.0 in a beta version is available at https://RBP2GO-2-Beta.dkfz.de.

696 The scripts for the main analysis, the production of the figures and the RBP2GO Shiny App are directly
697 available on the RBP2GO-2-Beta database under the DOWNLOAD menu from the sidebar.

698 **SUPPLEMENTARY DATA**

699 Supplementary data comprises the following items:

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

703 **Supplementary Figure S2, related to Figure 2.** Scatterplot of the RBP2GO scores of human proteins
704 versus their expression in HeLa cells in the four main groups of proteins (non-RBPs without RBD, non-
705 RBPs with RBD, RBPs without RBD and RBPs with RBD). R represents the Pearson's correlation
706 coefficient. The data used to generate this figure can be found in the Supplementary Table S5.

707 **Supplementary Figure S3, related to Figure 3** Analysis of the distribution of the RBDs in the proteins
708 of the RBP2GO database species (extended to all species). (**A**) Boxplots representing the distribution
709 of the RBP2GO score in proteins without RBD (light grey) and with RBD (dark grey). (**B**) Same as in
710 (**A**), but separated into non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs
711 without RBD (light orange) and RBPs with RBD (dark orange). (**C**) Proportion of RBD-containing
712 proteins for each interval of the RBP2GO score for the well-studied species. The red line represents a
713 linear regression (up to an RBP2GO score of 50). R represents the Pearson's correlation coefficient.
714 (**D**) Proportion of RBD-containing proteins for each interval of the RBP2GO score for the less-studied
715 species. The red line represents a linear regression (up to a RBP2GO score of 50). R represents the
716 Pearson's correlation coefficient. (**E**) Boxplots representing the distribution of the number of RBDs per
717 protein in non-RBPs (blue) and RBPs (orange). *, **, *** and **** correspond to *P*-values < 0.05, 0.01,
718 0.001 and 0.0001, respectively, ns corresponds to non-significant, as resulting from a Wilcoxon rank
719 sum test.

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

729 **Supplementary Figure S5, related to Figure 4.** Analysis of the selected Rfam IDs in the proteins of
730 the RBP2GO database species. (**A**) Boxplots representing the distribution of the RBP2GO score in
731 proteins without Rfam ID (no Rfam ID, light grey) and with Rfam ID (Rfam ID, dark grey). (**B**) Boxplots

representing the distribution of the RBP2GO score in 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) in the four well-studied species. (**C**) Same as in (**B**) for the less-studied species. (**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). After a score of 50, it does not appear to be linear anymore. R represents the Pearson's correlation coefficient. (**E**) Boxplots representing the distribution of the RBP2GO score in 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. (**A**) Analysis of the disordered fraction in the proteins of the RBP2GO database species (in human and mouse). 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. (**B**) Same as in (**A**) extended to all other species. *, **, *** 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 RBP2GO score of the proteins vs the presence of IDR in the proteins of the RBP2GO database species (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 5. Analysis of the LCR content of the proteins. (**A**) Boxplots showing the distribution of the five groups of amino acids (- charged, + charges, aromatic, uncharged non-polar and uncharged polar) in the four main groups of proteins: non-RBPs without RBD

766 (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light orange) and RBPs with RBD
767 (dark orange). (**B**) Frequency of the motifs FGG, RGG and YGG in the four main groups of proteins as
768 above in (A). *, **, *** and **** correspond to *P*-values < 0.05, 0.01, 0.001 and 0.0001, respectively, ns
769 corresponds to non-significant, as resulting from a Wilcoxon rank sum test.

770 **Supplementary Figure S9, related to Figure 8.** Pearson's correlation coefficients between the three
771 components of the RBP2GO composite score and the RBP2GO score. The coefficients are depicted
772 from blue (low) to red (high).

773 **Supplementary Figure S10, related to Figure 8.** Boxplots representing the distribution of the three
774 components and the RBP2GO composite score in the proteins as detected using various proteome-
775 wide strategies (method category).

776 **Supplementary Figure S11, related to Figure 8.** Violin plots showing the distribution of the three
777 components and the RBP2GO composite score in the ribosomal proteins as compared to the other
778 proteins. **** correspond to *P*-values < 0.0001, as resulting from a Wilcoxon rank sum test.

779 **Supplementary Figure S12, related to Figure 8.** Comparative analysis of RBP2GO, HydRA and
780 RBPbase. (**A**) Venn diagram depicting the overlap between the RBP2GO high-confidence RBPs, the
781 HydRA RBPs and the RBPbase superset of RBPs. (**B**) Boxplot depicting the RBP2GO composite score
782 for these three groups of RBPs (only proteins with available RBP2GO composite score and HydRA
783 score). (**C**) HydRA score for the same three groups of RBPs as in (**B**). *, ** and *** correspond to *P*-
784 values < 0.05, 0.01 and 0.001, respectively, ns corresponds to non-significant, as resulting from a
785 Wilcoxon rank sum test.

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

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

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

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

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

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

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

799 **Supplementary Table S8, related to Figure 7.** List of InterPro IDs identified as RBD candidates and
800 their enrichment (three last columns TRUE).

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

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

805 **Supplementary Table S11, related to Figure 8.** Results of the random forest analysis, including the
806 reference lists of proteins used to determine the importance of each parameter.

807 **Supplementary Table S12, related to Figure S10.** Analysis of the RBP2GO composite score within
808 the groups of RBPs detected using various methods.

809 **Supplementary Table S13, related to Figure 8.** List of high-confidence human RBPs.

810 **Supplementary Table S14, related to Figure 8B.** List of the updated proteins of the RBP2GO
811 database with the updated RBDs, RBP2GO score and RBP2GO composite score.

812 **Supplementary Table S15, related to Figure 2A.** Distribution of RBDs in the RBP2GO species into
813 the three protein categories: RBPs only, non-RBPs only, and both (RBPs and non-RBPs).

814 **Supplementary Table S16, related to Figure 3 to 5 and Supplementary Figures S3 to S7.** Left
815 (columns A:P): distribution of RBDs and Rfam IDs in the RBP2GO proteins. Right (columns T:AI):
816 distribution of the RBDs and IDRs in the RBP2GO proteins.

817 **Supplementary Table S17, related to Figure S3.** Analysis of the proteins with RBD and RBP2GO
818 score > 50.

819 **Supplementary Table S18, related to Figure 4.** Distribution of Rfam IDs in the RBP2GO species into
820 the three protein categories: RBPs only, non-RBPs only, and both (RBPs and non-RBPs).

821 **Supplementary Table S19, related to Figure 6.** List of non-RBPs with RBD that were validated as
822 RBPs by literature search.

823 **Supplementary Table S20, related to the Discussion.** List of RBPs detected in the Ec (top table) and
824 Sc (bottom table) proteome-wide studies. The table also include the total number of RBPs detected in
825 Ec and Sc, as well as the total number of RBD-containing proteins (highlighted in yellow).

826 **AUTHOR CONTRIBUTIONS**

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

833 **ACKNOWLEDGEMENT**

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

838 **FUNDING**

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

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

845 **REFERENCES**

846 1. Kishore, S., Luber, S. and Zavolan, M. (2010) Deciphering the role of RNA-binding proteins in
847 the post-transcriptional control of gene expression. *Brief Funct Genomics*, **9**, 391-404.
848 2. Castello, A., Fischer, B., Hentze, M.W. and Preiss, T. (2013) RNA-binding proteins in
849 Mendelian disease. *Trends Genet*, **29**, 318-327.
850 3. Gebauer, F., Schwarzl, T., Valcarcel, J. and Hentze, M.W. (2021) RNA-binding proteins in
851 human genetic disease. *Nat Rev Genet*, **22**, 185-198.
852 4. Gandhi, M., Gross, M., Holler, J.M., Coggins, S.A., Patil, N., Leupold, J.H., Munschauer, M.,
853 Schenone, M., Hartigan, C.R., Allgayer, H. *et al.* (2020) The lncRNA lincNMR regulates
854 nucleotide metabolism via a YBX1 - RRM2 axis in cancer. *Nat Commun*, **11**, 3214.
855 5. Klingenberg, M., Gross, M., Goyal, A., Polycarpou-Schwarz, M., Miersch, T., Ernst, A.S.,
856 Leupold, J., Patil, N., Warnken, U., Allgayer, H. *et al.* (2018) The Long Noncoding RNA
857 Cancer Susceptibility 9 and RNA Binding Protein Heterogeneous Nuclear Ribonucleoprotein
858 L Form a Complex and Coregulate Genes Linked to AKT Signaling. *Hepatology*, **68**, 1817-
859 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., Vries, I.S.N.-d., Simons, N., Eichelbaum, K., Föhr, 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. *Molecular & Cellular 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*, 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.T. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression: The RRM domain, a plastic RNA-binding platform. *FEBS Journal*, **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. Järvelin, A.I., Noerenberg, M., Davis, I. and Castello, A. (2016) The new (dis)order in RNA regulation. *Cell Communication and Signaling*, **14**, 9.
31. Burgute, B.D., Peche, V.S., Steckelberg, A.-L., Glöckner, 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 Research*, **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., Clementel, D., Dosztányi, Z. and Tosatto, S.C.E. (2021) MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavors in proteins. *Bioinformatics*, **36**, 5533-5534.
35. 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.
36. UniProt Consortium. (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*, **51**, D523-D531.
37. Bekker-Jensen, D.B., Kelstrup, C.D., Batth, T.S., Larsen, S.C., Haldrup, C., Bramsen, J.B., Sørensen, K.D., Høyer, S., Ørntoft, T.F., Andersen, C.L. *et al.* (2017) An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Systems*, **4**, 587-599.e584.
38. 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. *F1000Research*, **6**, 1025.
39. Predictors, C., DisProt, C., Necci, M., Piovesan, D. and Tosatto, S.C.E. (2021) Critical assessment of protein intrinsic disorder prediction. *Nature Methods*, **18**, 472-481.
40. Piovesan, D., Necci, M., Escobedo, N., Monzon, A.M., Hatos, A., Mičetić, I., Quaglia, F., Paladin, L., Ramasamy, P., Dosztányi, Z. *et al.* (2021) MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Research*, **49**, D361-D367.
41. Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
42. Kassambara, A. (2023) ggpubr: 'ggplot2' Based Publication Ready Plots.
43. Rezwani, M., Pourfathollah, A.A. and Noorbakhsh, F. (2022) rbioapi: user-friendly R interface to biologic web services' API. *Bioinformatics*, **38**, 2952-2953.
44. 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.
45. 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.
46. 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.
47. 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.
48. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P. *et al.* (2021) The STRING database in 2021: customizable

protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*, **49**, D605-D612.

49. Jin, W., Brannan, K.W., Kapeli, K., Park, S.S., Tan, H.Q., Gosztyla, M.L., Mujumdar, M., Ahdout, J., Henroid, B., Rothamel, K. *et al.* (2023) HydRA: Deep-learning models for predicting RNA-binding capacity from protein interaction association context and protein sequence. *Mol Cell*, **83**, 2595-2611 e2511.
50. Dunker, A.K., Babu, M.M., Barbar, E., Blackledge, M., Bondos, S.E., Dosztányi, 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 Disordered Proteins*, **1**, e24157.
51. 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.
52. 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.
53. 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. *Frontiers in Pharmacology*, **10**, 140.
54. Gene Ontology, C. (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res*, **49**, D325-D334.
55. 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.
56. 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.
57. 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.
58. 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.
59. Fierro-Monti, I. and Mathews, M.B. (2000) Proteins binding to duplexed RNA: one motif, multiple functions. *Trends Biochem Sci*, **25**, 241-246.
60. Hendrickson, G.D., Kelley, D.R., Tenen, D., Bernstein, B. and Rinn, J.L. (2016) Widespread RNA binding by chromatin-associated proteins. *Genome Biol*, **17**, 28.
61. 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.
62. 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.
63. 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.
64. Castello, A., Hentze, M.W. and Preiss, T. (2015) Metabolic Enzymes Enjoying New Partnerships as RNA-Binding Proteins. *Trends in Endocrinology & Metabolism*, **26**, 746-757.
65. 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.
66. 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 Biology*, **9**, 190096.
67. 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. *Molecular Cell*, **doi:10.1016/j.molcel.2022.05.019**.
68. 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. *Molecular Human Reproduction*, **14**, 331-336.
69. 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

1038 of Splicing Networks Reveals a Mechanism for Recognition of Autism-Misregulated Neuronal
1039 Microexons. *Molecular Cell*, **72**, 510-524.e512.
1040 70. Shen, H., Zheng, X., Luecke, S. and Green, M.R. (2010) The U2AF35-related protein Urp
1041 contacts the 39 splice site to promote U12-type intron splicing and the second step of U2-type
1042 intron splicing. *Genes & Dev*, **24**, 2389-2394.
1043 71. Dittmar, K.A., Jiang, P., Park, J.W., Amirikian, K., Wan, J., Shen, S., Xing, Y. and Carstens,
1044 R.P. (2012) Genome-Wide Determination of a Broad ESRP-Regulated Posttranscriptional
1045 Network by High-Throughput Sequencing. *Molecular and Cellular Biology*, **32**, 15.
1046 72. Morgan, M. (2017) mRNA 3′ uridylation and poly(A) tail length sculpt the mammalian
1047 maternal transcriptome. **548**, 347-351.
1048 73. Ansel, K.M., Pastor, W.A., Rath, N., Lapan, A.D., Glasmacher, E., Wolf, C., Smith, L.C.,
1049 Papadopoulou, N., Lamperti, E.D., Tahiliani, M. *et al.* (2008) Mouse Eri1 interacts with the
1050 ribosome and catalyzes 5.8S rRNA processing. *MOLECULAR BIOLOGY*, **15**, 8.
1051 74. Neelamraju, Y., Hashemikhabir, S. and Janga, S.C. (2015) The human RBPome: From genes
1052 and proteins to human disease. *Journal of Proteomics*, **127**, 61-70.
1053 75. Jain, B.P. and Pandey, S. (2018) WD40 Repeat Proteins: Signalling Scaffold with Diverse
1054 Functions. *The Protein Journal*, **37**, 391-406.
1055 76. Iwasaki, S., Kobayashi, M., Yoda, M., Sakaguchi, Y., Katsuma, S., Suzuki, T. and Tomari, Y.
1056 (2010) Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small
1057 RNA duplexes. *Mol Cell*, **39**, 292-299.
1058 77. Lau, C.K., Bachorik, J.L. and Dreyfuss, G. (2009) Gemin5-snRNA interaction reveals an RNA
1059 binding function for WD repeat domains. *Nat Struct Mol Biol*, **16**, 486-491.
1060 78. Pereira, B., Billaud, M. and Almeida, R. (2017) RNA-Binding Proteins in Cancer: Old Players
1061 and New Actors. *Trends in Cancer*, **3**, 506-528.
1062 79. Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y. and
1063 Yokoyama, S. (1999) Structural basis for recognition of the tra mRNA precursor by the Sex-
1064 lethal protein. *Nature*, **398**, 579-585.
1065 80. Shchepachev, V., Bresson, S., Spanos, C., Petfalski, E., Fischer, L., Rappsilber, J. and
1066 Tollervey, D. (2019) Defining the RNA interactome by total RNA-associated protein
1067 purification. *Mol Syst Biol*, **15**, e8689.
1068 81. The Gene Ontology, C. (2019) The Gene Ontology Resource: 20 years and still GOing
1069 strong. *Nucleic Acids Res*, **47**, D330-D338.
1070 82. Wang, C., Uversky, V.N. and Kurgan, L. (2016) Disordered nucleiome: Abundance of intrinsic
1071 disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and
1072 Archaea. *PROTEOMICS*, **16**, 1486-1498.
1073 83. Ho, W.L. and Huang, J.r. (2022) The return of the rings: Evolutionary convergence of
1074 aromatic residues in the intrinsically disordered regions of RNA‐binding proteins for liquid‐
1075 liquid phase separation. *Protein Science*, **31**, e4317.
1076 84. Loughlin, F.E., Lukavsky, P.J., Kazeeva, T., Reber, S., Hock, E.-M., Colombo, M., Von
1077 Schroetter, C., Pauli, P., Cléry, A., Mühlemann, O. *et al.* (2019) The Solution Structure of FUS
1078 Bound to RNA Reveals a Bipartite Mode of RNA Recognition with Both Sequence and Shape
1079 Specificity. *Molecular Cell*, **73**, 490-504.e496.
1080 85. Varadi, M., Zsolyomi, F., Guharoy, M. and Tompa, P. (2015) Functional Advantages of
1081 Conserved Intrinsic Disorder in RNA-Binding Proteins. *PLOS ONE*, **10**, e0139731.
1082 86. Zeke, A., Schád, É., Horváth, T., Abukhairan, R., Szabó, B. and Tantos, A. (2022) Deep
1083 structural insights into RNA‐binding disordered protein regions. *WIREs RNA*, e1714.
1084 87. Das, R.K. and Pappu, R.V. (2013) Conformations of intrinsically disordered proteins are
1085 influenced by linear sequence distributions of oppositely charged residues. *Proceedings of*
1086 *the National Academy of Sciences*, **110**, 13392-13397.

1087 **FIGURE LEGENDS**

1088 **Figure 1.** Selection of the RNA-binding domains. (**A**) Flow-chart showing the selection process of the
1089 RBD candidates and depicting the three-step selection procedure based on the ratios between RBPs

1090 and non-RBPs in the species available in the RBP2GO database (22). The starting lists of RBD
1091 candidates and the final list of selected RBDs are found in Supplementary Table S1 and Supplementary
1092 Table S3, respectively. (**B**) Boxplots depicting the distribution of the RBP2GO score of the proteins from
1093 the groups with no candidate RBDs (grey, no RBD), unselected candidate RBDs (orange) and selected
1094 RBDs (green) in Hs (*Homo sapiens*), Mm (*Mus musculus*), Sc (*Saccharomyces cerevisiae*) and Dm
1095 (*Drosophila melanogaster*). The numbers below the boxplots indicate the size of the groups. *, ** and
1096 *** correspond to *P*-values < 0.05, 0.01 and 0.001, respectively, as resulting from a Wilcoxon rank sum
1097 test.

1098 **Figure 2.** Distribution of the selected RBDs in the proteins of the RBP2GO database. (**A**) Number of
1099 selected RBDs present in each species and grouped according to the presence of the RBD either only
1100 in RBPs (orange), in both RBPs and non-RBPs (gray) or only in non-RBPs (blue). The dashed line
1101 represents the separation between the species with most (top) and fewer studies available. (**B**)
1102 Percentage of the proteome of the respective species comprised by RBPs (orange) or RBD-containing
1103 proteins (grey). The numbers at the end of the bars indicate the size of the respective proteomes in
1104 total. (**C**) Proportion in % of non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs
1105 without RBD (light orange) and RBPs with RBD (dark orange) in the different proteomes. (**D**) Fraction
1106 of RBPs among all proteins containing the indicated RBDs in the listed species. The percentages are
1107 depicted from black (high) to light grey (low). n.e. = non-existent, *i.e.,* the given domain is absent from
1108 the species. (**E**) Violin plots representing the protein expression level in HeLa cells according to a mass
1109 spectrometry experiment (37) in the four protein groups: non-RBPs without RBD (light blue), non-RBPs
1110 with RBD (dark blue), RBPs without RBD (light orange) and RBPs with RBD (dark orange). ***
1111 correspond to *P*-values < 0.001, ns corresponds to non-significant, as resulting from a Wilcoxon rank
1112 sum test.

1113 **Figure 3.** Correlation of the RBDs in the proteins of the well-studied species of the RBP2GO database
1114 with the RBP2GO score. (**A**) Boxplots representing the distribution of the RBP2GO score in proteins
1115 without RBD (light grey) and with RBD (dark grey). (**B**) Same as in (**A**), but separated into non-RBPs
1116 without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light orange) and RBPs
1117 with RBD (dark orange). (**C**) Boxplots representing the RBD content fraction per protein in non-RBPs
1118 with RBD (blue) and RBPs with RBD (orange). (**D**) Boxplots representing the distribution of the
1119 RBP2GO score per number of RBDs in non-RBPs (blue) and in RBPs (orange) in *Homo sapiens* (HS).
1120 *** corresponds to *P*-values < 0.001 as resulting from a Wilcoxon rank sum test.

1121 **Figure 4.** Distribution of the selected Rfam IDs in the proteins of the RBP2GO database. (**A**) Number
1122 of selected Rfam IDs present in each species and grouped according to the presence of the Rfam ID
1123 either only in RBPs (orange), in both RBPs and non-RBPs (gray) or only in non-RBPs (blue).

1124 (**B**) Boxplots representing the distribution of the RBP2GO score in proteins without Rfam ID (light grey)
1125 and with Rfam ID (dark grey). (**C**) Boxplots representing the distribution of the RBP2GO score in non-
1126 RBPs without RBD nor Rfam ID (light blue), non-RBPs with RBD but without Rfam ID (dark blue), non-
1127 RBPs without RBD but with Rfam ID (light green), non-RBPs with RBD and Rfam ID (dark green), RBPs
1128 without RBD nor Rfam ID (light orange), RBPs with RBD but without Rfam ID (dark orange), RBPs
1129 without RBD but with Rfam ID (light purple) and RBPs with RBD and Rfam ID (dark purple). *, ** and
1130 *** correspond to *P*-values < 0.05, 0.01 and 0.001, respectively, ns corresponds to non-significant, as
1131 resulting from a Wilcoxon rank sum test.

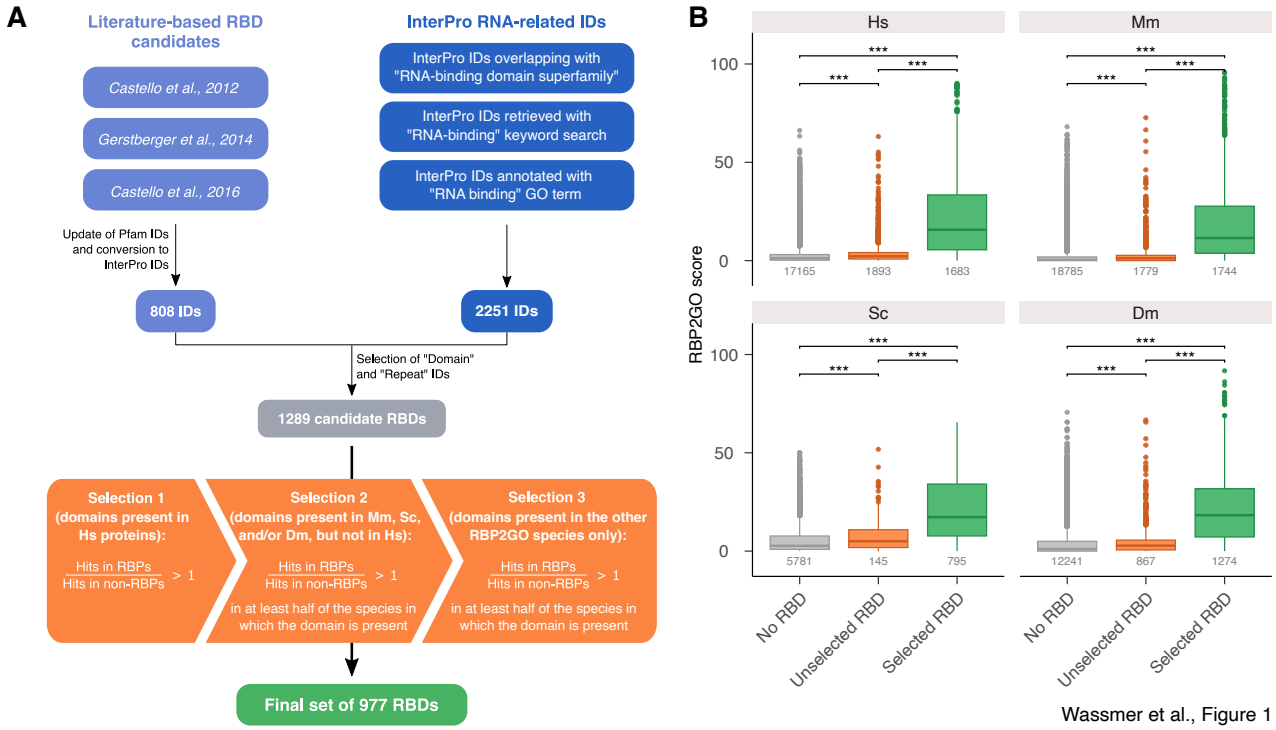
1132 **Figure 5.** Analysis of the disordered fraction in the human and murine proteins of the RBP2GO
1133 database. (**A**) Boxplots representing the distribution of RBP2GO score in eight groups of proteins: non-
1134 RBPs without RBD nor IDR (light green), non-RBPs without RBD but with IDR (light blue), non-RBPs
1135 with RBD but without IDR (dark green), non-RBPs with RBD and IDR (dark blue), RBPs without RBD
1136 nor IDR (light orange), RBPs without RBD but with IDR (light red), RBPs with RBD but without IDR
1137 (dark orange) and RBPs with RBD and IDR (dark red). (**B**) Boxplots representing the distribution of
1138 LCRs in non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light
1139 orange) and RBPs with RBD (dark orange). *, ** and *** correspond to *P*-values < 0.05, 0.01 and 0.001,
1140 respectively, ns corresponds to non-significant, as resulting from a Wilcoxon rank sum test.

1141 **Figure 6.** Discovery of candidates for RBP aspirants using the RBD content of the non-RBP proteins.
1142 (**A**) Number of RBPs and RBD-containing non-RBPs (numbers indicated in blue for this group) for
1143 species with a high (≥9, top bar) or a low (<9, bottom bar) number of RBP screen studies. (**B**) Heatmap
1144 showing the most enriched GO terms for molecular function in the RBD-containing non-RBPs as
1145 compared to the proteome of each species. The GO terms were classified by mean fold enrichment,
1146 from top to bottom. The terms are related among other molecular functions to metabolism (blue), to
1147 DNA and RNA (green), and to chromatin regulation (orange). (**C**) Same as in (**B**) for the most enriched
1148 GO terms for biological processes.

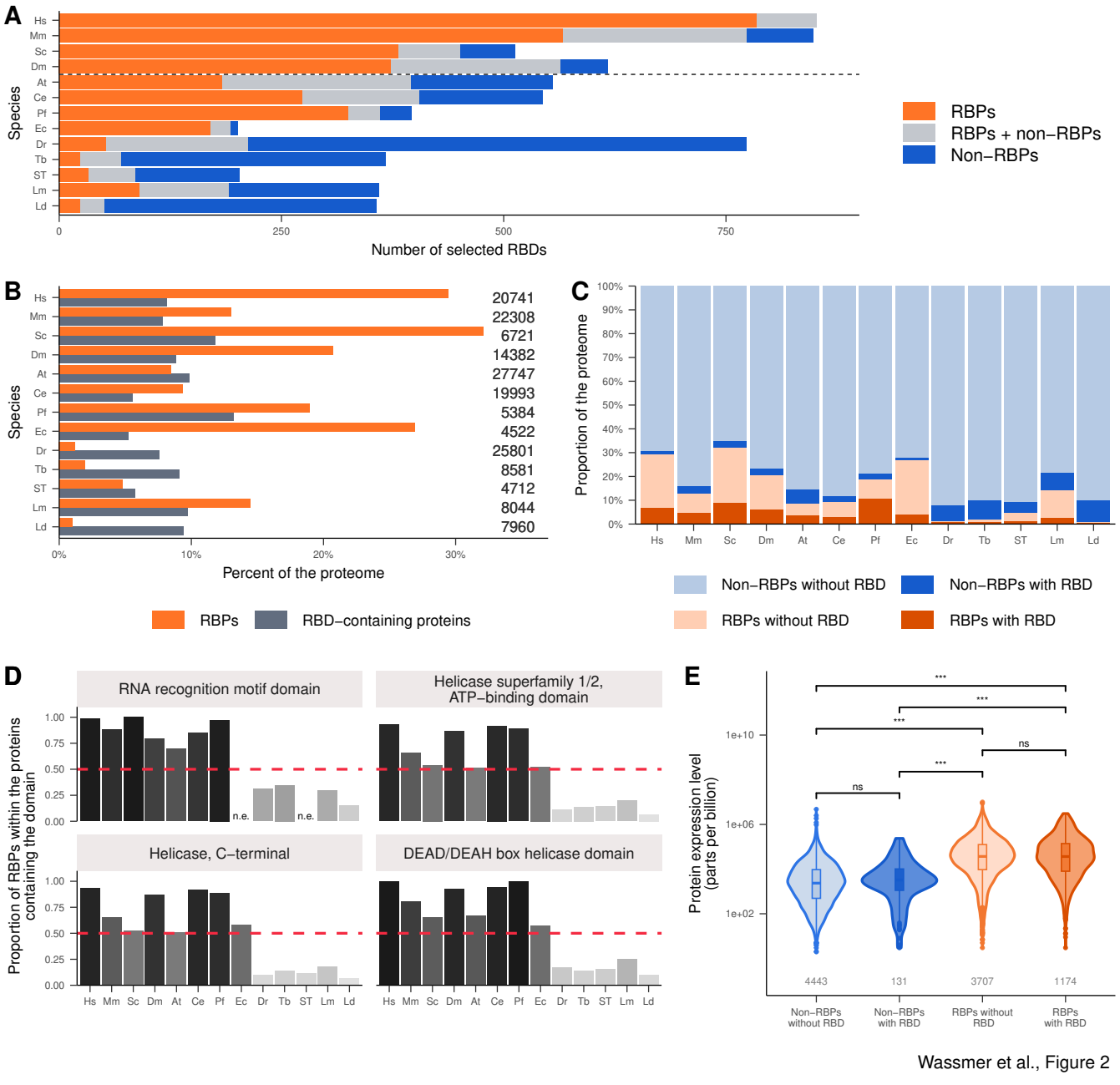
1149 **Figure 7.** Discovery of new RBD candidates in RBPs lacking RBDs. (**A**) Volcano plots showing the
1150 enrichment and its adjusted *P*-values for each of the InterPro domains present in the RBPs without
1151 RBD within the top 20% of RBP2GO scores, with an RBP/non-RBP ratio > 1, as compared to three
1152 different control groups (RBPs without RBD within the bottom 20% RBP2GO scores; non-RBPs without
1153 RBD; whole proteome). The *P*-values were calculated using a Fisher's exact test and were adjusted
1154 with the FDR method. (**B**) Enrichment and *P*-values for the 15 InterPro IDs selected as candidate RBDs,
1155 see also Supplementary Table S8. Some InterPro IDs are related to chaperones (orange), cytoskeletal-
1156 protein binding (green), or importins (blue). (**C**) Proportion of proteins without RBD (grey), with selected
1157 RBDs (dark green) or new RBD candidates (light green) in non-RBPs or RBPs. (**D**) Boxplots showing

1158 the distribution of the RBP2GO score for the proteins without RBD (grey), containing a new RBD
1159 candidate (light green), or a selected RBD (dark green). *** corresponds to *P*-values < 0.001 as resulting
1160 from a Wilcoxon rank sum test. The numbers given in grey represent the size of the protein groups. (**E**)
1161 Fold enrichment of RBPs over non-RBPs across all proteins containing one of the new RBD candidates.
1162 The *P*-values were calculated using a Fisher's exact test. (**F**) Boxplots showing the distribution of the
1163 log2 of the normalized mean proportion of the RNA-binding peptides overlapping the domain + $1x10^{-2}$.
1164 The domains were categorized as not selected RBDs (grey, see Figure 1), new RBD candidates (light
1165 green) and selected RBDs (dark green, see Figure 1). Only domain present in at least 5 proteins were
1166 taken into account (see Material and Methods section for more details). The number above each box
1167 represents the median values and the number under each box represents the size of the group. The *P*-
1168 values were calculated using a Wilcoxon rank sum test. (**G**) Schematic representations of three proteins
1169 (grey) with Armadillo repeats (blue) and the overlapping RNA-binding peptides (red).

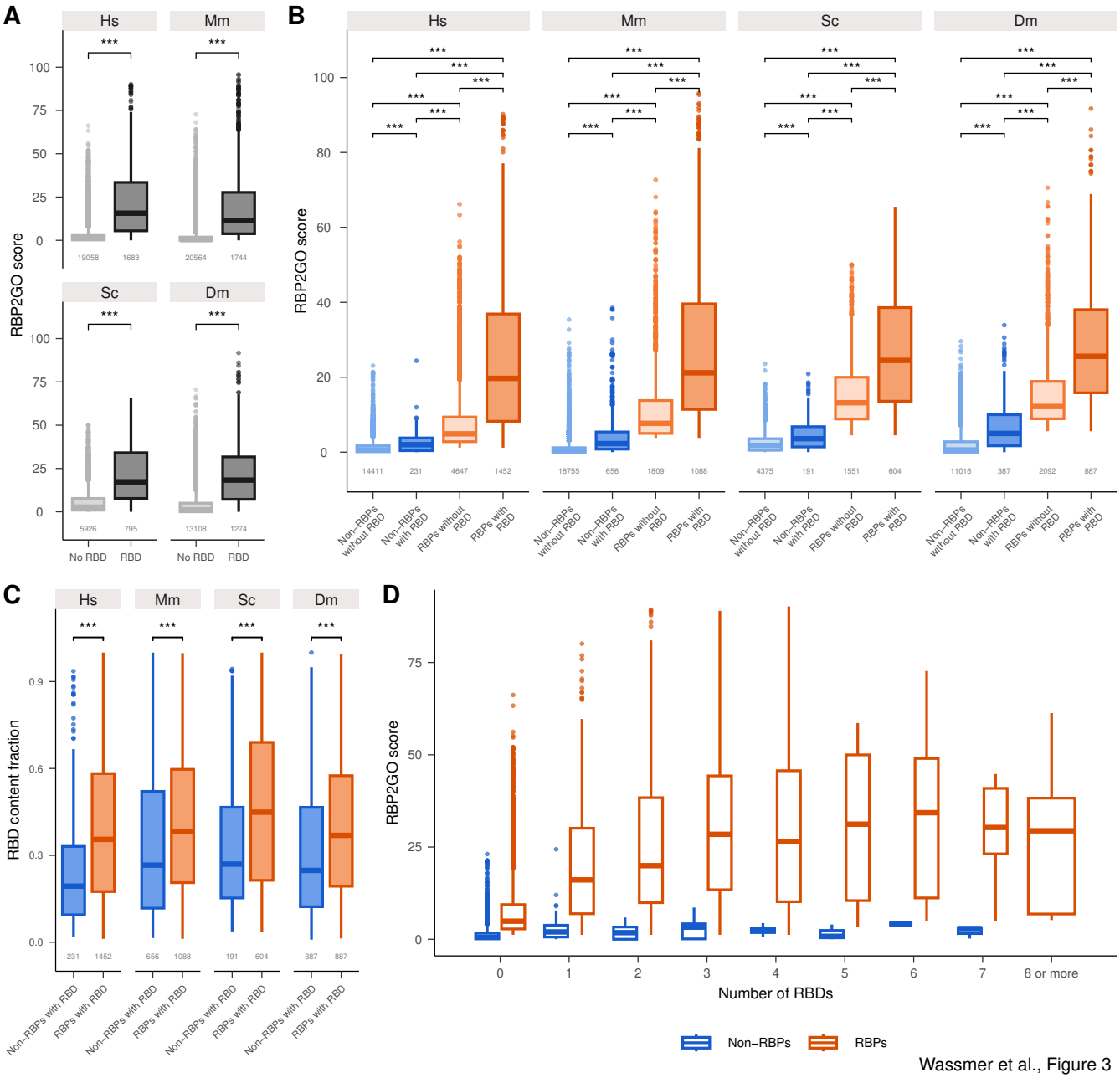
1170 **Figure 8.** Development of the RBP2GO composite score and update of the RBP2GO database. (**A**)
1171 Schematic representation of the compilation of the RBP2GO composite score. The quality factors as
1172 well as their attribution rules are explained in the Supplementary Table S10. The maximum values are
1173 50 for component 1 and 25 for each component 2 and 3. The RBP2GO composite score takes values
1174 between 0 and 100. (**B**) Violin plots representing the distribution of the RBP2GO score (orange) and
1175 the RBP2GO composite score (green) for non-RBPs without RBD nor Rfam ID, for non-RBPs with RBD
1176 and/or Rfam ID, for RBPs without RBD nor Rfam ID and for RBP with RBD and/or Rfam ID. (**C**)
1177 Screenshot of the sidebar of the advanced RBP2GO database (https://RBP2GO-2-Beta.dkfz.de)
1178 showing the added menu for domains and families in orange. (**D**) Screenshot of the additional search
1179 function dedicated to the protein domains (found under the SEARCH DOMAINS of the sidebar as shown
1180 in **C**). (**E**) Screenshot of the new "Domain information" tab added to the protein search section of each
1181 species in the RBP2GO database. This new tab displays all information pertaining to the RBDs and
1182 Rfam IDs of the selected protein, as well as information on its disorder content. It includes links to the
1183 InterPro (33) and MobiDB (40) databases. For each InterPro ID, a table indicates the type of the ID as
1184 "Domain", "Repeat" or "Family" and if it is linked to "RNA-binding".



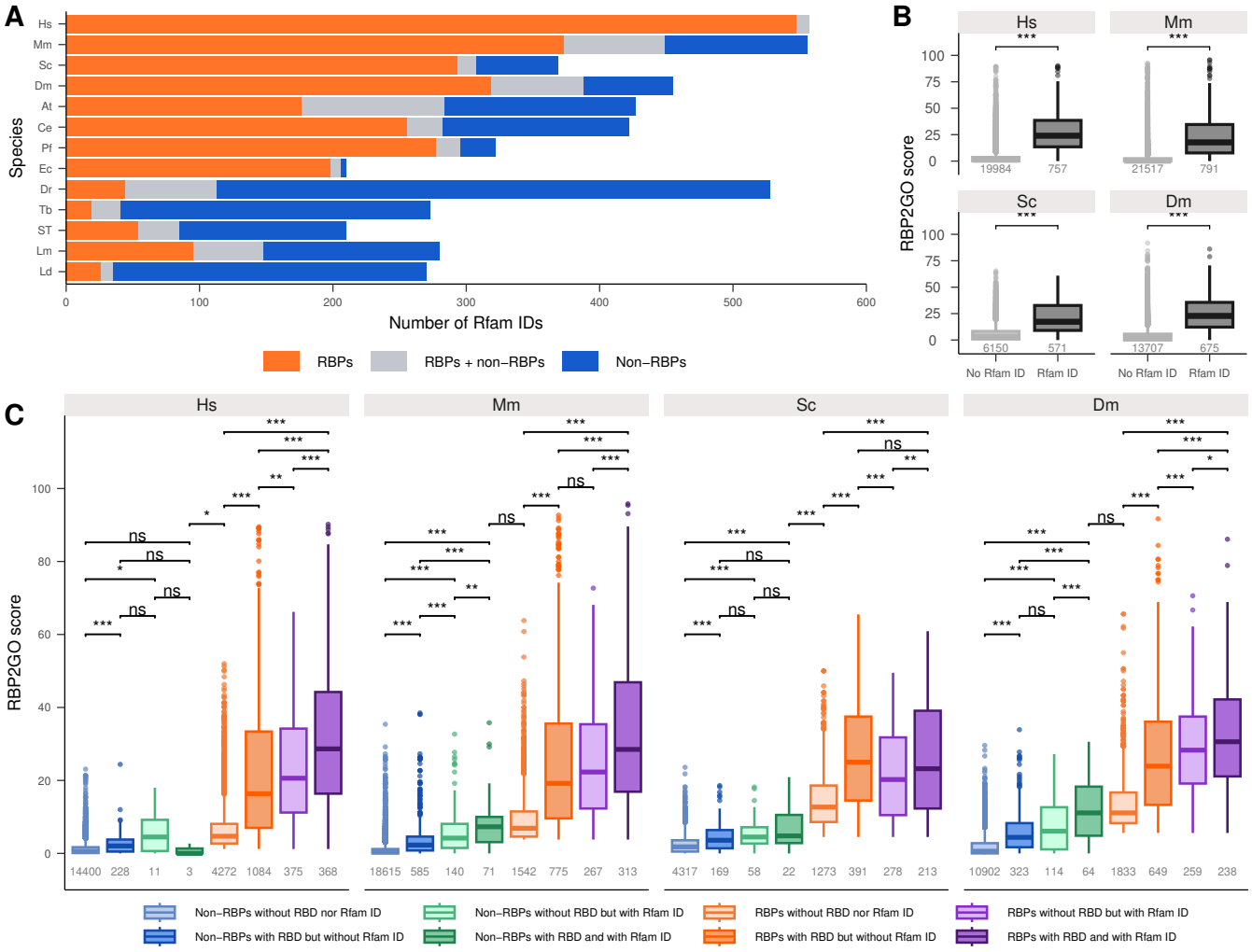
Wassmer et al., Figure 1



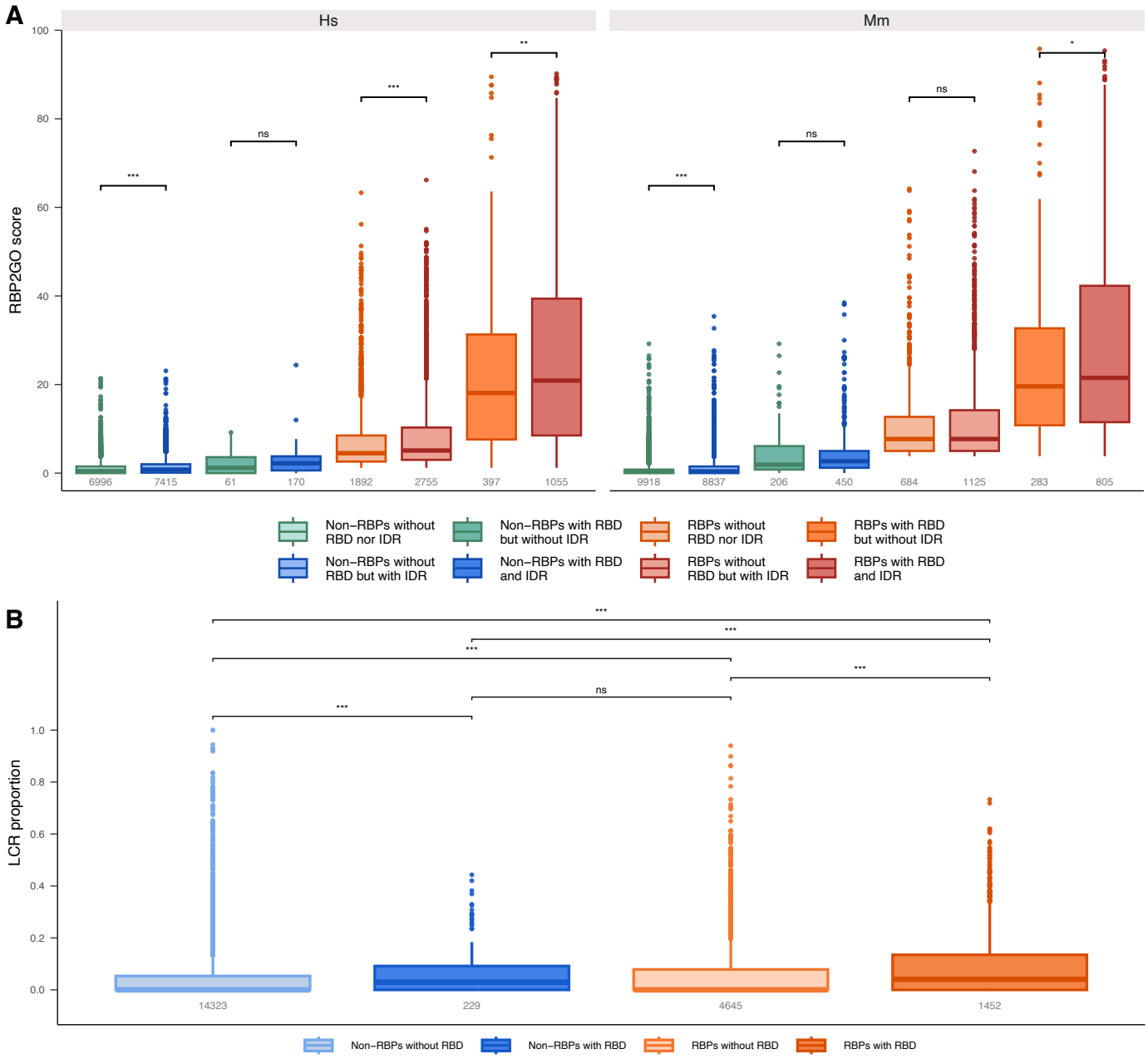
Wassmer et al., Figure 2



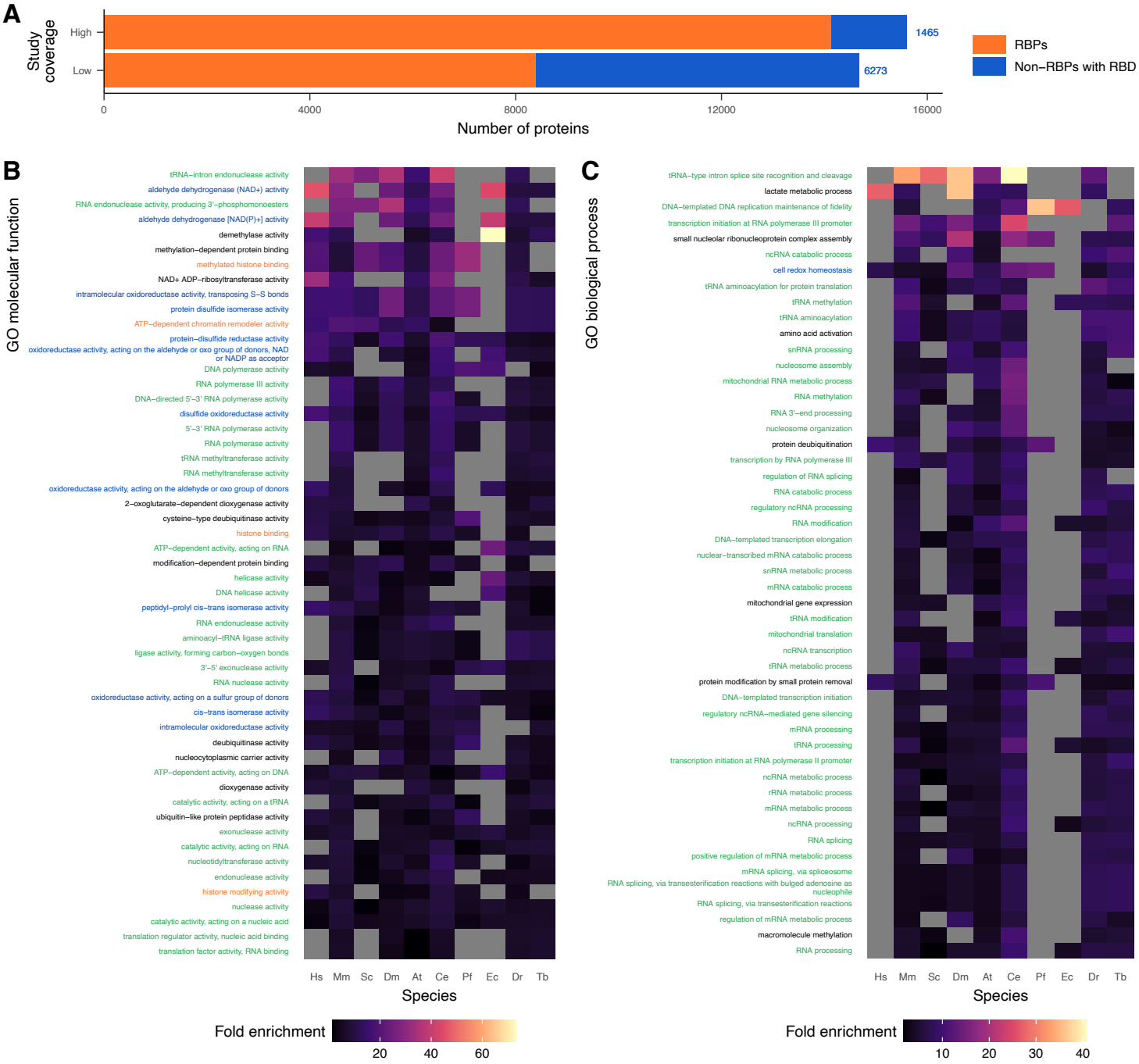
Wassmer et al., Figure 3



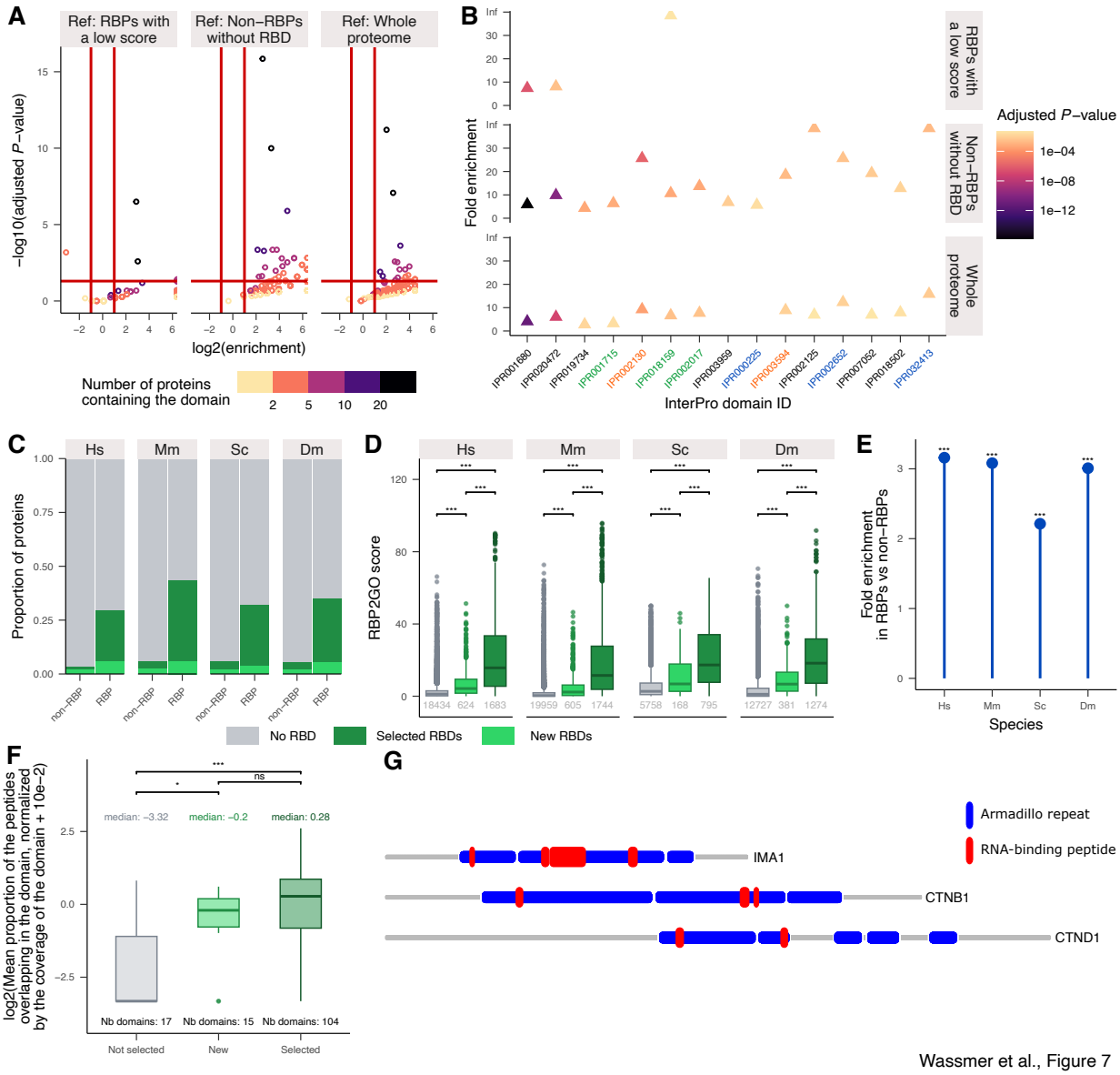
Wassmer et al., Figure 4



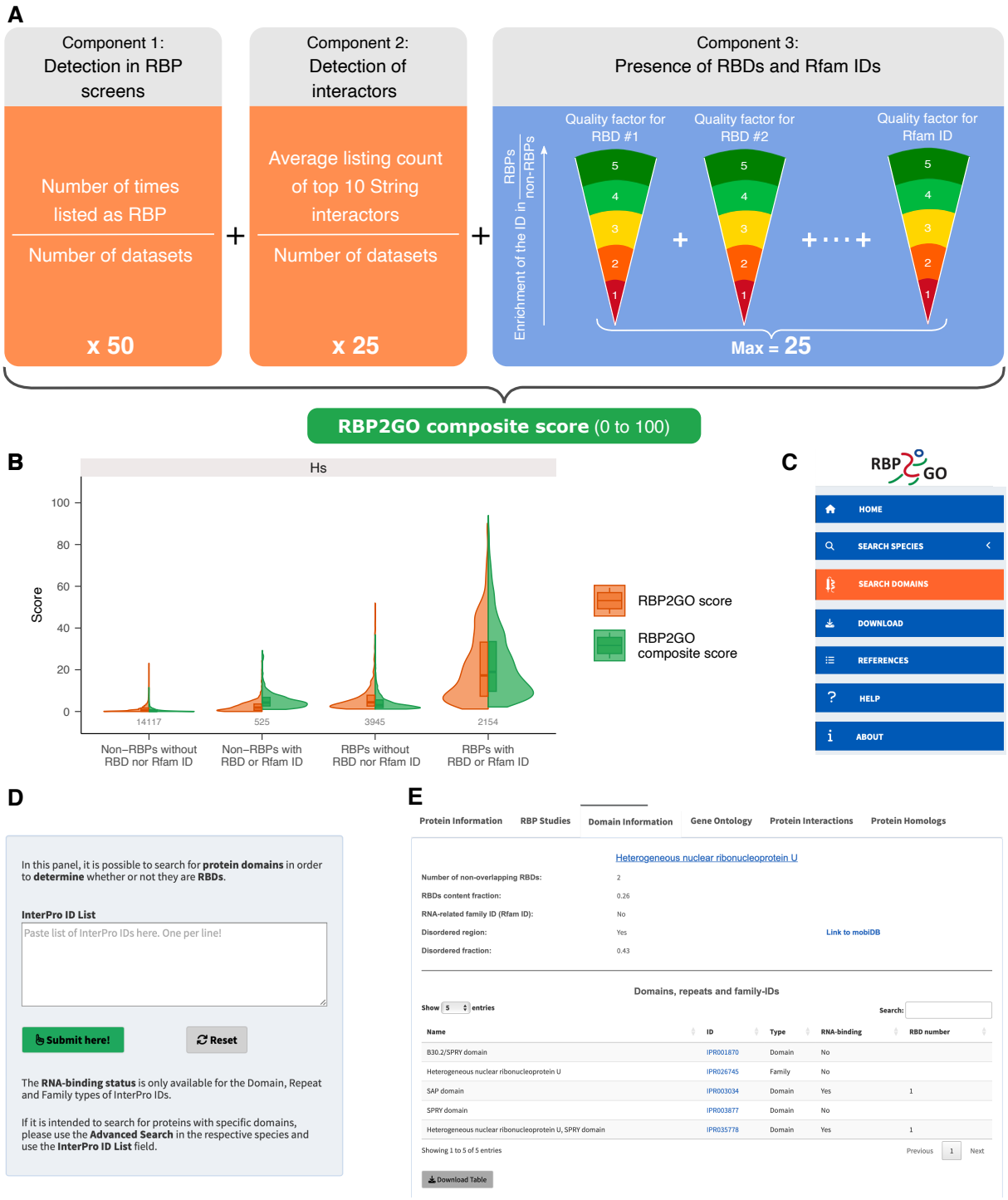
Wassmer et al., Figure 5



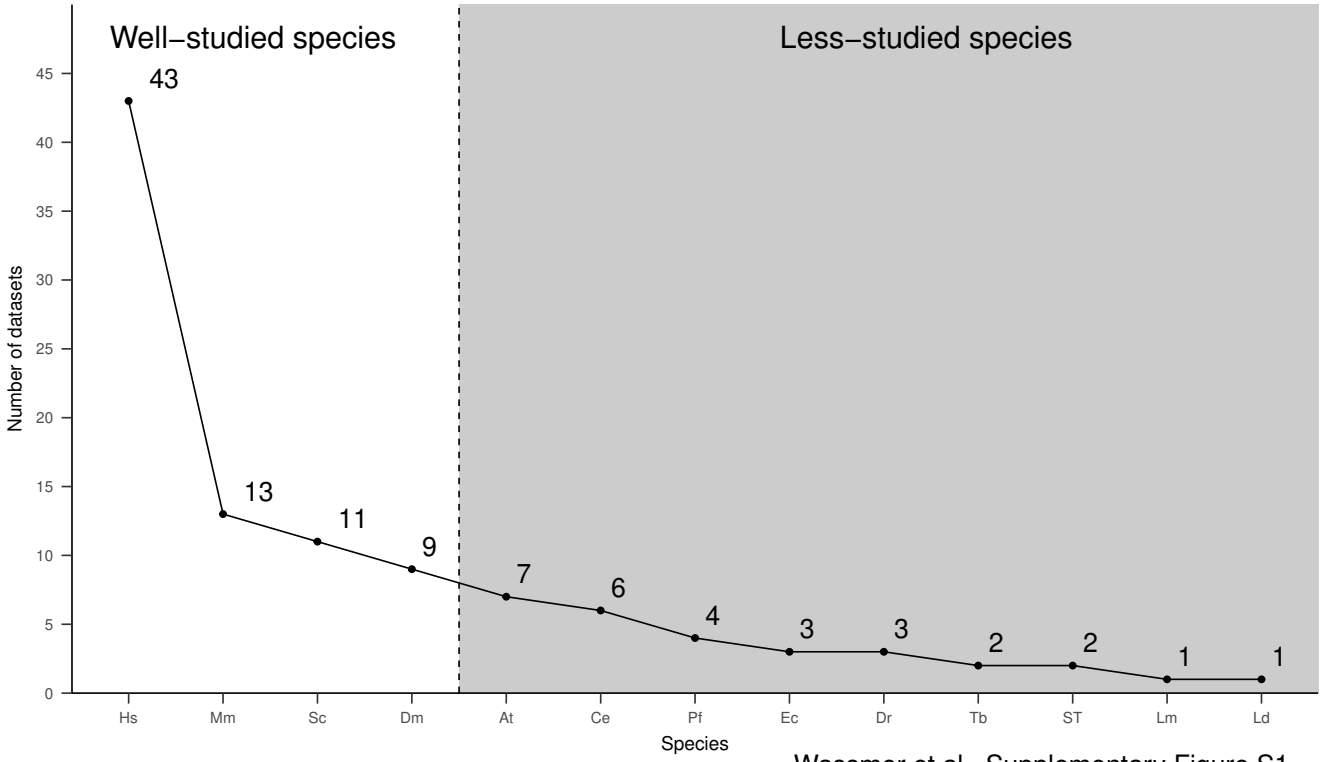
Wassmer et al., Figure 6



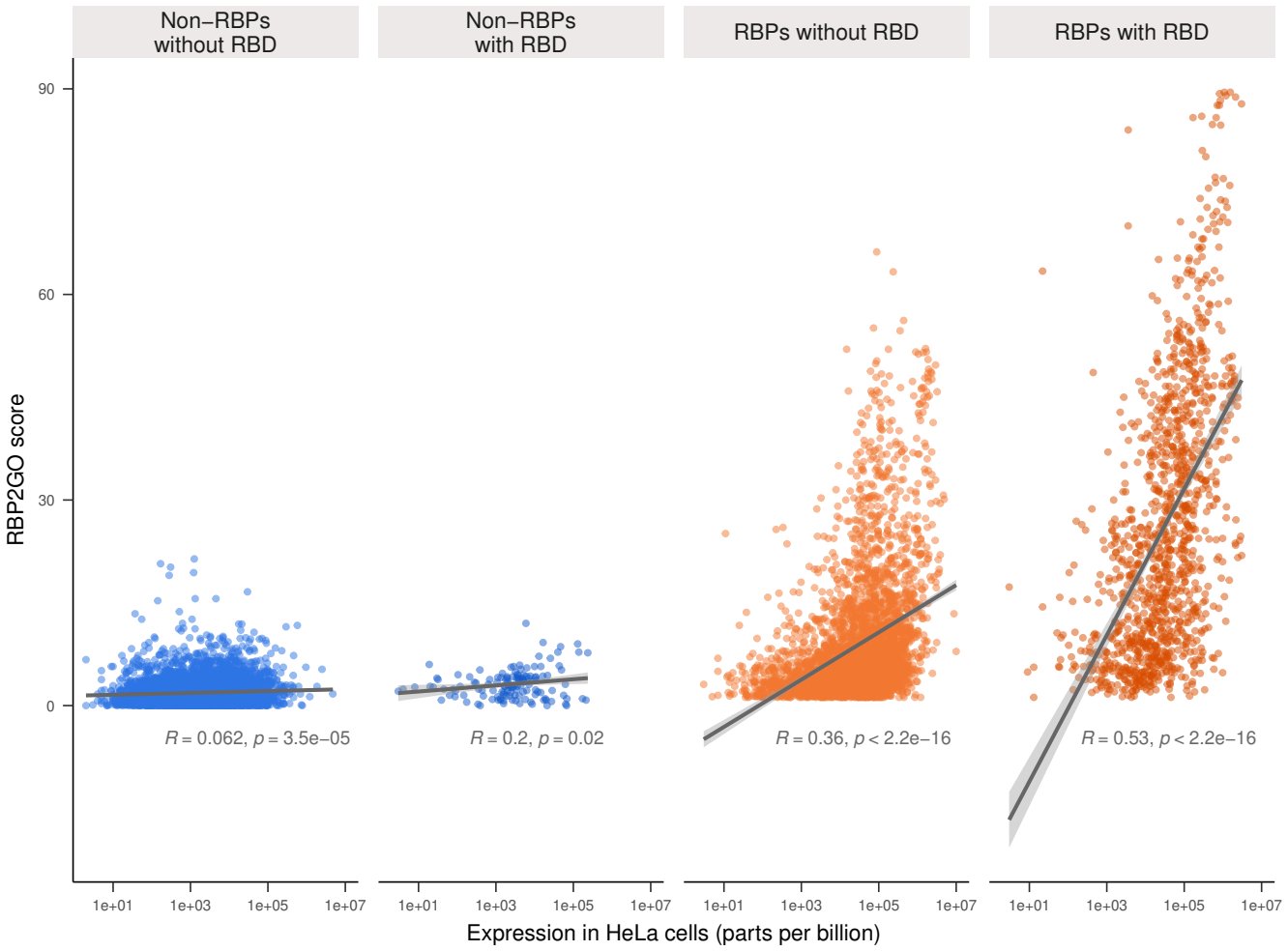
Wassmer et al., Figure 7



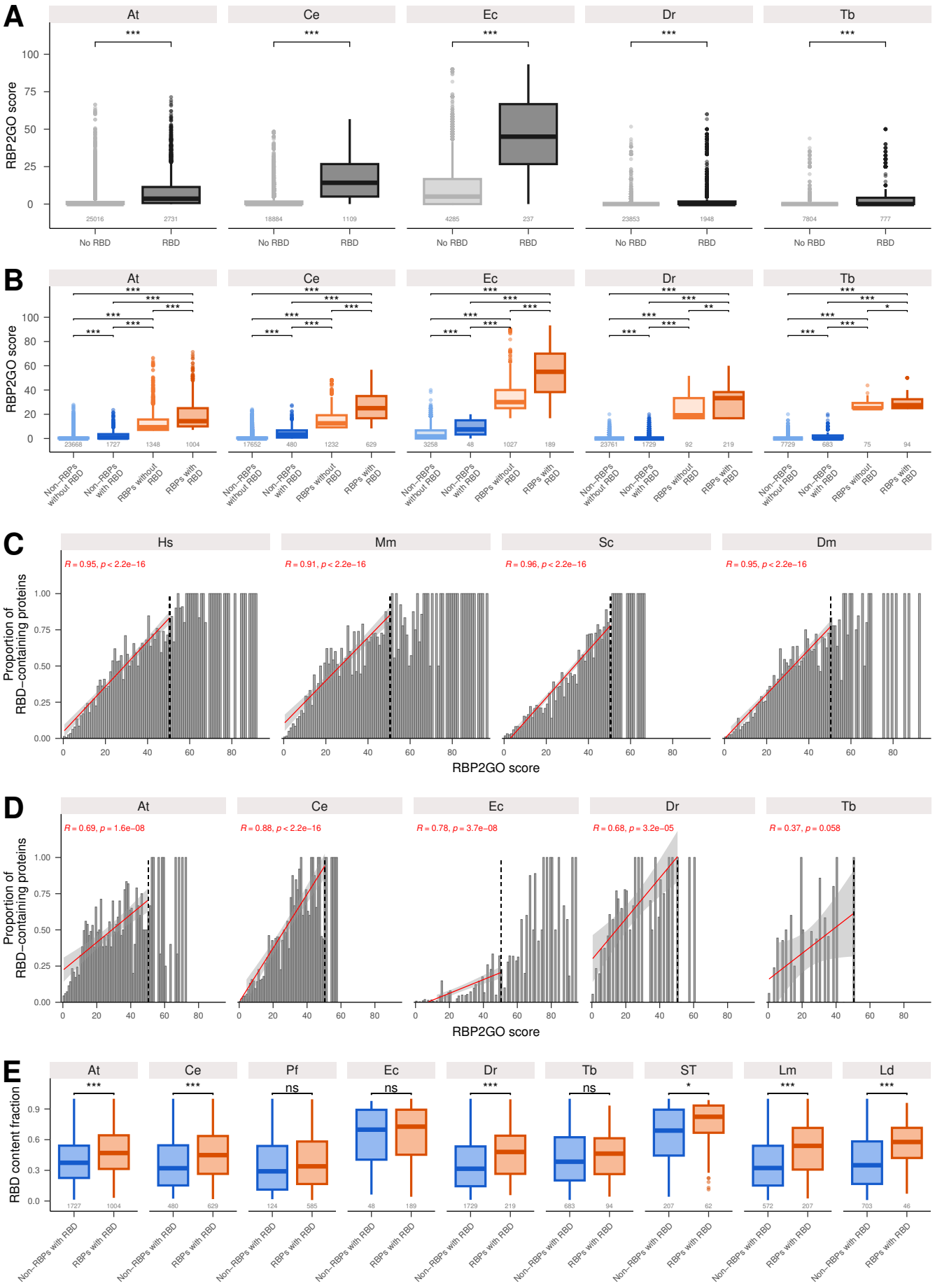
Wassmer et al., Figure 8



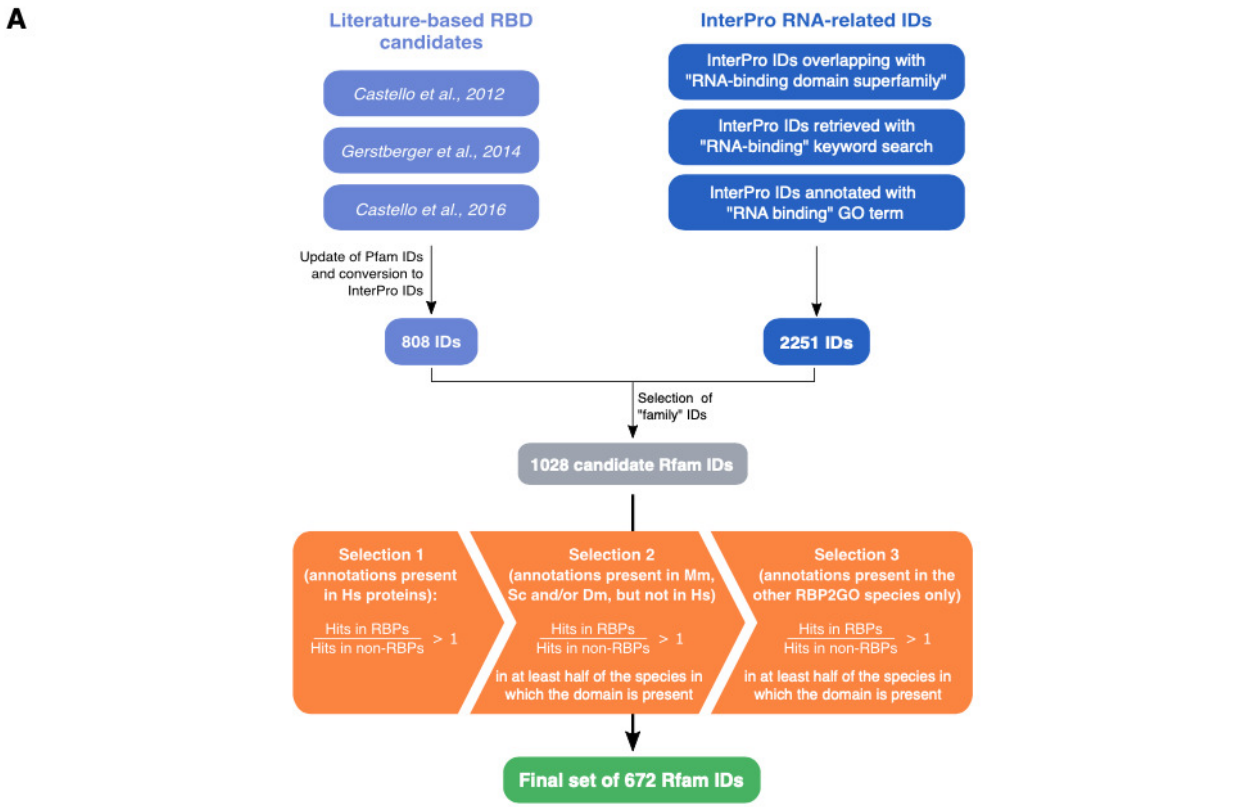
Wassmer et al., Supplementary Figure S1

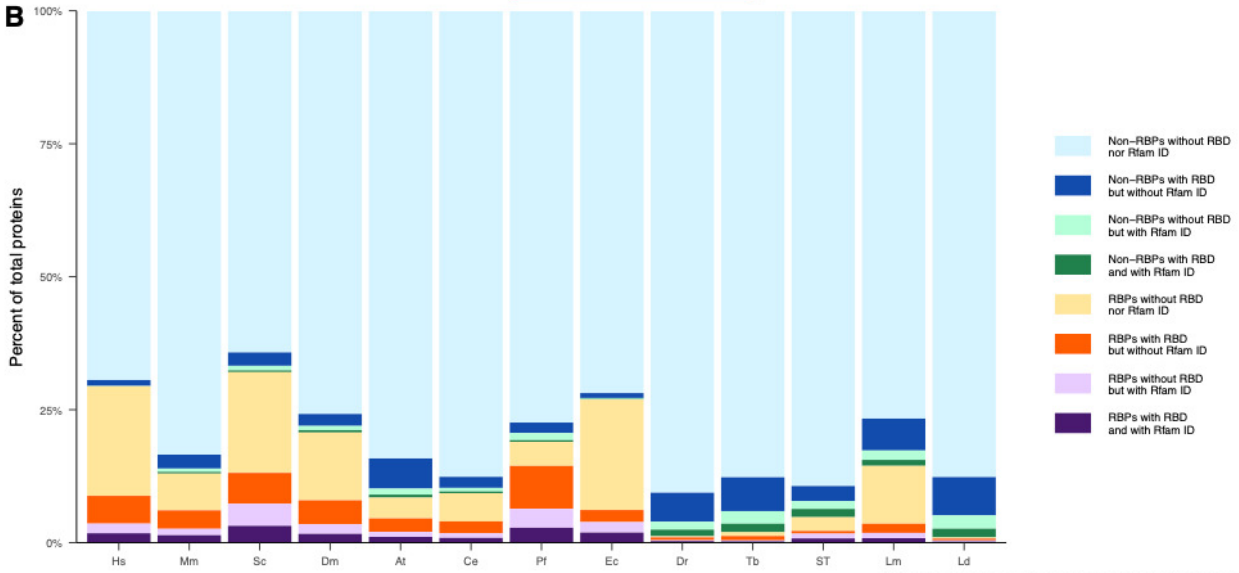


Wassmer et al., Supplementary Figure S2

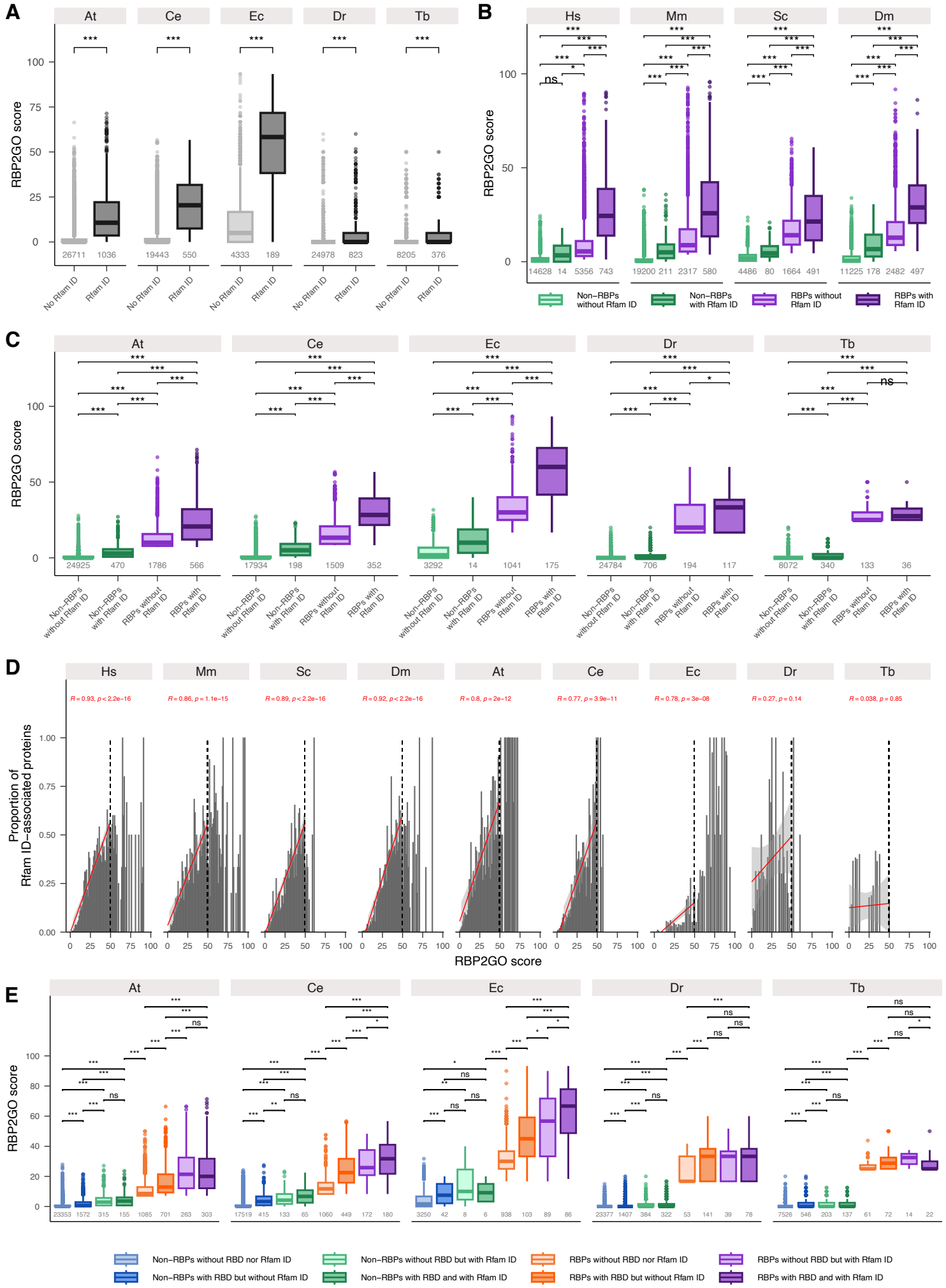


Wassmer et al., Supplementary Figure S3

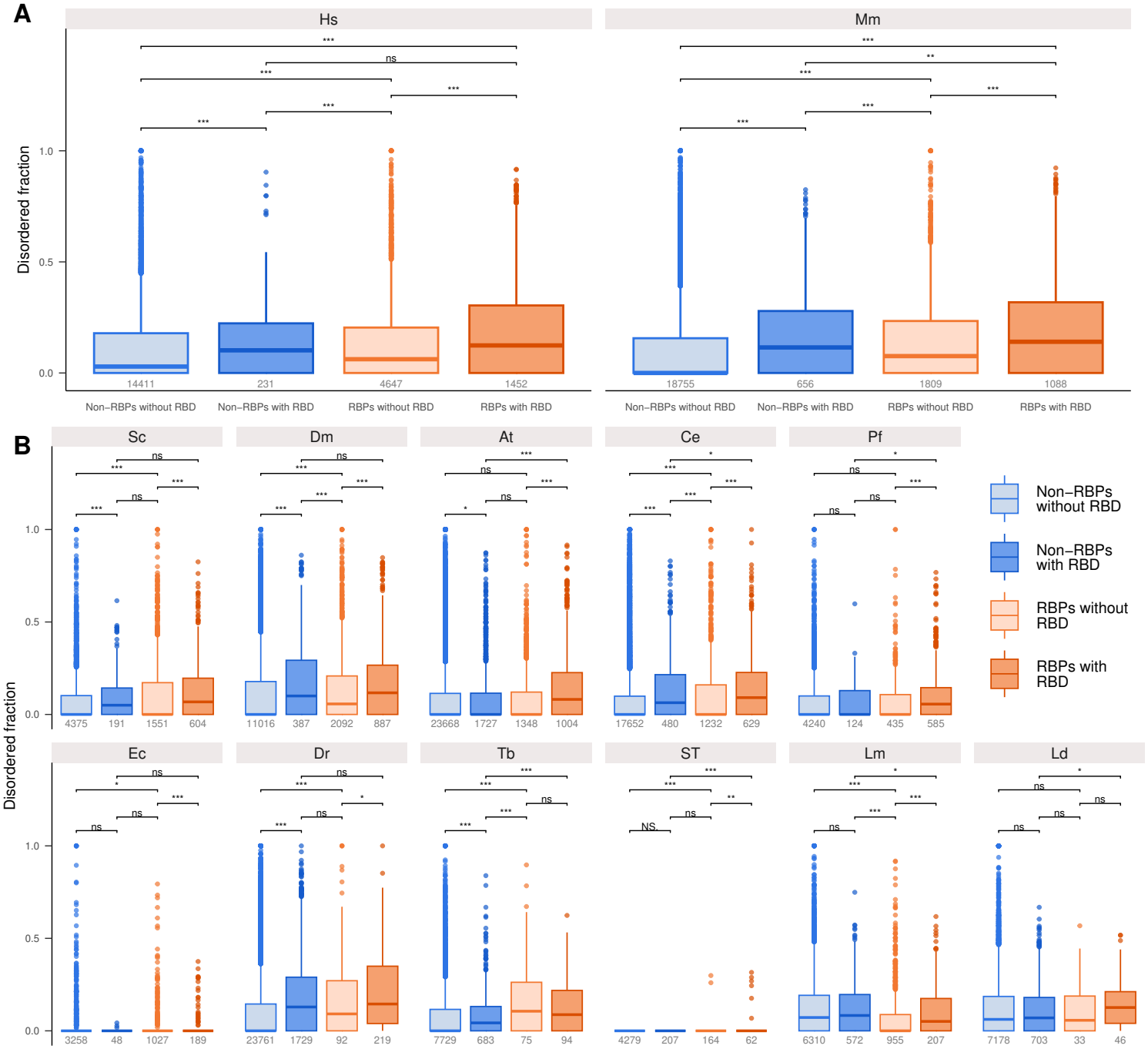




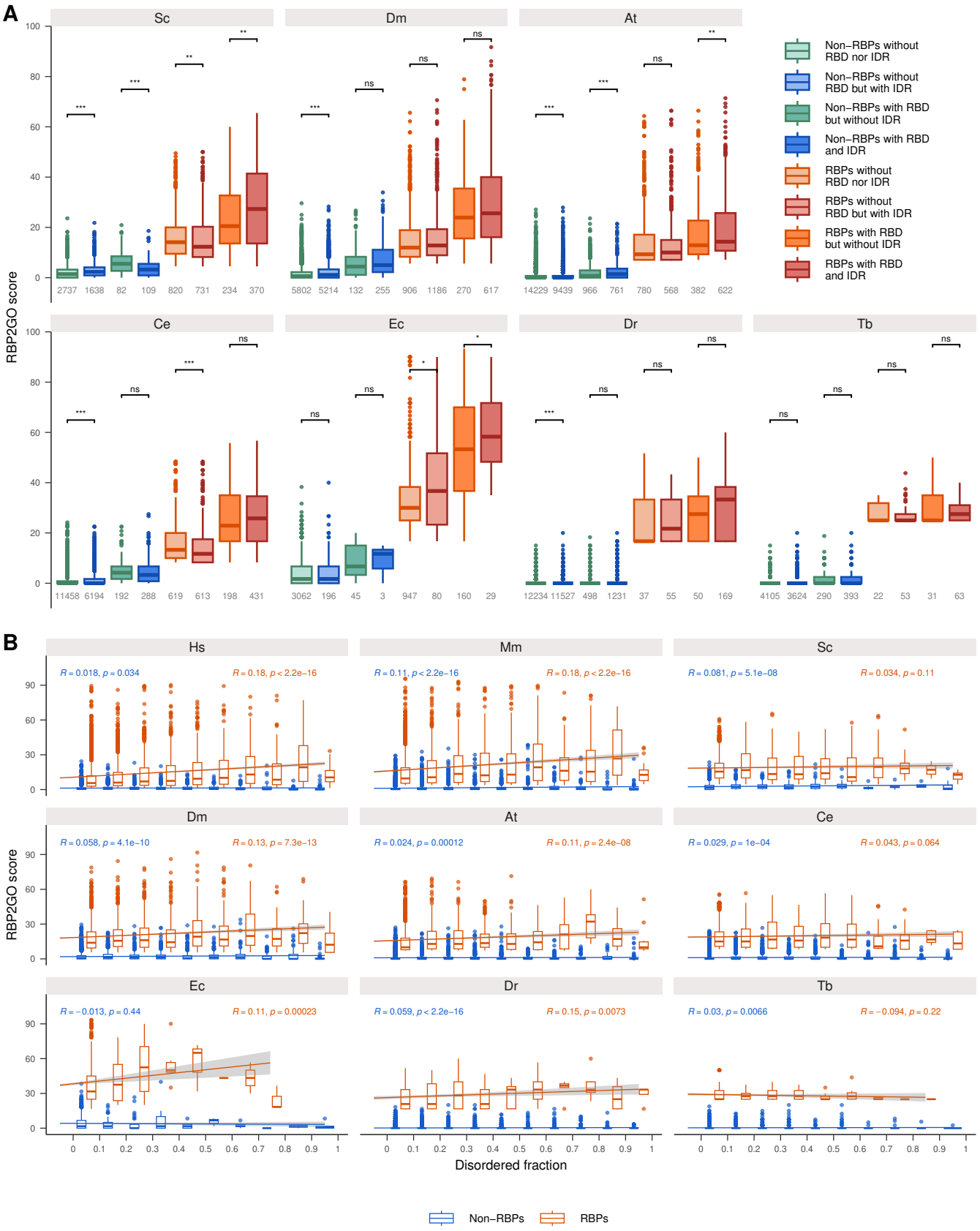
Wassmer et al., Supplementary Figure S4



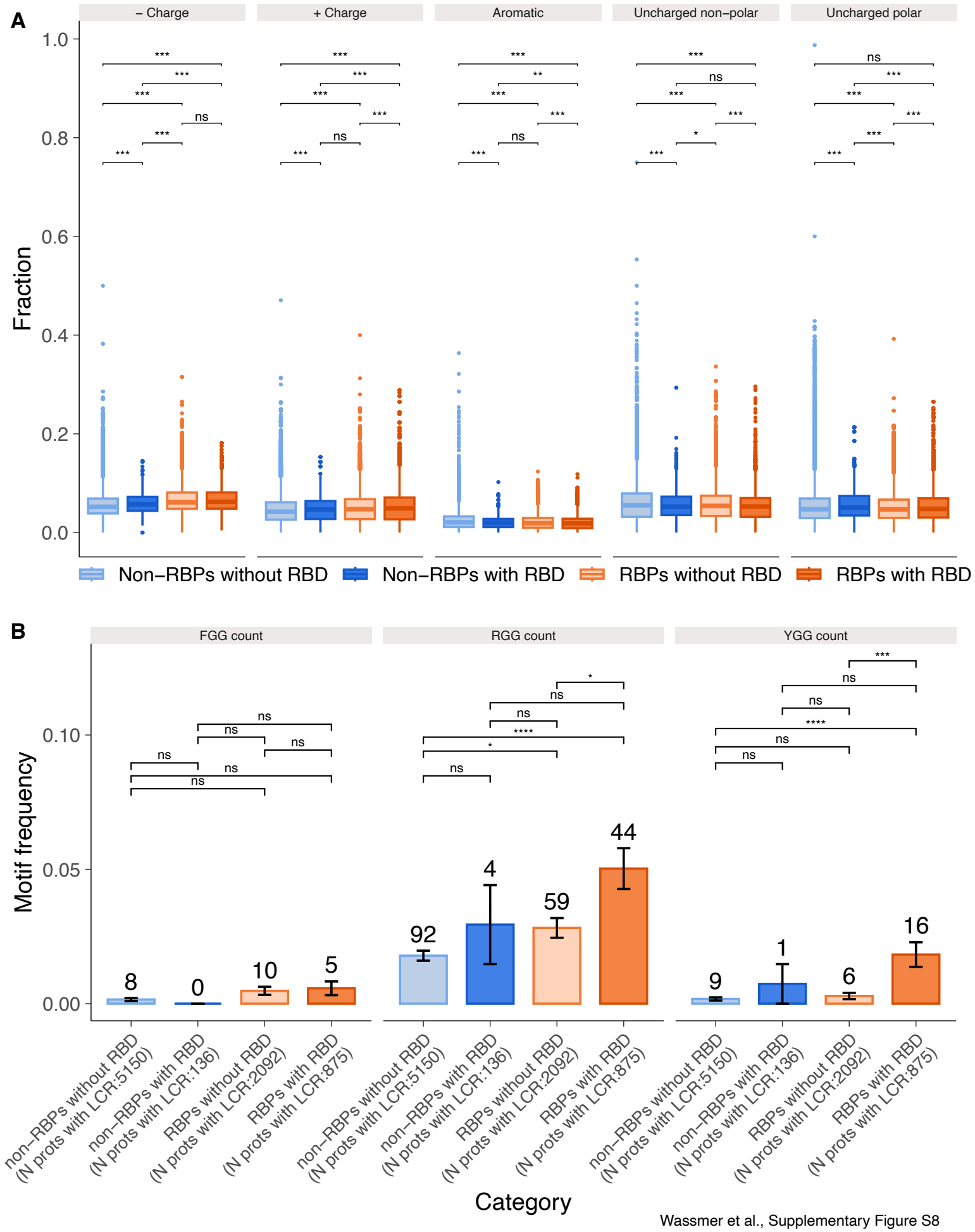
Wassmer et al., Supplementary Figure S5



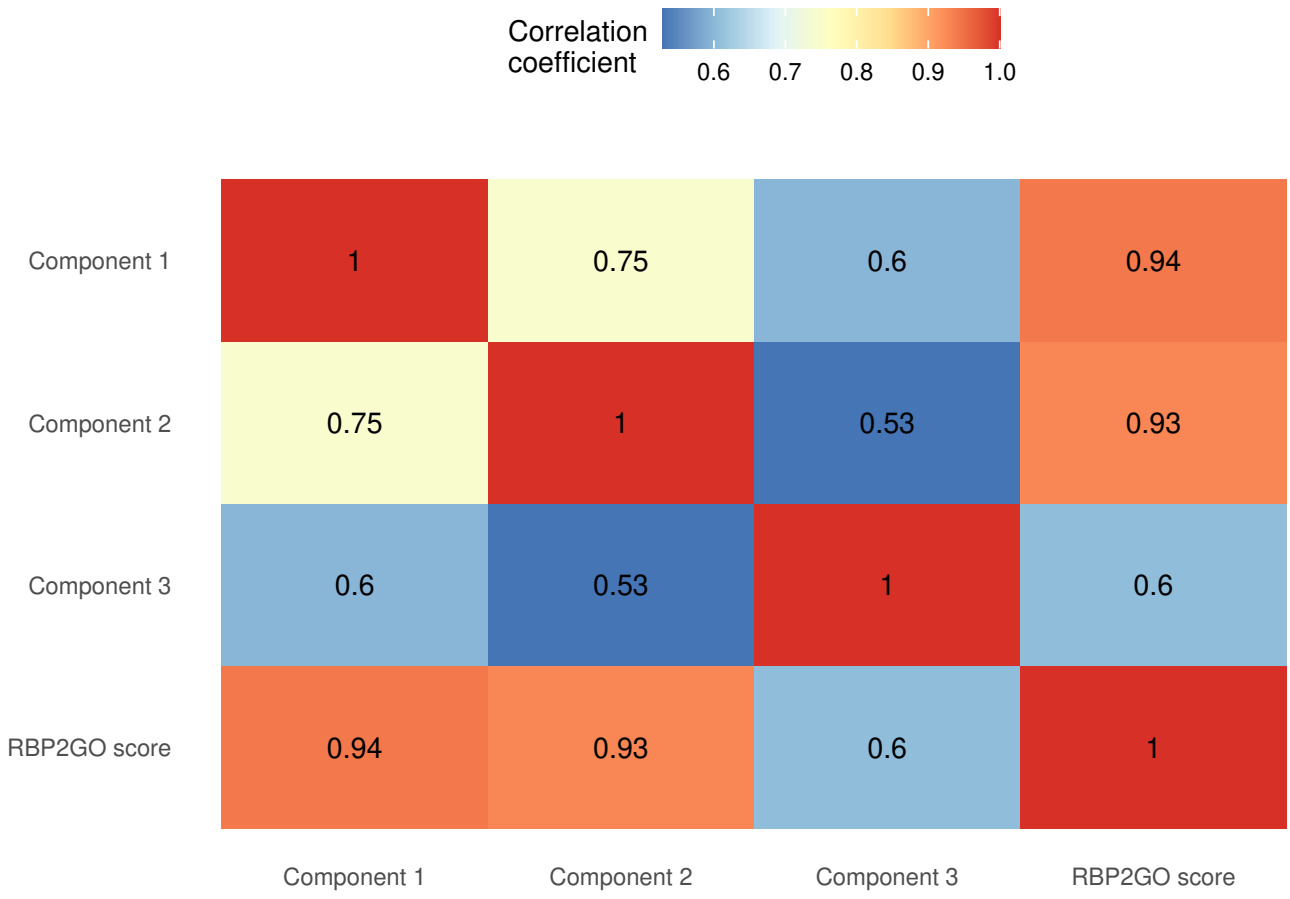
Wassmer et al., Supplementary Figure S6



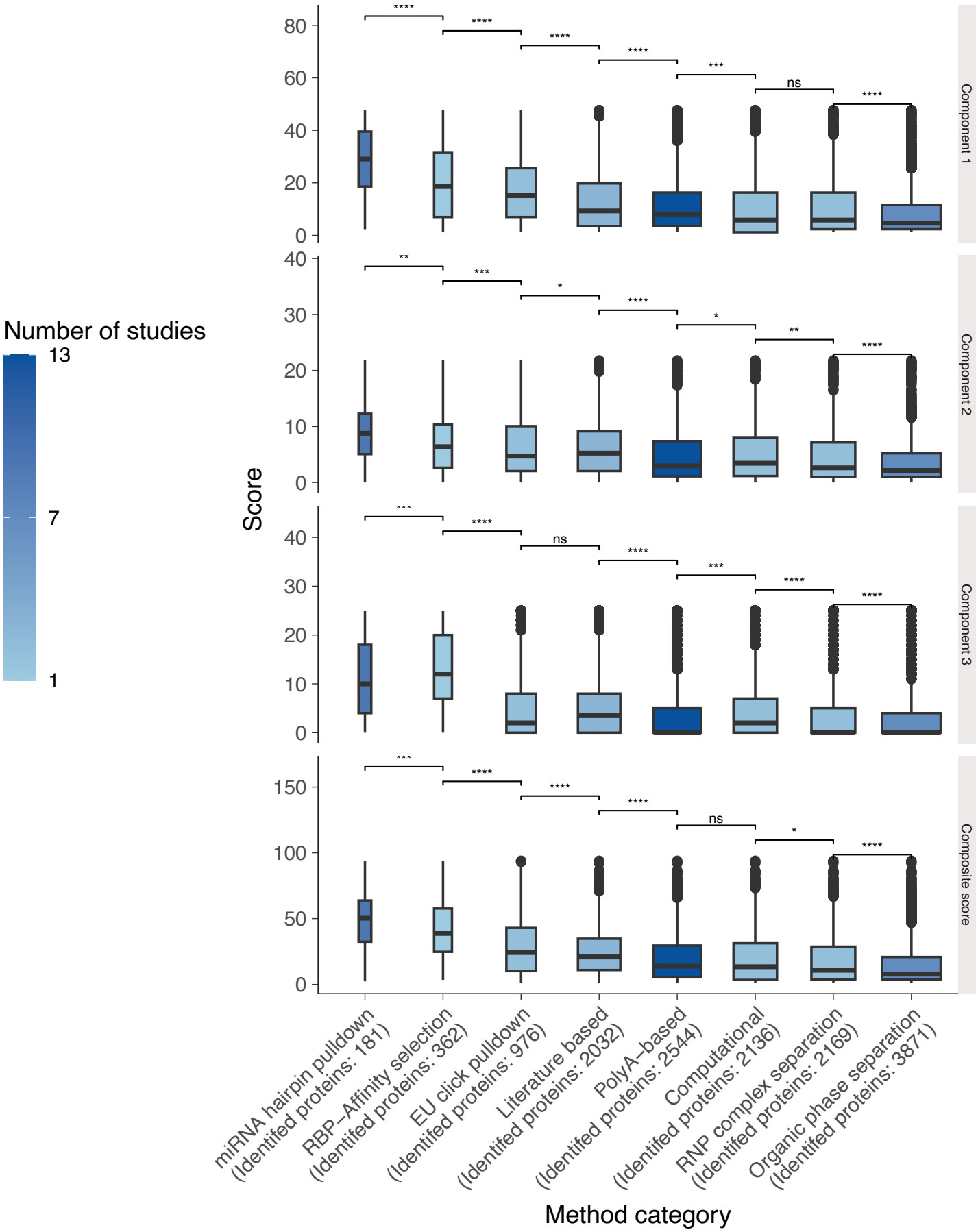
Wassmer et al., Supplementary Figure S7



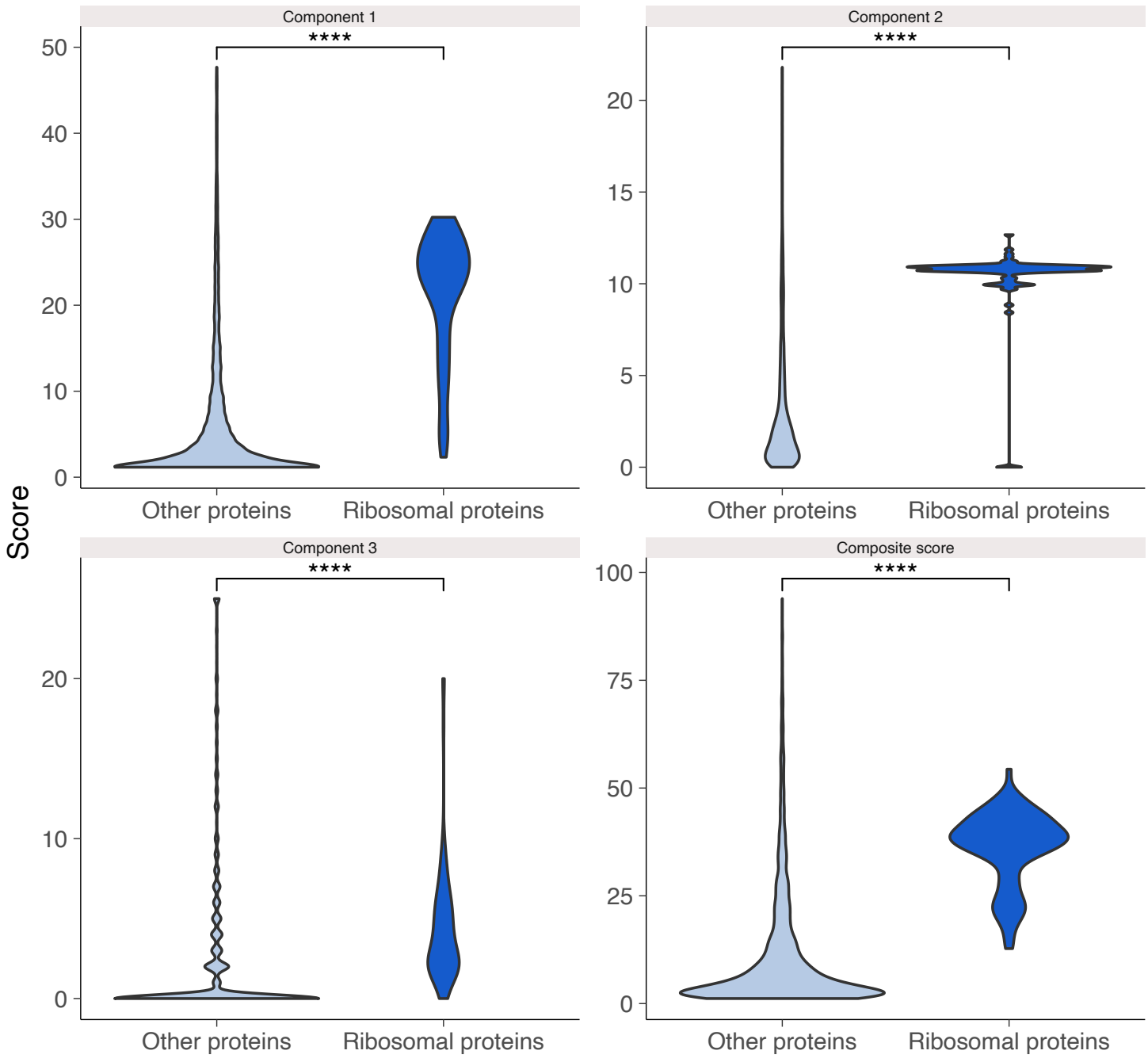
Wassmer et al., Supplementary Figure S8



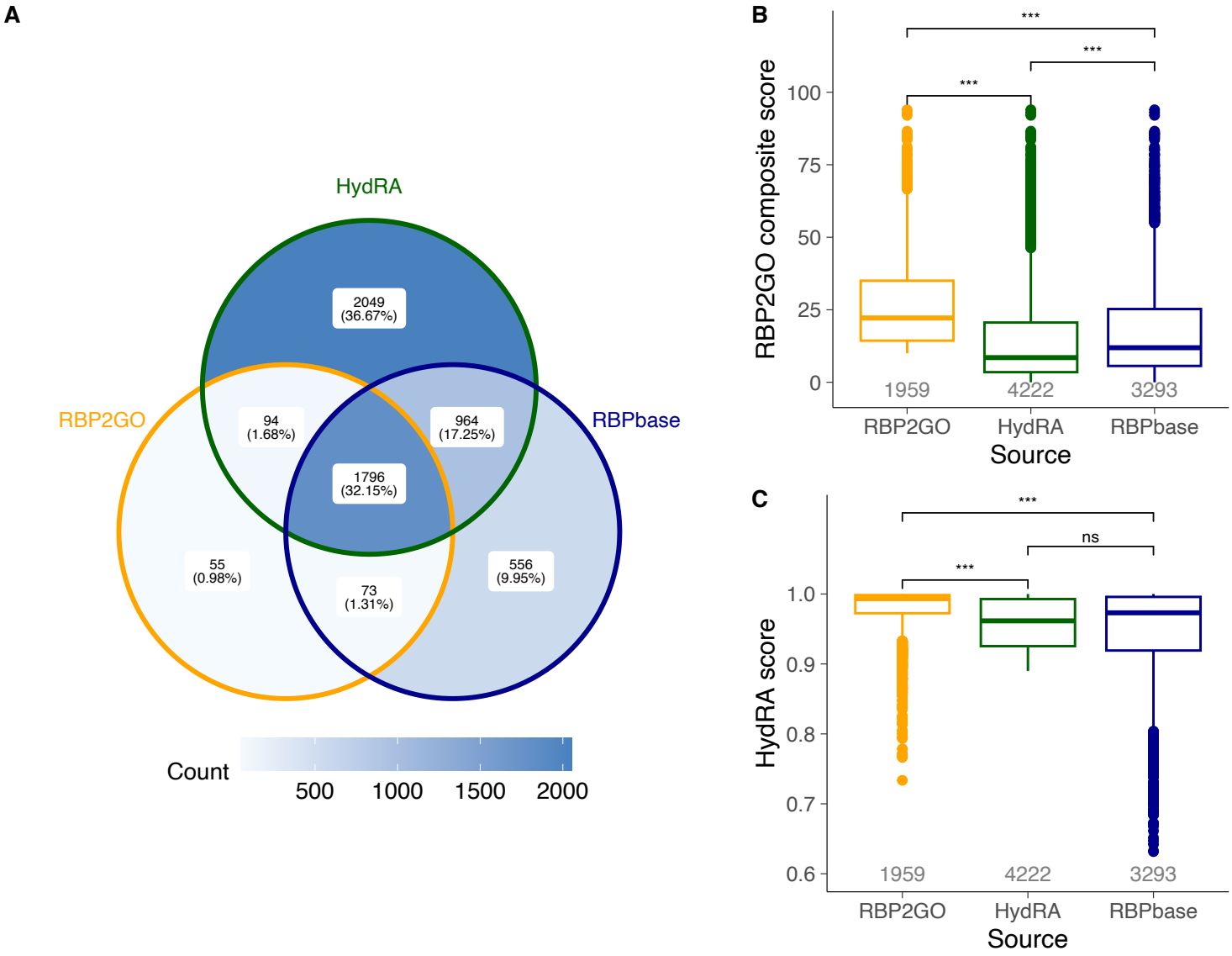
Wassmer et al., Supplementary Figure S9



Wassmer et al., Supplementary Figure S10



Wassmer et al., Supplementary Figure S11



Wassmer et al., Supplementary Figure S12